

INTRODUCTION TO MACHINE LEARNING IN ASTRONOMY

David Kirkby, UC Irvine

*LSSTC Data Science Fellows Program
Caltech, January 2017*

ABOUT ME

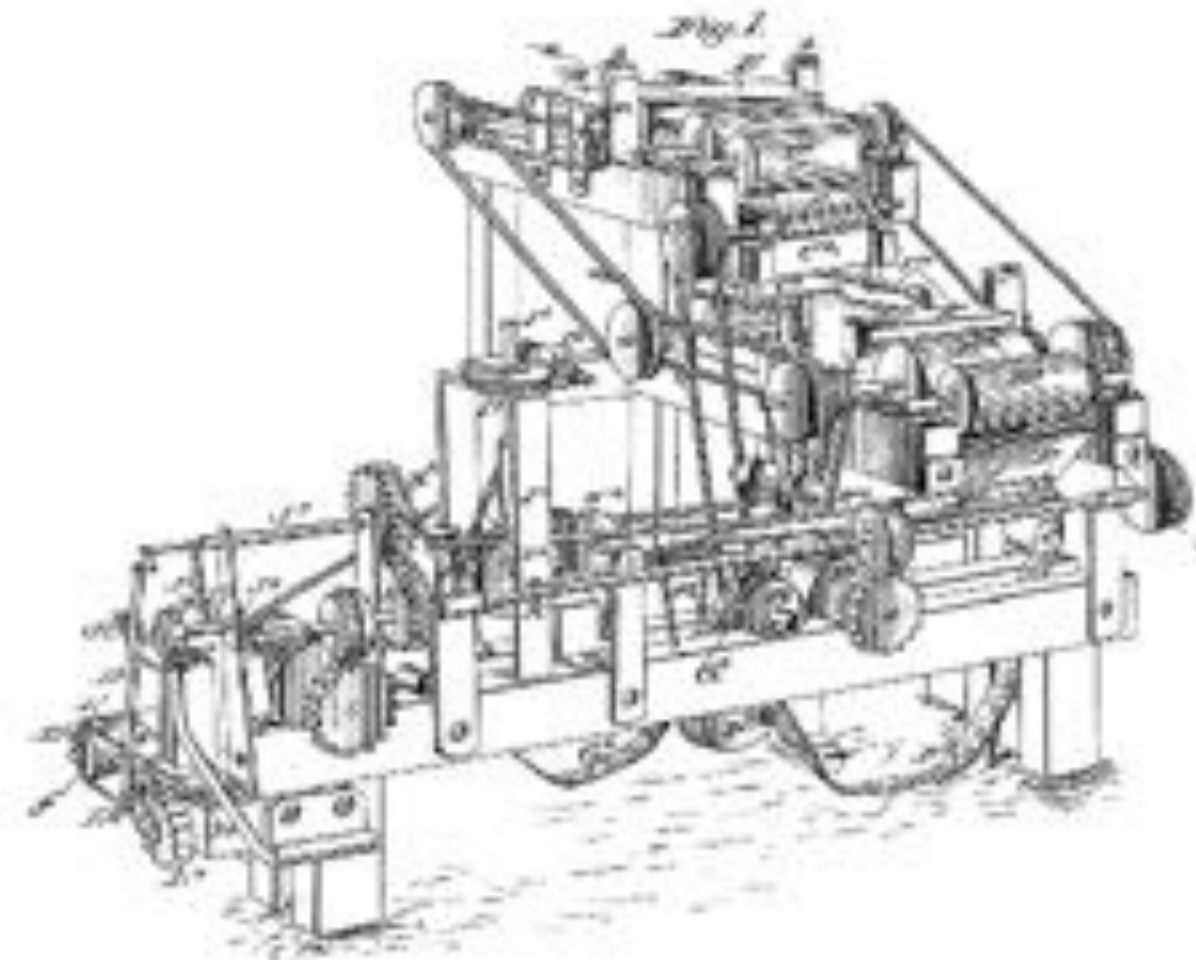
.....

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



MACHINE LEARNING

?



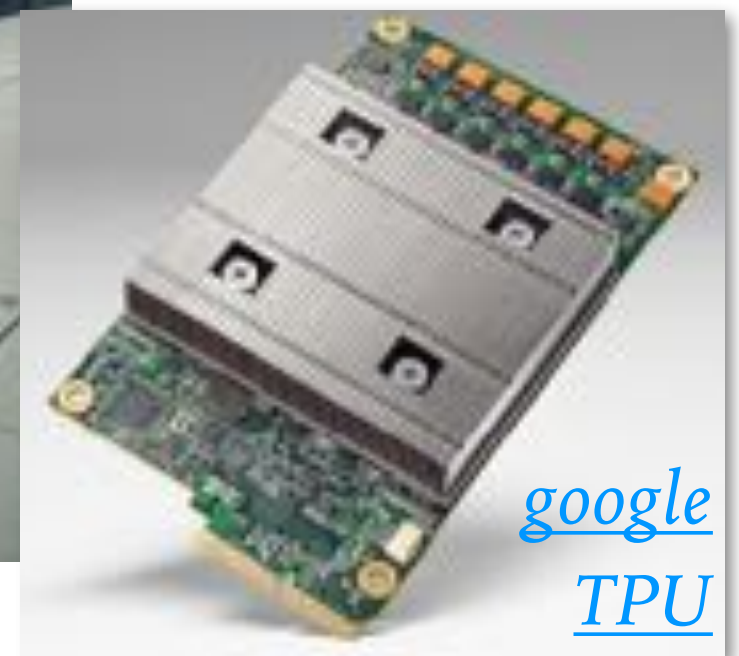
MACHINE LEARNING



```
100 return
101
102 compute : Place clusters with shape (k, n_features)
103           (distance from the last iteration of Kmeans)
104
105 select : integer matrix with shape (n, k_features)
106           (labelled as the code of the cluster the
107           j-th observation is closest to)
108
109 compute : float
110           The final value of the sum of squares (sum of squared distances to
111           the cluster centroid for all observations in the training set)
112
113 select : float, int
114           Number of iterations corresponding to the best results.
115           Returned only if 'max_iter' is set to True.
116
117
118 if n_init >= 10:
119     raise ValueError("Number of local maxima must be greater than 10")
120     # Default to 10
121     random_state = check_random_state(random_state)
122
123 if max_iter >= 0:
124     raise ValueError("Number of iterations should be a positive integer")
125     # Default to 300
126     max_iter = 300
127
128 best_iter = 0
129 k = np.float64(n_clusters)
130 tol = 1e-6
131
132 # If the distances are precomputed every job will create a copy of them
133 # in memory, to avoid this, we store them in memory only
134 # and use this if the number of observations is under 10000, if
135 # not we use a little under 10000 if they are of type double.
136 if precompute_distances < 'auto':
137     k_indices = k_indices
138     precompute_distances = 30, clusters = k_indices + 1000
139 else:
140     precompute_distances = 'auto'
141
142 # If the number of observations is under 10000, we use a little under 10000
143 # if not we use a little under 10000 if they are of type double.
144 # address of each of k for very accurate distance computation
145 if not np.iscomplexobj(X) or np.iscomplexobj(X).dtype == 'complex128':
146     k_indices = k_indices
147 else:
148     # The data was already done above
149     k_indices = k_indices
150
151 if np.iscomplexobj(X):
152     dist = check_array(X, dtype=[np.complex64, np.complex128],
153                        order='C', dtype=[np.complex64, np.complex128])
154
155     dist = k_indices
156     if k_indices is None:
157         # Default initial cluster position vector
158         # (performing only one job to avoid initial of k_clusters
159         # & k_init, but returning, k_init)
160         k_init = 0
161
162 # precompute squared norm of data points
163 k_squared_norms = np.zeros(k, dtype=[np.float64, np.float64])
164
165 best_labels, best_iter, best_cost = None, None, None
166 if k_clusters >= 0:
167     # If the number of clusters is a single cluster, all will process
168     # the right result
169     algorithm = 'k-means'
```



*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.

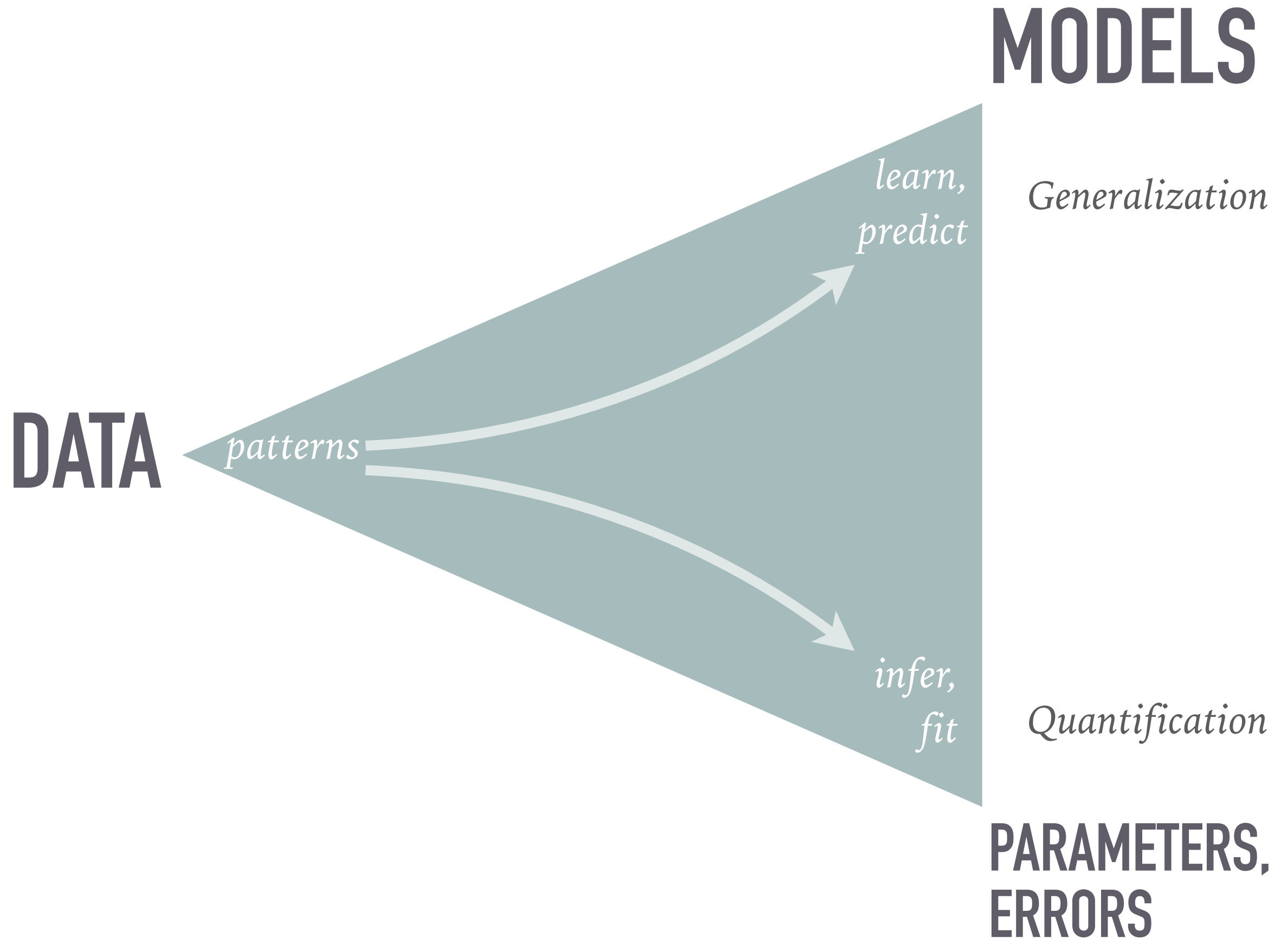
Predict the sky brightness for tomorrow night's observing.

Estimate a galaxy's redshift from its LSST magnitudes.

Describe the relationship between supernovae distance and redshift.

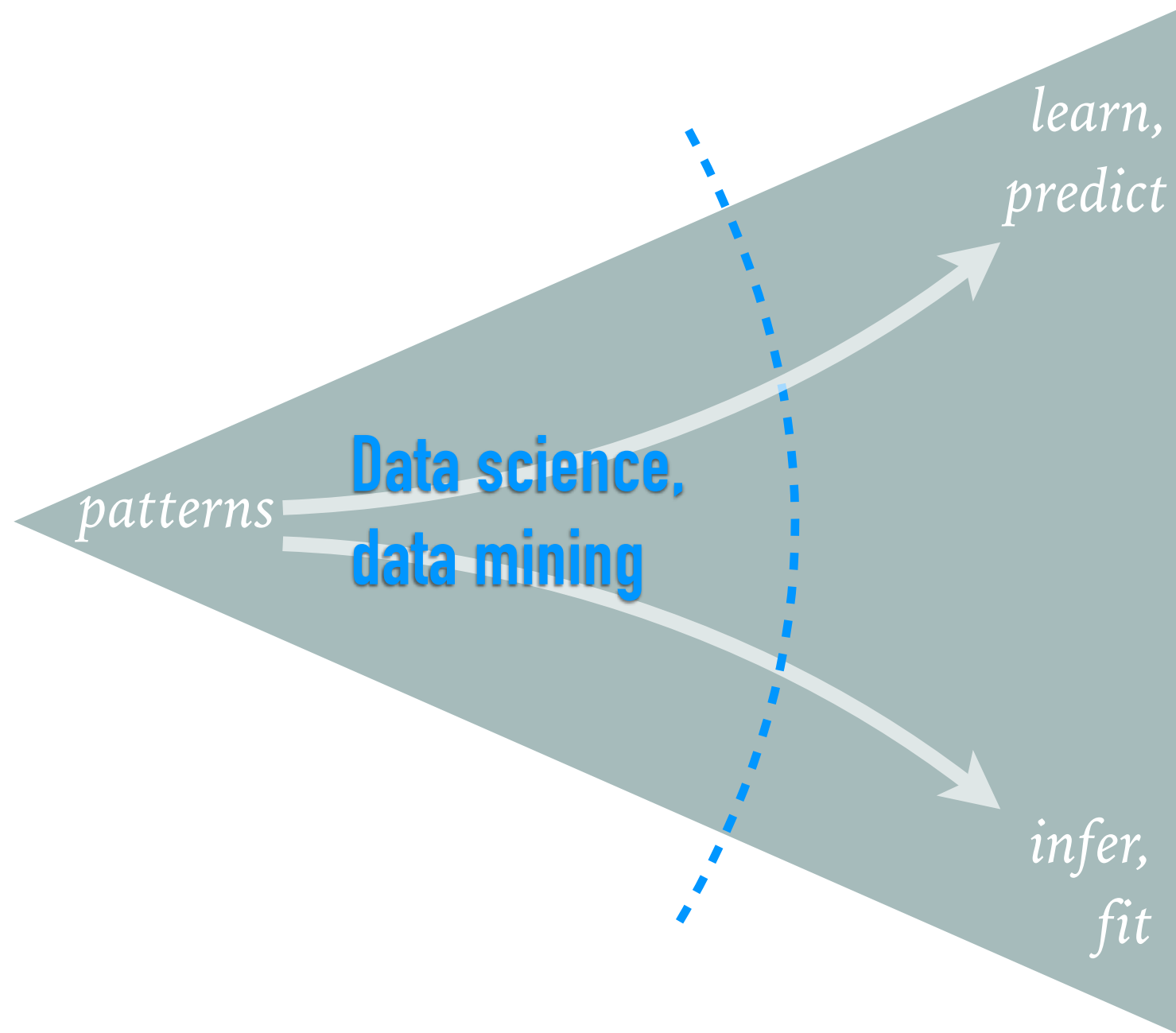
ACTIVITY: DEFINE YOUR TERMS

- 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



DATA

MODELS



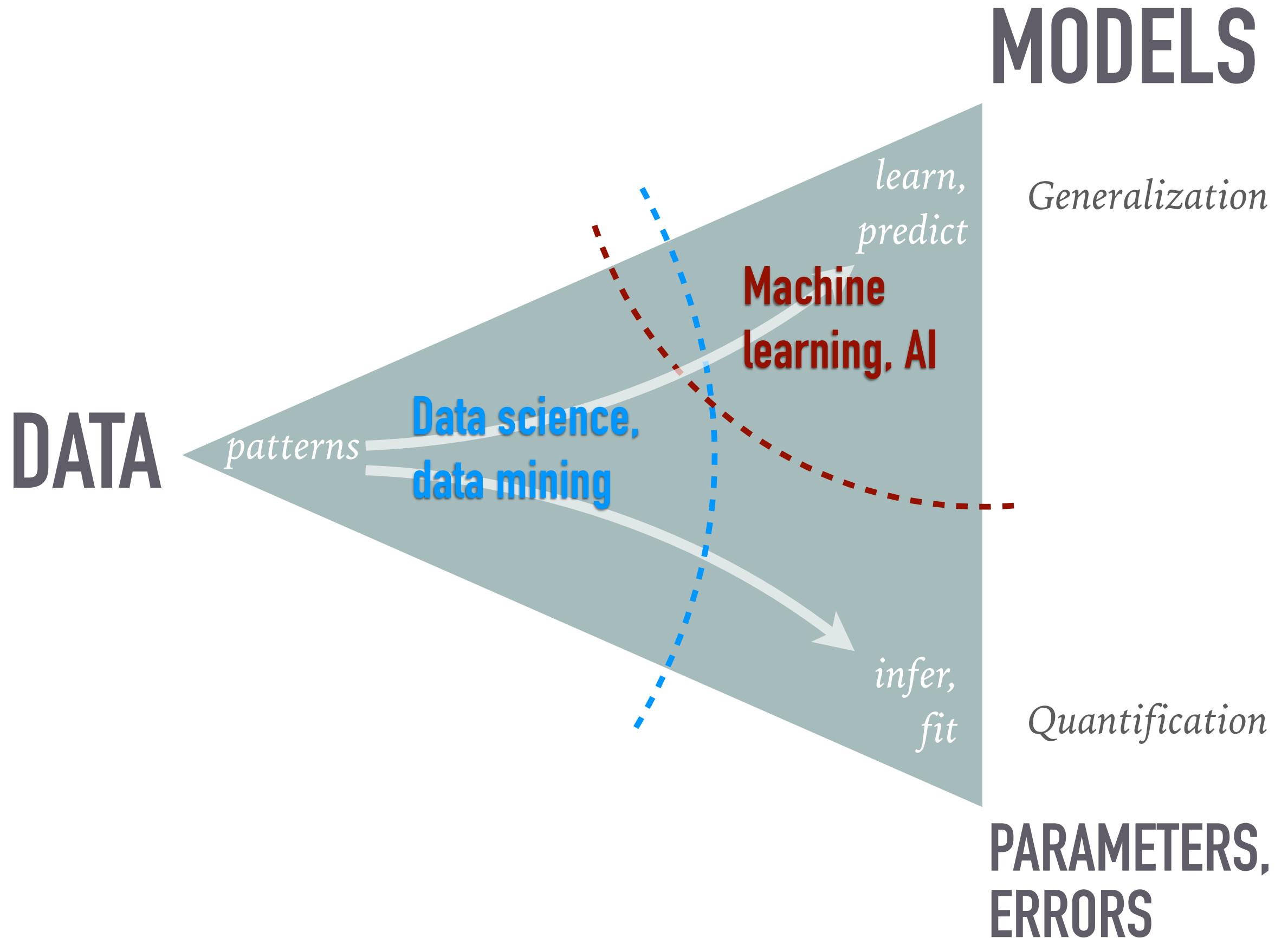
*learn,
predict*

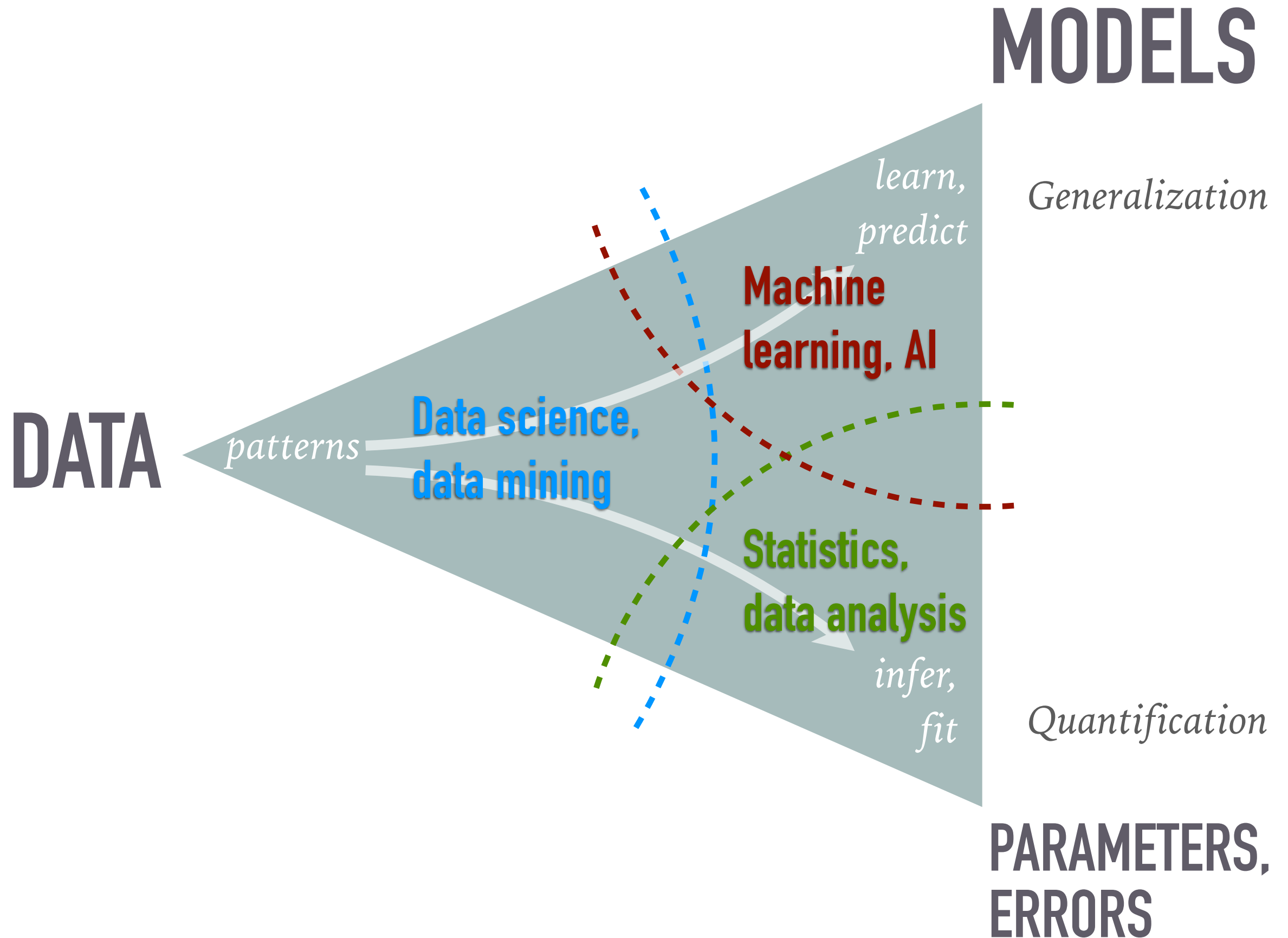
Generalization

*infer,
fit*

Quantification

**PARAMETERS,
ERRORS**





Software engineering

Data engineering

*The rise of the “data engineer”
A day in the life of a data engineer*

DATA

patterns

**Data science,
data mining**

**Machine
learning, AI**

**Statistics,
data analysis**

*infer,
fit*

MODELS

Generalization

Quantification

**PARAMETERS,
ERRORS**

Data mining and statistics: what’s the connection?

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 data?
- Is a galaxy catalog data?
- Is a histogram a model?
- Do you need a model to calculate an average?
- Does your research focus more on data or models?

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”

| | x | y | z | a | b | c |
|-------------------------------------|---|---|---|---|---|---|
| <i>rows ~ samples, observations</i> | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

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

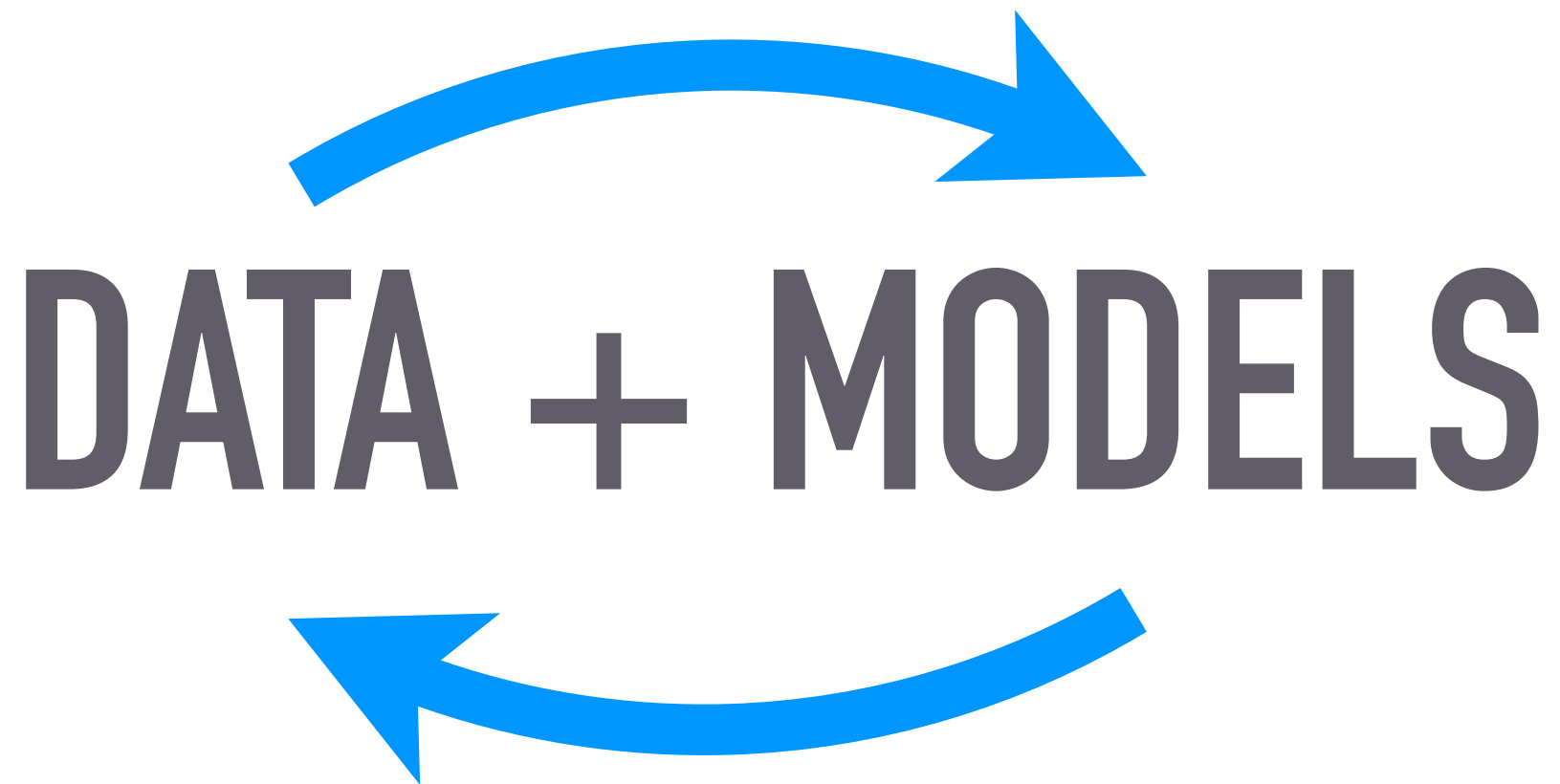
Astronomical models:

- 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 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?

- $P(\text{english} \mid \text{DATA})$?

- What do we need to assume?

- MODEL =

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

- PRIOR = “20% of astronomers are english”

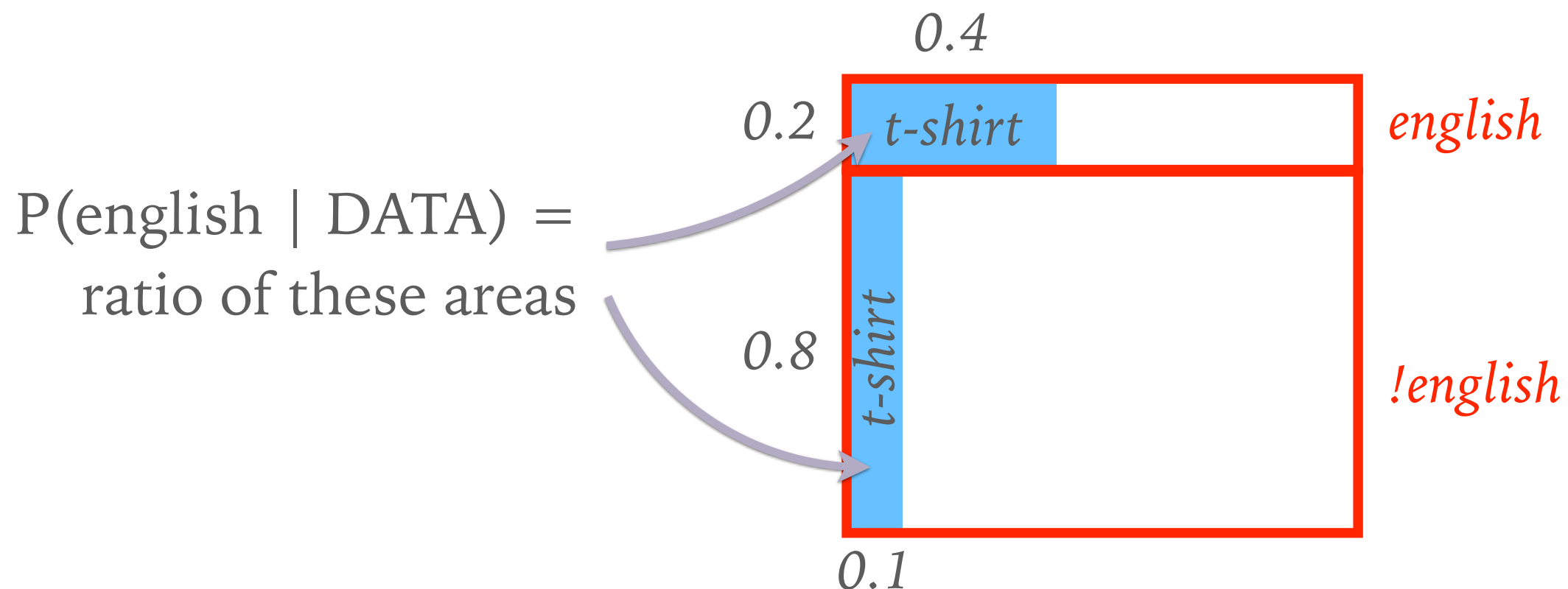
ACTIVITY: BAYESIAN REASONING

➤ PRIOR = “20% of astronomers are english”

➤ MODEL =

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

➤ DATA = “wearing an ENGLAND tshirt”



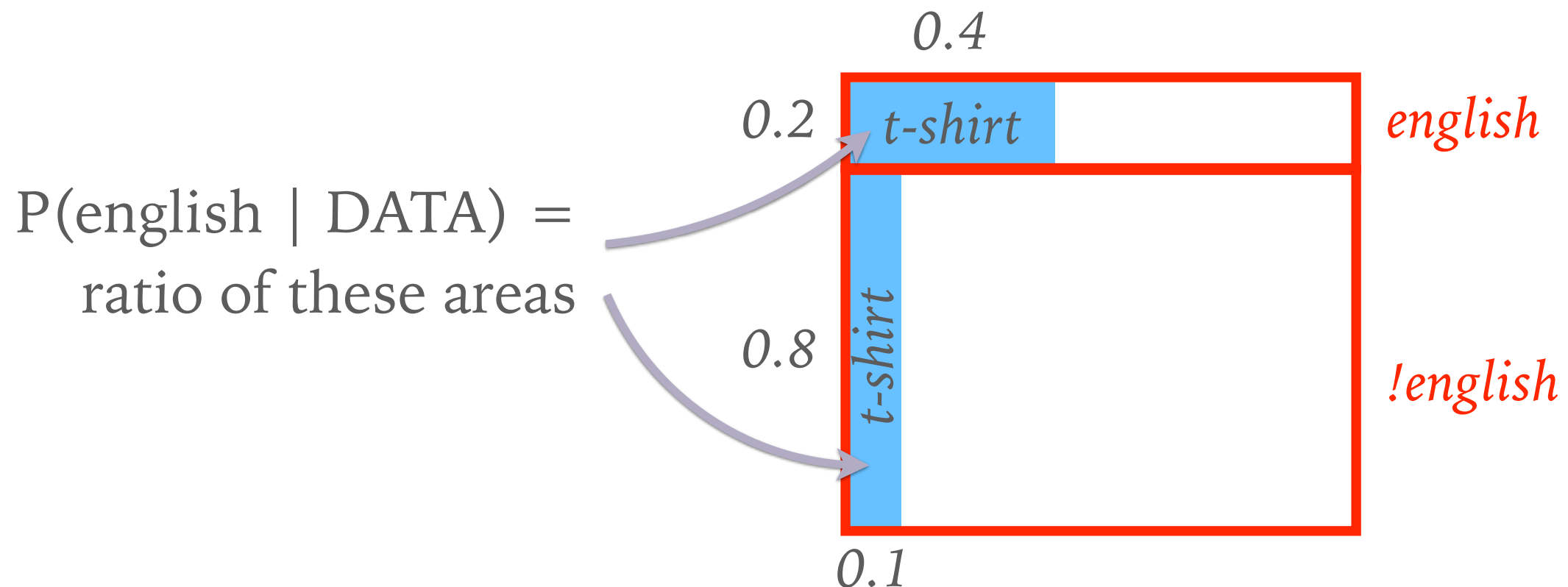
ACTIVITY: BAYESIAN REASONING

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

$0.5 \qquad \qquad \qquad 0.4 \qquad \qquad \qquad 0.2 \qquad \qquad \qquad 0.16$

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

$0.16 \qquad \qquad 0.4 \qquad \qquad 0.2 \qquad \qquad 0.1 \qquad \qquad 0.8$



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.

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

$P(A) = 0.200$ or 20.0%

$P(B) = 0.200$ or 20.0%

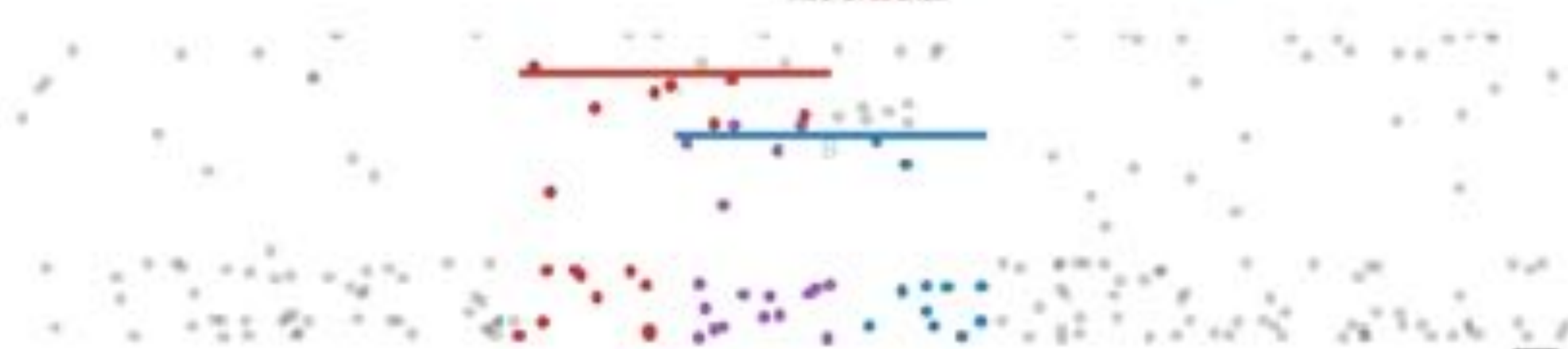
$P(A \cap B) = 0.100$ or 10.0%

$P(B|A) = 0.500$ or 50.0%

If we have a ball and we know it hit the **red** shelf, there's a 50.0% chance it also hit the **blue** shelf.

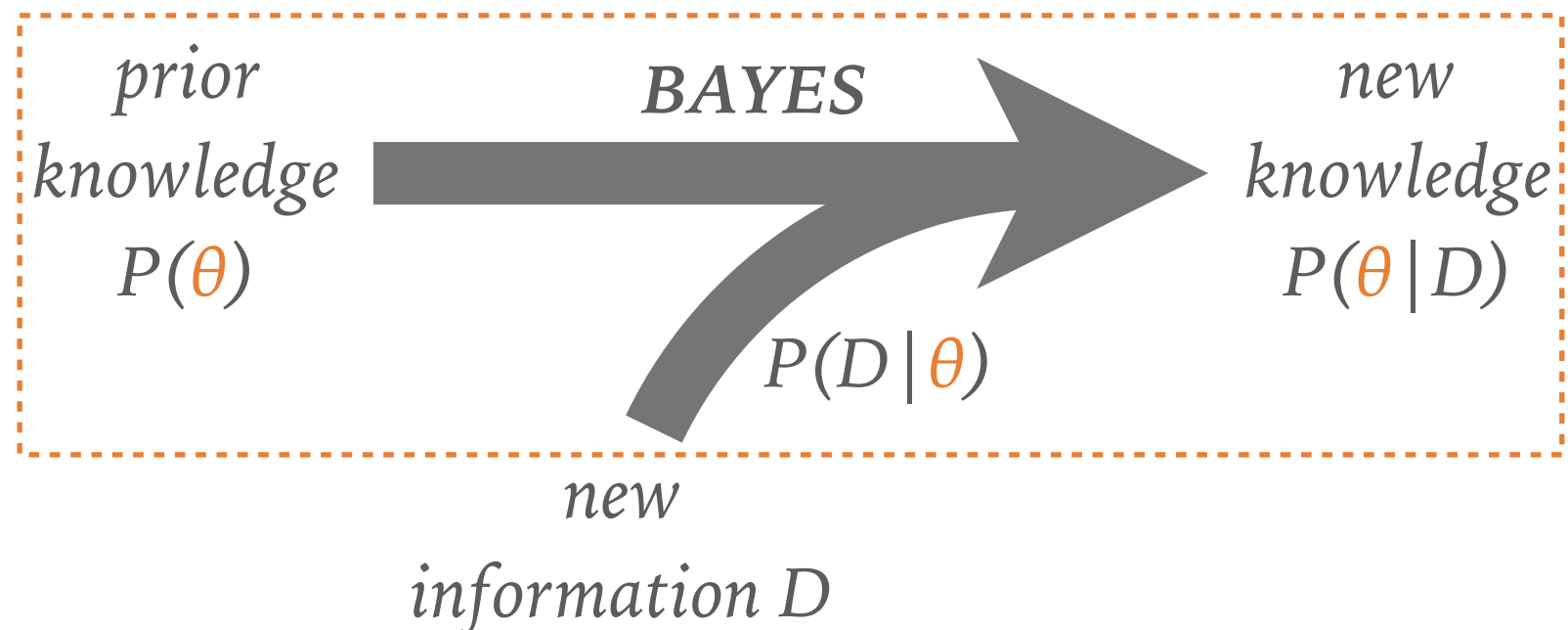
$P(A|B) = 0.500$ or 50.0%

If we have a ball and we know it hit the **blue** shelf, there's a 50.0% chance it also hit the **red** shelf.



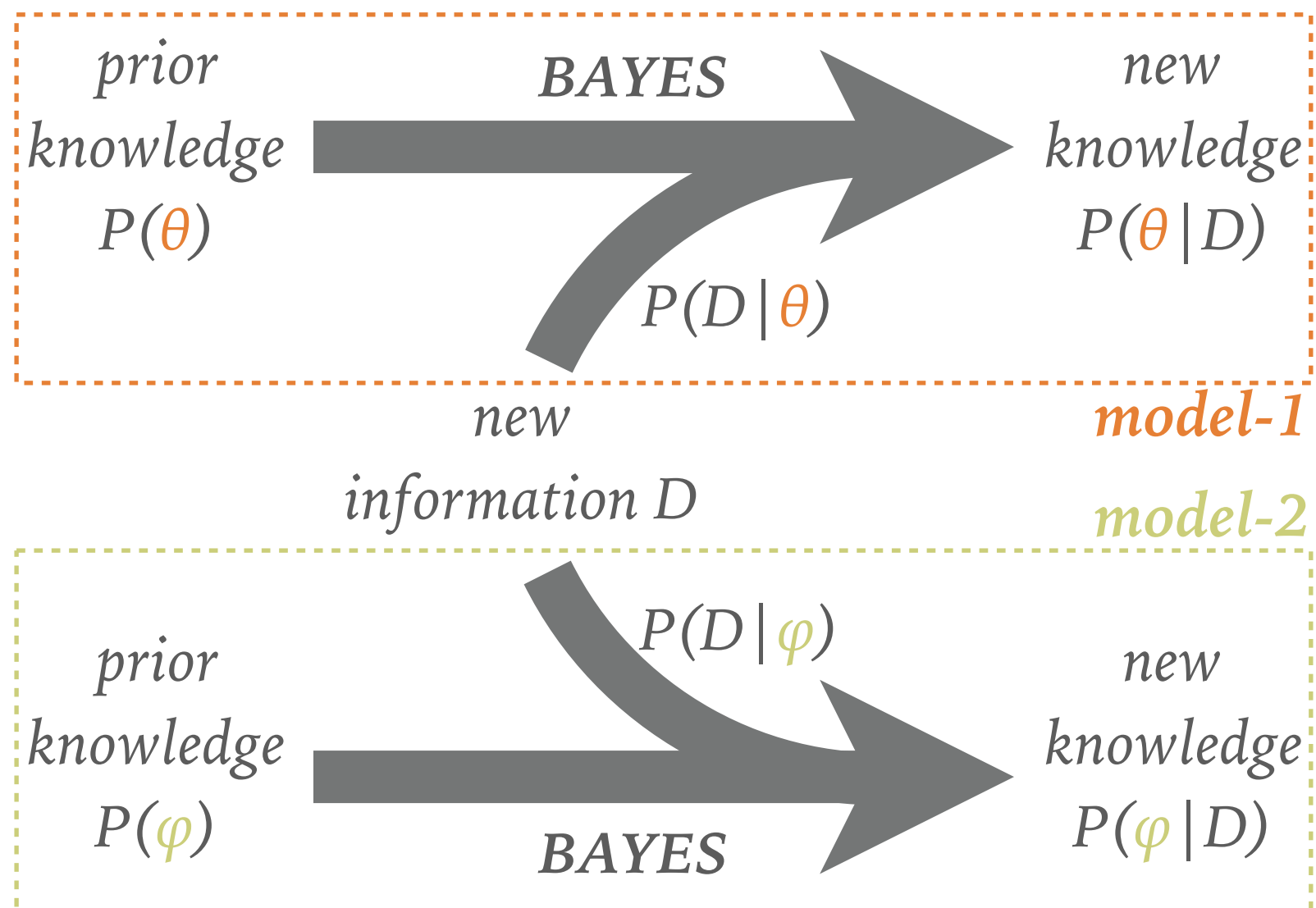
HOW DO WE USE BAYES' THEOREM?

- 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** :

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

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

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

(I am skipping many lines of probability calculus here)

BAYESIAN MODEL COMPARISON

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

- 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:

$$\frac{P(\textcolor{brown}{M}_1 | D)}{P(\textcolor{brown}{M}_2 | D)} = \frac{P(D | \textcolor{brown}{M}_1) P(\textcolor{brown}{M}_1)}{P(D | \textcolor{brown}{M}_2) P(\textcolor{brown}{M}_2)} \quad \begin{array}{l} \textit{Model} \\ \textit{priors} \end{array}$$

Odds ratio *Bayes factor*

BAYESIAN MODEL COMPARISON

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

$$\frac{P(\mathbf{M}_1 | D)}{P(\mathbf{M}_2 | D)} = \frac{P(D | \mathbf{M}_1) P(\mathbf{M}_1)}{P(D | \mathbf{M}_2) P(\mathbf{M}_2)} \quad \begin{array}{l} \text{Model} \\ \text{priors} \end{array}$$

Bayes factor

$$\frac{P(D | \mathbf{M}_1)}{P(D | \mathbf{M}_2)} \propto \frac{(\text{fraction of } \mathbf{M}_1 \text{ param. space favored by } D)}{(\text{fraction of } \mathbf{M}_2 \text{ param. space favored by } D)}$$

Bayes factor *Occam factor*

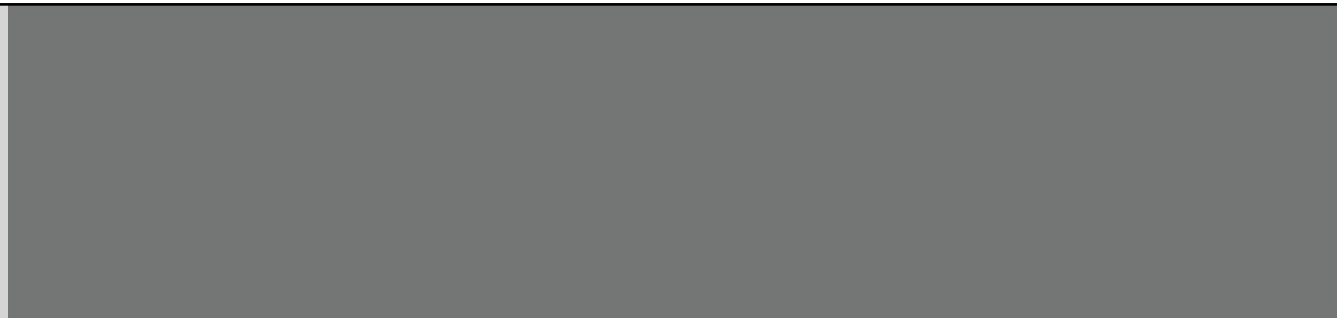
OCCAM FACTOR



How many large galaxies?

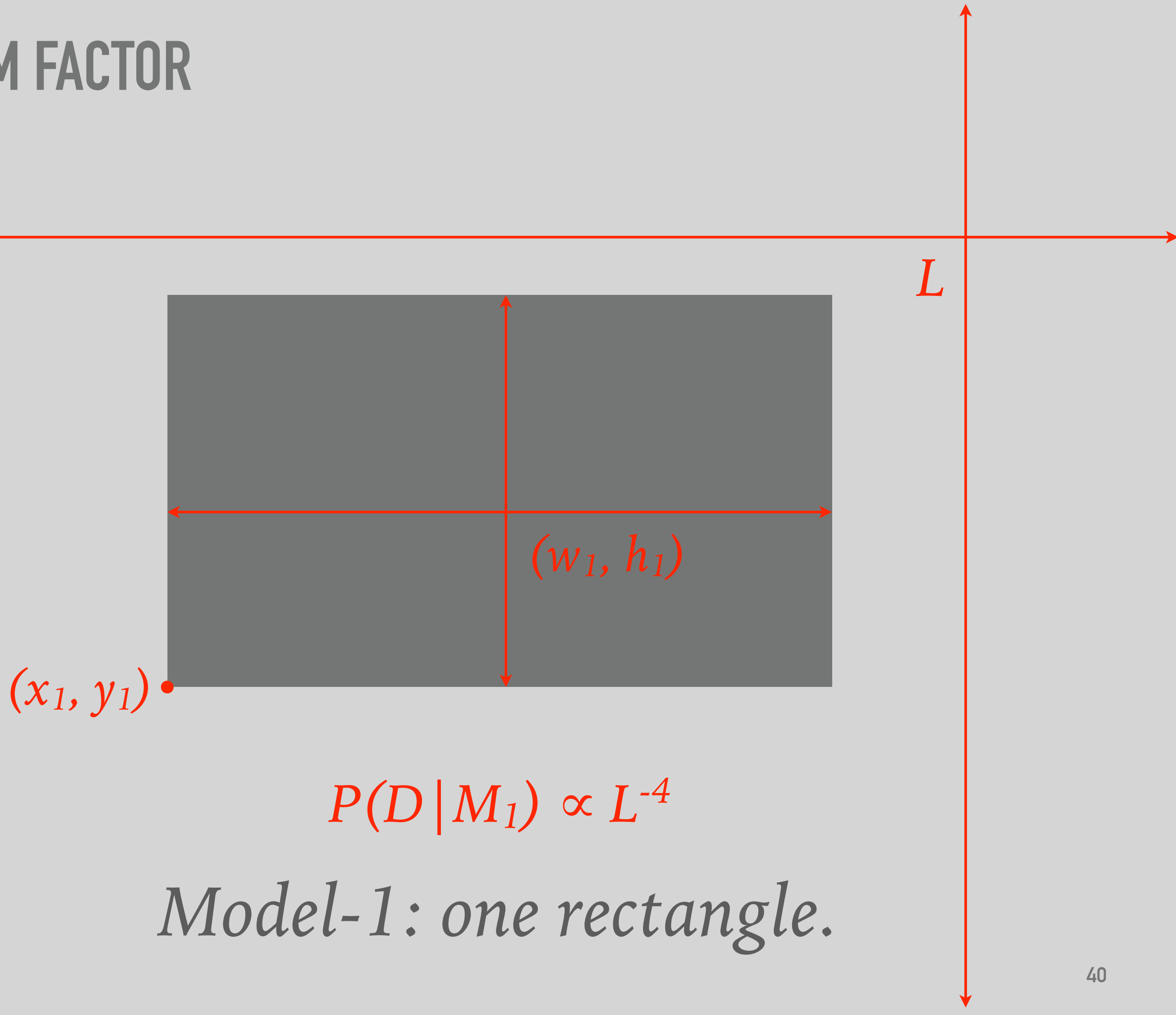
OCCAM FACTOR

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



How many rectangles?

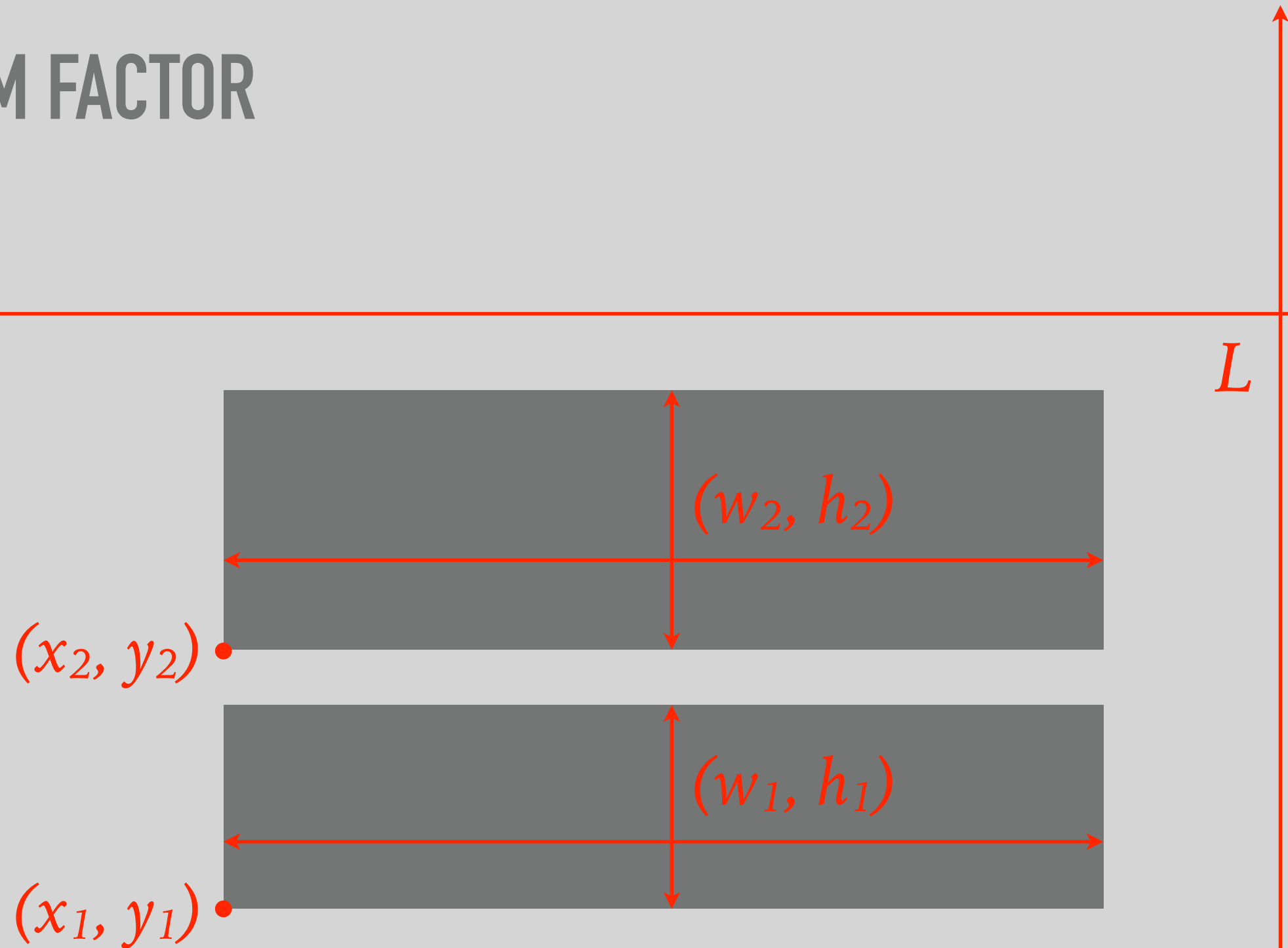
OCCAM FACTOR



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

Model-1: one rectangle.

OCCAM FACTOR



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

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!



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

How many rectangles?

TYPES OF LEARNING

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

[illegible][illegible]

A 6x6 grid with columns labeled x, y, z, a, b, c. The first three columns (x, y, z) are purple, and the last three (a, b, c) are red. A large red question mark is in the bottom right corner.

test

TYPES OF PROBLEM

➤ Classification.

Supervised

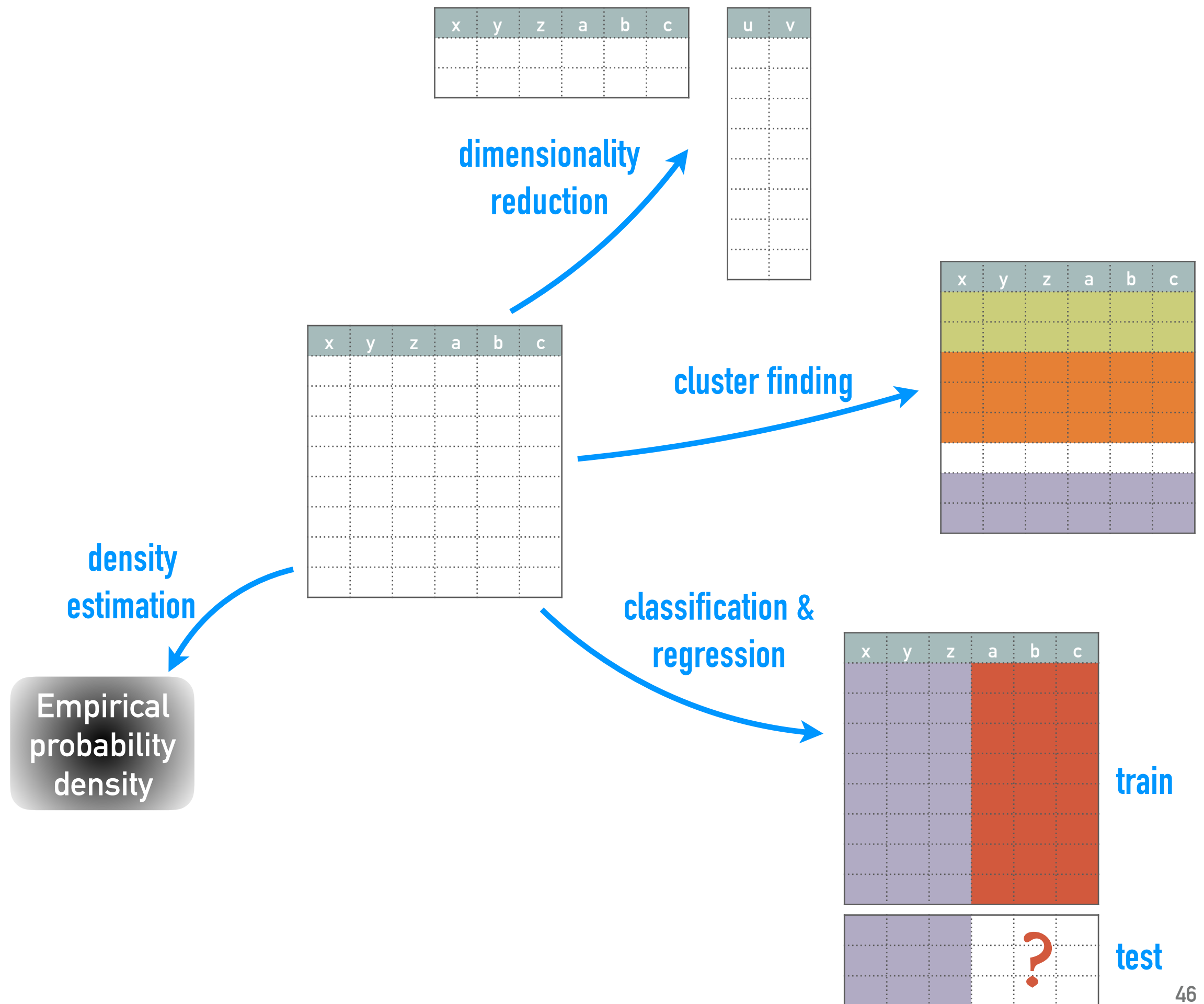
➤ Regression.

➤ Cluster finding.

➤ Density estimation.

Unsupervised

➤ Dimensionality reduction.



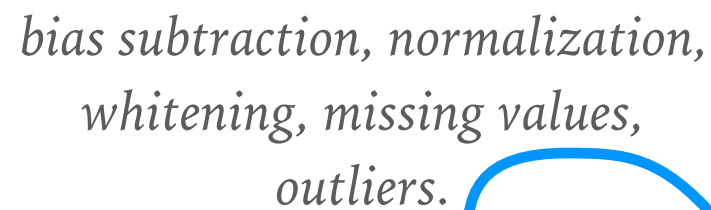
GROUP ACTIVITY: REASONING ABOUT ML PROBLEMS

- 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



preprocessing

dimensionality reduction

cluster finding

density estimation

Empirical probability density

maximum
likelihood

Theoretical probability density

2-point statistics

$\xi(r)$, $P(k)$

regression & classification

train

test

RECURRING THEMES OF ML

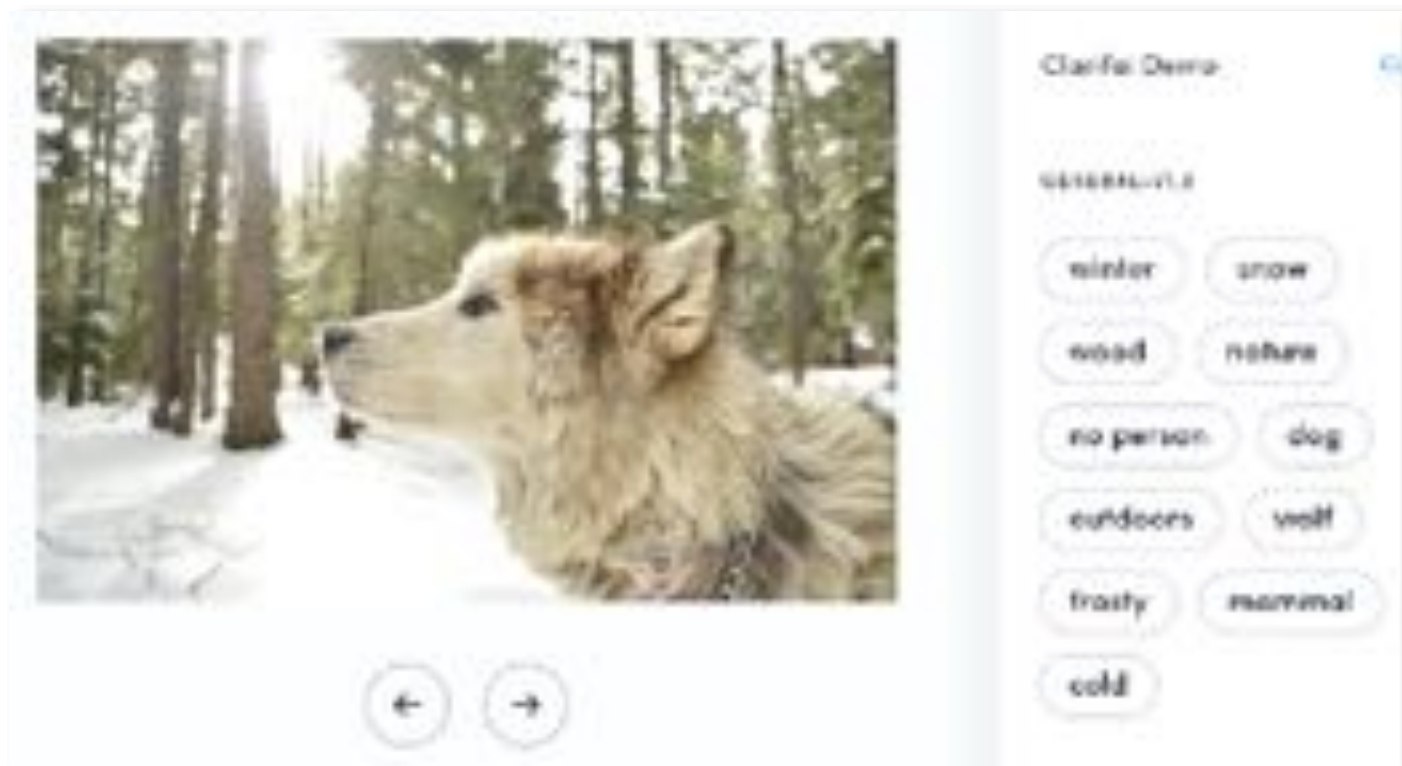
- Neighbors
- Kernels
- Ensembles
- Regularization
- Entropy

THE MACHINE-LEARNING ZOO



<http://insightdatascience.com>

BLEEDING EDGE: DEEP LEARNING



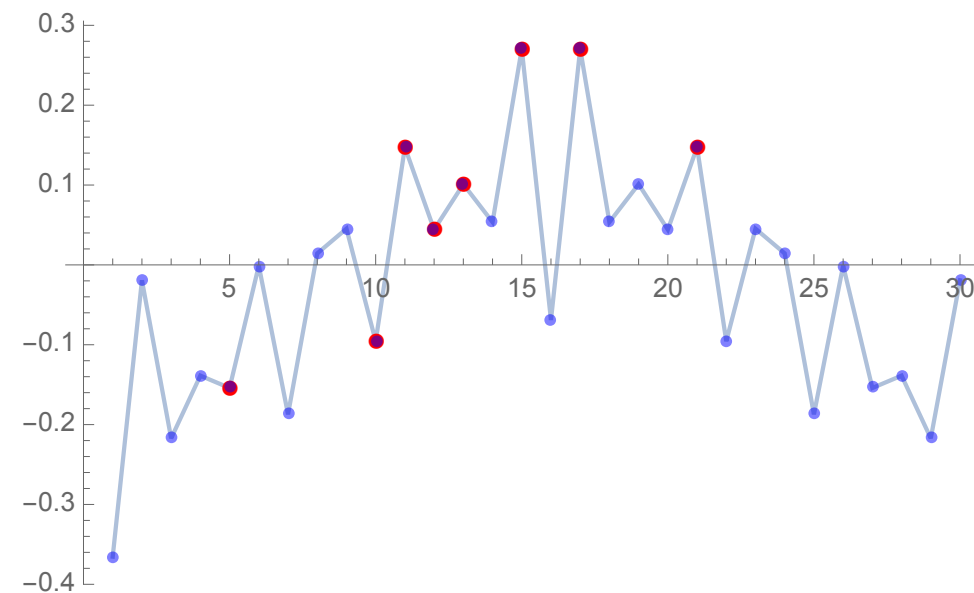
Convolutional
neural networks
for image
classification



Recurrent neural networks for natural language semantics and translation.

BLEEDING EDGE: COMPRESSIVE SENSING

An introduction to
compressive sensing



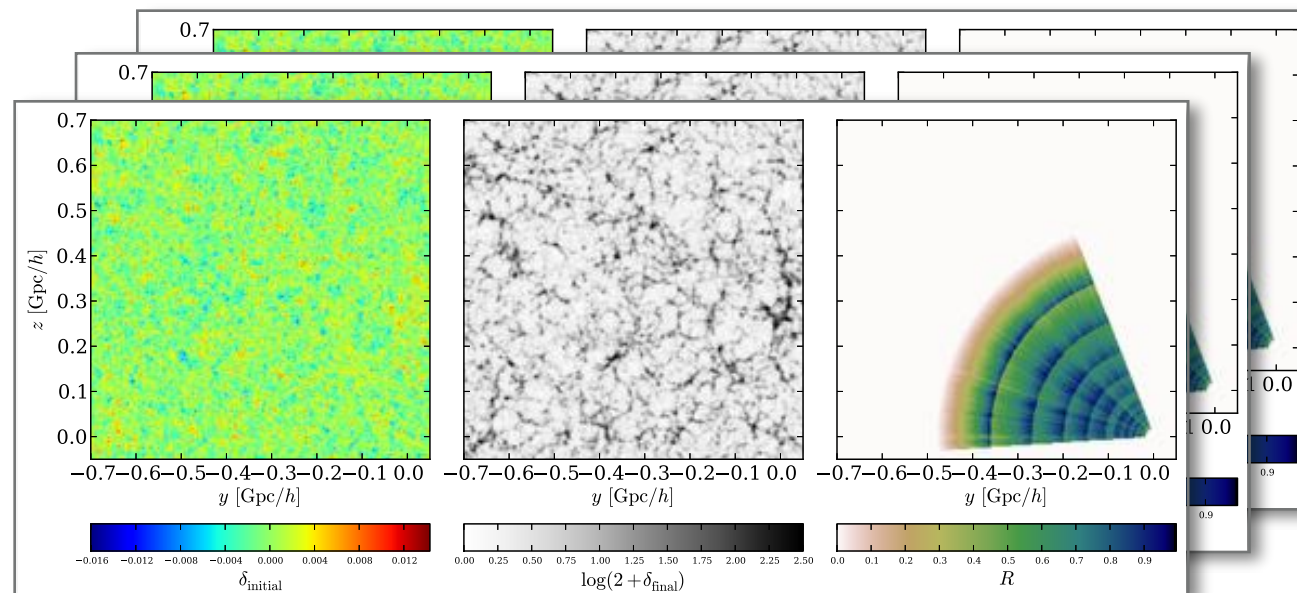
1 x 64Kpix



1300 x 1pix

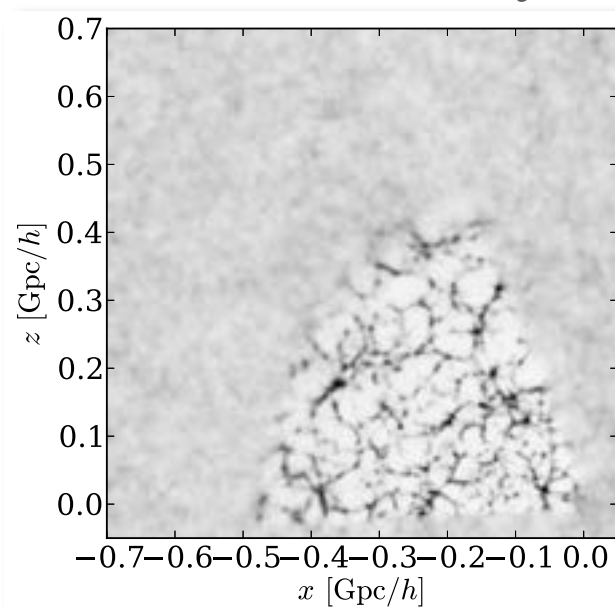
Compressive Imaging:
A New Single-Pixel Camera

BLEEDING EDGE: HAMILTONIAN MONTE CARLO

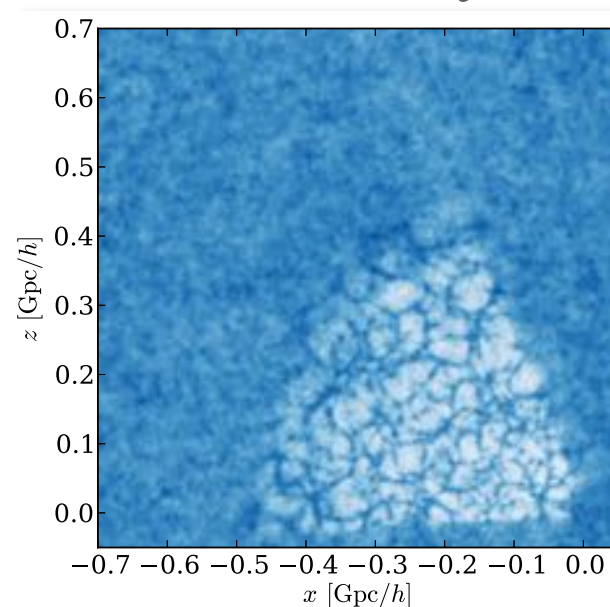


*Past and present cosmic structure
in the SDSS DR7 main sample*

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.