Machine Learning: Data Foundation + Algorithms & Applications

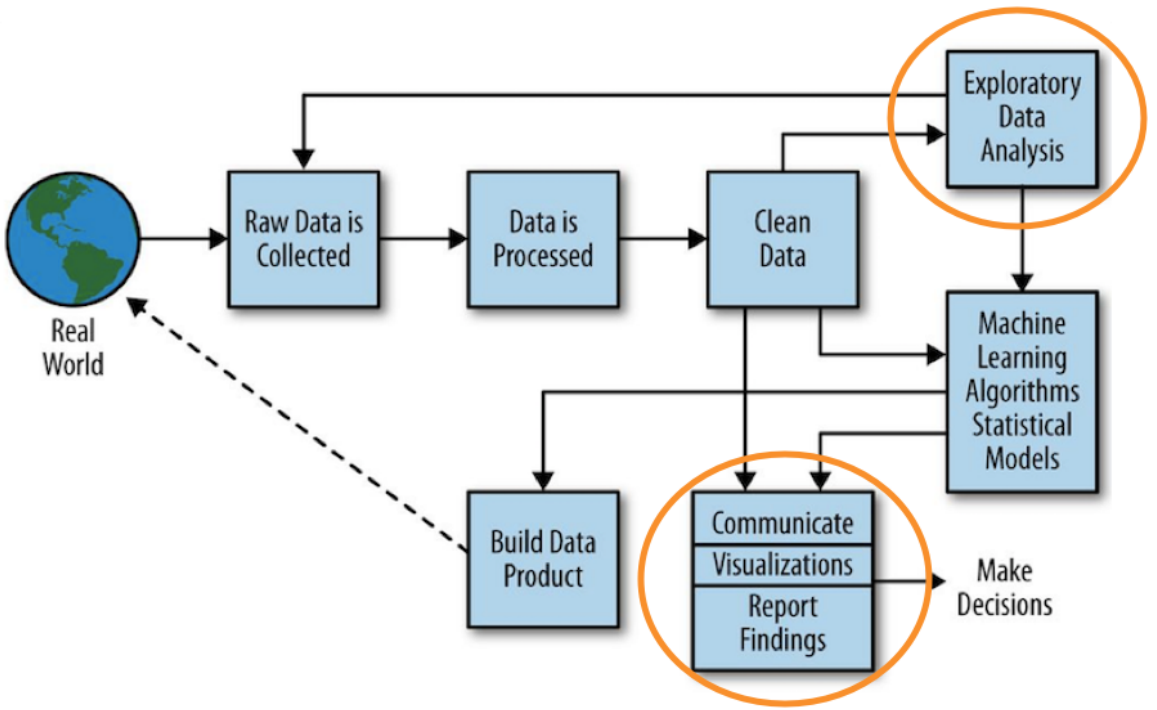
Day 2

Visualization

Objective

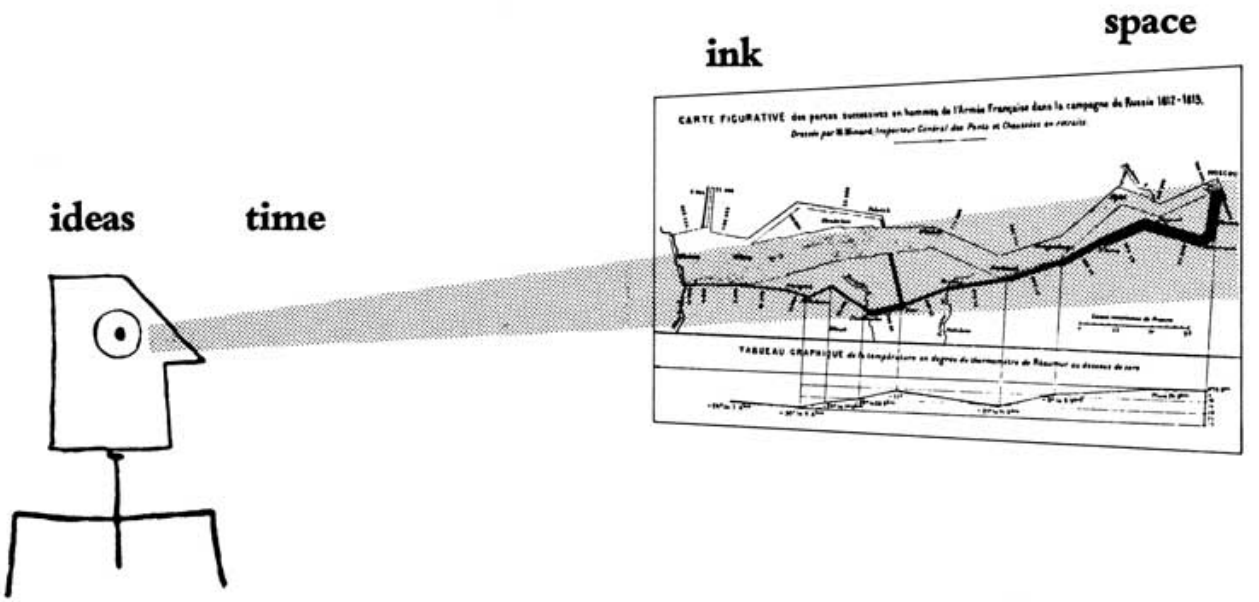
- Lay down some basic visualization theory
- Learn about the various types of data visualization tools
- Learn some basic Python techniques

Data Science Pipeline



"Doing Data Science"

GOAL: communicate the MOST ideas, in the LEAST amount of time, with the LEAST amount of ink, in the LEAST amount of space



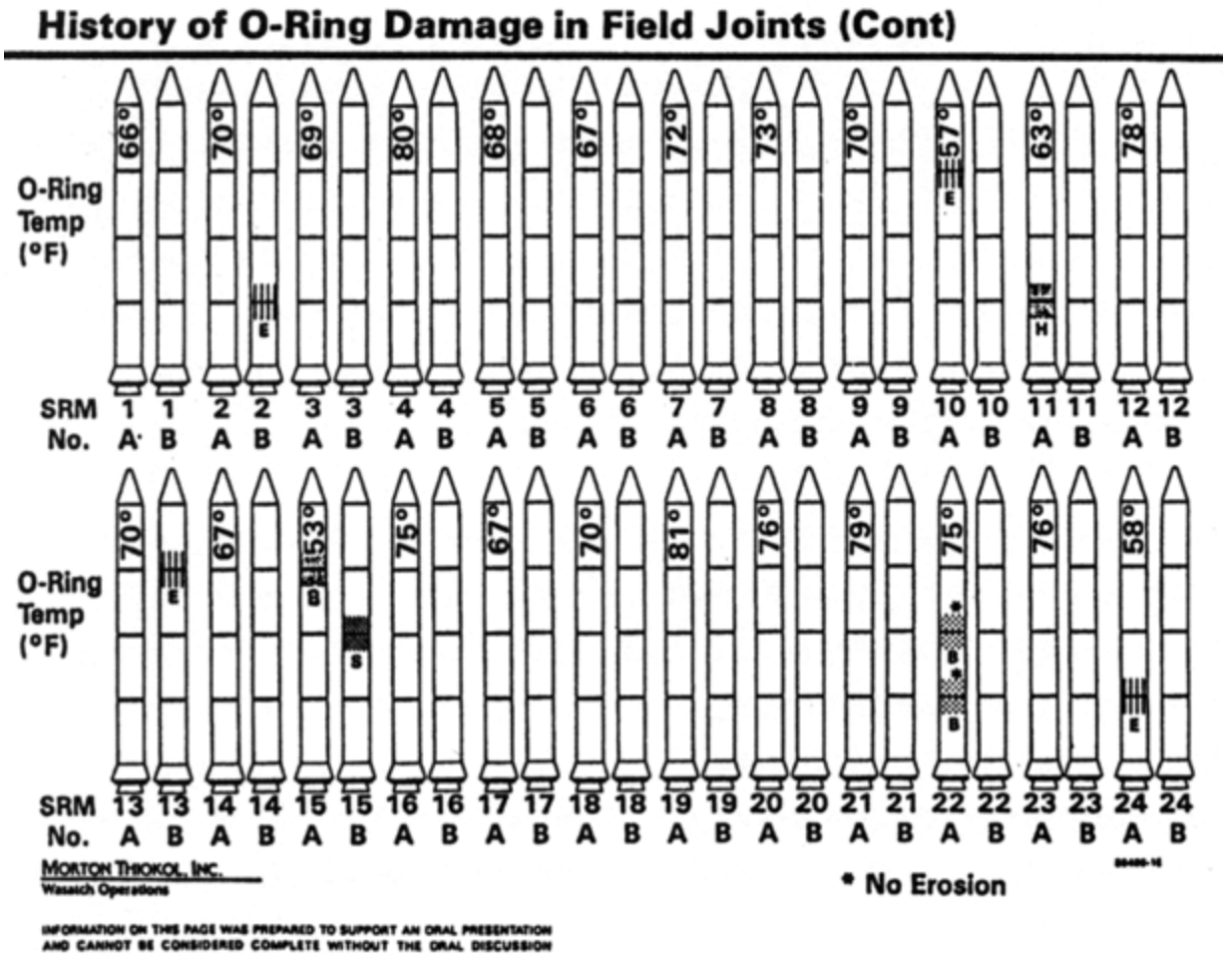
Edward Tufte

Charles Minard's graphic of Napoleon's Russian campaign of 1812

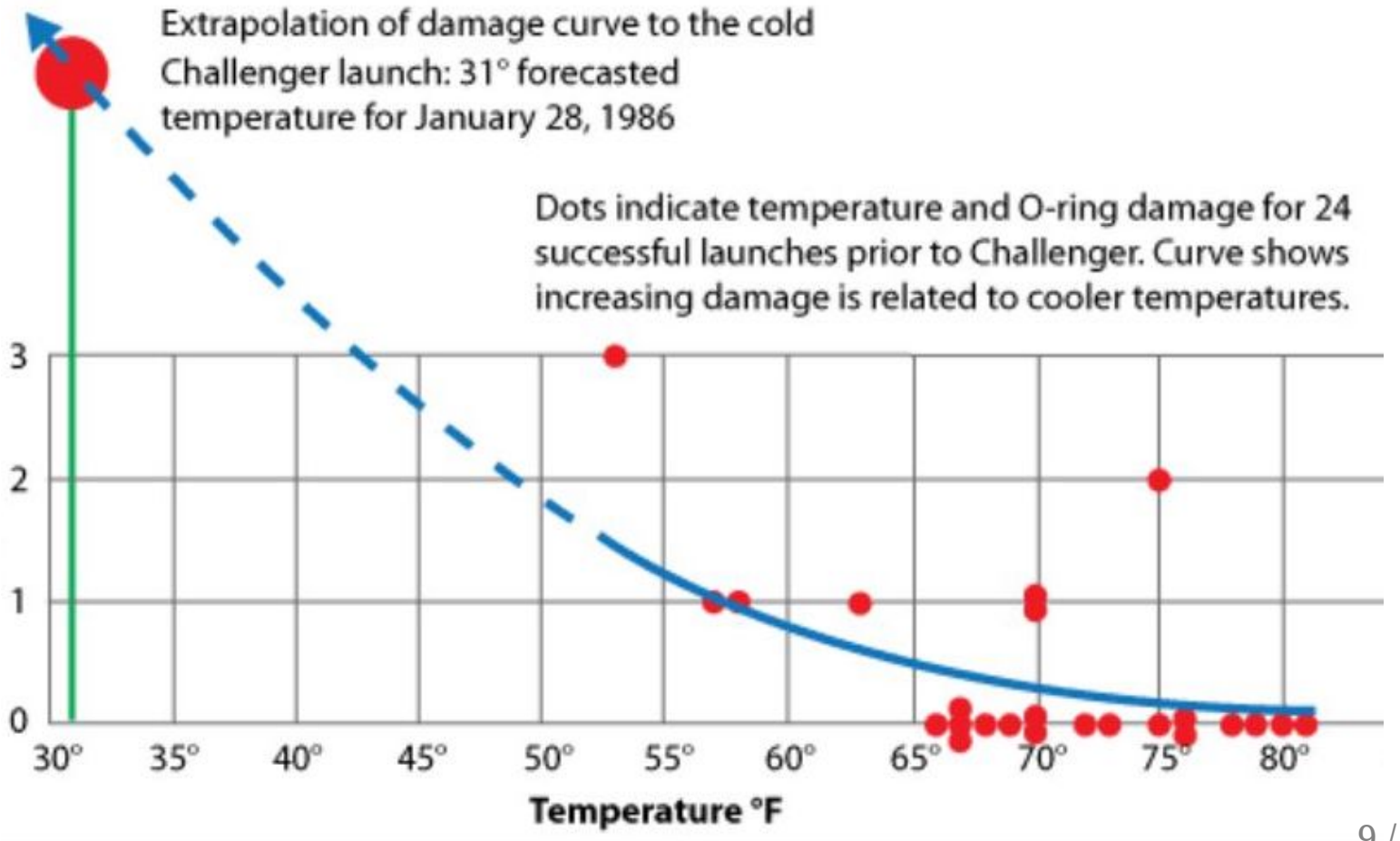
- Number of men present represented by widths of the colored zones
  - brown designates men who enter Russia, the black those who leave it
  - 1 mm = ten thousand men

Space Shuttle Challenger Explosion: January 28, 1986

Hiding the Story with "Chart Junk"?



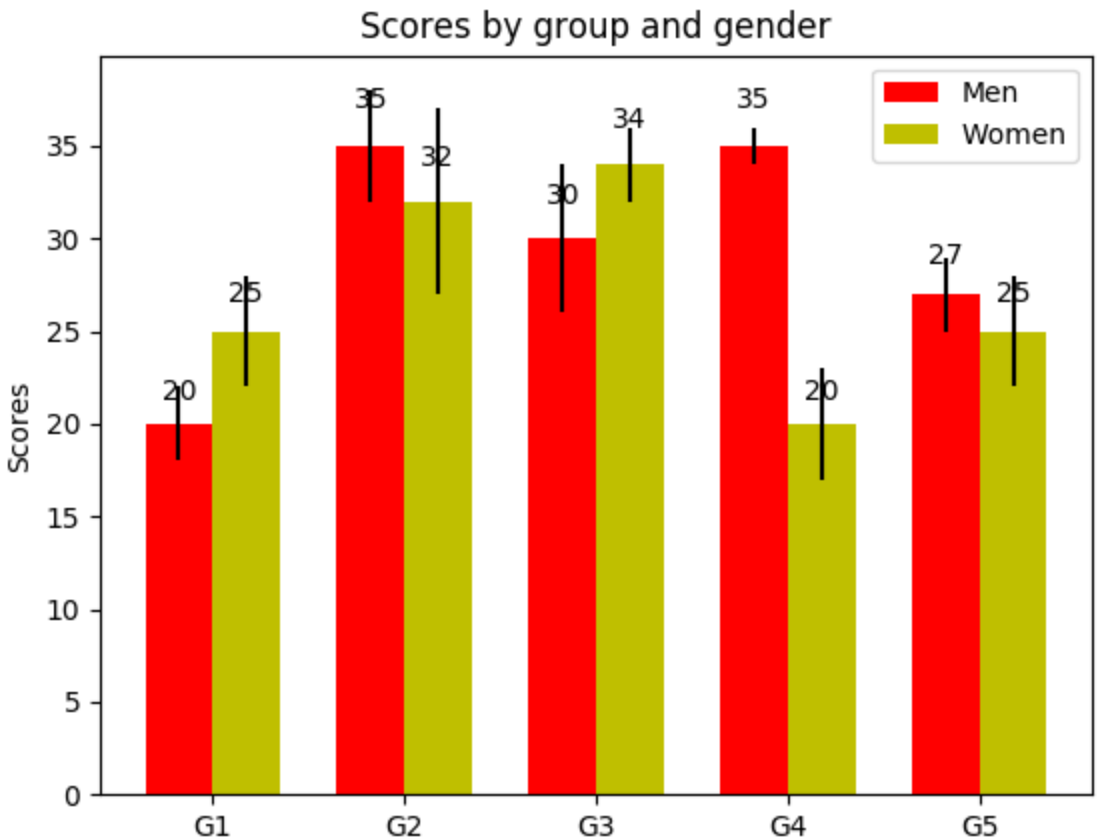
Challenger Data Redrawn by Tufte



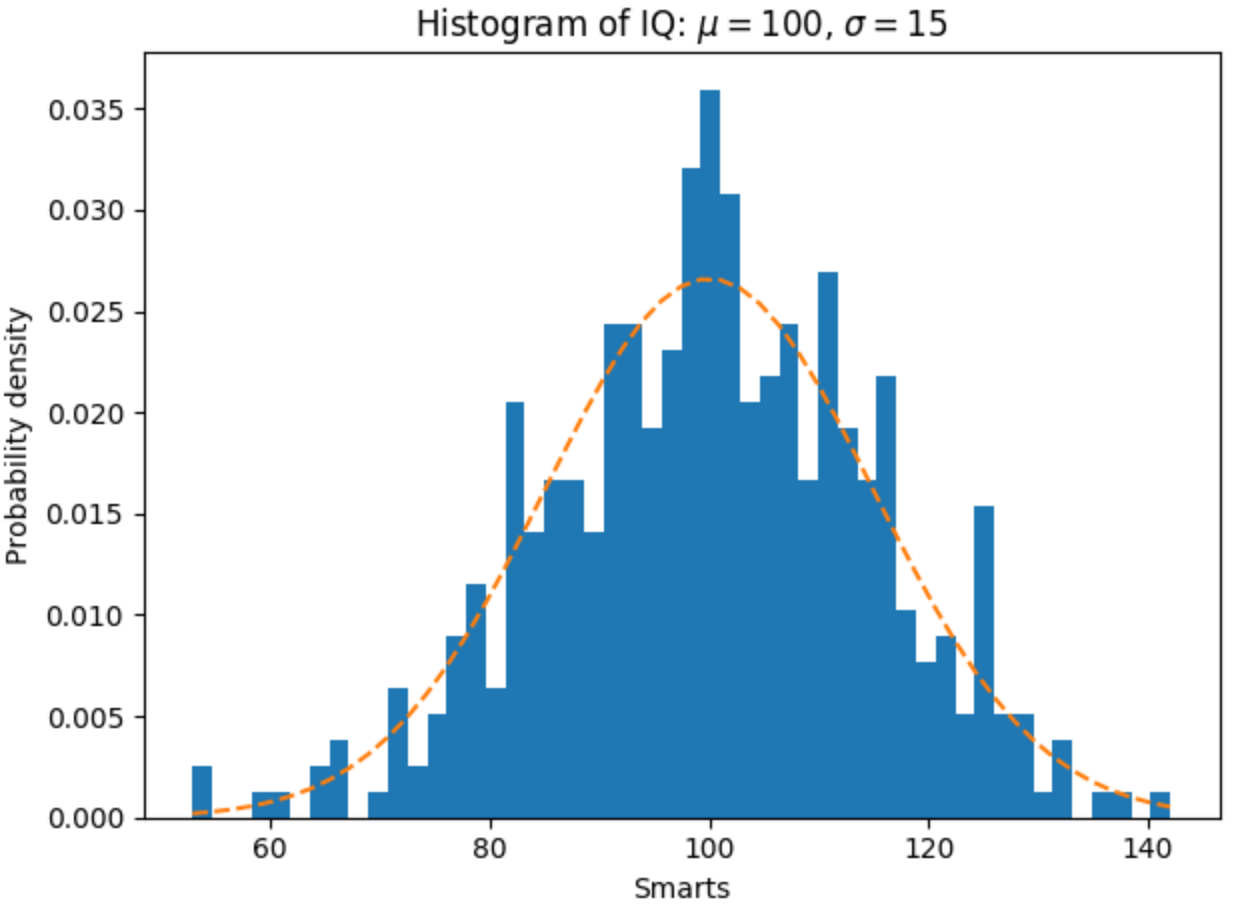

BY THE YEAR
2018
2019
2020
ALL
INFOR
MATION
ALL
NONE
2019
WILL
BE
6
YEARS
FROM
NOW
(72 MONTHS)
COMMUNICATED
YES!
IN
THIS
CLEAR
AND
CONCISE
26
0
T
FORMAT

Timeseries Data

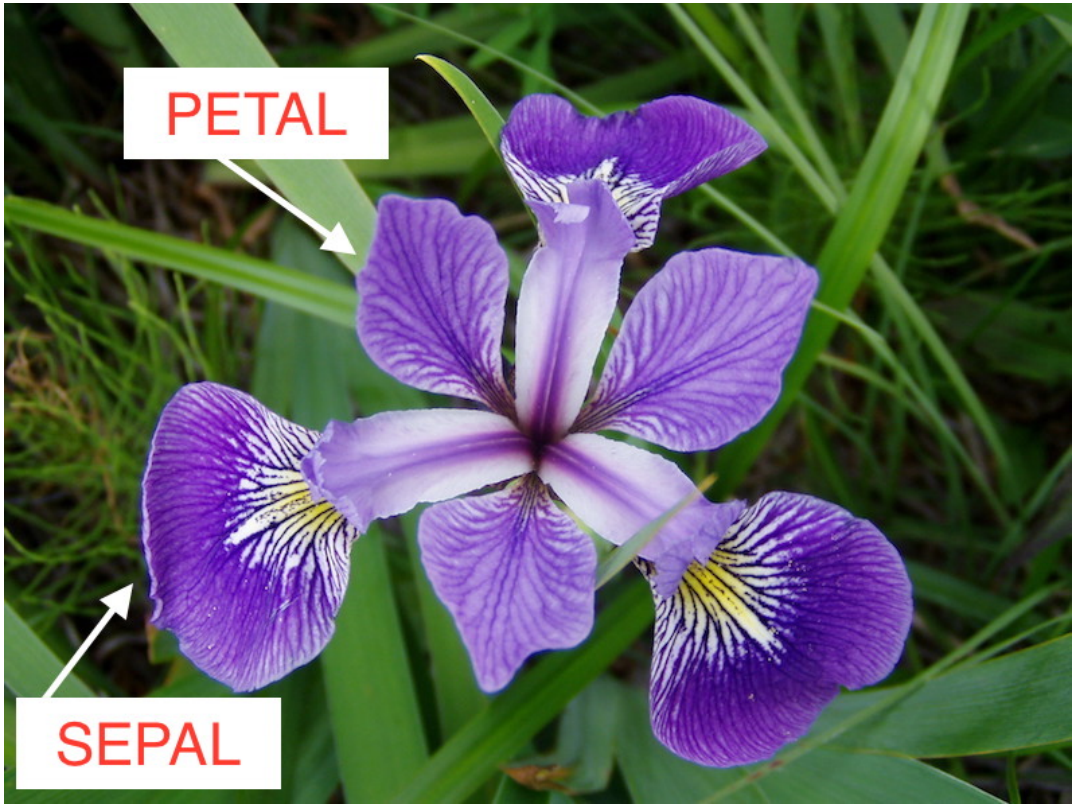
Bar Charts


Scores by group and gender
Men
Women
Scores
0
5
10
15
20
25
30
35
G1
G2
G3
G4
G5
20
25
35
32
30
34
35
20
27
25

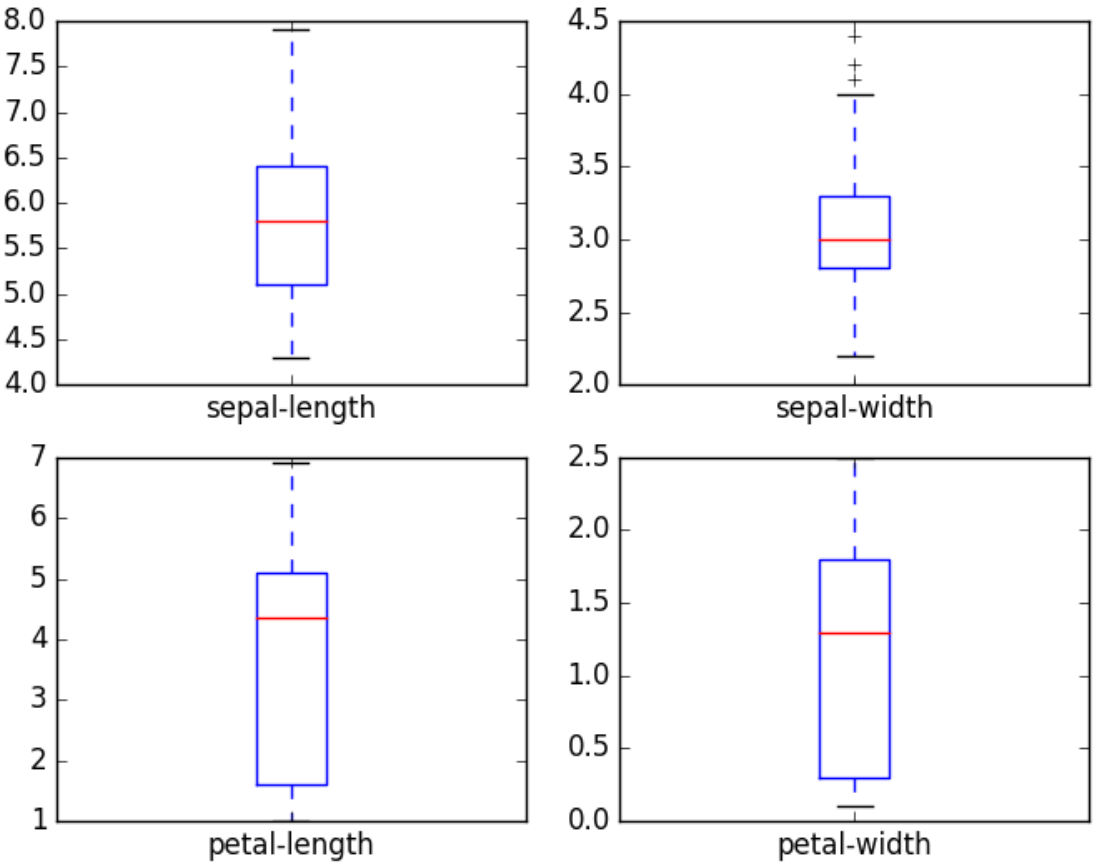
Histograms


Histogram of IQ: $\mu = 100$, $\sigma = 15$
Probability density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
60
80
100
120
140
Smarts

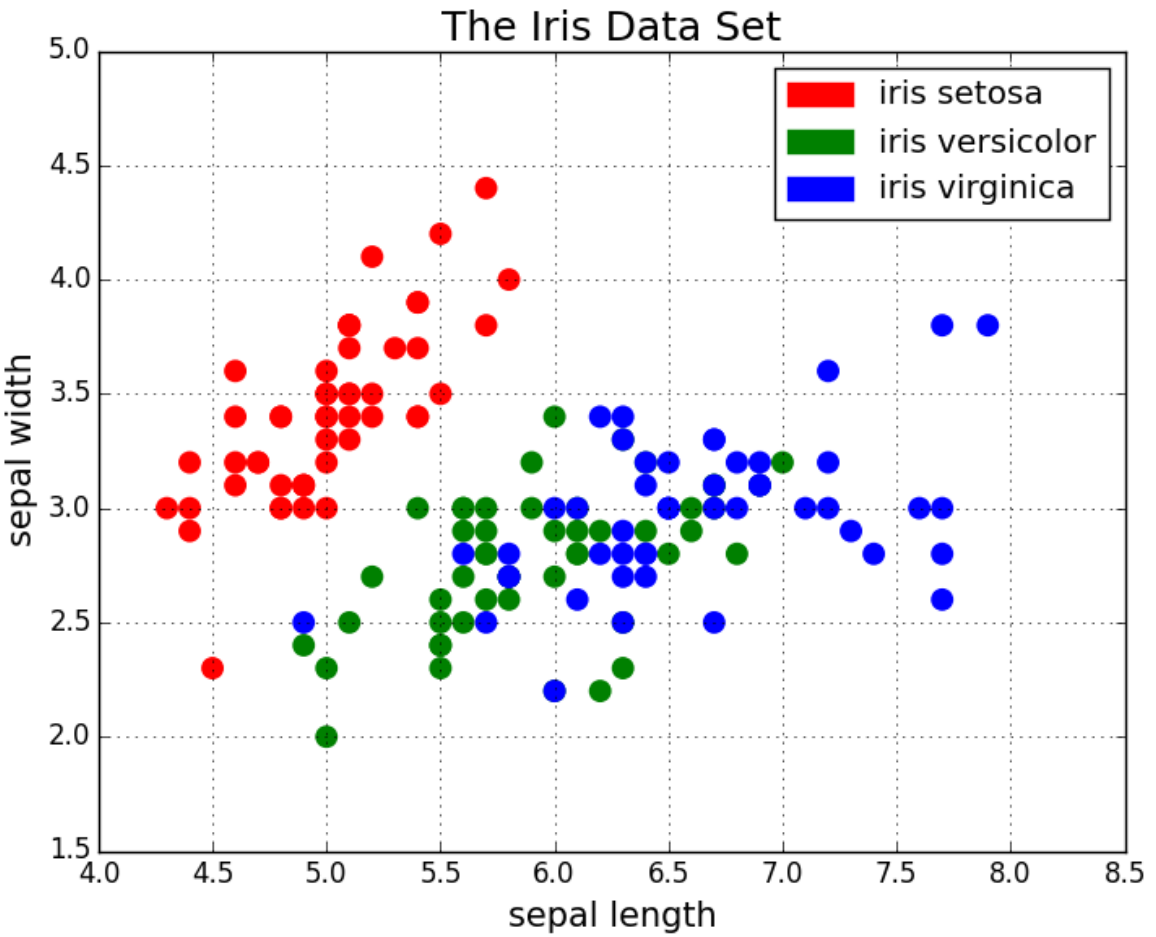
The Iris Data Set



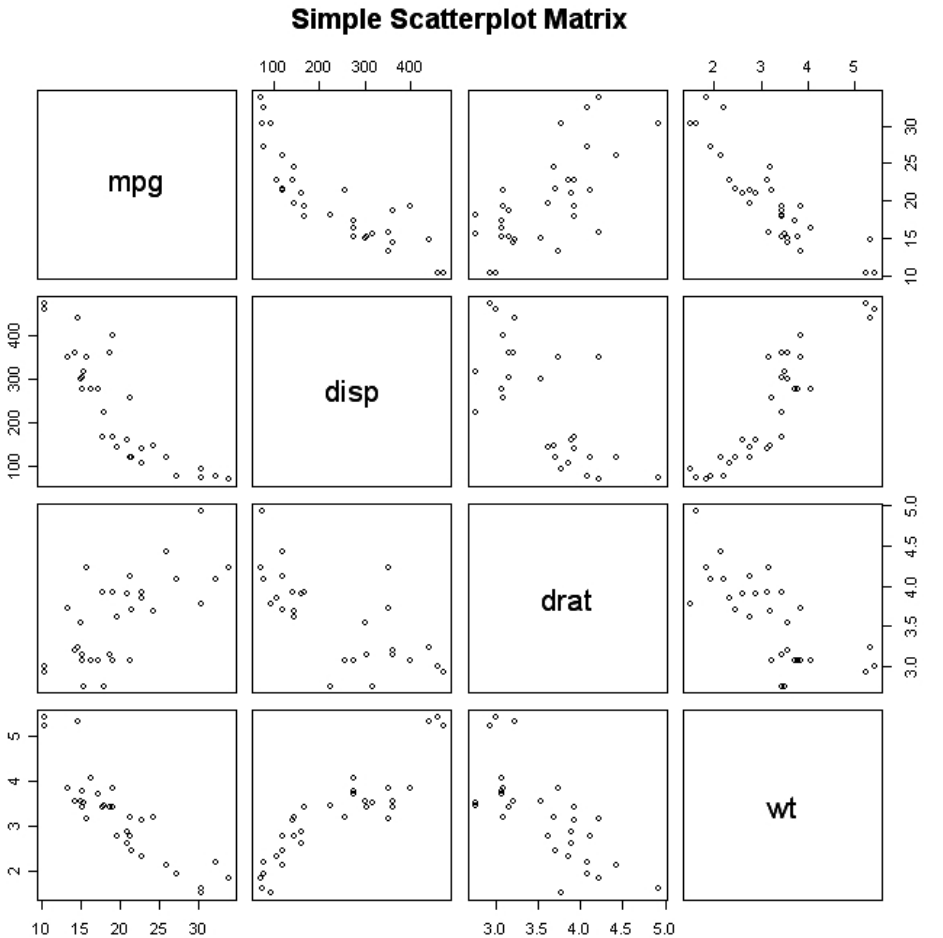
Box and Whiskers Plot



Scatter plots



Scatter Plot Matrices



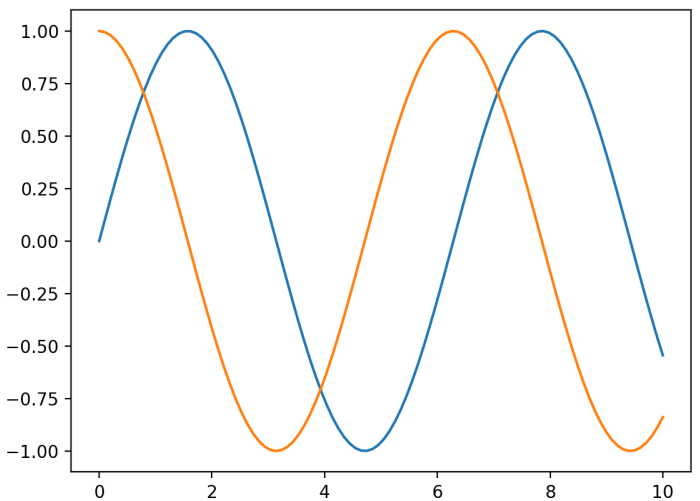
Demo: Data Visualization

Demo: Data Visualization

- there are screenshots in this presentation to maintain continuity, but let's open the notebook named **`Demo - Data Visualization.ipynb`** and go through it together, then there is an exercise to do on your own
- when done, click here to skip screenshots

```
# show.py

import matplotlib.pyplot as plt
import numpy as np

x = np.linspace(0, 10, 100)

plt.plot(x, np.sin(x))
plt.plot(x, np.cos(x))

plt.show()
```

```
> python show.py
```



Saving Image Files

In [5]:
```
fig.savefig('my_figure.png')
```

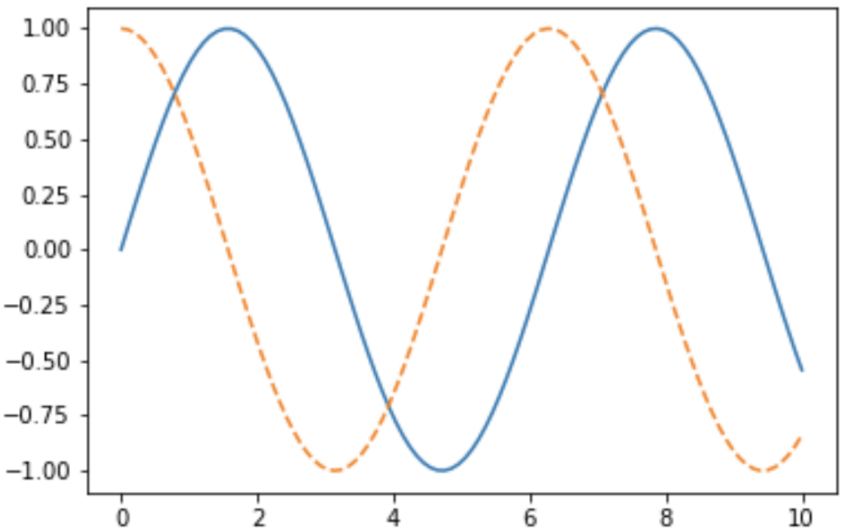
In [6]:
```
!ls -lh my_figure.png
```

```
-rw-r--r--  1 brian  staff    22K Jan 24 06:47 my_figure.png
```

Show Image Files

In [7]:
```
from IPython.display import Image
Image('my_figure.png')
```

Out[7]:



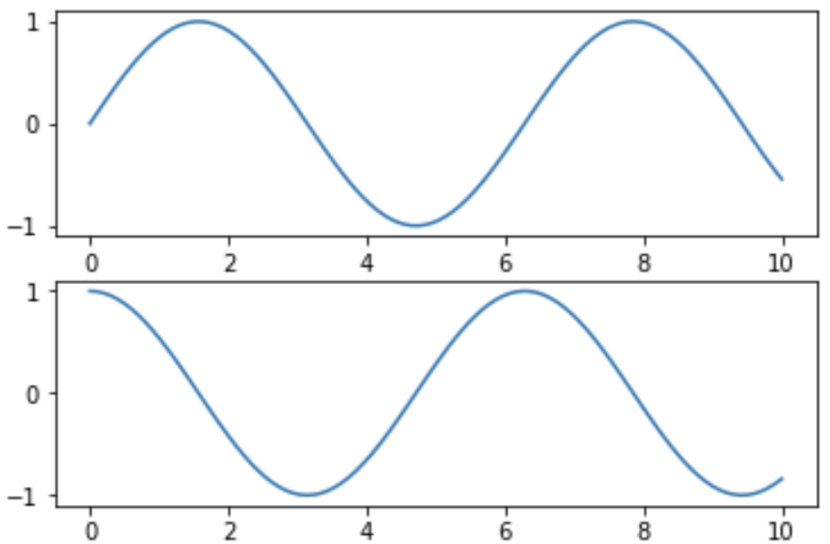
Show Available File Support

```
In [8]: fig.canvas.get_supported_filetypes()
```

```
Out[8]: {u'eps': u'Encapsulated Postscript',
         u'jpeg': u'Joint Photographic Experts Group',
         u'jpg': u'Joint Photographic Experts Group',
         u'pdf': u'Portable Document Format',
         u'pgf': u'PGF code for LaTeX',
         u'png': u'Portable Network Graphics',
         u'ps': u'Postscript',
         u'raw': u'Raw RGBA bitmap',
         u'rgba': u'Raw RGBA bitmap',
         u'svg': u'Scalable Vector Graphics',
         u'svgz': u'Scalable Vector Graphics',
         u'tif': u'Tagged Image File Format',
         u'tiff': u'Tagged Image File Format'}
```

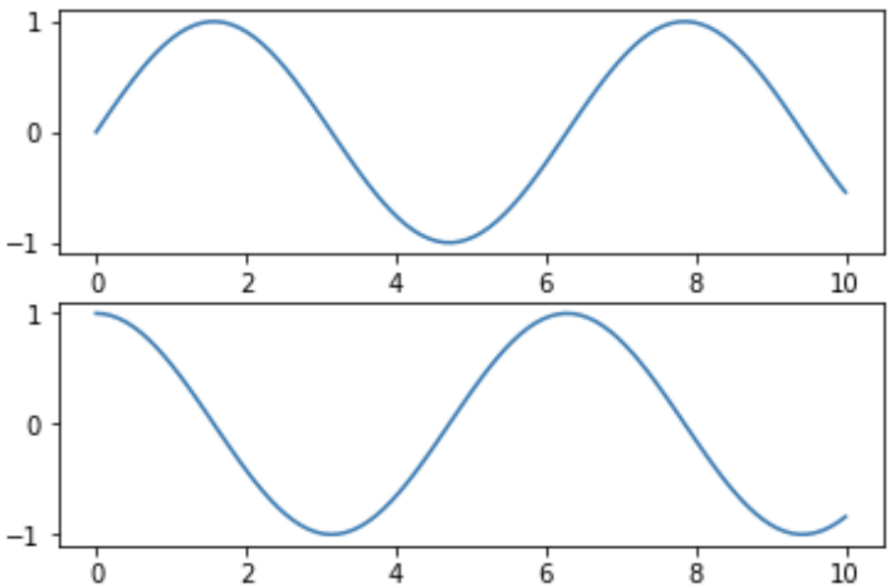
MATLAB-Style Interface

In [9]:
```
plt.figure()  # create a plot figure

# create the first of two panels and set current axis
plt.subplot(2, 1, 1) # (rows, columns, panel number)
plt.plot(x, np.sin(x))

# create the second panel and set current axis
plt.subplot(2, 1, 2)
plt.plot(x, np.cos(x));
```



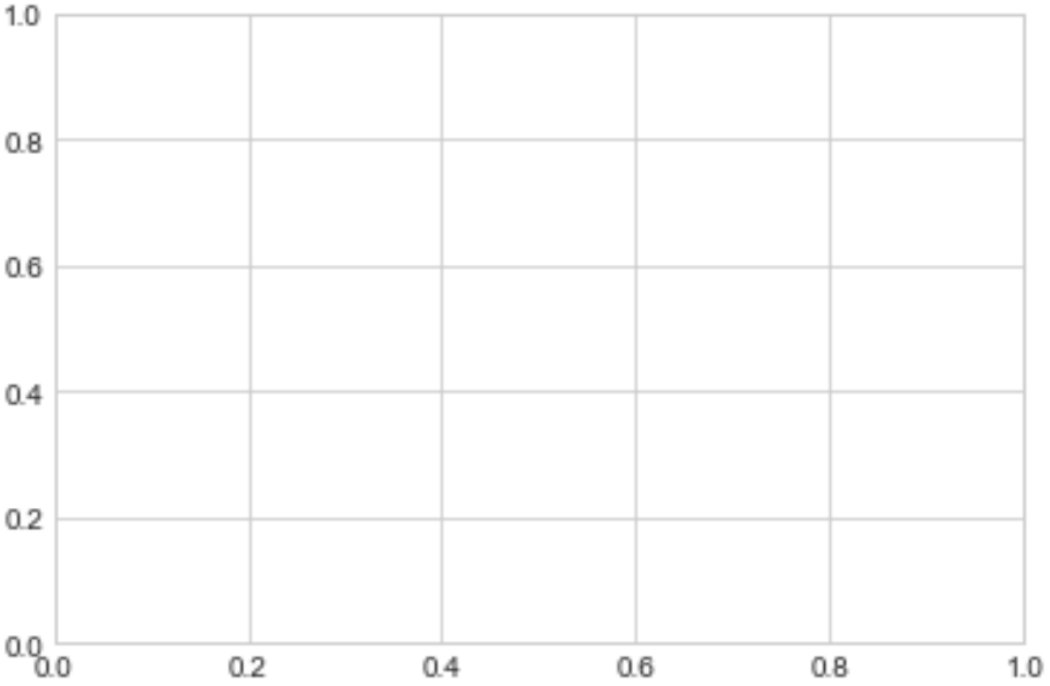
Object-Oriented Interface

In [10]:
```
# First create a grid of plots
# ax will be an array of two Axes objects
fig, ax = plt.subplots(2)

# Call plot() method on the appropriate object
ax[0].plot(x, np.sin(x))
ax[1].plot(x, np.cos(x));
```



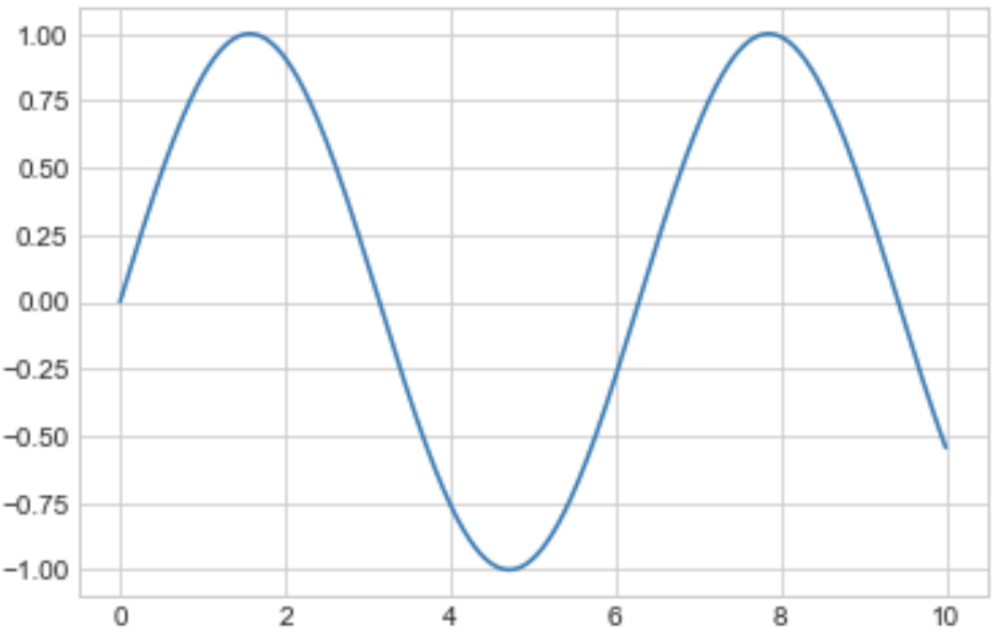
Set up a Grid

In [11]:
```
plt.style.use('seaborn-whitegrid')
fig = plt.figure()
ax = plt.axes()
```



Draw a Function

In [14]:
```
plt.style.use('seaborn-whitegrid')
fig = plt.figure()
ax = plt.axes()
x = np.linspace(0, 10, 1000)
ax.plot(x, np.sin(x));
```

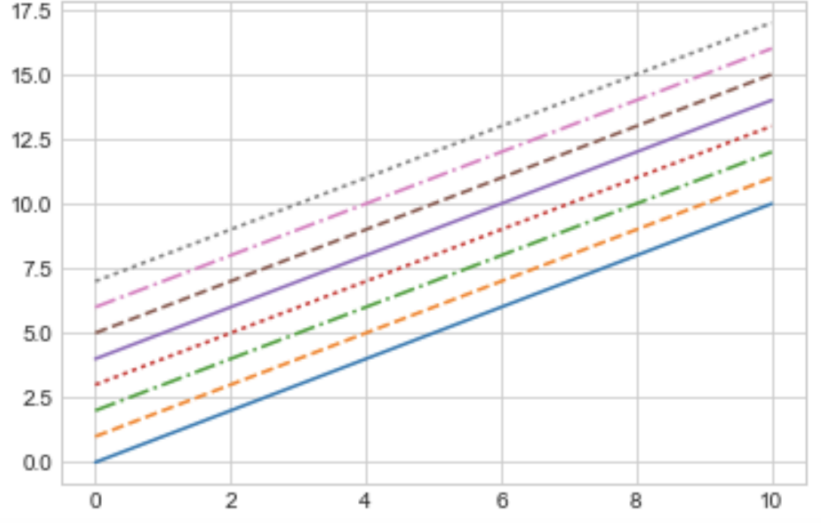


Ways to Specify Color

In [15]:
```
plt.plot(x, np.sin(x - 0), color='blue')        # specify color by name
plt.plot(x, np.sin(x - 1), color='g')           # short color code (rgbcmyk)
plt.plot(x, np.sin(x - 2), color='0.75')        # Grayscale between 0 and 1
plt.plot(x, np.sin(x - 3), color='#FFDD44')     # Hex code (RRGGBB from 00 to FF)
plt.plot(x, np.sin(x - 4), color=(1.0,0.2,0.3)) # RGB tuple, values 0 to 1
plt.plot(x, np.sin(x - 5), color='chartreuse'); # all HTML color names supported
```

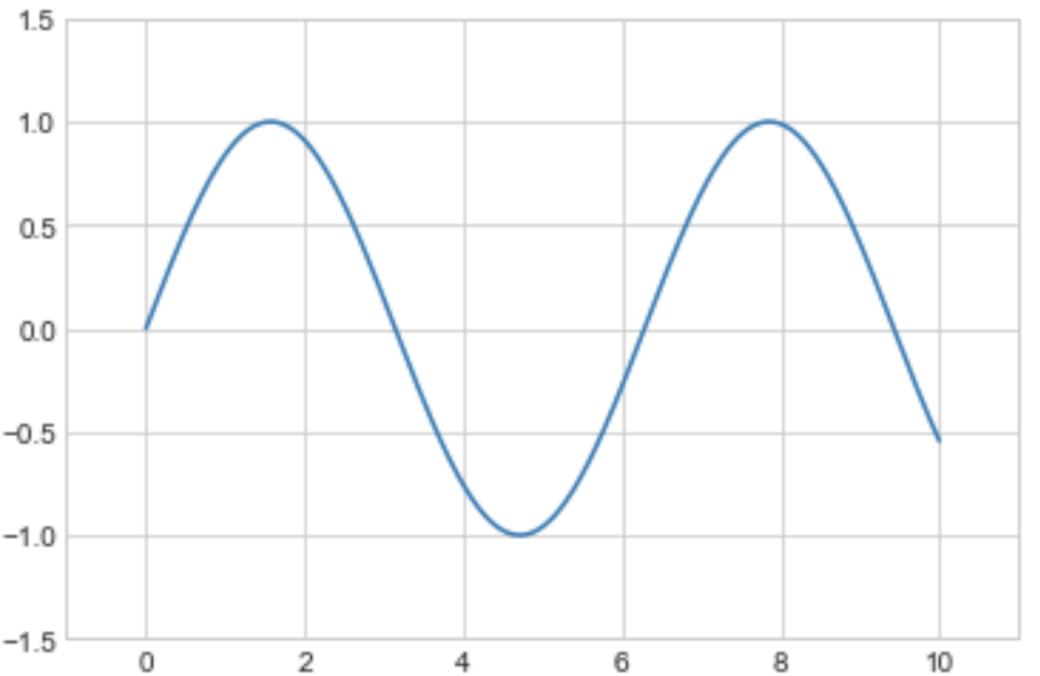
Ways to Specify Line Style

In [16]:
```
plt.plot(x, x + 0, linestyle='solid')
plt.plot(x, x + 1, linestyle='dashed')
plt.plot(x, x + 2, linestyle='dashdot')
plt.plot(x, x + 3, linestyle='dotted');

# For short, you can use the following codes:
plt.plot(x, x + 4, linestyle='-')  # solid
plt.plot(x, x + 5, linestyle='--') # dashed
plt.plot(x, x + 6, linestyle='-.') # dashdot
plt.plot(x, x + 7, linestyle=':');  # dotted
```



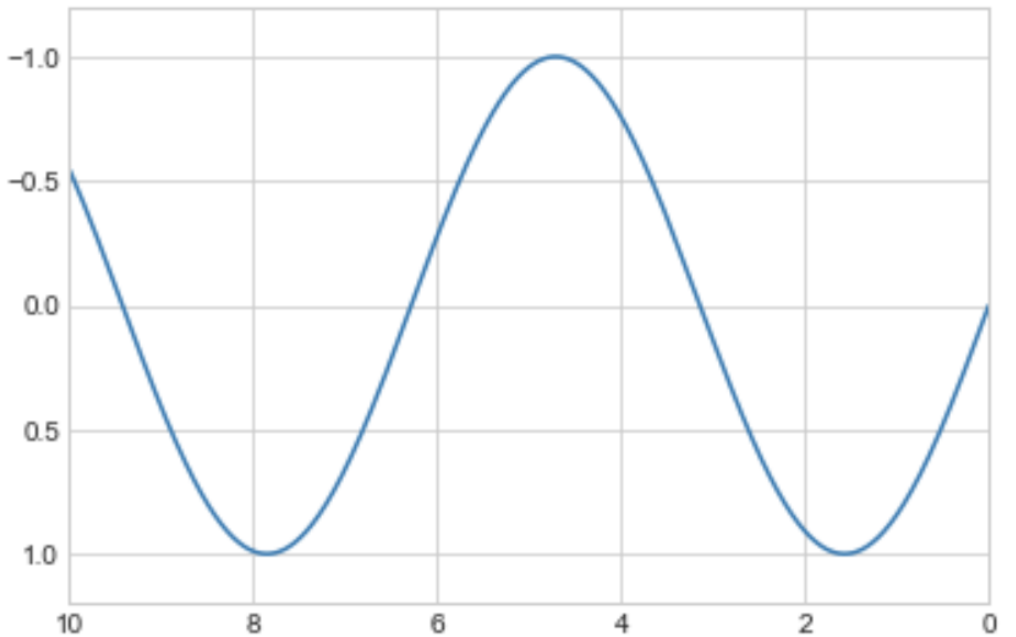
Setting Axes Limits

In [17]:
```
plt.plot(x, np.sin(x))

plt.xlim(-1, 11)
plt.ylim(-1.5, 1.5);
```



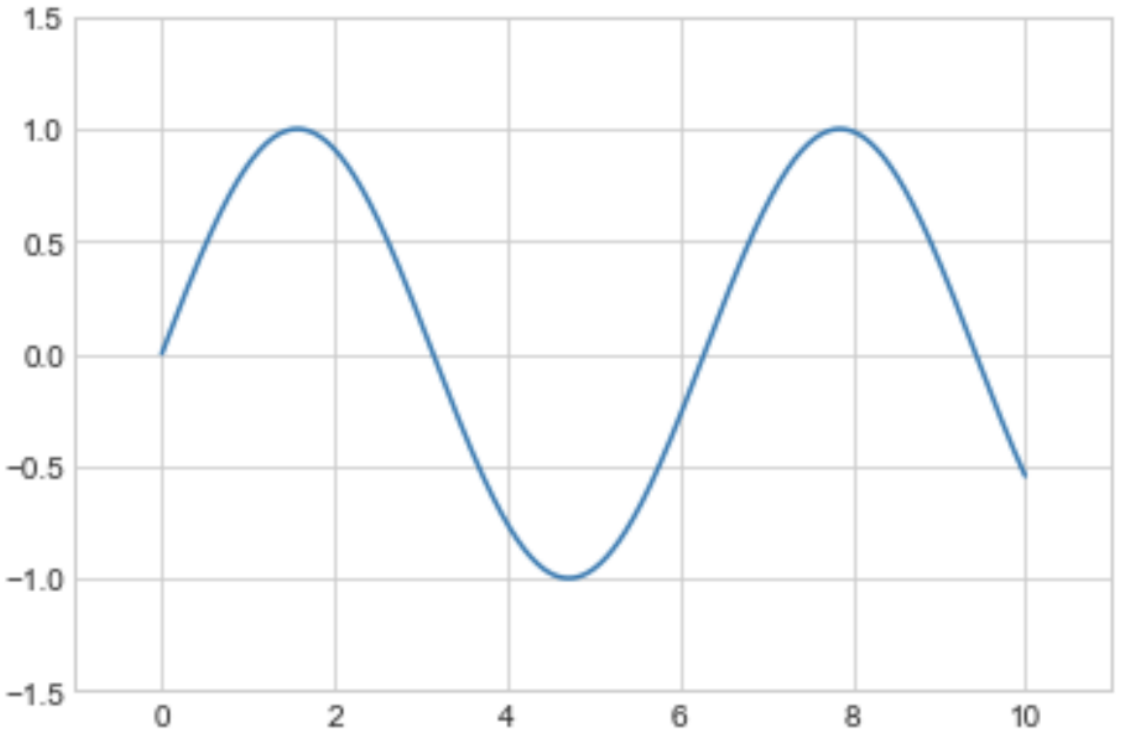
Flipping the Axes Limits

In [18]:
```
plt.plot(x, np.sin(x))

plt.xlim(10, 0)
plt.ylim(1.2, -1.2);
```



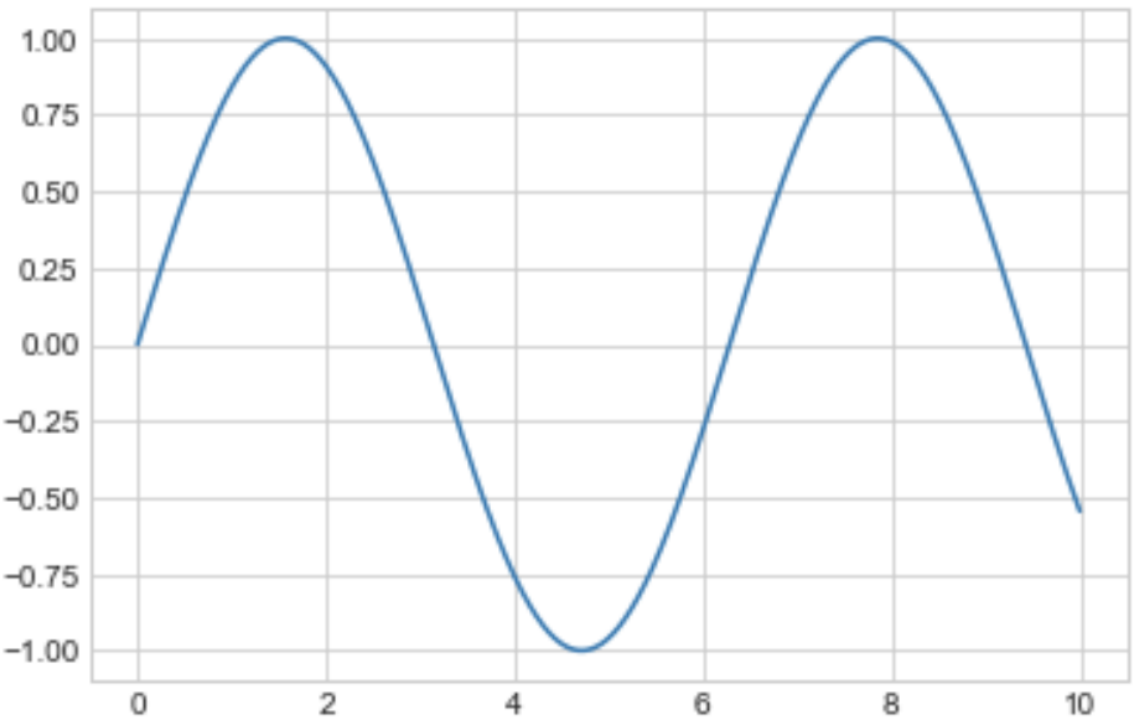
Axis

In [19]:
```
plt.plot(x, np.sin(x))
plt.axis([-1, 11, -1.5, 1.5]);
```



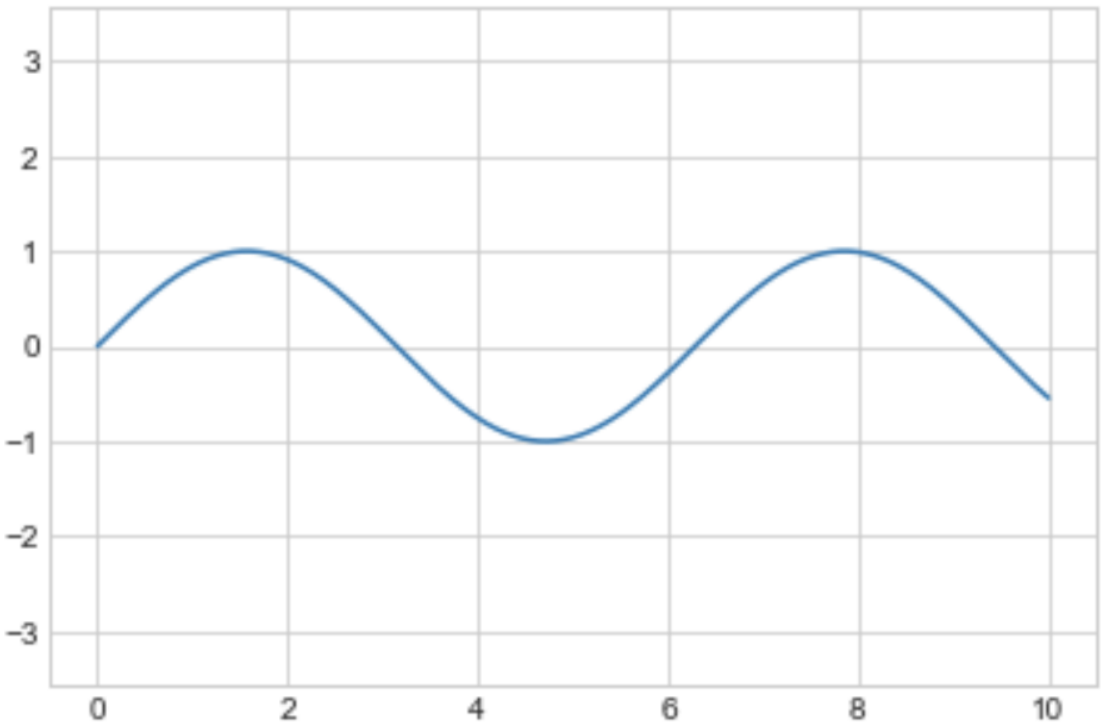
Tight Fit

In [20]:
```
plt.plot(x, np.sin(x))
plt.axis('tight');
```



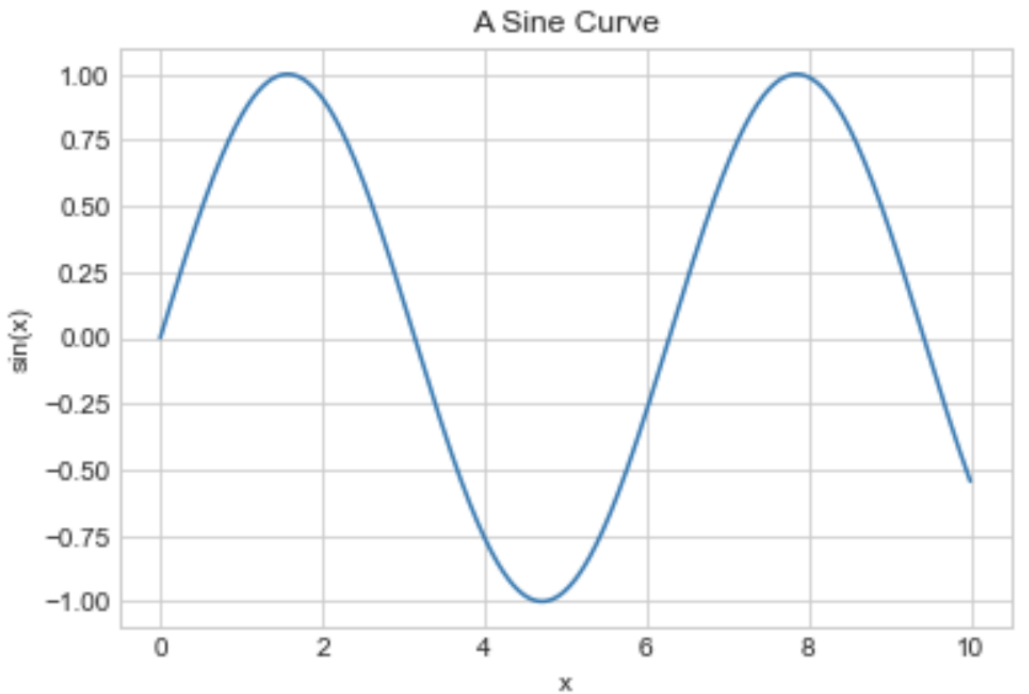
Equal Aspect Ratio

In [21]:
```
plt.plot(x, np.sin(x))
plt.axis('equal');
```



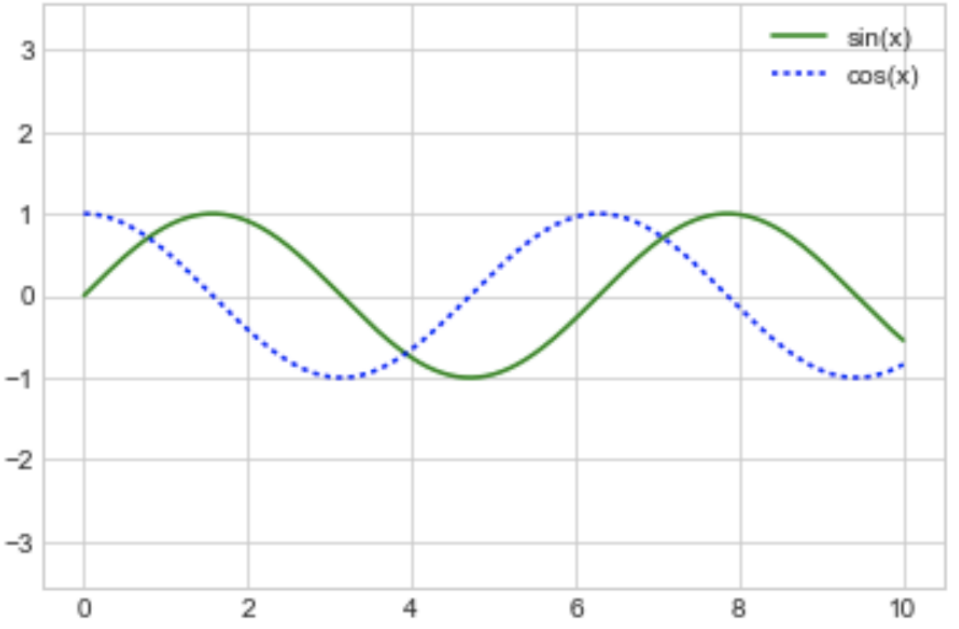
Labels

In [22]:
```
plt.plot(x, np.sin(x))
plt.title("A Sine Curve")
plt.xlabel("x")
plt.ylabel("sin(x)");
```



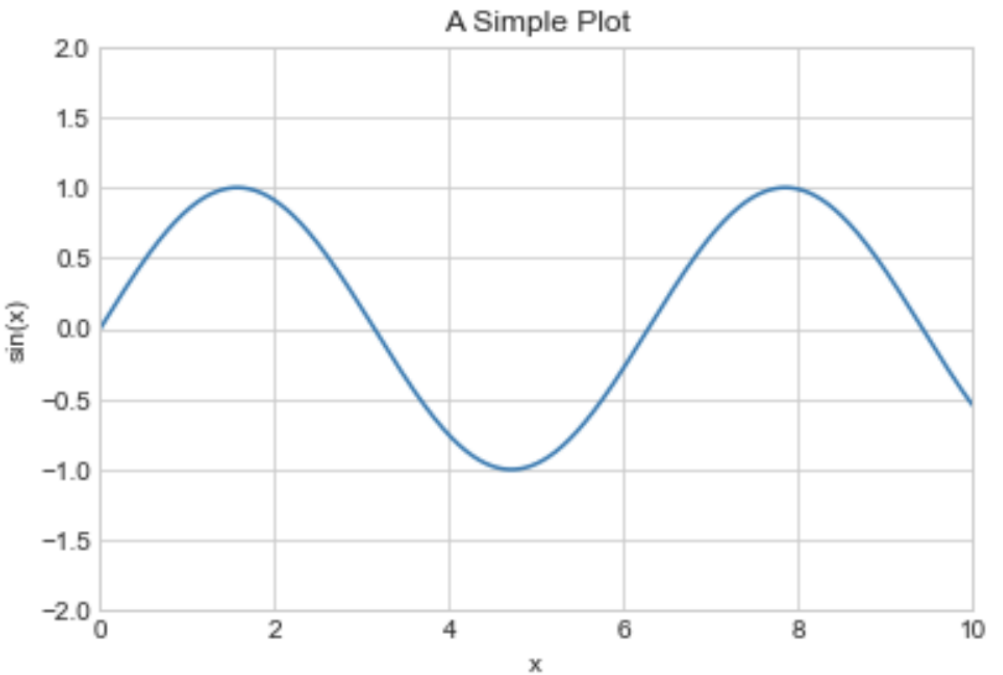
Legends

In [23]:
```
plt.plot(x, np.sin(x), '-g', label='sin(x)')
plt.plot(x, np.cos(x), ':b', label='cos(x)')
plt.axis('equal')

plt.legend();
```



Object-Oriented Interface

In [24]:
```
ax = plt.axes()
ax.plot(x, np.sin(x))
ax.set(xlim=(0, 10), ylim=(-2, 2),
       xlabel='x', ylabel='sin(x)',
       title='A Simple Plot');
```

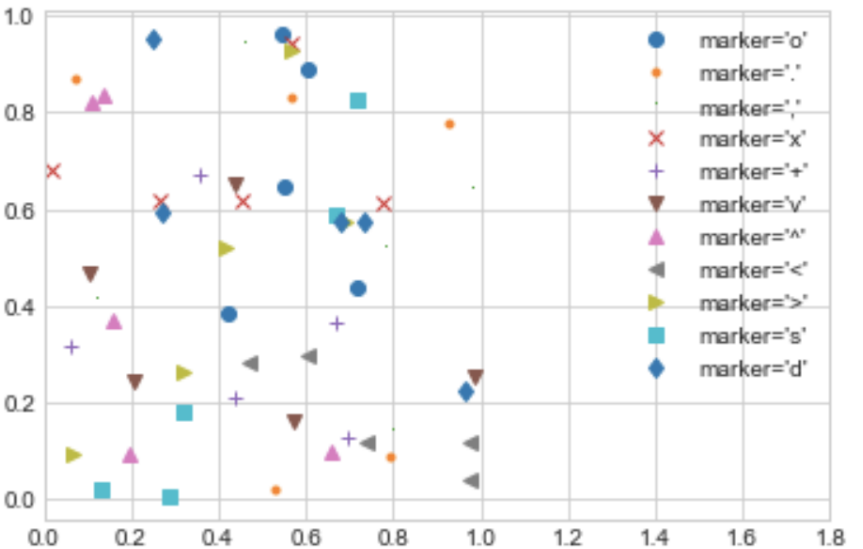


Interface Differences

MATLAB-Style	OO Style
plt.xlabel()	ax.set_xlabel()
plt.ylabel()	ax.set_ylabel()
plt.xlim()	ax.set_xlim()
plt.ylim()	ax.set_ylim()
plt.title()	ax.set_title()

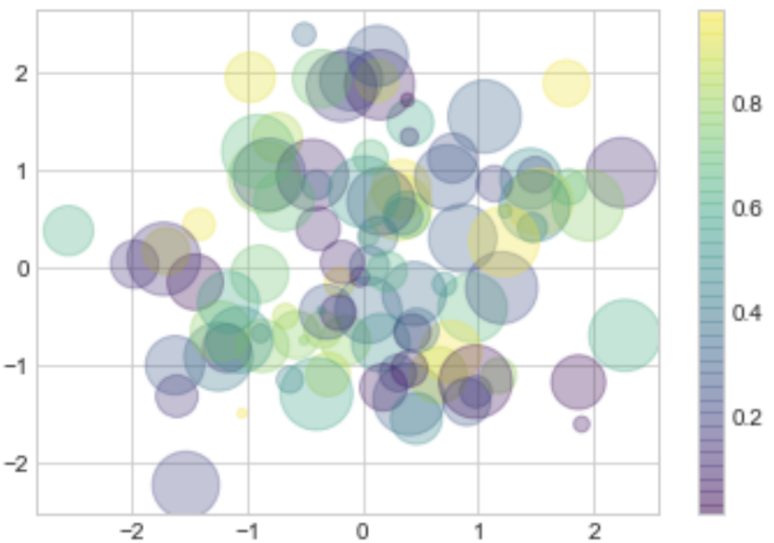
Specifying Markers

In [25]:
```
rng = np.random.RandomState(0)
for marker in ['o', '.', ',', 'x', '+', 'v', '^', '<', '>', 's', 'd']:
    plt.plot(rng.rand(5), rng.rand(5), marker,
             label="marker='{0}'".format(marker))
plt.legend(numpoints=1)
plt.xlim(0, 1.8);
```



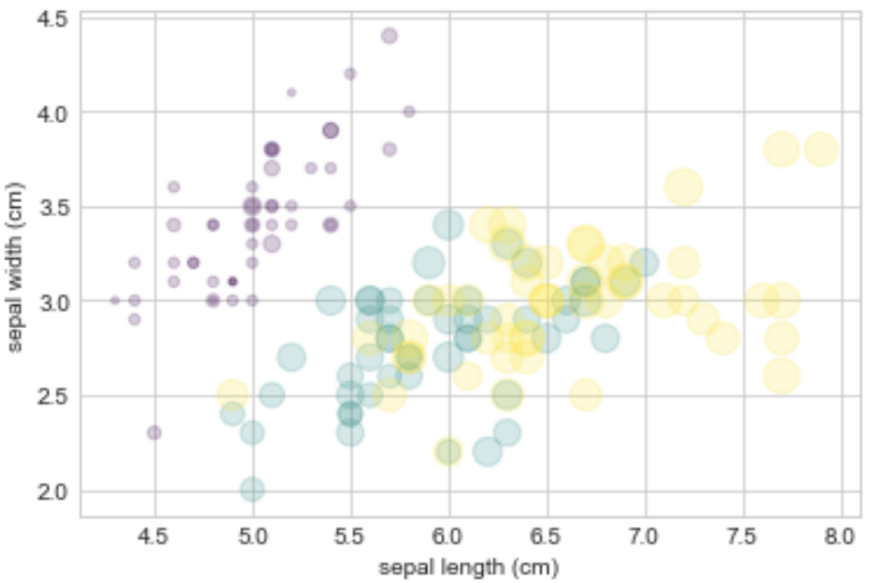
Scatterplot with Colors and Sizes

In [26]:
```
rng = np.random.RandomState(0)
x = rng.randn(100)
y = rng.randn(100)
colors = rng.rand(100)
sizes = 1000 * rng.rand(100)

plt.scatter(x, y, c=colors, s=sizes, alpha=0.3,
            cmap='viridis')
plt.colorbar();  # show color scale
```



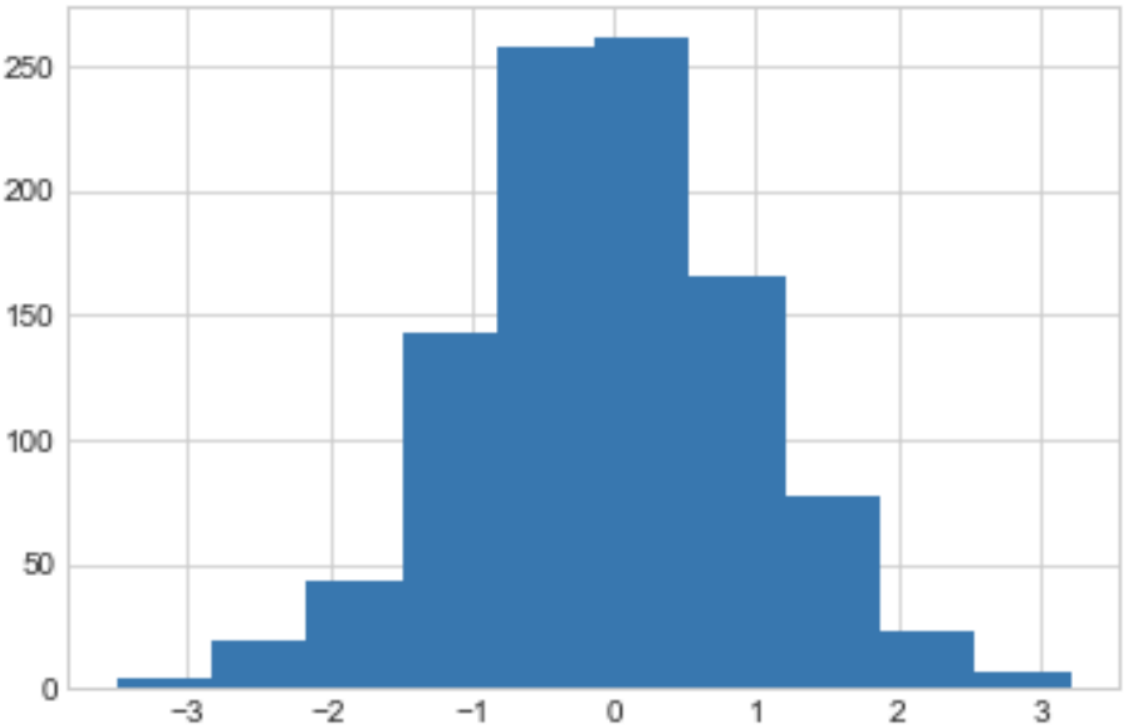
Visualizing Multiple Dimensions

In [27]:
```
from sklearn.datasets import load_iris
iris = load_iris()
features = iris.data.T

plt.scatter(features[0], features[1], alpha=0.2,
            s=100*features[3], c=iris.target, cmap='viridis')
plt.xlabel(iris.feature_names[0])
plt.ylabel(iris.feature_names[1]);
```



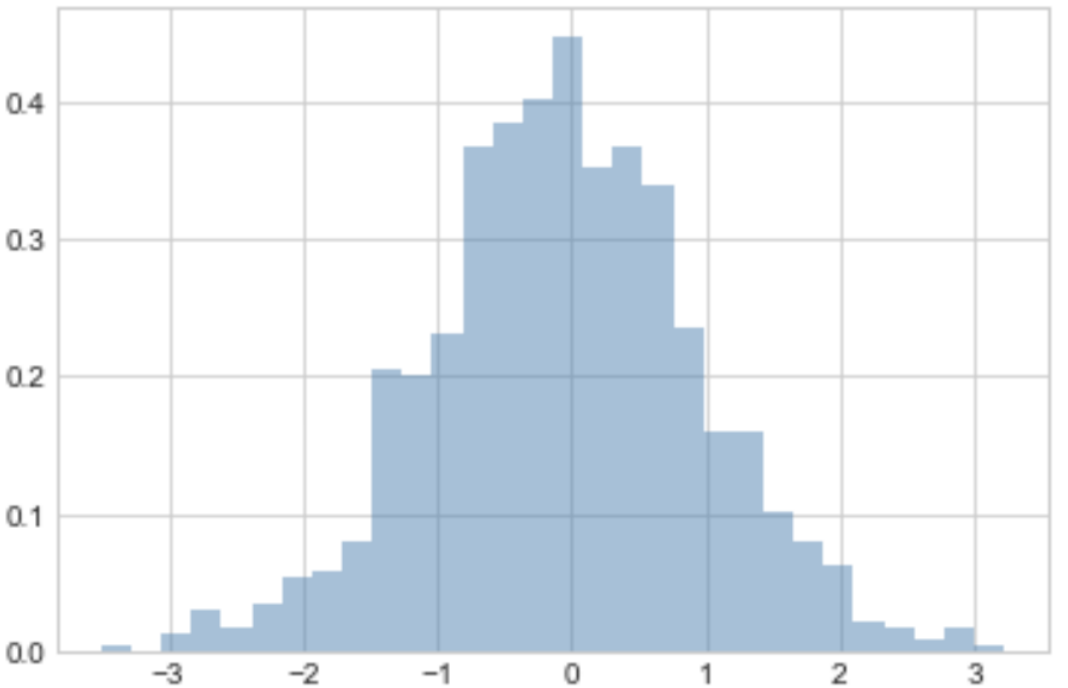
Histograms

In [28]:
```
data = np.random.randn(1000)
plt.hist(data);
```



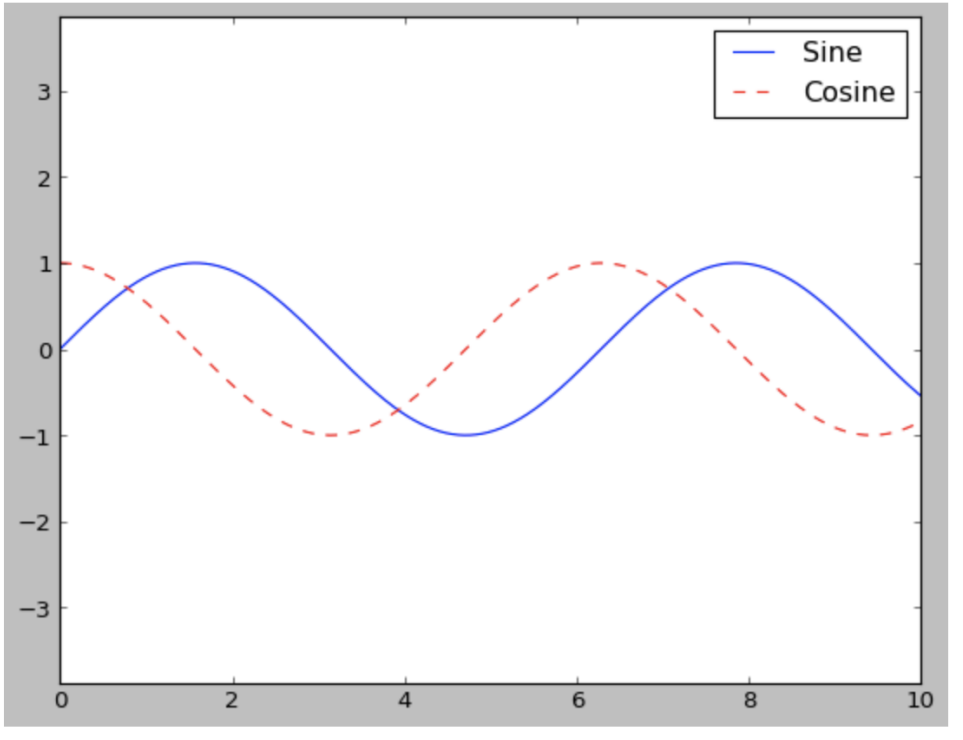
Customizing Histograms

In [29]:
```
plt.hist(data, bins=30, normed=True, alpha=0.5,
         histtype='stepfilled', color='steelblue',
         edgecolor='none');
```



Customizing Legends

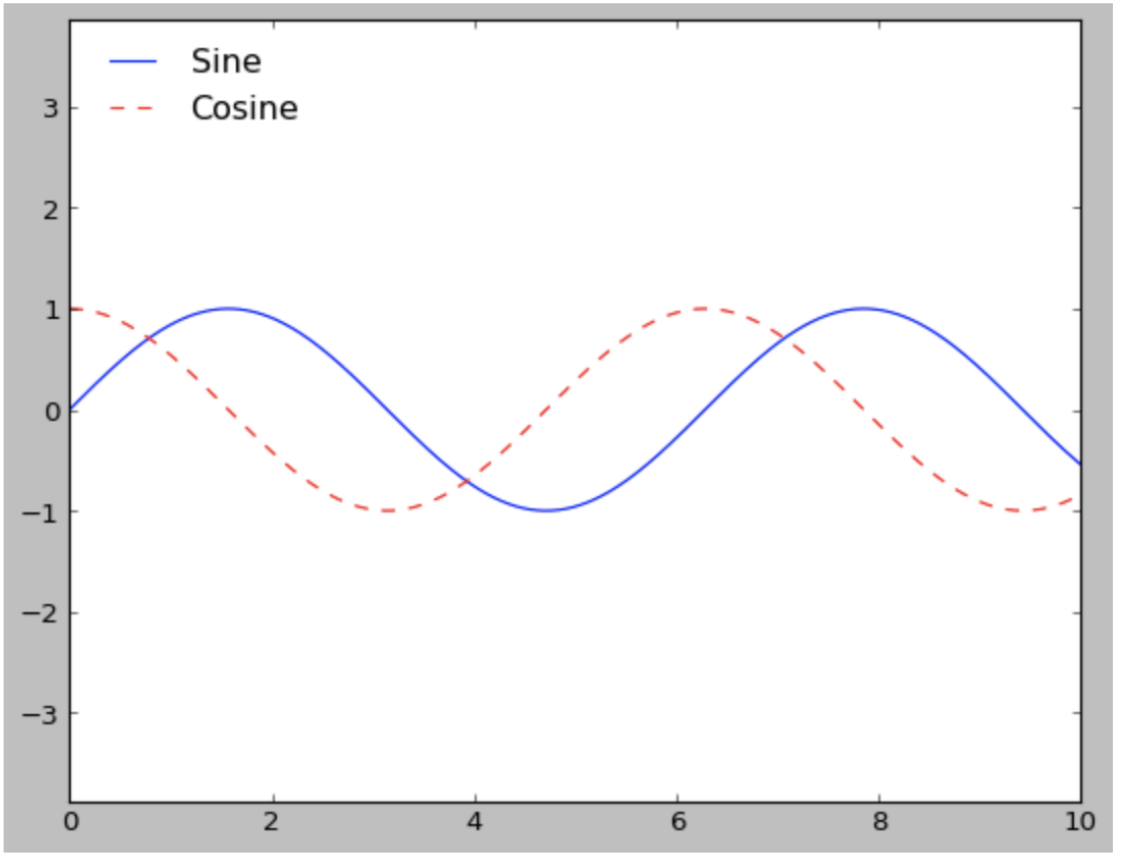
In [33]:
```
plt.style.use('classic')
x = np.linspace(0, 10, 1000)
fig, ax = plt.subplots()
ax.plot(x, np.sin(x), '-b', label='Sine')
ax.plot(x, np.cos(x), '--r', label='Cosine')
ax.axis('equal')
leg = ax.legend();
```



Customizing Legends

In [34]:
```
ax.legend(loc='upper left', frameon=False)
fig
```

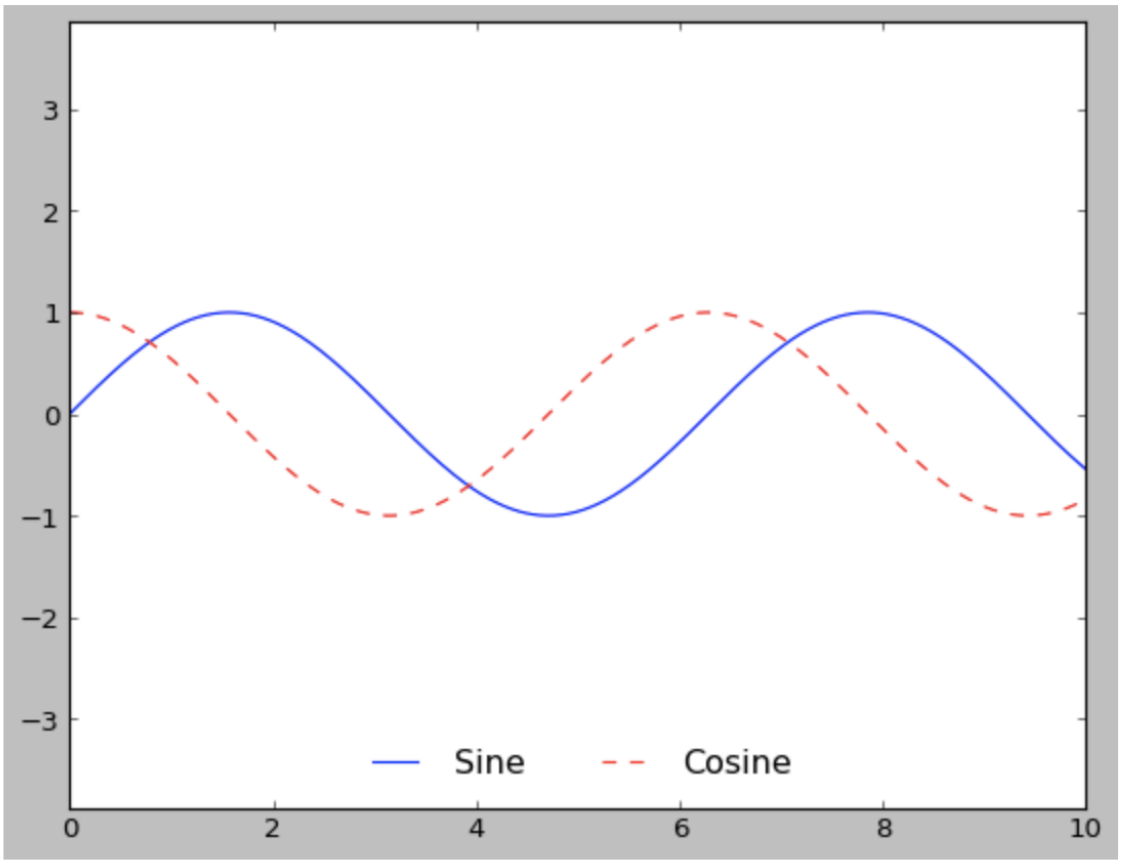
Out[34]:



Customizing Legends

In [35]:
```
ax.legend(frameon=False, loc='lower center', ncol=2)
fig
```

Out[35]:



Exercise: Data Visualization

(open the notebook named **`Exercise 5 - Data Visualization.ipynb`**)

Data Science

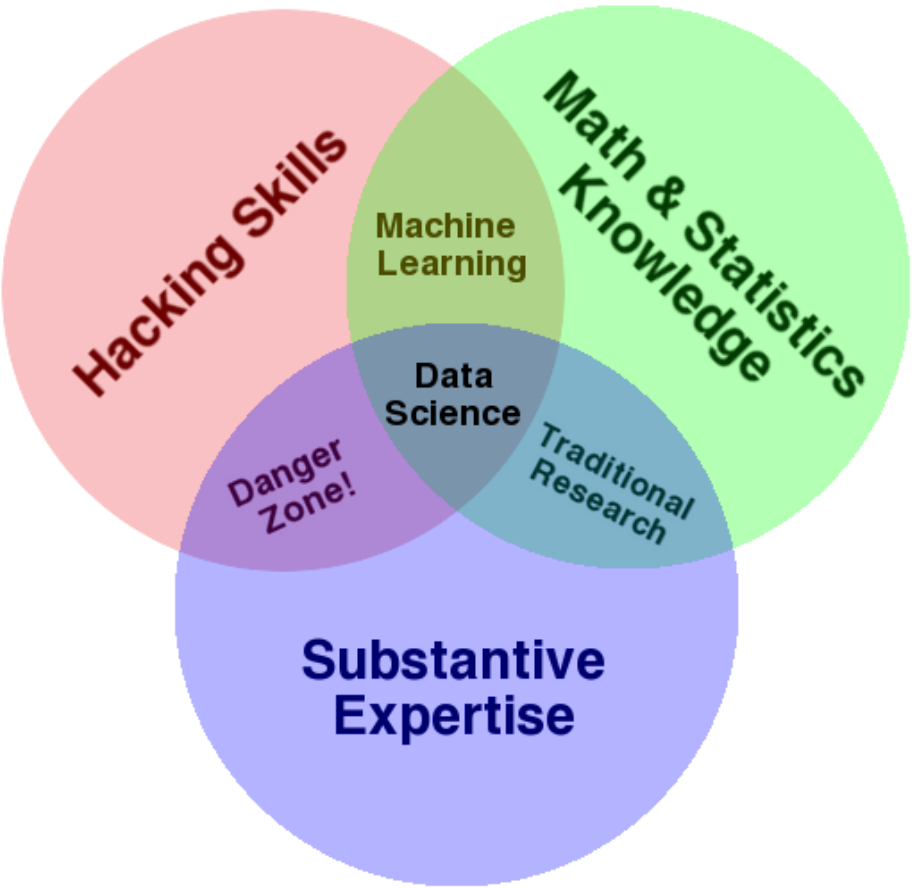
Objective

- Understand the various sides of Data Science
- Consider the activities of a data scientist
- Understand the issues of personal bias and narratives

"Data scientist: n. person who is better at statistics than any software engineer and better at software engineering than any statistician."

Josh Wills

Data Science

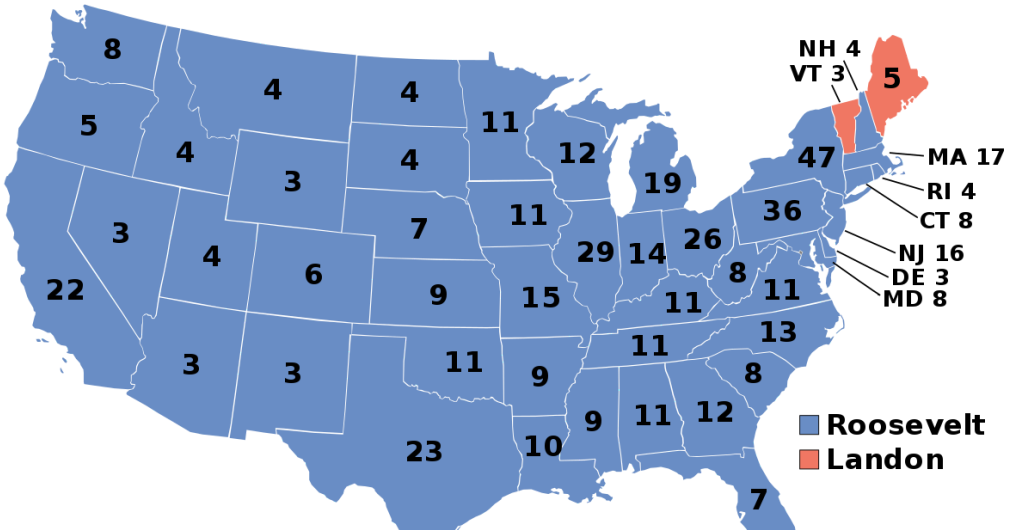


"Drew Conway"

Why is the discipline so important?

Famous example of the consequences of bad data: 1936 U.S. Presidential Election

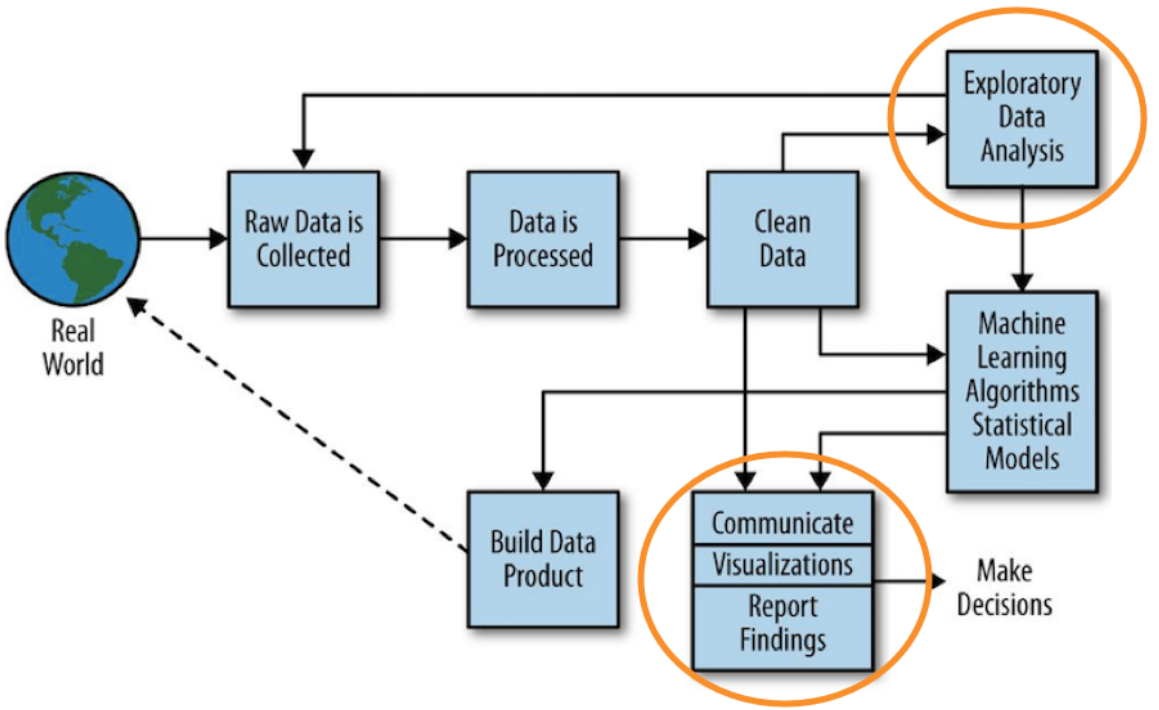
- Literary Digest conducted a mailed survey of 10 million (!) people in order to predict the winner of the 1936 U.S. presidential election
- participants chosen from telephone directories, lists of magazine subscribers, club rosters
- 2.5 million (!) people responded
- overwhelmingly believed Alf Landon would beat Franklin Roosevelt
- George Gallup polled 50,000 people and got it right...why?



"Long before worrying about how to convince others, you first have to understand what's happening yourself."

Andrew Gelman

Data Science Pipeline



"Doing Data Science"

"Naïve realism, also known as direct realism or common sense realism, is a philosophy of mind rooted in a theory of perception that claims that the senses provide us with direct awareness of the external world."

http://en.wikipedia.org/wiki/Naïve_realism

1951 Princeton/Dartmouth Football Game

- Storied rivalry
- Princeton's star player's nose was broken
- Another Princeton player retaliated and broke Dartmouth player's leg
- Princeton won (13-0)
- Both teams blamed the other side

"They Saw a Game"

- Albert Hastorf (Dartmouth) and Hadley Cantril (Princeton) decided to study this.
- They showed the game again to students from both schools
- Asked them to notice infractions, penalties, fill out a questionnaire about the game
- Princeton students 'saw' twice as many infractions by Dartmouth players than Dartmouth students did
- Dartmouth students saw a 'rough but fair' game
- **Two versions of the Truth**

"In brief, the data here indicate that there is no such 'thing' as a 'game' existing 'out there' in its own right which people merely 'observe.' The game 'exists' for a person and is experienced by him only insofar as certain happenings have significances in terms of his purpose."

Hastorf and Cantril

"Everything that has ever happened to you has happened inside your skull."

David McRaney, "You Are Now Less Dumb"

Compare the Students

- All male
- Mostly similar ethnically and socioeconomically
- Geographically similar
- Similar in age
- Similar basic cultural and religious beliefs
- Only difference…Went to different schools

The Dress

What colors are in this dress?

The Dress

When there is uncertainty, our brains choose a way to perceive.
But we are often not aware we made a choice.

The Dress Controversy

"It’s a real problem, though, when politicians, CEOs, and other people with the power to change the way the world works start bungling their arguments for or against things based on self-delusion generated by imperfect minds and senses."

David McRaney, "You Are Now Less Dumb"

"Narratives are meaning transmitters. They are history-preservation devices. They create and maintain cultures, and they forge identities that emerge out of the malleable, imperfect memories of life events."

David McRaney, "You Are Now Less Dumb"

"Your narrative bias makes it nearly impossible for you to really absorb the information from the outside world without arranging it into causes and effects."

David McRaney, "You Are Now Less Dumb"

"Your ancestors invented the scientific method because the common belief fallacy renders your default strategies for making sense of the world generally awful and prone to error."

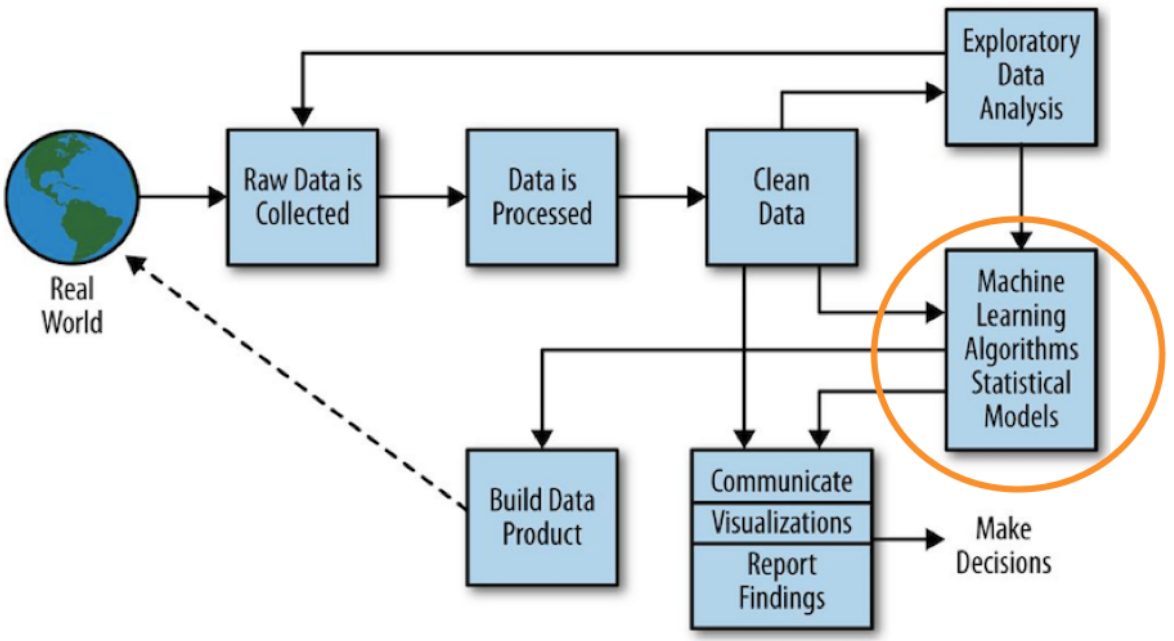
David McRaney, "You Are Now Less Dumb"

Data-Directed

Objective

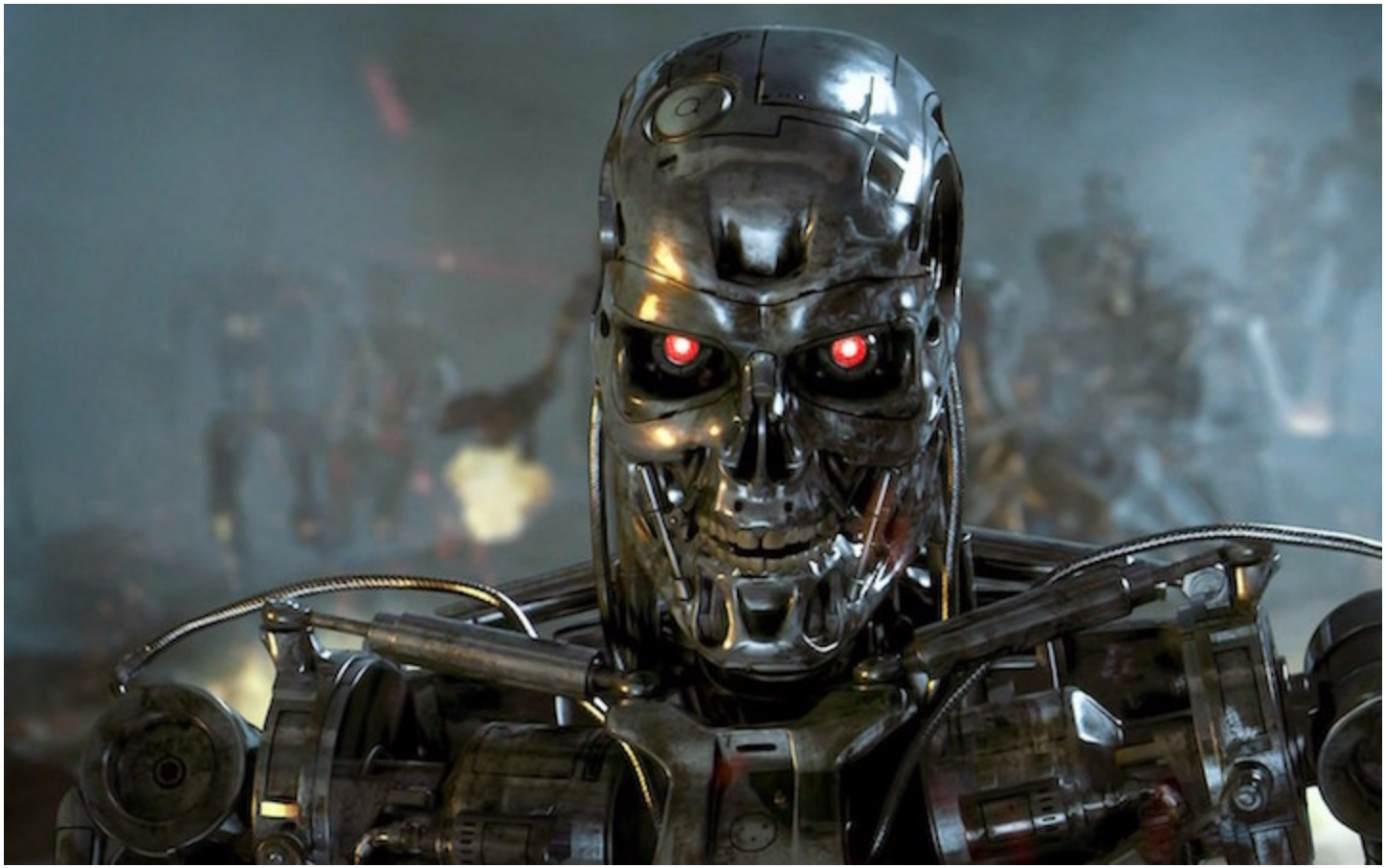
- Understand the basics of machine learning and how it can direct our activities
- Consider the issues of generalizing from observations of the world
- Explore *supervised* and *unsupervised* learning strategies
- Be introduced to a handful of basic and widely-used techniques
- Learn how to use them on basic problems

Data Science Pipeline

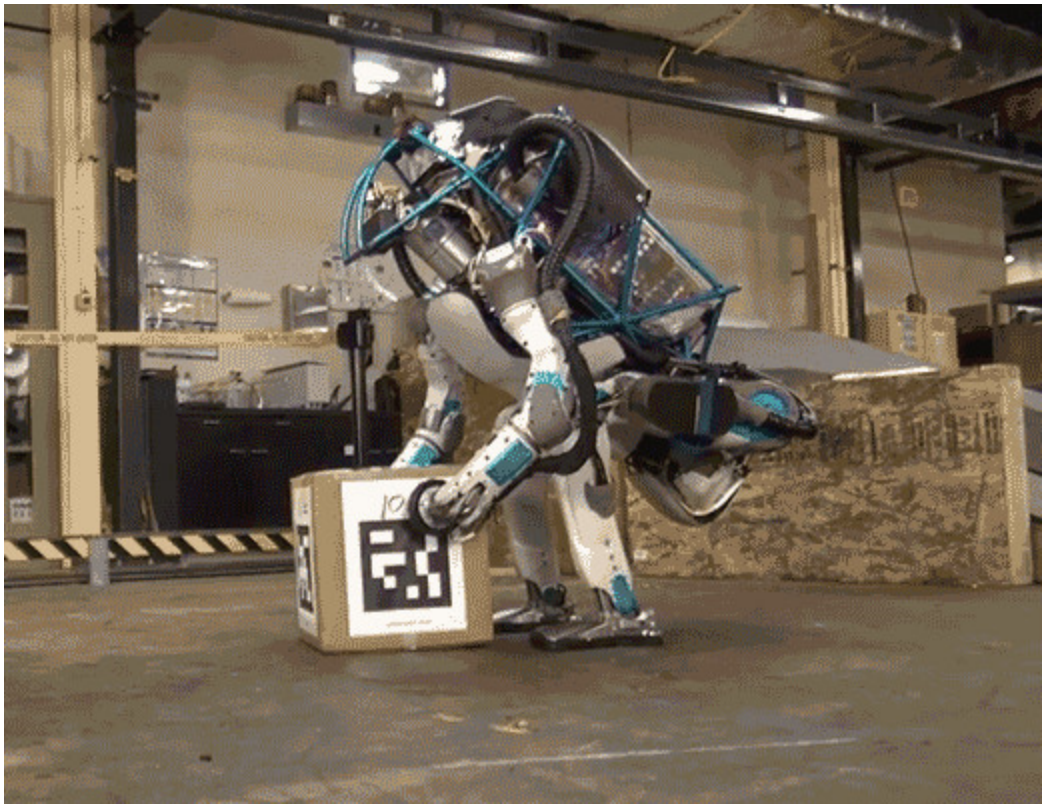


"Doing Data Science"

Is this what happens when robots learn?



Or is it more like this?

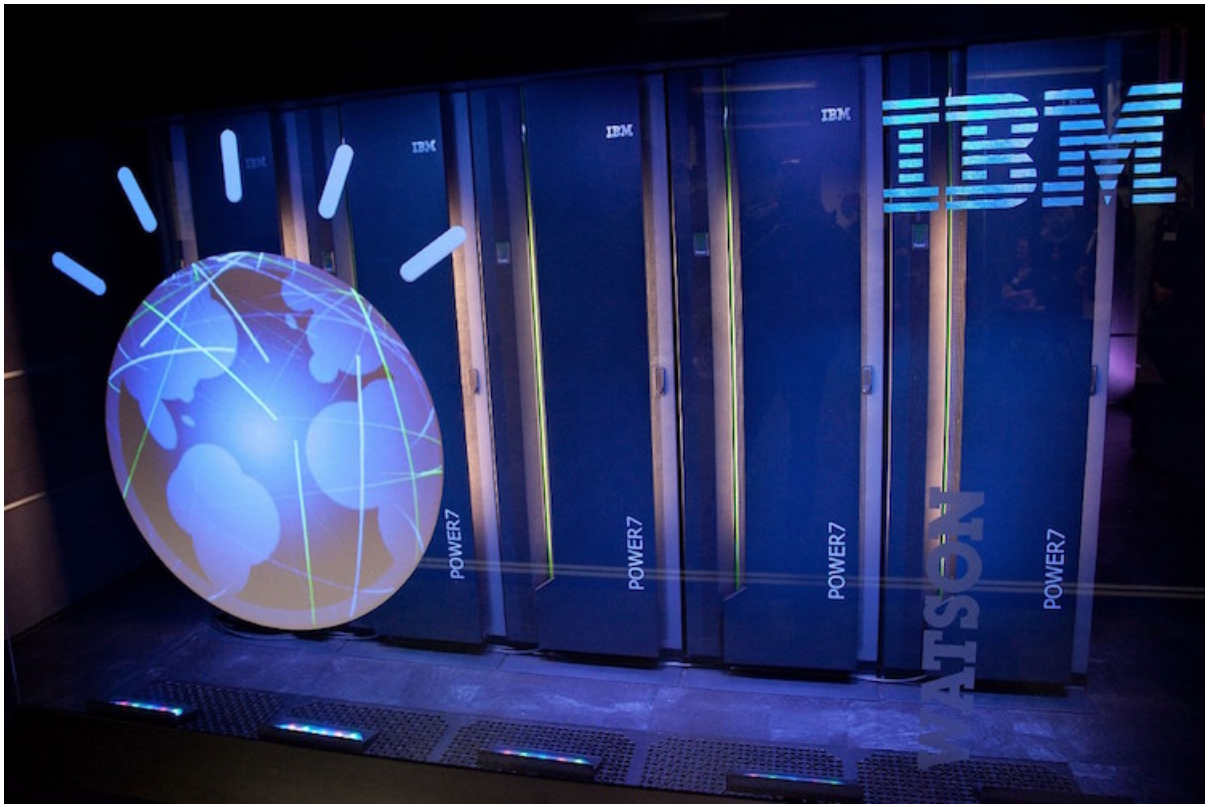


https://youtu.be/rVlhMGQgDkY?t=85

Societal Impact

- New technologies always bring new jobs (and take them away)
- ...and in some cases these technologies is distrupting faster than we can retrain displaced workers
  - JP Morgan has software to find anomalies in contracts–can do in seconds what took lawyers 360,000 hours
  - cancer detection, such as this and this
- How can software perform better than someone with decades worth of education and experience?
  - (Note that in the second example above, radiologists were outperformed by software built by non-experts)
- Self-driving cars and long-haul trucks
  - not only will drivers be out of work, think of the truck stops
- No easy answers, but it's clear we have to consider these issues!

IBM's Question Answering System–Watson



https://www.youtube.com/watch?v=WFR3lOm_xhE&t=20

B.F. Skinner

- trained pigeons to do various tasks including play ping pong by rewarding them for the behavior he wanted them to exhibit (operant conditioning)
- also trained pigeons to be superstitious–what does that mean and what does it have to do with machine learning?

"The term machine learning refers to the automated detection of meaningful patterns in data."

Source: Shavel-Shwartz and Ben-David, "Understanding Machine Learning: From Theory to Algorithms"

Advancement of Machine Learning

- Stock market prediction in the 1980s
- Mining corporate databases in the 1990s (direct marketing, CRM, credit scoring, fraud detection)
- E-commerce (personalization, click analysis)
- 9/11 brought interest to applying ML to fighting terror
- Web 2.0 (social networks, sentiment analysis, etc.)
- Science (molecular biologists and astronomers were early adopters)
- Housing bust freed up a lot of talent
- Big Data

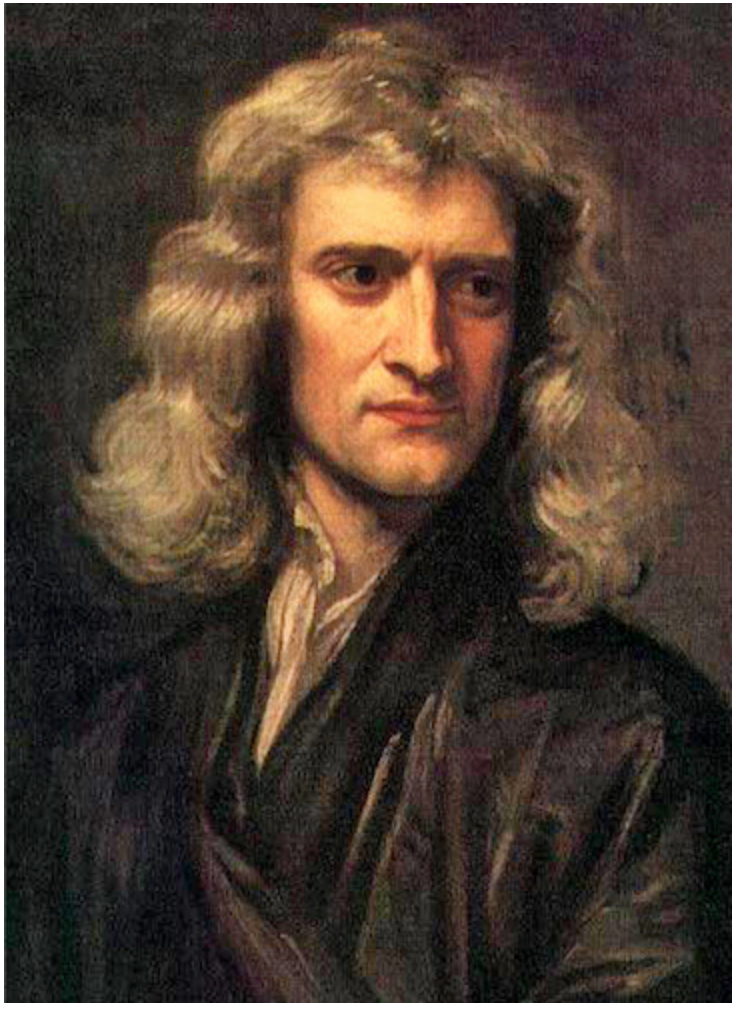
David Hume

Inductivist Turkey

Tycho Brahe

Johannes Kepler

Isaac Newton



Machine Learning Approaches

Supervised Learning

Given input variables x and an output variable Y

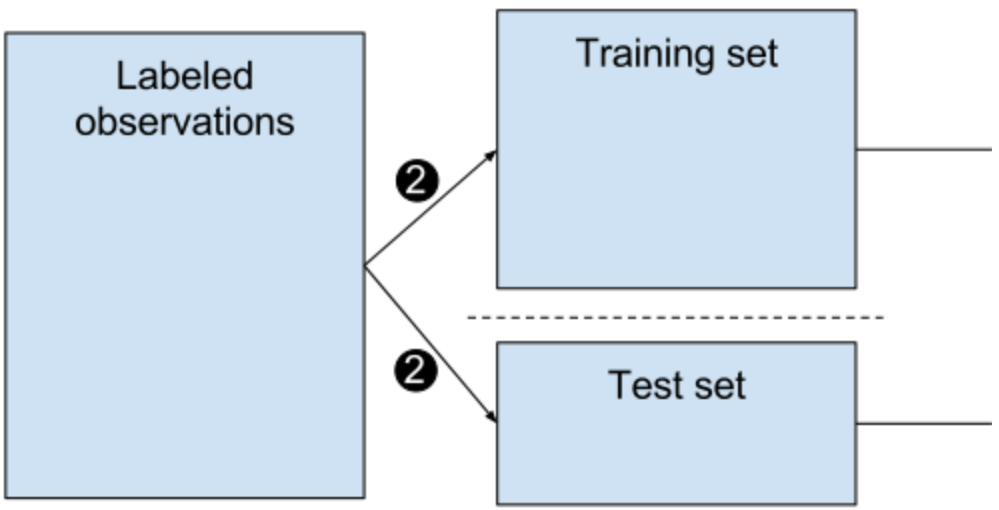
$$Y = f(x)$$

To train the model, the inputs and outputs are known and are used to determine f(x). The trained model f(x) outputs one of two types of outputs: classification and regression.

- Classification, i.e., classifying input into two or more categories (e.g., benign vs. malignant tumors)
- Regression - the output variable takes continuous values (e.g., how much would we expect this house to sell for, based on square footage, location, etc.)

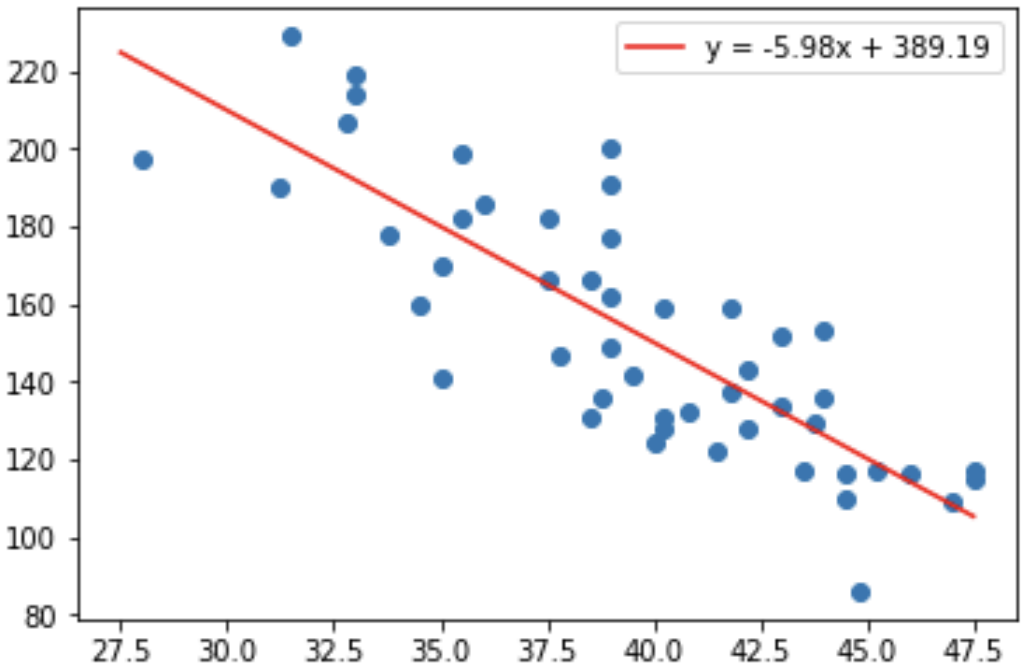
Supervised Learning

partition dataset into training set and test set



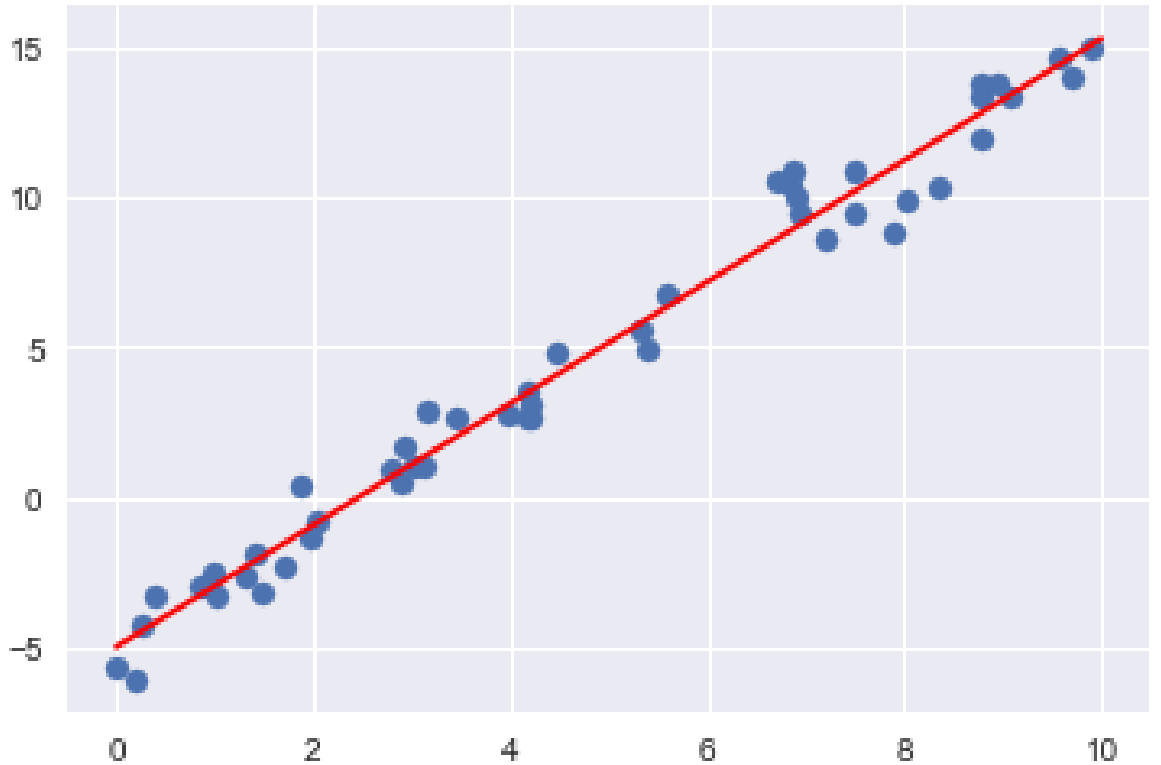
Linear Regression

relating your input (sometimes called "independent variable") to your output ("dependent variable") by fitting a straight line through your data



Ordinary Least Squares

- Simple Linear Regression
- Fit a straight line through the observed points
- Minimizes the sum of square residuals (errors) of the model



Hypothetical Business: The Zappos of Pants

- Where is the business risk?
  - too many returns
- If customers don't know their inseam, it'll be hard for them to measure
- We'd like to ask questions of our customers and use that as a proxy for their inseam
  - ...but we don't know what to ask
- So we gather a bunch of data from a population, which needs to be representative of the population we want to market to
- ...then what?
  - look for a relationship between one of the variables and inseam

Linear Regression

Pros	Cons
Common approach for numeric data	Strong assumptions about the data (linearity)
Easily interpretable	Sensitive to outliers
Estimates strength and size of relationships among features and outcomes	Only numeric features

Demo: Linear Regression

Demo: Linear Regression

- let's open the notebook named **`Demo - Linear Regression.ipynb`** and go through it together

Exercise: Linear Regression

(open the notebook named **`Exercise 6 - Linear Regression.ipynb`**)

Naive Bayes

Naive Bayes

- Family of algorithms to produce probabilistic classifiers based on Bayes Theorem
- Bayes Theorem is named after the Reverend Bayes an 18th century statistician and philosopher
- Never published in his lifetime

Features

- Determines probability based upon context, and the probabilities are updated as more information is received
- Starting probabilities can be arbitrarily set
- Requires relatively little training data
- Often used for text/document classification
- Assumes independence of the features

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(spam|Viagra) = \frac{P(Viagra|spam) \cdot P(spam)}{P(Viagra)}$$

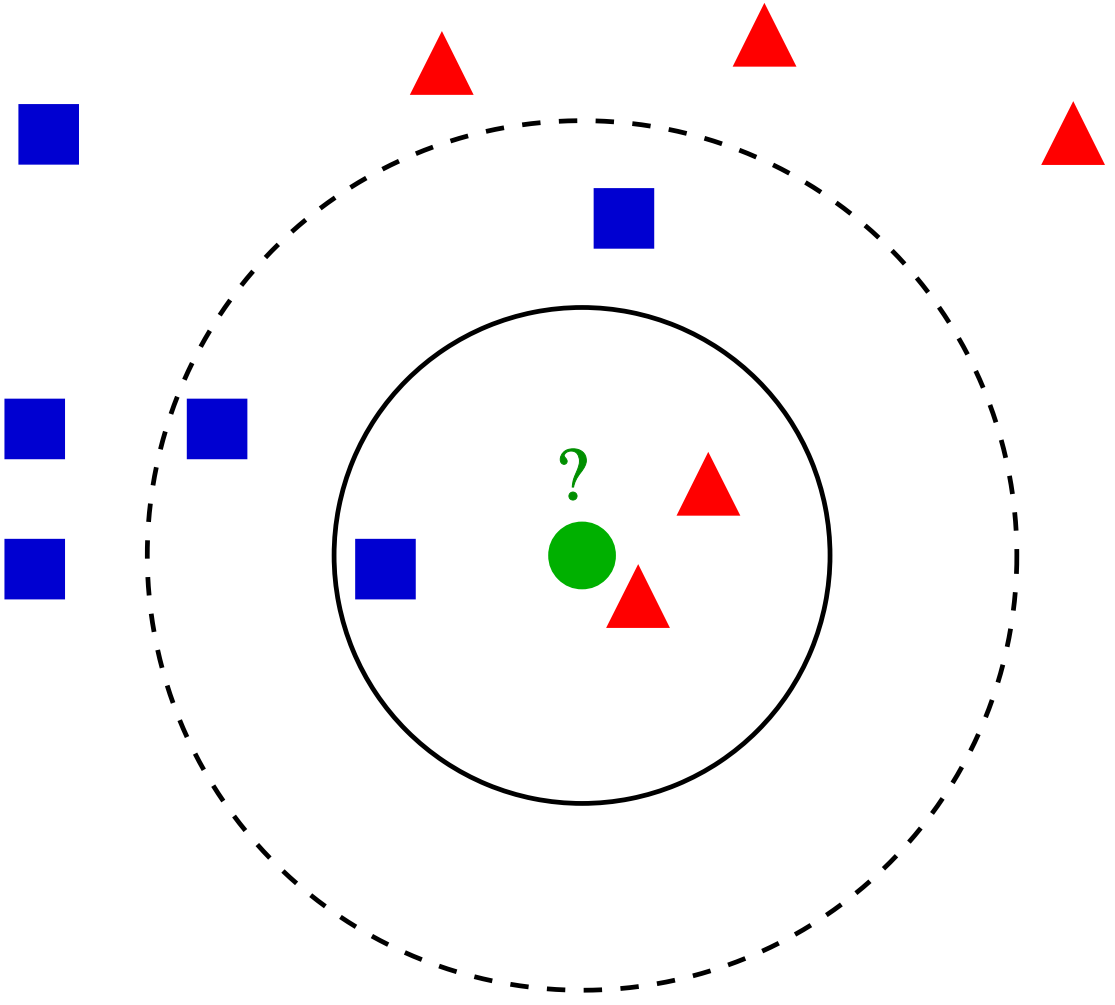
Naive Bayes

Pros	Cons
Simple and effective	Assumption of the independence of features is usually wrong
Does well with noisy and missing data	Doesn't work well with lots of numeric features
Works well with arbitrary sizes of training data	Estimated probabilities aren't as reliable as the classifications
Easy to produce the estimated probability for predictions	

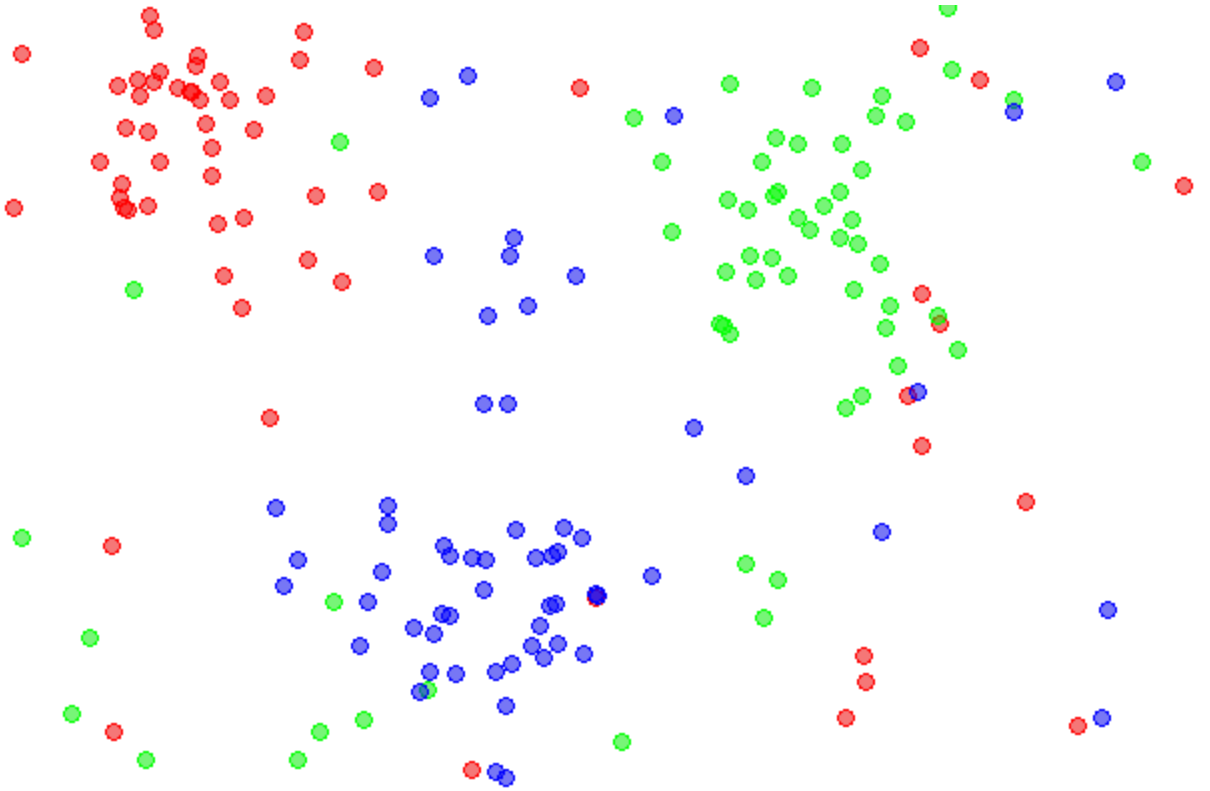
k-Nearest Neighbors

k-Nearest Neighbors

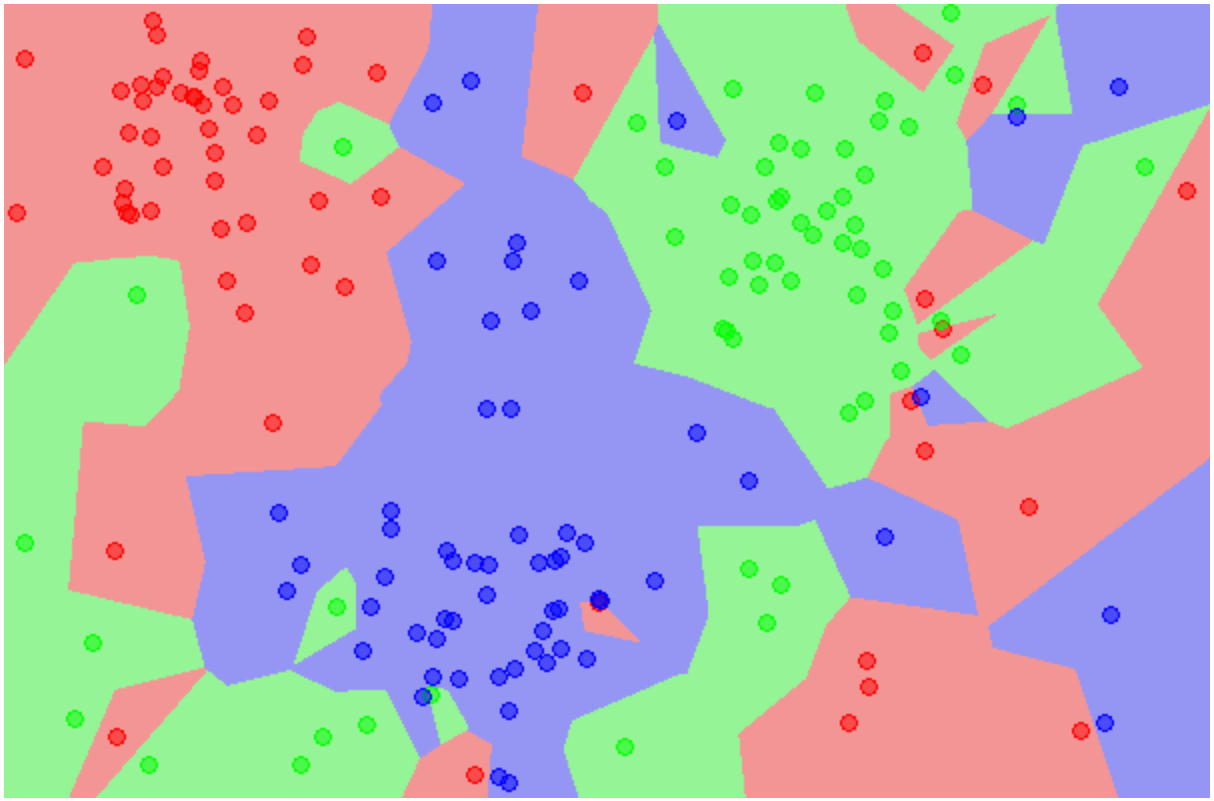
- k-NN Classification
  - output is class membership
  - object is classified by a majority vote of its neighbors
  - for k = 1, then the object is simply assigned to the class of that single nearest neighbor
- k-NN Regression
  - output is the property value for the object
  - this value is the average of the values of its k nearest neighbors



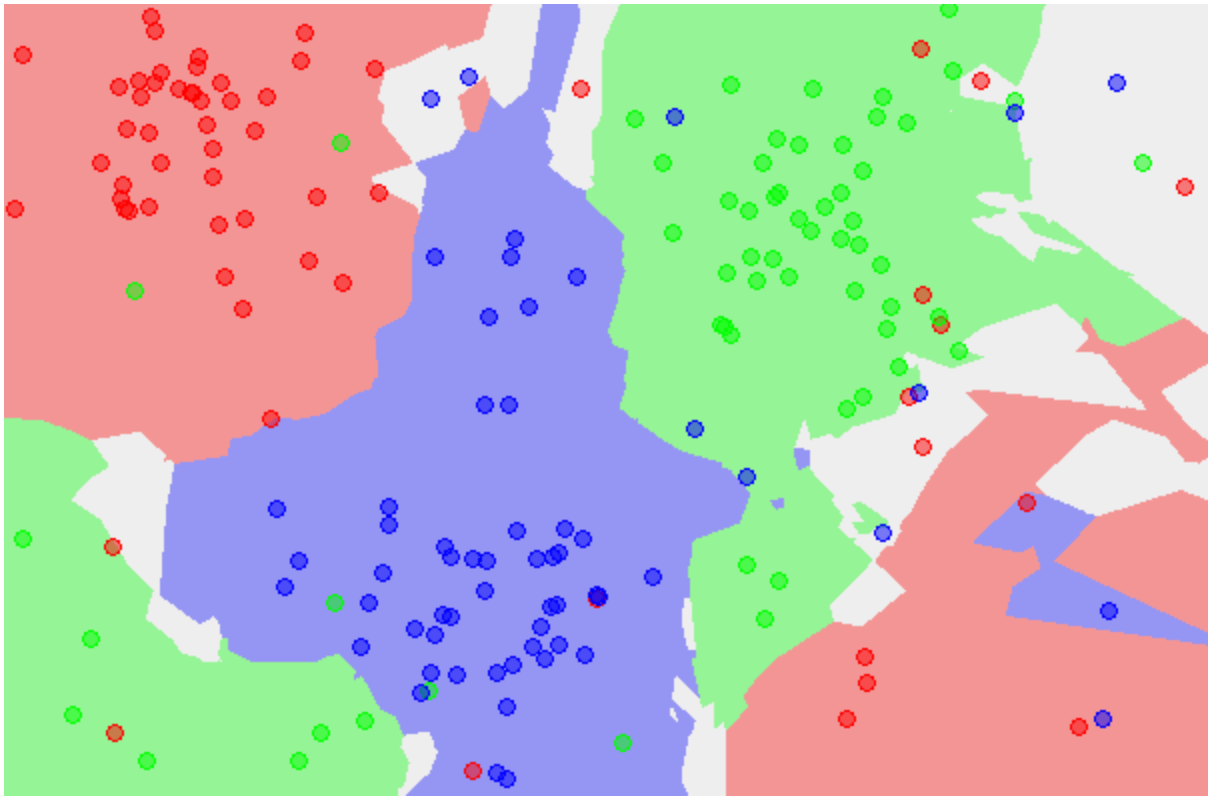
http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#/media/File:KnnClassification.svg



http://commons.wikimedia.org/wiki/File:Data3classes.png



http://commons.wikimedia.org/wiki/File:Map1NN.png



http://commons.wikimedia.org/wiki/File:Map5NN.png

k-Nearest Neighbors

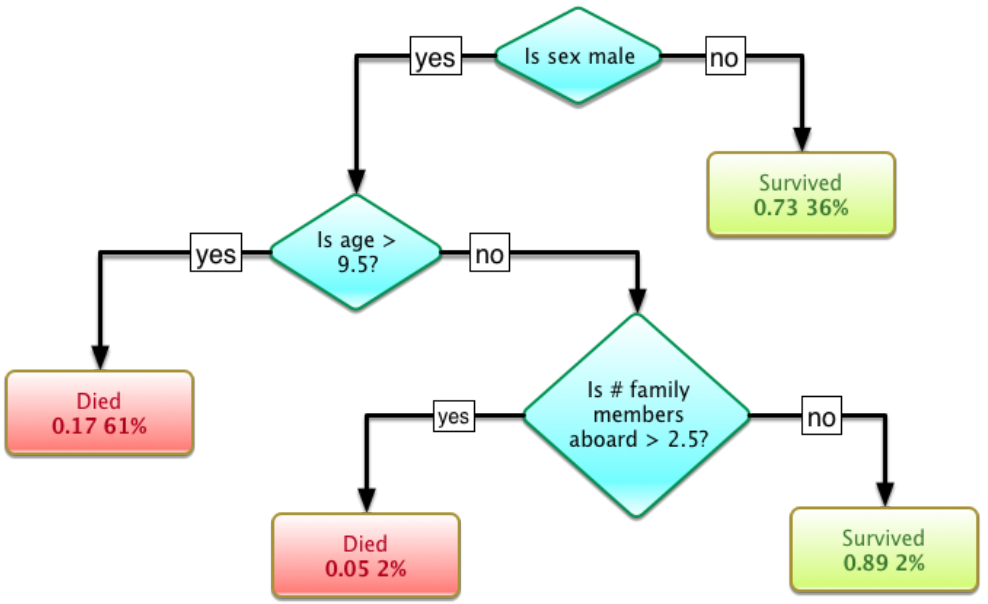
Pros	Cons
Simple and effective	Does not produce a model, but does produce a function
Makes no assumptions about the data distribution	Efficacy affected by choice of k
Fast training phase	Slow classification phase

Decision Trees

Decision Trees

- Tree-based classifier (like a bunch of *if-then* statements)
- Models the relationships between features and outputs
- Easy to explain to users
- Can be turned into external representation (i.e., a picture)
- Supports both classification and regression
- Builds a tree where each node divides the set of items based on the value of a feature
- The feature and feature value are chosen based upon which one "best" splits the set of items
- The "best" split can be determined by several approaches, two common ones are: *gini impurity* (default in SciKit-Learn) and *information gain*
    - Gini Impurity seeks to maximize the homogeneity of the subnodes
    - Information Gain seeks to minimize the entropy of the subnodes

Decision Tree: Titanic Dataset



Decision Tree

Pros	Cons
Useful classifier for most problems	Can be biased toward feature splits with several levels
Automated learning process	Easy to misfit the model
Supports numeric, nominal and missing data	Small changes in the training data can have been impact on decision logic
Works with large and small data sets	Large trees may be hard to interpret
Easily interpreted and efficient	

Demo: Decision Trees

- there are screenshots in this presentation to maintain continuity, but let's open the notebook named **`Demo - Decision Trees.ipynb`** and go through it together
- when done, click here to skip screen shots

```
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier
```

```
iris = load_iris()
X = iris.data[:, 2:]
y = iris.target
```

```
tree_clf = DecisionTreeClassifier(max_depth=2)
tree_clf.fit(X, y)
```

```
from sklearn.tree import export_graphviz
export_graphviz(tree_clf, out_file="iris_tree.dot",
                feature_names=iris.feature_names[2:],
                class_names=iris.target_names,
                rounded=True,
                filled=True)
```

```
# dot -Tpng iris_tree.dot -o iris_tree.png
```

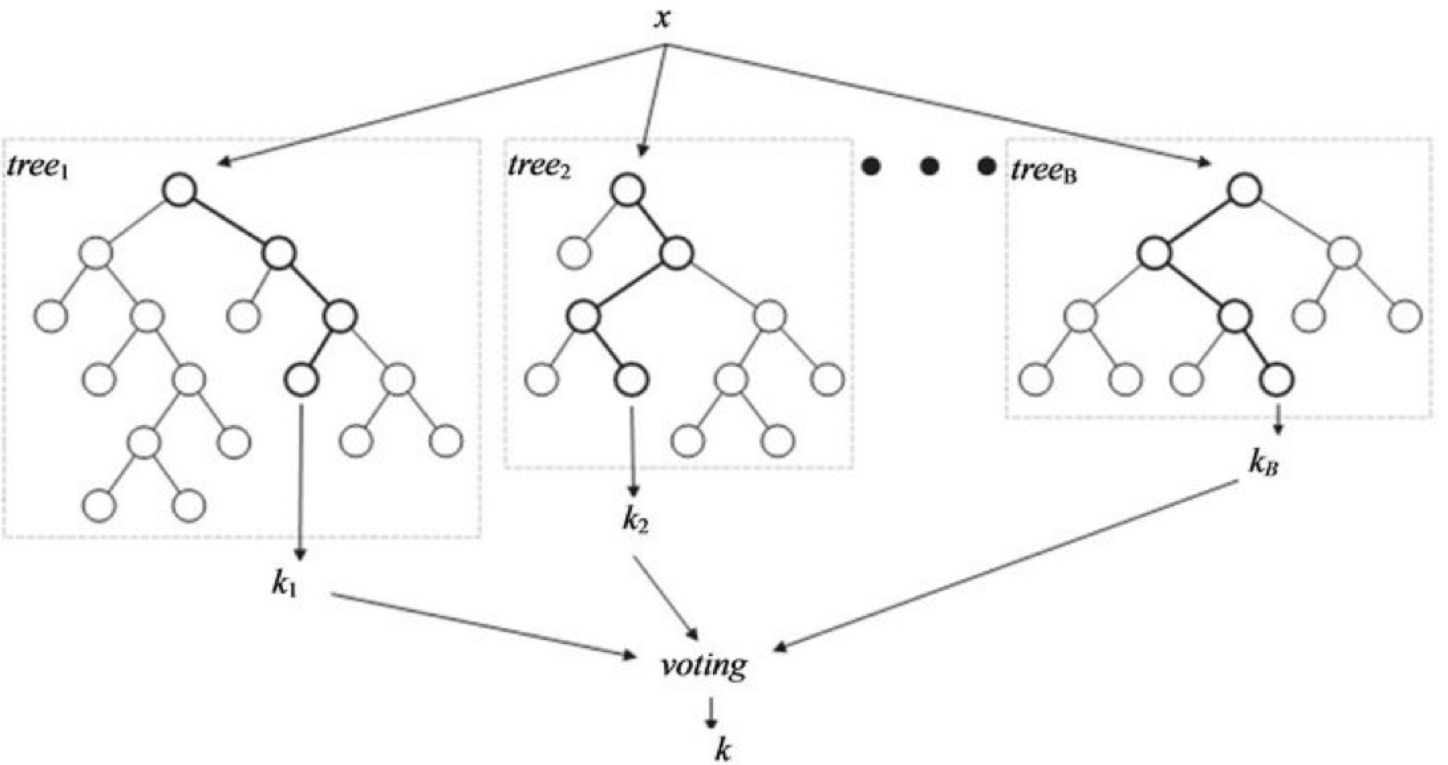
```
tree_clf.predict_proba([[5, 1.5]])
```

```
tree_clf.predict([[5, 1.5]])
```

Exercise: Decision Trees

(open the notebook named **`Exercise 7 - Decision Trees.ipynb`**)

Random Forests



http://file.scirp.org/Html/6-9101686_31887.htm

Demo: Random Forests

- there are screenshots in this presentation to maintain continuity, but let's open the notebook named **`ML - Decision Tree 2.ipynb`** and go through it together

Random Forests

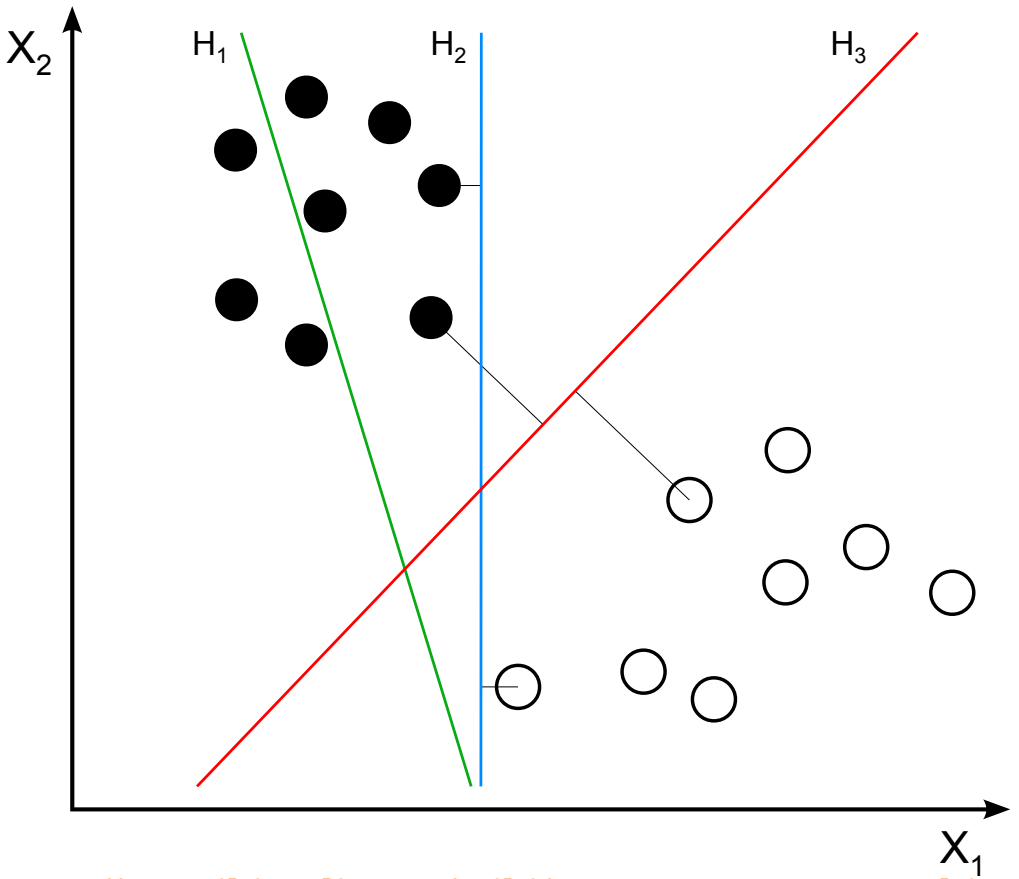
Pros	Cons
Supports both classification and regression.	Computationally expensive.
Averages out potential bias from Decision Trees.	Less interpretable than Decision Tree.

Support Vector Machines

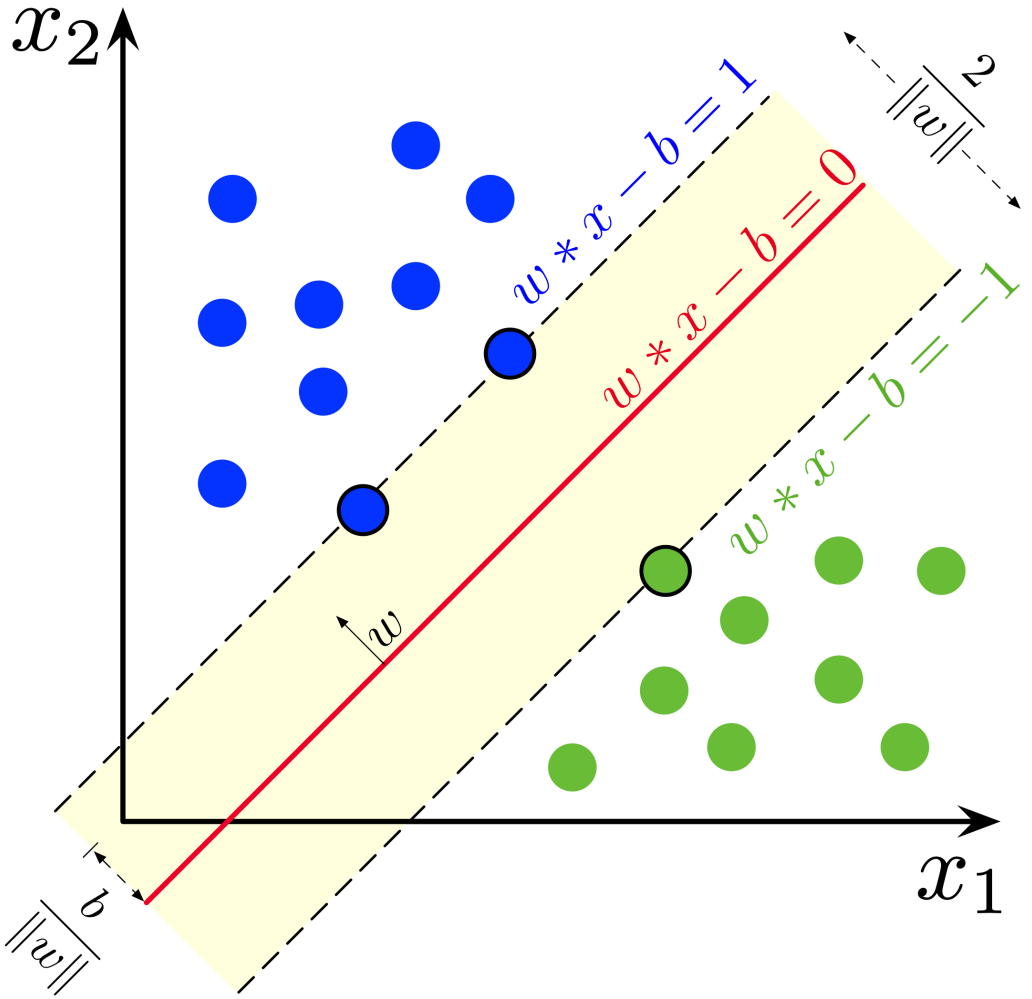
Support Vector Machines

- given a set of training examples, each marked as belonging to one of two categories, a *Support Vector Machine* builds a model that assigns new examples to one category or the other
- (works with more than two classes, but for simplicity, let's consider only two classes for the time being)
- the idea is to find the best separation between items in different classes
- also handles regression, but classification is more common

Which vector (hyperplane) separates the classes with the maximum margin?

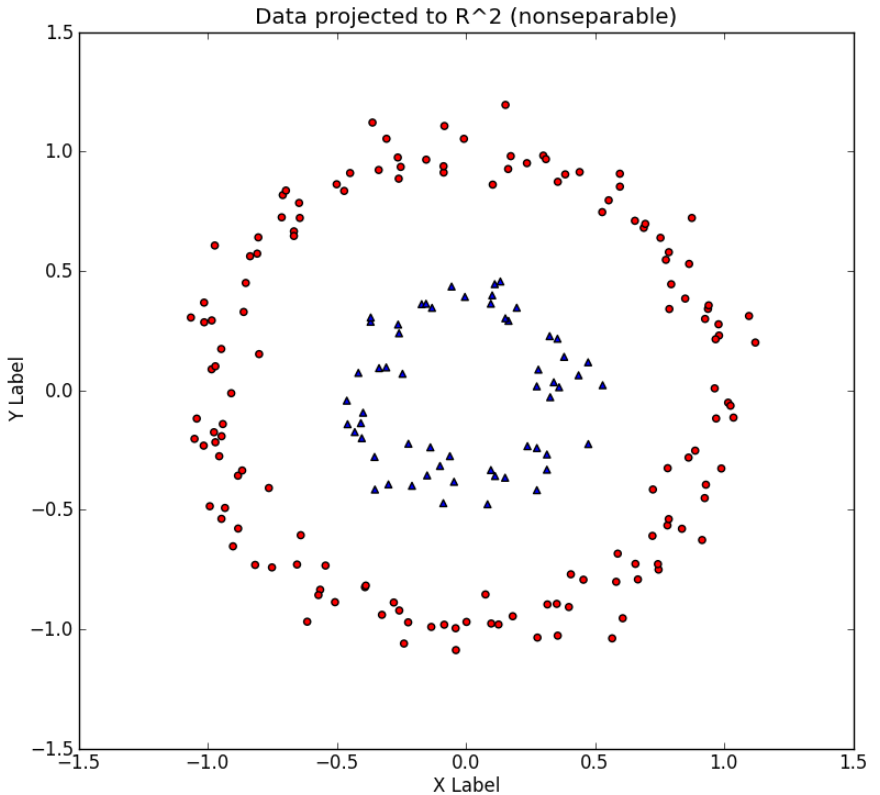


https://en.wikipedia.org/wiki/Support_vector_machine

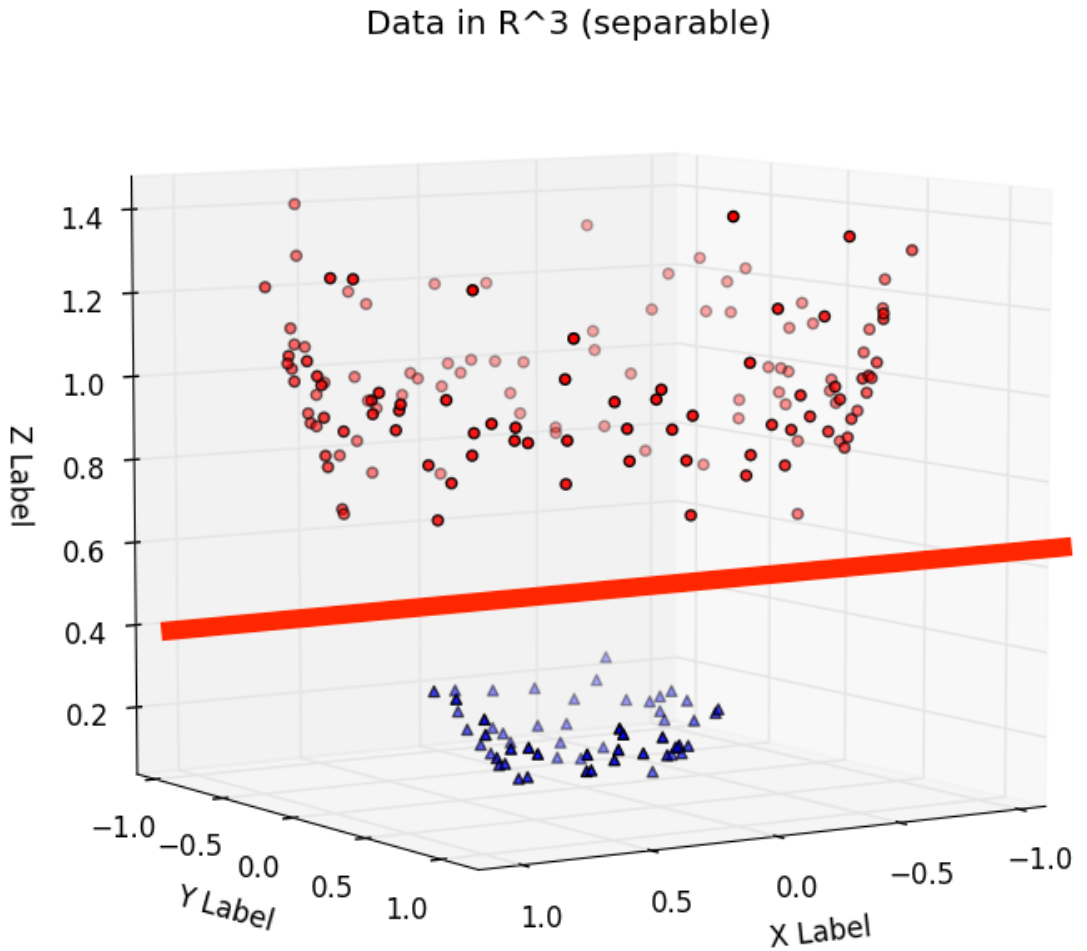


https://en.wikipedia.org/wiki/Support_vector_machine

What if the data are not linearly separable?



- we can use a "kernel trick" to transform the data into a higher-dimensional space in which the data *are* separable...



visualization

Demo: Support Vector Machines

- let's open the notebook named **`ML - Support Vector Machine.ipynb`** and go through it

Support Vector Machines

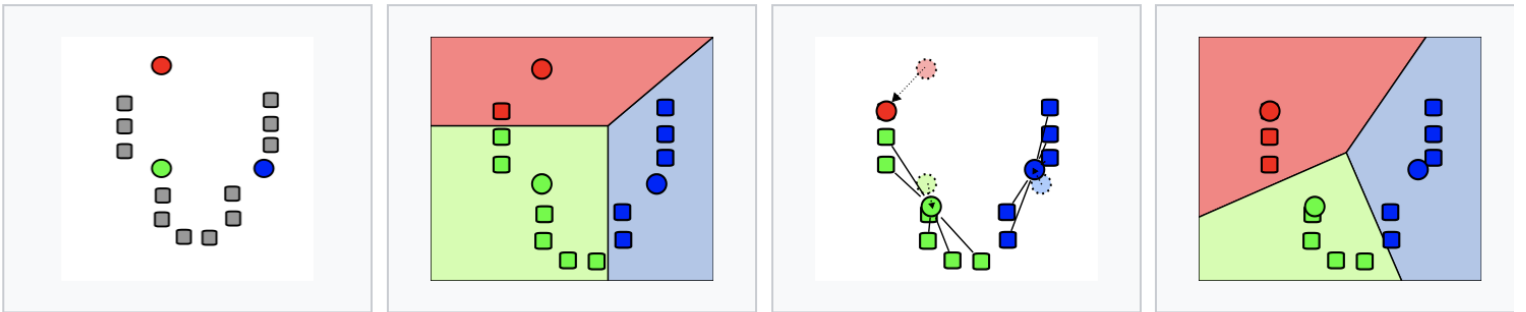
Pros	Cons
Similiar performance to logistic regression on linear boundaries.	Kernel choice can make it susceptible to overfitting.
Can handle non-linear boundaries.	Hard to interpret in high dimensions.
Handles high dimensional data	

Unsupervised Learning

Given only input variables x, find some underlying structure.

- Clustering
- Association rules

k-Means Clustering



- Select k-means randomly
- Associate every cluster with a mean
- The centroid of each cluster becomes the new mean
- Repeat until convergence

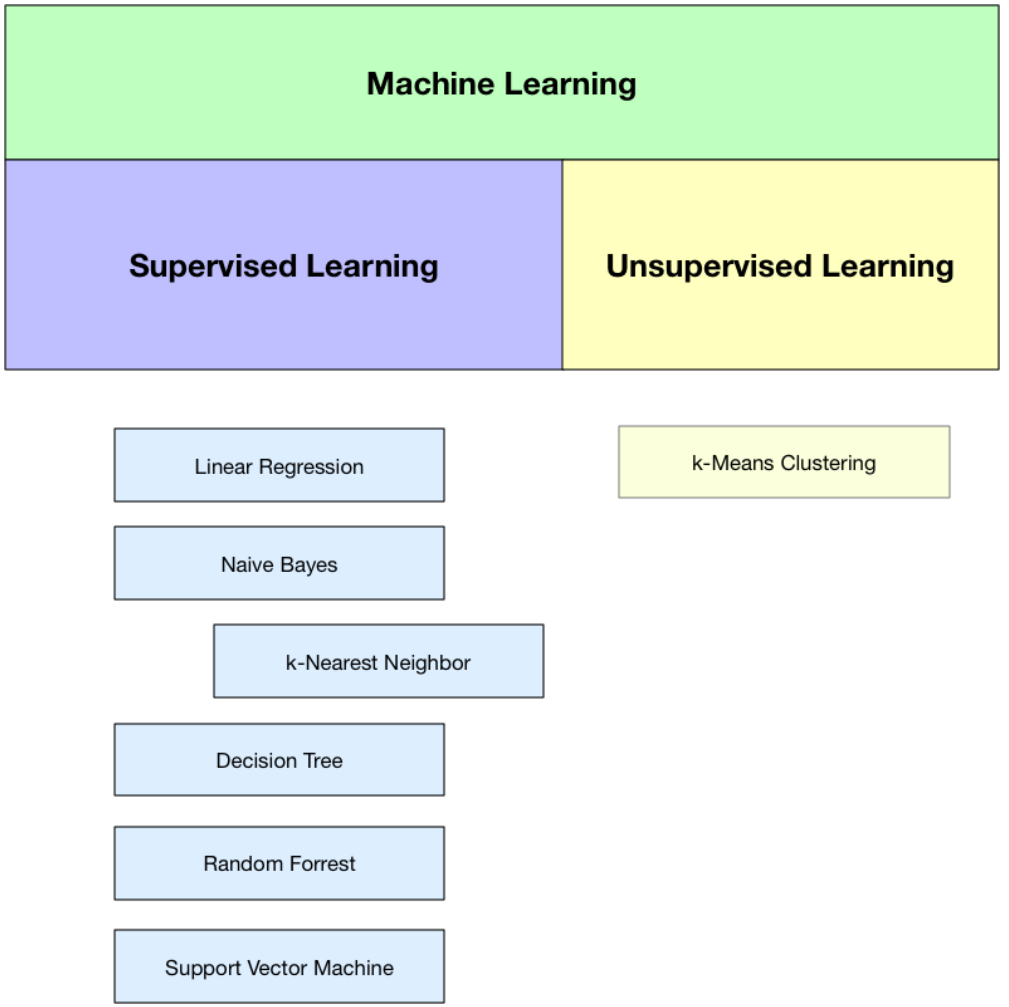
https://en.wikipedia.org/wiki/K-means_clustering

k-Means Clustering

Pros	Cons
Linear complexity O(n)	Random initialization may impact repeatability
Simple Euclidean distance calculation	Clusters don't really mean anything

Demo: k-Means Clustering

- let's open the notebook named **`ML - k-Means Clustering.ipynb`** and go through it


Machine Learning
Supervised Learning
Unsupervised Learning
Linear Regression
Naive Bayes
k-Nearest Neighbor
Decision Tree
Random Forrest
Support Vector Machine
k-Means Clustering

Pod Risk Assessment Demo

Pod Risk Assessment

- let's explore a real-world example of *salesforce* already benefiting from applying ML activities to its data
- but first, we'd like to acknowledge the help of additional sales*f*orce employees:
  - Lauren Valdivia
  - Kyle Gilson
  - Seung Soo Park
- some assumptions
  - your customers trust your infrastructure (which of course hosts their data) is *reliable*
  - in addition, they trust that you proactively assess the *health* of your infrastructure
- Questions:
  - What do we gain from being proactive?
  - What's the leading indicator that there's going to be a problem?
- Conclusion:
  - ...we want to make this whole problem BORING

Why is this so difficult?

- your system is multi-tier
  - app, db, search, storage
- it's also multi-tenant
  - most of your customers are sharing infrastructure with other companies, some of who may grow really fast (e.g., Uber)

Definitions

Org (Organization) – A deployment of Salesforce with a defined set of licensed users. Your organization includes all data and metadata, standard and custom objects, applications and security access. Your org is separate from all other orgs.

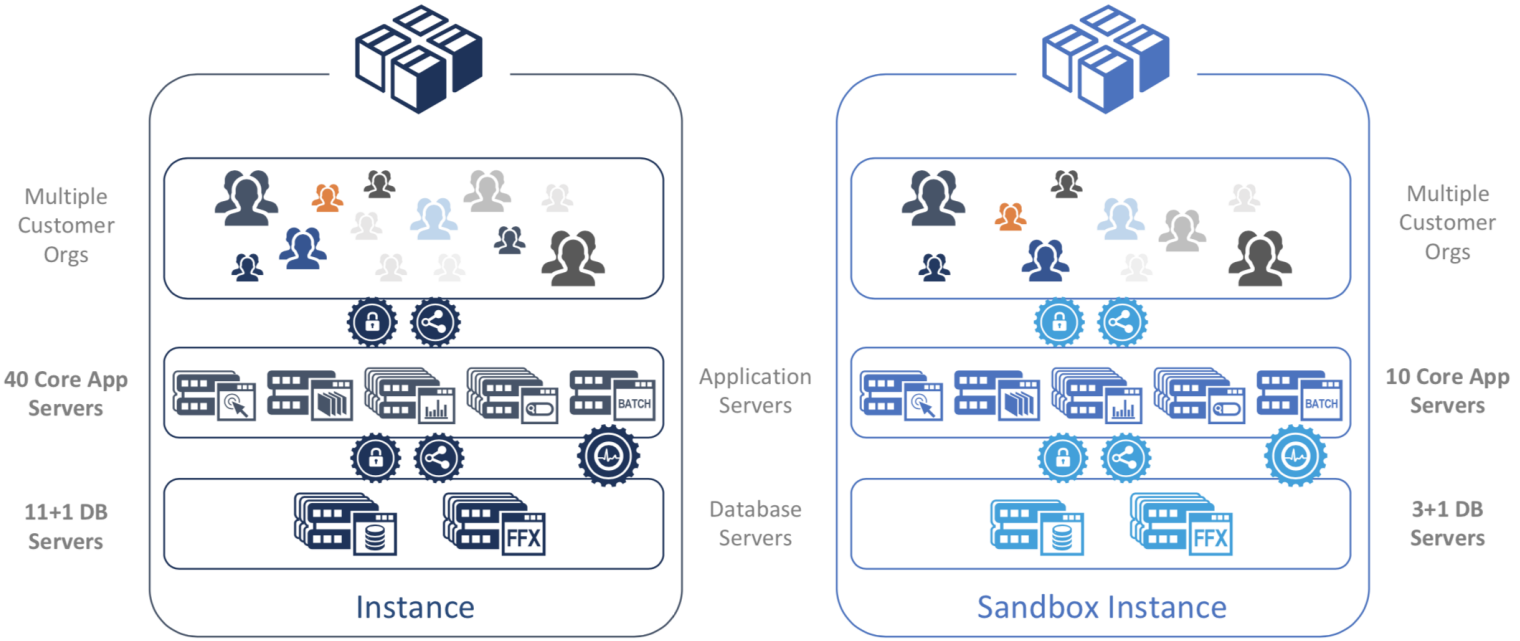
Core Instance – An instance is the collection of systems that provide service for a customer's Org. Salesforce's innovative multitenancy allows each Instance to support thousands of Orgs. There are two Instance types – Core (Production), Sandbox

Sandbox Instance – A sandbox allows you to create multiple copies of your Org in separate environments for development, testing, or training purposes without compromising the data and applications in your Salesforce product Org.

Instance Group – An Instance Group is a group of Instances that are supported by the same set of shared services (network and data) to provide additional service isolation within a single Data Center. Instance Groups can contain varying numbers of Core and Sandbox Instances.

Data Center – A Data Center is the physical location that houses one or more Instance Groups. It also includes the redundant shared services (power, cooling, networking) required to operate the Instance Groups and the Instances within those groups.

Core and Sandbox Instances

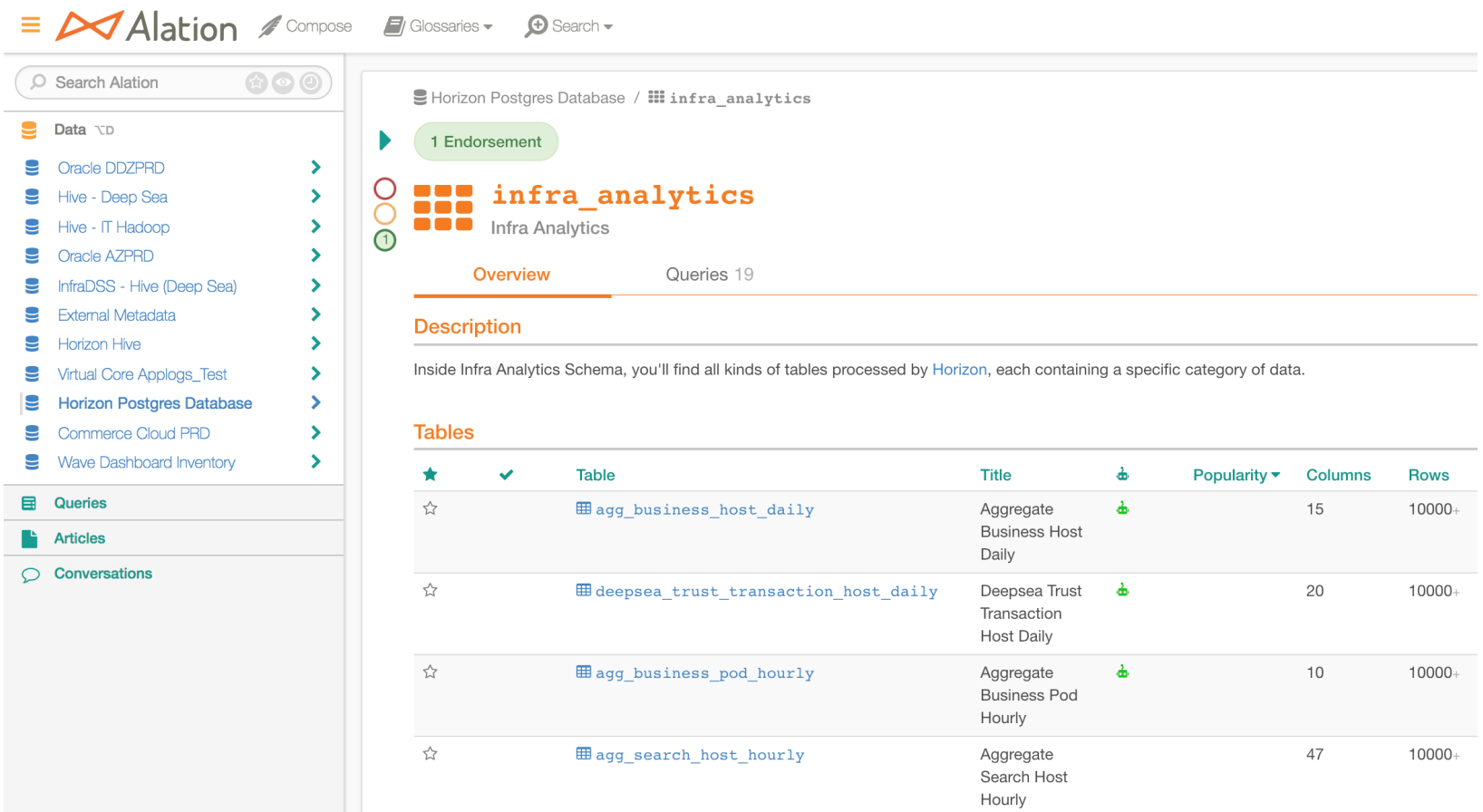


- A LOT of moving parts!
- Understanding which bits of data will be meaningful isn't easy

Where is the data?

- managed by the Infra Analytics team
- currently used to support use cases around capacity planning, performance engineering, and customer workload analyses
- the data catalog contains curated core app log data and core system data
- the team will continue populating this data catalog with discoverable sales*f*orce infrastructure data including but not limited to Marketing Cloud data, Commerce Cloud data, and so forth
- sales*f*orce uses **`Alation`**, a data cataloging tool, to store the data
    - all employees have access to this tool via Horizon (but your instructors do not)
    - see **`src/AlationHorizonDoc.pdf`** for details on how to access the data

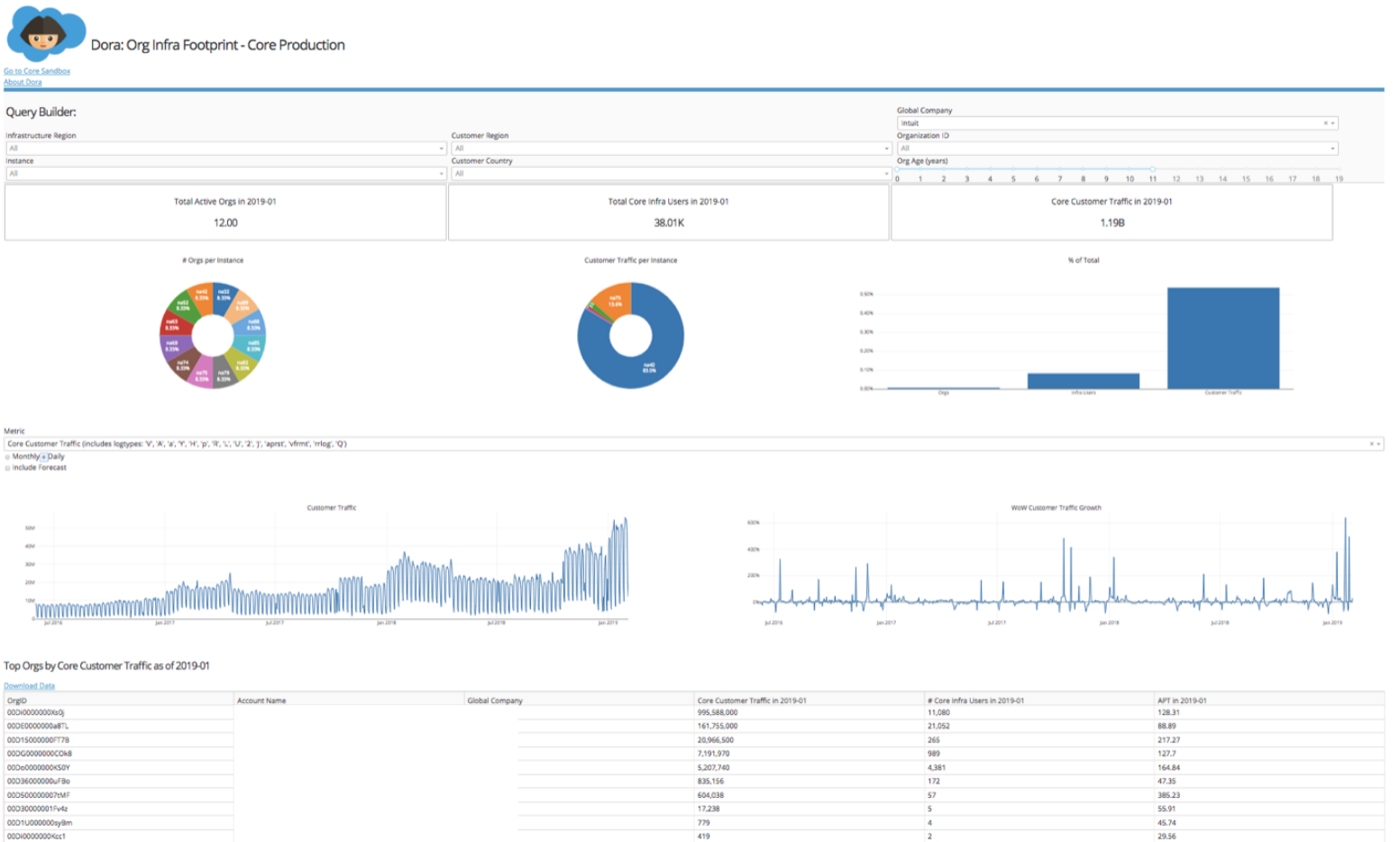
Alation


Alation
Compose
Glossaries
Search
Search Alation
Data
Oracle DDZPRD
Hive - Deep Sea
Hive - IT Hadoop
Oracle AZPRD
InfraDSS - Hive (Deep Sea)
External Metadata
Horizon Hive
Virtual Core Applogs_Test
Horizon Postgres Database
Commerce Cloud PRD
Wave Dashboard Inventory
Queries
Articles
Conversations
Horizon Postgres Database / infra_analytics
1 Endorsement
infra_analytics
Infra Analytics
Overview
Queries 19
Description
Inside Infra Analytics Schema, you'll find all kinds of tables processed by Horizon, each containing a specific category of data.
Tables
Table
Title
Popularity
Columns
Rows
agg_business_host_daily
Aggregate Business Host Daily
15
10000+
deepsea_trust_transaction_host_daily
Deepsea Trust Transaction Host Daily
20
10000+
agg_business_pod_hourly
Aggregate Business Pod Hourly
10
10000+
agg_search_host_hourly
Aggregate Search Host Hourly
47
10000+

Dora the Data Explorer

- data science team found that most of their time was consumed by providing curated datasets and repeating many simple data analyses (or at least very similar ones) for various stakeholders
- in an effort to reduce time-to-insight for themselves, they built Dora…

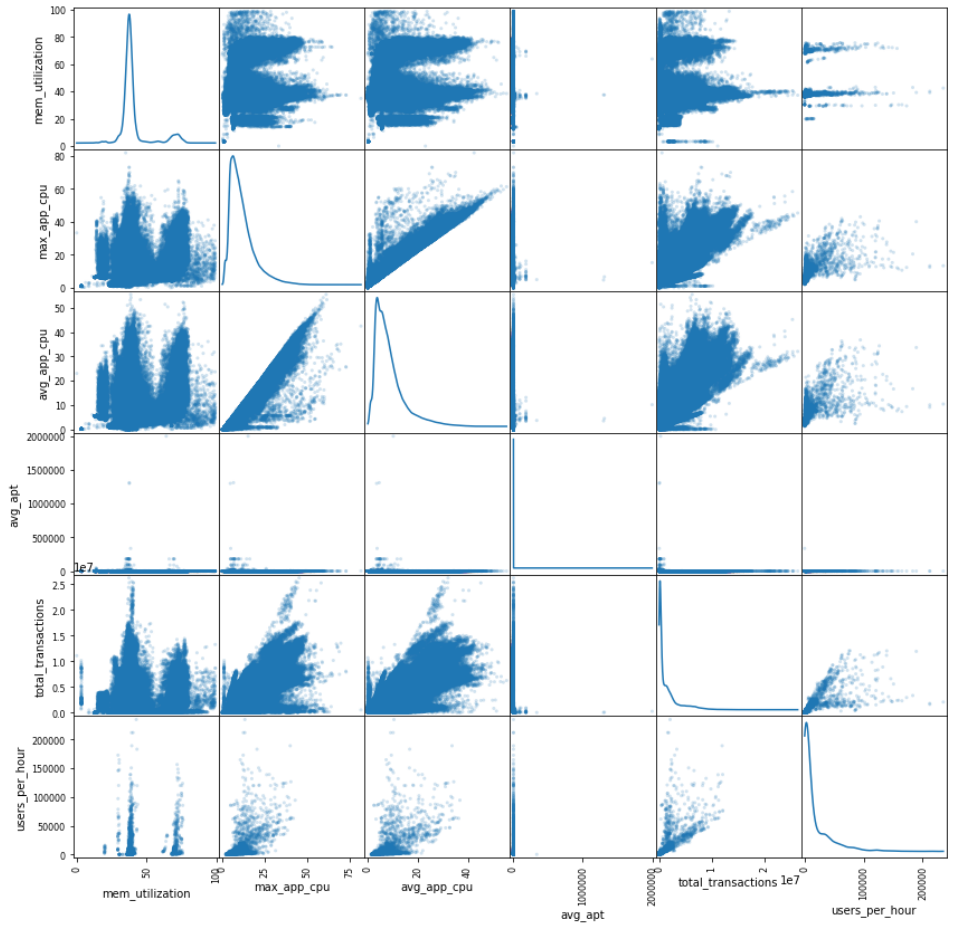
Dora the Data Explorer



How to solve this problem?

- goal was to distill these data down to a time to live
- they're not trying to explain the population
- this needs to be useful for the individual, i.e., a specific pod which is going to die
- the first thing to do is…

Try to find connections!

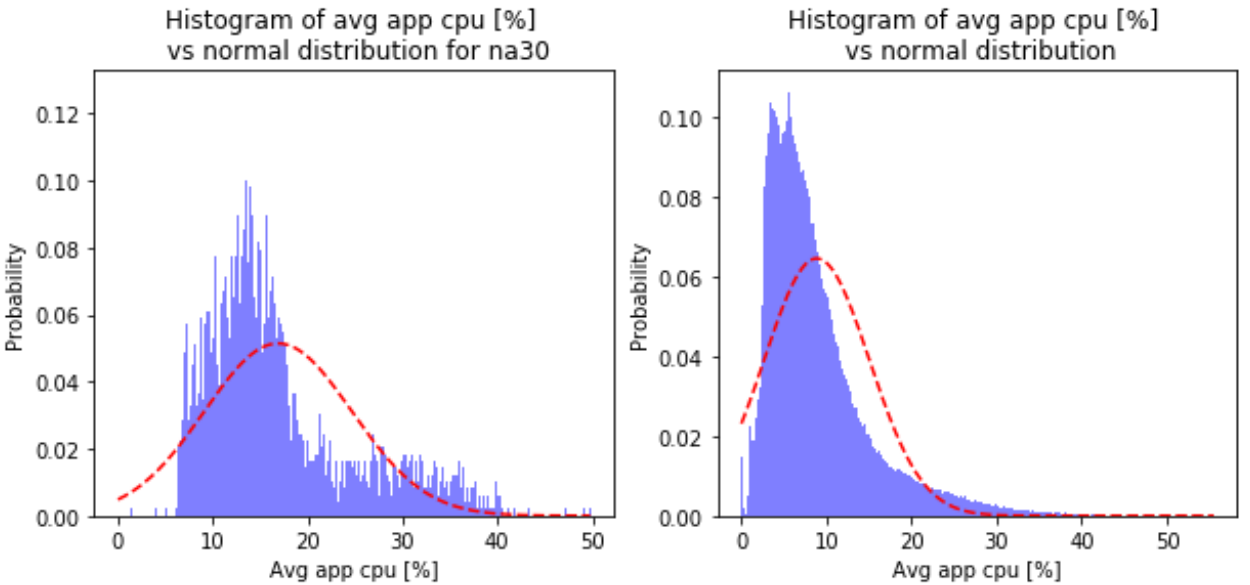


Which data ended up being relevant?

- App Tier: app cpu
- DB Tier: db cpu
- SAN Tier: db size and db file sequential read latency
- Combinations of these help determine the overall health of a pod
- Attempt to attach TTLs to system metrics (# transactions, average page time (APT))

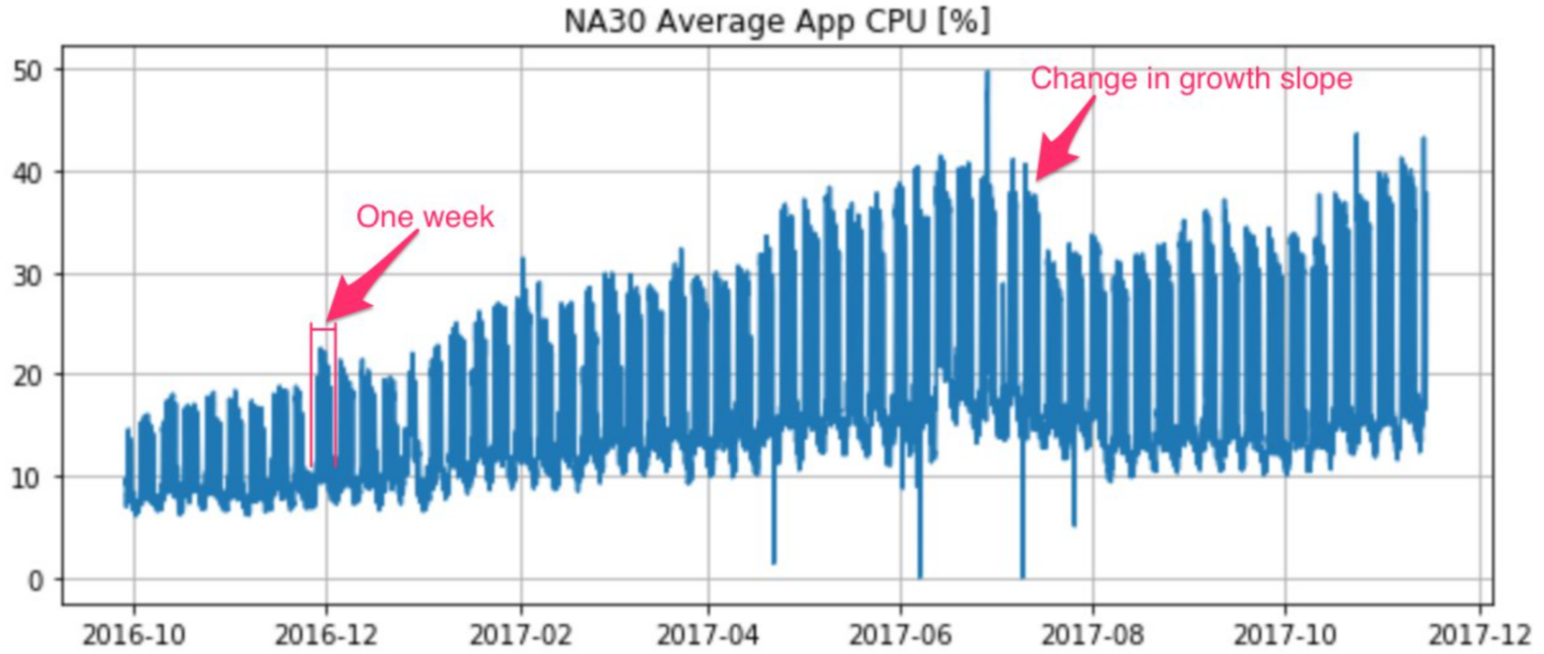
Now Apply Some Models...

- but it turns out that some of these data do not have normal distributions...



- ...so linear regression won't work here

Seasonality of data



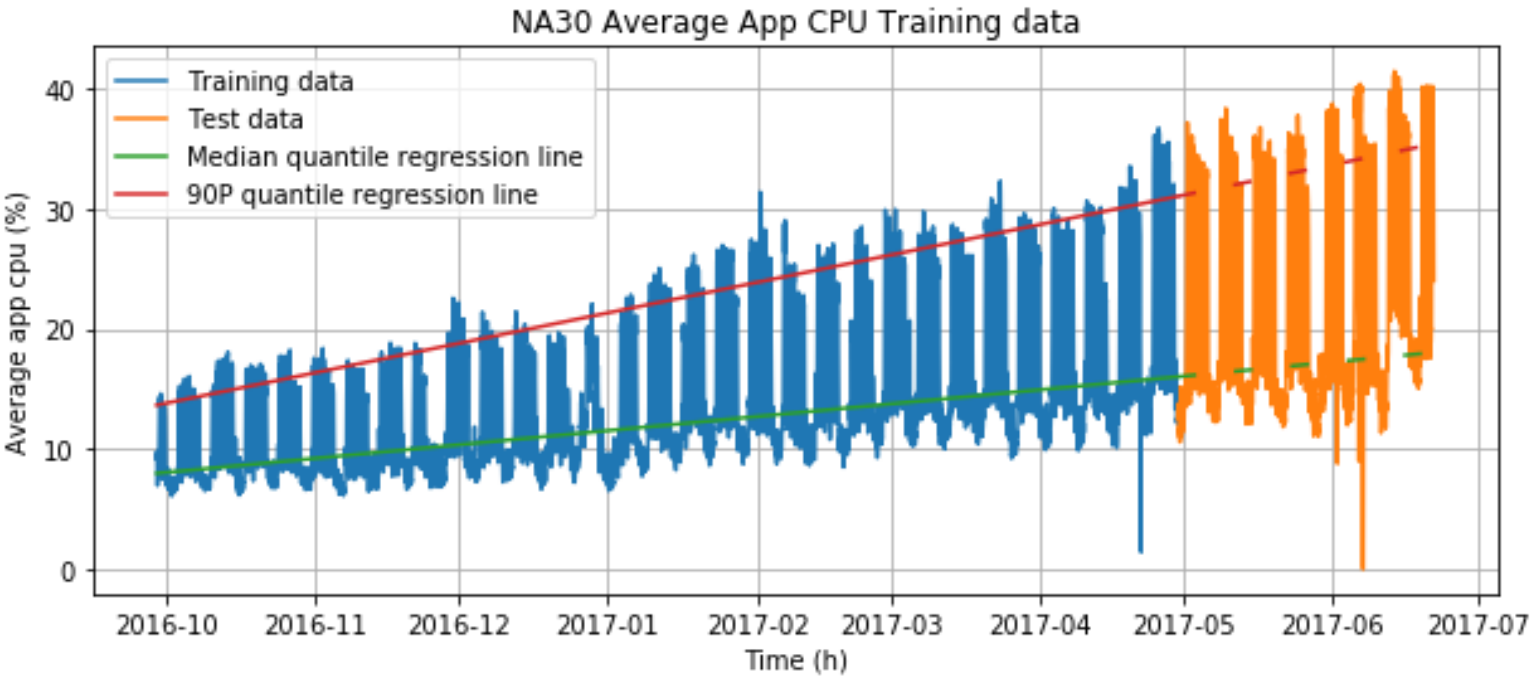
- Why not just average this out?

Other Problems

```
In [6]: data['max_db_cpu_user'].dropna().describe()
```

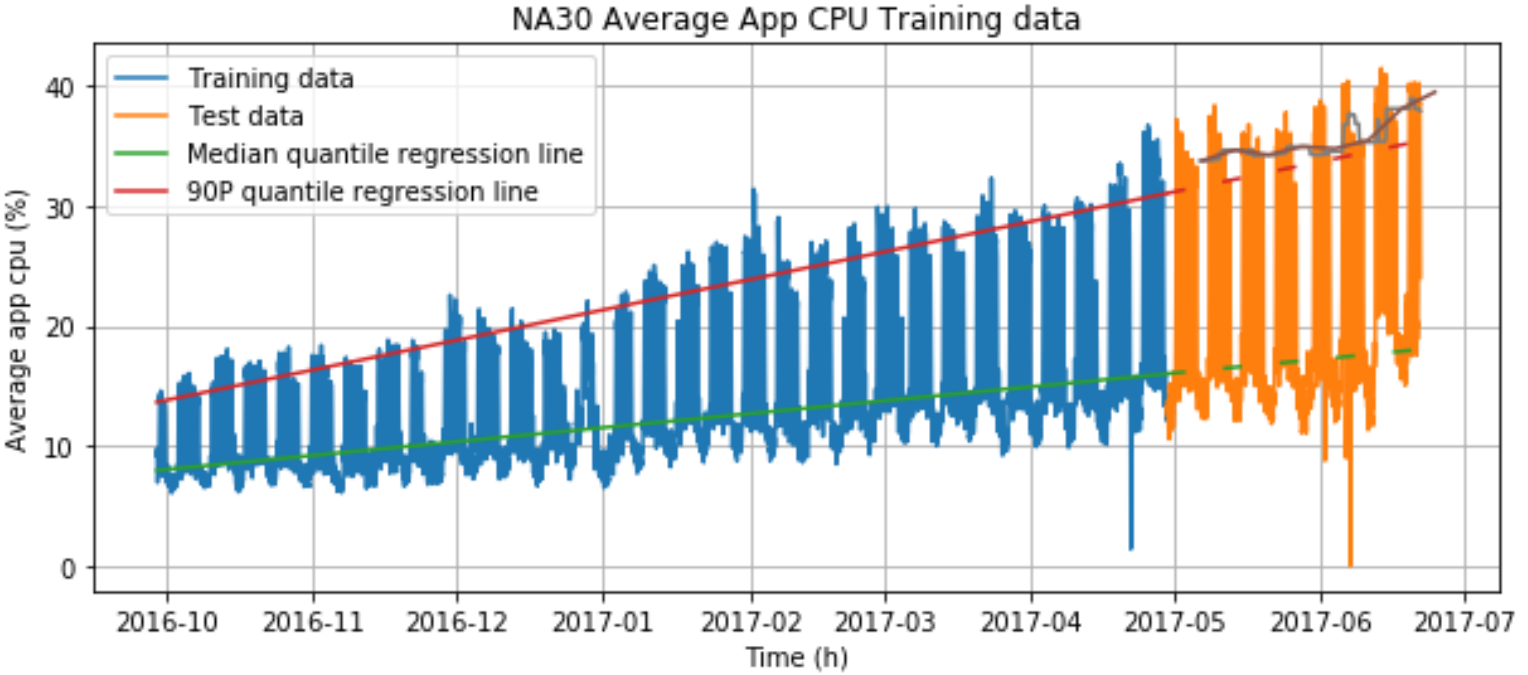
```
Out[6]: count     50913.000000
        mean       1504.906890
        std       25605.315296
        min           0.197738
        25%          12.048861
        50%          18.999514
        75%          29.516888
        max      759605.900000
        Name: max_db_cpu_user, dtype: float64
```

Quantile Regression



- 90th quantile regression vs. median quantile regression
- the 90th quantile ends up being a good predictor

Error Measurements



Complications

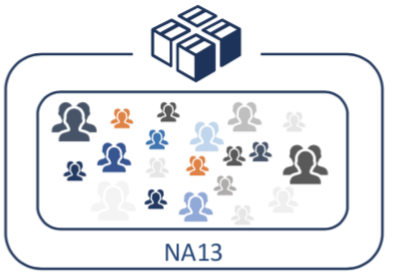
- What if there are some metrics which are missing from our model?
  - part of irreducible error–if you are not measuring for the things that are factoring into reality, you will always be off by a little bit
- **Example:** NA30 ran into capacity bottlenecks due to MQ (Message Queue) and emergency capacity additions had to be performed
  - TTL for NA30 was showing 3 months due to dB CPU
  - we were not showing a lower TTL due to not having an MQ metric as an input to our model
  - MQ is going to be added in the near future to Pod Risk Assessment

Complications (cont'd)

- What if we hit a new bottleneck?
  - Capacity can hit earlier than expected due to unplanned demand growth, changes in types of workload, code regressions, etc.
- **Example: Uber**
  - We had Uber in one pod and the transactions accelerated. Before we onboard a customer we have no idea how their demand is going to grow or what their usage patterns will be in terms of both their workload seasonality and the types of workload.

Mitigating Problem Pods–Instance Split

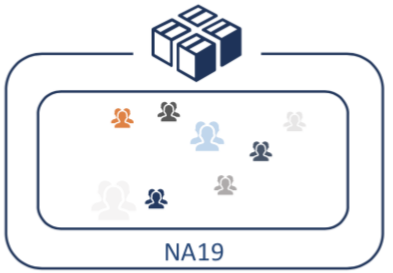
Before Split



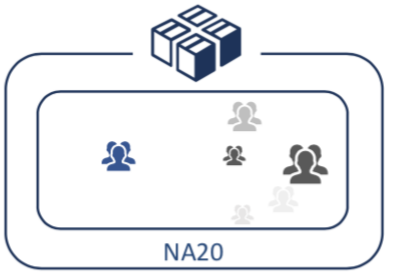
120M Transactions
3.9K Active Orgs
APT: 228ms
Peak App CPU: 20%
Peak DB CPU: 40%

After Split

50M Transactions
1.3K Active Orgs
APT: 149ms
Peak App CPU: 7%
Peak DB CPU: 15%

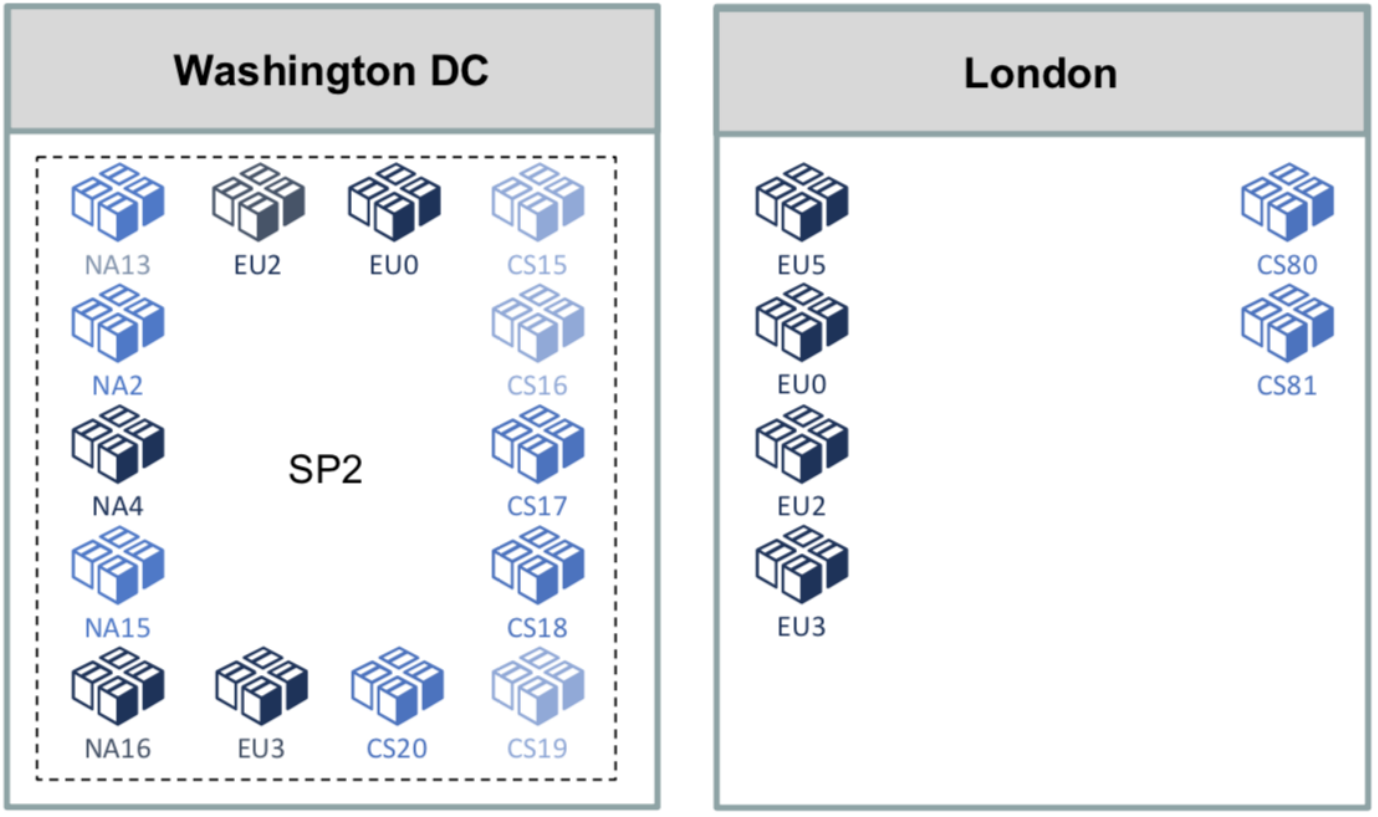


35M Transactions
1.2K Active Orgs
APT: 150ms
Peak App CPU: 4%
Peak DB CPU: 13%

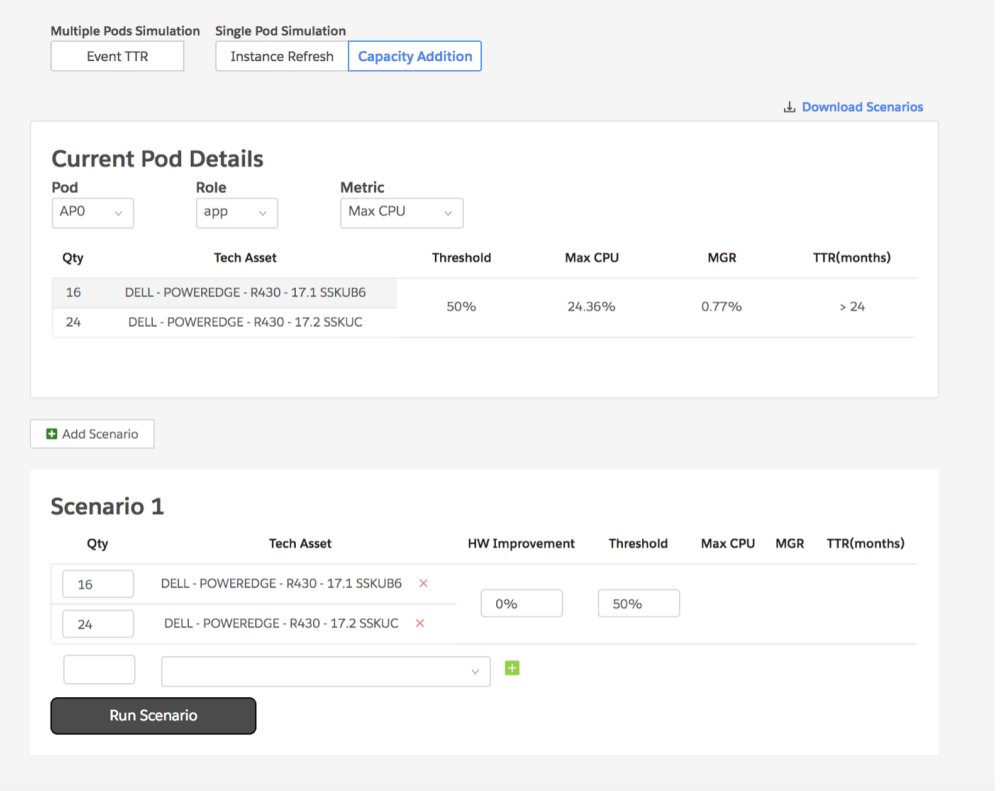


40M Transactions
1.4K Active Orgs
APT: 150ms
Peak App CPU: 5%
Peak DB CPU: 15%

Mitigating Problem Pods–Geo Migration



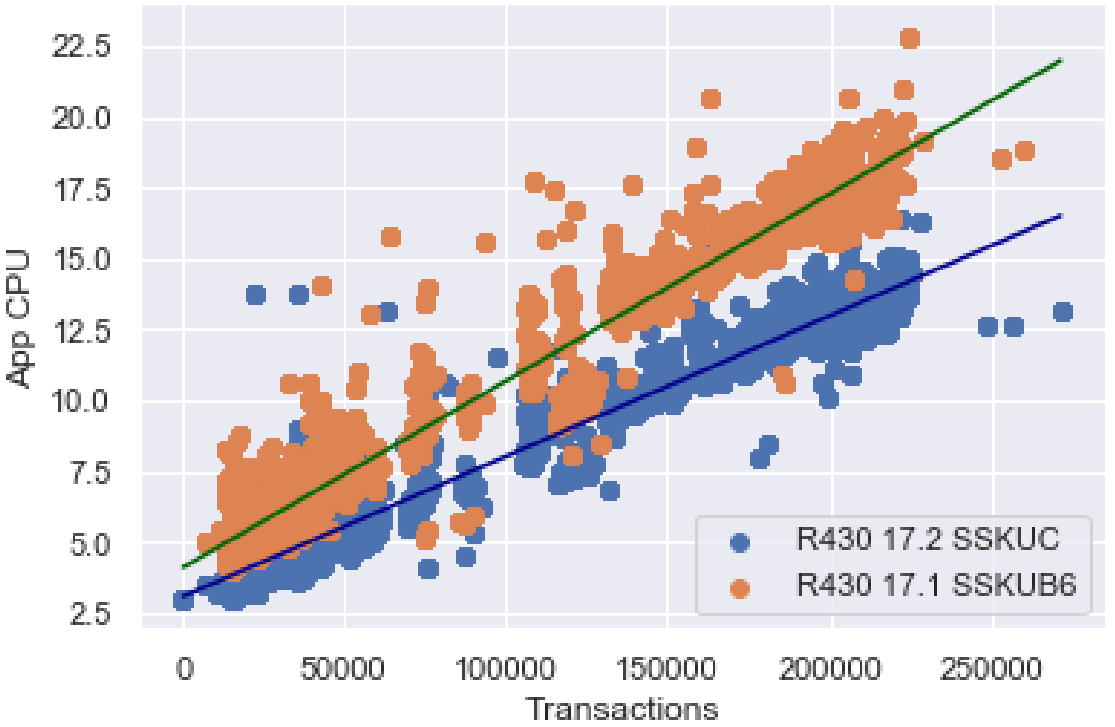
Mitigating Problem Pods–Add Capacity



- capacity simulator tool allows you to play with different scenarios using different/additional hardware, also can recommend remediations

Calculating Hardware Gain

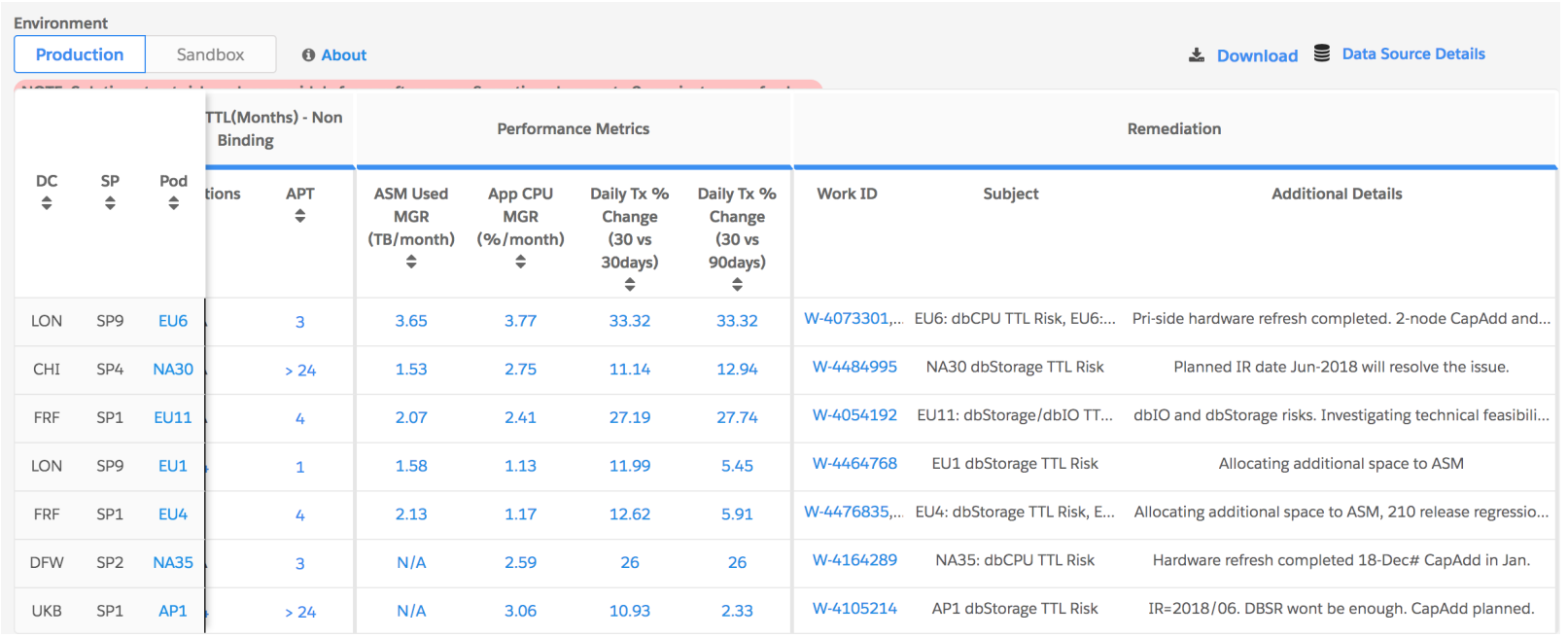
- Capacity Simulator tool can determine the hardware gain, i.e., how much capacity can be gained by switching to/adding new hardware?



- (see **`src/SF-CapacitySimulatorHWGain.ipynb`**)

Dashboard

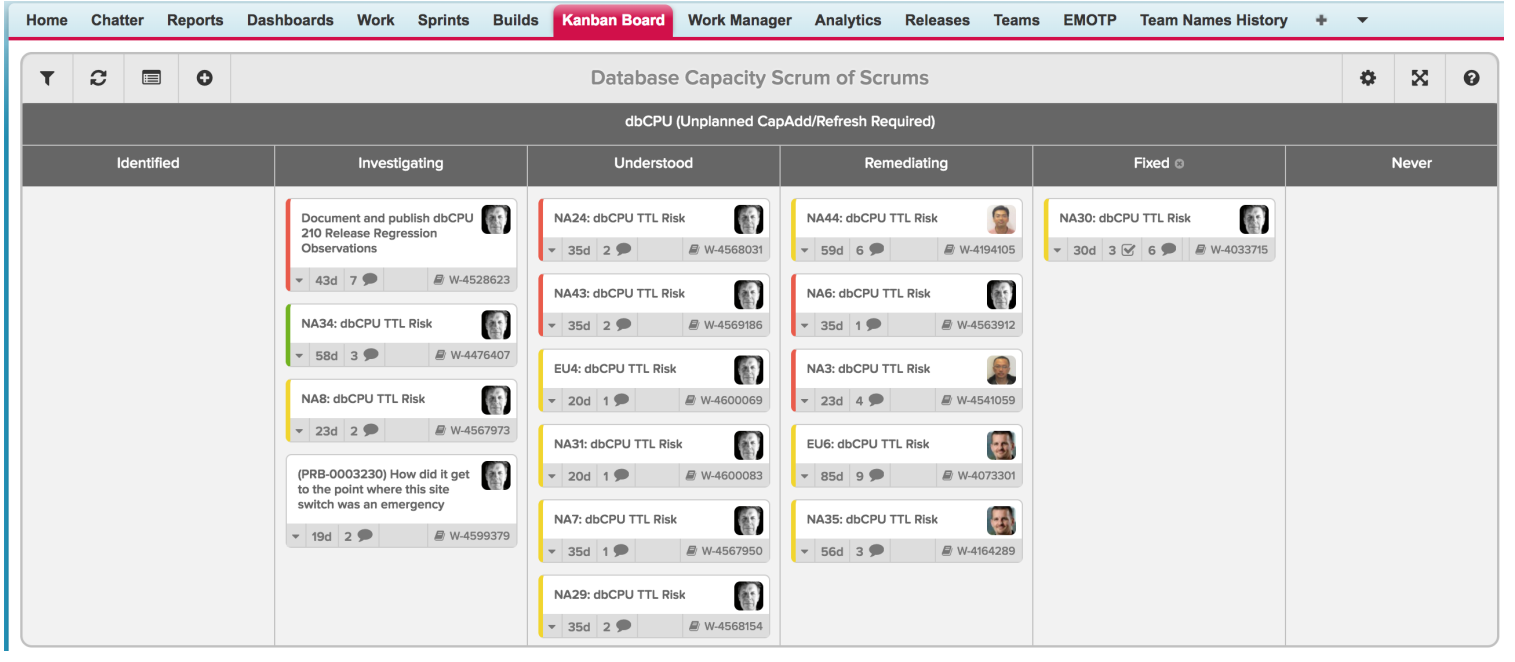
- there is a dashboard which shows the pods and instances and their TTL



DC	SP	Pod	TTL(Months) - Non Binding: tions	TTL(Months) - Non Binding: APT	Performance Metrics: ASM Used MGR (TB/month)	Performance Metrics: App CPU MGR (%/month)	Performance Metrics: Daily Tx % Change (30 vs 30days)	Performance Metrics: Daily Tx % Change (30 vs 90days)	Remediation: Work ID	Remediation: Subject	Remediation: Additional Details
LON	SP9	EU6		3	3.65	3.77	33.32	33.32	W-4073301,...	EU6: dbCPU TTL Risk, EU6:...	Pri-side hardware refresh completed. 2-node CapAdd and...
CHI	SP4	NA30		> 24	1.53	2.75	11.14	12.94	W-4484995	NA30 dbStorage TTL Risk	Planned IR date Jun-2018 will resolve the issue.
FRF	SP1	EU11		4	2.07	2.41	27.19	27.74	W-4054192	EU11: dbStorage/dbIO TT...	dbIO and dbStorage risks. Investigating technical feasibili...
LON	SP9	EU1		1	1.58	1.13	11.99	5.45	W-4464768	EU1 dbStorage TTL Risk	Allocating additional space to ASM
FRF	SP1	EU4		4	2.13	1.17	12.62	5.91	W-4476835,...	EU4: dbStorage TTL Risk, E...	Allocating additional space to ASM, 210 release regressio...
DFW	SP2	NA35		3	N/A	2.59	26	26	W-4164289	NA35: dbCPU TTL Risk	Hardware refresh completed 18-Dec# CapAdd in Jan.
UKB	SP1	AP1		> 24	N/A	3.06	10.93	2.33	W-4105214	AP1 dbStorage TTL Risk	IR=2018/06. DBSR wont be enough. CapAdd planned.

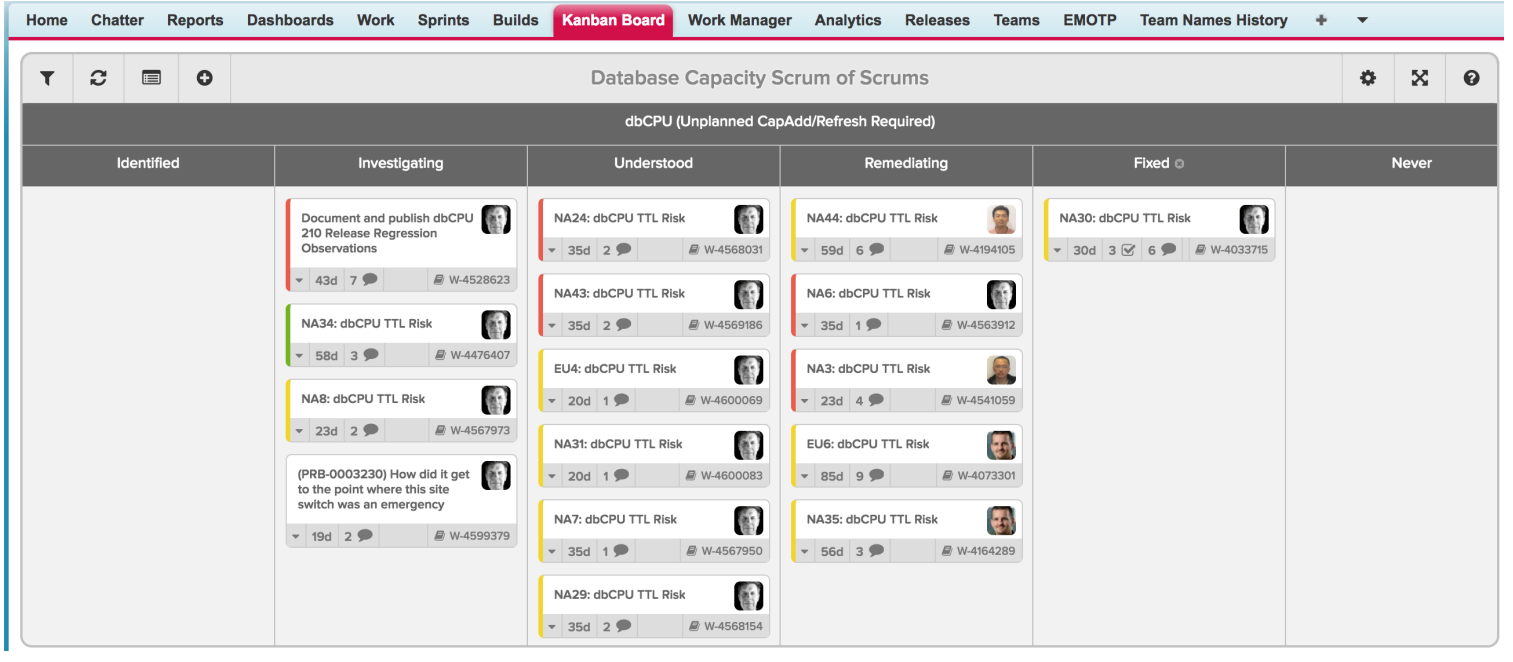
Ops Kanban Board

- this info has been added to Ops Kanban board, in order to make this problem boring, as was the goal



Ops Kanban Board

- this info has been added to Ops Kanban board, in order to make this problem boring, as was the goal



- engineers can see which pods need immediate attention, and which ones don't, and can prioritize their time accordingly