

# Intro Machine Learning

## 01 Overview

William Trimble, Amy Nussbaum  
Winter 2023



THE UNIVERSITY OF  
**CHICAGO**

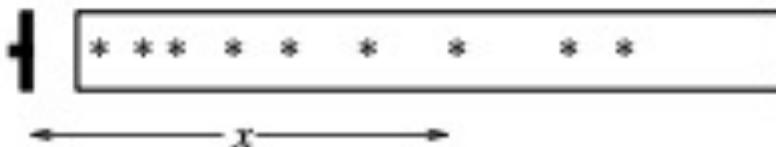
# Administrivia

- Week 1: read chapters 2 and 3 of Information Theory, Inference, and Learning Algorithms.
- Today: whirlwind overview
- Thursday: Bayesian estimation of continuous parameters
- Homework 1, heavy on prerequisites (calculus, linear algebra, some programming) is posted.



## Exercise 3.3. [3, p. 48] Inferring a decay constant

Unstable particles are emitted from a source and decay at a distance  $x$ , a real number that has an exponential probability distribution with characteristic length  $\lambda$ . Decay events can be observed only if they occur in a window extending from  $x = 1\text{ cm}$  to  $x = 20\text{ cm}$ .  $N$  decays are observed at locations  $\{x_1, \dots, x_N\}$ . What is  $\lambda$ ?



# Who are we?

## Instructors:

- William Trimble [wtrimbl@uchicago.edu](mailto:wtrimbl@uchicago.edu)



- Amy Nussbaum [anussbaum@uchicago.edu](mailto:anussbaum@uchicago.edu)



## TA's:

- JungHo Lee [jungholee@uchicago.edu](mailto:jungholee@uchicago.edu) Graduate student in Statistics



- Qiming Wang <[qmwang@uchicago.edu](mailto:qmwang@uchicago.edu)> PhD student CS



Andriy Burkov

# THE HUNDRED-PAGE MACHINE LEARNING BOOK



Andriy Burkov

David J. C. MacKay

# Information Theory, Inference, and Learning Algorithms



David MacKay

Springer Texts in Statistics

Gareth James  
Daniela Witten  
Trevor Hastie  
Robert Tibshirani

# An Introduction to Statistical Learning

with Applications in R

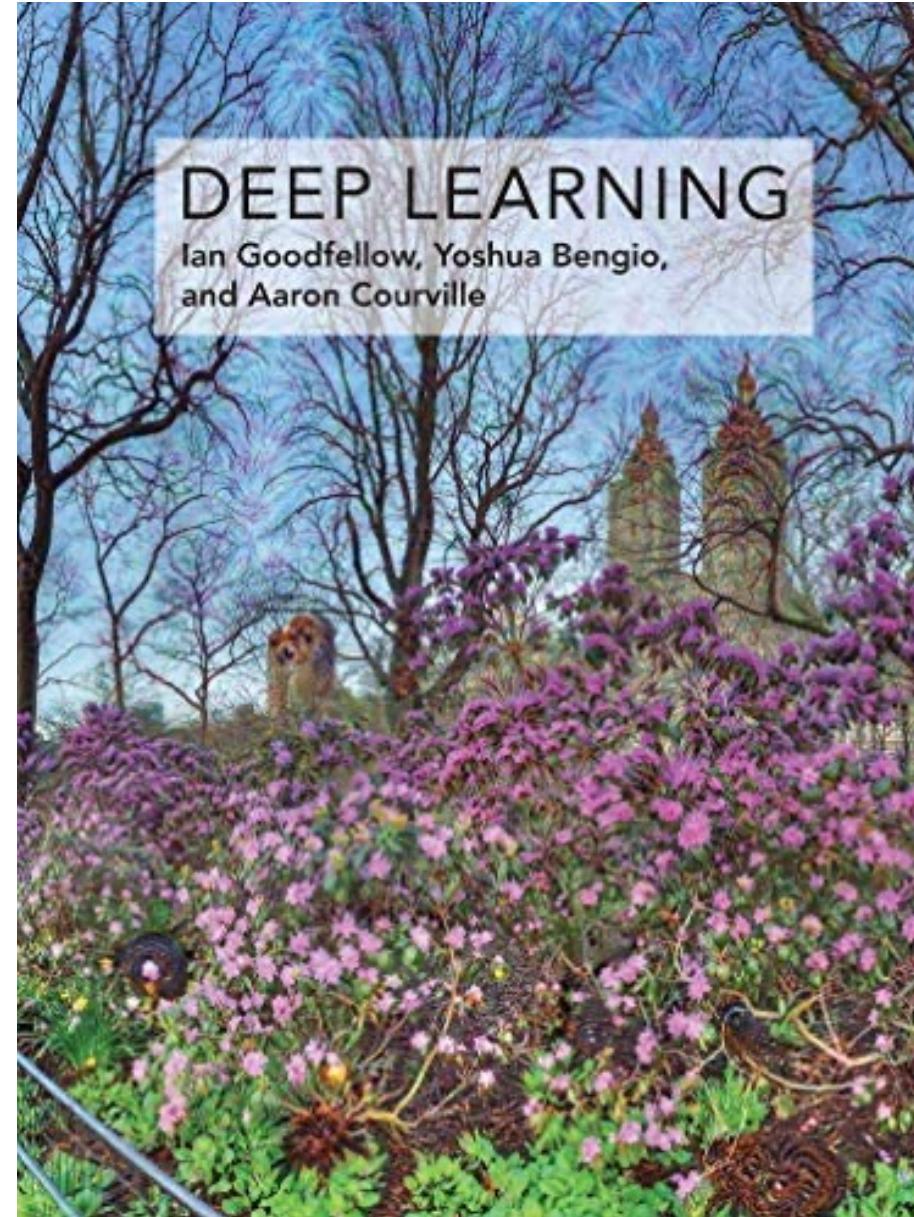
*Second Edition*

 Springer

Copyrighted Material

# DEEP LEARNING

Ian Goodfellow, Yoshua Bengio,  
and Aaron Courville



Gareth James

Ian Goodfellow

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL  
THEY START LOOKING RIGHT.



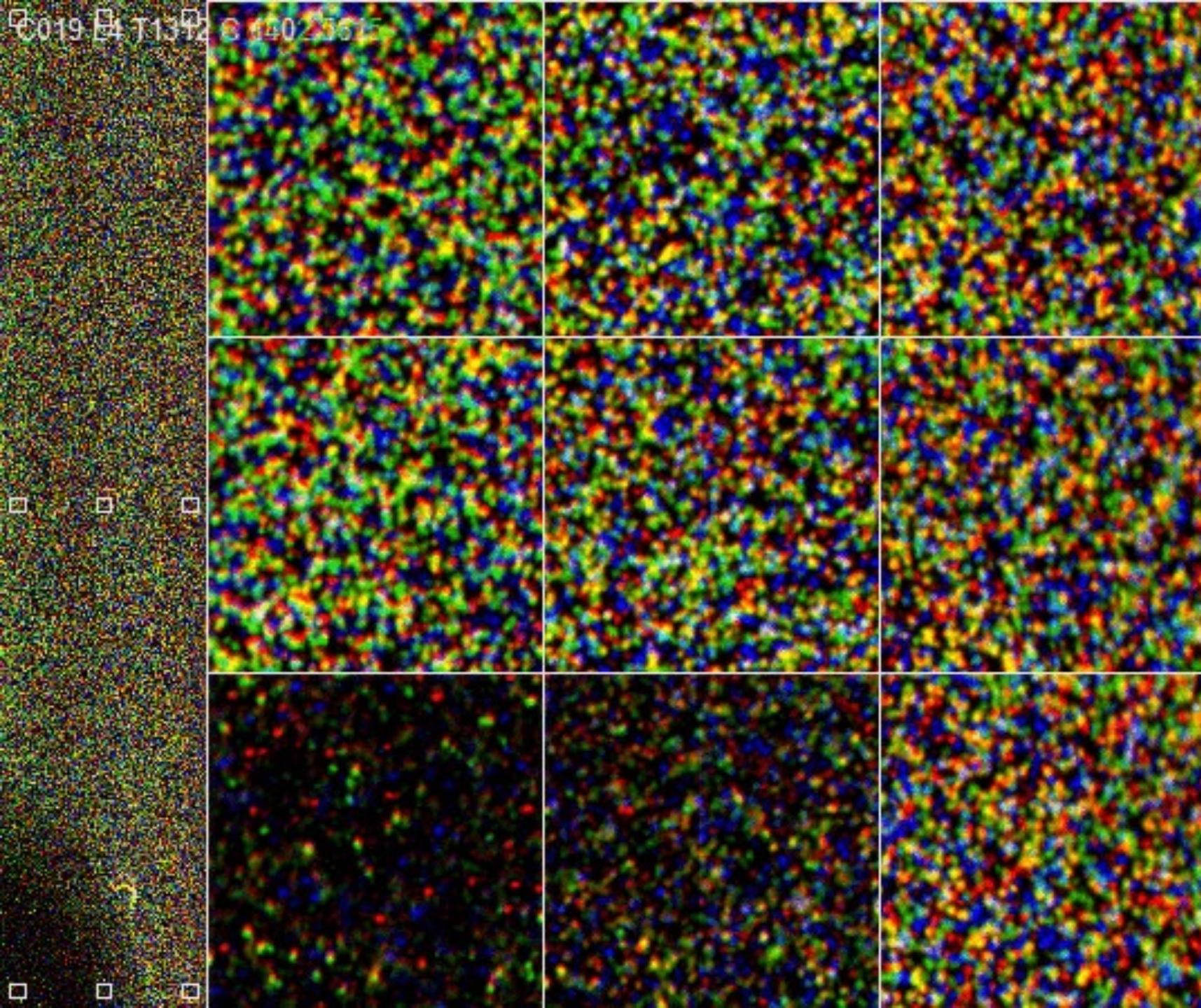
Why is this funny?

# What is Machine Learning?

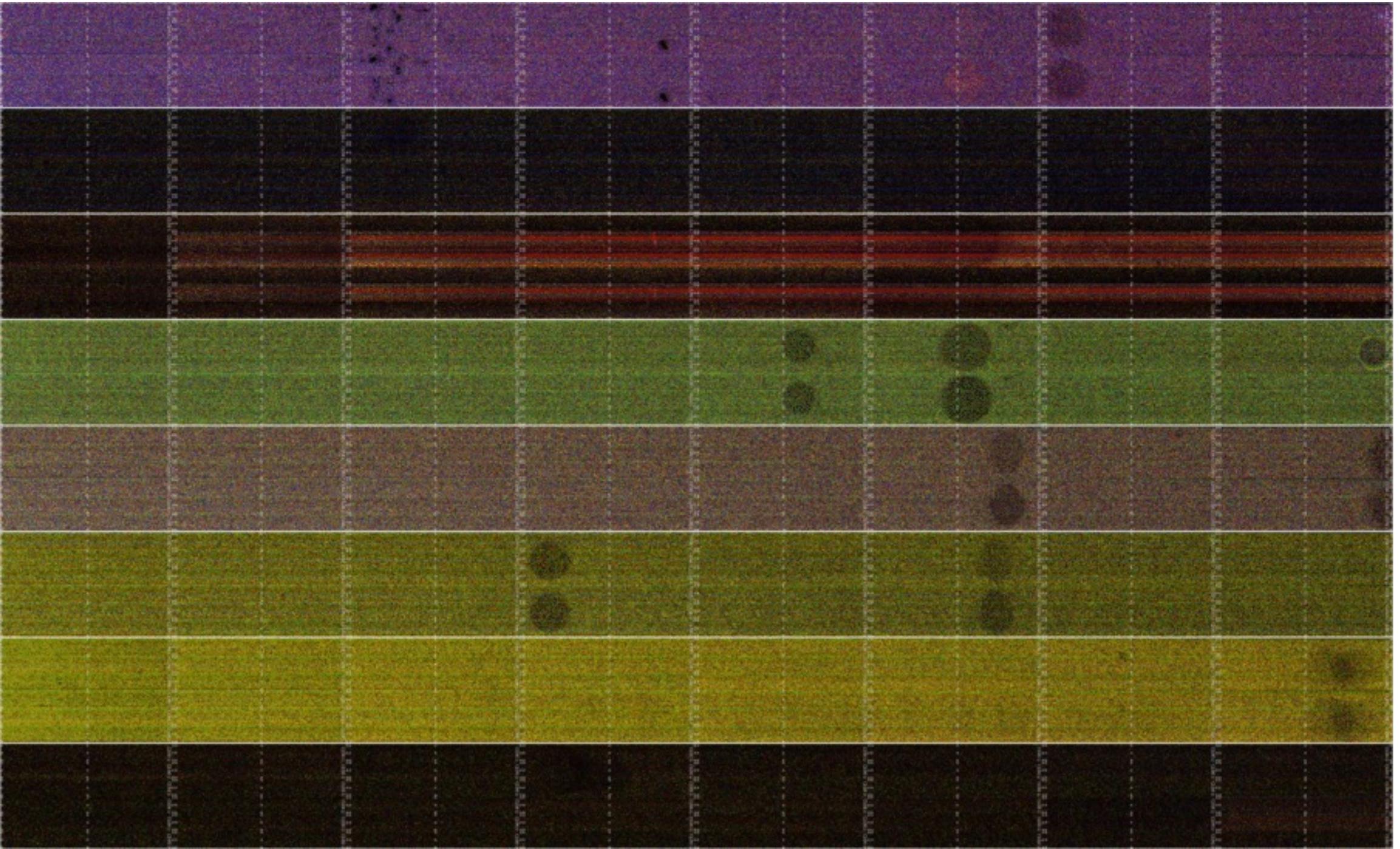
- That part of data analysis concerned with automated decisionmaking
- “Modeling” “parameter estimation” are related but narrower
- Why data? If the task can be solved without data (say, by applying thresholds) you don’t need ML



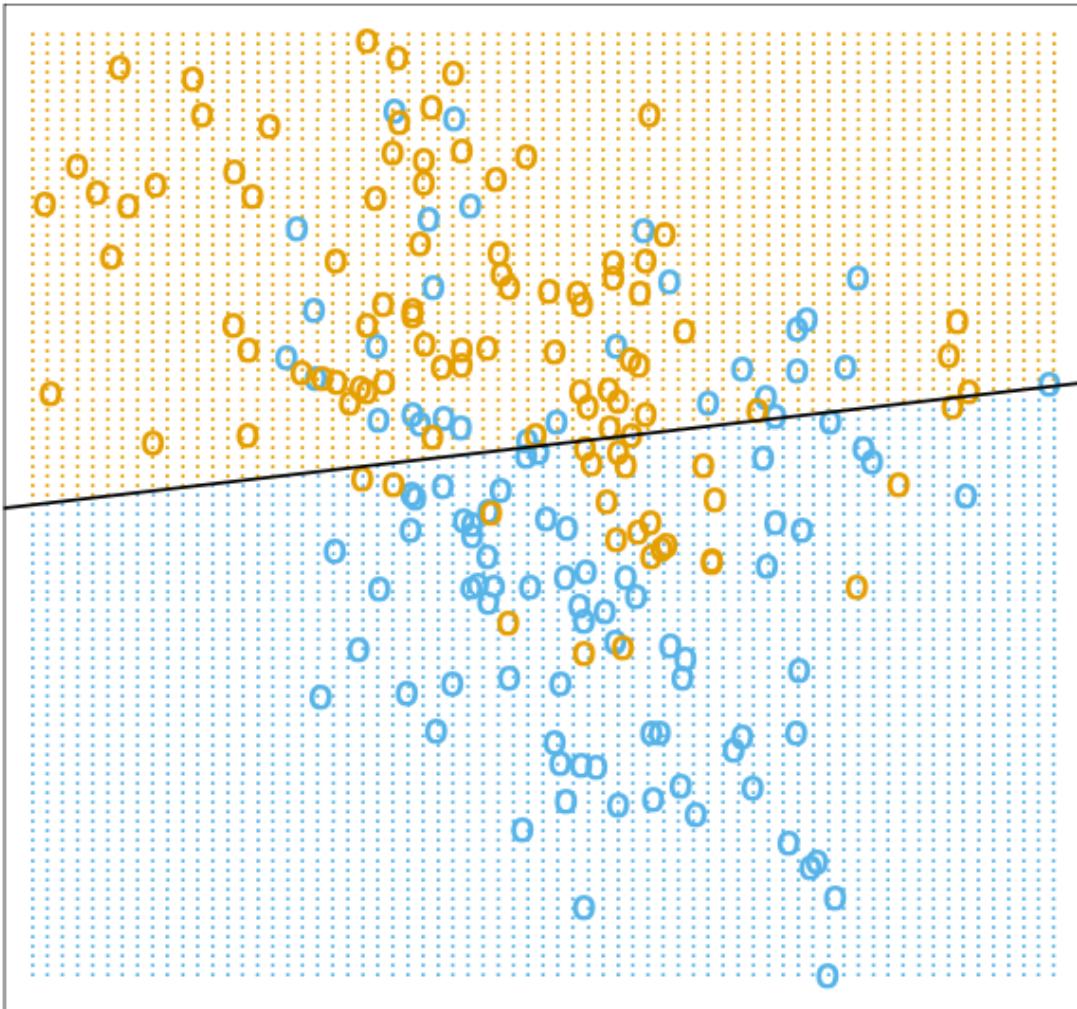
This machine does not learn



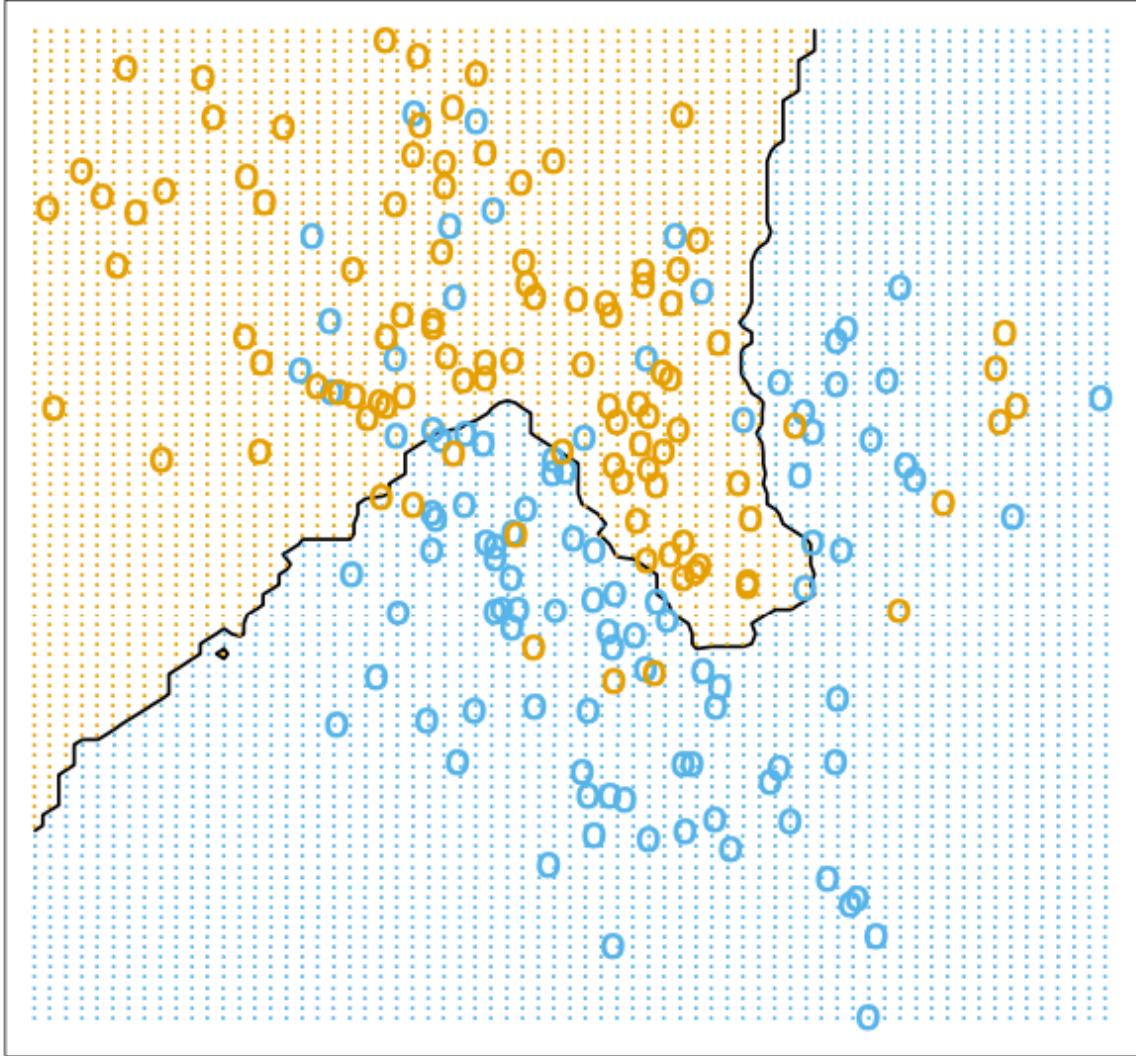
- So-called NGS sequencing data: one of the data firehoses



## Linear Regression of 0/1 Response

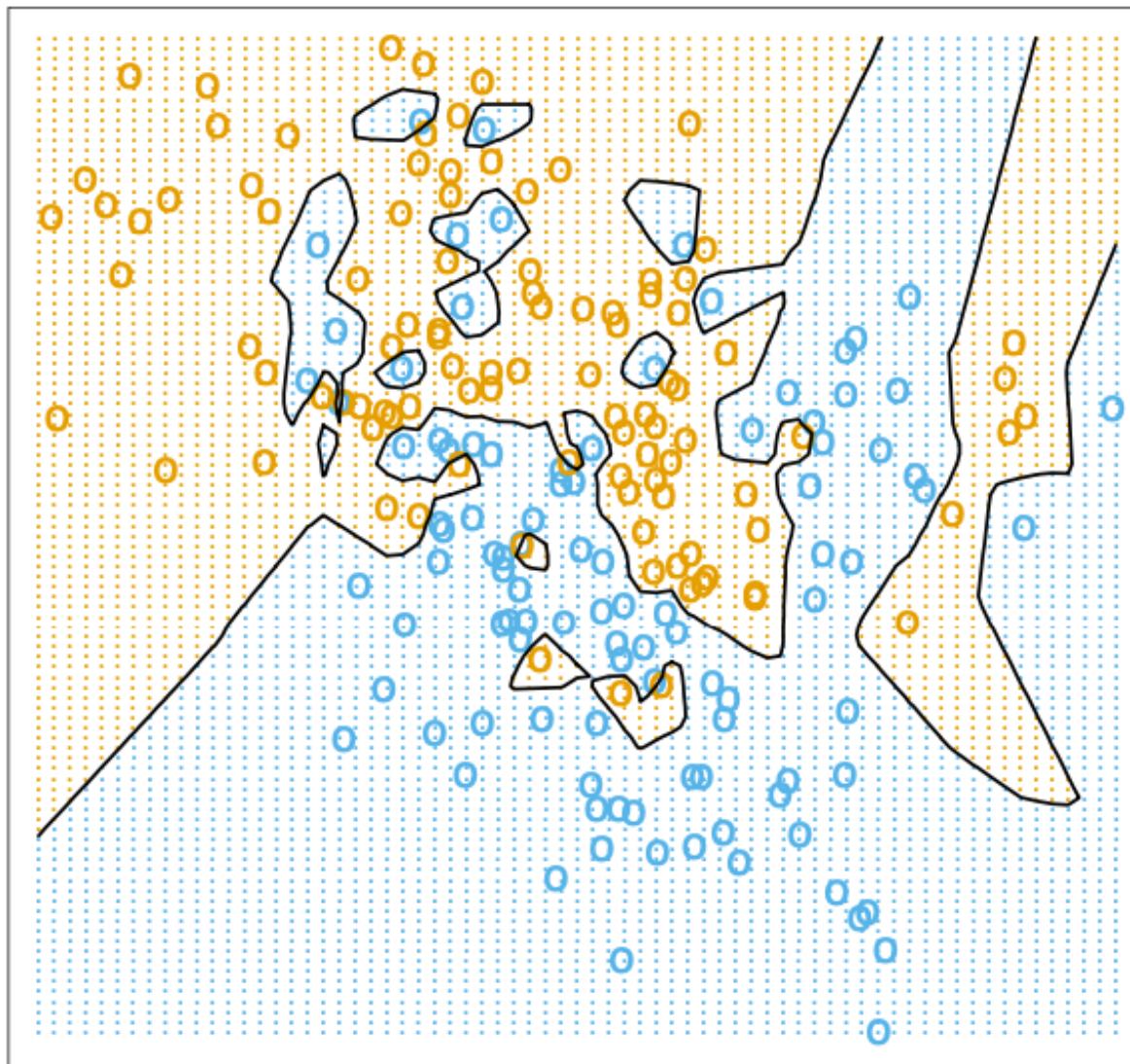


**FIGURE 2.1.** A classification example in two dimensions. The classes are coded as a binary variable—**BLUE** = 0, **ORANGE** = 1—and then fit by linear regression. The line is the decision boundary defined by  $x^T \hat{\beta} = 0.5$ . The orange shaded region denotes that part of input space classified as **ORANGE**, while the blue region is classified as **BLUE**.

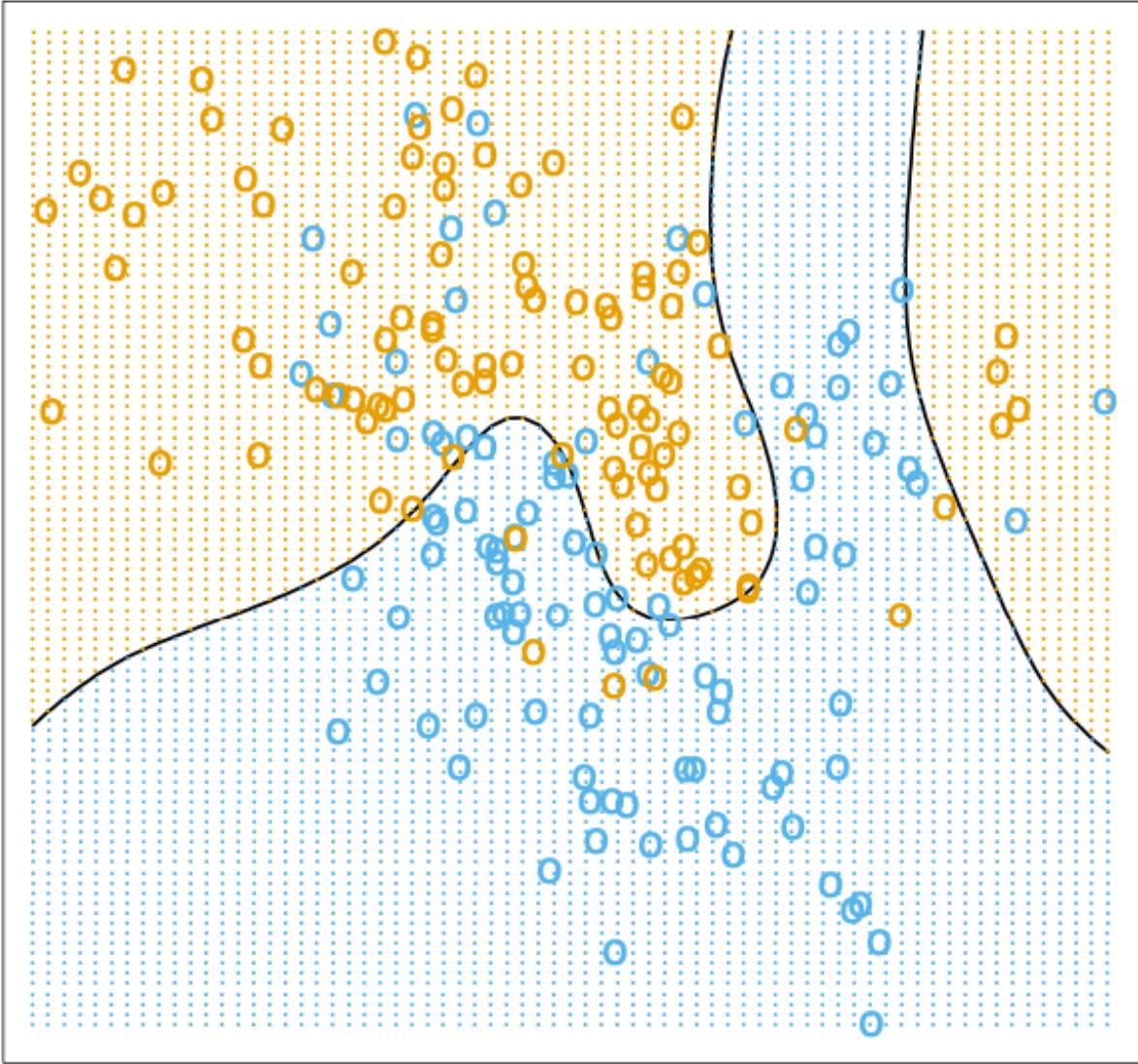


**FIGURE 2.2.** The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (**BLUE** = 0, **ORANGE** = 1) and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.

### 1-Nearest Neighbor Classifier



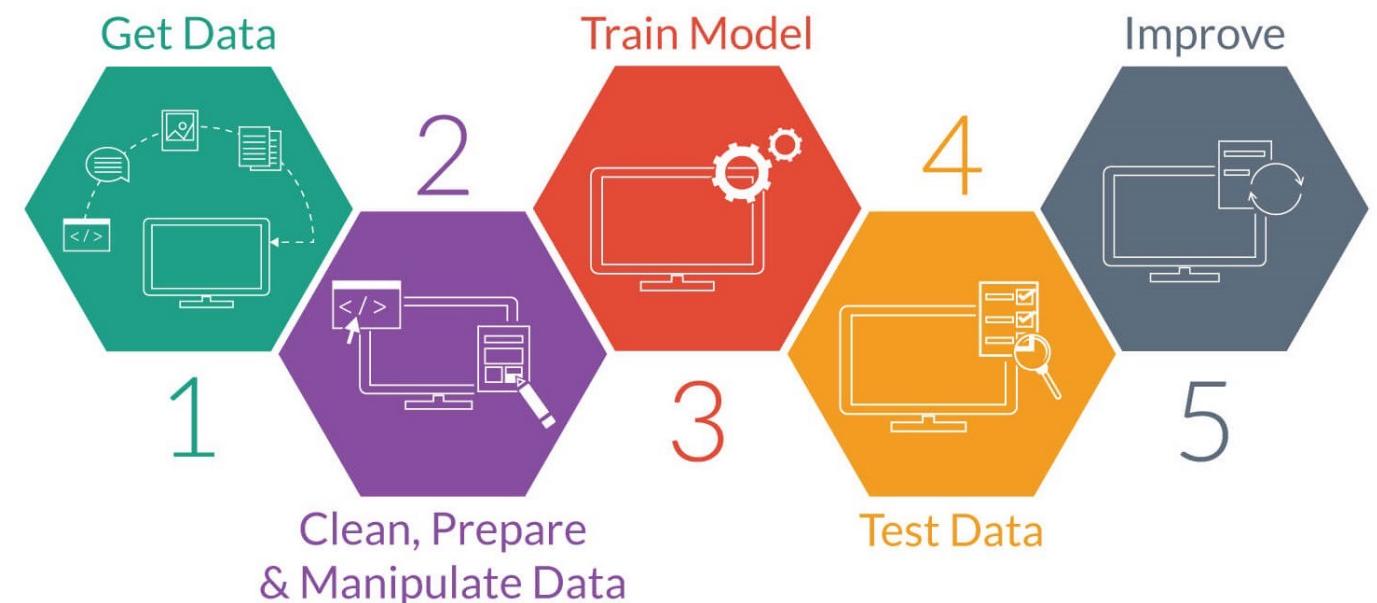
**FIGURE 2.3.** The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (**BLUE** = 0, **ORANGE** = 1), and then predicted by 1-nearest-neighbor classification.



*The optimal Bayes decision boundary for the simulation example. Since the generating density is known for each class, this boundary can be calculated exactly.*

# The steps

- Define the Task (input space and output space)
- Data preprocessing (feature extraction & wrangling)
- Measure performance
- Train (or “learn”) the model
- Victory ?!?



# Task

- Regression -- predicting some values of interest
  - Imputation of missing values
- Classification – put data points into categories
- Classification with missing inputs
  - collection of many classifiers, each for different set of missing data items?
  - generate a probability density for everything, marginalize over missing inputs
- Transcription – encode symbols in an image or audio waveform
- Structured output – answer very specific questions, like map annotation, write computer programs or captions
- Anomaly detection – fraud detection, for instance
- Noise detection / removal / error correction
- Density estimation; fitting (parameters of interest)

# Task

- Regression -- predicting some values of interest
  - Imputation of missing values
- Classification – put data points into categories
- Classification with missing inputs
  - collection of many classifiers, each for different set of missing data items?
  - generate a probability density for everything, marginalize over missing inputs
- Transcription – encode symbols in an image or audio
- Structured output – answer very specific questions, like map annotation, write computer programs or captions
  - generation of textures, waveforms
- Anomaly detection – fraud detection, for instance
  - Fake news; security applications
- Noise detection / removal / error correction
  - Geodesy; remote sensing; astronomy
- Density estimation; fitting (parameters of interest); matching
  - Goodfellow, CTR, fingerprints, DNA

# Measure Performance

- How well does the model answer our question?
- Measure “accuracy”
- Maybe fold in “costs for different sorts of errors to guide decision thresholds; manage our training in light of expected frequencies
- Measure accuracy on the training set and on data not used for training “testing” or “holdout” sets.
- Calculus helps us for continuous variables / continuous inputs
- Brute force or Markov-chain Monte Carlo for discrete variables
- Space of parameters is often exceedingly large and complex; we usually have to resort to randomized algorithms and truncated representations.

CIFAR-10 Confusion Matrix											
True Class	airplane	923	4	21	8	4	1	5	5	23	6
	automobile	5	972	2					1	5	15
	bird	26	2	892	30	13	8	17	5	4	3
	cat	12	4	32	826	24	48	30	12	5	7
	deer	5	1	28	24	898	13	14	14	2	1
	dog	7	2	28	111	18	801	13	17		3
	frog	5		16	27	3	4	943	1	1	
	horse	9	1	14	13	22	17	3	915	2	4
	ship	37	10	4	4		1	2	1	931	10
	truck	20	39	3	3			2	1	9	923
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											
Column Labels											
Row Labels											

# Train (or “learn”) the model

- Unsupervised algorithms usually try to distill the empirical probability density into something that can be succinctly communicated.
  - This is the essence of compression
  - Clustering, finding groups of similar items, recommendation
  - Dimension reduction, embedding, big data visualization
- Supervised algorithms: training data with gold-standard labels or desired outcomes
- Semi-supervised: we can only afford to label some of the data, but have access to a larger population of unlabeled data for modeling its density

# Linear classifier

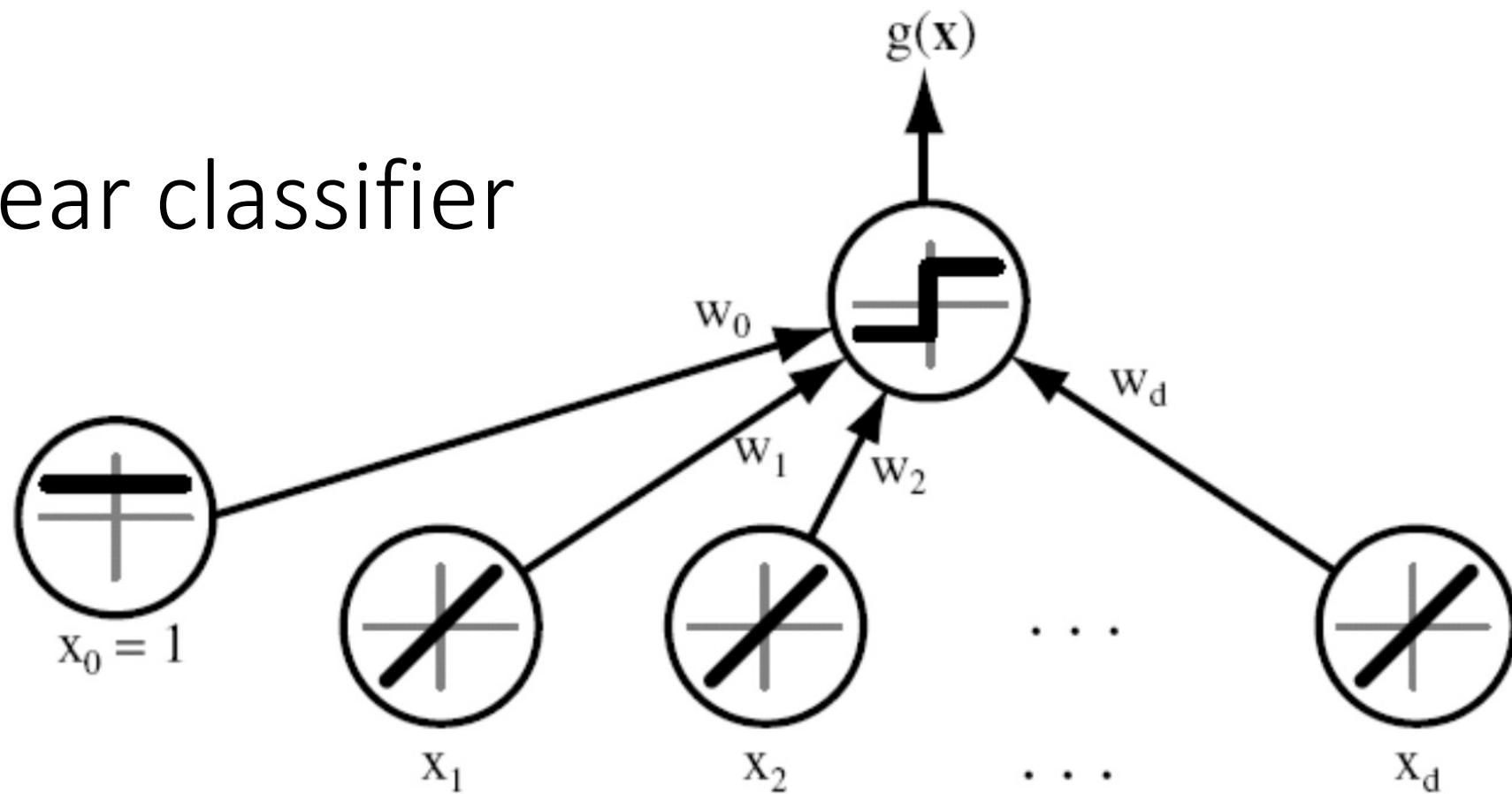
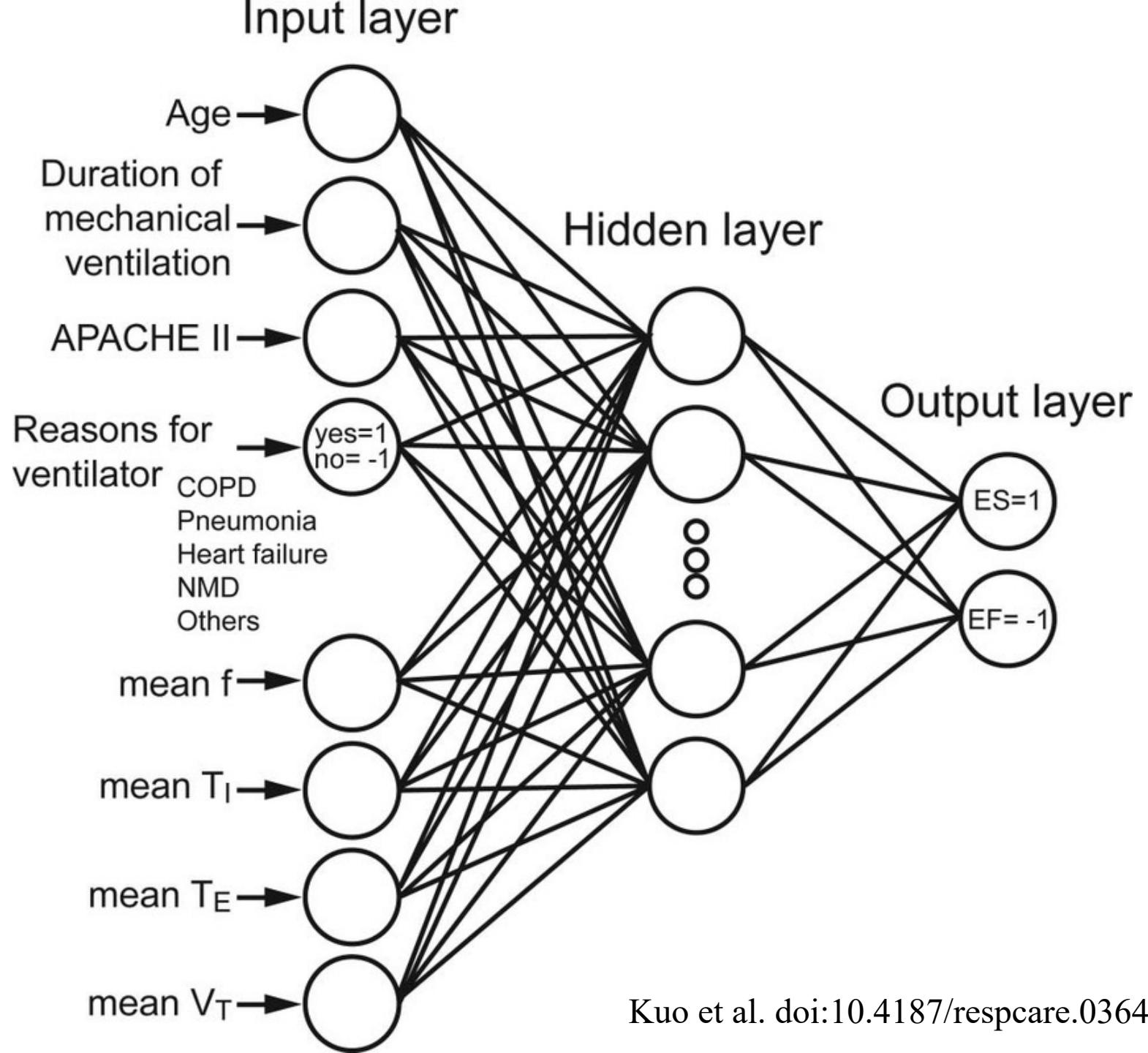


Figure 5.1: A simple linear classifier having  $d$  input units, each corresponding to the values of the components of an input vector. Each input feature value  $x_i$  is multiplied by its corresponding weight  $w_i$ ; the output unit sums all these products and emits a +1 if  $\mathbf{w}^t \mathbf{x} + w_0 > 0$  or a -1 otherwise.

# Neural networks

Linear classifiers  
+nonlinear sauce  
= arbitrary function machines



## DOGBERT CONSULTS

YOU NEED A DASH-  
BOARD APPLICATION  
TO TRACK YOUR  
KEY METRICS.

scottadams@acl.com

[www.dilbert.com](http://www.dilbert.com)

THAT WAY YOU'LL HAVE  
MORE DATA TO IGNORE  
WHEN YOU MAKE YOUR  
DECISIONS BASED ON  
COMPANY POLITICS.

WILL THE  
DATA BE  
ACCURATE?

OKAY,  
LET'S  
PRETEND  
THAT  
MATTERS.

© Scott Adams, Inc./Dist. by UFS, Inc.

5-6-07 © 2007 Scott Adams, Inc./Dist. by UFS, Inc.

DOGBERT CONSULTS

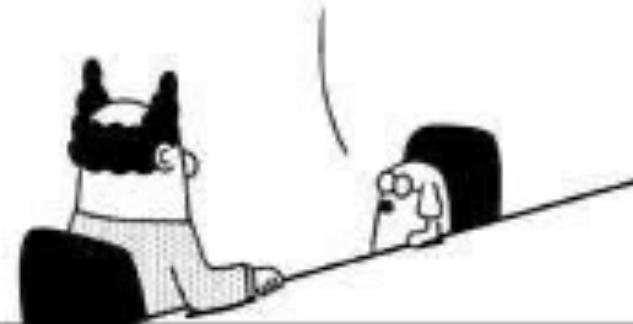
YOU NEED A DASH-  
BOARD APPLICATION  
TO TRACK YOUR  
KEY METRICS.



scottadams@acl.com

www.dilbert.com

THAT WAY YOU'LL HAVE  
MORE DATA TO IGNORE  
WHEN YOU MAKE YOUR  
DECISIONS BASED ON  
COMPANY POLITICS.



© 2007 Scott Adams, Inc./Dist. by UFS, Inc.

WILL THE  
DATA BE  
ACCURATE?

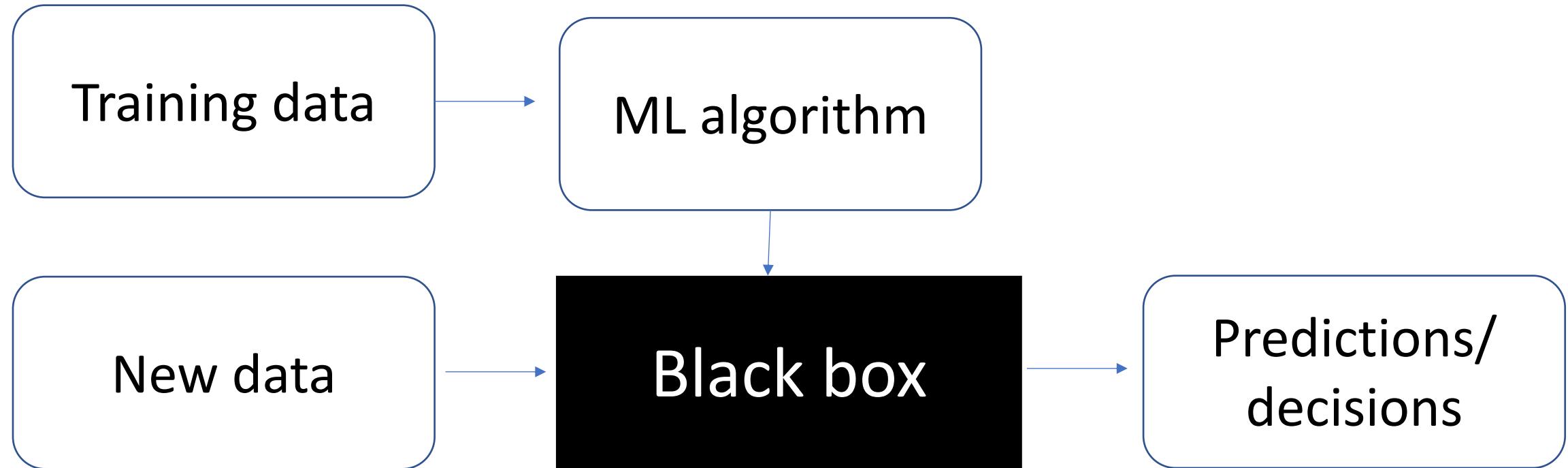
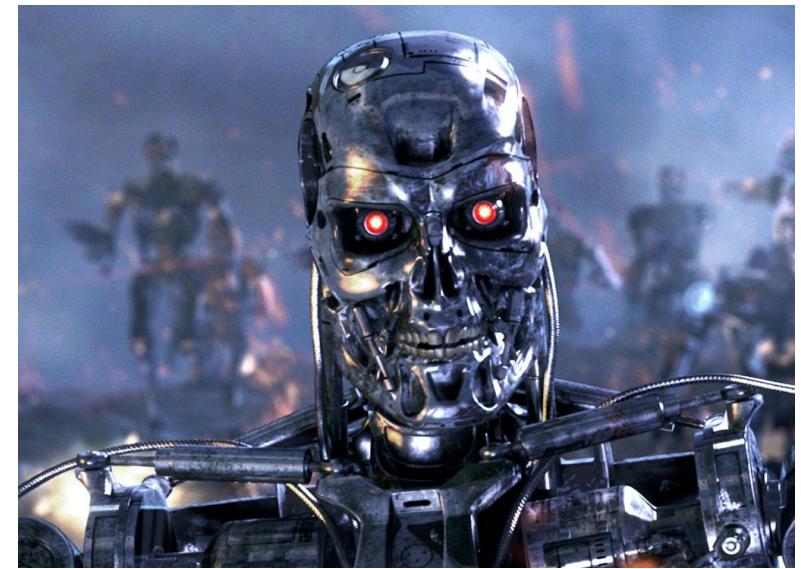


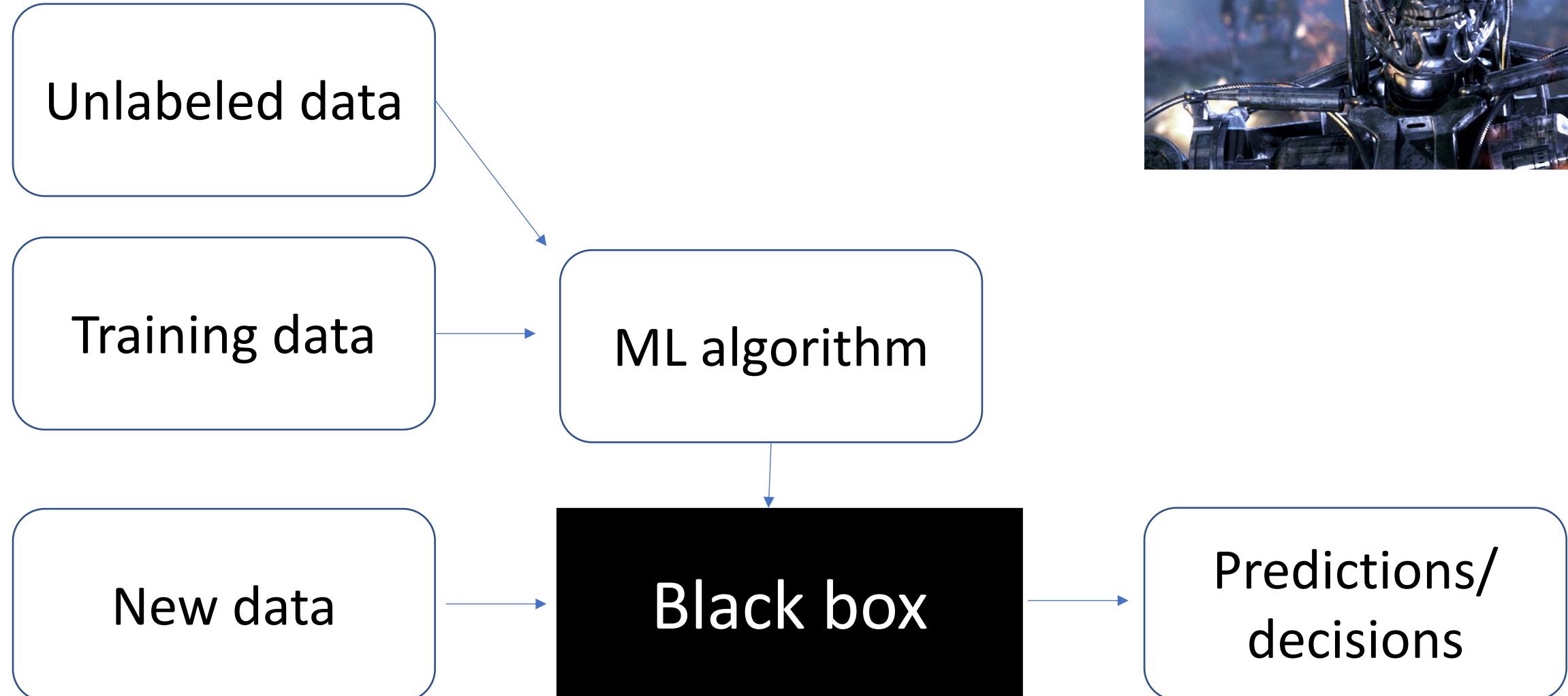
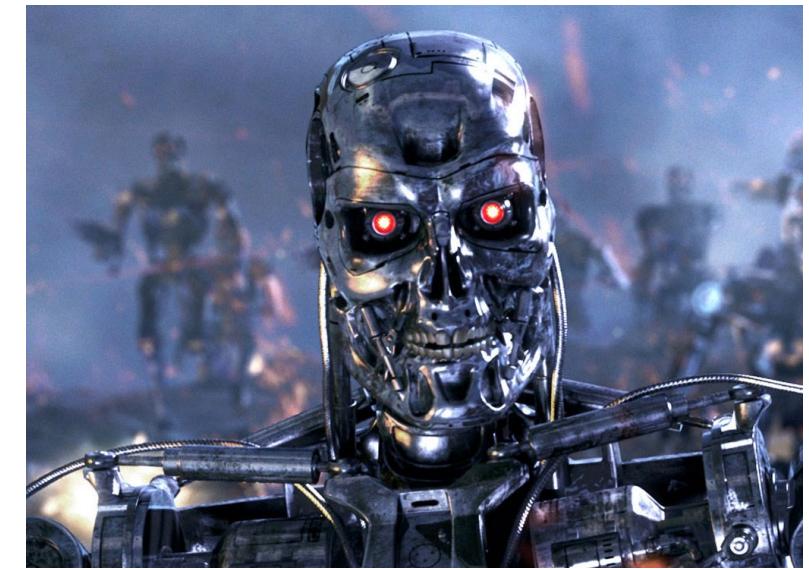
OKAY,  
LET'S  
PRETEND  
THAT  
MATTERS.

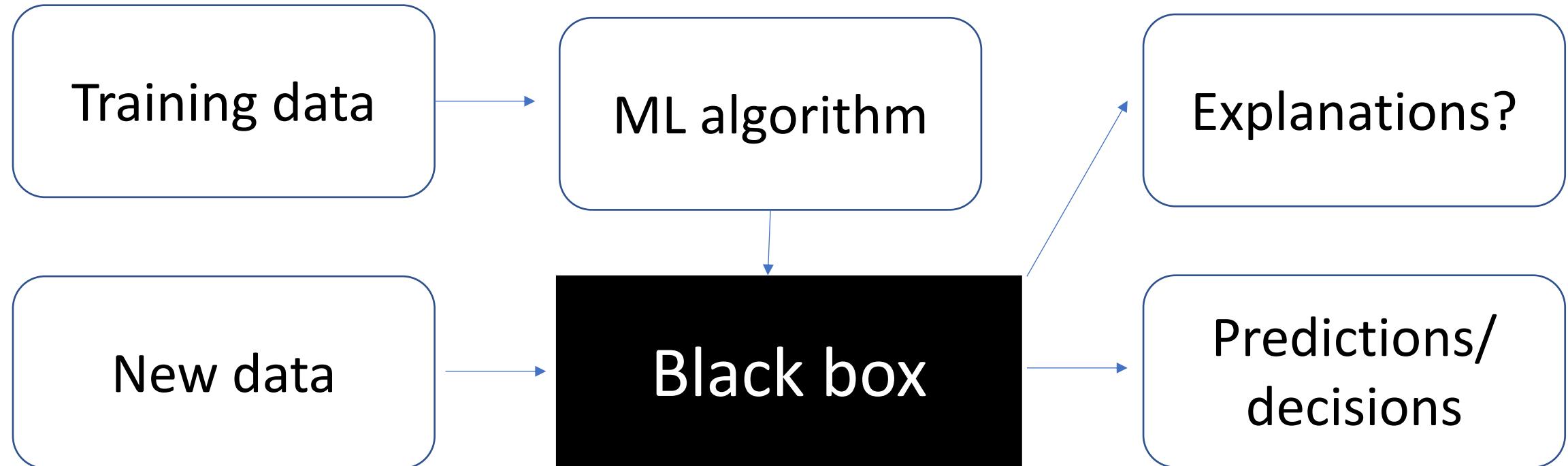
New data

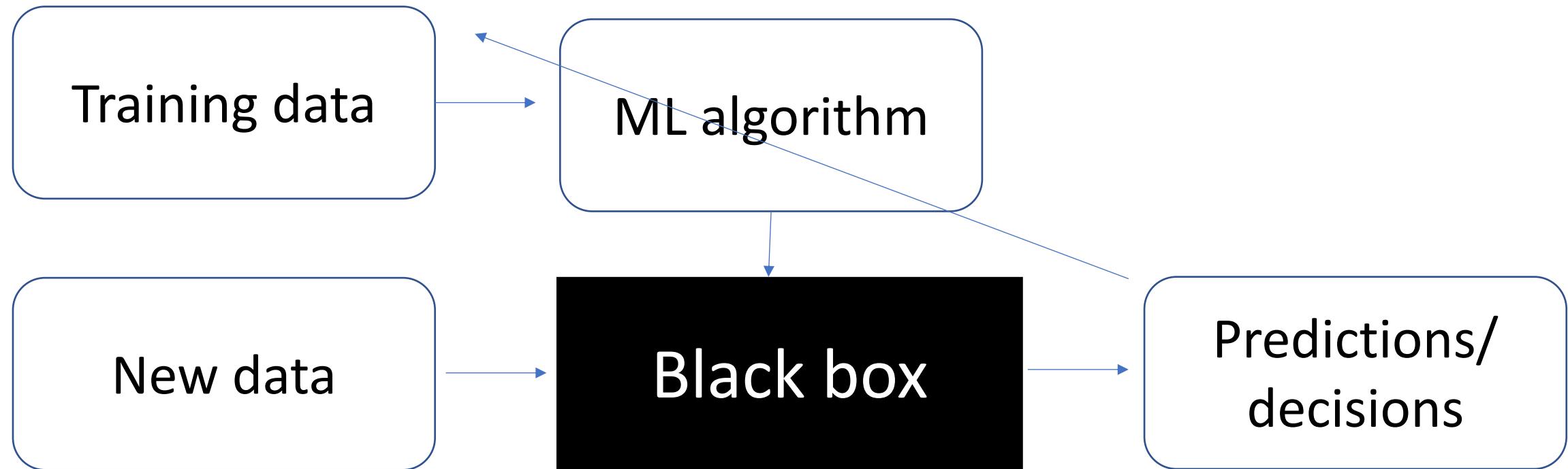
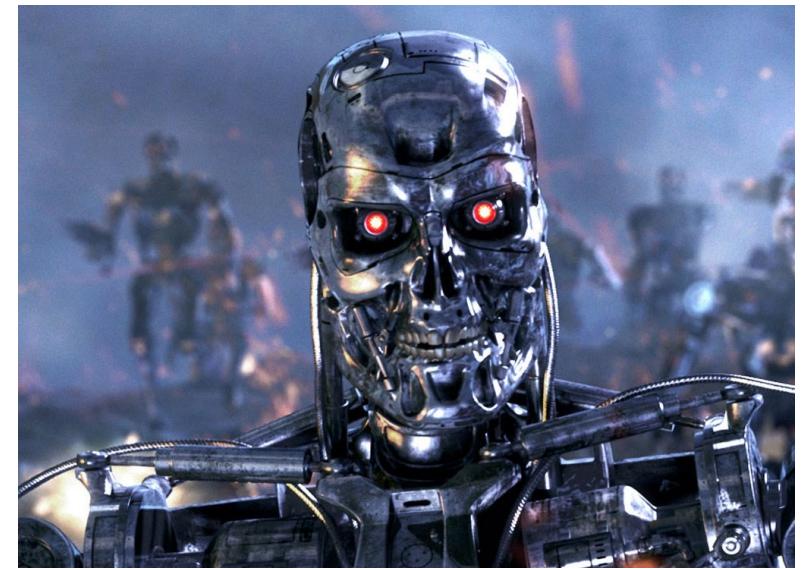
Black box

Predictions/  
decisions

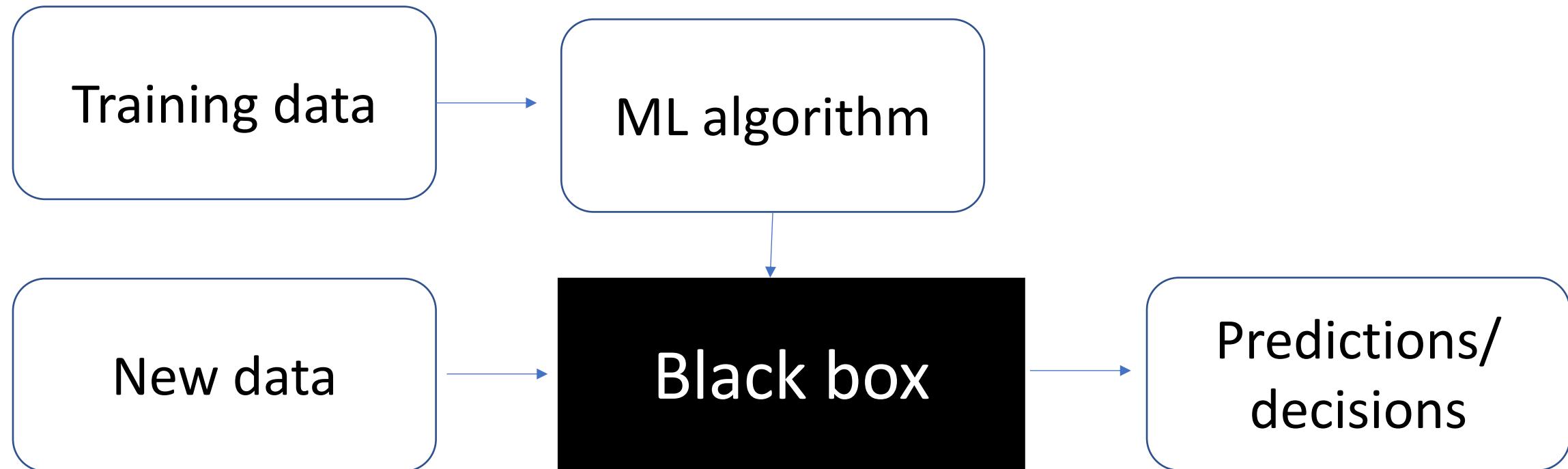








# Where do the costs lie?



# Where do the costs lie?



**Big Data Borat**  
@BigDataBorat

...

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

8:47 PM · Feb 26, 2013 · Twitter Web Client

---

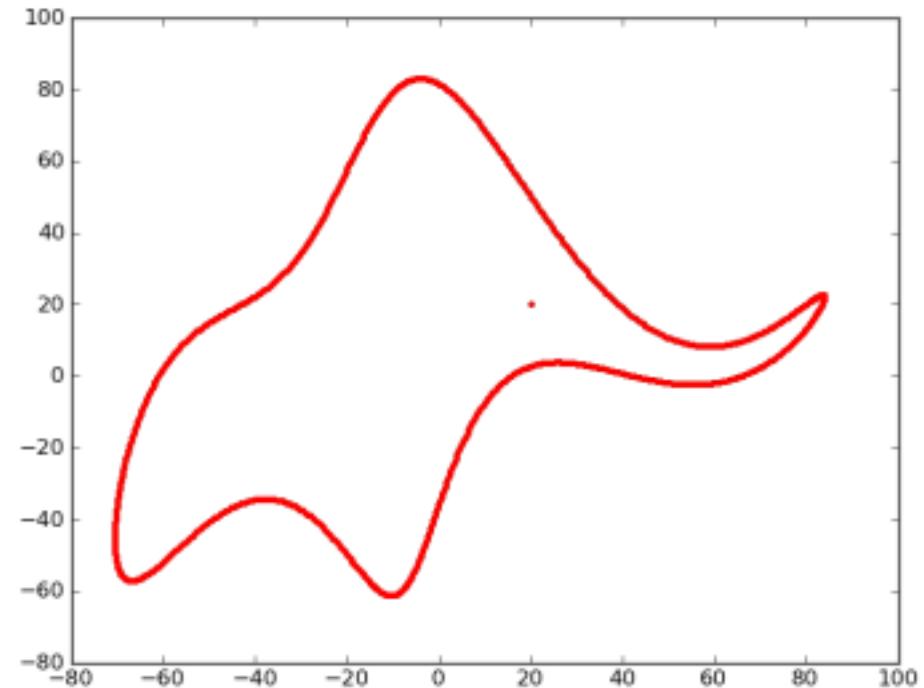
544 Retweets 25 Quote Tweets 402 Likes

Most of the cost is in getting worthwhile, relevant, usable training data.

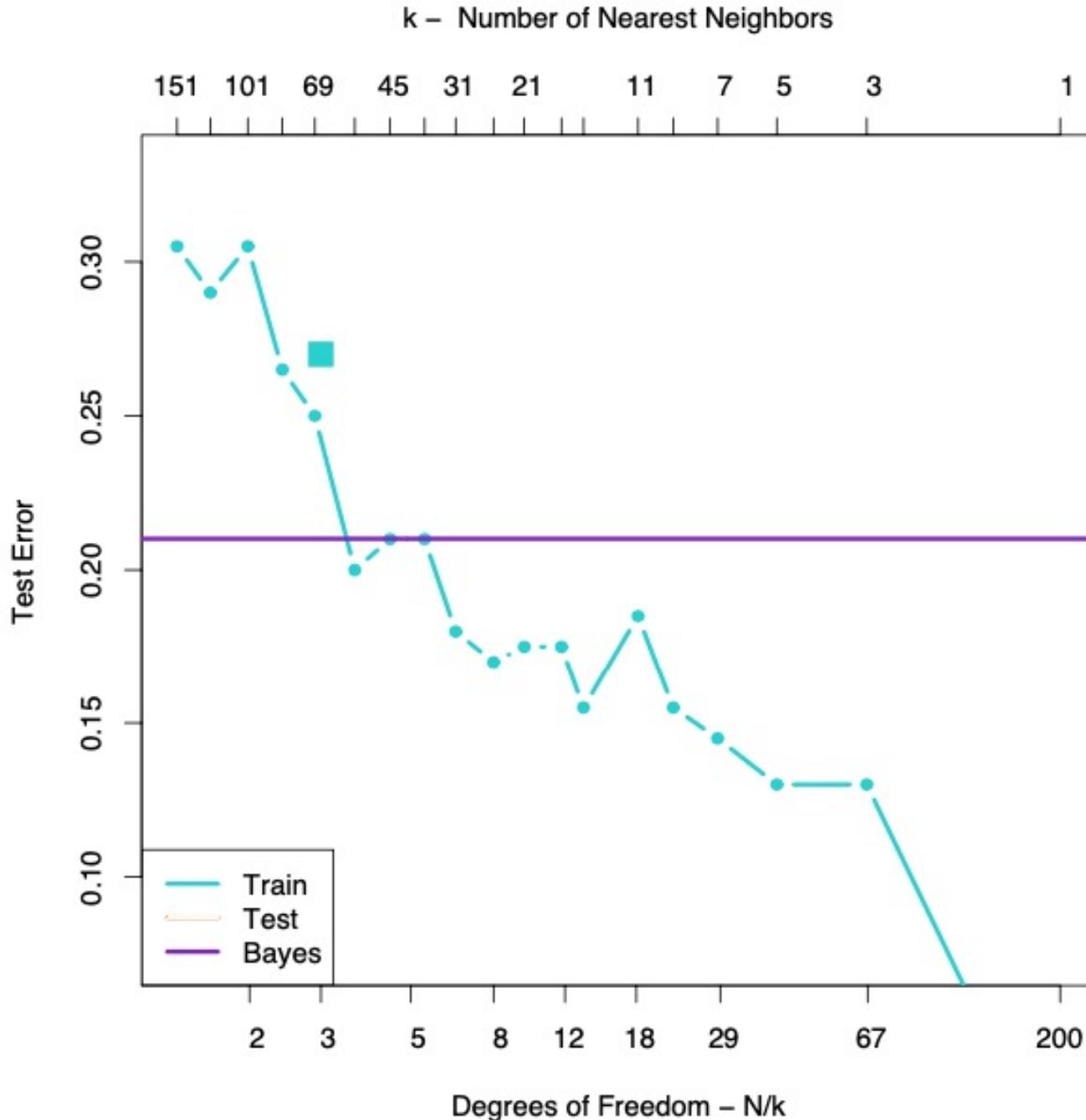
With an ocean of free parameters, we can fit anything.

**“with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.”**

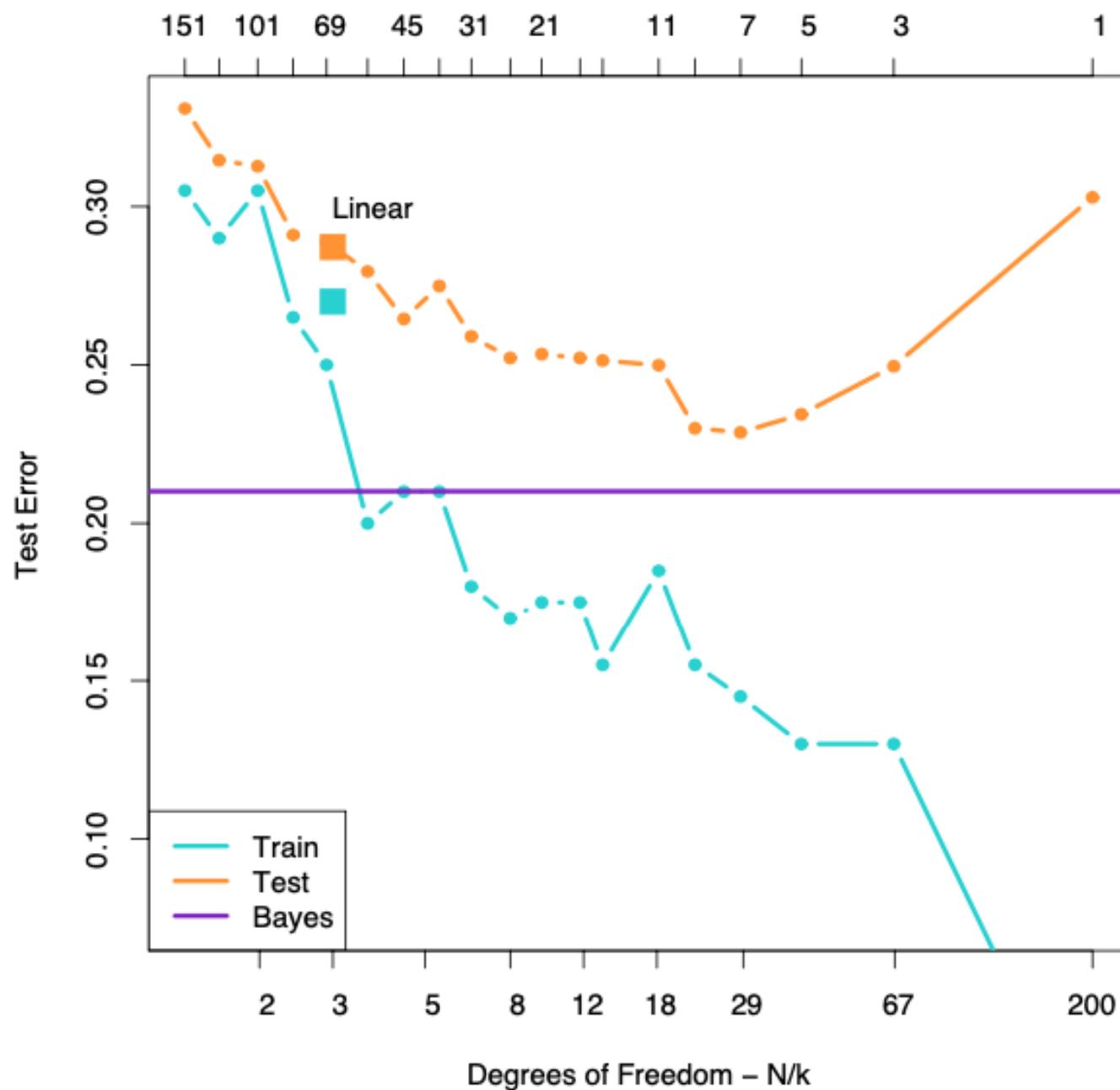
– John von Neumann, as told to Enrico Fermi, as told by Freeman Dyson.



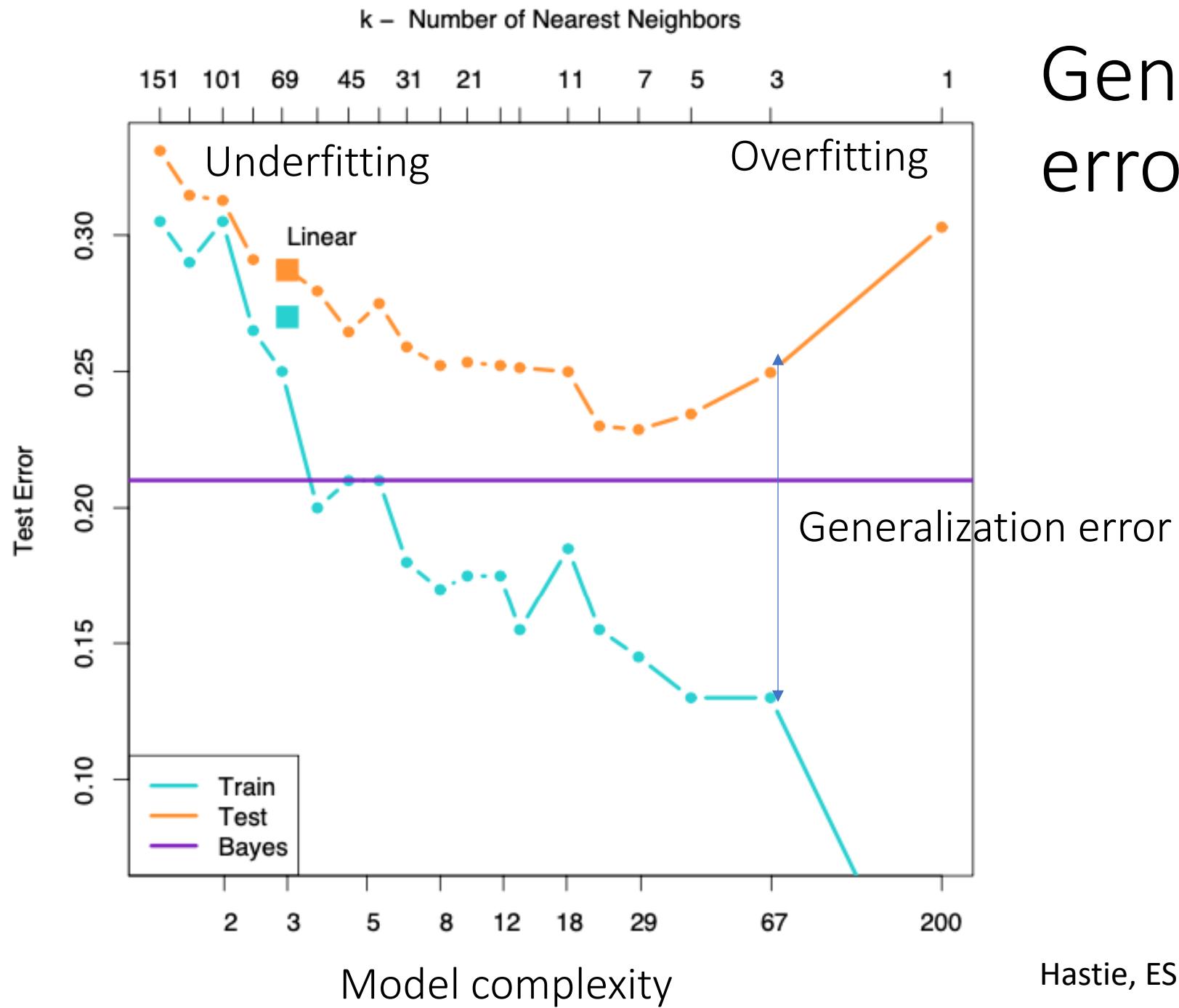
# Model complexity



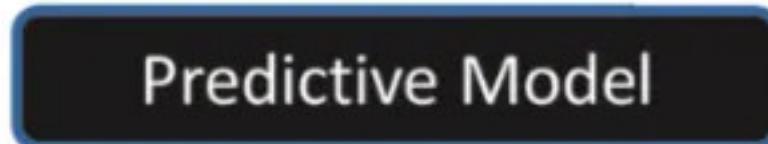
$k$  – Number of Nearest Neighbors



# Generalization error



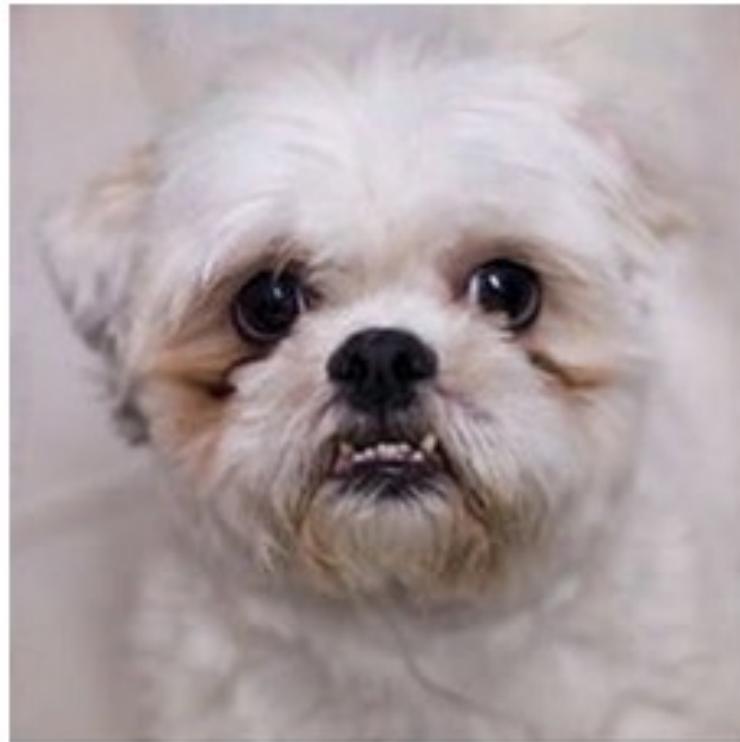
# Generalization error



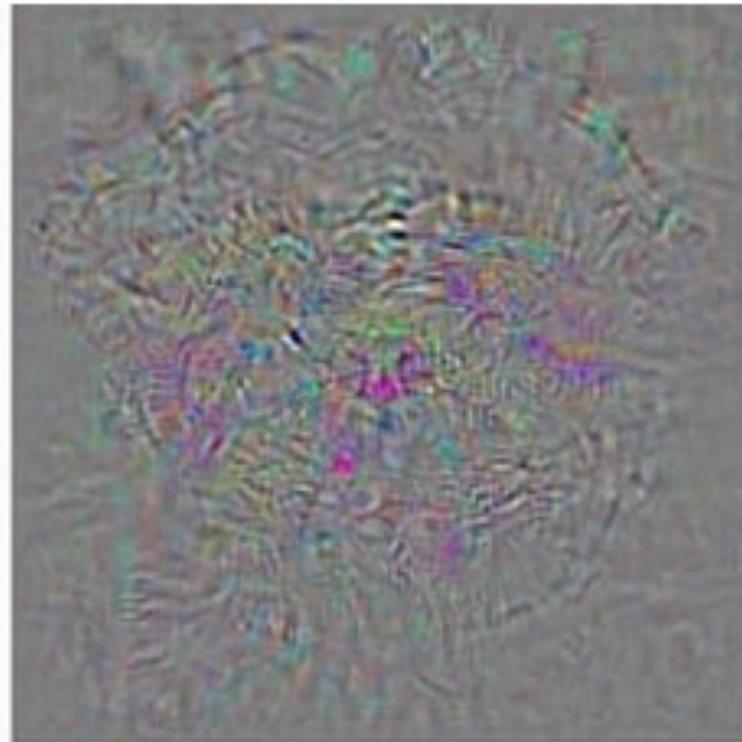
Cat  
(Conf. = 0.96)

# ML systems don't make mistakes the way that we do...

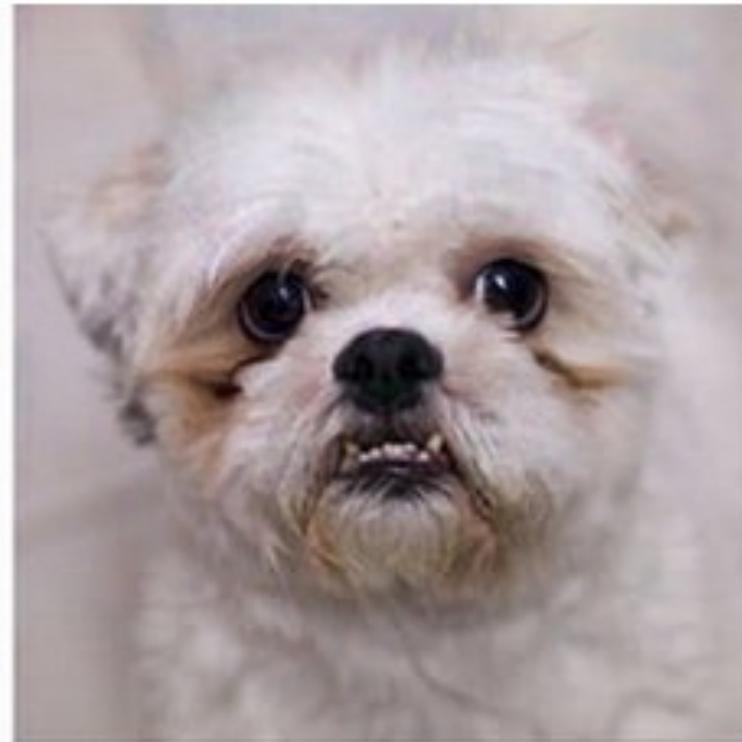
Maliciously constructed inputs can lead machine-decision makers astray



dog

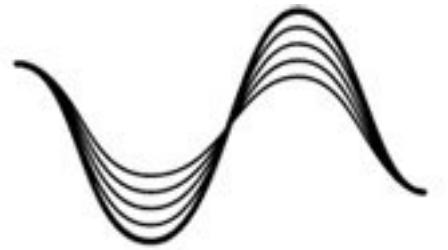


+noise



ostrich

# Data, ML systems are tools that benefit some more than others



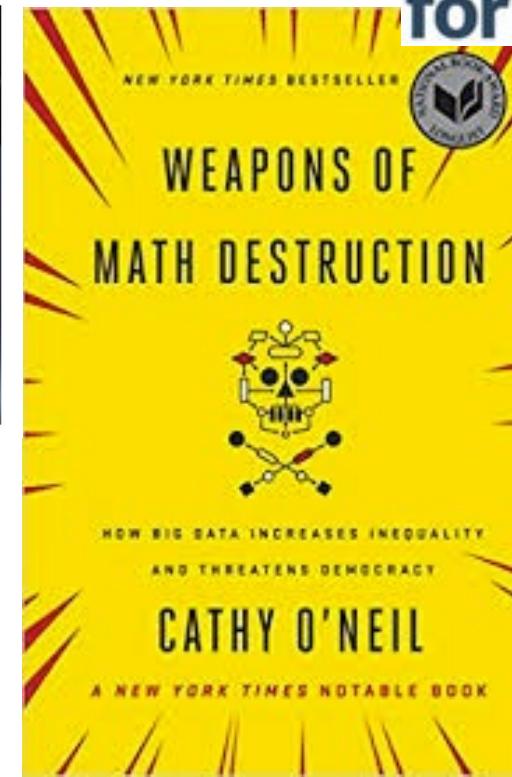
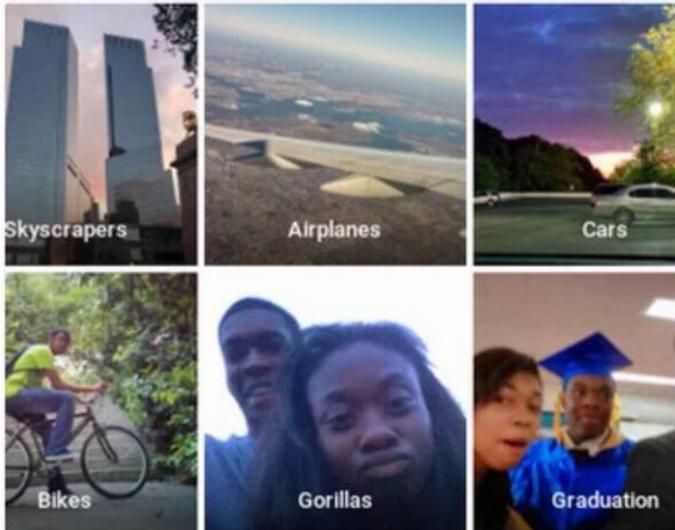
## Researchers Tape Speed Limit Sign to Make Teslas Accelerate to 85 MPH

By Ryan Whitwam on February 19, 2020 at 1:01 pm | [Comments](#)

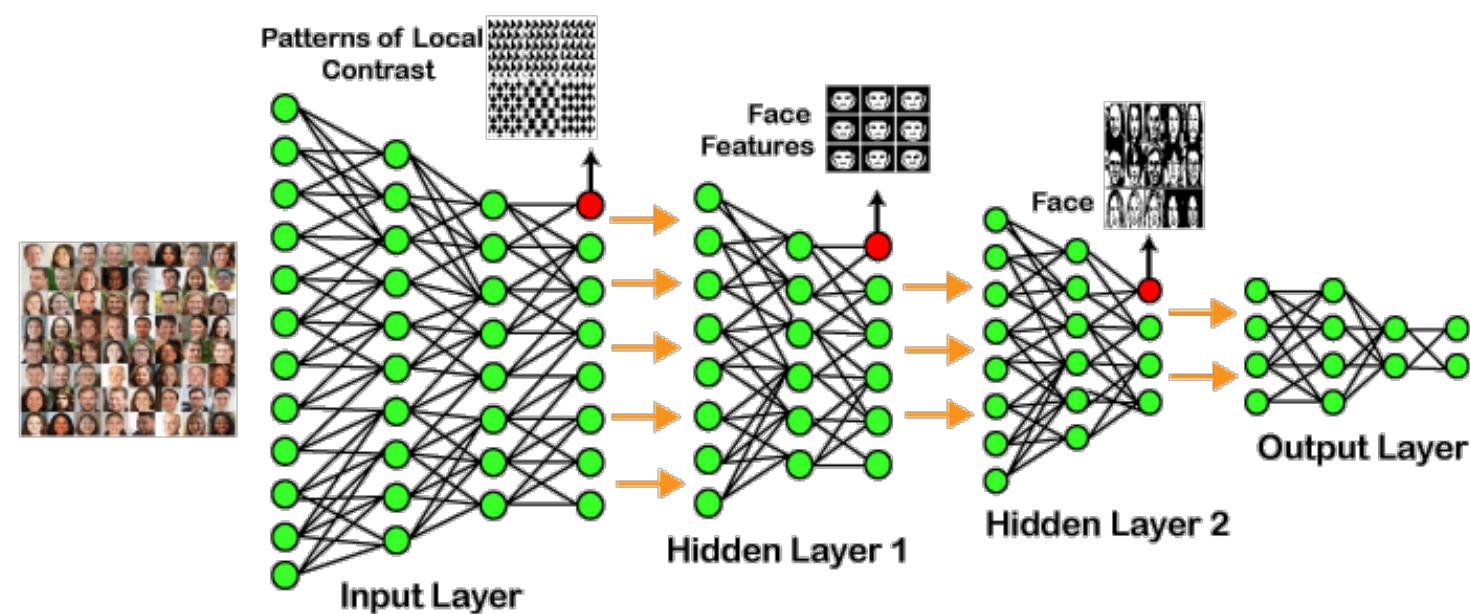


Jacky Alciné  
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

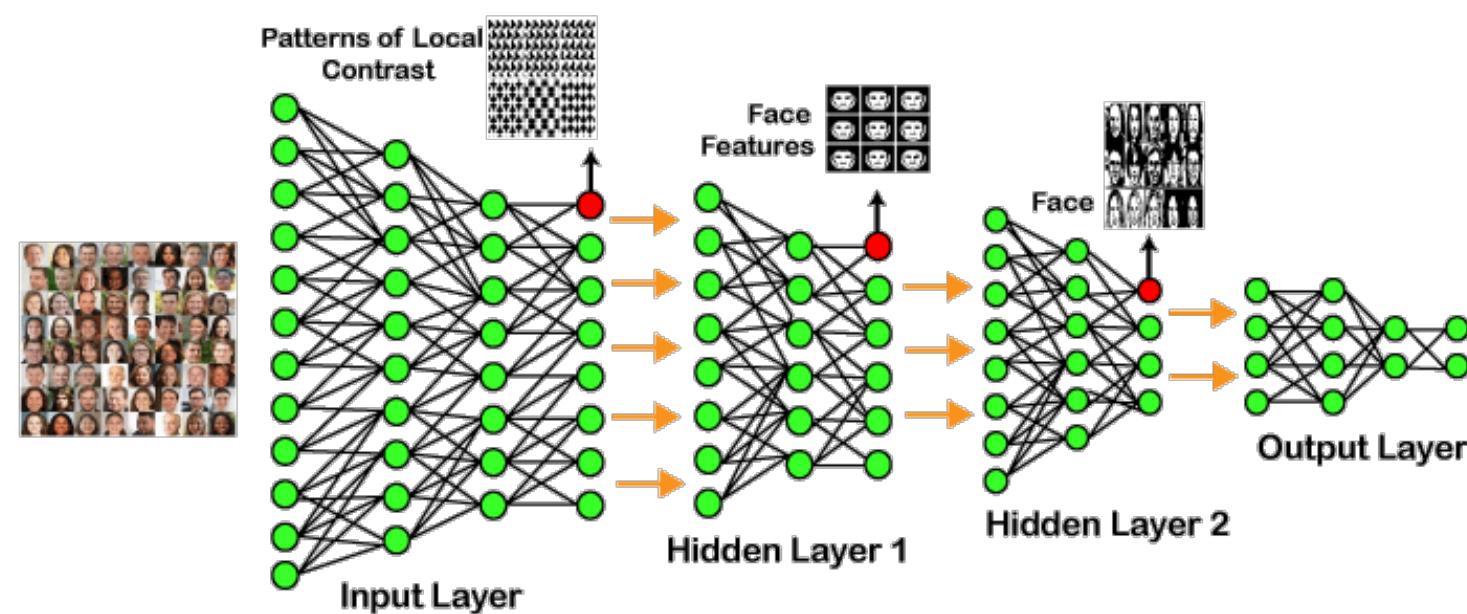


# What is good?



- We can simulate things, right? What is a good algorithm / model / procedure?

# What is good?



- We can simulate things, right? What is a good algorithm / model / procedure?
- Desired properties: generalize to data that perhaps aren't like the training data ???
- Desired properties: get good accuracy with comparatively little training data. !!!

# The toolbox

The toolbox has different kinds of things.

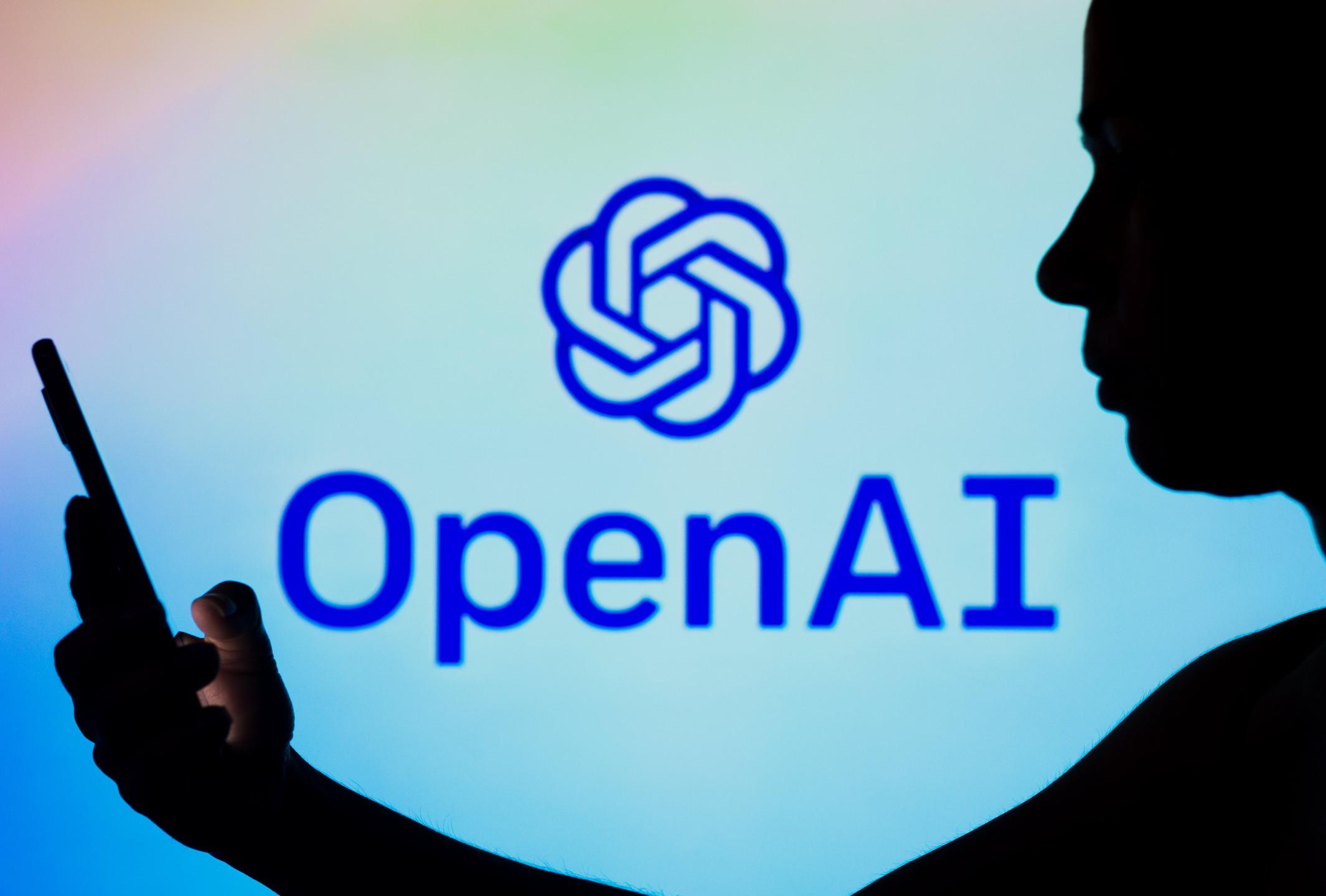
“Traditional” algorithms that work (E-M, backpropagation, simulated annealing)

Learning procedures sometimes correspond to Bayesian inference. Often can only draw samples via Markov Chain Monte Carlo sampling.

“Nonparametric” methods that look a little like adhockery

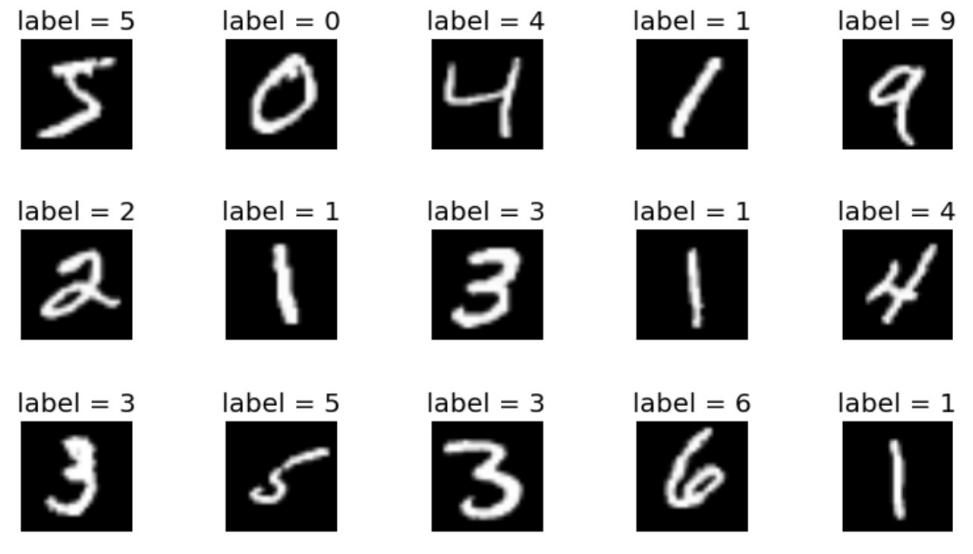
Optimization, stochastic gradient descent



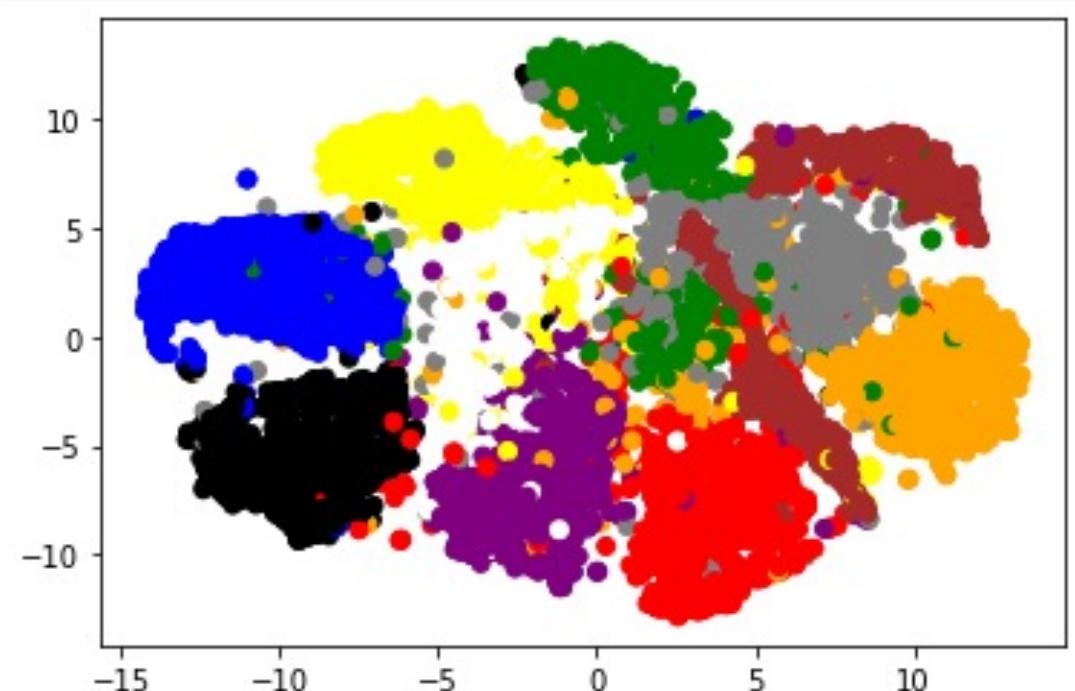


# The Plan

- Linear classifiers
- Nonparametric methods, clustering
- Optimization
- Evaluation, choosing algorithm
- Overfitting and regularization
- Neural networks
- Intro Deep Learning
- I will shoot for half math, half hacking



MNIST data set



# Administrivia

Expect turnover / attrition first week.

Homework (70%).

- Final project(30%).

# HW1



## Exercise 3.3. [9, p. 48] Inferring a decay constant

Unstable particles are emitted from a source and decay at a distance  $x$ , a real number that has an exponential probability distribution with characteristic length  $\lambda$ . Decay events can be observed only if they occur in a window extending from  $x = 1\text{ cm}$  to  $x = 20\text{ cm}$ .  $N$  decays are observed at locations  $\{x_1, \dots, x_N\}$ . What is  $\lambda$ ?

