Explainable artificial intelligence in medicine

Presented to the Ottawa Hospital Research Institute
May 8, 2019

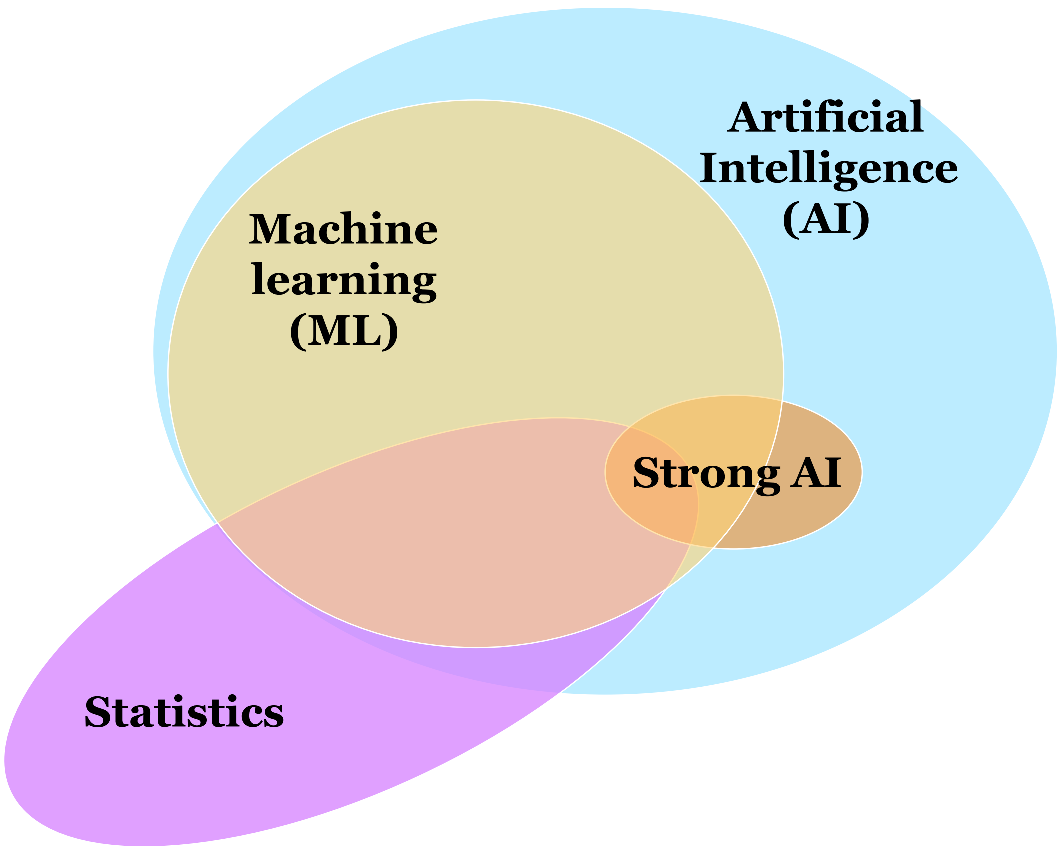
André Carrington, Ph.D, M.Math, P.Eng, CISSP

amcarrin uwaterloo.ca
amcarrin gmail.com

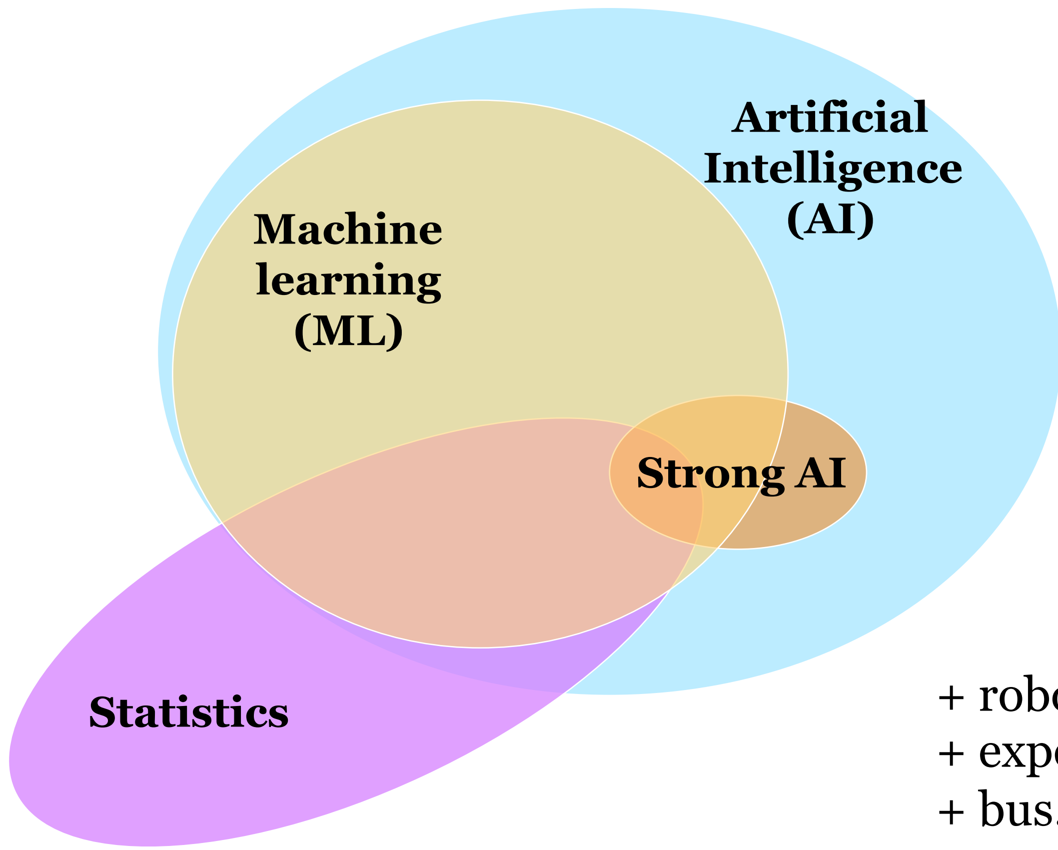
Agenda

- Context
- **Explainable AI**

AI versus statistics



AI versus statistics



+ robotics (control theory)
+ expert systems
+ bus. intelligence
+ operations research (game theory)

Which medical tasks can use AI?

- screening (for a single disease)
- prognosis (re cancer remission)
- …

AI in medicine

- **Detection**
- **Anomaly detection**
- **Prognosis**
- **Association (with features)**
- **Similarity search**
- **Influence (of cases)**
- …a subset of my list of examples

AI in medicine

- Detection
- Anomaly detection
- Prognosis
- Association (with features)
- Similarity search
- Influence (of cases)
- **Subgroup identification**
- **Modelling**
- **Scheduling/planning**
- **Treatment**
- **Simulation**
- **Diagnosis***

AI in medical imaging

- Detection
- Anomaly detection
- Prognosis
- Association (with features)
- Similarity search
- Influence (of cases)
- Subgroup identification
- Modelling
- Scheduling/planning
- Treatment
- Simulation
- Diagnosis*

- **Image segmentation**
- **Image registration**
- **Image denoising**
- **Image inpainting**
- **Surface reconstruction**
- **Image atlas creation**

AI in medicine

- Detection
- Anomaly detection
- Prognosis
- Association (with features)
- Similarity search
- Influence (of cases)
- Subgroup identification
- Modelling
- Scheduling/planning
- Treatment
- Simulation
- Diagnosis*

- Image segmentation
- Image registration
- Image denoising
- Image inpainting
- Surface reconstruction
- Image atlas creation
- **Genetic sequence alignment**
- **Protein sequencing**
- **Factor analysis**
- **Dimension reduction**

AI in medicine

- **Detection**: Do I have chronic kidney disease? (a specific disease)
  Which parts of an image indicate malignancy?
- **Anomaly detection**: Is there an unusual pattern of symptoms in the city?
- **Prognosis**: How long will I live with stage 4 lung cancer? Will I survive to year 5?
- **Association (with features)**: Which predictors matter?
- **Similarity search**: Which cases are similar?
- **Influence (of cases)**: Which cases influence the prediction most?

AI in medicine

- **Detection**: Do I have chronic kidney disease? (a specific disease)
  Which parts of an image indicate malignancy?
- **Anomaly detection**: Is there an unusual pattern of symptoms in the city?
- **Prognosis**: How long will I live with stage 4 lung cancer? Will I survive to year 5?
- **Association (with features)**: Which predictors matter?
- **Similarity search**: Which cases are similar?
- **Influence (of cases)**: Which cases influence the prediction most?
- **Subgroup identification**: What are the subgroups/clusters in data?
- **Modelling**: What is the best model of organ function?
- **Scheduling/planning**: What is the best schedule/plan for resource use? wait times?
- **Treatment**: Which therapy is best for me? (precision med; single/next step)
- **Simulation**: What is the best care pathway? multiple dose / longitudinal response?
- **Diagnosis***: I do not feel well, what is the problem? What tests should be ordered?
  (possibly any disease)

AI in medical imaging

- **Image segmentation**: finding borders or cell counting
- **Image registration**: aligning scans from different modalities
- **Image denoising**: removing noise
- **Image inpainting**: estimating an obstructed view
- **Surface reconstruction**: estimating a 3D surface from 2D images

AI in medicine

- **Image segmentation**: finding borders or cell counting
- **Image registration**: aligning scans from different modalities
- **Image denoising**: removing noise
- **Image inpainting**: estimating an obstructed view
- **Surface reconstruction**: estimating a 3D surface from 2D images
- **Image atlas creation**: creating an average/representative image/map, e.g., brain
- **Genetic sequence alignment**: align gene sequences for comparison
- **Protein sequencing**: identify the sequence of proteins
- **Factor analysis**: transforming data into independent factors
- **Dimension reduction**: reducing data into less factors

For those tasks what are the objectives of doctors?

- …

For those tasks what are the measurable objectives of doctors?

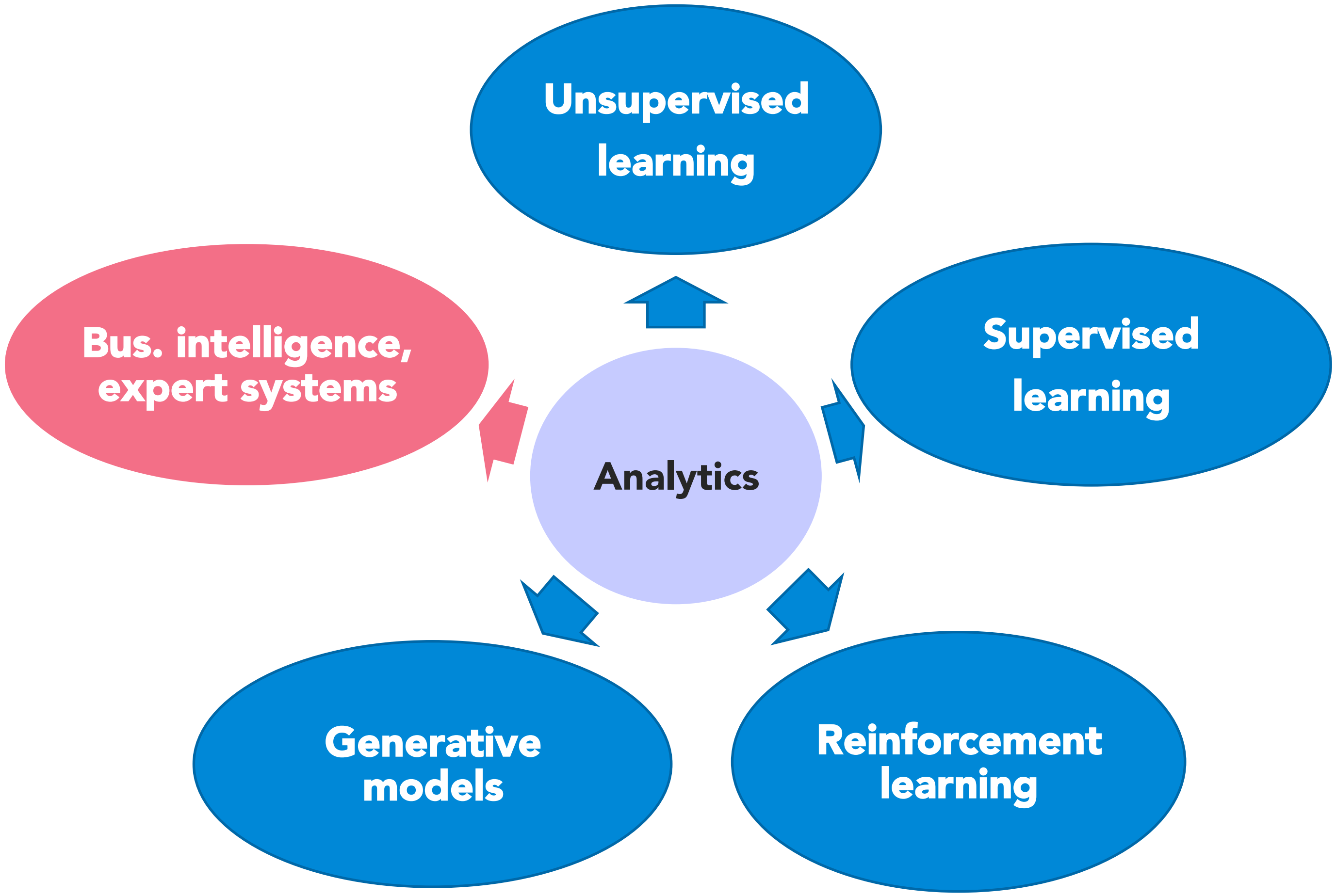
- cost-benefit
- discrimination
- calibration (or goodness of fit)
- probability of error for individual predictions
- parsimony
- interpretability
- understanding why (not how) it predicts outcome y for pt X
- understanding how they work & how they fail

Prediction: why (not how)

- Why was outcome y predicted for patient X?
  - feature A ?
  - patient Y ?
  - priors in data ?
  - model ? (a kind of prior)

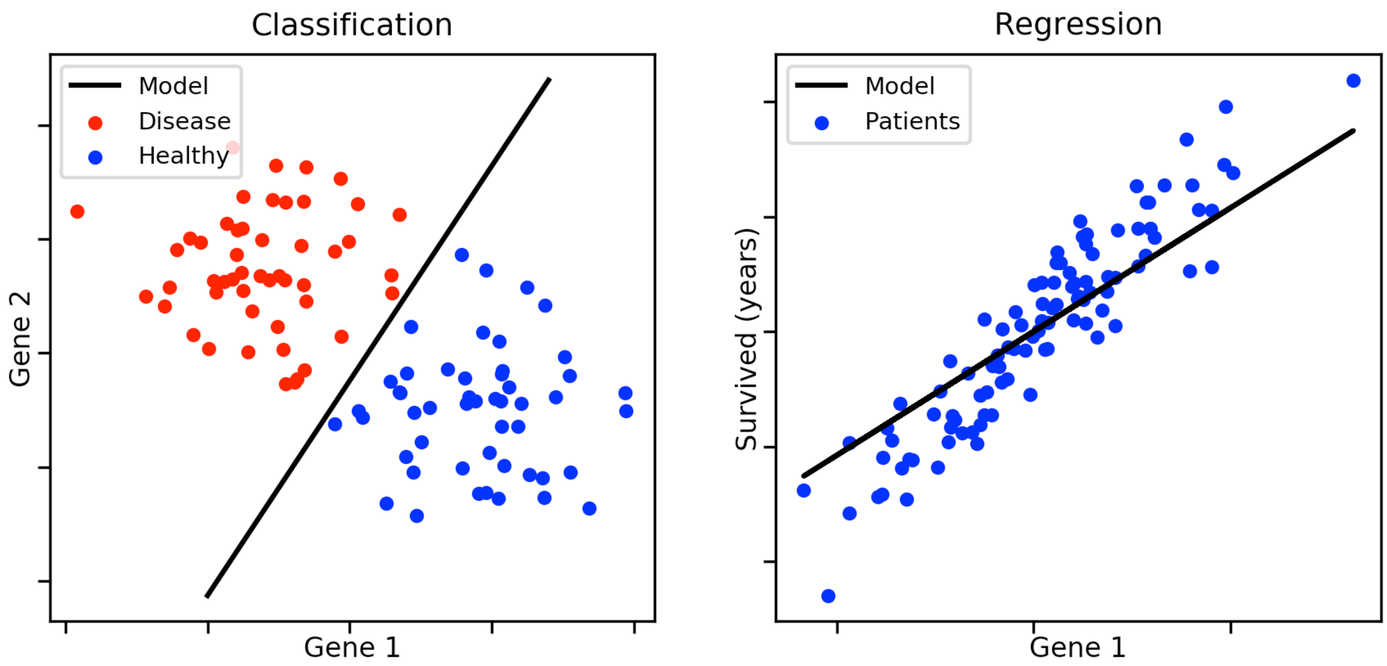
- To augment our thinking, explain to others, increase use

Analytics & learning (types)



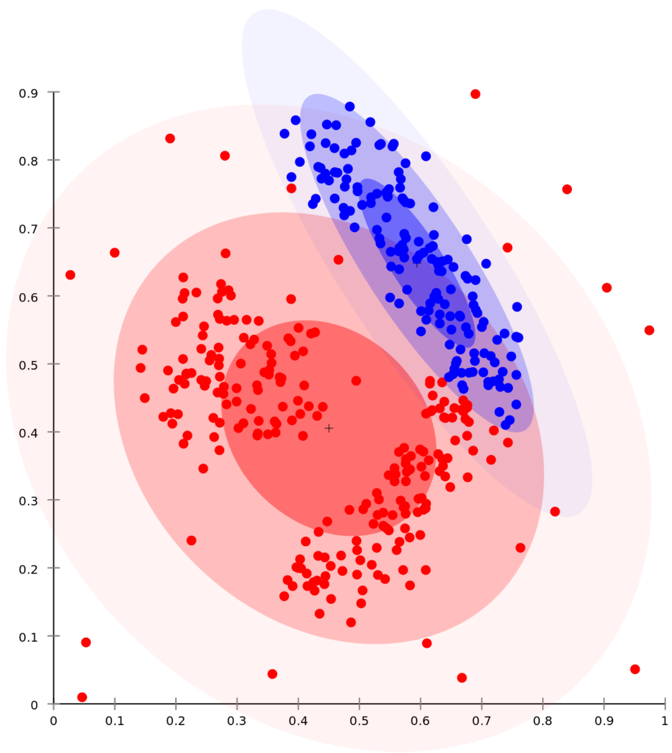
Supervised learning

- Imitating input/outcome patterns in data, with binary or continuous outcomes as ground truth.



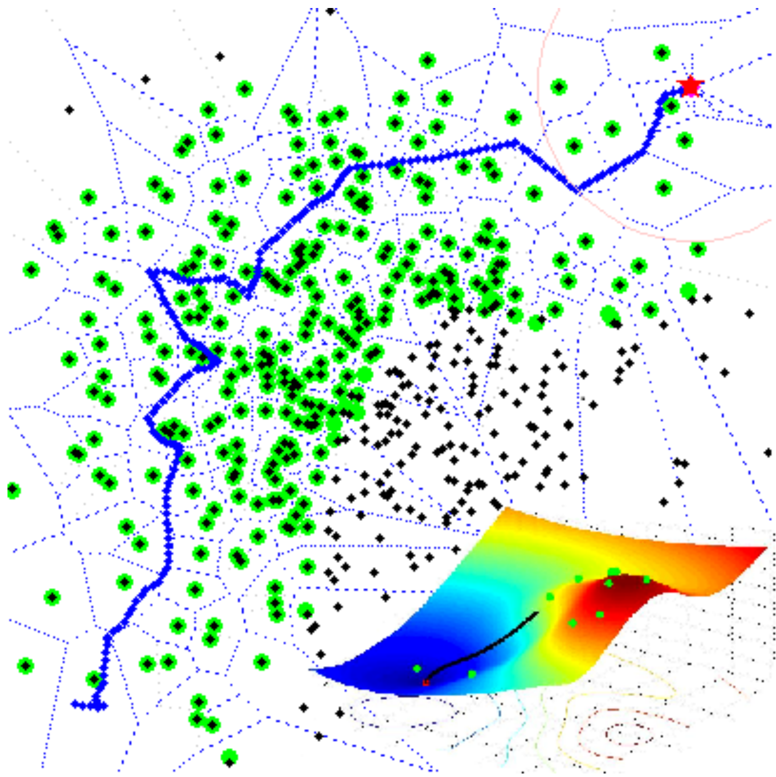
Unsupervised learning

- Organizing or transforming structure in data without ground truth.



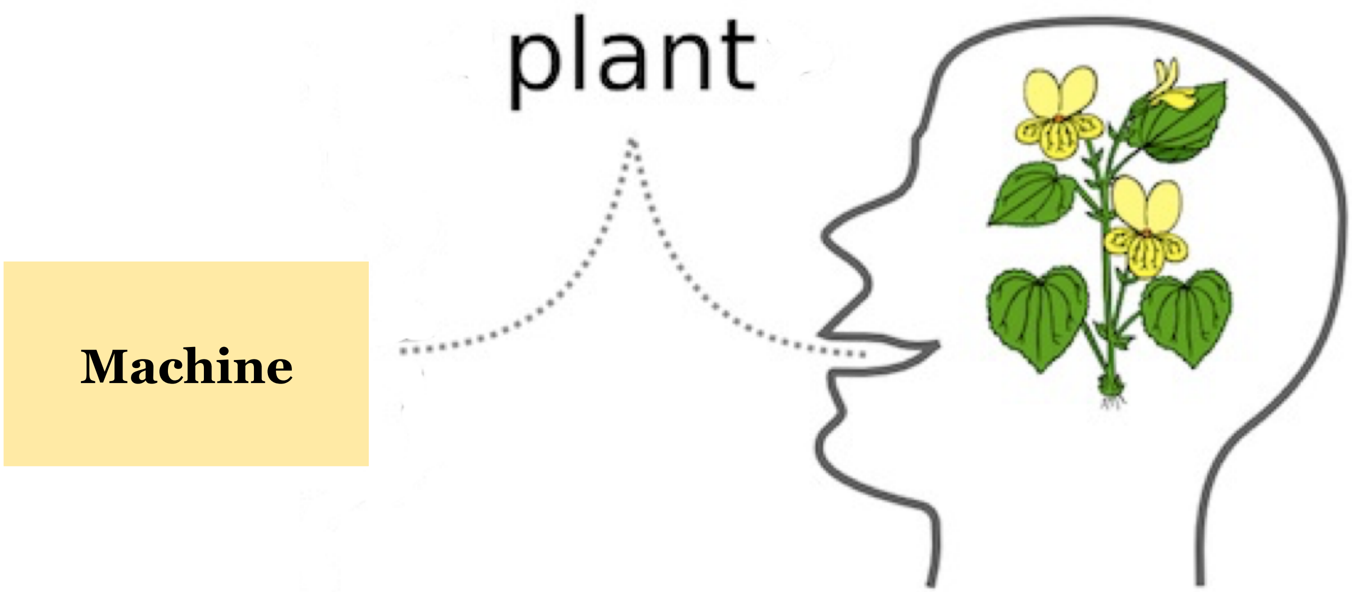
Reinforcement learning

- Navigating or planning a sequence of actions to maximize a reward/objective

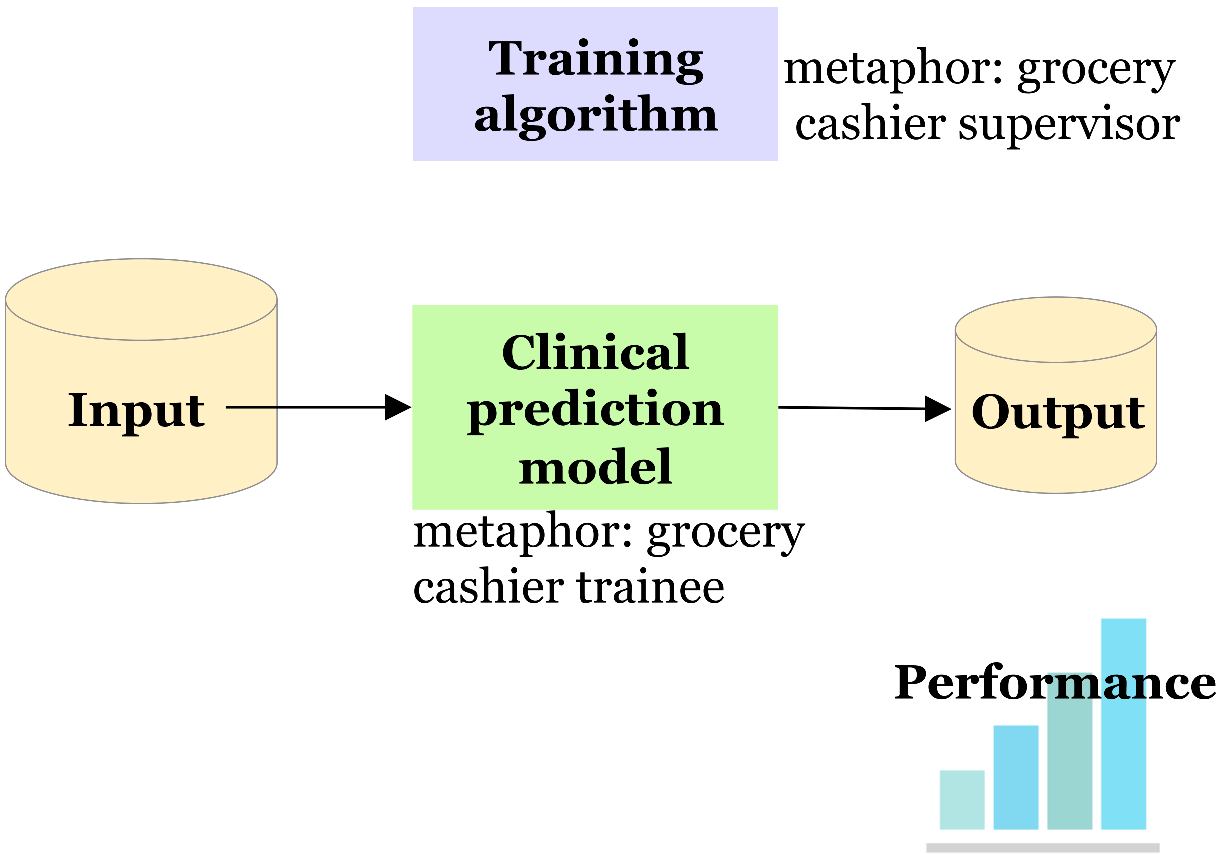


Generative models

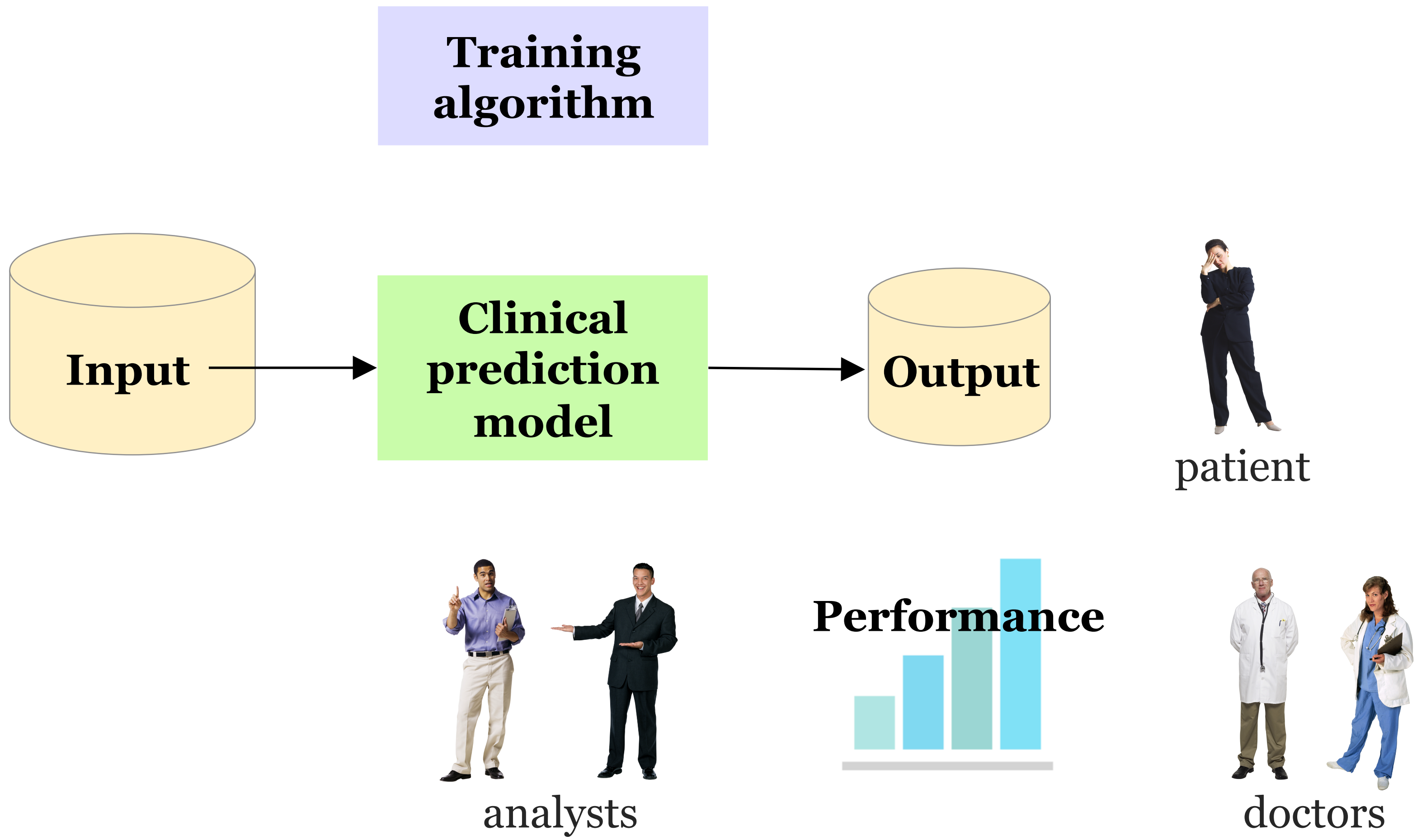
- Can generate the inputs. Can generate outputs that seem realistic, e.g., natural speech synthesis.



What needs explanation?



To whom do we explain?



Explainable AI

https://www.newyorker.com/cartoon/a19697

Why we need explainable AI

- To **understand** why a machine detects, recommends or predicts to effectively augment human decision-making
- To foster trust and use by doctors and patients
- For fairness, accountability and transparency in life decisions

Why we need explainable AI

- To understand why a machine detects, recommends or predicts to effectively augment human decision-making
- To foster trust and use by doctors and patients
- For fairness, accountability and transparency in life decisions
- To meet by EU law and general ethics
- To avoid law suits and maintain goodwill
- To understand how to improve accuracy/fit in subgroups

Fairness: goodness of fit, #samples

- gof = calibration
- Poor fit (accuracy) for subgroups with few samples
  - Race – melanoma rare for dark-skin
  - Pregnant women – clinical trial exclusion
  - Children – clinical trial exclusion
  - Elderly – clinical trial exclusion

Illustrating the previous point, how can I estimate the last row outcome?

wgt	height	pulse	age	sex	ACR	ckd
156	63	77	28	F	47	Y
150	65	60	46	F	219	Y
154	66	65	22	M	34	N
160	68	60	37	F	18	N

wgt	height	pulse	age	sex	ACR	ckd
166	70	82	31	M	33	?

Model interpretability ≈ XAI

Lipton (2016) describes two categories:
(different in timing/step and approach)

1. Transparency
2. Post-hoc interpretability, i.e., explanations

Model interpretability, XAI 1

1. Transparency – Lipton (2016) describes 3 parts
   a) Decomposability know influence of parts in data & model
   b) Simulatability mentally simulate & compute
   c) Algorithmic transparency know loss function behaviour

Model interpretability, XAI 2

1. Transparency – Lipton (2016) describes 3 parts
   a) Decomposability
   b) Simulatability
   c) Algorithmic transparency

2. Post hoc interpretability – Lipton (2016) describes 3 parts
   a) Natural language explanation rules, top words
   b) Visualization saliency maps
   c) Explanation by example similar case, class prototype

XAI for text and imaging

- Text: features ***are transparent***, e.g., topics, bag of words, n-grams, words
- Imaging: saliency maps ***are sometimes intuitive*** explanations
- Imaging: highly-engineered features (e.g., PCA) ***are sometimes intuitive***, e.g.:
  - lips smiling, e.g. MVU* (Weinberger *et al.*, 2006)
  - angle of face, e.g. LLE** (Ghodsi, 2006)

* maximum variance unfolding (MVU)
** local linear embedding (LLE)

XAI for numeric data

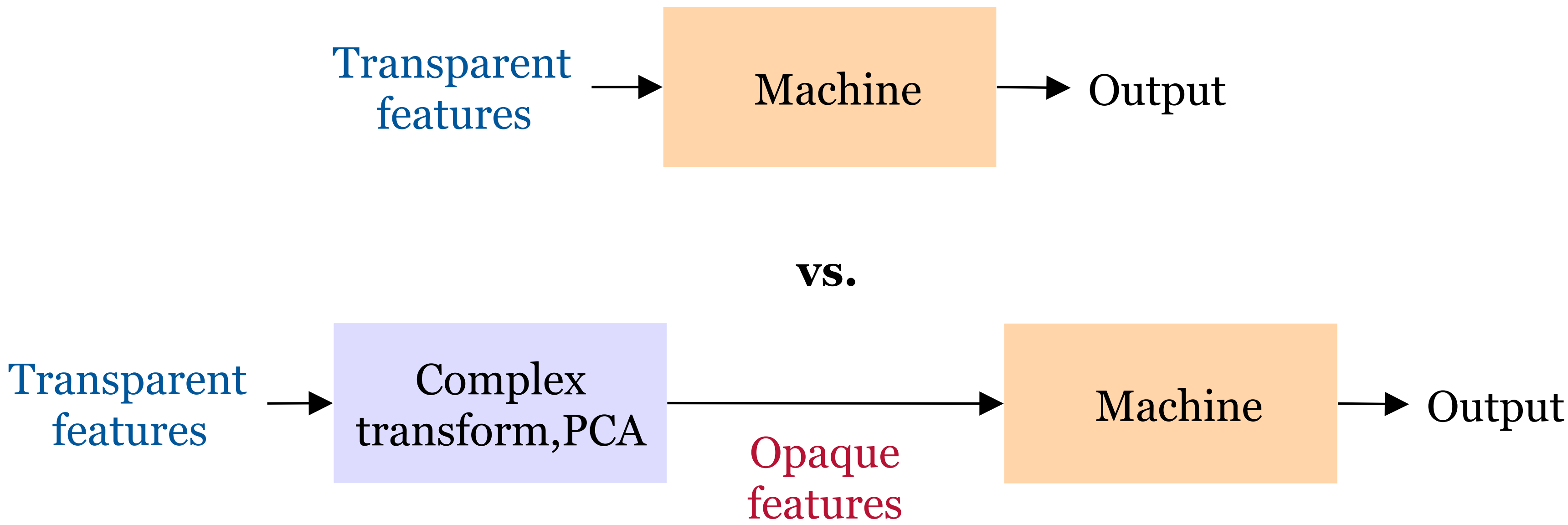
- Combinations of independent numeric features are ***usually not intuitive***

  3*height + diastolic blood pressure + 0.5*weight

- What does that mean? Is that clinically valid?
- Suppose it is risk. What kind of risk? Different from others?
- Concerns with physician numeracy (Estrada *et al.*, 1999; Hanoch *et al.*, 2010) and patient numeracy

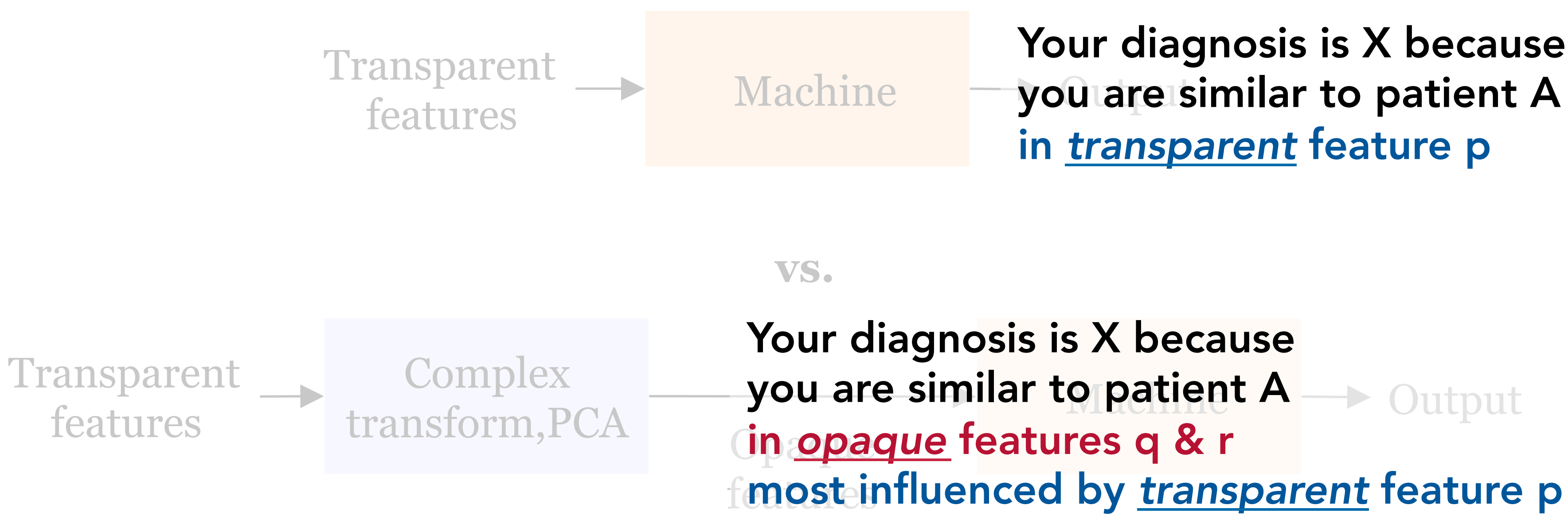
The need for transparent features

- To interpret output in one step or "inline"
- Holistic vs. piece-wise understanding.



The need for transparent features

- To interpret output in one step or "inline"
- Holistic vs. piece-wise understanding.



Your diagnosis is X because you are similar to patient A in *transparent* feature p

Your diagnosis is X because you are similar to patient A in *opaque* features q & r most influenced by *transparent* feature p

Transparent features defined

Carrington (2018) defines transparent features for **independent** Reals as transformations of originals we can mentally simulate in a set that avoids collinearity

Transparent	Not Transparent
shift, scale, flip, magnitude (abs)	shear, rotate
invert (1/x), square, order of magnitude (log x)	PCA, ICA, FA, MDS, t-SNE, ISOMAP, KPCA etc.
squash (tanh), bin, top-code, bottom-code	random projections

A false dichotomy

Lipton (2017) discusses two options:

- Linear models with highly-engineered features vs.
- Deep models with transparent* features

and trade-offs between them.

*Lipton refers to "raw or lightly processed" features

There are more options

Lou, Caruana *et al.* (2012) categorize models as:

1. Linear — most intelligible
2. Generalized linear models, GLM
3. Additive
4. Generalized additive models, GAM
5. Full complexity (deep) — least intelligible

Examples

2. Generalized linear models, GLM
   - Logistic regression
   - Piecewise linear models or splines, MARS (Friedman, 1991)

4. Generalized additive models, GAM (Hastie & Tibshirani, 1990)
   - Fractional polynomial regression (Royston & Altman, 1994)
   - Transparent kernels+support vector machines (Carrington, 2018)
     - which can be used in deep kernel learning (Wilson, 2014)

Explainable models (assumption)

Explainable:

- Decision trees, rules
- Bayesian networks
- Logistic regression

False positives in model assumptions

Explainable:

- Decision trees, rules
- Bayesian networks
- Logistic regression
- except for (Lipton, 2016; Carrington, 2018)
  - too many features, nodes, levels
  - collinear features
  - opaque features

False negatives in model assumptions

Explainable:

- Decision trees, rules
- Bayesian networks
- Logistic regression
- Support vector machines (Barbella *et al.*, 2016; Poulin *et al.*, 2006; Carrington, 2018)
- Neural networks (Montavon *et al.*, 2017)
- Random forests (Breiman, 2001)

Be wary of the trade-off assumption

- Assumed trade-off:

  Accuracy vs. interpretability (or explainability)

- For some problems, big data trump models (Banko & Brill, 2001) and simple models trump complex ones (Halevy, Norvig & Pereira, 2009).
- Logistic regression outperforms random forests in prediction of CVD mortality & heart failure type (Austin, 2012 & 2013).
- Explainable/finite kernels in SVM* outperform the infinite Gaussian RBF kernel on four heterogeneous clinical data sets without images or text (Carrington, 2014).

*support vector machines

If the accurate/explainable trade-off were true, then…

1. Plots of accuracy versus explainability (# support vectors) would show a negative sloped trend: linear or exponential
2. More explainable (finite) kernels, e.g., Mercer sigmoid, would achieve less accuracy than the infinite Gaussian RBF

- Neither of these phenomena show in the following plots (Carrington, 2018).
- The relation between accuracy, kernel width and SVM box constraint is more complicated (Ben-Hur *et al.*, 2010).

Accuracy vs interpretability (SV): Hep

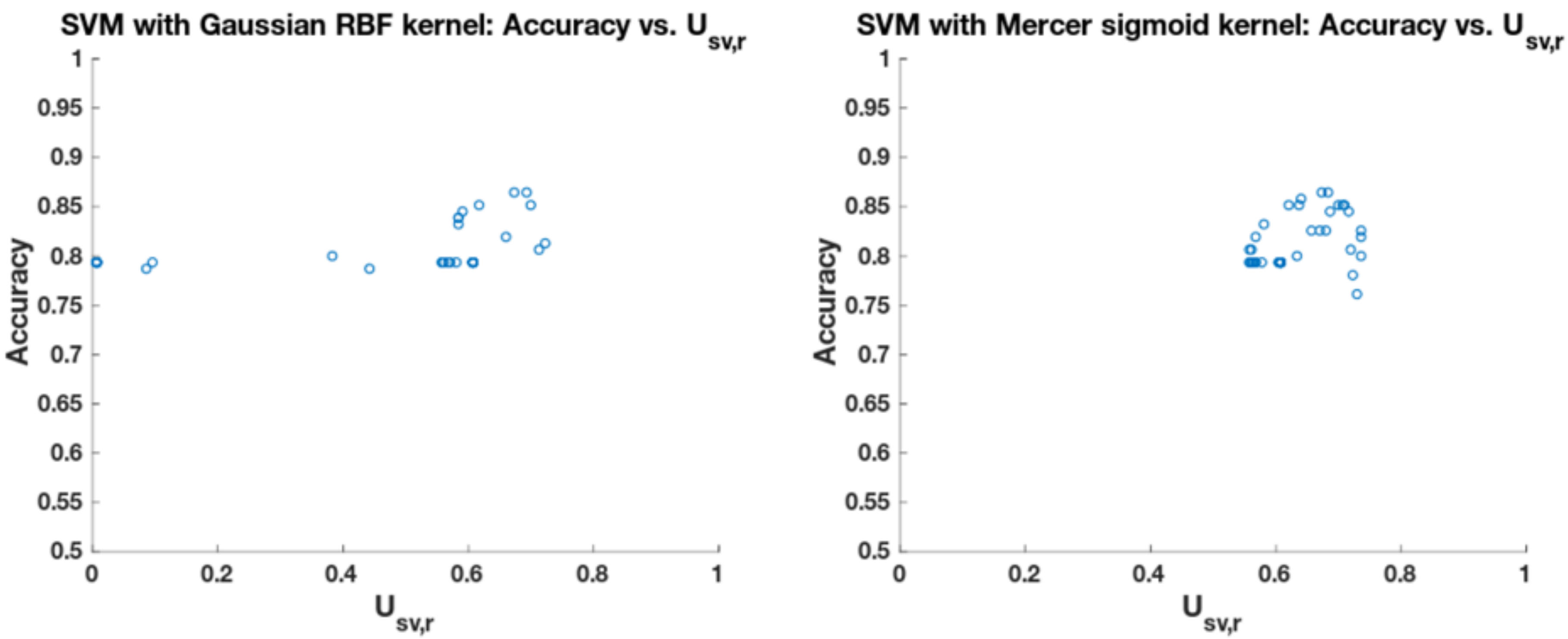
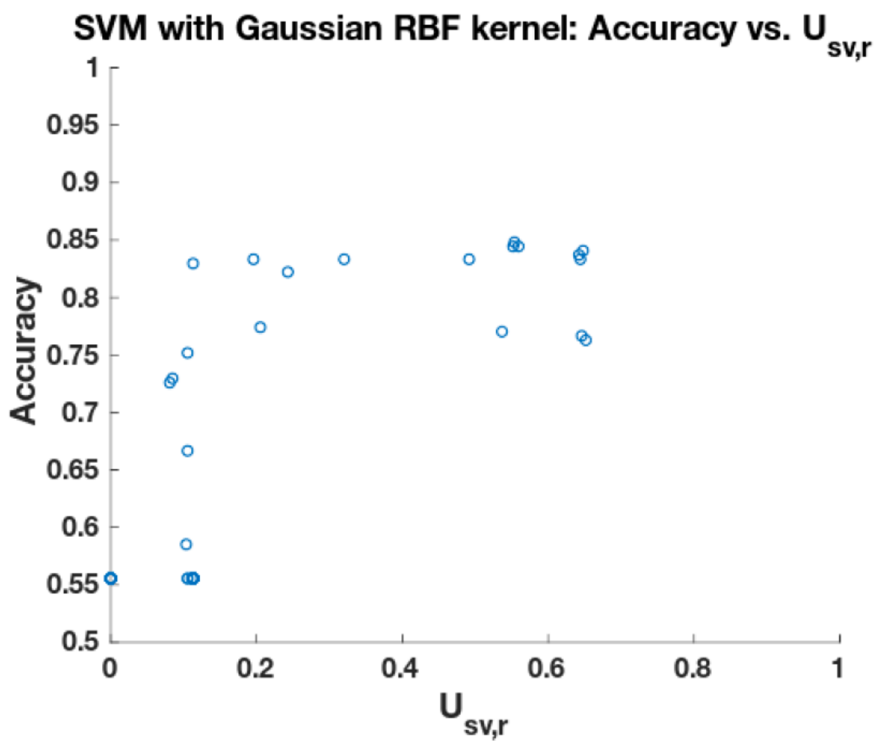


Figure 6.3: In classification with the Hepatitis data set there is a less than 5% sacrifice in inherent model interpretability for the highest accuracy.

Accuracy vs interpretability (SV): Heart



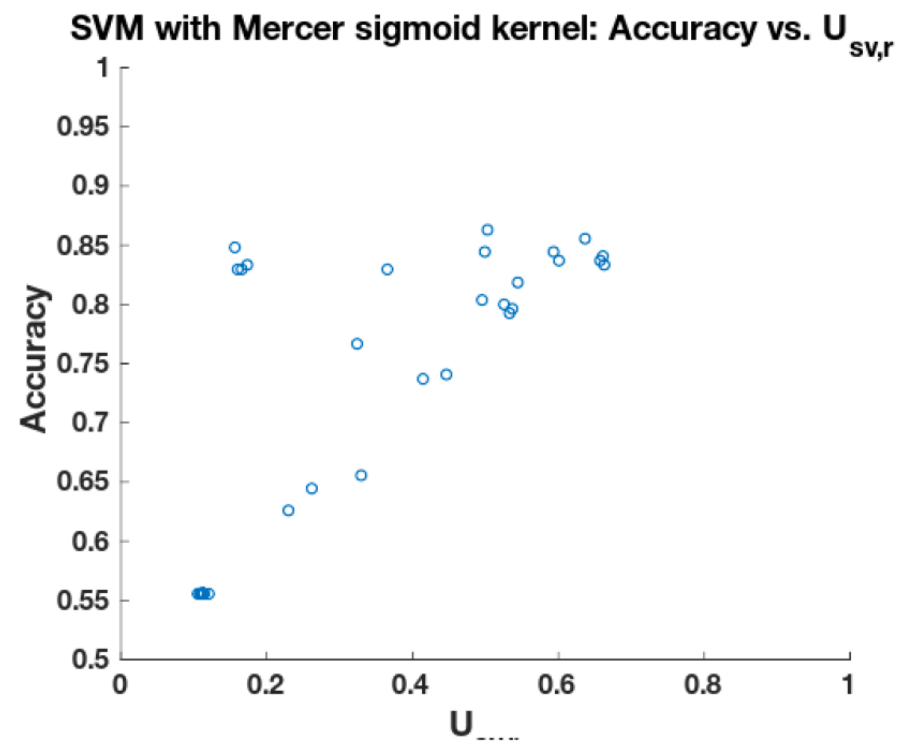


Figure 6.4: In classification with Statlog Heart data there are points with high accuracy and high inherent model interpretability, with minimal sacrifice, 1% and 2%, respectively.

Accuracy vs interpretability (SV): Liver (classically incorrect nonclinical target)

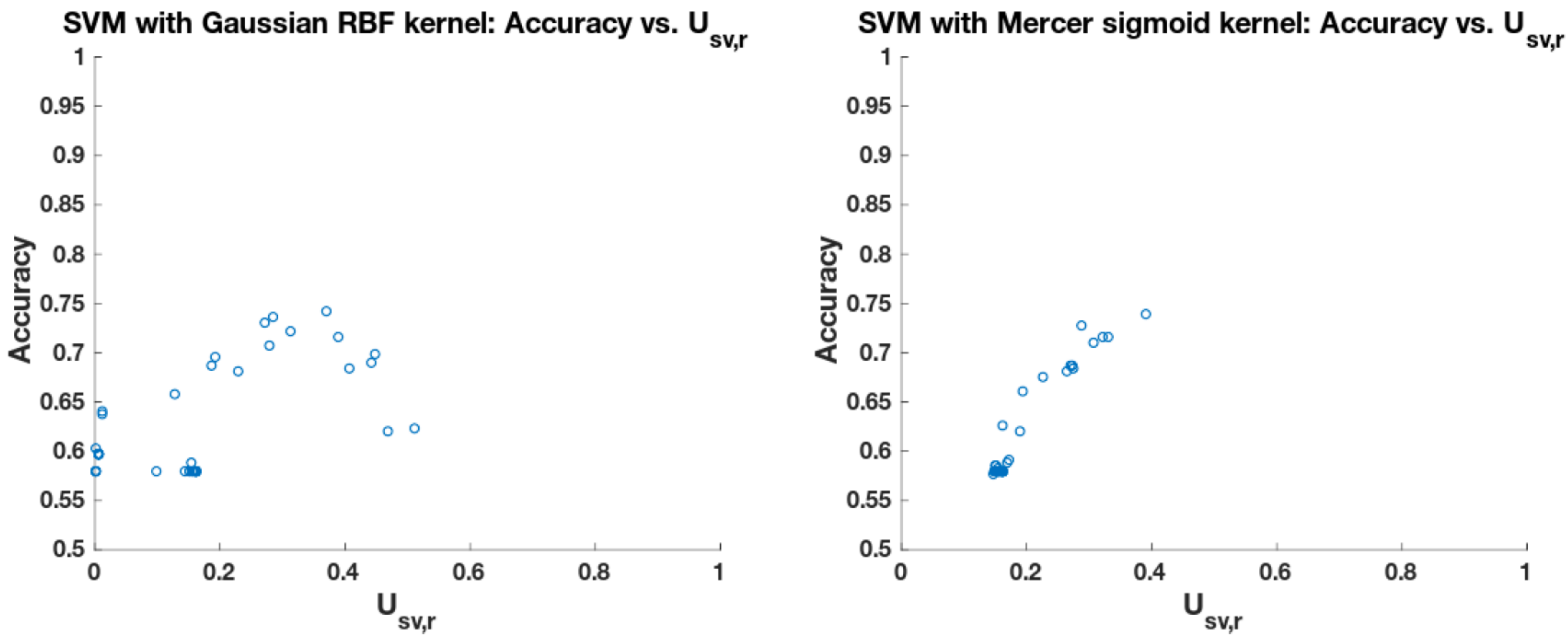
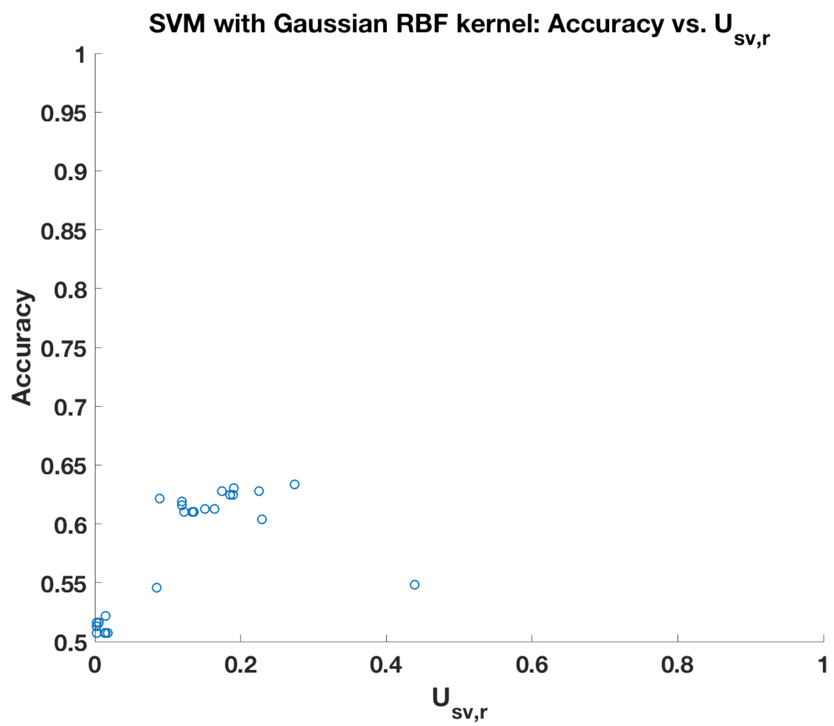
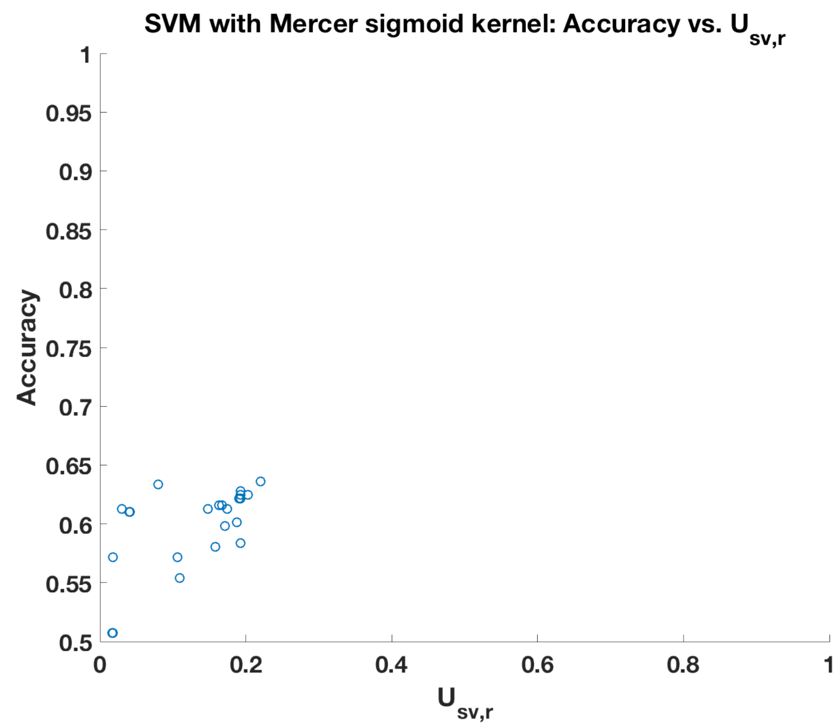


Figure 6.5: In classification with the Bupa liver data set there is a 20% and 0% sacrifice, respectively, in inherent model interpretability for the highest accuracy.

Accuracy vs interpretability (SV): Liver (rarely used correct clinical target)

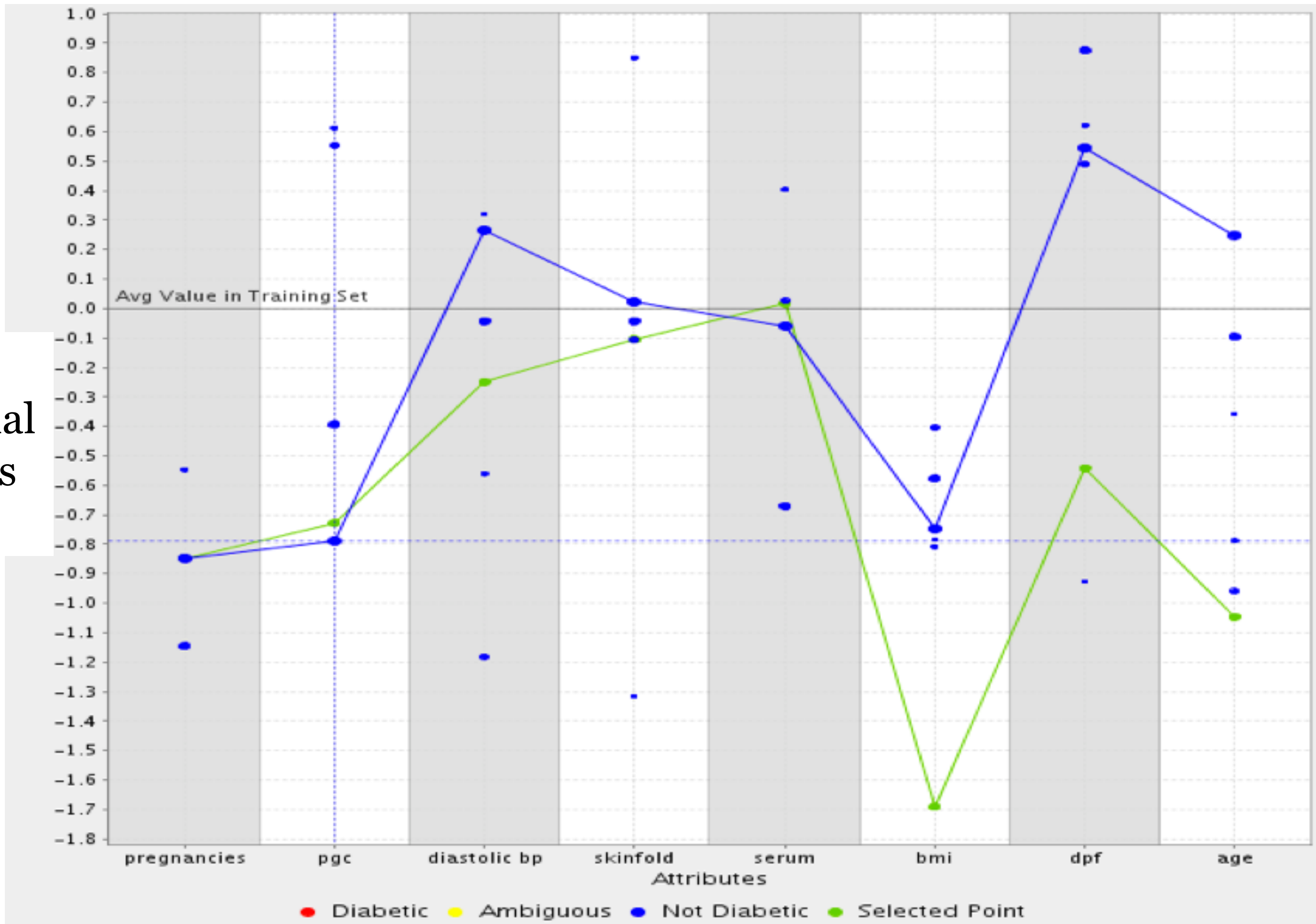




Explaining SVM results (Barbella *et al.*, 2009)

plot for selected patient (green)

most influential instances (blue)



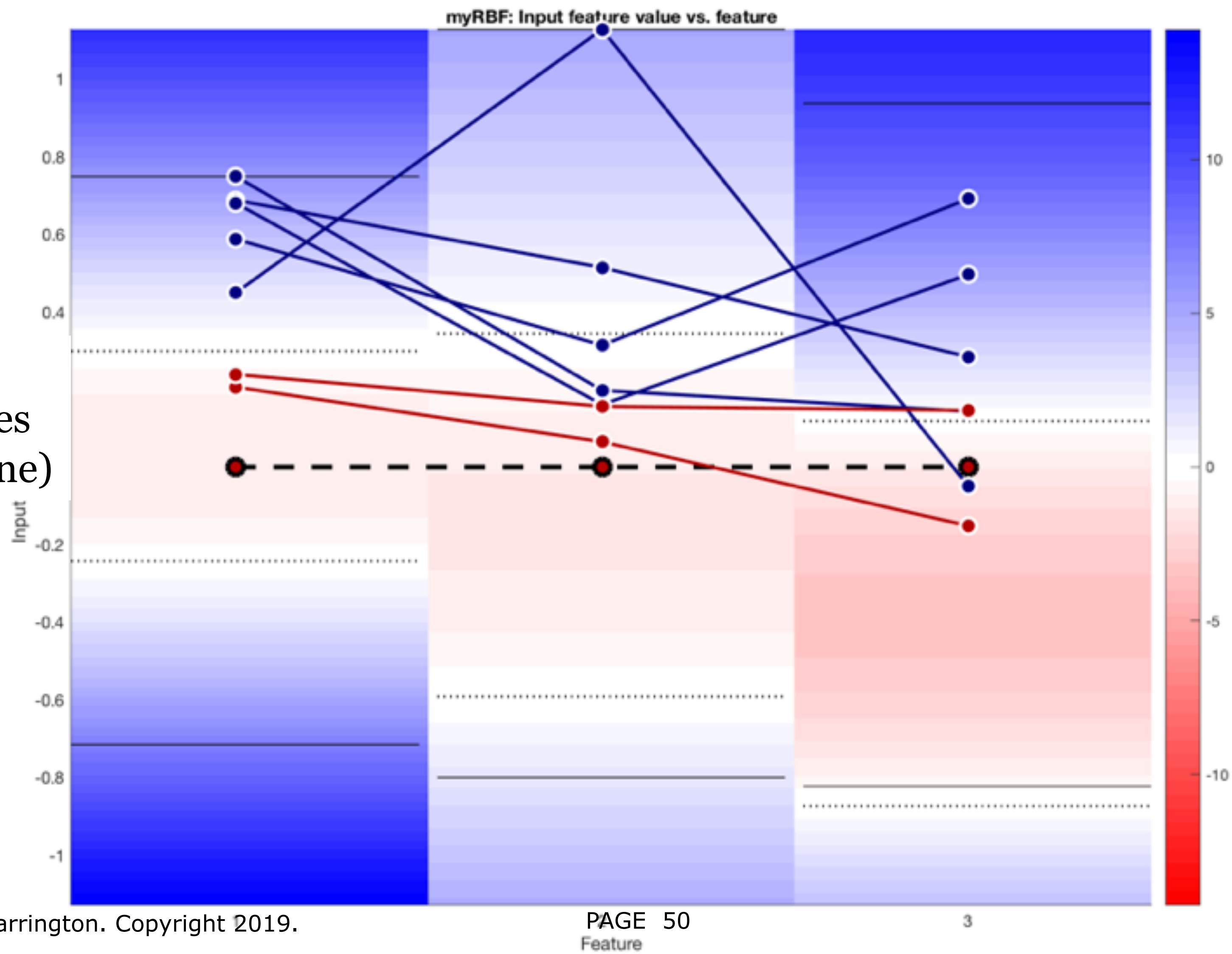
no class boundary

no counter-factuals (red)

An improved view (Carrington, 2018)

gradients show influence toward classes

class boundaries (dotted line)



includes counter-factuals

includes data limits

Explanations (Miller, 2017)

- In social science, explanations are:
  - **Contrastive** – why A and not B?
  - **Selected** (vs. complete)
  - **Causal** (vs. probabilistic)
  - **Social** – involving the beliefs of explainer & explainee

References

1. Lipton ZC (2016). The mythos of model interpretability. arXiv preprint arXiv:1606.03490.
2. Weinberger KQ, Saul LK (2006). An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In AAAI, volume 6, pages 1683–1686.
3. Ghodsi A (2006). Dimensionality reduction a short tutorial. University of Waterloo.
4. Estrada C, Barnes V, Collins C, Byrd JC (1999, Aug). Health literacy and numeracy. JAMA. 282(6):527.
5. Hanoch Y, Miron-Shatz T, Cole H, Himmelstein M, Federman AD (2010 Jul). Choice, numeracy, and physicians-in-training performance: The case of Medicare Part D. Health Psychology. 29(4):454.
6. Carrington, AM (2018). Kernel methods and measures for classification with transparency, interpretability and accuracy in health care. UWSpace. (Doctoral dissertation, University of Waterloo).
7. Lou Y, Caruana R, Gehrke J (2012). Intelligible models for classification and regression. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 150–158. ACM.
8. Friedman JH (1991). Multivariate adaptive regression splines. The annals of statistics. 19(1):1-67.
9. Hastie T, Tibshirani R (1990). Generalized additive models. Chapman and Hall, CRC Press.
10. Royston P, Altman DG (1994, Sep). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. In Journal of the Royal Statistical Society: Series C (Applied Statistics) 43(3):429-53.

References

11. Wilson AG (2014). *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes* (Doctoral dissertation, University of Cambridge).

12. Barbella D, Benzaid S, Christensen JM, Jackson B, Qin XV, Musicant DR (2009, Jul). Understanding Support Vector Machine Classifications via a Recommender System-Like Approach. In DMIN (pp. 305-311).

13. Poulin B, Eisner R, Szafron D, Lu P, Greiner R, Wishart DS, Fyshe A, Pearcy B, MacDonell C, Anvik J (2006, Jul). Visual explanation of evidence with additive classifiers. In Proceedings Of The National Conference On Artificial Intelligence (Vol. 21, No. 2, p. 1822). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press.

14. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR (2017, May). Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition. 65:211-22.

15. Breiman L (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science, 16(3):199–231.

16. Banko M, Brill E (2001, July). Scaling to very very large corpora for natural language disambiguation. In Proceedings of the 39th annual meeting on association for computational linguistics (pp. 26-33). Association for Computational Linguistics.

17. Halevy A, Norvig P, Pereira F (2009). The unreasonable effectiveness of data. Google.

References

18. Austin PC, Lee DS, Steyerberg EW, Tu JV (2012). Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? Biom J. 54(5):657-73.
19. Austin PC, Tu JV, Ho JE, Levy D, Lee DS (2013, Apr). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of clinical epidemiology. 66(4):398-407.
20. Carrington AM, Fieguth PW, and Chen HH (2014). A new mercer sigmoid kernel for clinical data classification. In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, pages 6397–6401.
21. Ben-Hur A, Weston J (2010). A user's guide to support vector machines. In Data mining techniques for the life sciences (pp. 223–239). Springer.
22. Miller T (2017) Explanation in artificial intelligence: Insights from the social sciences. arXiv preprint arXiv:1706.07269.

Questions?

André Carrington
amcarrin uwaterloo.ca
amcarrin gmail.com