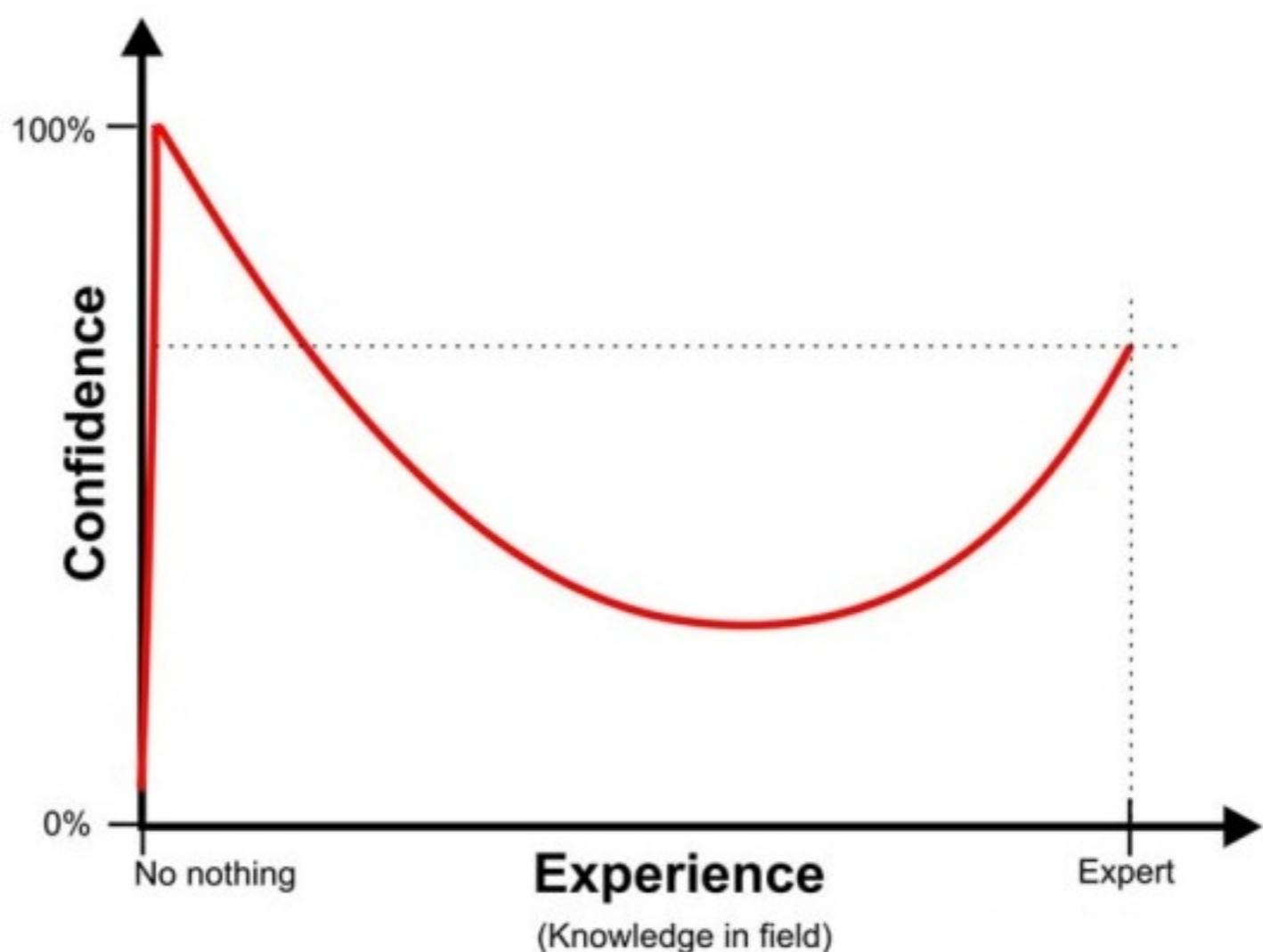


Work to Machine!

DIAS ML course by Martin Topinka

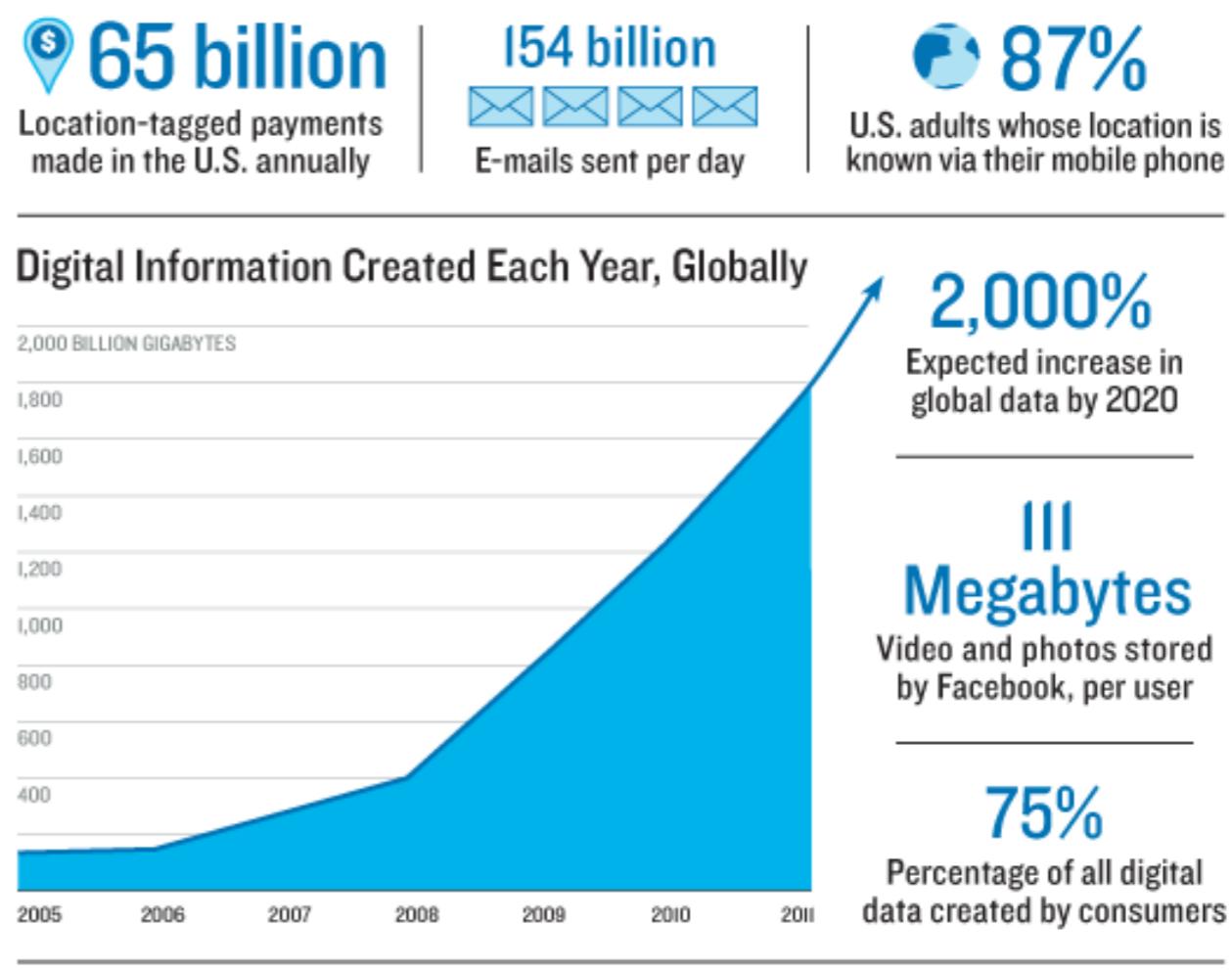


Introduce yourself and your goals



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

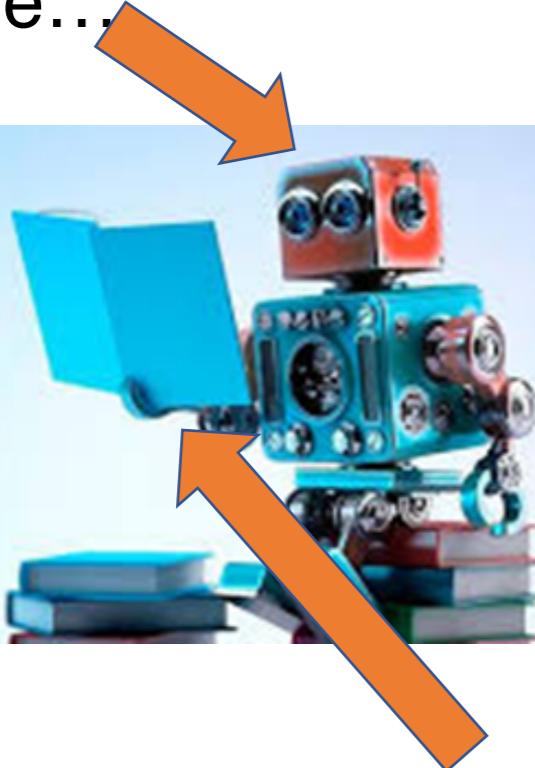
30 terabytes of data nightly

Final volume of raw image data = 60 PB

Final image collection (DR11) = 0.5 EB

Final catalog size (DR11) = 15 PB

Machine...



... Learning?!?

'AI IS THE NEW ELECTRICITY'



"Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform in the next several years."

Andrew Ng

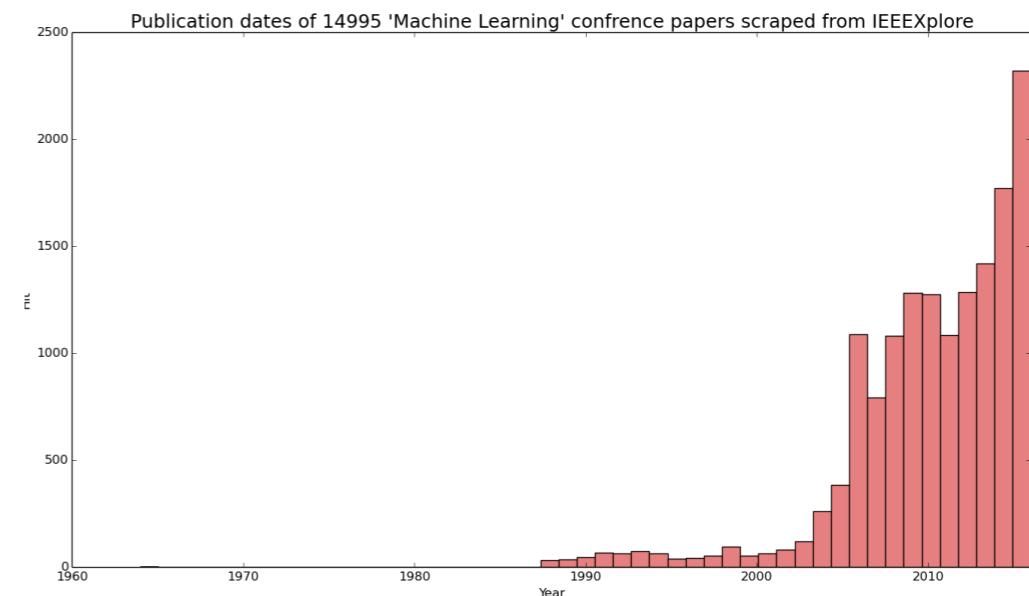
Former chief scientist at Baidu, Co-founder at Coursera

CBINSIGHTS

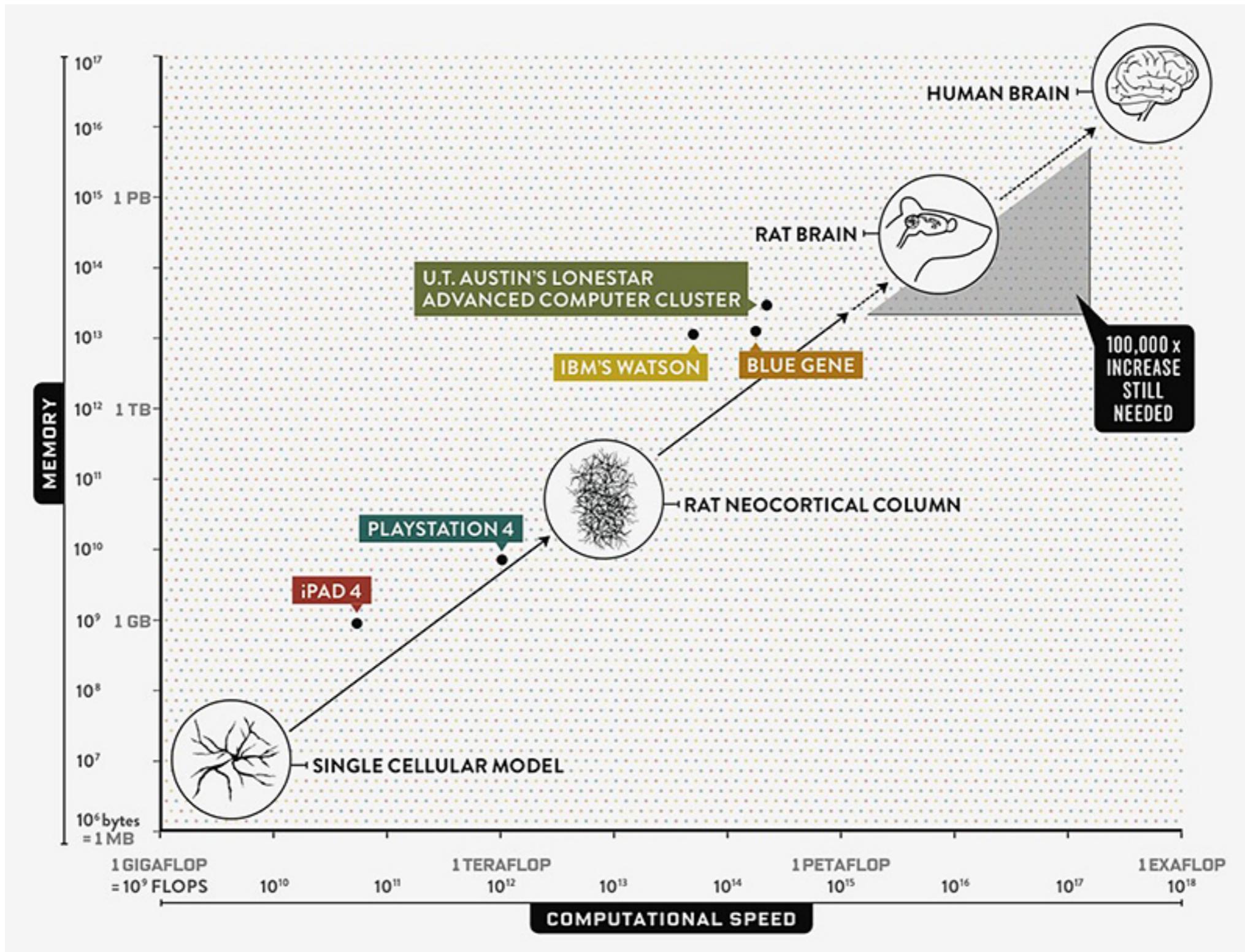
source: <https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity>

www.cbinsights.com

7



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



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

Machine Learning in the Sky

- Machine Learning owes a lot to astronomy: least-square regression for orbital parameter estimation (Legendre-Laplace-Gauss)

Data Big Bang in Astronomy too:

10^9 object photometric catalogs from USNO, 2MASS, SDSS...

10^{6-8} spectroscopic catalogs from SDSS, LAMOST...

10^{6-7} multi-wavelength source catalogs from WISE, eROSITA...

10^9 object x 10^2 epochs surveys like **LSST**, DES, PTF, CRTS, SNF, VVV, Pan-STARRS, Stripe 82

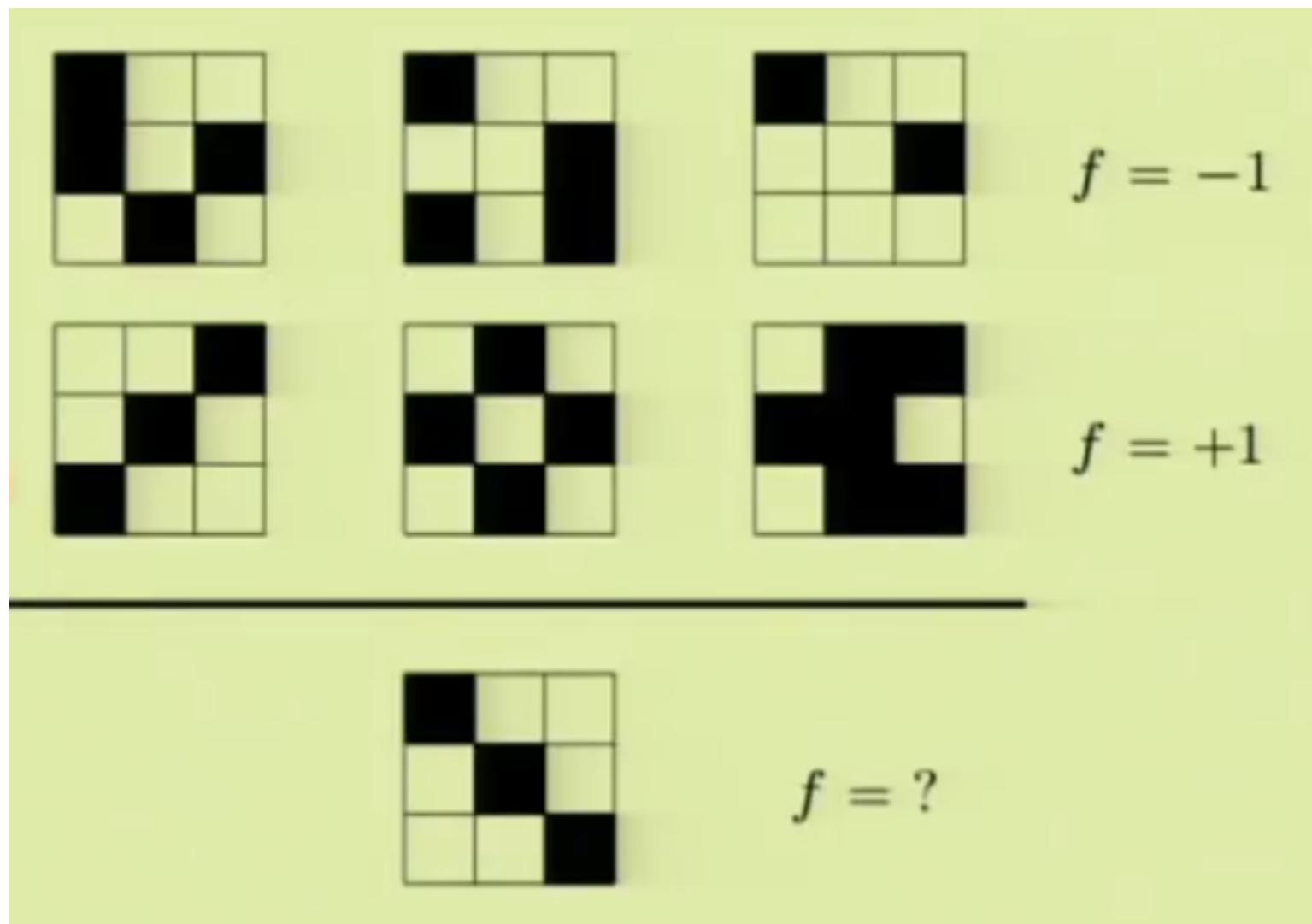
Spectral-image datacubes from VLA, ALMA, IUFs, ... JWST ...

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

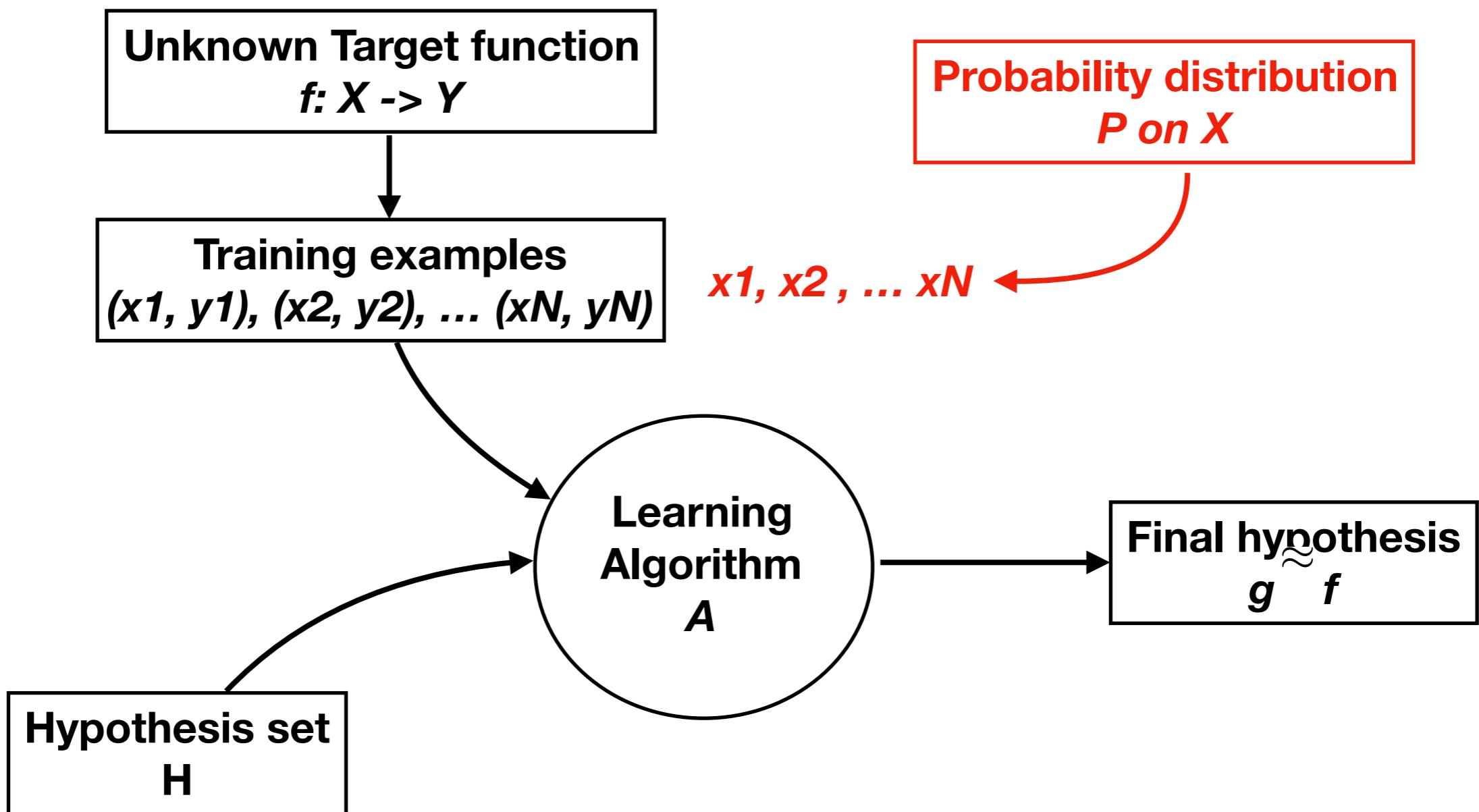
Essentials of Learning

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

Target function...

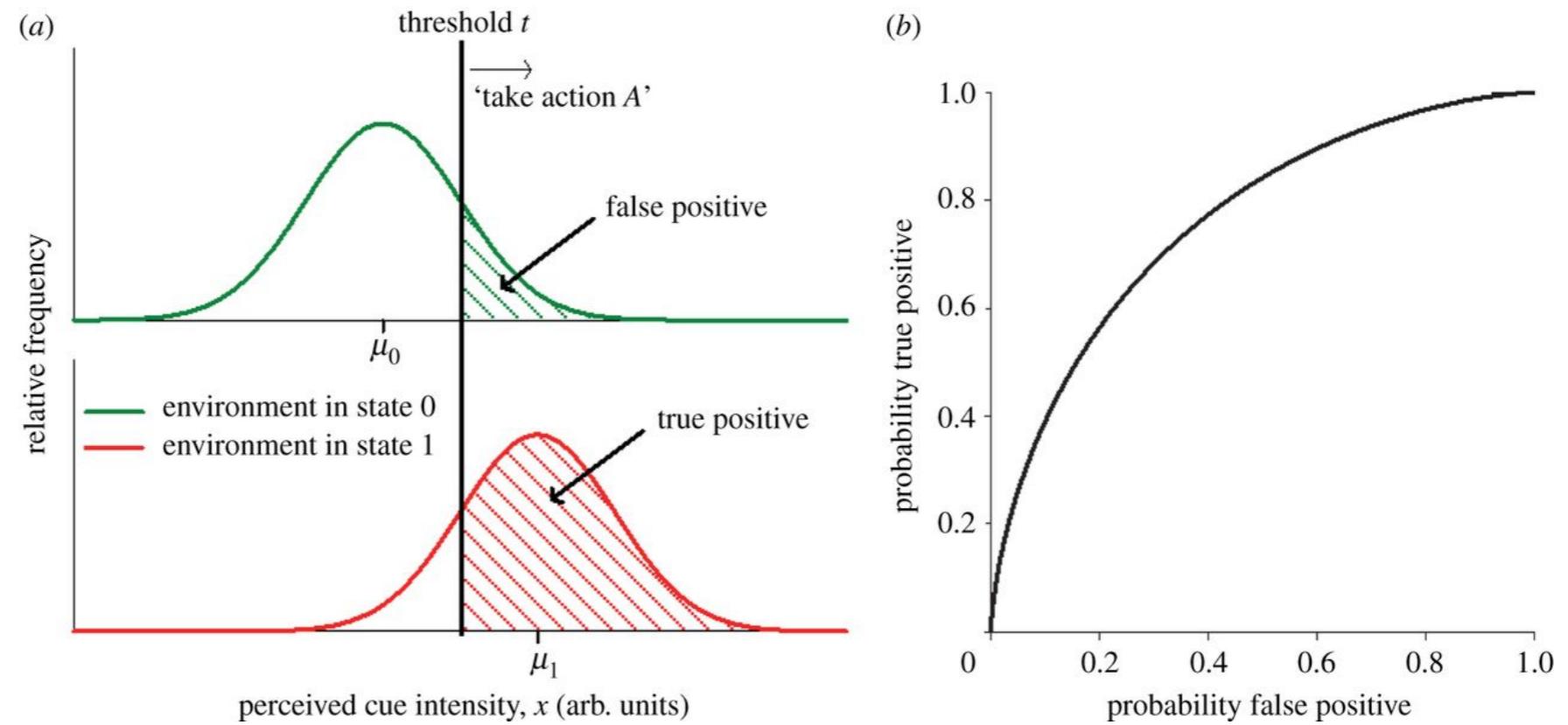


Learning Diagram

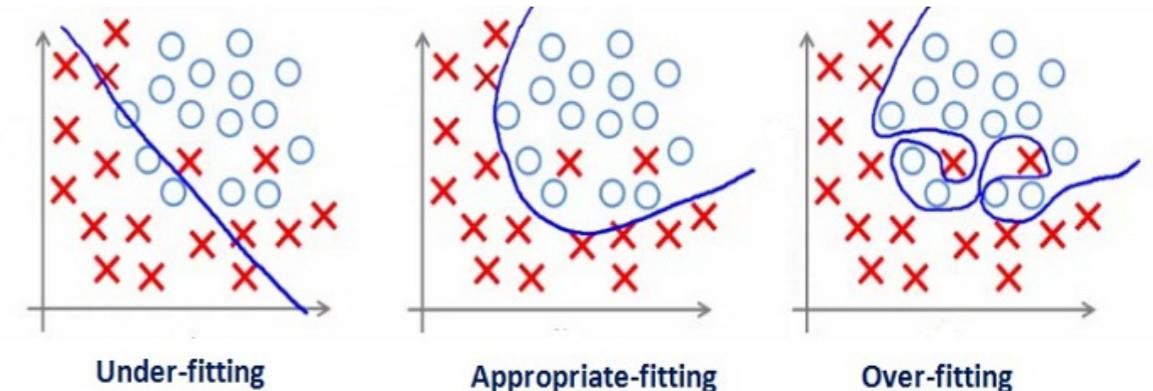


Loss function

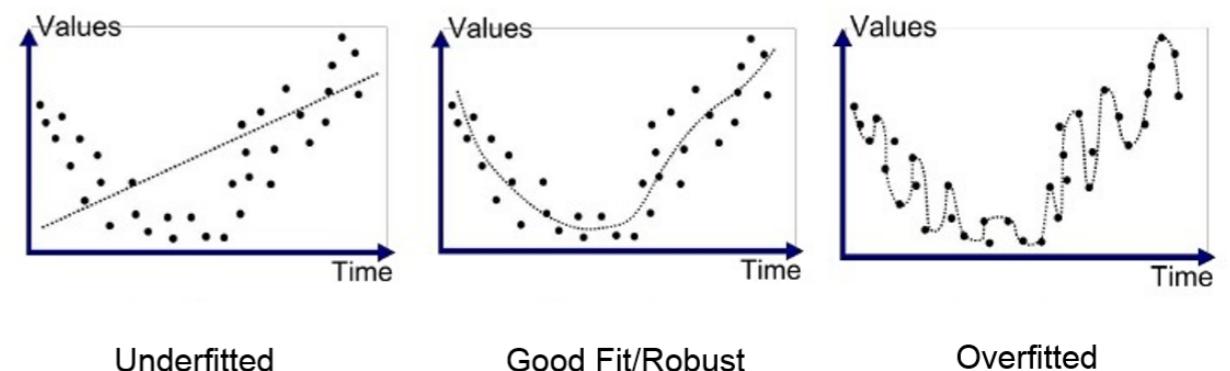
- imbalance classes



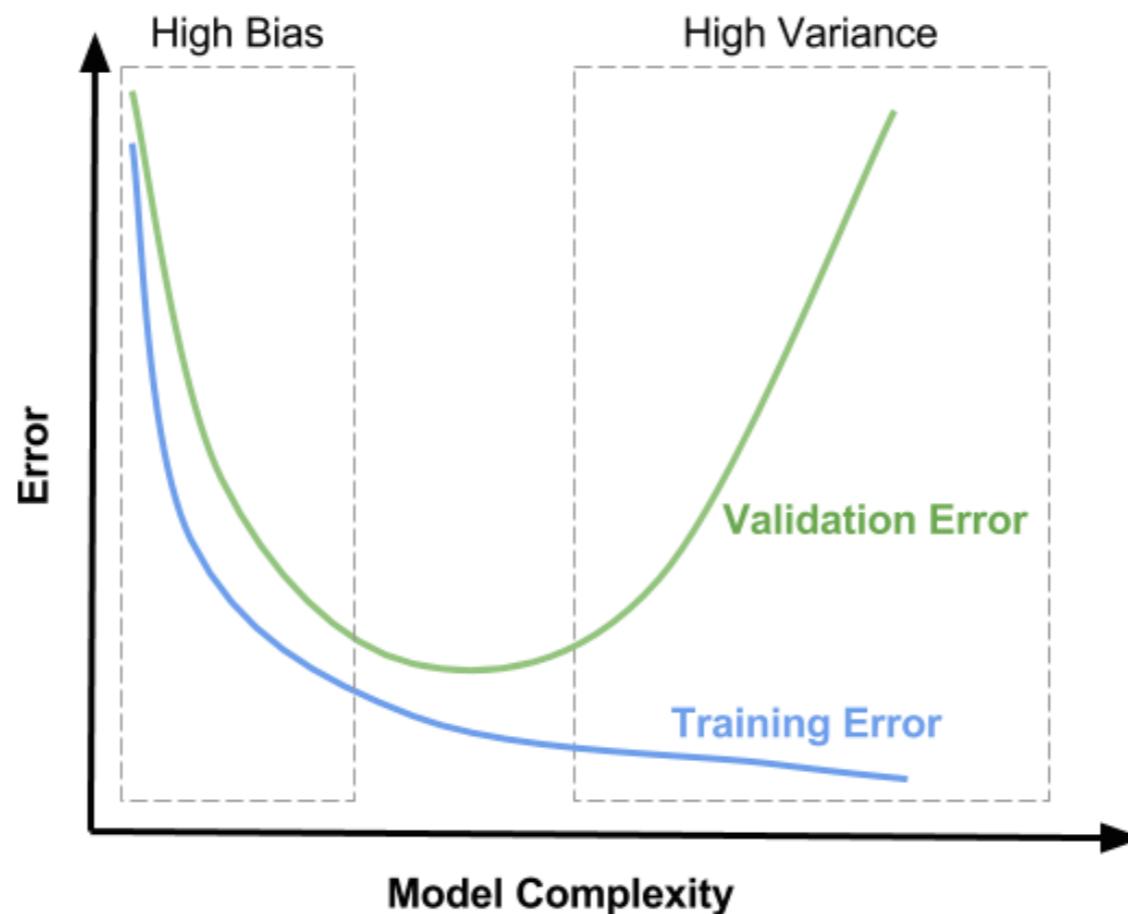
Bias - Variance Trade-off



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



(cross)-Validation



| | Remedies |
|---------------|---|
| High Bias | <ul style="list-style-type: none"> • Train longer • Increase model complexity <ul style="list-style-type: none"> • more features • more parameters, • richer architecture |
| High Variance | <ul style="list-style-type: none"> • Get more data • Decrease model complexity <ul style="list-style-type: none"> • 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



Dog?

50% dog, 50% cat?



Machines are lazy and love shortcuts

Correlations != Reasoning



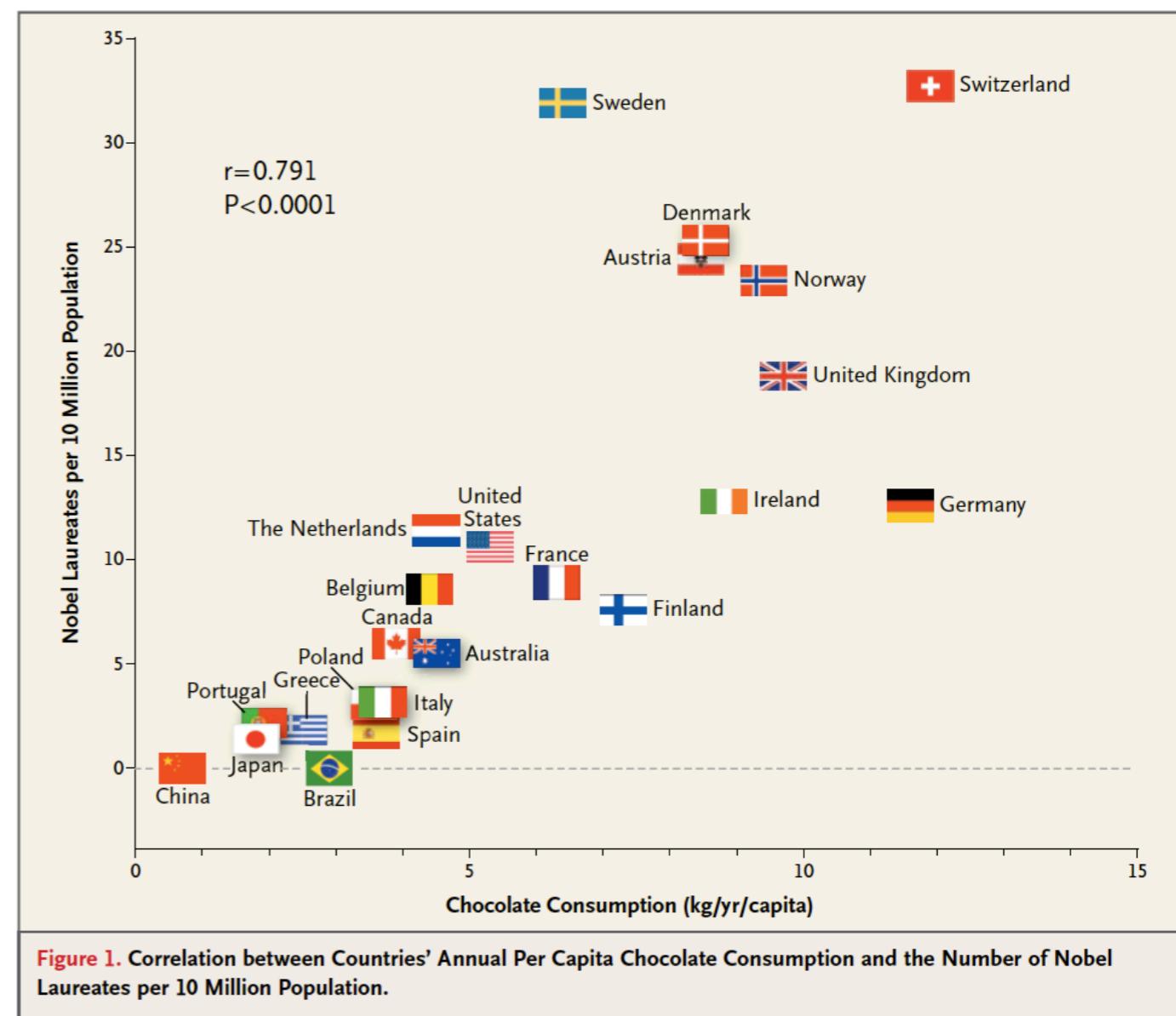
Justin King

@Justinkingnews

Headline: "Women who own horses live longer"

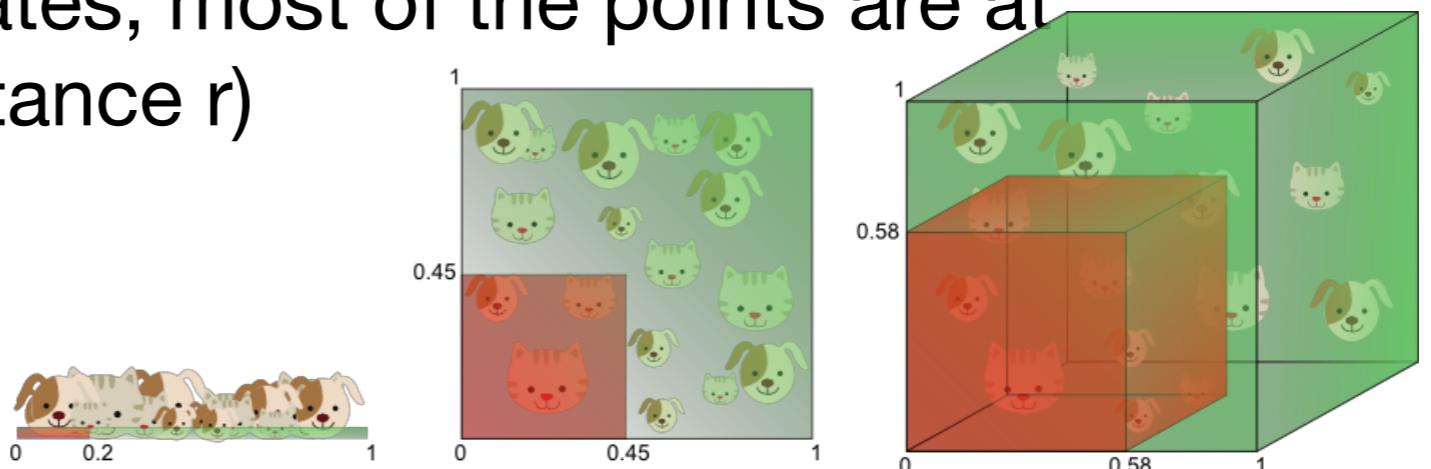
Implied correlation: horses make you live longer.

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



Curse of Dimensionality

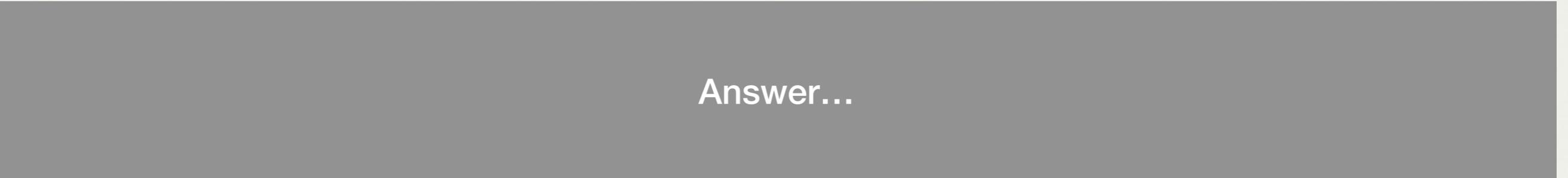
- Too many features
 - Expensive to store
 - Slowing down computation
 - Subject to *Dimensionality curse*
 - Sample space gets harder and harder to fill as dimensions grow
 - A reason why too many features lead to overfitting as data become **sparse**
-
- “*If people can see in multi-dimensions we would 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...

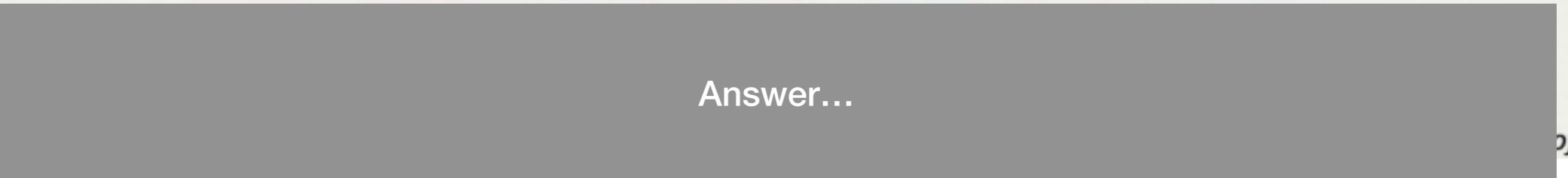


Value encodes continuous variables (less well)

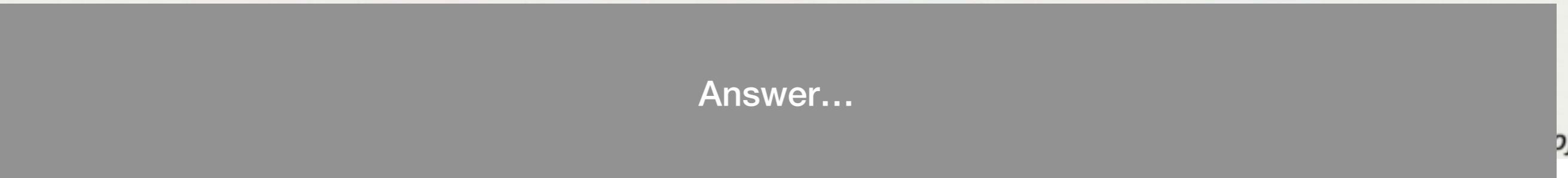
How Many Colors?



Hue encodes nominal variables

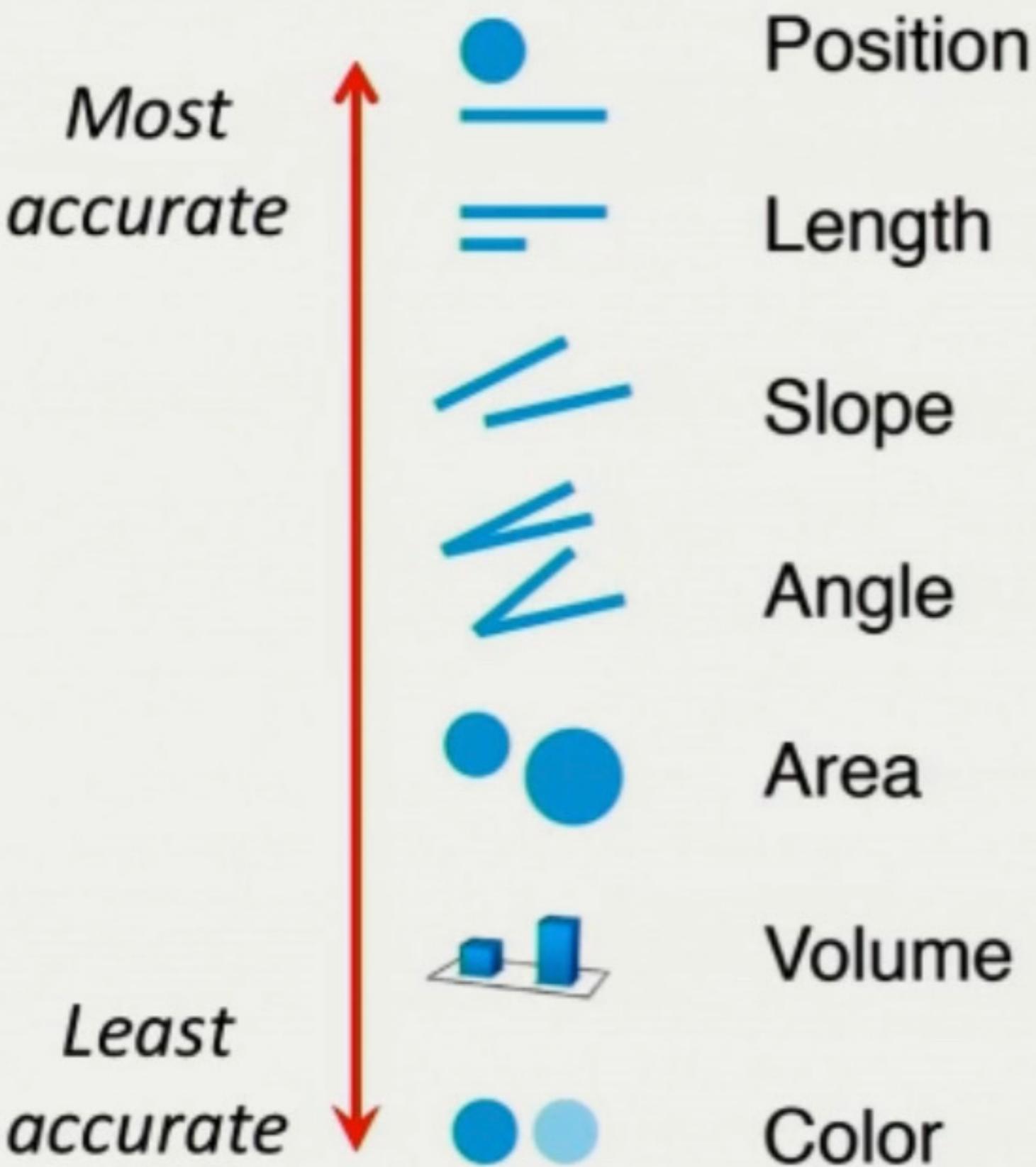


Answer...



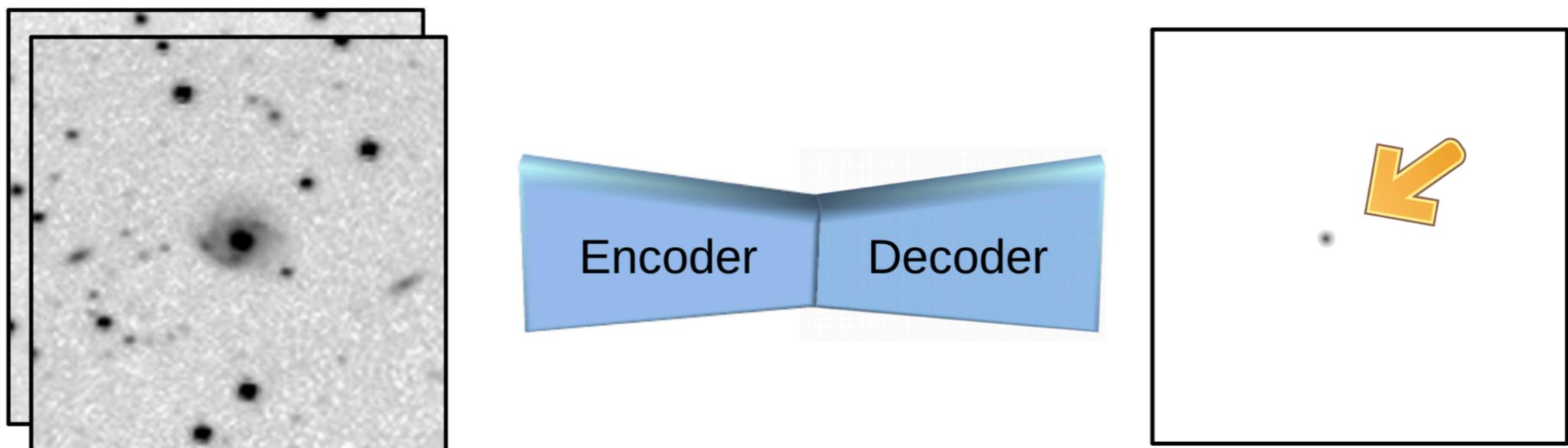
(off)

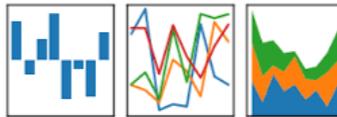
Relative accuracy of the visualisation space axes:

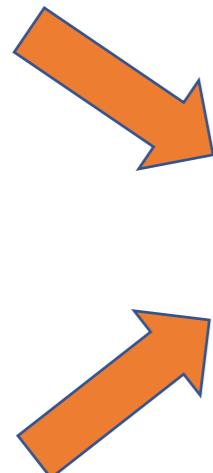


Hands on Sessions

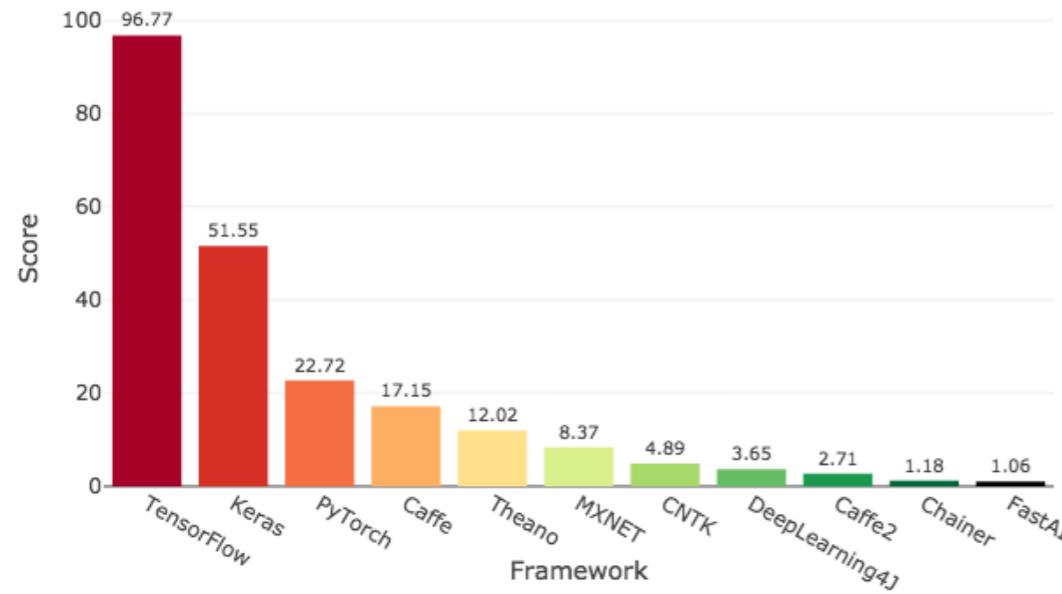
- Galaxy/Star classification (with PCA)
- Regression Photometric redshift with Random Forests
- “voodoo lecture” on Deep Learning with Convolutional Neural Networks on denoising/transient finding



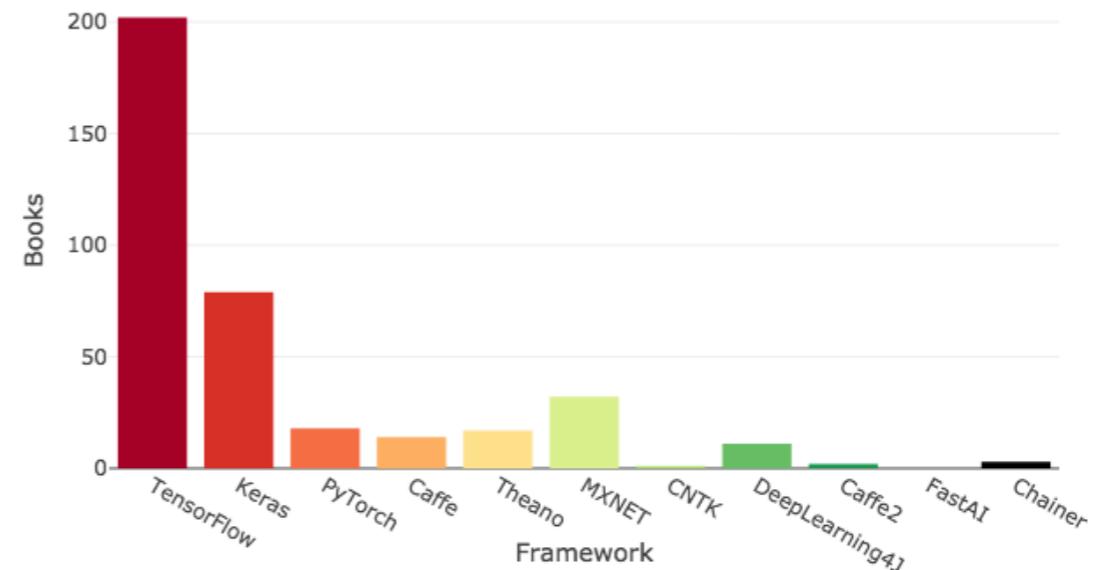
| Name | Use | Logo |
|--------------|---|---|
| Pandas | Data Analysis $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$ | pandas  |
| Spark | Distributed Computing |  |
| Scikit-learn | Machine Learning Toolbox |  |
| Keras | Deep Learning |  |
| TensorFlow | Deep Learning |  |
| Open-cv | Computer Vision |  |



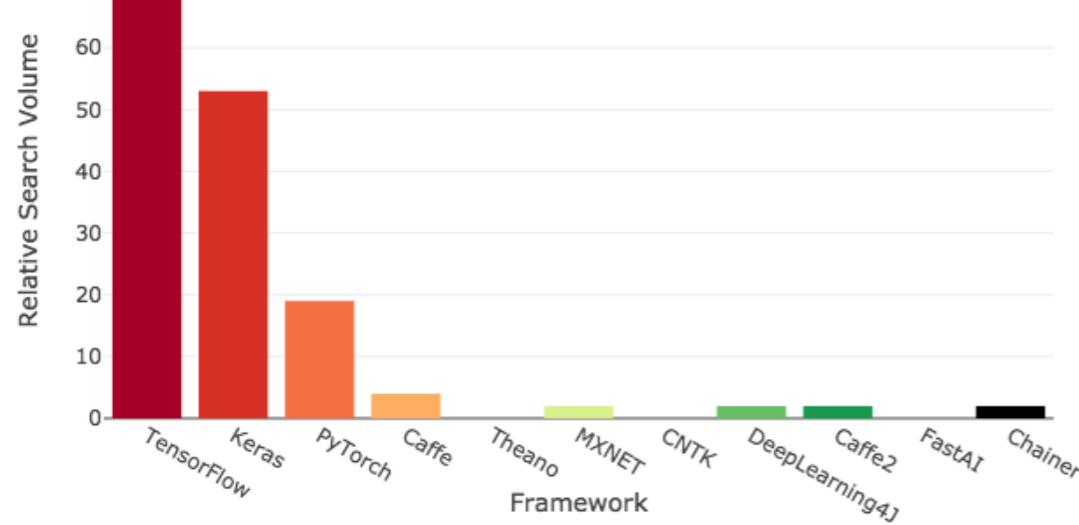
Deep Learning Framework Power Scores 2018



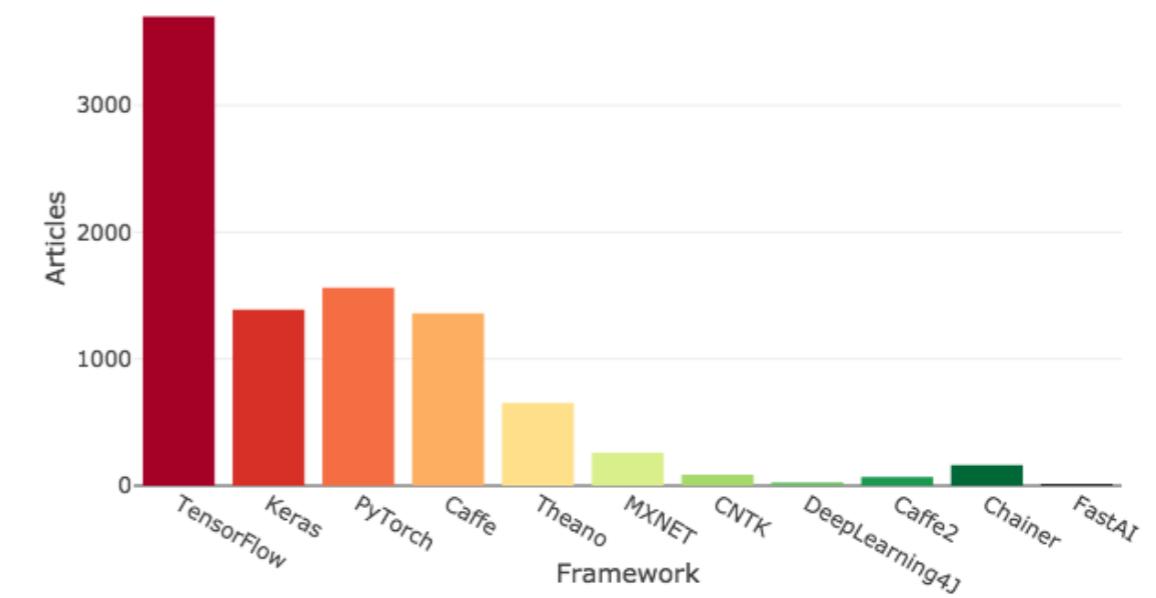
Amazon Books



Google Search Volume



ArXiv Articles



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* (may induce a downgrade to Python 3.6, but it's ok)
`conda install xxx`
- Cheat sheets...

ML seminar

- Aka astro-ph
- an algorithm
- publications?
- 1x every 2 weeks or month?