GENERAL ASSEMBLY

DATA SCIENCE

10 WEEK PART TIME COURSE

Week 10 - Course review of later modules
Monday 12th December 2016

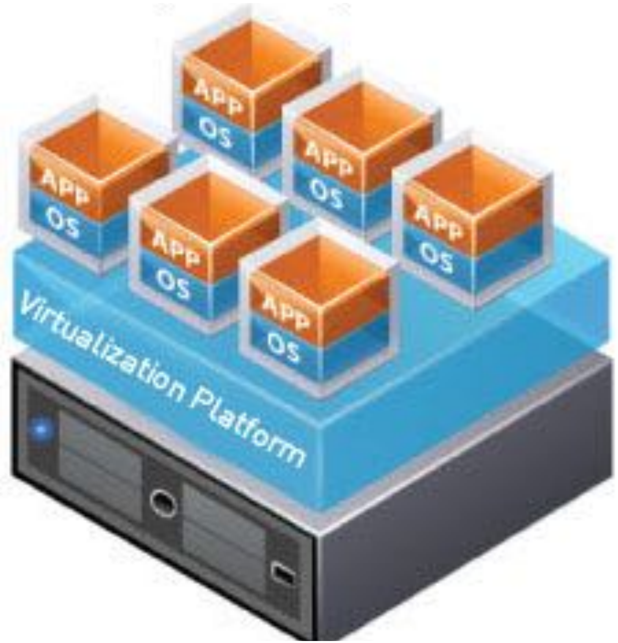
AGENDA

1. Course Review - Later Modules
    1. Cloud Computing
    2. Natural Language Processing
    3. Graphs & Network Analysis
    4. Time Series
    5. Causality
    6. Neural Networks
2. Presentations

DATA SCIENCE PART TIME COURSE

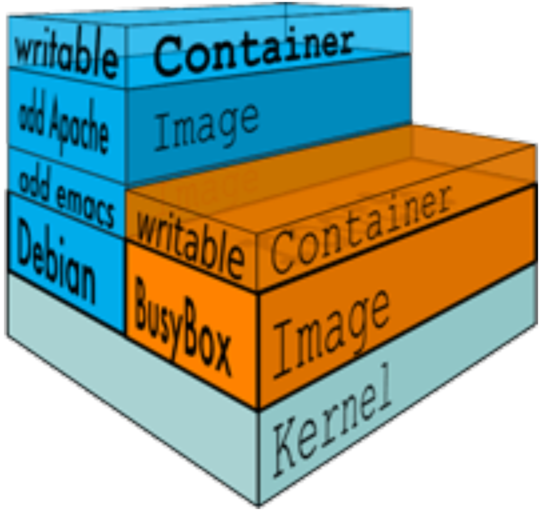
CLOUD COMPUTING

CLOUD


APP
OS
APP
OS
APP
OS
APP
OS
APP
OS
APP
OS
Virtualization Platform

DOCKER – WHAT IS IT?

Docker containers wrap up a piece of software in a complete filesystem that contains everything it needs to run: code, runtime, system tools, system libraries – anything you can install on a server. This guarantees that it will always run the same, regardless of the environment it is running in.



- Lightweight
- Open
- Secure

SQL/NoSQL

SQL

- Traditional rows and columns data
- Strict structure / Primary Keys
- Entire column for each feature
- Industry standard

NoSQL

- No well defined data structure
- Works better for unstructured data
- Cheaper hardware
- Popular among Startups

SQL/NoSQL - Examples

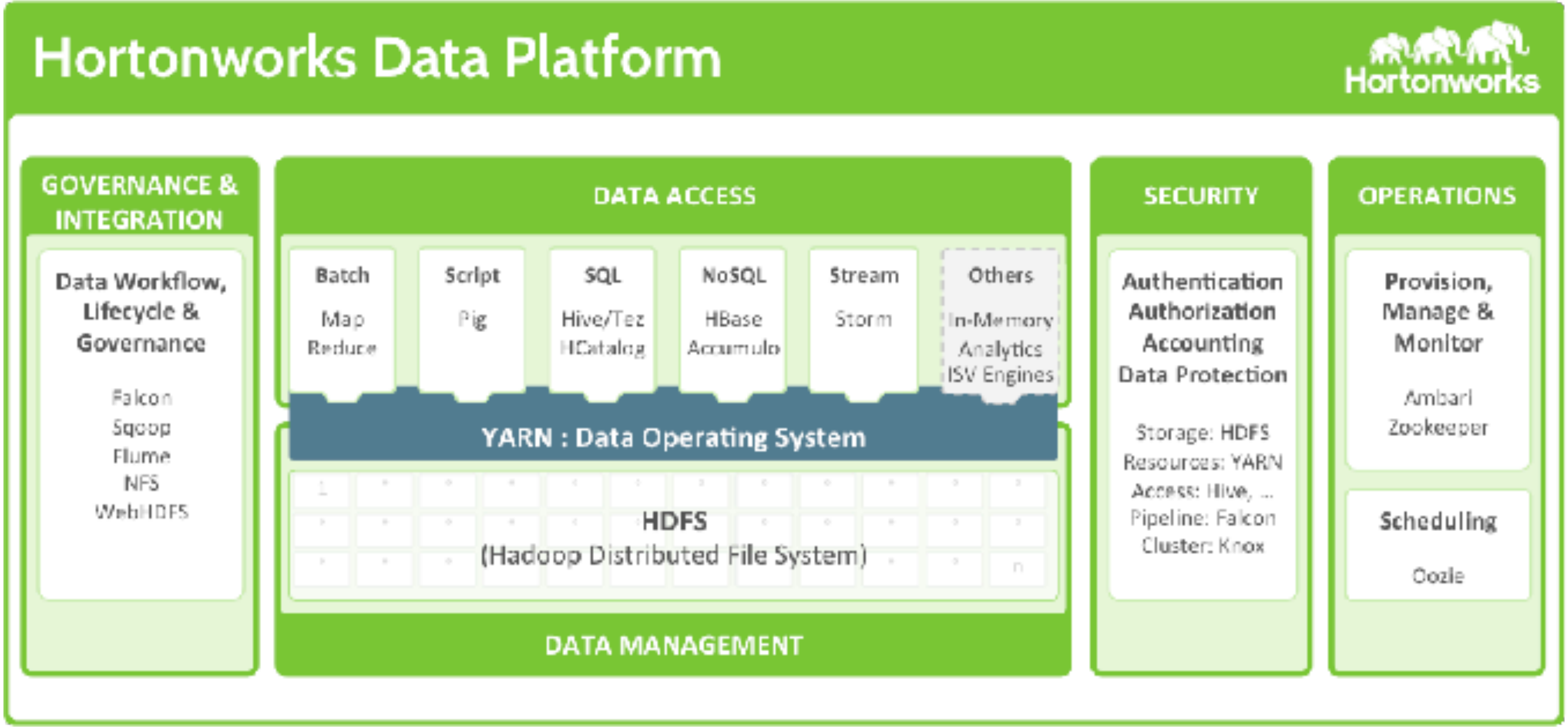
SQL

- MySQL
- Oracle
- Postgres
- SQLite
- SQLServer
- Redshift

NoSQL

- MongoDB
- CouchDB
- Redis
- Cassandra
- Neo4j
- HBase

Hortonworks Stack



Hortonworks Data Platform

Hortonworks

GOVERNANCE & INTEGRATION

Data Workflow, Lifecycle & Governance

Falcon
Sqoop
Flume
NFS
WebHDFS

DATA ACCESS

Batch
Map Reduce

Script
Pig

SQL
Hive/Tez
HCatalog

NoSQL
HBase
Accumulo

Stream
Storm

Others
In-Memory
Analytics
ISV Engines

YARN : Data Operating System

HDFS
(Hadoop Distributed File System)

DATA MANAGEMENT

SECURITY

Authentication
Authorization
Accounting
Data Protection

Storage: HDFS
Resources: YARN
Access: Hive, ...
Pipeline: Falcon
Cluster: Knox

OPERATIONS

Provision, Manage & Monitor

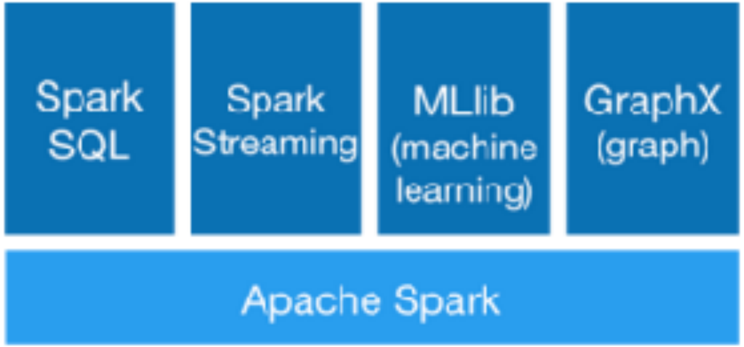
Ambari
Zookeeper

Scheduling

Oozie

SPARK – WHAT IS IT?

Spark is a fast and general processing engine compatible with Hadoop data. It can process data in HDFS, HBase, Cassandra, Hive, and any Hadoop InputFormat. It is designed to perform both batch processing (similar to MapReduce) and new workloads like streaming, interactive queries, and machine learning.



Spark SQL

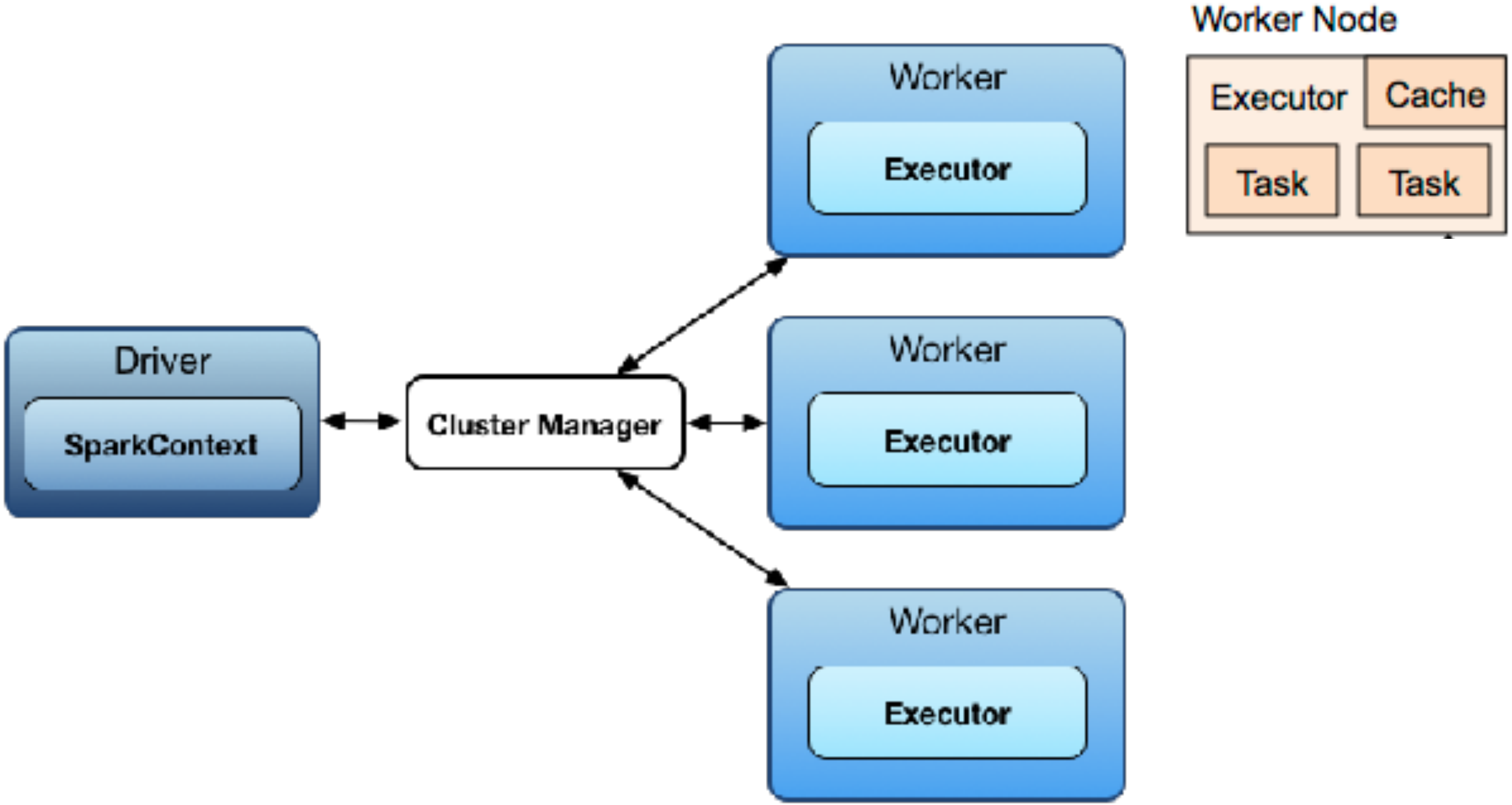
Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

SPARK CLUSTER

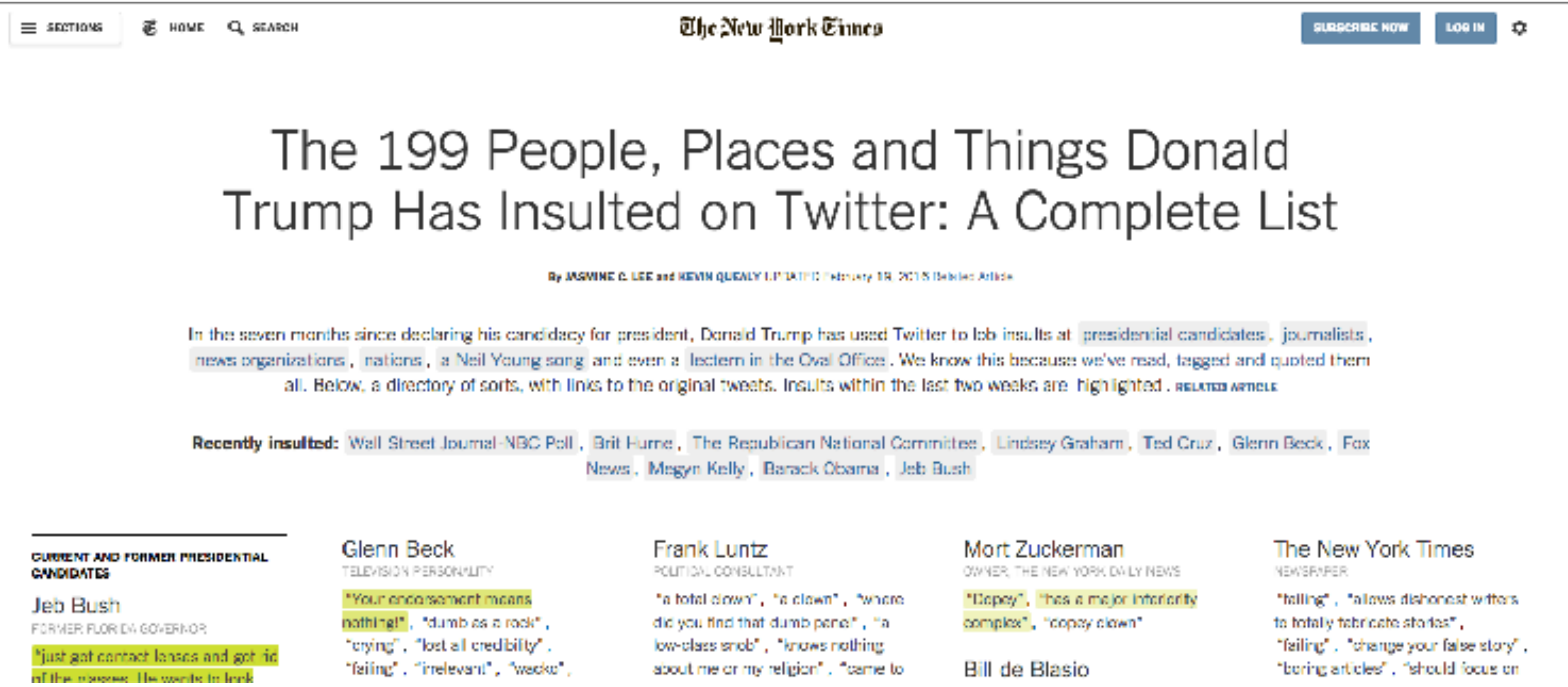


DATA SCIENCE PART TIME COURSE

NATURAL LANGUAGE PROCESSING

WHAT IS NATURAL LANGUAGE PROCESSING?

- 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.
- Natural Language Processing is the algorithms and processing we program to interpret human language.
- It allows us to extract meaning from text as it appears in emails, articles, tweets, journal articles, books, speech, advertisements, etc in the dialect it was created in.



The 199 People, Places and Things Donald Trump Has Insulted on Twitter: A Complete List

By JASMINE C. LEE and KEVIN QUEALY UPDATED February 18, 2016 Related Article

In the seven months since declaring his candidacy for president, Donald Trump has used Twitter to lob insults at presidential candidates, journalists, news organizations, nations, a Neil Young song and even a lectern in the Oval Office. We know this because we've read, tagged and quoted them all. Below, a directory of sorts, with links to the original tweets. Insults within the last two weeks are highlighted. RELATED ARTICLE

Recently insulted: Wall Street Journal-NBC Poll, Brit Hume, The Republican National Committee, Lindsey Graham, Ted Cruz, Glenn Beck, Fox News, Megyn Kelly, Barack Obama, Jeb Bush

CURRENT AND FORMER PRESIDENTIAL CANDIDATES

Jeb Bush
FORMER FLORIDA GOVERNOR

"just got contact lenses and got rid of the glasses. He wants to look

Glenn Beck
TELEVISION PERSONALITY

"Your endorsement means nothing!", "dumb as a rock", "crying", "lost all credibility", "failing", "irrelevant", "wacko",

Frank Luntz
POLITICAL CONSULTANT

"a total clown", "a clown", "where did you find that dumb panel", "a low-class snob", "knows nothing about me or my religion", "came to

Mort Zuckerman
OWNER, THE NEW YORK DAILY NEWS

"Dopey", "has a major inferiority complex", "dopey clown"

Bill de Blasio

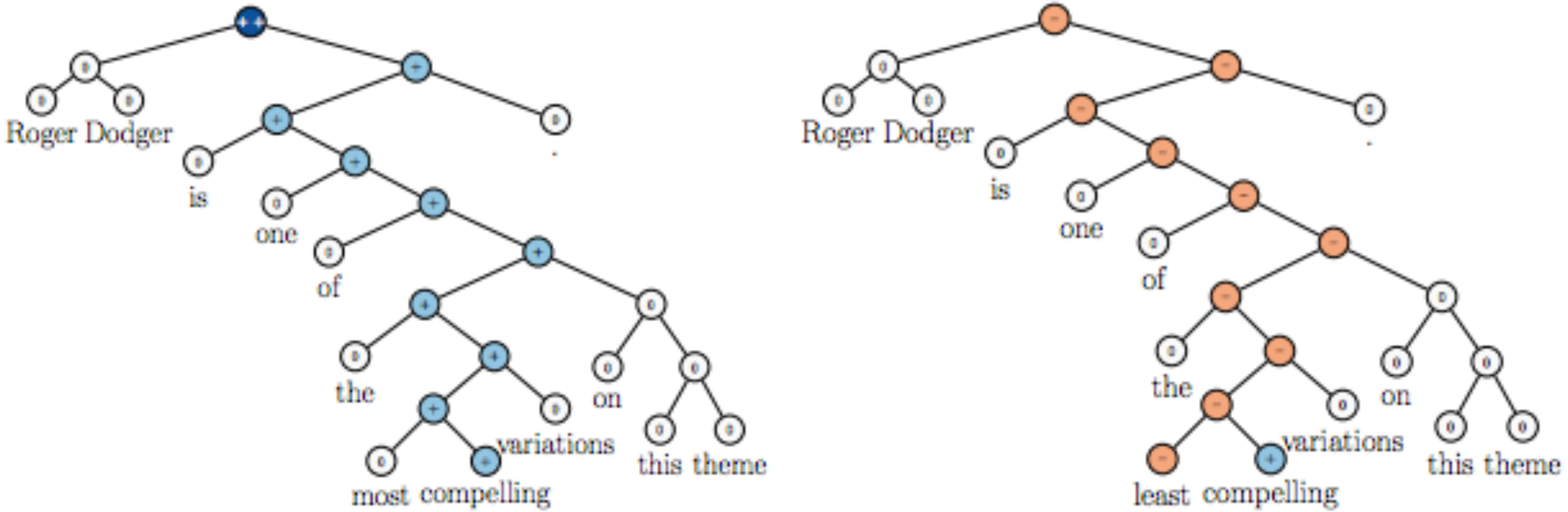
The New York Times
NEWSPAPER

"failing", "allows dishonest writers to totally fabricate stories", "failing", "change your false story", "boring articles", "should focus on

SOME TERMS

- Corpus, a large collection of text used for training (e.g. Gutenberg collection or scraping websites)
- Part-of-Speech tagging, understanding the nature of a word, is it a verb or a noun?
- Lexical Analysis, breaking down the structure of text (ie, Document -> Paragraph -> Sentence -> Words).
- Symbolic approach, using rules from language to parse text (can be manually written).
- Statistical approach, a sequence labelling problem, we try to infer the properties of a word by the words around it.

EXAMPLE – SENTIMENT ANALYSIS



WHAT IS TEXT ANALYSIS?

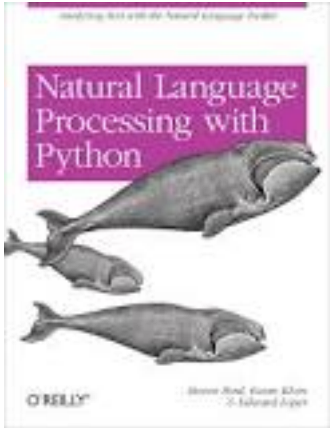
- Entity Extraction
- Sentiment Analysis
- Keyword Extraction
- Concept Tagging
- Relation Extraction
- Taxonomy Classification
- Author Extraction
- Language Detection
- Text Extraction
- Microformats Parsing
- Feed Detection
- Linked Data Support

AlchemyAPI™

An IBM Company

NATURAL LANGUAGE TOOLKIT (NLTK) – Python

- 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

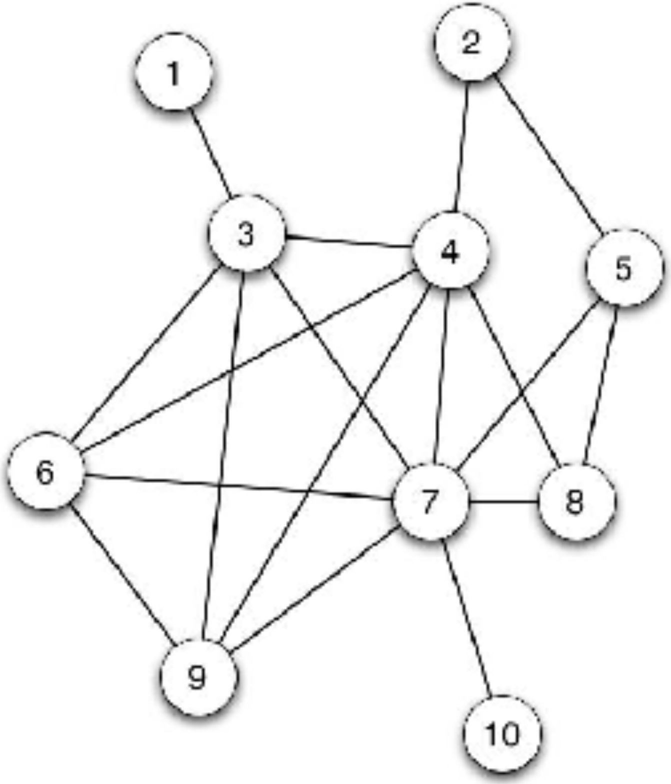
GRAPHS & NETWORK ANALYSIS

WHAT IS A GRAPH?

A graph consists of a nodes (or vertices) and are connected by edges.

For example the nodes may represent people and the edges are there if a friendship exists.

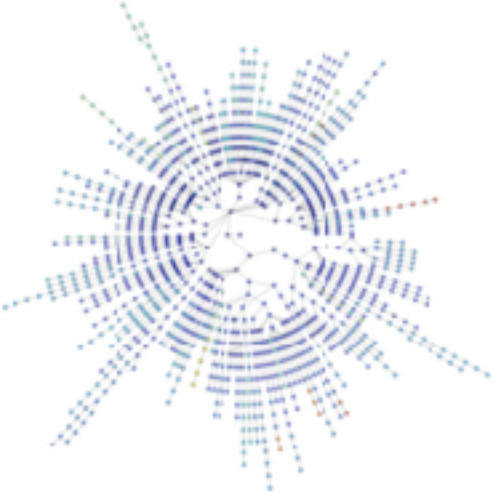
How many nodes and edges are there?

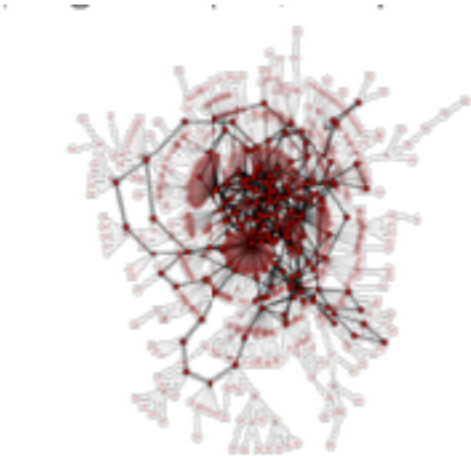


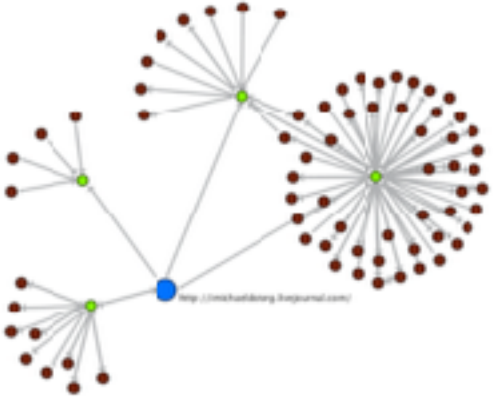
WHAT ARE THE METRICS OF A GRAPH?

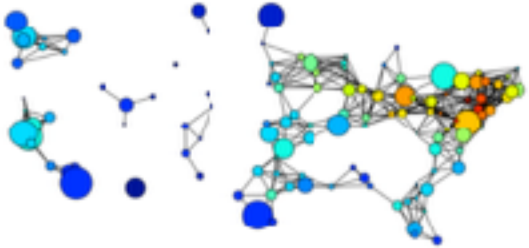
- Degree Centrality - number of edges a node has
- Closeness Centrality - the reciprocal of the sum of the shortest path distances from one node to all n-1 other nodes. Since the sum of distances depends on the number of nodes in the graph, closeness is normalized by the sum of minimum possible distances n-1. Higher values of closeness indicate higher centrality
- Betweenness Centrality - the sum of the fraction of all-pairs shortest paths that pass through the node v
- Eigenvector centrality - computes the centrality for a node based on the centrality of its neighbours
- Page Rank - count the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites

HOW DO WE VISUALISE A GRAPH?









DATA SCIENCE PART TIME COURSE

TIME SERIES

FORECASTING: PRINCIPLES & PRACTICE

1 Getting started

2 The forecaster's toolbox

3 Judgmental forecasts

4 Simple regression

5 Multiple regression

6 Time series decomposition

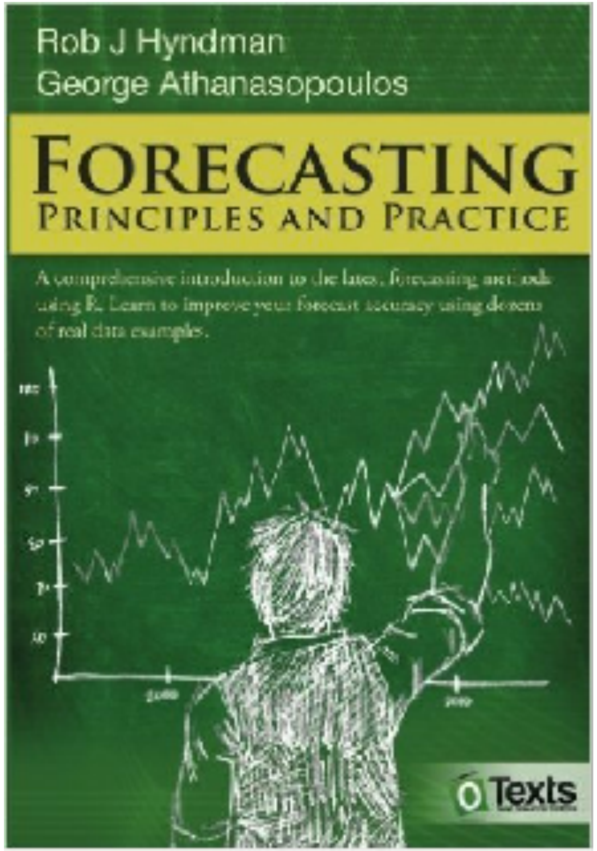
7 Exponential smoothing

8 ARIMA models

9 Advanced forecasting methods

10 Data

12 Using R


Rob J Hyndman
George Athanasopoulos
FORECASTING
PRINCIPLES AND PRACTICE
A comprehensive introduction to the latest forecasting methods using R. Learn to improve your forecast accuracy using dozens of real data examples.
OTexts

WHAT IS A TIME SERIES?

A time series is a series of data that is observed sequentially over time.

Examples include:

- Weekly Rainfall
- Daily Stock price of Atlassian
- Quarterly oil import figures

TIME SERIES COMPONENTS

There are three time series components we will use to describe a time series

- Trend
- Seasonal
- Cyclic

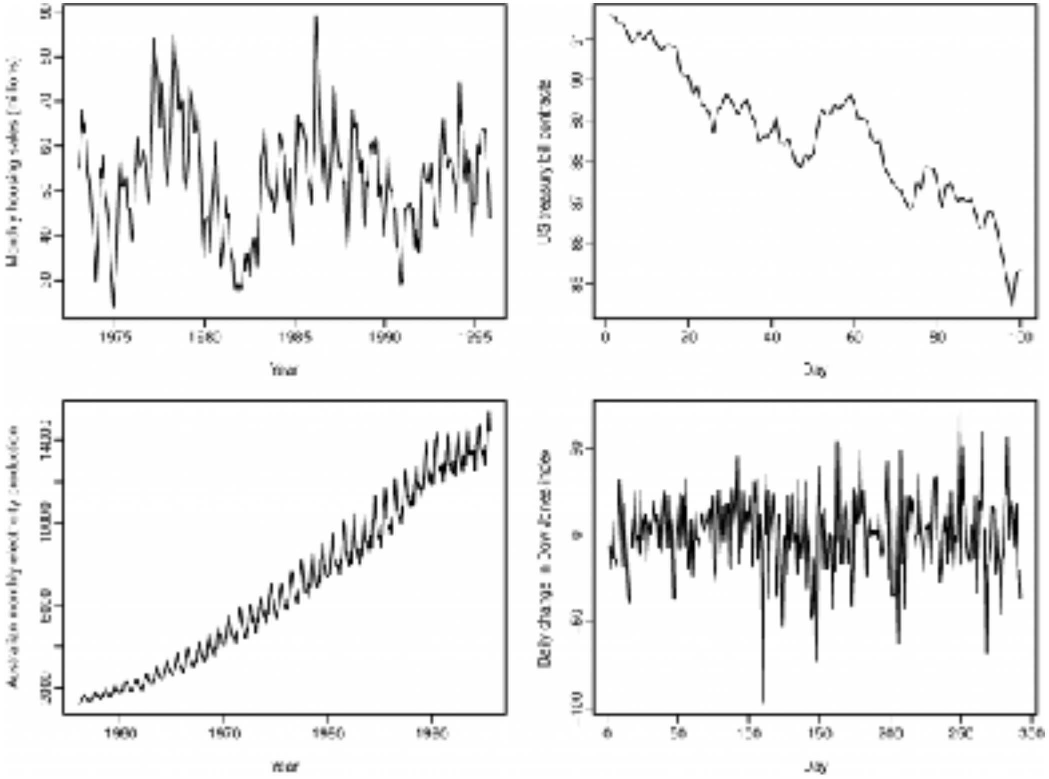
TIME SERIES COMPONENTS

Top Left: strong seasonality within each year, as well as some strong cyclic behaviour with period about 6–10 years. No Trend

Top Right: no seasonality, but an obvious downward trend. If we had more data we may be able to observe a cycle

Bottom Left: strong increasing trend, with strong seasonality. No cycle

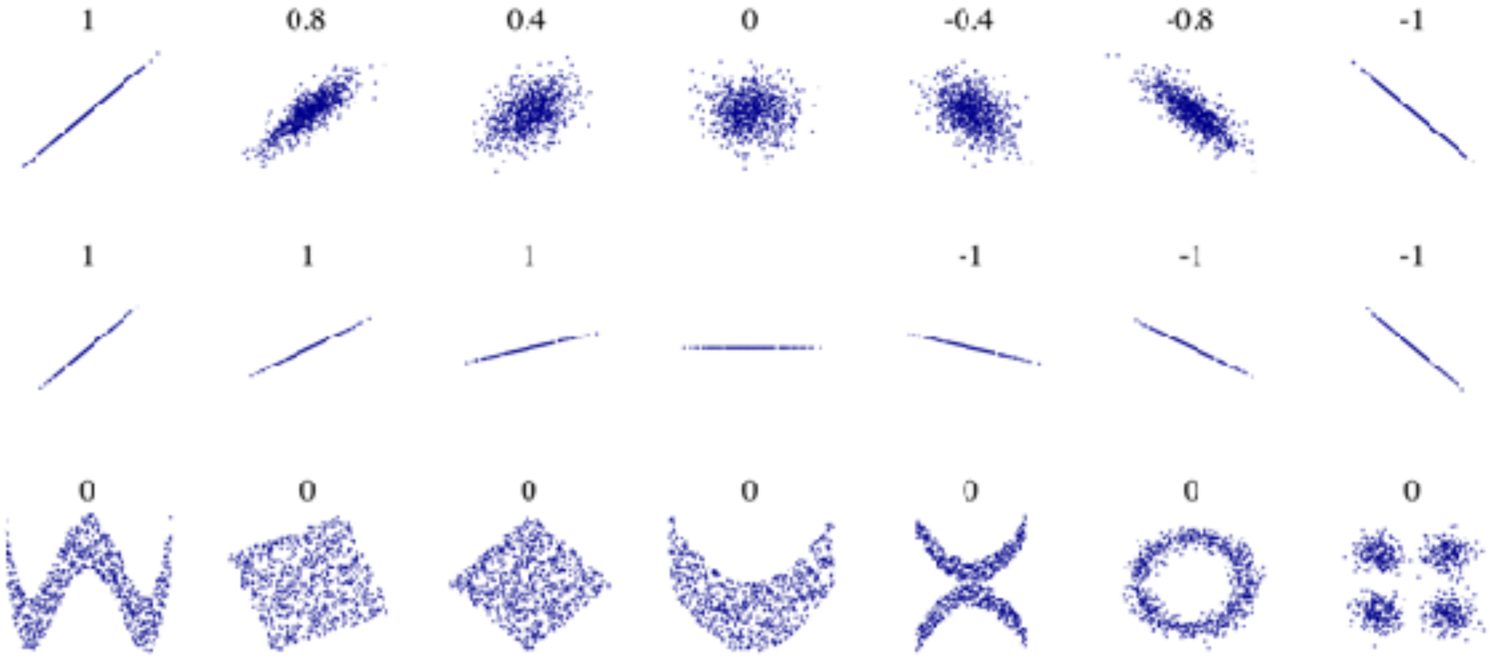
Bottom Right: no trend, seasonality or cyclic behaviour



DATA SCIENCE PART TIME COURSE

CAUSALITY

WHAT IS CORRELATION

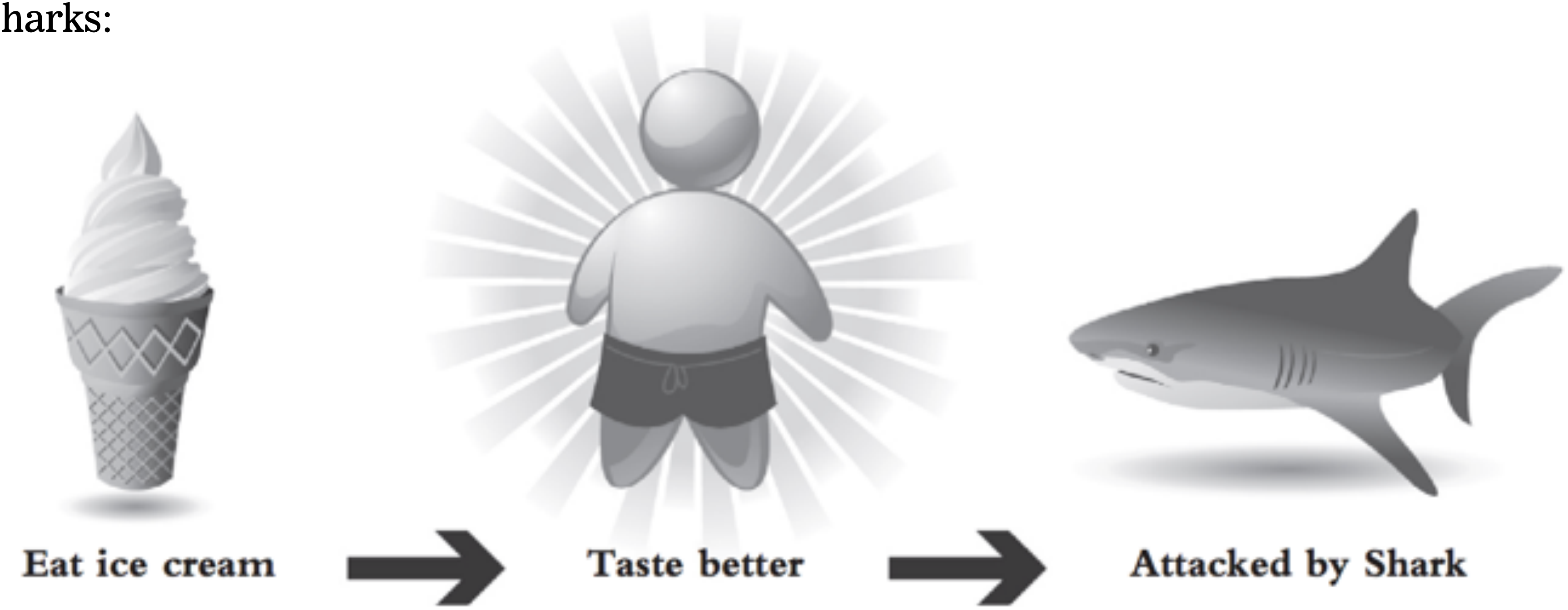


CORRELATION DOES NOT IMPLY CAUSATION

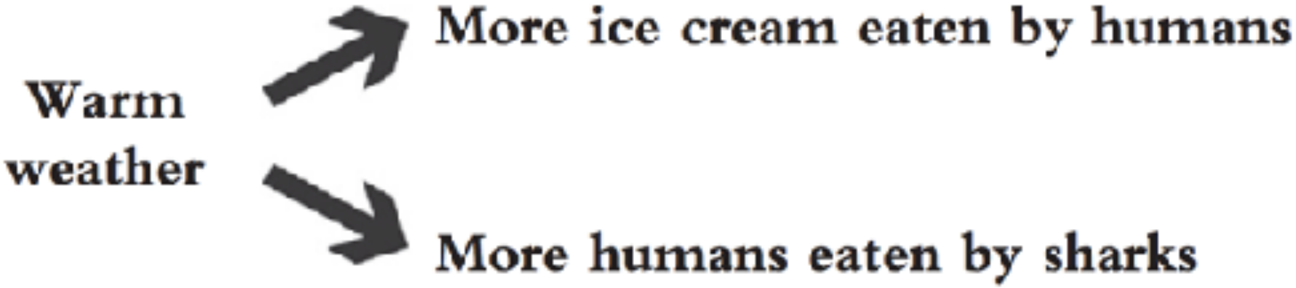
Correlation is a sign of a potential causal connection, and we can use it as a guide to further investigation (for example, trying to understand what the causal chain might be).

CAUSAL ANALYSIS

Increased ice cream sales correspond with increased shark attacks. Why do you think that is? A causal explanation could be that eating ice cream makes us taste better to sharks:

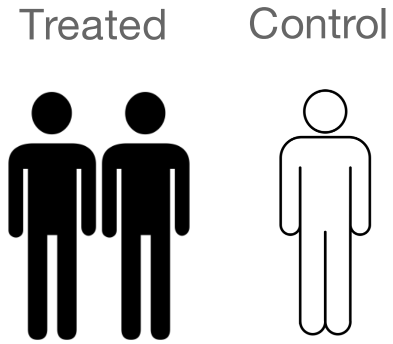


Another explanation is that, rather than one being caused by the other, they are both caused by the same thing. On cold days, people eat less ice cream and also swim less; on warm days, they do the opposite:



Randomised Clinical Trials

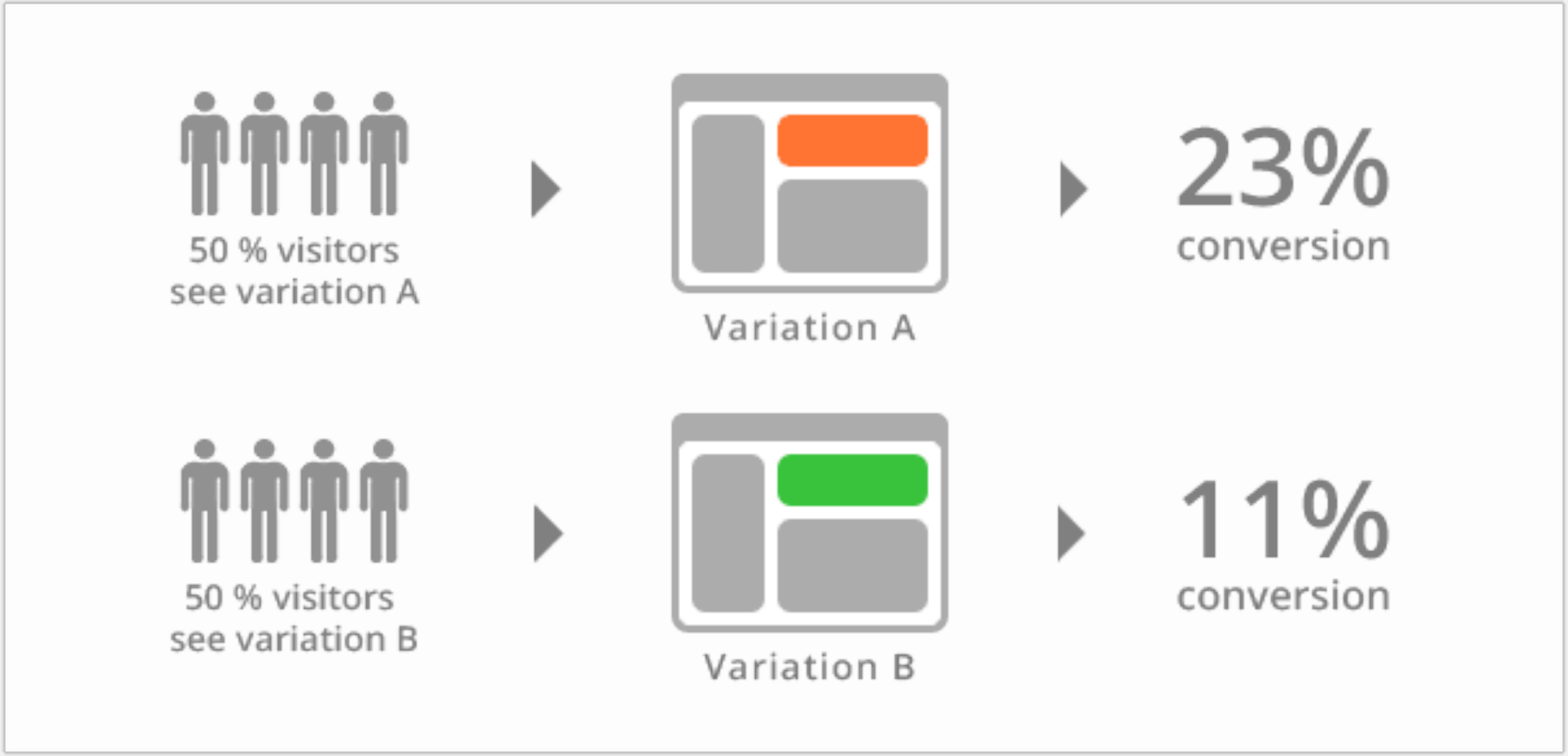
We randomly assign some group of people to receive a "treatment" and others to be in the "control" group —that is, they don't receive the treatment.



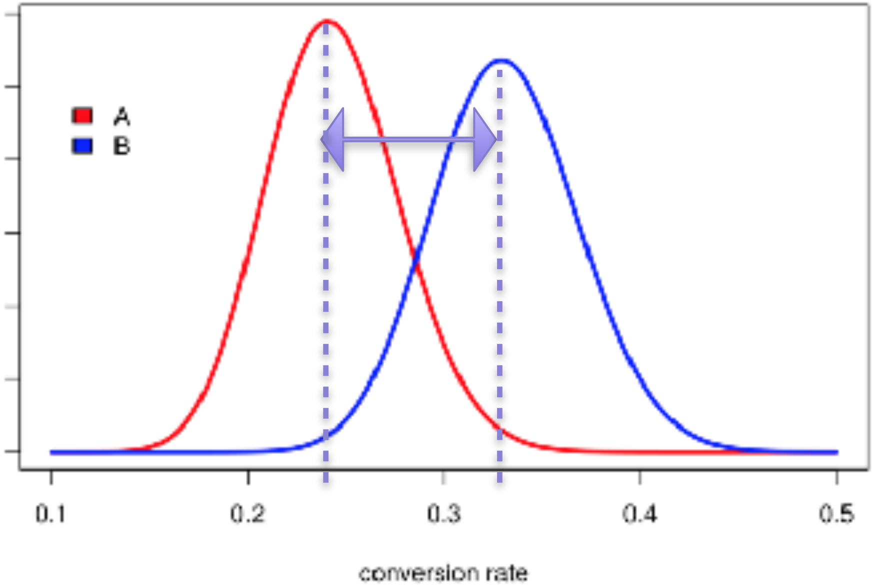
Treated Control

Response: (X + 0.5)% X%

A/B Testing


50 % visitors see variation A
Variation A
23% conversion
50 % visitors see variation B
Variation B
11% conversion

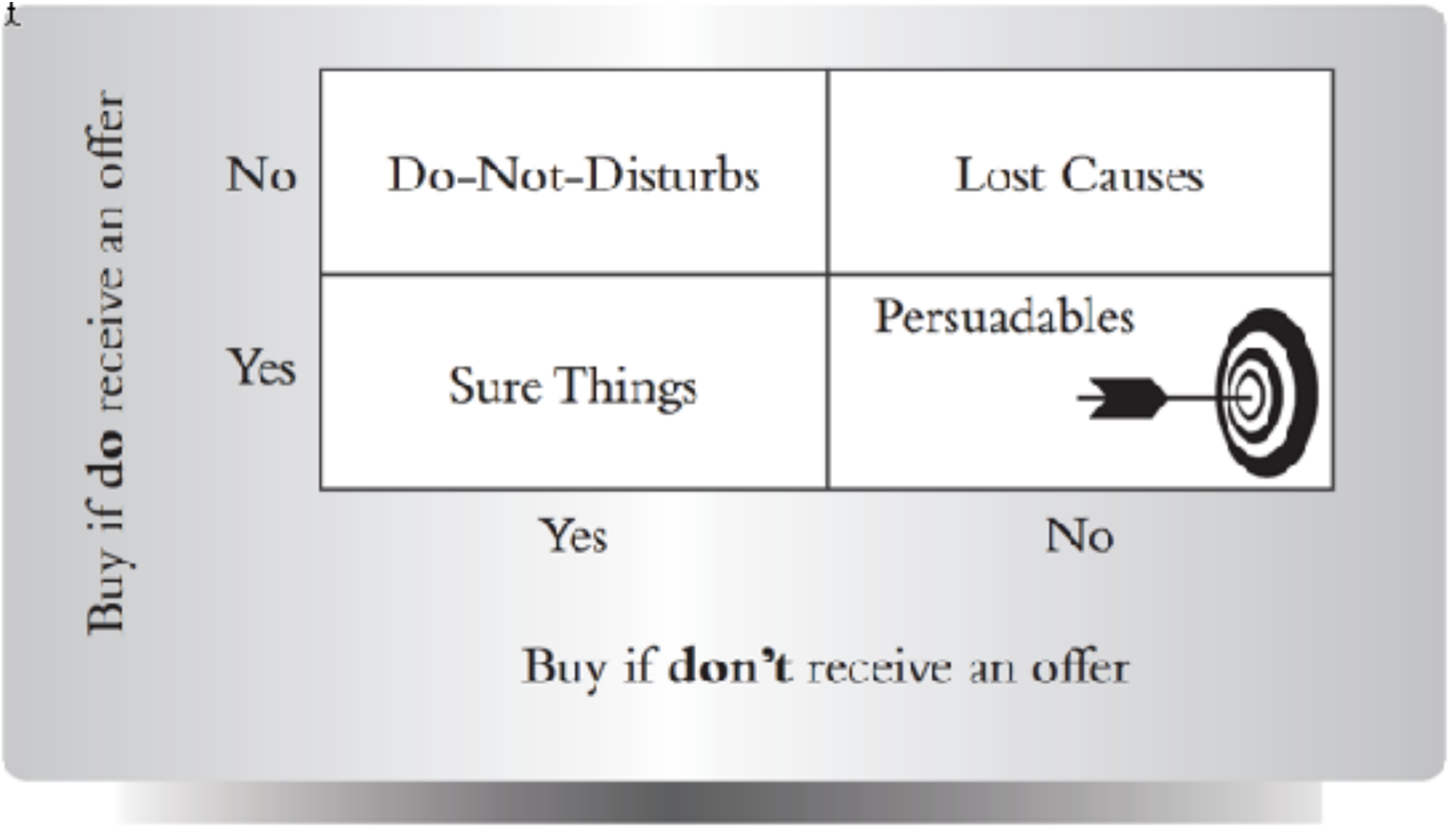
The difference in response rate between a treated group and a randomised control group



A - Treatment

B - Control

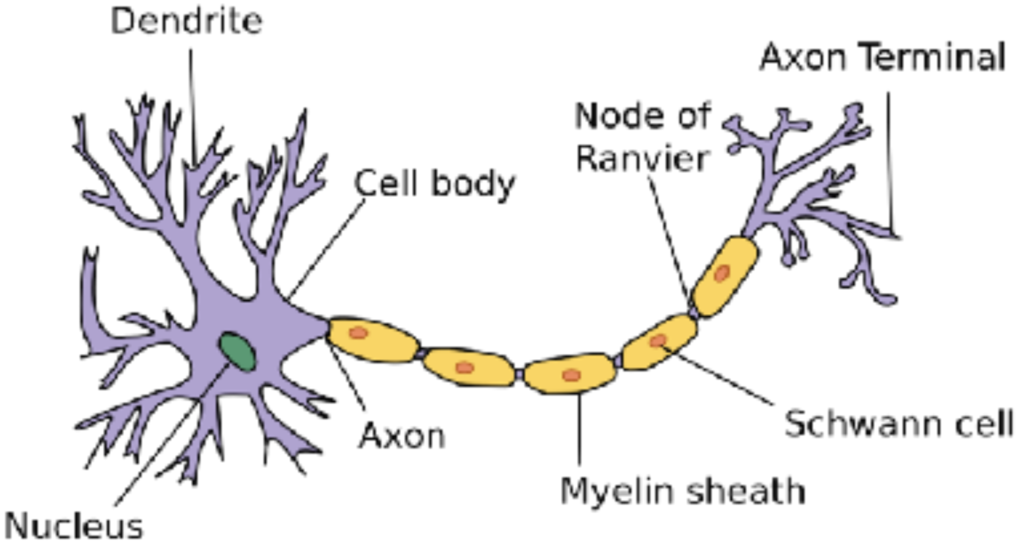
PURSUASION



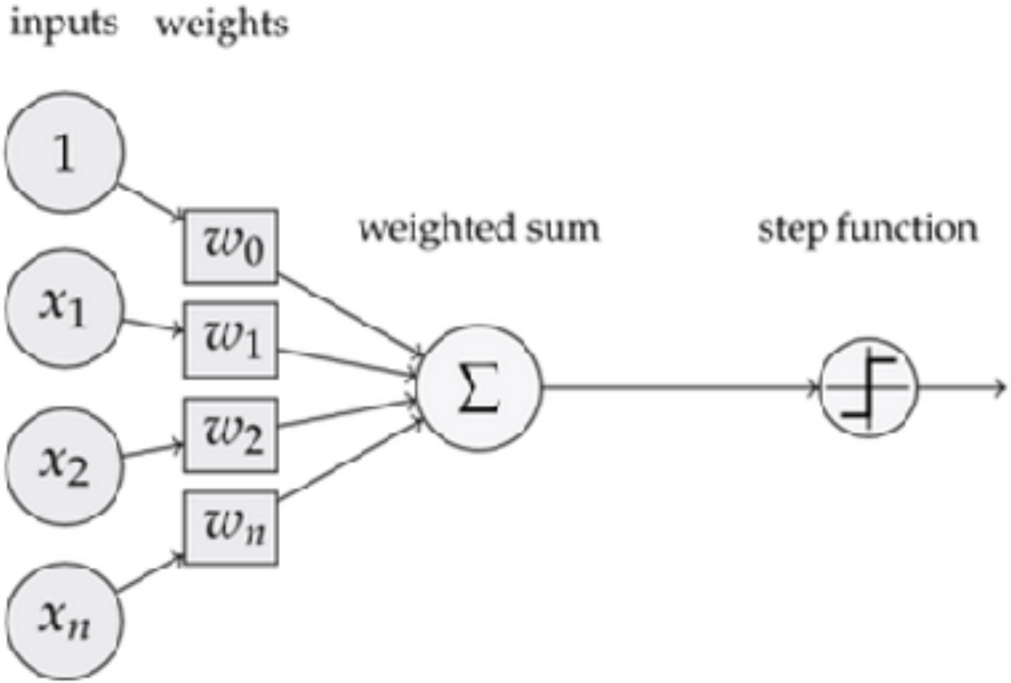
		Buy if **don't** receive an offer	
		Yes	No
Buy if **do** receive an offer	No	Do-Not-Disturbs	Lost Causes
	Yes	Sure Things	Persuadables

NEURAL NETWORKS

WHAT IS A NEURAL NETWORK?

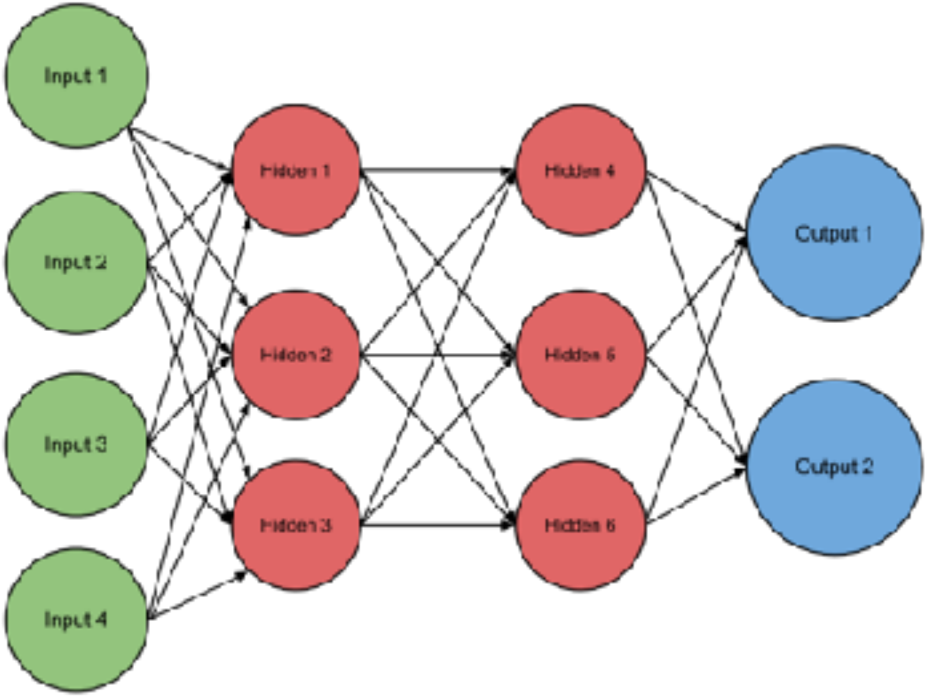


WHAT IS AN ARTIFICIAL NEURAL NETWORK?



NETWORK STRUCTURE

Hidden layers often have fewer neurons than the input layer to force the network to learn compressed representations of the original input.



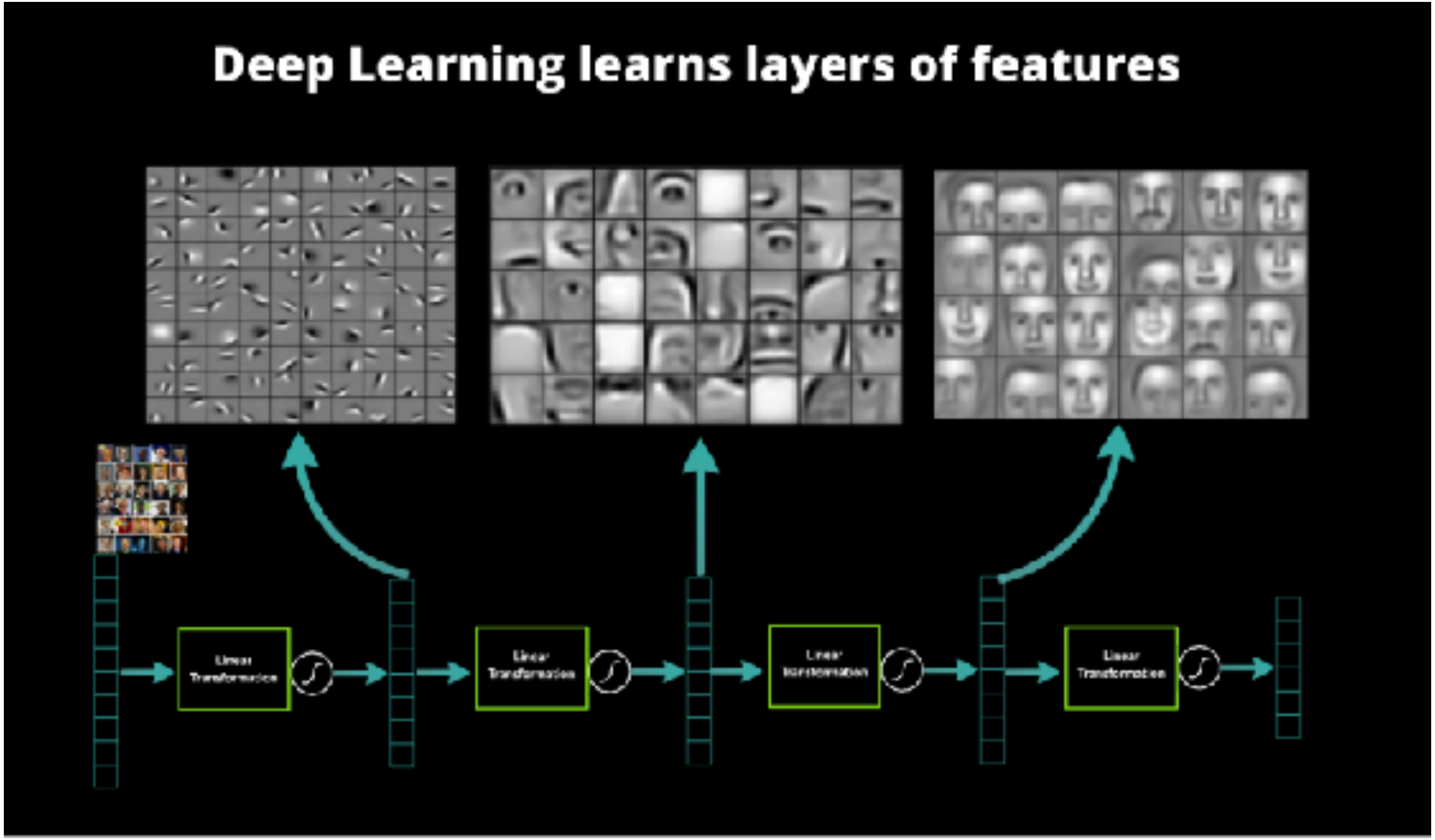
BACK-PROPOGATION

Back-Propogation is a two-pass algorithm.

The Forward pass fixes the current weights and the predicted values are calculated.

The Backward pass calculates the errors on the output layer and are then back-propagated to give the errors at the hidden layer units.

DEEP LEARNING



Deep Learning learns layers of features

When is deep learning the right choice of algorithm?

When the input features are dense:

- Images
- Videos
- Audio
- Text

Some recent interesting applications:

- Colorisation of Black and White Images.
- Adding Sounds To Silent Movies.
- Automatic Machine Translation.
- Object Classification in Photographs.
- Automatic Handwriting Generation.
- Character Text Generation.
- Image Caption Generation.
- Automatic Game Playing.

ARTIFICIAL NEURAL NETWORK TECHNOLOGIES

Caffe

torch

theano

TensorFlow™

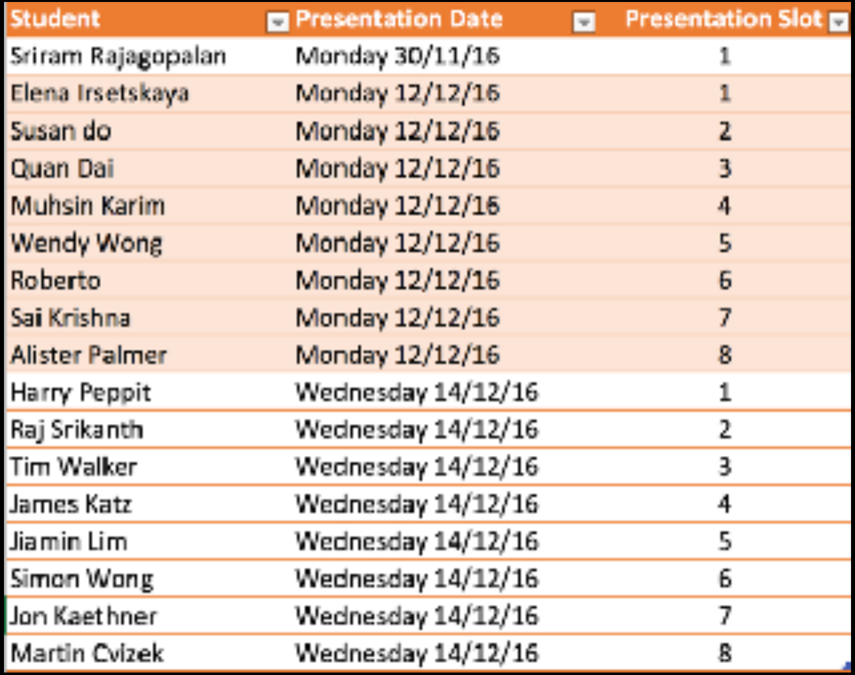
TensorFlow™ is an open source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them.

PRESENTATIONS

- 10 mins presentation with 5 mins for questions
  - What did you do?
  - What were the results?
  - What did you achieve?
  - What did you learn?
  - What else will you try in the future?
  - Appendix with any interesting findings
- On your own laptop or mine, your choice

DATA SCIENCE

PRESENTATIONS



Student	Presentation Date	Presentation Slot
Sriram Rajagopalan	Monday 30/11/16	1
Elena Irsetskaya	Monday 12/12/16	1
Susan do	Monday 12/12/16	2
Quan Dai	Monday 12/12/16	3
Muhsin Karim	Monday 12/12/16	4
Wendy Wong	Monday 12/12/16	5
Roberto	Monday 12/12/16	6
Sai Krishna	Monday 12/12/16	7
Alister Palmer	Monday 12/12/16	8
Harry Peppit	Wednesday 14/12/16	1
Raj Srikanth	Wednesday 14/12/16	2
Tim Walker	Wednesday 14/12/16	3
James Katz	Wednesday 14/12/16	4
Jiamin Lim	Wednesday 14/12/16	5
Simon Wong	Wednesday 14/12/16	6
Jon Kaethner	Wednesday 14/12/16	7
Martin Cvizek	Wednesday 14/12/16	8