

# Lecture Notes for **Machine Learning in Python**

Professor Eric Larson  
**Introduction, Syllabus, Data Types**

# Class Logistics and Agenda

- Agenda:
  - Introductions
  - Syllabus and Course Overview
  - What is Machine Learning?
  - Types of Data
  - Numpy Demo
- My approach to this course:
  - Programming
  - Math
  - **Applications** and **Analytics**

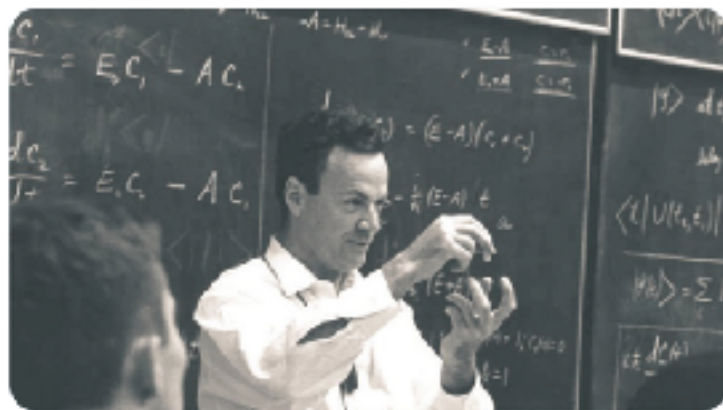
# Introductions & Course Syllabus



**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.



# Introductions

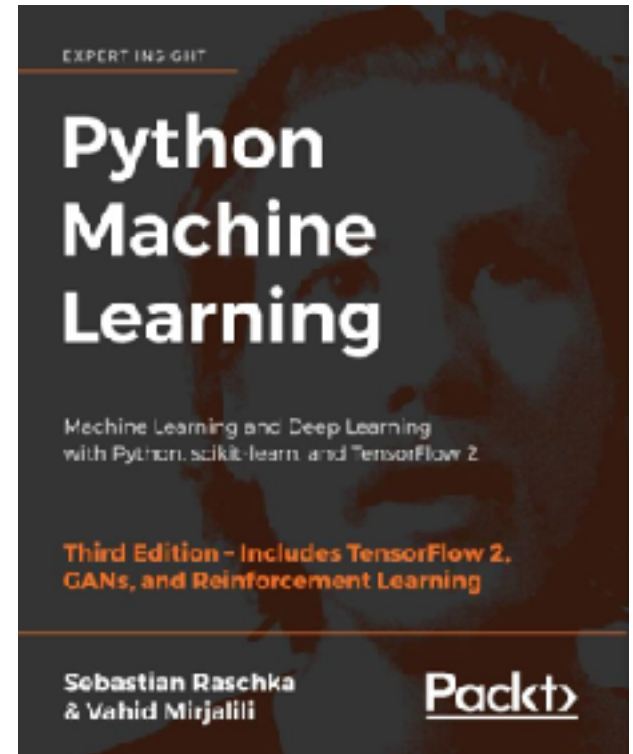
- Me
  - Eric 👍
  - Dr. Larson 👍
  - Prof. Larson 👍
  - Other 👎
- You
  - Name
  - Where you grew up
  - Department
  - Grad/Undergrad
  - Something true or false

## **Choices:**

High School  
College Job  
ML Instructor  
Marvel  
Star Wars  
Kids

# FAQ

- Text:
  - **Recommended:** Python Machine Learning, Raschka & Mirjalili, Second edition
- Use Canvas for posted course material
- Prerequisites:
  - Linear algebra, calculus (multivariate)
  - Basic statistics and probability
  - Basic python programming
- Version of python: 3.X
  - Install through Anaconda
  - Use conda environments
  - JupyterLab (or notebooks)
- Most Used Libraries: Numpy, Pandas, Scikit-Learn, Matplotlib, Keras with Tensorflow



# How will participation be graded?

- Participation will be graded in the course:
  - **Distance students** will answer these questions via **canvas upload**
  - upload “over” the last submission
    - must upload the questions throughout semester for full credit
- In Class Students:
  - Choose to respond to the question:
    - Do you think this will work?
      - A: **Yes** this is going to work
      - B: This is **not** going to work
      - C: I needed a card.
      - D: I added an Alias to my card

# Canvas Syllabus

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

# 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 statment (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!



# Machine Learning Overview



**Steve Fawley**

@stevehasnoweave

My psych professor asked if we'd heard of Pavlov. I said "it rings a bell." No one laughed, I'm too witty for this class.

# 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](https://www.whatis.techtarget.com/definition/machine-learning)

*About this result • Feedback*

- Beware:
  - full of imprecise words
  - words that play on our understanding of “learning” and consciousness

# Machine Learning

One Small Piece of Artificial Intelligence

Data Mining

ML

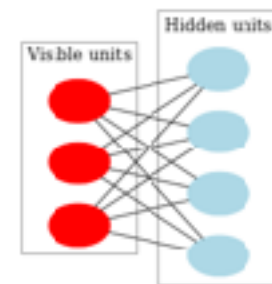
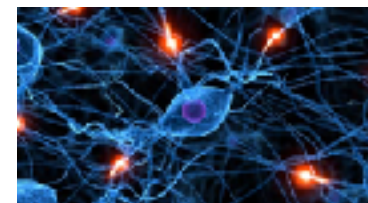
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables
- Description Methods
  - Find human-interpretable patterns that describe the data.

ML

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

# Abridged History of Machine Learning

- Historically builds from disciplines statistics and computer science (algorithms)
- At present: Its really just algorithms for optimizing weights 🤖
- **1952**: Arthur Samuel IBM creates checker program
- **1957**: Rosenblatt, Neural Network Perceptron
- **1967**: Nearest Neighbor Pattern Recognition
- **1970's**: AI Winter
- **1990's**: Volley of “New” Machine learning Algorithms
- **2001**: Breiman's Random Forests
- ~**2004**: Modern Support Vector Machines with Kernels
- **2005**: Second AI Winter
- ~**2010**: Deep Learning Convolutional Networks
- **2015**: Deep Learning becomes buzz word,
- **Modern Day**: you hear about it and take this course



# Problem Types in Machine Learning





kaggle

Customer Solutions

Competitions

Community ▾

Active Competitions

		<b>Click-Through Rate Prediction</b> Predict whether a mobile ad will be clicked	21 days 1512 teams \$15,000
		<b>National Data Science Bowl</b> Predict ocean health, one plankton at a time	56 days 430 teams \$175,000
		<b>Driver Telematics Analysis</b> Use telematic data to identify a driver signature	56 days 686 teams \$30,000

# Example Classification: Malware

- Classify files as malware based on structure, size, and naming.
- Approach:
  - ♦ Use already classified malware files
  - ♦ Must translate name to set of features
  - ♦ **{malware, not malware}** decision forms the **class attribute**
  - ♦ Collect various malware examples and a number of safe files, providing labels for each and a set of features

## Training Set

TID	Name	Size	Class
1	erte.dll	916 b	not
2	fufu.bin	1M	yes
3	exe.exe	1G	not
4	ex.py	113 b	not

## Unknown

<i>TID</i>	<i>Name</i>	<i>Size</i>
<b>1</b>	asdf.dll	11b

# Example Regression: Housing Price

- Predict a value of a given *continuous valued* variable based on the values of other variables
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Predicting House Sales

## Training Set

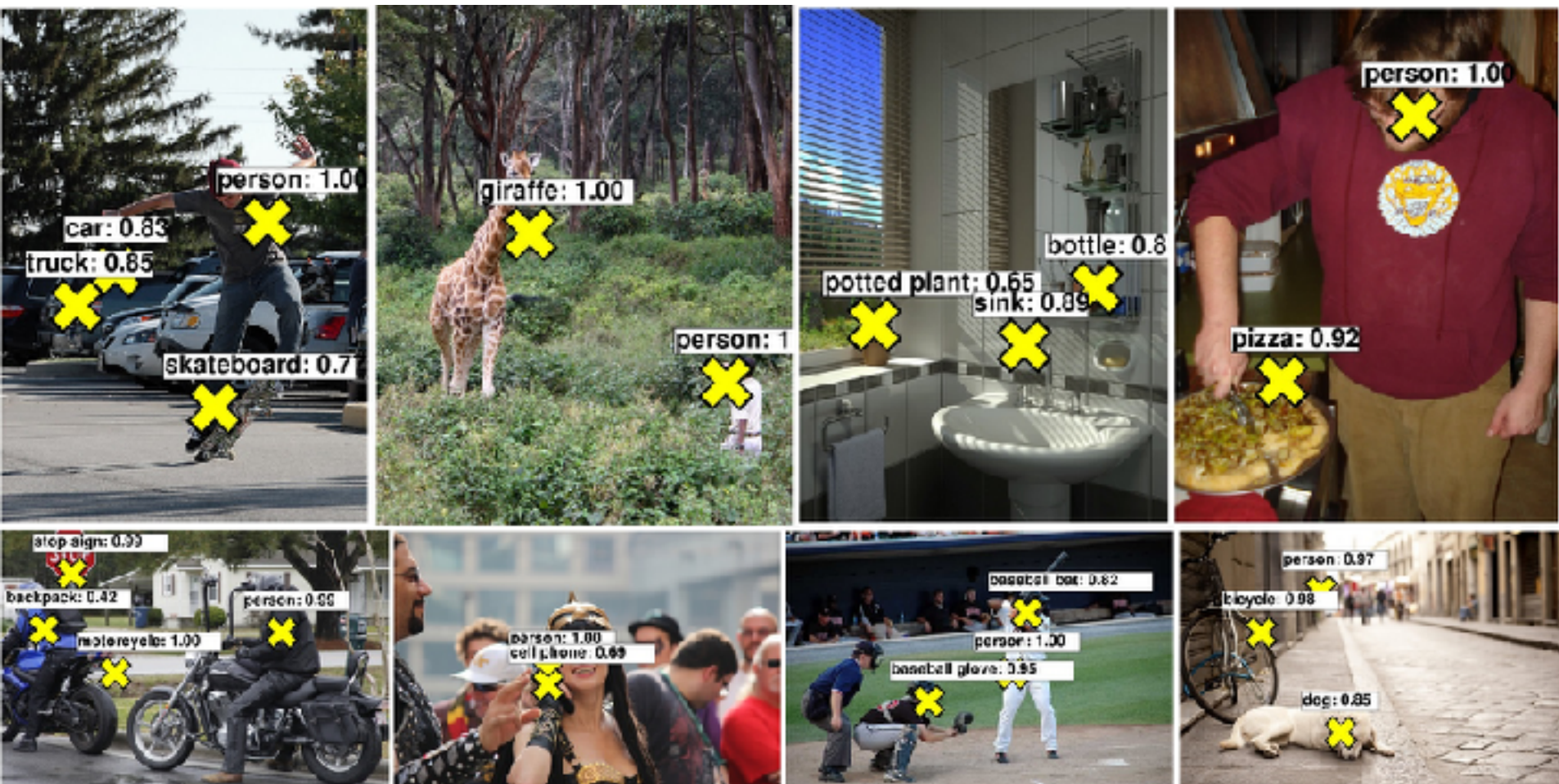
<i>TI</i>	<i># Rms</i>	<i>Sq Ft</i>	<i>Zip</i>	<i>Price</i>
<b>1</b>	2	1125	74012	150K
<b>2</b>	2	2525	75155	200k
<b>3</b>	10	4678	90210	3M
<b>4</b>	4	2678	75154	350k

## Unknown

<i>TI</i>	<i># Rms</i>	<i>Sq Ft</i>	<i>Zip</i>
<b>1</b>	2	2200	75115



# Example Classifying: Objects in Images



## Image Net:

- 14 million images
- 200 Labeled Categories
- 1000 Location Labels

## Attributes:

- Images



# Self Test

- **A. Classification**  
**B. Regression**  
**C. Not Machine Learning**
- Dividing up customers by potential profitability?
- Extracting frequency of sound?

# Types of Data and Categorization

Optimist



The glass is  
half full

Pessimist



The glass is  
half empty



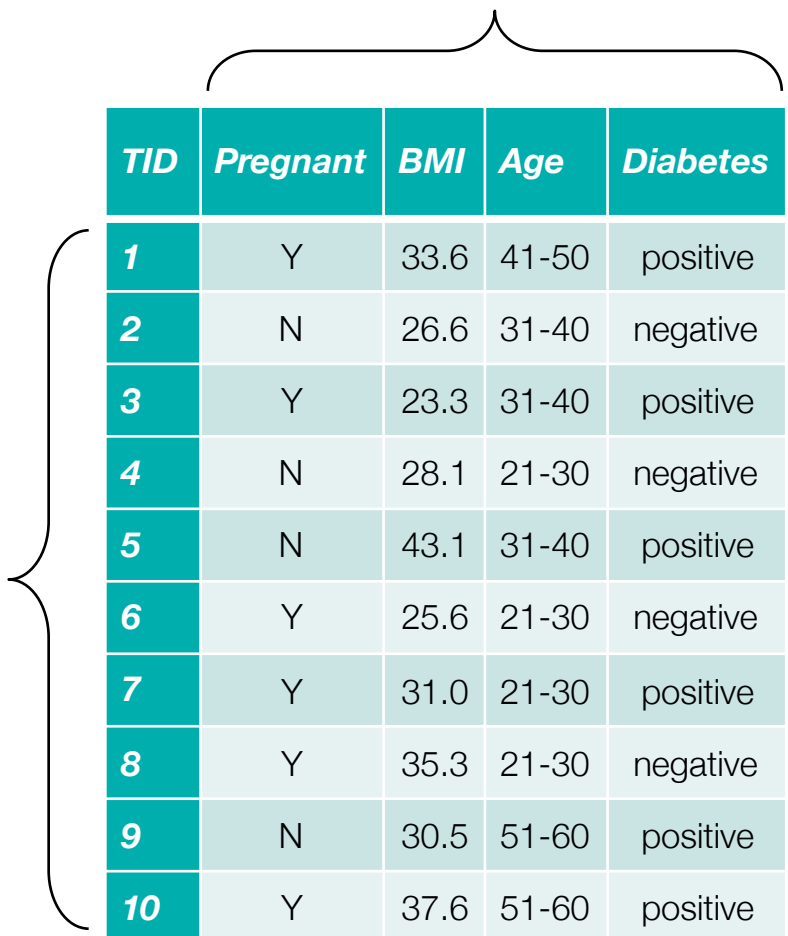
Closed  
as subjective

# Table Data

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

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

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



<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Diabetes</i>
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 Type Representation

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

# Self Test

- Are these **A. ordinal, B. nominal, or C. binary**?
  - military rank
    - ordinal
  - coat check number
    - nominal

# Before Next Lecture

- Before next class:
  - install python on your laptop
  - install anaconda distribution of python
- Look at Python primer if you need review
  - I made ~4 hours of YouTube content...
  - <https://www.youtube.com/playlist?list=PL7IPdRN5E0YKCnVI-fvx8jOOCWVeGTsrV>

**If time:  
Jupyter Notebooks**



`01_Numpy and Pandas Intro.ipynb`

# Lecture Notes for **Machine Learning in Python**

Professor Eric Larson  
**Numpy, Pandas, Document Features**



# Class Logistics and Agenda

- Canvas? Anaconda Installs?
- Distance transfers?
- Agenda:
  - Numpy
  - Data Quality
  - Attributes Representation
    - documents
  - The Pandas eco-system
    - loading and manipulating attributes
- Needing some more help?
  - **fast.ai** has great, free resources

## **“Finish” Jupyter Notebooks and Numpy**

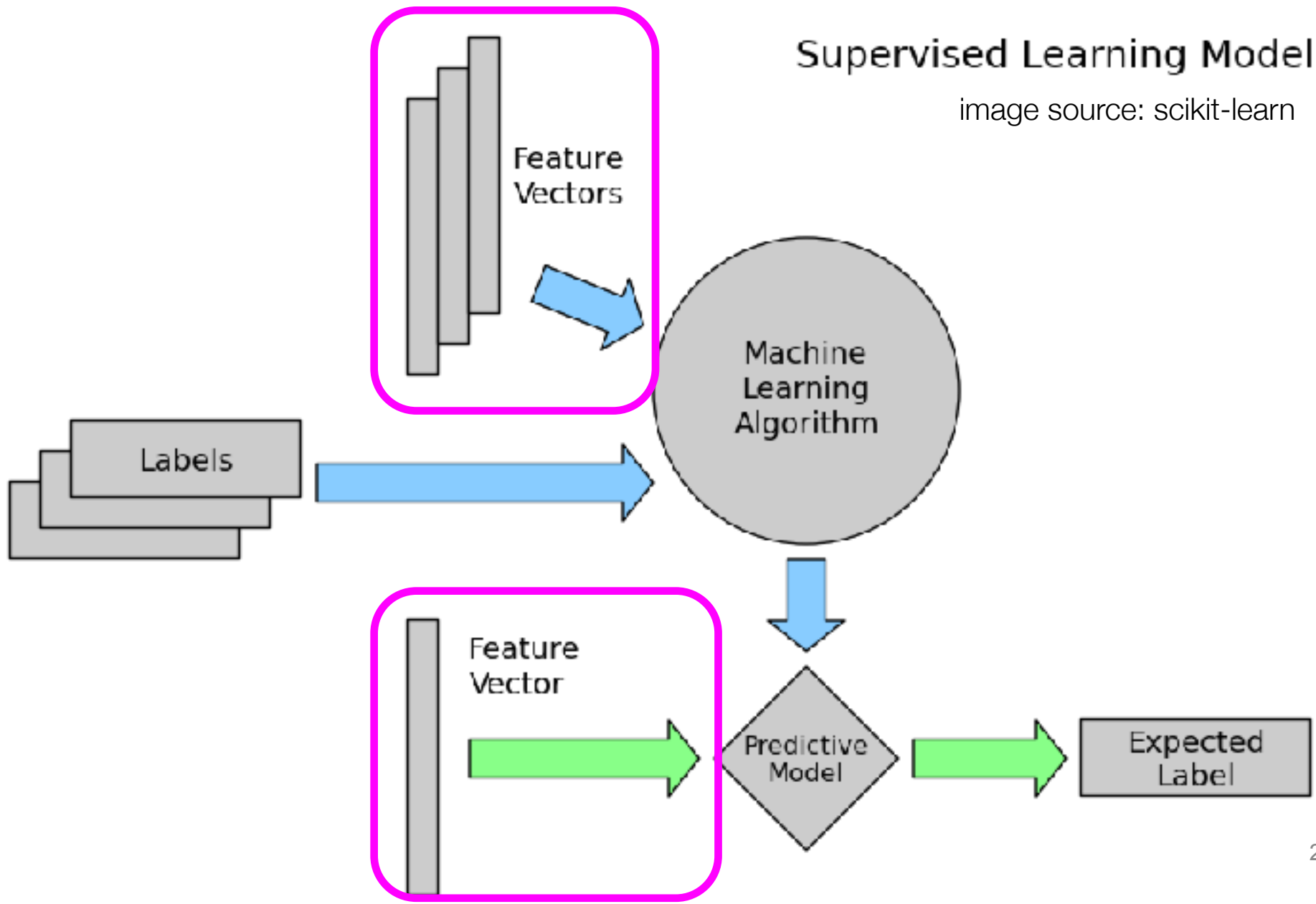


`01_Numpy and Pandas Intro.ipynb`

# Data Quality



# Review of Feature Data



# Data Quality Problems

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

<i>TID</i>	<i>Hair Color</i>	<i>Height</i>	<i>Age</i>	<i>Arrested</i>
<b>1</b>	Brown	5'2"	23	no
<b>2</b>	Hazel	1.5m	12	no
<b>3</b>	Bl	5	999	no
<b>4</b>	Brown	5'2"	23	no

# 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

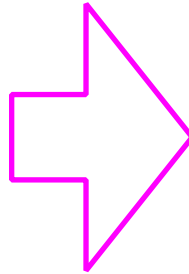
How?

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



<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Diabetes</i>
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

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

Mode: none, can't impute

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

Mode: 21-30



# K-Nearest Neighbors Imputation

<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Diabetes</i>
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

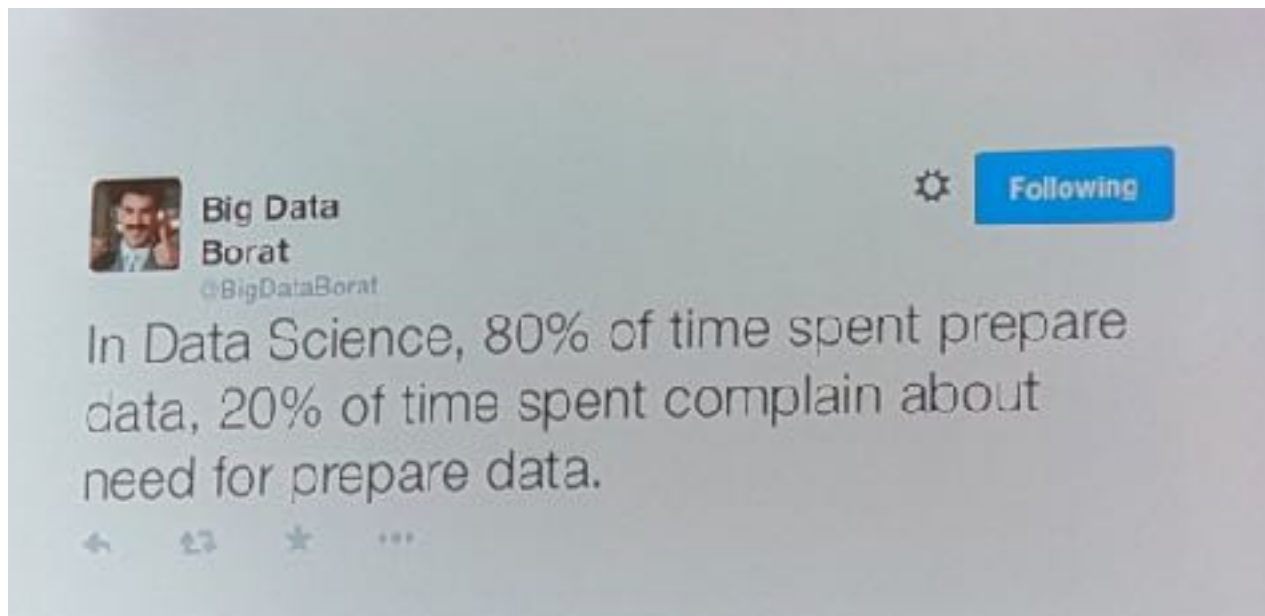
For K=3, find 3 closest neighbors

<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Diabetes</i>	<i>Dist</i>
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	$(4.8+1)/2$

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



# Feature Type Representation Review

	Attribute	Representation Transformation	Comments
Discrete	Nominal	Any permutation of values  <b>one hot encoding</b>	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., $\text{new\_value} = f(\text{old\_value})$ where $f$ is a monotonic function.  <b>integer</b>	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	$\text{new\_value} = a * \text{old\_value} + b$ where $a$ and $b$ are constants  <b>float</b>	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	$\text{new\_value} = a * \text{old\_value}$ <b>float</b>	Length can be measured in meters or feet.

# Data Tables as Variable Representations

Table

<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Eye Color</i>	<i>Diabetes</i>
<b>1</b>	Y	33.6	41-50	brown	positive
<b>2</b>	N	26.6	31-40	hazel	negative
<b>3</b>	Y	23.3	31-40	blue	positive
<b>4</b>	N	28.1	21-30	brown	inconclusive
<b>5</b>	N	43.1	31-40	blue	positive
<b>6</b>	Y	25.6	21-30	hazel	negative

Internal Rep.

<i>TID</i>
<b>1</b>
<b>2</b>
<b>3</b>
<b>4</b>
<b>5</b>
<b>6</b>

# Data Tables as Variable Representations

Table

<i>TID</i>	<i>Pregnant</i>	<i>BMI</i>	<i>Age</i>	<i>Eye Color</i>	<i>Diabetes</i>
<b>1</b>	Y	33.6	41-50	brown	positive
<b>2</b>	N	26.6	31-40	hazel	negative
<b>3</b>	Y	23.3	31-40	blue	positive
<b>4</b>	N	28.1	21-30	brown	inconclusive
<b>5</b>	N	43.1	31-40	blue	positive
<b>6</b>	Y	25.6	21-30	hazel	negative

Internal Rep.

<i>TID</i>	<i>Binary</i>	<i>Float</i>	<i>Ordinal</i>	<i>Object</i>	<i>Diabetes</i>
<b>1</b>	1	33.6	2	hash(0)	1
<b>2</b>	0	26.6	1	hash(1)	0
<b>3</b>	1	23.3	1	hash(2)	1
<b>4</b>	0	28.1	0	hash(0)	2
<b>5</b>	0	43.1	1	hash(2)	1
<b>6</b>	1	25.6	0	hash(1)	0

# Bag of words model

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

Bag of Words

Vocabulary						
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

# Feature Hashing

what happens when we get more words?

TID	Slee	Fati	Wei	Ras	First	Sigh	Why	Fox	Bro	Lazy	Dog	Etc	Stev
1	0	1	0	0	2	0	0	0	0	1	0	2	0
2	1	1	0	0	1	1	0	0	4	0	1	3	0
3	1	1	0	2	1	1	1	0	1	0	0	1	0

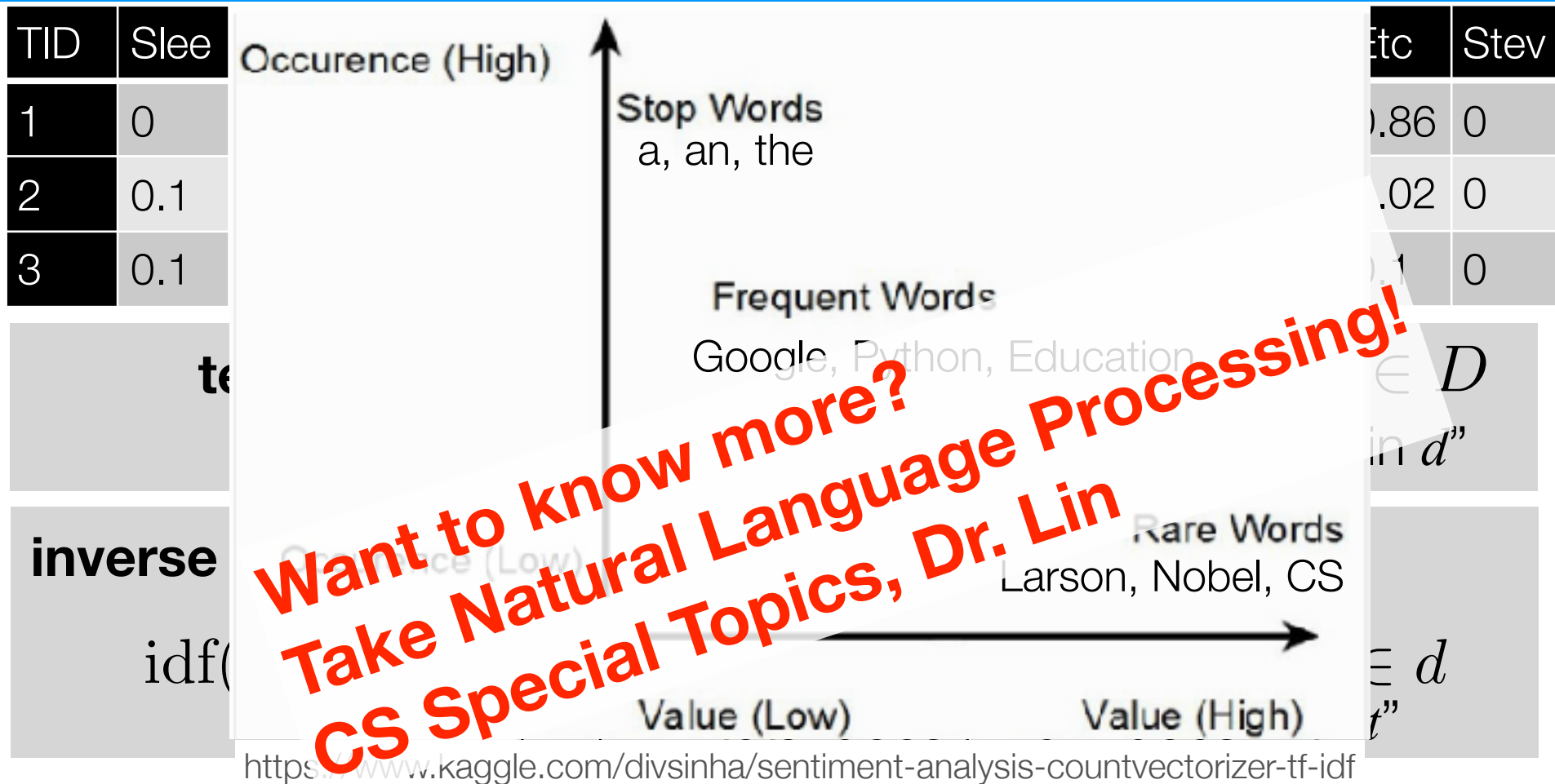
or we could have a hashing function,  $h(x) = y$

TID	$h(x)=1$	$h(x)=2$	$h(x)=3$	$h(x)=4$	$h(x)=5$	$h(x)=6$
1	0	1	0	1	2	0
2	1	1	4	0	2	1
3	2	1	1	2	1	1

multiple words mapped to one feature

(want to minimize collisions or make collisions meaningful)

# Term-Frequency, Inverse-Document-Frequency



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

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



Pandas and Imputation  
Scikit-Learn



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:
  - install seaborn
  - install plotly (or bokeh if you want)
  - look at pandas table data and look at additional tutorials
- Next Week: Data Visualization