6.862 Applied Machine Learning: Introduction

Feb 10, 2020

Welcome to 6.862!

- Idea: use ML in your research while learning it
- This class is open to your creativity to implement an interesting project. So, **have fun**!

Today:

- What is this class about?
- Organization
- Some simple basics

Your Team

Stefanie Jegelka
stefje@mit.edu

Andrew Rouditchenko
roudi@mit.edu

Nathan Hunt
nhunt@mit.edu

Overview

- class entails all of 6.036 (70%) plus long project (30%)
- **office hours** (mandatory):
  - regularly meet with instructors in office hours to discuss progress/hurdles
  - sign up for individual slots, **meet once in each Phase**

now-Mar 1	**Phase I** *(3 weeks)*
02/17	**pre-proposal due**
02/27	**full project proposal due**
Mar 2 - Apr 4	**Phase II** *(4 weeks + spring break)*
04/02	**intermediate reports due**
Apr 5 - Apr 26	**Phase III** *(3 weeks)*
Apr 27 - May 7	**Phase IV: meetings optional** *(2 weeks)*
May 7	**final project report due**

Pre-proposal

1 paragraph, 5% of grade

- title + abstract, think about this before your first OH meeting
- high level idea of project, **data** you will be using (you must have it now)
- *think about*: what question do I want to answer? What data probably has that information?

Proposal

1-2 pages, 15% of grade

- What do you want to do? What question are you answering?
- Motivation and formulation as ML problem
- What data will you use? Specific description of data
- What methods will you try and compare?
- What computational resources will you use? (time, feasibility)
- Brief summary of related work
- Project plan: at least 4 steps per team member, deadlines
- Risks: what may be more difficult than expected? Mitigation?

Intermediate Progress Report

3 pages, 20% of grade

- What have you done so far? What worked out / did not? If not, troubleshooting, alternative paths?
- timeline for the remaining time: at least 2 steps per team member
- describe the general layout of at least one plot that will appear in the final report. What are the axes? What will the reader learn from it?

Final Report

4 pages, technical content: 40%, presentation/writing/clarity 20%

- Progress and results of project: what did you do? What did you find?
- **Interpret and discuss** your results: why did results come out this way? What do they say about your research question?
- Include **evidence** for each claim you make (graphs, tables, sensitivity analysis, etc).
- Experiments reproducible

Logistics

- Lectures, labs, psets, exams: 6.036.
  Make sure you have access to the 6.036 MITx website
- Report submissions: 6.036 website
- Questions about homeworks, labs, exams, etc related to 6.036:
  6.036 Piazza / office hours
- **Sign up for one 6.862 office hour (20 minutes) in each phase.
  Starting this week.**

Formulating a project

1. **What and why?** Research question - what is the contribution?
    - what question are you trying to answer? what are you trying to predict?

Formulating a project

1. **What and why?** Research question - what is the contribution?

2. **Phrase it as a Machine Learning problem**. E.g.
   - classification
   - regression
   - clustering
   - recommender system
   - reinforcement learning …

often, there is more than one way!

Full 6.036 lecture notes will be available on Stellar, just for 6.862.
If you use methods from later parts/outside of class, read ahead and start them before they are discussed in class.

Formulating a project

1. **What and why?** Research question - what is the contribution?

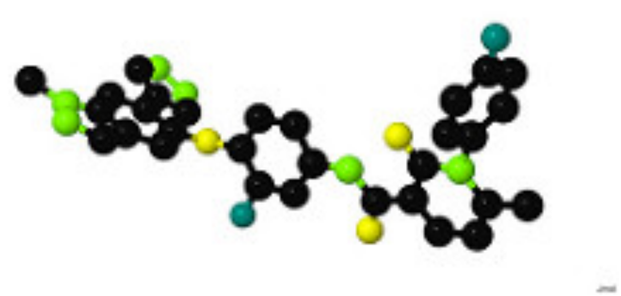
2. **Phrase it as a Machine Learning problem**.

3. **What data** do I need? What do I have?
   labels, amount, noisy, missing entries,…

4. **What methods** to try? What exists? Need to develop new ones?
   try several. Remember baselines, features, …

Supervised Learning (Classification, Regression)

- data points $x^{(1)}, \ldots, x^{(n)}$, labels $y^{(1)}, \ldots, y^{(n)}$

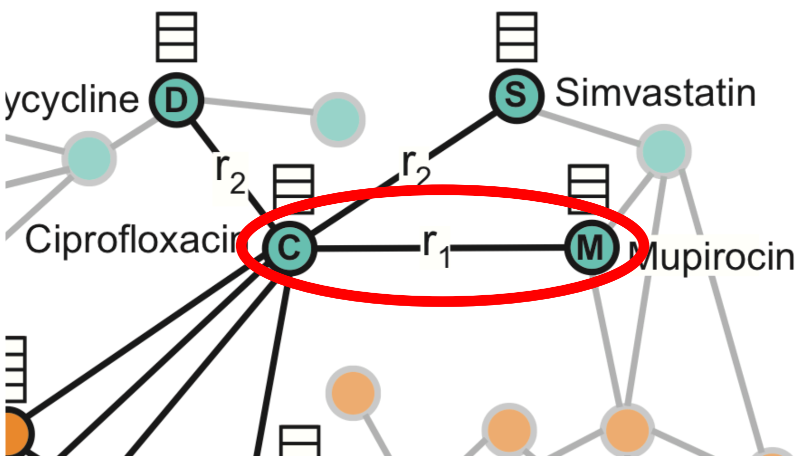
$x^{(i)} =$ 

$y^{(i)} = $ "toxic"

$y^{(i)} = \text{drug efficacy}$

Supervised Learning (Classification, Regression)

- data points $x^{(1)}, \dots, x^{(n)}$, labels $y^{(1)}, \dots, y^{(n)}$

$x^{(i)} =$ 

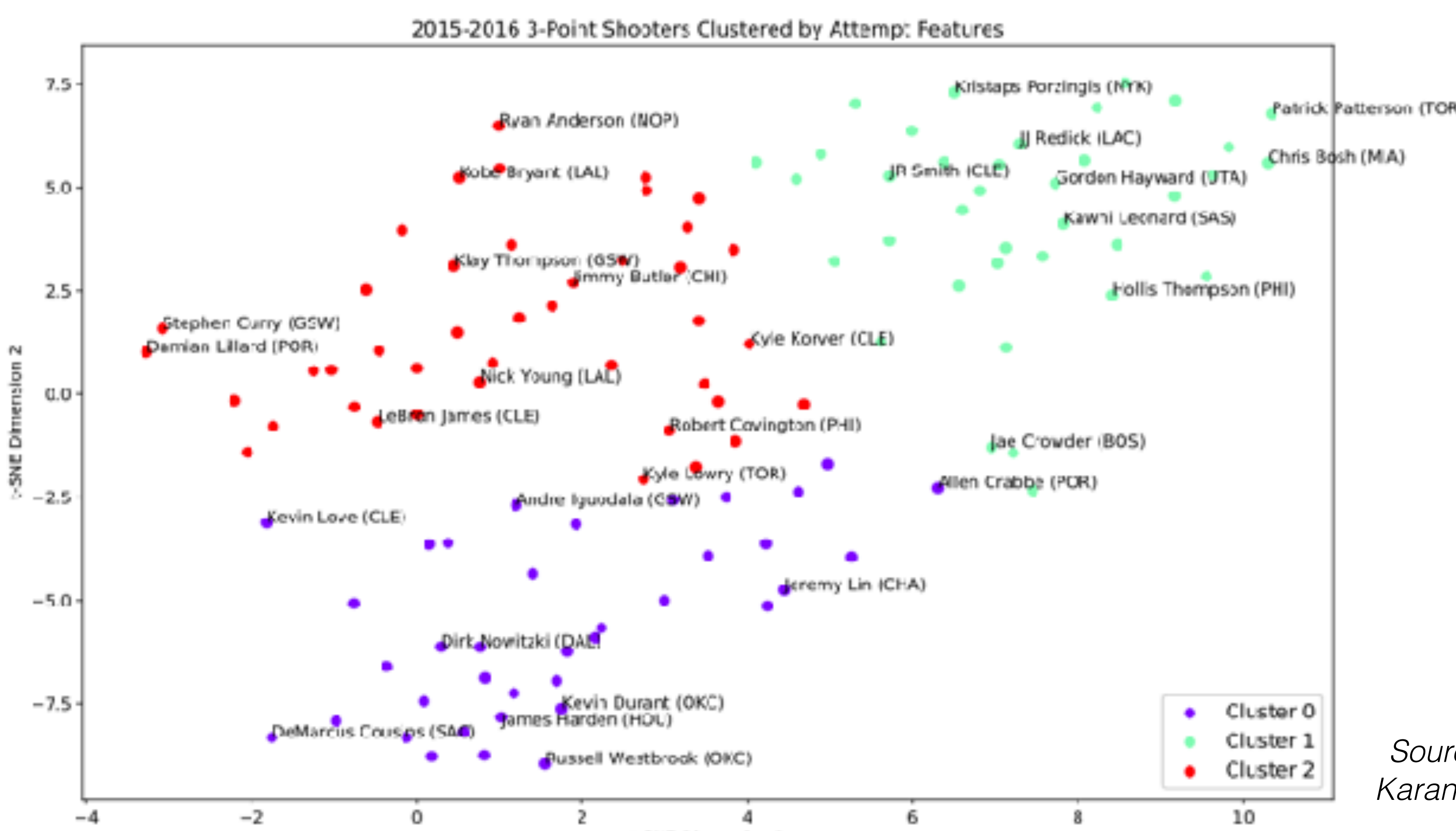
$y^{(i)} \in \{\text{edge, no edge}\}$

Clustering

- data points $x^{(1)}, \ldots, x^{(n)}$



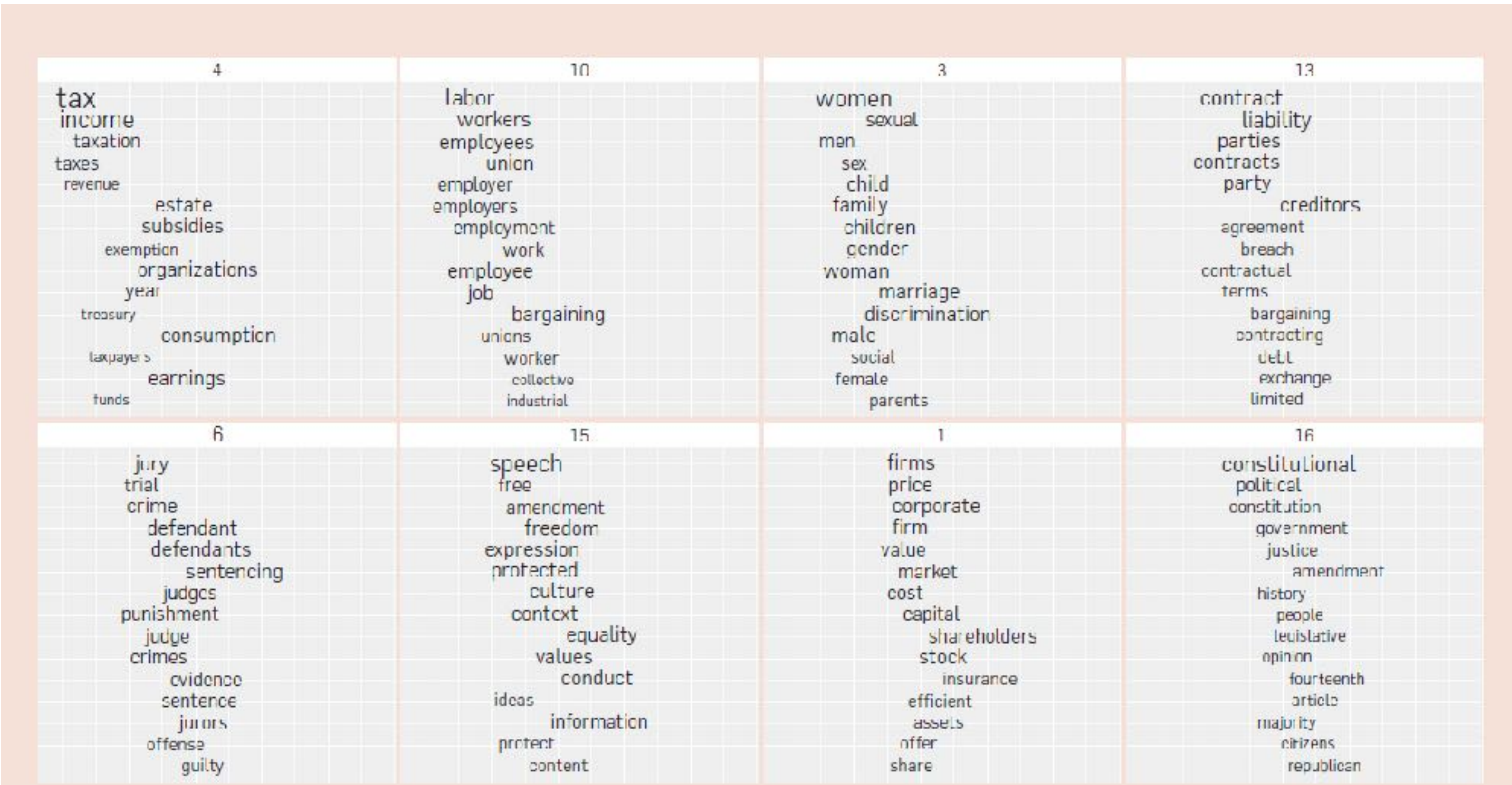


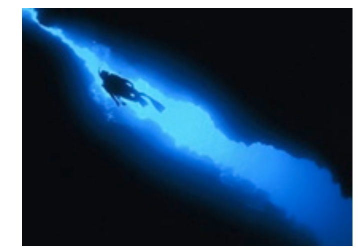
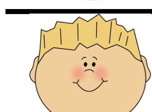
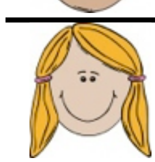
Source: IDS.012 class project by Nate Bailey, Karan Bhuwalka, Hin Lee, Tim Zhong

Feature Learning / Topic Modeling

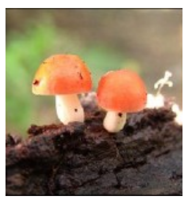
- data points $x^{(1)}, \ldots, x^{(n)}$

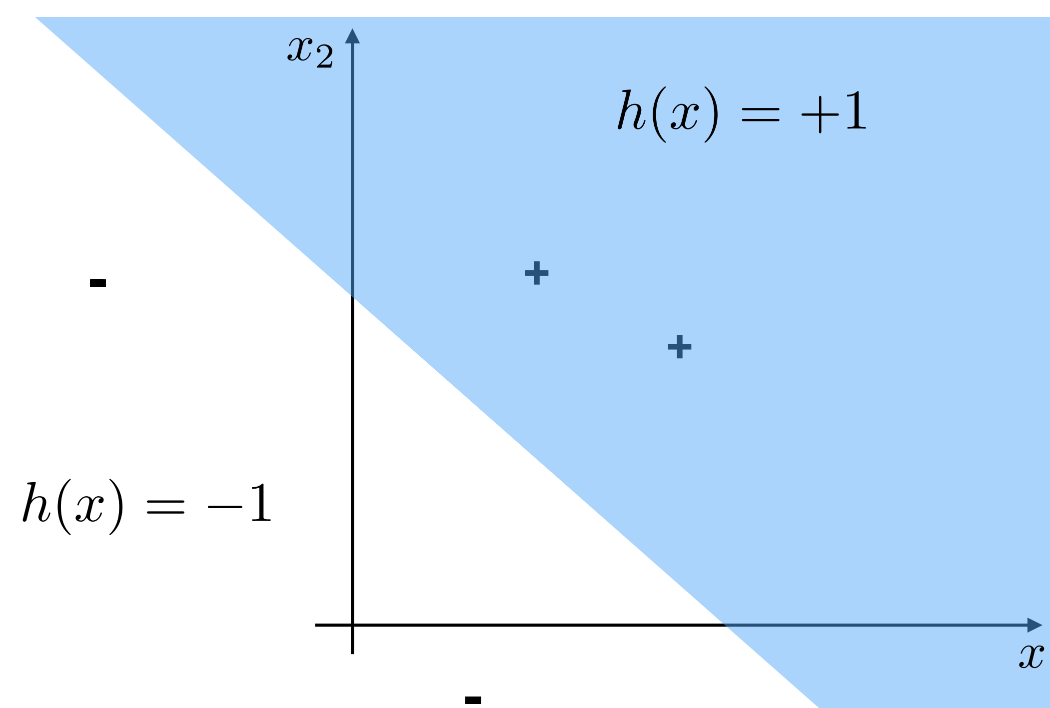


Collaborative Filtering

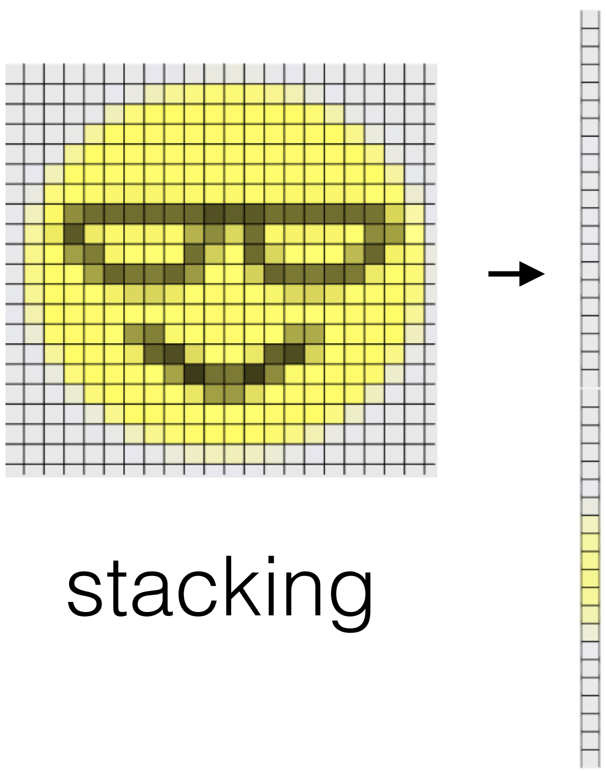
	Alpine Spa 	Himalayas 	Hawaii 	Scuba 
	20	??	16	??
	??	2	18	10
	13	1	??	??
	??	13	??	19
	18	??	10	??
	??	??	0	16

Features

Image		**Category**
$h\Big($  $;\theta\Big)$	$=$	mushroom
		cherry
...		...

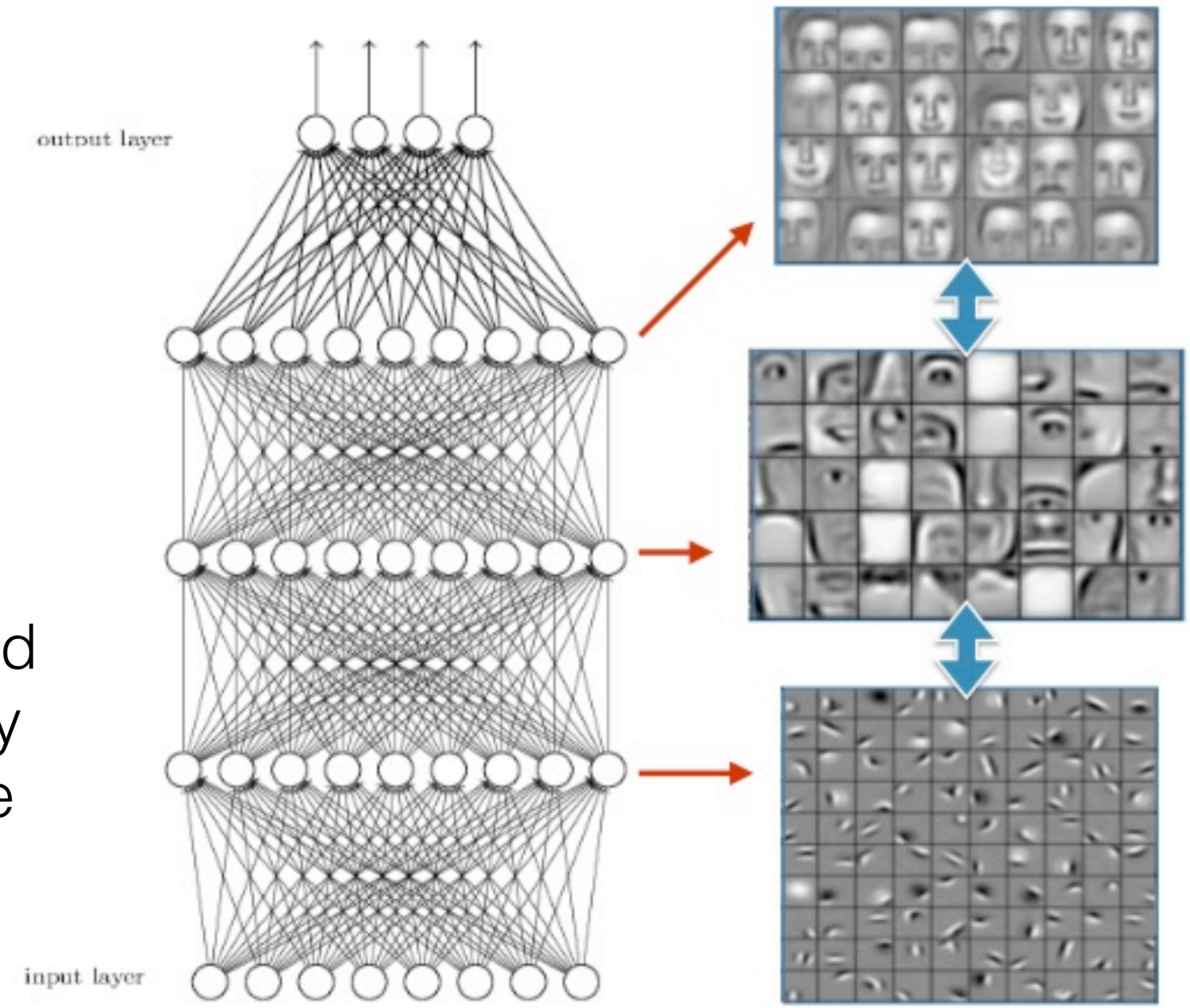


Features: examples for images

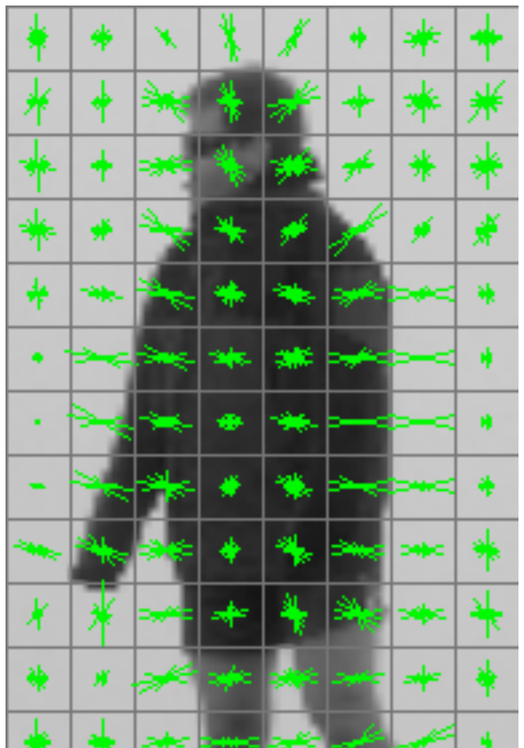


stacking

deep network

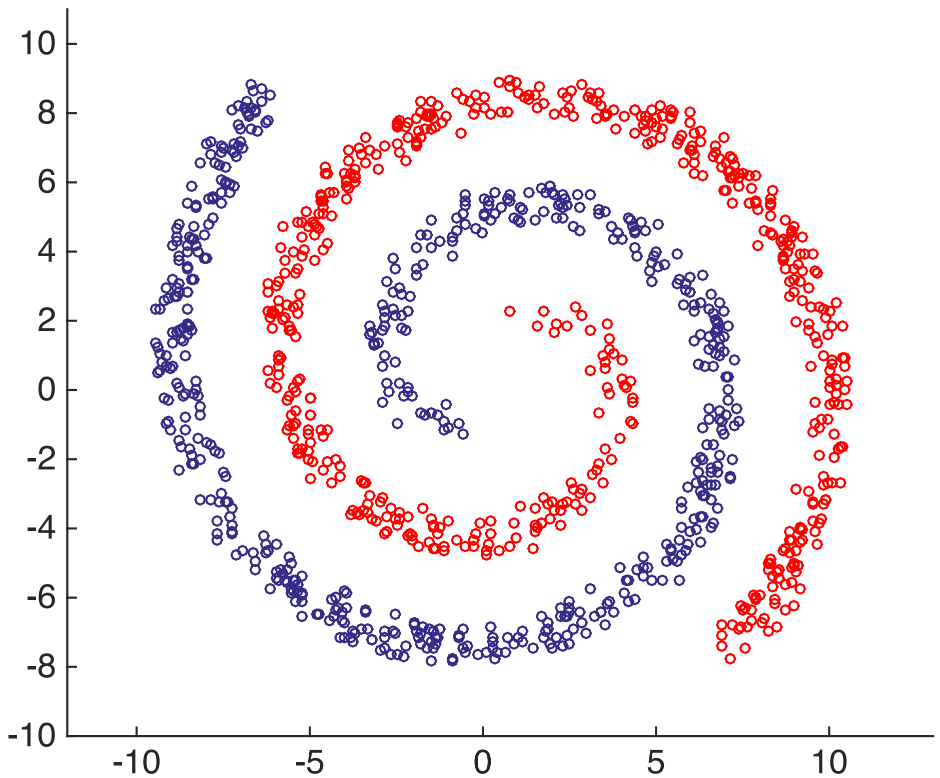


HOG

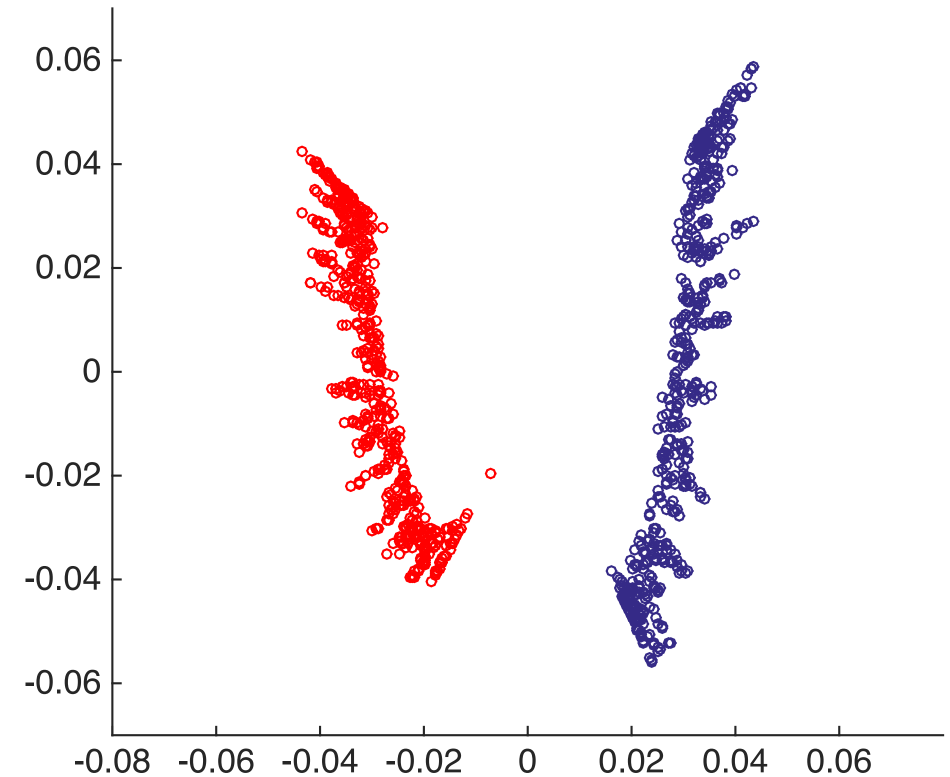


… and many more

The Power of Features



… or …



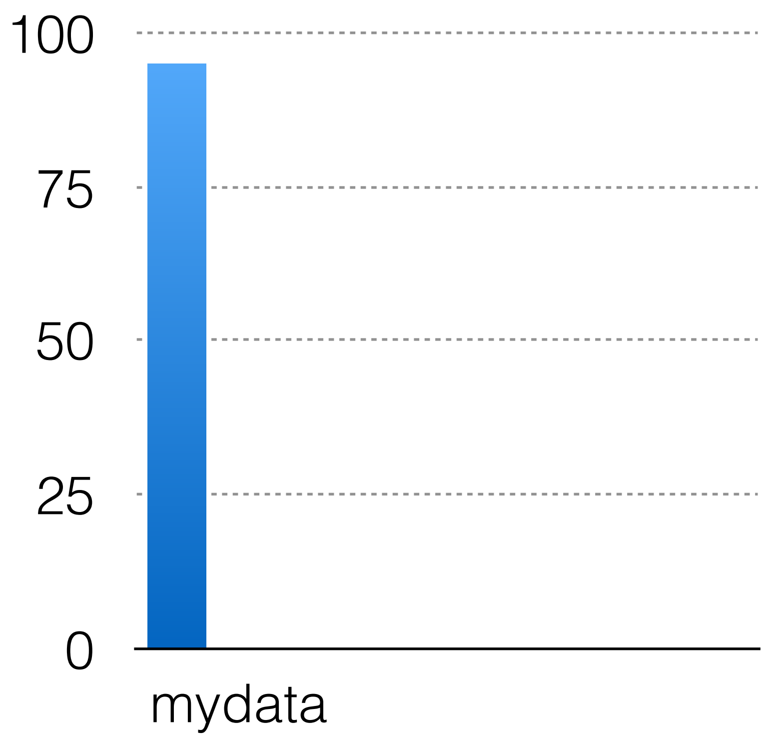
- above: “spectral embedding”
- typical procedures in statistics: make it look Gaussian
  - log-transform
  - powers
  - exponentiation
  - normalization

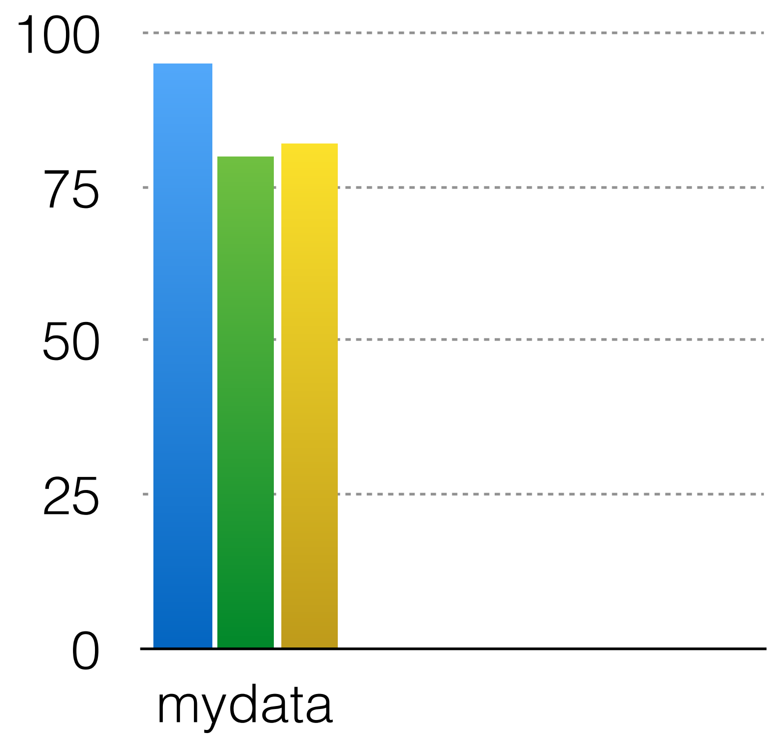
A few things to be aware of ...

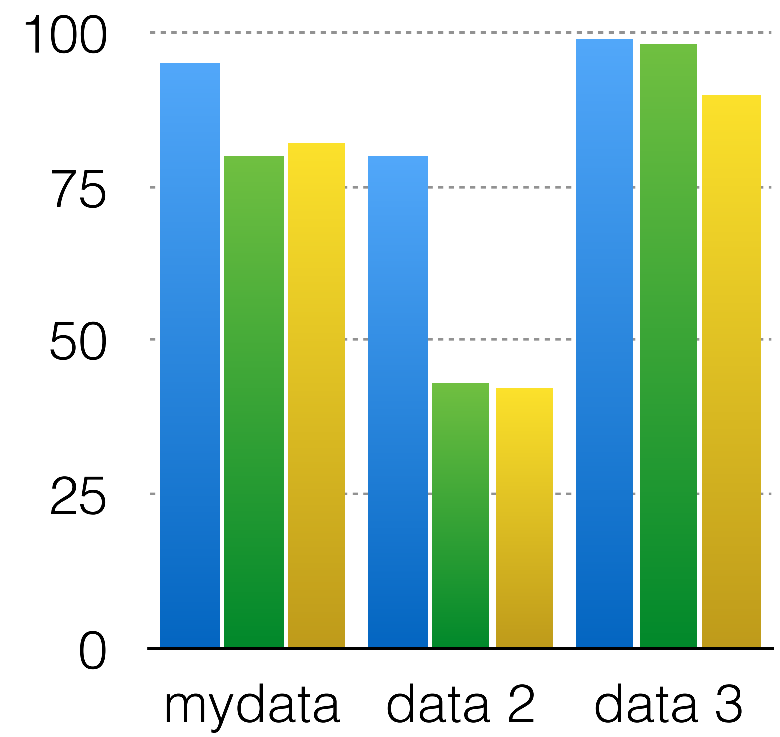
Error and evaluation

- our aim for a good prediction method: generalize!
- Which error?
  Training and testing subset: train on training set, error on test set
- Tuning parameters?
  the one with lowest error?
  cross-validation

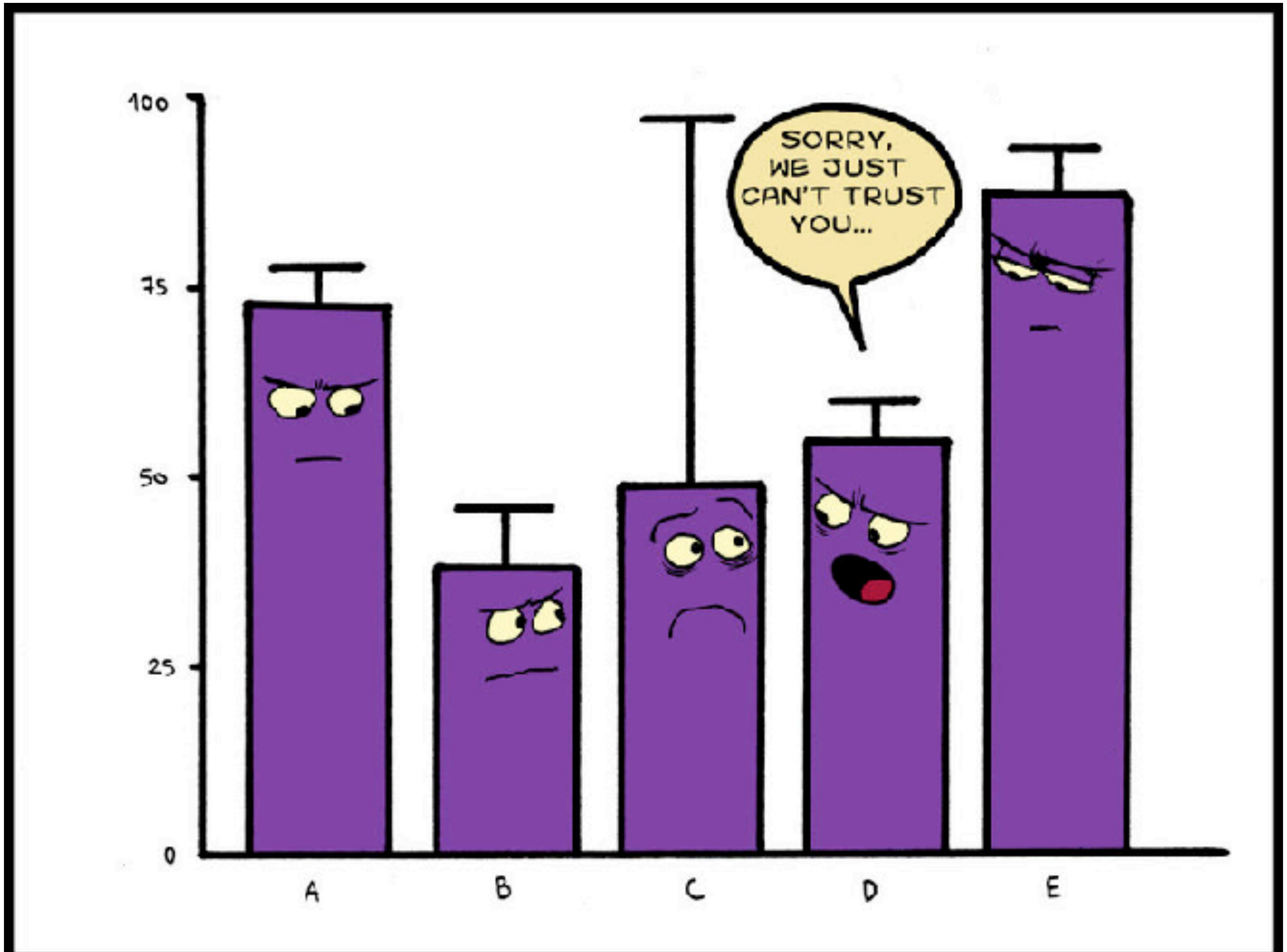
Is my method good?







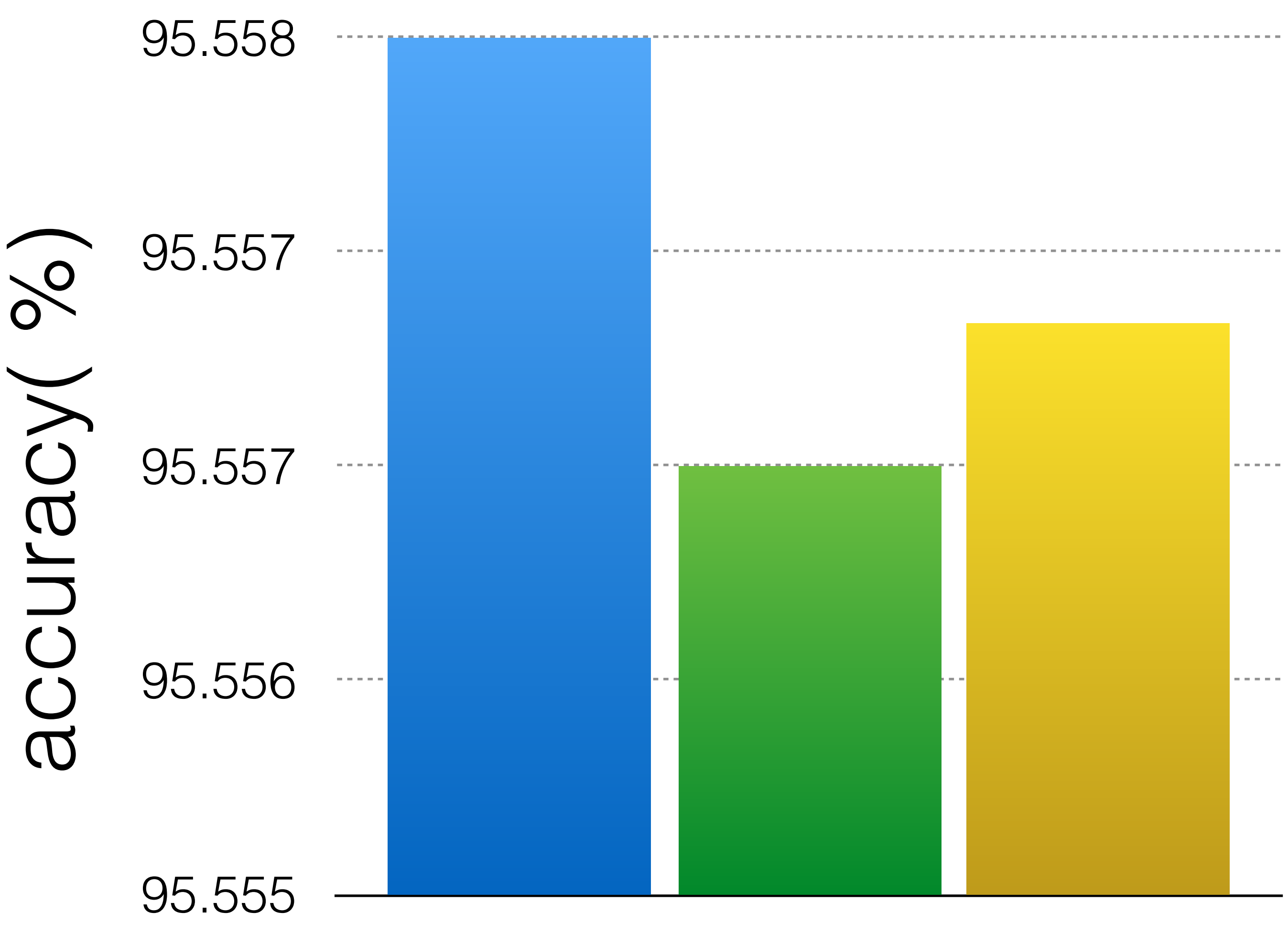
- Baseline & comparison methods, different datasets (dependent on problem)
- When working with one dataset: try different methods
  When developing a new method: compare to “standard”, different data
- error bars
- read related work (Google scholar, NIPS, ICML, KDD, AISTATS, UAI, CVPR, ICCV, NAACL, …)



100
75
50
25
0
SORRY, WE JUST CAN'T TRUST YOU...
A
B
C
D
E

Is my method good?

Mind the y axis!



Examples of projects

- CNN approaches to molecular representations for predicting chemical properties
- Characterizing and predicting air traffic delays
- Do signaling networks in brain tumors differ from those of normal brain tissue, and if so, how?
- Analysis of browsing behavior to predict mental state of users
- …

Add-ons: other advice

- Make sure you have the computation infrastructure needed. If your data is large, you should still be able to process it.
- Real data can be messy, it can have outliers, missing values etc. Document, remember and describe any preprocessing step you do. This is part of the analysis!
- You must have your data when you hand in the pre-proposal. E.g., if your data involves human subjects and you need to collect it, you would need approval for that, and that can take time