INTRODUCTION TO MACHINE LEARNING IN ASTRONOMY

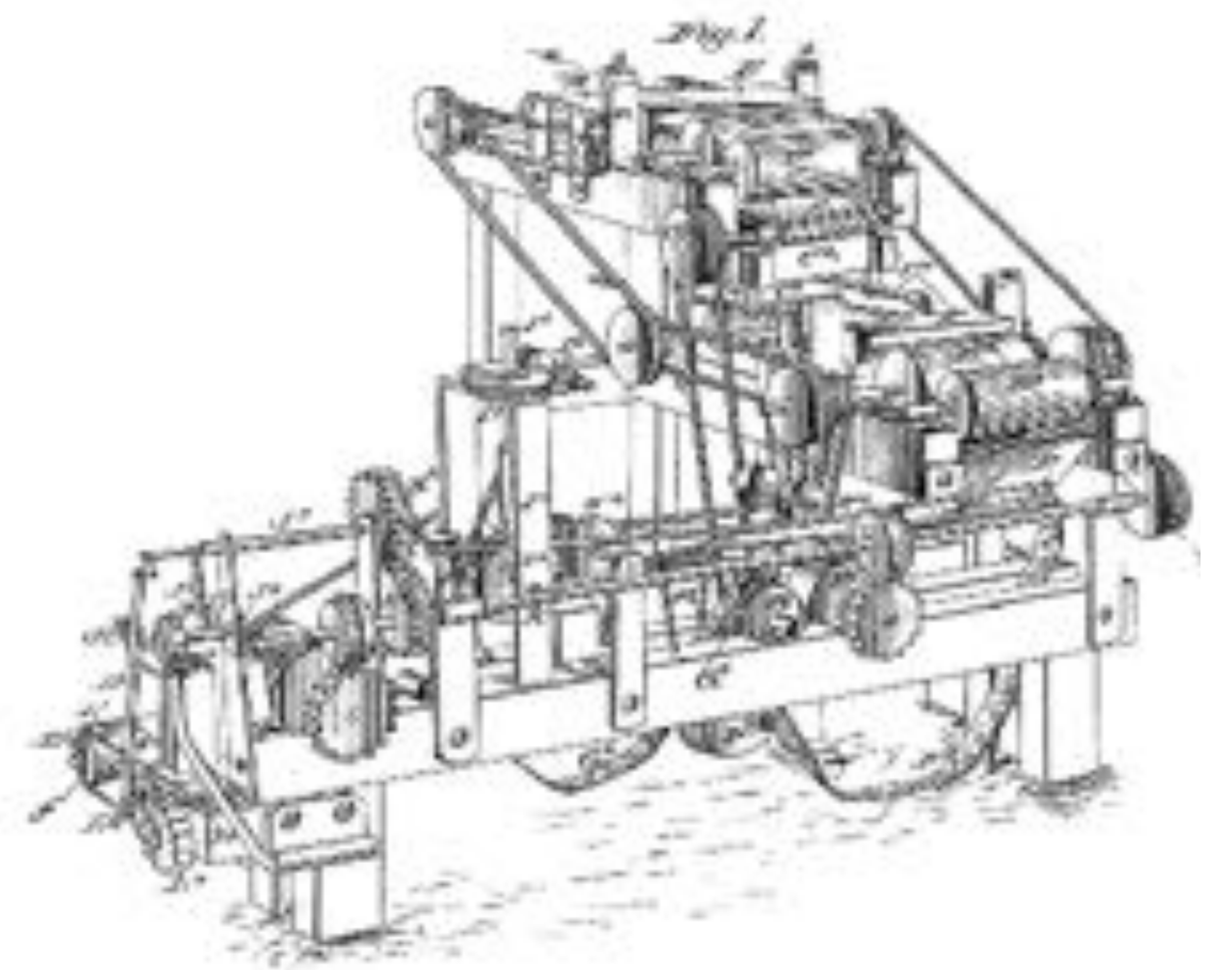
David Kirkby, UC Irvine

LSSTC Data Science Fellows Program
Caltech, January 2017

ABOUT ME

- Particle Physics
- Cosmology
  - Redshift Surveys (BAO)
    - SDSS (BOSS, eBOSS)
    - DESI
  - Imaging Surveys (WL)
    - LSST
- Data Science, Statistics

MACHINE LEARNING
?



MACHINE LEARNING

GPU +
cuda/opencl

[google TPU]

MACHINE LEARNING

e.g.

Suggest a missing word in a sentence.

Identify a specific person in a photo.

Drive a car automatically.

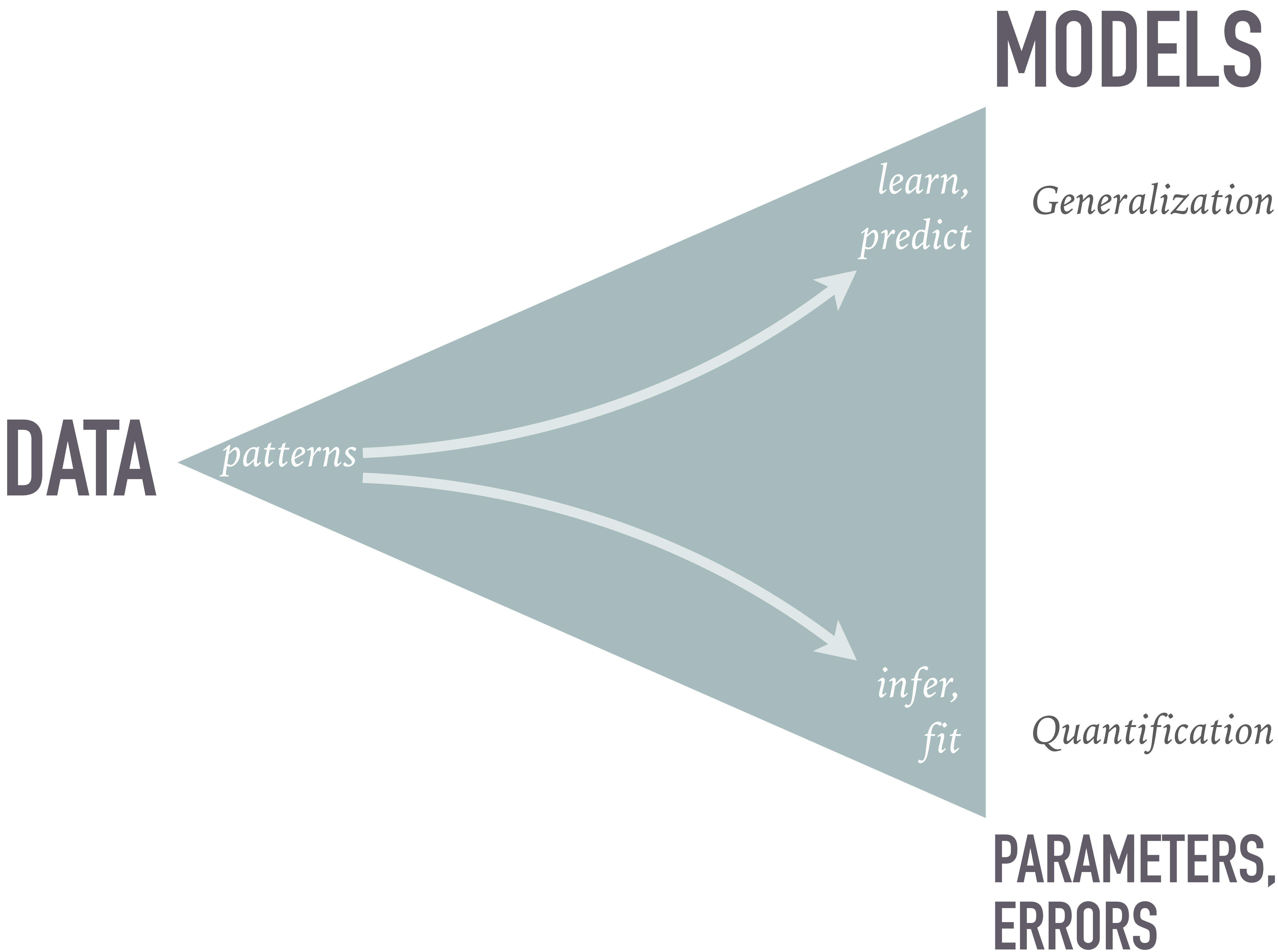
Predict the sky brightness for tomorrow night's observing.

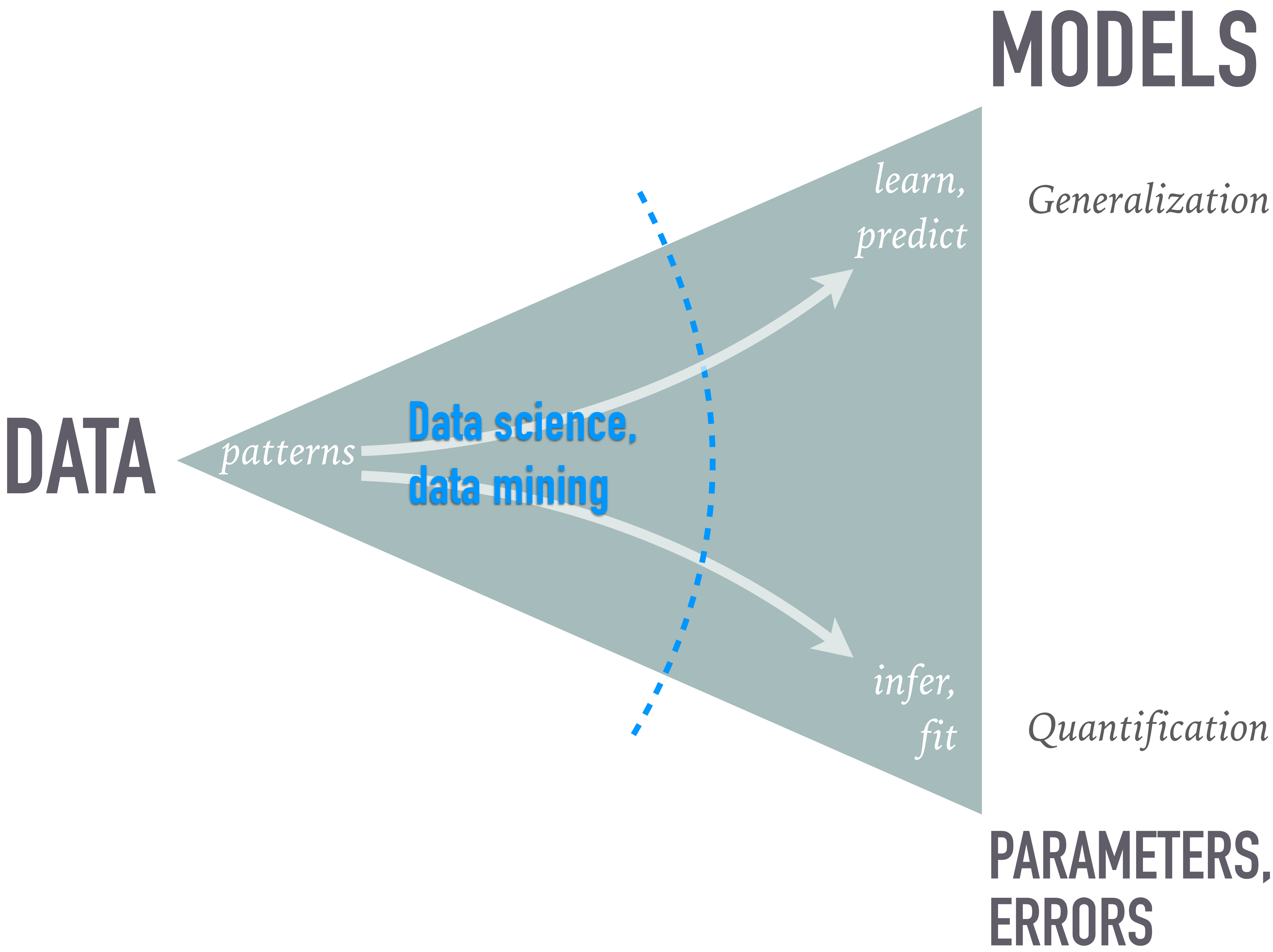
Estimate a galaxy's redshift from its LSST magnitudes.

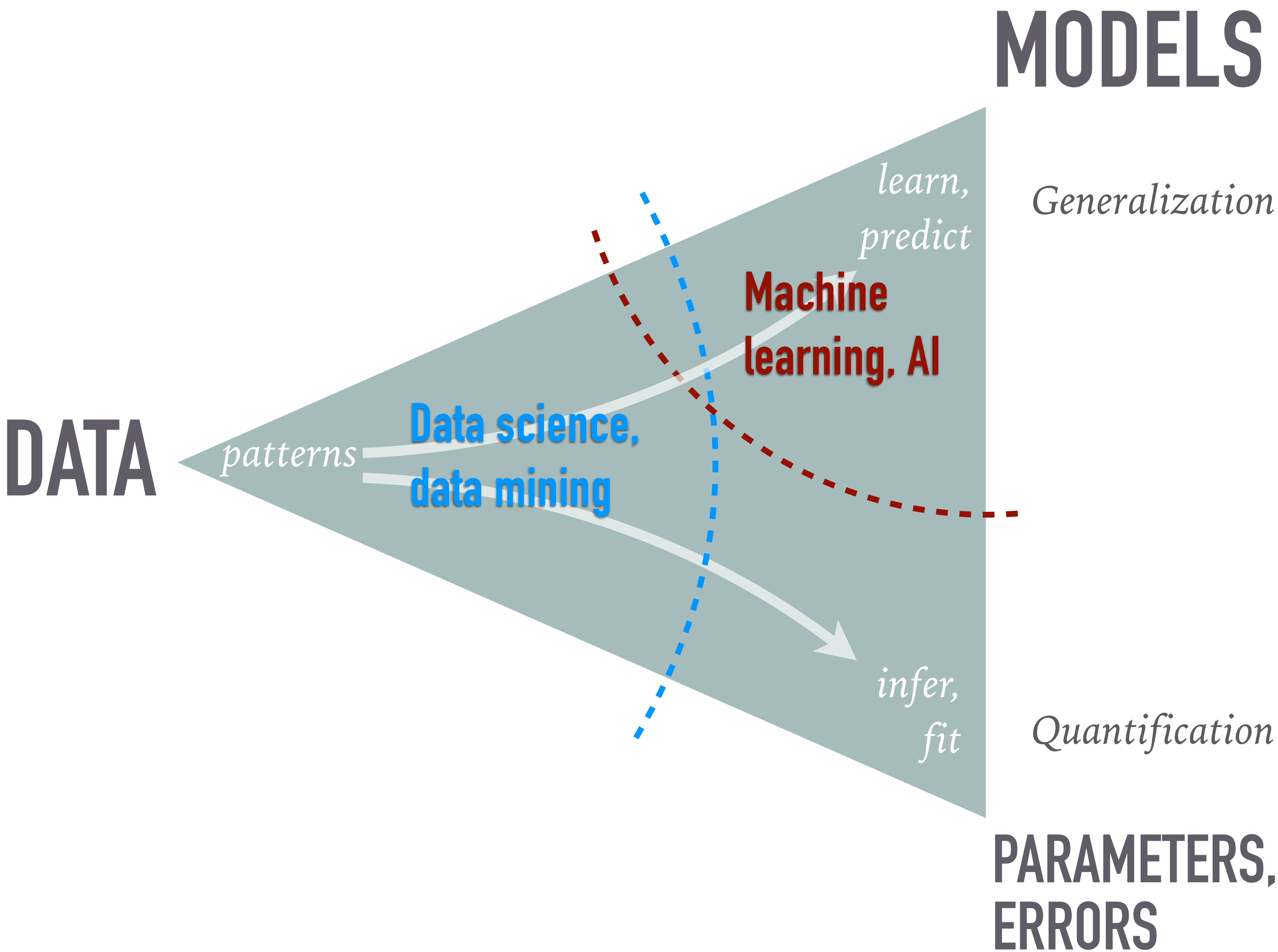
Describe the relationship between supernovae distance and redshift.

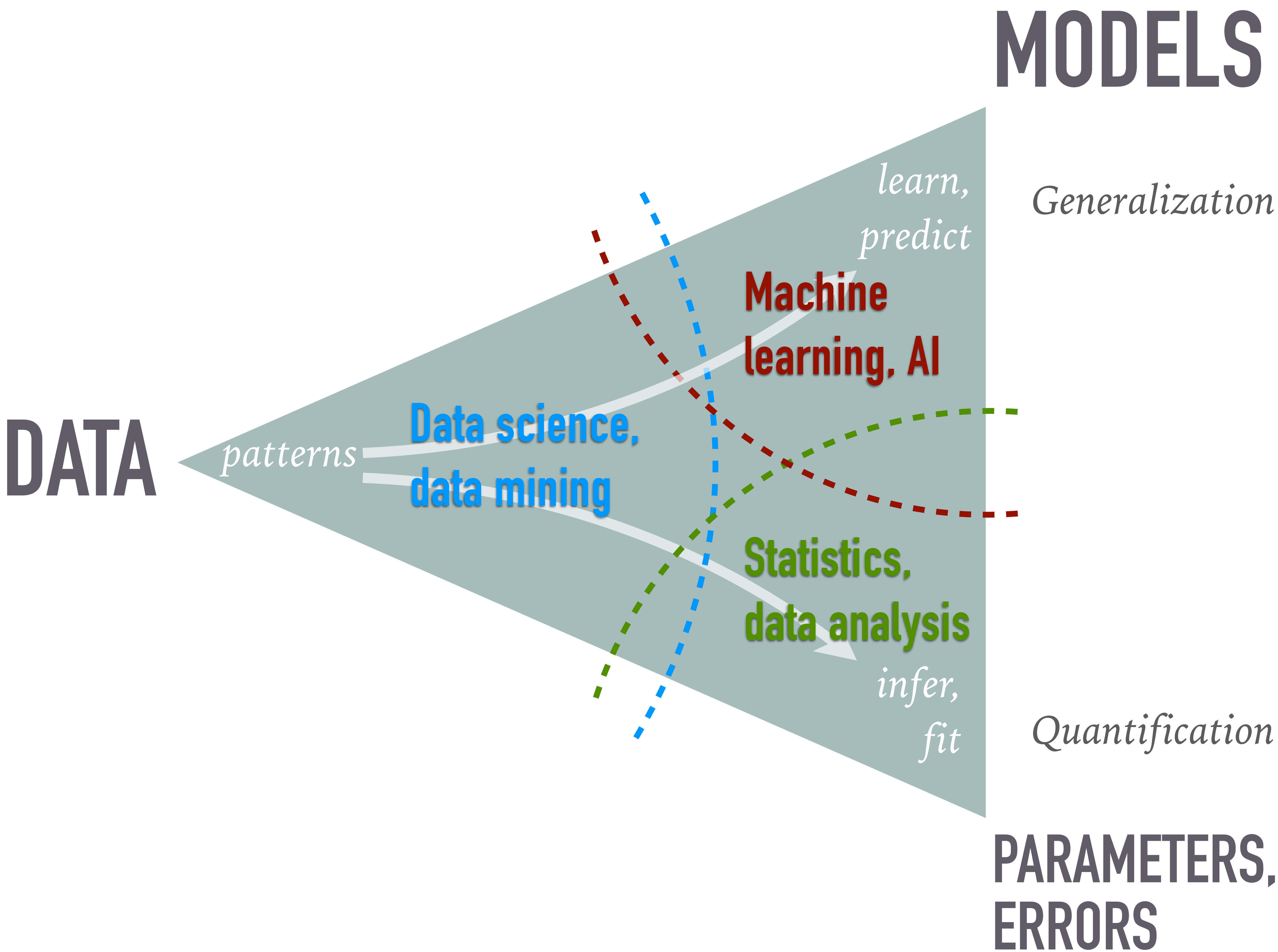
ACTIVITY: DEFINE YOUR TERMS

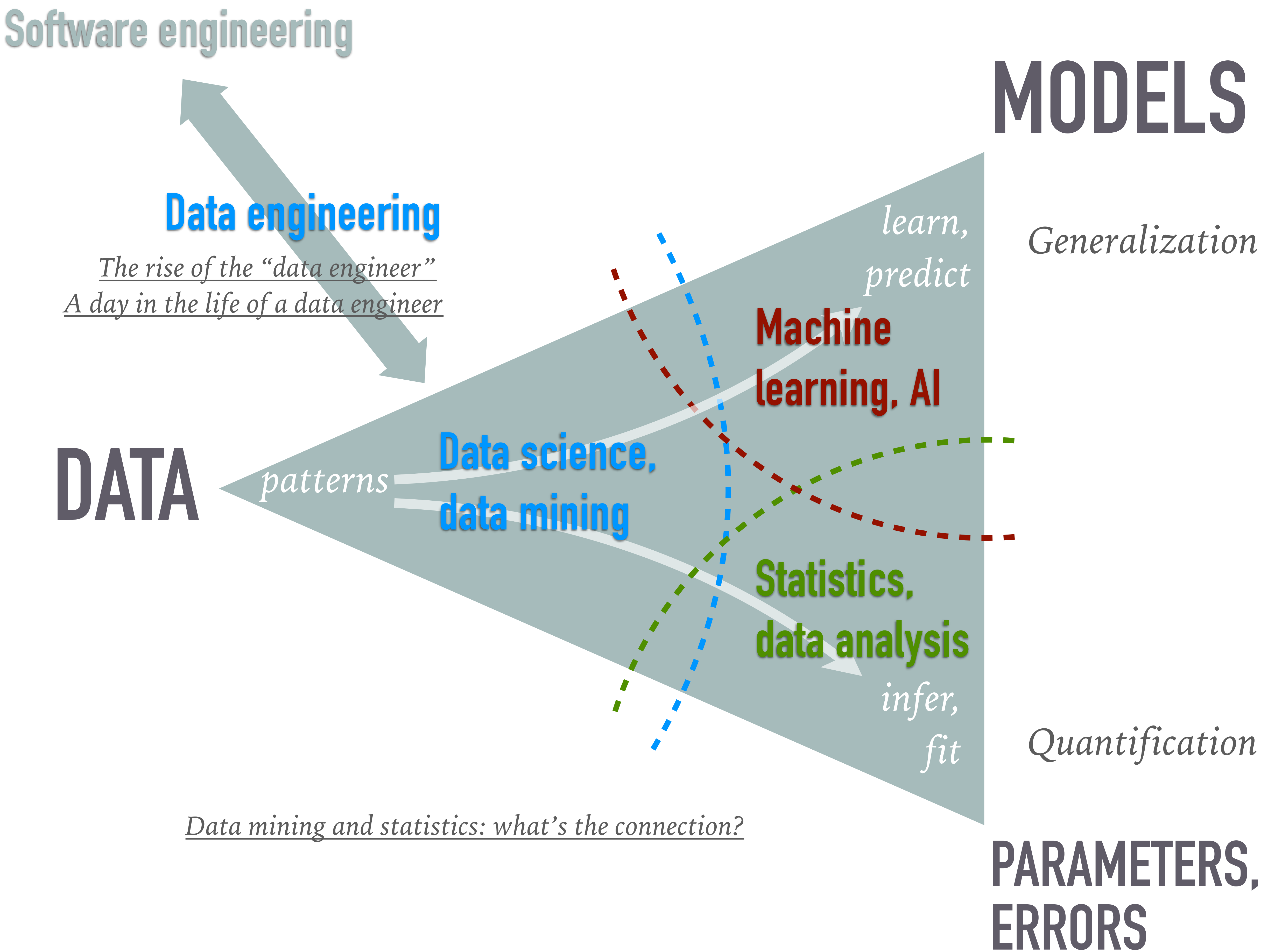
- What is the relationship between machine learning and statistics?
- What is the difference between a "data scientist" and a "data engineer"?
- Rank these tasks in order of importance for your research:
  - estimating model parameters
  - finding patterns in data
  - predicting new data


MODELS
learn,
predict
Generalization
DATA
patterns
infer,
fit
Quantification
PARAMETERS,
ERRORS


MODELS
learn,
predict
Generalization
DATA
patterns
Data science,
data mining
infer,
fit
Quantification
PARAMETERS,
ERRORS


MODELS
learn,
predict
Generalization
Machine
learning, AI
DATA
patterns
Data science,
data mining
infer,
fit
Quantification
PARAMETERS,
ERRORS


MODELS
Generalization
learn,
predict
Machine
learning, AI
DATA
patterns
Data science,
data mining
Statistics,
data analysis
infer,
fit
Quantification
PARAMETERS,
ERRORS


Software engineering
MODELS
Data engineering
The rise of the "data engineer"
A day in the life of a data engineer
learn,
predict
Generalization
Machine
learning, AI
DATA
patterns
Data science,
data mining
Statistics,
data analysis
infer,
fit
Quantification
Data mining and statistics: what's the connection?
PARAMETERS,
ERRORS

Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August
python	*R*
conference talk	*journal article*

http://statweb.stanford.edu/~tibs/stat315a/glossary.pdf

What's your projection into this state space?

```
['computer engineer',
 'computer scientist',
 'data engineer',
 'data scientist',
 'lsst engineer',
 'lsst scientist']
```

```
['%s %s' % (a,b) for a in ('computer', 'data', 'lsst')
                 for b in ('engineer', 'scientist')]
```

WHAT IS SPECIAL ABOUT MACHINE LEARNING IN ASTRONOMY?

- We are data producers, not data consumers:
  - Experiment / survey design.
  - Optimization of statistical errors.
  - Control of systematic errors.
- Our models are usually traceable to an underlying physical theory:
  - Models constrained by theory and observations.
  - Parameter values linked to universal constants of nature.
- A parameter error estimate is just as important as its value:
  - Prefer methods that handle input data errors and provide error estimates.

MACHINE LEARNING

=

DATA + MODELS

ROADMAP

- ✓ Introduction
- ➤ Data in astronomy
- ➤ Models in astronomy
- ➤ Statistical context of ML
- ➤ Types of learning, problems, solutions
- ➤ The bleeding edge of ML

ACTIVITY: REASONING ABOUT DATA AND MODELS

- Is a CCD raw image <u>data</u>?
- Is a galaxy catalog <u>data</u>?
- Is a histogram a <u>model</u>?
- Do you need a <u>model</u> to calculate an average?
- Does your research focus more on <u>data</u> or <u>models</u>?

DATA + MODELS

"Data" are a finite set of measurements:

- ➤ *e.g., spreadsheet, FITS table, …*
- ➤ *numeric / categorical / mixed?*
- ➤ *ordered? (special role of time, stochastic processes & MCMC)*
- ➤ *independent? identically distributed?*
- ➤ *measurement errors (implicit / explicit)*
- ➤ *binned / un-binned?*
- ➤ *similarity measure?*

columns ~ "features"

rows ~ samples, observations

x	y	z	a	b	c

ASTRONOMICAL DATA + MODELS

Astronomical data:

- measure physical processes (units!)
- low level: images, spectra, time series.
- high level: catalogs.
- results from careful experimental / survey design.
- has quantifiable statistical and systematic uncertainties.

DATA + MODELS

"Models" specify the probability of different outcomes:

- *explicit: probability density function.*
- *implicit: algorithm to generate random outcomes ("forward" model).*
- *observables vs parameters (vs hyper-parameters vs …).*
- *integrability: required to calculate normalized probabilities.*
- *differentiability: required to find most probable (uphill) direction.*
- *hierarchical construction.*
- *variance - bias tradeoffs.*
- *regularization: bias towards "sensible" interpretations.*

DATA + ASTRONOMICAL MODELS

Astronomical models:

- ➤ are usually traceable to an underlying physical theory.
- ➤ parameters often related to universal constants of nature.
- ➤ often have known distribution of measurement errors.
- ➤ need to account for instrumental effects (calibration).
- ➤ prefer models that:
  - ➤ handle input data errors (often via weights),
  - ➤ provide error estimates.

ACTIVITY: REASONING ABOUT ML PROBLEMS

- Rank these ML problems in order of "difficulty".
- List a few possibly relevant data features.

1. *Suggest a missing word in a sentence.*
2. *Identify a specific person in a photo.*
3. *Drive a car automatically.*
4. *Predict the sky brightness for tomorrow night's observing.*
5. *Estimate a galaxy's redshift from its LSST magnitudes.*
6. *Describe the relationship between supernovae distance and redshift.*

A model is trained on, fit to, or inferred from data.

DATA + MODELS

Data is a realization of some model

(but probably not the one you are using).

HOW TO BUILD A MODEL?

Exploratory data analysis & visualization:

- single feature: histogram of PDF, CDF.
- two features: scatter plot.
- multiple features:
  - pair-wise corner plot.
  - local embedding (tSNE, …).

DATA + MODELS

THE LANGUAGE OF ML IS STATISTICS (NOT PYTHON!)

- Key ideas:
  - Bayes theorem.
  - Occam’s razor.

- Given some data:
  - Infer probabilities assuming a model.
  - Compare alternative models.

ACTIVITY: BAYESIAN REASONING

English or not english?

- Write down your best guess: YES/NO.

ACTIVITY: BAYESIAN REASONING

English or not english?

- Write down your best guess: YES/NO.
- Think about the probability that the answer is YES. Write down a number.

ACTIVITY: BAYESIAN REASONING

English or not english?

- Write down your best guess: YES/NO.
- Think about the probability that the answer is YES. Write down a number.
- Discuss your reasoning with your neighbor, then update your answer.

ACTIVITY: BAYESIAN REASONING

- What do we know?
  - DATA = “wearing an ENGLAND t-shirt”
- What question are we asking?
  - P(english | DATA) ?
- What do we need to assume?
  - MODEL =

```
if english:
  prob[DATA] = 1/3
else:
  prob[DATA] = 1/5
```

  - PRIOR = “20% of astronomers are english”

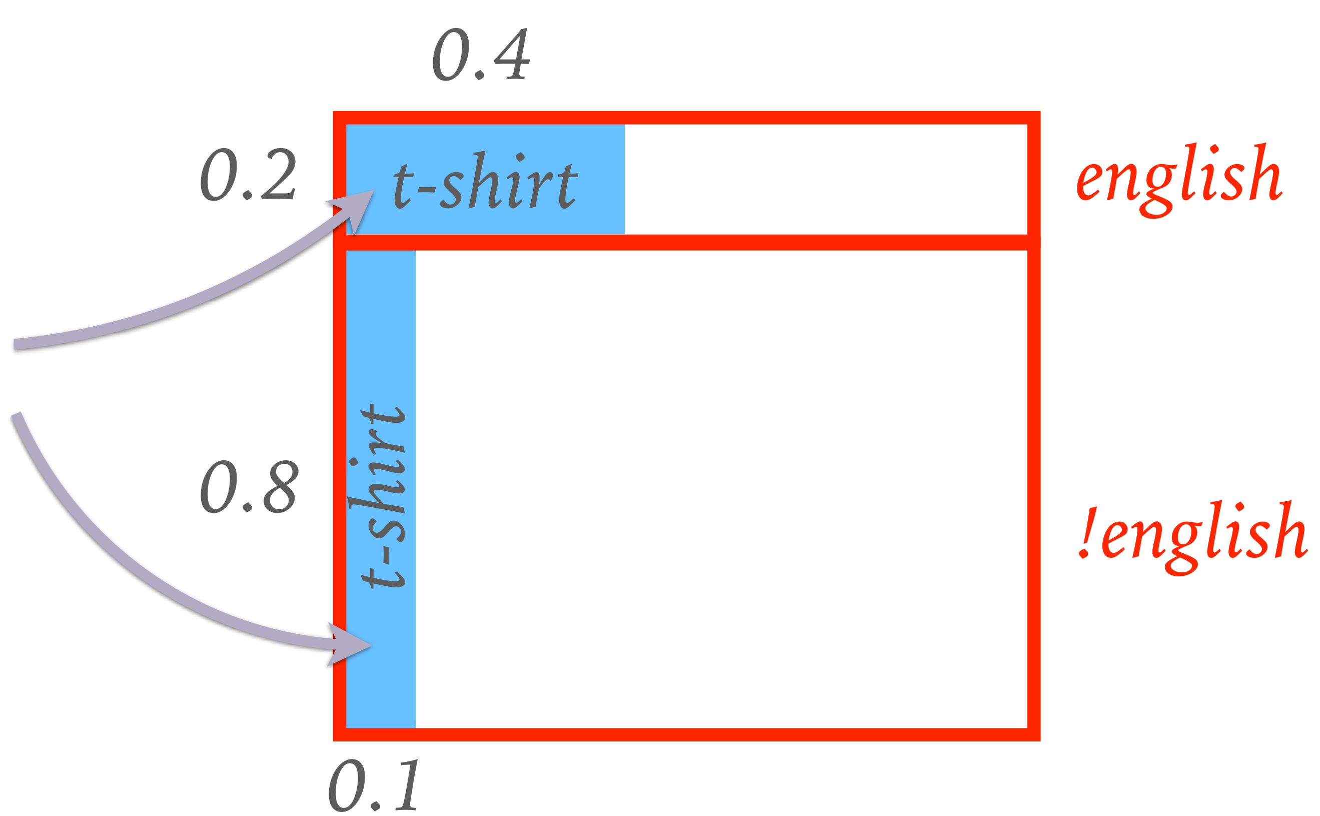
ACTIVITY: BAYESIAN REASONING

- PRIOR = "20% of astronomers are english"
- MODEL =

```
if english:
  prob[DATA] = 0.4
else:
  prob[DATA] = 0.1
```

- DATA = "wearing an ENGLAND tshirt"

P(english | DATA) = ratio of these areas



ACTIVITY: BAYESIAN REASONING

0.4 0.2

$$P(\text{english} \mid \text{tshirt}) = \frac{P(\text{tshirt} \mid \text{english})\ P(\text{english})}{P(\text{tshirt})}$$

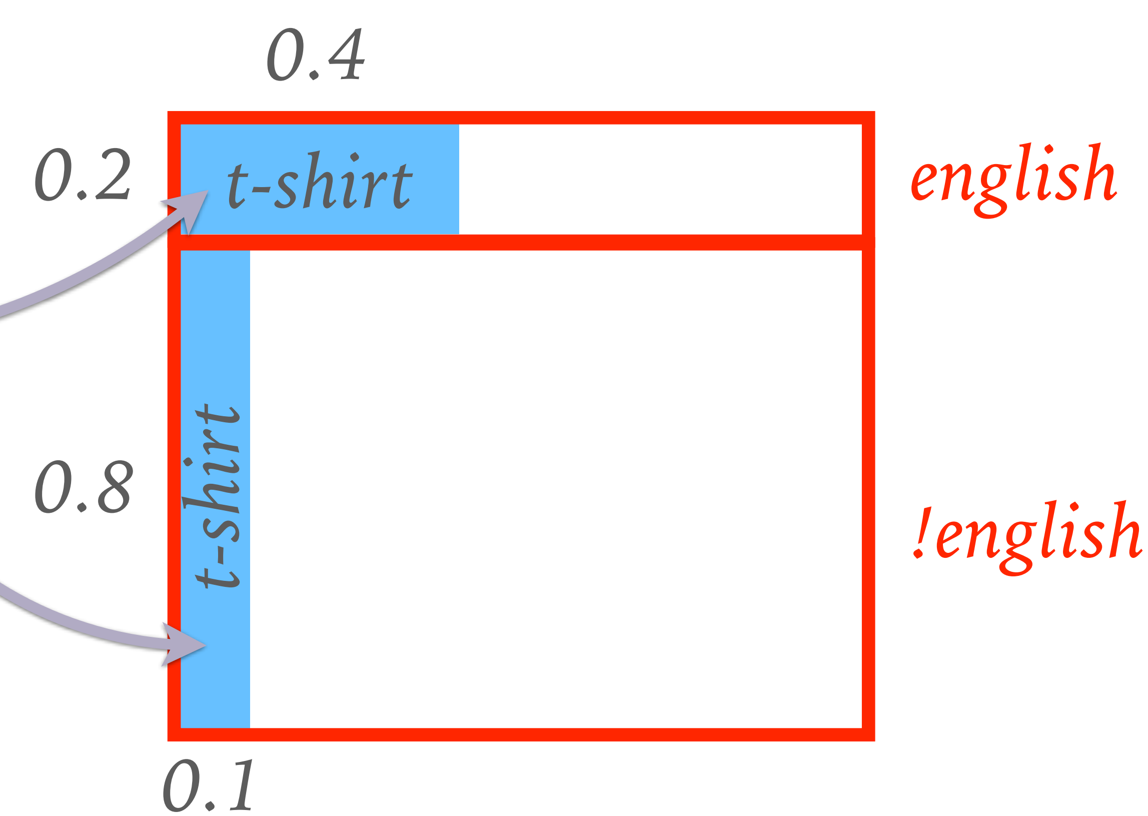
0.5

0.16

$$P(\text{tshirt}) = P(\text{tshirt} \mid \text{english})\ P(\text{english}) + P(\text{tshirt} \mid \text{!english})\ P(\text{!english})$$

0.16 0.4 0.2 0.1 0.8

P(english | DATA) =
ratio of these areas

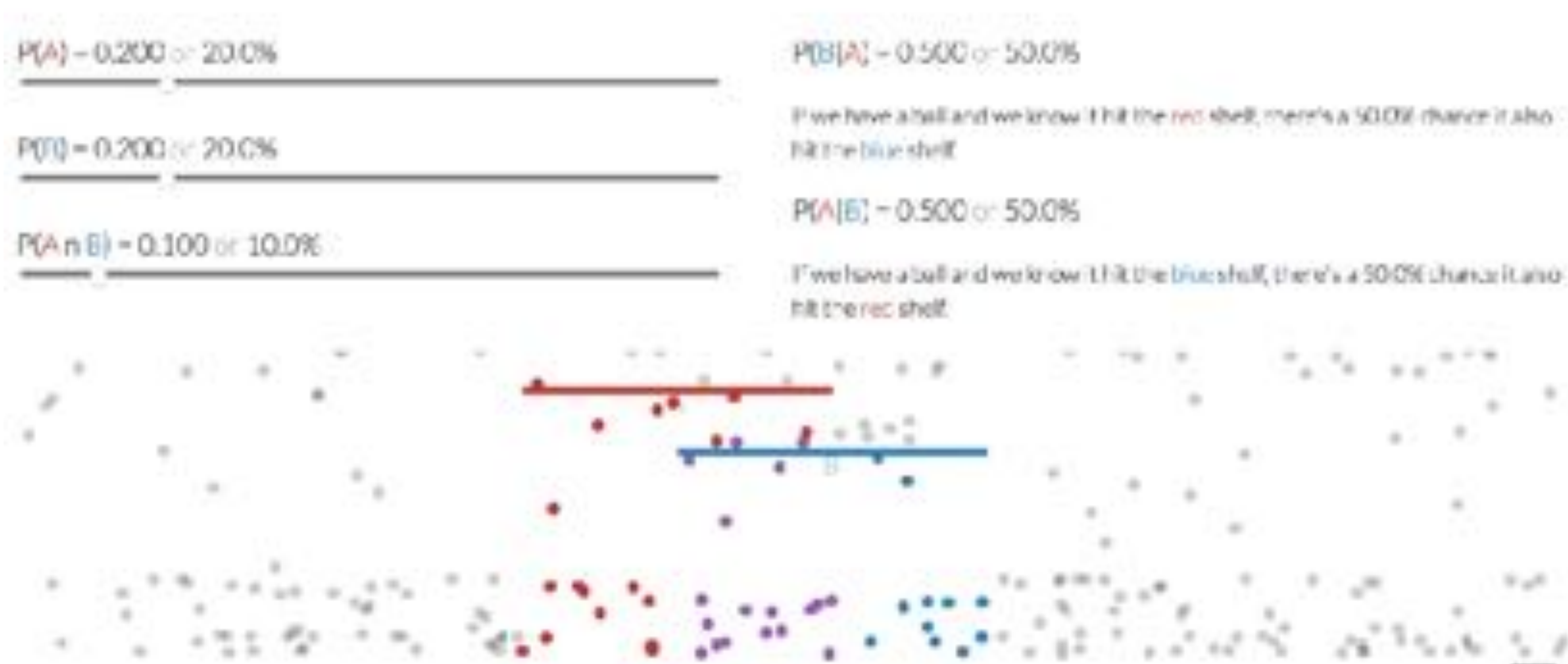


BAYES' THEOREM

http://setosa.io/ev/conditional-probability/

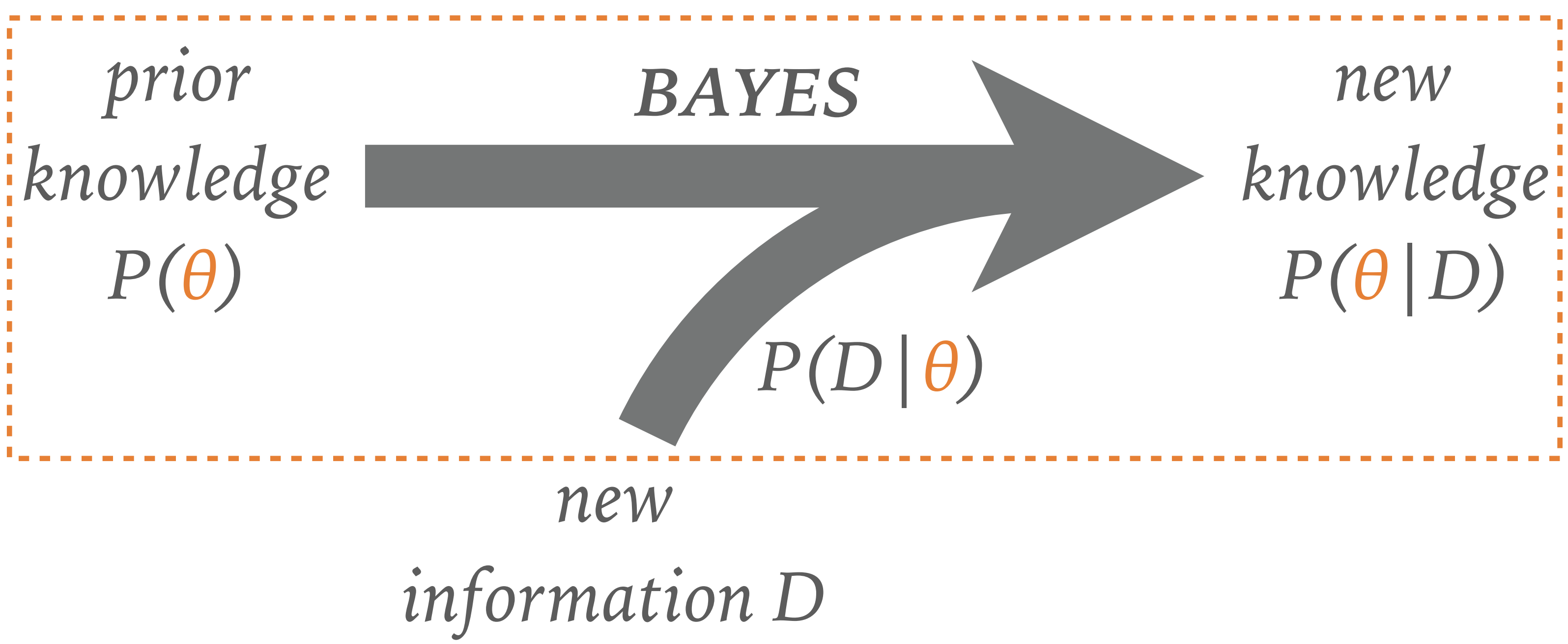
- The theorem has two ingredients:
  - The definition of conditional probability for outcomes.
  - Unified treatment of observables (data) and parameters (model) as outcomes.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$



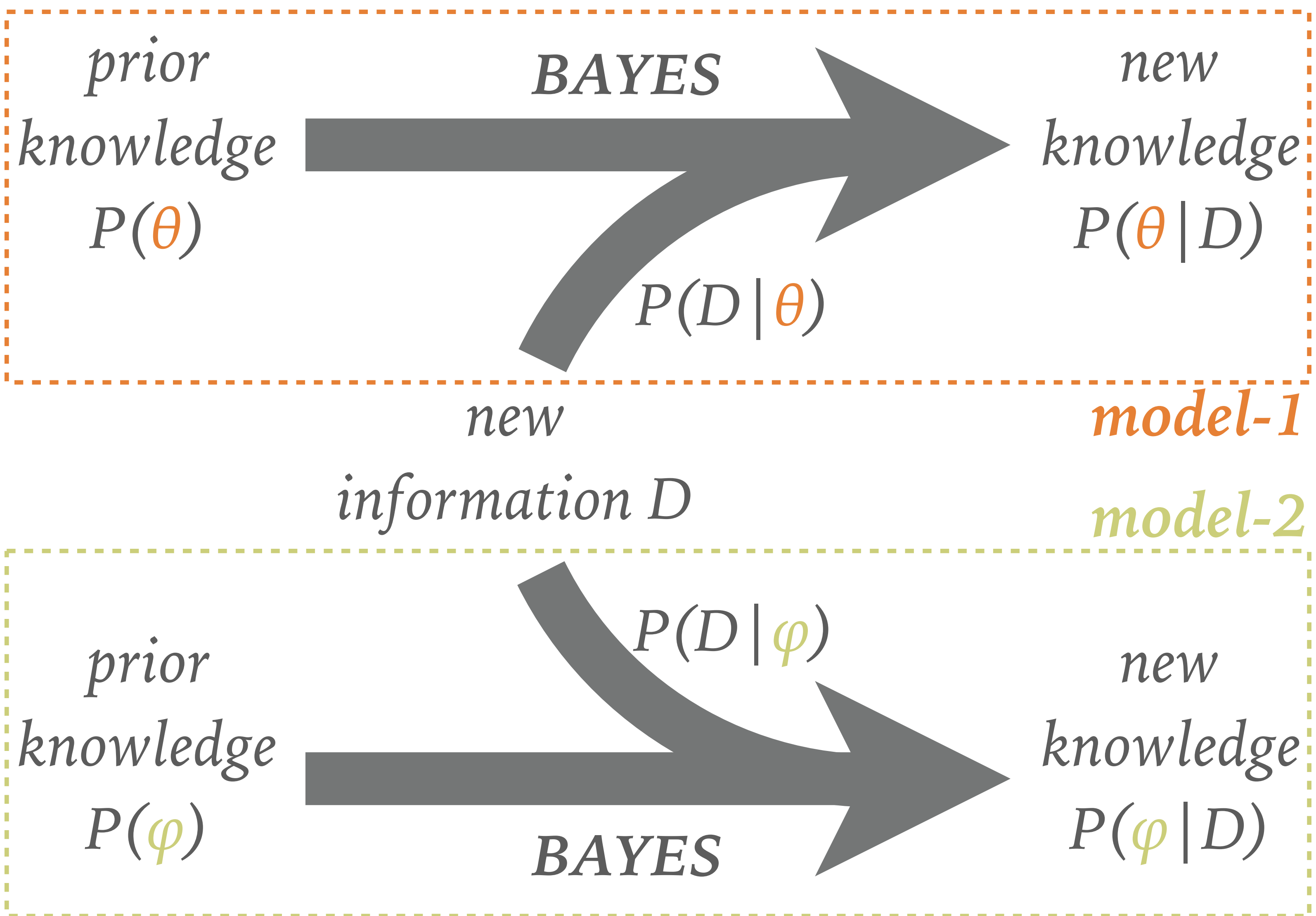
HOW DO WE USE BAYES' THEOREM?

- To update our knowledge based on new information.
  - Must specify a model and priors!



HOW DO WE USE BAYES' THEOREM?

- To update our knowledge based on new information.
  - Must specify a model and priors!
- To compare alternative models that explain the same data.



BAYESIAN MODEL COMPARISON

- We normally use Bayes' rule for the (posterior) probability of data D given specified parameters θ and model M:

$$P(\theta|D,M) = \frac{P(D|\theta,M)\,P(\theta,M)}{P(D,M)}$$

- In order to turn this into a statement about the model without specifying the parameters, we need to marginalize (integrate) them out:

$$P(M|D) = \frac{P(D|M)\,P(M)}{P(D)}$$

(I am skipping many lines of probability calculus here)

BAYESIAN MODEL COMPARISON

$$P(M|D) = \frac{P(D|M)\,P(M)}{P(D)}$$

- The denominator P(D) can only be evaluated if you can fully specify all possible models!
  - Generally cannot make statements about the absolute (posterior) probability of a single model.
  - However, P(D) cancels in probability ratios:

$$\underbrace{\frac{P(M_1|D)}{P(M_2|D)}}_{\textit{Odds ratio}} = \underbrace{\frac{P(D|M_1)}{P(D|M_2)}}_{\textit{Bayes factor}}\;\underbrace{\frac{P(M_1)}{P(M_2)}}_{\textit{Model priors}}$$

BAYESIAN MODEL COMPARISON

- How is the "naturalness" of a model taken into account?
  - Model priors.
  - Occam factor.

$$\frac{P(M_1|D)}{P(M_2|D)} = \underbrace{\frac{P(D|M_1)}{P(D|M_2)}}_{\textit{Bayes factor}} \underbrace{\frac{P(M_1)}{P(M_2)}}_{\textit{Model priors}}$$

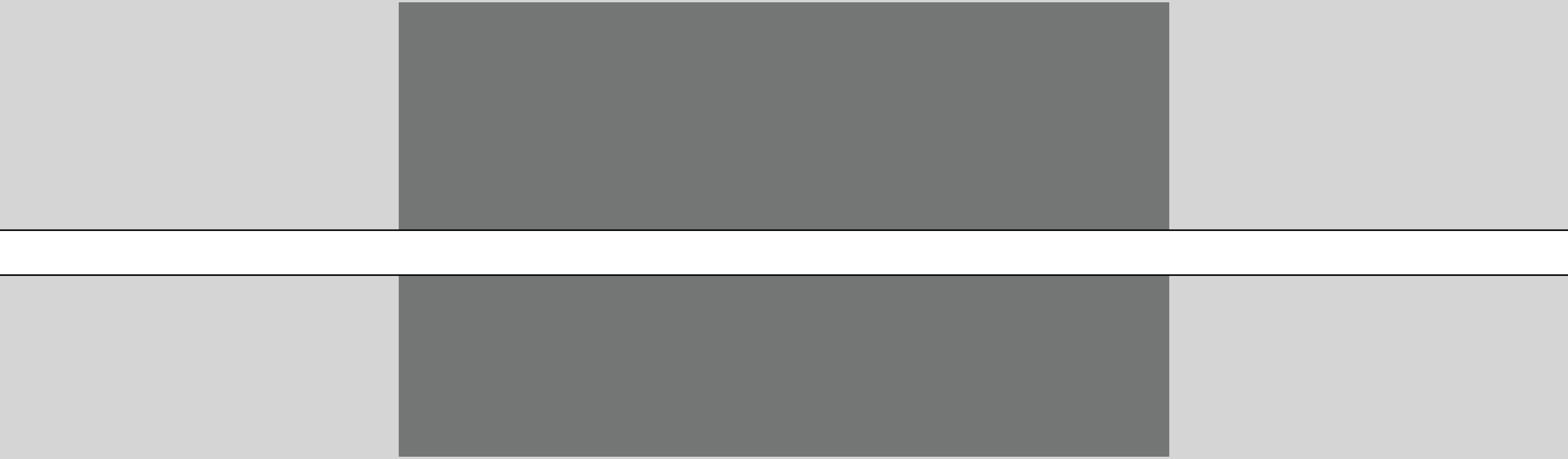
$$\underbrace{\frac{P(D|M_1)}{P(D|M_2)}}_{\textit{Bayes factor}} \propto \underbrace{\frac{(\textit{fraction of } M_1 \textit{ param. space favored by D})}{(\textit{fraction of } M_2 \textit{ param. space favored by D})}}_{\textit{Occam factor}}$$

OCCAM FACTOR

How many large galaxies?

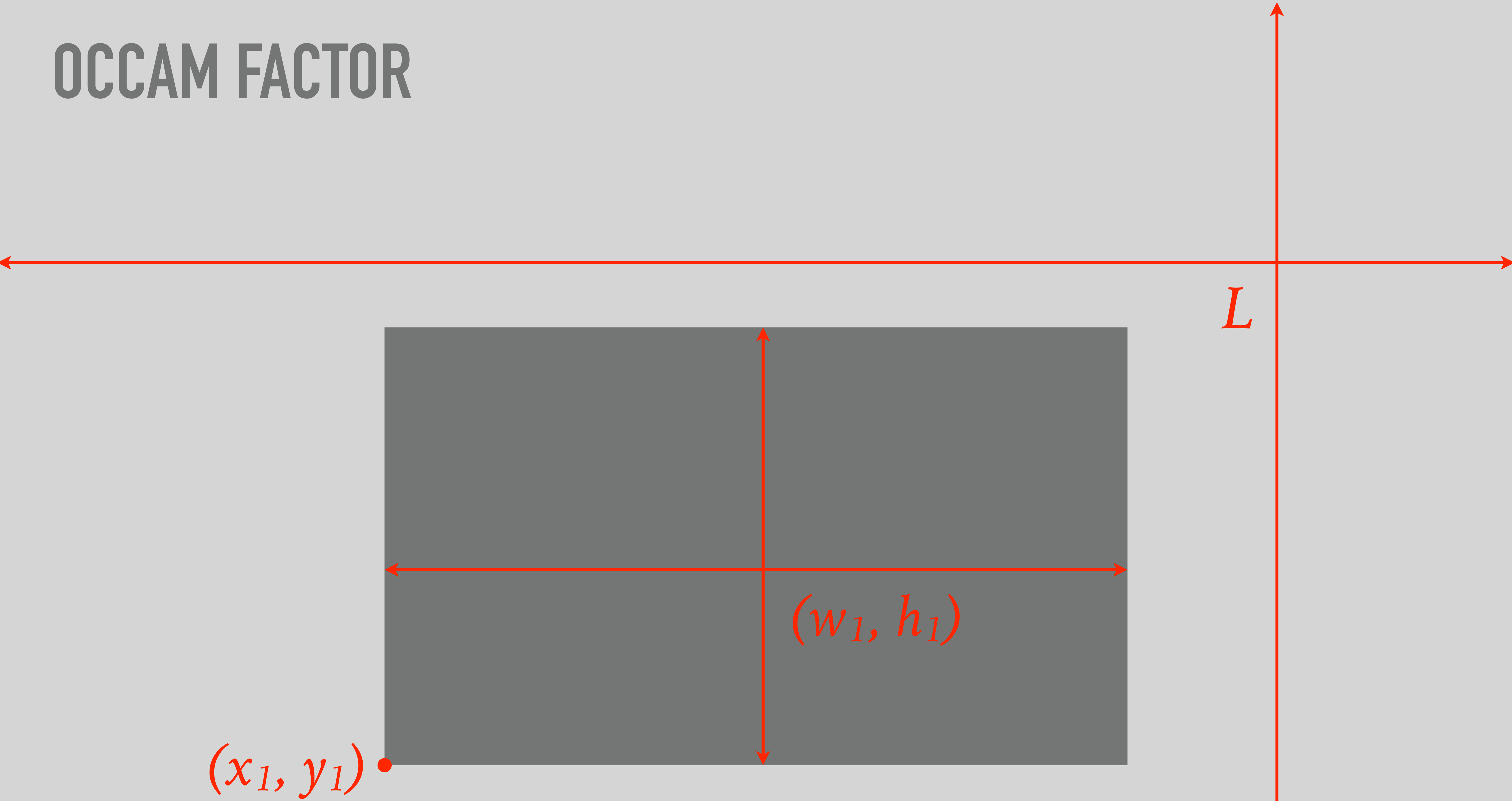
OCCAM FACTOR

- *Is it possible there are two rectangles?*
- *Why are two rectangles an unnatural model?*



How many rectangles?

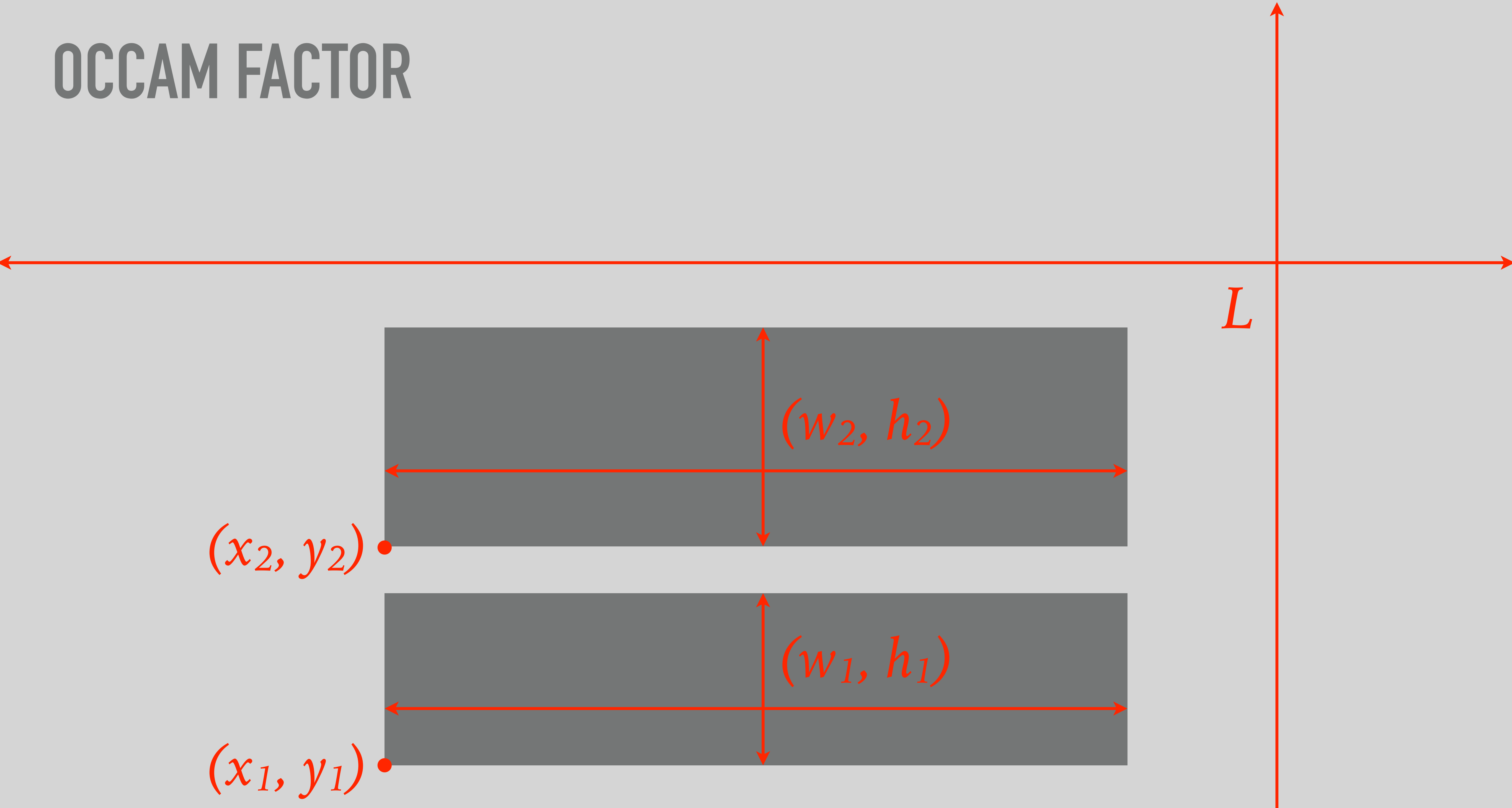
OCCAM FACTOR



$$P(D|M_1) \propto L^{-4}$$

Model-1: one rectangle.

OCCAM FACTOR



$$P(D \mid M_2) \propto L^{-8}$$

Model-2: two rectangles.

OCCAM FACTOR

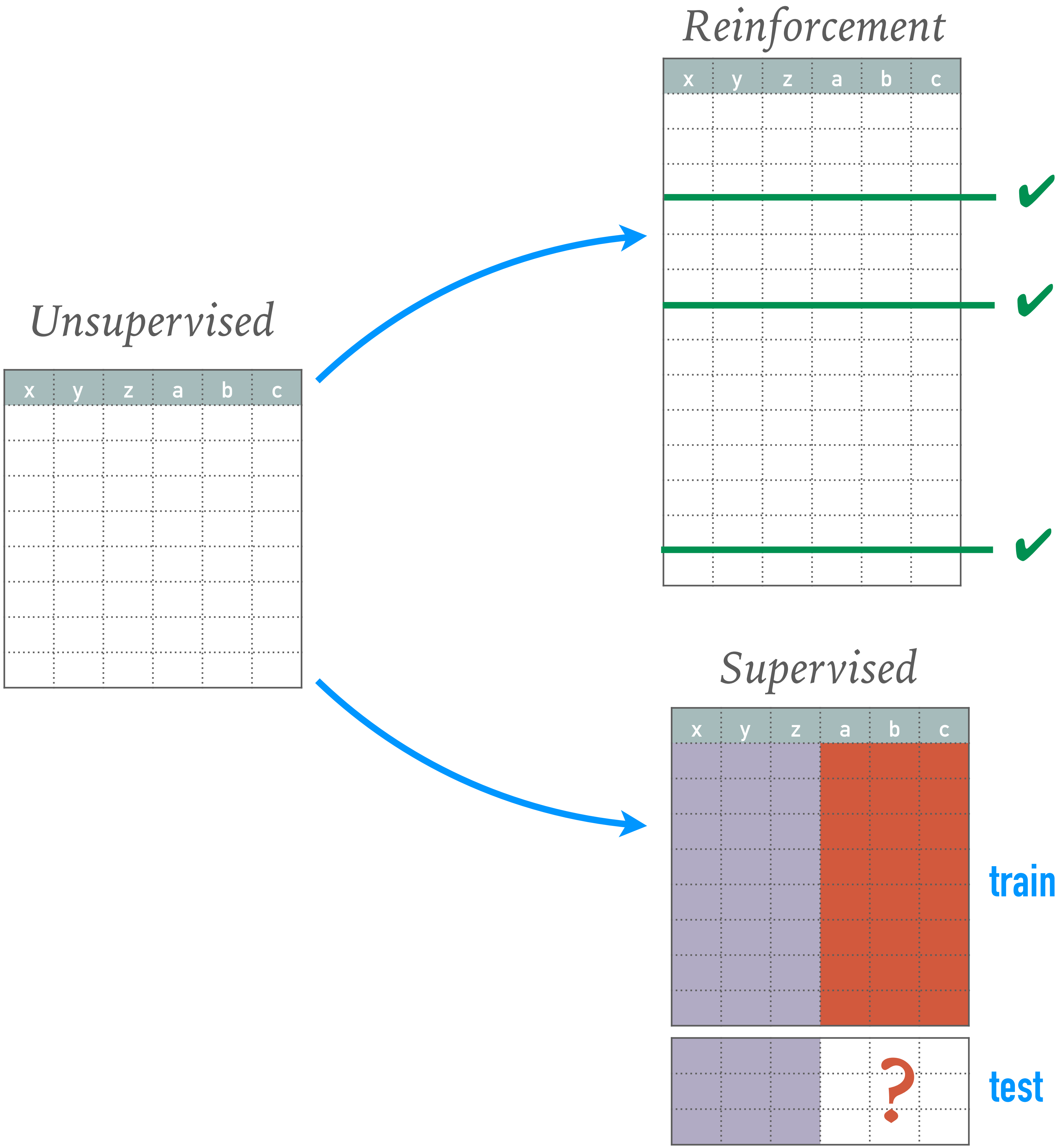
- Two rectangles are possible but extremely unlikely, even if we believe that one vs two rectangles are equally likely a-priori!

$P(D|M_2) \,/\, P(D|M_1) \propto L^{-4} \ll 1$ *Occam factor*

How many rectangles?

TYPES OF LEARNING

- Supervised
- Un-supervised
- Reinforcement
  - Video games, GO (pong example)
  - LSST observing strategy?



Reinforcement
x y z a b c
Unsupervised
x y z a b c
Supervised
x y z a b c
train
test

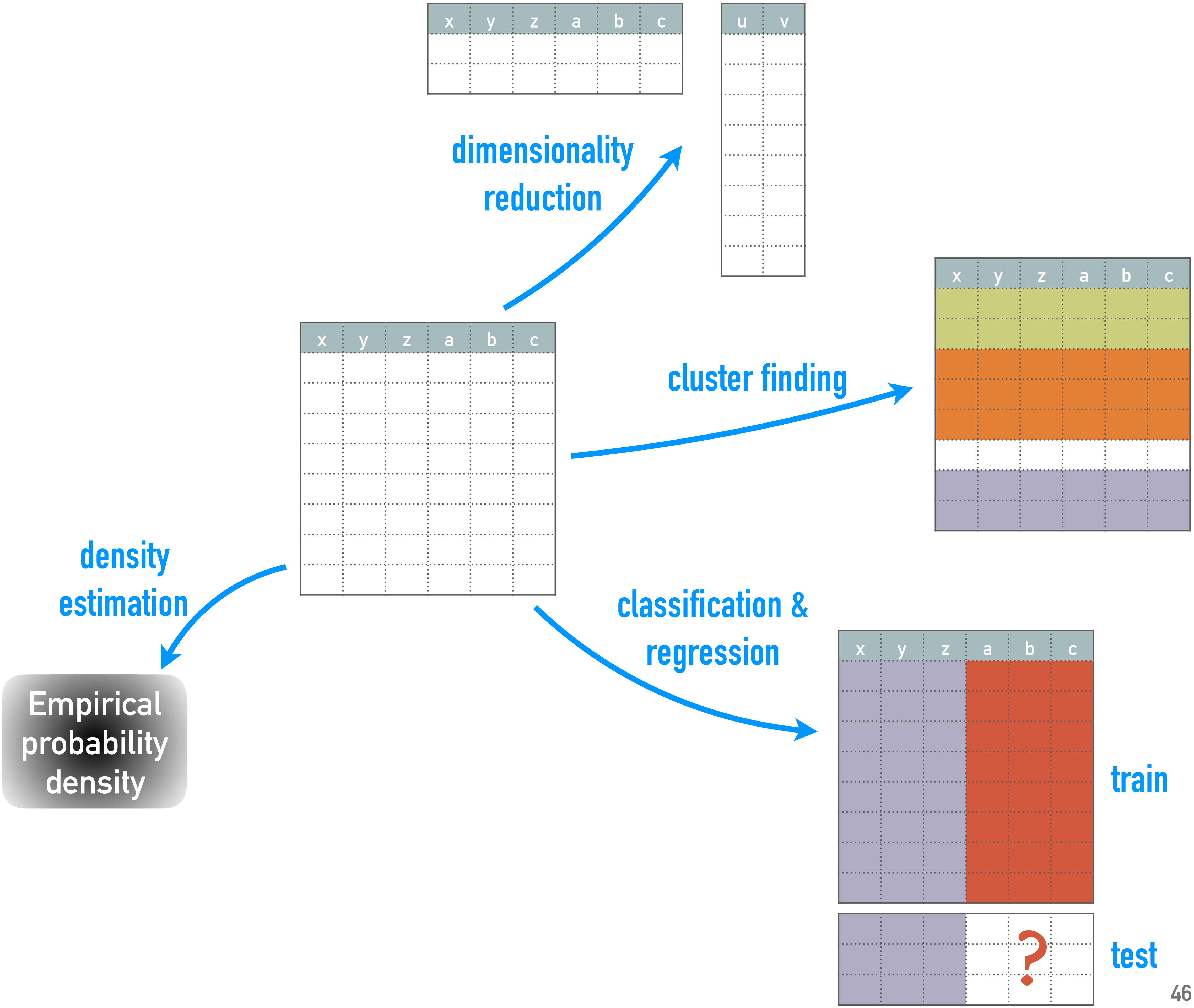
TYPES OF PROBLEM

- Classification.
- Regression.

Supervised

- Cluster finding.
- Density estimation.
- Dimensionality reduction.

Unsupervised



x y z a b c
u v
dimensionality reduction
cluster finding
density estimation
classification & regression
Empirical probability density
train
test
?

GROUP ACTIVITY: REASONING ABOUT ML PROBLEMS

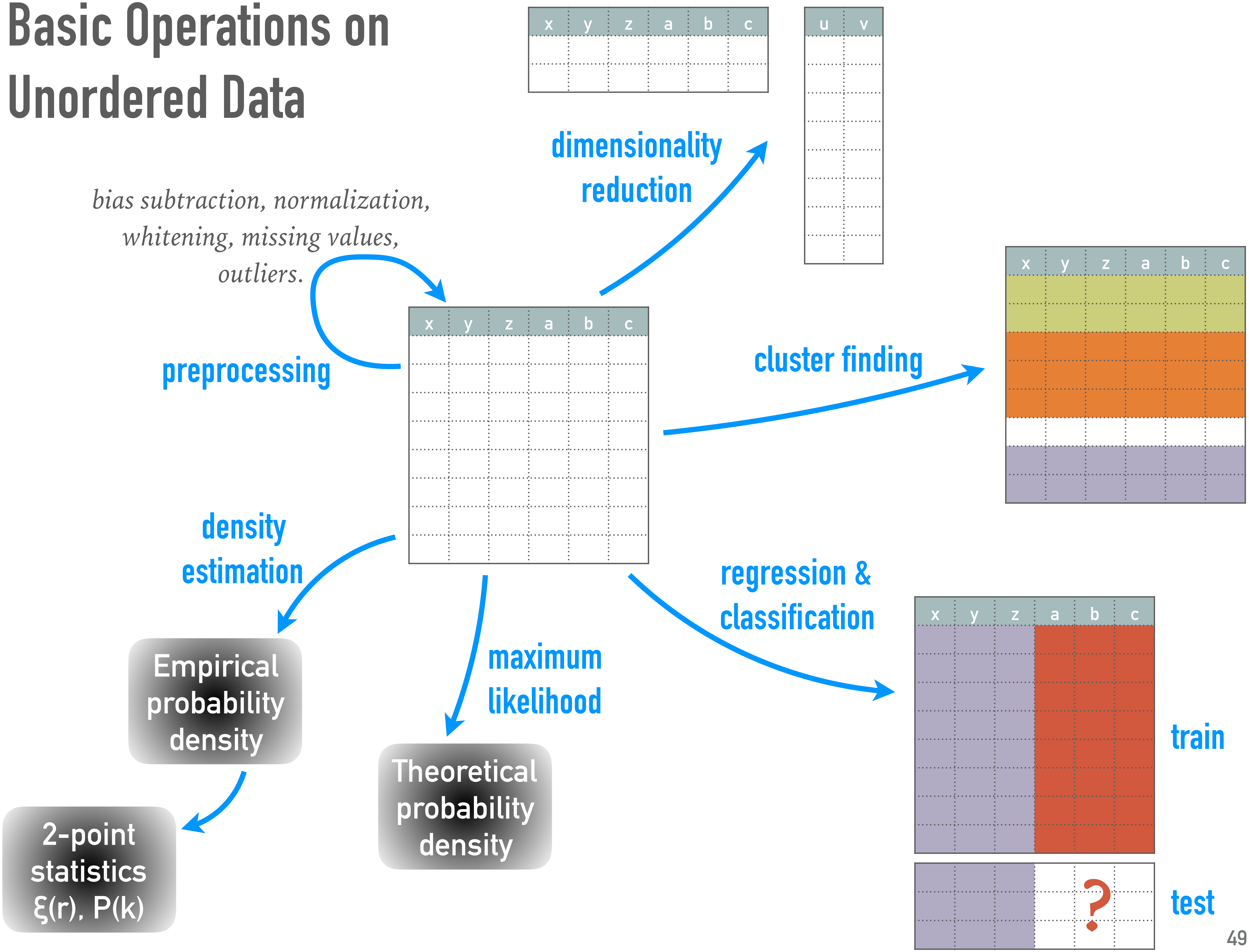
- What type of learning is best suited for these tasks?
- What type of problem best describes each task?

1. *Suggest a missing word in a sentence.*
2. *Identify a specific person in a photo.*
3. *Drive a car automatically.*
4. *Predict the sky brightness for tomorrow night's observing.*
5. *Estimate a galaxy's redshift from its LSST magnitudes.*
6. *Describe the relationship between supernovae distance and redshift.*

TYPES OF SOLUTION

- Fundamental problem: P(D) is difficult to evaluate.
- Exact solution:
  - enumerate all possible outcomes (do it when you can!)
- Approximate solution:
  - Analytic / Deterministic:
    - maximum likelihood (best-fit parameters).
    - Laplace’s approximation (parabolic errors on best-fit params).
    - variational inference (exact results for an approx. posterior).
  - Sampling:
    - Markov-chain MC (approx. results for an exact posterior).

Basic Operations on Unordered Data



RECURRING THEMES OF ML

- Neighbors
- Kernels
- Ensembles
- Regularization
- Entropy

THE MACHINE-LEARNING ZOO

MODELS

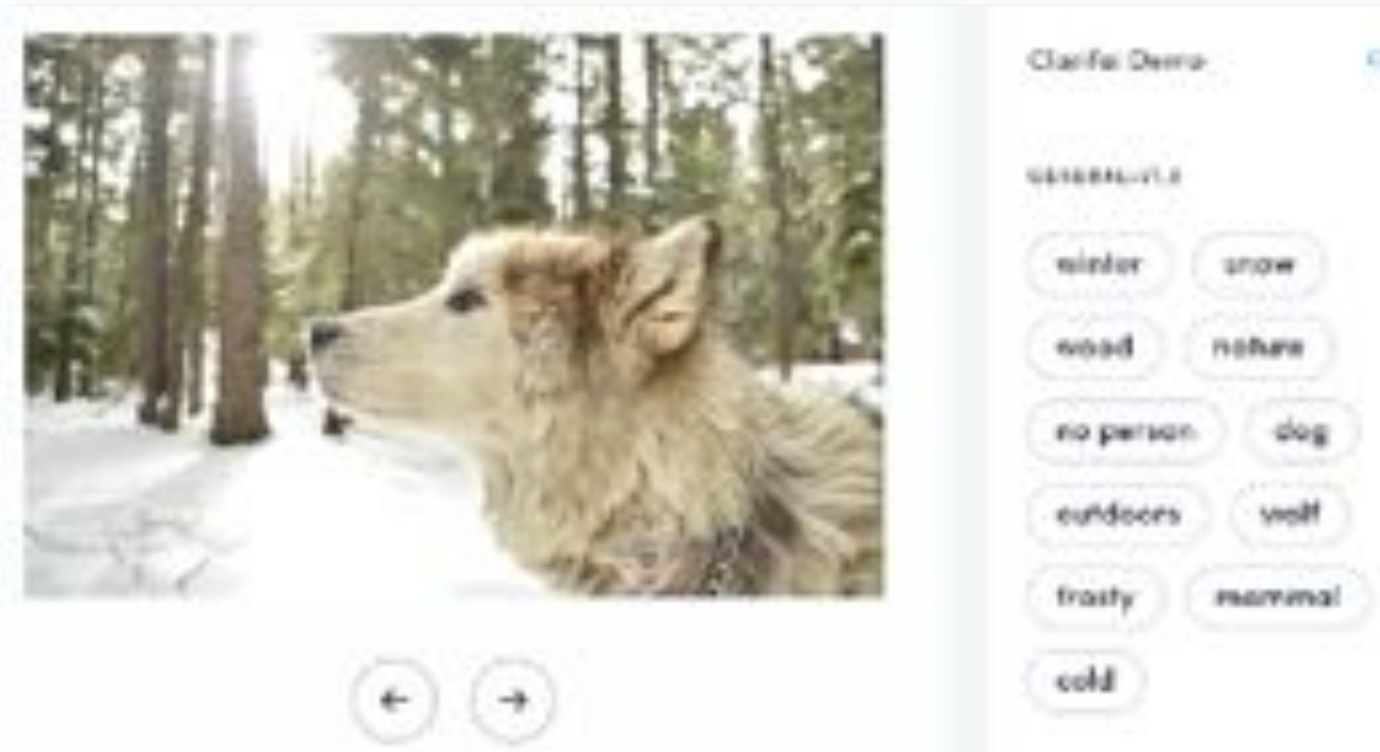
Science *Online Services* *Financial Services*

DATA

PEOPLE

http://insightdatascience.com

BLEEDING EDGE: DEEP LEARNING

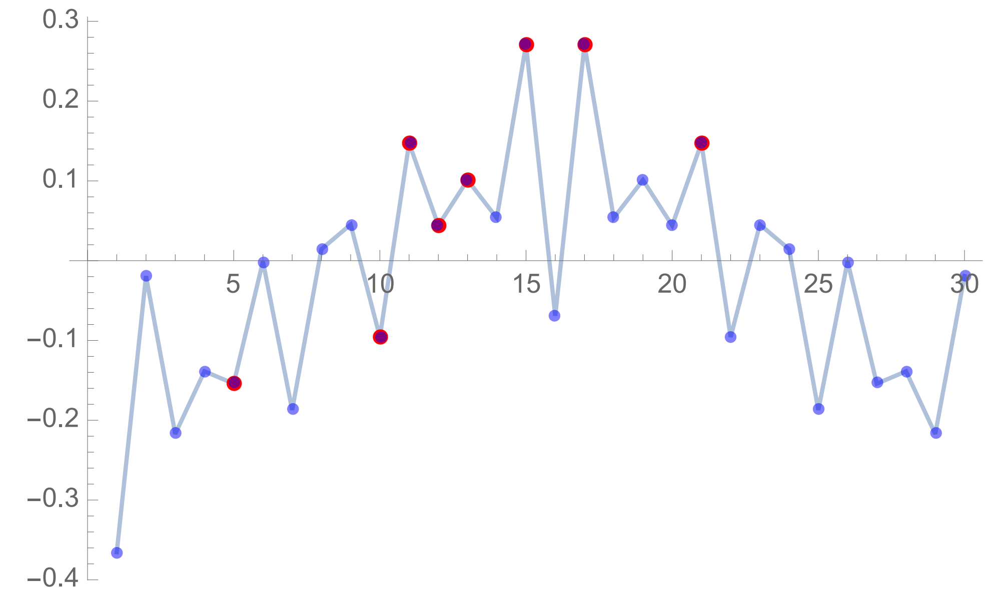


Convolutional neural networks for image classification

Recurrent neural networks for natural language semantics and translation.

BLEEDING EDGE: COMPRESSIVE SENSING

An introduction to compressive sensing

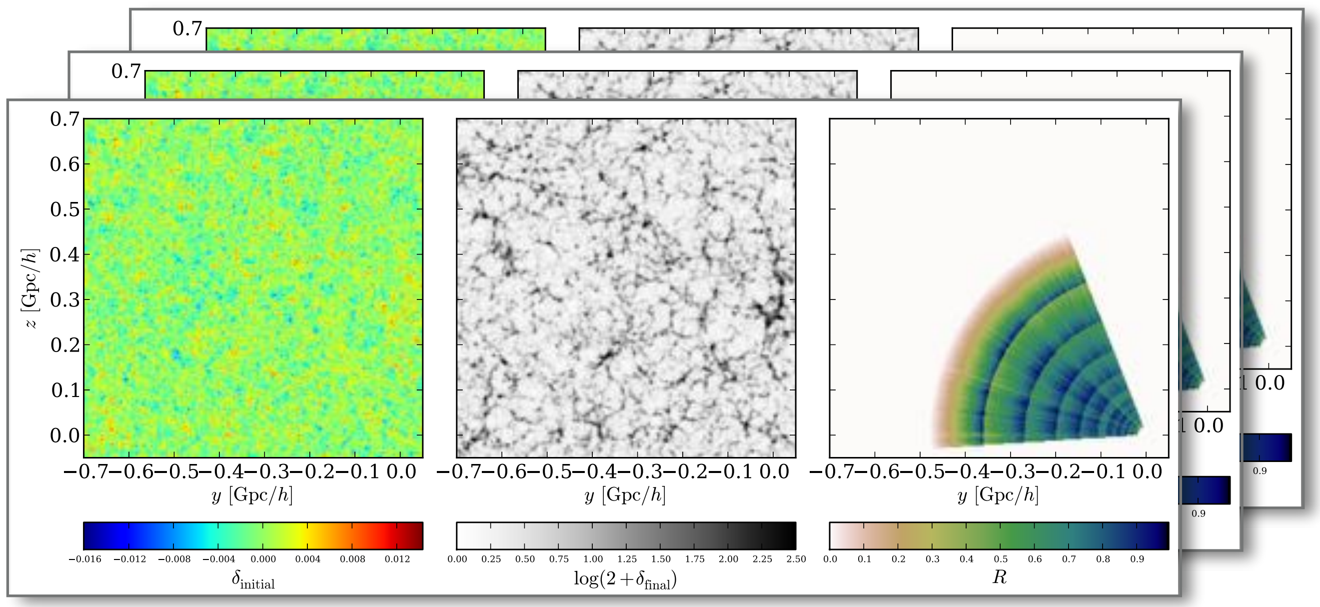


1 x 64Kpix

1300 x 1pix

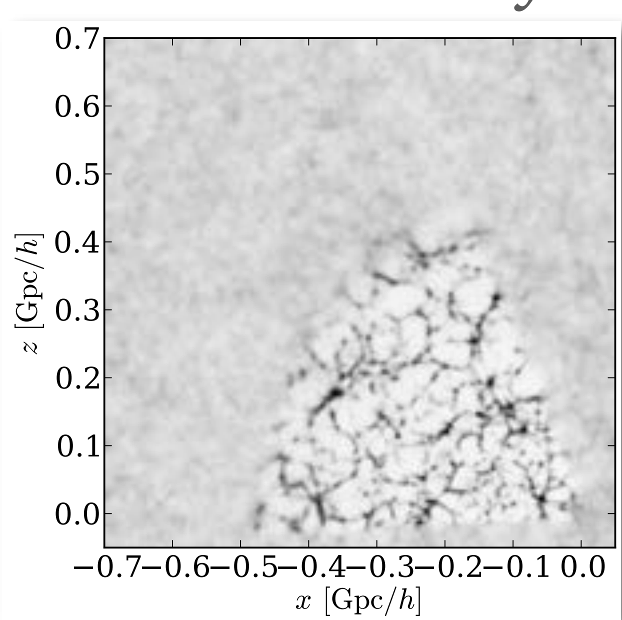
Compressive Imaging: A New Single-Pixel Camera

BLEEDING EDGE: HAMILTONIAN MONTE CARLO

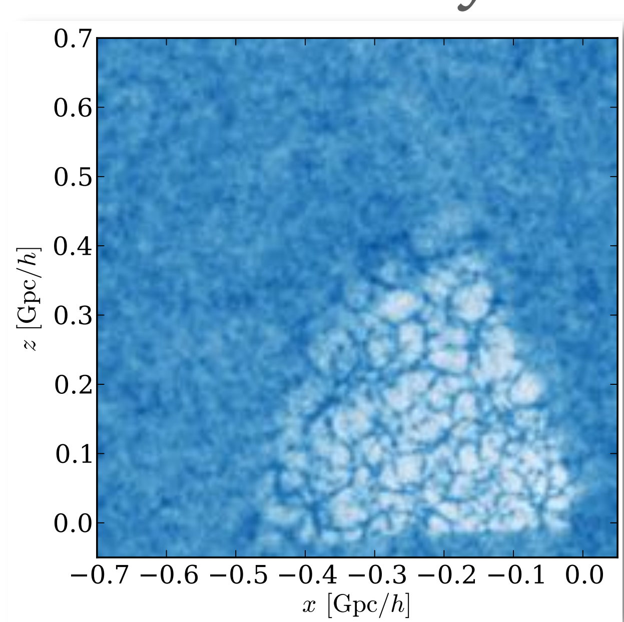


Past and present cosmic structure in the SDSS DR7 main sample

mean density



rms density



LESSONS FROM THE BLEEDING EDGE

- We need to develop & share high-quality building blocks:
  - standard data sets.
  - state of the art pre-trained solutions to low-level tasks.
- Be bold.
- Be persistent.