

Discussion

Spyros Samothrakis
Research Fellow, IADS
University of Essex

March 20, 2017

About

Applications

Society

Tools

System tips

LAST LECTURE

- ▶ Revision lecture based on Lecture one
- ▶ We will go again through what we have discussed already, and put everything into context
- ▶ The Beginning Is the End Is the Beginning....

BETTER SCIENCE THROUGH DATA

Hey, Tony, Stewart Tansley, and Kristin M. Tolle. “Jim Gray on eScience: a transformed scientific method.” (2009).

- ▶ Thousand years ago: empirical branch
 - ▶ You observed stuff and you wrote down about it
- ▶ Last few hundred years: theoretical branch
 - ▶ Equations of gravity, equations of electromagnetism
- ▶ Last few decades: computational branch
 - ▶ Modelling at the micro level, observing at the macro level
- ▶ Today: data exploration
 - ▶ Let machines create models using vast amounts of data

MIXING STATISTICS, PHILOSOPHY OF SCIENCE AND MACHINE LEARNING

- ▶ Wu, C. F. J. “Statistics= data science.” (1997).
- ▶ Breiman, Leo. “Statistical modeling: The two cultures (with comments and a rejoinder by the author).” *Statistical Science* 16.3 (2001): 199-231.
- ▶ Science is the epistemology of causation
- ▶ Data science is basically science on arbitrary data
 - ▶ But quite often we only care about predictions
- ▶ Possibly a re-branding of data mining, machine learning, artificial intelligence, statistics

BETTER BUSINESS THROUGH DATA

- ▶ There was a report by Mckinsey

Manyika, James, et al. “Big data: The next frontier for innovation, competition, and productivity.” (2011).

- ▶ Urges everyone to monetise “Big Data”
- ▶ Use the data provided within your organisation to gain insights
- ▶ Has some numbers as to how much this is worth
- ▶ Proposes a number of methods, most of them associated with machine learning and databases

MORE IS DIFFERENT

- ▶ Anderson, Philip W. “More is different.” *Science* 177.4047 (1972): 393-396.
- ▶ The idea of emergence
- ▶ You put stuff together, you go from physics to chemistry
- ▶ ...from chemistry to biology
- ▶ ...from biology to psychology and zoology
- ▶ ...from psychology to sociology
- ▶ “quantity changes into quality”

IBM's INFOGRAPHIC

40 ZETTABYTES

[43 TRILLION GIGABYTES]
of data will be created by
2020, an increase of 300
times from 2005



Volume SCALE OF DATA

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day



Most companies in the
U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

The New York Stock Exchange
captures
**1 TB OF TRADE
INFORMATION**
during each trading session



By 2016, it is projected
there will be
**18.9 BILLION
NETWORK
CONNECTIONS**
— almost 2.5 connections
per person on earth

Velocity ANALYSIS OF STREAMING DATA



Modern cars have close to
100 SENSORS
that monitor items such as
fuel level and tire pressure

The FOUR V's of Big Data

From traffic patterns and music downloads to web
history and medical records, data is recorded,
stored, and analyzed to enable the technology
and services that the world relies on every day.
But what exactly is big data, and how can these
massive amounts of data be used?

As a leader in the sector, IBM data scientists
break big data into four dimensions: **Volume,**
Velocity, Variety and Veracity.

Depending on the industry and organization, big
data encompasses information from multiple
internal and external sources such as transactions,
social media, enterprise content, sensors and
mobile devices. Companies can leverage data to
adapt their products and services to better meet
customer needs, optimize operations and
infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data,
with 1.9 million in the United States.



As of 2011, the global size of
data in healthcare was
estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]



**30 BILLION
PIECES OF CONTENT**
are shared on Facebook
every month



By 2014, it's anticipated
there will be
**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**



Variety DIFFERENT FORMS OF DATA

**4 BILLION+
HOURS OF VIDEO**
are watched on
YouTube each month



400 MILLION TWEETS
are sent per day by about 200
million monthly active users



1 IN 3 BUSINESS LEADERS

don't trust the information
they use to make decisions



**27% OF
RESPONDENTS**

in one survey were unsure of
how much of their data was
inaccurate



Veracity UNCERTAINTY OF DATA

Poor data quality costs the US
economy around
\$3.1 TRILLION A YEAR



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MPTTEC, SAS

IBM

CLASSIC SCIENCE

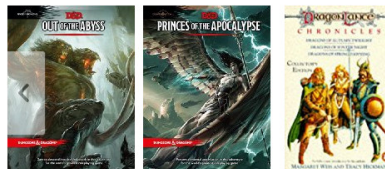
- ▶ The original data science field
- ▶ SKA (The Square Kilometer Array) ~ 4.6 EB expected (i.e. 4.6×10^6 TB), (Zhang, Yanxia, and Yongheng Zhao. “Astronomy in the Big Data Era.” *Data Science Journal* 14 (2015).)¹
- ▶ Bioinformatics
- ▶ Medical science



¹<http://datascience.codata.org/article/10.5334/dsj-2015-011>

RECOMMENDER SYSTEMS

- ▶ One of the most popular applications of data science
- ▶ Propose products to customers based on past history
- ▶ Almost all online vendors do it
- ▶ Made popular by the Netflix prize

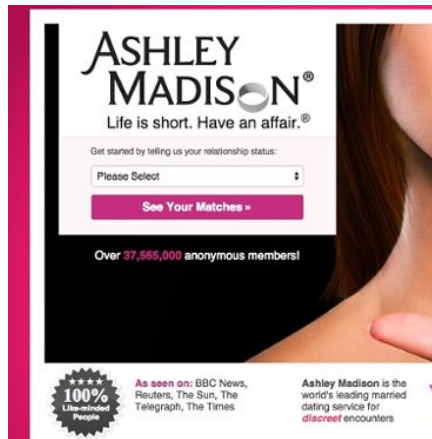


Digital Cameras best sellers [See more](#)



DATA JOURNALISM

- ▶ Wikileak style data dumps are everywhere
- ▶ The Ashley-Madison Affair, 2015
- ▶ “Just three in every 10,000 female accounts on infidelity website are real”
- ▶ “The website claims 5.5 million of its 37 million accounts are ‘female’ ”



²<http://www.independent.co.uk/life-style/gadgets-and-tech/news/ashley-madison-hack-just-three-in-every-10000-female-accounts-on-infidelity-website-are-real-10475310.html>

FINANCE & INSURANCE

- ▶ Predict stock prices (Hedge Funds)
- ▶ Insurance models
- ▶ Credit score
- ▶ In fact, a lot of trading that currently happens is algorithmic trading²
- ▶ Sudden drops in share prices often caused by defective algorithms



³<http://www.bbc.com/news/business-34264380>

POLITICS (CURRENT)

“...This included a) integrating data from social media, online advertising, websites, apps, canvassing, direct mail, polls, online fundraising, activist feedback, and some new things we tried such as a new way to do polling (about which I will write another time) and b) having experts in physics and machine learning do proper data science in the way only they can – i.e. far beyond the normal skills applied in political campaigns...”

Dominic Cummings's (Head of *Vote Leave*) Blog³

⁴<https://dominiccummings.wordpress.com/2016/10/29/on-the-referendum-20-the-campaign-physics-and-data-science-vote-leaves-voter-intention-collection-system-vics-now-available-for-all/>

POLITICS (HISTORICAL)

- ▶ New Yorker - THE PLANNING MACHINE: Project Cybersyn and the origins of the Big Data nation⁴
- ▶ Cybersyn / Chile during Alliente's rule, co-designed by Stafford Beer
- ▶ Plan was to use data fed directly from each industry to automate production



⁵<http://www.newyorker.com/magazine/2014/10/13/planning-machine>

QUESTION ANSWERING

- ▶ e.g. Antol, Stanislaw, et al. “VQA: Visual question answering.” Proceedings of the IEEE International Conference on Computer Vision. 2015.⁶
- ▶ Input can be videos, websites, et
- ▶ Think google



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

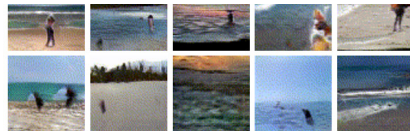
⁶http://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf

DIGITAL MARKETING

- ▶ Is a new product I just created well received by our customers?
- ▶ Is a new marketing campaign e-mail sent detrimental to our efforts?
- ▶ What is the content a chain of e-mails should have?
- ▶ Customer segmentation
- ▶ What adverts should I present to a user?

CREATIVE ARTIFICIAL INTELLIGENCE (RECIPES, MUSIC, ART, TEXT)

- ▶ e.g. Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. “Generating videos with scene dynamics.” Advances In Neural Information Processing Systems. 2016.⁶
- ▶ Generate an artefact
 - ▶ Generate videos
 - ▶ Generate text
 - ▶ Generate music



Train Station



⁶http://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf

GAME PLAYING

- ▶ We recently have seen a resurgence of game playing machines
- ▶ A computer GO programme finally outperformed top humans (AlphaGO)
- ▶ No-limit heads up poker (matches still played as we speak!)
- ▶ New labs are opening from major game companies dealing with game AI
- ▶ Though directly related, game analytics

ARTIFICIAL INTELLIGENCE

- ▶ Everything we have seen so far are basically applications of Artificial Intelligence and Machine Learning
- ▶ Inductive reasoning from a limited amount of examples
 - ▶ Structured learning
 - ▶ One-shot models
- ▶ Deductive reasoning
 - ▶ From concepts to data
 - ▶ Platonic forms

SOME SAMPLE DATA

- ▶ `takes_off_road`: owner takes the vehicle off road
- ▶ `company_vehicle`: it belongs to a business
- ▶ `is_over_30`: age of vehicle is over 30
- ▶ `regular_service`: is the vehicle serviced regularly?
- ▶ `brake_down`: will it break down within three months of our inspection date?

<code>takes_off_road</code>	<code>company_vehicle</code>	<code>is_over_30</code>	<code>regular_service</code>	<code>brake_down</code>
0	1	1	0	1
0	0	1	1	0
1	1	1	1	1
0	1	1	0	1
0	0	1	0	0
0	1	0	0	0
1	0	0	1	0
1	1	1	1	1
1	0	0	1	1
0	1	1	0	1
1	0	0	1	0
1	1	0	0	0
0	0	0	0	0

PREDICTIONS

- ▶ The most common data science operation
- ▶ Can you predict if a car will break down given the data, and if yes with what probability?
- ▶ Can you learn a model, that if provided with a tuple $\langle takes_off_road, company_vehicle, is_over_30, regular_service \rangle$ predict $break_down$?
- ▶ The tuple represents a vehicle
- ▶ Columns are called *features*
- ▶ If we call the model M , can you learn $P(C|D; M)$
- ▶ You might have seen this as *supervised learning*
- ▶ You can also try to predict if a vehicle was taken off-road, given that it broke down

CLUSTERING

- ▶ Another very common request
- ▶ Imagine there is some hidden property in the data, another feature that we have not observed
 - ▶ This feature groups together vehicles
 - ▶ Again we are looking for $P(C|D; M)$, but C is a fictional/latent variable
- ▶ Unsupervised learning
- ▶ The probabilistic intuition I provided is not unique

INFERRING WHAT-IF SCENARIOS FROM THE DATA

- ▶ Say your vehicle broke down
- ▶ What would have happened if you have not driven if off-road?
- ▶ Have a look at the data - what can you say?
- ▶ Do you have enough data of the needed type?
- ▶ Causality from observational data
 - ▶ Super hard, but super important
 - ▶ Think of smoking!

ACQUIRING NEW DATA

- ▶ We can't really answer what would happen to vehicle from the data collected already
- ▶ We might need to set a controlled experiment where:
 - ▶ We find vehicles of similar characteristics
 - ▶ Drive them off-road
 - ▶ See if they break down
 - ▶ What is the optimal way of doing such a procedure?
- ▶ Causality from experimental data - mostly what science is all about
 - ▶ **Science is the epistemology of causality**

ANOMALY DETECTION

- ▶ If we are given a new vehicle, can we say if it is “special” in a way?
- ▶ Maybe it's the only vehicle with certain features
- ▶ Maybe it's a unique vehicle
- ▶ Somehow we need to find bizarre samples that do not conform to expect norm
- ▶ Multiple formal definitions

GENERATE NEW DATA

- ▶ Can I generate fictional vehicles and their properties?
- ▶ Mathematically, learn $P(D;M)$, a model of the data
- ▶ You can then use your plausible, but fictional vehicles for entertainment
- ▶ “Learning to draw before learning to see”
 - ▶ $P(D, C; M) = P(C|D)P(D)$
 - ▶ $P(D|C; M)$

DIMENSIONALITY REDUCTION

- ▶ Maybe we only need some feature combination above
- ▶ Maybe some features only carry noise with them - they are irrelevant
- ▶ For example, how important the *car_colour* feature would be?
- ▶ What happens if we learn based on irrelevant features?
- ▶ Spurious correlations are everywhere
- ▶ Kicking out useless features might make the model more interpretable

LINKING WITH OTHER DATA/COLLECTING LABELS

- ▶ What if the data we have is not enough?
- ▶ In our example, model make is not provided
- ▶ Can we inquire data providers to find that?
- ▶ How expensive would that be?
- ▶ How easy is to label the data?
 - ▶ Active learning
 - ▶ Labelled data often very expensive

MAKING DECISIONS FROM DATA

- ▶ Now that we have a model
- ▶ Let's say you know that a vehicle will break down after three months with a certain probability
 - ▶ How much do we charge for insurance on it?
 - ▶ Should we even sell insurance to the owner?
 - ▶ What is the risk of actually selling insurance?
- ▶ We are missing another model (that of the customer)
 - ▶ Do we actually need the model?
 - ▶ Do customer preferences change over time?
- ▶ Bandits, reinforcement learning

SOME NOTES

- ▶ *“If you torture the data enough, nature will always confess.”*
 - ▶ *Disputed*
- ▶ *“If you torture the data long enough, it will confess to anything.”*
 - ▶ Huff, D. “How to lie with statistics (illust. I. Geis).” NY: Norton (1954).
- ▶ *Lies, damned lies, and statistics*
 - ▶ *Disputed*

STARTUP MAYHEM

MACHINE INTELLIGENCE 3.0

ENTERPRISE INTELLIGENCE

VISUAL

Ortalix Insight Planet
darfai deepvision
cortica iQvision
SPACE KNOW iQvision
netra deepomatic

AUDIO

Gridspace TalkIQ
nexusia
CAPIO Expect Labs
Clover
Qurbulax
payit archive

SENSOR

PREDIX GYOT MAANA
Sentient PLANET OS
UPTAKE IBM
thingwork KODAK Aluatum

INTERNAL DATA

PRIMER DEWATSON
Dyars Palantir ADAMO
Alation Sapho Outlier
Digital Reasoning

MARKET

mattermark Quid
Datafux PREMISE
Bottlenuse MOTIVA
enigma CIBERWITS
OTRACK predata

CUSTOMER SUPPORT

DigitalGenius Kasisto
ELOQUENT T/S/SEA
ACTIONIO zendesk
Pinct CLARABRIDGE

SALES

collective sense
fuse/machines AVISO
salesforce INSIDE
SALES clari
Zennight

MARKETING

MENTIGO Lattice RADIUS
Lifeline PERSADO
Brightfunnel retention
COCONOR AURA Cmapa

SECURITY

CYCLANCE DARKTRACE
ZIMPERIUM dependant
Sentinel DEMISTO
graphistry drawbridge
SignalSense AppZen

RECRUITING

textio envilo
Wade & Wendy h
unitive SpringRole
GIGSTER HireVue

ENTERPRISE FUNCTIONS

AUTONOMOUS SYSTEMS

GROUND NAVIGATION

drive.ai AdaWorks
ZOOX
USER Google TIRSLA
Autonomous Auto Robotics

AERIAL

SKYDIO SHIELD AI
Airware DJI
DroneDeploy
plio.ai SKYWATCH

INDUSTRIAL

JAYBRIDGE OSARO
CLEARPATH fetch
KING 3 D
HARVEST
nubot robotics

PERSONAL

amazon alexa
Cortana ALO
facebook
Siri Repika

AGENTS

PROFESSIONAL

butter.ai POPS SKIFFLAG
clara x.ai slack
talia Zoom sudo

TECHNOLOGY STACK

AGENT ENABLERS

OCTANE.AI howdy Maluba KITT.AI
OpenAI Gym Kasisto AUTOMAT
semanticmachines

DATA SCIENCE

DOMINO SPARKBEYOND rapidminer
kaggle DataRobot yhat AYASDI
data iku seldon yscope bigml

MACHINE LEARNING

CognitiveScale GoogleML context relevant
Dyars HyperScience n2o logic minds.ai H2o.ai
SCALES INFERENCE sparkognition loop
deepsource reactive skyminde bonsai

NATURAL LANGUAGE

agolo FLYLIE LEXALYTICS
Narrative Science spaCy LUMINOSO
cortical.io MonkeyLearn

DEVELOPMENT

SIGOPT HyperOpt fuzzy okite
rainforest lobe Anodot
Signifai LAYER 6 bonsai

DATA CAPTURE

CrowdFlower diffbot CrowdAI import
Paxala DATASET amazon mechanical turk enigma
WorkFusion DATALOGUE TRIFACTA parsehub

OPEN SOURCE LIBRARIES

Keras Chainer CNTK TensorFlow Caffe
H2O DEEPLARNING4J theano torch
DSSTNE scikit-learn AzureML neon
MXNet DMTK Spark PaddlePaddle WEKA

HARDWARE

KNUPATH TENSTORRENT Cirascale
NVIDIA nvidia nervana Movius
terisilica GoogleTPU 10 Labs
Cerebras Isosemi

RESEARCH

OpenAI Inria Numenta ELEMENT vicarious
KNOGIN Numenta Kmera Systems Cogitai

AGRICULTURE

BLUE DRIVER mavix
tule TRACE Pinct
Droptree AGRI-DATA
Cortica

EDUCATION

KNEWTON volley
gradescope
CTI COURSERA
UDACITY edX school

INDUSTRIES

INVESTMENT

Bloomberg KNEWTON sentiment
SENTIUM KENSHC
alpha sense Dotominr
CIBERWITS Quidai

LEGAL

blue J BEAGLE
Everlaw RAVEL
Seal ROSS
LEGAL ROBOT

LOGISTICS

NAUTO Acenta
PRETECK
Routific clearmat
MARBLE PITSTOP

INDUSTRIES CONT'D

MATERIALS

zymergen Citrine
Eigen Innovations
BRIGHT MACHINE
nanosystems
CALCULABIO

RETAIL FINANCE

TALA finance
Lendo earnest
affirm MIRADOR
wealthfront Betterment

PATIENT

PULSE CareZone
ZEPHYRUS HEALTH
Oncoda SEPTENIA
Atomwise Numerate

HEALTHCARE

BUTTERY 3SCAN
ARTERYS enlitic
BAYLABS imago
Google DeepMind

BIOLOGICAL

CarbonX color GRAIL
deep genomics RECURSION
LUMINIST illuminate
Atomwise verity

THE LAW

“We summarize the potential impact that the European Union’s new General Data Protection Regulation will have on the routine use of machine learning algorithms. Slated to take effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors) which “significantly affect” users. The law will also effectively create a **right to explanation**, whereby a user can ask for an explanation of an algorithmic decision that was made about them. We argue that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation”

Goodman, Bryce, and Seth Flaxman. “European Union regulations on algorithmic decision-making and a” right to explanation“.” arXiv preprint arXiv:1606.08813 (2016).

THE SOCIAL IMPACT OF AI/MACHINE LEARNING

“We examine how susceptible jobs are to computerisation. To assess this, we begin by implementing a novel methodology to estimate the probability of computerisation for 702 detailed occupations, using a Gaussian process classifier. Based on these estimates, we examine expected impacts of future computerisation on US labour market outcomes, with the primary objective of analysing the number of jobs at risk and the relationship between an occupation’s probability of computerisation, wages and educational attainment. According to our estimates, about 47 percent of total US employment is at risk. We further provide evidence that wages and educational attainment exhibit a strong negative relationship with an occupation’s probability of computerisation”

- Not sure I believe them, but read the article

Frey, Carl Benedikt, and Michael A. Osborne. “The future of employment: how susceptible are jobs to computerisation.” Technological Forecasting and Social Change (2014).

OVERALL ON DATA AND SOCIETY

- ▶ Think about how much of your life you spend online
 - ▶ Not just on a computer, but mobile phones, GPS signals etc., car sensors
 - ▶ Soon your fridge and coffee machine (IoT)
- ▶ Tons of data flying around
 - ▶ They are being used to make decisions on a micro level (i.e. about you)
- ▶ Regulations are set in place
- ▶ New El-Dorado?

LINUX VM

- ▶ Download the VM for this module
- ▶ External link https://docs.google.com/uc?id=0B_kDfEzMuWD6ZGJFU1VfeEY3TnM&export=download
- ▶ The VM contains all (or most) of what you need if you are to create a successful python project
- ▶ Username/password is `mlvm/mlvm`
- ▶ You will have a USB stick where you should copy the VM folder (after you un-rar the archive)
- ▶ More about this on the labs

PYTHON

- ▶ Python is the language of this module
- ▶ You are expected to be competent python programmers (or willing to put the extra effort)
- ▶ Python has evolved to be one of the two “data science” languages (the other is **R**)
- ▶ Python has/is:
 - ▶ An excellent list of features coming from functional programming
 - ▶ A huge number of related libraries
 - ▶ Easy to learn
 - ▶ Object oriented programming capabilities
 - ▶ Can be extended via *C* trivially
 - ▶ A massive amount of related libraries

IPYTHON/JUPITER

- ▶ A better command line interface to python
- ▶ Has something called a “notebook”
 - ▶ A notebook combines code + natural language
- ▶ See [here](#) for a very nice example

PYCHARM SHORTCUTS

► Double shift - meta-shortcut!

PyCharm Default Keymap



Editing

Ctrl + Space	Basic code completion (the name of any class, method or variable)
Ctrl + Alt + Space	Class name completion (the name of any project class independently of current imports)
Ctrl + Shift + Enter	Complete statement
Ctrl + P	Parameter info (within method call arguments)
Ctrl + Q	Quick documentation lookup
Shift + F1	External Doc
Ctrl + mouse over code	Brief Info
Ctrl + F1	Show descriptions of error or warning at caret
Alt + Insert	Generate code...
Ctrl + O	Override methods
Ctrl + Alt + T	Surround with...
Ctrl + /	Comment/uncomment with line comment
Ctrl + Shift + /	Comment/uncomment with block comment
Ctrl + W	Select successively increasing code blocks
Ctrl + Shift + W	Decrease current selection to previous state
Ctrl + Shift + J	Select till code block end/start
Alt + Enter	Show intention actions and quick-fixes
Ctrl + Alt + L	Reformat code
Ctrl + Alt + O	Optimize imports
Ctrl + Alt + I	Auto-indent line(s)
Tab / Shift + Tab	Indent/unindent selected lines
Ctrl + X or Shift + Delete	Cut current line or selected block to clipboard
Ctrl + C or Ctrl + Insert	Copy current line or selected block to clipboard
Ctrl + V or Shift + Insert	Paste from clipboard
Ctrl + Shift + V	Paste from recent buffers...
Ctrl + D	Duplicate current line or selected block
Ctrl + Y	Delete line at caret
Ctrl + Shift + J	Smart line join
Ctrl + Enter	Smart line split
Shift + Enter	Start new line
Ctrl + Shift + U	Toggle case for word at caret or selected block
Ctrl + Delete	Delete to word end
Ctrl + Backspace	Delete to word start

PyCharm Default Keymap



Running

Alt + Shift + F10	Select configuration and run
Alt + Shift + F9	Select configuration and debug
Shift + F10	Run
Shift + F9	Debug
Ctrl + Shift + F10	Run context configuration from editor
Ctrl + Alt + R	Run manage.py task

Debugging

F8	Step over
F7	Step into
Shift + F8	Step out
Alt + F9	Run to cursor
Alt + F8	Evaluate expression
Ctrl + Alt + F8	Quick evaluate expression
F9	Resume program
Ctrl + F8	Toggle breakpoint
Ctrl + Shift + F8	View breakpoints

Navigation

Ctrl + N	Go to class
Ctrl + Shift + N	Go to file
Ctrl + Alt + Shift + N	Go to symbol
Alt + Right/Left	Go to next/previous editor tab
F12	Go back to previous tool window
Esc	Go to editor (from tool window)
Shift + Esc	Hide active or last active window
Ctrl + Shift + F4	Close active run/messages/find/... tab
Ctrl + G	Go to line
Ctrl + E	Recent files popup
Ctrl + Alt + Left/Right	Navigate back/forward
Ctrl + Shift + Backspace	Navigate to last edit location
Alt + F1	Select current file or symbol in any view
Ctrl + B or Ctrl + Click	Go to declaration
Ctrl + Alt + B	Go to implementation(s)
Ctrl + Shift + I	Open quick definition lookup

(From jetbrains blog)

JUPITER/IPYTHON NOTEBOOK SHORTCUTS

The Jupyter Notebook has two different keyboard input modes. **Edit mode** allows you to type `x` code/text into a cell and is indicated by a green cell border. **Command mode** binds the keyboard to notebook level actions and is indicated by a grey cell border.

Command Mode (press `Esc` to enable)

Enter :	enter edit mode	B :	insert cell below
Shift - Enter :	run cell, select below	X :	cut selected cell
Ctrl - Enter :	run cell	C :	copy selected cell
Alt - Enter :	run cell, insert below	Shift - V :	paste cell above
Y :	to code	V :	paste cell below
M :	to markdown	Z :	undo last cell deletion
R :	to raw	D,D :	delete selected cell
1 :	to heading 1	Shift - M :	merge selected cells
2 :	to heading 2	S :	Save and Checkpoint
3 :	to heading 3	Ctrl - S :	Save and Checkpoint
4 :	to heading 4	L :	toggle line numbers
5 :	to heading 5	O :	toggle output
6 :	to heading 6	Shift - O :	toggle output scrolling
K :	select cell above	Esc :	close pager
Up :	select cell above	Q :	close pager
J :	select cell below	H :	show keyboard shortcut help dialog
Down :	select cell below	I,I :	Interrupt kernel
Shift - K :	extend selection above	0,0 :	restart kernel
Shift - J :	extend selection below	Shift - Space :	scroll up
A :	insert cell above		

(From stackoverflow)

NUMPY

- ▶ Numpy is possibly the most important library in Python for numerical computing
- ▶ Provides vector and matrix operations on top of *arrays*
- ▶ Almost every other library manipulates numpy arrays underneath

SCIPY

- ▶ A scientific computing framework
- ▶ Linear Algebra
- ▶ Optimisation
- ▶ Statistics
- ▶ Clustering

SCIKIT-LEARN

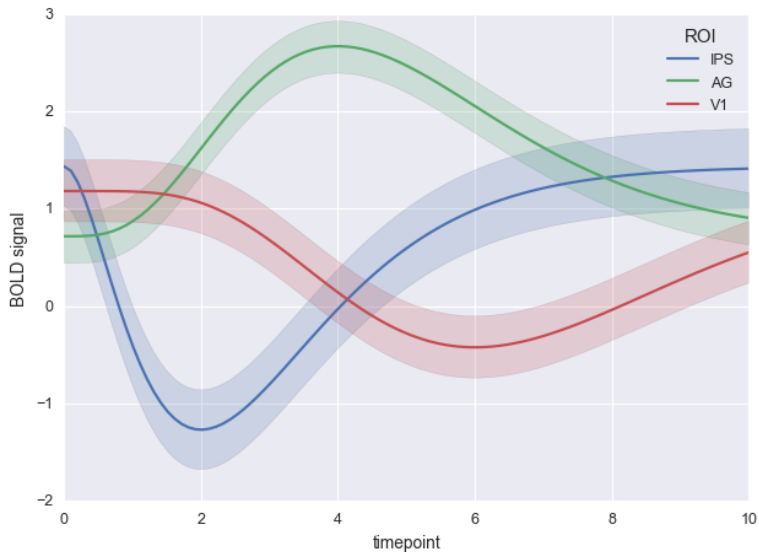
- ▶ A machine learning framework
- ▶ Includes almost everything, apart from neural networks
- ▶ We are going to use it extensively
- ▶ Super-fast trees
- ▶ Excellent documentation
- ▶ `http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html`
 - ▶ Cross validation for time series

KERAS

- ▶ A neural networks framework
- ▶ Very popular
- ▶ Uses theano or tensorflow underneath
- ▶ We will use this as well
- ▶ Though notice this is not a module on neural networks
 - ▶ But you can delve into this if you want
 - ▶ Not trivial, but not super hard either
 - ▶ Again, a lot of examples and online tutorials

MATPLOTLIB, SEABORN

► Standard visualisation tools



PANDAS

- ▶ *R* had dataframes
 - ▶ Essentially, a very SQL-like table-like data structure
- ▶ “DataFrame is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. It is generally the most commonly used pandas object”
- ▶ You can manipulate these, and it helps a lot with cleaning up and re-shaping your data
- ▶ This is a big part of data science!
 - ▶ Data munging/data wrangling

APACHE SPARK

- ▶ The clustering framework
- ▶ You need it when you have tons of data to process
- ▶ Has its own machine learning library (mllib), which we are not going to use
 - ▶ But it makes sense to use it if your data doesn't fit in memory
 - ▶ Can be used with 3rd party modules in conjunction with sk-learn
- ▶ Sits on top of HDFS (which we are going to install and use later on)

GITHUB

- ▶ All your code for your project will need to be publicly available
- ▶ Create a github account if you don't have one
- ▶ Two directories (`/src`, `/pdf`)
 - ▶ One for the pdf of the project
 - ▶ One for the code
 - ▶ If you have an ipython ipnb it should go here
- ▶ Add a README.md as well!