Data-Intensive Distributed Computing

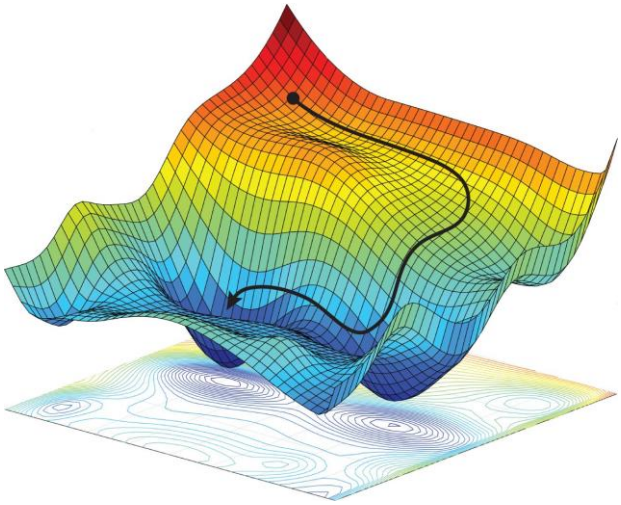
CS 431/631 451/651 (Fall 2021)

Part 7: Data Mining (2/4)

Ali Abedi

These slides are available at https://www.student.cs.uwaterloo.ca/~cs451

Stochastic Gradient Descent



2

Stochastic Gradient Descent

Gradient Descent

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \frac{1}{n} \sum_{i=0}^{n} \nabla \ell(f(\mathrm{x}_i; \theta^{(t)}), y_i)$$

Stochastic Gradient Descent (SGD)

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \nabla \ell(f(\mathrm{x}; \theta^{(t)}), y)$$

3

Stochastic Gradient Descent

Gradient Descent

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \frac{1}{n} \sum_{i=0}^{n} \nabla \ell(f(\mathrm{x}_i; \theta^{(t)}), y_i)$$

Considers all training instances in every iteration

Stochastic Gradient Descent (SGD)

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \nabla \ell(f(\mathrm{x}; \theta^{(t)}), y)$$

Considers a random instance in every iteration

Stochastic Gradient Descent

Batch

Gradient Descent

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \frac{1}{n} \sum_{i=0}^{n} \nabla \ell(f(\mathrm{x}_i; \theta^{(t)}), y_i)$$

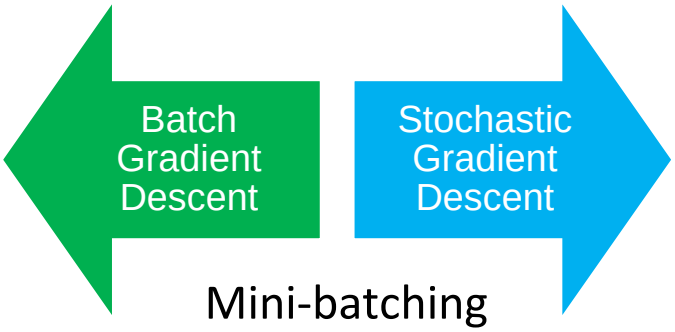
Considers all training instances in every iteration

Online

Stochastic Gradient Descent (SGD)

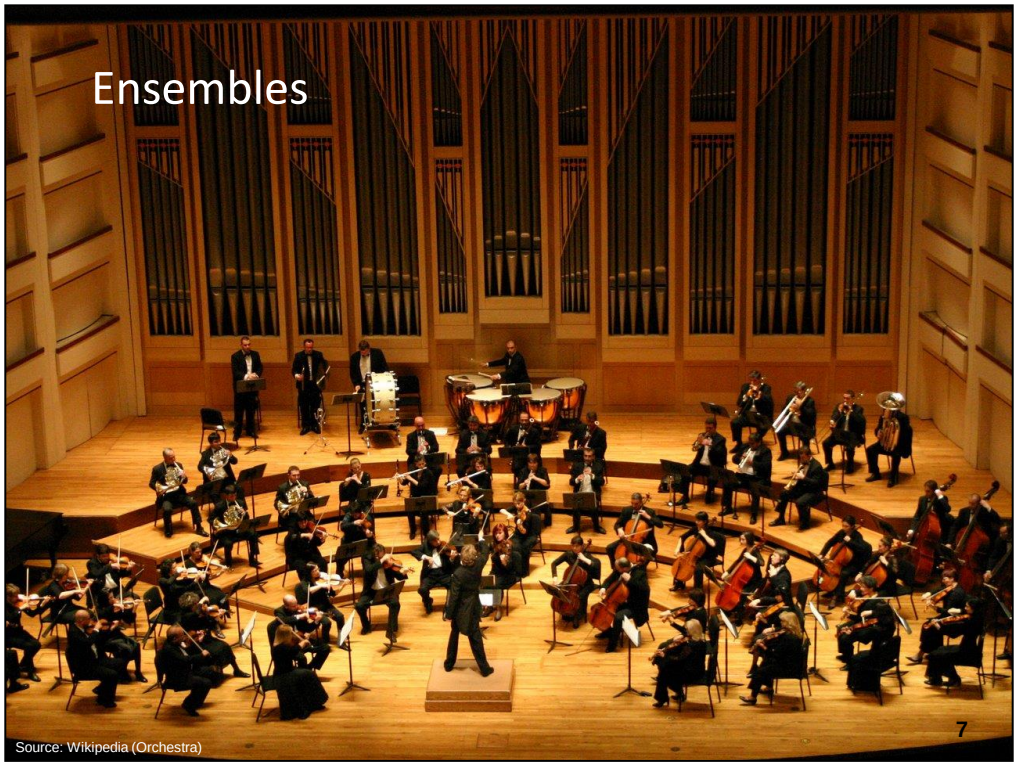
$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \nabla \ell(f(\mathrm{x}; \theta^{(t)}), y)$$

Considers a random instance in every iteration



Mini-batching

Considers a random subset of instances in every iteration


Ensembles
Source: Wikipedia (Orchestra)
7

Ensemble Learning

Learn multiple independent models, combine results from different models to make prediction

Common implementation:

Train classifiers on different input partitions of the data

Embarrassingly parallel!

Combining predictions:

Majority voting

Model averaging

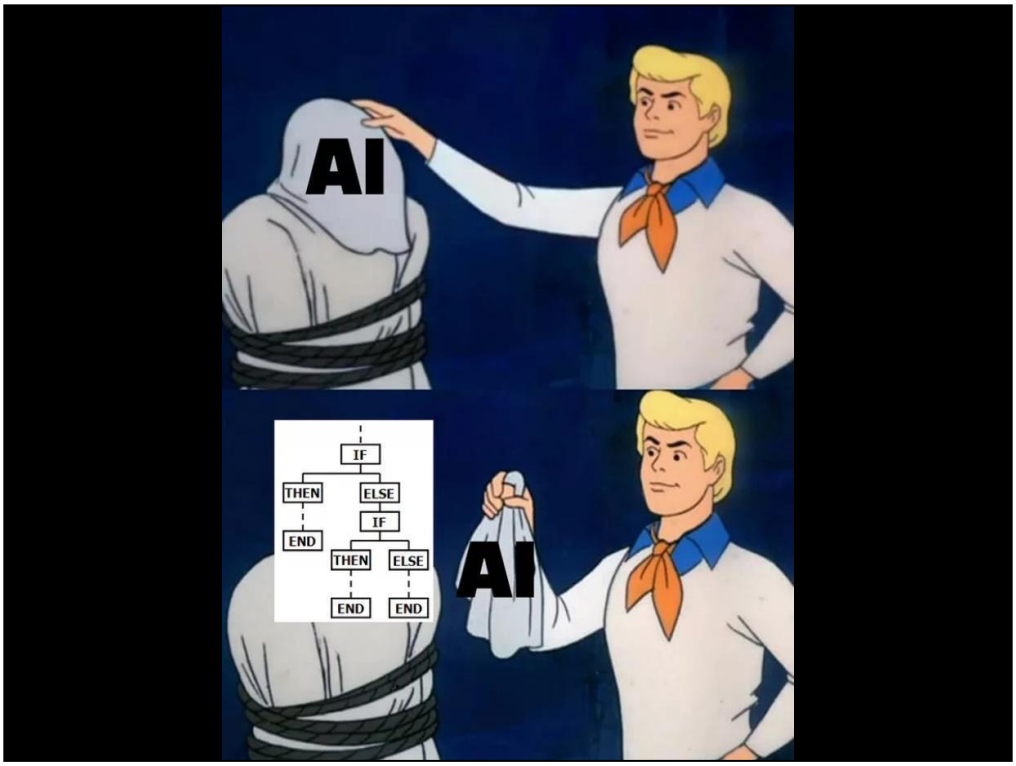
Ensemble Learning

Learn multiple independent models, combine results from different models to make prediction

Why does it work?

If errors uncorrelated, multiple classifiers being wrong is less likely

Reduces the variance component of error

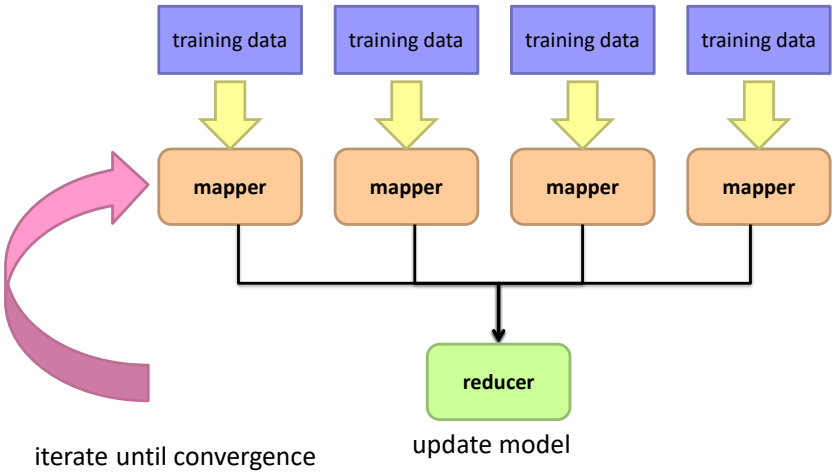

AI
IF
THEN
ELSE
END
IF
THEN
ELSE
END
END
AI

MapReduce Implementation

Reminder

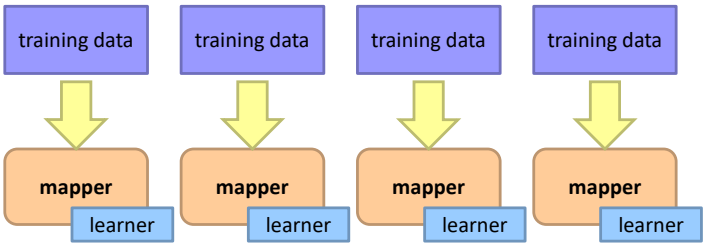
Gradient Descent

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \frac{1}{n} \sum_{i=0}^{n} \nabla \ell(f(\mathrm{x}_i; \theta^{(t)}), y_i)$$



Stochastic Gradient Descent

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \nabla \ell(f(\mathrm{x}; \theta^{(t)}), y)$$



No iteration!

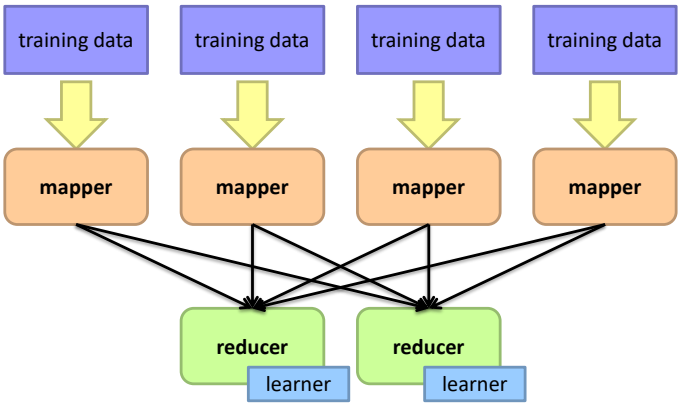
13

This is great because we no longer need iterations!

Mappers go through the record and apply the stochastic gradient descend rule on that record and update the model. This process continues for all records

Stochastic Gradient Descent

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \nabla \ell(f(\mathrm{x}; \theta^{(t)}), y)$$



No iteration!

MapReduce Implementation

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \nabla \ell(f(\mathrm{x}; \theta^{(t)}), y)$$

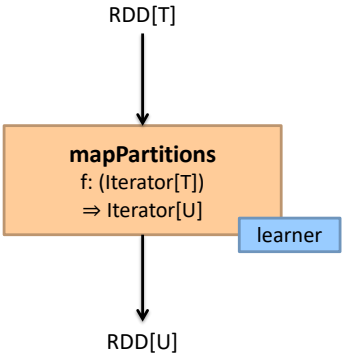
How do we output the model?

Option 1: write model out as "side data"

Option 2: emit model as intermediate output

What about Spark?

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \nabla \ell(f(\mathrm{x}; \theta^{(t)}), y)$$



In practice …

Data scientists usually use provided transformations in Spark ML

```
val model = LinearRegressionWithSGD.train(parsedData, numIterations, stepSize)

val prediction = model.predict(point.features)
```

Sentiment Analysis Case Study

Binary polarity classification: {positive, negative} sentiment


Use the "emoticon trick" to gather data


Data

Test: 500k positive/500k negative tweets from 9/1/2011
Training: {1m, 10m, 100m} instances from before (50/50 split)

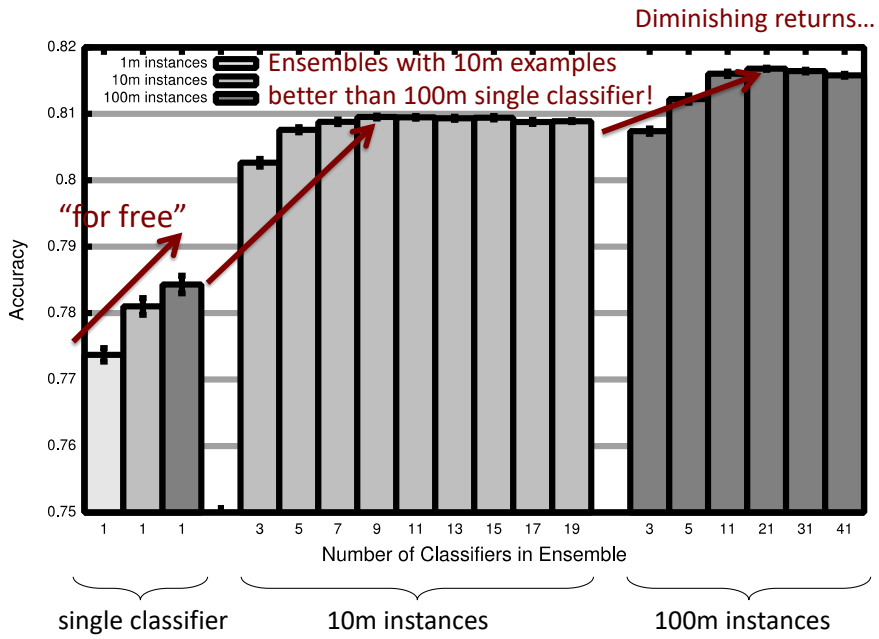
Features:

Sliding window byte-4grams

Models + Optimization:

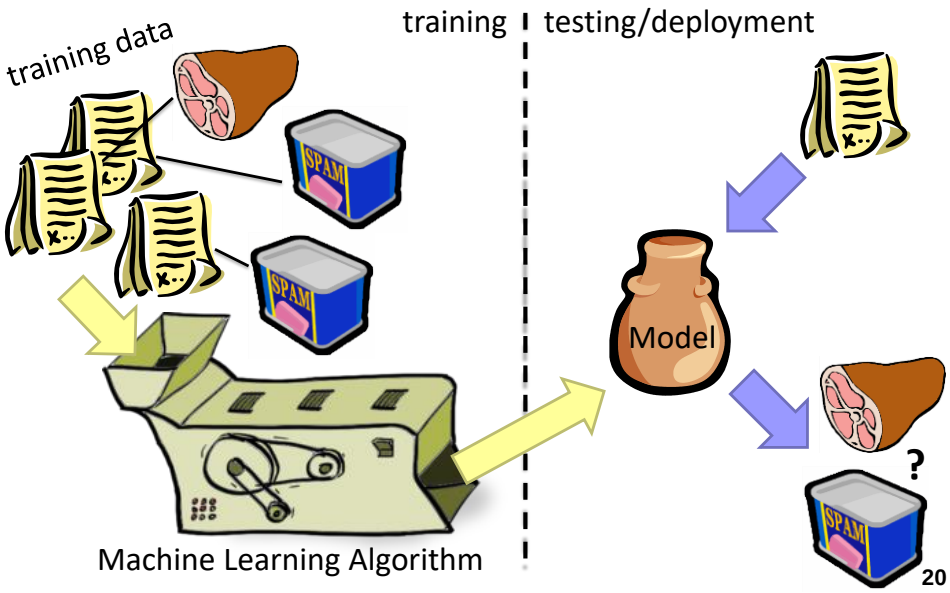
Logistic regression with SGD (L2 regularization)
Ensembles of various sizes (simple weighted voting)

Source: Lin and Kolcz. (2012) Large-Scale Machine Learning at Twitter. SIGMOD.


Diminishing returns…
1m instances
10m instances
100m instances
Ensembles with 10m examples
better than 100m single classifier!
"for free"
Accuracy
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0.75
1 1 1 3 5 7 9 11 13 15 17 19 3 5 11 21 31 41
Number of Classifiers in Ensemble
single classifier
10m instances
100m instances

19

Supervised Machine Learning



20

Evaluation

How do we know how well we're doing?

Why isn't this enough?

Induce: $f : X \to Y$

Such that loss is minimized

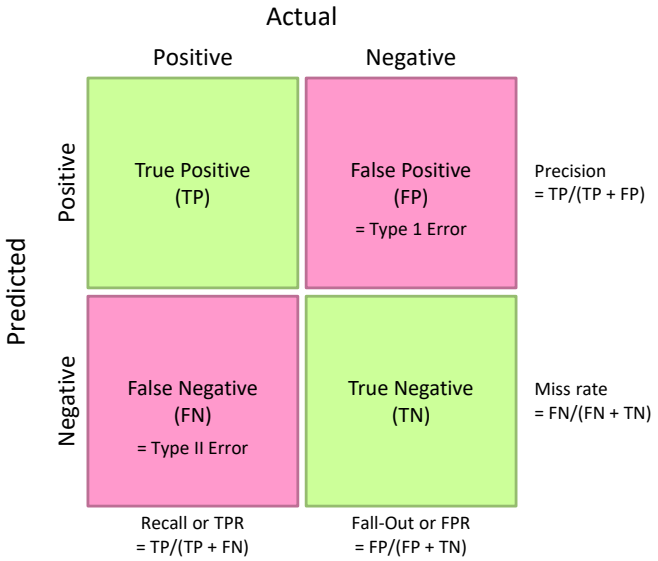
$$\arg\min_{\theta} \frac{1}{n} \sum_{i=0}^{n} \ell(f(x_i; \theta), y_i)$$

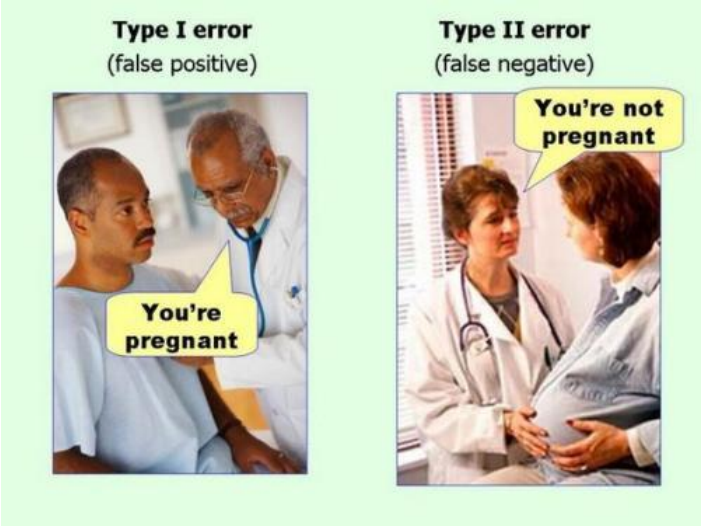
We need end-to-end metrics!

Obvious metric: accuracy

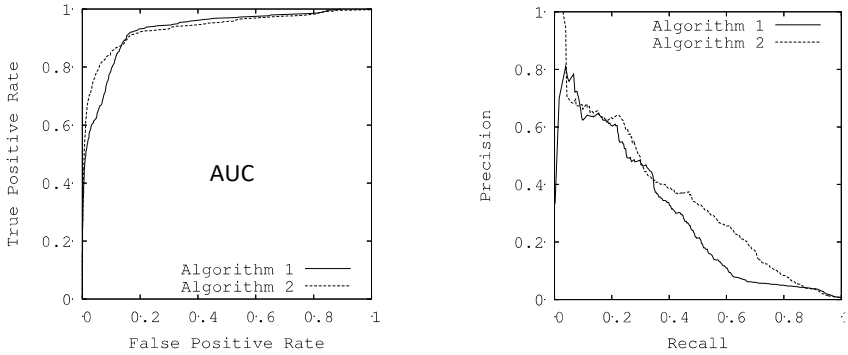
Why isn't this enough?

Metrics


Actual
Positive
Negative
Predicted
Positive
Negative
True Positive
(TP)
False Positive
(FP)
= Type 1 Error
Precision
= TP/(TP + FP)
False Negative
(FN)
= Type II Error
True Negative
(TN)
Miss rate
= FN/(FN + TN)
Recall or TPR
= TP/(TP + FN)
Fall-Out or FPR
= FP/(FP + TN)


Type I error
(false positive)
You're pregnant
Type II error
(false negative)
You're not pregnant

ROC and PR Curves

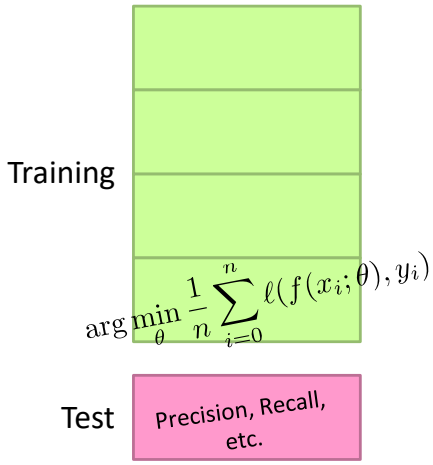


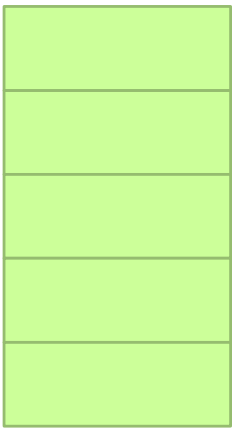
Source: Davis and Goadrich. (2006) The Relationship Between Precision-Recall and ROC curves

24

A **receiver operating characteristic curve**, or **ROC curve**, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

Training/Testing Splits

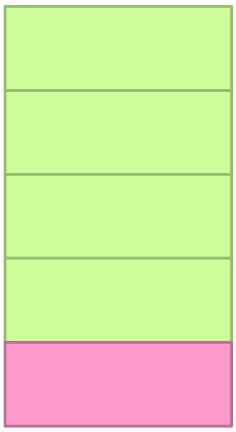




What happens if you need more?

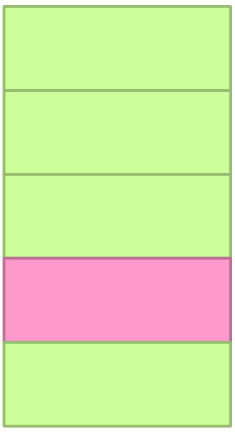
Cross-Validation

Training/Testing Splits



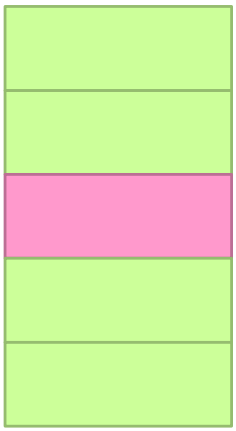
Cross-Validation

Training/Testing Splits



Cross-Validation

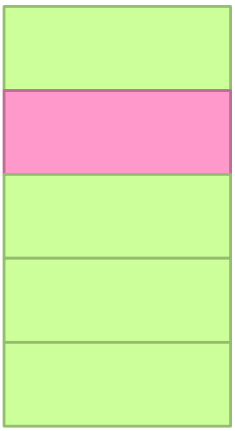
Training/Testing Splits



Cross-Validation

28

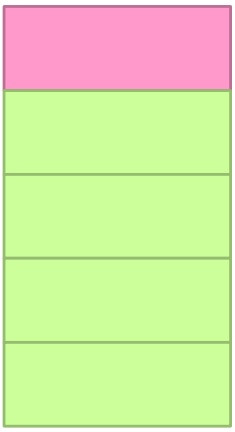
Training/Testing Splits



Cross-Validation

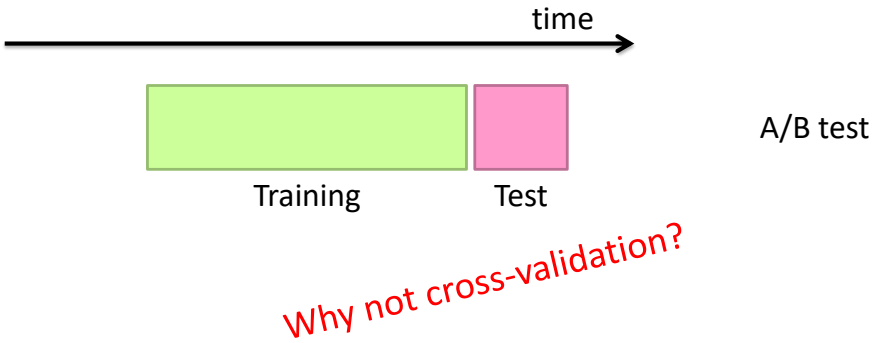
29

Training/Testing Splits



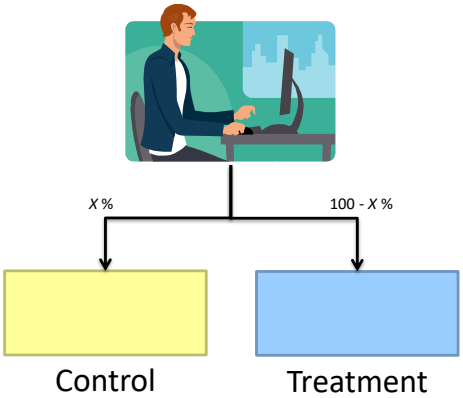
Cross-Validation

Typical Industry Setup



31

A/B Testing



Gather metrics, compare alternatives

A/B Testing: Complexities

Properly bucketing users

Novelty

Learning effects

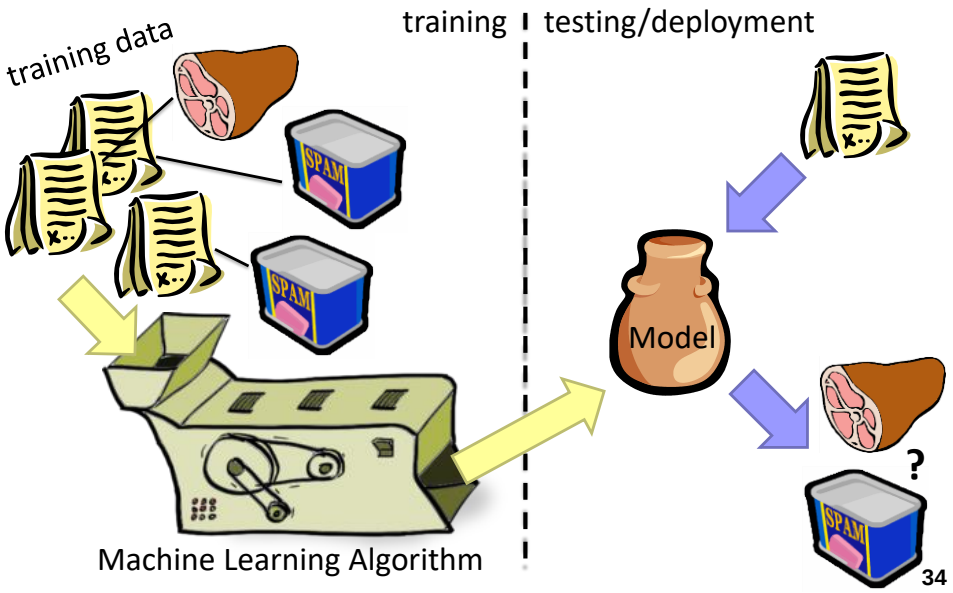
Long vs. short term effects

Multiple, interacting tests

Nosy tech journalists

…

Supervised Machine Learning



34

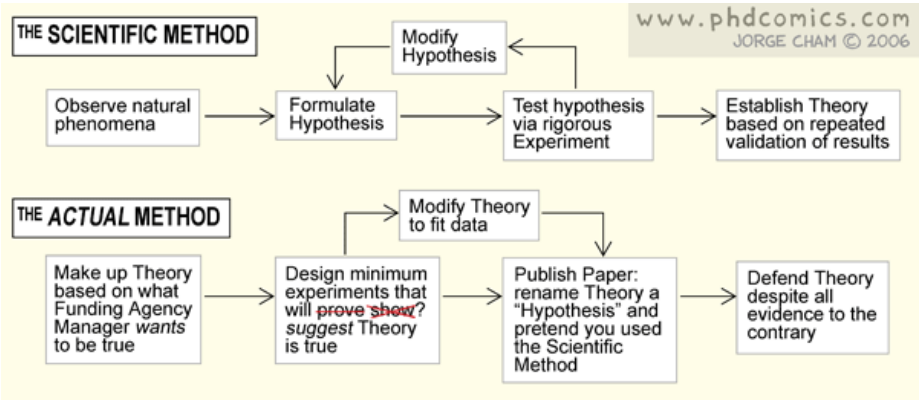
Applied ML in Academia

Download interesting dataset (comes with the problem)

Run baseline model
Train/Test

Build better model
Train/Test

Does new model beat baseline?
Yes: publish a paper!
No: try again!


www.phdcomics.com
JORGE CHAM © 2006
THE SCIENTIFIC METHOD
Observe natural phenomena
Formulate Hypothesis
Modify Hypothesis
Test hypothesis via rigorous Experiment
Establish Theory based on repeated validation of results
THE ACTUAL METHOD
Make up Theory based on what Funding Agency Manager wants to be true
Design minimum experiments that will prove show? suggest Theory is true
Modify Theory to fit data
Publish Paper: rename Theory a "Hypothesis" and pretend you used the Scientific Method
Defend Theory despite all evidence to the contrary

"You can't keep adjusting the data to prove that you would be the best Valentine's date for Scarlett Johansson."

Harvard Business Review

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

38

Fantasy

- Extract features
- Develop cool ML technique
- #Profit

Reality

- What's the task?
- Where's the data?
- What's in this dataset?
- What's all the f#$!* crap?
- Clean the data
- Extract features
- "Do" machine learning
- Fail, iterate…

Dirty secret: very little of data science is about machine learning per se!

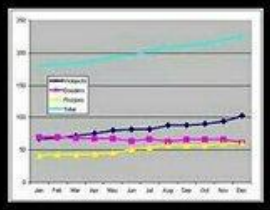
39

Data Scientist

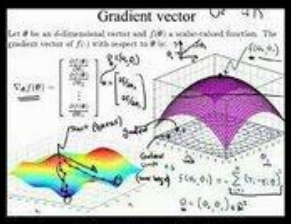
What my friends think I do

What my mom thinks I do

What society thinks I do

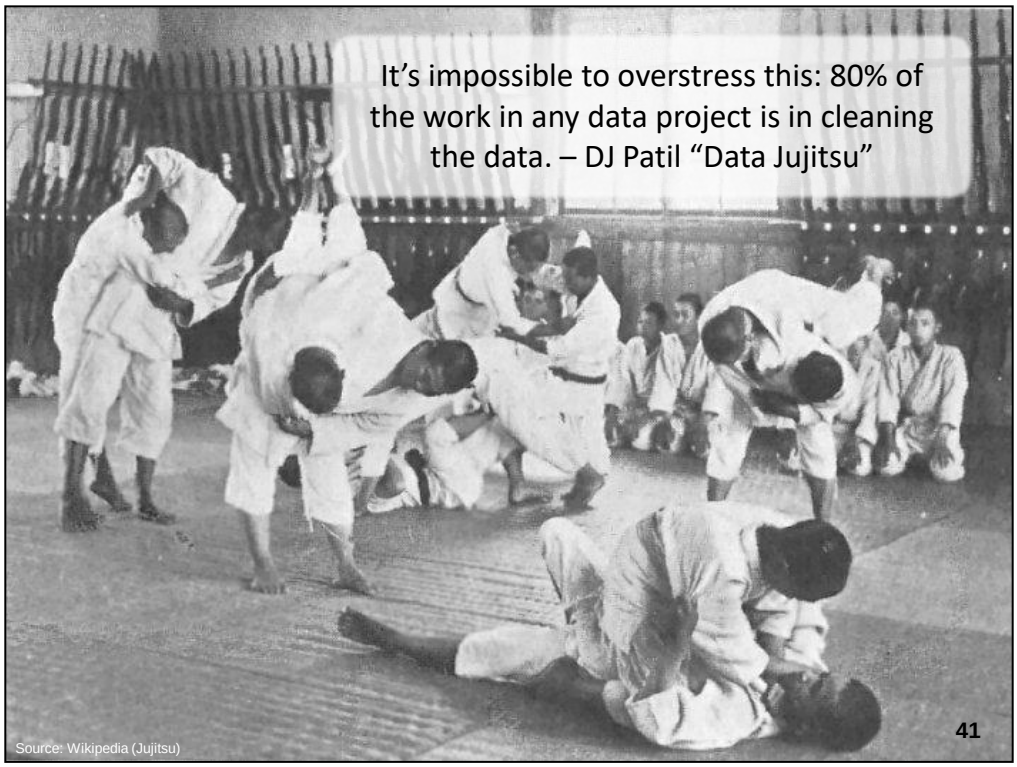


What my boss thinks I do



What I think I do

What I actually do


It’s impossible to overstress this: 80% of the work in any data project is in cleaning the data. – DJ Patil “Data Jujitsu”
Source: Wikipedia (Jujitsu)
41

The New York Times

SUBSCRIBE NOW LOG IN

SECTIONS HOME SEARCH

TECHNOLOGY

For 'Big Data' Scientists, Hurdle to Insights Is 'Janitor Work'

By STEVE LOHR AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist.

Peter DaSilva for The New York Times

On finding things…

On naming things…

uid UserId

userId

CamelCase

userid

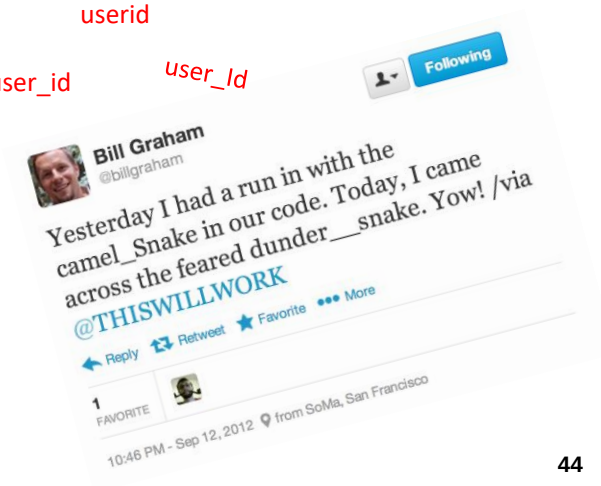
smallCamelCase

user_id user_Id

snake_case

camel_Snake

dunder__snake



On feature extraction…

```
^(\\w+\\s+\\d+\\s+\\d+:\\d+:\\d+)\\s+
([^@]+?)@(\\S+)\\s+(\\S+):\\s+(\\S+)\\s+(\\S+)
\\s+((?:\\S+?,\\s+)*(?:\\S+?))\\s+(\\S+)\\s+(\\S+)
\\s+\\[([^\\]]+)\\]\\s+\"(\\w+)\\s+([^\"\\\\]*
(?:\\\\.[^\"\\\\]*)*)\\s+(\\S+)\"\\s+(\\S+)\\s+
(\\S+)\\s+\"([^\"\\\\]*(?:\\\\.[^\"\\\\]*)*)
\"\\s+\"([^\"\\\\]*(?:\\\\.[^\"\\\\]*)*)\"\\s*
(\\d*-[\\d-]*)?\\s*(\\d+)?\\s*(\\d*\\.[\\d\\.]*)?
(\\s+[-\\w]+)?.*$
```

An actual Java regular expression used to parse log message at Twitter circa 2010

Friction is cumulative!

Data Plumbing… Gone Wrong!

[scene: consumer internet company in the Bay Area…]

It's over here…

Well, it wouldn't fit, so we had to shoehorn…

Hang on, I don't remember…

Uh, bad news. Looks like we forgot to log it…

Okay, let's get going… where's the click data?

Well, that's kinda non-intuitive, but okay…

Oh, BTW, where's the timestamp of the click?

[grumble, grumble, grumble]

Frontend Engineer

Develops new feature, adds logging code to capture clicks

Data Scientist

Analyze user behavior, extract insights to improve feature

46

Fantasy

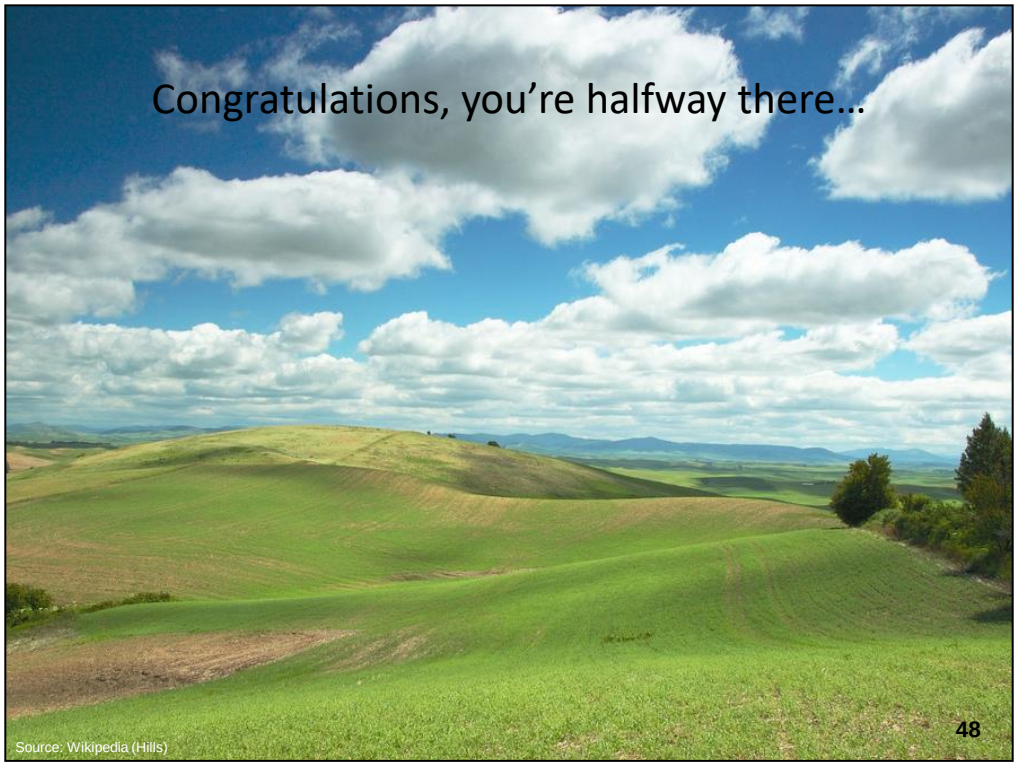
- Extract features
- Develop cool ML technique
- #Profit

Reality

- What's the task?
- Where's the data?
- What's in this dataset?
- What's all the f#$!* crap?
- Clean the data
- Extract features
- "Do" machine learning
- Fail, iterate…

Finally works!

47


Congratulations, you're halfway there…
Source: Wikipedia (Hills)
48

Congratulations, you're halfway there…

Does it actually work?
A/B testing

Is it fast enough?

Good, you're two thirds there…

49


Productionize
50
Source: Wikipedia (Oil refinery)

Productionize

What are your jobs' dependencies?

How/when are your jobs scheduled?

Are there enough resources?

How do you know if it's working?

Who do you call if it stops working?

Infrastructure is critical here!
(plumbing)


Takeaway lesson:
Most of data science isn't glamorous!
Source: Wikipedia (Plumbing)
52