Lecture Notes for

Machine Learning in Python

⟦👨‍🏫, 👨‍💻, 🐍, 👨‍🔬⟧

Professor Eric Larson

Introduction, Syllabus, Data Types

Class Logistics and Agenda

- Agenda:
  - Course Overview
  - ~~Introductions/Cards~~
  - Syllabus
  - What is Machine Learning?
  - Types of Data
  - Numpy/Pandas Demo

- My approach to this course:
  - Programming
  - Math
  - **Applications** and **Analytics**

Class Overview, by topic

Table Data Visualization

Numpy, Pandas, Seaborn
Overviews with some in-depth discussion

Dimension Reduction and Image Processing

Scikit-learn, Scikit Image,
Intuition only, Some mathematics

Linear and Logistic Regression

Numpy, Recreate API for Scikit-learn
Detailed mathematics for simple optimization
intuition for advanced optimization

Neural Networks and Back Prop.

Numpy
Detailed mathematics for NN operations

Wide and Deep Networks | **Convolutional Networks** | **Sequential Networks**

Keras, Tensorflow
Intuition, Detailed implement.

Ethics in Language Models

ConceptNet
Case studies

Class Overview, by assignment

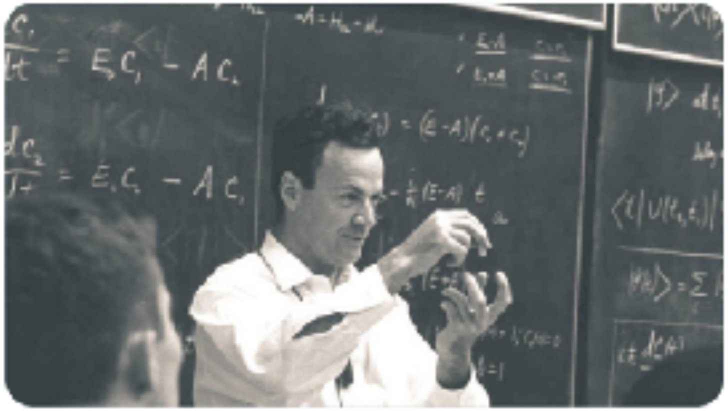
- **Lab One**: Visualize data and extract some features
- **Lab Two**: Analyze Images, Use dimensionality Reduction
- **Lab Three**: Program Logistic Regression in style of Sci-kit Learn
- **Lab Four**: Program NN Back propagation from Scratch, implement Adaptive Gradient Techniques
  - Use given dataset for this lab
- **Lab Five**: Wide and Deep networks
- **Lab Six**: Classify Images with Convolutional Networks
- **Lab Seven**: Classify Text with Sequential Networks

All Assignments posted on Canvas, with Rubric
Everything is a team assignment except quizzes, participation
You CANNOT makeup late quizzes, participation

Introductions & Course Syllabus

Richard Feynman @ProfFeynman · 12h

Don't just teach your students to read.
• Teach them to question what they read, what they study.
• Teach them to doubt.
• Teach them to think.
• Teach them to make mistakes and learn from them.
• Teach them how to understand something.
• Teach them how to teach others.



Richard Feynman @ProfFeynman · 21h

You cannot get educated by this self-propagating system in which people study to pass exams, and teach others to pass exams, but nobody knows anything.

You learn something by doing it yourself, by asking questions, by thinking, and by experimenting. 🧠

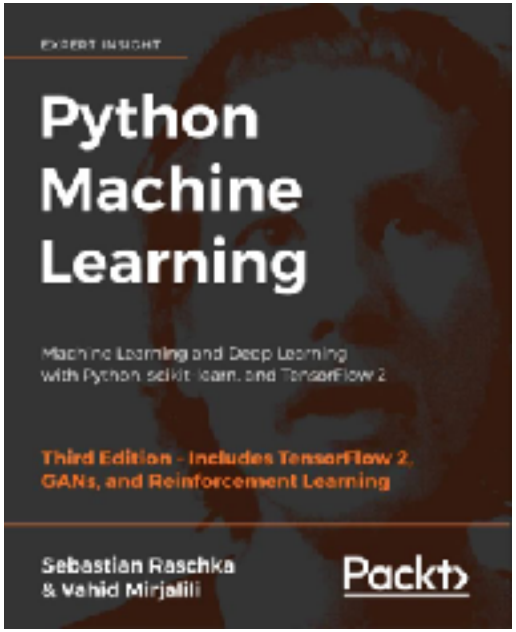
Introductions

- Me
    - Dr. Larson 👍
    - Prof. Larson 👍
    - PhD students: (Eric) 👀
    - Other 👎
- You
    - Name Department
    - Grad/Undergrad
    - Something true or false

Limited Introduction because of Class Size

FAQ

- Text:
  - **Recommended**: Python Machine Learning, Raschka & Mirjalili, Third Edition



- Use Canvas for posted course material
- Prerequisites:
  - Linear algebra & calculus (multivariate)
  - Basic statistics and probability
  - Basic OO programming, some python
- Version of **python 3.8**
  - Install through **Anaconda** and **pip**
  - Use **conda** environments
  - JupyterLab (or **notebook**)
- Most Used Libraries: Numpy, Pandas, Scikit-Learn, Matplotlib, Seaborn, Tensorflow
- Use OIT Data Science Workshops

Canvas Syllabus

- Lab Assignments
- Flipped Assignments
- Grading Rubrics
- Participation
- Course Schedule
- Difference between 5000 and 7000

How will participation be graded?

- Participation will be graded in the course:
  - **Distance students** will answer these questions via **canvas upload** (same for Zoom)
  - upload "over" the last submission
    - must upload the questions each week for full credit
- In Class Students:
  - Live question answering (mostly attendance):
    - ~~Do you think this will work?~~
      - ~~A: **Yes** this is going to work.~~
      - ~~B: This is **not** going to work.~~
      - ~~C: My name was not on my card.~~
      - ~~D: I (will/did) add an Alias to my card.~~

Is this plagiarism in this class?

- Copying code/text from another source without citing it
  - A. Yes, plagiarism!
  - B. No, its fine!
- Copying code/text from another source, citing at the end of the assignment in a blanket statement (but not making it clear which part of the assignment was from another source)?
  - A. Yes, plagiarism!
  - B. No, its fine!
- Copying code, citing the source directly next to the code, and commenting on what parts were changed?
  - A. Yes, plagiarism!
  - B. No, its fine!
- Copying text directly and citing the source with the text, but not placing the text in quotes.
  - A. Yes, plagiarism!
  - B. No, its fine!

Is this plagiarism in this class?

- Using ChatGPT or other LLM that generates text/code/responses?
  - A. Yes, plagiarism!
  - B. No, its fine!
  - C. It might be okay, but people should:
    1) acknowledge when using it, include prompt
    2) add your own comments to code (not just generated)
    3) check the accuracy and reliability
    4) not use text word for word, only as an outline or exemplar of a possible answer
    5) consent to be graded with an LLM

Don't use a LLM at the detriment of your own understanding. Don't use a LLM because your are unsure of your own understanding

Machine Learning Overview

What is Machine Learning?

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can change when exposed to new data.

What is machine learning? - Definition from WhatIs.com
whatis.techtarget.com/**definition/machine-learning**

About this result • Feedback

- **Beware of this definition**:
    - full of imprecise, loaded words:
        - intelligence, learning
    - ignores social structures, ethics, deployment, and that all results are interpreted by a human
- **My definition**: a way to optimize model parameters for recognizing complex patterns in data

Machine Learning

One Small Piece of Artificial Intelligence

Data Mining

ML

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables
- Description Methods
  - Find human-interpretable patterns that describe the data.

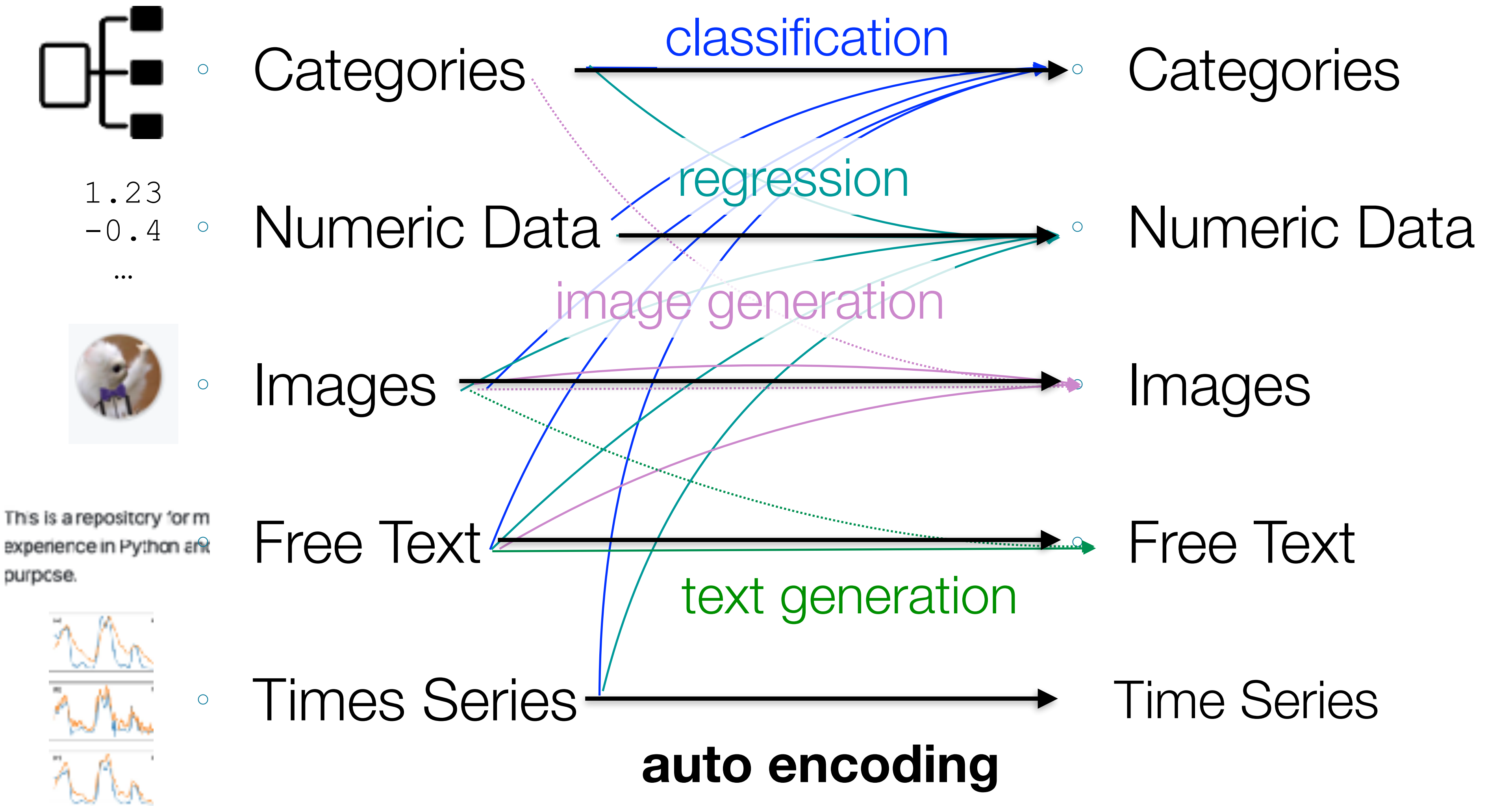
ML

- Classification
- Regression
- Deviation Detection
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery

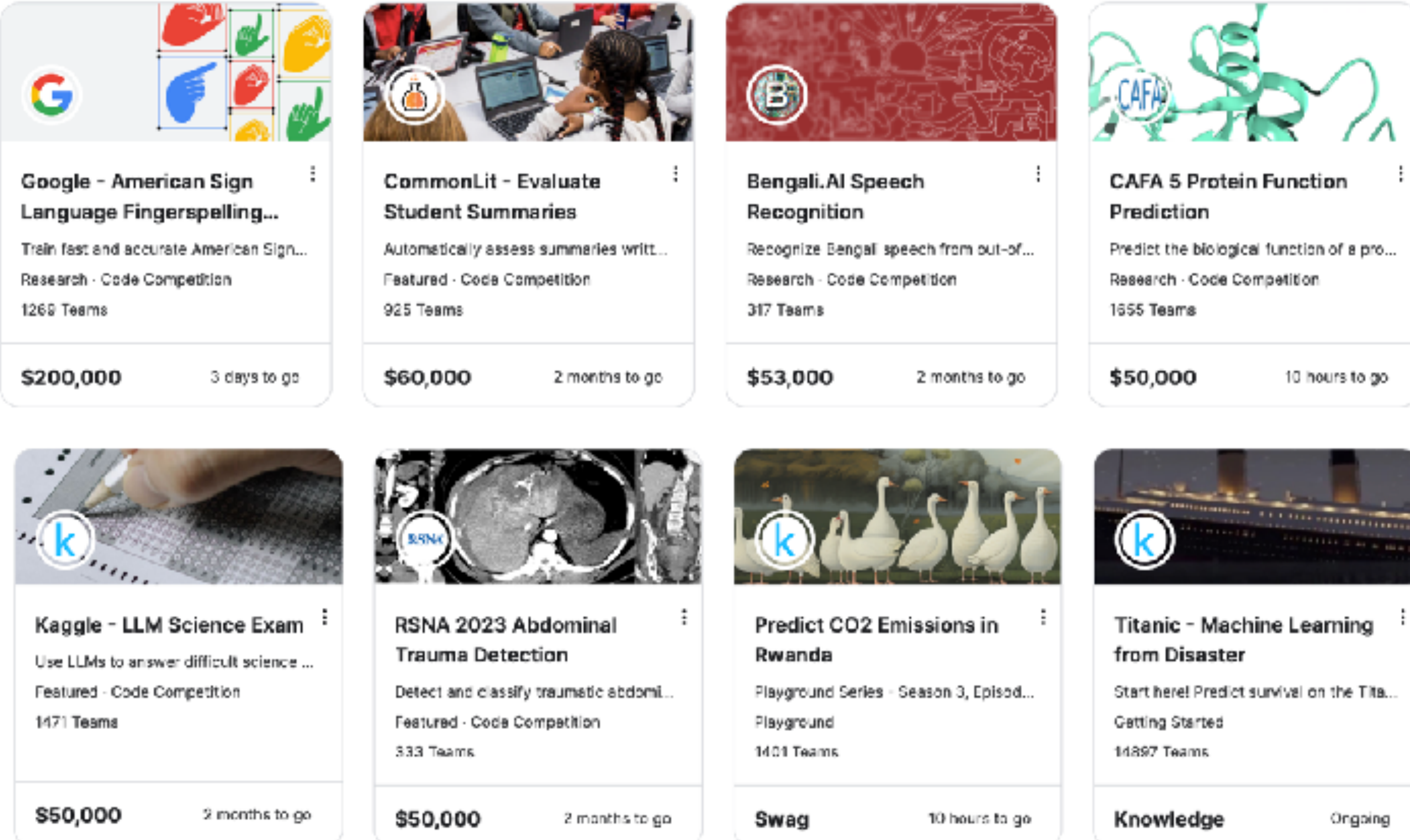
section 1, manipulated from Tan et al. Introduction to Data Mining

Problem Types in Machine Learning

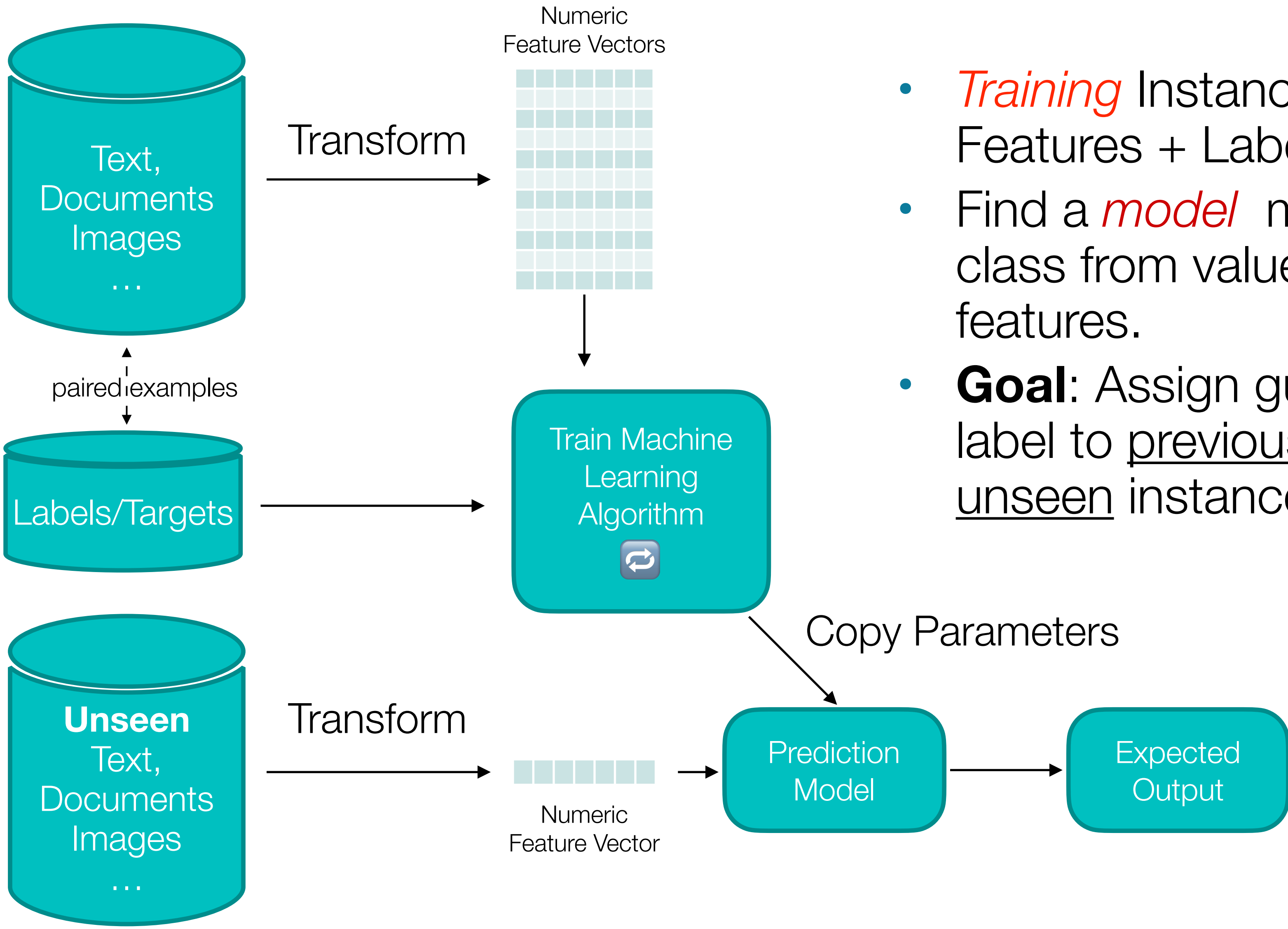
- Inputs
- Outputs



Problem Types in Machine Learning



Classification and Regression, Supervised



- *Training* Instances: Features + Labels
- Find a *model* mapping class from values of features.
- **Goal**: Assign guessed label to <u>previously unseen</u> instances

Some Popular Datasets

ImageNet

1M+

224 x 224 Color Image
⬇
1000 Classes
(prominent object)

MNIST

60k

24 x 24 Grey Image
⬇
10 Classes (digits)

Adult

# feature	original feature
1	age
2	workclass
3	final weight
4	education
5	ed_num
6	marital_status
7	occupation
8	relationship
9	race
10	sex
11	capital_gain
12	capital_loss
13	hours x week
14	country

5k

Census Demographics
⬇
Binary (salary > 50k?)

CoCo

200k Images

Large, Multi-sized Images
⬇
Location, Size, 80 Objects

Boston Housing

1. CRIM
2. ZN
3. INDUS
4. CHAS
5. NOX
6. RM
7. AGE
8. DIS
9. RAD
10. TAX
11. PTRATIO
12. B
13. LSTAT
14. MEDV

House/Neighborhood
Descriptions
⬇
House Price
$$

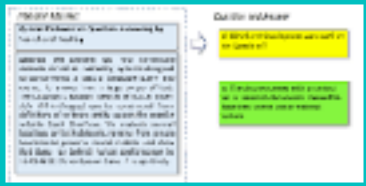
500 Examples

Translation

Language A
⬇
Language B

Many datasets

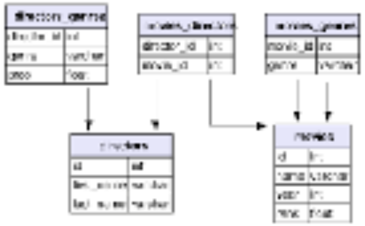
SQuAD



Question
⬇
Answer

100k+

Imdb



Movie/Actors/Director/+
⬇
Critic/Audience rating

50k reviews

Self Test

- **A. Classification**
  B. Regression
  C. Not Machine Learning
- **D. Machine Learning Generation**

- Dividing up customers by potential profitability?

- Extracting frequency of sound?

Before Next Lecture

- Before next class:
  - install python (3.8 preferred) on your laptop
  - install anaconda distribution of python
  - use virtual environments (`conda env`)

- Look at Python primer if you need review
  - I made ~4 hours of YouTube content…
  - https://www.youtube.com/playlist?list=PL7lPdRN5E0YKCnVl-fvx8jOOCWVeGTsrV

Opening Demo: Jupyter Notebooks

`01_Numpy and Pandas Intro.ipynb`

Lecture Notes for

Machine Learning in Python

Professor Eric Larson

Introduction, Syllabus, Data Types

Lecture Notes for

Machine Learning in Python

⟦👨‍🏫, 👨‍💻, 🐍, 👨‍🔬⟧

Professor Eric Larson

Table Data using Numpy, Pandas

Class Logistics and Agenda

- Canvas? Anaconda Installs?
- In-person versus Zoom and other classes
- Agenda:
  - Data Encodings
  - Demo: Table Data, Numpy
  - Data Quality
  - Attributes Representation
    - documents
  - The Pandas eco-system
    - loading and manipulating attributes

Class Overview, by topic

Table Data Visualization

Numpy, Pandas, Seaborn
Overviews with some in-depth discussion

Dimension Reduction and Image Processing

Scikit-learn, Scikit Image,
Intuition only, Some mathematics

Linear and Logistic Regression

Numpy, Recreate API for Scikit-learn
Detailed mathematics for simple optimization
intuition for advanced optimization

Neural Networks and Back Prop.

Numpy
Detailed mathematics for NN operations

Wide and Deep Networks

Convolutional Networks

Recurrent Networks

Keras, Tensorflow
Intuition, Detailed implement.

Ethics in Language Models

ConceptNet
Case studies

Opening Demo:
Jupyter Notebooks

`01_Numpy and Pandas Intro.ipynb`

Types of Data and Categorization

Table Data

- **Table Data**: Collection of data **instances** and their **features**
  - **Python:** Pandas Dataframe
  - **R:** Data.frame
  - **Matlab:** Table Class
  - **C++**: Trick Question

Attributes, columns, variables, fields, characteristics, **Features**

Objects, records, rows, points, **samples**, cases, entities, **instances**

TID	Pregnant	BMI	Age	Diabetes
1	Y	33.6	41-50	positive
2	N	26.6	31-40	negative
3	Y	23.3	31-40	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	21-30	negative
9	N	30.5	51-60	positive
10	Y	37.6	51-60	positive

Feature Vector Representation

	Attribute	Representation Transformation	Comments
Discrete	**Nominal**	Permutation of values only. **one hot encoding or hash function**	If all **employee ID** numbers were reassigned, would it make any difference?
Discrete	**Ordinal**	Order must be preserved `new_value = f(old_value)` where f is a monotonic function. **integer**	An attribute encompassing the notion of **good, better best** can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Continuous	**Interval**	`new_value =f(old_value) + b` f is monotonic through origin **float**	Thus, the **Fahrenheit** and **Celsius** temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Continuous	**Ratio**	`new_value = f(old_value)` f is monotonic through origin **float**	**Length** can be measured in meters or feet, but **zero is zero**

from Tan et al. Introduction to Data Mining

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
4	N	28.1	21-30	brown	inconclusive
5	N	43.1	31-40	blue	positive
6	Y	25.6	21-30	hazel	negative

Internal Rep.

TID
1
2
3
4
5
6

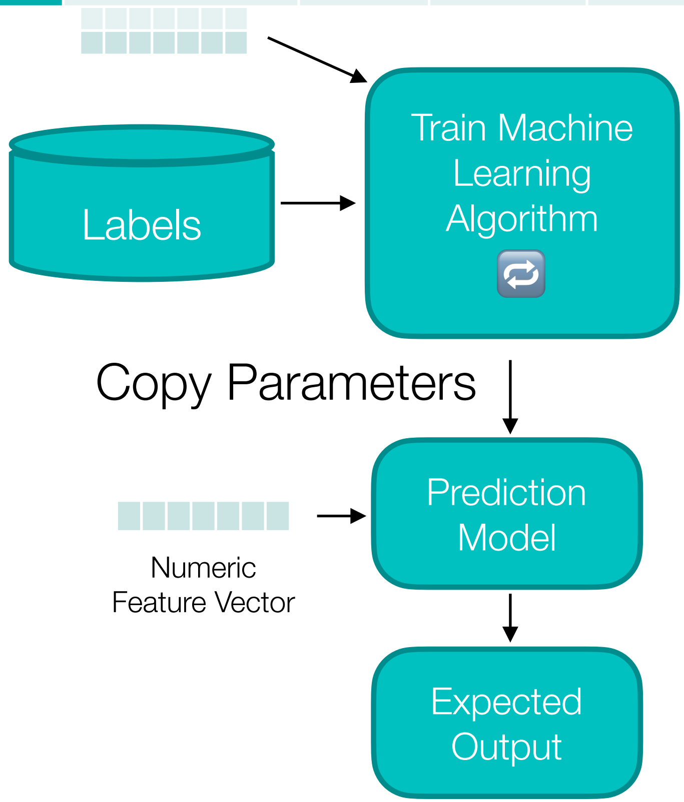
"Finish" Jupyter Notebooks

`01_Numpy and Pandas Intro.ipynb`

Data Quality

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



- 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

Handling Issues with Data Quality

- **Eliminate** Instance or Feature
- **Ignore** the Missing Value During Analysis Replace with all possible values (talk about later)
- **Impute** Missing Values

How?

Stats?
mean
median
mode

Imputation

- When is it probably fine to impute missing data:
  - (A) When there is not much missing data
  - (B) When the missing feature is mostly predictable from another feature
  - (C) When there is not much missing data for each subgroup of the data
  - (D) When it is the class you want to predict

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

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

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

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.

For Next Lecture

- Before next class:
  - verify installation of seaborn, plotly, (and/or bokeh if you want)
  - look at pandas table data and additional tutorials

- Next time: Documents, Data Imputation Demo

Lecture Notes for

Machine Learning in Python

Professor Eric Larson

Table Data using Numpy, Pandas

Lecture Notes for

Machine Learning in Python

[👨‍🏫, 👨‍💻, 🐍, 👨‍🔬]

Professor Eric Larson

Data Quality and Imputation

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

Class Overview, by topic

Table Data Visualization
Numpy, Pandas, Seaborn
Overviews with some in-depth discussion

Dimension Reduction and Image Processing
Scikit-learn, Scikit Image,
Intuition only, Some mathematics

Linear and Logistic Regression
Numpy, Recreate API for Scikit-learn
Detailed mathematics for simple optimization
intuition for advanced optimization

Neural Networks and Back Prop.
Numpy
Detailed mathematics for NN operations

Wide and Deep Networks | **Convolutional Networks** | **Recurrent Networks**
Keras, Tensorflow
Intuition, Detailed implement.

Ethics in Language Models

ConceptNet
Case studies

Last Time

Data Quality Problems

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

TID	Hair Color	Height	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

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

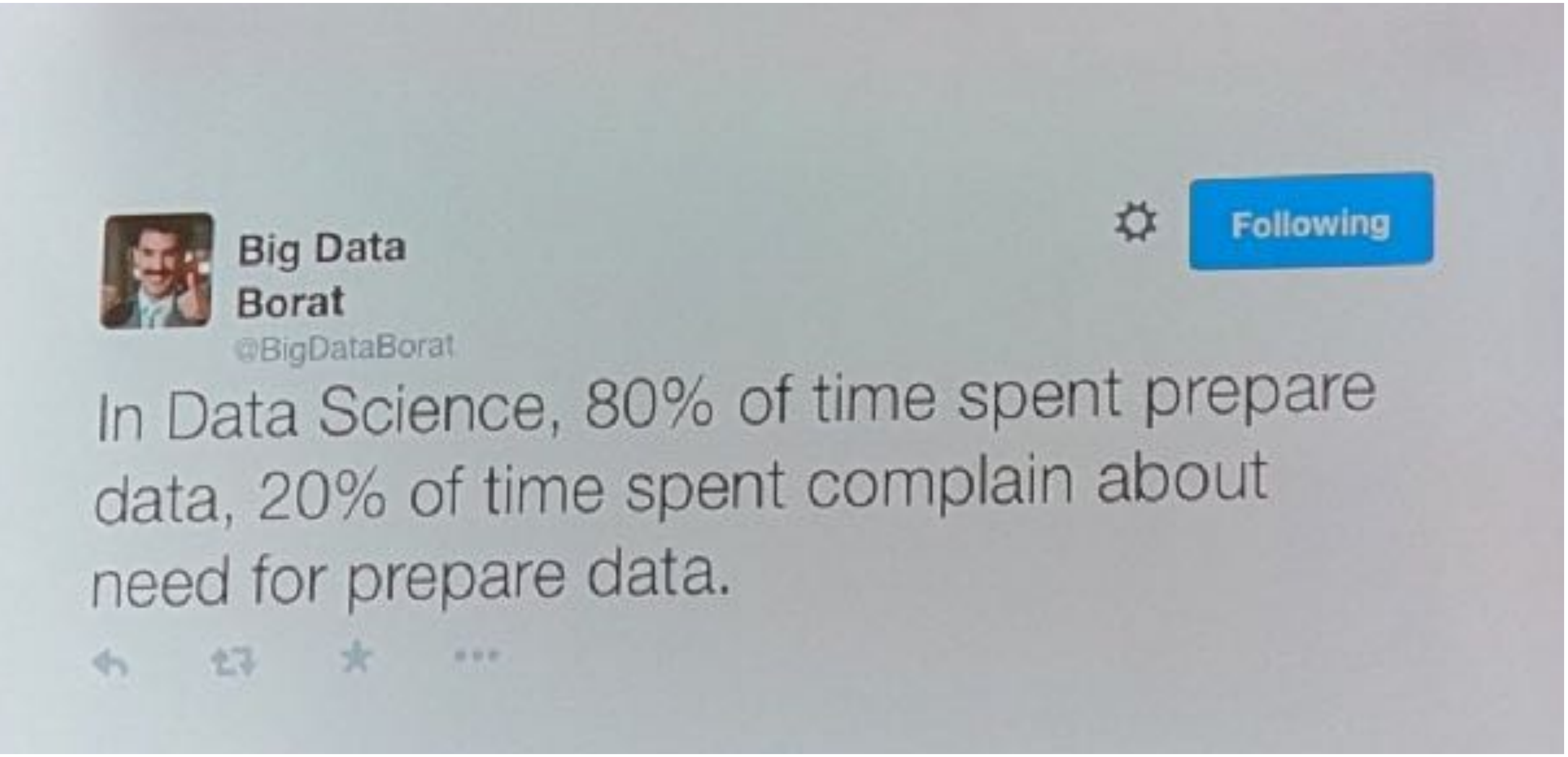
TID	Pregnant	BMI	Age	Diabetes	Distance
3	Y	23.3	?	positive	0
6	Y	25.6	21-30	negative	(0 + 2.3 + 1)/3
2	N	26.6	31-40	negative	(1 + 3.3 + 1)/3
4	?	28.1	21-30	negative	(4.8 + 1)/2

Imputed Age: 21-30

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.

Data Representation and Documents



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
4	N	28.1	21-30	brown	inconclusive
5	N	43.1	31-40	blue	positive
6	Y	25.6	21-30	hazel	negative

Internal Rep.

TID
1
2
3
4
5
6

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
4	N	28.1	21-30	brown	inconclusive
5	N	43.1	31-40	blue	positive
6	Y	25.6	21-30	hazel	negative

Internal Rep.

TID	Binary	Float	Ordinal	Object	Diabetes
1	1	33.6	2	hash(0)	1
2	0	26.6	1	hash(1)	0
3	1	23.3	1	hash(2)	1
4	0	28.1	0	hash(0)	2
5	0	43.1	1	hash(2)	1
6	1	25.6	0	hash(1)	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	Slee		tc	Stev
1	0		.86	0
2	0.1		.02	0
3	0.1		.1	0

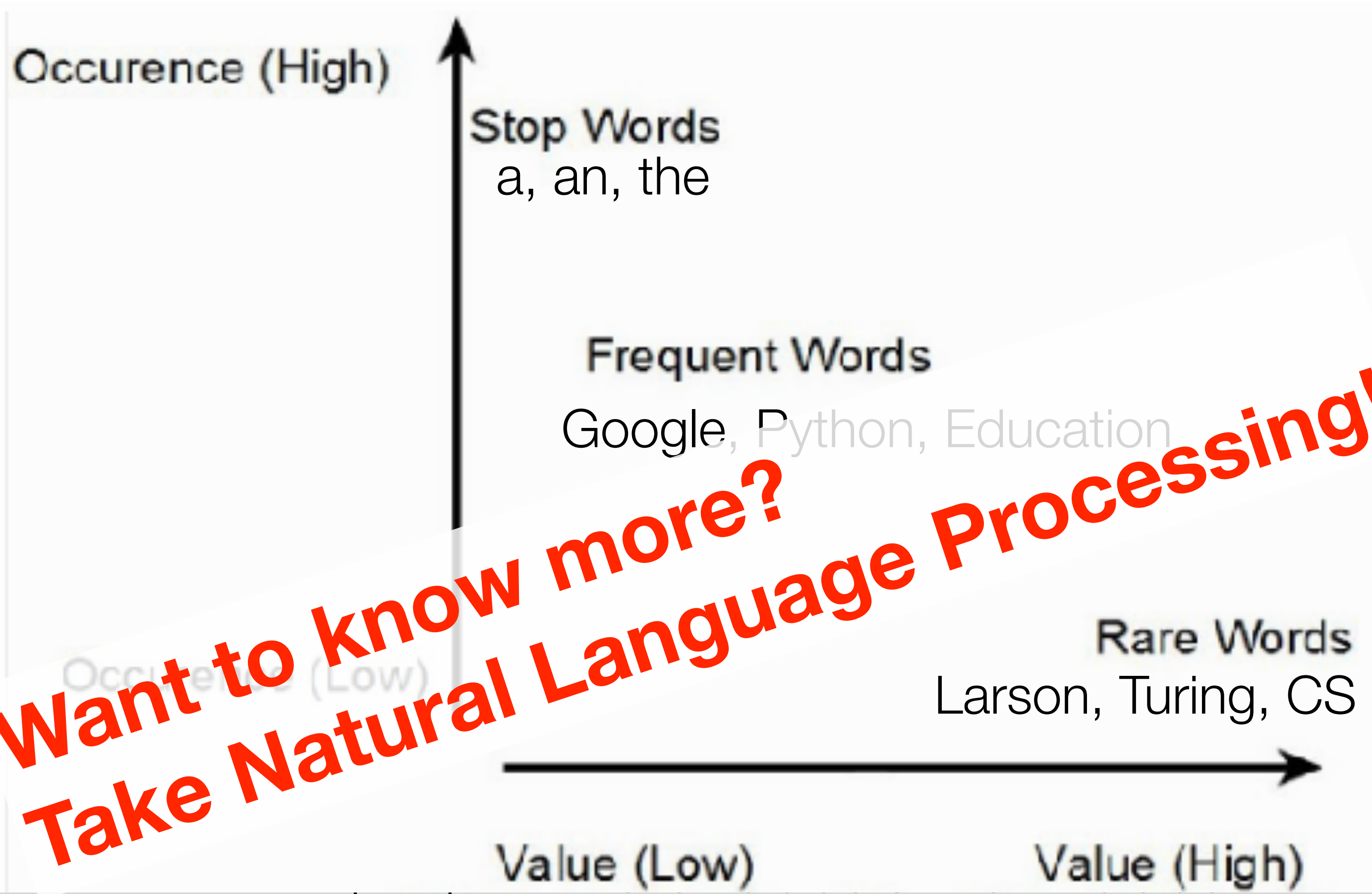
te $\in D$

in d"

inverse

idf($\in d$

t"



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

Want to know more?
Take Natural Language Processing!

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

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

For Next Lecture

- Before next class:
  - verify installation of seaborn, plotly, (and/or bokeh if you want)
  - look at pandas table data and additional tutorials

- Next time: Data Visualization

Lecture Notes for

Machine Learning in Python

Professor Eric Larson

Data Quality and Imputation