

Deep Explanations in Machine Learning via Interpretable Visual Methods

Boris Kovalerchuk¹
Muhammad Aurangzeb Ahmad^{2,3}
Ankur Teredesai^{2,3}

[1] Dept. of Computer Science, Central
Washington University

[2] Dept. of Computer Science and Systems,
University of Washington Tacoma

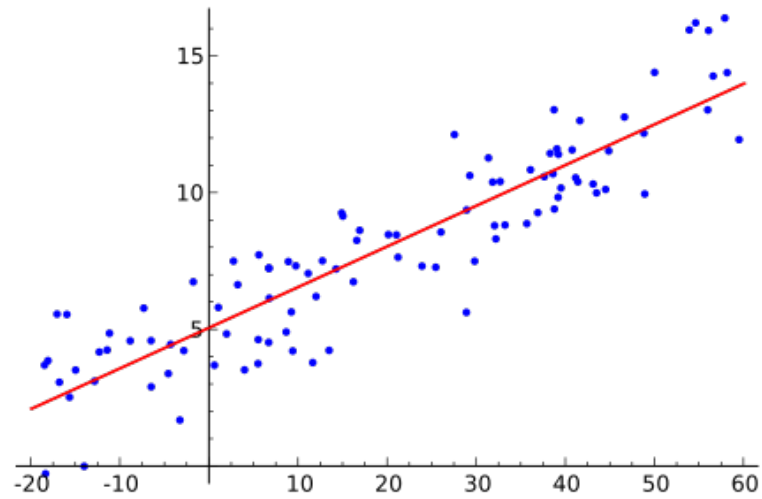
[3] KenSci Inc.

Outline

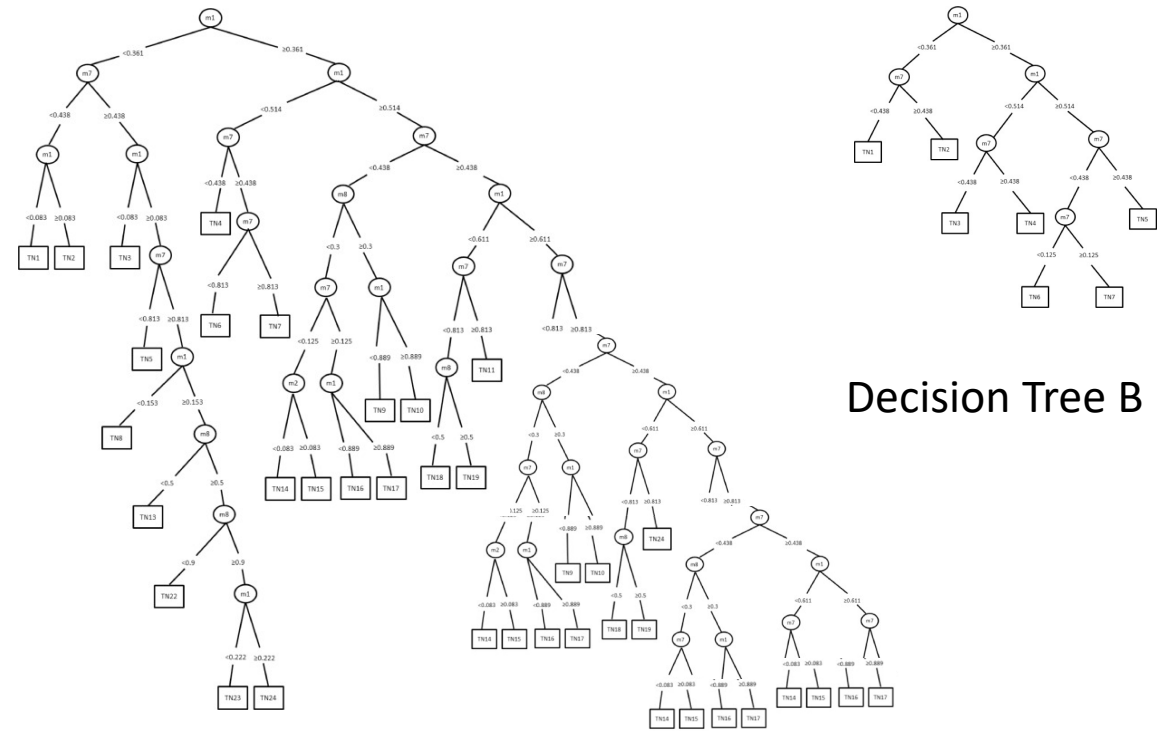
- I. Foundations of Interpretability
- II. Discovering Visual Interpretable Models
- III. Limits of Visual Interpretability in Deep Learning
- IV. User-centric Views of Interpretability of Visual Methods
- V. Open Problems and Current Research Frontiers

Foundations of Interpretability

What is Interpretability?



$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n$$



Decision Tree A

Decision Tree B

Types of Machine Learning Models



There is an explanation about how the model is making predictions Examples: Decision Trees, Regression Models etc.



There is no explanation with respect to how the model is making predictions Examples: SVMs, Random Forest, Gradient Descent Models etc.

Explainable AI / Interpretable Machine Learning

- Explainable AI or interpretable machine learning: Giving **explanations** of AI/machine learning models to **humans** with **domain knowledge**
- Explanation: Why is the prediction being made?
- Explanation to Human: The explanation should be comprehensible to humans in (i) natural language (ii) easy to understand representations
- Domain Knowledge: The explanation should make sense to a domain expert

[Craik 1967, Doshi-Velez 2014]

$$\begin{aligned}
 \mathcal{L}_{SM} = & -\frac{1}{2}\partial_\nu g_\mu^a \partial_\nu g_\mu^a - g_{s_j}^{abc} \partial_\mu g_\nu^a g_\mu^b g_\nu^c - \frac{1}{4}g_s^2 f^{abc} f^{ade} g_\mu^b g_\nu^c g_\mu^d g_\nu^e - \partial_\nu W_\mu^+ \partial_\nu W_\mu^- \\
 & - M^2 W_\mu^+ W_\mu^- - \frac{1}{2}\partial_\nu Z_\mu^0 \partial_\nu Z_\mu^0 - \frac{1}{2c_w^2} M^2 Z_\mu^0 Z_\mu^0 - \frac{1}{2}\partial_\mu A_\nu \partial_\mu A_\nu - ig_{cw} (\partial_\nu Z_\mu^0 (W_\mu^+ W_\nu^- - W_\nu^+ W_\mu^-) - \\
 & Z_\nu^0 (W_\mu^+ \partial_\mu W_\nu^- - W_\mu^- \partial_\mu W_\nu^+) + Z_\mu^0 (W_\nu^+ \partial_\nu W_\mu^- - W_\nu^- \partial_\nu W_\mu^+)) - ig_{sw} (\partial_\nu A_\mu (W_\mu^+ W_\nu^- - \\
 & W_\nu^+ W_\mu^-) - A_\nu (W_\mu^+ \partial_\mu W_\nu^- - W_\mu^- \partial_\mu W_\nu^+) + A_\mu (W_\nu^+ \partial_\nu W_\mu^- - W_\nu^- \partial_\nu W_\mu^+)) - \\
 & \frac{1}{2}g^2 W_\mu^+ W_\nu^+ W_\mu^- W_\nu^- + \frac{1}{2}g^2 W_\mu^+ W_\nu^- W_\mu^+ W_\nu^- + g^2 c_w^2 (Z_\mu^0 W_\nu^+ Z_\nu^0 W_\mu^- - Z_\nu^0 Z_\mu^0 W_\nu^+ W_\mu^-) + \\
 & g^2 s_w^2 (A_\mu W_\nu^+ A_\nu W_\mu^- - A_\mu A_\nu W_\nu^+ W_\mu^-) + g^2 s_w c_w (A_\mu Z_\nu^0 (W_\mu^+ W_\nu^- - W_\nu^+ W_\mu^-) - \\
 & 2A_\mu Z_\mu^0 W_\nu^+ W_\nu^-) - \frac{1}{2}\partial_\mu H \partial_\mu H - 2M^2 \alpha_h H^2 - \partial_\mu \phi^+ \partial_\mu \phi^- - \frac{1}{2}\partial_\mu \phi^0 \partial_\mu \phi^0 - \\
 & \beta_h \left(\frac{2M^2}{g^2} + \frac{2M}{g} H + \frac{1}{2}(H^2 + \phi^0 \phi^0 + 2\phi^+ \phi^-) \right) + \frac{2M^4}{g^2} \alpha_h - g\alpha_h M (H^3 + H\phi^0 \phi^0 + 2H\phi^+ \phi^-) - \\
 & \frac{1}{8}g^2 \alpha_h (H^4 + (\phi^0)^4 + 4(\phi^+ \phi^-)^2 + 4(\phi^0)^2 \phi^+ \phi^- + 4H^2 \phi^+ \phi^- + 2(\phi^0)^2 H^2) - gM W_\mu^+ W_\mu^- H - \\
 & \frac{1}{2}g \frac{M}{c_w} Z_\mu^0 Z_\mu^0 H - \frac{1}{2}ig (W_\mu^+ (\phi^0 \partial_\mu \phi^- - \phi^- \partial_\mu \phi^0) - W_\mu^- (\phi^0 \partial_\mu \phi^+ - \phi^+ \partial_\mu \phi^0)) + \\
 & \frac{1}{2}g (W_\mu^+ (H \partial_\mu \phi^- - \phi^- \partial_\mu H) + W_\mu^- (H \partial_\mu \phi^+ - \phi^+ \partial_\mu H)) + \frac{1}{2}g \frac{1}{c_w} (Z_\mu^0 (H \partial_\mu \phi^0 - \phi^0 \partial_\mu H) + \\
 & M (\frac{1}{c_w} Z_\mu^0 \partial_\mu \phi^0 + W_\mu^+ \partial_\mu \phi^- + W_\mu^- \partial_\mu \phi^+) - ig \frac{g_w}{c_w} M Z_\mu^0 (W_\mu^+ \phi^- - W_\mu^- \phi^+) + ig s_w M A_\mu (W_\mu^+ \phi^- - \\
 & W_\mu^- \phi^+) - ig \frac{1-2c_w^2}{2c_w} Z_\mu^0 (\phi^+ \partial_\mu \phi^- - \phi^- \partial_\mu \phi^+) + ig s_w A_\mu (\phi^+ \partial_\mu \phi^- - \phi^- \partial_\mu \phi^+) - \\
 & \frac{1}{2}g^2 W_\mu^+ W_\mu^- (H^2 + (\phi^0)^2 + 2\phi^+ \phi^-) - \frac{1}{2}g^2 \frac{1}{c_w} Z_\mu^0 (H^2 + (\phi^0)^2 + 2(2s_w^2 - 1)2\phi^+ \phi^-) - \\
 & \frac{1}{2}g^2 \frac{s_w^2}{c_w} Z_\mu^0 \phi^0 (W_\mu^+ \phi^- + W_\mu^- \phi^+) - \frac{1}{2}ig^2 \frac{s_w^2}{c_w} Z_\mu^0 H (W_\mu^+ \phi^- - W_\mu^- \phi^+) + \frac{1}{2}g^2 s_w A_\mu \phi^0 (W_\mu^+ \phi^- + \\
 & W_\mu^- \phi^+) + \frac{1}{2}ig^2 s_w A_\mu H (W_\mu^+ \phi^- - W_\mu^- \phi^+) - g^2 \frac{s_w^2}{c_w} (2c_w^2 - 1) Z_\mu^0 A_\mu \phi^+ \phi^- - g^2 s_w^2 A_\mu A_\mu \phi^+ \phi^- + \\
 & \frac{1}{2}ig_s \lambda_{ij}^a (\bar{q}_i^\alpha \gamma^\mu q_j^\beta) g_\mu^a - \bar{e}^\lambda (\gamma^\mu + m_e^\lambda) e^\lambda - \bar{\nu}^\lambda (\gamma^\mu + m_\nu^\lambda) \nu^\lambda - \bar{u}_j^\lambda (\gamma^\mu + m_u^\lambda) u_j^\lambda - \bar{d}_j^\lambda (\gamma^\mu + m_d^\lambda) d_j^\lambda + \\
 & ig s_w A_\mu \left(-(\bar{e}^\lambda \gamma^\mu e^\lambda) + \frac{2}{3}(\bar{u}_j^\lambda \gamma^\mu u_j^\lambda) - \frac{1}{3}(\bar{d}_j^\lambda \gamma^\mu d_j^\lambda) \right) + \frac{ig}{4c_w} Z_\mu^0 \{ (\bar{\nu}^\lambda \gamma^\mu (1 + \gamma^5) \nu^\lambda) + (\bar{e}^\lambda \gamma^\mu (4s_w^2 - \\
 & 1 - \gamma^5) e^\lambda) + (\bar{d}_j^\lambda \gamma^\mu (\frac{2}{3}s_w^2 - 1 - \gamma^5) d_j^\lambda) + (\bar{u}_j^\lambda \gamma^\mu (1 - \frac{8}{3}s_w^2 + \gamma^5) u_j^\lambda) \} + \\
 & \frac{ig}{2\sqrt{2}} W_\mu^+ \left((\bar{\nu}^\lambda \gamma^\mu (1 + \gamma^5) U^{lep}{}_{\lambda e} e^\lambda) + (\bar{u}_j^\lambda \gamma^\mu (1 + \gamma^5) C_{\lambda e} d_j^\lambda) \right) + \\
 & \frac{ig}{2\sqrt{2}} W_\mu^- \left((\bar{e}^\lambda U^{lep}{}_{\lambda \nu} \nu^\lambda) + (\bar{d}_j^\lambda C_{\lambda \nu}^1 \gamma^\mu (1 + \gamma^5) u_j^\lambda) \right) + \\
 & \frac{ig}{2M\sqrt{2}} \phi^+ \left(-m_e^\lambda (\bar{\nu}^\lambda U^{lep}{}_{\lambda e} (1 - \gamma^5) e^\lambda) + m_\nu^\lambda (\bar{\nu}^\lambda U^{lep}{}_{\lambda \nu} (1 + \gamma^5) \nu^\lambda) + \right. \\
 & \left. \frac{ig}{2M\sqrt{2}} \phi^- \left(m_e^\lambda (\bar{e}^\lambda U^{lep}{}_{\lambda \nu} (1 + \gamma^5) \nu^\lambda) - m_\nu^\lambda (\bar{\nu}^\lambda U^{lep}{}_{\lambda e} (1 - \gamma^5) e^\lambda) - \frac{g}{M} m_\nu^\lambda H (\bar{\nu}^\lambda \nu^\lambda) - \right. \right. \\
 & \left. \left. \frac{g}{M} m_\nu^\lambda H (\bar{e}^\lambda e^\lambda) + \frac{ig}{2} m_\nu^\lambda \phi^0 (\bar{\nu}^\lambda \gamma^5 \nu^\lambda) - \frac{ig}{2} m_\nu^\lambda \phi^0 (\bar{e}^\lambda \gamma^5 e^\lambda) - \frac{1}{4} \bar{\nu}_\lambda M_{\lambda \kappa}^R (1 - \gamma^5) \bar{\nu}_\kappa - \right. \right. \\
 & \left. \left. \frac{1}{4} \bar{\nu}_\lambda M_{\lambda \kappa}^R (1 - \gamma^5) \bar{\nu}_\kappa + \frac{ig}{2M\sqrt{2}} \phi^+ \left(-m_\nu^\lambda (\bar{u}_j^\lambda C_{\lambda \kappa} (1 - \gamma^5) d_j^\lambda) + m_\nu^\lambda (\bar{u}_j^\lambda C_{\lambda \kappa} (1 + \gamma^5) d_j^\lambda) \right) + \right. \right. \\
 & \left. \left. \frac{ig}{2M\sqrt{2}} \phi^- \left(m_\nu^\lambda (\bar{d}_j^\lambda C_{\lambda \kappa}^1 (1 + \gamma^5) u_j^\lambda) - m_\nu^\lambda (\bar{d}_j^\lambda C_{\lambda \kappa}^1 (1 - \gamma^5) u_j^\lambda) - \frac{g}{M} m_\nu^\lambda H (\bar{u}_j^\lambda u_j^\lambda) - \frac{g}{M} m_\nu^\lambda H (\bar{d}_j^\lambda d_j^\lambda) + \right. \right. \\
 & \left. \left. \frac{ig}{2} m_\nu^\lambda \phi^0 (\bar{u}_j^\lambda \gamma^5 u_j^\lambda) - \frac{ig}{2} m_\nu^\lambda \phi^0 (\bar{d}_j^\lambda \gamma^5 d_j^\lambda) \right) \right)
 \end{aligned}$$

Standard Model Lagrangian

Definitions

Definition (Comprehensibility, $C(S, P)$)

The comprehensibility of a definition (or program) P with respect to a human population S is the mean accuracy with which a human s from population S after brief study and without further sight can use P to classify new material sampled randomly from the definition's domain

Definition (Inspection time $T(S, P)$)

The inspection time T of a definition (or program) P with respect to a human population S is the mean time a human s from S spends studying P before applying P to new material

Definition (Textual complexity, $Sz(P)$)

The textual complexity Sz of a definition or definite program P is the sum of the occurrences of predicate symbols, functions symbols and variables found in P

How interpretable are interpretable models?

- Domain
 - Problems in healthcare e.g., risk of mortality have more stringent requirement than retail e.g., placement of ads [Ahmad 2018]
- Soundness
 - An explanation is *Sound* if it adheres to how the model works [Kuleza 2014]
- Completeness
 - An explanation is *Complete* if it encompasses the complete extent of the model [Kuleza 2014]
- Modality
 - Until recently most interpretable large *predictive* time series models were not really interpretable [Schlegel 2019]

Domain specificity of interpretations

- Describing the trained ML model in terms of *domain ontology* without using terms that are foreign to the domain where the ML task must be solved [Kovalerchuk, 2020]
- The explanations/interpretation has to make sense to the domain expert who is going to use the ML model
- The same data may have different meaning in different domains e.g., ratings in Uber vs. Amazon



User centricity of interpretations

- Explanations need to be in the right language and in the right context [Doshi-Velez 2014, Druzdzel 1996]
- ELI5 Principle: Explain it like I am 5 (years old)
- Making domain sense may require sacrificing or deemphasizing model fidelity
- Explanations should be role-based - A physician requires different explanations as compared to a staffing planner in a hospital

Ante-Hoc vs. Post Hoc Models

Ante-Hoc (Internally Interpreted)

- Models where the predictive model and the explanation model is the same
- ML model explained in terms of interpreted elements of their structure not only inputs

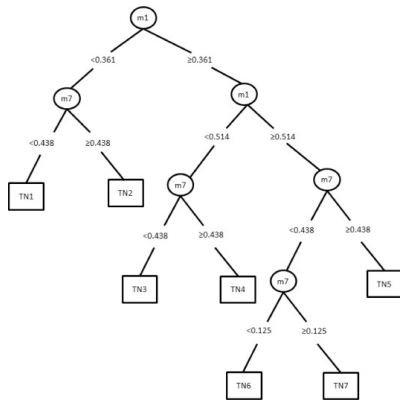
Post-Hoc (Externally Interpreted)

- Models where the predictive model and the explanation model are different
- ML model explained in terms of interpretable input data and variables, but without interpreting the model structure

Explicit vs. Implicit Interpretations

Explicit Interpretations

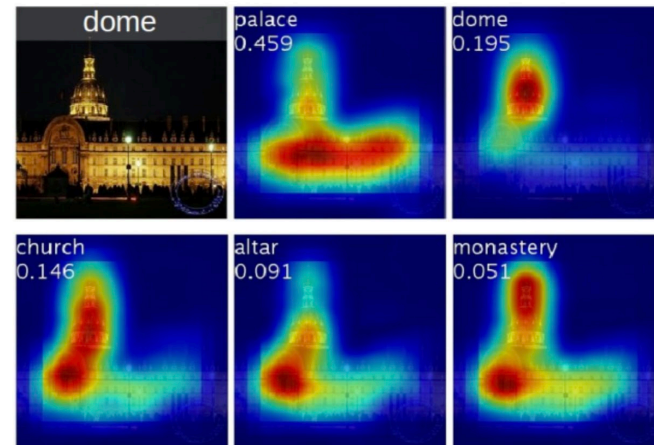
The interpretation is explicitly from the model output



Decision Trees

Implicit Interpretations

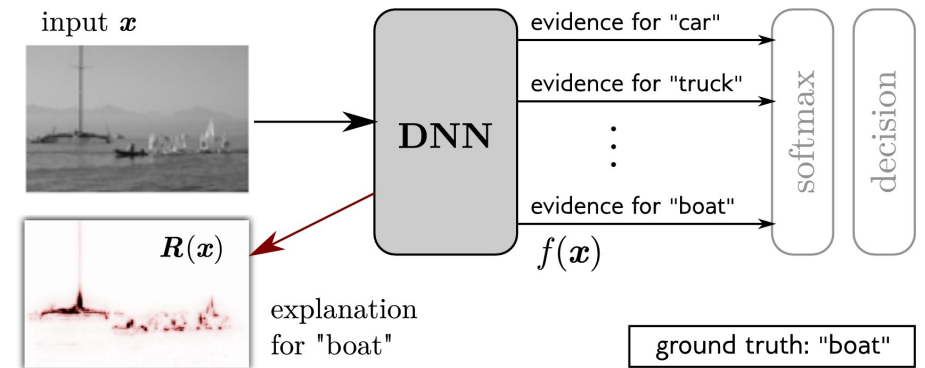
The interpretation needs to be derived after the application of additional domain knowledge



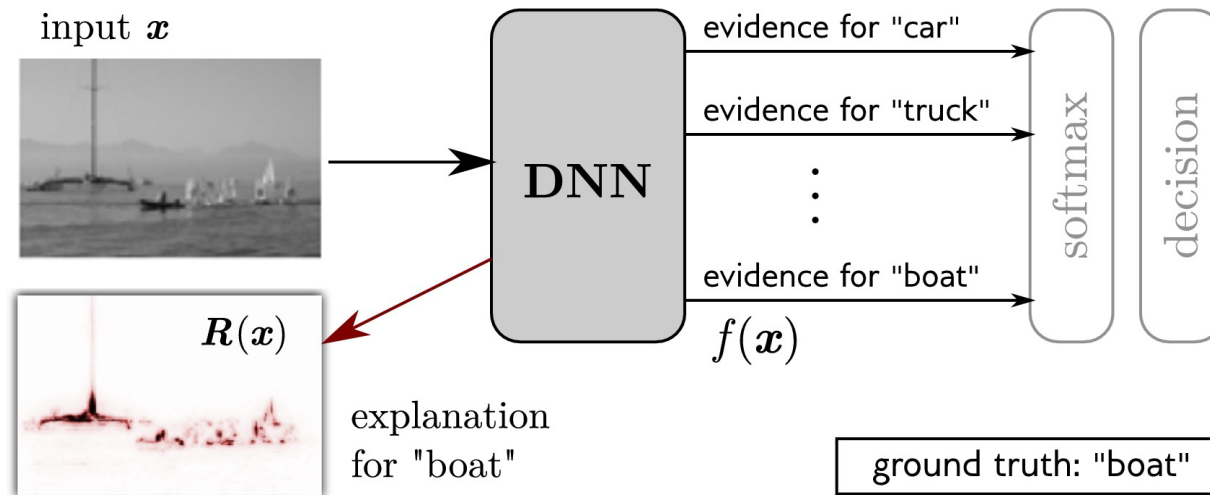
Heatmaps

Implicit Explanations

- Consider recognizing a boat in an image
Recognize an image as boat based on a group of pixels that look like a mast
- This explanation is not applicable to another boat in the same image since that boat has no mast and requires its own explanation. Such conceptual explanations cannot be derived from DNN models
- In contrast, in medical imaging, if a radiologist cannot explicitly match DNN dominant pixels with the domain concepts such as tumor, these pixels will not serve as an explanation for the radiologist.

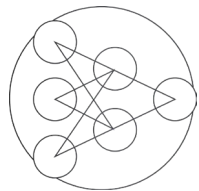


Using black-box models to explain black box models?

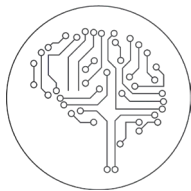


The dominant/salient pixels of the image represent the mast as a distinct feature of the boat relative to a car and a truck. This is a result of human knowledge what is a mast that is not explicitly present in the image

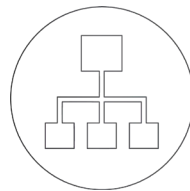
Tutorial Scope



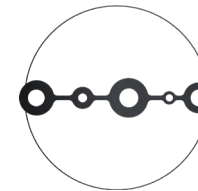
Bayesian Models



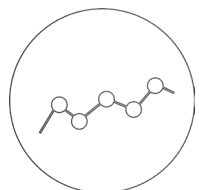
Neural Nets



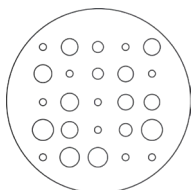
Ensemble Models



Markov Models



Statistical Models



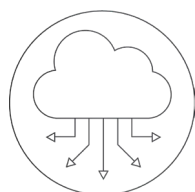
Graphical Models



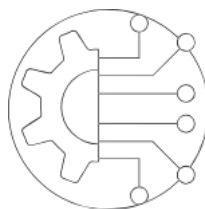
Reinforcement
Learning



Natural Language
Processing



Expert Systems

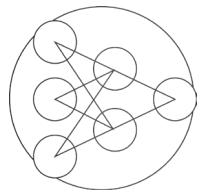


Supervised Learning

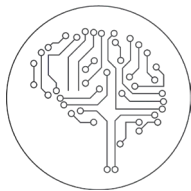


Recommendation
Systems

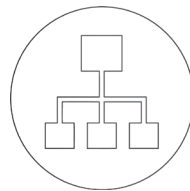
Tutorial Scope



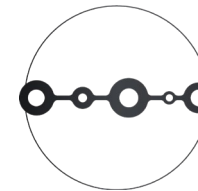
Bayesian Models



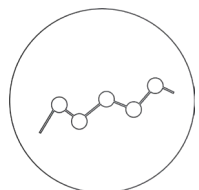
Neural Nets



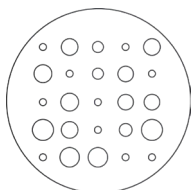
Ensemble Models



Markov Models



Statistical Models



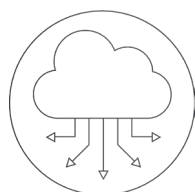
Graphical Models



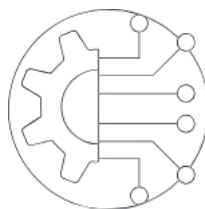
Reinforcement Learning



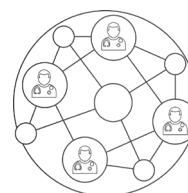
Natural Language Processing



Expert Systems



Supervised Learning



Recommendation Systems

Discovering Visual Interpretable Models

What is visual interpretability?

if	bruises=no, odor=not-in-(none,foul)	then probability that the mushroom is edible = 0.00112
else if	odor=foul, gill-attachment=free,	then probability that the mushroom is edible = 0.0007
else if	gill-size=broad, ring-number=one,	then probability that the mushroom is edible = 0.999
else if	stalk-root=unknown,	then probability that the mushroom is edible = 0.996
else if	stalk-surface-above-ring=smooth,	then probability that the mushroom is edible = 0.0385
else if	bruises=foul, veil-color=white,	then probability that the mushroom is edible = 0.995
else if	stalk-shape=tapering,	then probability that the mushroom is edible = 0.986
else if	ring-number=one,	then probability that the mushroom is edible = 0.958
else if	habitat=paths,	then probability that the mushroom is edible = 0.001
else	(default rule)	

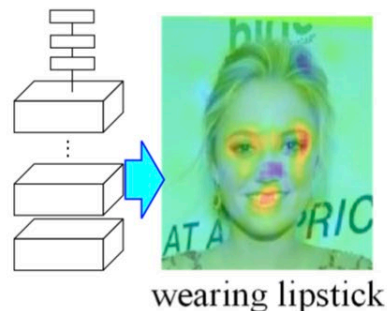
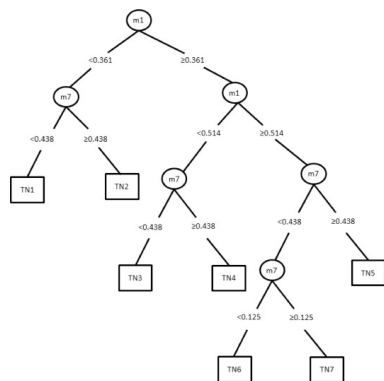
[Yang 2017]

Non-Visual Methods

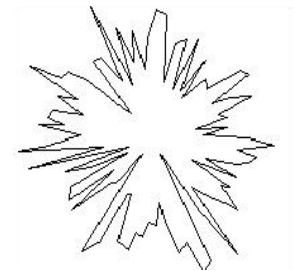
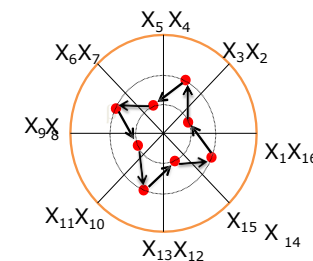
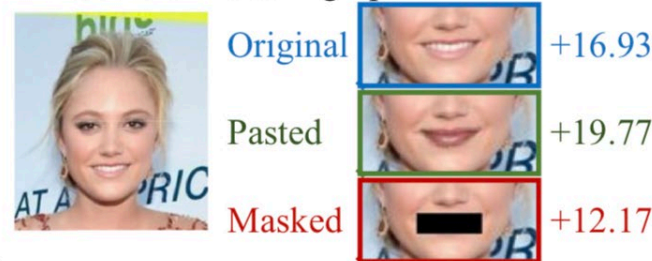
- **if** hemiplegia and age > 60
 - **then** stroke risk 58.9% (53.8%–63.8%)
- **else if** cerebrovascular disorder
 - **then** stroke risk 47.8% (44.8%–50.7%)
- **else if** transient ischaemic attack
 - **then** stroke risk 23.8% (19.5%–28.4%)
- **else if** occlusion and stenosis of carotid artery without infarction
 - **then** stroke risk 15.8% (12.2%–19.6%)
- **else if** altered state of consciousness and age > 60
 - **then** stroke risk 16.0% (12.2%–20.2%)
- **else if** age ≤ 70
 - **then** stroke risk 4.6% (3.9%–5.4%)
- **else** stroke risk 8.7% (7.9%–9.6%)

[Letham 2015]

Visual Methods

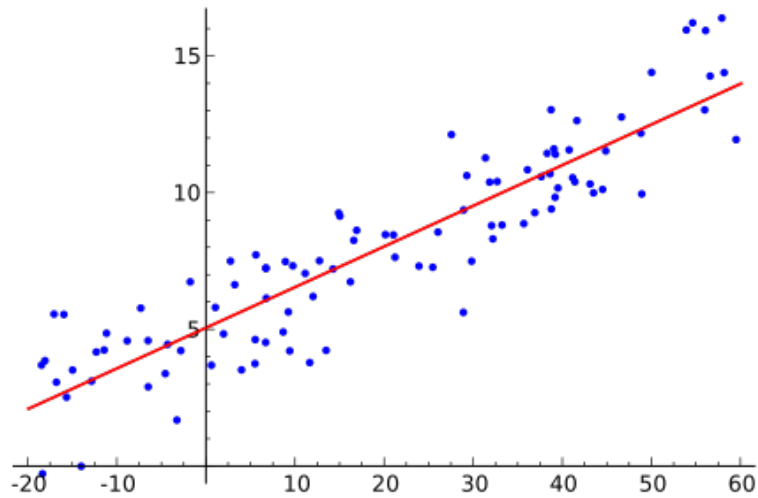


Score of “wearing lipstick”



The Allure of Visual Methods

Visual Understanding



Textual Understanding

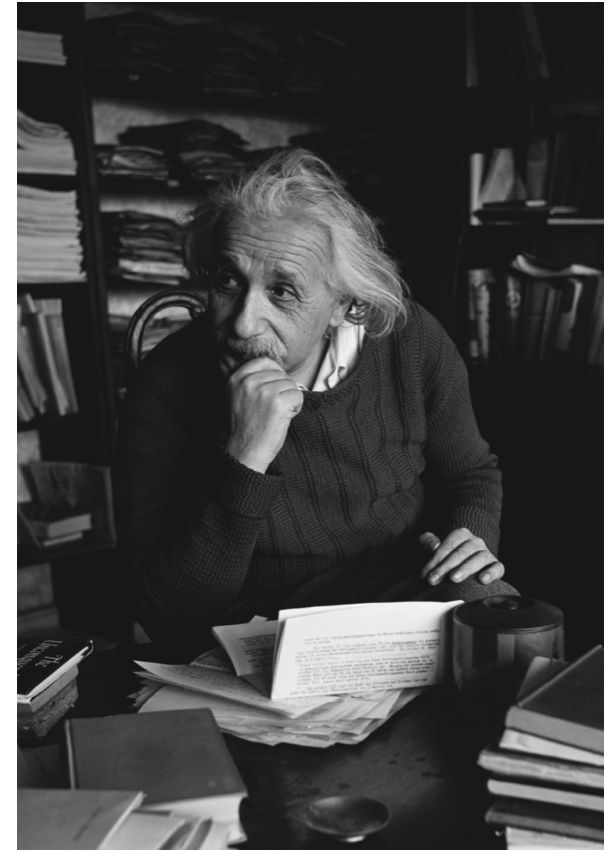
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n$$

Understanding the algorithm
may not mean be sufficient

$$a_j^l = \sigma \left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l \right)$$

Why Visual Thinking?

- A lot of creative thinking is visual
- Scientists who declared the fundamental role that images played in their most creative thinking: Bohr, Boltzmann, Einstein, Faraday, Feynman, Heisenberg, Helmholtz, Herschel, Kekule, Maxwell, Poincare, Tesla, Watson, Watt etc.
- Albert Einstein: “The words or the language, as they are written or spoken, do not seem to play any role in my mechanism of thought.”



[Thagard & Cameron, 1997; Hadamard, 1954; Shepard & Cooper, 1982]

Pre-History of Visual Thinking

Chinese and Indians knew a visual proof of the Pythagorean Theorem in 600 B.C. before it was known to the Greeks [Kulpa 1994]



Madhura Meenakshi Temple

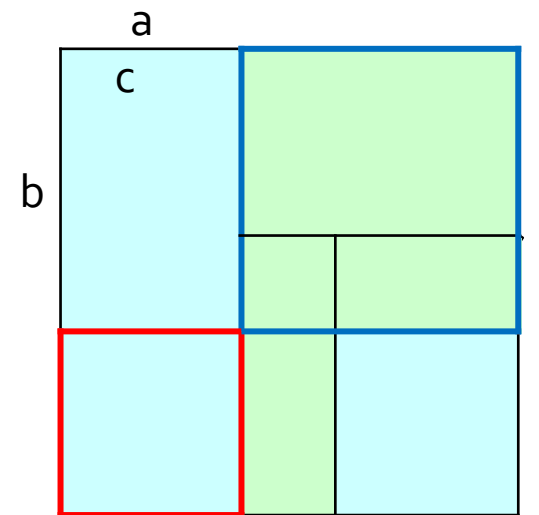
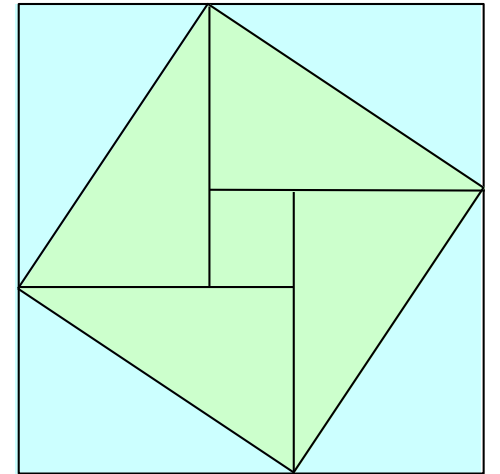


Lingxiao Pagoda of Zhengding

Pre-History of Visual Thinking

$(a+b)^2$ (area of the large square) = $a^2+b^2+ab+ab=(a+b)^2$
 $a^2+b^2=(a+b)^2$ (area of the large square) - $2ab$ (4 light green triangles) = c^2 (area of inner darker green square)

Thus, we follow this tradition -- moving from visualization of solution to finding solution visually with modern data science tools



Approaches to discovering visual methods

- We are moving from visualization of solution to finding solution visually
- Why Visual?
 - To leverage human perceptual capabilities
- Why interactive?
 - To leverage human abilities to adjust tasks on the fly
- Why Machine Learning?
 - To leverage analytical discovery that are outside of human abilities
 - We cannot see patterns in multidimensional data by a naked eye

Approaches beyond visualization of existing models

- Components of approaches:
 - Visual methods for n-D data representation
 - Visual methods for model discovery in visual n-D data representations
 - Methods to interpret visual data representations and models that are not internally interpretable
- Visual methods for 2-D/3-D representation of n-D data
 - Reversible/lossless/interpretable: Parallel Coordinates, Radial Coordinates, General Line Coordinates, Shifted Paired Coordinates, Collocated Paired Coordinates, and others.
 - Non-reversible/lossy/with challenging interpretation: PCA, MDF, RadVis, Manifolds, t-SNE and others

What is Visual Discovery?

x	y	class
1	0.5	1
1.1	6	2
2	1.5	1
2.2	5	2
2.8	2.8	1
3	4	2
3.5	3.3	1
4	3.8	1
4	2.6	2
4.5	4.7	1
5	1.8	2
5	5	1
5.5	5.5	1
6	0.8	2

What would be the best guess about a line fitting this data?

What is Visual Discovery?

x	y	class
1	0.5	1
1.1	6	2
2	1.5	1
2.2	5	2
2.8	2.8	1
3	4	2
3.5	3.3	1
4	3.8	1
4	2.6	2
4.5	4.7	1
5	1.8	2
5	5	1
5.5	5.5	1
6	0.8	2

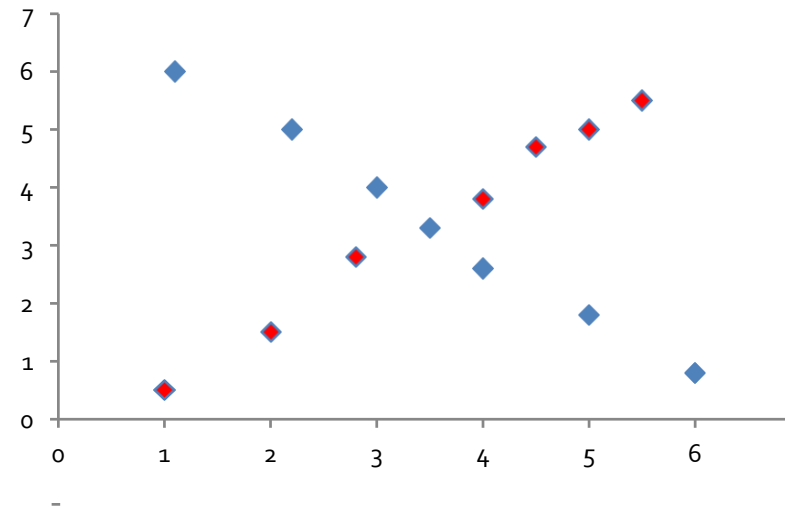
What would be the best guess about a fitting this data?

A simple linear discrimination function

What is Visual Discovery?

x	y	class
1	0.5	1
1.1	6	2
2	1.5	1
2.2	5	2
2.8	2.8	1
3	4	2
3.5	3.3	1
4	3.8	1
4	2.6	2
4.5	4.7	1
5	1.8	2
5	5	1
5.5	5.5	1
6	0.8	2

=

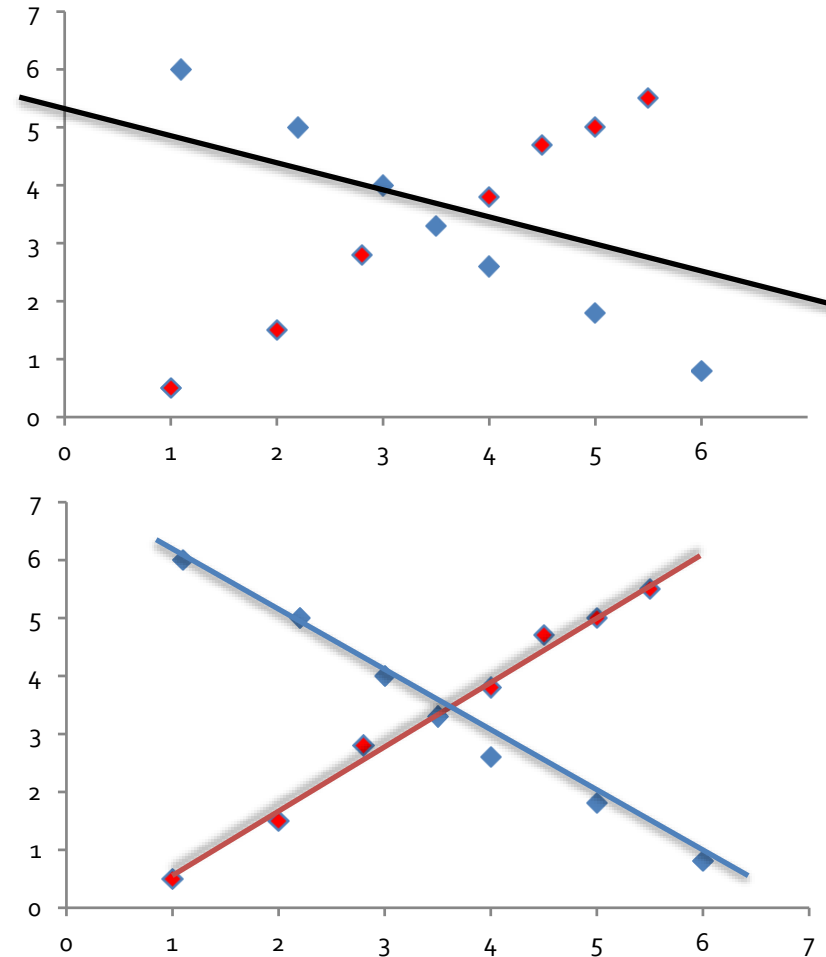


In contrast a quick look at these data, immediately gives a visual insight of a correct model class of “crossing” two linear functions

Visual Discovery in 2-D

x	y	class
1	0.5	1
1.1	6	2
2	1.5	1
2.2	5	2
2.8	2.8	1
3	4	2
3.5	3.3	1
4	3.8	1
4	2.6	2
4.5	4.7	1
5	1.8	2
5	5	1
5.5	5.5	1
6	0.8	2

==

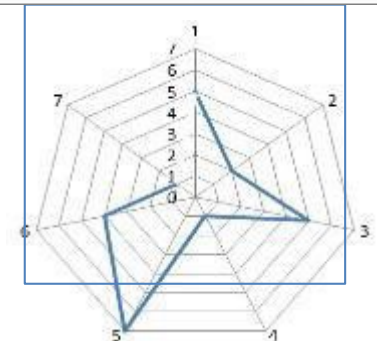
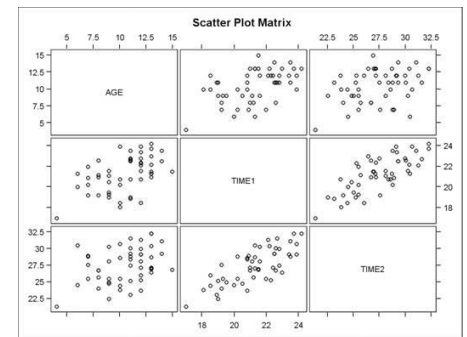
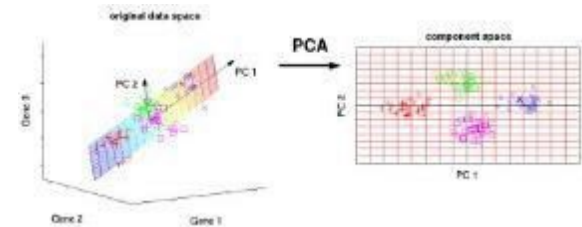


How to do visual discovery in n -dimensions?

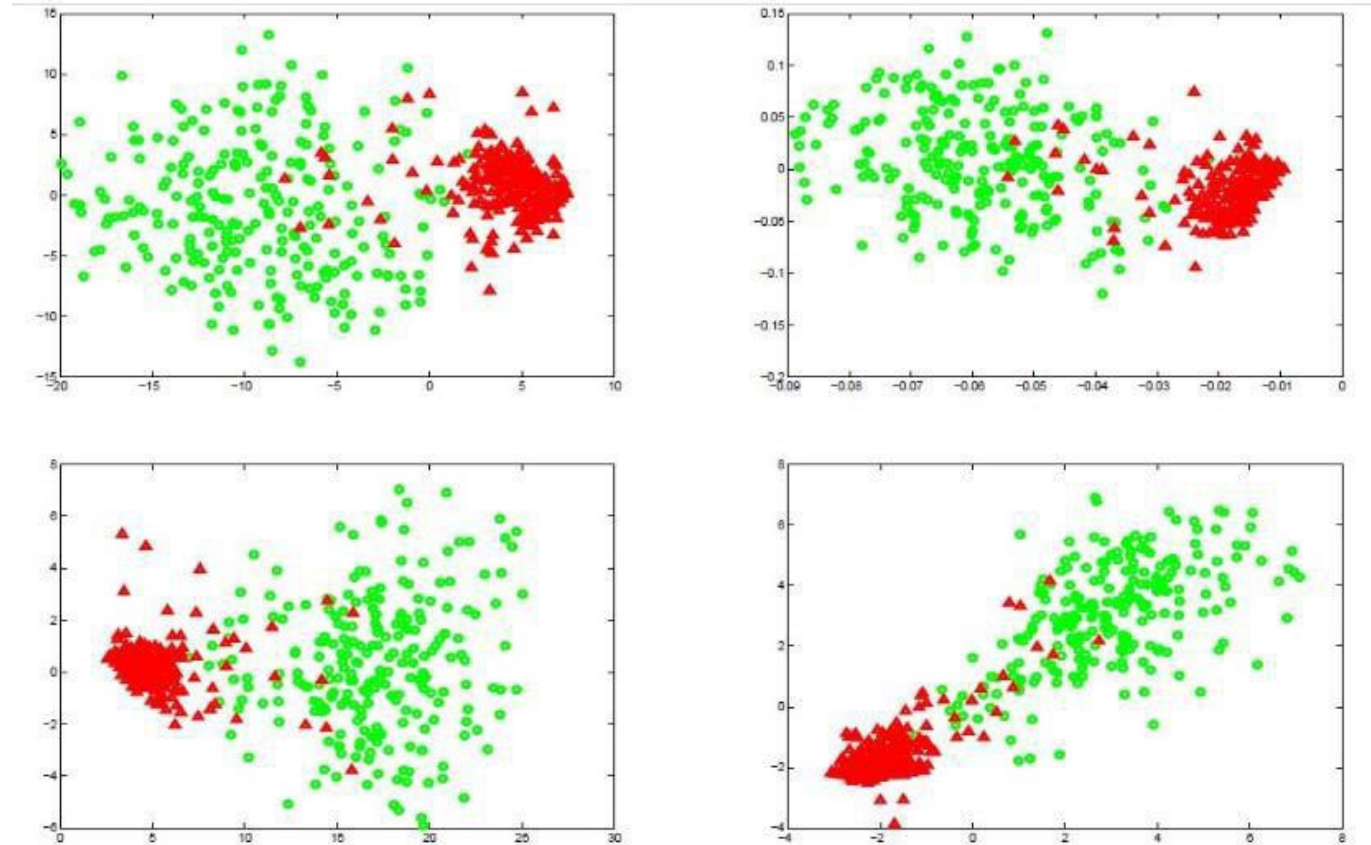
ID	FD1	FD2	FD4	FD5	FD6	FD10	FD12	FD15	FD16	FD18	FD20	FD22	FD23	FD24	FD25	FD26	FD27	FD28
1	0	0	2.749807	9.826302	4.067554	0	0	0	5.244006	0	2.743422	0	0	0	0	6.254963	0	0
2	11.51334	9.092989	0	12.46223	0	7.597155	0	0	8.940897	0	0	0	4.268456	0	0	0	0	1.309903
3	10.27931	0	2.075787	0	4.042145	0	0	0.477713	3.97378	0	0	2.477745	0	0	0	5.583099	0	7.418219
4	0	18.31495	0	0	0	0	0	0	4.472742	4.671682	0	7.248355	12.11645	0	0	0	6.030322	0
5	14.12261	15.1236	9.695051	0	0.915031	0	0	6.086389	9.139287	0	0	0	8.931774	0	0	0	0	0
6	0	0	5.405394	0	0	2.951092	0	3.797284	4.576391	0	0	0	0	0	2.763756	0	0	2.562996
7	0	0	0	8.068472	0	3.267916	0	0	5.09157	6.082168	0	0	5.42044	0	0	4.431955	0.415844	2.73227
8	6.169271	4.918356	5.566813	0	0	4.884737	5.168666	0	5.189289	0	0	0	2.49011	0	4.750784	2.994664	0	0
9	11.64548	0	0	12.16663	0	8.407408	0	0	0	0	0	0	4.289772	0	0	4.652006	0	0
10	9.957874	7.829115	0	0	0	0	0	0	7.082694	8.388349	0	0	0	0	0	4.706276	0	0.705345
11	9.994487	12.3192	3.058695	0	0	0	6.111047	0.380701	3.904454	0	2.573056	0	0	0	0	5.610187	0	0
12	0	8.446147	7.506574	0	0	5.846259	7.362241	6.557457	7.627757	9.05184	0	0	0	0	6.646436	0	0	0
13	13.65315	18.11681	2.457055	0	8.218276	0	5.689919	0	4.45029	3.213032	5.992753	0	11.56691	0	0	7.734966	0	0
14	0	0	0	8.710629	0	0	0	0	6.466624	0	0	0	3.865449	0	5.339944	3.943355	0	0
15	11.08665	0	0	12.57808	0	8.377558	0	9.269582	0	10.28637	0	0	4.141793	0	0	4.953615	0	0.433766
16	0	0	7.32989	9.848915	0	0	6.639803	0	0	0	0	0	0	0	0	4.288343	0	0
17	0	0	8.49376	0	0	0	7.403671	9.346368	0	0	0	0	0	0	0	0	0	0
18	9.52255	0	0	10.30969	0	0	6.508697	0	0	9.04743	0	0	3.113288	0	7.667032	0	0	0
19	0	9.237608	3.488988	7.443493	0	0	0	0	0	0	0.921821	1.305681	0	0	0	4.447716	0	4.174564
20	0	16.78071	2.745921	0	5.606468	0	7.824948	0	0	4.807075	4.454489	0	0	0	0	7.226364	0	10.62363
21	0	0	8.18506	0	0.469365	4.241147	0	5.823779	0	0	0	0	0	0	0	6.475445	0	4.49432
22	9.609696	12.07202	0	6.483721	0	0	0	0	0	1.554688	0	5.446015	0	0	0	0	0	9.85667
23	10.71318	0	0	11.44685	0	8.097867	0	8.832153	8.646919	0	0	0	0	0	0	4.705225	0	0
24	6.625456	0	3.686915	6.715843	0.187058	0	3.735899	3.55698	0	0	0	0	0	0	2.996381	3.700704	0	0
25	9.794333	0	0	9.788224	0	4.599581	0	0	0	0	0	0	0	0	0	4.694789	0	0 ¹⁰
26	10.25995	0	0	9.531824	0	1.156152	6.604298	0	0	0	0	0	6.346496	0	1.300262	0	1.869395	4.265034

Multi-dimensional data visualization

- In high-dimensions one cannot comprehensively see data
- Methods for lossless and interpretable visualization of n-D data in 2-D are required
- Often multidimensional data are visualized by **lossy dimension reduction** (e.g., PCA) or by **splitting** n-D data to a set of low dimensional data (pairwise correlation plots)
- While splitting is useful it destroys integrity of n-D data and leads to a shallow understanding complex n-D data
- An alternative for deeper understanding of n-D data is visual representations of n-D data in low dimensions without splitting and loss of information is **graphs not 2-D points** e.g., Parallel and Radial coordinates



Example: WBC



- Benign and malignant cancer cases overlap
- Interpretation of dimensions is difficult. Non-reversible lossy methods: 9-D to 2-D

g. 4. Wisconsin data set, top row: MDS and PCA, bottom row: FDA and SVM

[Maszczyk 2008]

Johnson-Lindenstrauss Lemma

Given $0 < \varepsilon < 1$, a set X of m points in \mathbb{R}^N , and a number $n > 8 \ln(m)/\varepsilon^2$, there is a linear map $f : \mathbb{R}^N \rightarrow \mathbb{R}^n$ such that

$$(1 - \varepsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \varepsilon)\|u - v\|^2$$

for all $u, v \in X$.

- Only a small number of arbitrary n-D points can be mapped to k-D points of a smaller dim k that preserve n-D distances with relatively small deviations
- Reason: the 2-D visualization space does not have enough neighbors with equal distances to represent the same n-D distances in 2-D.
- Result: the significant corruption of n-D distances in 2-D visualization

Different Formulations of the Lemma

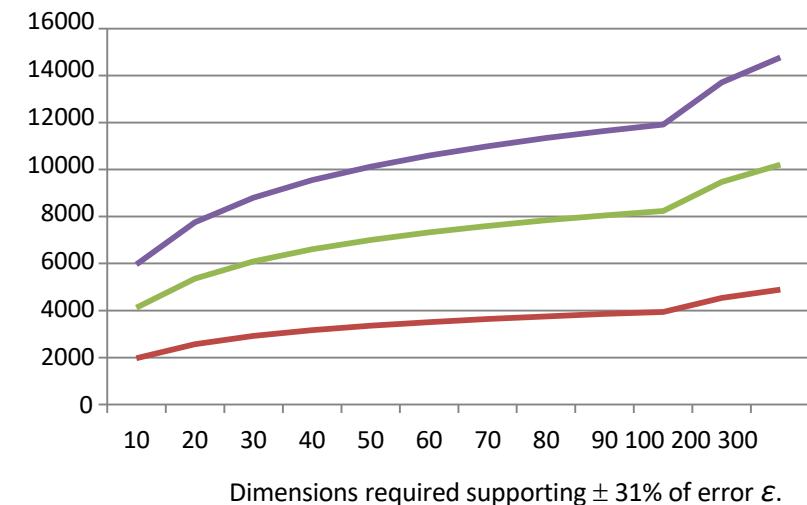
- Defines the possible dimensions $k < n$ such that for any set of m points in R^n there is a mapping $f: R^n \rightarrow R^k$ with “similar” distances in R^n and R^k between mapped points. This similarity is expressed in terms of error $0 < \varepsilon < 1$.
- For $\varepsilon = 0$ these distances are equal. For $\varepsilon=1$ the distances in R^k are less or equal to $\sqrt{2} S$, where S is the distance in R^n . This means that distance s in R^k will be in the interval $[0, 1.42S]$.
- In other words, the distances will not be more than 142% of the original distance, i.e., it will not be much exaggerated. However, it can dramatically diminish to 0

Theoretical limits: Preserve n-D in 2-D

- Johnson-Lindenstrauss Lemma shows that to keep distance errors within about 30% for just 10 arbitrary high-dimensional points, we need over 1,900 dimensions, and over 4,500 dimensions for 300 arbitrary points
- Visualization methods do not meet these requirements

Number of arbitrary points in high-dimensional space	Sufficient dimension with formula (3.1)	Sufficient dimension with formula (3.2)_	Insufficient dimension with formula (3.3)
10	1974	2145	1842
20	2568	2791	2397
30	2915	3168	2721
40	3162	3436	2951
50	3353	3644	3130
60	3509	3813	3275
70	3642	3957	3399
80	3756	4081	3506
90	3857	4191	3600
100	3947	4289	3684
200	4541	4934	4239
300	4889	5312	4563

Dimensions to support $\pm 31\%$ of error ($\epsilon=0.1$).



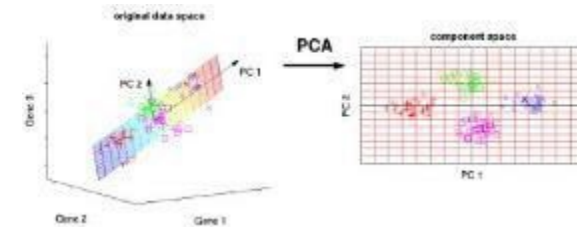
Approaches to Convert n-D data to 2-D data

- Lossy approach

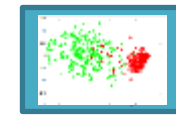
- Lossy conversion to 2-D (dimension reduction, DR)
- Point to point (n-D point to 2-D point)
- Visualization in 2-D
- Interactive discovery of 2-D patterns in visualization

- Lossless approach

- Lossless conversion (visualization) to 2-D (n-D data fully restorable from its visualization)
- Interactive discovery of 2-D patterns on graphs in visualization



n-D data



2-D data &
2-D patterns



n-D data and
n-D patterns



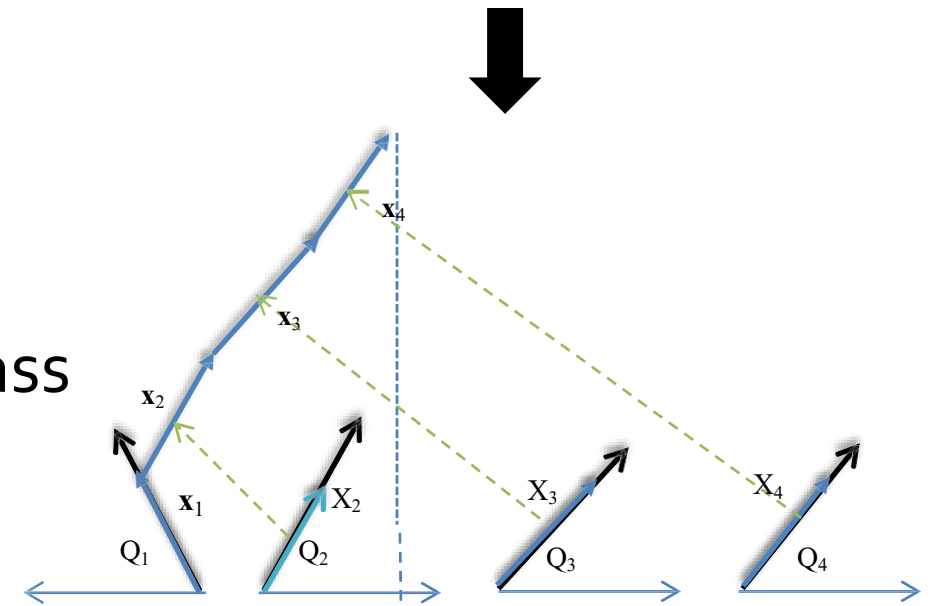
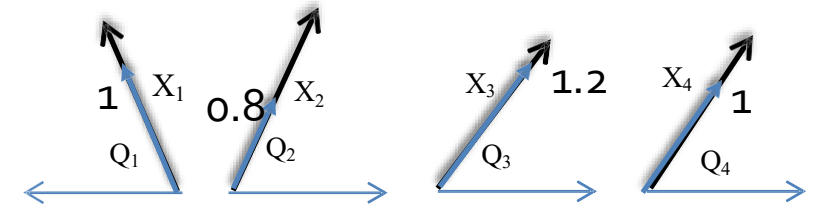
2-D data & n-D
patterns

GLC-L Algorithms for Visualization

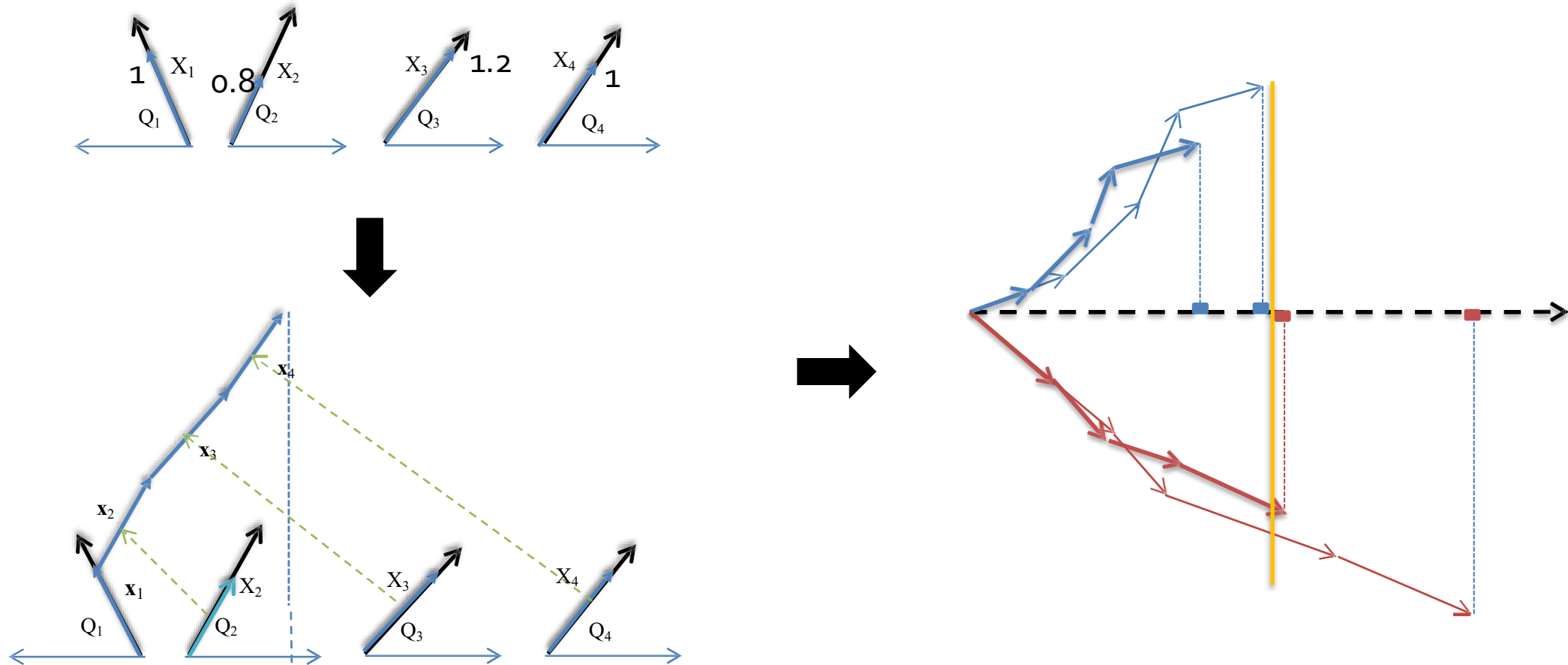
Given: 4-D point $(x_1, x_2, x_3, x_4) = (1, 0.8, 1.2, 1)$

Algorithm

- 4 coordinate lines at different angles Q_1 - Q_4
- Values shown as blue lines (vectors)
- Shifting and stacking blue lines
- Projecting the last point to U line
- Do the same for other 4-D points of blue class
- Do the same for 4-D points of red class
- Optimize angles Q_1 - Q_4 to separate classes (yellow line)



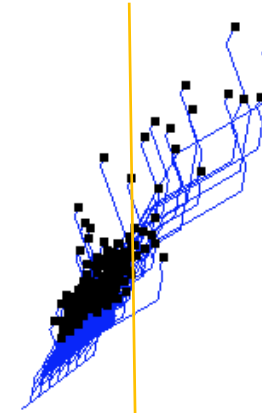
GLC-L Algorithms for Visualization



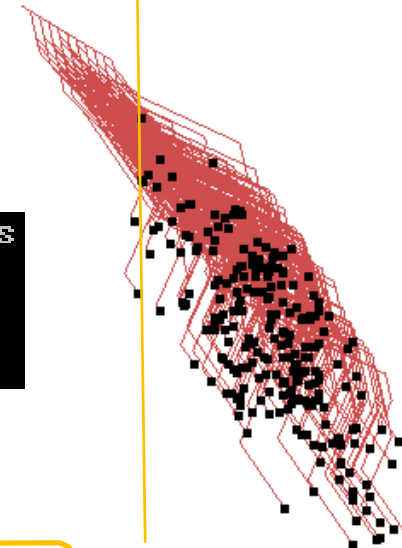
9-D Wisconsin Breast Cancer

- Critical in Medical diagnostics and many other fields
- Explanation of patterns and understanding patterns
- Lossless visual means Reversible/restorable
- Only one malignant (red case) on the wrong side

444 benign (blue) cases



239 malignant (red) cases

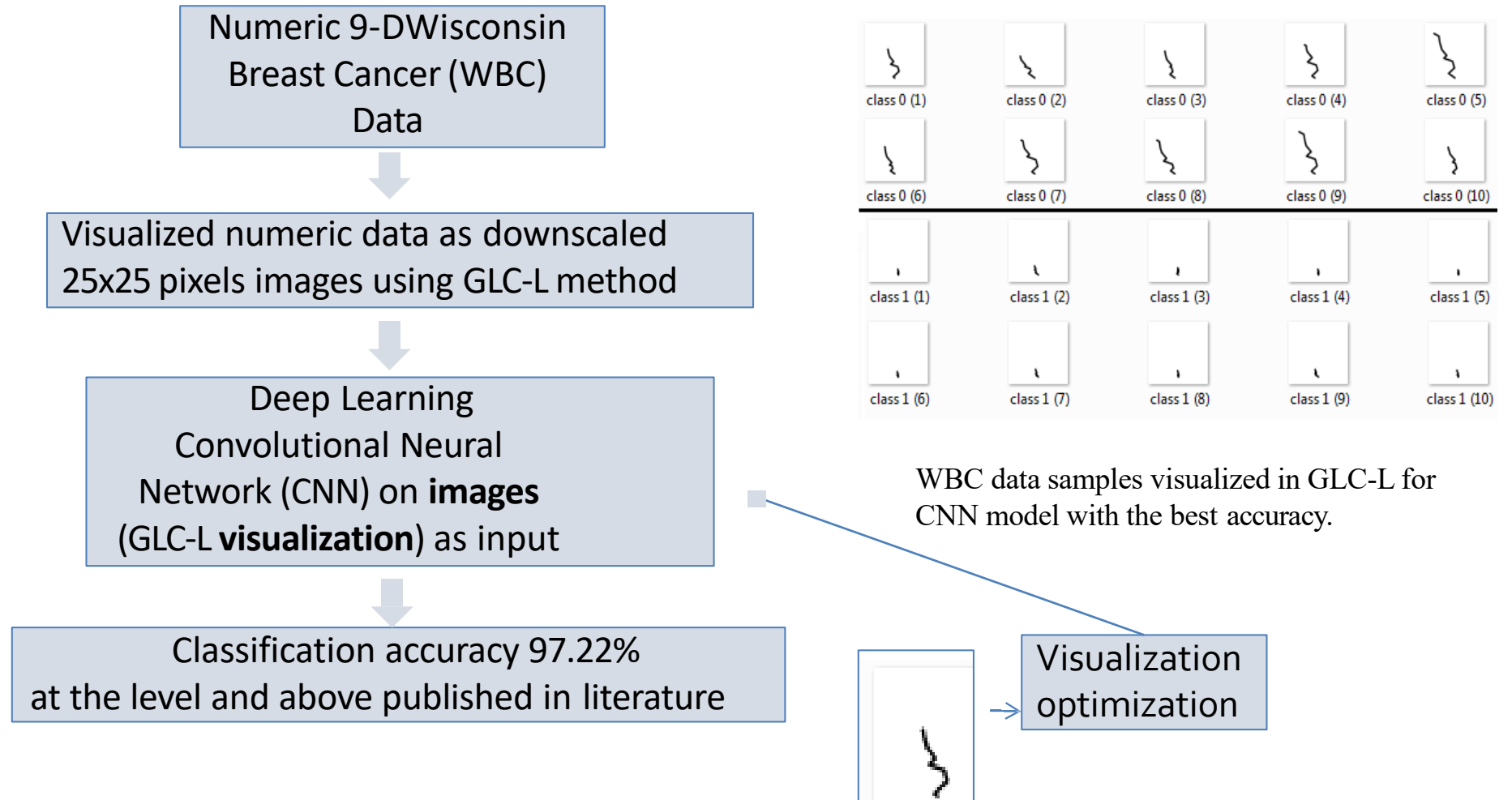


Real Class	Predicted Class 1	Predicted Class 2
1	424	20
2	1	238

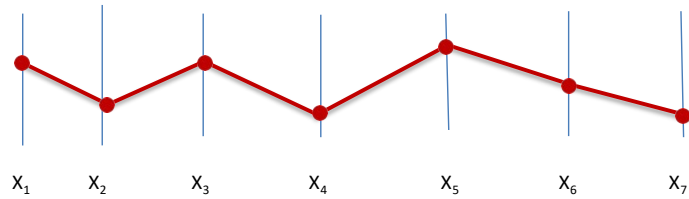
Accuracy is: 96.9253%



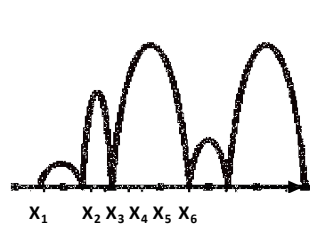
Avoiding Occlusion with Deep Learning on WBC data



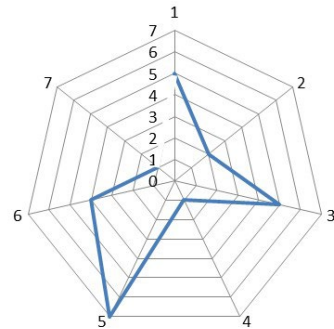
General Line Coordinates



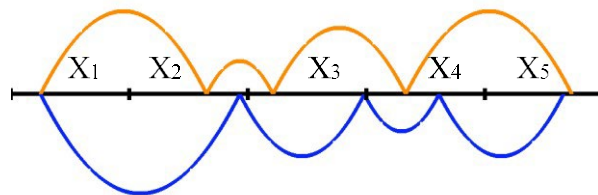
7-D point $D=(5,2,5,1,7,4,1)$ in Parallel Coordinates



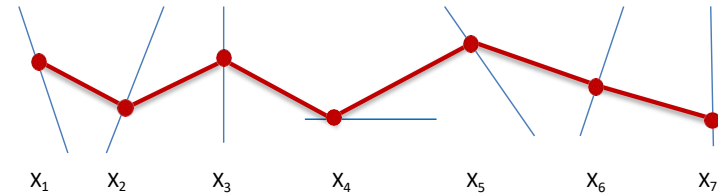
6-D $(5,4,0,6,4,10)$ point in In-line Coordinates



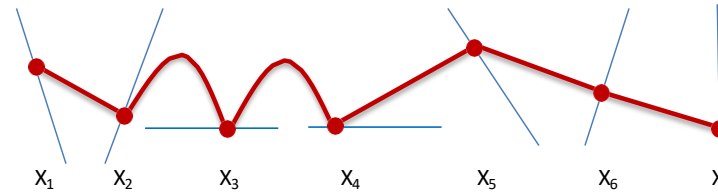
7-D point $D=(5,2,5,1,7,4,1)$ in Radial Coordinates.



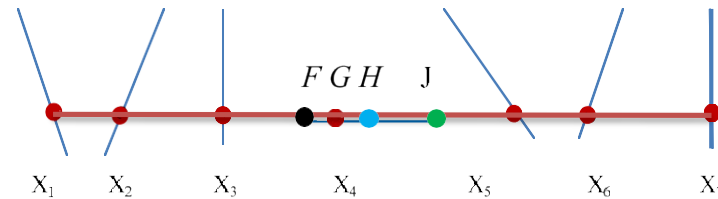
Two 5-D points of two classes in Sequential In-Line Coordinates.



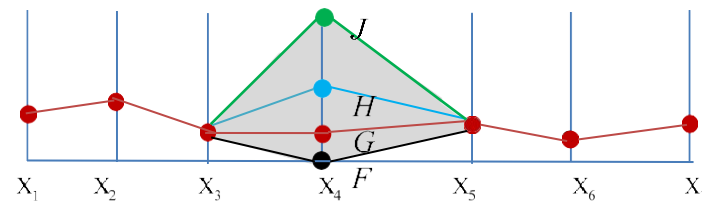
(a) 7-D point D in General Line Coordinates with straight lines.



(b) 7-D point D in General Line Coordinates with curvilinear lines.



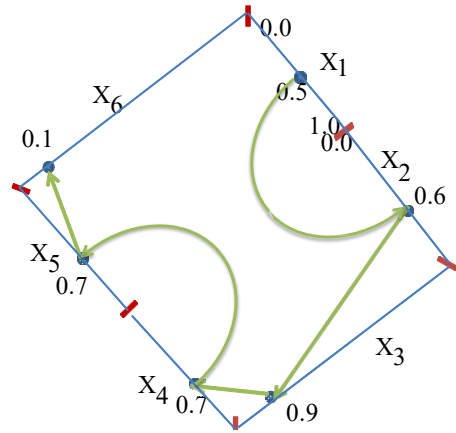
(c) 7-D points $F-J$ in General Line Coordinates that form a simple straight line.



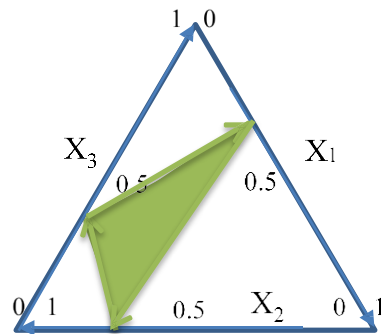
(d) 7-D points $F-J$ in Parallel Coordinates that do not form a simple straight line.

7-D points in General Line Coordinates with different directions of coordinates X_1, X_2, \dots, X_7 in comparison with Parallel Coordinates.

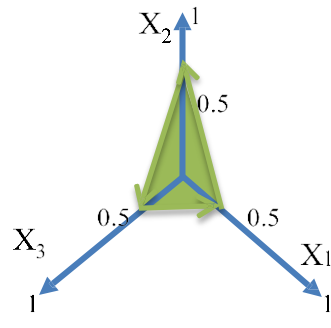
General Line Coordinates



n-Gon (rectangular) coordinates with 6-D point (0.5, 0.6, 0.9, 0.7, 0.7, 0.1).

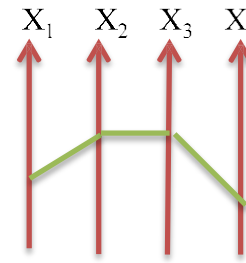


(a) Point A in in 3-Gon coordinates.

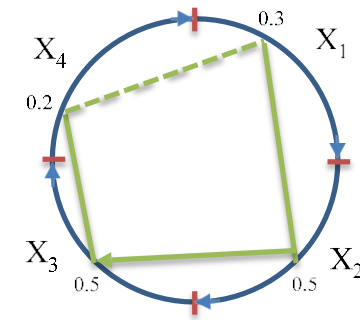


(b) Point A in in radial coordinates.

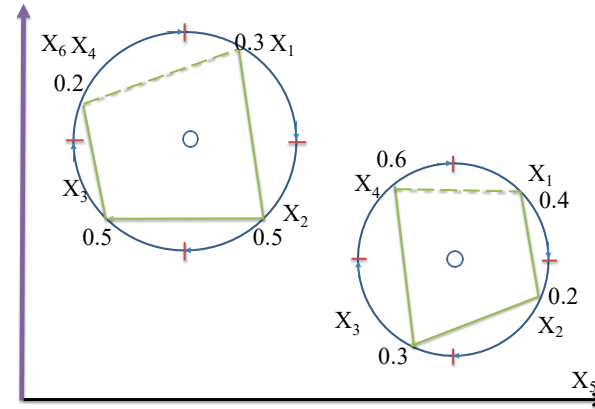
3-D point A=(0.3, 0.7, 0.4) in 3-Gon (triangular) coordinates and in radial coordinates.



(a) Parallel Coordinates display.

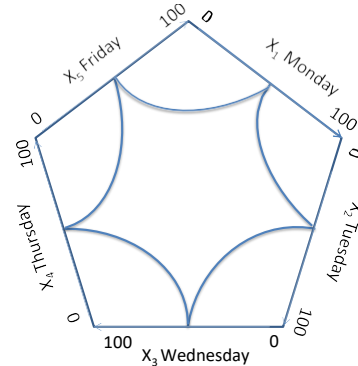


(b) Circular Coordinates display.

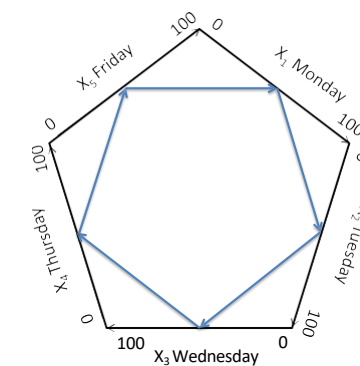


(c) Spatially distributed objects in circular coordinates with two coordinates X_5 and X_6 used as a location in 2-D and X_7 is encoded by the sizes of circles.

Figure 2.5. Examples of circular coordinates in comparison with parallel coordinates.



(a) Example in n-Gon coordinates with curvilinear edges of a graph.



(b) Example in n-Gon coordinates with straight edges of a graph.

Figure 2.6 Example of weekly stock data in n-Gon (pentagon) coordinates.

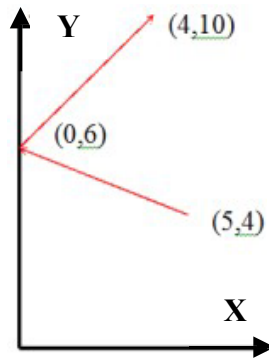
General Line Coordinates (GLC): 2-D

Type	Characteristics
2-D General Line Coordinates (GLC)	Drawing n coordinate axes in 2-D in variety of ways: curved, parallel, unparallel, collocated, disconnected, etc.
Collocated Paired Coordinates (CPC)	Splitting an n-D point x into pairs of its coordinates $(x_1, x_2), \dots, (x_{n-1}, x_n)$; drawing each pair as a 2-D point in the collocated axes; and linking these points to form a directed graph. For odd n coordinate x_n is repeated to make n even.
Basic Shifted Paired Coordinates (SPC)	Drawing each next pair in the shifted coordinate system by adding (1,1) to the second pair, (2,2) to the third pair, (i-1, i-1) to the i-th pair, and so on. More generally, shifts can be a function of some parameters.
2-D Anchored Paired Coordinates (APC)	Drawing each next pair in the shifted coordinate system, i.e., coordinates shifted to the location of a given pair (anchor), e.g., the first pair of a given n-D point. Pairs are shown relative to the anchor easing the comparison with it.
2-D Partially Collocated Coordinates (PCC)	Drawing some coordinate axes in 2D collocated and some coordinates not collocated.
In-Line Coordinates (ILC)	Drawing all coordinate axes in 2D located one after another on a single straight line.
Circular and n-Gon coordinates	Drawing all coordinate axes in 2D located on a circle or an n-Gon one after another.
Elliptic coordinates	Drawing all coordinate axes in 2D located on ellipses.
GLC for linear functions (GLC-L)	Drawing all coordinates in 2D dynamically depending on coefficients of the linear function and value of n attributes.
Paired Crown Coordinates (PWC)	Drawing odd coordinates collocated on the closed convex hull in 2-D and even coordinates orthogonal to them as a function of the odd coordinate.

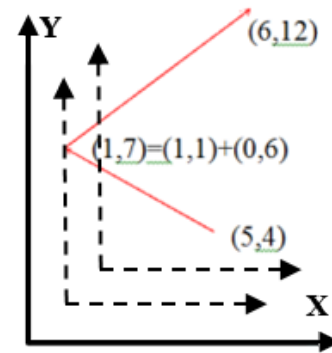
General Line Coordinates (GLC): 3-D

Type	Characteristics
3-D General Line Coordinates (GLC)	Drawing n coordinate axes in 3-D in variety of ways: curved, parallel, unparallel, collocated, disconnected, etc.
Collocated Tripled Coordinates (CTC)	Splitting n coordinates into triples and representing each triple as 3-D point in the same three axes; and linking these points to form a directed graph. If $n \bmod 3$ is not 0 then repeat the last coordinate X_n one or two times to make it 0.
Basic Shifted Tripled Coordinates (STC)	Drawing each next triple in the shifted coordinate system by adding (1,1,1) to the second tripple, (2,2,2) to the third tripple (i-1, i-1,i-1) to the i-th triple, and so on. More generally, shifts can be a function of some parameters.
Anchored Tripled Coordinates (ATC) in 3-D	Drawing each next triple in the shifted coordinate system, i.e., coordinates shifted to the location of the given triple of (anchor), e.g., the first triple of a given n-D point. Triple are shown relative to the anchor easing the comparison with it.
3-D Partially Collocated Coordinates (PCC)	Drawing some coordinate axes in 3-D collocated and some coordinates not collocated.
3-D In-Line Coordinates (ILC)	Drawing all coordinate axes in 3D located one after another on a single straight line.
In-Plane Coordinates (IPC)	Drawing all coordinate axes in 3D located on a single plane (2-D GLC embedded to 3-D).
Spherical and polyhedron coordinates	Drawing all coordinate axes in 3D located on a sphere or a polyhedron.
Ellipsoidal coordinates	Drawing all coordinate axes in 3D located on ellipsoids.
GLC for linear functions (GLC-L)	Drawing all coordinates in 3D dynamically depending on coefficients of the linear function and value of n attributes.
Paired Crown Coordinates (PWC)	Drawing odd coordinates collocated on the closed convex hull in 3-D and even coordinates orthogonal to them as a function of the odd coordinate value.

Reversible Lossless Paired Coordinates

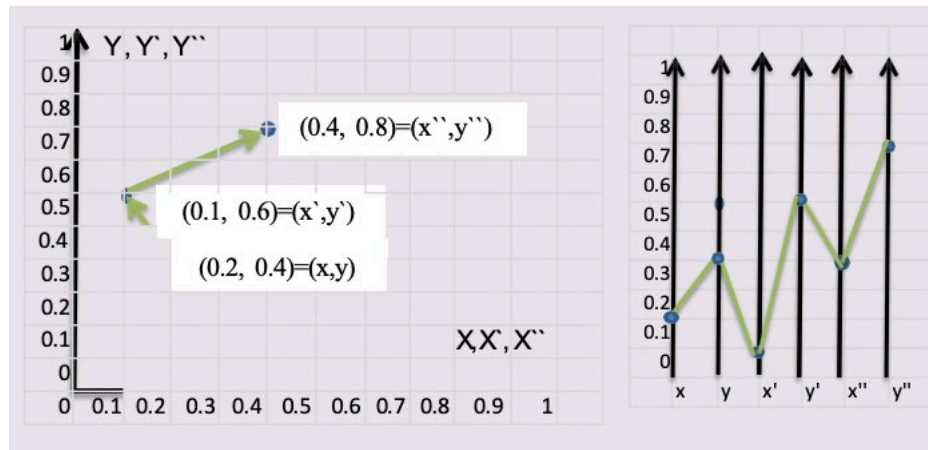


(a) Collocated Paired Coordinates



(b) Shifted Paired Coordinates.

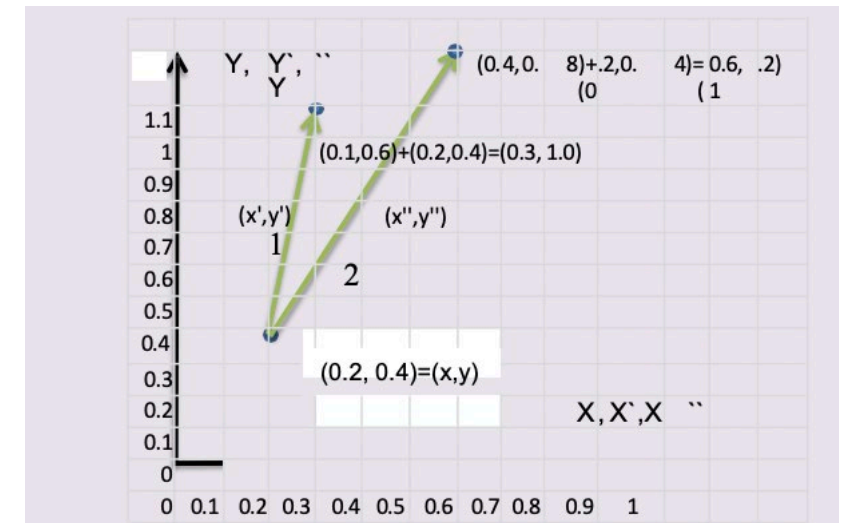
6-D point $(5,4,0,6,4,10)$ in Paired Coordinates.



(a) Collocated Paired Coordinates

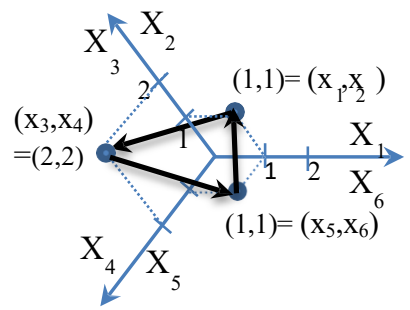
(b) Parallel Coordinates

State vector $\mathbf{x} = (x, y, x', y', x'', y'') = (0.2, 0.4, 0.1, 0.6, 0.4, 0.8)$ in Collocated Paired and Parallel Coordinates.

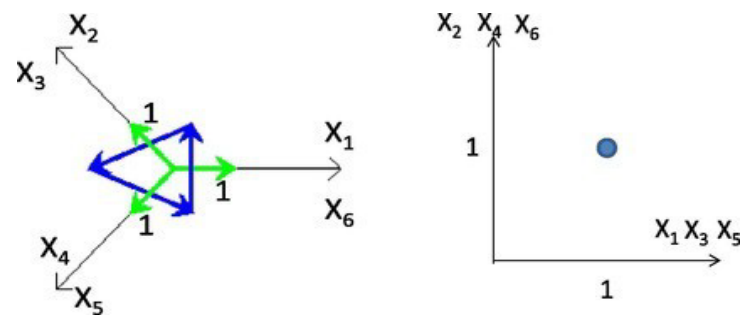


6-D point $\mathbf{x} = (x, y, x', y', x'', y'') = (0.2, 0.4, 0.1, 0.6, 0.4, 0.8)$ in Anchored Paired Coordinates with numbered arrows.

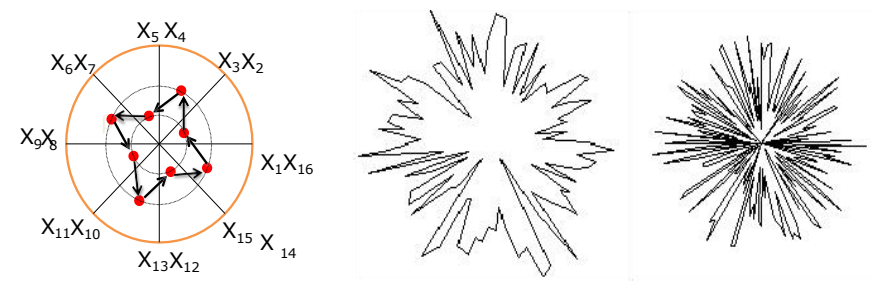
Reversible lossless Paired Coordinates



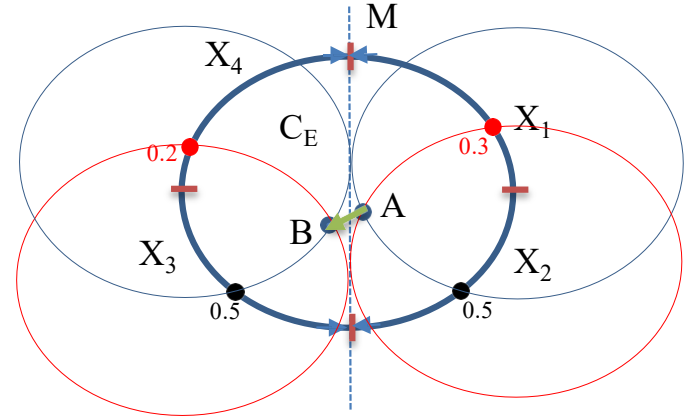
6-D point as a closed contour in 2-D where a 6-D point $x=(1,1,2,2,1,1)$ is forming a triangle from the edges of the graph in Paired Radial Coordinates with non-orthogonal Cartesian mapping.



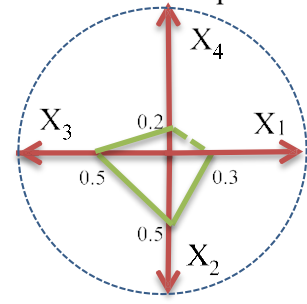
6-D point $(1, 1, 1, 1, 1, 1)$ in two X_1 - X_6 coordinate systems (left – in Radial Collocated Coordinates, right – in Cartesian Collocated Coordinates).



(a) 16-D point $(1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2)$ in Partially Collocated Radial Coordinates with Cartesian encoding, (b) CPC star of a 192-D point in Polar encoding, (c) the same 192-D point as a traditional star in Polar encoding.

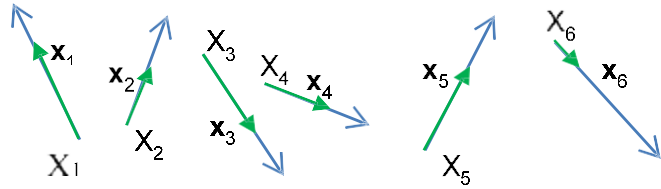


4-D point $P=(0.3,0.5,0.5,0.2)$ in 4-D Elliptic Paired Coordinates, EPC-H as a green arrow. Red marks separate coordinates in the Coordinate ellipse.

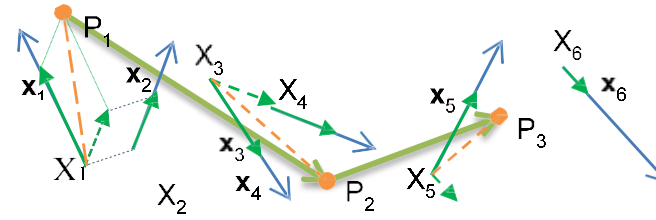


4-D point $P=(0.3,0.5,0.5,0.2)$ in Radial Coordinates.

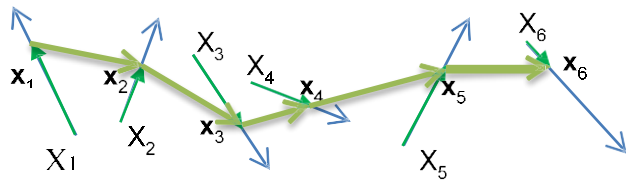
Graph construction algorithms in GLC



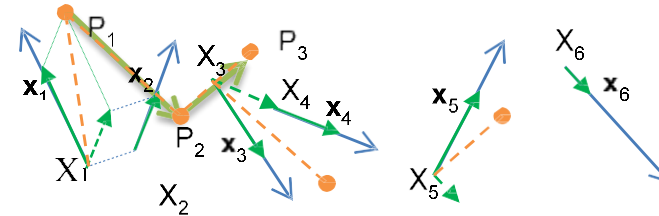
Six coordinates and six vectors that represent a 6-D data point $(0.75, 0.5, 0.7, 0.6, 0.7, 0.3)$



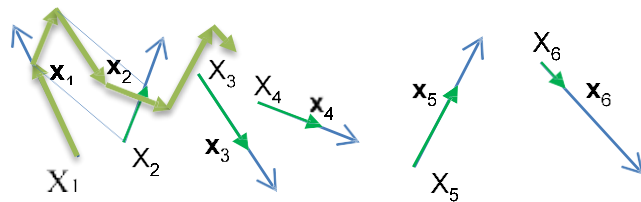
6-D data point $(0.75, 0.5, 0.7, 0.6, 0.7, 0.3)$ in GLC-CC1



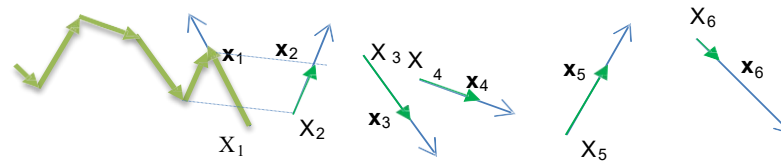
6-D data point $(0.75, 0.5, 0.7, 0.6, 0.7, 0.3)$ in GLC-PC.



6-D data point $(0.75, 0.5, 0.7, 0.6, 0.7, 0.3)$ in GLC-CC2



6-D data point $(0.75, 0.5, 0.7, 0.6, 0.7, 0.3)$ in GLC-SC1.



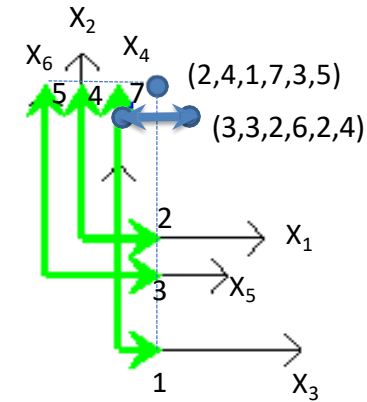
6-D data point $(0.75, 0.5, 0.7, 0.6, 0.7, 0.3)$ in GLC-SC2

Math, theory and pattern simplification methodology: Statements

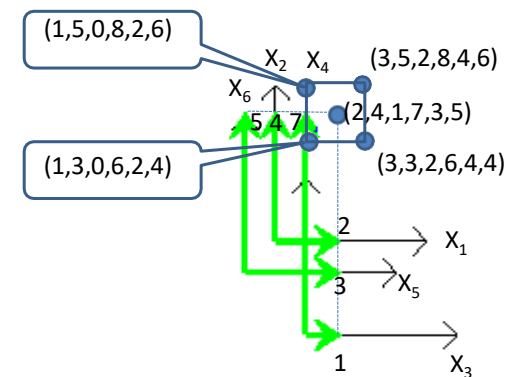
- Statement 1. Parallel Coordinates, CPC and SPC preserve L^p distances for $p=1$ and $p=2$, $D(x,y) = D^*(x^*,y^*)$.
- Statement 2 (n points lossless representation). If all coordinates X_i do not overlap then GLC-PC algorithm provides bijective 1:1 mapping of any n -D point x to 2-D directed graph x^* .
- Statement 3 (n points lossless representation). If all coordinates X_i do not overlap then GLC-PC and GLC-SC1 algorithms provide bijective 1:1 mapping of any n -D point x to 2-D directed graph x^* .
- Statement 4 ($n/2$ points lossless representation). If coordinates X_i , and X_{i+1} are not collinear in each pair (X_i, X_{i+1}) then GLC-CC1 algorithm provides bijective 1:1 mapping of any n -D point x to 2-D directed graph x^* with $\lceil n/2 \rceil$ nodes and $\lceil n/2 \rceil - 1$ edges.
- Statement 5 ($n/2$ points lossless representation). If coordinates X_i , and X_{i+1} are not collinear in each pair (X_i, X_{i+1}) then GLC-CC2 algorithm provides bijective 1:1 mapping of any n -D point x to 2-D directed graph x^* with $\lceil n/2 \rceil$ nodes and $\lceil n/2 \rceil - 1$ edges.

Math, theory and pattern simplification methodology: Statements

- Statement 6 (n points lossless representation). If all coordinates X_i do not overlap then GLC-SC2 algorithm provides bijective 1:1 mapping of any n -D point x to 2-D directed graph x^* .
- Statement 7. GLC-CC1 preserves L^p distances for $p=1$, $D(x,y) = D^*(x^*,y^*)$.
- Statement 8. In the coordinate system X_1, X_2, \dots, X_n constructed by the Single Point algorithm with the given base n -D point $x=(x_1, x_2, \dots, x_n)$ and the anchor 2-D point A , the n -D point x is mapped one-to-one to a single 2-D point A by GLC-CC algorithm.
- Statement 9 (locality statement). All graphs that represent nodes N of n -D hypercube H are within square S

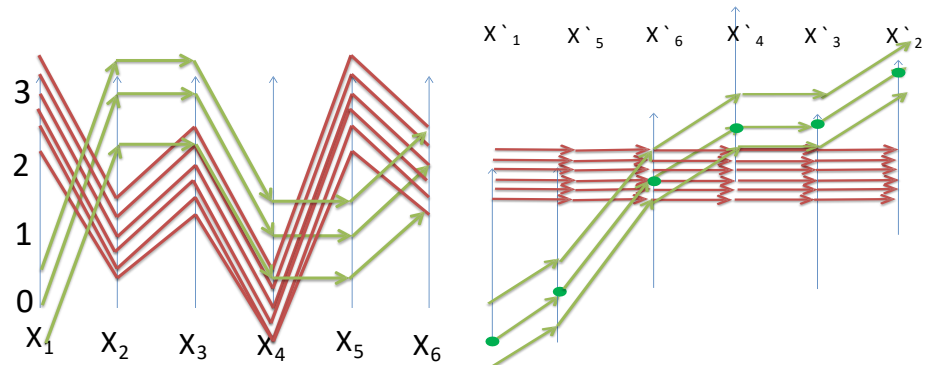


6-D points $(3,3,2,6,2,4)$ and $(2,4,1,7,3,5)$ in X_1 - X_6 coordinate system build using point $(2,4,1,7,3,5)$ as an anchor.



Data in Parameterized Shifted Paired Coordinates. Blue dots are corners of the square S that contains all graphs of all n -D points of hypercube H for 6-D base point $(2,4,1,7,3,5)$ with distance 1 from this base point.

Adjustable GLCs for decreasing occlusion and pattern simplification

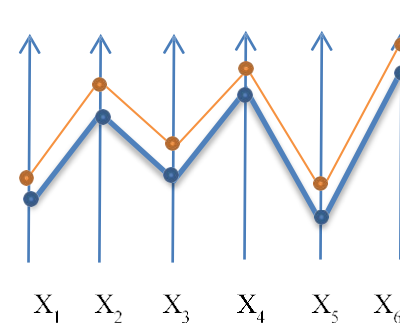


(a) Original visual representation of the two classes in the Parallel Coordinates,

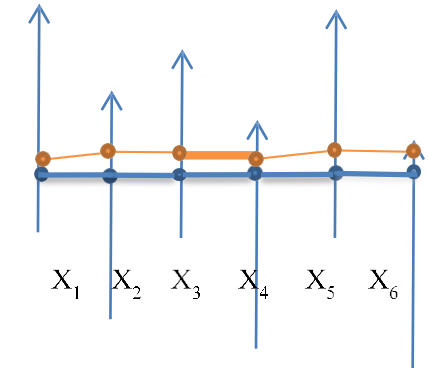
(b) Simplified visual representation after the shifting and reordering of the Parallel Coordinates.

Simplification of the visual representation by the shifting and reordering of the Parallel Coordinates

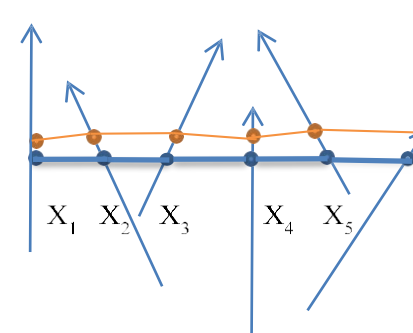
Non-preattentive vs. preattentive visualizations (linearized patterns): 6-D point A = (3, 6, 4, 8, 2, 9) in blue, and 6-D point B = (3.5, 6.8, 4.8, 8.5, 2.8, 9.8) in orange in Traditional, Shifted Parallel Coordinates, and GLC



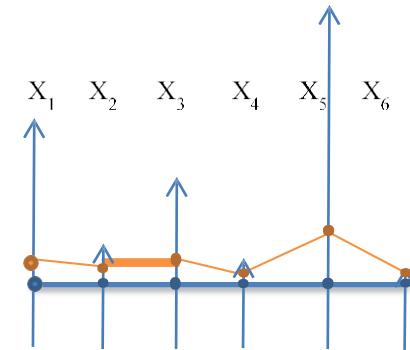
(a) Data in Parallel Coordinates – non-preattentive representation.



(b) Data in the Shifted Parallel Coordinates - preattentive representation.

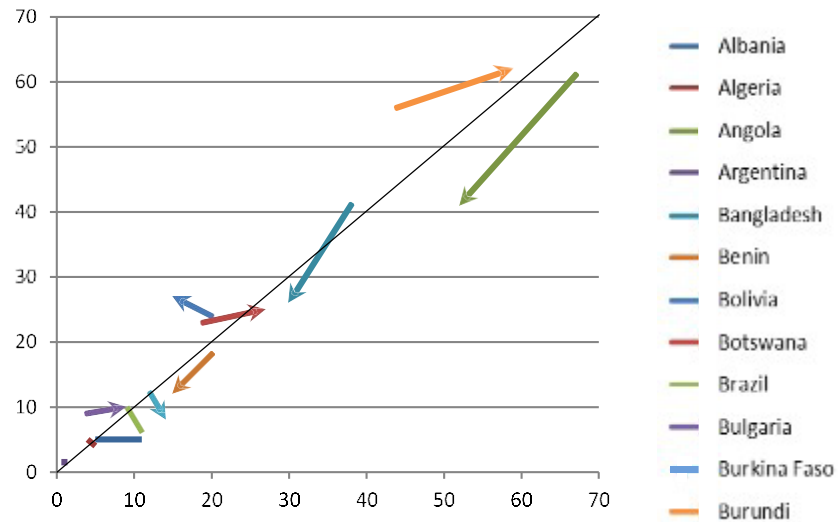


(c) Data in the Shifted General Line Coordinates - preattentive representation.

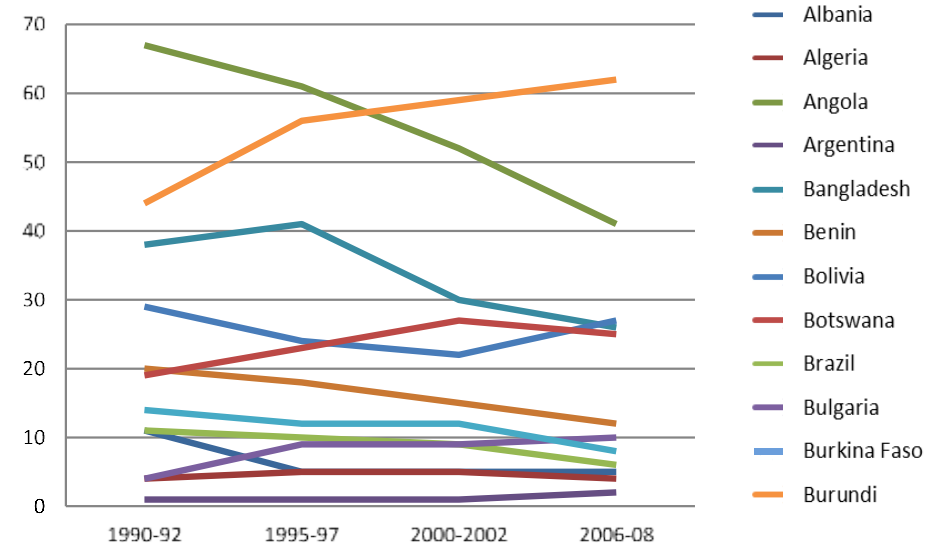


(d) Data in the scaled Parallel Coordinates – preattentive representation

Case Studies: World Hunger data



4-D data: representation of prevalence of undernourished in the population (%) in Collocated Paired Coordinates



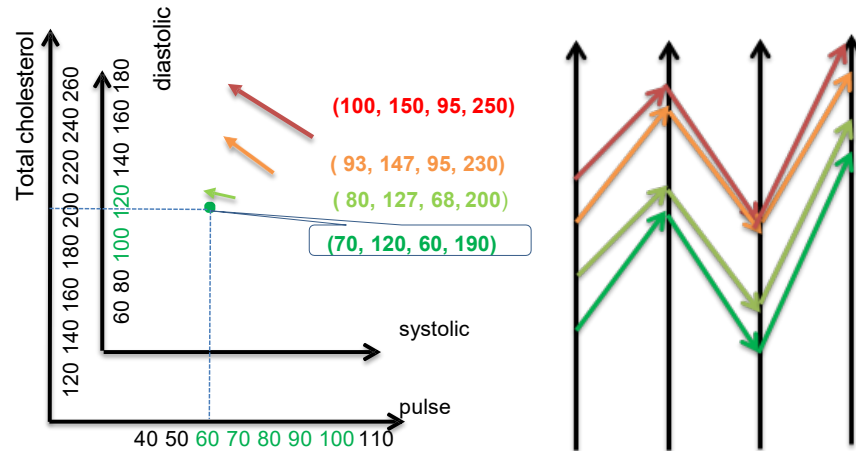
4-D data: representation of prevalence of undernourished in the population (%) in traditional time series (equivalent to Parallel Coordinates for time series)

The Global Hunger Index (GHI) for each country measures as,

$$\text{GHI} = (\text{UNN} + \text{UW5} + \text{MR5}) / 3,$$

where UNN is the proportion of the population that is Undernourished (in %), UW5 is the prevalence of Underweight in children under age of five (in %), and MR5 is the Mortality rate of Children under age five (in %).

Case Studies: Health Monitoring with PC and CPC



a) PSPC: The green dot is the desired goal state, the red arrow is the initial state, the orange arrow is the health state at the next monitoring time, and the light green arrow is the current health state of the person.

4-D Health monitoring visualization in PSPC (a) and Parallel Coordinates (b) with parameters: systolic blood pressure, diastolic blood pressure, pulse, and total cholesterol at four time moments.

(b) the same data as in (a) in Parallel Coordinates.

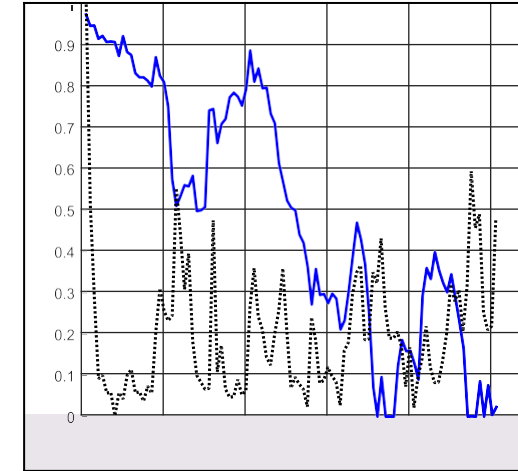
The colors show the progress to the goal.

- Dark green dot – goal.
- Yellow and light green – closer to the goal point.
- Red arrow – initial health status.

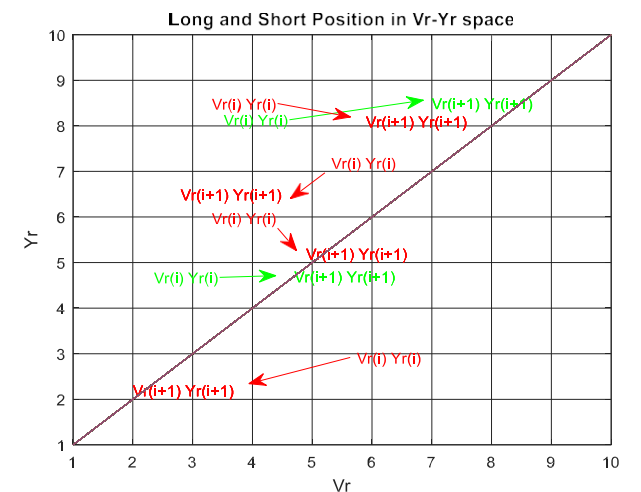
- Experiments – people quickly grasp how to use this health monitor.
- This health monitor is expandable.
- Two more indicators is another pair of shifted Cartesian Coordinates.
 - The goal is the same dark green 2-D dot
 - Each graph has two connected arrows.
- Graphs closer to the goal are smaller.

Case studies: Knowledge Discovery and Machine Learning for Investment Strategy with CPC

- The CPC visualization shows arrows in (V_r, Y_r) space of volume V_r and relative main outcome variable Y_r
- This is a part of the data shown as traditional time series with *time* axis.
- CPC has no time axis. The arrow direction shows time.
- The arrow beginning is the point in the space $(V_{r,i}, Y_{r,i})$, and its head is the next time point in the collocated space $(V_{r,i+1}, Y_{r,i+1})$.
- CPC give the inspiration idea for building a trading strategy in contrast with time series figure without it.
 - It allows finding the areas with clusters of two kinds of arrows.
 - The arrows for the long positions are green arrows.
 - The arrows for the short positions, are red.
 - Along the Y_r axis we can observe a type of change in Y in the current candle. if $Y_{r,i+1} > Y_{r,i}$ then $Y_{i+1} > Y_i$ the right decision in i -point is a long position opening. Otherwise, it is a short position.
 - Next, CPC shows the effectiveness a decision in the positions.
 - The very horizontal arrows indicates small profit
 - A more vertical arrows indicates the larger profit.
- In comparison with traditional time series, the CPC bring the additional knowledge about the potential of profit in selected area of parameters in (V_r, Y_r) space.



Comparison of two time series: relative outcome Y_r and relative volume in every one hundred period.



Some examples of arrows which show points of

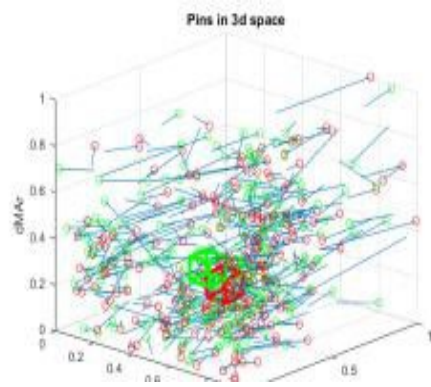


Figure 8.16. Pins in 3-D space: two cubes found in (Y_r, dMA_r, V_r) space with the maximum asymmetry between long and short positions.

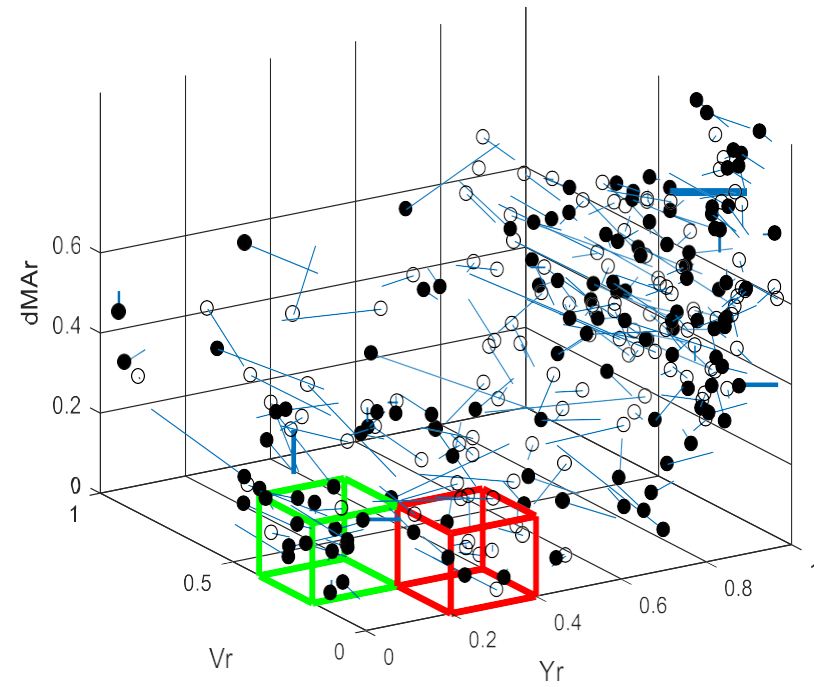


Figure 8.18. Two determined cubes in $Y_r-dMA_r-V_r$ space with the maximum asymmetry between long and short positions for the new grid resolution.

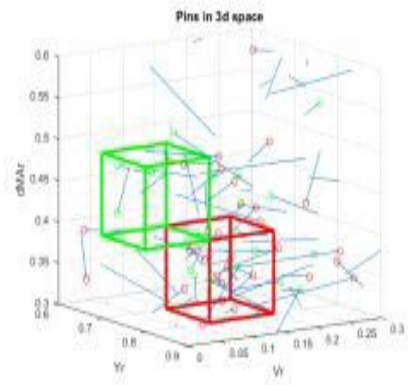
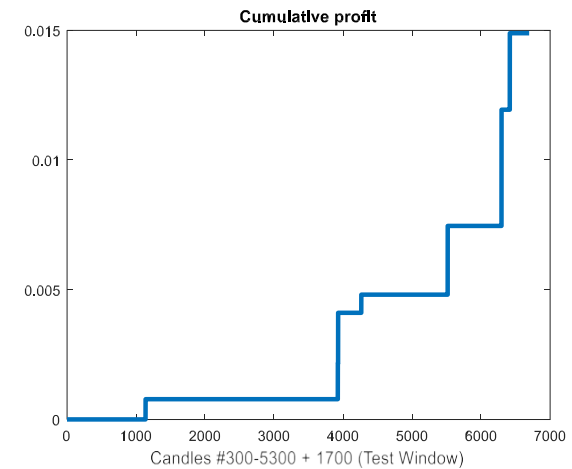
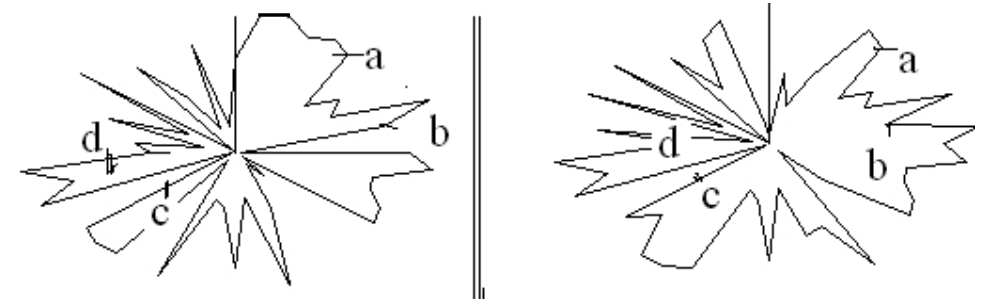
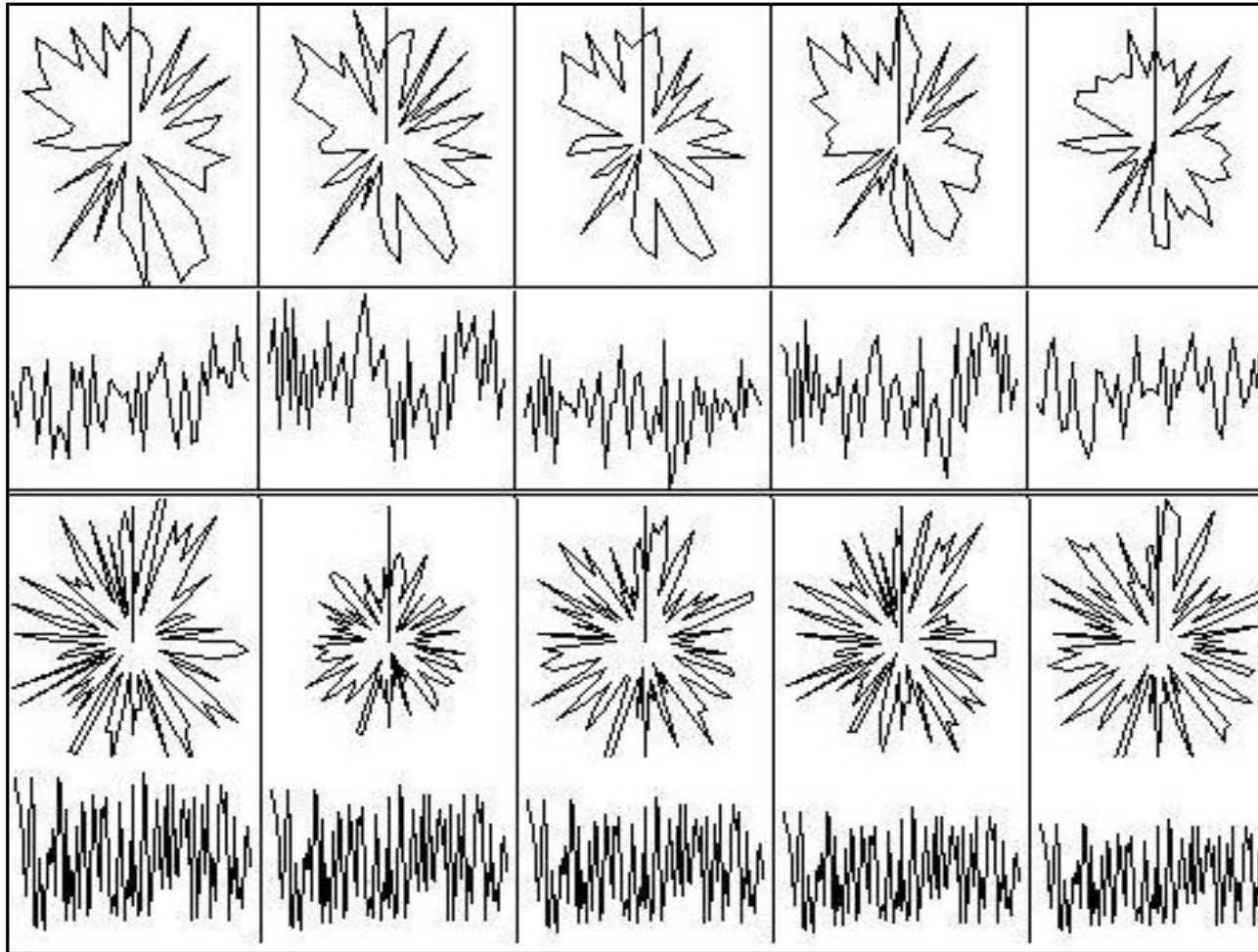


Figure 8.17. The zoomed cubes with the best asymmetry from Figure 8.16. The upper cube with green circles is selected for long positions lower cube with red circles is for short positions. For better visibility, the viewpoint is changed from Figure 8.16.



Lossless Visualization of 48-D and 96-D data



Two stars with identical shape fragments on intervals [a,b] and [d,c] of coordinates.



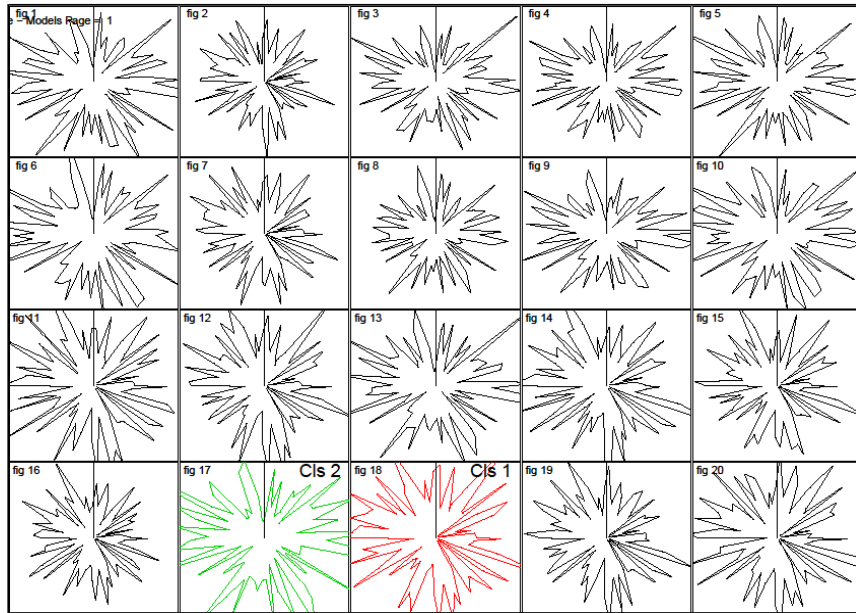
Samples of some class features on Stars for n=48



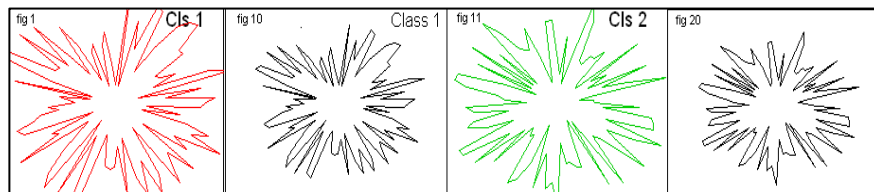
Samples of some class features on PCs for n=48

Visual Patterns-- combinations of attributes

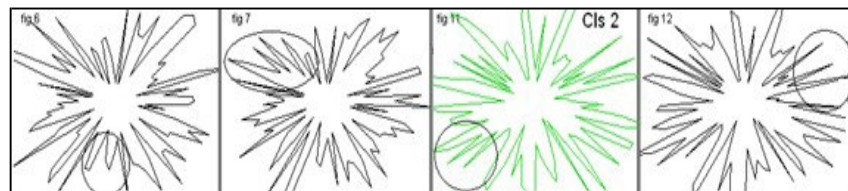
Examples of corresponding figures: stars (row 1) and PCs lines (row 2) for five



Twenty 160-D points of 2 classes represented in star CPC with noise 10% of max value of normalized coordinates (max=1) and with standard deviation 20% of each normalized coordinate.



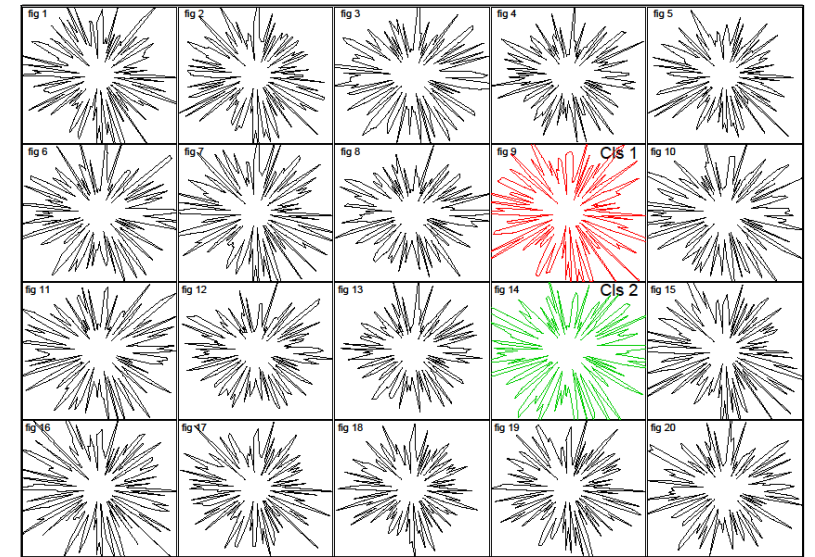
(a) Initial 100-D points without noise for Class (Hyper-tube) #1 and Class (Hyper-tube) #2



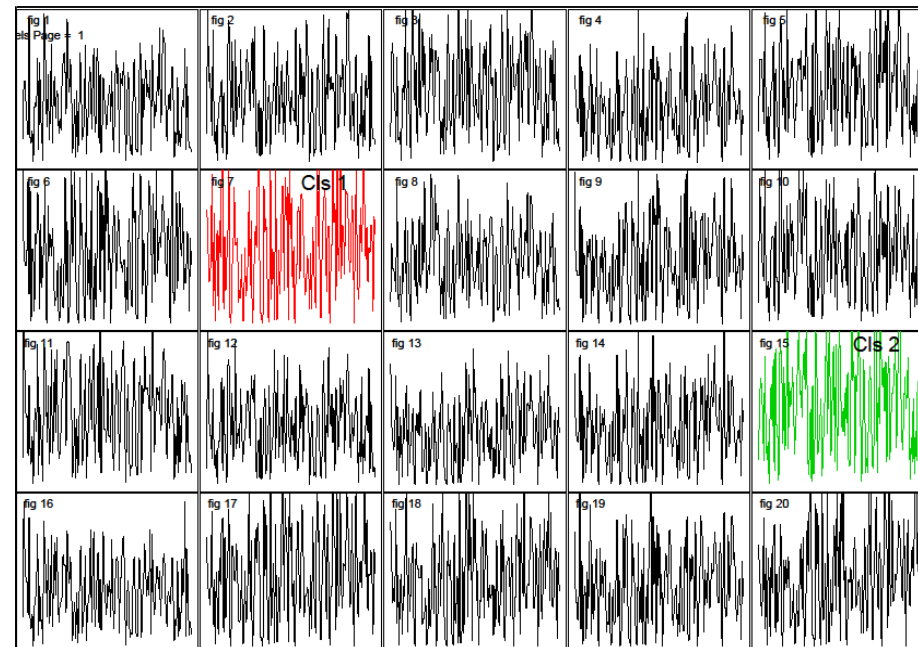
(b) 100-D points with multiplicative noise: circled areas are the same as in upper star.

Figure 6.10. Samples of 100-D data in Star CPC used to make participants familiar with the task.

Discovering high-dimensional interpretable patterns in 160-Dimensions!

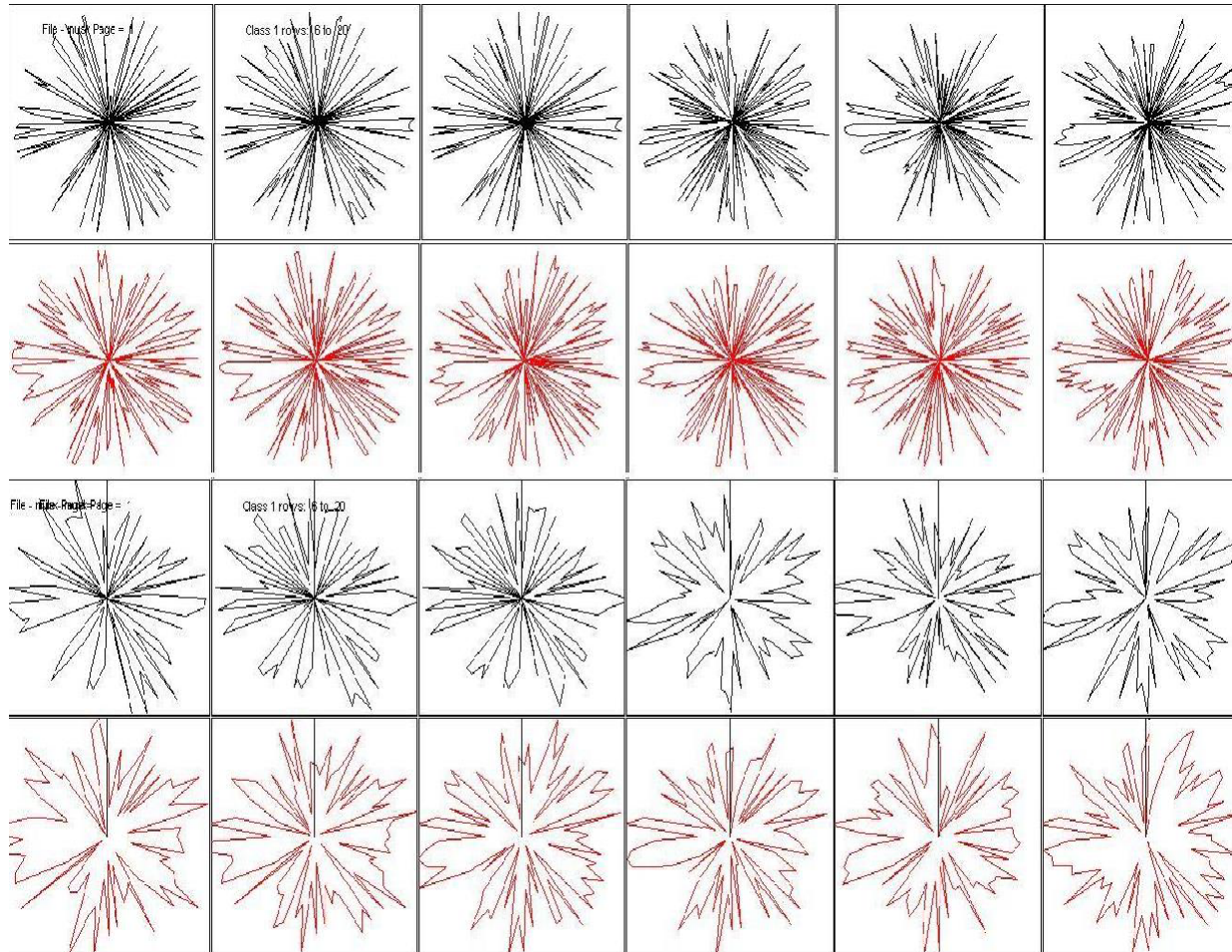


Twenty 160-D points of 2 classes represented in Radial Coordinates with noise 10% of max value of normalized coordinates (max=1) and with standard deviation 20% of each normalized coordinate.



Twenty 160-D points of 2 classes represented in Parallel Coordinates with noise 10% of max value of normalized coordinates (max=1) and with standard deviation 20% of each normalized coordinate

Human abilities to discover patterns in high-D data



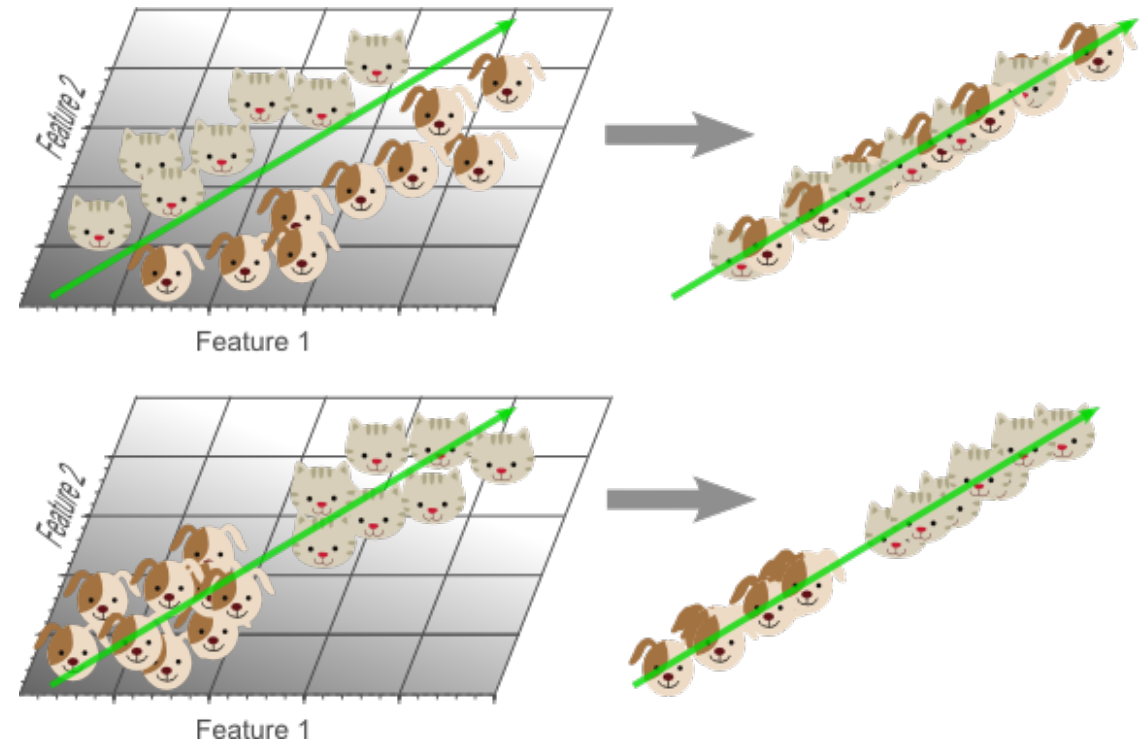
- The expected classifiable dimensions are in $[160, 192]$ interval for the Radial Coordinates
- Due to advantages of Star CPC over Radial Coordinates, these limits must be higher for Star CPC and lower for Parallel Coordinates
- Finding bounds for linear-hyper-tubes most likely will be also limits for non-linear hyper-tubes due to their higher complexity
- Traditional 170-D stars: class “musk” (first row) and class “non-musk chemicals” (second row). CPC 170-D stars from the same dataset: class “musk” (third row) and class “non-musk chemicals” (forth row).



Nine 170-dimensional points of two classes in Parallel Coordinates

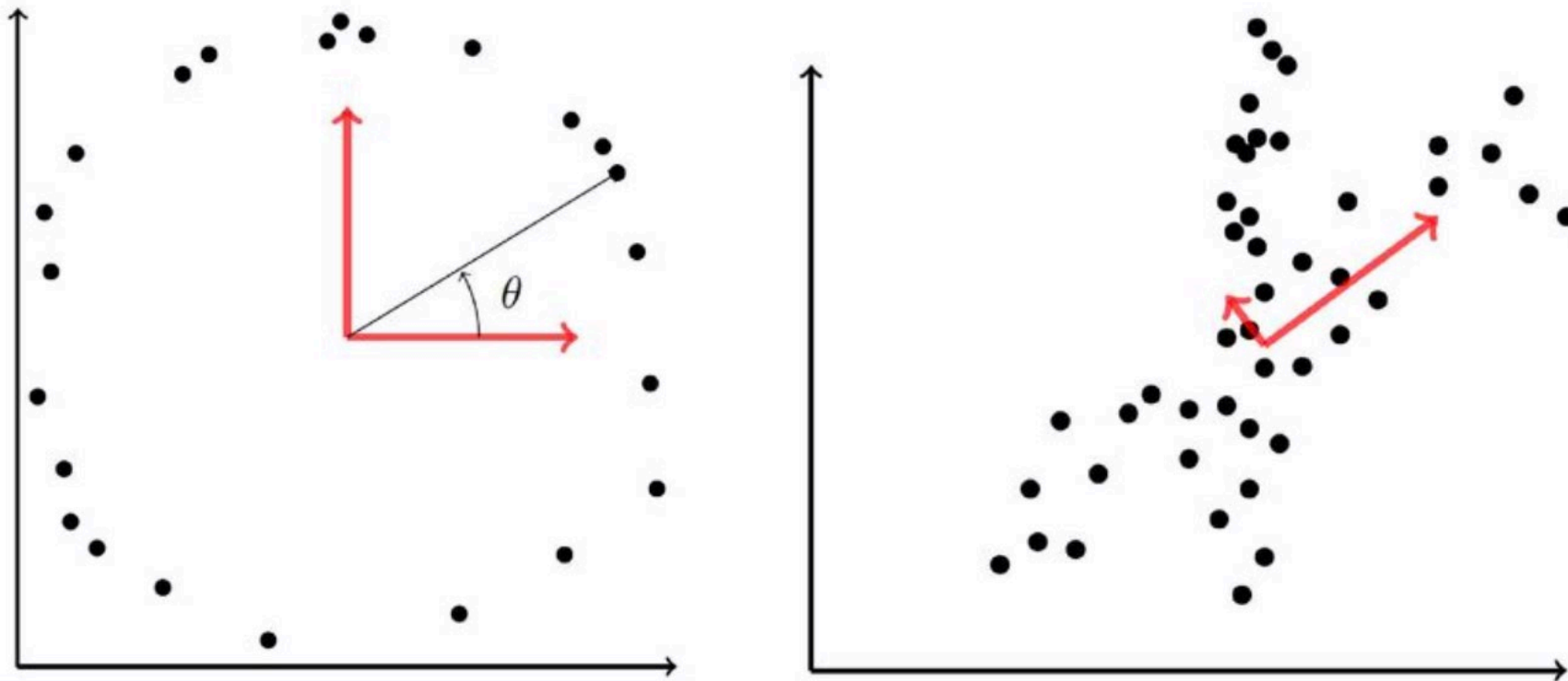
Lossy Methods Revisited: PCA

- Account for variance of data in as few dimensions as possible (using linear projection)
- First PC is the projection direction that maximizes the variance of the projected data
- Second PC is the projection direction that is orthogonal to the first PC and maximizes variance of the projected data



PCA can occasionally fail

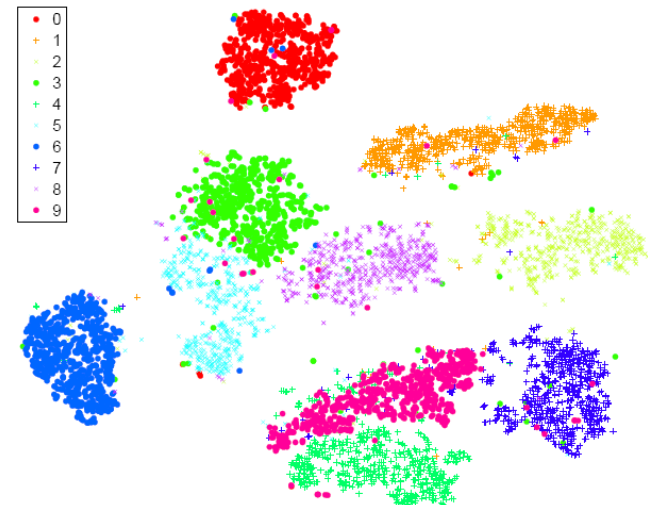
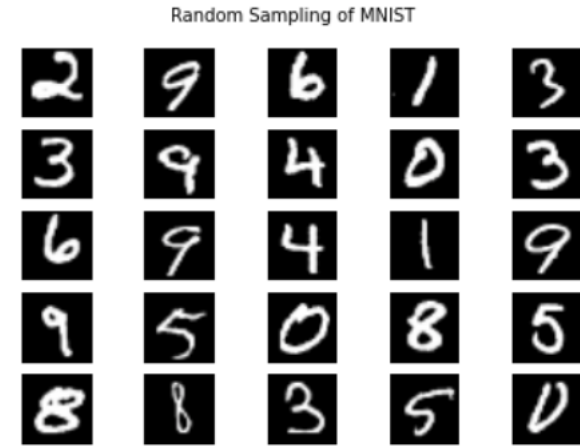
- Angle has the most information but this is not captured by the basis
- Sometimes most important information is not orthogonal



Lossy Methods Revisited: t-SNE

- Visualizes high-dimensional data in a 2- or 3-dimensional map
- Better than most techniques at creating a single map that reveals structure at many different scales
- Particularly good for high-dimensional data that lie on several different, but related, low-dimensional manifolds.

Example: images of objects from multiple classes seen from multiple viewpoints.



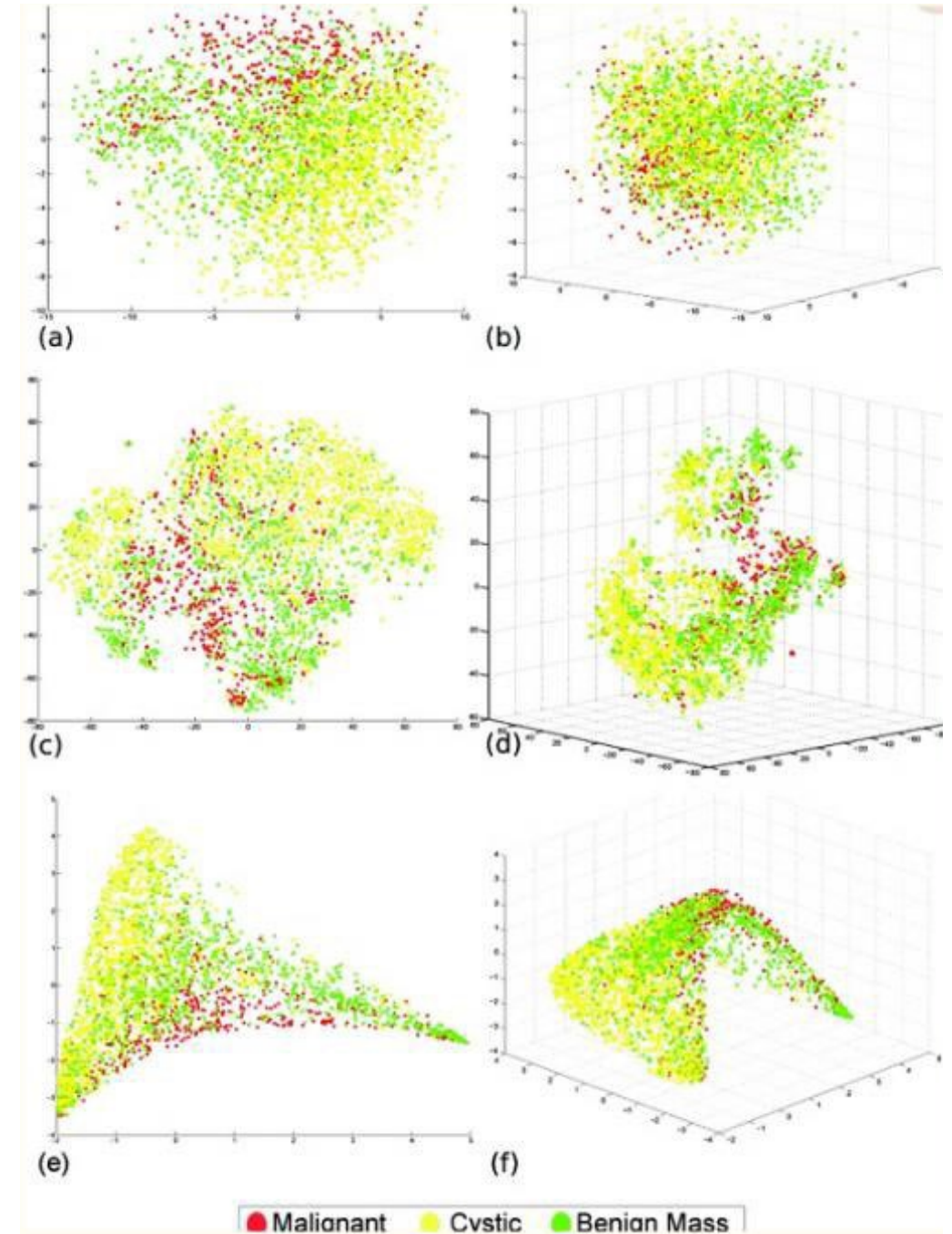
Lossy Visualization

(a) PCA, first two principal components

(b) first three principal components, 3D PCA.

(c) 2D and (d) 3D visualization of the nonlinear reduction mapping using t-SNE

(e) 2D and (f) 3D visualization of the nonlinear mapping using Laplacian eigenmaps

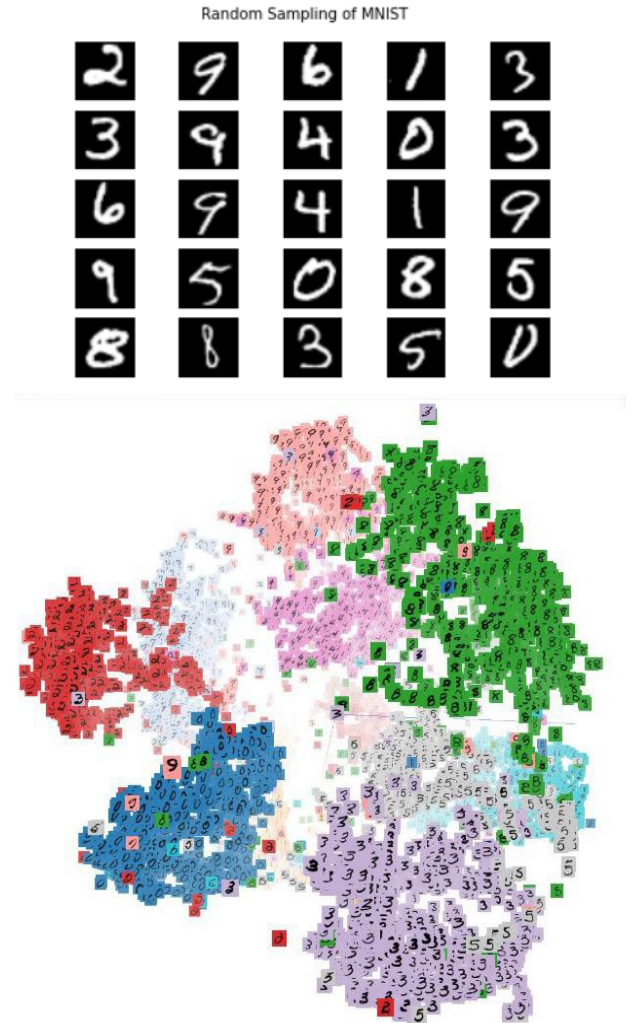


Lossy Visualization in ML Models

- Methods like T-SNE, PCA and others do not preserve all information of initial features (they are lossy visualizations of n -D data)
- They convert n interpretable features to 2-3 artificial features that have no direct interpretation
- General Line Coordinates is an alternative that preserves all n -D information when occlusion/clutter in visualization is suppressed that was successfully done in [Kovalerchuk 2014-2019]




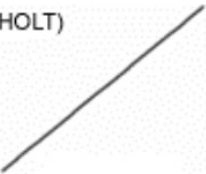
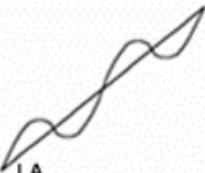




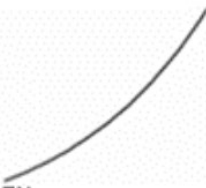
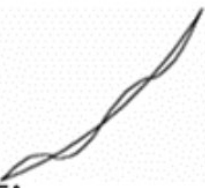
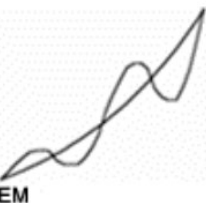
Do PCA and t-SNE *actually reveal* multi-dimensional relationships?

- MNIST is visualized in t-SNE as clusters colored by their associated digit labels so that “those images with high similarity in their original feature space are placed close to each other in the 2D/3D space
- In this manner, “one can easily identify and which digit images are outliers (and thus confusing as another digit)”
- t-SNE author Maaten warned about such statements-
- t-SNE may not assign meaning, to point densities, in clusters. The outlier and dense areas, visible in t-SNE, may not be them, in the original n-D space.
- In addition, the 2-D attributes, generated by t-SNE, do not have direct domain interpretation



Interpreting Time Series via Reversible Methods

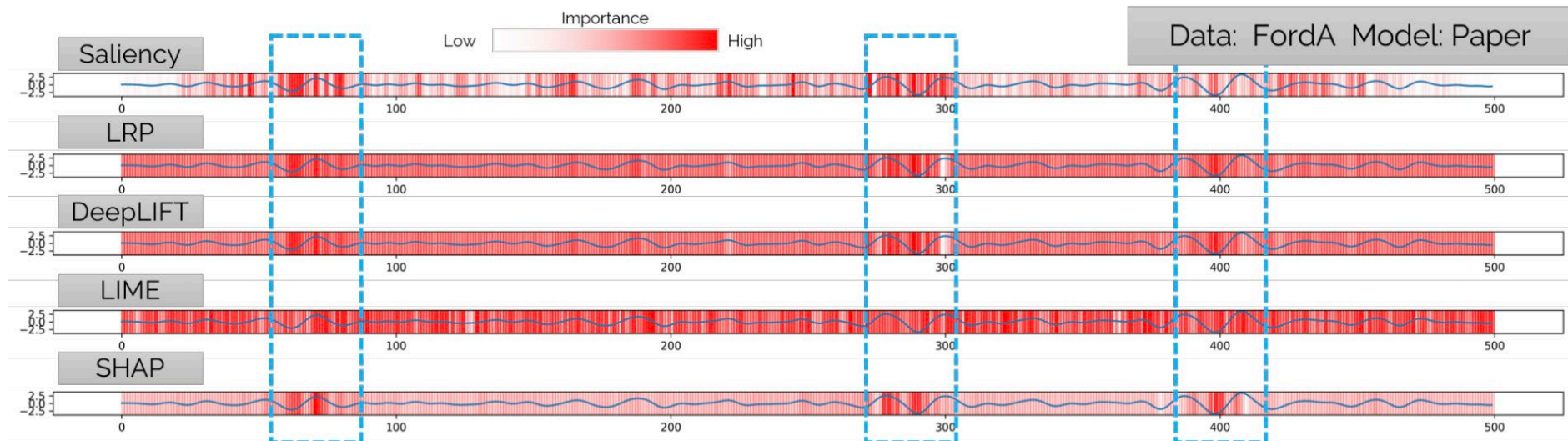
- Trends in retrospective time series data are relatively straightforward to understand
- How do we understand predictions of time series in data?

	Nonseasonal	Additive Seasonal	Multiplicative Seasonal
Constant Level	(SIMPLE)  NN	 NA	 NM
Linear Trend	(HOLT)  LN	 LA	(WINTERS)  LM
Damped Trend (0.95)	 DN	 DA	 DM
Exponential Trend (1.05)	 EN	 EA	 EM

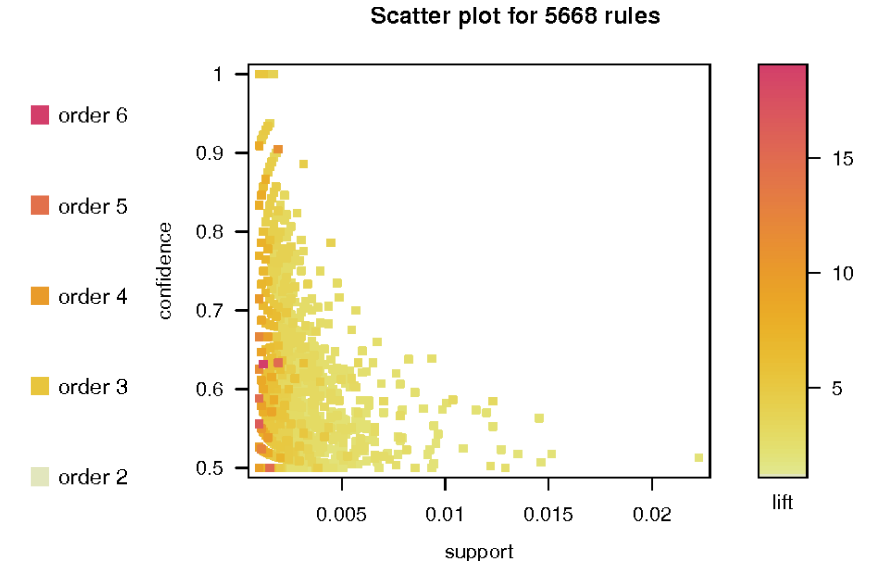
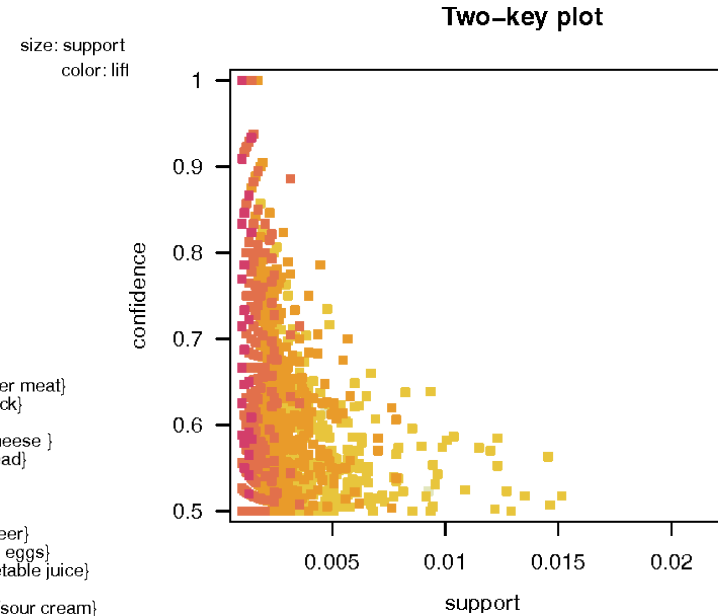
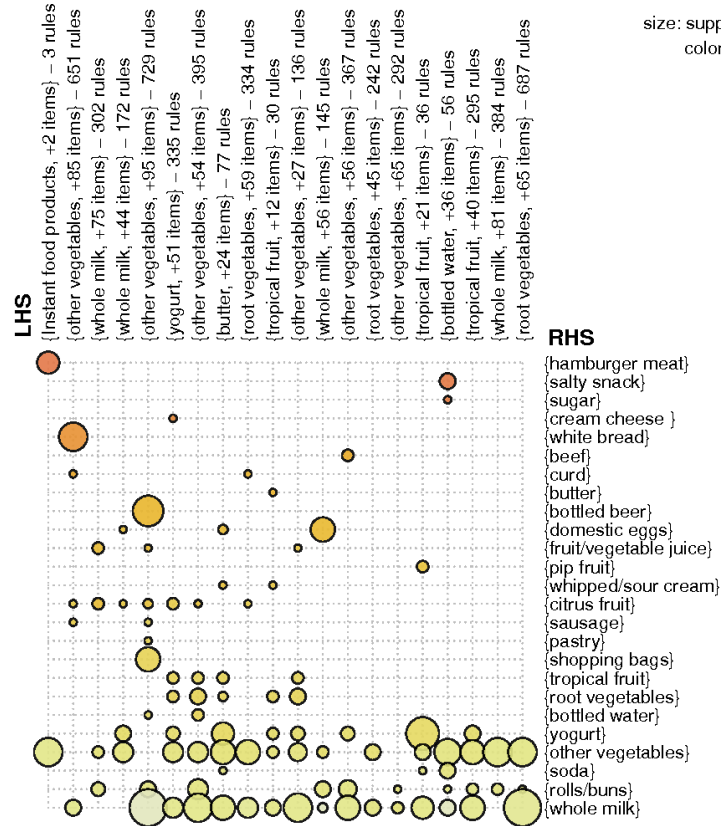
Interpreting Time Series via Reversible Methods

- Similar to the saliency masks on images, a heatmap can be created based on the relevance produced by XAI methods
- Create a visualization with this heatmap enriching a line plot of the original time series. Together with domain knowledge, an expert can inspect the produced explanation visualizations

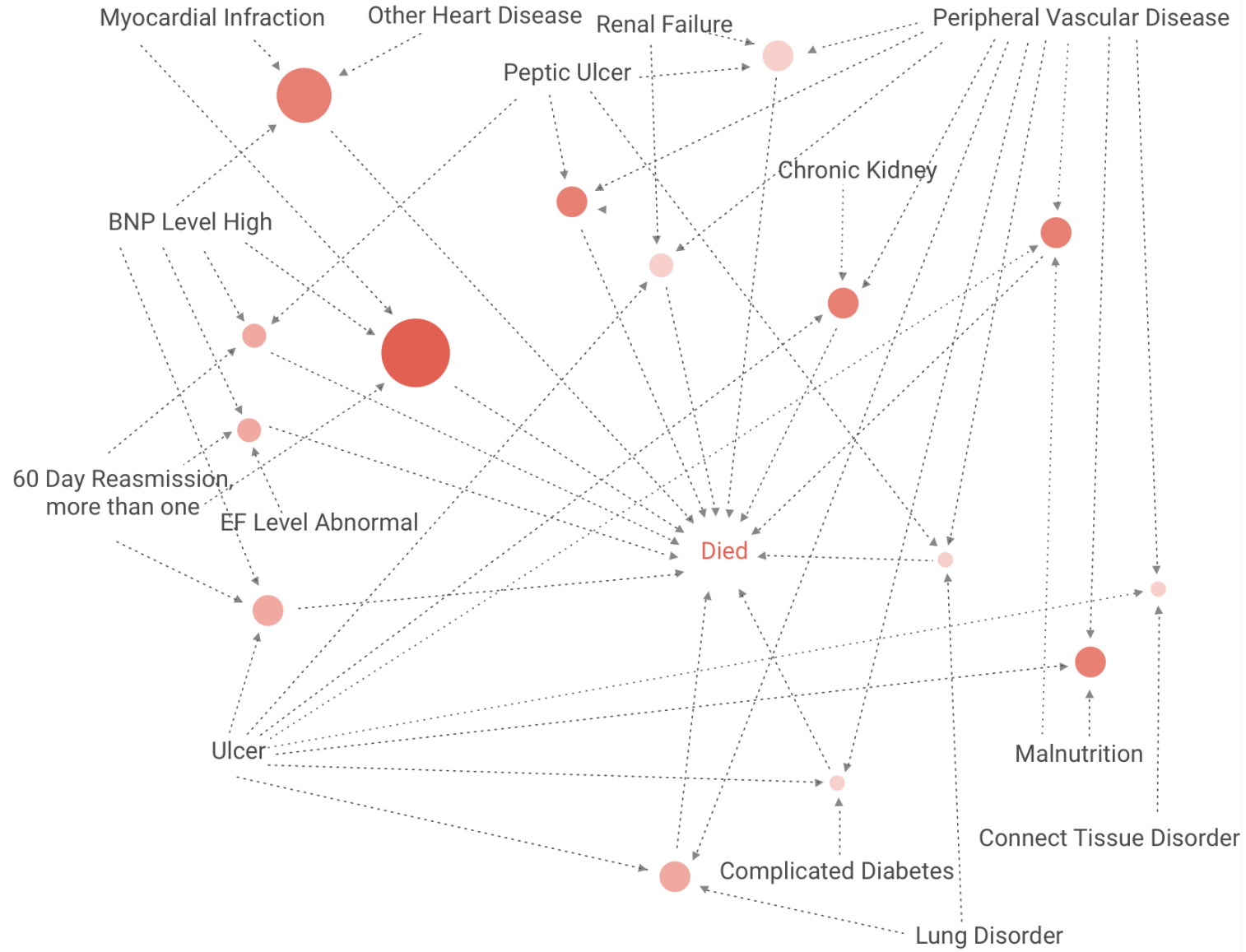
[Schlegel 2019]



Visualizing Association Rules

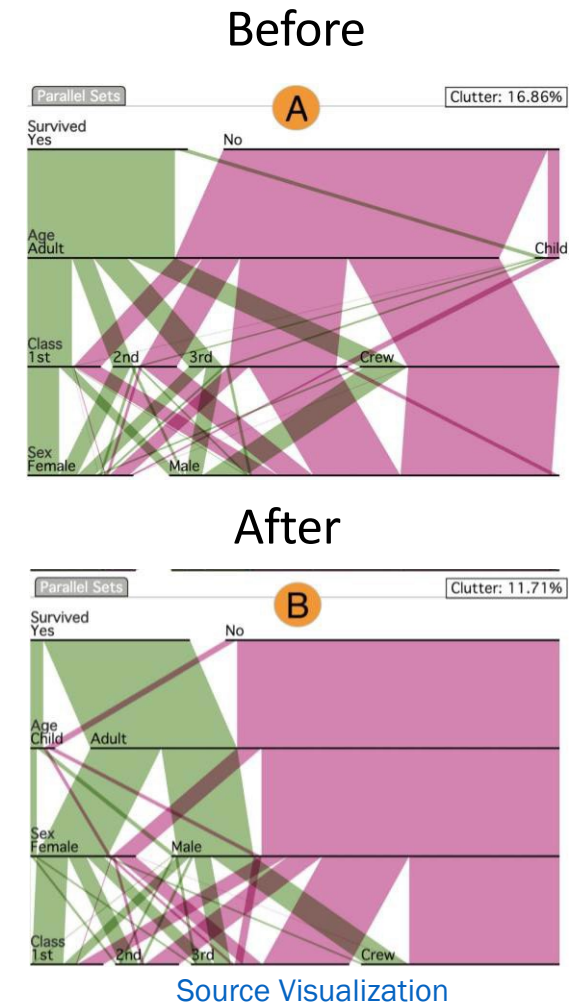


- Matrix with rows as Left Hand Side, LHS itemsets of rules and columns as Right Hand side (RHS) itemsets of rules
- Scalability for many LHS and RHS
- Readability of small cells having many items

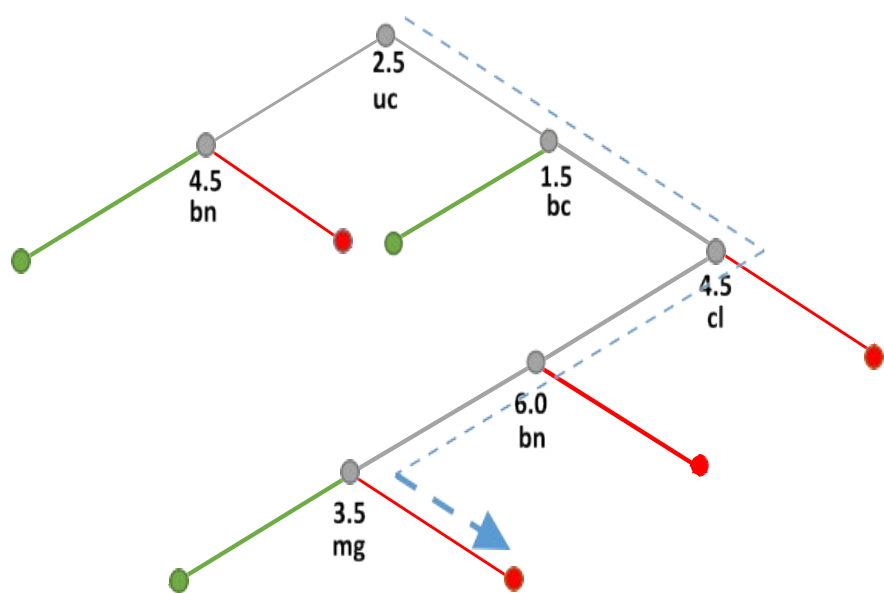


Visualizing ARules using Parallel Sets

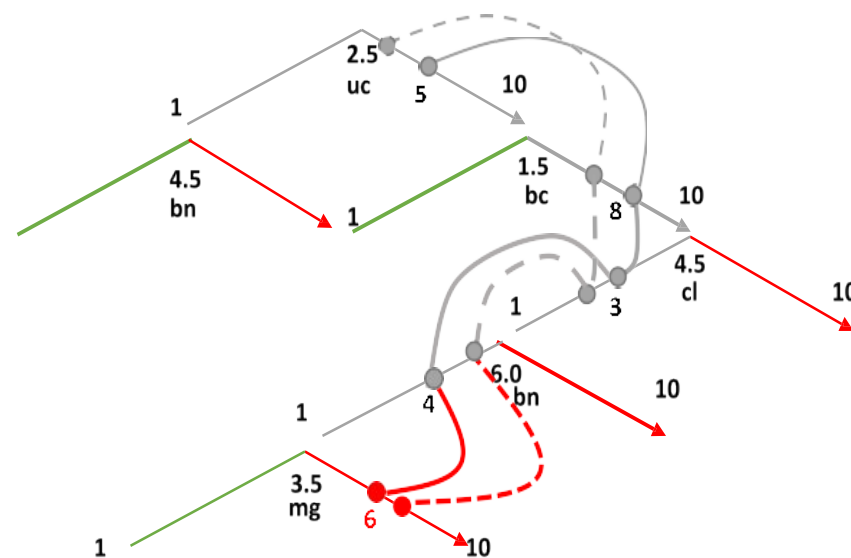
- Approach:
 - Discovering Association Rules.
 - Deleting dimensions irrelevant to AR.
 - Feeding rules to two coordinated rule visualizations (called ARTable and ParSets),
- User interactions
 - Visually explore rules in ARTable
 - Find interesting rules, dimensions, and categories in ARTable
 - Create and optimize the layout of ParSets
 - Validate interesting rules
 - Explore details of rules in ParSets using domain knowledge



Folded Coordinate Decision Tree (FC-DT)



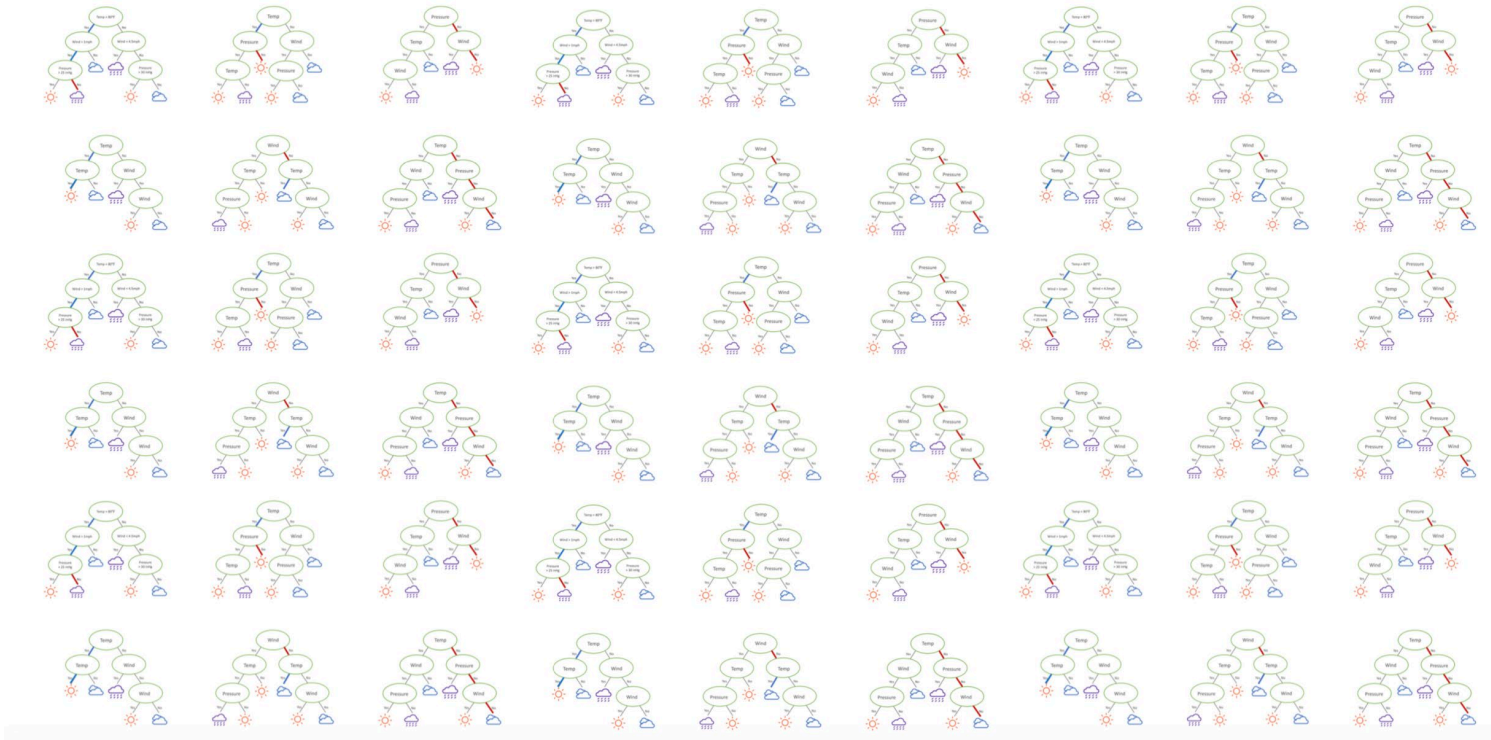
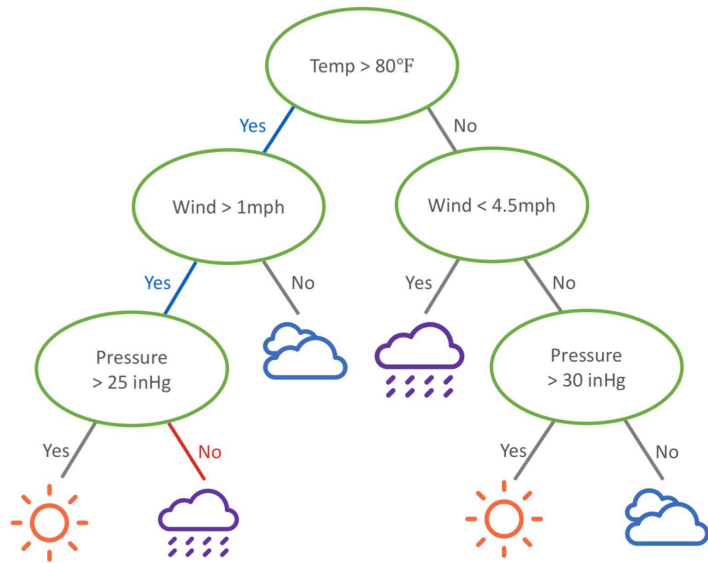
(a) Traditional visualization of WBC data decision tree. Green edges and nodes indicate the benign class and red edges and nodes indicate the malignant class.



(b) DT with edges as Folded Coordinates in disproportional scales. The curved lines are cases that reach the DT malignant edge with different certainties due to the different distances from the threshold node.

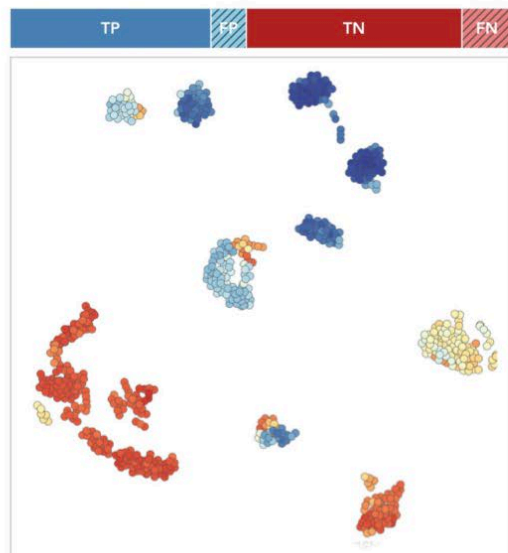
iForest: Interpreting Random Forests via Visual Analytics

Making sense of Random Forests



Load Data: Titanic Upload CSV

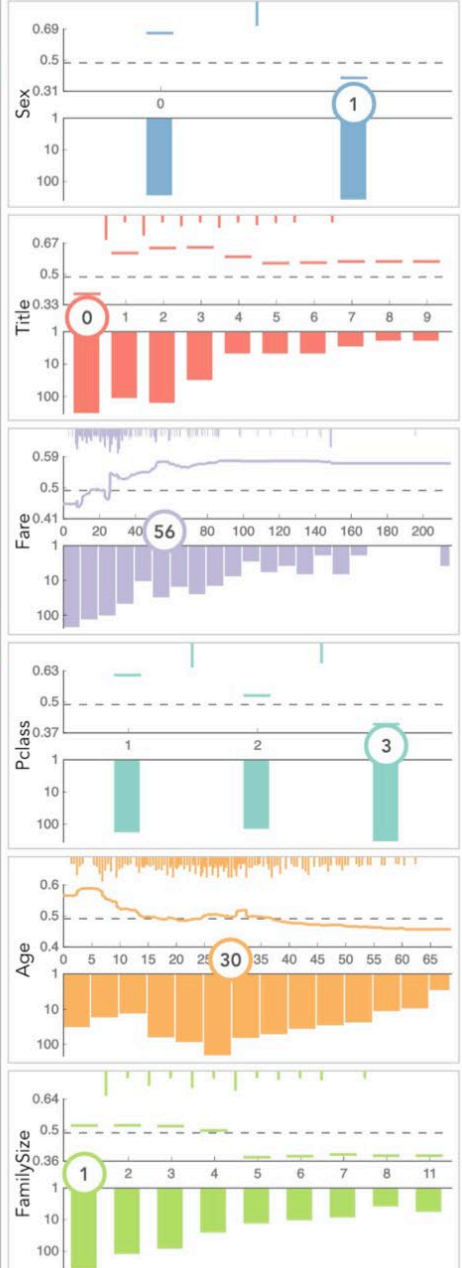
Data Overview



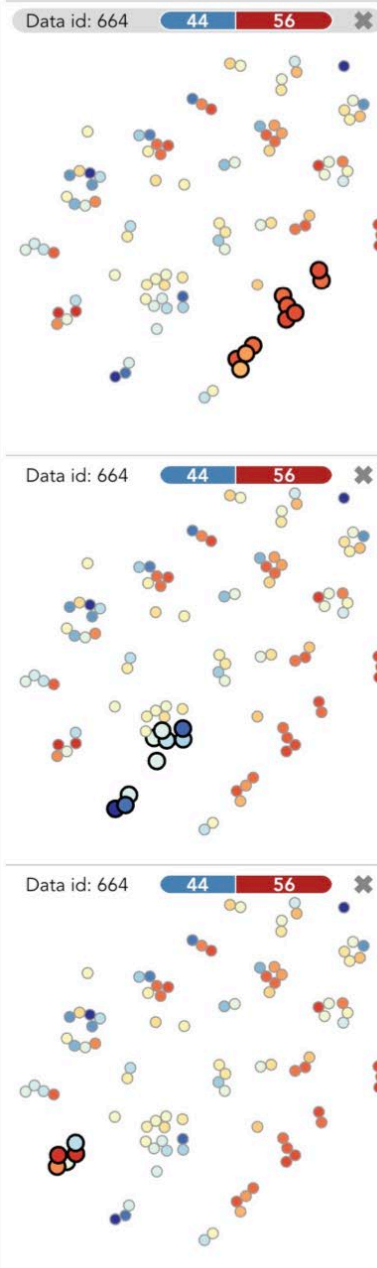
Data ID Search

nid	Sex	Title	Fare	Pclass	Age
0	1	0	15.85	3	32
1	1	0	7.75	3	65
2	1	0	13	2	30
3	1	0	0	3	19
4	1	0	8.05	3	29.70
5	1	0	10.50	2	32
6	0	2	25.47	3	29.70
7	1	0	32.32	1	61
8	1	0	8.16	3	19
9	1	0	27.72	1	29.70
10	1	0	41.58	2	25
11	1	3	25.47	3	29.70
12	1	6	26.55	1	60
13	1	0	13	2	50
14	0	1	151.55	1	25
15	1	0	7.78	3	29.70
16	1	0	7.90	3	29.70
17	1	0	9.48	3	29
18	1	0	7.88	3	29

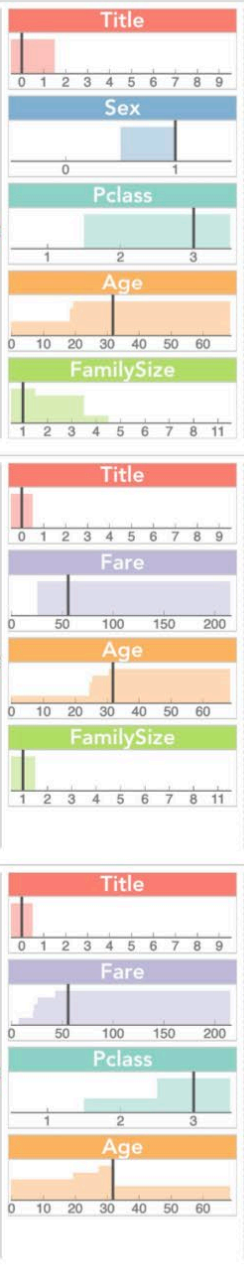
Feature View Local Scale



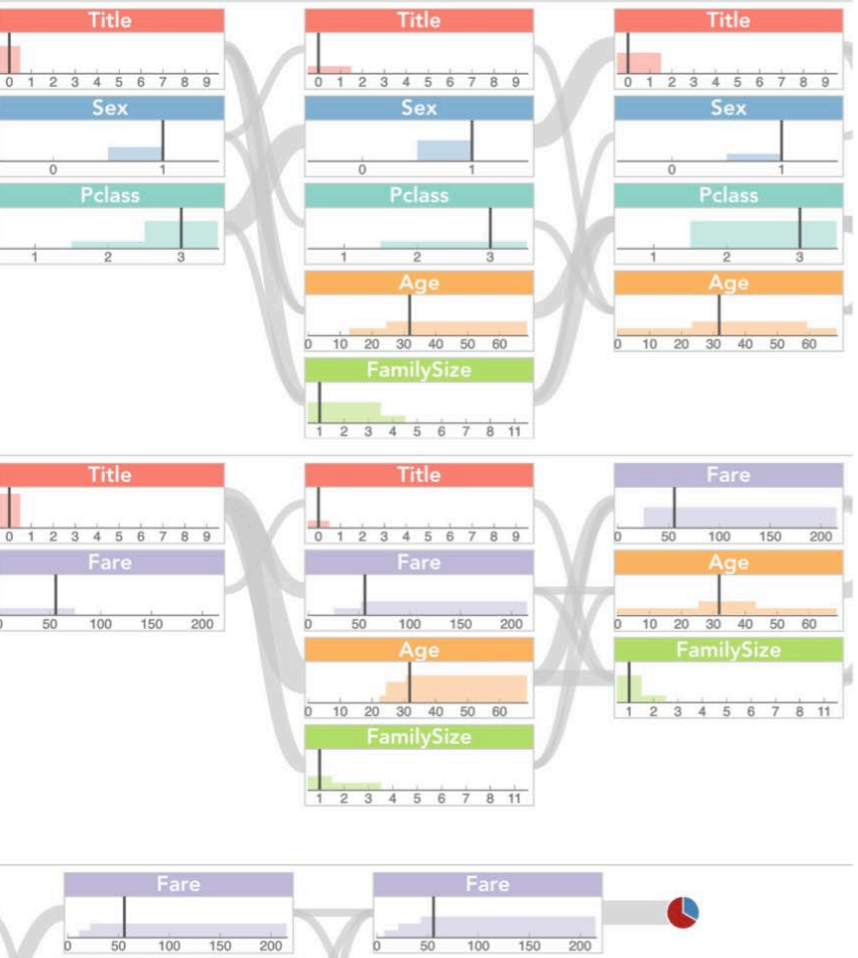
Decision Path View



Feature Summary



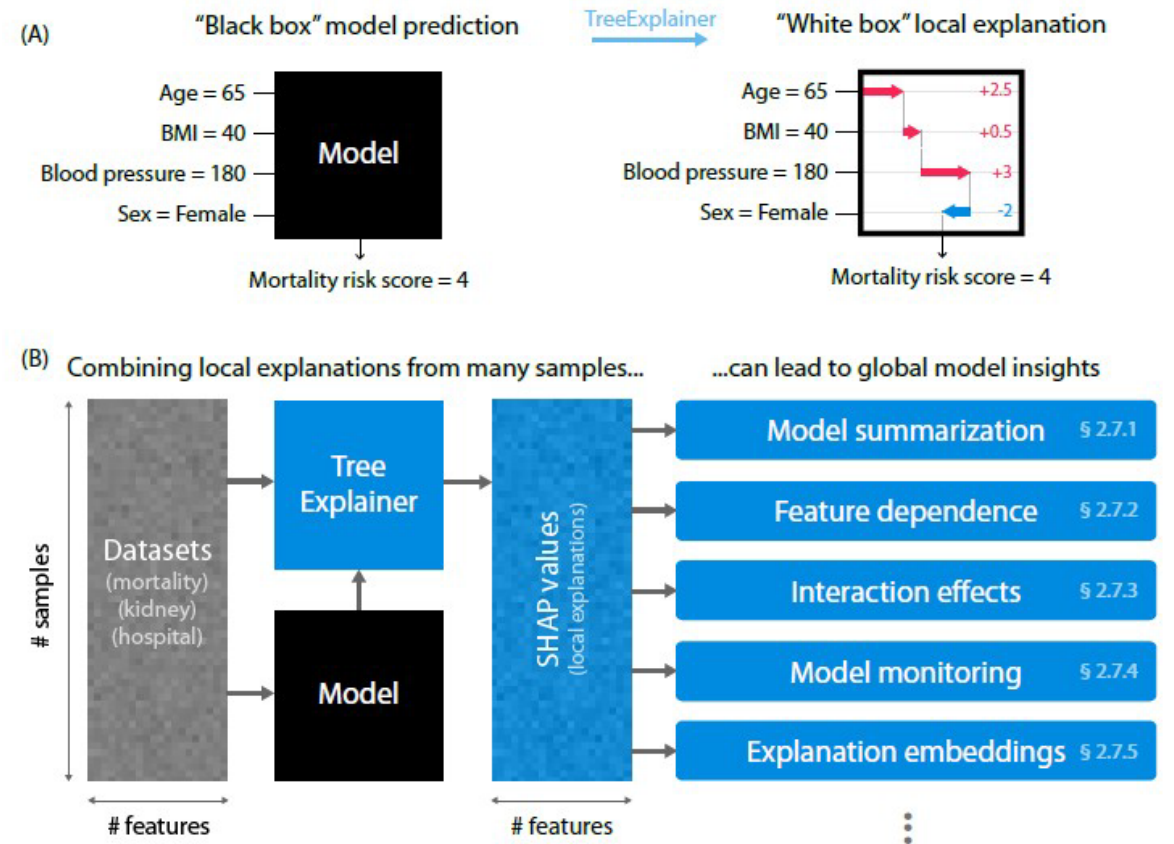
Decision Path Details



iForest

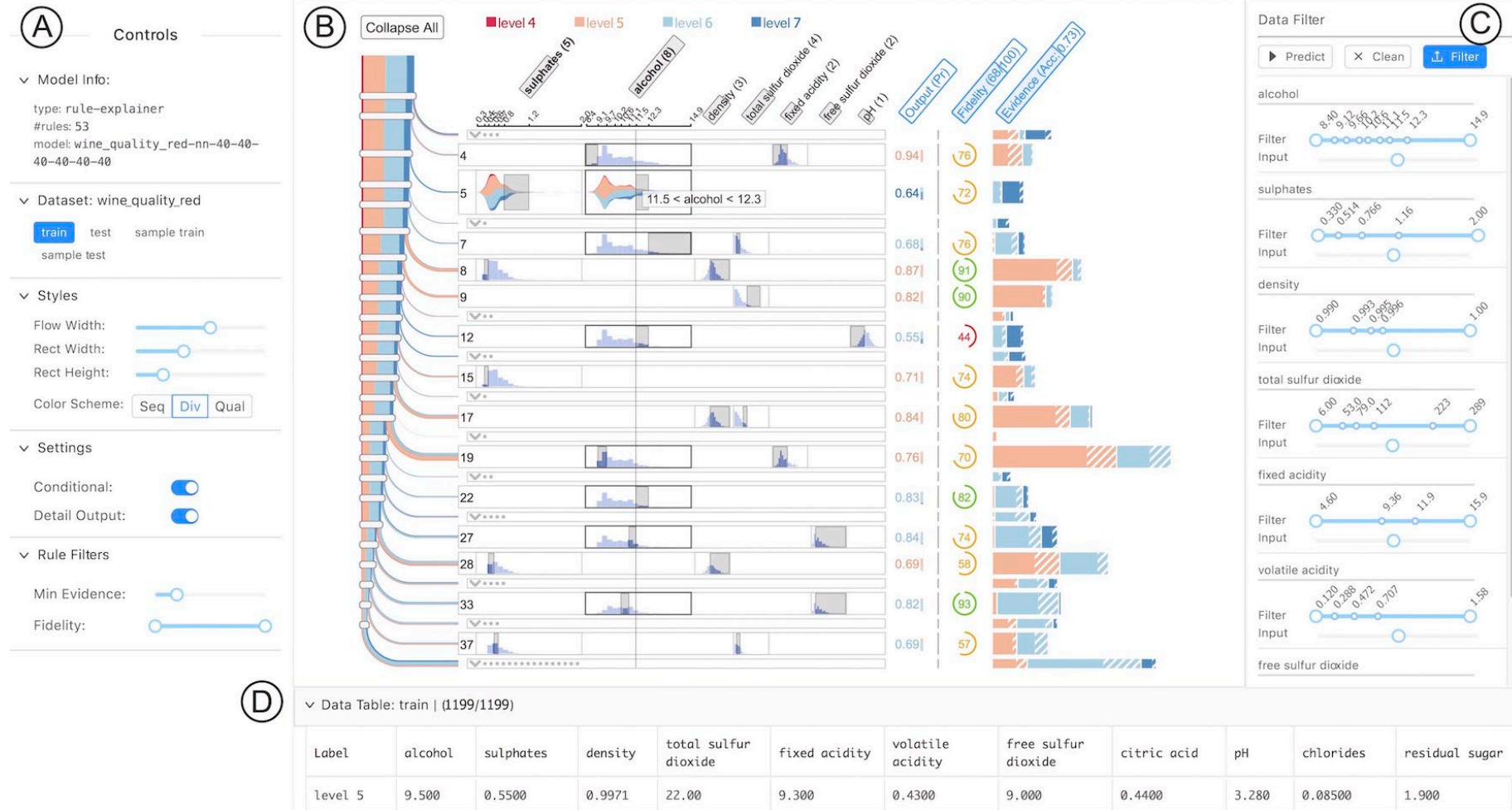
TreeExplainer for Tree Based Models

- The polynomial time algorithm to compute optimal explanations based on game theory
- An explanation that directly measures local feature interaction effects. Tools for understanding global model structure based on combining local explanations of each prediction
- TreeExplainer matches human intuition across a benchmark of 12 user study scenarios



Simple visualization with Local explanations based on TreeExplainer to understand global model structure [Lundberg 2019]

Rule Matrix



RuleMatrix: Visualizing and Understanding Classifiers using Rules

Brief overview of Visualization Methods in Deep Learning

Understanding Deep Learning via Generalization Analysis

Empirical observations

- Convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data.
- It occurs even after replacing the true images by completely unstructured random noise.
- Here the learning must be impossible and should show up during training, e.g., by not converging or slowing down.

Theoretical results

- Large neural networks can express any labeling of the training data.
- Theorem: There exists a two-layer neural network with $2n+d$ weights that can represent any function on a sample of size n in d dimensions.
- These models are in principle rich enough to memorize the training data.

Understanding Deep Learning via Generalization Analysis

- Explanation for such accurate models by known heatmap activation methods can be constructed, but what will be its value?
- To distinguish it from a meaningful explanation we need to analyze the generalization process and errors beyond training data.
- How to distinguish between the models trained on the true labels that are potentially explainable and models trained on random labels (high generalization error) that should not be meaningfully explainable?

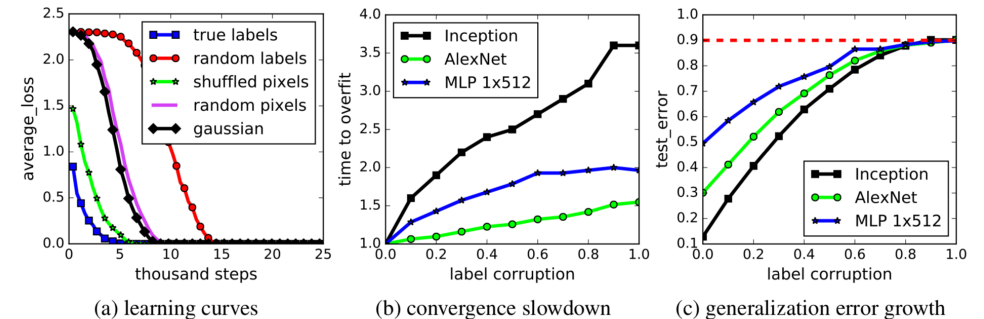
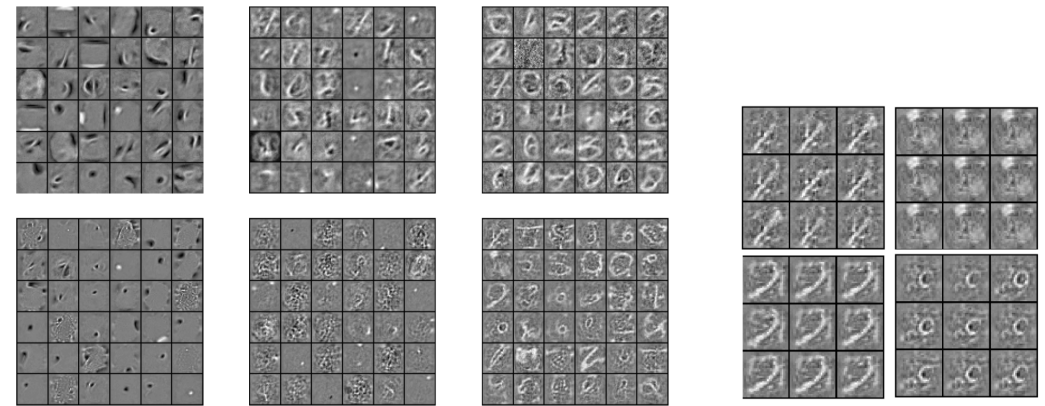


Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

Activation maximization (AM)

- Activation maximization is an analysis framework that searches for an input pattern that produces a maximum model response for a quantity of interest
- Response of individual units in the network. Like the analysis of individual neurons in the brain by neuroscientists, this approach has limitations



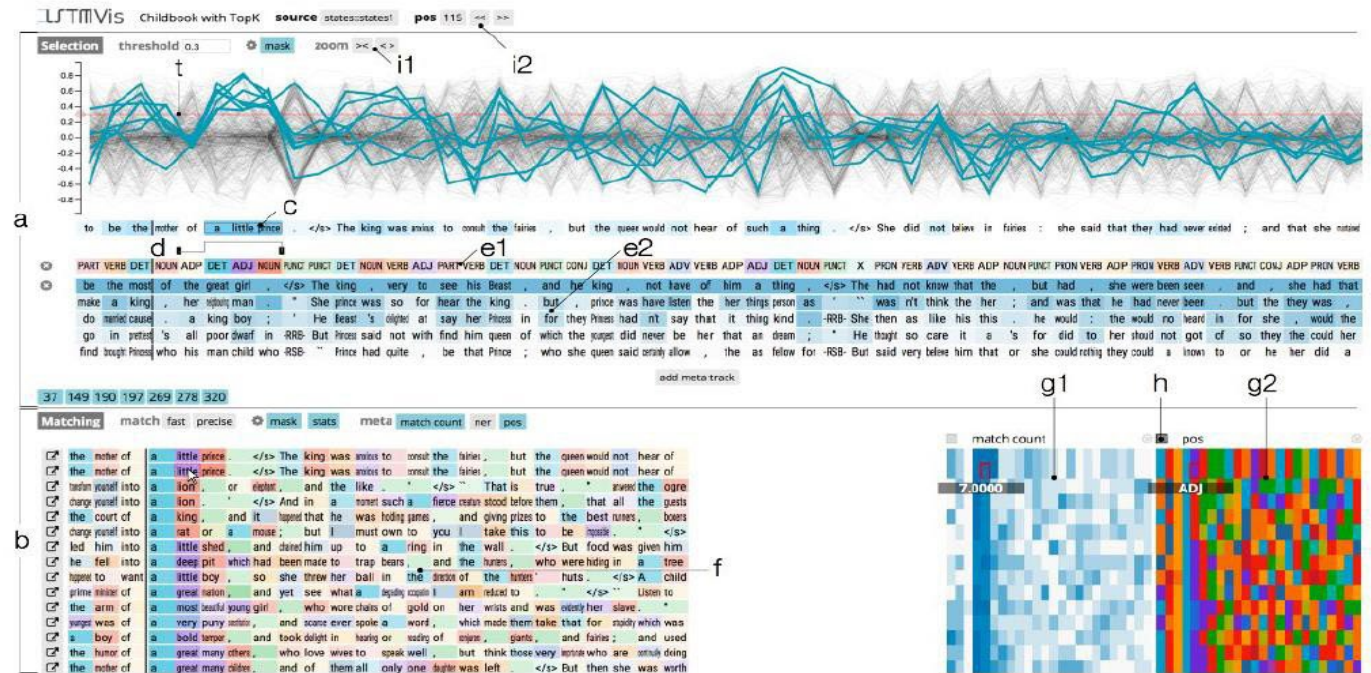
Activation maximization applied on MNIST.

Activation patterns of individual hidden nodes

- LSTMVis: Interactive exploration of the learnt behavior of hidden nodes
- A user selects a phrase, e.g., "a little prince," and specifies a threshold the system
- shows hidden nodes with activation values greater than the threshold and
- finds other phrases for which the same hidden nodes are highly activated.
- Given a phrase in a document, the line graphs in the top panel visualize the activation patterns of hidden nodes over the phrase
- Several other works with a similar idea -- activation and heatmap.

LSTMVis System: Open questions

- In the nearest neighbor explanation assumes the most similar case. No explanation of why the activation makes sense
- Where are relations between salient element is captured in this visualization?
- How to measure that the explanation is right?
- Visual tools are limited by Heatmap and Parallel coordinates

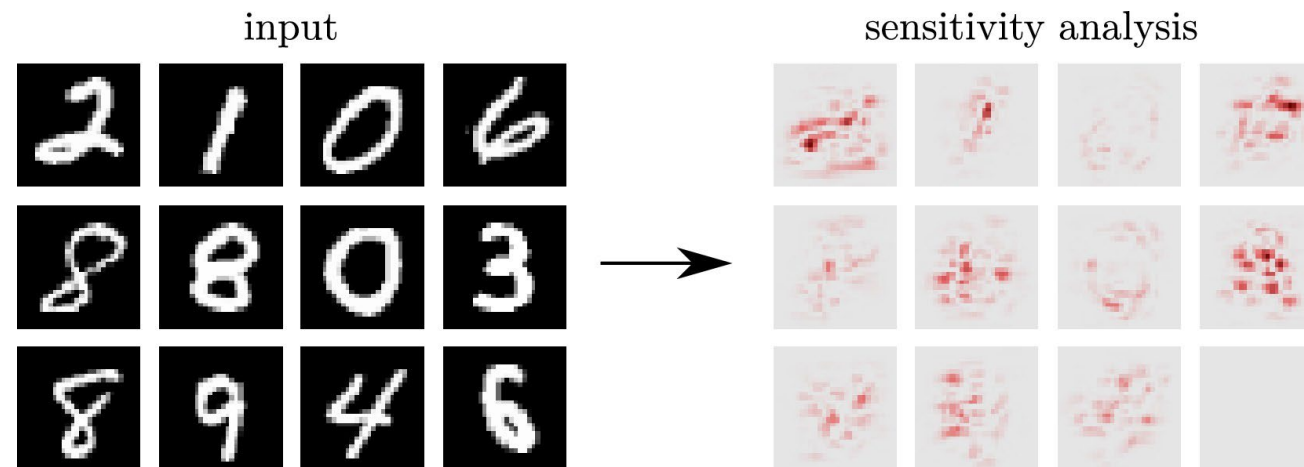


Sensitivity Analysis

- Identify the most important input features based on the model's locally evaluated gradient or some other local measure of variation
- The most relevant input features are those to which the output is most sensitive
- Sensitivity Analysis does not produce an explanation of the function value $f(x)$ itself, but rather a variation of it *i.e.*, what makes this image more/less a car?”, rather than the more basic question “what makes this image a car?”.
 - Example: Image-specific class saliency map, highlighting the areas of the given image, discriminative with respect to the given class

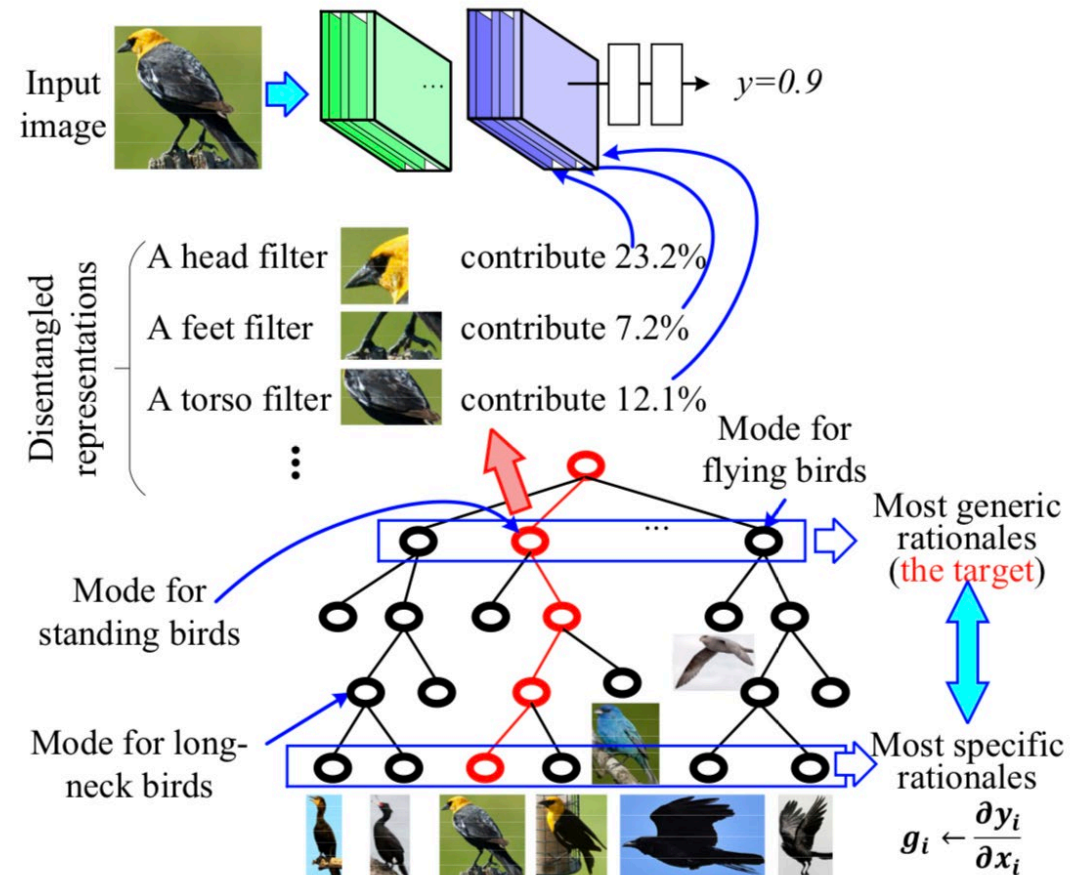
Sensitivity Analysis: Example

- Sensitivity analysis applied to a convolutional DNN trained on MNIST, and resulting explanations (heatmaps) for selected digits
- Heatmaps are spatially discontinuous and scattered, and do not focus on the actual class-relevant features
- This inadequate behavior can be attributed to the nature of sensitivity analysis



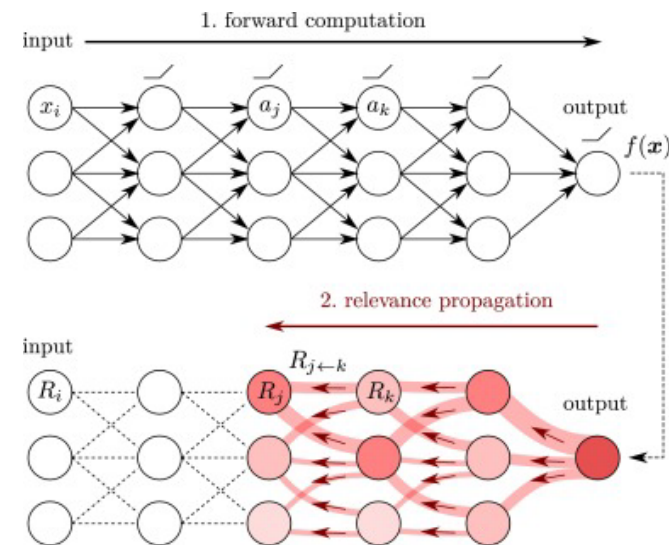
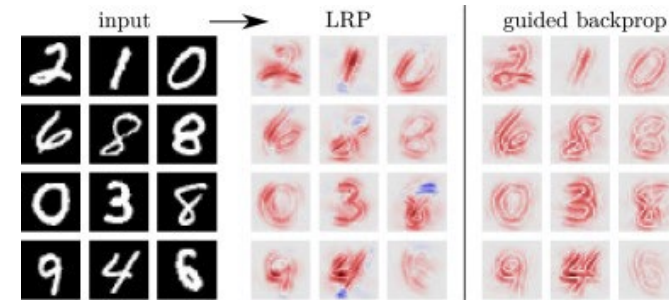
Decision Trees for Deep Learning Models

- Learn a decision tree, which clarifies the specific reason for each prediction made by the CNN at the semantic level
- Decision tree decomposes feature representations into elementary concepts of object parts
- The decision tree shows which object parts activate which filters for the prediction and how much each object part contributes to the prediction score



Layer-wise relevance propagation (LRP)

- Technique for explaining predictions
- The LRP technique is rooted in a conservation principle, where each neuron receives a share of the network output, and redistributes it to its predecessors in equal amount, until the input variables are reached
- For LRP to produce good explanations, the number of fully connected layers should be kept low, as LRP tends for these layers to redistribute relevance to too many lower-layer neurons (loose selectivity)



[Bach 2015]

Learning deep features

- Challenge: Scene recognition performance is lower than that for object recognition
- Reasons: Current deep features trained from ImageNet are not sufficiently competitive
- Approach: Methods to compare the density and diversity of image datasets
- CNN to learn deep features for scene recognition
- Heatmap Visualization of the CNN layers' responses to show differences in the internal representations of object-centric and scene-centric networks.

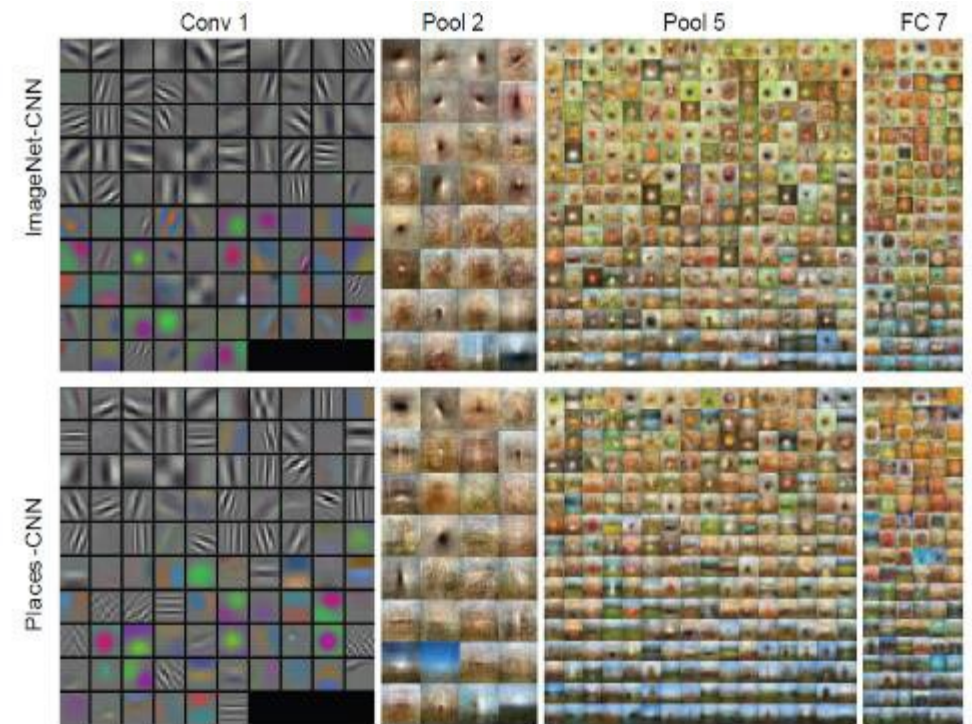
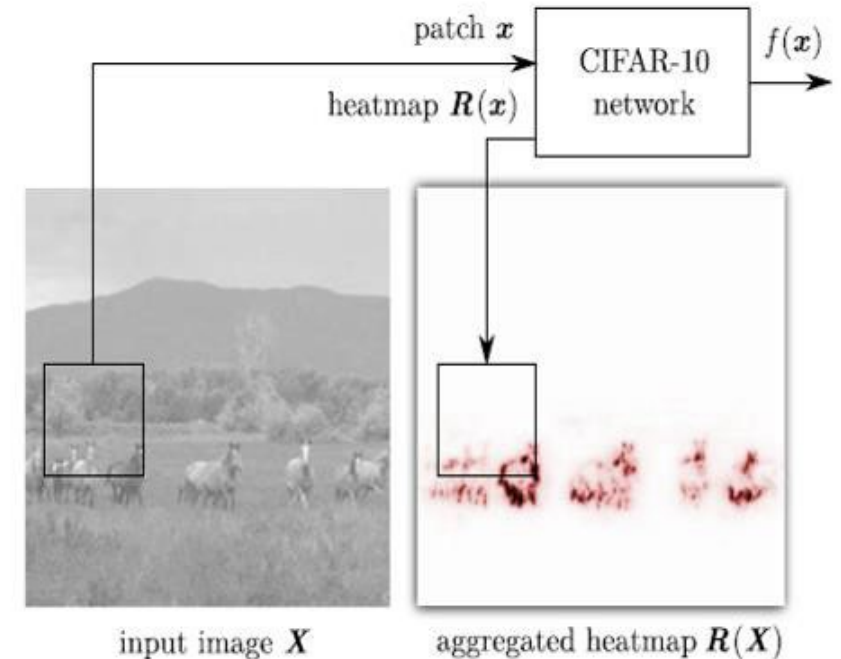


Figure 5: Visualization of the units' receptive fields at different layers for the ImageNet-CNN and Places-CNN. Conv 1 units contains 96 filters. The Pool 2 feature map is $13 \times 13 \times 256$; The Pool 5 feature map is $6 \times 6 \times 256$; The FC 7 feature map is 4096×1 . Subset of units at each layer are shown.

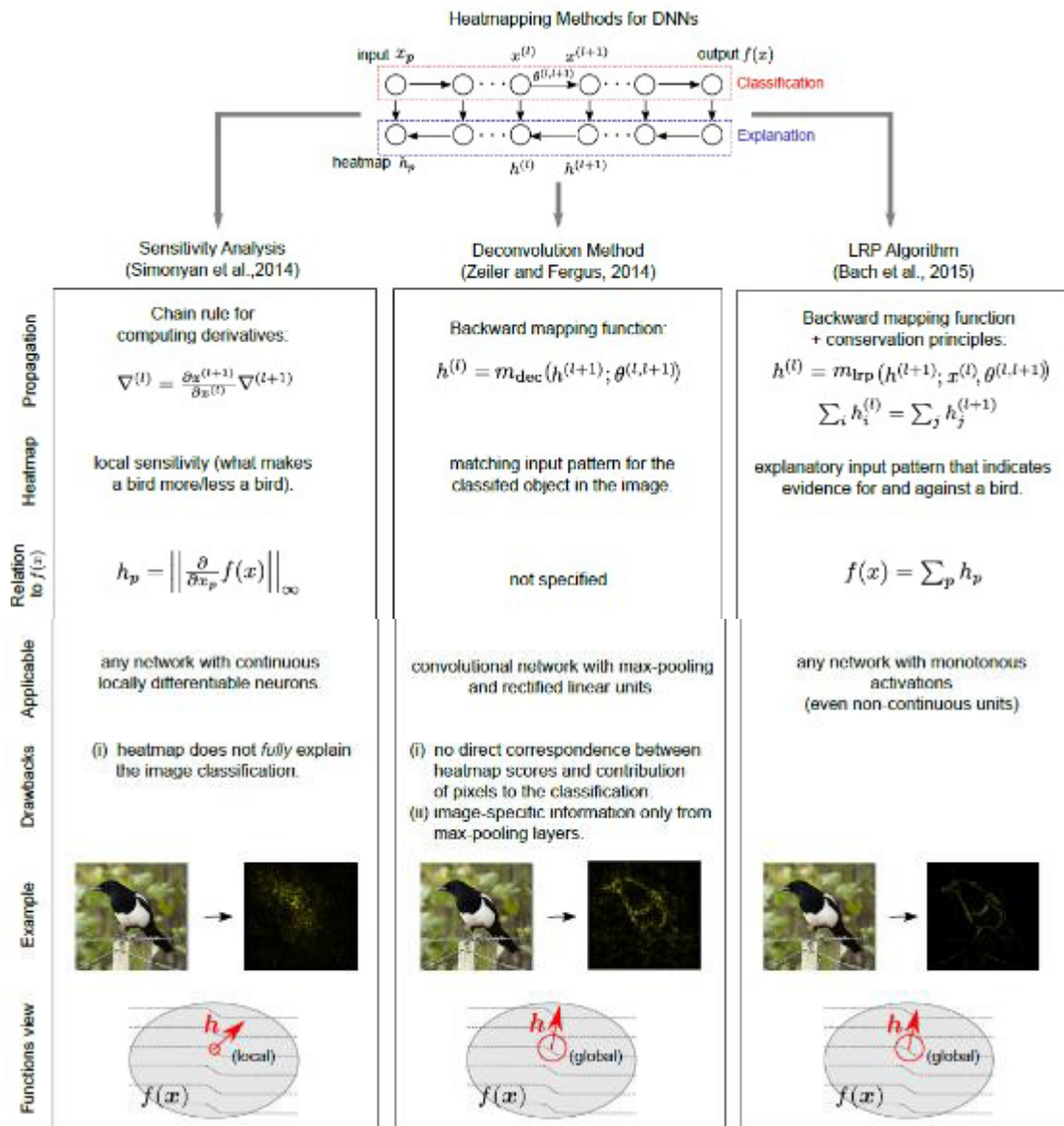
Visualization of Image Features in Heat Maps

- CIFAR-10 classification benchmark problem is to classify RGB 32x32 pixel images across 10 categories
- CIFAR-10 is a multi-layer network with alternating convolutions and nonlinearities followed by fully connected layers and softmax classifier
- 1M learnable parameters 19.5M multiply-add operations to compute inference on a single image



Highlighting in a large image pixels that are relevant for the CIFAR-10 class "horse", using the sliding window technique.

Comparison of the three heatmap computations



- Sensitivity heatmaps (local explanations) measure change of the class when specific pixels are changed based on partial derivatives. Applicable to architectures with differentiable units
- Deconvolution method (“autoencoder”) applies a convolutional network g to the output of another convolutional network f . Network g “undoes” f
- Layer-wise Relevance Propagation (LRP) exactly decomposes the classification output $f(x)$ into pixel relevancies by observing the layer-wise evidence for class preservation (conservation principle) Applicable to generic architectures (including with non-continuous units) -- does not use gradients

Example: Heat maps visualization and explanation of deep learning for pulmonary tuberculosis



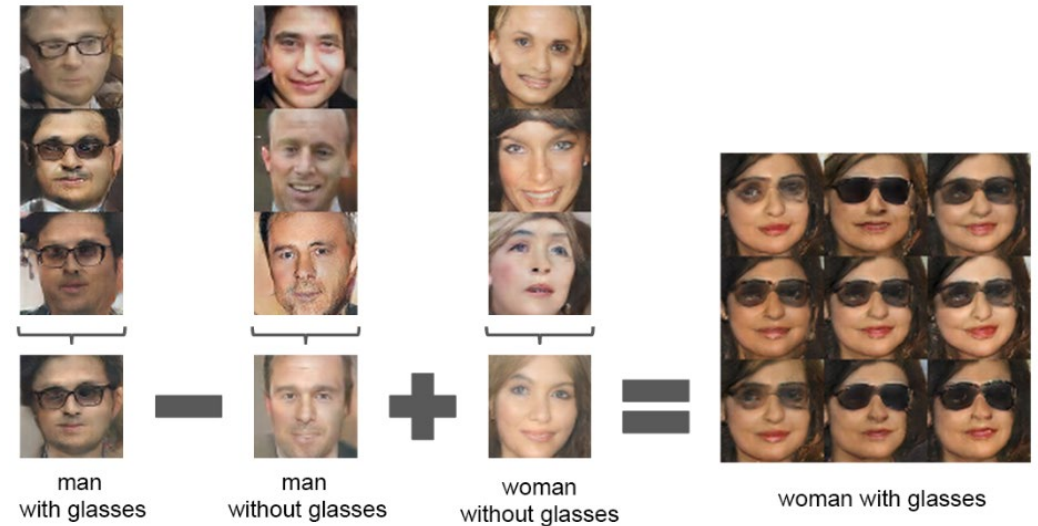
Left: Chest radiograph with pathologically proven active TB.

Right: The same radiograph with a heat map overlay of a strongest activations from the 5th convolutional layer from GoogLeNet-TA classifier. The red and light blue regions in the upper lobes -- areas activated by the deep neural network. (areas where the disease is present) The dark purple background -- areas that are not activated.

Generative Adversarial Networks (GANs)

Visualization

- Visualization and understanding of GANs is largely missing.
- How does a GAN represent our visual world internally?
- What causes the artifacts in GAN results?
- How do architectural choices affect GAN learning?



GANs Visualization

- A framework to visualize and understand GANs at the unit, object, and scene level
- Step 1: identify interpretable units closely related to object concepts with a segmentation-based network dissection.
- Step 2: quantify their causal effect by measuring interventions to control objects in the output
- Step 3: examine the contextual relationship between these units and their surrounding by inserting the discovered objects into new images

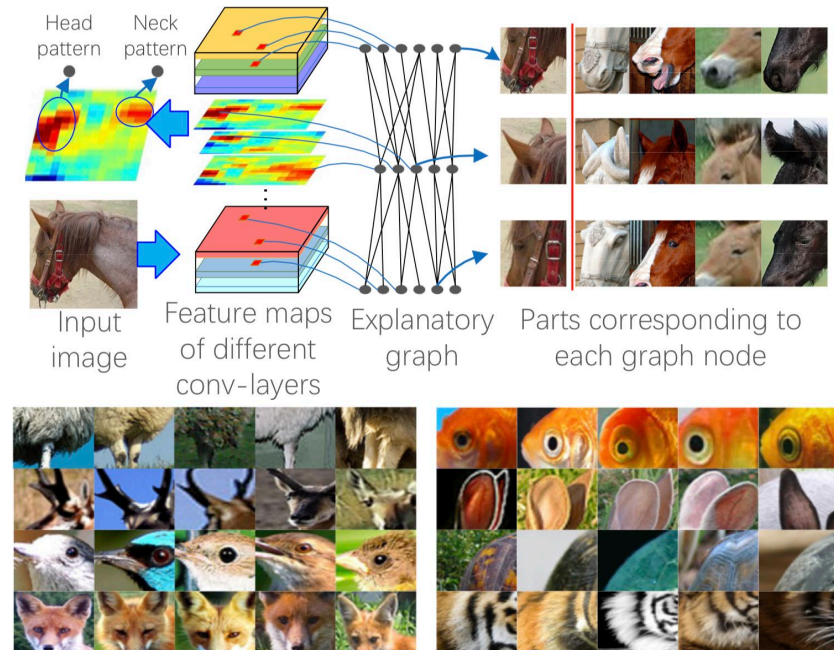


Figure 8: (a) We show two example units that are responsible for visual artifacts in GAN results. There are 20 units in total. By ablating these units, we can fix the artifacts in (b) and significantly

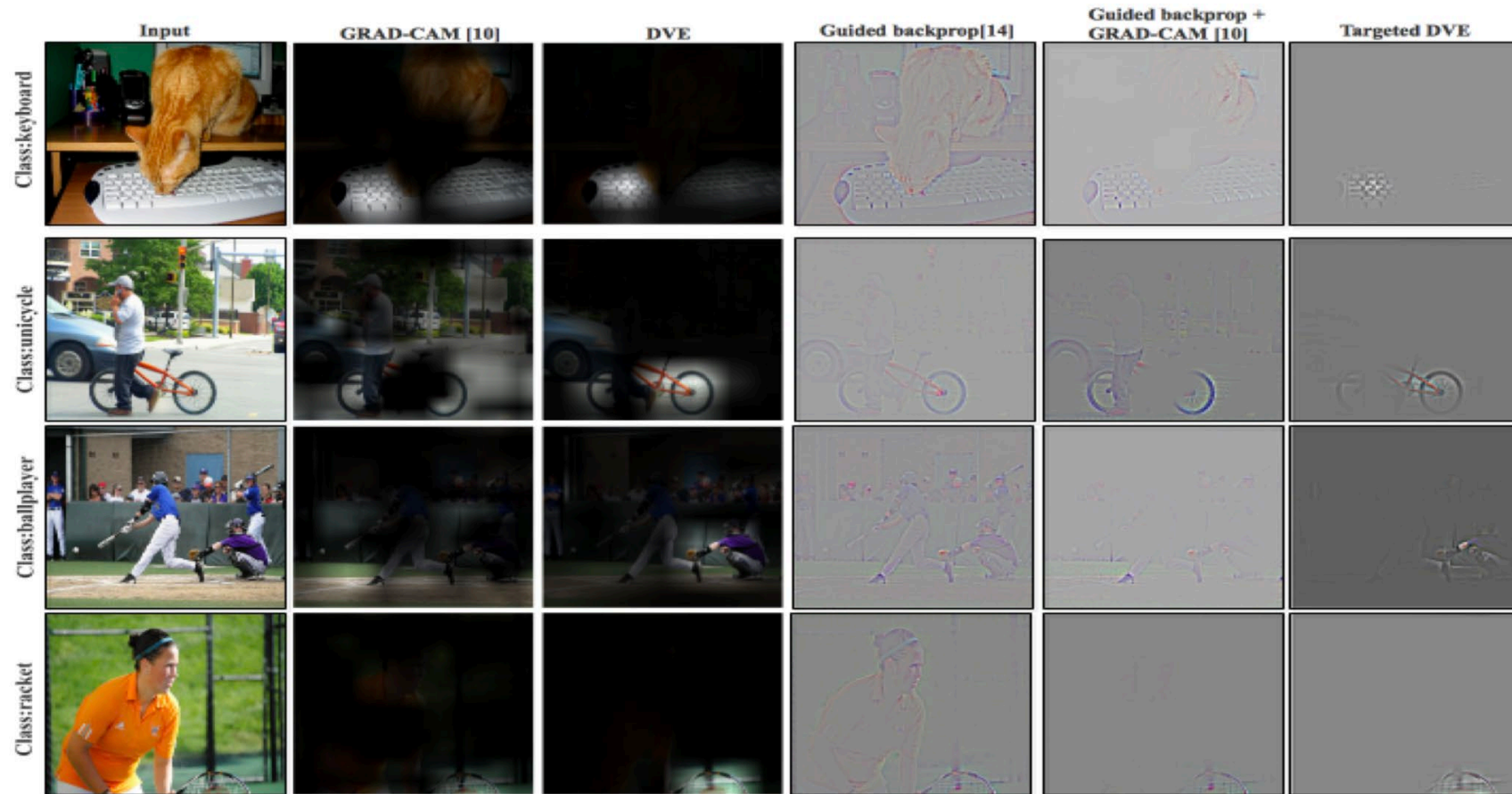
[Bau 2018]

Explanatory Graphs

- Represents the knowledge hierarchy hidden in conv-layers of a CNN
- The explanatory graph has multiple layers. Each graph layer corresponds to a specific conv-layer of a CNN
- Each filter in a conv-layer may represent the appearance of different object parts
- Think of these as compression of feature maps of conv-layers
- Just like a dictionary, each input image can only trigger a small subset of part patterns (nodes) in the explanatory graph

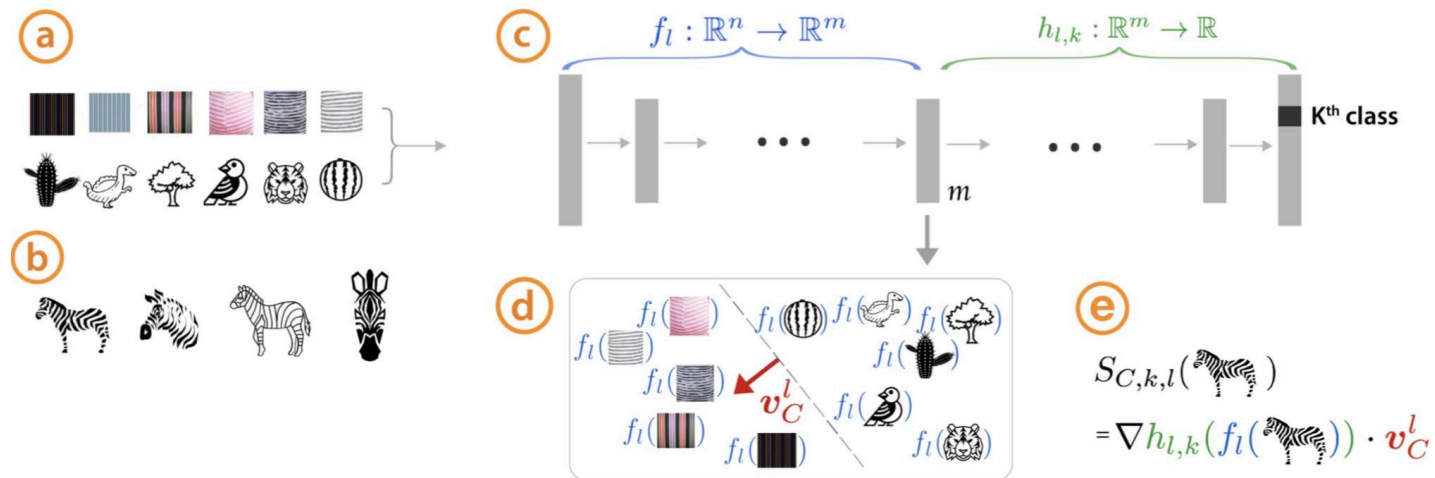


Deep Visual Explanations



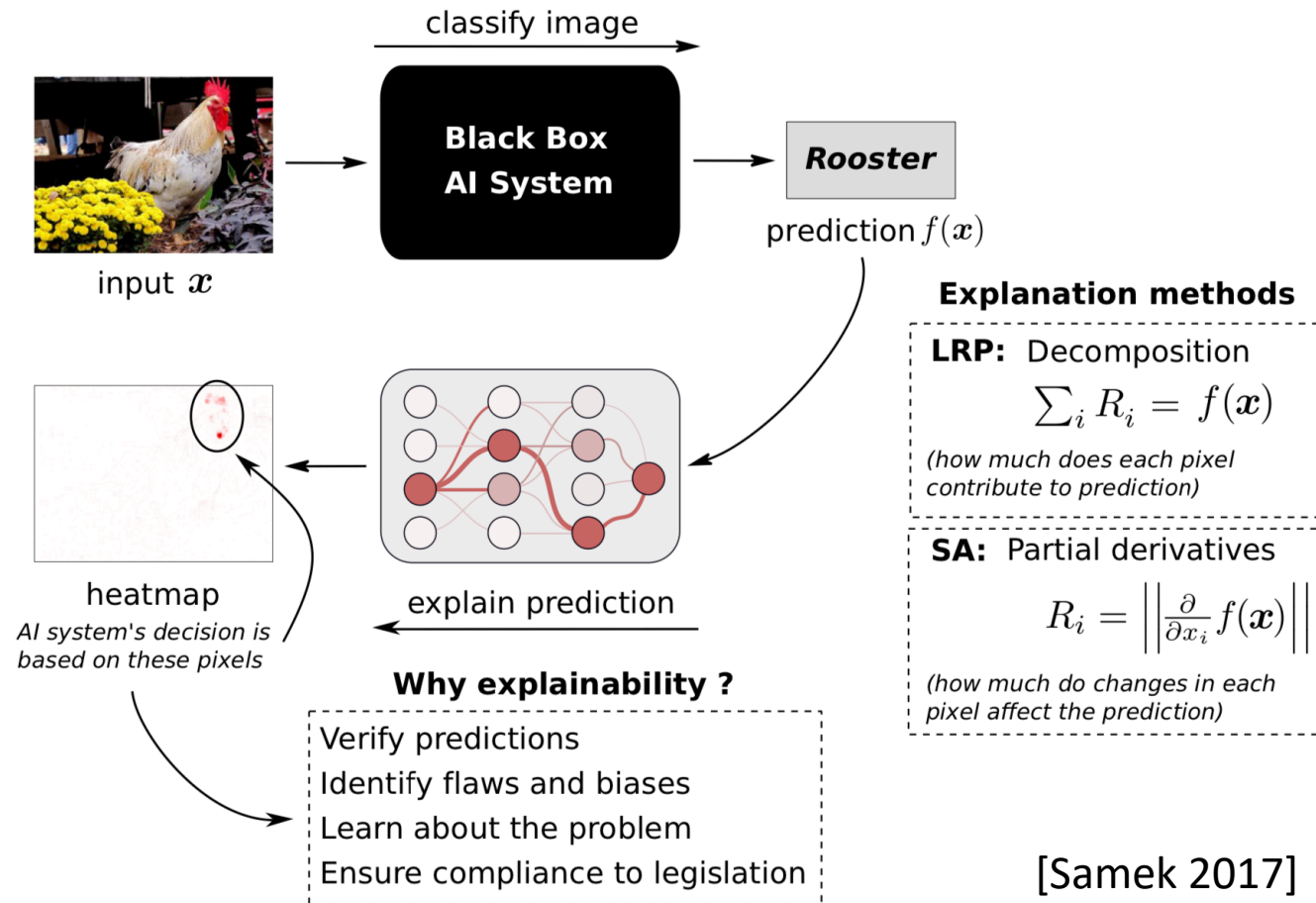
Concept Activation Vectors (CAV)

- Given a set of examples representing a concept of human interest, find a vector in the space of activations of layer L that represents this concept
- To find such a vector consider the activations in layer L produced by input examples that in the concept set versus random examples



Limits of Visual Interpretability in Deep Learning

Visual Methods for Interpretability in Deep Learning

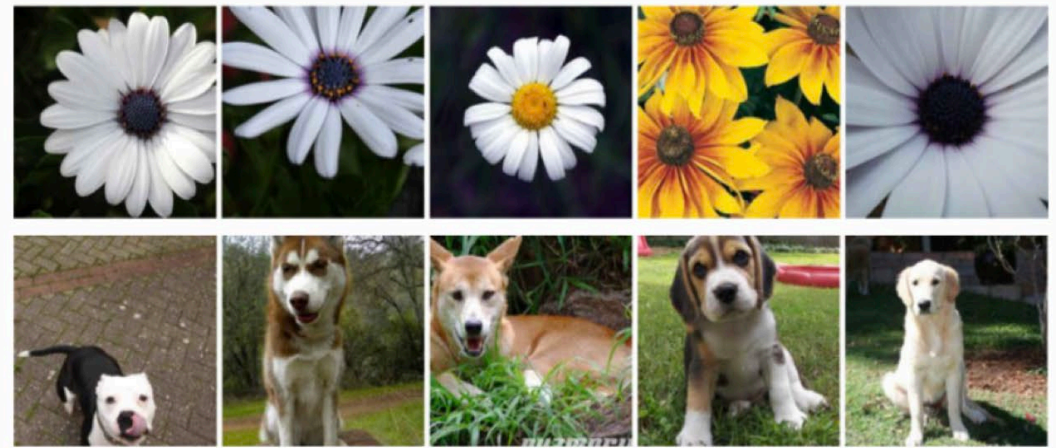


Are the concepts discovered by deep learning explainers real?

- No distinction between individual high-level unit 'concepts' and random linear combinations of high-level unit 'concepts'
- It is the space of relations rather than the individual units that contains the semantic information in network?



Single Neuron

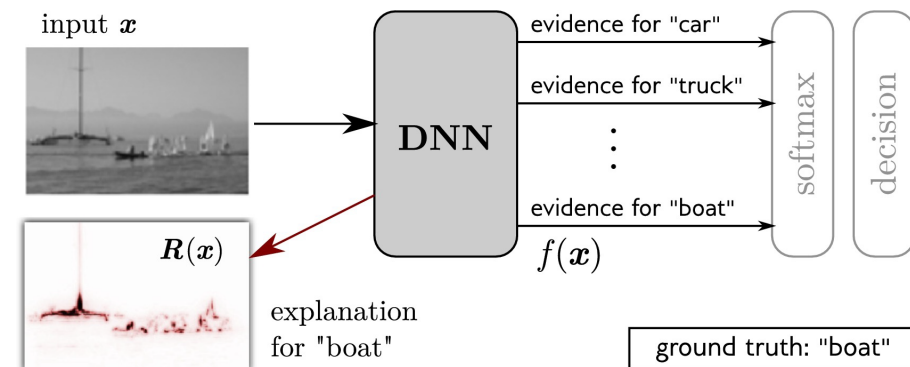


Random Projection

[Szegedy 2014]

Most methods for explainability in deep learning are incomplete

- Most explanations in deep learning are **implicit and incomplete** requiring a human giving a meaning to salient/dominant elements
- In the most example a human recognizes a mast in these pixels. In addition, this explanation can be **local** and **case specific**. In the boat example, another boat in the same image has no mast and requires its own explanation to be recognized as a boat



Insights from Adversarial Learning

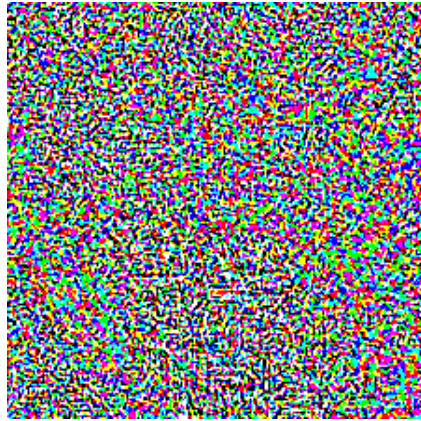
- What does adversarial learning reveal about what deep learning models are learning?
- Humans impose semantics on ML models



When a panda is a gibbon



Classified as panda

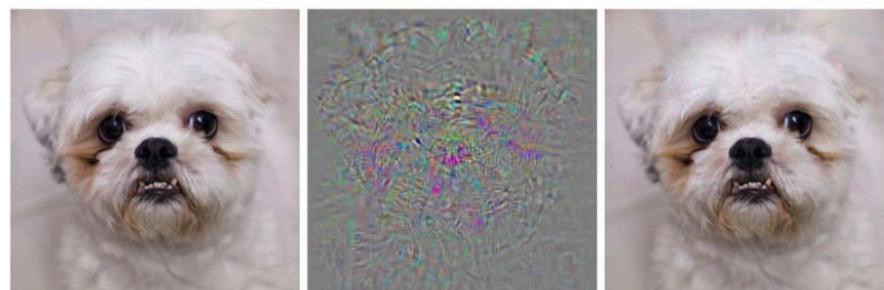
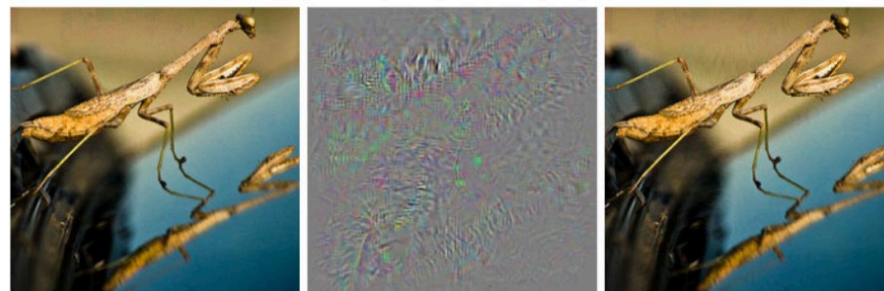
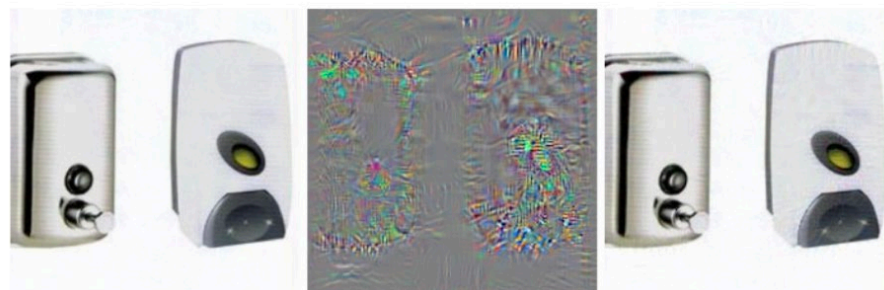
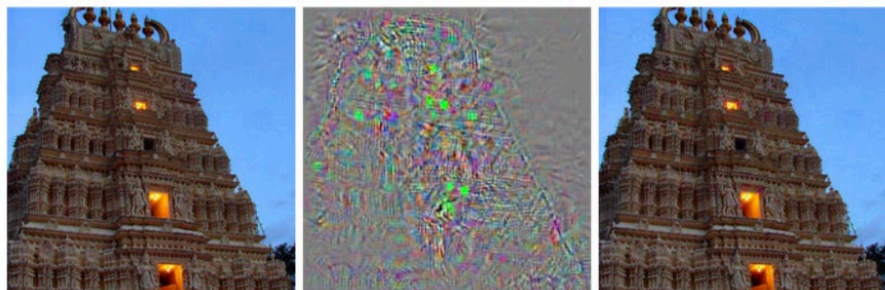
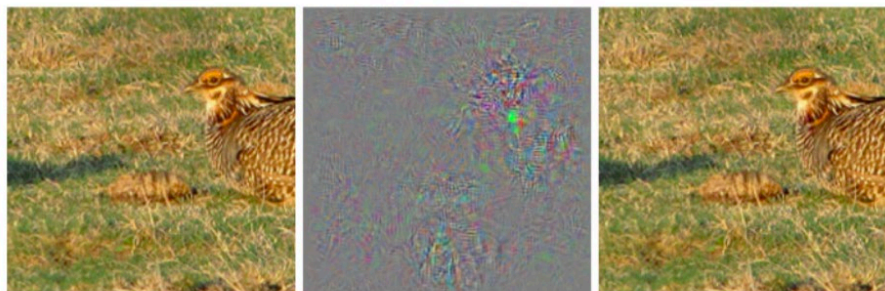
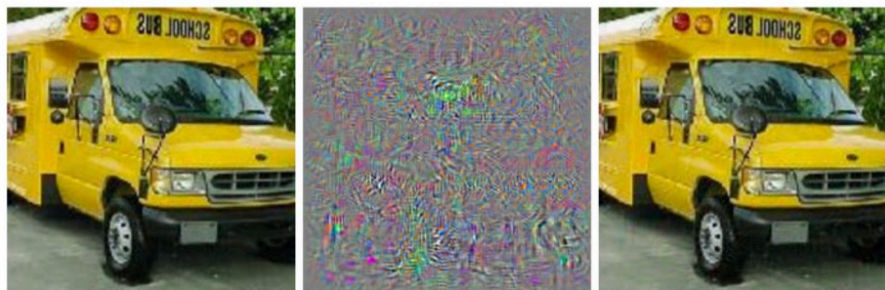


Small adversarial noise



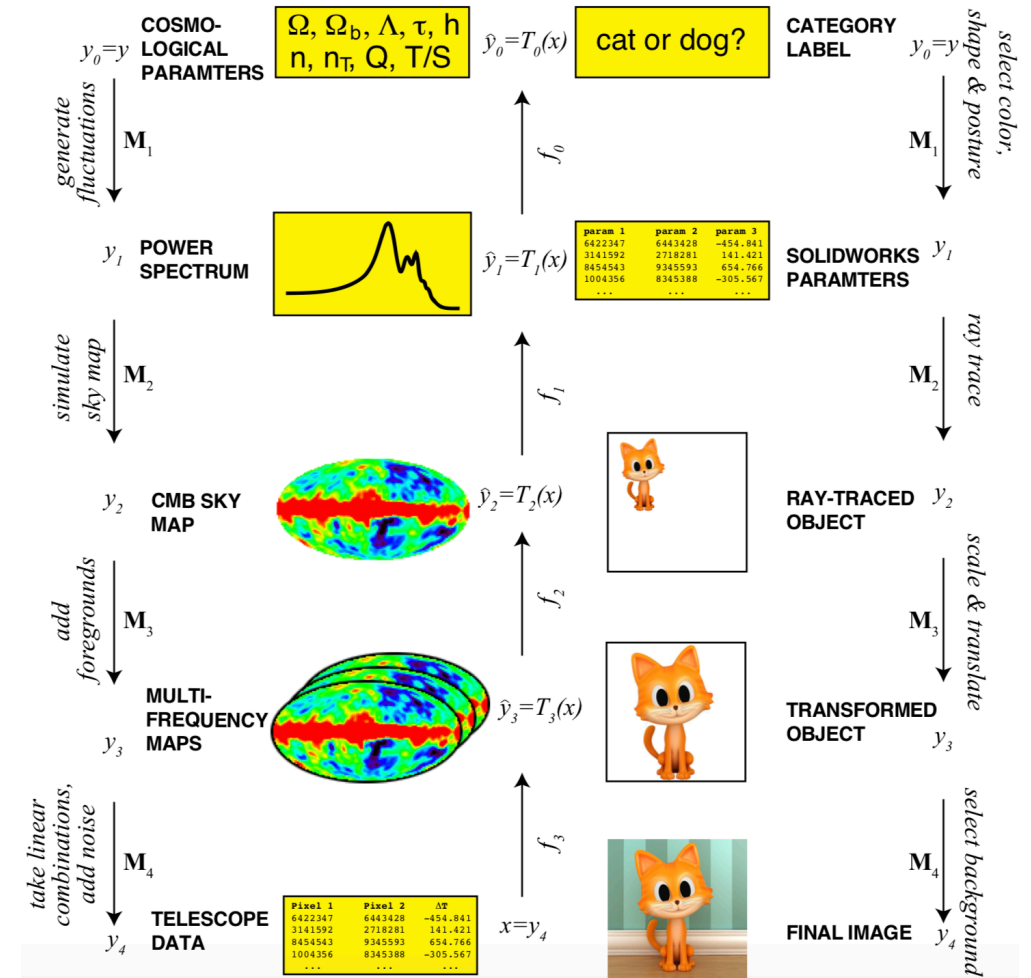
Classified as gibbon

Behold the Ostriches!



Hidden layers and Semantic Hierarchy

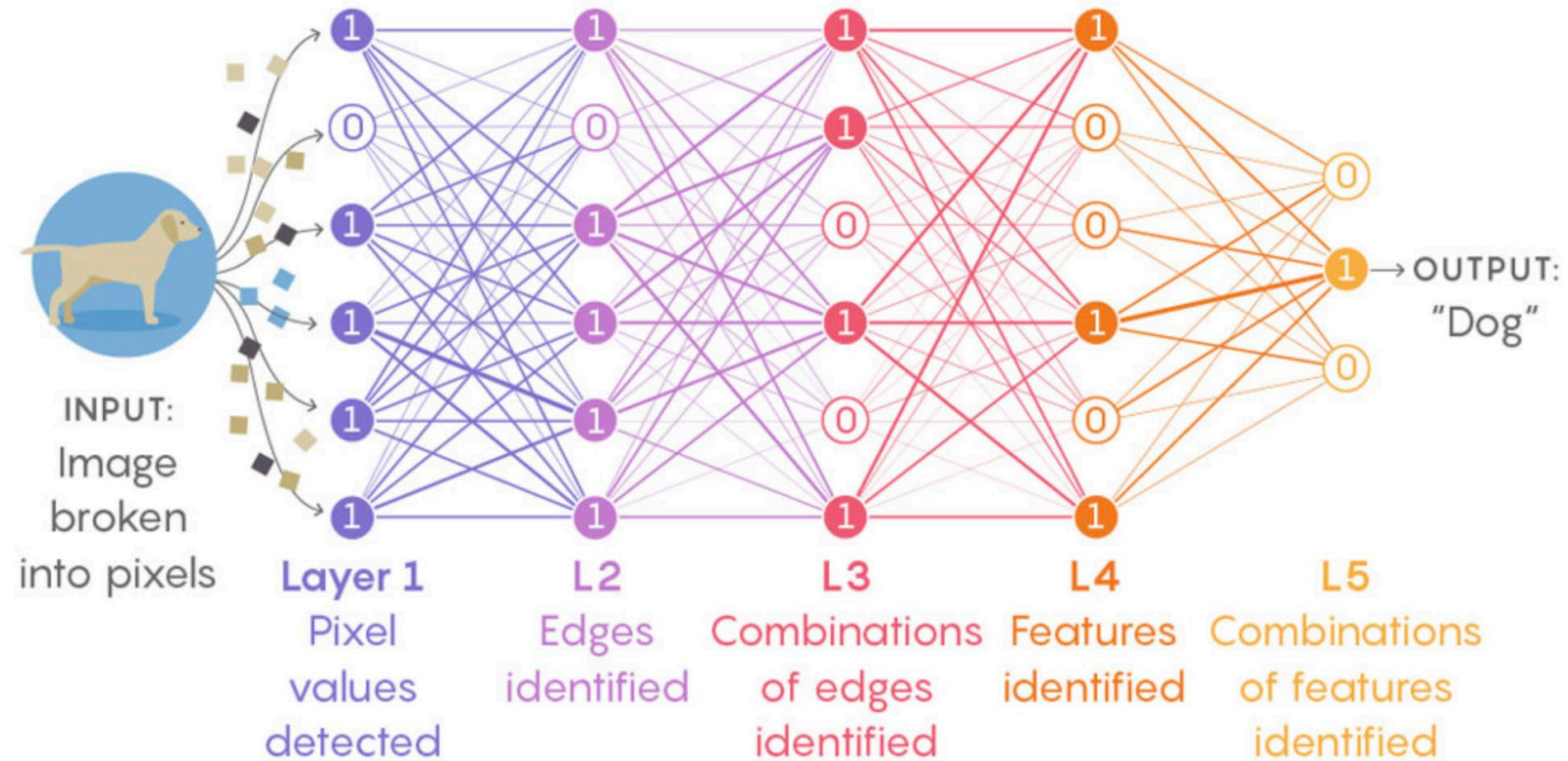
- The success of DL is not just because of “mathematics but also on physics, which favors certain classes of exceptionally simple probability distributions that deep learning is uniquely suited to model”
- Given a multivariate polynomial and any generic non-linearity, a neural network with a fixed size and a generic smooth activation function can indeed approximate the polynomial highly efficiently
- Success of deep learning possibly related to hierarchical and compositional generative processes in physics



Information Bottleneck

- Information Bottleneck: A distortion function that measures how well Y is predicted from a compressed representation T compared to its direct prediction from X
- “error back-propagation, pushes the layers of any deep neural network - one by one - to the information bottleneck optimal tradeoff between sample complexity and accuracy, for large enough problems. This happens in two distinct phases”
- The first, the network memorizes training examples with a lot of irrelevant details with respect to the labels
- The second phase the layers "forget" irrelevant details of the inputs, which dramatically improves the generalization ability of the network

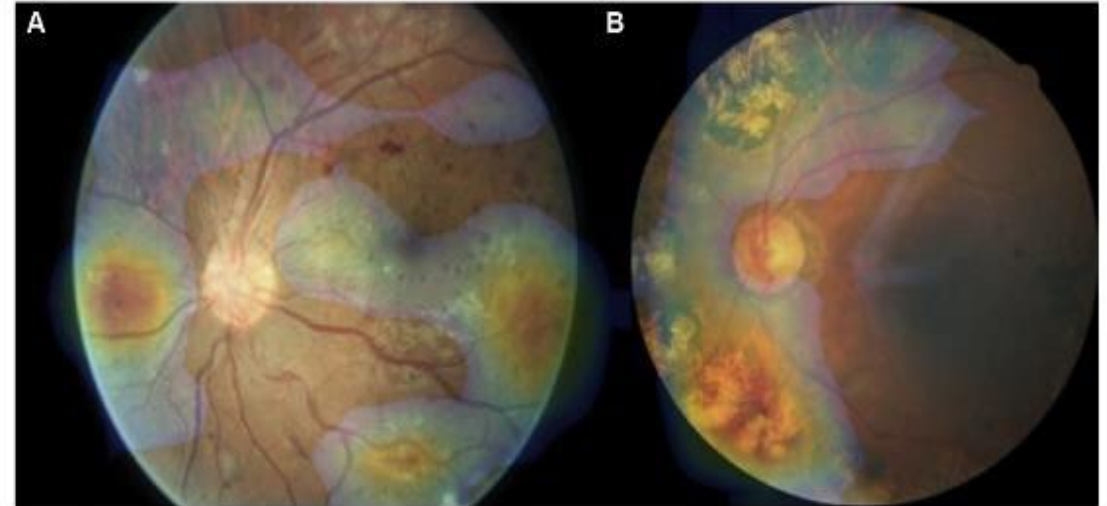
Information Bottleneck



[Tishby 2000, Shwartz-Ziv 2017]

Data quality and Generalization

- A common method of combining results of Deep Learning (DL) from images with visualization is discovering classification model for images using a DL algorithm, identifying informative deep features
- Visualizing identified deep features on the original image.
- Issue – Are visualized features always explainable?



[Gargeya 2017]

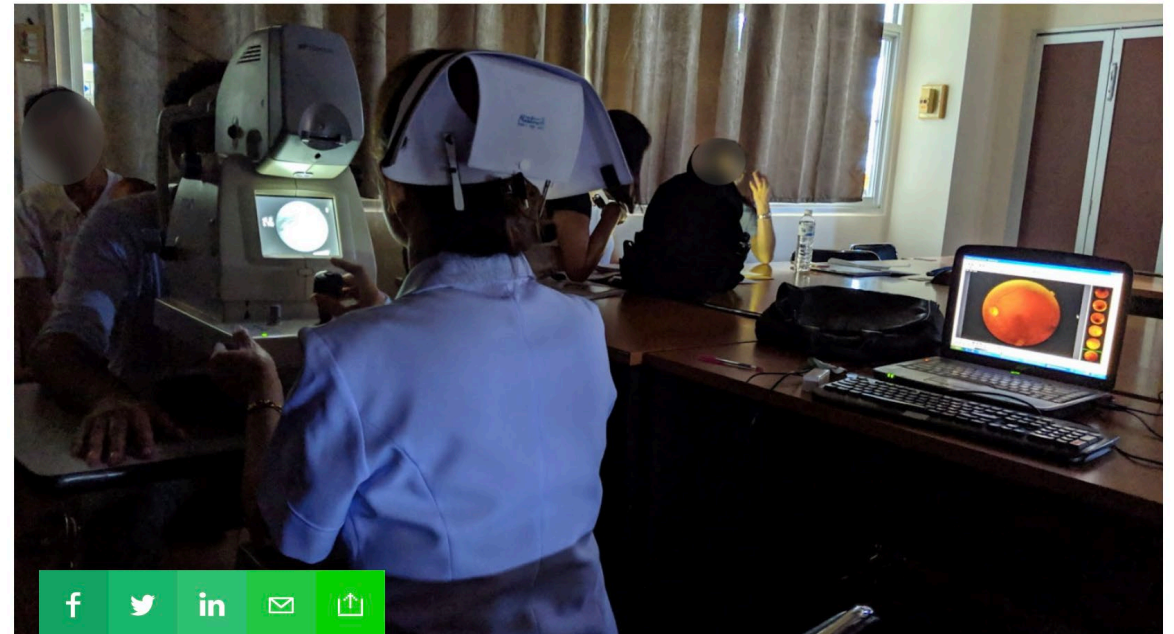
Data quality and Generalization

- If an image has a bit of blur or a dark area, the system will reject it
- Clinics in the study often experienced slower and less reliable connections. In one clinic, the internet went out for a period of two hours during eye screening, reducing the number of patients screened from 200 to only 100.
- Fewer people in this case received treatment because of an attempt to leverage this technology

Google medical researchers humbled when AI screening tool falls short in real-life testing

Devin Coldewey @techcrunch / 5:03 pm EDT • April 27, 2020

Comment



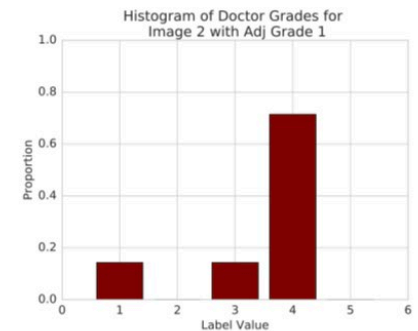
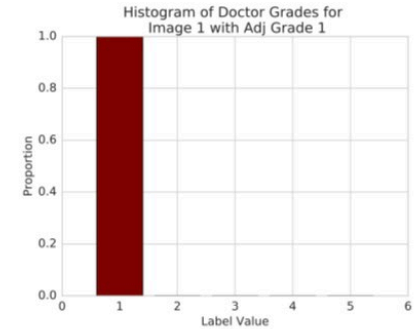
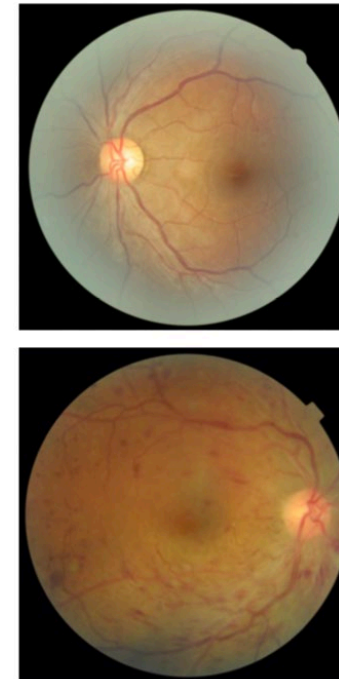
User-centric Views of Interpretability of Visual Methods

What is User Centric Interpretability?

- The participation of end users in the design of machine learning tools is imperative - to better understand how the end users will utilize the output components
- The notions of interpretability that the designer and the user have may be different
- In many cases the user's *expectation* of Interpretability or explainability are centered on actionability

The Problem of Ground Truth

- Data may not be of good quality because experts may not agree on definitions of labels e.g., diagnosis in radiology
- Requires further follow-up, pathologic diagnosis, or clinical outcomes to achieve ground truth
- It is estimated that 2%–20% of radiology reports contain demonstrable errors



For the top image, all doctors agreed that the grade should be 1, while there was a significant spread for the bottom image

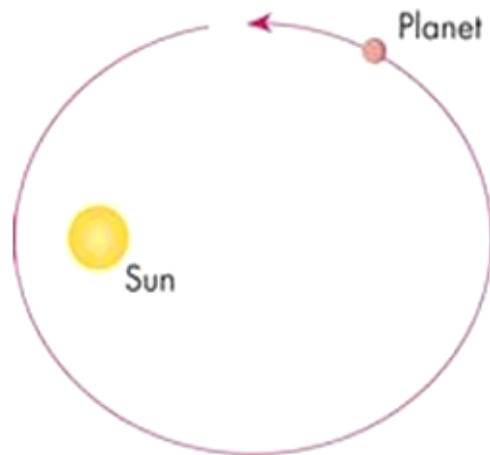
[Raghu 2019, Willeminck 2020]

Interpretations are often Incomplete

- How do we make sense of cases where it is possible to explain a model without completely understanding it?
- Can we use black-boxes to understand black-boxes?
- What does it mean for an explanation to be complete?
- Many Examples from the History of Science

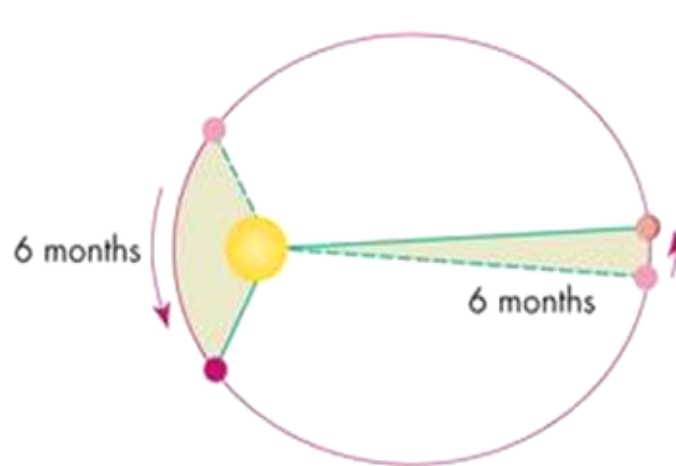
Kepler's Laws of Planetary Motion

- Kepler law's (1619) provided a elliptic mathematical approximation of planetary motions but not a why explanation for it
- Newton's theory of gravitation provided an explanation almost 70 years later (1687)



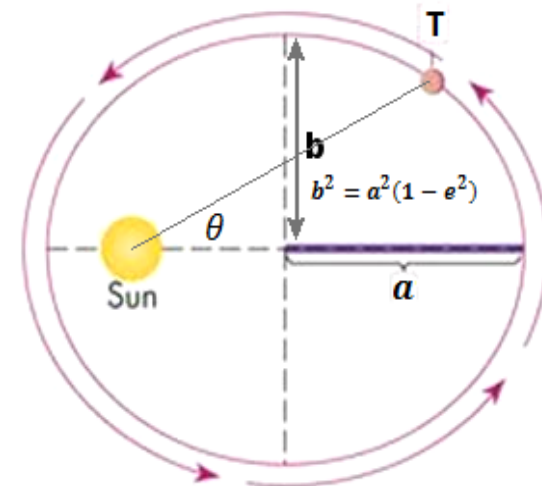
(1)

The orbits are ellipses



(2)

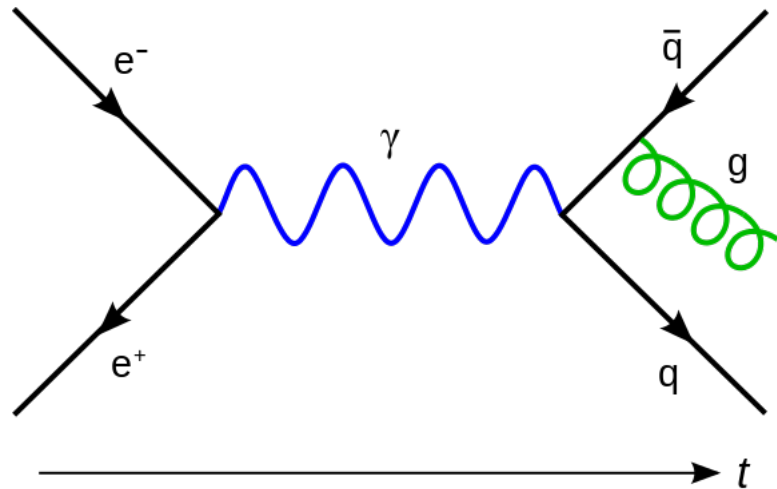
Equal areas in equal time



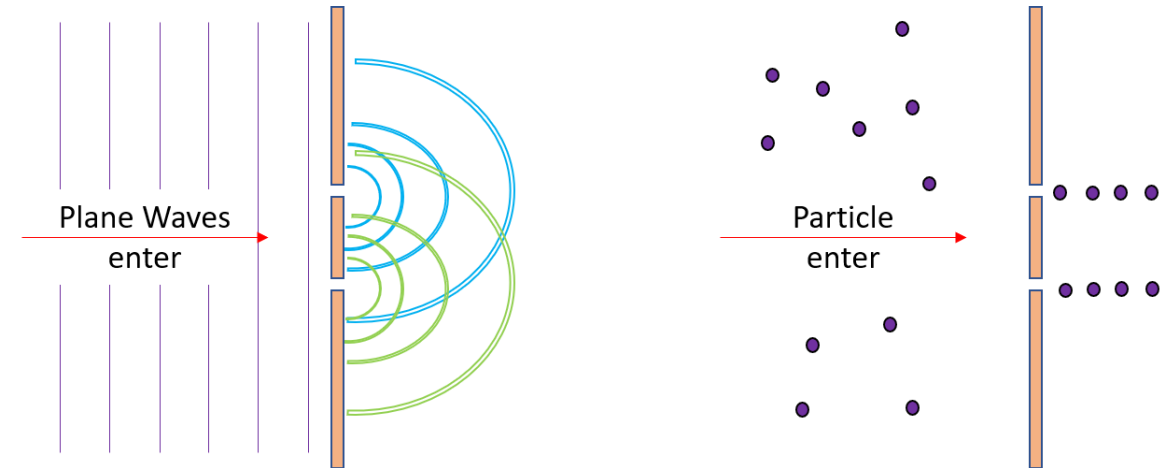
(3)

T = time to complete orbit
 $T^2 \propto a^3$ a = semi-major axis

Nobody Understands Quantum Mechanics!



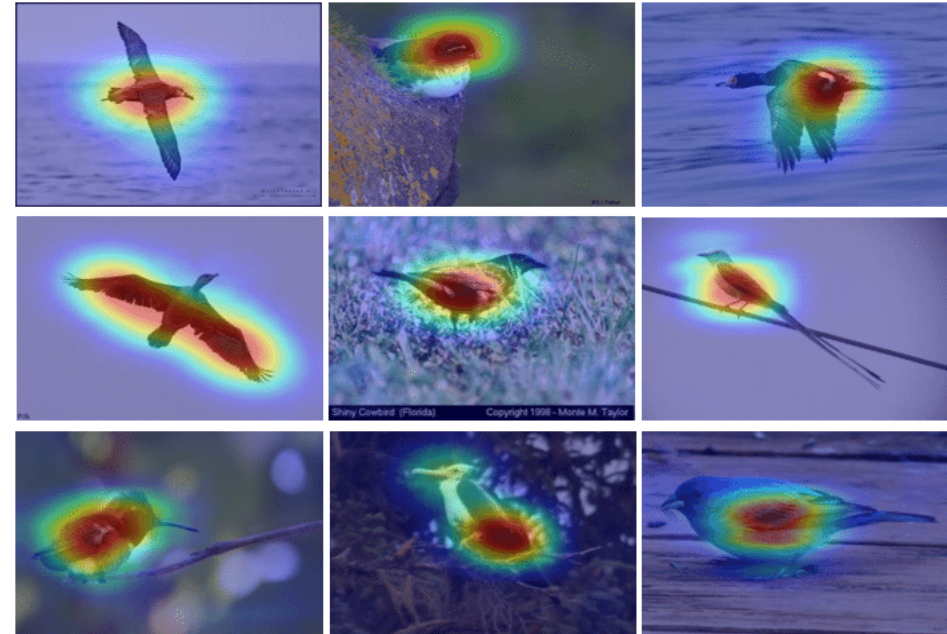
In this Feynman diagram, an electron (e^-) and a positron (e^+) annihilate, producing a photon (γ , represented by the blue sine wave) that becomes a quark–antiquark pair (quark q , antiquark \bar{q}), after which the antiquark radiates a gluon (g , represented by the green helix).



What does it mean for something to be a particle and wave at the same time?

Why do Saliency Based Methods Work

- Visual Summarization
- Attenuation to human gaze
- Low cognitive overload
- Plausible justification



When Saliency does not work



- What is the model learning when it learns lipstick?

When Saliency does not work

- Not always possible to extract semantics from feature maps



Human vs. Algorithmic semantics revisited

- Better performing DL models have higher proportions of deep neurons highly predictive of human gaze
- The predictive neurons are attuned to clear semantic categories such as animals (dogs, cats), objects (motorbike, ball) and parts (head, hair)
- This hints that saliency, as experienced by humans, likely involves high-level world knowledge in addition to low-level perceptual cues
- Computational approach to improve DL: minimizing the distance between the predicted saliency maps and the ground truth recorded by human gaze

What needs to be interpretable when we interpret ML models

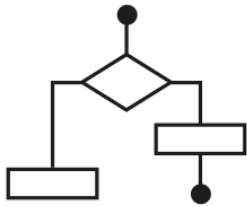
- Interpretability is a system wide phenomenon, features, parameters, and even insight delivery must be interpretable
- Satisfying is needed rather than always having model fidelity
- “All models are wrong. Some models are useful.” - Box
- Interpretability often does not require completeness
- A satisfactory explanation of the decision process of the underlying model is often required

Systems view of interpretability

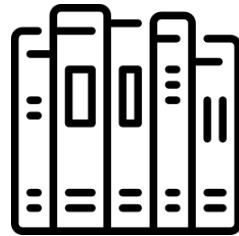
AI Solution



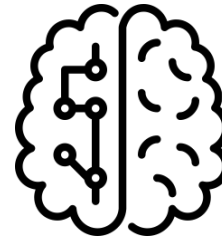
Features



Algorithm



Model Parameters



Model

Each element constituent of the solution process needs to be explainable for the solution to be truly explainable [Lipton 2016]

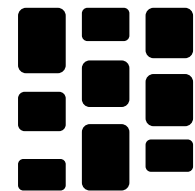
User



Cognitive Capacity

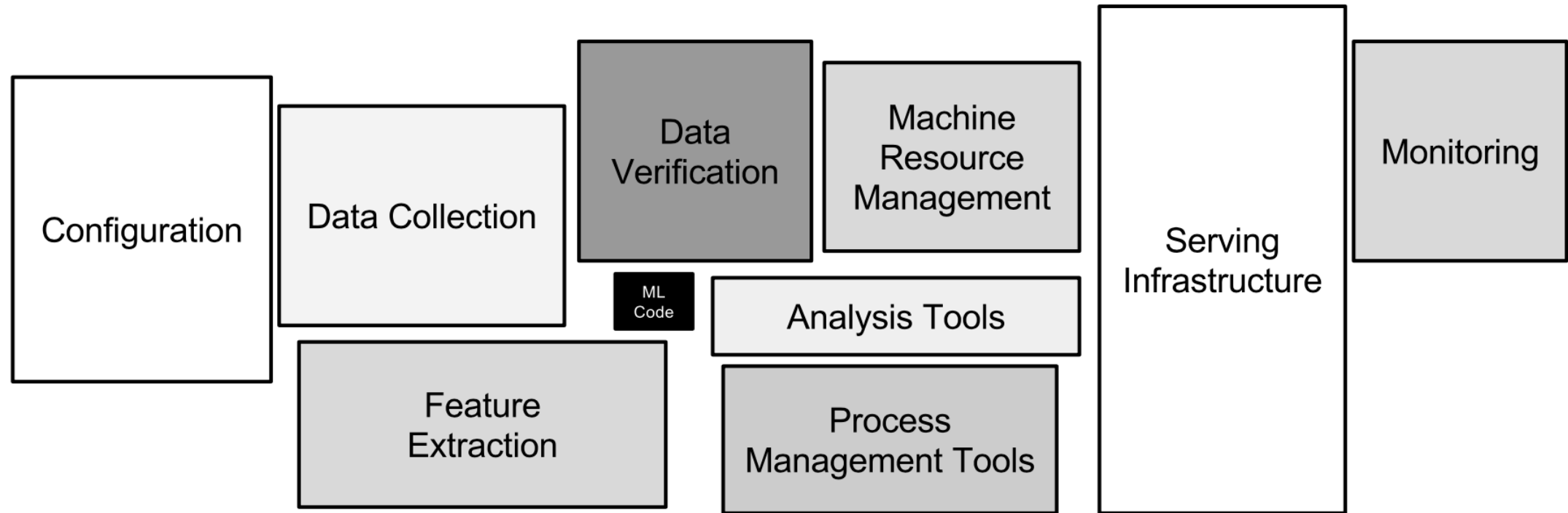


Domain Knowledge



Explanation Granularity

Operationalizing Interperable ML



Only a small fraction of real-world machine learning systems actually constitutes machine learning code [Sculley 2015].

Open Problems and Current Research Frontiers

Explanation Fidelity in Visual Methods

- Many, if not most, explanations are wrong, while some explanations are useful. Requiring absolute fidelity in interpretable ML is unwarranted, given the complexity of models involved
- What are the “good enough” models that allow debugging? Does the explanation capture the space of phenomenon to be explained?
- The *right* explanation is not necessarily the ‘correct’ explanation. Context and use cases determine what level of fidelity is required for the explanations

Right for the Right Reasons Model

- Models can be right for the wrong reasons [Ross 2017]
- Use domain knowledge to constrain explanations
- Training models with input gradient penalties

Input gradients `+soc.religion.christian` `+alt.atheism`

From: USTS012@uabdpo.dpo.uab.edu
Subject: Should teenagers `pick` a `church` parents `don't` attend?
Organization: UTexas Mail-to-News Gateway
Lines: 13

Q. Should teenagers have the `freedom` to choose what `church` they go to?

My `friends` teenage kids do not like to go to `church`.
If left `up` to them they would sleep, `but` that's not `an` option.
They `complain` that they have no `friends` that go there, yet `don't`
`attempt` to make `friends`. They `mention` not respecting their Sunday
school teacher, and usually `find` a way to miss Sunday school `but`
do make `it` to the `church` service, (after their `parents` are thoroughly
disgusted) I might `add`. A never ending battle? It can just ruin your
`whole` day if `you` let it.

Has `anyone` had this problem and how did `it` get resolved?

f

LIME `+soc.religion.christian` `+alt.atheism`

From: USTS012@uabdpo.dpo.uab.edu
Subject: Should teenagers `pick` a `church` parents `don't` attend?
Organization: UTexas Mail-to-News Gateway
Lines: 13

Q. Should teenagers have the `freedom` to choose what `church` they go to?

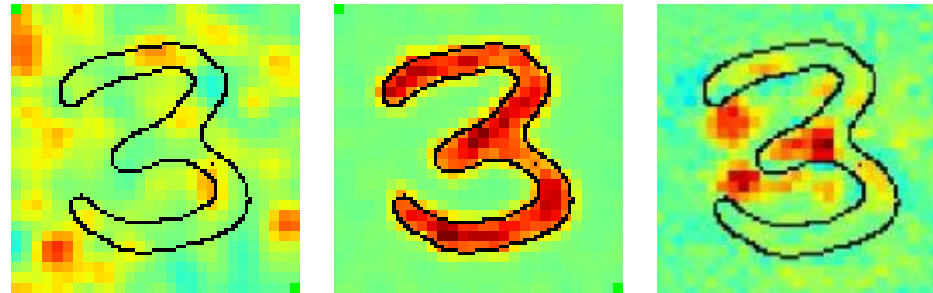
My `friends` teenage kids do not like to go to `church`.
If left up to them they would sleep, but that's not an option.
They `complain` that they have no `friends` that go there, yet `don't`
`attempt` to make `friends`. They `mention` not respecting their Sunday
school teacher, and usually find a way to miss Sunday school but
do make it to the `church` service, (after their `parents` are thoroughly
disgusted) I might add. A never ending battle? It can just ruin your
whole day if you let it.

Has anyone had this problem and how did it get resolved?

f

[Lakkaraju, Bach & Leskovec, 2016]

Evaluation of visual methods



- Comparison of three heatmaps for digit '3'.
- L: The randomly generated heatmap – no interpretable information
- C: The segmentation heatmap – shows the whole digit without relevant parts, say, for distinguishing '3' from '8' or '9'.
- R: A relevance heatmap shows parts of the image used by the classifier.
- Reflects human intuition on differences between '3', '8' and '9' and other digits

Domain vs. non-Domain validation

- How do we validate explanations if complete fidelity is not required?
- The interpretation must make sense within the ontology of the domain
- Outside of the domain, the method needs to operate within the constraints imposed by formal methods when applicable
- Validation is a domain focused question, but can one create cross-domain general methods for validation?

Cognitive Limitations

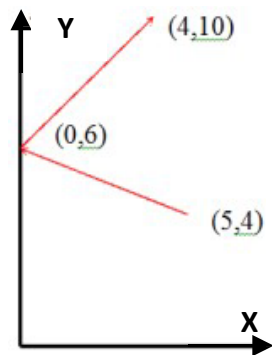
- Machine Learning is used in problems where the size of the data and/or the number of variables is too large for humans to analyze
- What if the most parsimonious model is indeed too complex for humans to analyze or comprehend?
- Ante-Hoc explanations may be impossible and post-hoc explanations would be 'incorrect'
- "[Humans] make a decision first, and then you ask, and then they generate an explanation and that may not be the true explanation."
– Peter Norvig

Cross-Domain Pollination

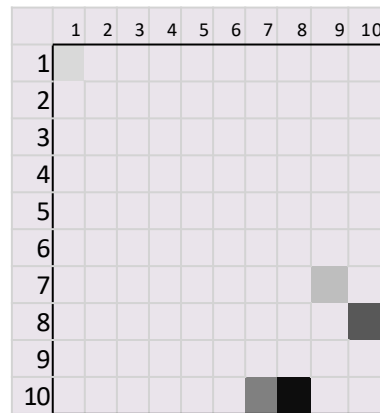
- Case study: WBC data with the Collocated Paired Coordinates (CPC-R) algorithm, for converting non-image data to images, and CNN algorithms for discovering the classification model in these images.
- Each image represents a single WBC data case, as a set of squares with a different level of intensities and colors
- The CPC-R algorithm is a modification of Collocated Paired Coordinates (CPC) algorithm
- The CPC-R algorithm, instead of connecting pairs (x_1, x_2) by arrows, uses the grey scale intensity from black for (x_1, x_2) and very light grey for (x_{n-1}, x_n) for cells. Alternatively, intensity of a color is used. This order of intensities allows full restoration

Cross-Domain Pollination

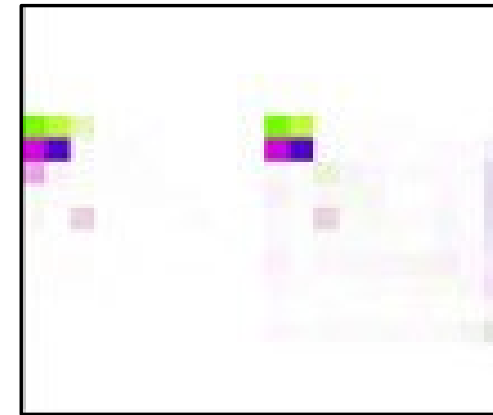
- Figure (a) shows the basic CPC-R image design and Figure (b) shows a more complex design of images, where a colored CPC-R visualization of a case is superimposed with mean images of the two classes, which are put side by side, creating double images.
- The advantage of CPC-R is in lossless visualization of n-D cases, and the ability to overlay them using heatmap with salient points discovered by the CNN model, for model explanation



6-D point (5,4,0,6,4,10) in Collocated Paired Coordinates.



(a) 10-D point (8, 10, 10, 8, 7,10, 9,7,1,1) in CPC-R.



(b) Visualization in colored CPC-R of a case superimposed with mean images of two classes put side by side.

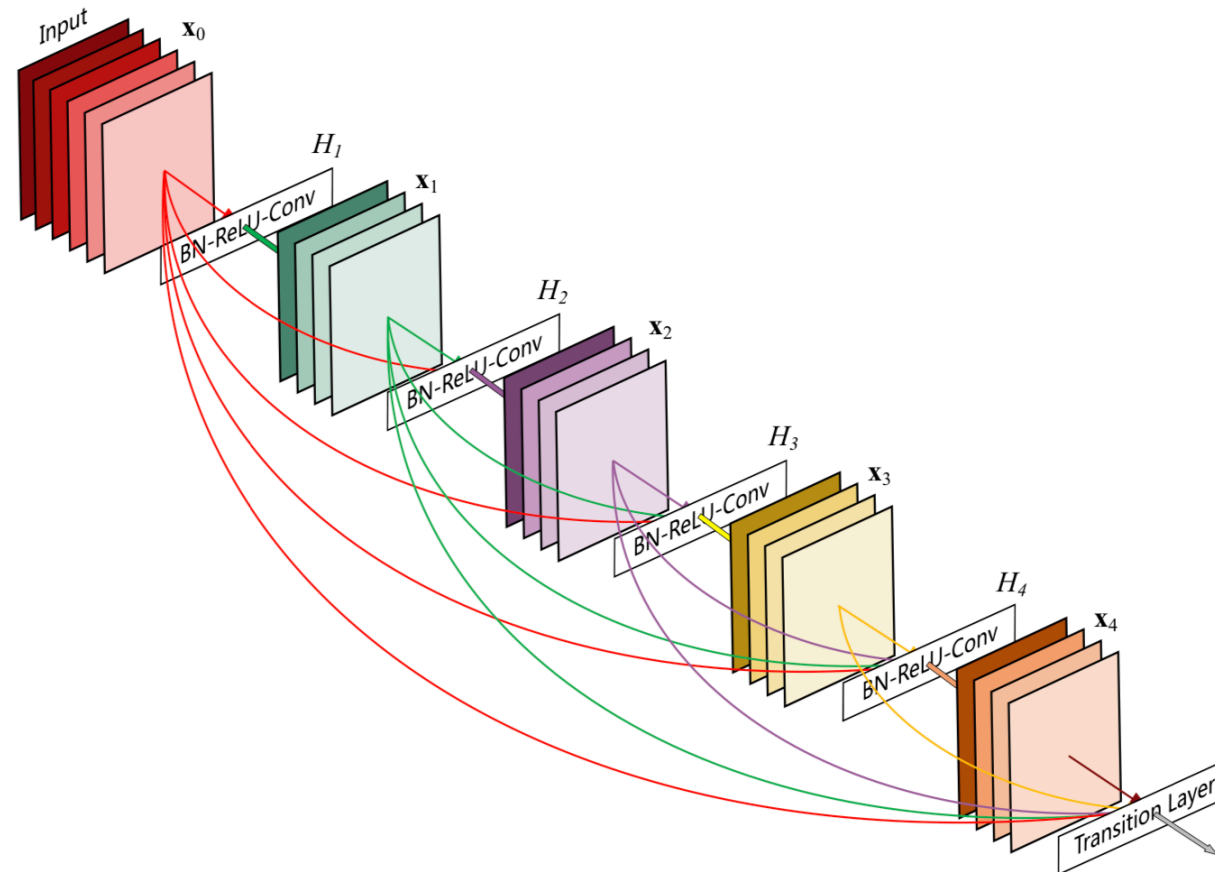
CPC-R visualization of non-image 10-D points.

Explanations are only as good as the model

- If there are varying degrees of fidelity in the interpretation then how do we add guards in implementation of interpretable models in the real world?
- Examples where the performance of the expert declines after results from a DL system are shown to them
- Users of ML systems are tempted to doubt their own judgement when information from a decision support system is shown

How to deal with extremely complex models

- What should explanation for very complex model look like?



Future Directions

- Creating simplified explainable models with prediction that humans can actually understand.
- “Downgrading” complex Deep Learning models for humans to understand them.
- Expanding visual and hybrid explanation models.
- Further developing explainable Graph Models.
- Further developing ML model in First Order Logic (FOL) terms of the domain ontology.
- Generating models with the sole purpose of explanation.
- Post-training rule-extraction.

Future Directions

- Expert-in-the-loop in the training and testing stages with auditing models to check generalizability of models to wider real-world data.
- Rich semantic labeling of a model's features that users can understand.
- Estimating the causal impact of a given feature on model prediction accuracy.
- Using new techniques such as counter-factual probes, generalized additive models, generative adversarial network technique for explanations.
- Further developing heatmap visual explanations of CNN by Gradient-weighted Class Activation Mapping and other methods with highlighting the salient image areas.
- Adding explainability to DL architectures by layer-wise specificity of the targets at each layer

References

- Cashman D., Humayoun SR., SR. Heimerl F, Park K. et al. Visual Analytics for Automated Model Discovery. arXiv preprint arXiv:1809.10782. 2018
- Endert A., Ribarsky W., Turkay C., Wong BW, .et al, The state of the art in integrating machine learning into visual analytics. In: Computer Graphics Forum 2017, Vol. 36, No. 8, 458-486.
- Guo Y., Liu Y., Oerlemans A., Lao S., Wu S., S. Lew S., Deep learning for visual understanding: A review. Neurocomputing. 2016, 187:27-48.
- Kovalerchuk B., Grishin V. Adjustable General Line Coordinates for Visual Knowledge Discovery in n-D data, Information Visualization, 18(1),2019, 3-32.
- Kovalerchuk B., Visual Knowledge Discovery and Machine Learning, Springer, 2018.
- Kovalerchuk B., Gharawi A., Decreasing Occlusion and Increasing Explanation in Interactive Visual Knowledge Discovery, In: S. Yamamoto and H. Mori (Eds.) Human Interface and the Management of Information. Interaction, Visualization, and Analytics, LNCS 10904, Springer, pp. 505–526, 2018.
- Kovalerchuk B., Neuhaus N., Toward Efficient Automation of Interpretable Machine Learning. In: 2018 IEEE International Conference on Big Data, pp. 4933-4940, Seattle, Dec. 10-13, 2018 IEEE.
- Kovalerchuk B., Schwing J. (eds), Visual and Spatial Analysis: Advances in Visual Data Mining, Reasoning, and Problem Solving, Springer, 2005,
- Kovalerchuk B., Agarwal B., Solving Non-image Learning Problems by Collocated Visualization and Deep Learning, CWU technical report, 2019.
- Koutra D., Di Jin, Y. Ning, C. Faloutsos. Perseus: An Interactive Large-Scale Graph Mining and Visualization Tool. Proc..of the VLDB
- Neuhaus N, Kovalerchuk B., Interpretable Machine Learning with Boosting by Boolean Algorithm, IEEE Joint 2019 8th Intern. Conf. on Informatics, Electronics & Vision, Spokane, WA, 2019, 307-311.
- Sacha D., Kraus M., Keim DA., Chen M.. VIS4ML: An Ontology for Visual Analytics Assisted Machine Learning. IEEE TVCG 2019 Jan;25(1):385-95.
- C. Seifert, A. Aamir, A. Balagopalan, D. Jain, A. Sharma, S. Grottel, S. Gumhold, Visualizations of deep neural networks in computer vision: A survey. In: Transparent Data Mining for Big and Small Data 2017, 123-144. Springer,
- Turkay C., Laramée R, Holzinger A., On the challenges and opportunities in visualization for machine learning and knowledge extraction: A research agenda. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction 2017, 191-198, Springer.
- Samek W., Montavon G., Vedaldi A., et al., (Eds), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019
- Sietzen S., Waldner M., Interactive Feature Visualization in the Browser, 2nd Workshop on Visualization for AI Explainability, IEEE VIS., 2019 Canada
- Vig J., Deconstructing BERT: Visualizing the Inner Workings of Attention, Analyzing the Design Space for Visualizing Neural Attention in Text Classification, 2nd Workshop on Visualization for AI Explainability, IEEE VIS., 2019
- Burkhardt D., Gigante S., Krishnaswamy S., Selecting the right tool for the job: a comparison of visualization algorithms, 2nd Workshop on Visualization for AI Explainability, IEEE VIS., 2019
- Lipton Z. The Mythos of Model Interpretability, Commun. of the ACM, 2018, 61,36-43.
- Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digital Signal Processing. 2018 Feb 1;73:1-5.

References

- Parra D., Valdivieso H., Carvallo A., Rada G., Verbert K., Schreck T., Analyzing the Design Space for Visualizing Neural Attention in Text Classification, 2nd Workshop on Visualization for AI Explainability, IEEE VIS., 2019, <https://observablehq.com/@clpuc/Britton>
- Lundberg S., A unified approach to explain the output of any machine learning model, <https://github.com/slundberg/shap>
- He S, Borji A, Mi Y, Pugeault N. What catches the eye? Visualizing and understanding deep saliency models. arXiv preprint arXiv:1803.05753. 2018
- Cornia M, Baraldi L, Serra G, Cucchiara R. SAM: Pushing the Limits of Saliency Prediction Models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2018, 890-1892.
- Li Y, Mou X. Saliency detection based on structural dissimilarity induced by image quality assessment model. Journal of Electronic Imaging. 2019 Apr;28(2):023025. <https://arxiv.org/pdf/1905.10150>
- Vergari A, Di Mauro N, Esposito F. Visualizing and understanding sum-product networks. Machine Learning. 2019 Apr 15;108(4):551-73. <https://arxiv.org/pdf/1608.08266>
- Schlegel U, Arnout H, El-Assady M, Oelke D, Keim DA. Towards a rigorous evaluation of XAI Methods On Time Series., 2019 ICCV Workshop on Interpreting and Explaining Visual Artificial Intelligence Models, <https://arxiv.org/abs/1909.07082>
- Schlegel U, Arnout H, El-Assady M, Oelke D, Keim DA. Towards A Rigorous Evaluation Of XAI Methods On Time Series. 2019, arXiv:1909.07082.
- Gunning D. Explainable Artificial Intelligence (XAI). DARPA, 2017, <https://www.darpa.mil/attachments/XAIProgramUpdate.pdf>
- Teredesai, A., Eckert, C., Ahmad, MA., Kumar, V. Explainable Machine Learning Models for Healthcare AI September 26, 2018 ACM Seminar
- Ahmad, MA., Eckert, C., Teredesai, A., Kumar, V., Explainable Models for Healthcare AI KDD London, United Kingdom August 19-23, 2018
- Ahmad, MA., Eckert, C., Teredesai, A., Kumar, V., Machine Learning in Healthcare International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2018, August 29 - September 01, 2018 Washington, DC, USA.
- Ahmad, MA., Eckert, C., Teredesai, A., Kumar, V., Interpretable Machine Learning in Healthcare 2018 IEEE International Conference on Healthcare Informatics 4-7 June, New York, NY, USA.
- Ahmad, MA., Eckert, C., Teredesai, A., The Challenge of Imputation in Explainable Artificial Intelligence Models. Macau, China AISafety Workshop at IJCAI 2019
- Lu, J. and Ester, M., 2019. An Active Approach for Model Interpretation. arXiv preprint arXiv:1910.12207.
- Lakkaraju, H. and Bastani, O., 2019. " How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. arXiv preprint arXiv:1911.06473
- Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, ACM KDD 2019, Lin, H.W., Tegmark, M. and Rolnick, D., 2017. Why does deep and cheap learning work so well?. Journal of Statistical Physics, 168(6), pp.1223-1247.
- Wolchover, N. and Reading, L., 2017. New theory cracks open the black box of deep learning. Quanta Magazine, 3.
- Shwartz-Ziv, R. and Tishby, N., 2017. Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810.