GENERAL ASSEMBLY

DATA SCIENCE

11 WEEK PART TIME COURSE

Week 11 - The Home Stretch (Review)
Monday 30th May 2016

1. Marc Frankel - Guest Speaker
2. Course Review - Part 2
    1. Random Forests
    2. Cloud Computing
    3. Natural Language Processing
    4. Time Series
    5. Graphs & Network Analysis
    6. Neural Networks
    7. Communication
3. The Future
4. Presentations

Date	Week	Theme	Topics
Monday, March 21, 2016	1	Foundations	Introduction+Basics
Wednesday, March 23, 2016	1	Foundations	Basics
Easter Break			
Wednesday, March 30, 2016	2	Foundations	Visualisation
Monday, April 4, 2016	3	Foundations	Linear Regression
Wednesday, April 6, 2016	3	Foundations	Logistic Regression
Monday, April 11, 2016	4	Foundations	Model Evaluaiton
Wednesday, April 13, 2016	4	Intermediate	Regularisation
Monday, April 18, 2016	5	Intermediate	Clustering
Wednesday, April 20, 2016	5	Intermediate	Recomendations & Associations
Anzac Day			
Wednesday, April 27, 2016	6	Intermediate	Dimensionality Reduction
Monday, May 2, 2016	7	Intermediate	Decision Trees
Wednesday, May 4, 2016	7	Intermediate	Random Forests & Ensembling
Saturday, May 7, 2016	8	Review	Review, AWS, Text Analytics, Projects
Monday, May 9, 2016	8	Practical	Cloud Computing
Wednesday, May 11, 2016	8	Advanced	Natural Language Processing & APIs
Monday, May 16, 2016	9	Advanced	Time Series
Wednesday, May 18, 2016	9	Advanced	Less Technical Skills - Communication
Monday, May 23, 2016	10	Advanced	Neural Networks & Deep Learning
Wednesday, May 25, 2016	10	Advanced	Graphs & Network Analysis

DATA SCIENCE PART TIME COURSE

RANDOM FORESTS

Random Forests is a slight variation of bagged trees that has even better performance! Here's how it works:

- Exactly like bagging, we create an ensemble of decision trees using bootstrapped samples of the training set.
- However, when building each tree, each time a split is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is only allowed to use one of those m predictors.

However, when building each tree, each time a split is considered, a

random sample of m predictors

is chosen as split candidates from the

full set of p predictors.

The split is only allowed to use one of those

m predictors.

Notes:

- A new random sample of predictors is chosen for every single tree at every single split.
- For classification, m is typically chosen to be the square root of p. For regression, m is typically chosen to be somewhere between p/3 and p.

What's the point?

- Suppose there is one very strong predictor in the data set. When using bagged trees, most of the trees will use that predictor as the top split, resulting in an ensemble of similar trees that are "highly correlated".
- Averaging highly correlated quantities does not significantly reduce variance (which is the entire goal of bagging).
- By randomly leaving out candidate predictors from each split, Random Forests "decorrelates" the trees, such that the averaging process can reduce the variance of the resulting model.

Although bagging increases predictive accuracy, it decreases model interpretability because it's no longer possible to visualize the tree to understand the importance of each variable.

- To compute variable importance for bagged regression trees, we can calculate the total amount that the mean squared error is decreased due to splits over a given predictor, averaged over all trees.
- A similar process is used for bagged classification trees, except we use the Gini index instead of the mean squared error.

OUT-OF-BAG ERROR

Bagged models have a very nice property: out-of-sample error can be estimated without using the test set approach or cross-validation. How it works:

- On average, each bagged tree uses about two-thirds of the observations. For each tree, the remaining observations are called "out-of-bag" observations.
- For the first observation in the training data, predict its response using only the trees in which that observation was out-of-bag. Average those predictions (for regression) or take a majority vote (for classification).
- Repeat this process for every observation in the training data.
- Compare all predictions to the actual responses in order to compute a mean squared error or classification error. This is known as the out-of-bag error.

DATA SCIENCE PART TIME COURSE

CLOUD COMPUTING

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

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

NoSQL

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

JSON – WHAT IS IT?

JSON - JavaScript Object Notation

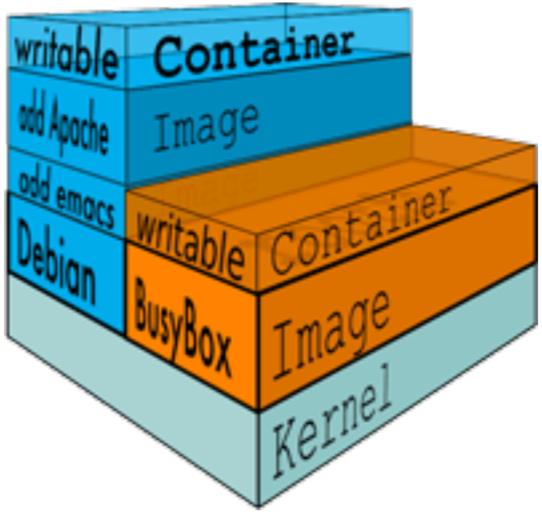
- Human readable data with attribute-value pairs.
- What is inside the curly brackets is an object
- In the object we declare variables with 'attribute' : 'value' pairs

```
var json = {
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "34 York St",
    "city": "Sydney",
    "state": "NSW",
    "postalCode": "2000"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "02 95999999"
    },
    {
      "type": "office",
      "number": "0431 111 111"
    }
  ],
  "children": [],
  "spouse": null
}
```

JSON – HOW DOES IT RELATE TO DATA SCIENCE?

- Webservices provide application programming interfaces (APIs) are now usually transferring data via JSON
- Underlying document databases like MongoDB
- Increasingly common data format

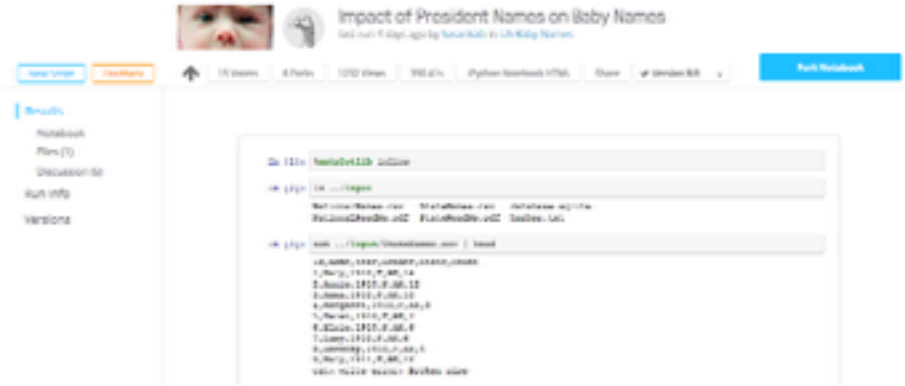
Docker containers wrap up a piece of software in a complete filesystem that contains everything it needs to run: code, runtime, system tools, system libraries – anything you can install on a server. This guarantees that it will always run the same, regardless of the environment it is running in.



- Lightweight
- Open
- Secure

DOCKER – HOW DOES IT RELATE TO DATA SCIENCE?

- Installing data science software can be a pain because of software dependencies and different OS environments. Docker helps solve this problem
- See Kaggle Scripts

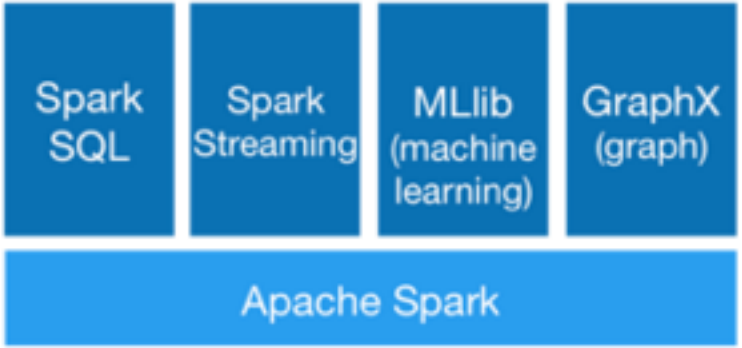


Amazon EC2

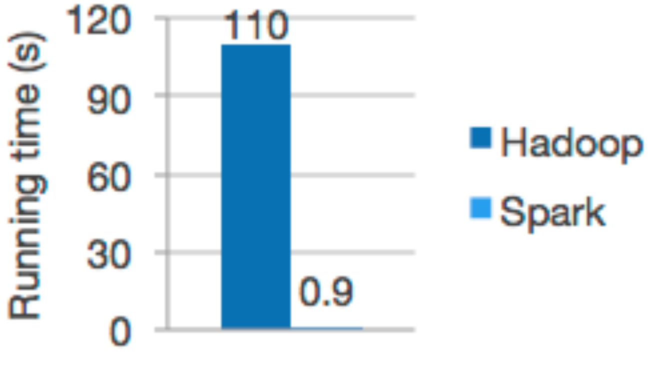
Spark

jupyter

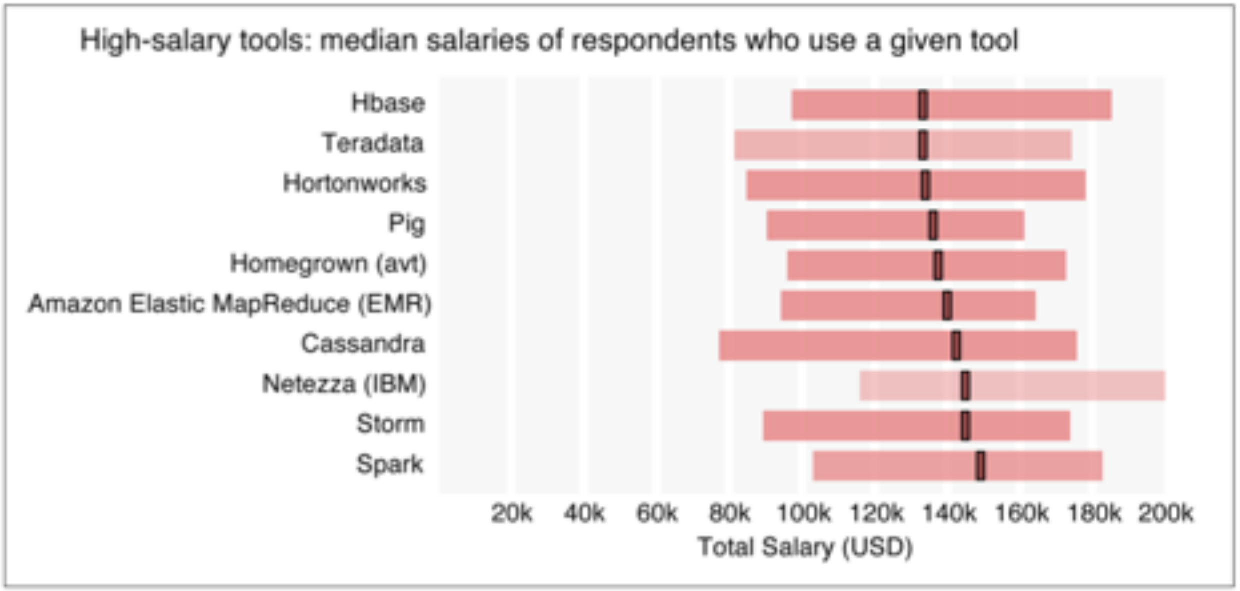
Spark is a fast and general processing engine compatible with Hadoop data. It can process data in HDFS, HBase, Cassandra, Hive, and any Hadoop InputFormat. It is designed to perform both batch processing (similar to MapReduce) and new workloads like streaming, interactive queries, and machine learning.



- MLlib is Spark's machine learning library. Its goal is to make practical machine learning scalable and easy. It consists of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as lower-level optimization primitives and higher-level pipeline APIs.
- GraphX in Spark for graphs and graph-parallel computation



Logistic regression in Hadoop and Spark



‘We can talk, but money talks, so talk more bucks’ - Jay-Z (Izzo - The Blueprint)

Spark revolves around the concept of a resilient distributed dataset (RDD), which is a fault-tolerant collection of elements that can be operated on in parallel.

There are two ways to create RDDs:

1. Parallelizing an existing collection in your driver program
2. Referencing a dataset in an external storage system, such as a shared filesystem, HDFS, HBase, or any data source offering a Hadoop InputFormat

Data types

Basic statistics

- summary statistics
- correlations
- stratified sampling
- hypothesis testing
- streaming significance testing
- random data generation

Classification and regression

- linear models (SVMs, logistic regression, linear regression)
- naive Bayes
- decision trees
- ensembles of trees (Random Forests and Gradient-Boosted Trees)
- isotonic regression

Collaborative filtering

- alternating least squares (ALS)

Clustering

- k-means
- Gaussian mixture
- power iteration clustering (PIC)
- latent Dirichlet allocation (LDA)
- bisecting k-means
- streaming k-means

Dimensionality reduction

- singular value decomposition (SVD)
- principal component analysis (PCA)

Feature extraction and transformation

Frequent pattern mining

- FP-growth
- association rules
- PrefixSpan

Evaluation metrics

PMML model export

Optimization (developer)

Big Data Analysis with Apache Spark

Learn how to apply data science techniques using parallel programming in Apache Spark to explore big data.

Berkeley
UNIVERSITY OF CALIFORNIA

Starts on August 10, 2016

Enroll Now

I would like to receive email from Berkeley and learn about its other programs.

About this course

0 Reviews 0/5

Organizations use their data to support and influence decisions and build data-intensive products and services, such as recommendation, prediction, and diagnostic systems. The collection of skills required by organizations to support these functions has been grouped under the term 'data science'.

This statistics and data analysis course will attempt to articulate the expected output of data scientists

See more

Sponsored by databricks

The production of this course would not have been possible without the generous contribution of Databricks

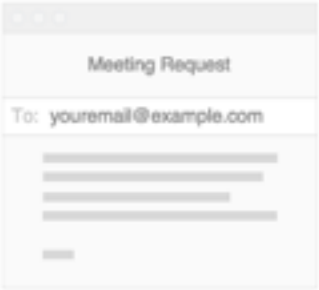
What you'll learn

Length:	4 weeks
Effort:	5-10 hours per week
Price:	FREE Add a Verified Certificate for $99
Institution:	UC BerkeleyX
Subject:	Data Analysis & Statistics
Level:	Intermediate
Languages:	English

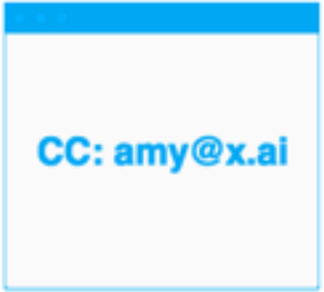
DATA SCIENCE PART TIME COURSE

NATURAL LANGUAGE PROCESSING AND APIs

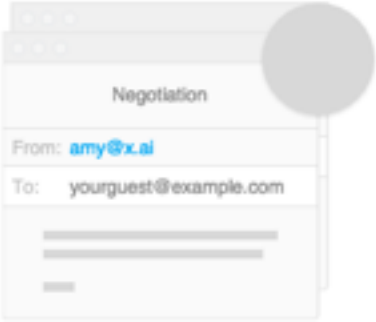
- Text is considered to be un-structured data. This means we don't have nice features we can use as inputs. We will have to construct them using a model or rules we know about language.
- Natural Language Processing is the algorithms and processing we program to interpret human language.
- It allows 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.



You receive a meeting request, but don't want to deal with the back and forth to get it scheduled



You Cc: Amy, handing the job over to her

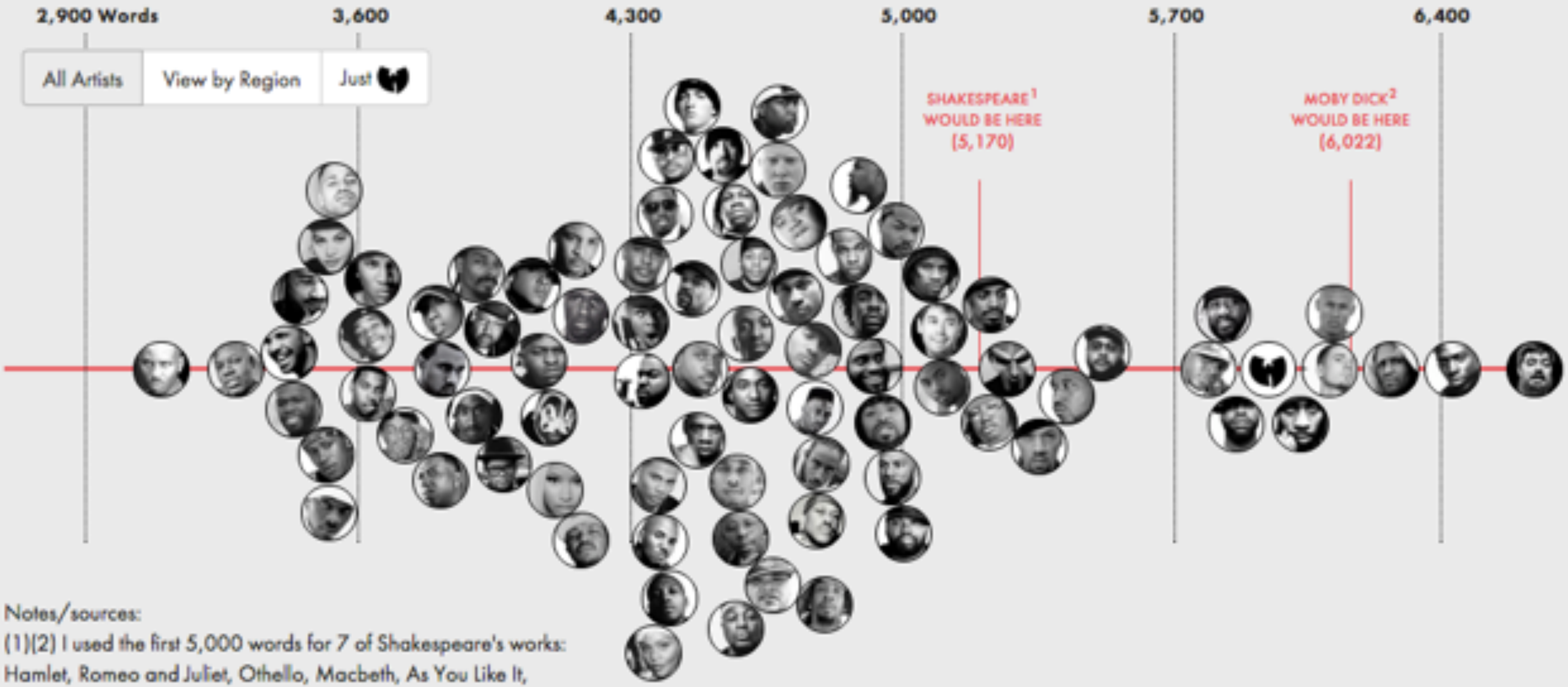


Amy emails with your guest to find the best time and location, knowing your schedule and preferences

Rank	Word Network Feature	Information Gain
1	Term frequency of the word *until*	0.621
2	Neighborhood size of the word *until*	0.611
3	Degree of the word *until*	0.610
4	Neighborhood size of the word *by*	0.576
5	Term frequency of the word *several*	0.574
6	Term frequency of the word *thus*	0.555
7	Degree of the word *thus*	0.553
8	Degree of the word *several*	0.544
9	Neighborhood size of the word *several*	0.543
10	Coreness of the word *thus*	0.538
11	Neighborhood size of the word *though*	0.524
12	Term frequency of the word *had*	0.509
13	Term frequency of the word *by*	0.507
14	Neighborhood size of the word *may*	0.505
15	Degree of the word *or*	0.499
16	Clustering coefficient of the word *said*	0.497
17	Coreness of the word *upon*	0.489
18	Coreness of the word *whom*	0.489
19	Degree of the word *by*	0.488
20	Neighborhood size of the word *returned*	0.484

Table 8: Ranking of term frequency and local word network features based on Information Gain, on Gutenberg data. We took 500 most frequent words on the whole dataset, and collected their term frequency, clustering coefficient, neighborhood size, coreness and vertex degree (for each document) in a single file. This ranking reflects the top 20 among 2,500 features in that file, along with their information gain values. Note that both term frequency as well as local word network features appeared at the top. Moreover, stopwords like *until*, *by*, *several* and *thus* are found to be important predictors of writing style.

OF UNIQUE WORDS USED WITHIN ARTIST'S FIRST 35,000 LYRICS



Notes/sources:
(1)(2) I used the first 5,000 words for 7 of Shakespeare's works: Hamlet, Romeo and Juliet, Othello, Macbeth, As You Like It, Winter's Tale, and Troilus and Cressida. For Melville, I used the first 35,000 words of Moby Dick.
All lyrics are provided by Rap Genius, but are only current to 2012. My lack of recent data prevented me from using quite a few current artists.
This data viz uses code by Amelia Bellamy-Royds's in this jsfiddle.

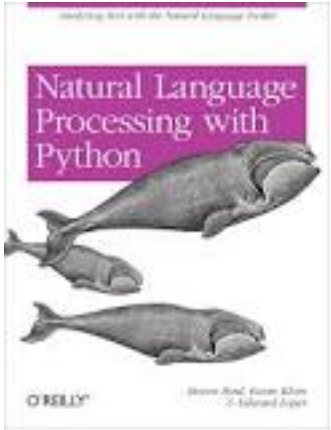
DATA SCIENCE PART TIME COURSE

TOOLS FOR 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

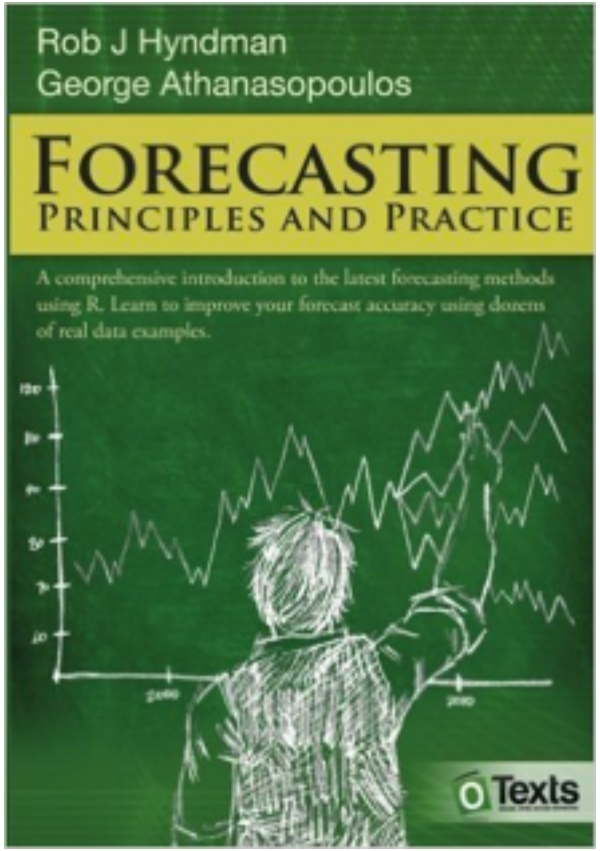
AlchemyAPI™
An IBM Company

- 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

TIME SERIES

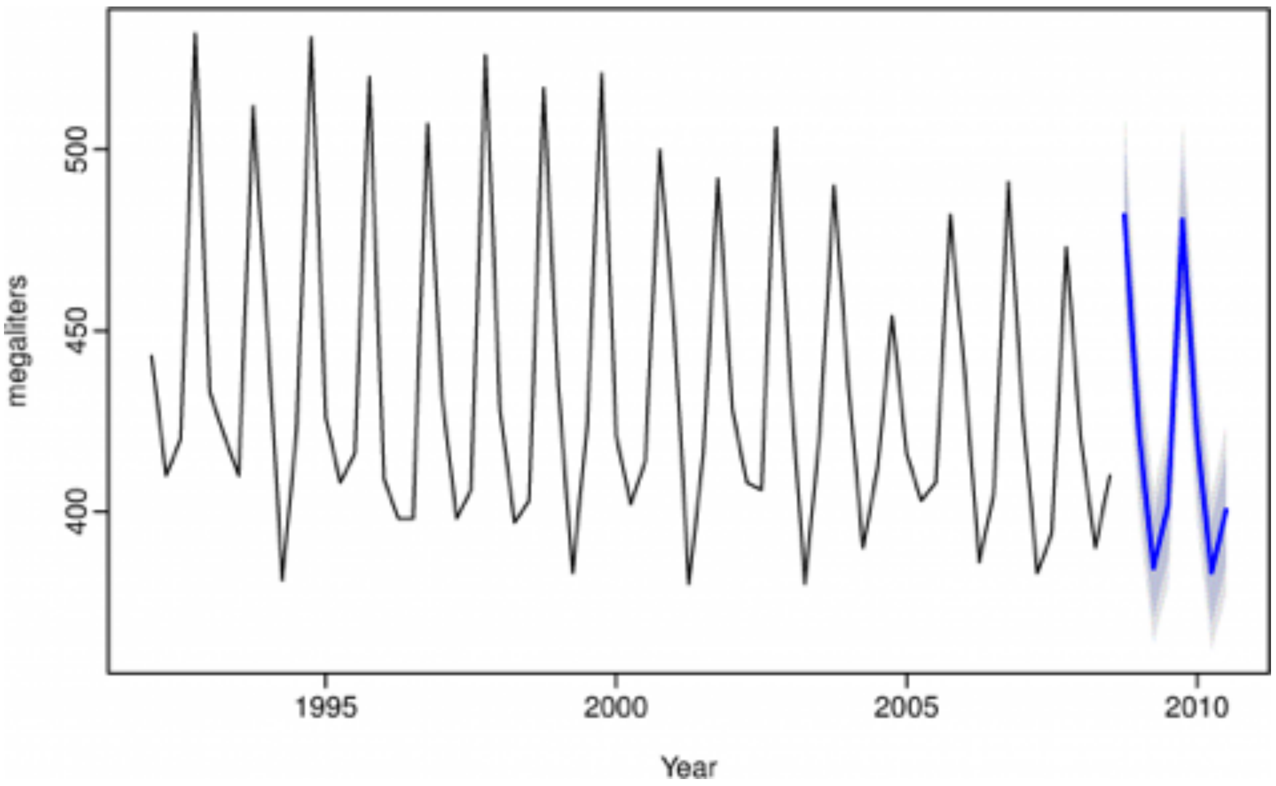

Rob J Hyndman
George Athanasopoulos
FORECASTING
PRINCIPLES AND PRACTICE
A comprehensive introduction to the latest forecasting methods using R. Learn to improve your forecast accuracy using dozens of real data examples.
OTexts

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


megaliters
500
450
400
1995
2000
2005
2010
Year

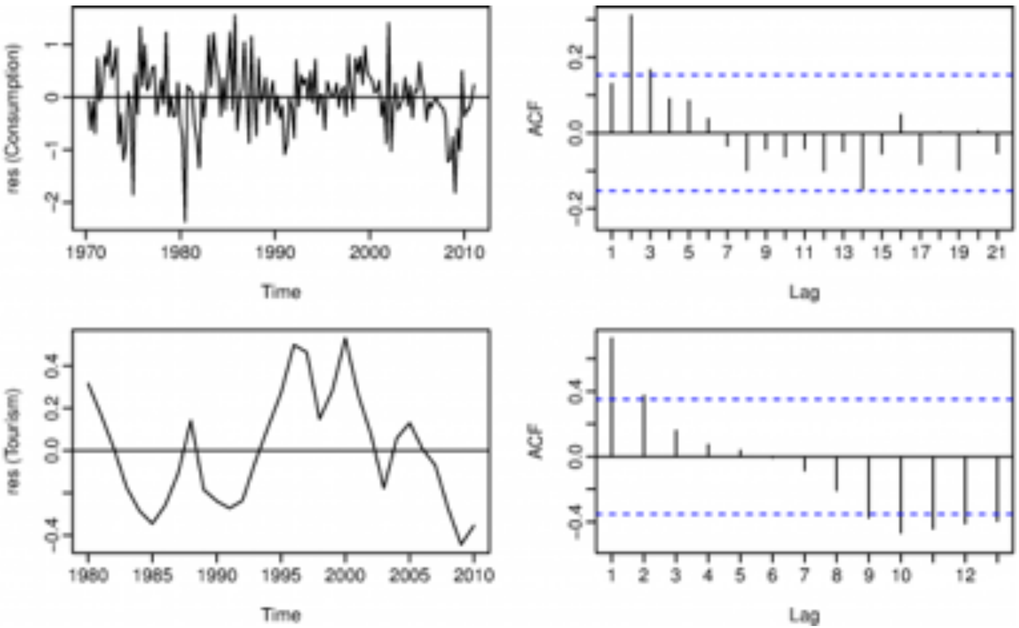
In other words, why wouldn't we just use linear regression and have the time variable as our X values?

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

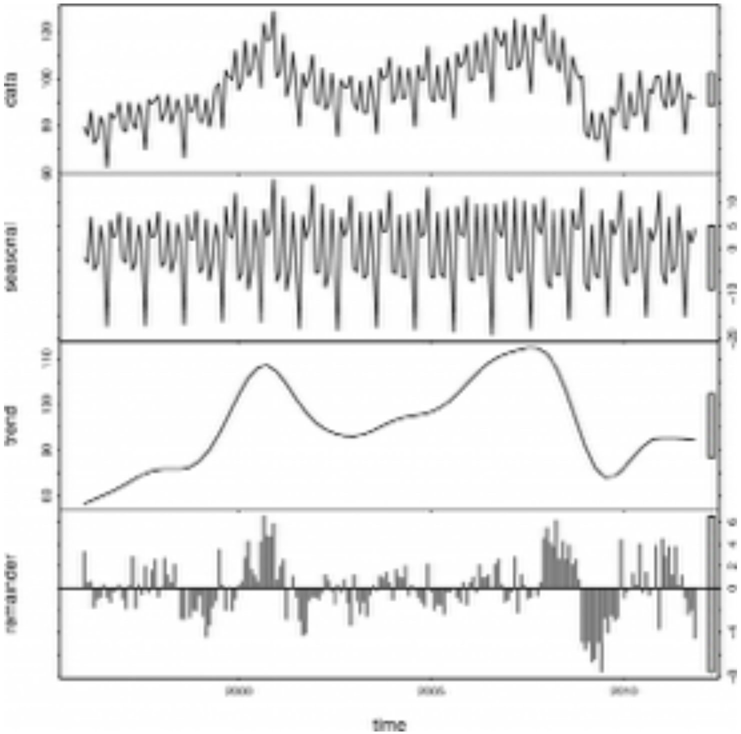
Recall some of the conditions of Linear Regression models:

- have mean zero; otherwise the forecasts will be systematically biased
- are not autocorrelated; otherwise the forecasts will be inefficient as there is more information to be exploited in the data
- are unrelated to the predictor variable; otherwise there would be more information that should be included in the systematic part of the model

With time series data it is highly likely that the value of a variable observed in the current time period will be influenced by its value in the previous period, or even the period before that, and so on…



Time Series Decomposition is a way to break down a time series into the Season, Trend (which includes the cycle) and Remainder.



We can consider these to be weighted averages of past observations. This means that the more recent the observation, the higher the weighting of that observation.

The Naive model is the case where the forecast is equal to the last observed value,

$$\hat{y}_{T+h|T} = y_T$$

What if we were to weight the observations to have decreasing weights as the observations got older? What would the equation look like?

If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model. ARIMA is an acronym for AutoRegressive Integrated Moving Average model ("integration" in this context is the reverse of differencing).

We call this an ARIMA(p,d,q) model, where

p= order of the autoregressive part;

d= degree of first differencing involved;

q= order of the moving average part.

GRAPHS & NETWORK ANALYSIS

Many types of real-world problems involve dependencies between observations.

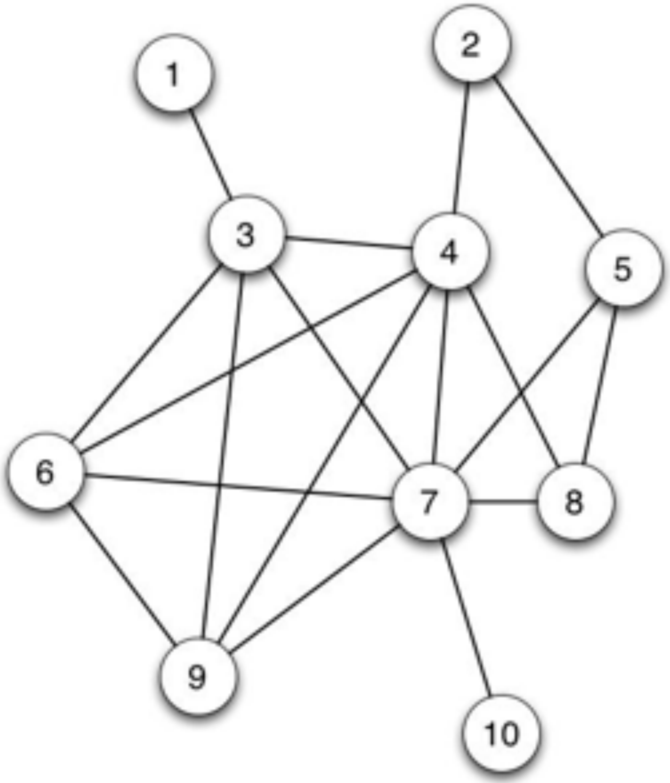
For example:

- Town planners are looking at vehicular flows through a city
- Sociologist want to understand how people influence others that they know (if at all)
- Biologists want to know how proteins regulate the actions of other proteins
- Information agencies want to discover groups of adversaries

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?



The criteria for finding good communities is similar to that for finding good clusters.

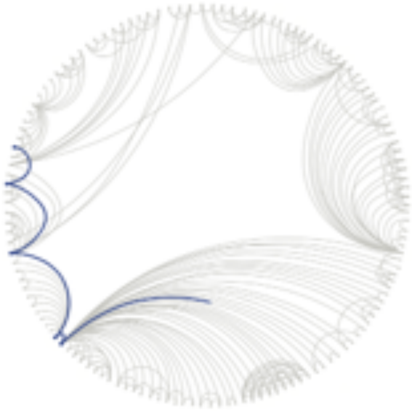
We want to maximize intra-community edges while minimizing inter-community edges.

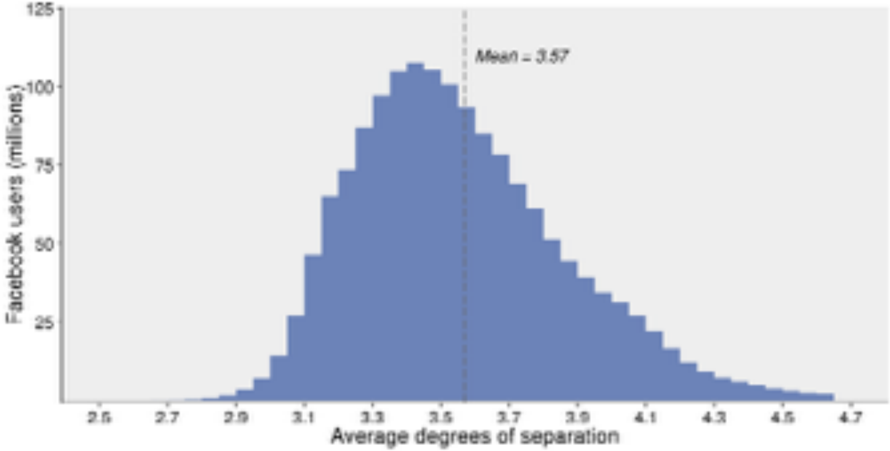
Formally, the algorithm tries to maximize the modularity of network, or the fraction of edges that fall within the community minus the expected fraction of edges if the edges were distributed by random. Good communities should have a high number of intra-community edges, so by maximizing the modularity, we detect dense communities that have a high fraction of intra-community edges.

How connected is the world?

Each person in the world (at least among the 1.59 billion people active on Facebook) is connected to every other person by an average of three and a half other people.

Rather than calculate it exactly, they estimate distances with statistical algorithms





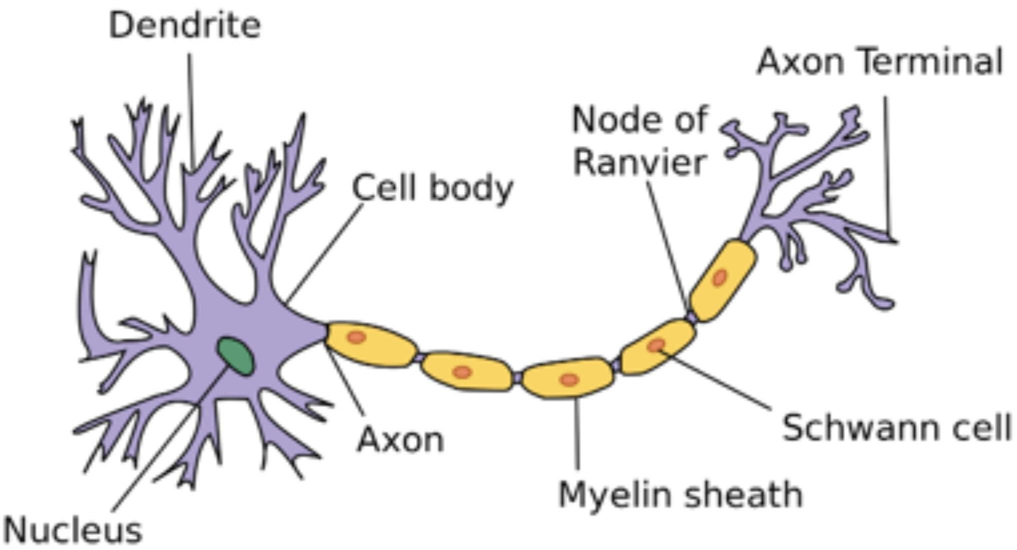
DATA SCIENCE PART TIME COURSE

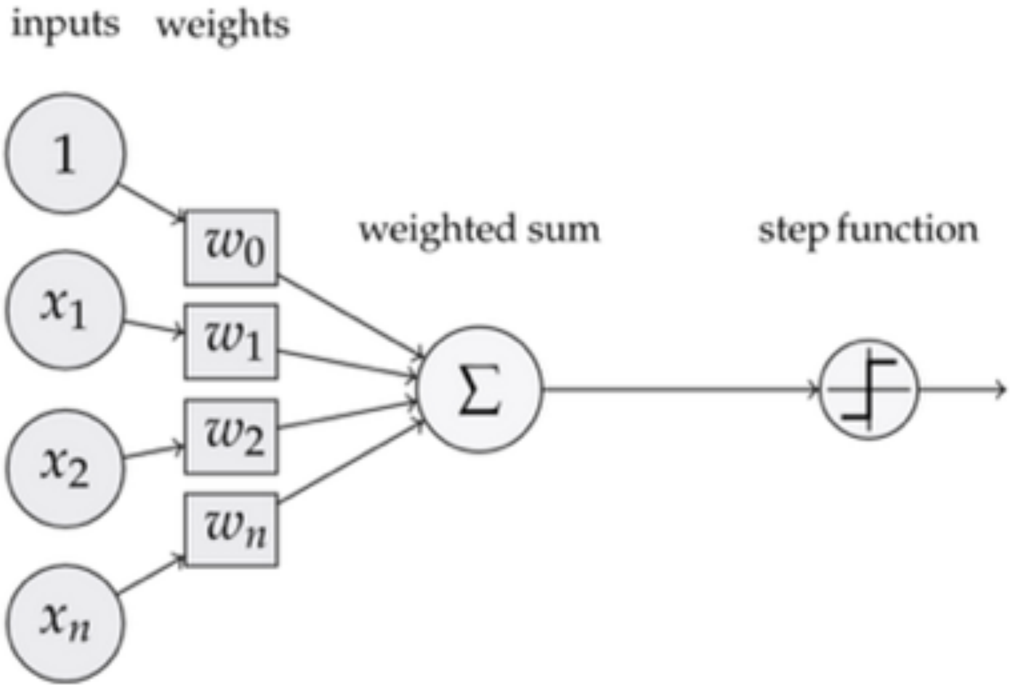
NEURAL NETWORKS

MEET THE MAN GOOGLE HIRED TO MAKE AI A REALITY

Brief Summary of the Panel Discussion at DL Workshop @ICML 2015

posted Jul 13, 2015, 5:27 AM by KyungHyun Cho [updated Jul 14, 2015, 11:04 PM]


Dendrite
Axon Terminal
Node of Ranvier
Cell body
Schwann cell
Axon
Myelin sheath
Nucleus


inputs
weights
1
x_1
x_2
x_n
w_0
w_1
w_2
w_n
weighted sum
Σ
step function

Three papers were published in 2006 that were breakthroughs for Deep Learning. They shared the following principles:

- Unsupervised learning of representations is used to (pre-)train each layer
- Unsupervised training of one layer at a time, on top of the previously trained ones. The representation learned at each level is the input for the next layer
- Use supervised training to fine-tune all the layers (in addition to one or more additional layers that are dedicated to producing predictions)

http://deepdreamgenerator.com/

COMMUNICATION

☐ EMAIL ONE OF THE GUEST PRESENTERS OF THIS CLASS

- Who's the audience?
- Clear and concise
- Know the business
- Always ask why
- What's the lever?
- Over-Communication is better than under-communication
- Listen hard
- Break bread

- How you present your work will determine if it gets implemented.
- If your work isn't implemented then it's worthless
- Start with the results and action then dive into how you got there.
- Can your audience understand what's in front of them?
- What questions do you think your audience will have?
- Peer Review

- Tailor it to the position you are applying for
- Relevant experience
- Format
  - Keep it to one page (double sided)
  - Use past tense
  - Keep any descriptions succinct
  - Avoid colour coding
  - Send it as a pdf (the best way to ensure there are no scaling issues)

- Don’t lie
- Brush up on technology you use before an interview
- Highlight the benefits of your analysis (e.g. x% reduction in customer churn which increased revenue by $y)
- What were some of the major projects you worked on?
- What was the purpose of them?
- What did you contribute?
- What technologies did you use to complete the project?

1. What was the last thing that you made for fun?

2. What's your favourite algorithm? Can you explain it to me?

3. Tell me about a data project you've done that was successful. How did you add unique value?

4. Tell me about something that failed. What would you change if you had to do it over again? ...

5. You clearly know a bit about our data and our work. When you look around, what's the first thing that comes to mind as "why haven't you done X"?! ...

THE FUTURE

POST COURSE

- Keep in touch via slack
- Provide a tailored learning pathway on what your preferences
- Provide feedback on CVs (which you can take or leave)
- Hopefully you'll get involved in the Sydney Data Science community (see the meet up channel)
- Keep working on your projects and find new ones to work on (e.g. Kaggle competitions)

PRESENTATIONS