Social Data Mining Introduction

Grzegorz Chrupała
@gchrupala

Instructors

- Lectures: Grzegorz Chrupała
  - Computational linguistics, applied machine learning
  - http://grzegorz.chrupala.me

- Practicals: Chris Emmery
  - Social network mining
  - https://cmry.github.io

Practical Matters

Lectures

- Attendance is expected
- Slides are not meant to be self-contained
  - Take notes!

Course forum

- Subscribe to the course forum on BlackBoard
- Ask any question regarding course content and organization
- Try to answer fellow students' questions
- Chris and me will be monitoring the forum

Assessment

- Final exam
- In-class tests

SDM

- Beginner: no prerequisites
- Broader
- Less technical detail
- Practicals with Orange

ML

- Programming in Python
- More focused
- More technical detail
- Practicals with Python

Overlap in content

Don't try to follow both at the same time

What is Data Mining

"Data mining is the **computational** process of discovering patterns in **large data sets** involving methods at the intersection of **artificial intelligence**, **machine learning**, **statistics**, and **database** systems."

https://en.wikipedia.org/wiki/Data_mining

Related Fields

- Statistics
  - branch of mathematics focused on data
- Machine Learning
  - branch of Computer Science studying learning from data
- Artificial Intelligence
  - Interdisciplinary field aiming to develop intelligent machines

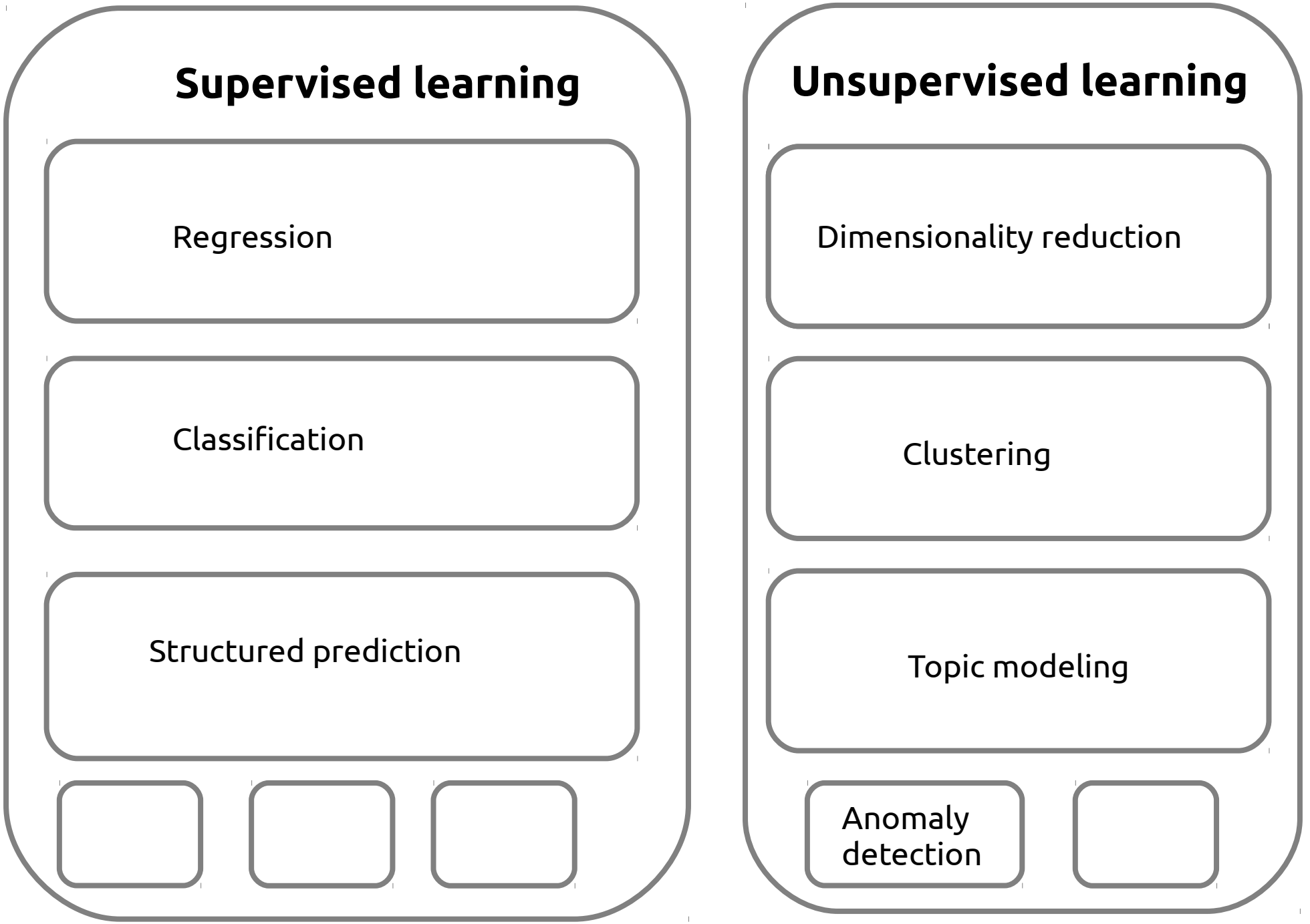
Key aspects

- Computation
- Large data sets

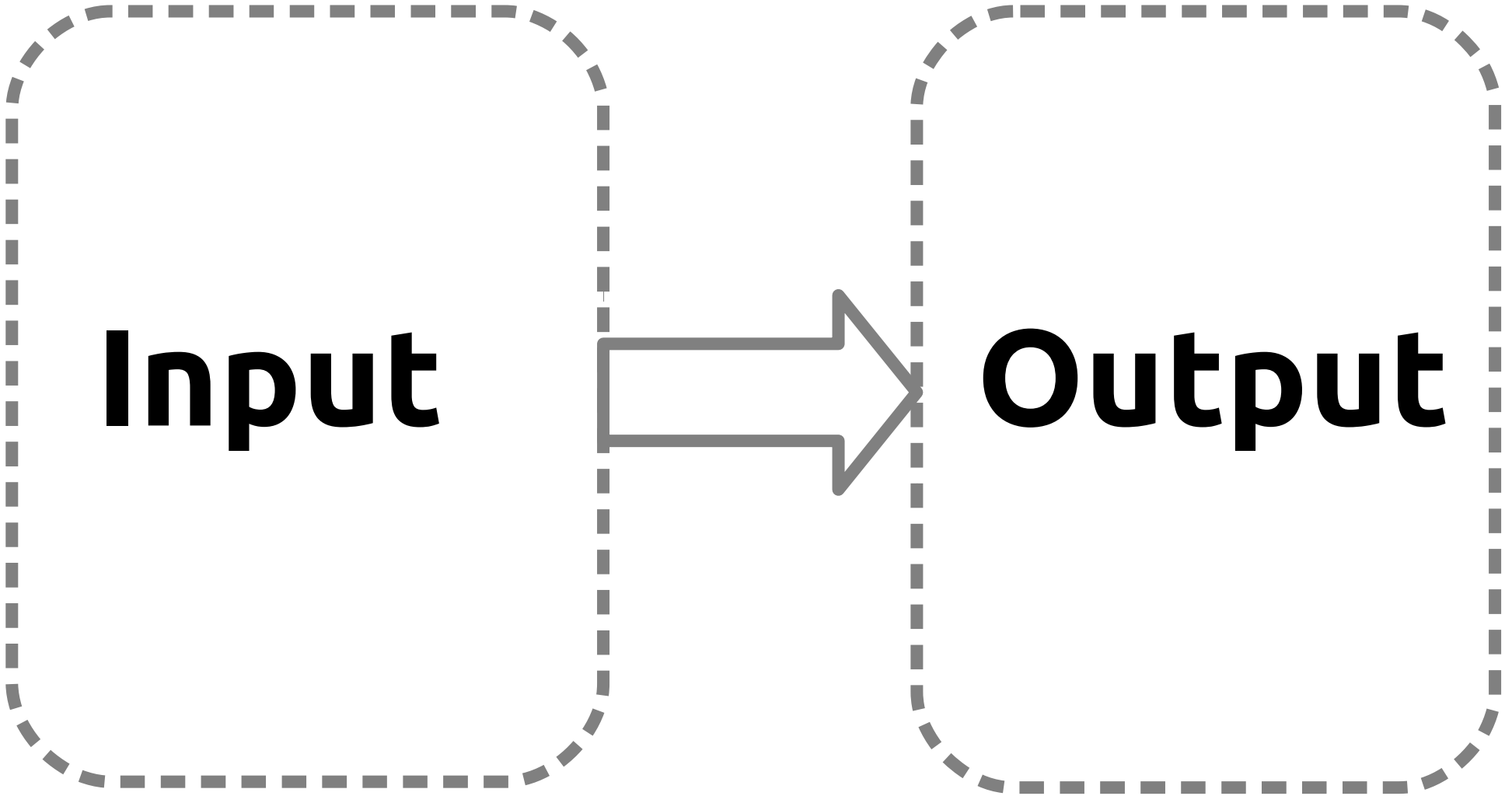
Computation enables analysis of large data sets

How large is large?

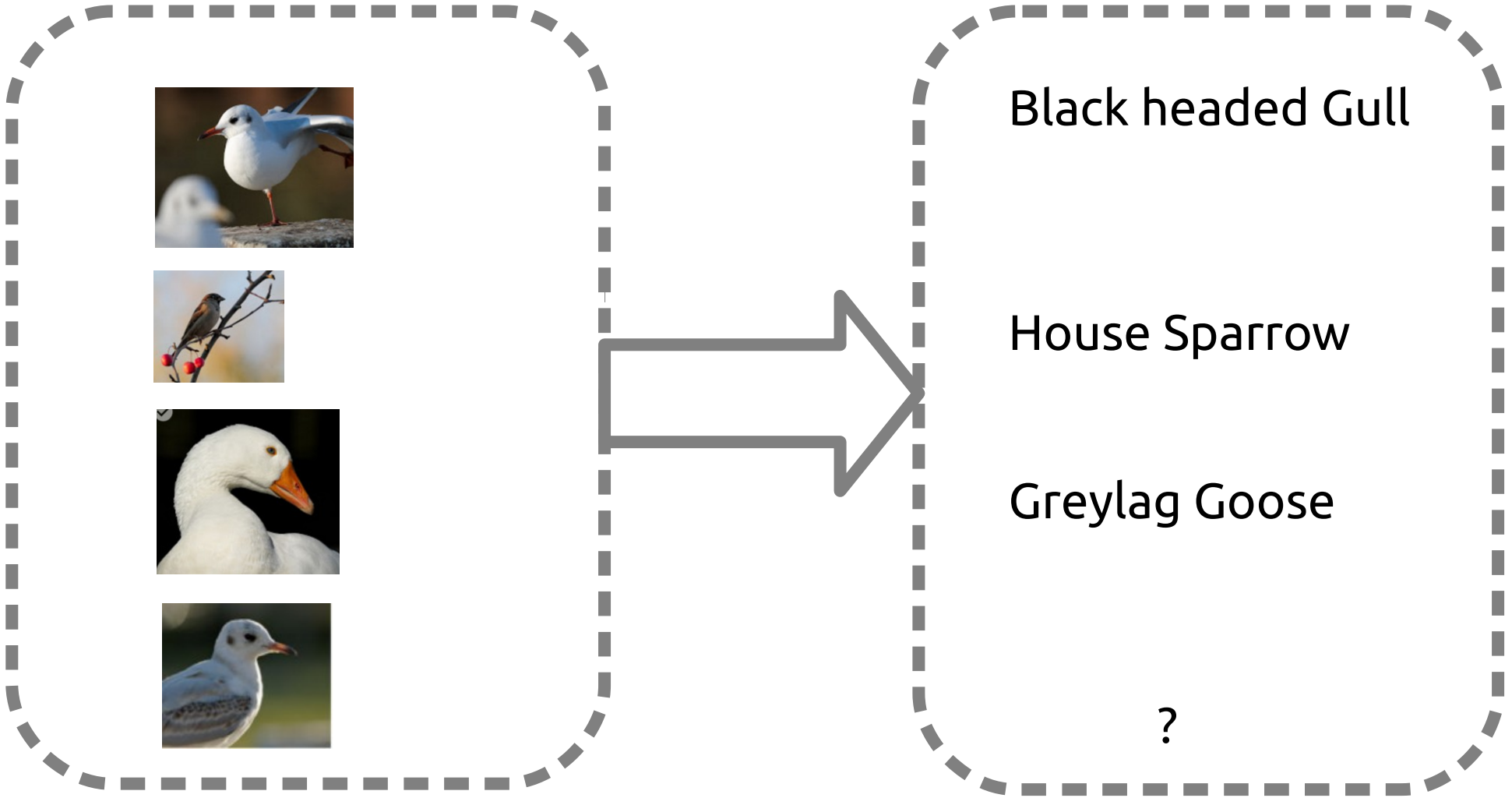
1)Too big for manual analysis
2)Too big to fit in RAM
3)Too big to store on disk
...


Supervised learning
Regression
Classification
Structured prediction
Unsupervised learning
Dimensionality reduction
Clustering
Topic modeling
Anomaly detection

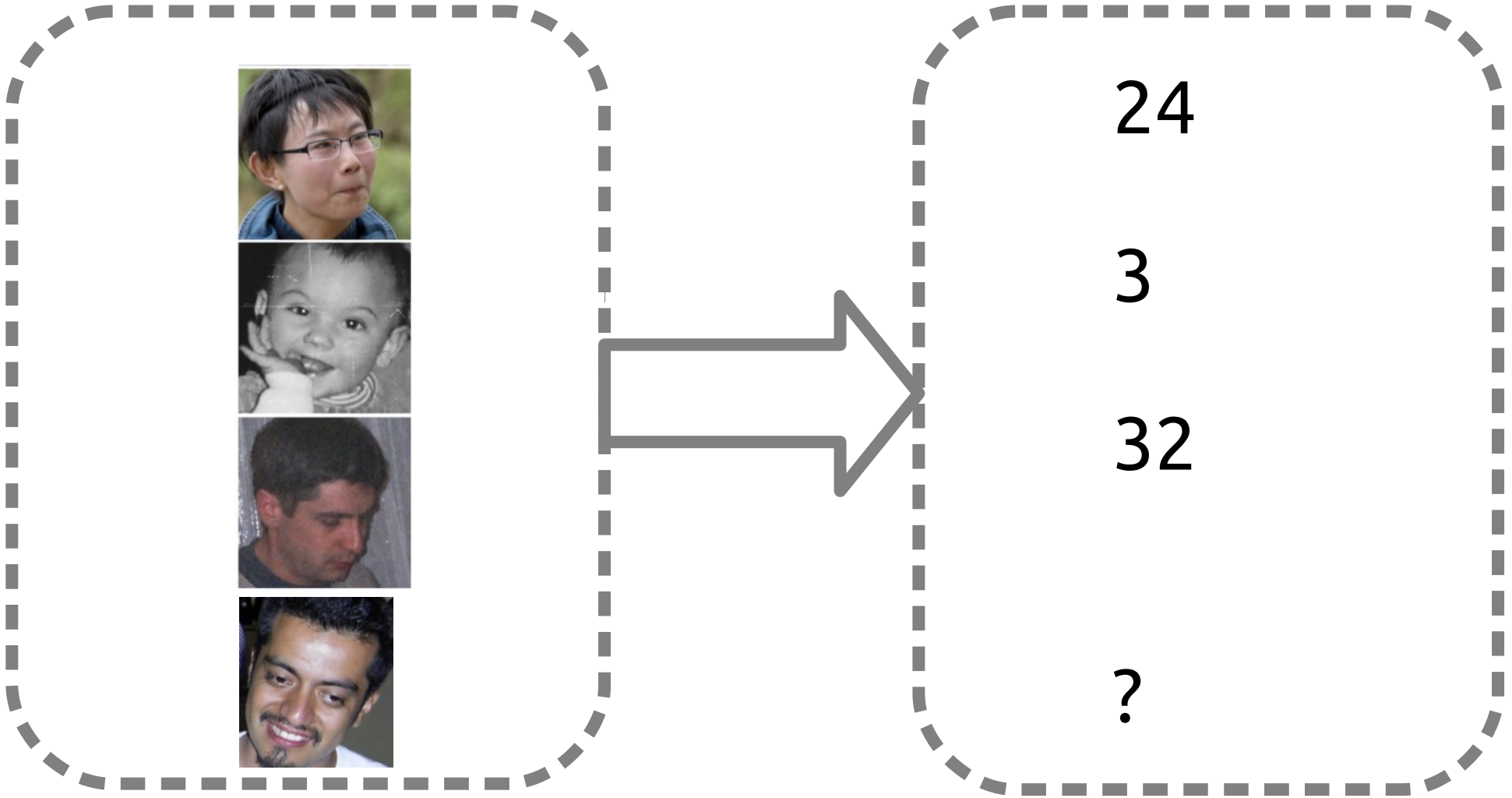
Supervised learning



Classification



Regression



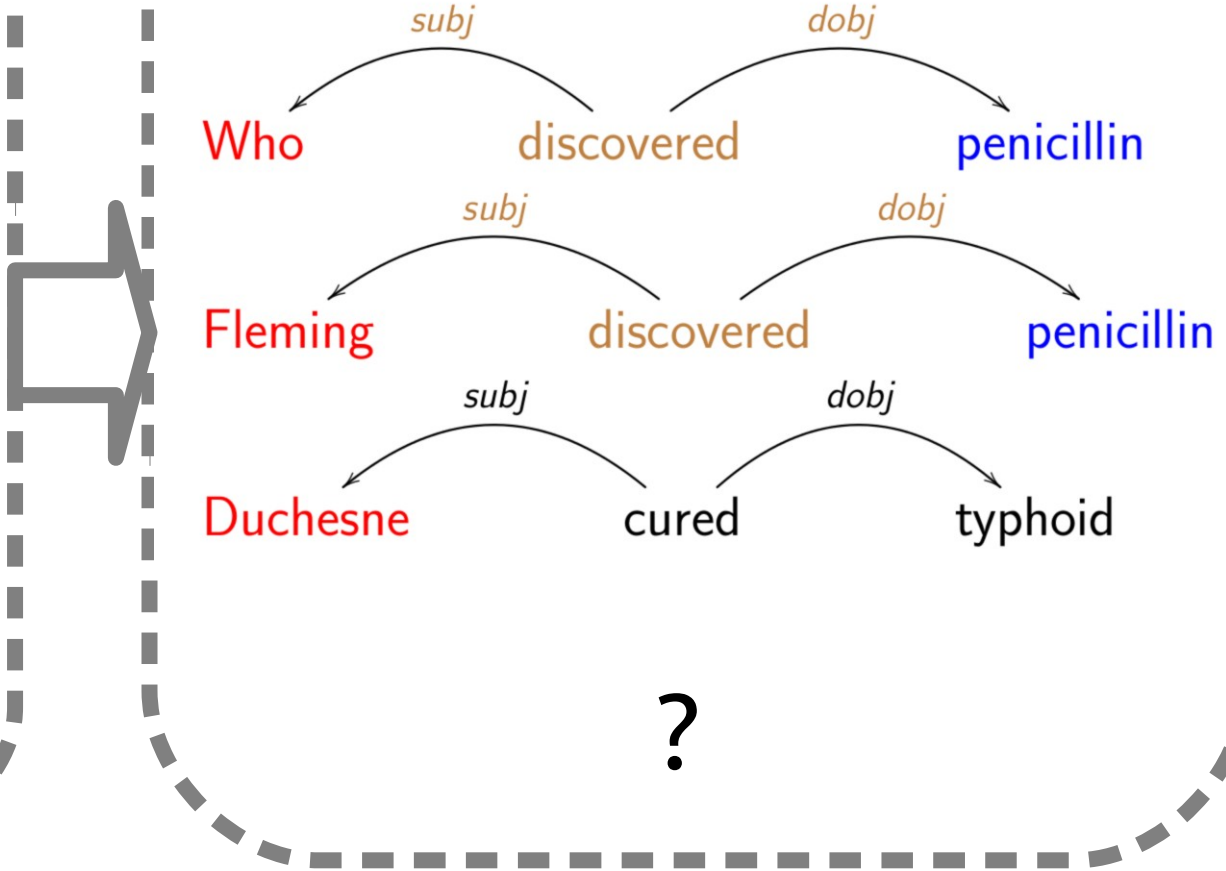
Structured prediction

Who discovered penicillin

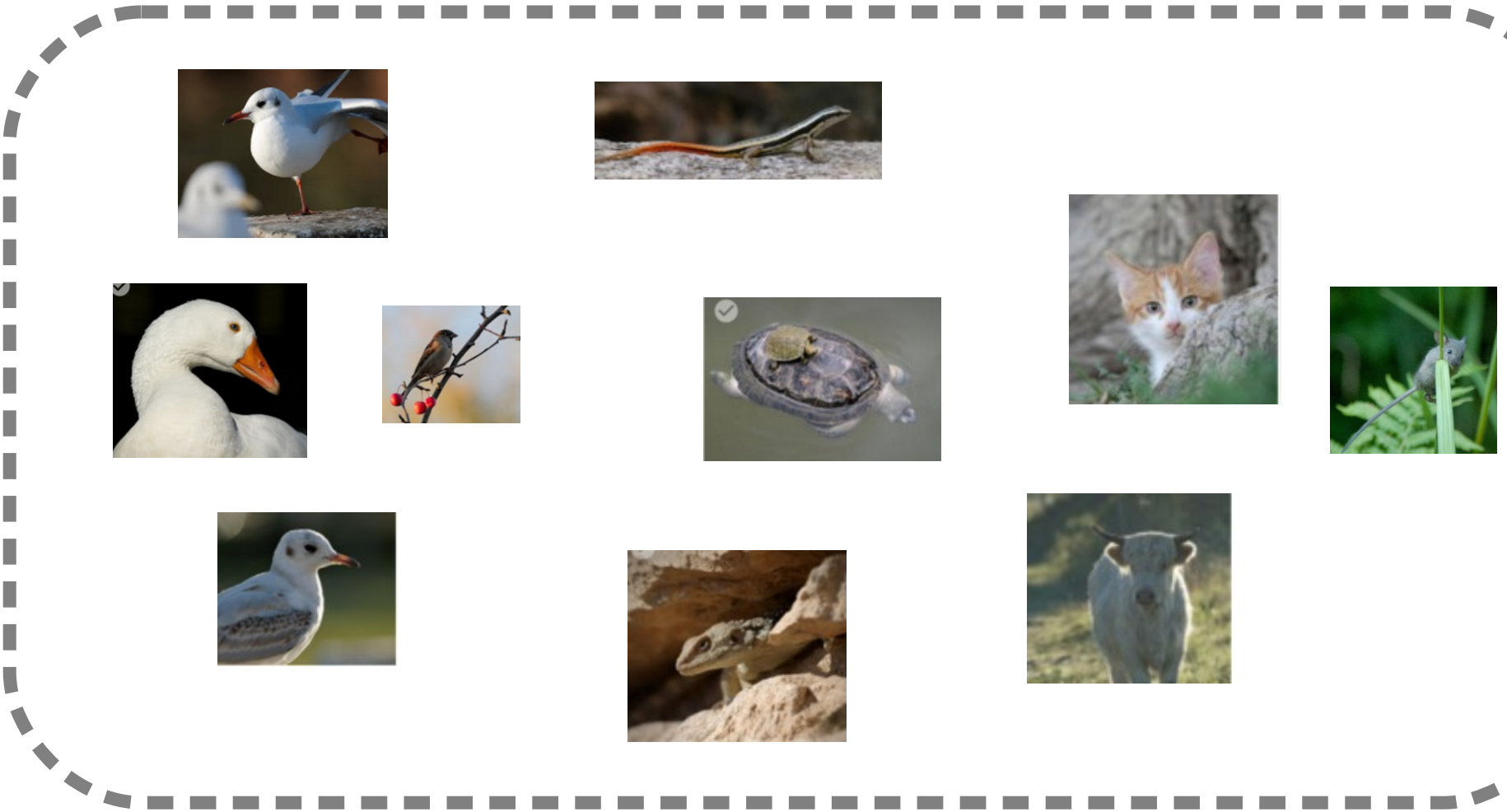
Fleming discovered penicillin

Duchesne cured typhoid

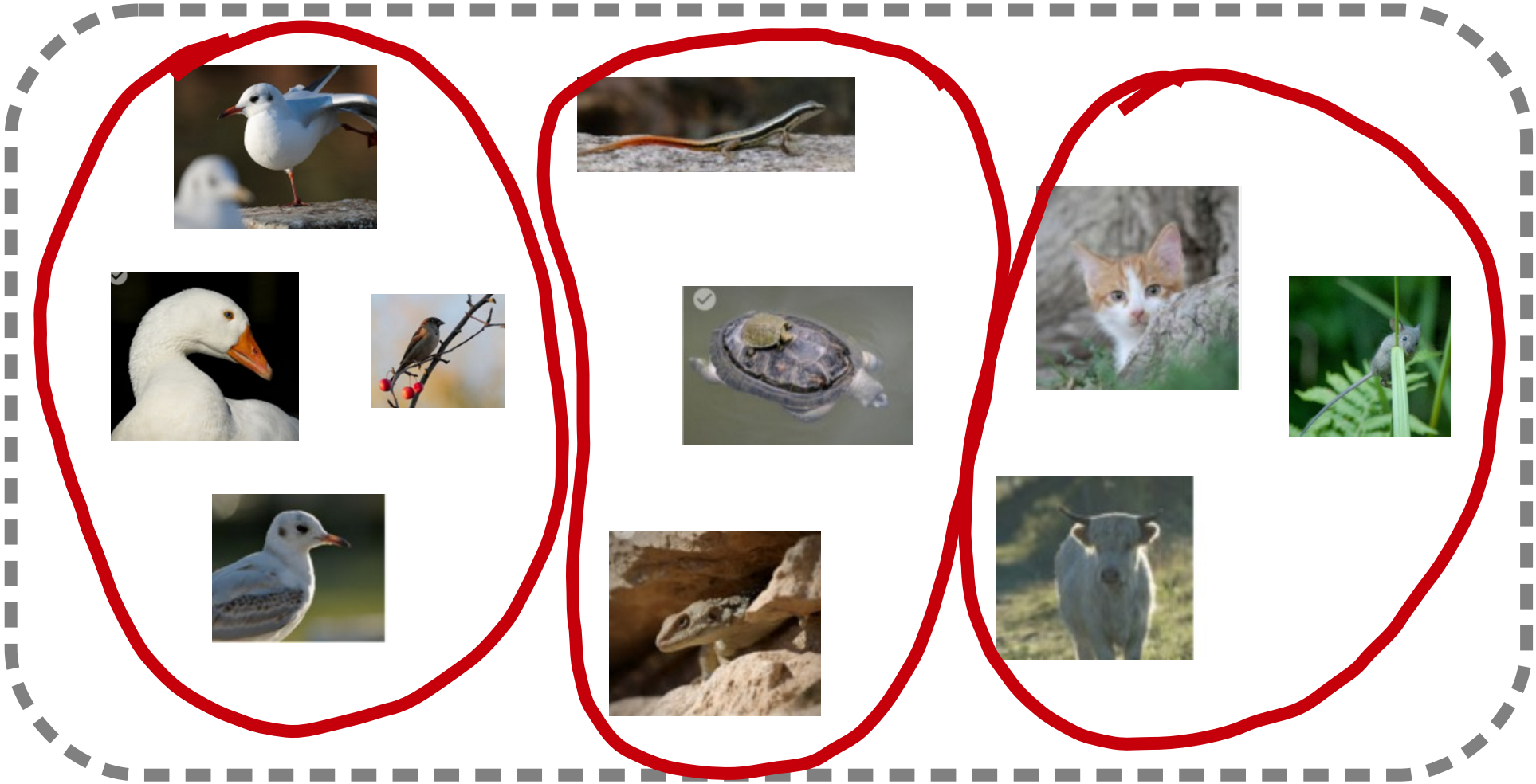
Penicillin kills bacteria



Clustering



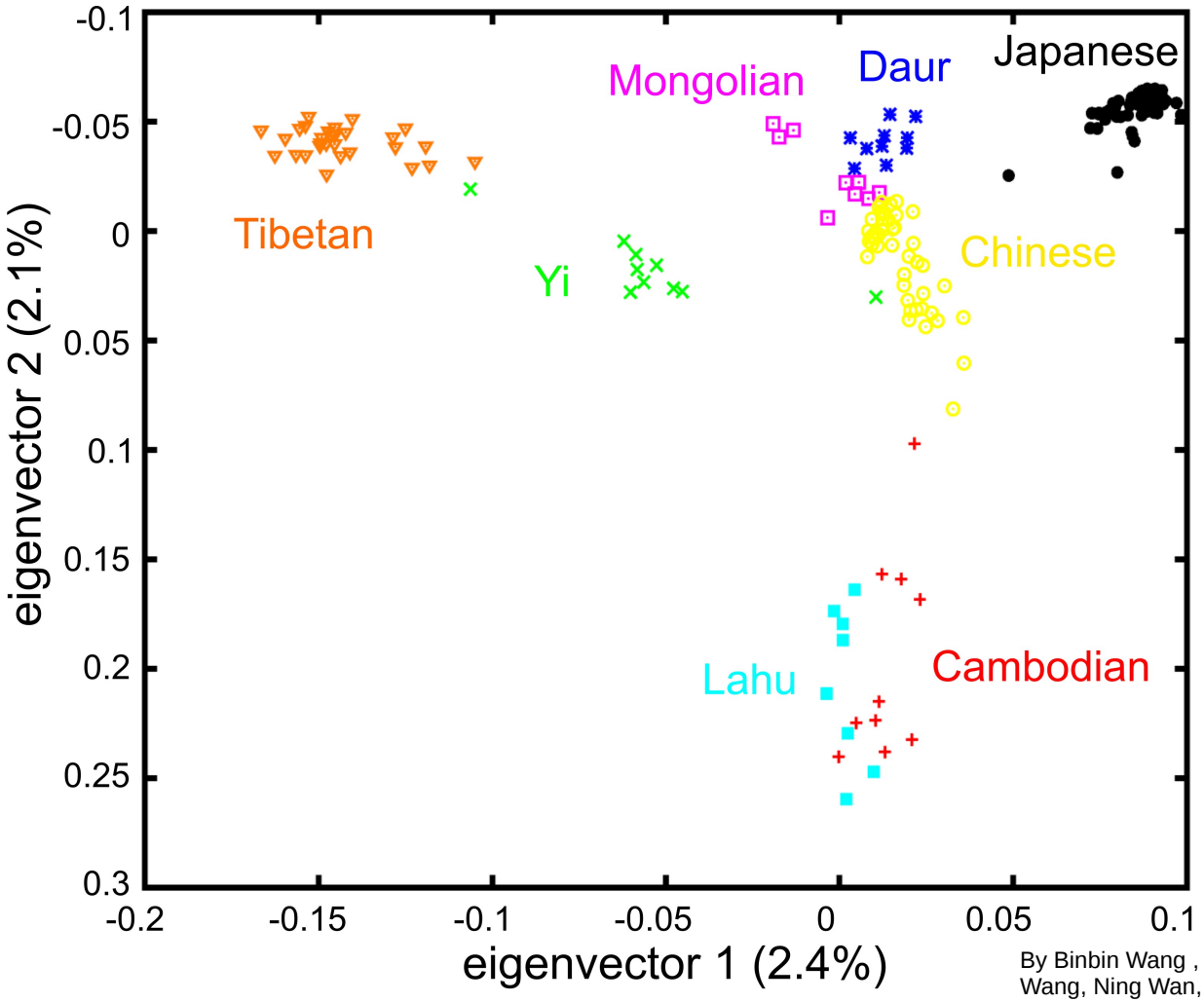
Clustering



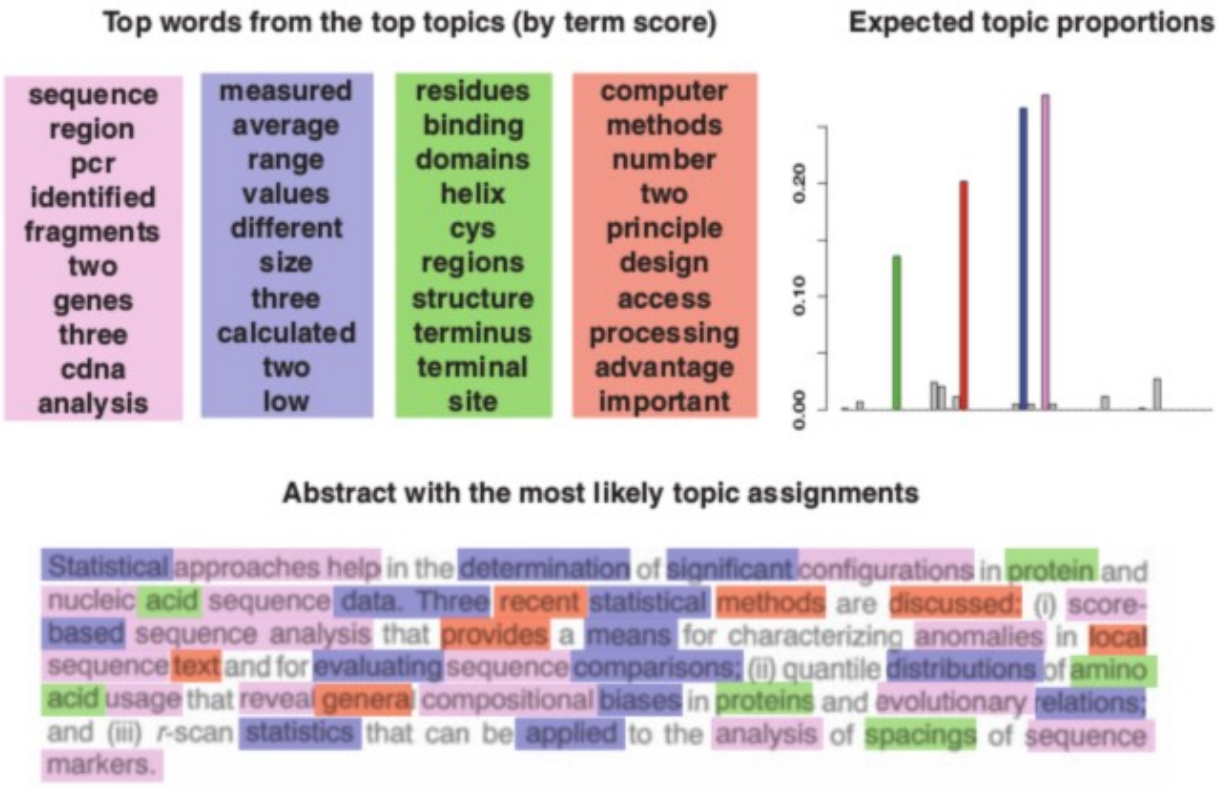
Anomaly detection

Anomaly detection

Dimensionality reduction



Topic modeling



Blei, D. M., & Lafferty, J. D. (2009). Topic models. Text mining: classification, clustering, and applications, 10(71), 34. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.463.1205&rep=rep1&type=pdf#page=96

Supervised learning Workflow

1. Collect data
2. Label examples
3. Choose example representation
4. Train model(s)
5. Evaluate

1. Collect data

- How do you select your sample?
- Reliability of measurement
- Privacy and other regulations

2. Label examples

- Annotation guidelines
- Measure inter-annotator agreement
- Crowdsourcing?

3. Representation

- Features – attributes describing examples
  - Numerical
  - Categorical
- Possibly convert to feature **vectors**

Feature vectors

- A vector is a fixed-size list of numbers
- Some learning algorithms require examples represented as vectors

$$[123, 189, 5, 123, 232, \ldots]$$

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2

aardvark	Danish	Dutch	it	its	lamb	language	minding	politics	stays	stops	…
0	0	1	0	1	0	1	1	1	0	1	

4. Train

- Keep some examples for final evaluation: test set
- Use the rest for
  - Learning: training set
  - Tuning: validation set

Tuning

- Learning algorithms can have settings (aka **hyperparameters**)
- For each value of hyperparam:
  - Apply algo to training set to learn
  - Check performance on validation set
- Choose best-performing setting

5. Evaluate

Check performance of tuned model on test set

Goal: estimate how well your model will do in the **real world**.

Keep evaluation realistic.

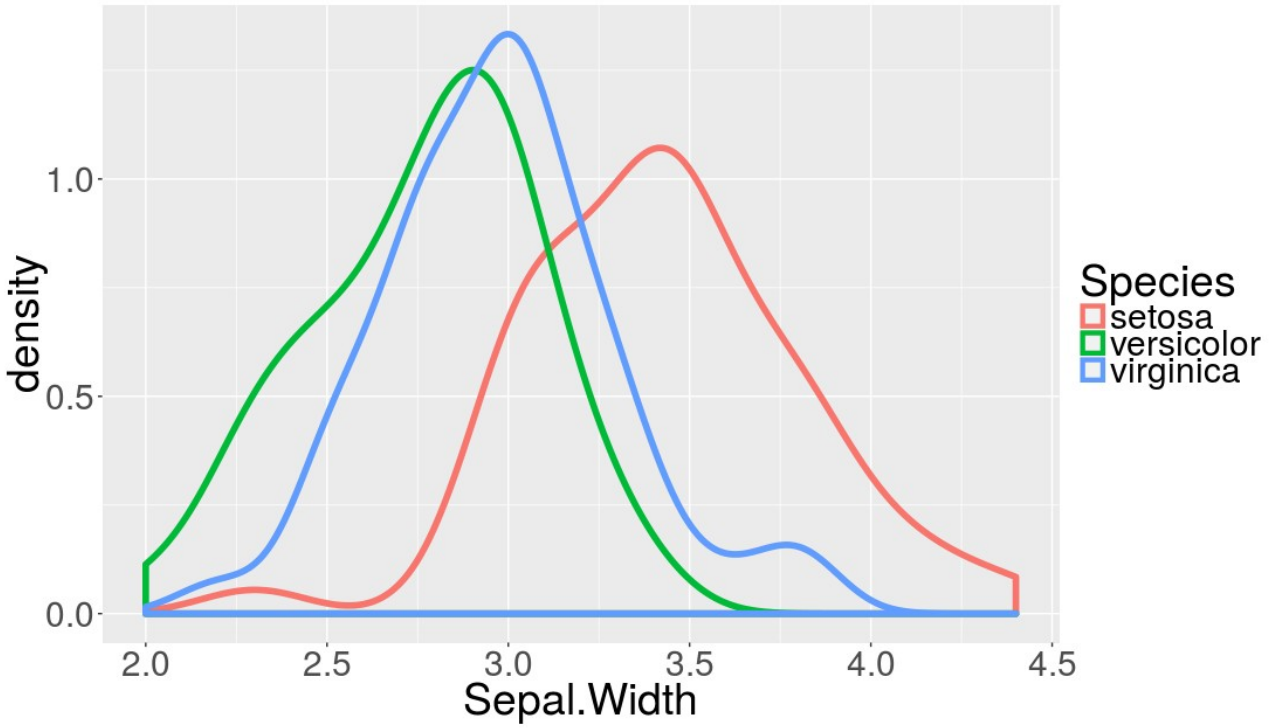
Learning to predict

- What makes prediction possible?
- Associations between features
  - Numerical: correlation coefficient
  - Categorical: mutual entropy
- Value of x_1 contains information about value of x_2

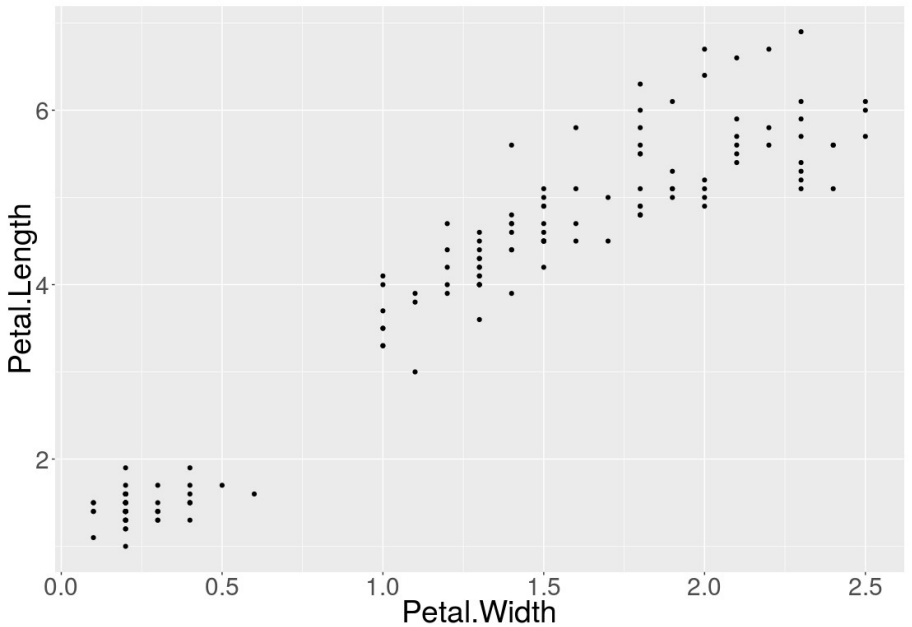
Iris

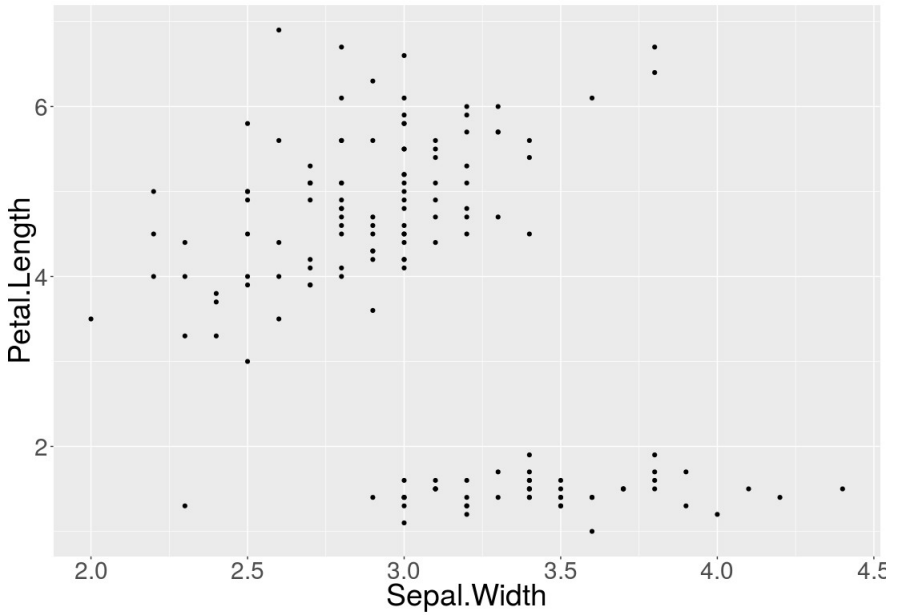
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
94	6.0	2.9	4.5	1.5	versicolor
95	5.4	3.0	4.5	1.5	versicolor
96	6.7	3.1	4.7	1.5	versicolor
97	6.0	2.2	5.0	1.5	virginica
98	6.3	2.8	5.1	1.5	virginica
99	6.3	3.3	4.7	1.6	versicolor

Iris: Sepal Width vs Species



Iris: Petal Length





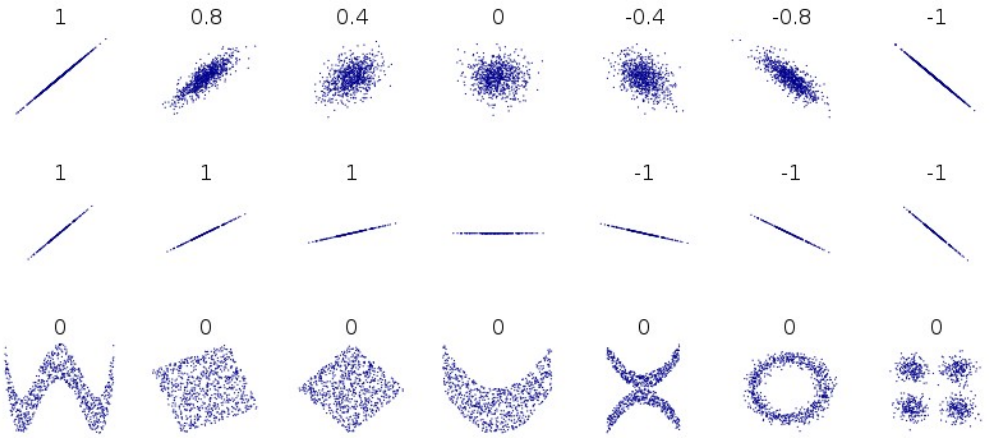
Pearson's correlation coefficient

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Numerator: **covariance**. To what extent the features change together.
- Denominator: **product of standard deviations.** Makes correlations independent of units.

Caveats

- Pearson's *r* only measures **linear** dependency
  - Other types of dependency can also be used for prediction!



- **Correlation** does not imply **causation**
  - but it may still enable prediction