GENERAL ASSEMBLY

DATA SCIENCE
10 WEEK PART TIME COURSE

Week 6 – Bonus Round
Saturday 1st July 2017


1P 95100 HI 95100
00000
RYU
40
BONUS
STAGE

1. Review of all material so far
2. Text Analysis
3. Web Scraping
4. Prepping Data
5. Other Questions / General Discussion
6. Projects

DATA SCIENCE PART TIME COURSE

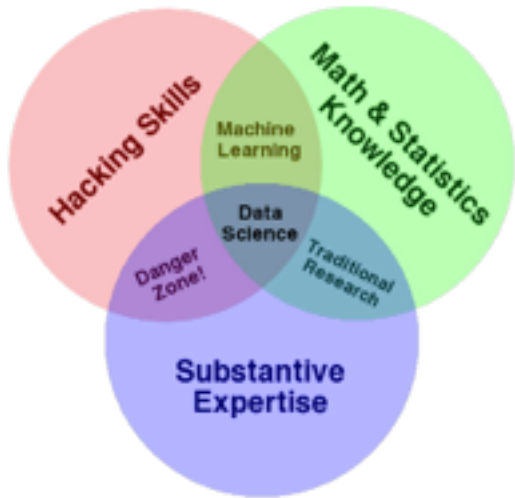
REVIEW

DATA SCIENCE PART TIME COURSE

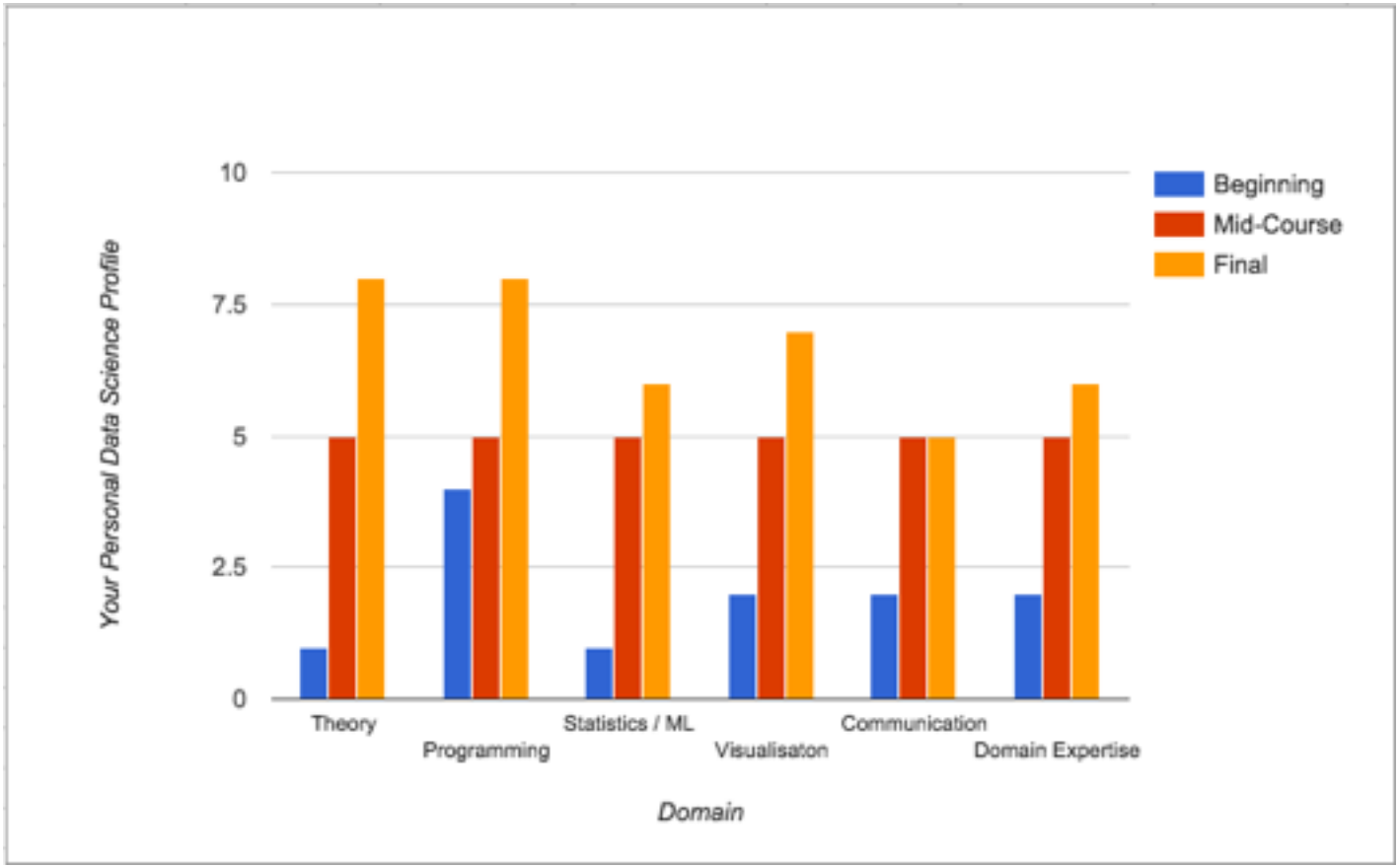
REVIEW - WEEK 1

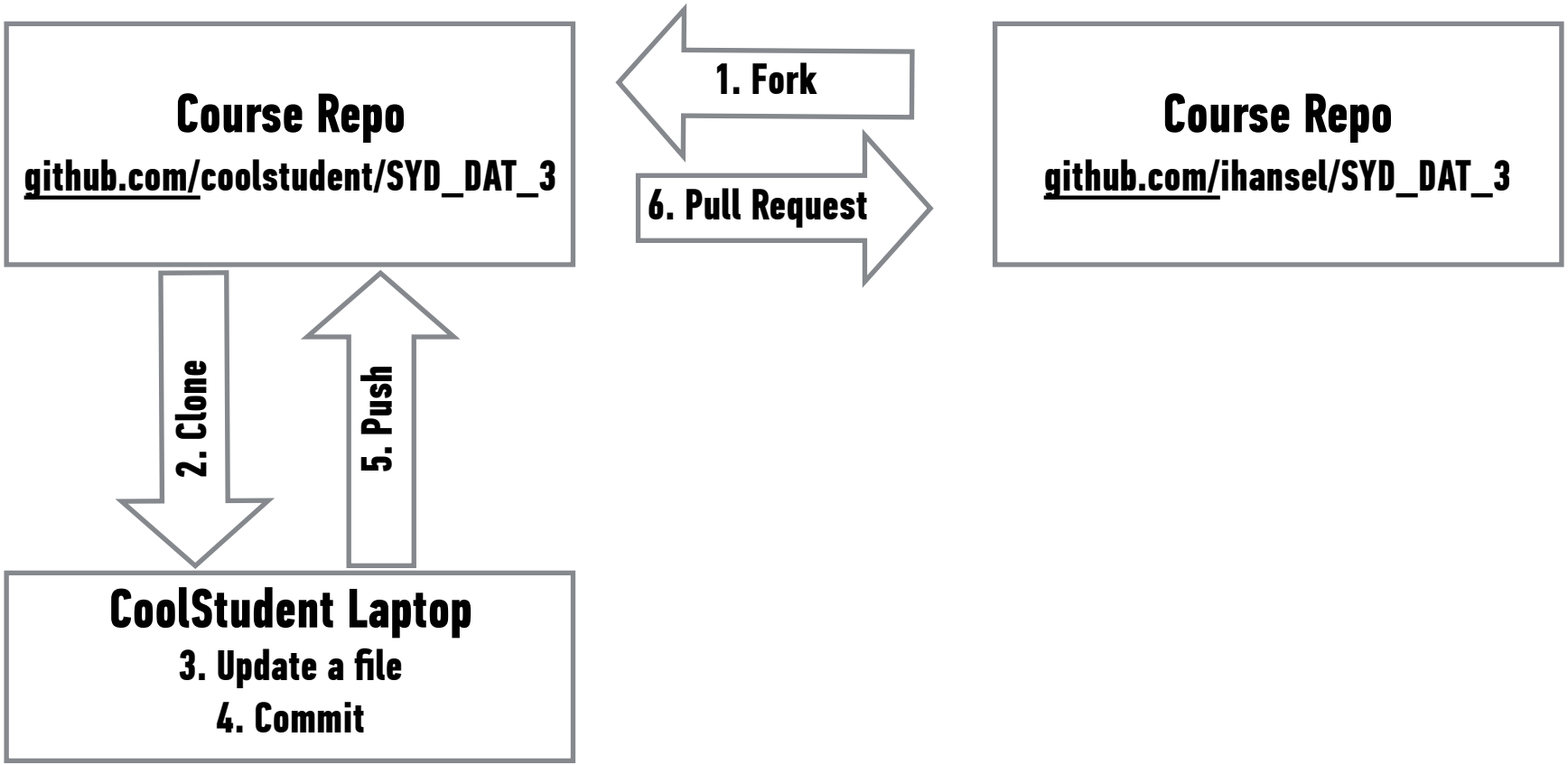
- Multidisciplinary Investigations
- Models and Methods for Data
- Computing with Data
- Pedagogy
- Tool Evaluation
- Theory

Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics
William S. Cleveland



Drew Conway's
Data Science Venn Diagram


10
7.5
5
2.5
0
Your Personal Data Science Profile
Beginning
Mid-Course
Final
Theory
Programming
Statistics / ML
Visualisaton
Communication
Domain Expertise
Domain


Course Repo
github.com/coolstudent/SYD_DAT_3
1. Fork
6. Pull Request
Course Repo
github.com/ihansel/SYD_DAT_3
2. Clone
5. Push
CoolStudent Laptop
3. Update a file
4. Commit

DATA SCIENCE PART TIME COURSE

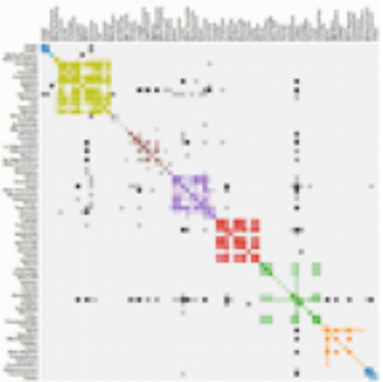
REVIEW - WEEK 2

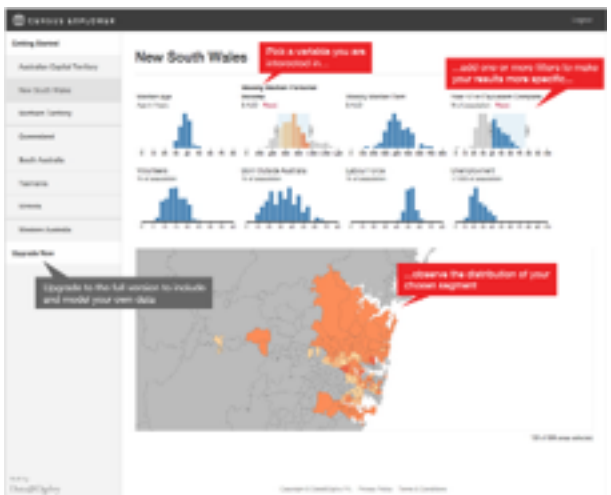
Reporting

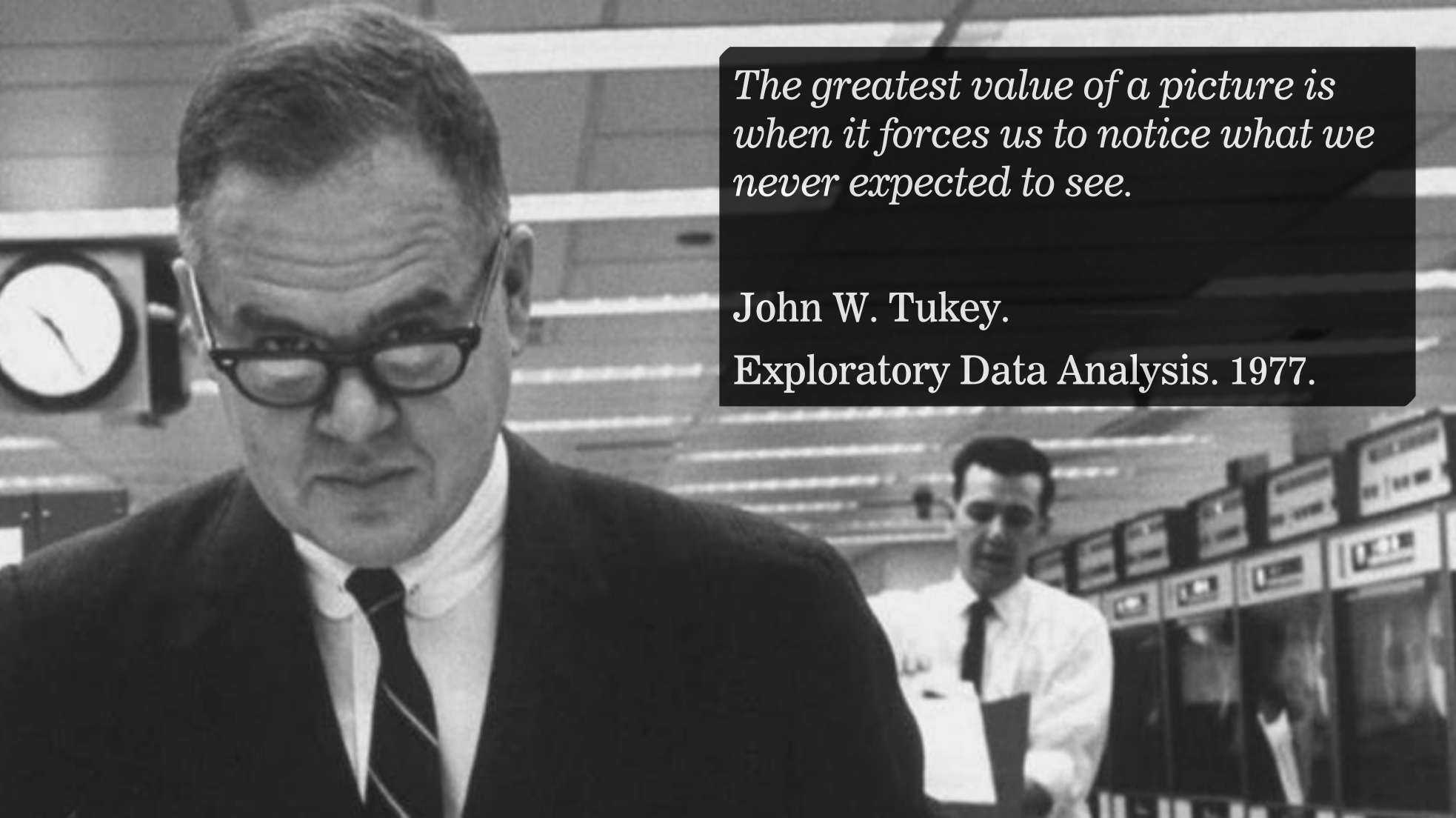
- Dashboards and Business Intelligence
- Know the questions you want answers to
- Can detect changes from the norm
- Good for taking a 30,000 foot view of the problem

Exploring

- Exploratory Data Analysis
- Combines multiple data sources for single view of a problem
- Technical analysis of data
- Combined with modelling allows for the discovery of new problems and solutions






The greatest value of a picture is when it forces us to notice what we never expected to see.
John W. Tukey.
Exploratory Data Analysis. 1977.

We want to predict some value, let's call it **y**, based on some observed data we have, let's call that **x**.

We will use statistical learning to estimate a function that approximates **y** based on the input, **x**.

y is also called; label, dependent variable, target

x is also called; predictor, independent variable, features

If the y variable is numeric then we have a regression problem - we are trying to predict a continuous number

If the y variable is a category (for example trying to predict a type of flower) the we have a classification problem - we are trying to classify what group that y belongs to.

We want to find some underlying structure or patterns in the data but in this case we don't have any labeled data.

So for example, if we have a large group of customers but would like to separate them into groups (or clusters) to better target them.

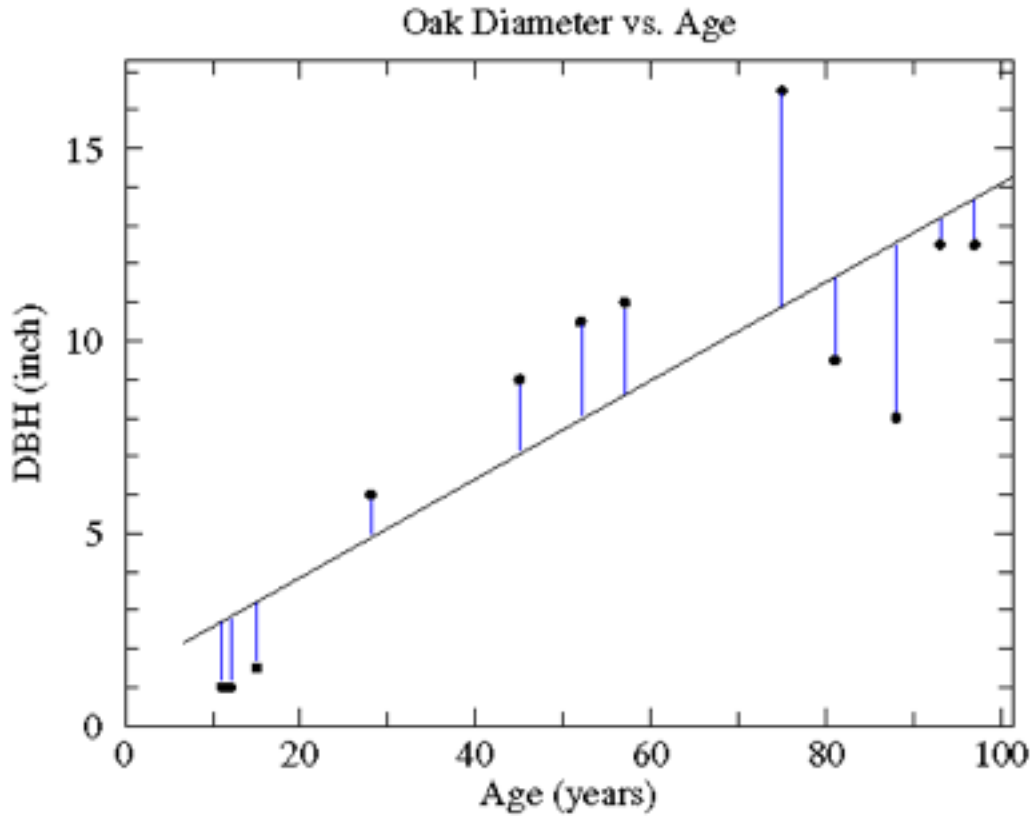
$$y = X\beta + \epsilon$$

- y = target variable
- X = input variable
- β = coefficients
- ϵ = error term

Note, one of our input variables can be 1 so we have an intercept parameter

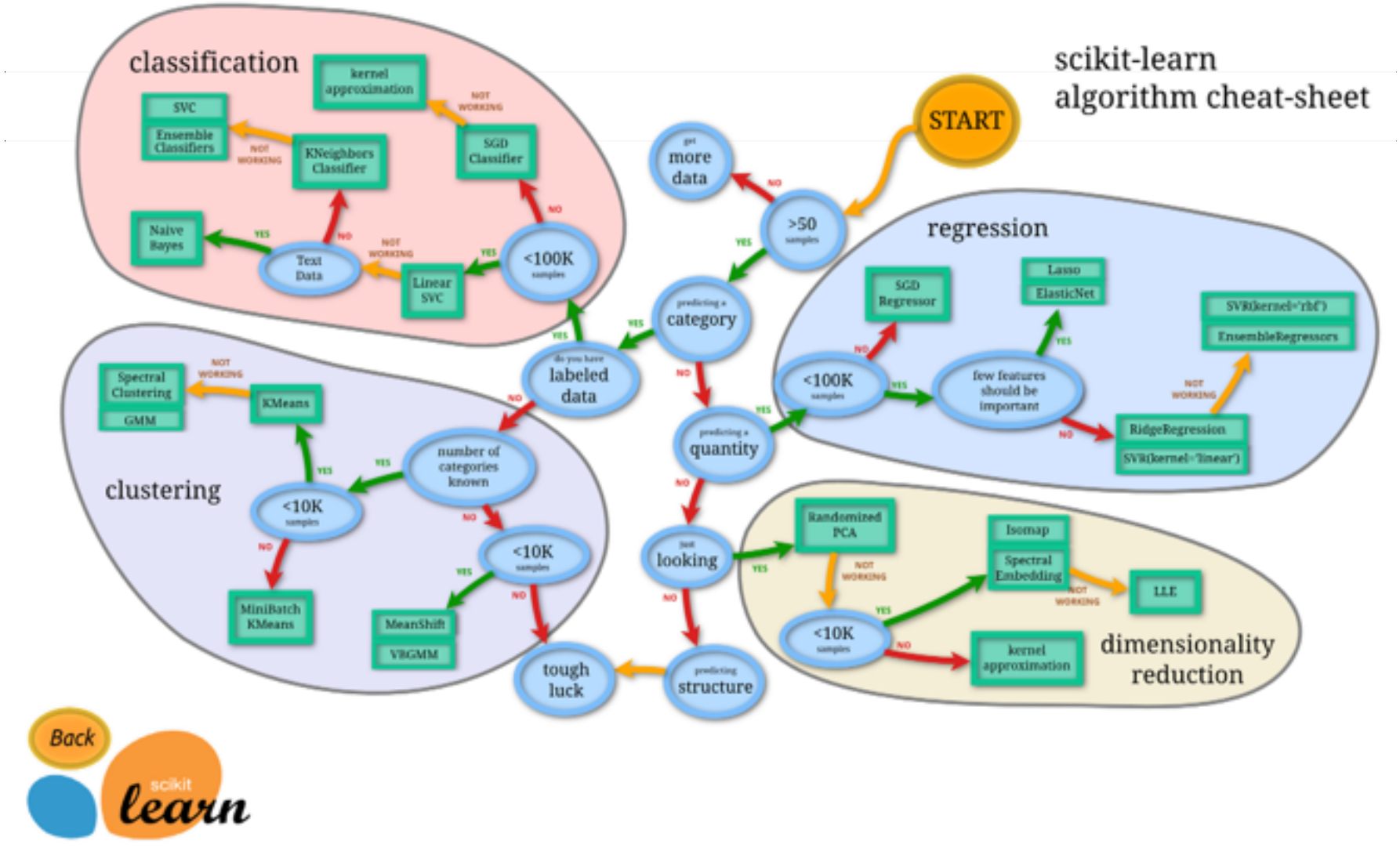
$$SS_{res} = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Basically, what we are trying to do is minimise the Residual Sum of Squares. This is the Sum of the squared difference between our observed value and the value from the model


Oak Diameter vs. Age
DBH (inch)
Age (years)
0
5
10
15
0
20
40
60
80
100

DATA SCIENCE PART TIME COURSE

REVIEW - WEEK 3


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
NO
YES
few features should be important
YES
NOT WORKING
RidgeRegression
NO
SVR(kernel='linear')
YES
predicting a category
YES
do you have labeled data
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
NO
<10K samples
YES
MiniBatch KMeans
MeanShift
VBGMM
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn

We want to build a classifier that correctly identifies which class our target variable y belongs to given our input variable x.

Why not use the linear regression model?

$$y = X\beta + \epsilon$$

- If we only have a binary response variable (0 or 1) it might make sense… BUT we can have our estimated value of $y > 1$ or $y < 0$ … which doesn't make sense.
- What of the case where we have more than one class? Linear regression cannot easily handle these cases.
- We want a classification method that can handle these cases and give us results we can easily interpret.

Q: What's wrong with training error?

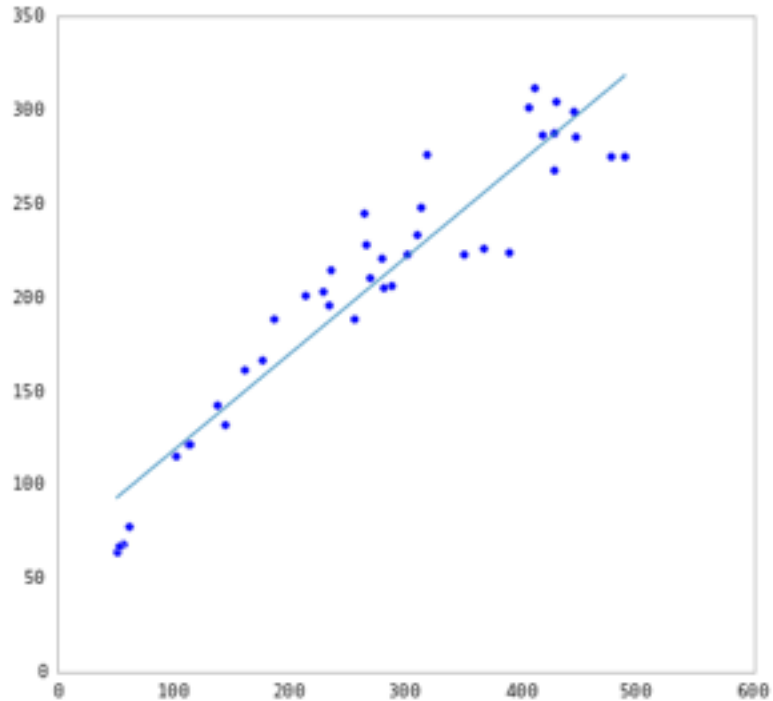
Thought experiment:

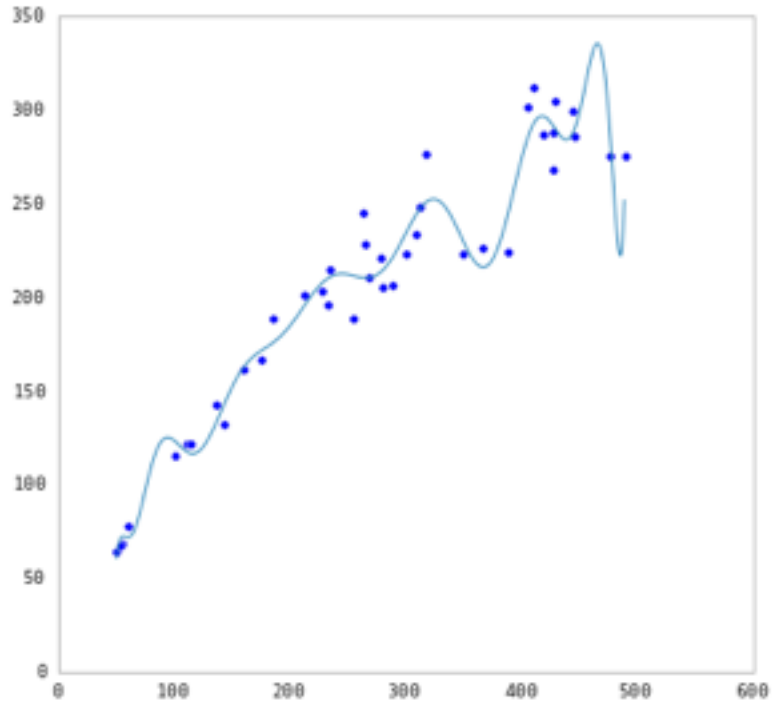
Suppose we train our model using the entire dataset.

Q: How low can we push the training error?
- We can make the model arbitrarily complex (effectively "memorizing" the entire training set).

A: Down to zero!

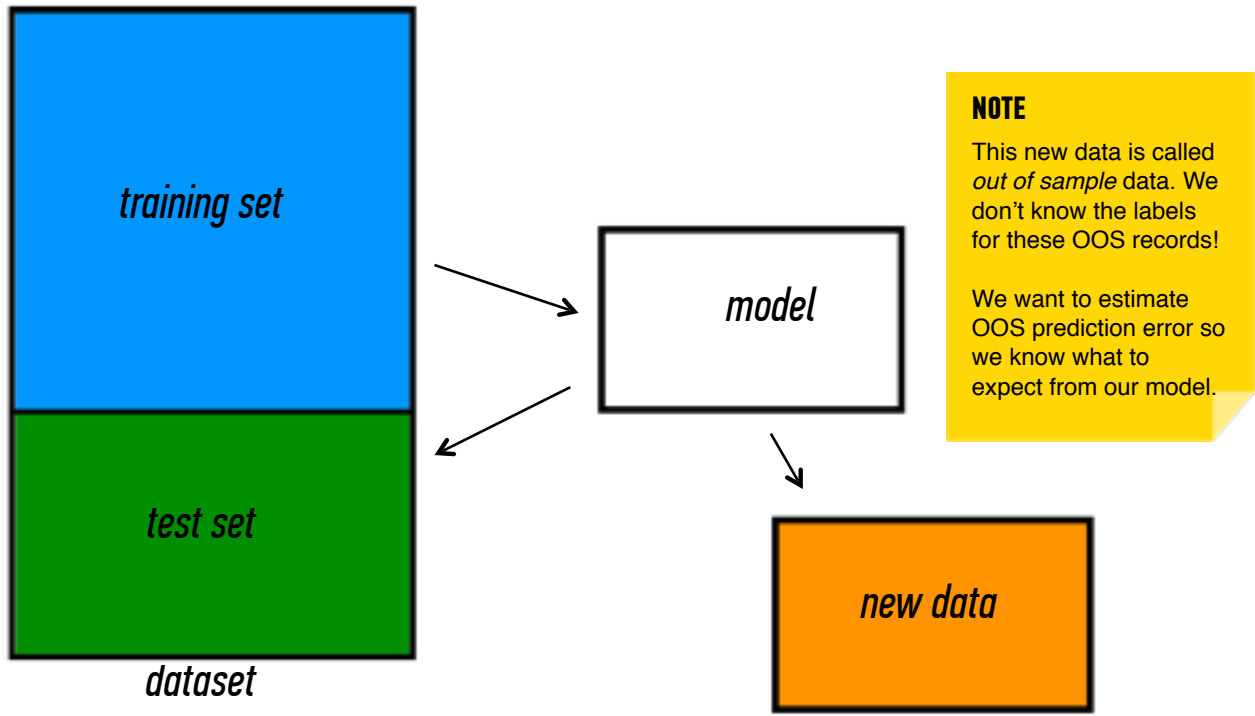
TRAINING ERROR





Q: How can we make a model that generalizes well?

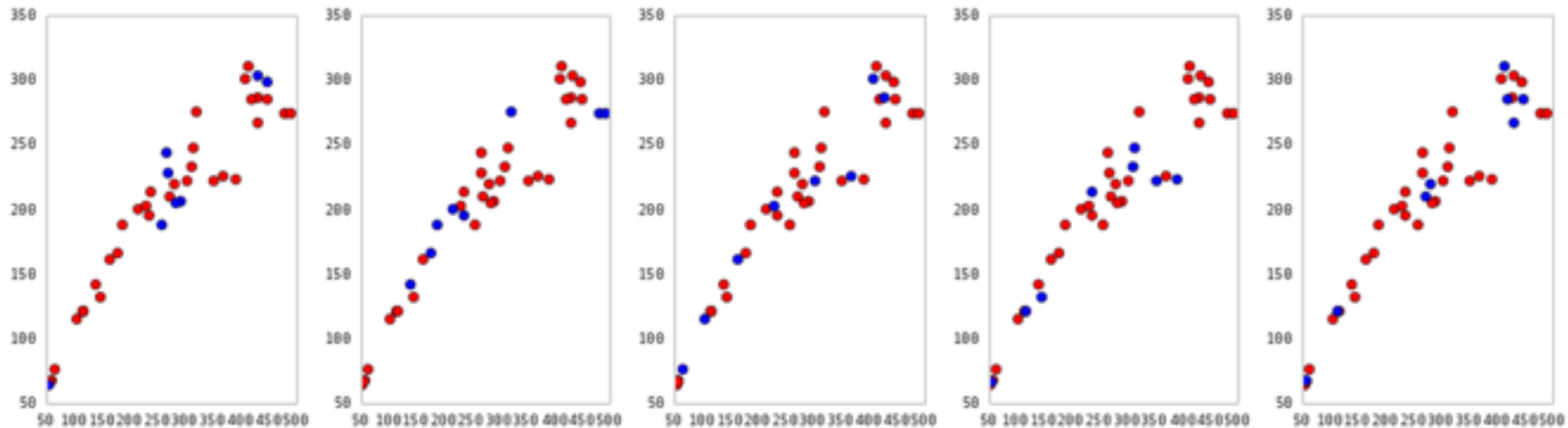
1) split dataset
2) train model
3) test model
4) parameter tuning
5) choose best model
6) train on all data
7) make predictions on new data



Steps for K-fold cross-validation:

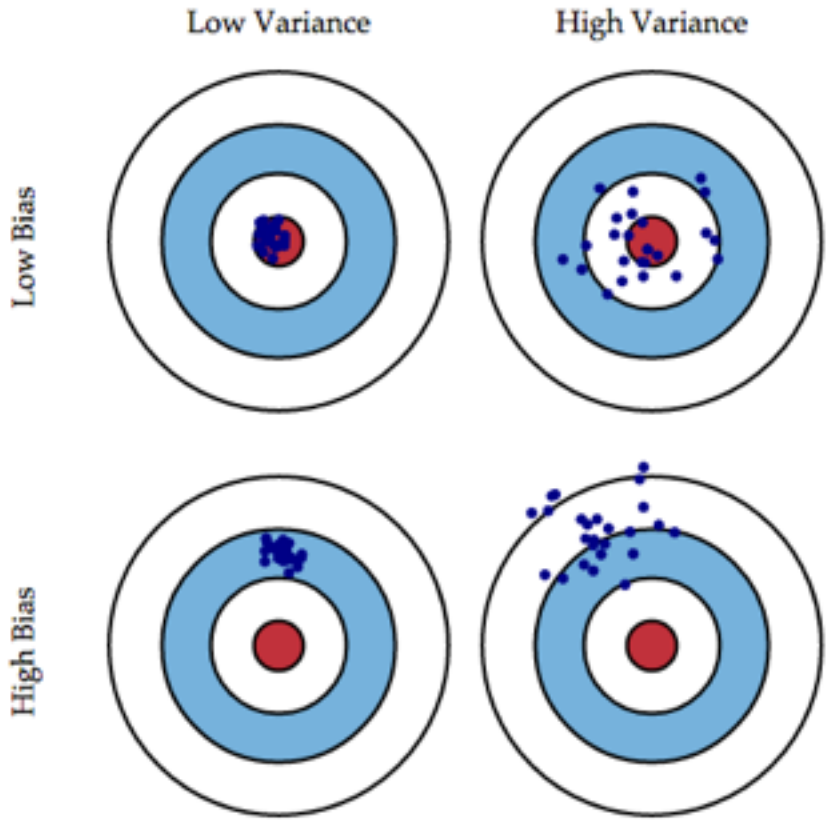
1) Randomly split the dataset into K equal partitions.
2) Use partition 1 as test set & union of other partitions as training set.
3) Calculate test set error.
4) Repeat steps 2-3 using a different partition as the test set at each iteration.
5) Take the average test set error as the estimate of OOS accuracy.

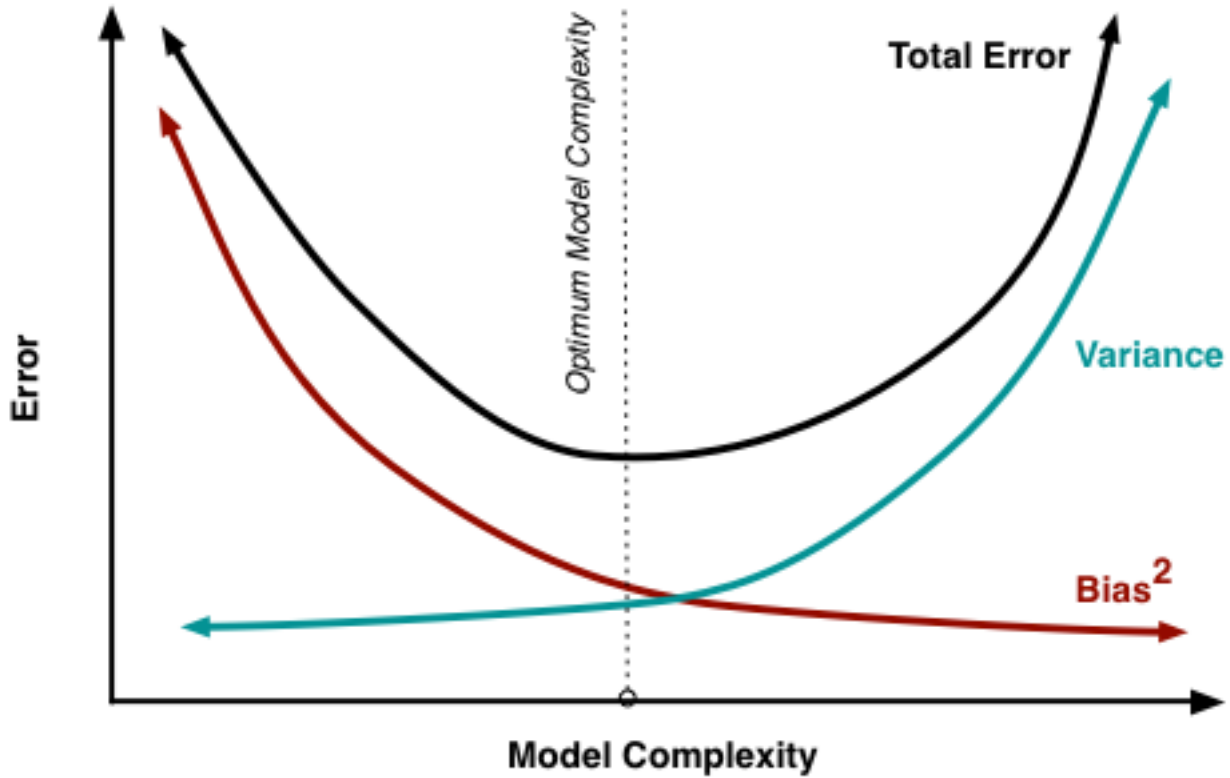
CROSS VALIDATION

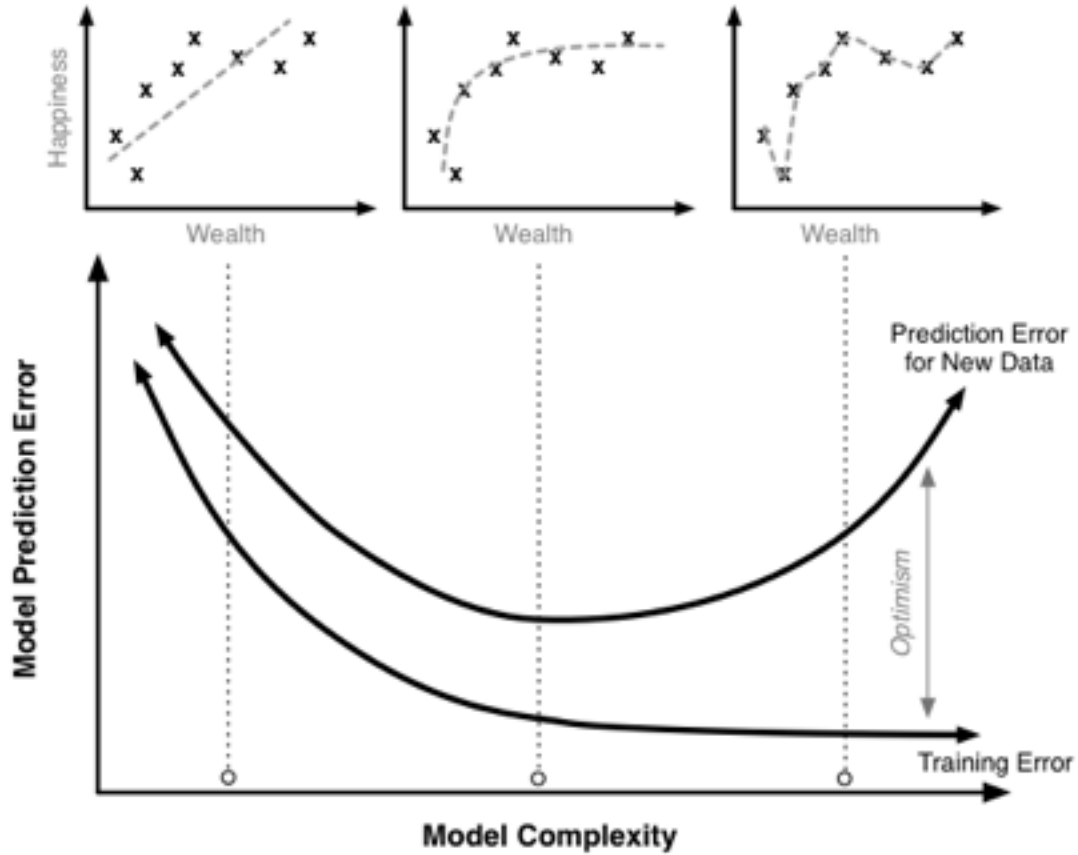


5-fold cross-validation: red = training folds, blue = test fold

Source: http://nbviewer.ipython.org/github/fonnesbeck/Bios366/blob/master/notebooks/Section6_3-Model-Selection-and-Validation.ipynb


Low Variance
High Variance
Low Bias
High Bias


Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity


Happiness
Wealth
Wealth
Wealth
Model Prediction Error
Prediction Error
for New Data
Optimism
Training Error
Model Complexity

DATA SCIENCE PART TIME COURSE

REVIEW - WEEK 4

We could fit a separate linear regression model for every combination of our features.

But what happens when we have a large number of features?

Computation time becomes a factor and we also need to consider that as we include more features we are increasing the chance we include a variable that doesn't add any predictive power for future data.

- A tuning parameter lambda (or sometimes alpha) imposes a penalty on the size of coefficients.
- Instead of minimizing the "loss function" (mean squared error), it minimizes the "loss plus penalty".
- A tiny alpha imposes no penalty on the coefficient size, and is equivalent to a normal linear model.
- Increasing the alpha penalizes the coefficients and shrinks them toward zero.

Ridge Regression is similar to least squares, except we include a penalty term,

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

the λ term is a tuning parameter. When it is zero we get least squares, as it increases the term, $\lambda\sum_{j=1}^{p}\beta_j^2$ (the shrinkage penalty) has more of an

impact and the coefficients will *approach* zero.

Lasso Regression is similar to Ridge Regression, except we have the absolute value of beta in our penalty term,

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

the λ term is a tuning parameter. When it is zero we get least squares, as it increases the term, $\lambda\sum_{j=1}^{p}|\beta_j|$ (the shrinkage penalty) has more of an

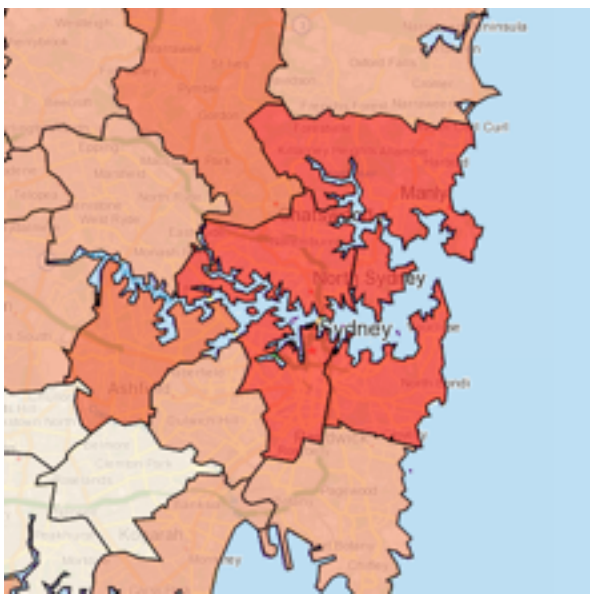
impact and the coefficients will **equal** zero.

Recall unsupervised learning is when we are trying to find interesting patterns or groups in our data. We don't have a variable we are trying to predict (a Y value).

Clustering aims to discover subgroups in our data where the points are similar to each other. So we have a collection of groups and all points belonging to the same group are similar. Points in different groups are different to each other.

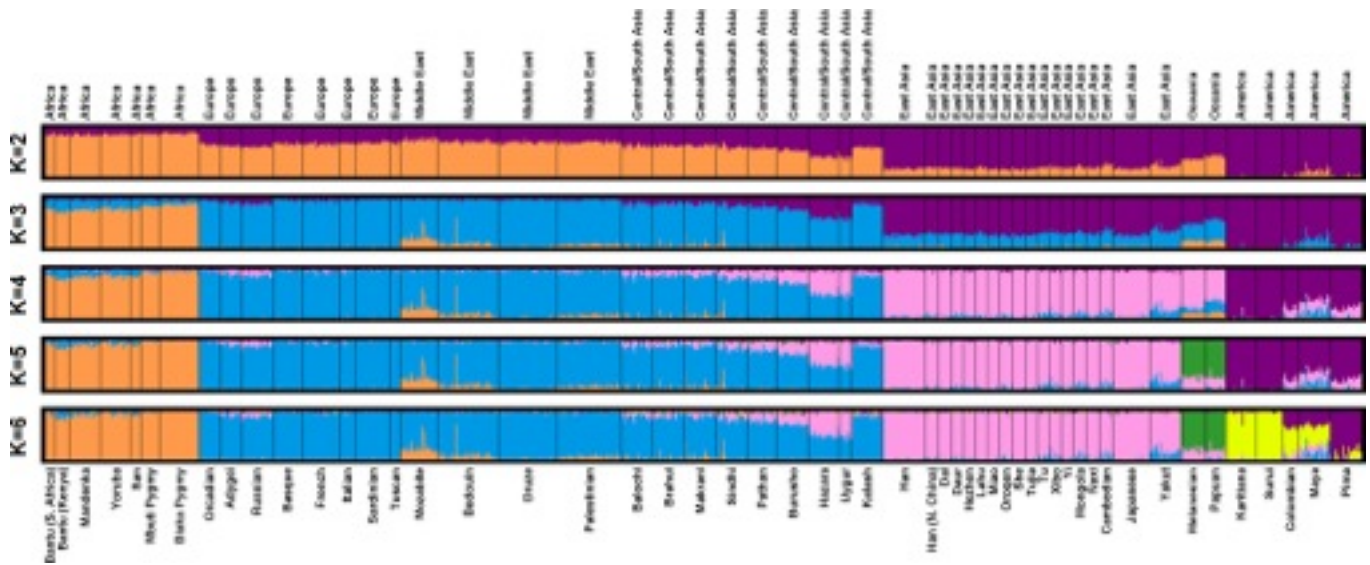
We have to decide what variables we will construct the groups on. What makes them different (or similar)?

Marketing teams might want to group customers into like groups as a way of summarising the data



Financial groups may want to group transactions into like groups as a way to find unusual payments

Genetics data can be clustered to identify ancestry

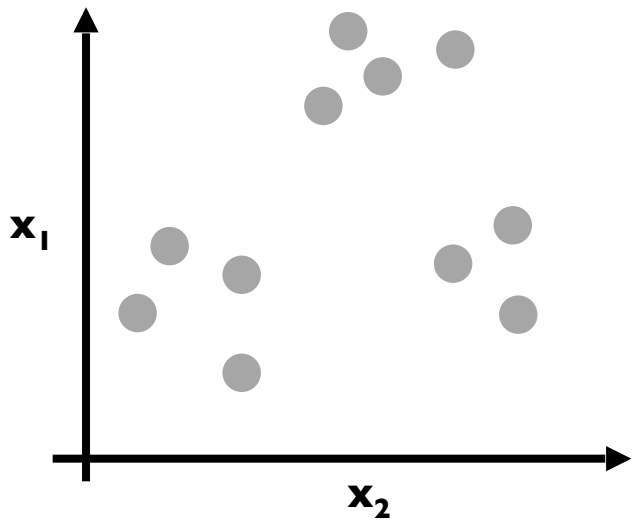


1) Choose k initial centroids

2) For each point:
   - find distance to each centroid
   - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met

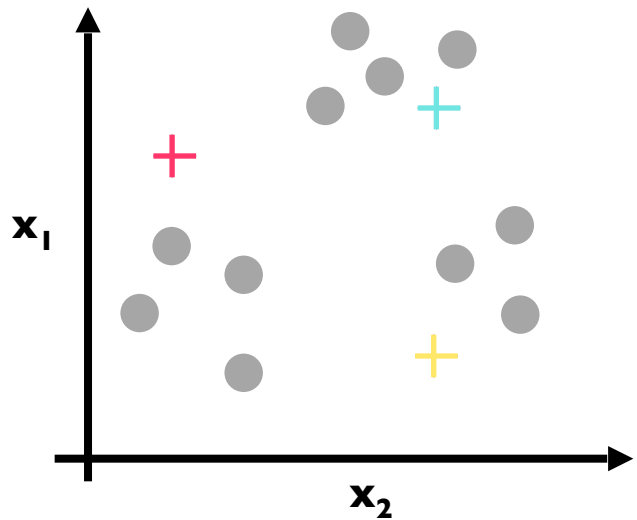


1) **Choose k initial centroids**

2) For each point:
   - find distance to each centroid
   - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met

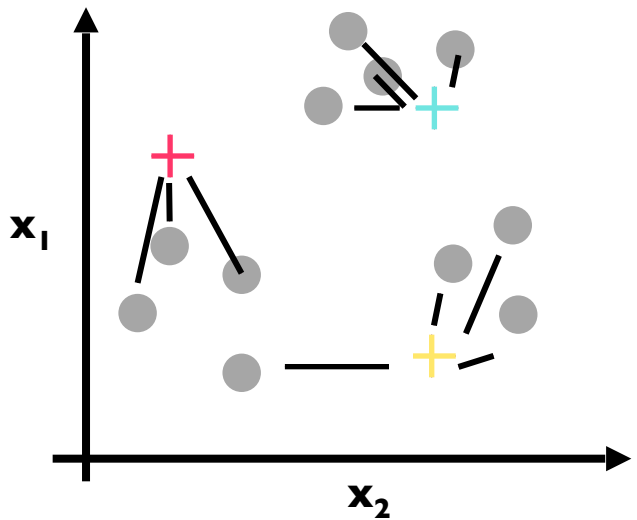


1) Choose k initial centroids

2) For each point:
   - **find distance to each centroid**
   - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met

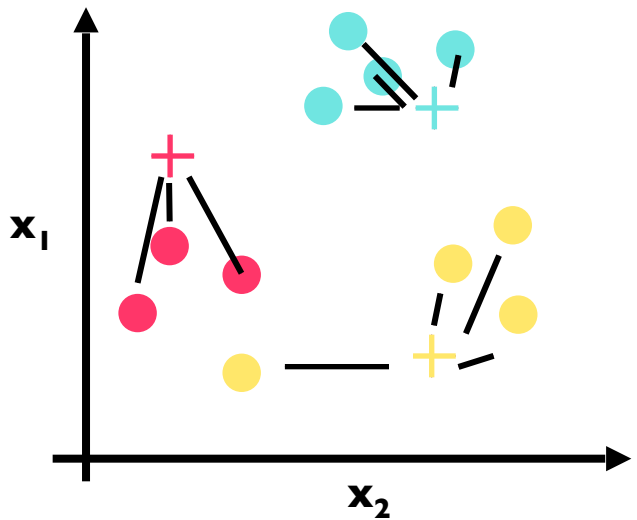


1) Choose k initial centroids

2) For each point:
   - find distance to each centroid
   - **assign point to nearest centroid**

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met

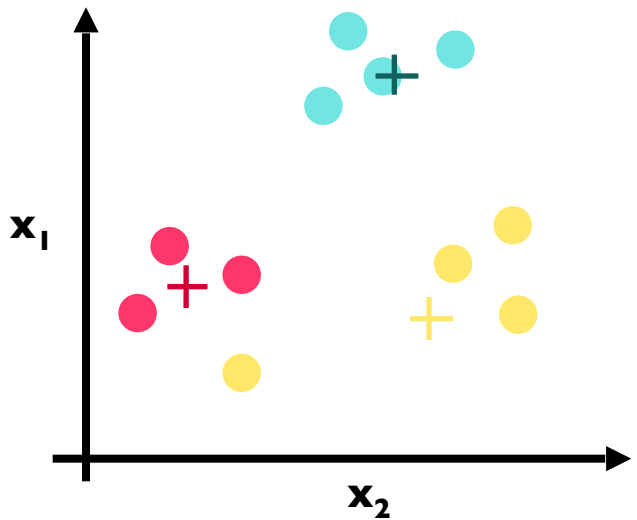


1) Choose k initial centroids

2) For each point:
   - find distance to each centroid
   - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met

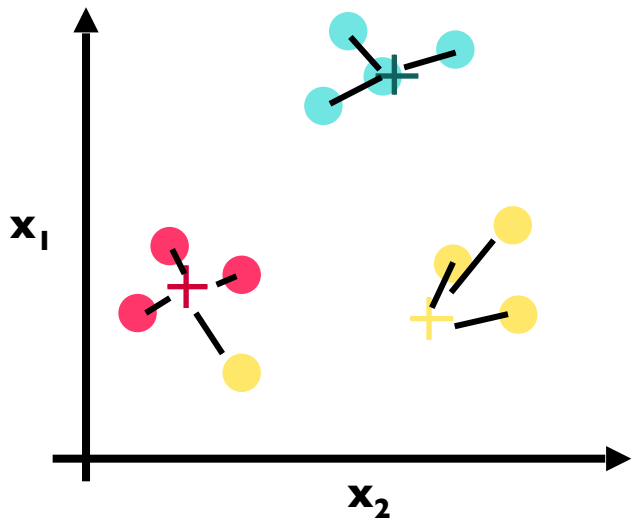


1) Choose k initial centroids

2) For each point:
   - find distance to each centroid
   - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met

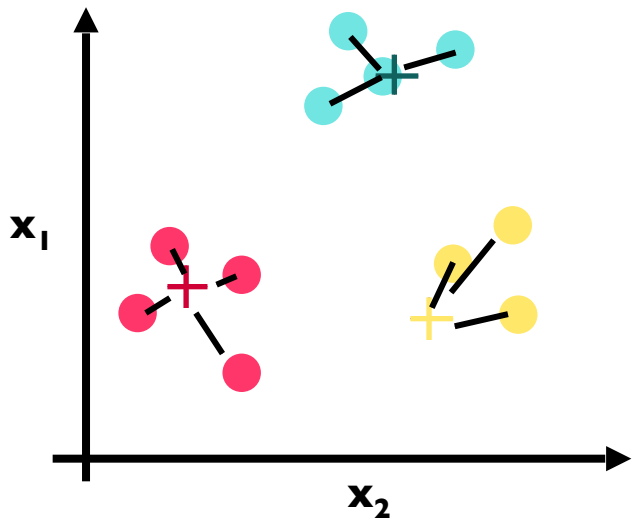


1) Choose k initial centroids

2) For each point:
   - find distance to each centroid
   - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met

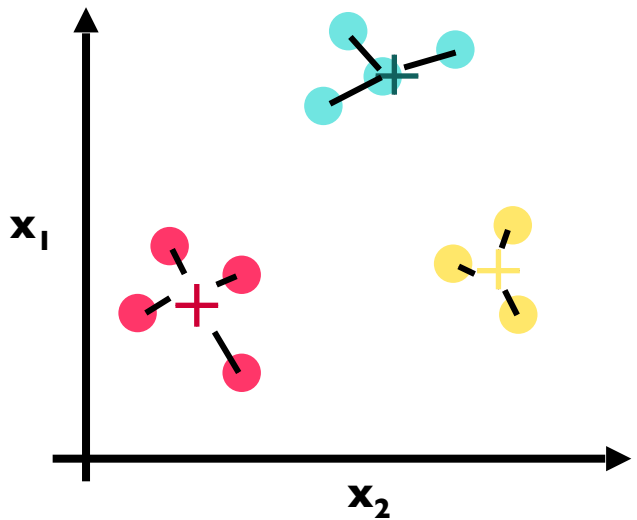


1) Choose k initial centroids

2) For each point:
   - find distance to each centroid
   - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met

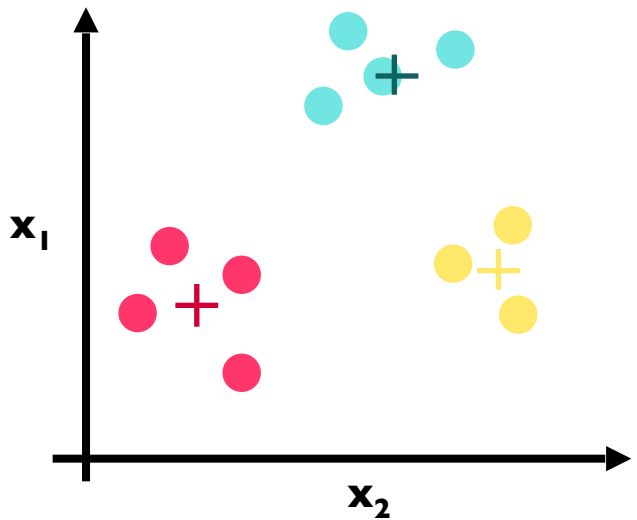


1) Choose k initial centroids

2) For each point:
   - find distance to each centroid
   - assign point to nearest centroid

3) Recalculate centroid positions

4) Repeat steps 2-3 until stopping criteria met



DATA SCIENCE PART TIME COURSE

REVIEW - WEEK 5

Recommendation engines aims to match users to things (movies, songs, items, events, etc) they might enjoy but have not yet tried.

The rating is produced by analysing other user/item ratings (and sometimes item characteristics) to provide personalised recommendations to users.

Content-based filtering begins by mapping each item into a feature space. Both users and items are represented by vectors in this space.

Item vectors measure the degree to which the item is described by each feature, and user vectors measure a user's preferences for each feature.

Ratings are generated by taking dot products of user & item vectors.

Content-based filtering has some difficulties:

- Must map items into a feature space (manual work)
- Recommendations are limited in scope (items must be similar to each other)
- Hard to create cross-content recommendations (eg books/music films...this would require comparing elements from different feature spaces)

“Customers who purchased X also purchased Y”

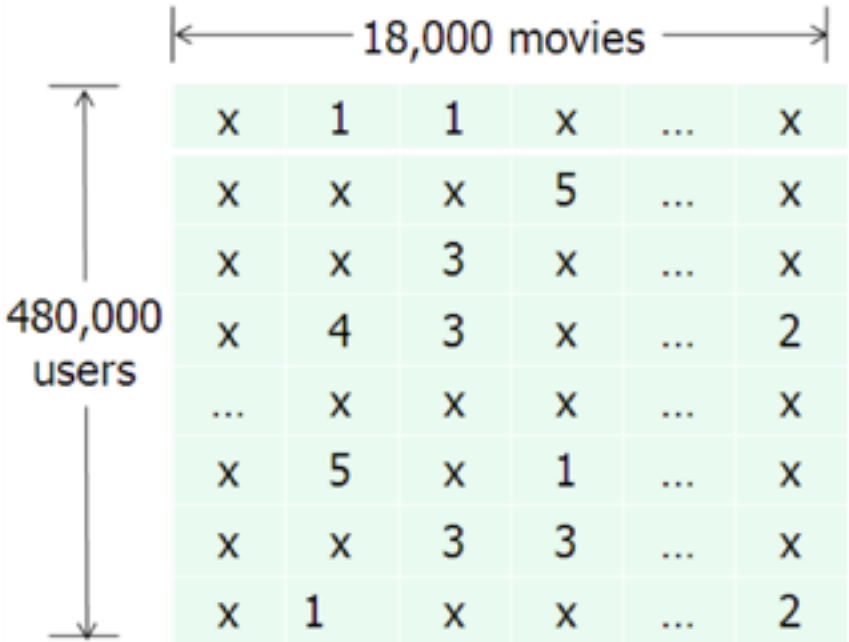
Someone with similar tastes to you will be able to recommend things you might like, e.g. people who watch ‘The Newsroom’ will probably enjoy ‘The Social Network’ because there is a large audience in common.

Collaborative filtering refers to a family of methods for predicting ratings where instead of thinking about users and items in terms of a feature space, we are only interested in the existing user-item ratings themselves.

In this case, our dataset is a ratings matrix whose columns correspond to items, and whose rows correspond to users.

This will be the general form of the data we analyse for collaborative filtering.

The method relies on previous user-item ratings (or feedback).



DATA SCIENCE PART TIME COURSE

SQL & PRODUCTIVITY TOOLS

Rules on structure make writing and retrieving data more reliable and efficient.

- Standardised business definitions
- Source of truth
- Fast read

Where databases are valued:

- Operational systems
- Reporting

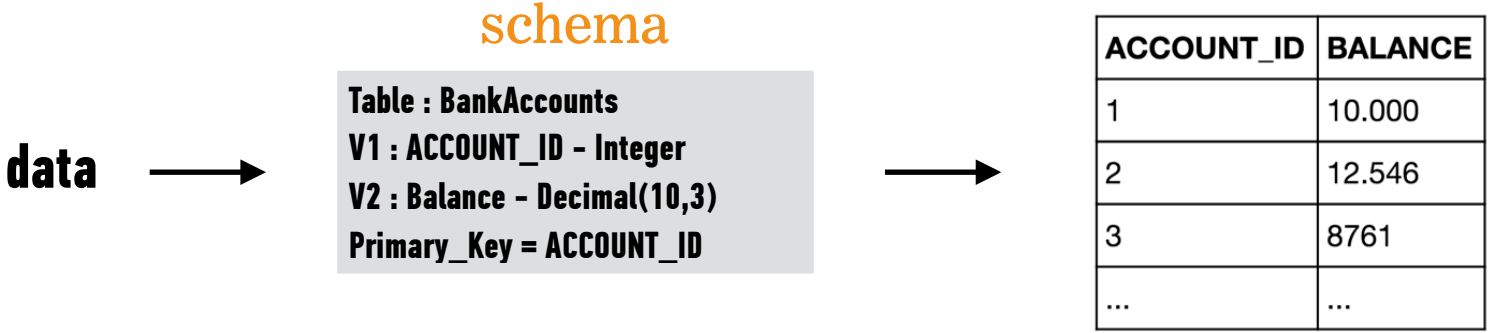
A relational database is a database based tabular data and links between data entities or concepts.

A relational database is organised into tables. Each table should correspond to one entity or concept.

ACCOUNT_ID	BALANCE
1	10.000
2	12.546
3	8761
...	...

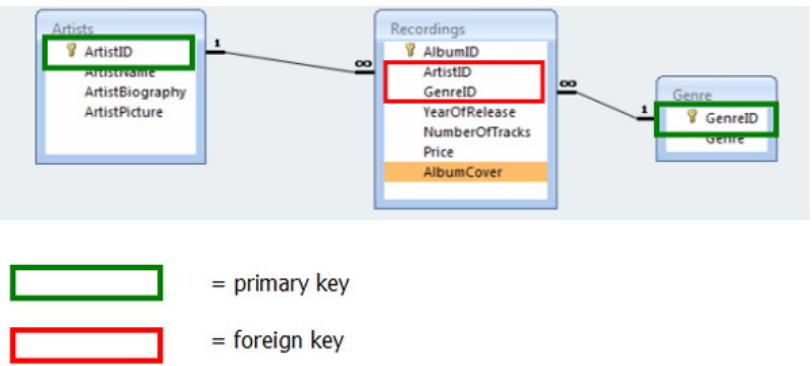
A table is made up rows and columns, similar to a Pandas dataframe or Excel spreadsheet.

A table also has a schema which is a set of rules for what goes in each table. These specify what columns are contained in the table and what type those columns are (text, integers, floats, etc.).



Each table typically has a primary key column. This column is a unique value per row and serves as the identifier for the row.

A table can have many foreign keys as well. A foreign key is a column that contains values to link the table to the other tables.

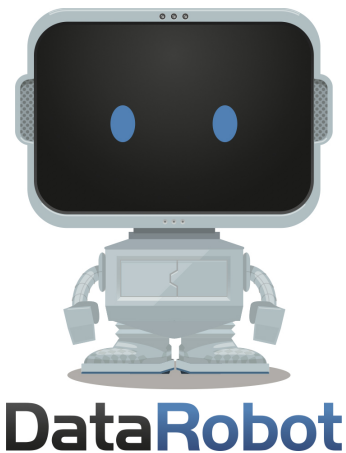


Workflow, collaboration, operationalising ⟷ Automatic Machine Learning - Auto ML

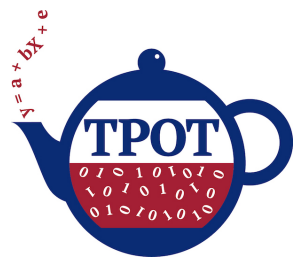
rapidminer

sas

scienceOPS



DataRobot



Auto-Sklearn

DATA SCIENCE PART TIME COURSE

MODELLING CASE STUDY

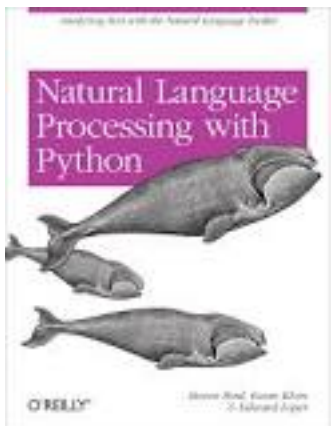
DATA SCIENCE PART TIME COURSE

TEXT ANALYSIS

- Text is considered to be un-structured data. This means we don't have nice features we can use as inputs. We will have to construct them using a model or rules we know about language.
- When analysing text we need to consider what we actually want to use the text for
  - Sentiment Analysis
  - Topic Labelling
  - Classification

- Entity Extraction
- Sentiment Analysis
- Keyword Extraction
- Concept Tagging
- Relation Extraction
- Taxonomy Classification
- Author Extraction
- Language Detection
- Text Extraction
- Microformats Parsing
- Feed Detection
- Linked Data Support

- NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.



DATA SCIENCE PART TIME COURSE

WEB SCRAPING

- Using a program to download and process a website
- Lots of websites have API's which can be used to collect data - START WITH THIS
- If they don't have an API you will need to download the material yourself (or get a freelancer to do it)

- urllib2: useful for fetching URLs and handling things like cookies and basic authentication
- BeautifulSoup: great at extracting information from a page
- Scrapy: Create web scraping routines

States and territories of Australia

From Wikipedia, the free encyclopedia

Australia (officially known as the Commonwealth of Australia) is a federation of six **states**, together with ten federal **territories**. The Australian mainland consists of five of the six federated states—the state of Tasmania is an island in close proximity to the mainland—and three of the federal territories. Aside from the Australian Antarctic Territory, which is Australia's claim to part of Antarctica, Australia is the world's sixth-largest country by total area.

Australian Antarctic Territory

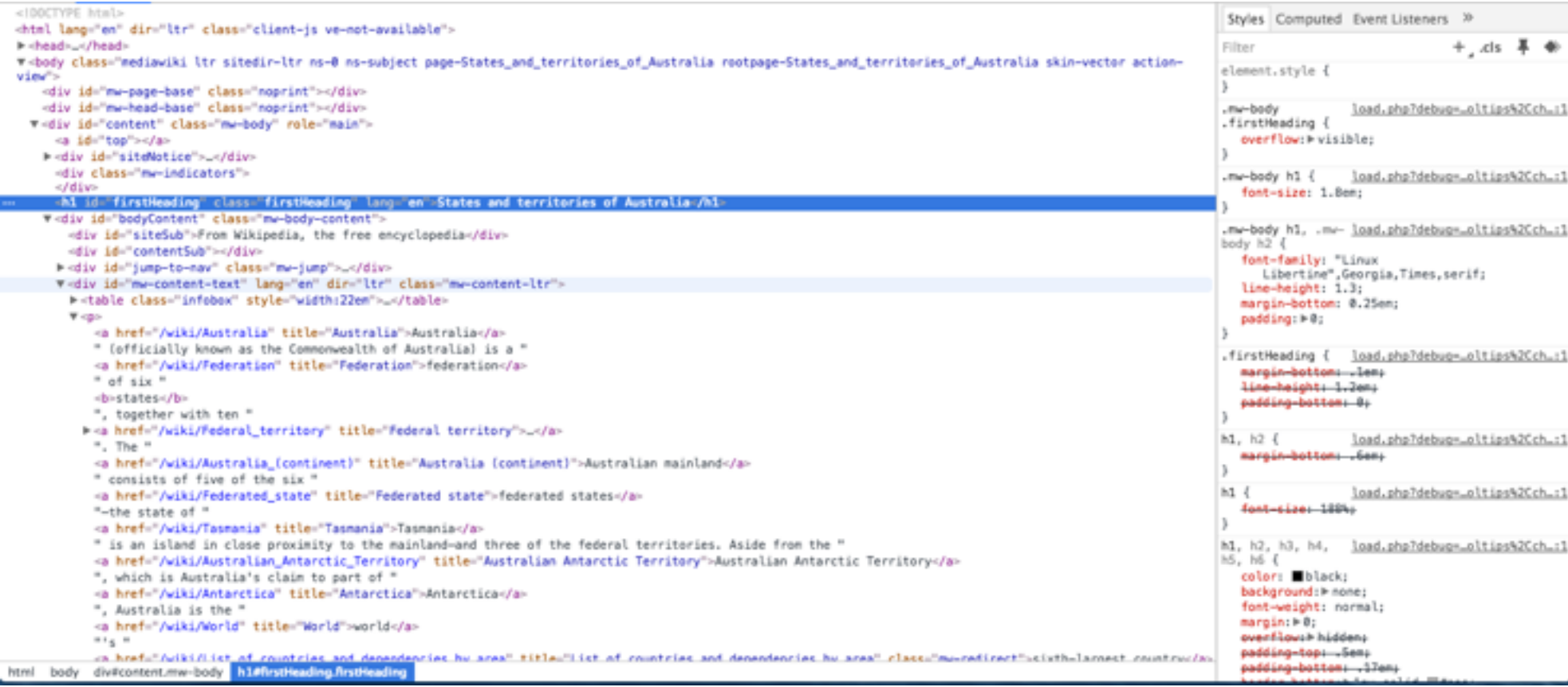
The three mainland territories are classified for some purposes as "internal" territories, the others as "external" territories. All states and the two largest internal territories are partially self-governing, as well as being represented in the federal parliament; the other territories are administered by the federal government. Three of the external territories are inhabited; the others are uninhabited, apart from non-permanent scientists.

Contents [hide]

Australian States and Territories

A clickable map of Australia's states, mainland territories and their capitals

Category Federated states (6)
Federal territories (3)
External territories (7)


<!DOCTYPE html>
<html lang="en" dir="ltr" class="client-js ve-not-available">
<head>…</head>
<body class="mediawiki ltr sitedir-ltr ns-0 ns-subject page-States_and_territories_of_Australia rootpage-States_and_territories_of_Australia skin-vector action-view">

…

<h1 id="firstHeading" class="firstHeading" lang="en">States and territories of Australia</h1>

From Wikipedia, the free encyclopedia

…

<table class="infobox" style="width:22em">…</table>
<p>
Australia
" (officially known as the Commonwealth of Australia) is a "
federation
" of six "
states
", together with ten "
…
". The "
Australian mainland
" consists of five of the six "
federated states
"—the state of "
Tasmania
" is an island in close proximity to the mainland—and three of the federal territories. Aside from the "
Australian Antarctic Territory
", which is Australia's claim to part of "
Antarctica
", Australia is the "
world
"'s "
html body div#content.mw-body h1#firstHeading.firstHeading
Styles Computed Event Listeners
Filter
.cls
element.style {
}
.mw-body .firstHeading {
overflow: visible;
}
load.php?debug=…oltips%2Cch…:1
.mw-body h1 {
font-size: 1.8em;
}
load.php?debug=…oltips%2Cch…:1
.mw-body h1, .mw-body h2 {
font-family: "Linux Liberitne",Georgia,Times,serif;
line-height: 1.3;
margin-bottom: 0.25em;
padding: 0;
}
load.php?debug=…oltips%2Cch…:1
.firstHeading {
margin-bottom: .1em;
line-height: 1.2em;
padding-bottom: 0;
}
load.php?debug=…oltips%2Cch…:1
h1, h2 {
margin-bottom: .6em;
}
load.php?debug=…oltips%2Cch…:1
h1 {
font-size: 188%;
}
load.php?debug=…oltips%2Cch…:1
h1, h2, h3, h4, h5, h6 {
color: black;
background: none;
font-weight: normal;
margin: 0;
overflow: hidden;
padding-top: .5em;
padding-bottom: .17em;

1. Figure out the URL (where to scrap from)
2. Query the website and fetch the HTML
3. Parse the website with something like Beautiful Soup
4. Query the dataset and extract the information of interest
5. Repeat

DATA SCIENCE PART TIME COURSE

PREPPING DATA


Big Data Borat
@BigDataBorat
Follow
In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.
RETWEETS
431
LIKES
171
6:47 PM - 26 Feb 2013

Every dataset is unique and will require it's own approach to cleaning and preparation.

You will need to do some summaries as a basic 'sniff test', were you expecting the distributions of numeric variables? Were you expecting to have an overall count of some categories?

How many NA's or missing variables do you have?

Am I using data from the future (relative to the observation)? This is information leakage and requires more thinking about the problem than data manipulation.

You usually don't want to exclude a variable because it has a missing value in the column

To include it most algorithms require all values be non-NA's

The best way to impute data is dependent on the problem (but it's usually not a good idea to fill it with the average).

An excellent library in R for missing data is Amelia

DATA SCIENCE – Week 7 Day 3

MORE LEARNING

- Udacity Deep Learning Course by Google – https://www.udacity.com/course/deep-learning--ud730
- Gaussian Process Models – http://www.gaussianprocess.org/gpml/
- Great Podcast – http://www.thetalkingmachines.com/

DATA SCIENCE – Week 7 Day 3

COURSE REMAINDER

- Cloud Computing
- Time Series
- Graph Analysis
- Neural Networks
- Natural Language Processing
- Communication Skills and Job Market Overview
- Project Presentations (Final 2 sessions)