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 selfcontained
 - 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

ML

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

- 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."

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

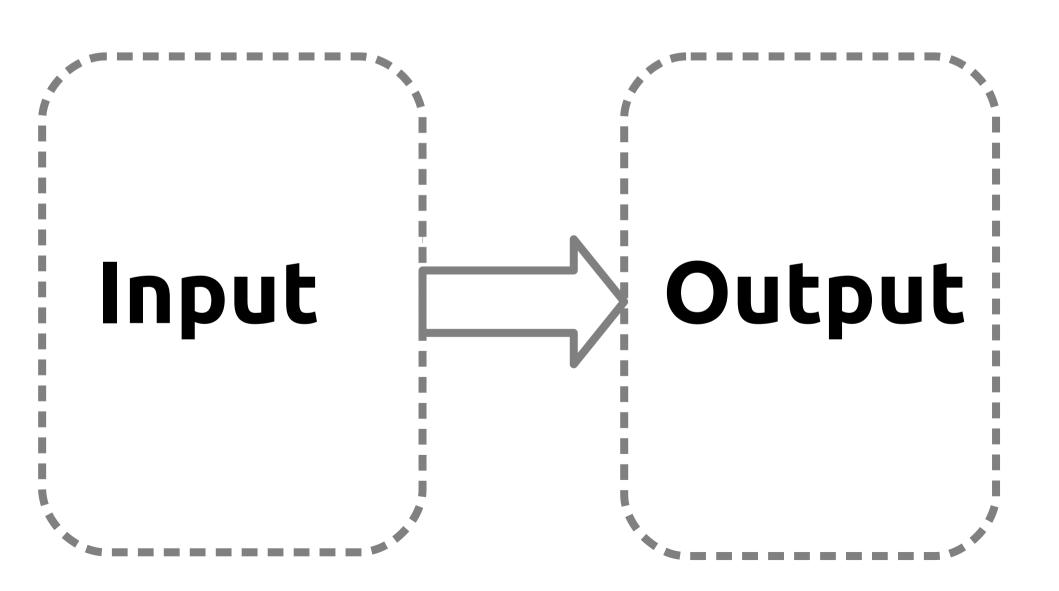
How large is large?

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

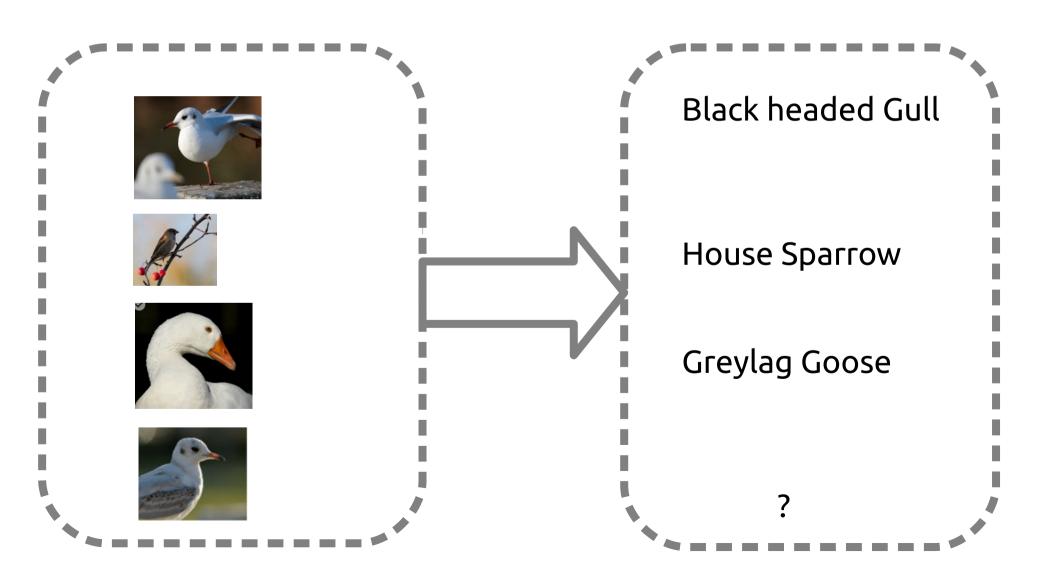
• • •

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

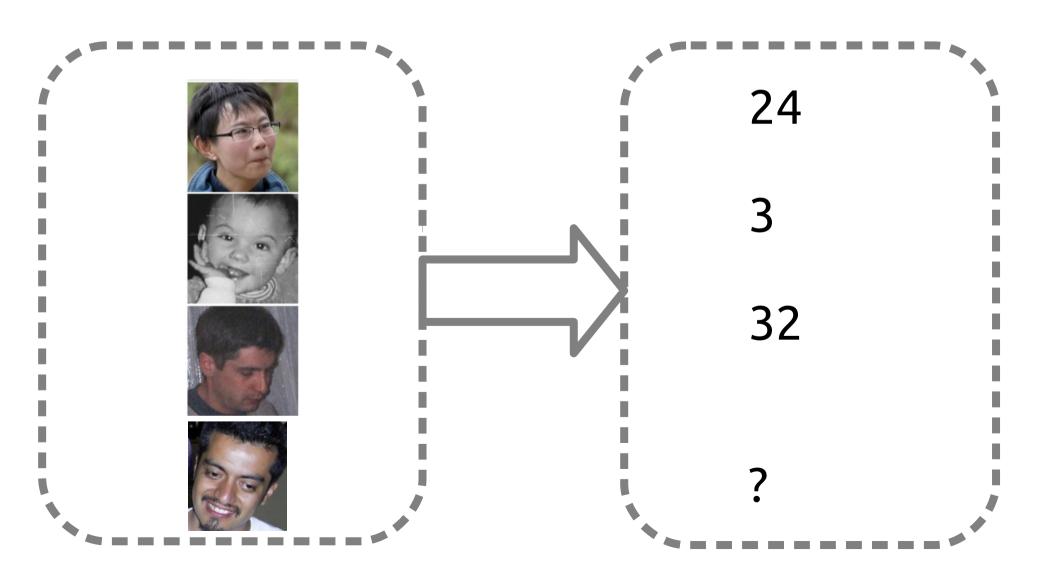
Supervised learning



Classification



Regression



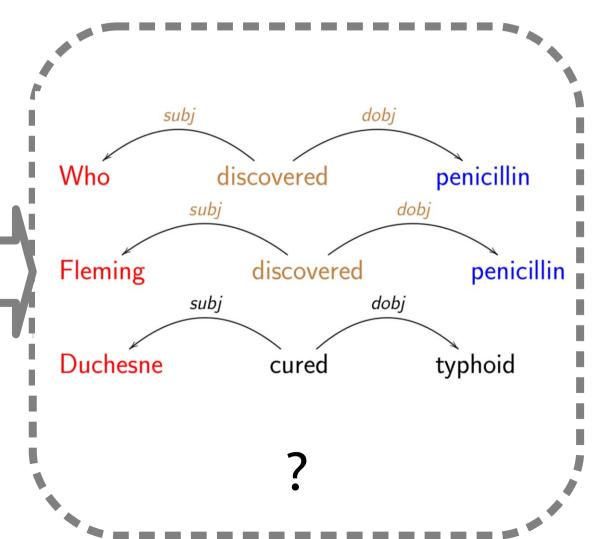
Structured prediction

Who discovered penicillin

Fleming discovered penicillin

Duchesne cured typhoid

Penicillin kills bacteria



Clustering













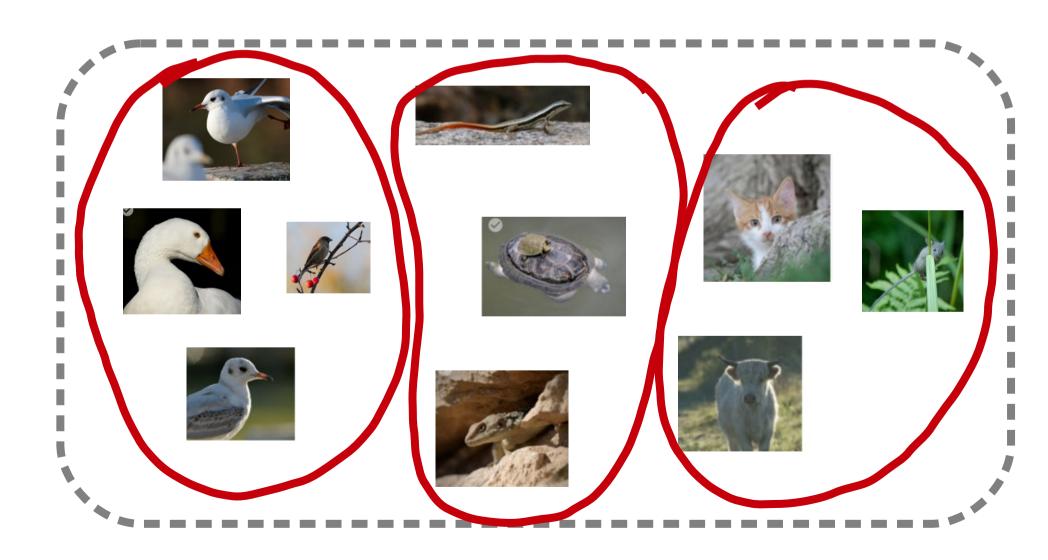




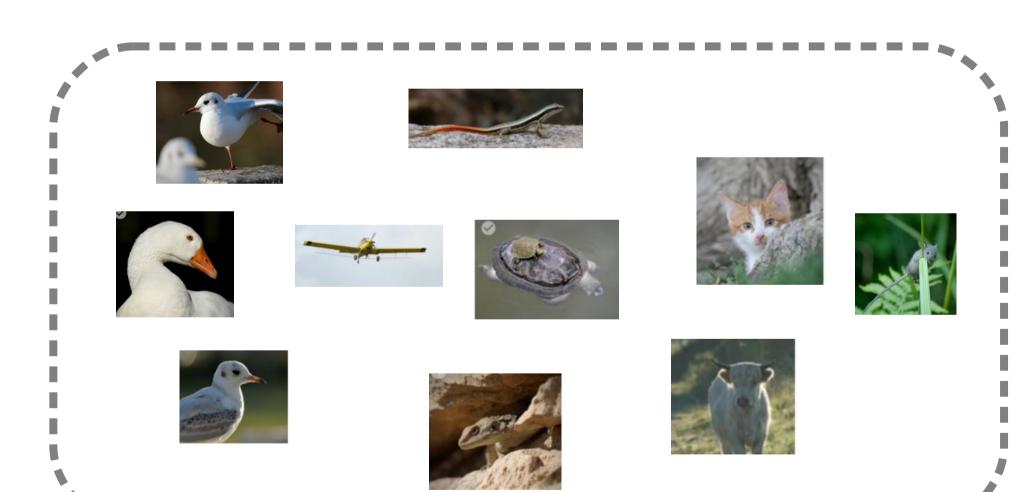




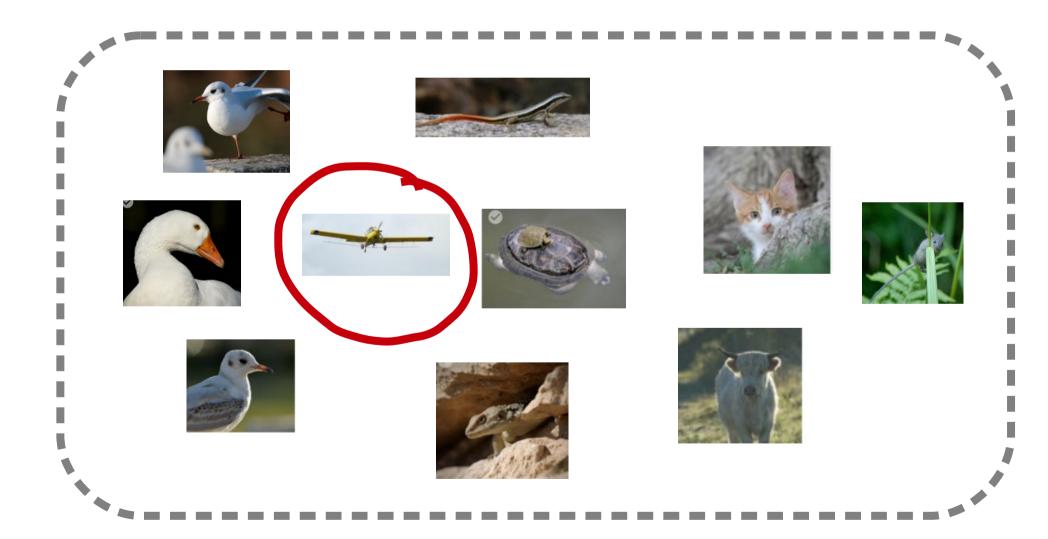
Clustering



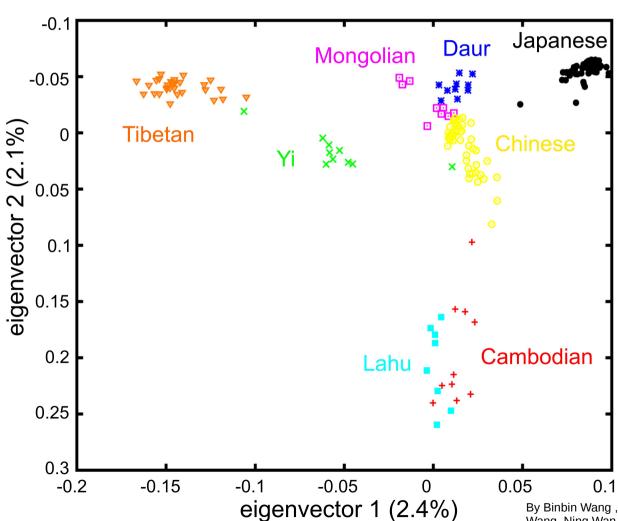
Anomaly detection



Anomaly detection



Dimensionality reduction



By Binbin Wang, Yong-Biao Zhang, Feng Zhang, Hongbin Lin, Xumin Wang, Ning Wan, Zhenqing Ye, Haiyu Weng, Lili Zhang, Xin Li, Jiangwei Yan, Panpan Wang, Tingting Wu, Longfei Cheng, Jing Wang, Duen-Mei Wang, Xu Ma, Jun Yu [CC BY 2.5 (http://creativecommons.org/licenses/by/2.5)], via Wikimedia Commons

Topic modeling

Top words from the top topics (by term score)

or and many top topics (a) term seems

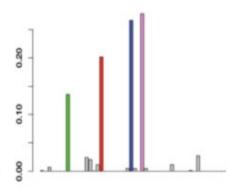
sequence region pcr identified fragments two genes three cdna analysis

measured
average
range
values
different
size
three
calculated
two
low

residues
binding
domains
helix
cys
regions
structure
terminus
terminal
site

computer
methods
number
two
principle
design
access
processing
advantage
important

Expected topic proportions



Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) r-scan statistics that can be applied to the analysis of spacings of sequence markers.

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

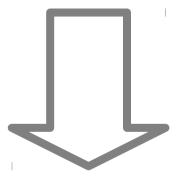
- A vector is a fixed-size list of numbers
- Some learning algorithms require examples represented as vectors





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





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 \dots 0 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.

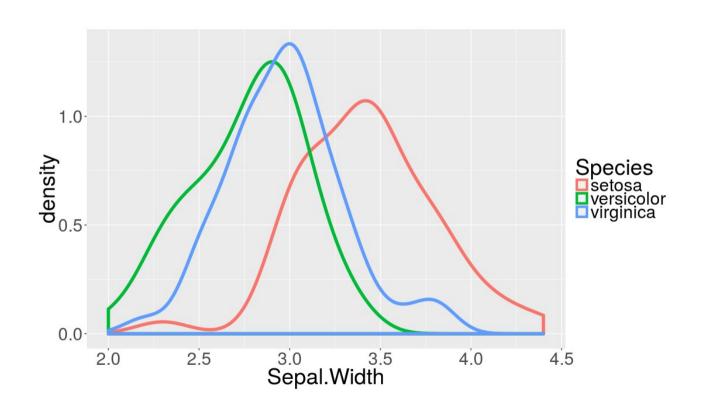
Learning to predict

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

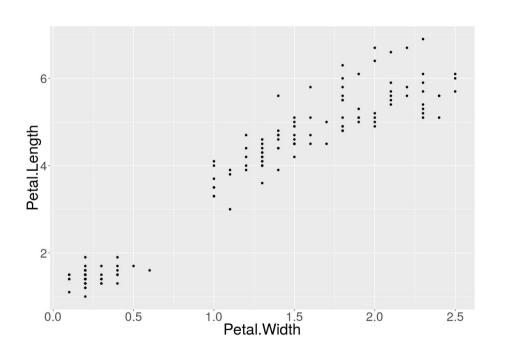


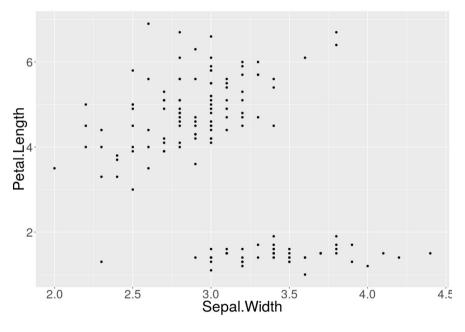
Species	Petal.Width	Petal.Length	Sepal.Width	Sepal.Length	
versicolor	1.5	4.5	2.9	6.0	94
versicolor	1.5	4.5	3.0	5.4	95
versicolor	1.5	4.7	3.1	6.7	96
virginica	1.5	5.0	2.2	6.0	97
virginica	1.5	5.1	2.8	6.3	98
versicolor	1.6	4.7	3.3	6.3	99

Iris: Sepal Width vs Species



Iris: Petal Length





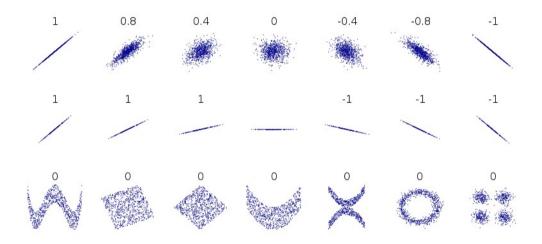
Pearson's correlation coefficient

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

- 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