Lecture Notes for

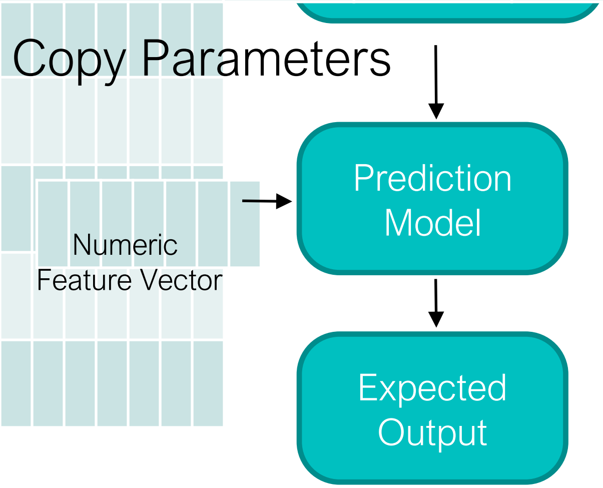
Machine Learning in Python

⟦👨‍🏫, 👨‍💻, 🐍, 👨‍🔬⟧

Data Quality and Imputation

Data Quality Problems

TID	***Hair Color***	***Hgt.***	***Age***	***Arrested***
1	Brown	5’2”	23	no
2	Hazel	1.5m	12	no
3	Bl	5	999	no
4	Brown	5’2”	23	no



Copy Parameters

Prediction Model

Numeric Feature Vector

Expected Output

- Missing
  - Easy to find, NaNs
- Duplicated
  - Easy to find, hard to verify
- Noise or Outlier
  - Hard to define / catch

Information is not collected (e.g., people decline to give their age and weight)

Features **not applicable** (e.g., annual income for children)

UCI ML Repository: 90% of repositories have missing data

Split-Impute-Combine

TID	Pregnant	BMI	Age	Diabetes
1	Y	33.6	41-50	positive
2	N	26.6	31-40	negative
3	Y	23.3	**?**	positive
4	N	28.1	21-30	negative
5	N	43.1	31-40	positive
6	Y	25.6	21-30	negative
7	Y	31.0	21-30	positive
8	Y	35.3	**?**	negative
9	N	30.5	51-60	positive
10	Y	37.6	51-60	positive

split: pregnant
split: BMI > 32

TID	Pregnant	BMI	Age	Diabetes
1	Y	>32	41-50	positive
8	Y	>32	**?**	negative
10	Y	>32	51-60	positive

Mode: none, can't impute

TID	Pregnant	BMI	Age	Diabetes
3	Y	<32	**?**	positive
6	Y	<32	21-30	negative
7	Y	<32	21-30	positive

Mode: 21-30

K-Nearest Neighbors Imputation

TID	Pregnant	BMI	Age	Diabetes
1	Y	33.6	41-50	positive
2	N	26.6	31-40	negative
3	Y	23.3	?	positive
4	?	28.1	21-30	negative
5	N	43.1	31-40	positive
6	Y	25.6	21-30	negative
7	Y	31.0	21-30	positive
8	Y	35.3	?	negative
9	N	30.5	51-60	positive
10	Y	37.6	51-60	positive

For K=3, find 3 closest neighbors

TID	Preg.	BMI	Age	Diabetes	Distance
3	Y	23.3	?	positive	0
6	Y	25.6	21-30	negative	(0 + 2.3 + 1)/3
2	N	26.6	31-40	negative	(1 + 3.3 + 1)/3
4	?	28.1	21-30	negative	(4.8+1)/2

Imputed Age: 21-30

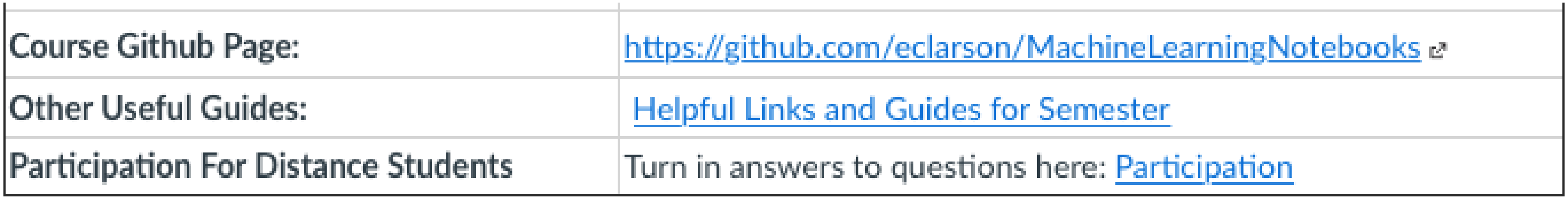
$$d_{i,j} = \frac{1}{|F_{valid}|} \sum_{f \in F_{valid}} \| f_i - f_j \|$$

How to calculate distance?

- Difference for valid features only
- May need to normalize ranges
- Or weight neighbors differently
- Or have min # of valid features
- Euclidean, city-block, etc.

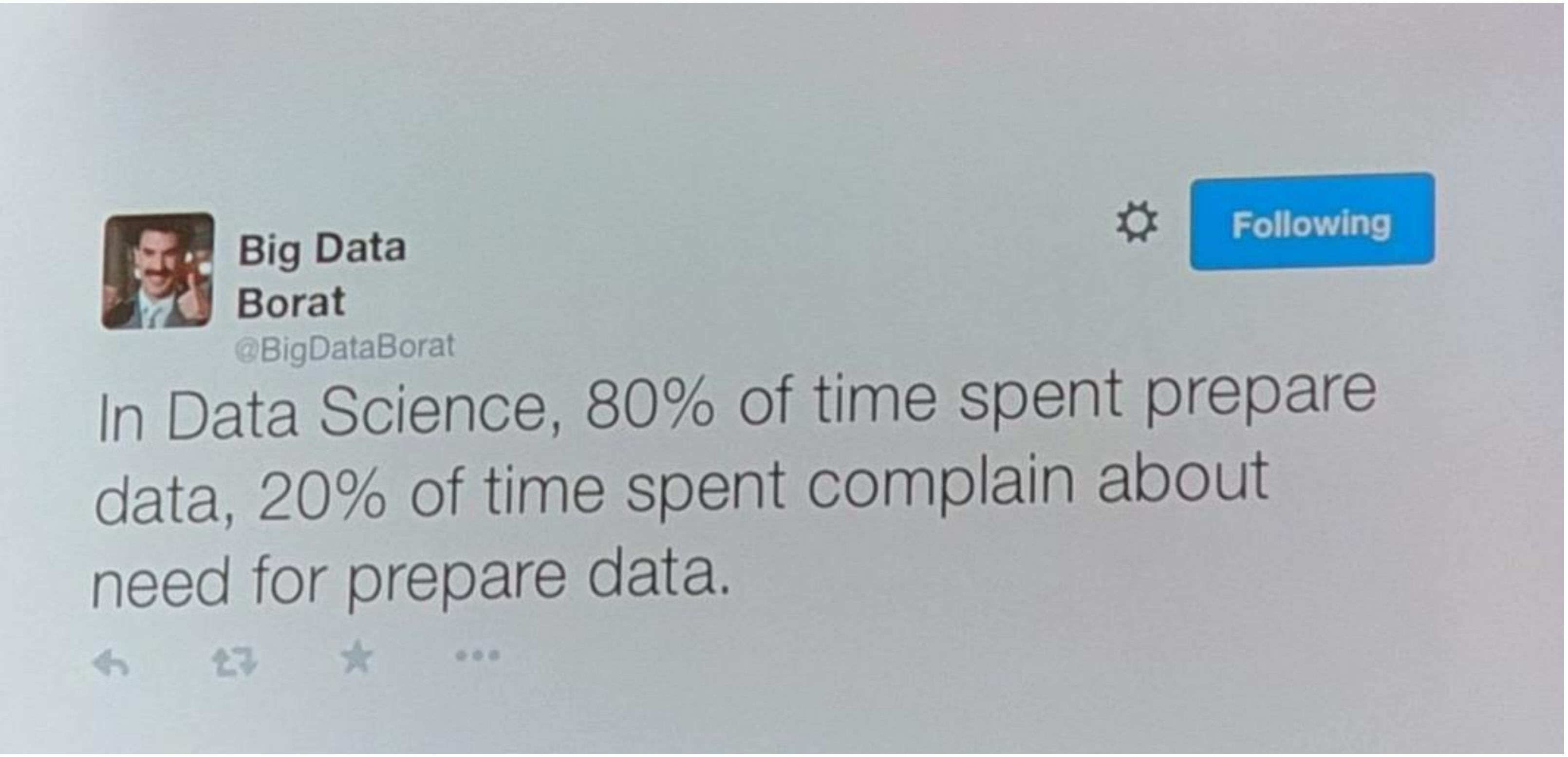
Class Logistics and Agenda

- Agenda:
  - Data Quality
  - Data Representations
  - Imputation methods
- Needing some more help?
  - **fast.ai** has great, free resources
  - canvas has links to various resources
  - your textbook is a great resource!



Course Github Page:	https://github.com/eclarson/MachineLearningNotebooks
Other Useful Guides:	Helpful Links and Guides for Semester
Participation For Distance Students	Turn in answers to questions here: Participation

Data Representation and Documents



Big Data Borat
@BigDataBorat

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

Data Tables as Variable Representations

Table

TID	Pregnant	BMI	Age	Eye Color	Diabetes
1	Y	33.6	41-50	brown	positive
2	N	26.6	31-40	hazel	negative
3	Y	23.3	31-40	blue	positive

Internal Rep.

TID							
1							
2							
3							
4							
5							
6	1	25.6	0	0	1	0	0

Bag of words model

TID	Pregnant	BMI	Chart Notes	Diabetes
1	Y	33.6	Complaints of fatigue wh…	positive
2	N	26.6	Sleeplessness and some…	negative
3	Y	23.3	First saw signs of rash o…	positive
4	N	28.1	Came in to see Dr. Steve…	inconclusive
5	N	43.1	First diagnosis for hospit…	positive
6	Y	25.6	N/A	negative

Vocabulary

Bag of Words

TID	Sleep	Fatigue	Weight	Rash	First	Sight
1	0	1	0	0	2	0
2	1	1	0	0	1	1
3	1	1	0	2	1	1

number of occurrences

Term-Frequency, Inverse-Document-Frequency

TID	Sleep	Etc	Stev
1	0	.86	0
2	0.1	.02	0
3	0.1	0.1	0

te

$\in D$

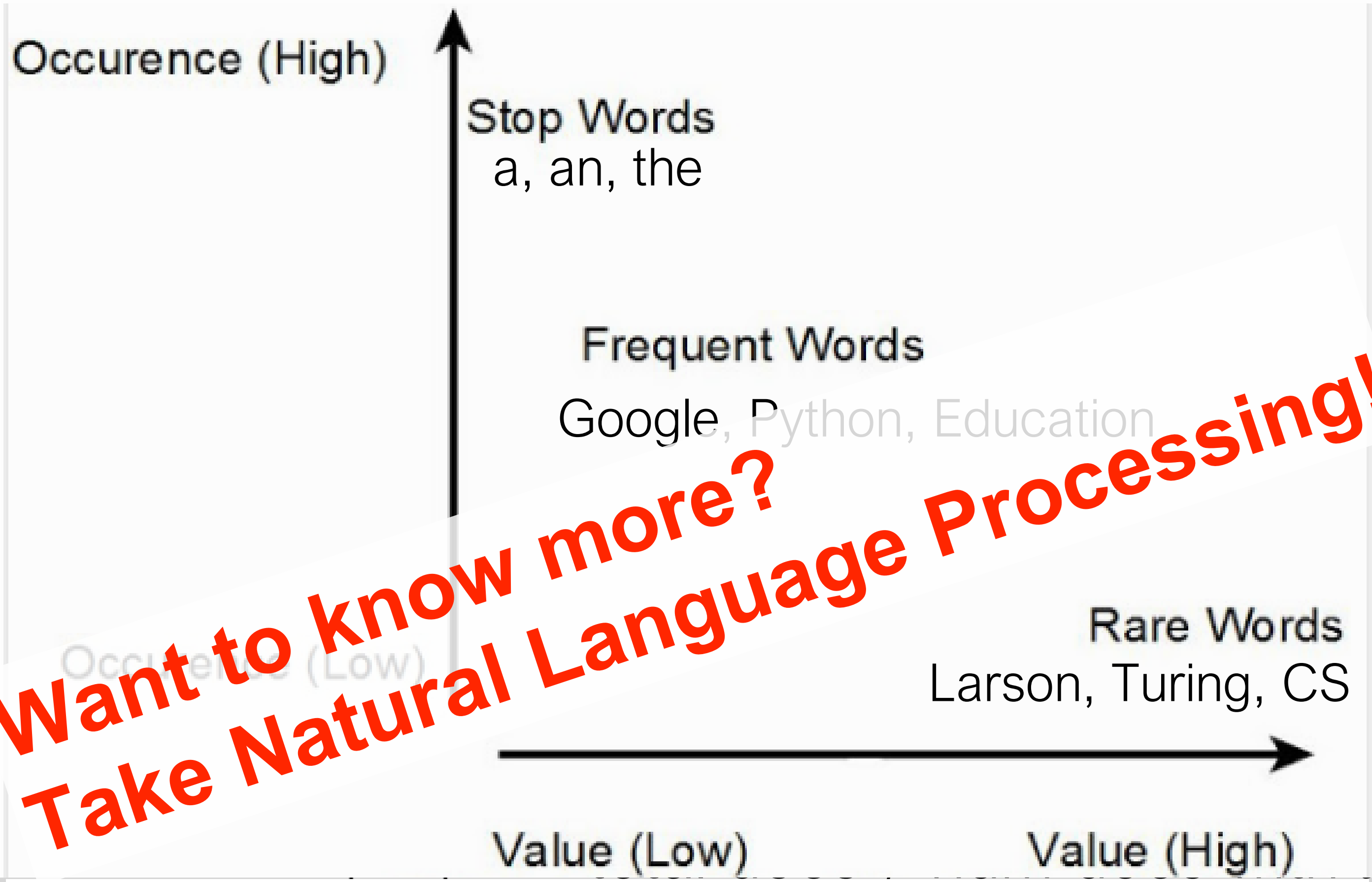
n d”

inverse

idf(

$\in d$

”



Want to know more?

Take Natural Language Processing!

https://www.kaggle.com/divsinha/sentiment-analysis-countvectorizer-tf-idf

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \cdot \text{idf}(t,d)$$

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \cdot (1 + \text{idf}(t,d))$$ smoothed

Demo

Pandas and Imputation

Scikit-Learn

Start the following:

03. Data Visualization.ipynb

Other Tutorials:

http://vimeo.com/59324550

http://pandas.pydata.org/pandas-docs/version/0.15.2/tutorials.html

Lecture Notes for

Machine Learning in Python

Professor Eric Larson

Data Quality and Imputation