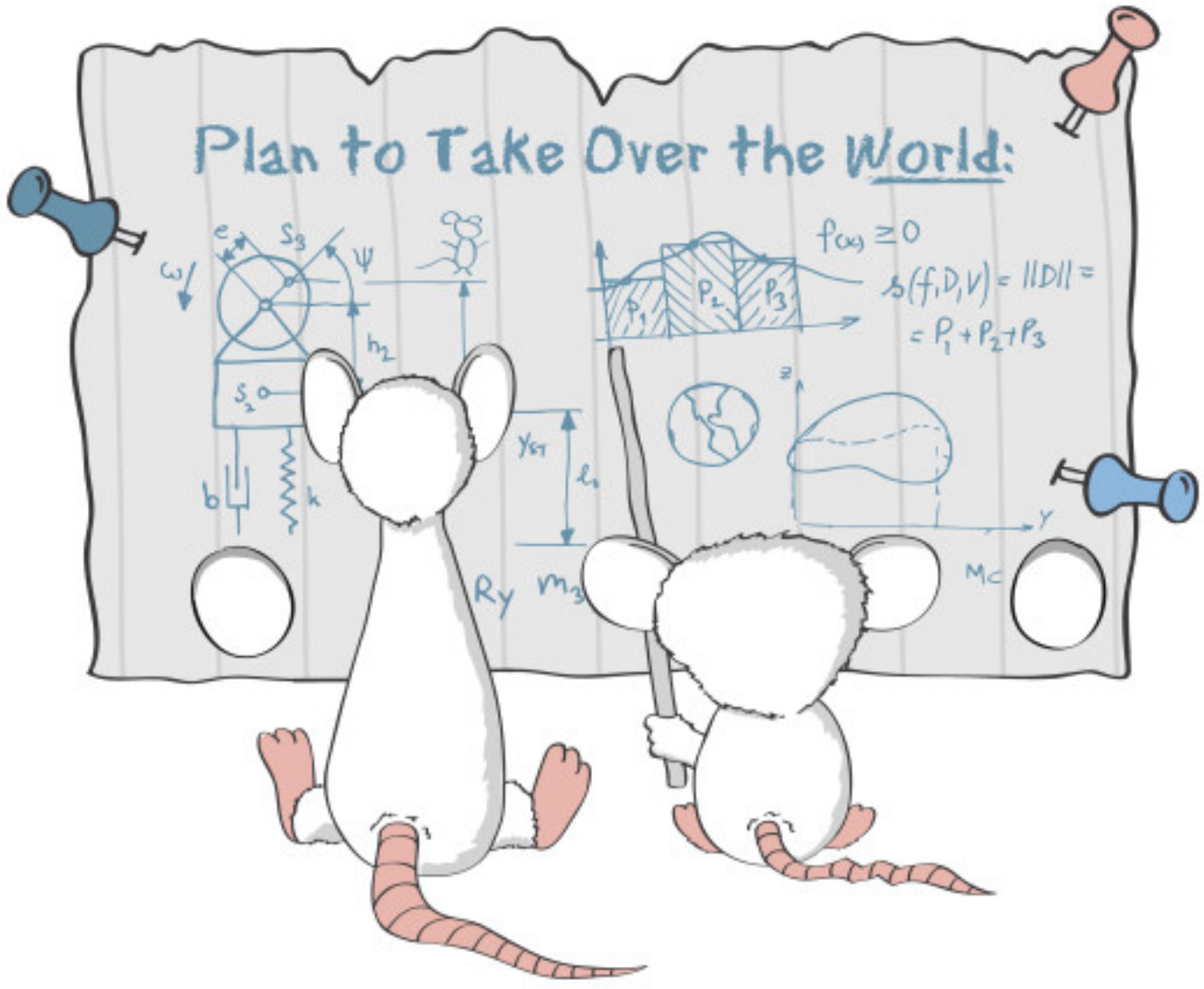
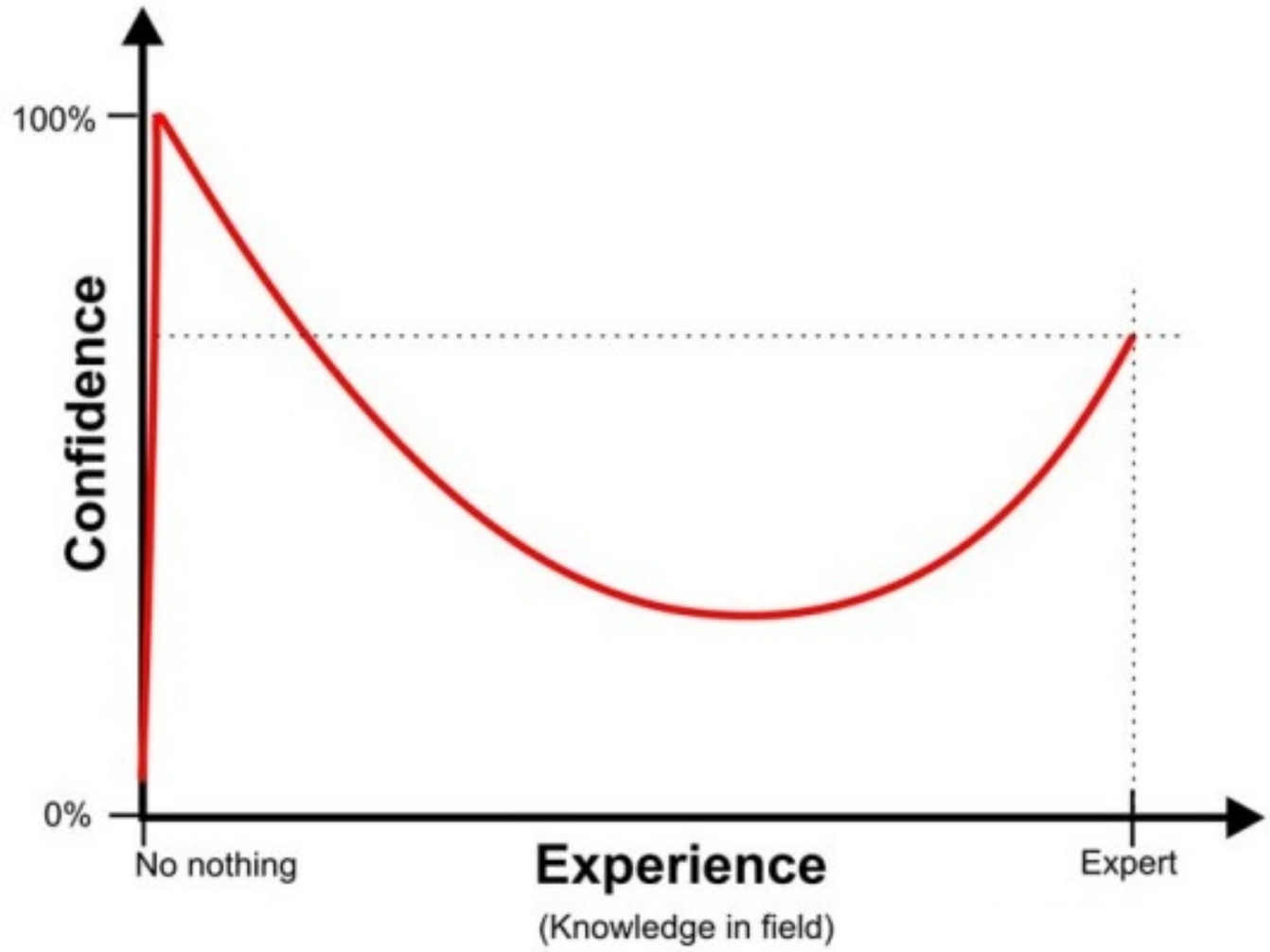


Work to Machine!

DIAS ML course by Martin Topinka

https://github.com/toastmaker/ml-dias

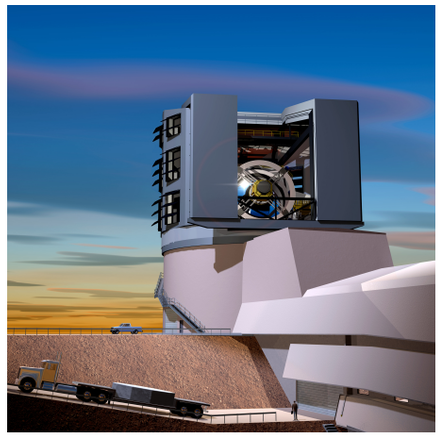

100%
Confidence
0%
No nothing
Experience
(Knowledge in field)
Expert

Introduce yourself and your goals

- Distinguish short GRBs flares from SGR flares from INTEGRAL
- Search for GRB afterglows in SDSS
- GRBs photometric redshift GROND
- Clustering of spectra (minimal spanning tree) in the distance-from-templates space in DFBS
- Search for M-stars, Carbon stars in DFBS
- JWST high-z: galaxy morphology classification
- JWST high-z: search for bars in galaxies
- JWST high-z: mock up catalog generation

Big Data in a Nutshell

- Data volume doubles every 18 months (Moore's law)
- LSST:
  - 10 years long movie of the sky
  - 30’000 GB/night (size of entire SDSS)
  - 1’000’000 transient alerts/night in differential imaging (faster and more reliable than catalog cross-matching)
  - 50% rubbish
  - Human time/attention does **not** scale :-(



Automated real-time classification is needed!

The Fourth Paradigm: Data-Intensive Scientific Discovery

Era of “Data-driven discoveries”

3Vs':

- **Volume (large archives)**
- **Velocity (continuous flow)**
- **Variety (complexity)**

LSST

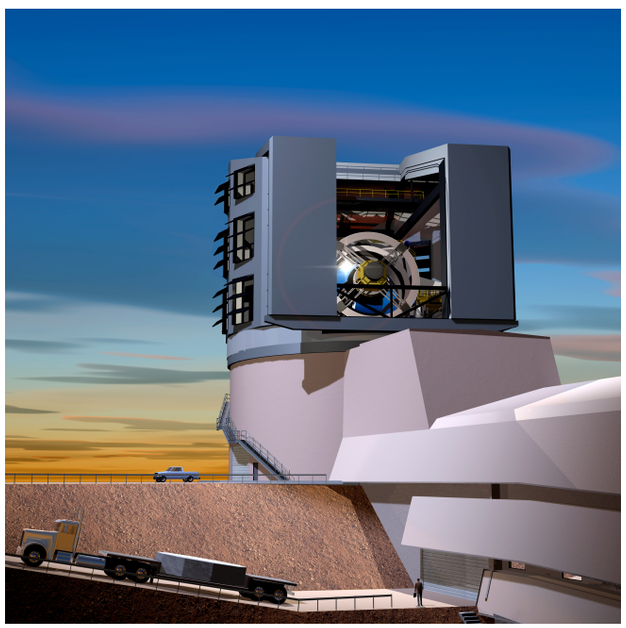
DR11 37 10^9 objects, 7 10^{12} sources,
5.5 million 3.2 Gigapixel images

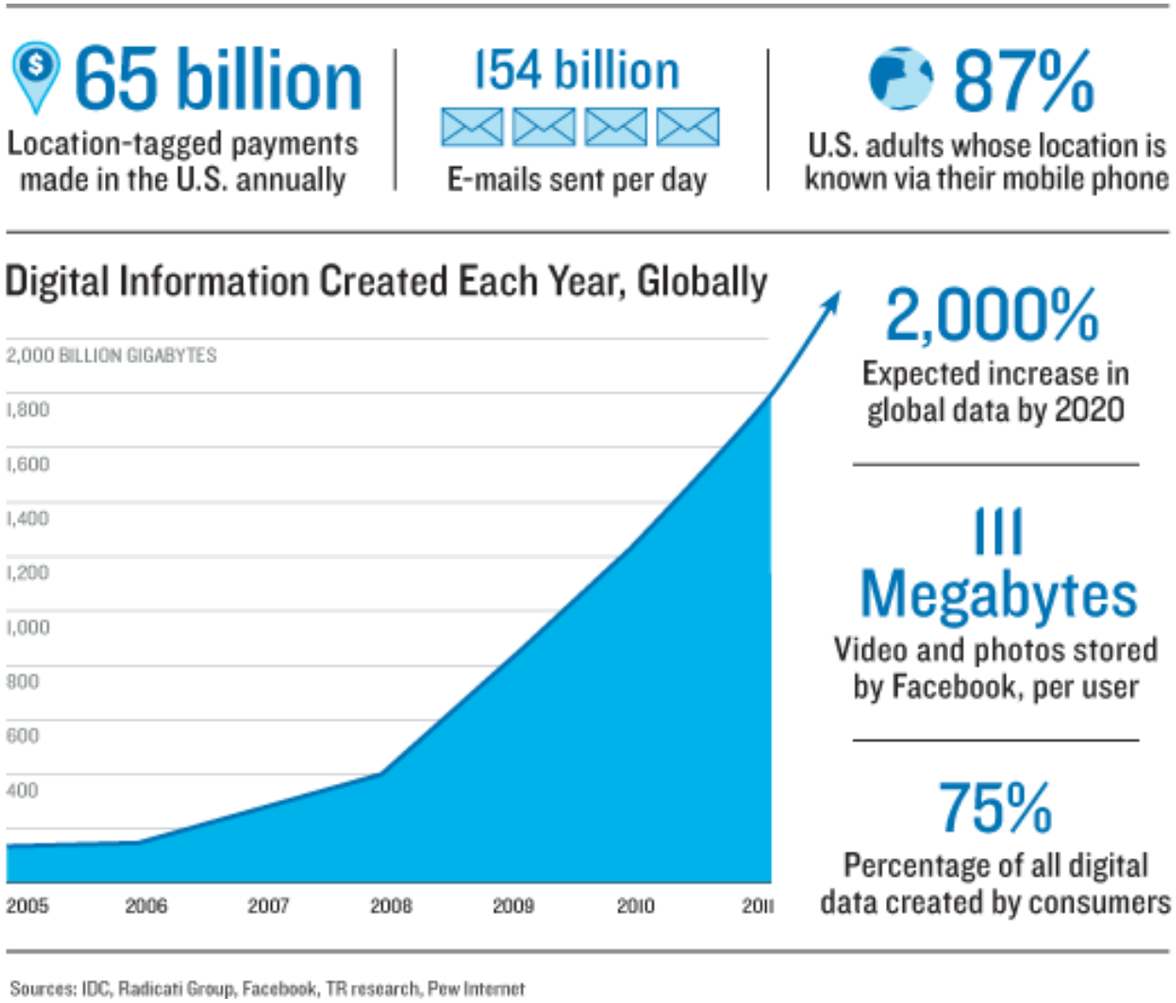
30 terabytes of data nightly

Final volume of raw image data = 60 PB

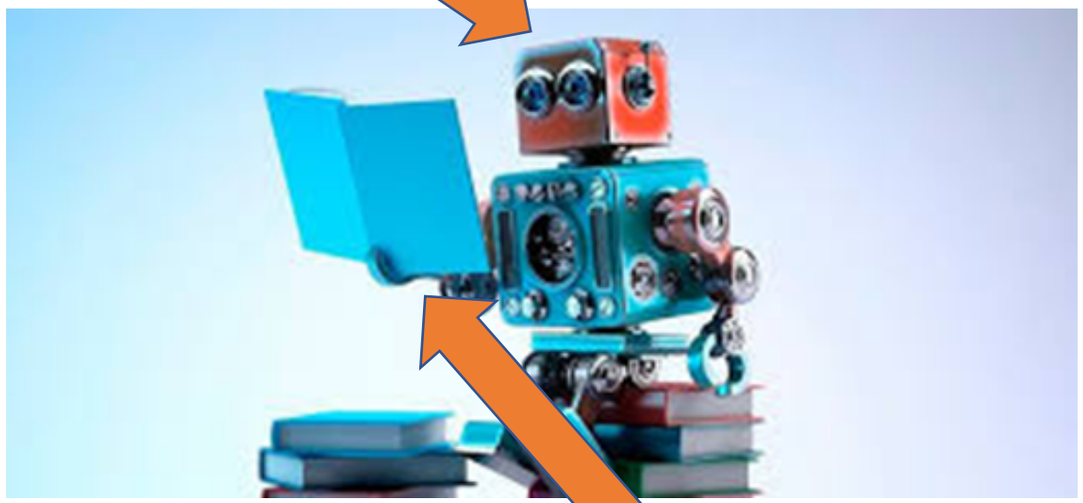
Final image collection (DR11) = 0.5 EB

Final catalog size (DR11) = 15 PB

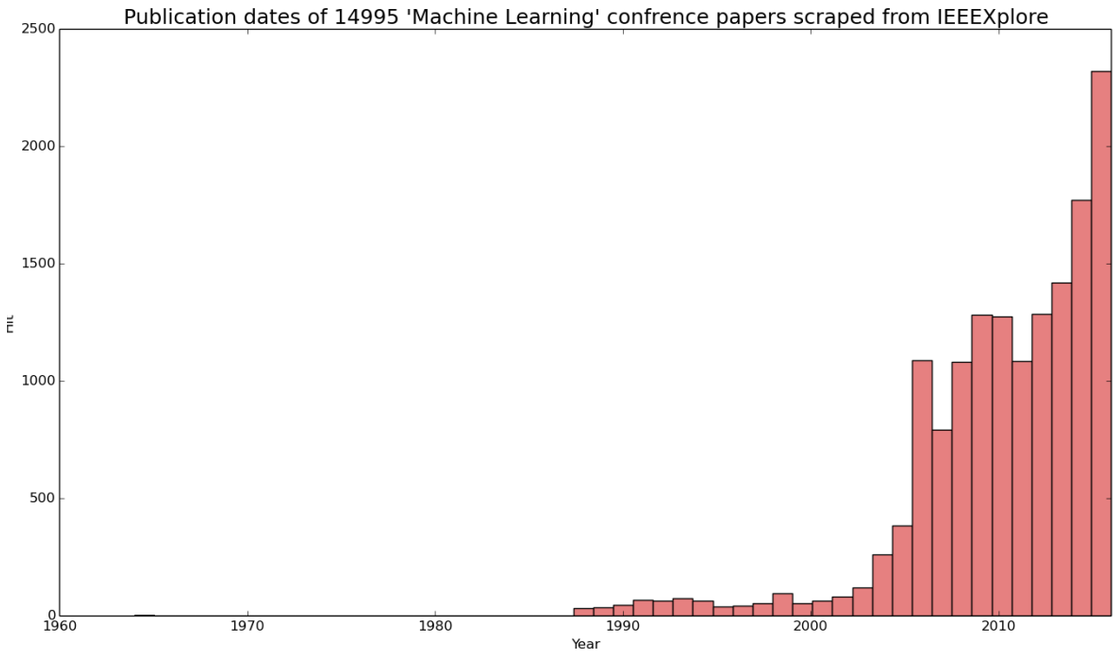




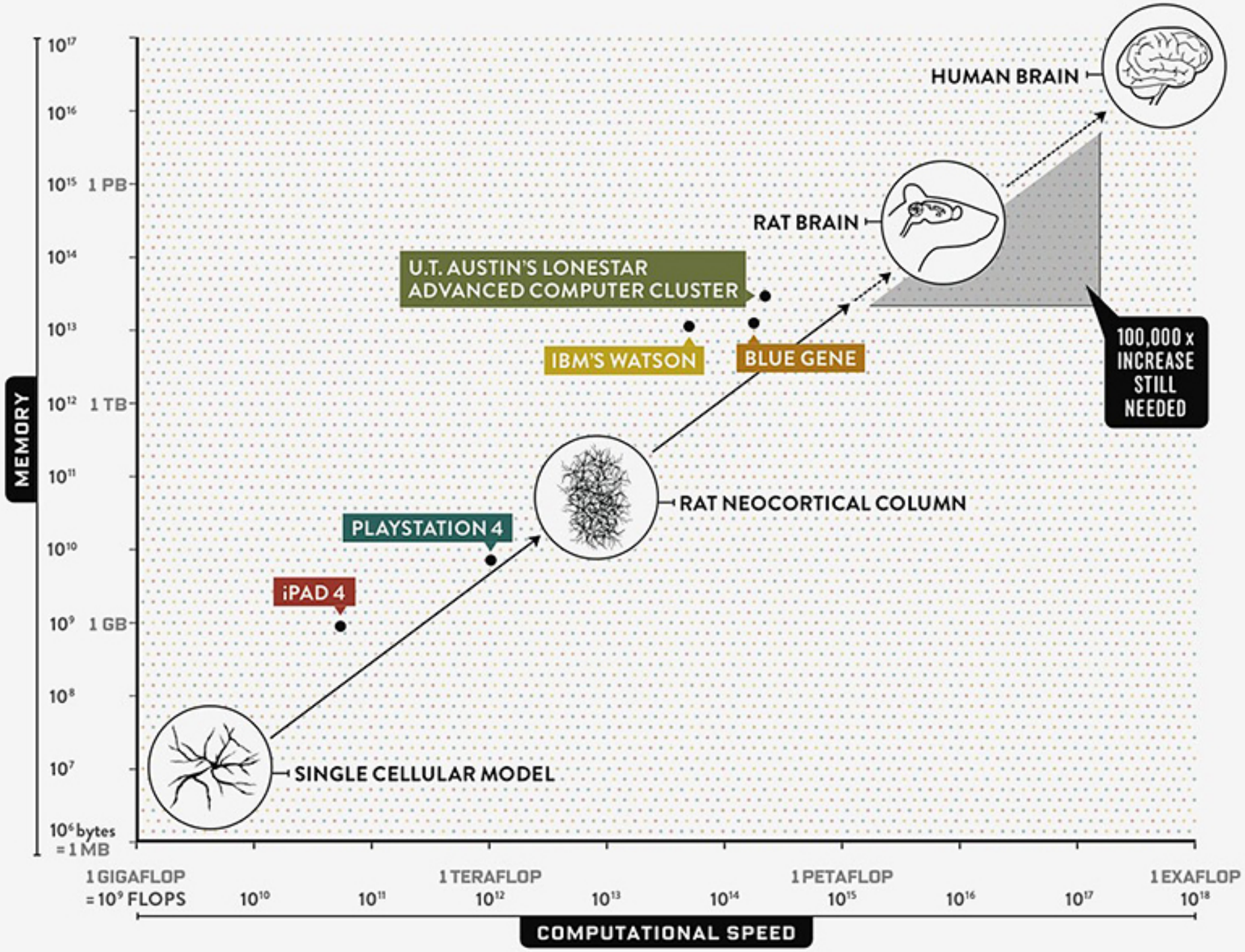
Machine…

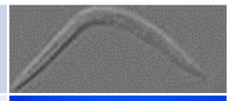
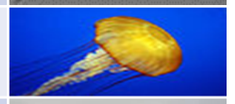


… Learning?!?



Machine learning algorithms can figure out how to perform tasks by generalising from examples (experience).


MEMORY
10^{17}
10^{16}
10^{15} 1 PB
10^{14}
10^{13}
10^{12} 1 TB
10^{11}
10^{10}
10^{9} 1 GB
10^{8}
10^{7}
10^{6} bytes = 1 MB
HUMAN BRAIN
RAT BRAIN
U.T. AUSTIN'S LONESTAR ADVANCED COMPUTER CLUSTER
IBM'S WATSON
BLUE GENE
100,000 x INCREASE STILL NEEDED
RAT NEOCORTICAL COLUMN
PLAYSTATION 4
iPAD 4
SINGLE CELLULAR MODEL
1 GIGAFLOP = 10^{9} FLOPS
10^{10}
10^{11}
1 TERAFLOP 10^{12}
10^{13}
10^{14}
1 PETAFLOP 10^{15}
10^{16}
10^{17}
1 EXAFLOP 10^{18}
COMPUTATIONAL SPEED

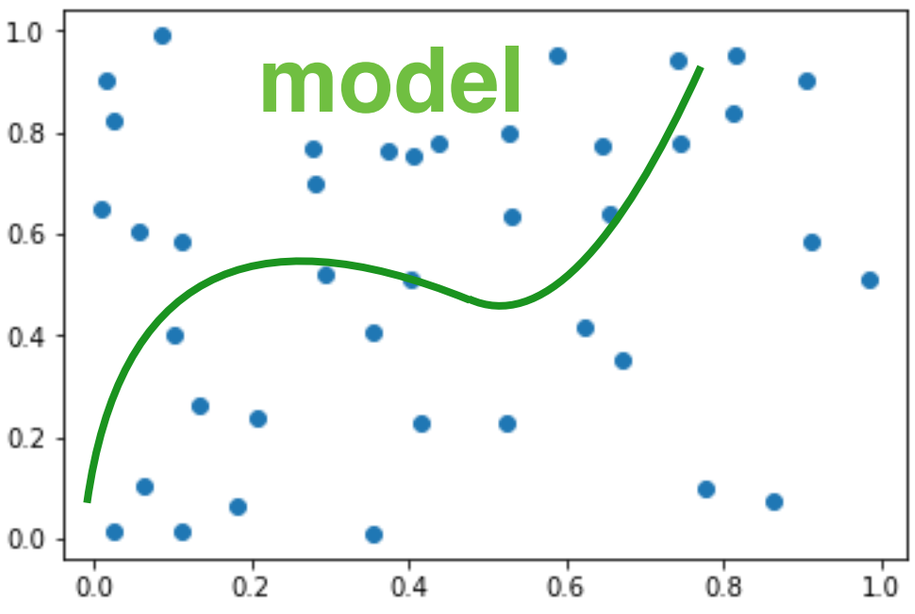
Name	# of neurons / # of synapses	Visuals
Caenorhabditis elegans	302	
Hydra vulgaris	5,600	
Homarus americanus	100,000	
Blatta Orientalis	1,000,000	
Nile Crocodile	80,500,000	
Digital Reasoning NN (2015)	**~86,000,000 (est.) / 1.6E11**	
Rattus Rattatouillensis	200,000,000	
Blue and yellow macaw	1,900,000,000	
Chimpanzee	28,000,000,000	
Homo Sapiens Sapiens	**86,000,000,000 / 1.5E14**	
African Elephant	257,000,000,000	

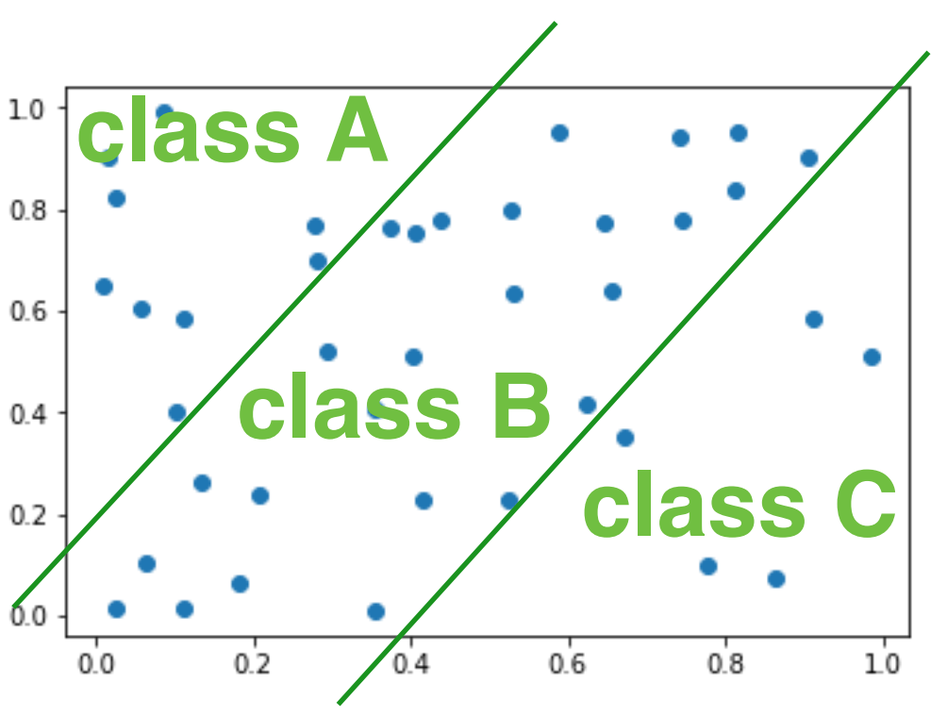
- **Supervised** - classification, regression (fitting)
- **Unsupervised** - clustering, graphs/trees, transformations, typically in multi-dim, outlier detection
- **Semi-supervised**, genetic algorithms, GANs...

The difference between a physicist and an astronomer:

The physicist sees random 2D data and draws a curved line in it saying it's the model that describes the data. The astronomer draws two parallel lines, saying these points belong to class A, these to class B and this is class C.

—- Andy Lawrence, private communication





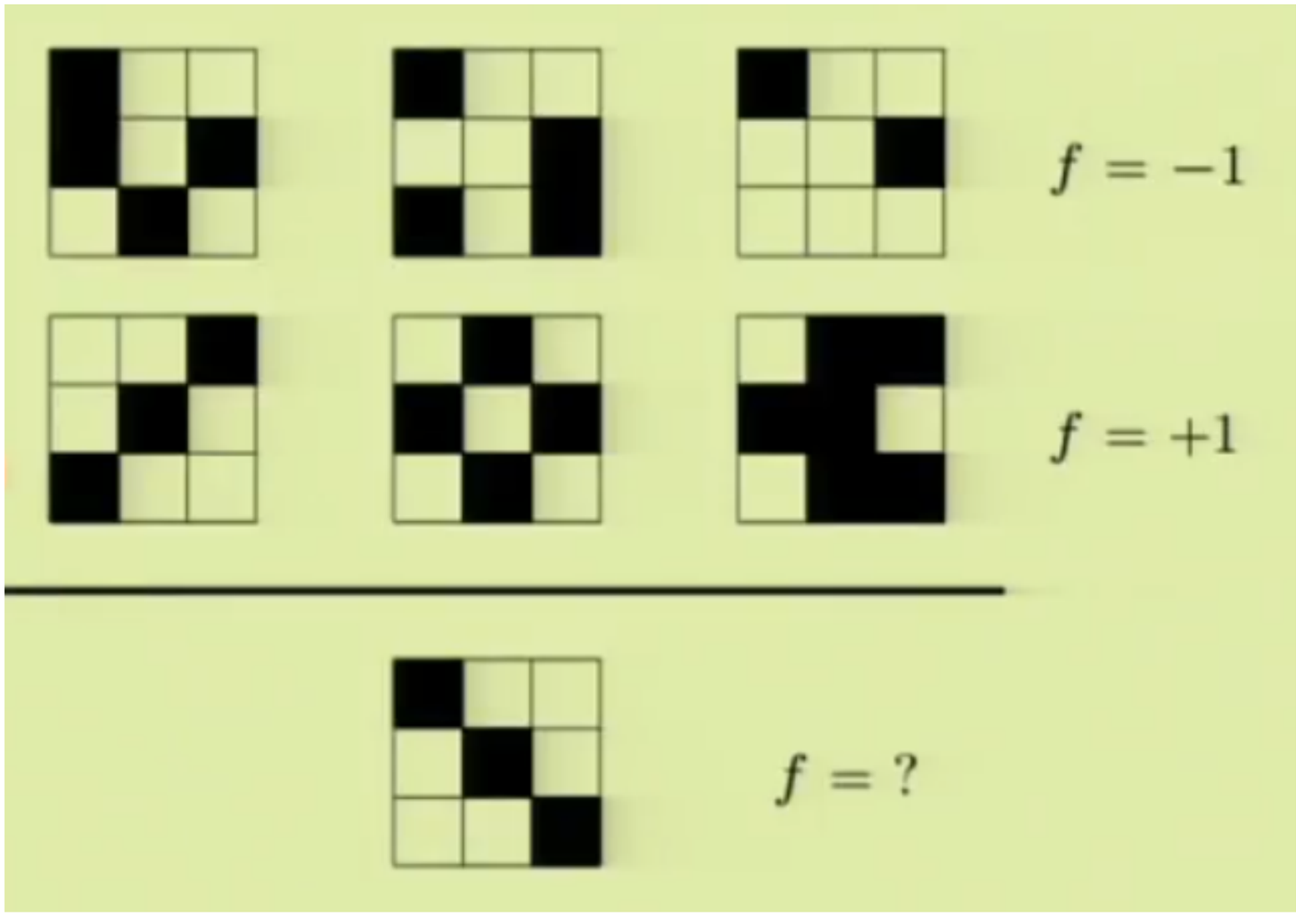
Some use-cases

- Self-driving cars
- Fraud detection in banks
- Spam/ham
- Object recognition in an image
- All sort of classification tasks when you know classes
- Clustering, new class detection, outlier detections, exploring multi-dim spaces, finding correlations, finding groups
- Replacing parts of a complicated simulation with ML engine
- Search for transit in expoplanets, time-series prediction
- Artificial real-looking catalogue building
- Denoising, data-compressing
- Feature selection

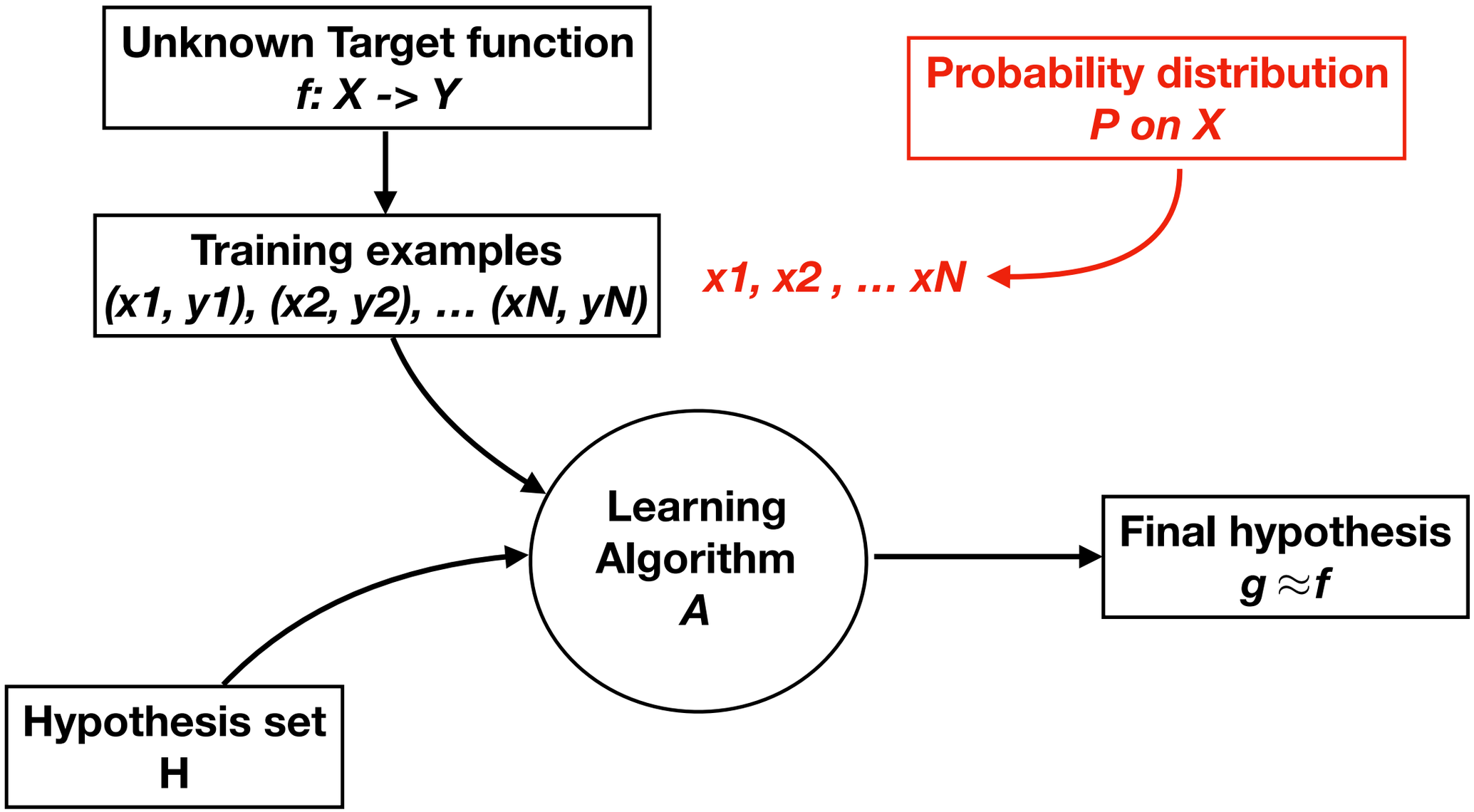
Essentials of Learning

- Pattern must exist
- Mapping “target function” is unknown or expensive to calculate
- We have the data (and computing resources…)
- Data sample is representative

Target function...

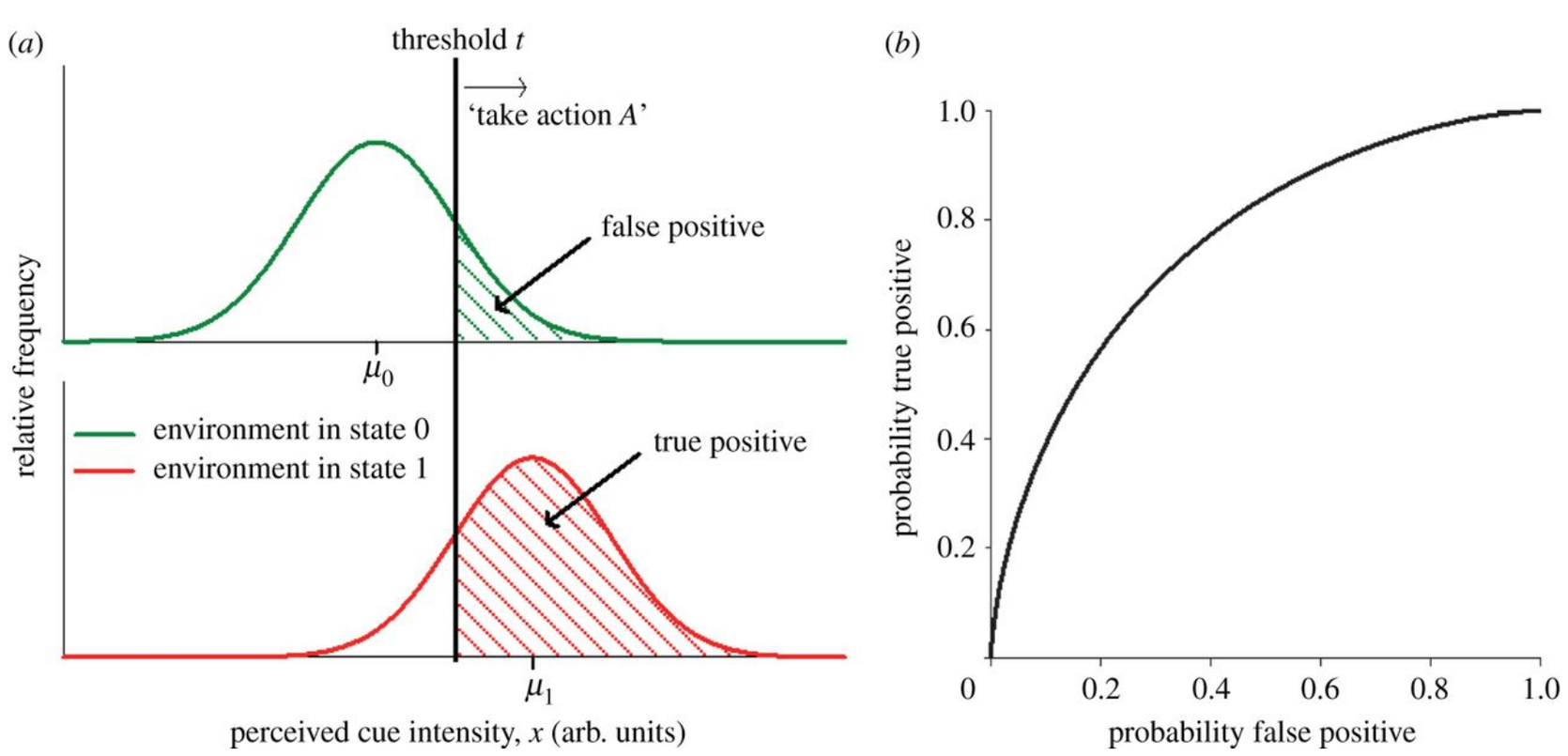


Learning Diagram



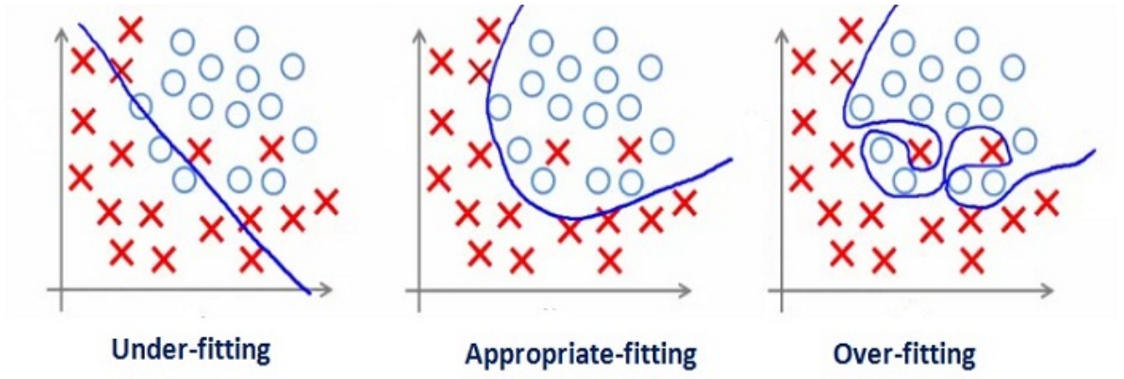
Loss function

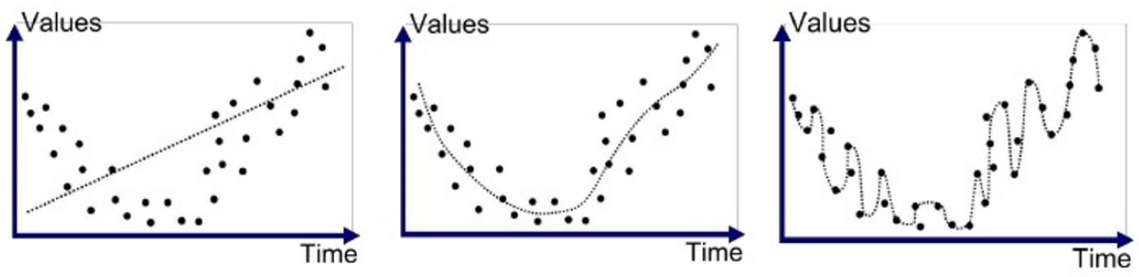
- penalty for being wrong
- imbalanced classes

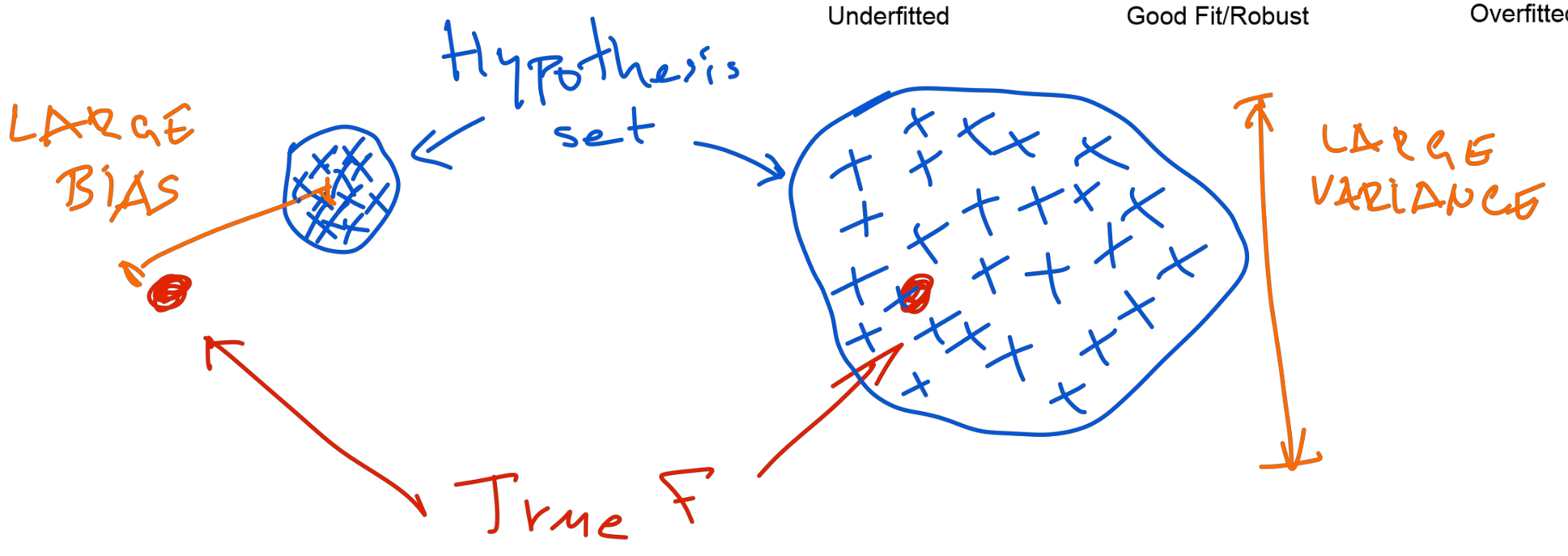


Bias - Variance Trade-off

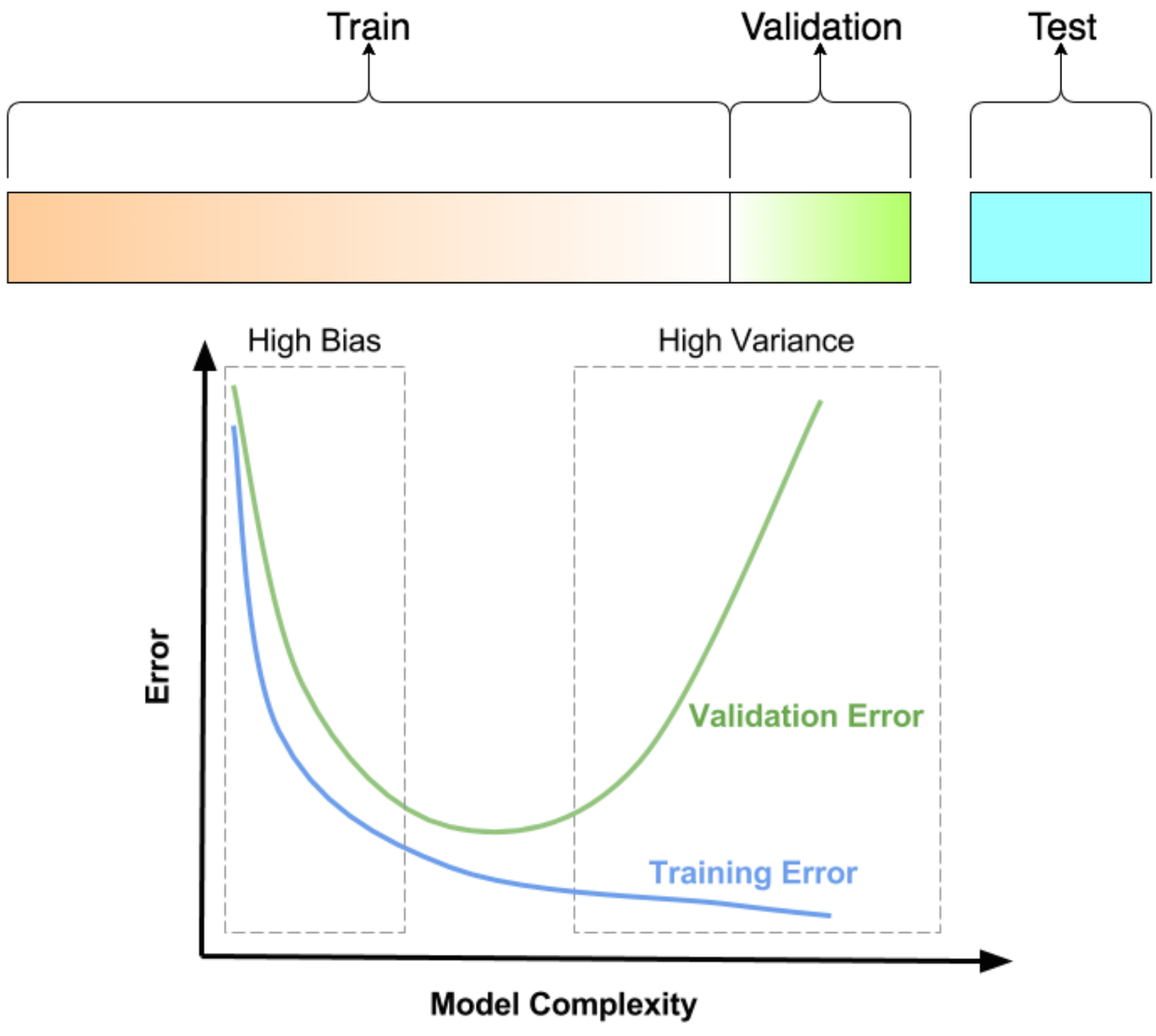
- Under-fitting/over-fitting
- in sample error vs out of sample error
- VC dimension





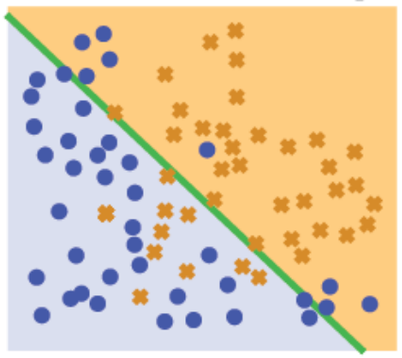


(cross)-Validation



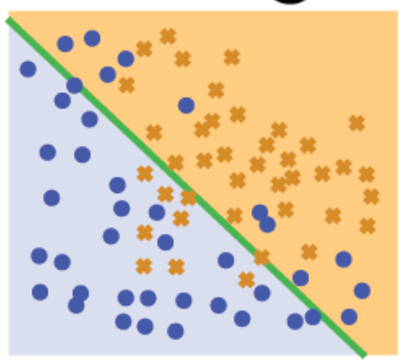
Model 1...

...on Training data. 1



30 10 error: 22.5%
32 8 acc. : 77.5%

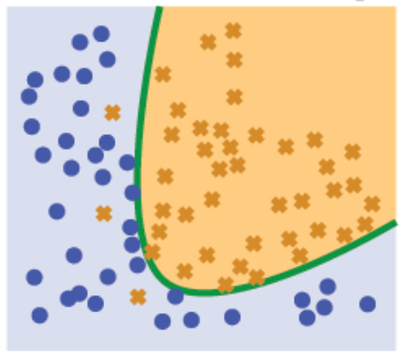
...on Test data. 4



32 8 error: 23.8%
29 11 acc. : 76.2%

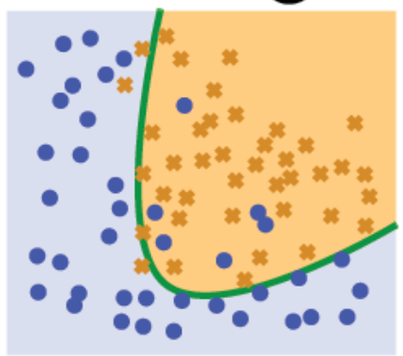
Model 2...

...on Training data. 2



37 3 error: 7.5%
37 3 acc. : 92.5%

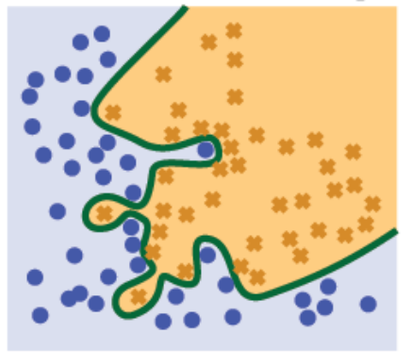
...on Test data. 5



37 3 error: 11.3%
34 6 acc. : 88.7%

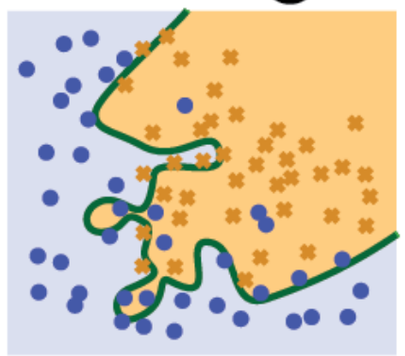
Model 3...

...on Training data. 3

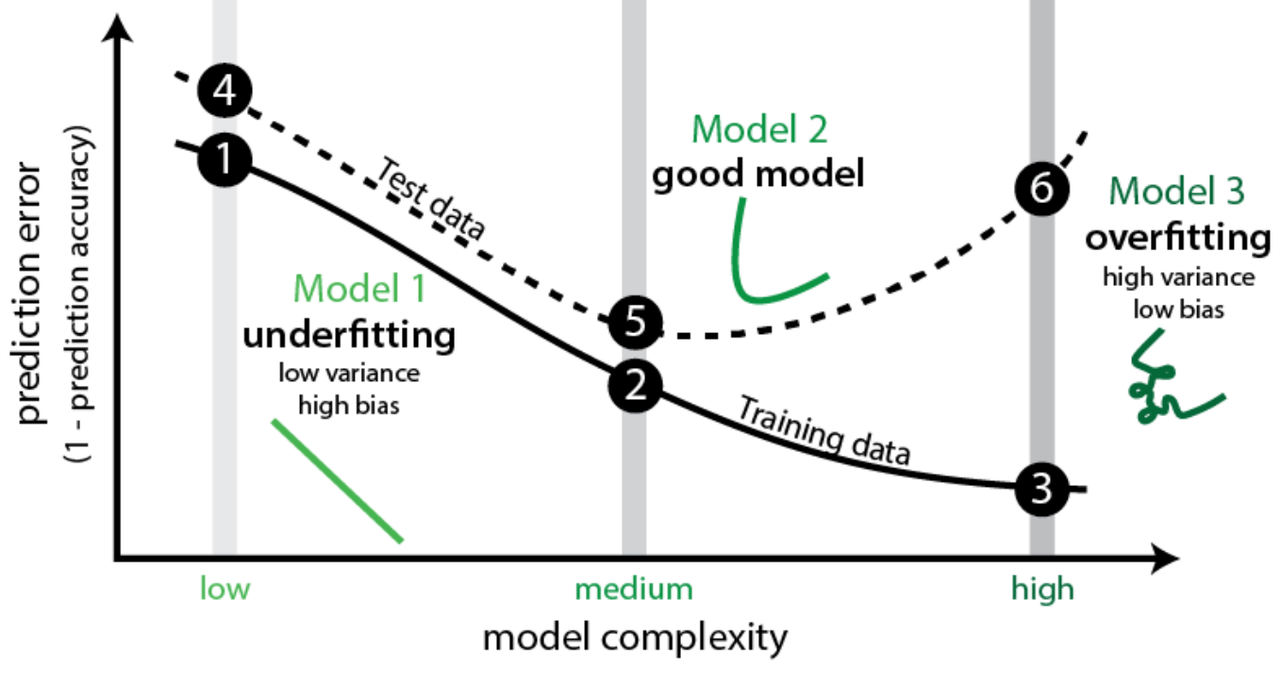


37 0 error: 0%
37 0 acc. : 100%

...on Test data. 6



34 6 error: 21.3%
29 11 acc. : 78.7%



	Remedies

High Bias	• Train longer
	• Increase model complexity
	  • more features
	  • more parameters,
	  • richer architecture

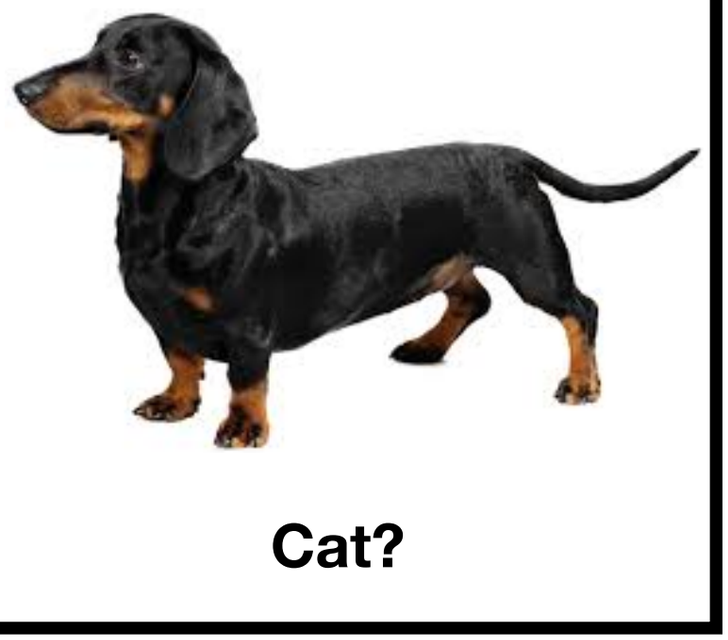
High Variance	• Get more data
	• Decrease model complexity
	  • less features
	  • less parameters,
	  • simpler architecture
	• Regularization
	• Early stopping
	• Drop-out

Data Augmentation:

- When more data are needed, make up new ones! (The way of the god.)
- Translate, rotate, flip, crop, lighten/darken, add noise, dephase, etc.

Cats

Dogs



Cat?

50% dog, 50% cat?

Machines are lazy and love shortcuts

Correlations != Reasoning

Justin King
@Justinkingnews

Headline: "Women who own horses live longer"

Implied correlation: horses make you live longer.

Reality: if you own a horse, you can probably afford health insurance.

9/10/18, 10:18 PM

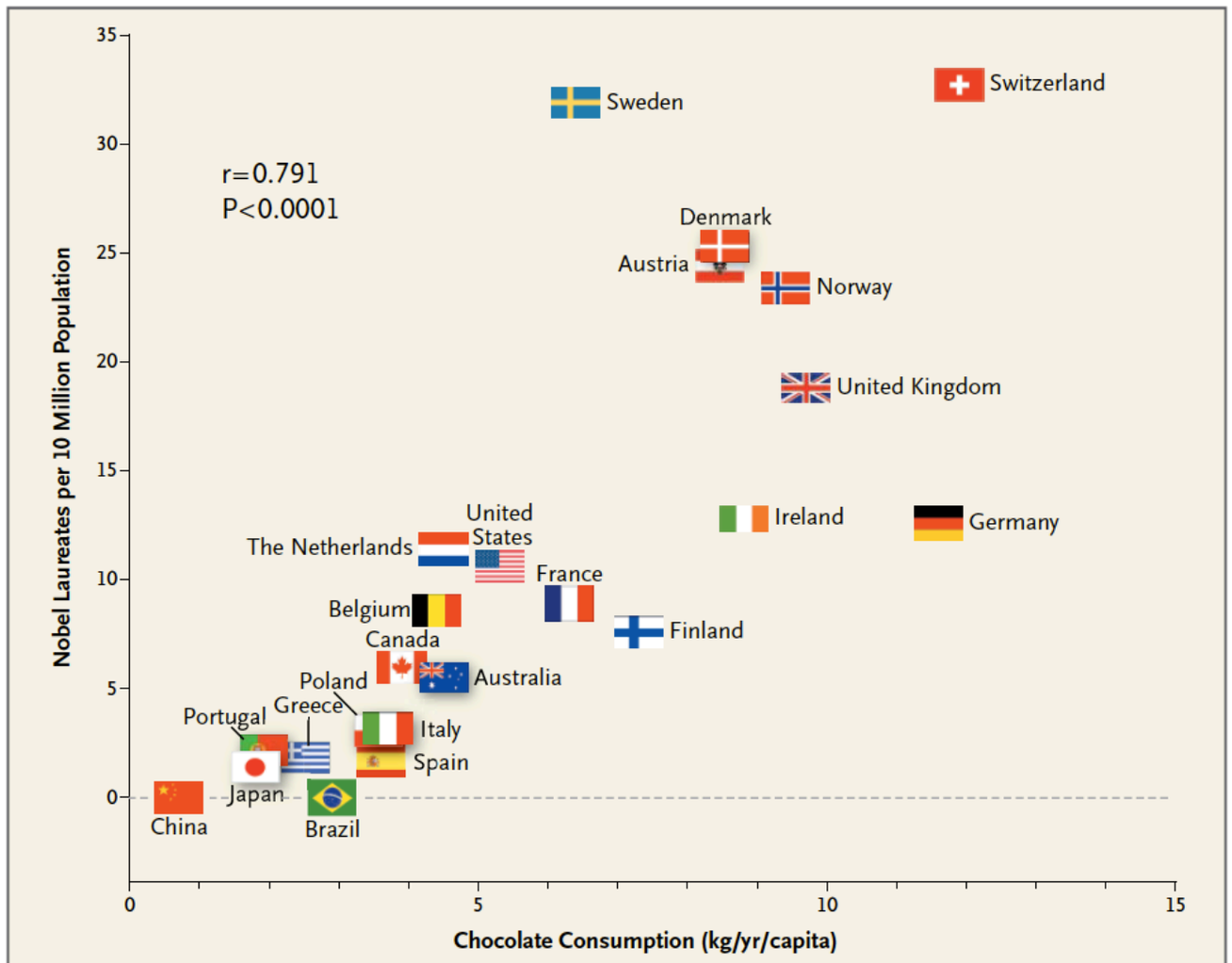
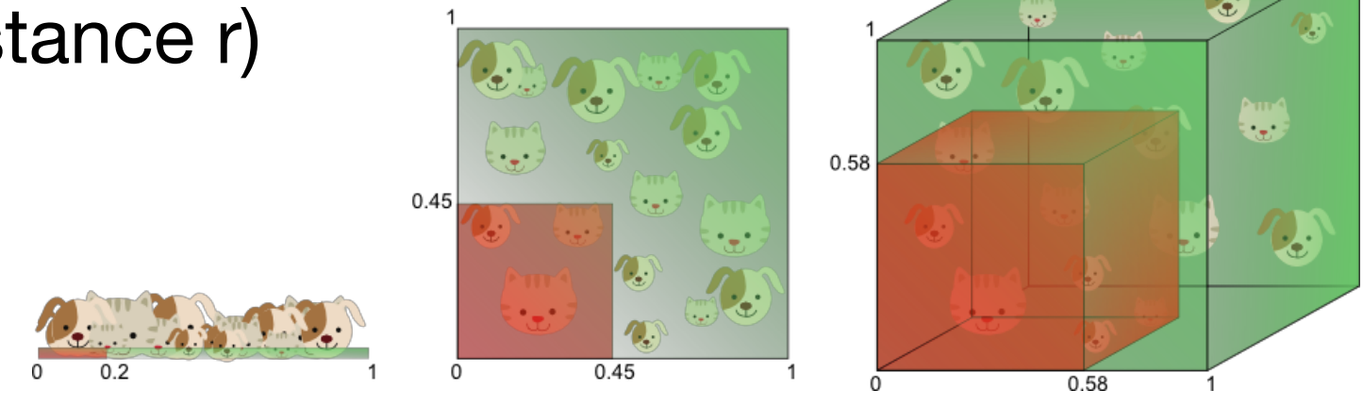


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Curse of Dimensionality

- Too many features
  - Expensive to store
  - Slowing down computation
  - Subject to *Dimensionality curse*
- Sample space gets harder and harder to fill as dimensions grow
- A reason why too many features lead to overfitting as data become **sparse**

- *“If people can see in multi-dimensions we would not need machine learning”*
- More and more data needed to fill the same % of space
- Distance measure degenerates, most of the points are at the surface of a sphere (distance r)



How Many Shades of Gray Can you Distinguish?

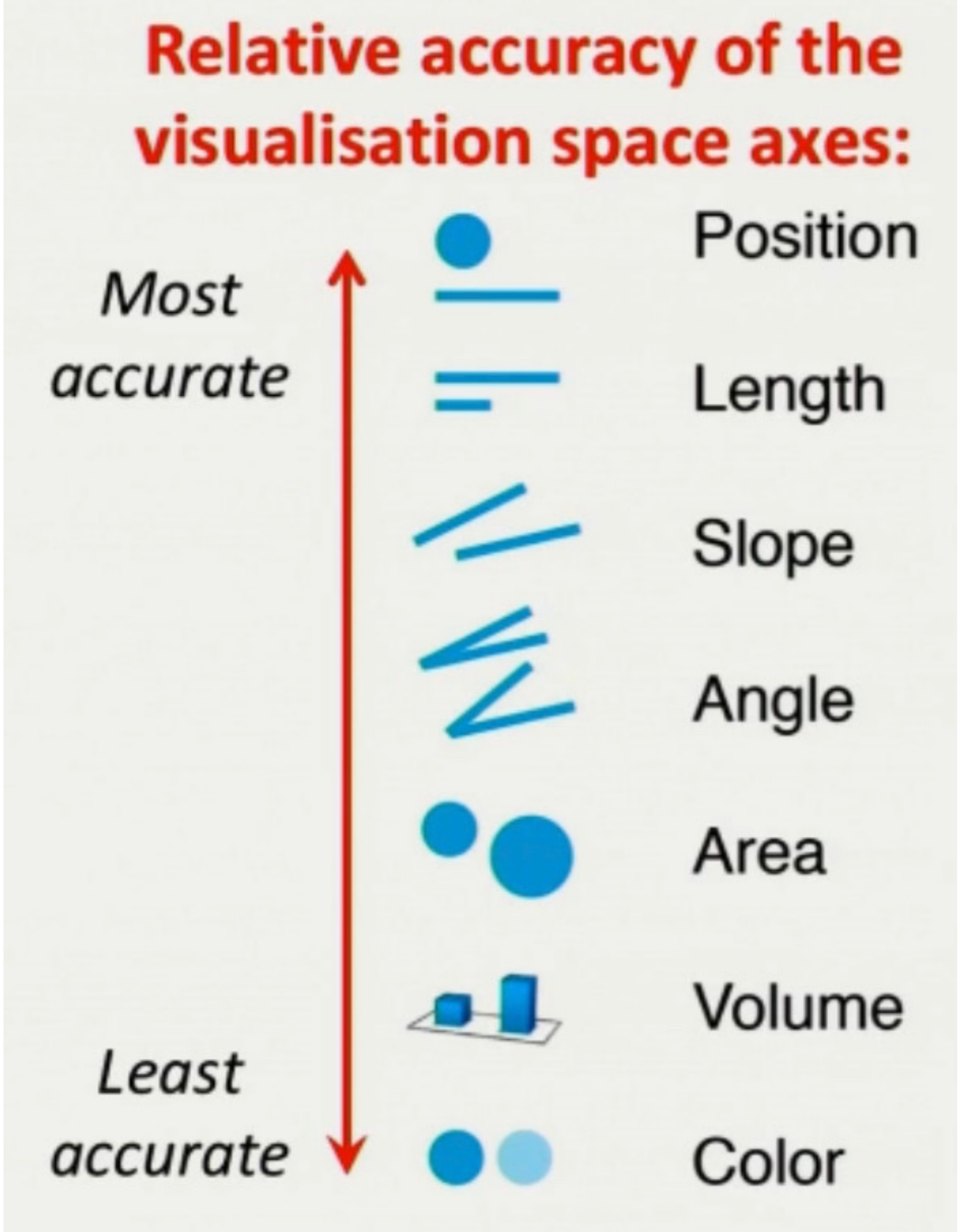
Answer…

Value encodes continuous variables (less well)

How Many Colors?

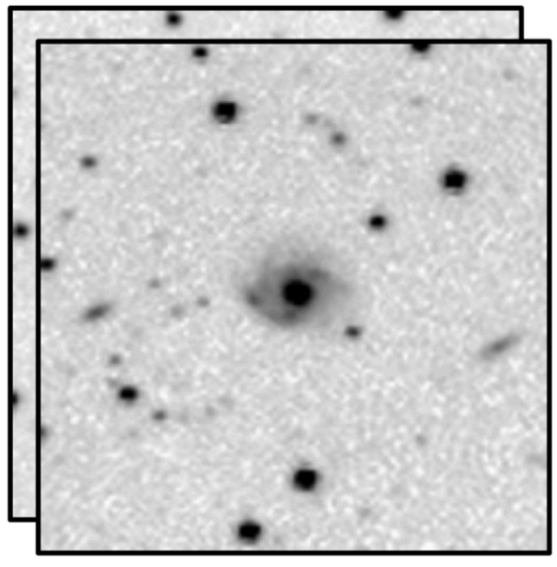
Hue encodes nominal variables

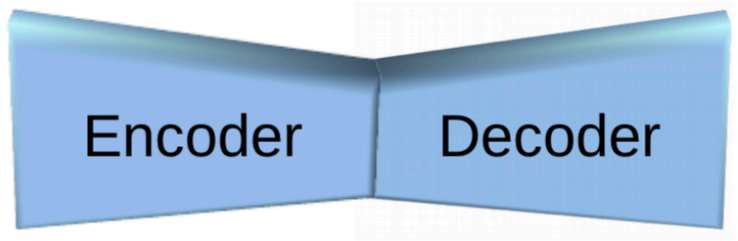
Answer…

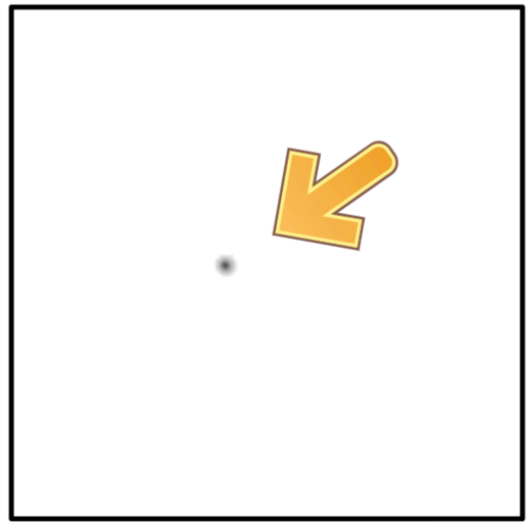

Relative accuracy of the visualisation space axes:
Most accurate
Position
Length
Slope
Angle
Area
Volume
Color
Least accurate

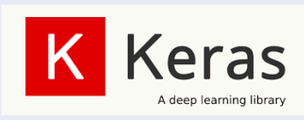
Hands on Sessions

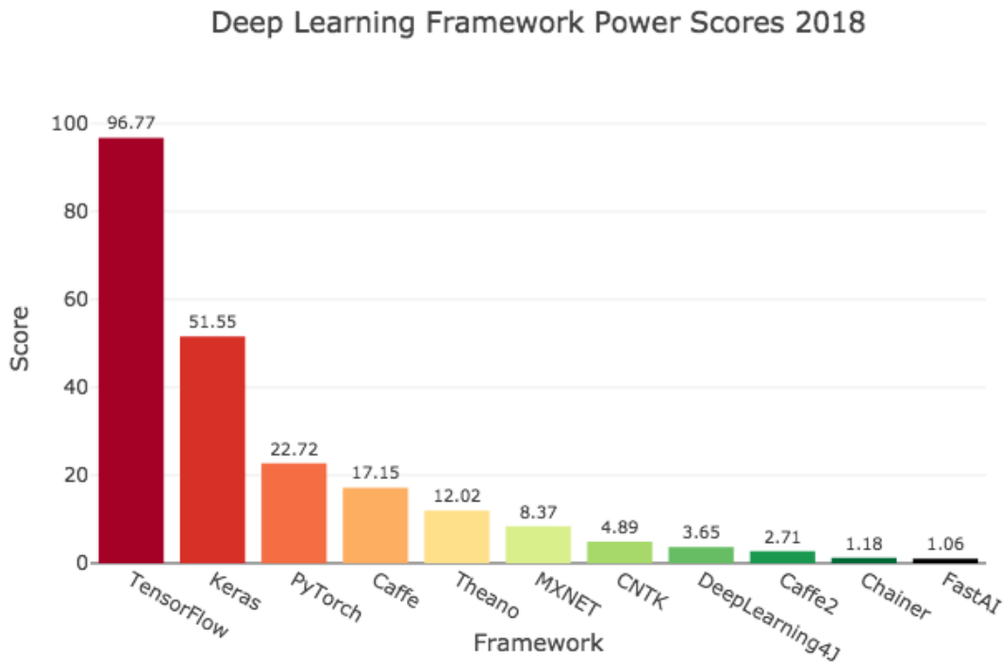
- *Thu 10 Jan 2019:*
    - Galaxy/Star classification (with PCA, feature selection, dim reduction)
    - Regression Photometric redshift with Random Forests
- *Mon 14 Jan 2019:*
    - “voodoo lecture” on Deep Learning with Convolutional Neural Networks on denoising/transient finding



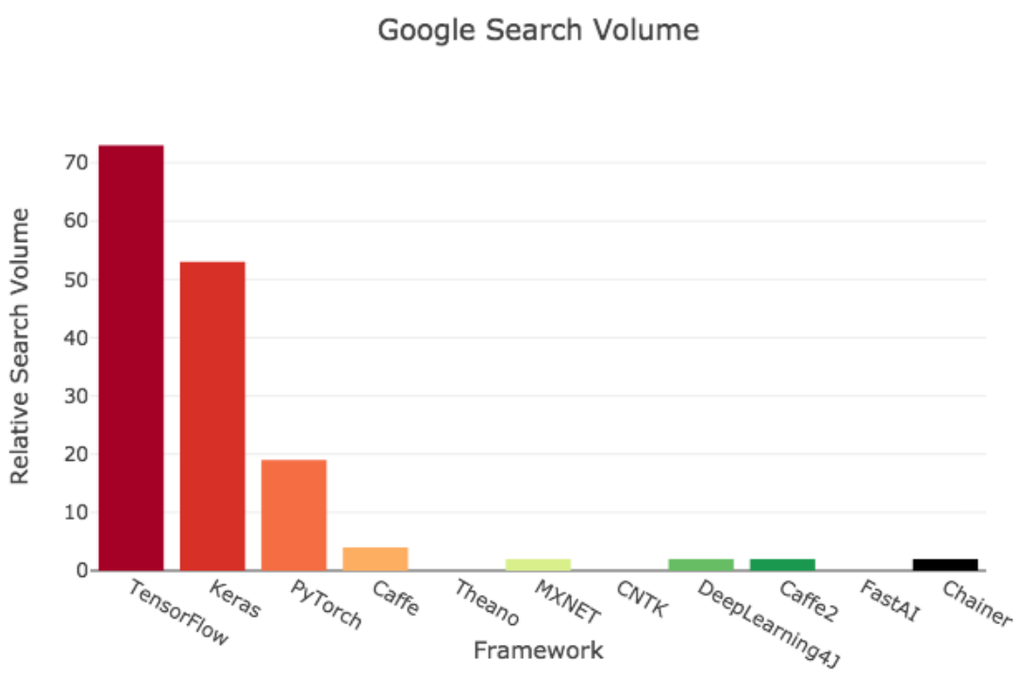


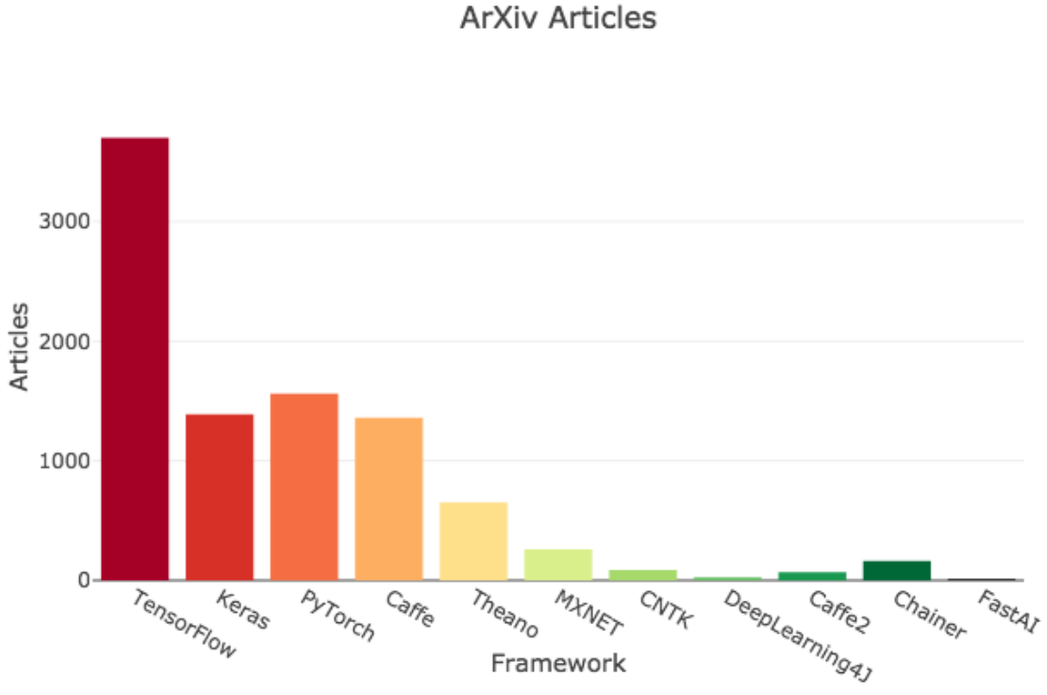


Name	Use	Logo
Pandas	Data Analysis	
Spark	Distributed Computing	
Scikit-learn	Machine Learning Toolbox	
Keras	Deep Learning	
TensorFlow	Deep Learning	
Open-cv	Computer Vision	


Deep Learning Framework Power Scores 2018
Score
Framework
TensorFlow
Keras
PyTorch
Caffe
Theano
MXNET
CNTK
DeepLearning4J
Caffe2
Chainer
FastAI
96.77
51.55
22.72
17.15
12.02
8.37
4.89
3.65
2.71
1.18
1.06


Amazon Books
Books
Framework
TensorFlow
Keras
PyTorch
Caffe
Theano
MXNET
CNTK
DeepLearning4J
Caffe2
FastAI
Chainer


Google Search Volume
Relative Search Volume
Framework
TensorFlow
Keras
PyTorch
Caffe
Theano
MXNET
CNTK
DeepLearning4J
Caffe2
FastAI
Chainer


ArXiv Articles
Articles
Framework
TensorFlow
Keras
PyTorch
Caffe
Theano
MXNET
CNTK
DeepLearning4J
Caffe2
Chainer
FastAI

Software prerequisites

- Python 3
- Anaconda anaconda.org with Python 3.7
- You may create an environment
  `conda create –name=ml python=3.6`
- *scikit-learn*, *pandas*, *scikit-image*, *seaborn*, *tensorflow*, *keras* (anaconda may suggest to downgrade some of the libraries to Python 3.6, but it's ok, the environment works as an independent container)
  `conda install xxx` or `conda install -c conda-forge xxx`
- `conda update conda ; conda update anaconda ; conda update xxx`
- Cheat sheets…

Intel…

- conda create -n intelpython --override-channels --channel intel python=3.5 intelpython scipy pydaal anaconda scikit-learn

Weak points

- Sometimes a black box, not straightforward to interpret and to verify by other than a try-and-fail method

- Hard to handle data with errors, error-bars and incomplete datasets

- Neural Networks requires **a lot** of data and CPU/graphical card accelerators . Simply does not work otherwise.

ML seminar

- Aka astro-ph
- an algorithm overview, usage
- DIAS ML group:
  - Helping each other across the schools (interdisciplinary)
  - publications together?
- The ML seminar to happen every 2 weeks. It will kick off in February after we go through the lectures!

References/Further Reading

- scikit-learn.org documentation, great gallery, tons of examples

- Andrew Ng Standford course (I passed it and it's really good, the code is in Matlab)
  `https://www.coursera.org/learn/machine-learning`

- Yaser Abu-Mostafa Caltech course — #1 course online (in my humble opinion, I passed it, it requires no coding, very good but simple math explanation)
  `https://work.caltech.edu/telecourse.html`