Most “Open-Source” AI **Isn’t**. And What We Can Do About That

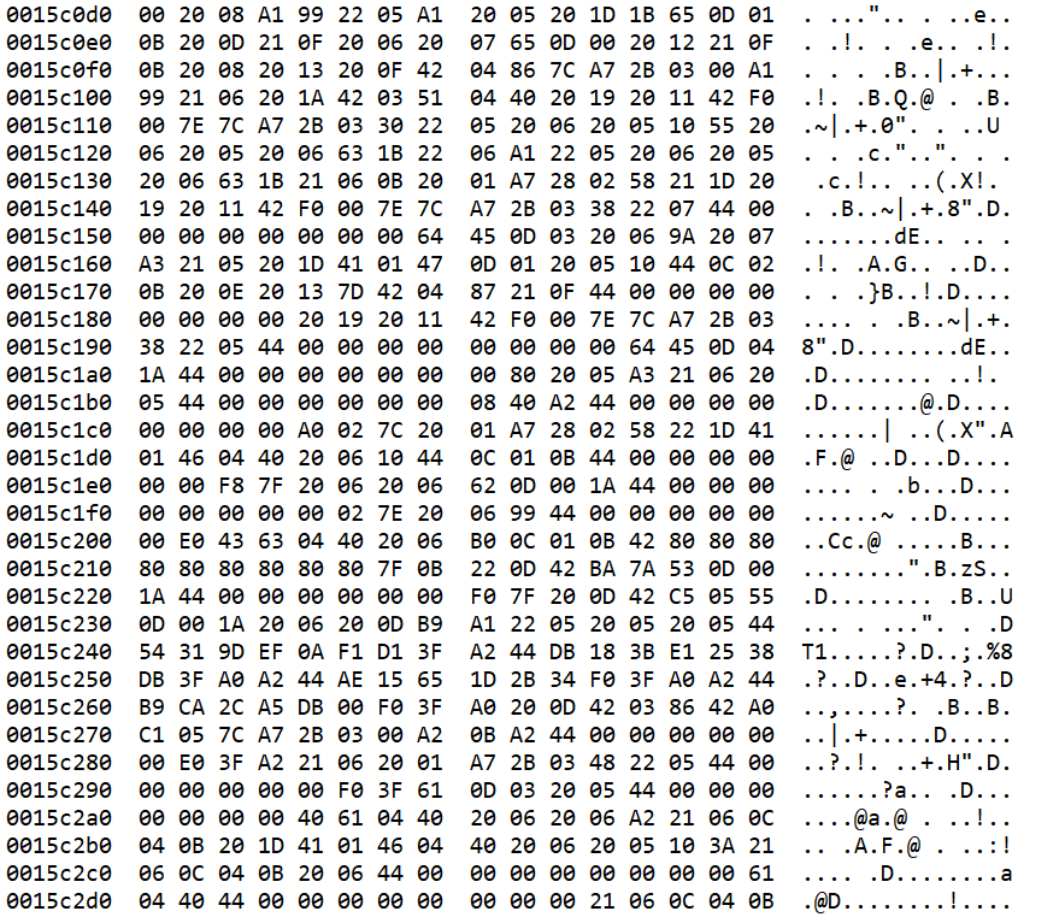
Christopher J. Hazard, PhD

HOW/SO™

A Story of Contract Software Development

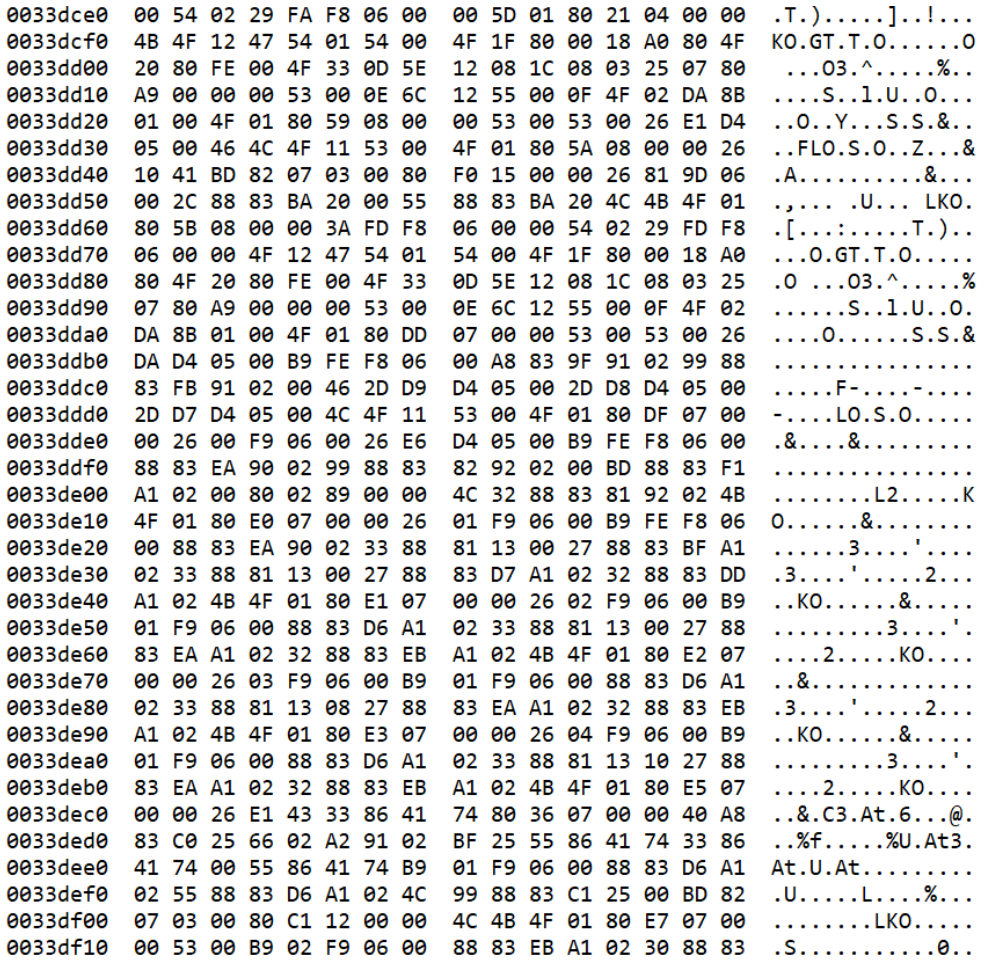
- Definition and requirements
- Framework selected
- Representative data given
- Contractor selected and paid

- You get: A binary executable
- You could poke at it, edit the assembly code, or ask the contractor to try to fix issues



A Story of Machine Learning Development

- Definition and requirements
- Framework selected
- Representative data given
- Compute selected and paid



- You get: A bunch of weights
- You could poke at it, adjust training data, loss function, and architecture to try to fix issues


They're the same picture.

What Is Free/Open-Source Software?

- Free use
- Free distribution
- Free modification and understanding!

The Four Essential Freedoms of Free Software – FSF

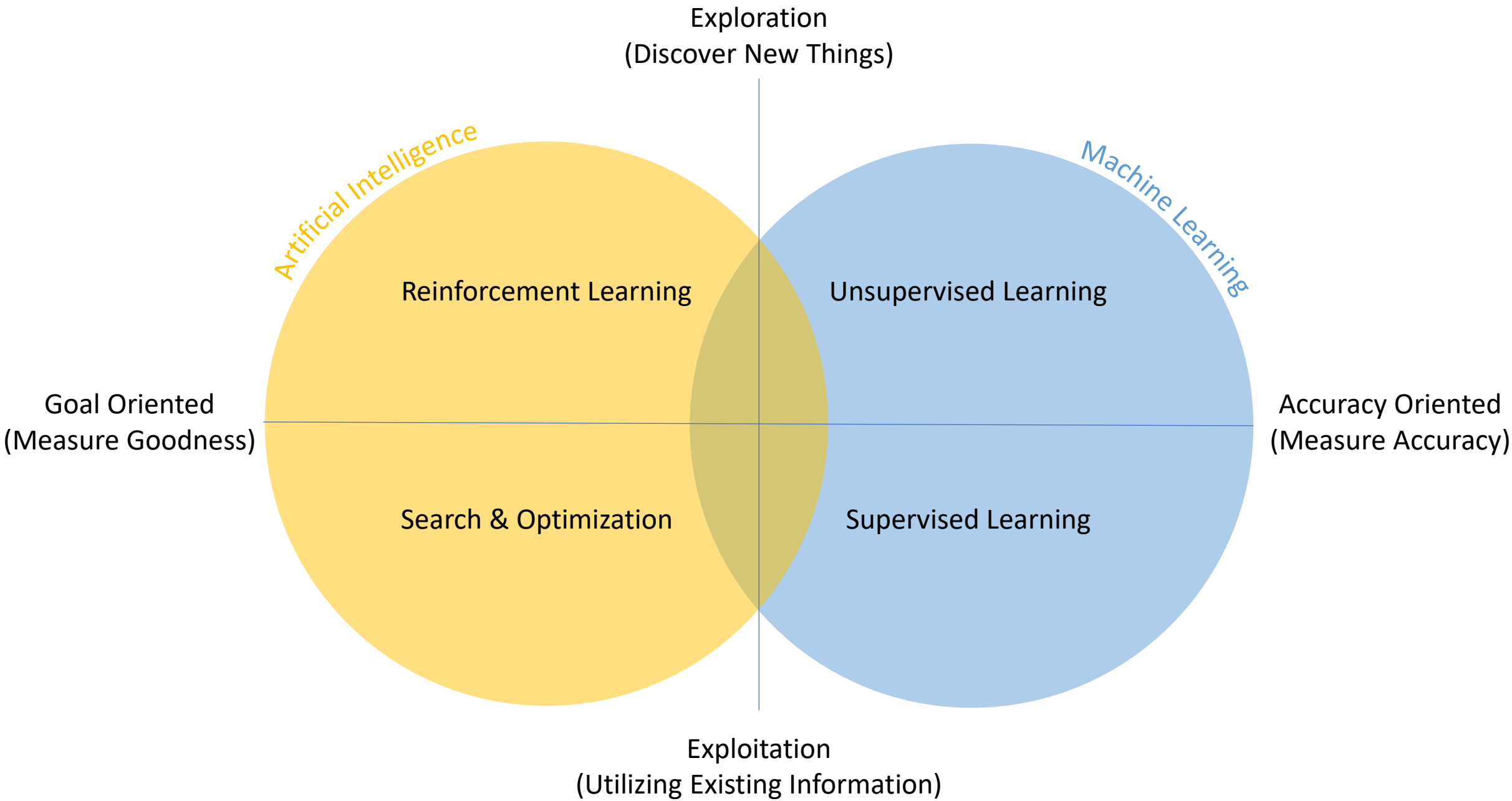
The freedom to study how the program works, and change it so it does your computing as you wish (freedom 1). Access to the source code is a precondition for this.

The Open Source Definition – OSI

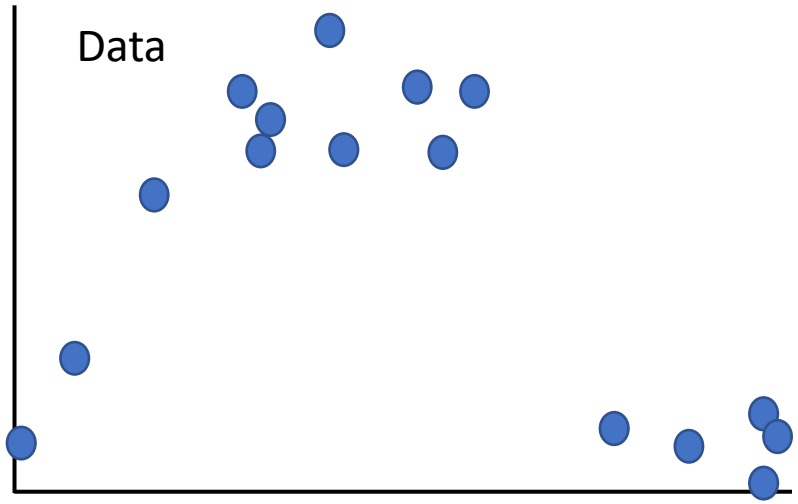
Source code: The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well-publicized means of obtaining the source code for no more than a reasonable reproduction cost preferably, downloading via the Internet without charge. **The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed.** Intermediate forms such as the output of a preprocessor or translator are not allowed.

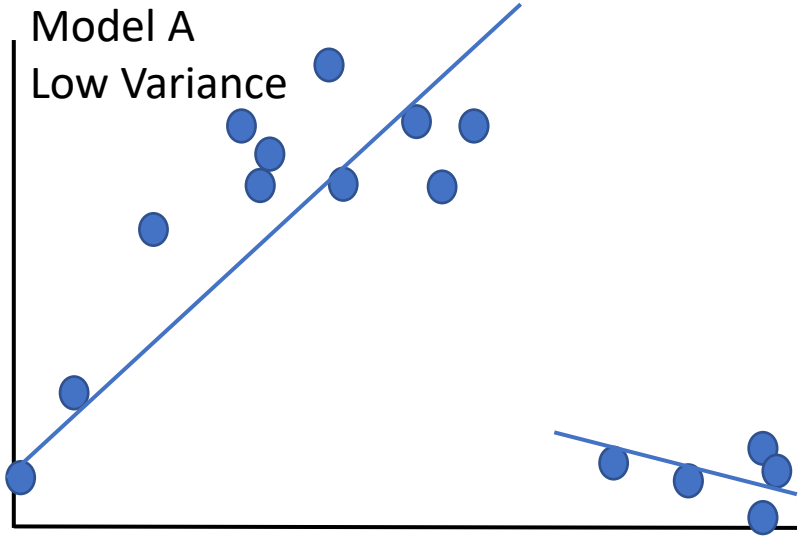
- ML: Programming with data
- ML: Compression + generalization

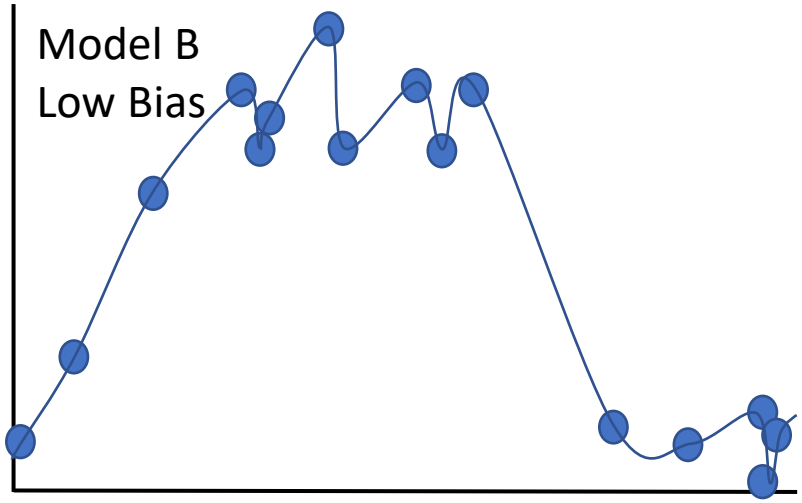
- AI: Hard computer science problems that haven’t been solved yet
- AI: Machines doing intelligent things

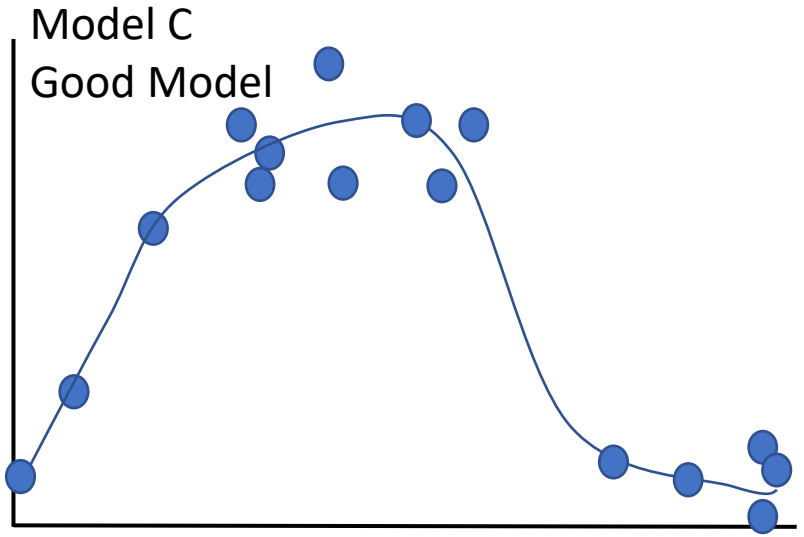

Exploration
(Discover New Things)
Artificial Intelligence
Machine Learning
Reinforcement Learning
Unsupervised Learning
Goal Oriented
(Measure Goodness)
Accuracy Oriented
(Measure Accuracy)
Search & Optimization
Supervised Learning
Exploitation
(Utilizing Existing Information)

Machine Learning: Function Approximators

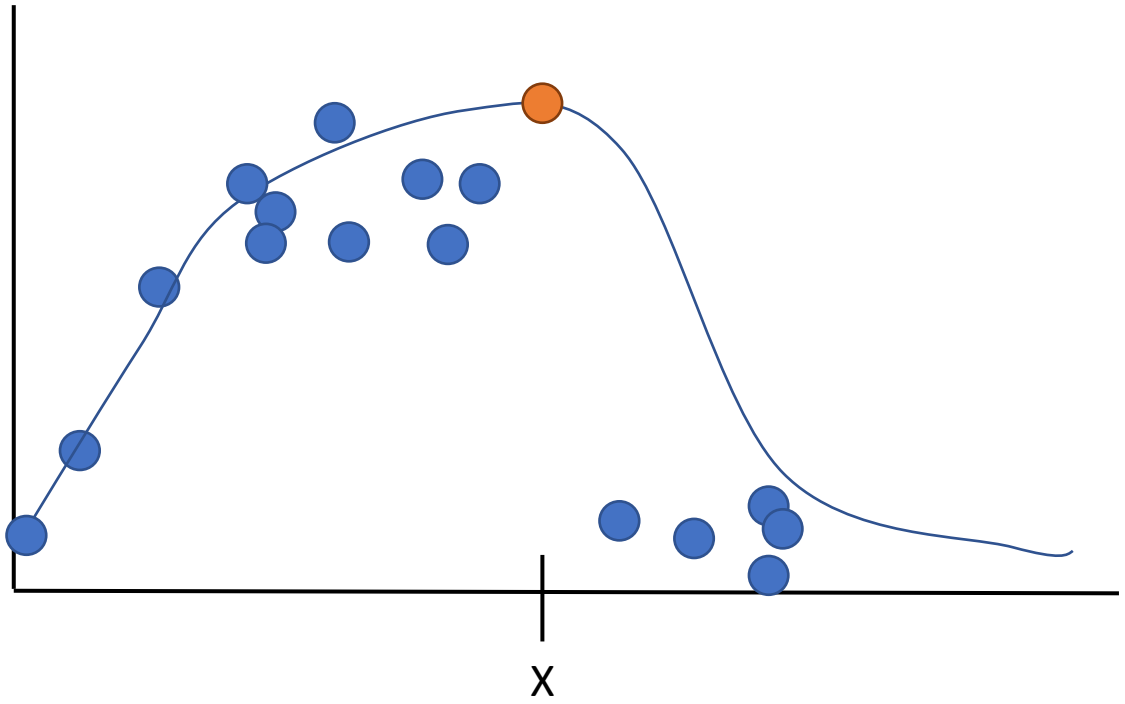




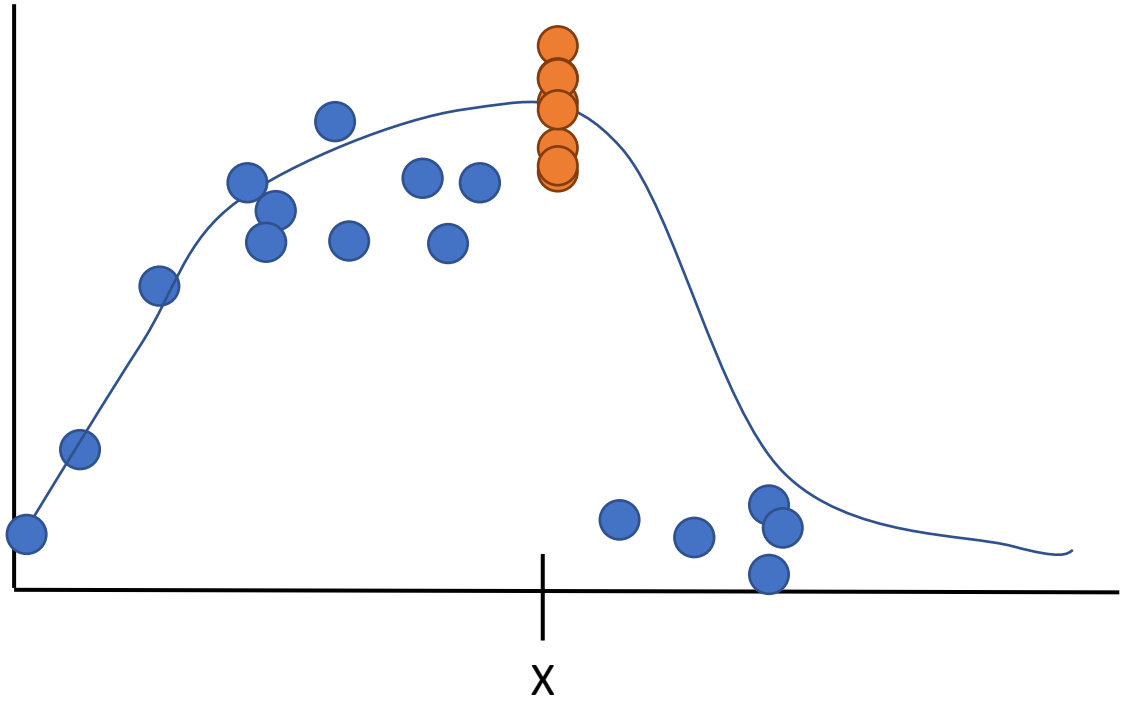




Discriminative

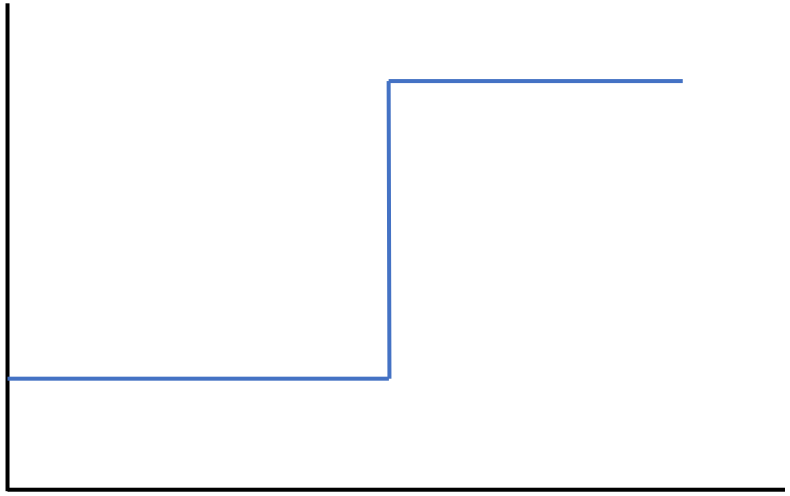


Generative

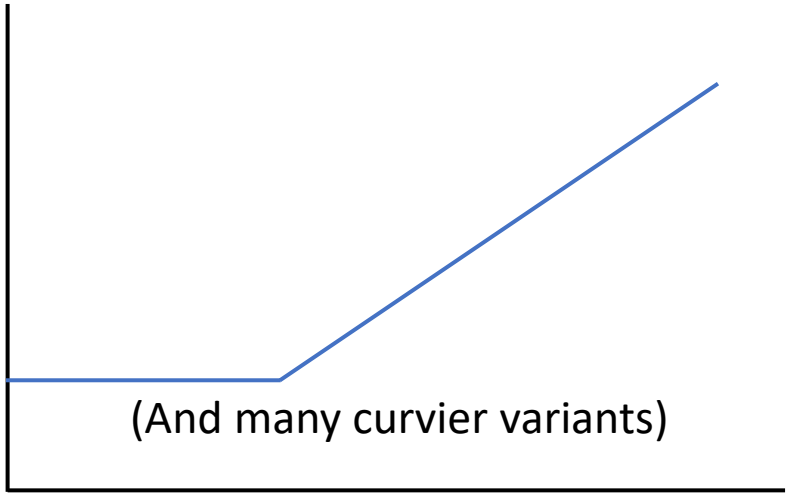


Programming With Data, Building Blocks

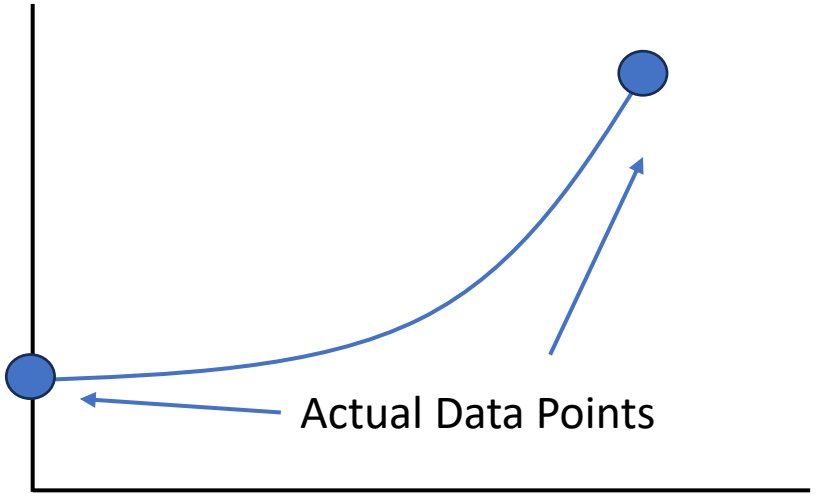
If-Then:
Decision Tree,
Gradient Boosted Random Forest



Artificial Neuron:
Neural Networks, Deep Learning



Case/Instance:
Instance Based Learning, kNN



What Does It Mean to Understand?

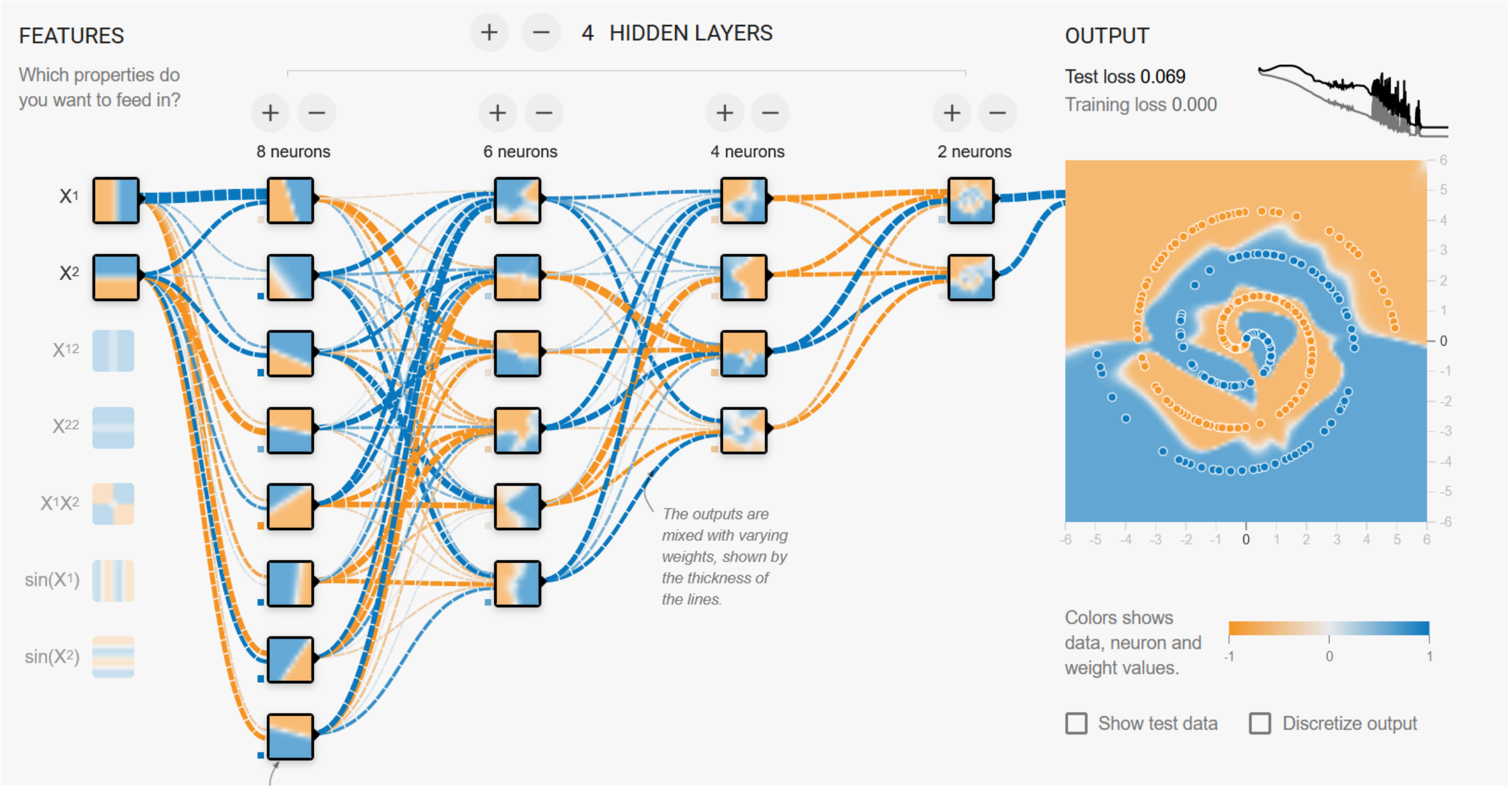
- Knowable
- Comprehendible
- Empirical
- Predictable
- Causal
- Counterfactual
- Communicable

Without understandability, we build **intellectual debt**

What Does Most AI/ML Offer?

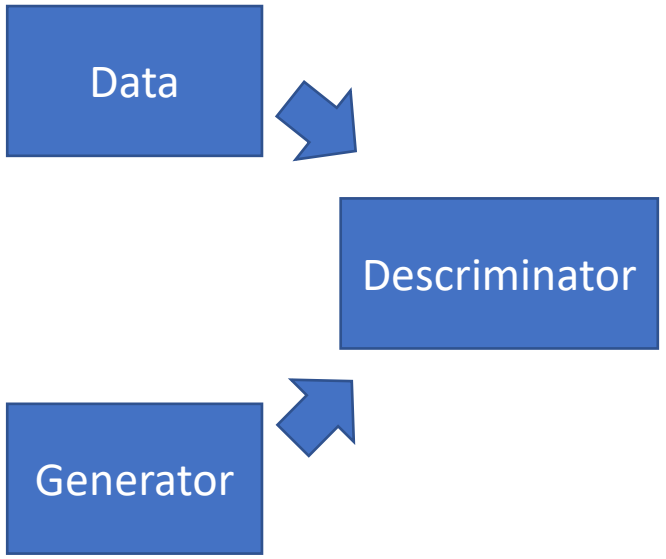
- Interpretable: Inner workings are understandable, can follow and reproduce the prediction or decision
  - Decision trees, linear regression, GAM, etc.
  - Being eroded to mean something more similar to “explainable”

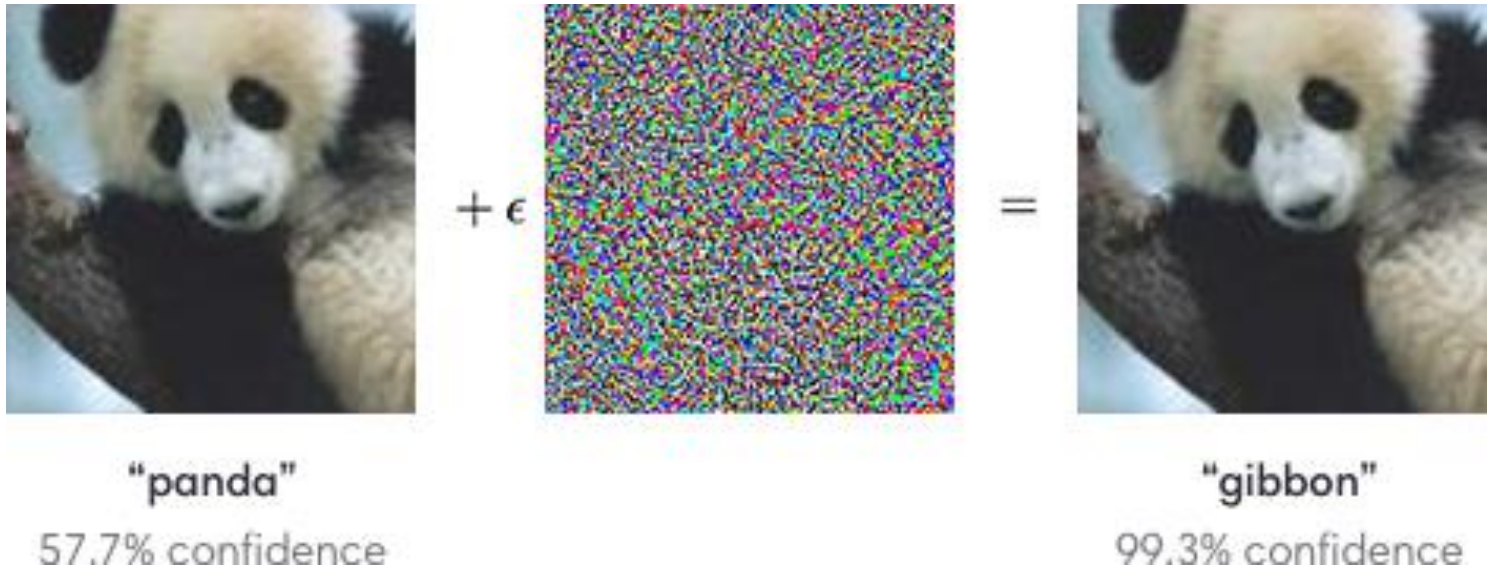
- Explainable: Ex post characterization of how a model behaves; a justification
  - SHAP, LIME, counterfactual values, surrogate models, feature importance, etc.



playground.tensorflow.org

GANs (Generative Adversarial Networks): Using AI to Attack AI results and explanations

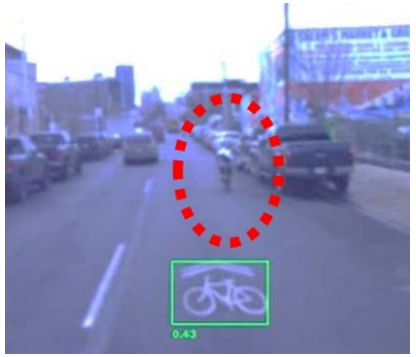




Goodfellow, 2014

Context and Biased Training Data

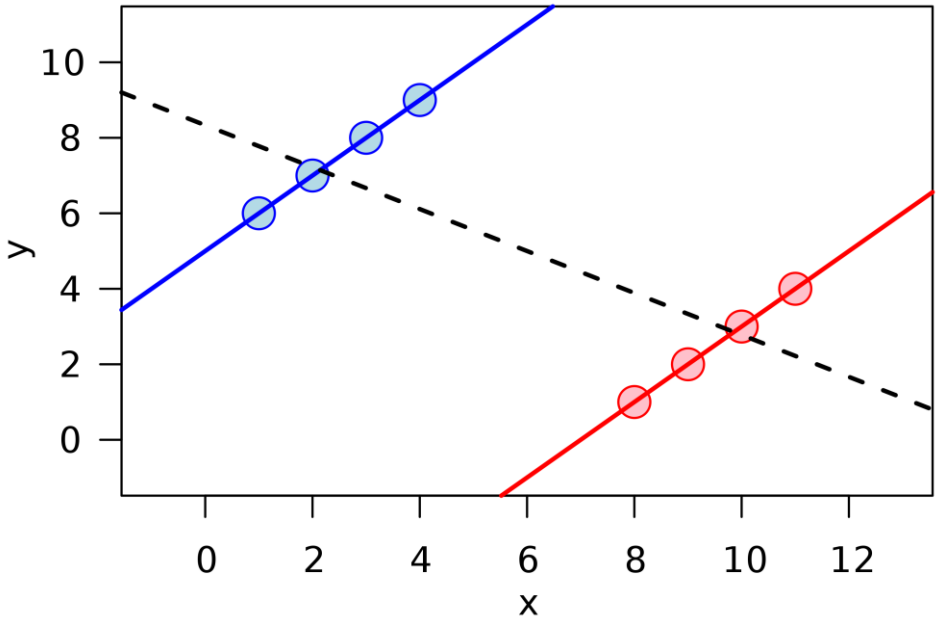
- The algorithm will optimize given the data
- Importance of loss functions: cost of error, symmetry
- Long tail of situations: Ensure sufficient coverage of the real world



Phillip Koopman, SSS 2019 & SafeAI 2019

What's Your AI/ML Model Really Doing?

- All sorts of bias: confirmation bias, Dunning-Kruger, loss aversion, projection bias, survivorship bias, group attribution error, etc.
- Stratified sampling is the answer?
  - Not entirely: Simpson's paradox
  - Sometimes practically impossible
- Testing on the training set
  - (Overfitting)
- Empirical results
  - Smith & Pell, BMJ: "Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials"
  - No evidence

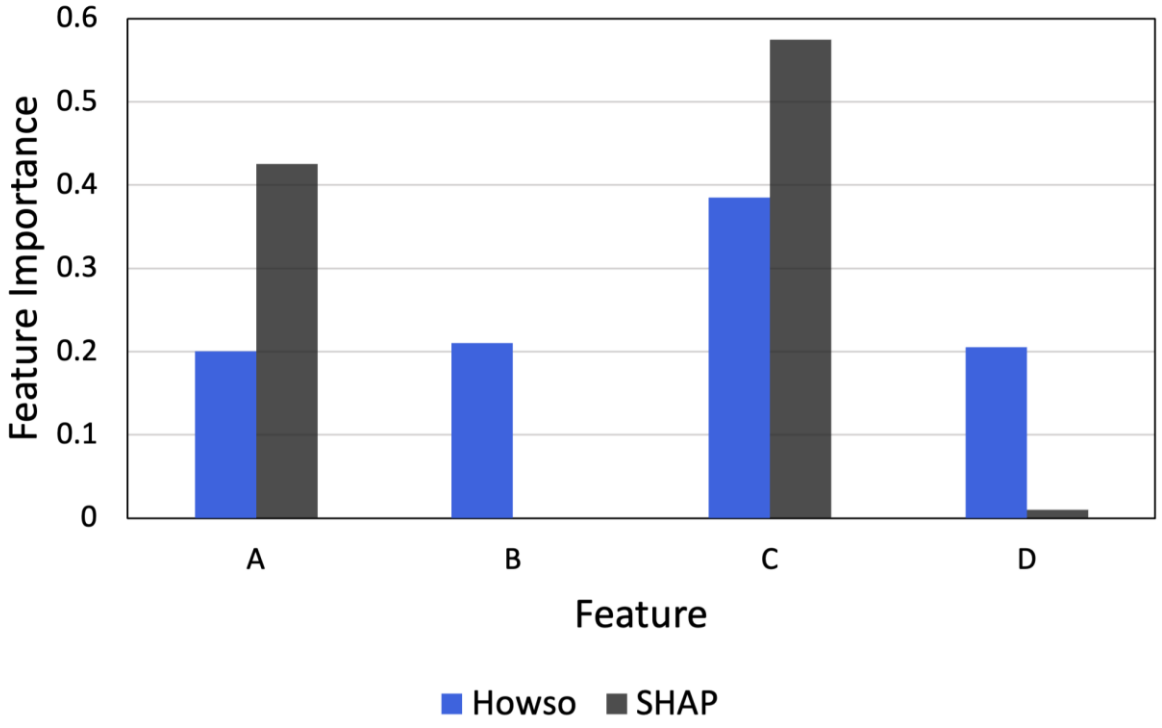


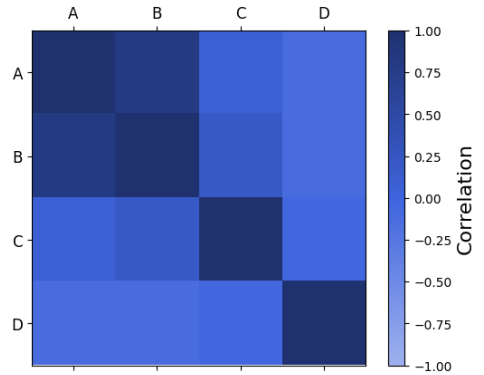
Because…

- “Excuse me, I have 5 pages. May I use the Xerox machine?”
- 60% allowed line skipping
- “Excuse me, I have 5 pages. May I use the Xerox machine, because I’m in a rush?”
- 94% allowed line skipping
- “Excuse me, I have 5 pages. May I use the Xerox machine, because I have to make copies?”
- 93% allowed line skipping

-Langer & Chanowitz, J Personality & Social Psychology, 1978

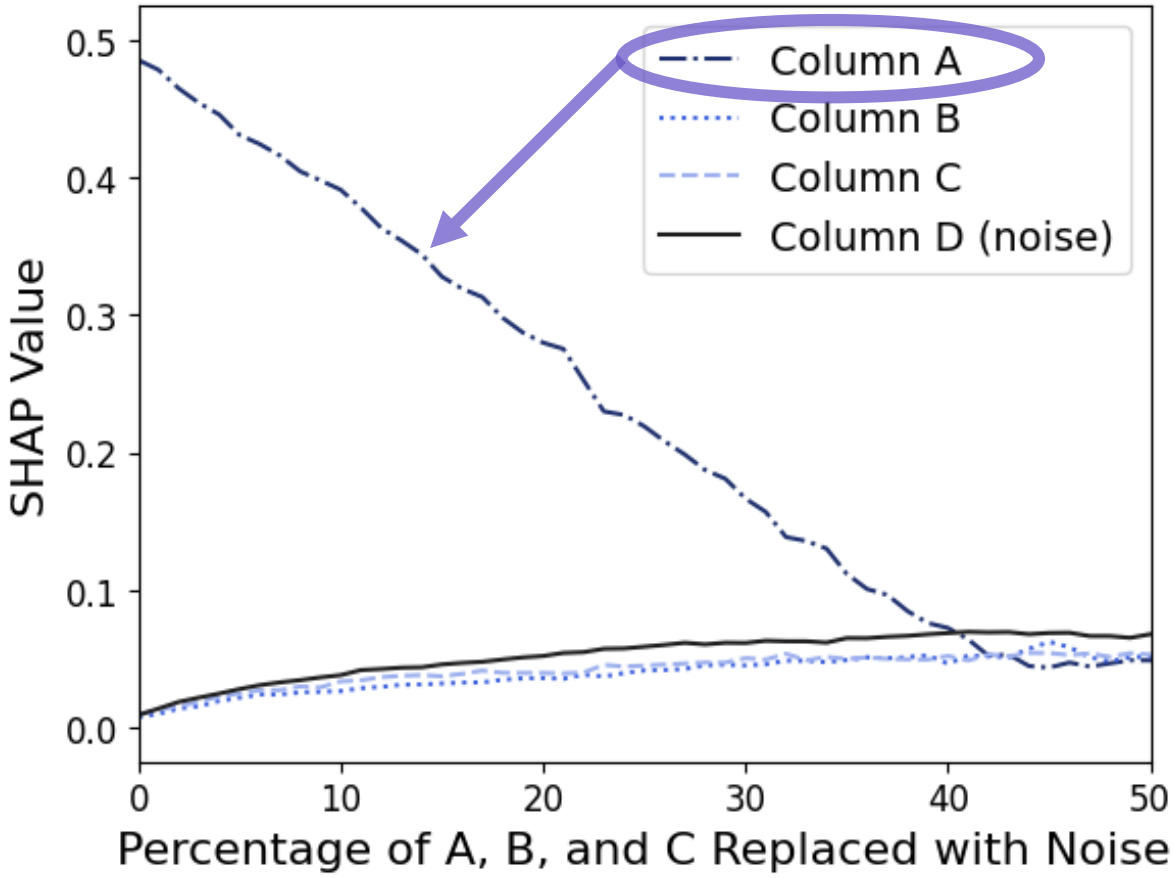
Feature Importance: When SHAP Fails

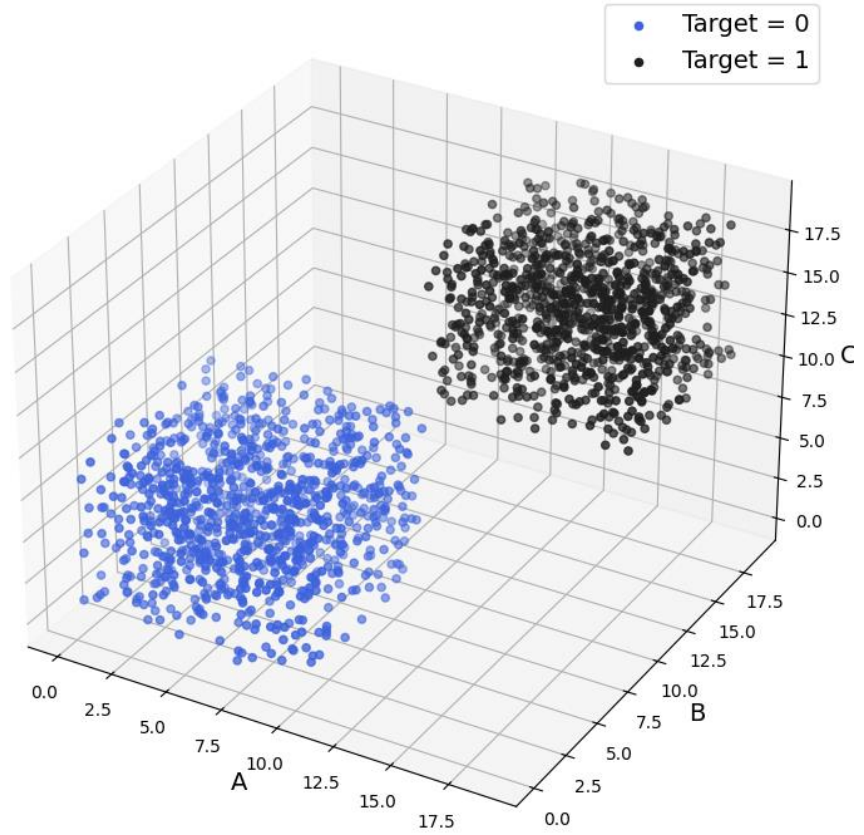




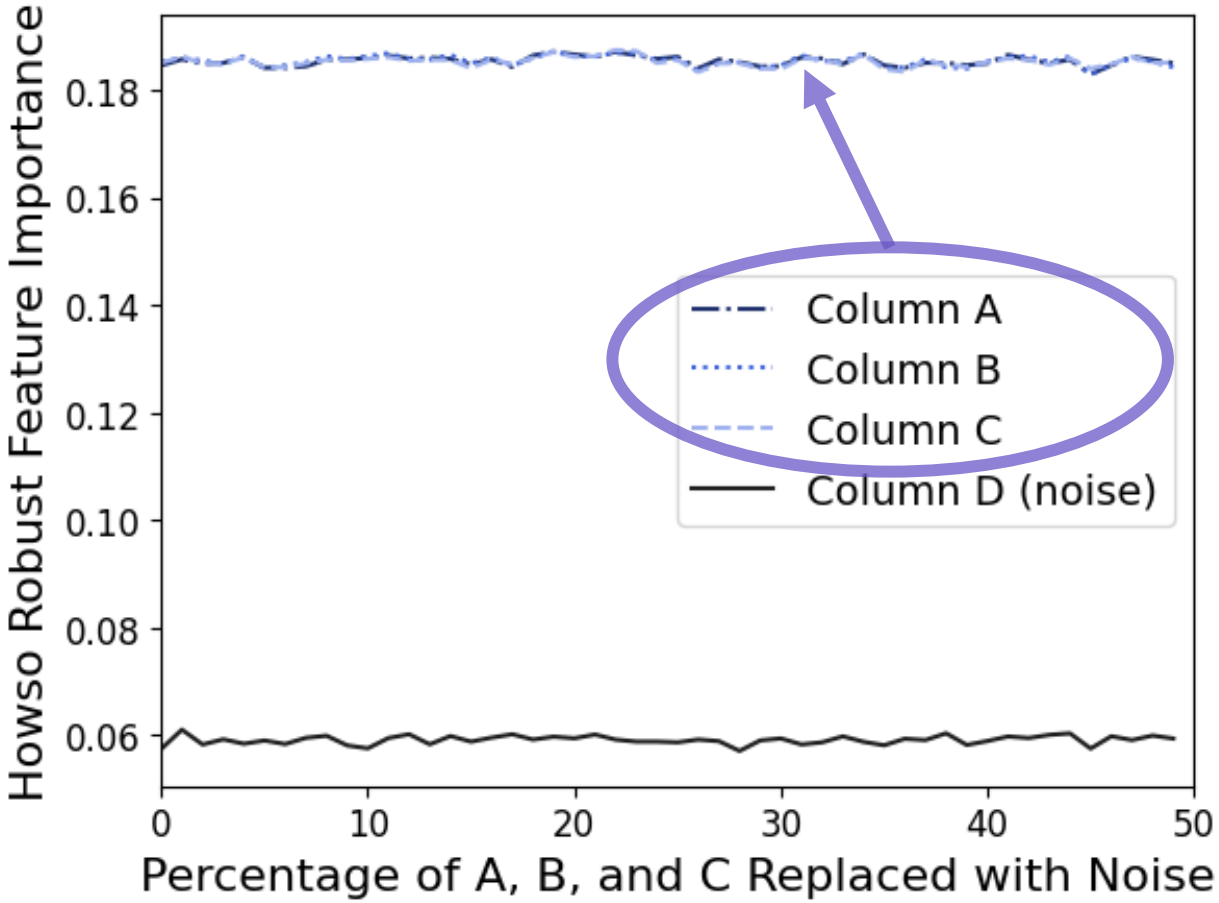
Target = A + B + C + D

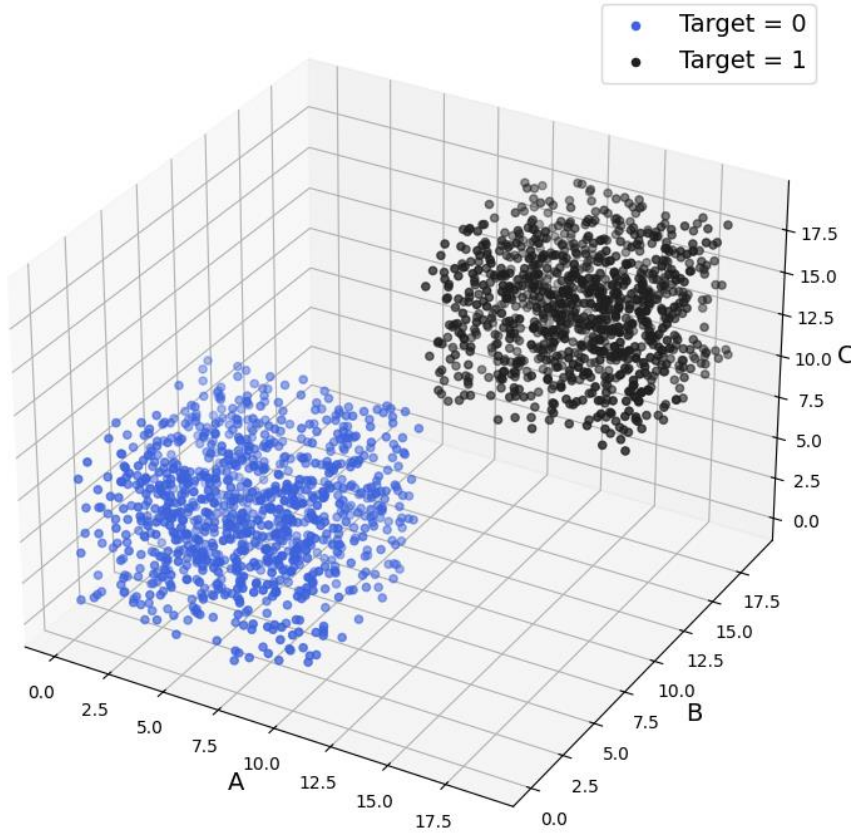
Feature Importance Can Be Misleading
The Status Quo of SHAP



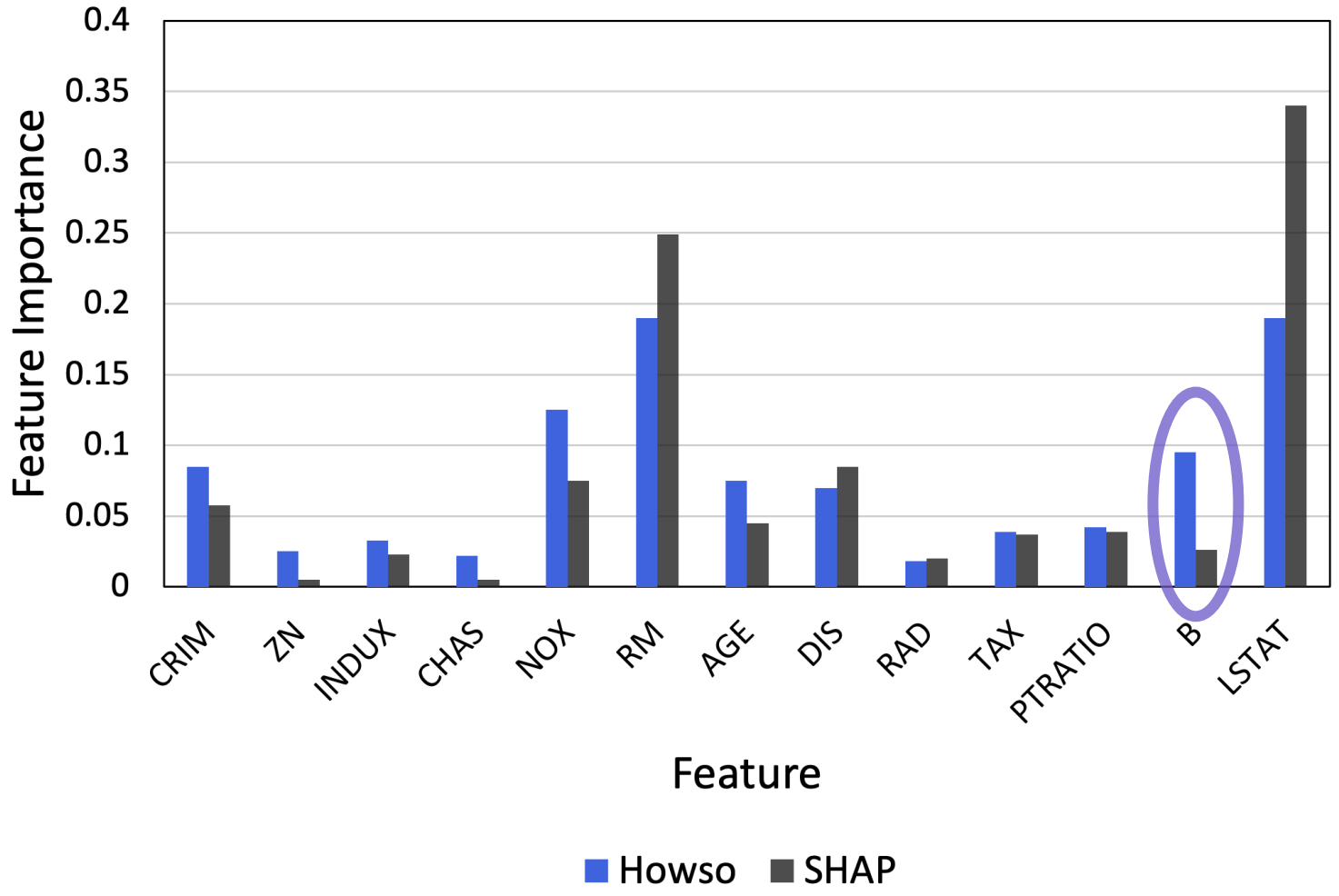


Robust Feature Contributions





Why Getting Feature Attribution Right Matters Know the Data, Not Just the Model



Two Essential Paths of Understanding

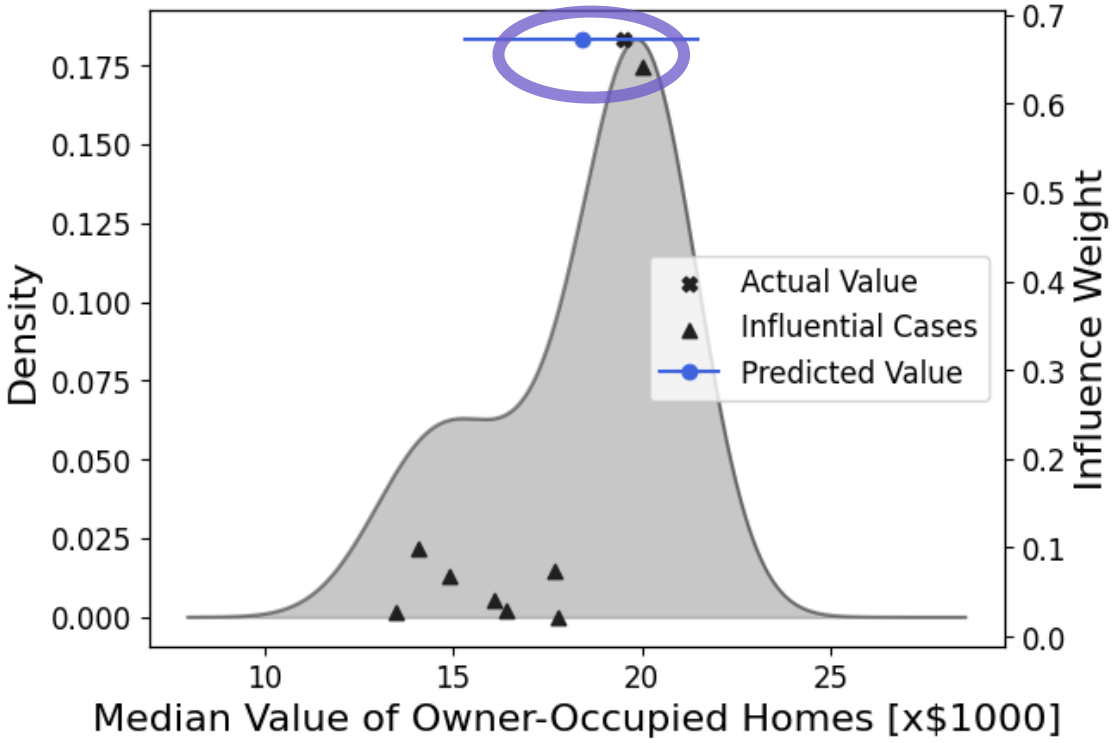
As a Consumer of Insights

- Justifiable
- Corroborate with experience & evidence
- Point to specifics (cases and features)
- Able to answer why and why not?
- Accurate

As a Provider of Insights

- Every step is clear
- Clearly defined units (e.g., ft/sec)
- Clearly defined operations
- Verifiable operations and outcome
- Verifiable provenance and lineage of data

Understandable ML Example

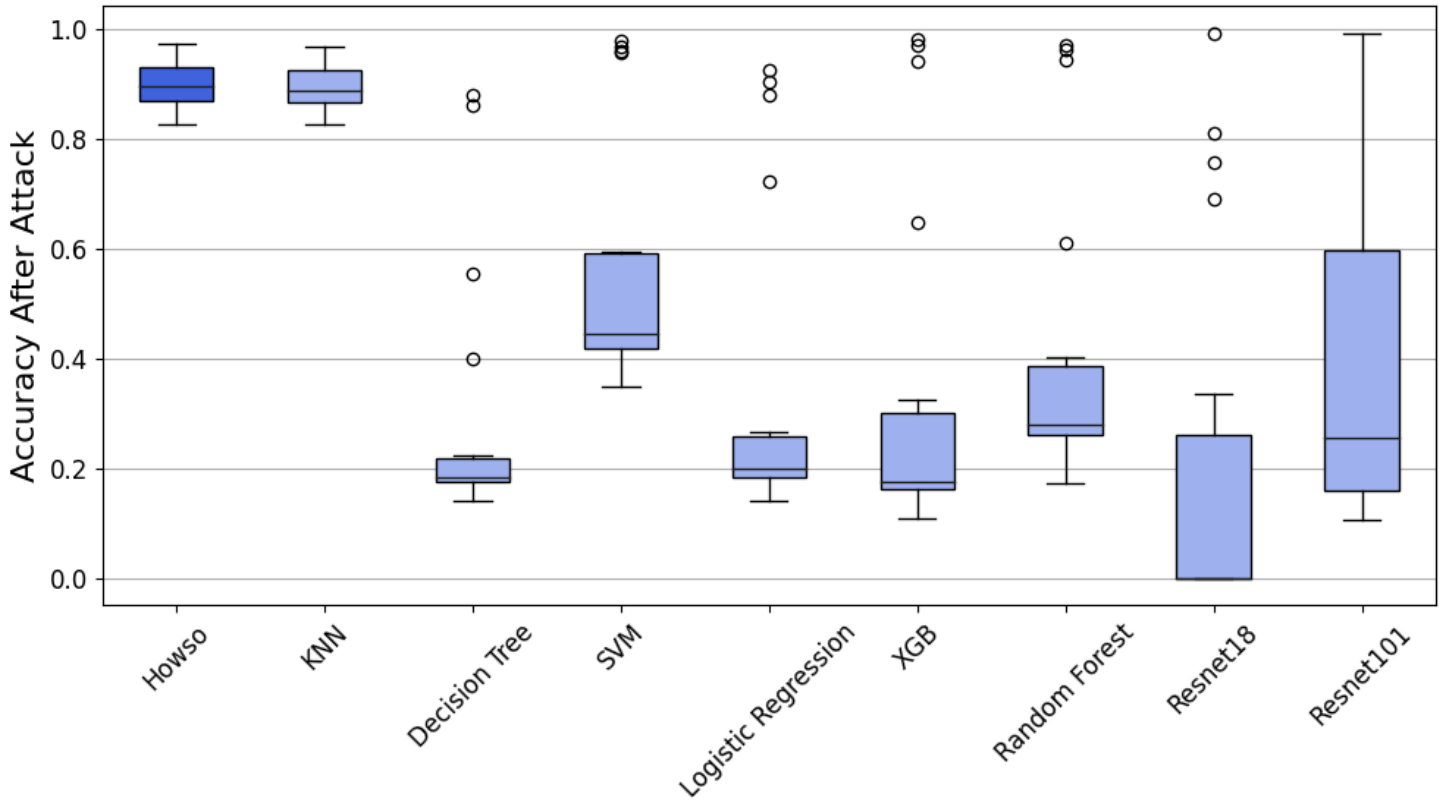


Results from Howso Engine™

All steps reasonably reproduceable manually from the data

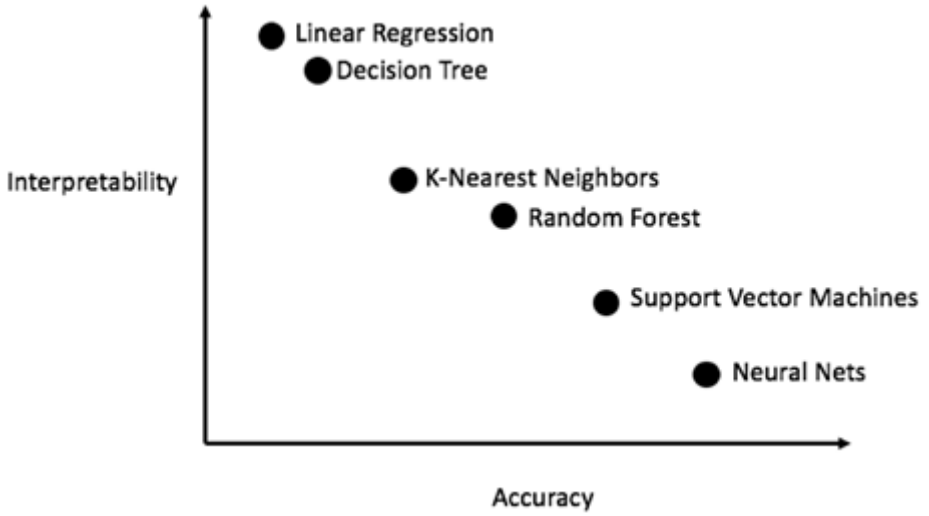
Know Your Data: Debuggability
Data Quality Matters

AI Robustness = Security



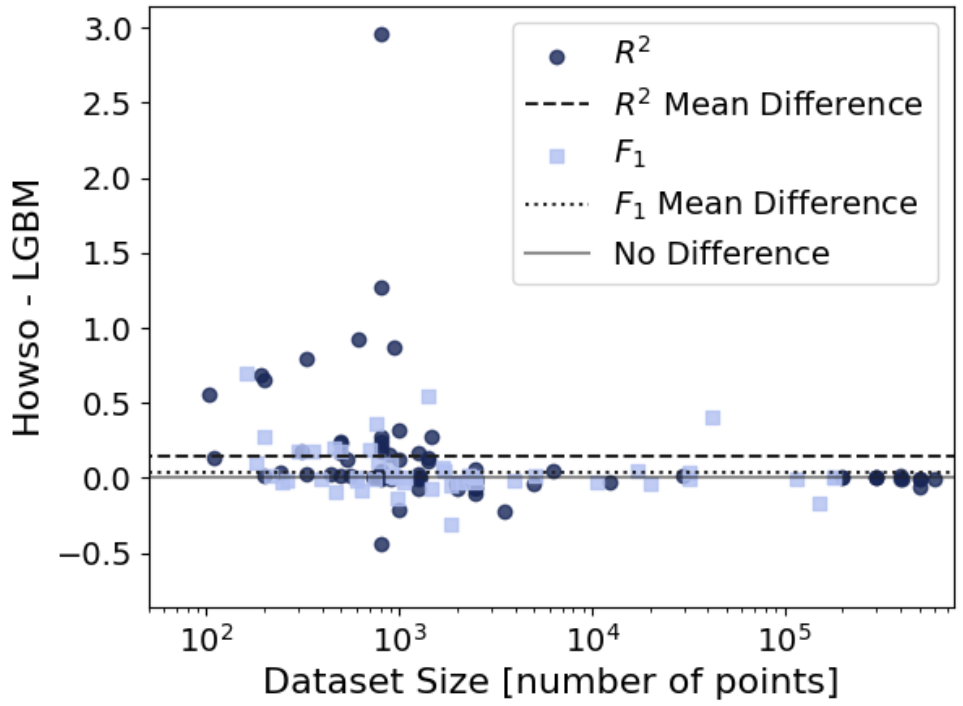
Interpretability & Explainability Myths

- “X is all you need!” e.g., feature importance, counterfactuals.
- Explanations are good enough
  - Problematic bias, adversarial models, high-cost mistakes with insufficient explanations, intellectual debt
- “My model gives me a probability value, so I can use that” without calibration
- Decision trees are accessible



From https://towardsdatascience.com/model-complexity-accuracy-and-interpretability-59888e69ab3d

Accuracy – Interpretability Not A Tradeoff Anymore



Influential Data Attribution

Local Feature Contributions	0.26	0	0	0	1.2	0.19	0.24	0.21	0	0	0	0.44	0.29	-
%Influence	**CRIM**	**ZN**	**INDUS**	**CHAS**	**NOX**	**RM**	**AGE**	**DIS**	**RAD**	**TAX**	**PTRATIO**	**B**	**LSTAT**	**Median Value of Home**
Actual	6.65	0.0	18.1	0.0	0.71	6.32	83.0	2.73	24	666	20.2	396.9	14.0	19.5
64.0%	6.80	0.0	18.1	0.0	0.713	6.08	84.4	2.72	24	666	20.2	396.9	14.7	20.0
9.82%	9.33	0.0	18.1	0.0	0.713	6.19	98.7	2.26	24	666	20.2	396.9	18.1	14.1
7.42%	3.69	0.0	18.1	0.0	0.713	6.38	88.4	2.57	24	666	20.2	391.4	14.7	17.7
6.86%	7.75	0.0	18.1	0.0	0.713	6.30	83.7	2.78	24	666	20.2	272.2	16.2	14.9
4.09%	5.09	0.0	18.1	0.0	0.713	6.30	91.8	2.37	24	666	20.2	385.1	17.3	16.1

Compression – What's in a Model?

- Diffusion models: 15k-to-1 to 50k-to-1

Original:

Generated:

- Significant memorization found in large models, even diffusion models (Carlini et al., 2023 https://arxiv.org/pdf/2301.13188.pdf)
- Some claim fair use for ML training, but memorization can occur
  - Differential privacy may be a plausible fair use
- ChatGPT ~7-to-1. Other LLMs???
- Where did the data come from? Was it rightfully used? Are you sure?
- If transferring the model, is that sufficient transformation?
  - Or do we want organizations only offering SaaS like search?

How Do Free & Open-Source Licenses Apply?

- Source code: Permissive or copyleft software license
- Documentation: Permissive or copyleft documentation license
- Data: Permissive documentation license
  (relatively new, e.g., https://cdla.dev/)

How Do Free & Open-Source Licenses Apply To Black Box?

- A black box model: ???
- A black box model trained on non-free material without differential privacy mechanisms that may have memorized some of the material: ????
- A black box model trained on AGPLv3 code without differential privacy that emits code derived from AGPLv3: ?????
- The output of one of the above blackbox models: ?????

How Do Free & Open-Source Licenses Apply To Instance-Based Learning?

- An instance-based learning model using data? The data license(s – assuming compatibility)!
- An instance-based learning model that ingests code and can do inference on code? The source code license
- The output of one of the above instance-based models: A license compatible to the data it was trained on

How Can We Fix This?

- Use AI/ML that is understandable, debuggable
- Use AI/ML that is attributable and/or causal
- Know the licenses of all code and data and use appropriately
- Know where you got the data and code from, as well as rights, consent, quality
- Use differential privacy, synthetic data, and other appropriate privacy mechanisms when the data cannot or should not be published
- Publish characteristics and performance
- Publish negative results! (Just like science)
- Consider some exception like the GCC Runtime Library Exception for output of AI/ML
- Let's not compromise open-source fundamentals for short-term practicality! We may have almost all the license structures everything we need right now.

Data & Trust Alliance

Proposed Data Provenance Standards

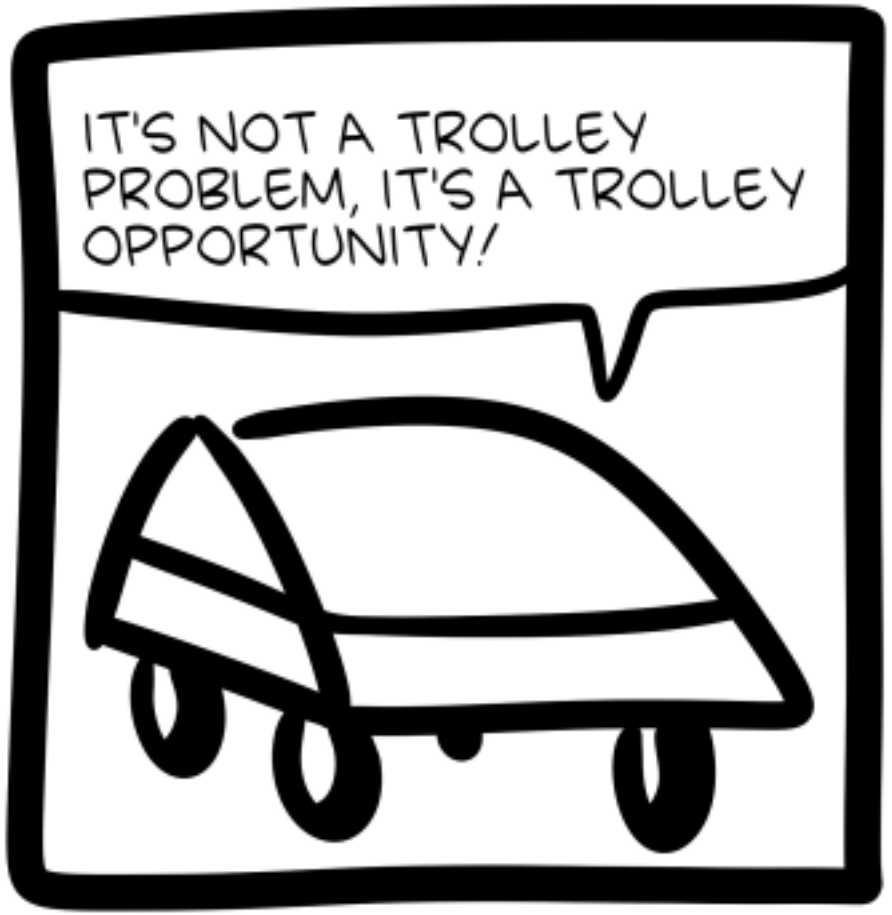
SET IDENTIFIER

Provenance Metadata Unique ID

A unique label identifying the provenance metadata of the current dataset

STANDARD	DESCRIPTION
Lineage	Identifiers or pointers to metadata representing the data which comprise the current dataset
Source	Identifies the origin (person, organization, system, device, etc.) of the current dataset
Legal rights	Identifies the legal or regulatory framework applicable to the current dataset, along with the required data attributions, associated copyright or trademark, and localization and processing requirements
Privacy and protection	Identifies any types of sensitive data associated with the current dataset and any privacy enhancing techniques applied
Generation date	Timestamp marking the creation of the current dataset
Data type	Identifies the data type contained in the current set, and provides insights into how the data is organized, its potential use cases, and the challenges associated with handling and using it
Generation method	Identifies how the data was produced (data mining, machine-generated, IoT sensors, etc.)
Intended use and restrictions	Identifies the intended use of the data and which downstream audiences should not be allowed access to the current dataset

Ethics!



GPT-4, probably, with the right prompt

https://www.smbc-comics.com/comic/decisions -- press the red button

Thank you!

HOW/SO™

github.com/howsoai