

# Introduction to Data Science

## Math 4100 / COMP 5360

### Lecture 29: Final Projects, Recap, Outlook



THE  
UNIVERSITY  
OF UTAH



IN CS, IT CAN BE HARD TO EXPLAIN  
THE DIFFERENCE BETWEEN THE EASY  
AND THE VIRTUALLY IMPOSSIBLE.

# Final Projects

# Project Peer Assessment

It is important to provide positive feedback to people who truly worked hard for the good of the team and to also make suggestions to those you perceived not to be working as effectively on team tasks. We ask you to provide an honest assessment of the contributions of the members of your team, including yourself. The feedback you provide should reflect your judgment of each team member's:

- Preparation – were they prepared during team meetings?
- Contribution – did they contribute productively to the team discussion and work?
- Respect for others' ideas – did they encourage others to contribute their ideas?
- Flexibility – were they flexible when disagreements occurred?

Final Project Peer Feedback

\* Required

Project Name \*

Your answer

Your Name \*

Your answer

How did you perform in the project, how much did you contribute?\*

1 2 3 4 5

Your teammate's assessment of your contributions and the accuracy of your self-assessment will be considered as part of your overall project score.

# **Best Projects: The Process**

Each staff member nominates 1/2 of their projects

All TAs meet, watch all videos, look at the notebooks, and discuss which ones get a nomination

Top Three Elicited in Discussion

# The Results

Three Winners!

120% of points

7 Honorable Mentions

110% of points

For all: listed in “Hall of Fame” on website

# The Honorable Mentions

In no particular Order

# **Analysis of Temporal Changes in Bird Population Sizes in the United States**

Robben Migacz, Michael Quigley

<https://youtu.be/dhP9RSj6teo>

# Get Me Outta Here

Elizabeth Maynard, Anne Senay

[https://www.youtube.com/watch?  
v=7YhTXPQWCFQ&feature=youtu.be](https://www.youtube.com/watch?v=7YhTXPQWCFQ&feature=youtu.be)

# **Correlating near surface wind speeds on ice sheet mass balance in West Antarctica**

Eric Johnson, Durban Keeler

[https://drive.google.com/file/d/1E\\_DKSimxT58BydEHJr64Dy\\_SXZN7onq1/view?  
usp=sharing](https://drive.google.com/file/d/1E_DKSimxT58BydEHJr64Dy_SXZN7onq1/view?usp=sharing)

# **Declassifying the Classifieds**

## **(Detecting Media Bias)**

Patrick Sullivan, Todd Christensen, Carlie Davis

[https://www.youtube.com/watch?  
v=FS4vAApXAgU&lc=UgzKM-iByg0OXqOfOUd4AaABA](https://www.youtube.com/watch?v=FS4vAApXAgU&lc=UgzKM-iByg0OXqOfOUd4AaABA)

# **Accident Predictions in Utah**

Kyle Cornwall, Shushanna Mkrtychyan

<https://vimeo.com/409643685>

# **Let's Talk About Women's Health: Finding Endometriosis in Narrative Text**

Amber Kiser

[https://www.youtube.com/watch?  
v=h6YsKKEvl34&feature=youtu.be](https://www.youtube.com/watch?v=h6YsKKEvl34&feature=youtu.be)

# Good Vibes?

Paul Mundt, Katie Jones

[https://www.youtube.com/watch?  
v=Wt8SnJ4hdvU&feature=youtu.be](https://www.youtube.com/watch?v=Wt8SnJ4hdvU&feature=youtu.be)

# The Winners

In no particular Order

# **Classification of the invasive plant *Phragmites australis* in the Howard Slough Waterfowl Management Area using TensorFlow and Neural Networks**

Troy Saltiel, Greg Lee

<https://www.dropbox.com/s/vcragber3zf5u7/Data-Science-Project.mp4?dl=0>

# KSL AutoMoDeals

Michael Paskett, Chantel Charlebois, Taylor Hansen

<https://youtu.be/KEYOVU5K76Q>

# Lyrical Analysis by Popularity

Etahn Burrows, Tanner Sims, Lass Omar

<https://www.youtube.com/watch?v=23HPKX0BToA>

What did we learn?

# Programming! Python!

What is a for loop? What is a function?

Jupyter Notebooks

Python Libraries:

numpy, scipy, statsmodels, scikit-learn,  
NLTK, networkx, TensorFlow, keras

Data Structures:

Lists, Dictionaries, Series, Pandas Data  
Frames

Data Cleanup, Aggregation,  
Reshaping!

Version Control - Git & GitHub

## Masking and Filtering

With pandas we can create boolean arrays that we can use to mask and filter new array that has "True" for every band formed after 1964:

```
In [68]: mask = bands_founded > 1964  
mask
```

```
Out[68]: Beatles      False  
Zeppelin     True  
Pink Floyd   True  
Pink Floyd   True  
Name: Bands founded, dtype: bool
```

This uses a technique called **broadcasting**. We can use broadcasting with various operations.

```
In [69]: # Not particularly useful for this dataset..  
founding_months = bands_founded * 12  
founding_months
```

```
Out[69]: Beatles      23520  
Zeppelin     23616  
Pink Floyd   23580  
Pink Floyd   24180  
Name: Bands founded, dtype: int64
```

We can use a boolean mask to filter a series, as we've seen before:

# Getting Data

Web Scraping

HTML / Inspector

Retrieving Websites

Parsing HTML

APIs - REST, JSON

Data Bases - SQL

## Scraping with BeautifulSoup

[BeautifulSoup](#) is a Python library design for computationally extracting data from html documents. It supports navigating in the DOM and retrieving exactly the data elements you need.

Let's start with a simple example using the [lyrics.html](#) file.

```
In [6]: from bs4 import BeautifulSoup  
  
# we tell BeautifulSoup and tell it which parser to use  
soup = BeautifulSoup(open("lyrics.html"), "html.parser")  
# the output corresponds exactly to the html file  
soup
```

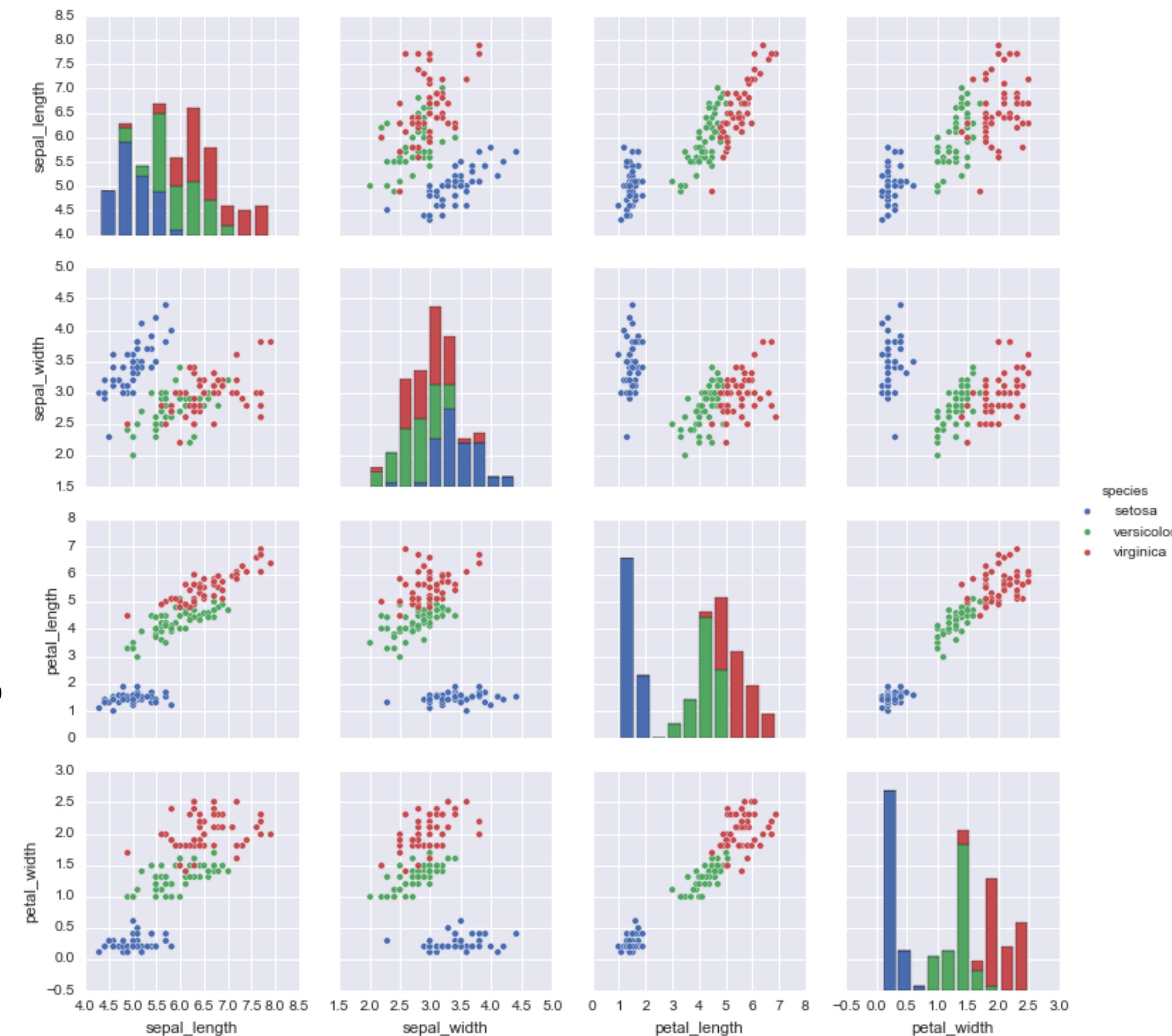
  

```
Out[6]: <!DOCTYPE html>  
  
<html lang="en">  
<head>  
<meta charset="utf-8">  
<title>Lyrics</title>  
</meta></head>  
<body>  
<article id="zep">  
<span class="date">Published: 2016-08-25</span>  
<span class="author">Led Zeppelin</span>  
<h1>Ramble On</h1>  
<div class="content">  
    Leaves are falling all around, It's time I was on my  
    way.
```

# Visualization

MatPlotLib, seaborn, and  
GraphX packages

Histograms, Scatterplot  
Matrices, Decision Trees,  
Force-Directed Graph Layouts



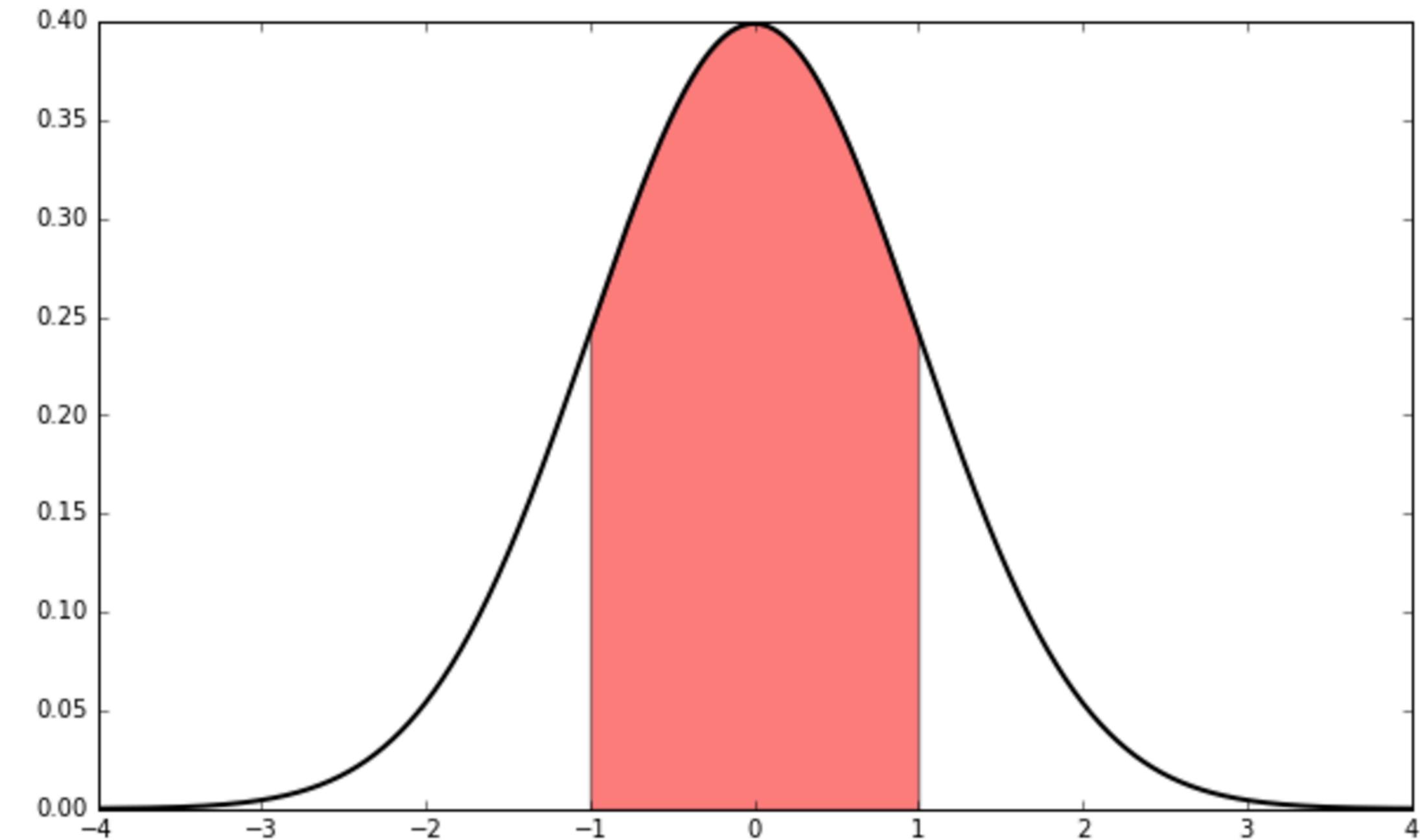
# Hypothesis Testing and Statistical Inference

Bernoulli, Binomial, and normal distributions

Hypothesis Significance Testing

A/B Testing

```
plt.plot(x, pdf, linewidth=2, color='k')
x2 = sc.arange(mu-sigma,mu+sigma,0.001)
plt.fill_between(x2, y1= norm.pdf(x2,loc=mu, scale=sigma), facecolor='red')
plt.show()
```



This integral can be computed using the *cumulative distribution function* (CDF)

$$F(x) = \int_{-\infty}^x f(x)dx.$$

# Time Series

Getting and Visualizing Stock Data

Moving averages

Trading Strategies

Pandas, Yahoo! Finance API

In [26]:

```
apple["50d"] = np.round(apple["Close"].rolling(window = 50, center = False).mean(), 2)
apple["200d"] = np.round(apple["Close"].rolling(window = 200, center = False).mean(), 2)

pandas_candlestick_ohlc(apple.loc['2016-01-04':'2016-08-07',:], otherseries = ["20d", "50d", "200d"])
```



# Regression

Linear Regression

Multilinear Regression

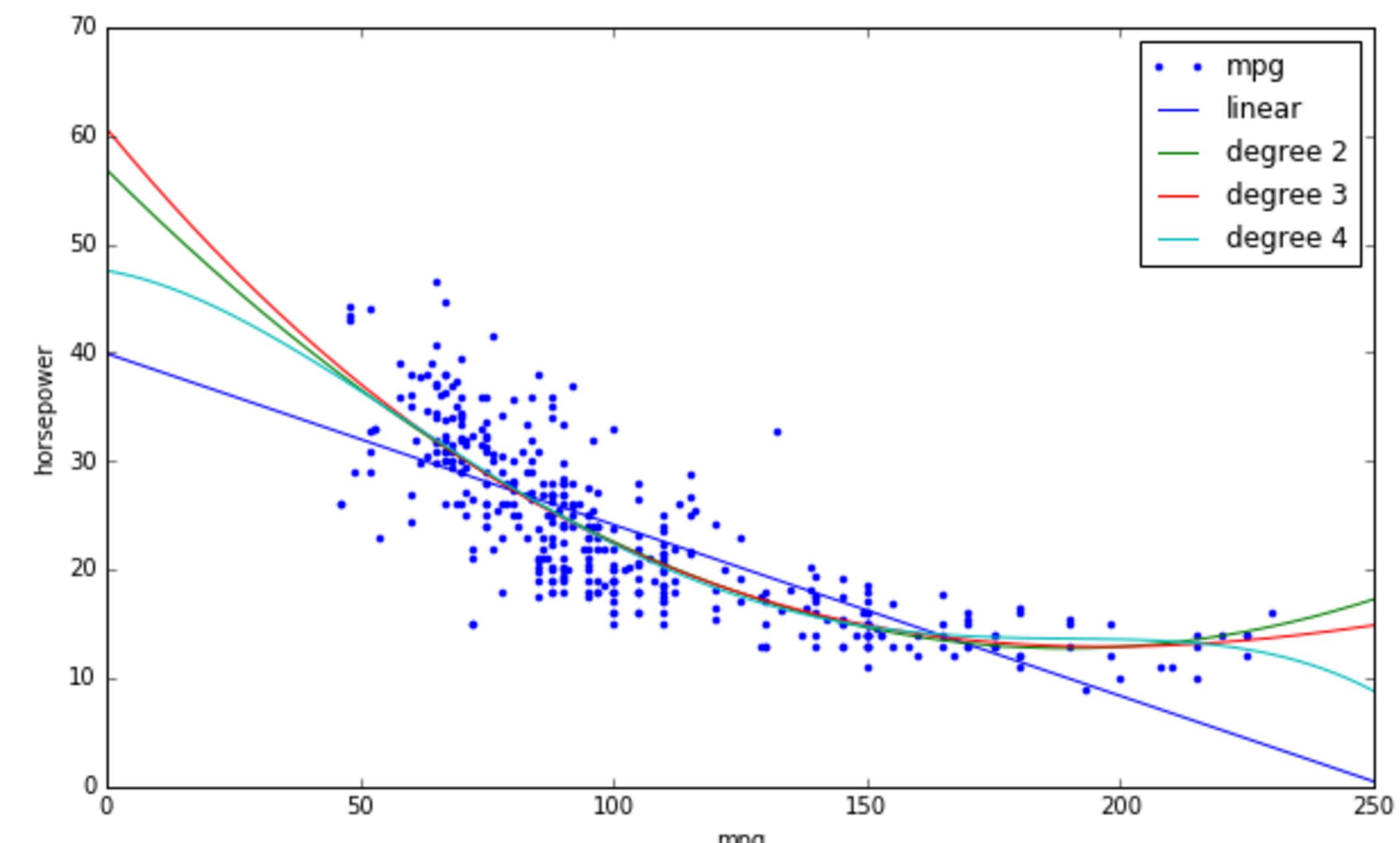
Nonlinear models

Confounders

Exploratory vs. inferential  
viewpoints

statsmodels, scikit-learn

OLS Regression Results						
Dep. Variable:	Rating	R-squared:	0.941			
Model:	OLS	Adj. R-squared:	0.940			
Method:	Least Squares	F-statistic:	3140.			
Date:	Fri, 23 Sep 2016	Prob (F-statistic):	4.70e-244			
Time:	15:28:22	Log-Likelihood:	-2019.2			
No. Observations:	400	AIC:	4044.			
Df Residuals:	397	BIC:	4056.			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[95.0% Conf. Int.]	
Intercept	145.3506	3.285	44.249	0.000	138.893	151.808
Income	0.0022	6.06e-05	36.057	0.000	0.002	0.002
Balance	0.2129	0.005	45.810	0.000	0.204	0.222
Omnibus:		63.810	Durbin-Watson:			1.879
Prob(Omnibus):		0.000	Jarque-Bera (JB):			92.125
Skew:		-1.168	Prob(JB):			9.89e-21
Kurtosis:		3.257	Cond. No.			9.95e+04



# Classification

Logistic Regression

k-Nearest Neighbors

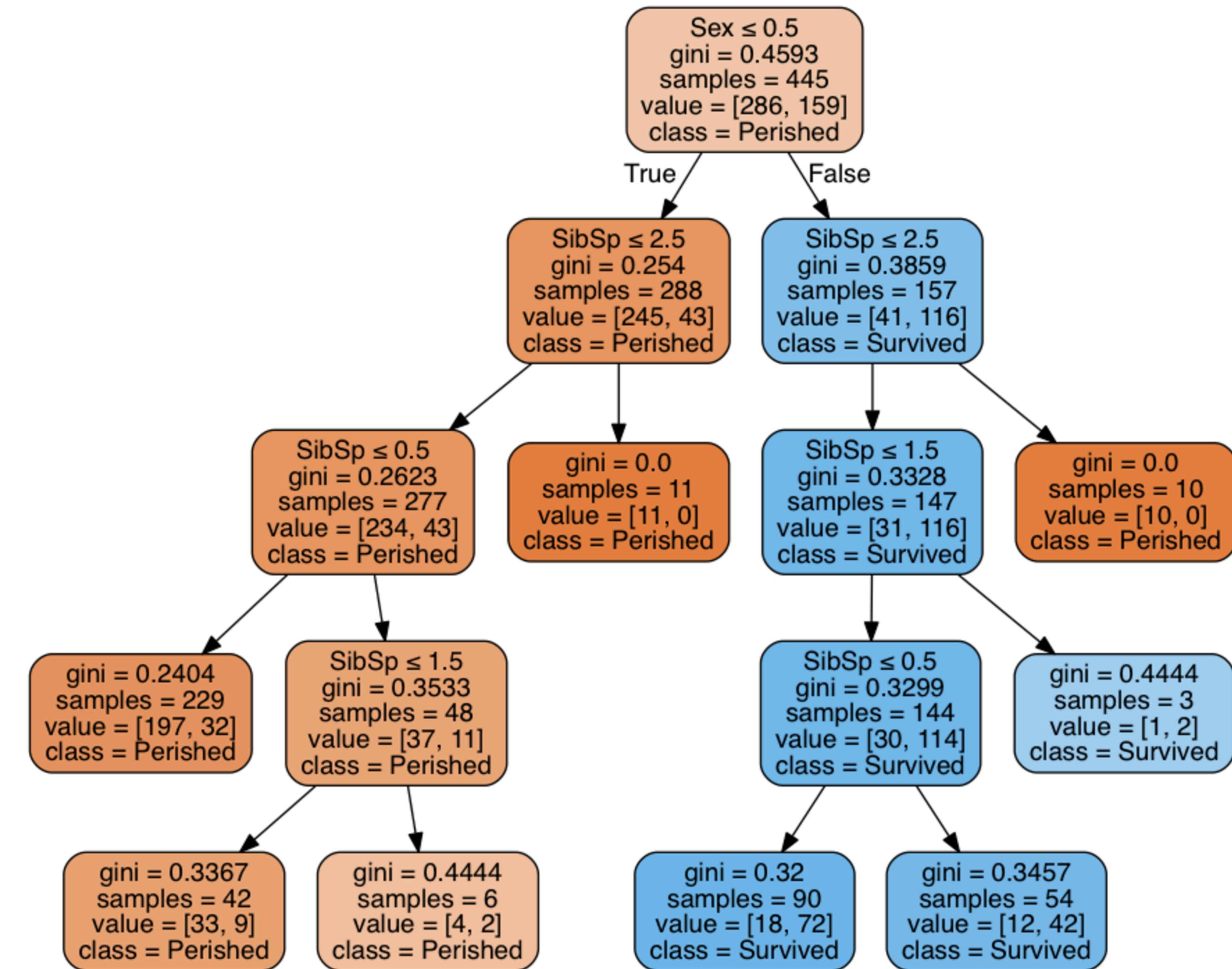
Decision Trees

Support Vector Machines

Neural Networks

generalizability and cross validation

scikit-learn, tensorflow



# Clustering

k-Means

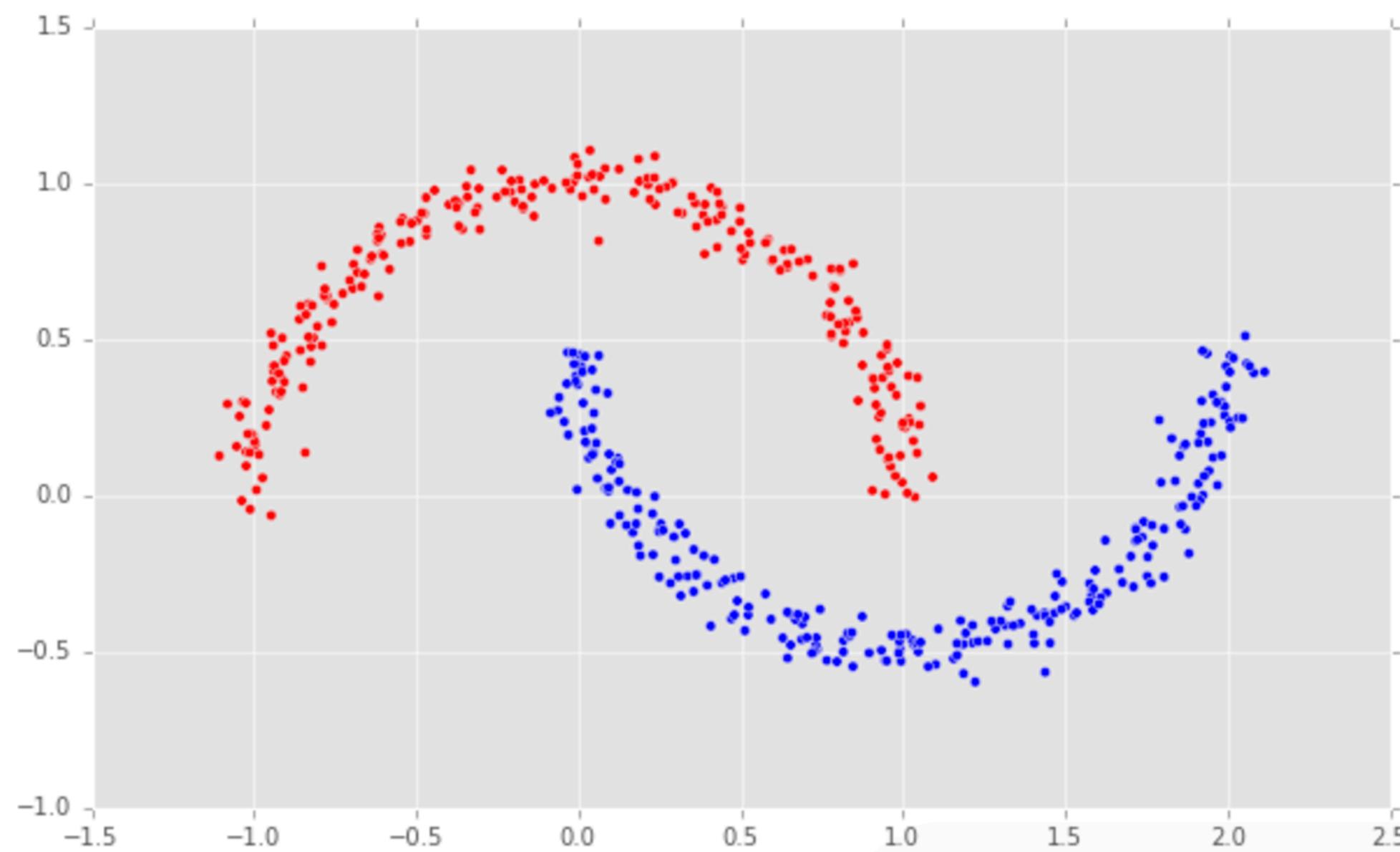
Hierarchical Clustering

DBScan

dendogram plots

scikit-learn

```
X, y = make_blobs(n_samples=n_samples, centers=5, random_state=1)
y_pred = KMeans(n_clusters=7).fit_predict(X)
plt.scatter(X[:, 0], X[:, 1], c=y_pred, marker="o", cmap=cmap)
plt.show()
```

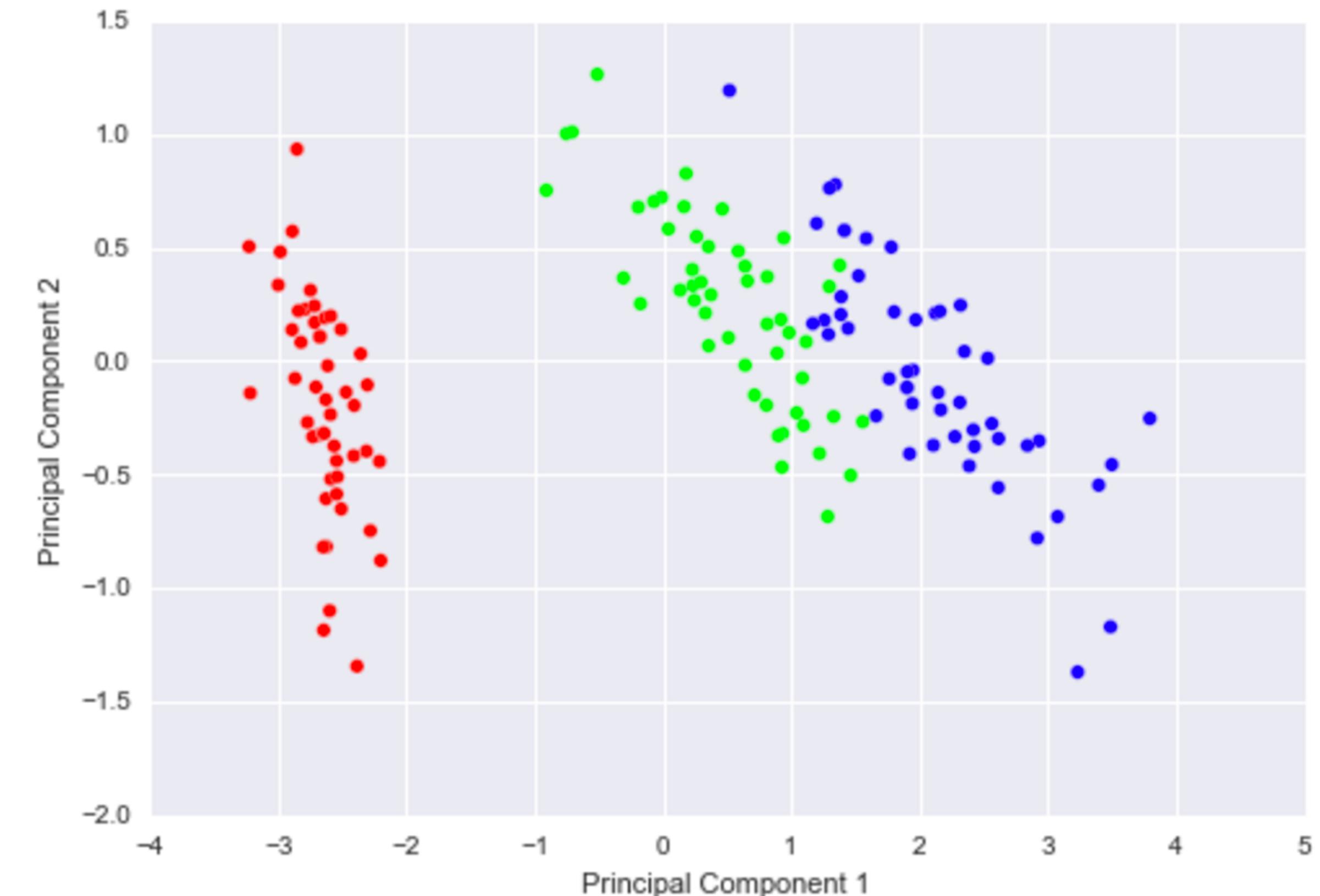


# Dimensionality Reduction

Principal Component Analysis  
(PCA)

method of dimensionality  
reduction for, e.g., visualization

scikit-learn



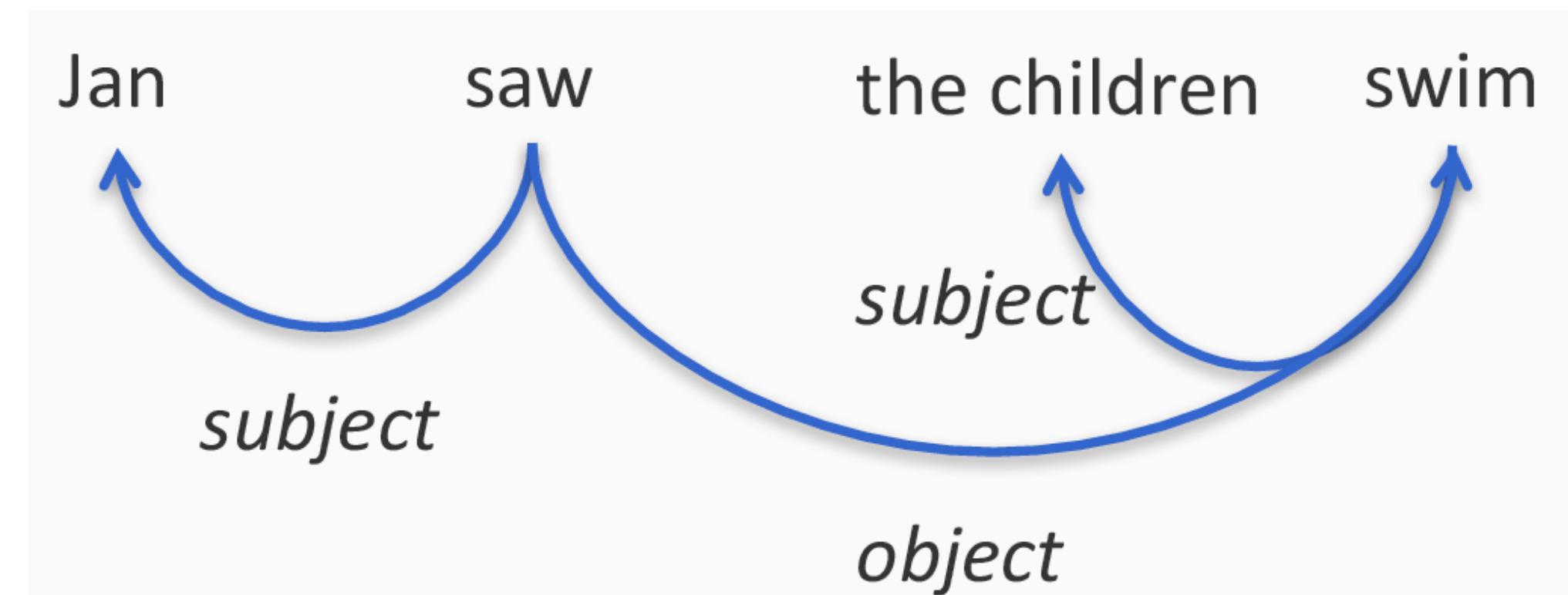
# Natural Language Processing

Guest lecture: Vivek Srikumar

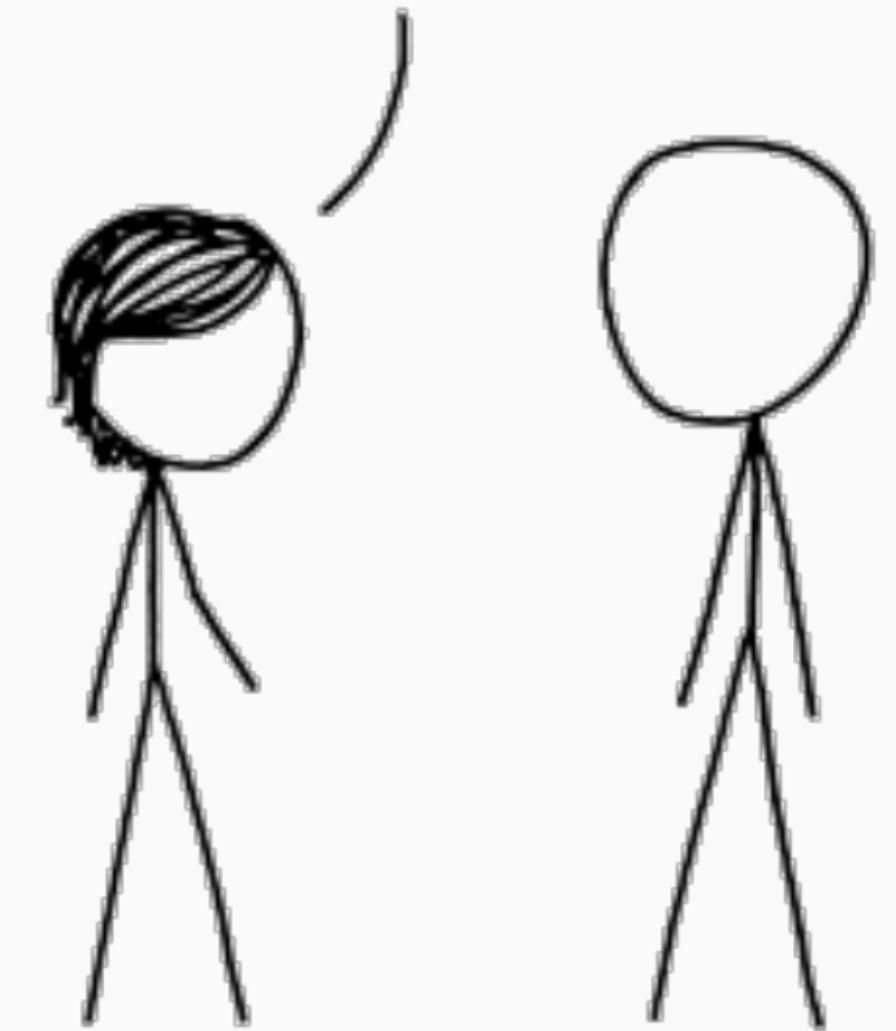
Semantic analysis

Regular expressions

Natural Language Toolkit (NLTK)



I DON'T MEAN TO GO ALL LANGUAGE NERD ON YOU, BUT I JUST LEGIT ADVERBED "LEGIT", VERBED "ADVERB", AND ADJECTIVED "LANGUAGE NERD."



<https://xkcd.com/1443/>

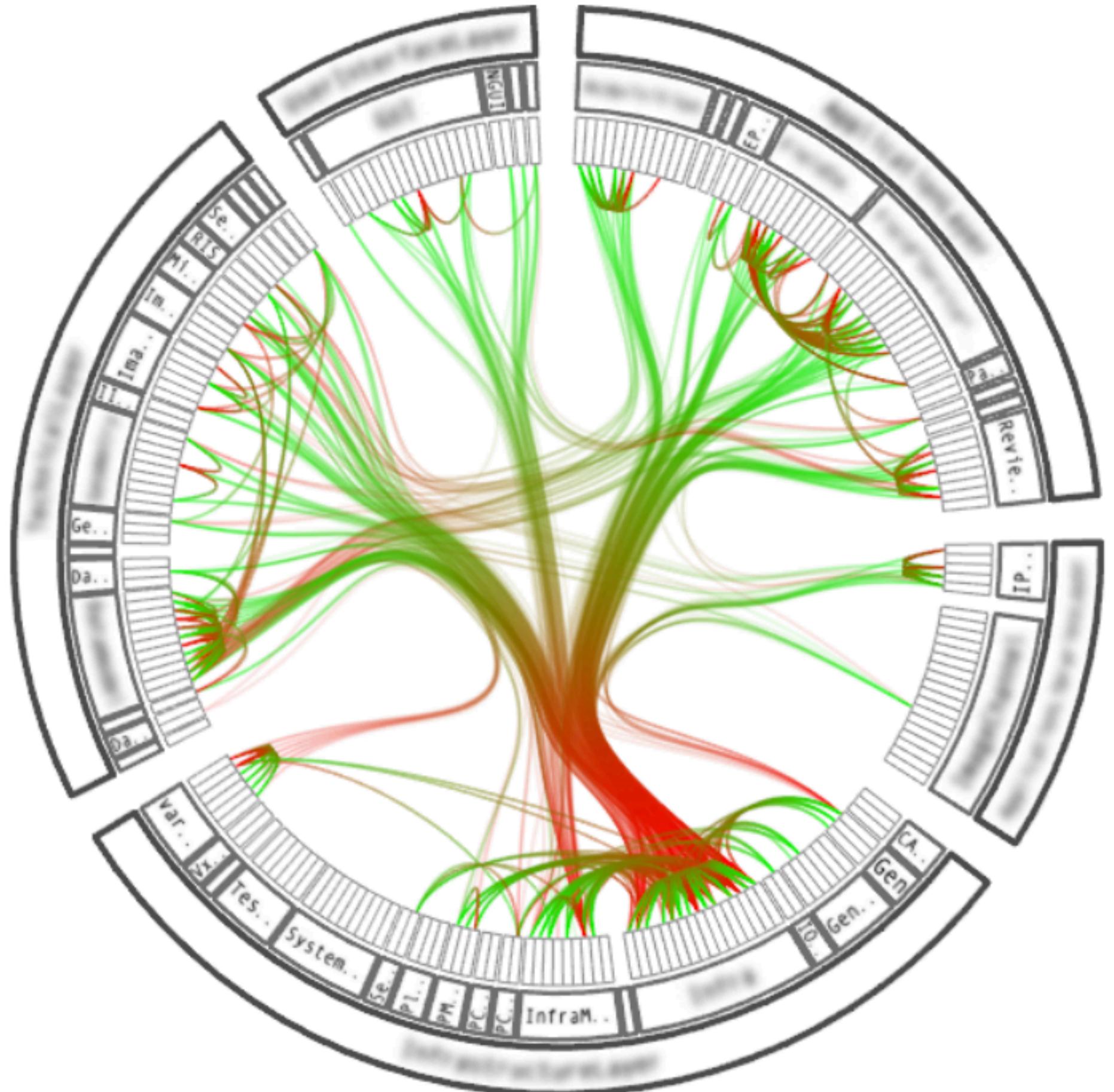
# Networks

# Centrality Measures

# PageRank

# Path Finding

# Graph Visualization



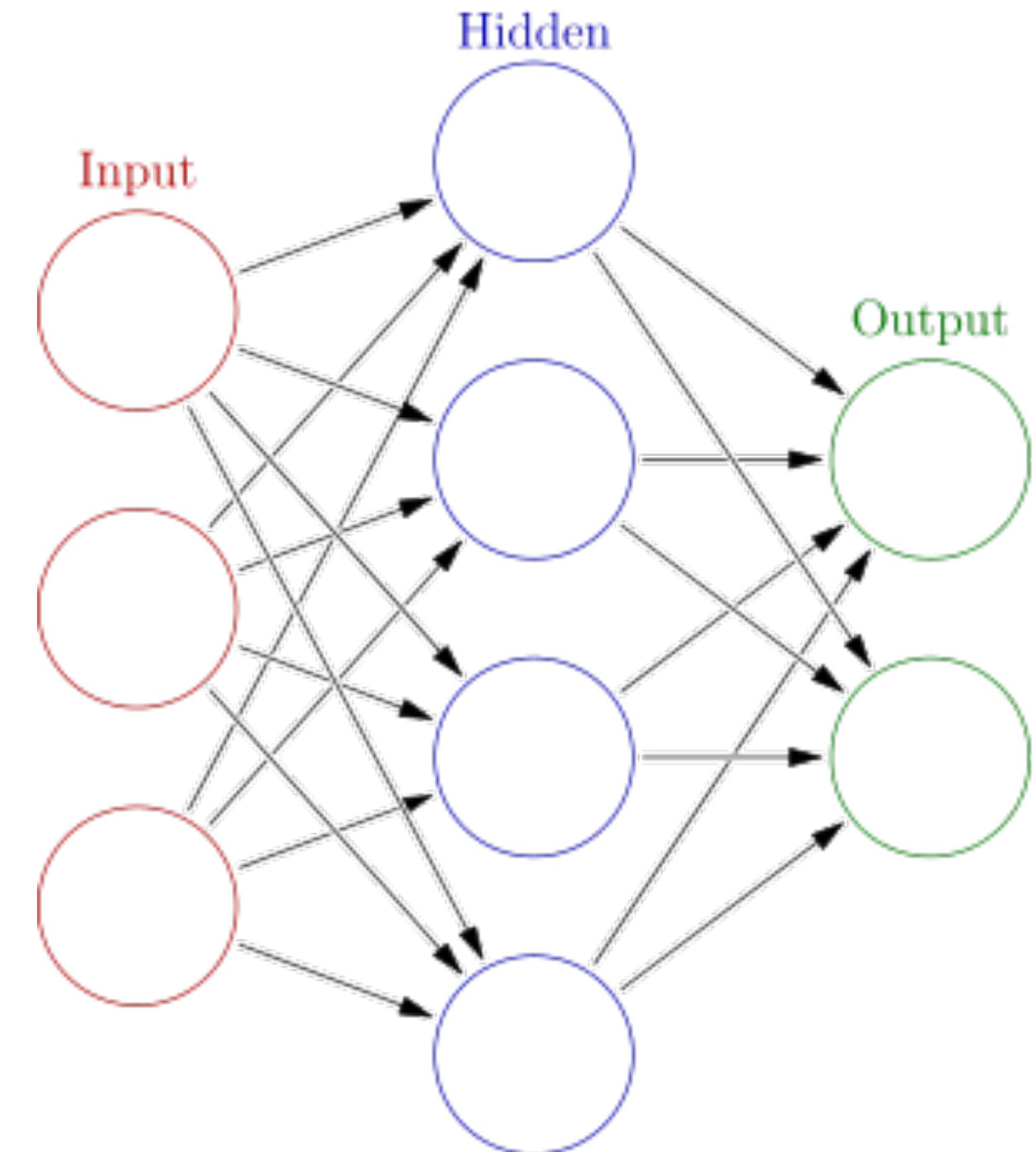
# Neural Networks / Deep Learning

Can be used for both classification and regression

Difficult to train

Notoriously difficult to interpret

scikit-learn, TensorFlow, keras



[Wikipedia](#)

# Ethics in Data Analysis

Guest lecture: dana boyd

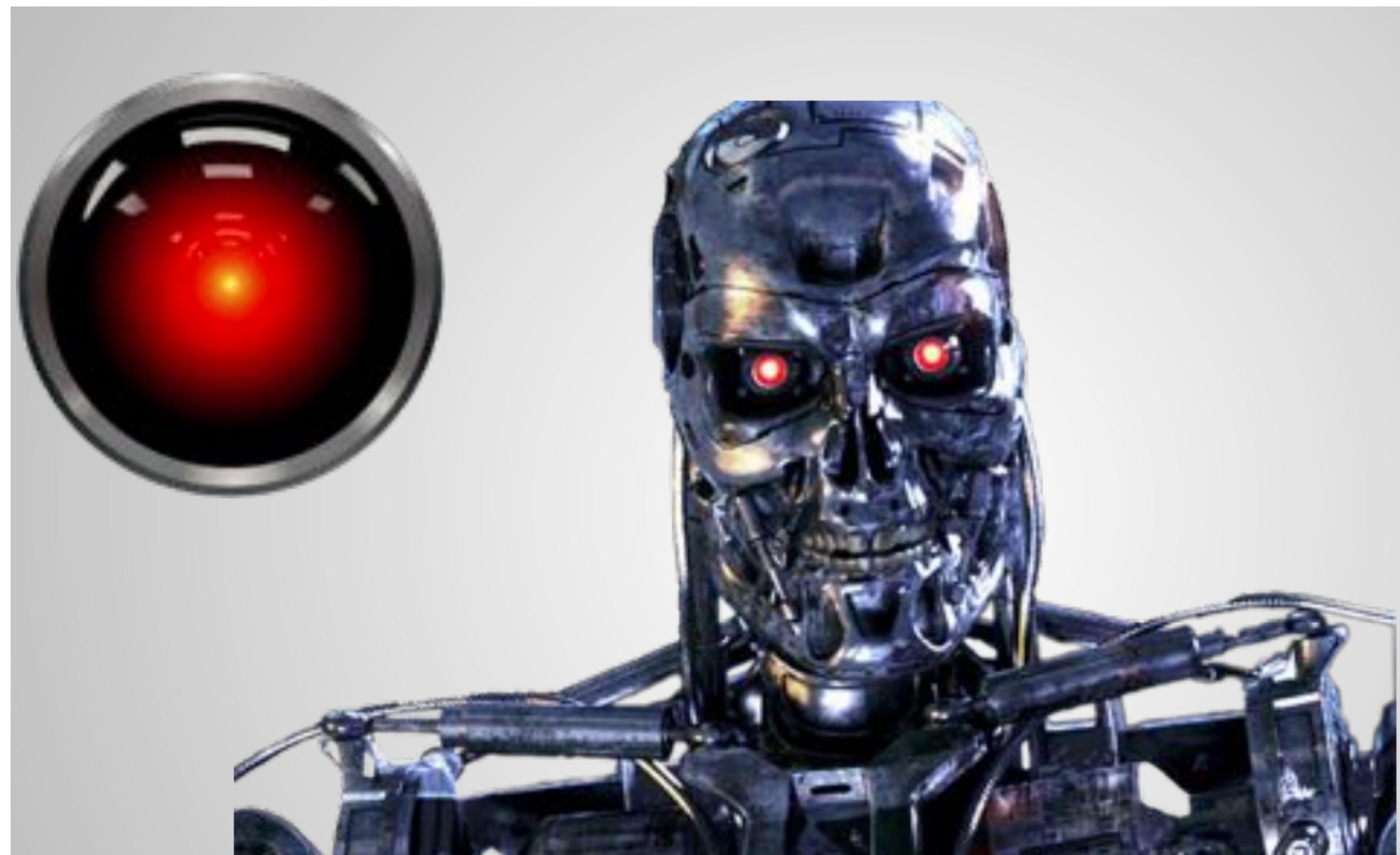
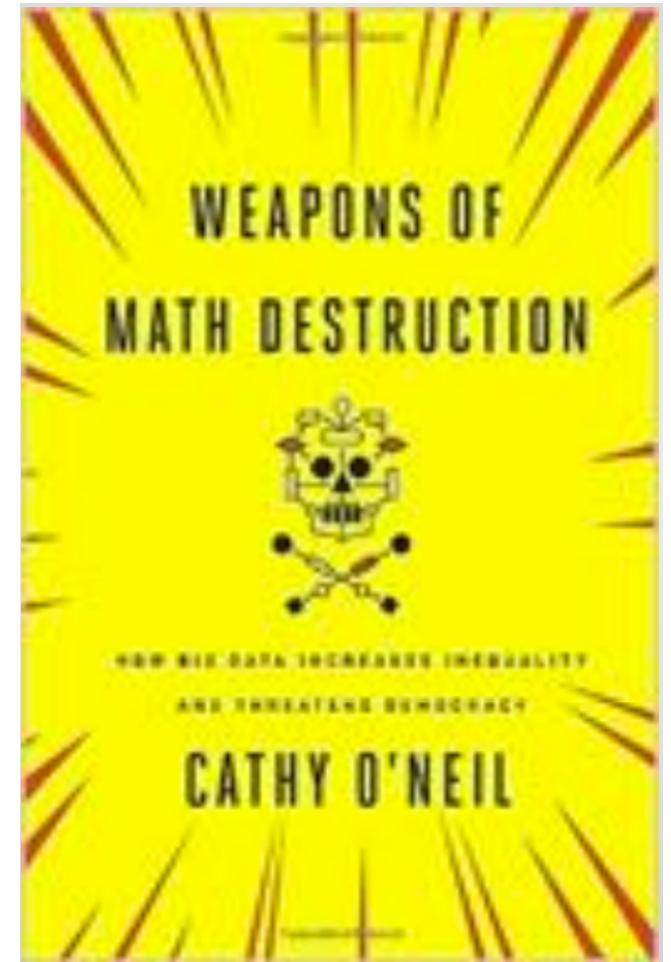
The age of automated decision making

Where does bias come from?

Data scientists are responsible for their algorithms

Algorithms should be interpretable and transparent

(Physical and digital) data is people



# Rating/Ranking/Elections

Least squares method for rating/ranking

Arrow's Impossibility Theorem

2016-17 SEASON

NORTH DIVISION	CONFERENCE	OVERALL
 4 Washington	8-1	11-1
 Washington State	7-2	8-4
 18 Stanford	6-3	9-3
 California	3-6	5-7
 Oregon State	3-6	4-8
 Oregon	2-7	4-8
SOUTH DIVISION	CONFERENCE	OVERALL
 8 Colorado	8-1	10-2
 11 USC	7-2	9-3
 20 Utah	5-4	8-4
 Arizona State	2-7	5-7
 UCLA	2-7	4-8
 Arizona	1-8	3-9

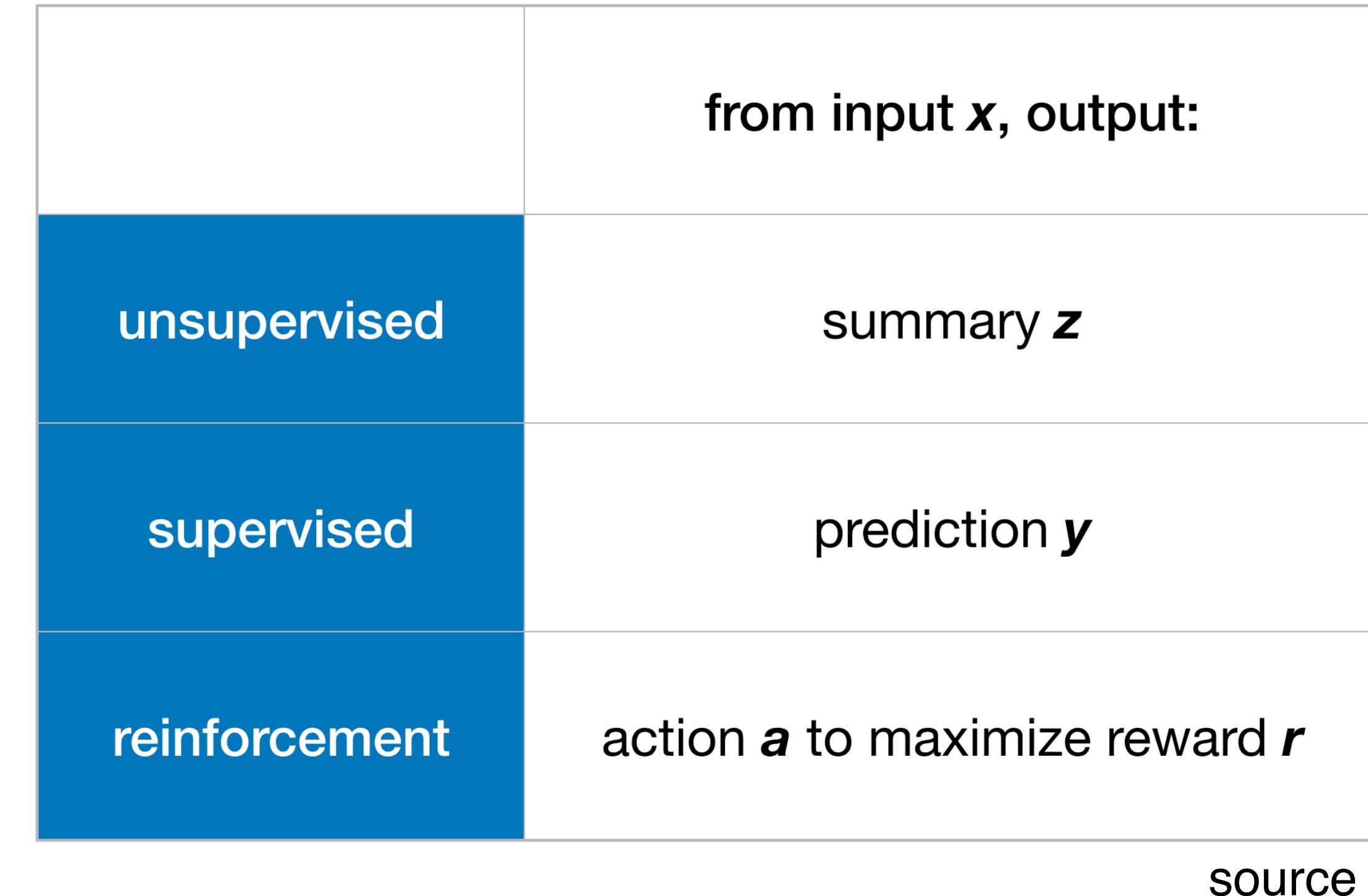
# What is Missing?

# Reinforcement learning

The third pillar of machine learning beyond unsupervised and supervised learning

Related to optimal control

This statistical model can evolve over time



# Recommender Systems

Example: In 2009, Netflix offered a \$1M prize to the winner of a competition to develop an algorithm that improves the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

Collaborative filtering: build a model from a user's past behavior (e.g. items previously purchased) and similar decisions made by other users.

One approach: matrix completion

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	0	3	0	3	0
User 2	4	0	0	2	0
User 3	0	0	3	0	0
User 4	3	0	4	0	3
User 5	4	3	0	4	0

A matrix of user/item ratings

# Probabilistic / Bayesian Methods

- data,  $\mathbf{X}$
- parameters in a probabilistic model,  $\theta$
- The *prior distribution* is the distribution of the parameters before any data is observed,  $p(\theta)$ .
- The *sampling distribution* or *likelihood function* is the distribution of the observed data conditional on its parameters,  $p(\mathbf{X} \mid \theta)$ .
- The *marginal likelihood* is the distribution of the observed data over the parameter(s),

$$p(\mathbf{X}) = \int_{\theta} p(\mathbf{X} \mid \theta)p(\theta)d\theta.$$

- The *posterior distribution* is the distribution of the parameter(s) after taking into account the observed data. This is computed using *Bayes' rule*,

$$p(\theta \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \theta)p(\theta)}{p(\mathbf{X})} \propto p(\mathbf{X} \mid \theta)p(\theta).$$

# Research Reproducibility

There is a “Replication Crisis”, particularly in psychology and medicine.

Reasons: Publication bias, cherrypicking, small N, irreproducible experiments, ....

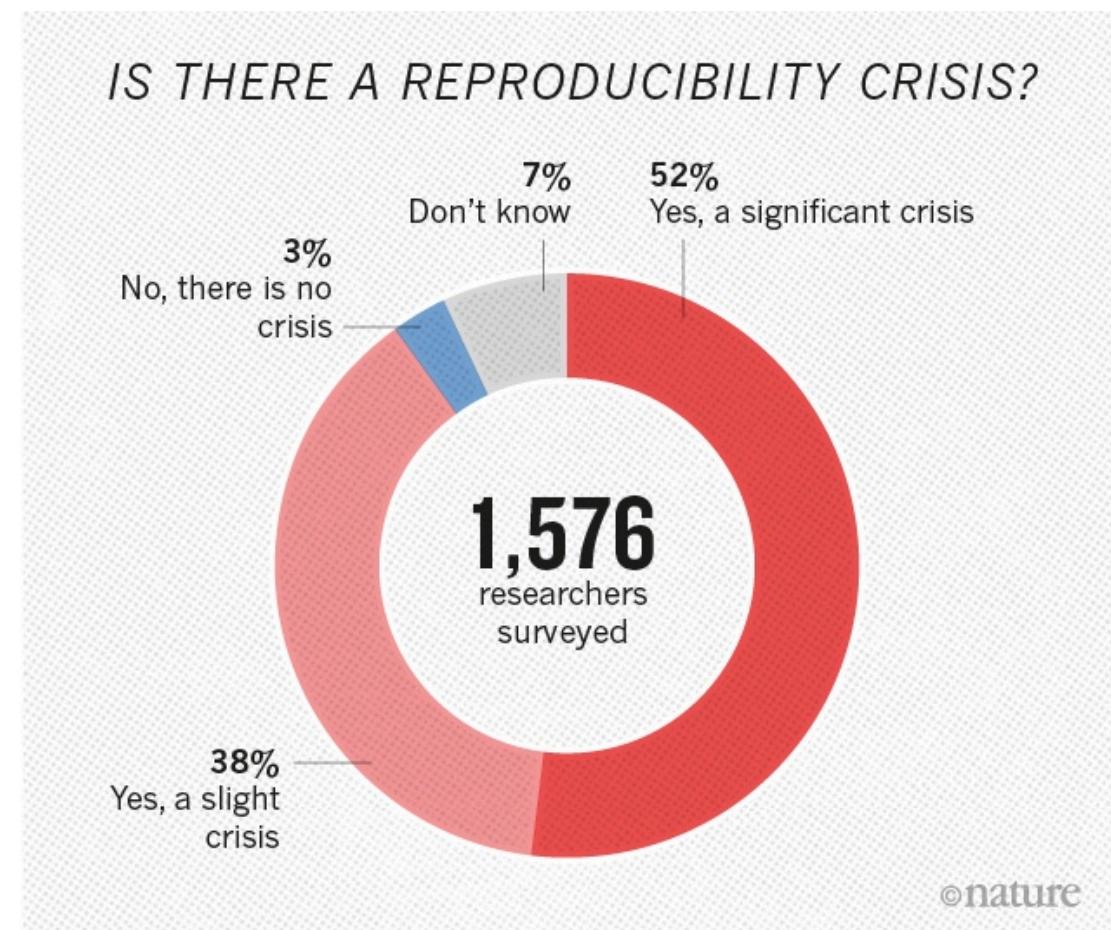
Reproducibility is critical, but not everything can be described in paper

One step: provide raw data and scripts (Jupyter Notebooks!) for analysis methods

Assumptions can be checked

Data can be analyzed with different assumptions/methods

Methods can be applied to different data



Where can you learn  
more?

# Further courses of interest

## Mathematics / Statistics

Math 1070: Introduction to Statistical Inference

Math 3070: Applied Statistics

Math 5075: Time Series Analysis

Math 5080/6824: Statistical Inference I

Math 5090/6828: Statistical Inference II

Math 5010/6805: Introduction to Probability

Math 6010: Linear Regression Analysis

Math 6020: Multilinear Models

Math 6070: Mathematical Statistics

Math 5770/6640: Introduction to Optimization

## Computer Science

CS 2420: Introduction to Algorithms & Data Structures

CS 4964: Math for Data (Foundations of Data Analysis)

CS 5130/6130: Computational Statistics

CS 5140/6140: Data Mining

CS 5340/6340: Natural Language Processing

CS 5350/6350: Machine Learning

CS 5630/6630: Visualization

# Masters in Statistics: MATHEMATICS

# MSTAT degree

## Master of Statistics Program: Data Science Option

<http://mstat.utah.edu/>

The Masters in Statistics (Mathematics) requires a thesis project. The degree consists of 36 credit hours of graduate level classes, 3 of which should be thesis hours.

### AFFILIATED FACULTY

Tom Alberts, Tom Fletcher (School of Computing), Lajos Horvath, Davar Khoshevisan, Braxton Osting, Jeff Phillips (School of Computing), Firas Rassoul-Agha, Suresh Venkatasubramanian (School of Computing).

### DATA SCIENCE OPTION

#### Prerequisites

**Math:** Calculus I-III (Math 1210, 1220, 2210), Linear Algebra (Math 2270), Probability (Math 5010), or equivalent coursework.

**CS:** Introduction to Algorithms and Data Structures (CS 2420), Algorithms (CS 4150), or equivalent coursework.

#### CORE CLASSES (to be completed in the first year of study)

Math 5080	Intro to Statistical Inference I
Math 5090	Intro to Statistical Inference II
CS 6140	Data Mining
CS 6350	Machine Learning

Electives: A total of 9 elective courses are required. Two must be taken from the Math elective list, and two must be taken from the CS elective list. The remaining electives may be taken from these lists, or from other departments on campus (subject to the approval of a student's advisor).

#### MATH ELECTIVES

Math 5030	Actuarial Mathematics
Math 5040-50	Stochastic Processes & Simulation I-II
Math 5600	Survey of Numerical Analysis
Math 5610-20	Introduction to Numerical Analysis I-II
Math 5650	Topics in Numerical Analysis
Math 5660	Parallel Numerical Methods
Math 5740	Mathematical Modeling
Math 5075	Time Series
Math 5770	Introduction to Optimization
Math 6010	Linear Models
Math 6030	Multivariate Models
Math 6040	Probability
Math 6070	Mathematical Statistics

#### CS ELECTIVES

CS 5530	Database System
CS 6150	Advanced Algorithms
CS 6190	Probabilistic Learning
CS 6300	Artificial Intelligence
CS 6340	Natural Language Processing
CS 6630	Visualization
CS 6961	Structured Prediction

# MS in Computing Track: Data Management and Analysis

Designed for full-time  
students

[http://www.cs.utah.edu/  
docs/Graduate/  
handbook16-17/  
datamang-2016-17.pdf](http://www.cs.utah.edu/docs/Graduate/handbook16-17/datamang-2016-17.pdf)

## MS IN COMPUTING: **DATA MANAGEMENT & ANALYSIS**

A student may pursue an MS with a (1) thesis option, or (2) a project option, or (3) a course-only option. The minimum number of credits for any of the three options is 30 from graduate level classes. A maximum of 6 project hours or 9 thesis hours is allowed to be included in the program of study for students in the project or the thesis option. A minimum of 6 hours of thesis research is required for the thesis option.

### TRACK FACULTY

Tom Fletcher, Lajos Horvath (Math), Chris Johnson, Sneha Kumar Kasera, Mike Kirby, Alexander Lex, Feifei Li, Miriah Meyer, Baxton Osting (Math), Valerio Pascucci, Bei Wang Phillips, Jeff Phillips (Track Director), Vivek Srikumar, Hari Sundar, Suresh Venkatasubramanian

### CORE CLASSES

Must take 4 core classes, at least one from each line.

CS 6140	Data Mining /or/ CS 6350 Machine Learning
CS 6150	Advanced Algorithms
CS 6530	Database Systems
CS 6630	Visualization

A average grade of B or greater is required for core classes.

**ELECTIVES:** Three courses from the following list are required: (or CS 6140/CS 6350 if not counted above, or appropriate classes by track faculty)

### ALGORITHMIC

CS 6160	Computational Geometry
CS 6170	Computational Topology
CS 7960	Models of Computation for Massive Data

# CS Big Data Certificate

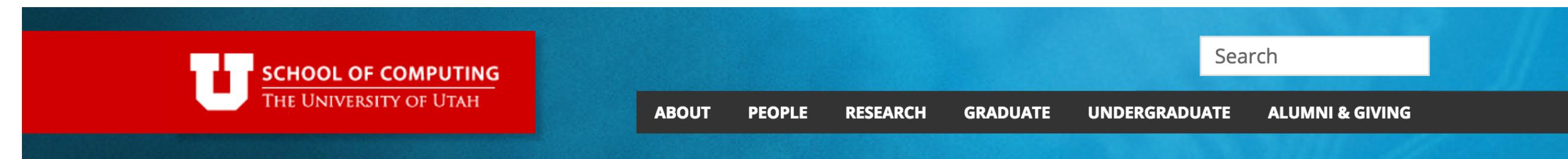
[www.cs.utah.edu/bigdata/](http://www.cs.utah.edu/bigdata/)

Designed for professionals  
working full-time

Goes through many higher  
level CS data courses

Undergrad required

Not a MS degree!



Home » Big Data Certificate

## Big Data Certificate

Big Data is impacting many areas of science, engineering, and industry; from analyzing troves of weather data to modeling traffic patterns to processing millions of online customers, it is the enormous data which is creating new opportunities and challenges. To tackle these challenges, one must have the training to store, manage, process and analyze data at these scales. But the challenges are beyond scale alone, the complexity of the data requires new powerful analytical techniques. Finally, it is crucial to have skills in communicating and interpreting the results of this analysis. A person trained in all of these skills is a big data scientist.

### Computing Degree

[Data Management and Analysis \(MS and PhD\)](#)

### Big Data Certificate Program

[Requirements \(PDF\)](#)

[Program Information \(PDF\)](#)

### Applying to the Big Data Certificate

Admissions to the Big Data Program – [here](#)

[Frequently Asked Questions](#)

### Videos

[Data Mining with Jeff Phillips](#)

[Visualization with Miriah Meyer](#)

[Getting to Know U](#)

### Research Centers & Groups

[Center for Extreme Data Management Analysis and Visualization \(CEDMAV\)](#)

[Data Group](#)

### Graduate

[Admissions](#)

[Graduate Program](#)

[Graduate Student Resources](#)

[Graduate Handbook](#)

[CES Program](#)

[Why Utah?](#)

[Course Descriptions](#)

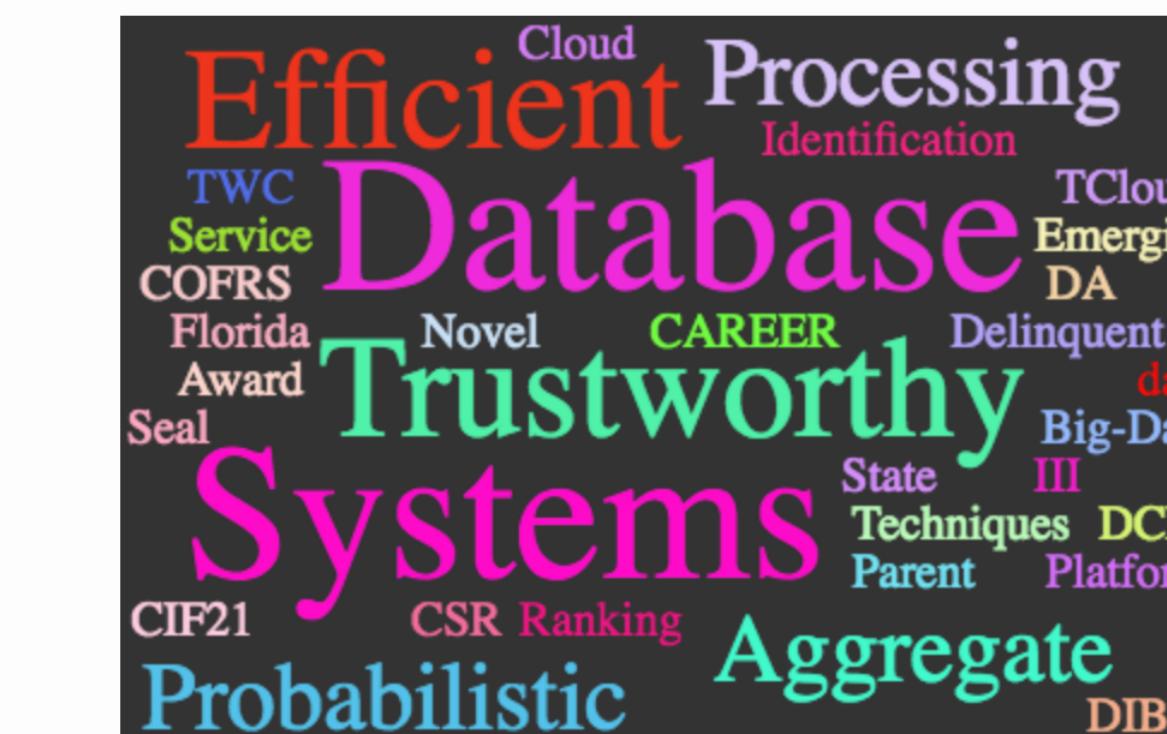
[TA Application](#)

[New Student Information](#)

# Data Group

Meets weekly to discuss methods and applications in data analysis

# Data Group Website



## Recent and Upcoming Seminars

February 8, 201

More Seminar

# <http://datascience.utah.edu/>

The screenshot shows the homepage of the Utah Center For Data Science. The header features a dark background with the center's logo on the left and navigation links for Members, Events, Programs, Seminar, Clubs, and Get Involved. Below the header is a large, semi-transparent white box containing the title "Utah Center For Data Science" and a descriptive paragraph about the center's mission. The main content area has a background image of snow-capped mountains. At the bottom, there are three white callout boxes with icons and text: "ACADEMIC PROGRAMS" (graduation cap icon), "DATA DAY" (calendar icon), and "DATA SCIENCE CLUB" (people icon).

**Utah Center For Data Science**

This center leads, organizes, and manages data science resources and research efforts at the University of Utah. Its members advance the fundamental principles and practice of data science through research, applications, and community engagement.

**ACADEMIC PROGRAMS**

The university has various different academic programs connected to data science.

**DATA DAY**

The 2019 Data Science Day will be held on October 18, 2019. More details to follow.

**DATA SCIENCE CLUB**

The data science club helps students learn about data science and machine learning through tutorial, presentations, and applications.

# Data Science Day @ Utah

Fall 2020

Data Science Day welcomes all students, staff, and faculty at the University of Utah to present a poster or demo.

Consider presenting your class projects!

---

[PIPELINE Career Expo](#)

---

[Posters and Demos](#)

---

Welcome address: Dan Reed, Senior Vice Preside

---

[Panel: Data Science in Industry](#)

---

[Data Science + X Talks](#)

---

[Keynote Talk](#)

---

[Poster Awards !!](#)

---

# How did it Go?

# Feedback Please!

Were your expectations met?

What else would you have liked to learn about?

Was the order good?

Did you feel prepared? Are the prerequisites appropriate?

Was it too much work? Was it too easy?

Too little math? Too much math?

Too little programming? Too much programming?

Did you like Python / Jupyter / GitHub?

Did you enjoy the project?

Should we have had exams?

# Course Evaluation

<https://goo.gl/lbhkEr>

Please Take 5 Min now to evaluate  
this course!

Evaluations are important for us to  
improve the course and our teaching!

# Thanks!

To you for participating and coming to lectures!

To our guest lecturers: Vivek Srikumar, Dana Boyd, and Jim Steenburgh!

To our TAs Haihan Lin, Youjia Zhou, and Jacob Jones!