APPLIED DATA SCIENCE
מדע נתונים יישומי

Dr. Omri Allouche
2017

APPLIED DATA SCIENCE

Class #224

Wednesdays 11am-2pm

No Tirgul Classes

WHO AM I?

Omri Allouche

PhD in Computational Ecology from HUJI

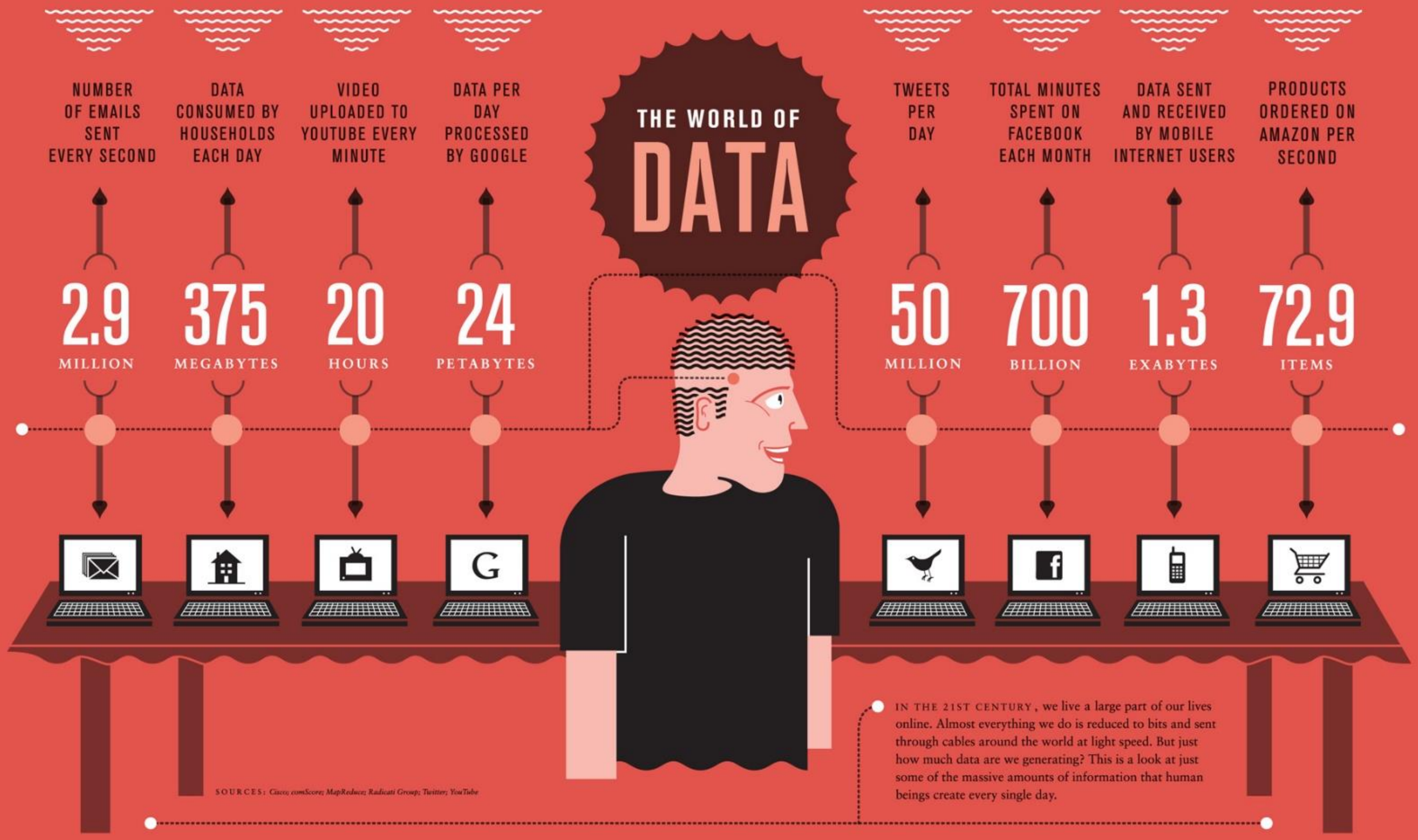
Senior Data Scientist @ Gong.io

In the past – startup entrepreneur, R&D team leader @ Elisra

No set office hours – coordinate by email

omri.allouche@gmail.com

THE STATE OF DATA



THE WORLD OF DATA

NUMBER OF EMAILS SENT EVERY SECOND: 2.9 MILLION

DATA CONSUMED BY HOUSEHOLDS EACH DAY: 375 MEGABYTES

VIDEO UPLOADED TO YOUTUBE EVERY MINUTE: 20 HOURS

DATA PER DAY PROCESSED BY GOOGLE: 24 PETABYTES

TWEETS PER DAY: 50 MILLION

TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH: 700 BILLION

DATA SENT AND RECEIVED BY MOBILE INTERNET USERS: 1.3 EXABYTES

PRODUCTS ORDERED ON AMAZON PER SECOND: 72.9 ITEMS

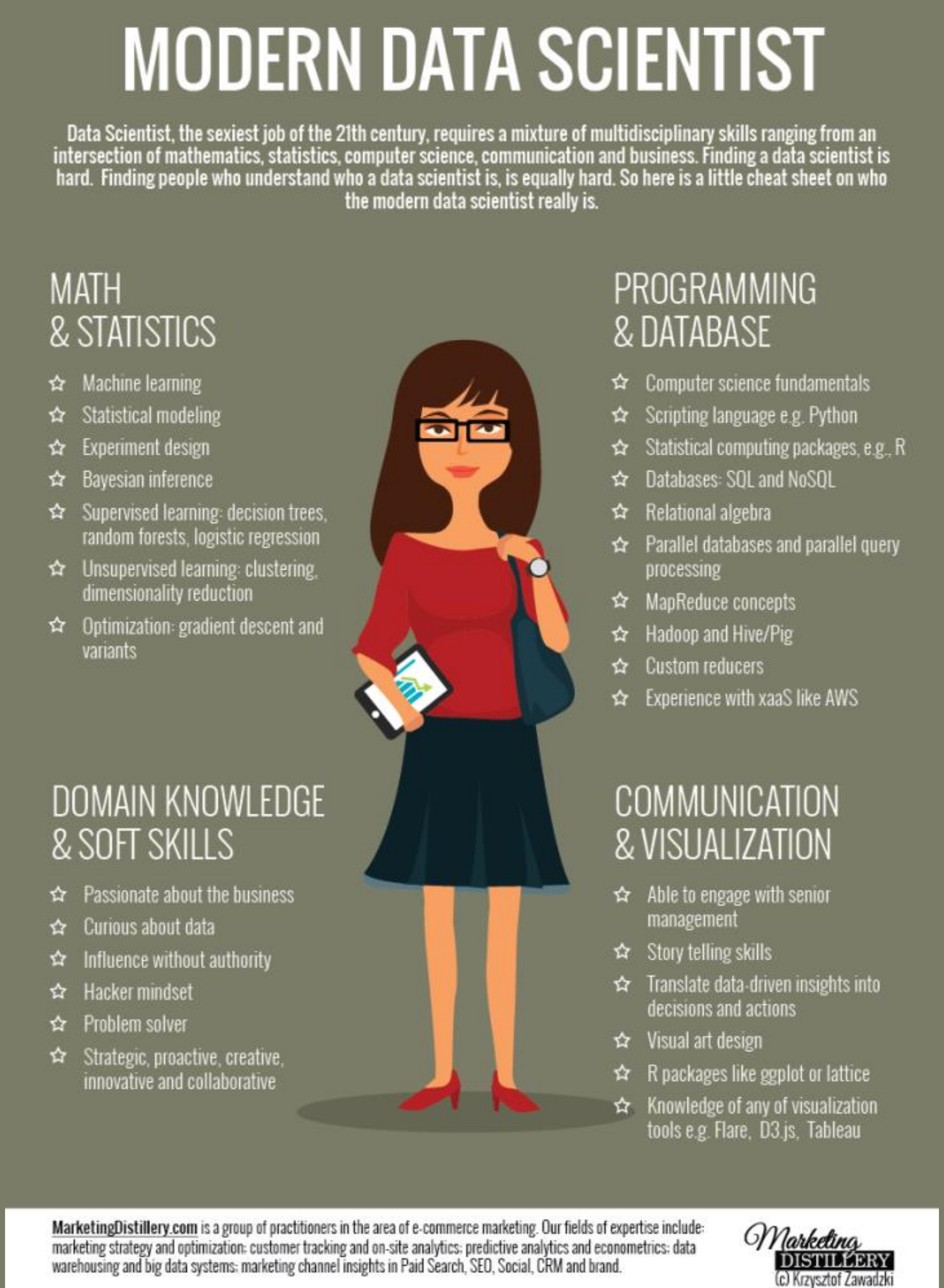
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

SOURCES: *Cisco; comScore; MapReduce; Radicati Group; Twitter; YouTube*

A COLLABORATION BETWEEN GOOD AND OLIVER MUNDAY

IN PARTNERSHIP WITH IBM

Source: bimeanalytics.com



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

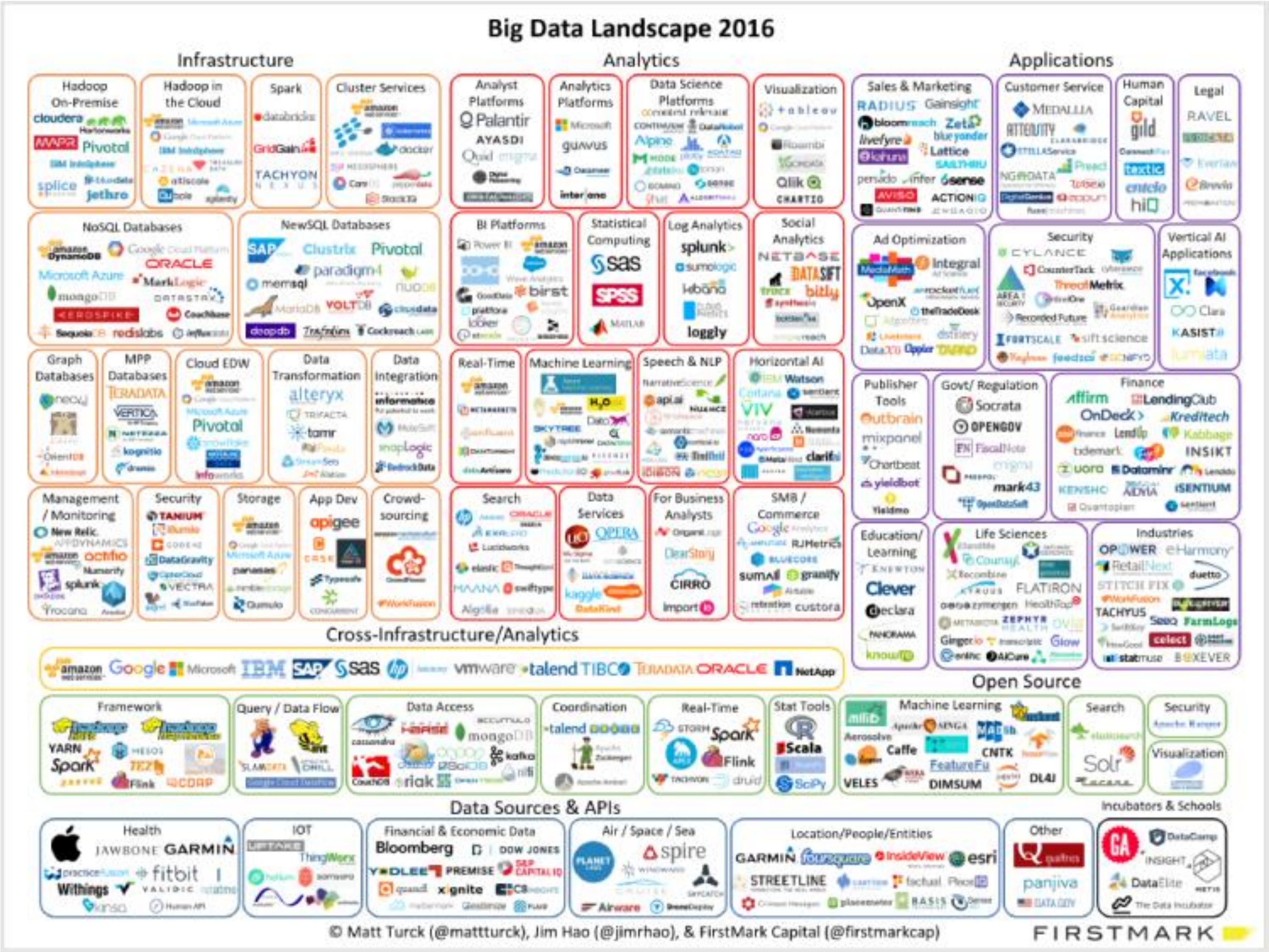
COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY
(c) Krzysztof Zawadzki

Source: marketingdistillery.com



Source: http://mattturck.com/2016/02/01/big-data-landscape/

A SINGLE COURSE WON'T MAKE YOU A (GOOD) DATA SCIENTIST

Data Science is booming

Pace of progress is unparalleled

There's no silver bullet

No real out-of-the-box solutions

A few rules, meant to be broken

You have to be able to learn on your own, constantly

YOU'RE LEARNING A NEW LANGUAGE

Immerse yourself in the new language

- Read. A lot
- Read. In English

Don’t be intimidated if you don’t understand everything, or even most of the things

Learn how you learn, and use it

- There are great video courses & lectures online (list follows)
- Weekly Meetups are held in Israel
- Many books & blog articles, for all levels
- Podcasts
- Kaggle competitions

YOU'RE LEARNING A NEW LANGUAGE

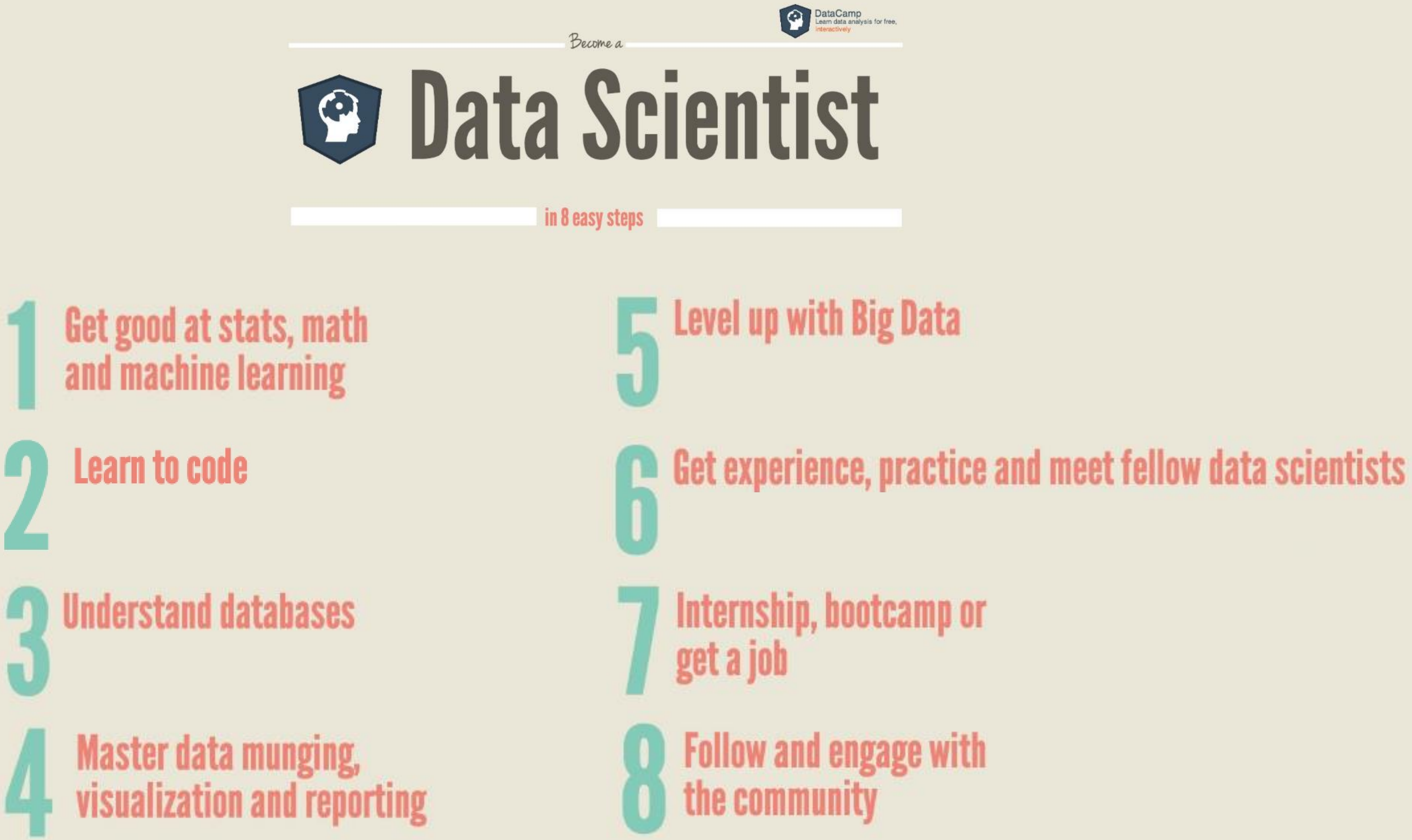
If you can't understand a concept from a specific source, try finding another one

Find a partner

Learn like a child. Experiment, and make mistakes

Learn a bit every day

Keep notes of what you don't know (or better yet, use the Spaced Repetition method)



DataCamp
Learn data analysis for free, interactively

Become a

Data Scientist

in 8 easy steps

1 Get good at stats, math and machine learning

2 Learn to code

3 Understand databases

4 Master data munging, visualization and reporting

5 Level up with Big Data

6 Get experience, practice and meet fellow data scientists

7 Internship, bootcamp or get a job

8 Follow and engage with the community

Source: https://www.datacamp.com/community/tutorials/how-to-become-a-data-scientist

COURSE SYLLABUS

1. Python for Data Analysis
2. Information Visualization
3. Exploratory Data Analysis
4. Applied Machine Learning

LECTURES PLAN

Lessons usually include

- a quick, non-mathematical review of an algorithm
- discussion of code for implementing it
- a hands-on exercise,
- discussion in class

We **will not** go into the mathematical details of the algorithms

We **will** discuss important concepts in data science and machine learning

You should come prepared for each lesson

- Read the reading list
- Freshen up on the algorithm that will be discussed

Main Assumption – we're all here because we want to

COURSE GRADING

Mostly based on a final project

Small part due to personal assessment - participation in class, bonus tasks etc.

No exam

Participation in class is mandatory

FINAL PROJECT

In groups of 2-3

Go through the complete data science process

You will acquire the data, explore it for interesting patterns, design your visualizations, run statistical analysis, and communicate the results

Use the tools you've learned during the course

You select the research question and dataset

We'll provide datasets from Israeli government authorities, HiTech companies

FINAL PROJECT

Phases:

- Form groups
- Submit a project proposal
- Periodic project reviews
- Final submission

Project should include:

- Code on GitHub - well written, well documented!
- Accompanying paper, describing the question, related work, methods, results and discussion, as a Jupyter Notebook
- A 2 minute Screencast / Presentation in Class. Instructions on creating a screencast can be found here

For ideas for projects, check out some of the projects done during Harvard's CS109 course in 2014 and 2015

PROJECT PAPER

Should include the following sections:

Overview and Motivation: Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.

Related Work: Anything that inspired you, such as a paper, a web site, or something we discussed in class.

Initial Questions: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis? - Data: Source, scraping method, cleanup, storage, etc.

Exploratory Data Analysis: What visualizations did you use to look at your data in different ways? What are the different statistical methods you considered? Justify the decisions you made, and show any major changes to your ideas. How did you reach these conclusions?

Final Analysis: What did you learn about the data? How did you answer the questions? How can you justify your answers?

Presentation: Present your final results in a compelling and engaging way using text, visualizations, images, and videos on your project web site.

YOU'RE LEARNING A NEW LANGUAGE

Immerse yourself in the new language

- Read. A lot
- Read. In English

Don’t be intimidated if you don’t understand everything, or even most of the things

Learn how you learn, and use it

- There are great video courses & lectures online (list follows)
- Weekly Meetups are held in Israel
- Many books & blog articles, for all levels
- Podcasts
- Kaggle competitions

YOU'RE LEARNING A NEW LANGUAGE

If you can't understand a concept from a specific source, try finding another one

Find a partner

Learn like a child. Experiment, and make mistakes

Learn a bit every day

Keep notes of what you don't know (or better yet, use the Spaced Repetition method)

PYTHON CRASH COURSE

PYTHON CRASH COURSE

This lesson provides a quick introduction to the Python language. It assumes you know the basic concepts of programming, including basic data types, conditionals, loops, functions etc.

For a more in-depth review of Python and programming, I recommend the following resources, which were the source for many of this course's material:

Python Data Science Handbook (Book, Paid)

Complete Python Bootcamp on Udemy (Video course, Paid). The course's Jupyter Notebooks are here.

Python 3 Essential Training on Lynda.com (Video course, Paid)

THE PYTHON LANGUAGE

Developed by Guido van Rossum in the early 1990s

Named after Monty Python

Open source general-purpose language

Becoming a standard in scientific computing

Easy to use and learn

Interpreted language: work with an evaluator for language expressions

Dynamically typed: variables do not have a predefined type

Rich, built-in collection types (Lists, Tuples, Dictionaries (maps), Sets)

Concise

Object Oriented, Procedural, Functional

Great interactive environment

Much slower than C/C++ (though wrappers exist)

DYNAMIC TYPING – THE KEY DIFFERENCE

Java: statically typed

Variables are declared to refer to objects of a given type

Methods use type signatures to enforce contracts

Python

Variables come into existence when first assigned to

A variable can refer to an object of any type

All types are (almost) treated the same way

Main drawback: type errors are only caught at runtime

THE BASICS OF PYTHON CODE

Indentation matters to the meaning of the code:

- Block structure indicated by indentation

The first assignment to a variable creates it

- Variable types don’t need to be declared
- Python figures out the variable types on its own

Assignment uses = and comparison uses ==

Logical operators are words (and, or, not) not symbols

WHITESPACE

A whitespace is meaningful in Python: especially indentation and placement of newlines.

Use a newline to end a line of code

Use \ when must go to next line prematurely

No braces { } to mark blocks of code in Python... Use consistent indentation instead

The first line with less indentation is outside of the block

The first line with more indentation starts a nested block

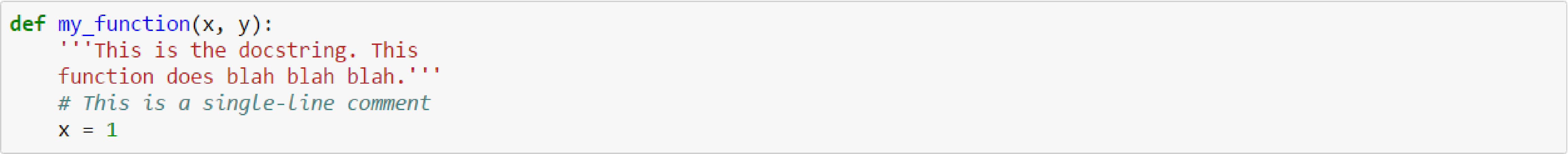
Often a colon appears at the start of a new block (e.g. for function and class definitions)

COMMENTS

Start comments with # – the rest of line is ignored.

Can include a “documentation string” as the first line of any new function or class that you define.

The development environment, debugger, and other tools use it: it’s good style to include one.



```
def my_function(x, y):
    '''This is the docstring. This
    function does blah blah blah.'''
    # This is a single-line comment
    x = 1
```

WHY PYTHON 3?

Cleaner and faster

Has inherent glitches and minor drawbacks fixed

It is the future!

Python 3 has been around for over 5 years

LET’S START PLAYING WITH SOME CODE

FOLLOW ALONG USING IBM DSX

For future lessons, you'll install your own Python environment locally

For this lesson, open http://datascience.ibm.com/

Sign in / sign up (free)

Create a new Project named "ADS" (use defaults settings)

Click "Add Notebooks" > "From URL"

Use "Python Part 1" for name, and https://github.com/omriallouche/applied-data-science/blob/master/Python%20Crash%20Course/Part%201.ipynb for url

A machine will start in the cloud, allowing you to run the code yourself

JUPYTER NOTEBOOK

Web app, served from localhost / web server

Runs a Python kernel in the background, and outputs its results

Combines markdown, code and code results

Becomes the standard for creating **reproducible data science**

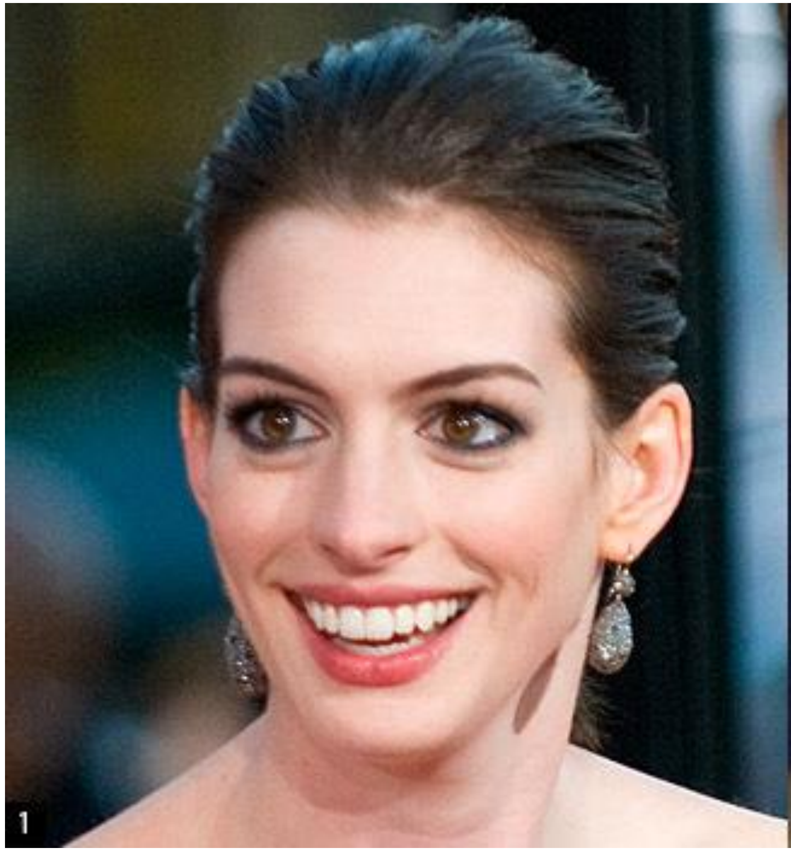
Saved as ipynb files

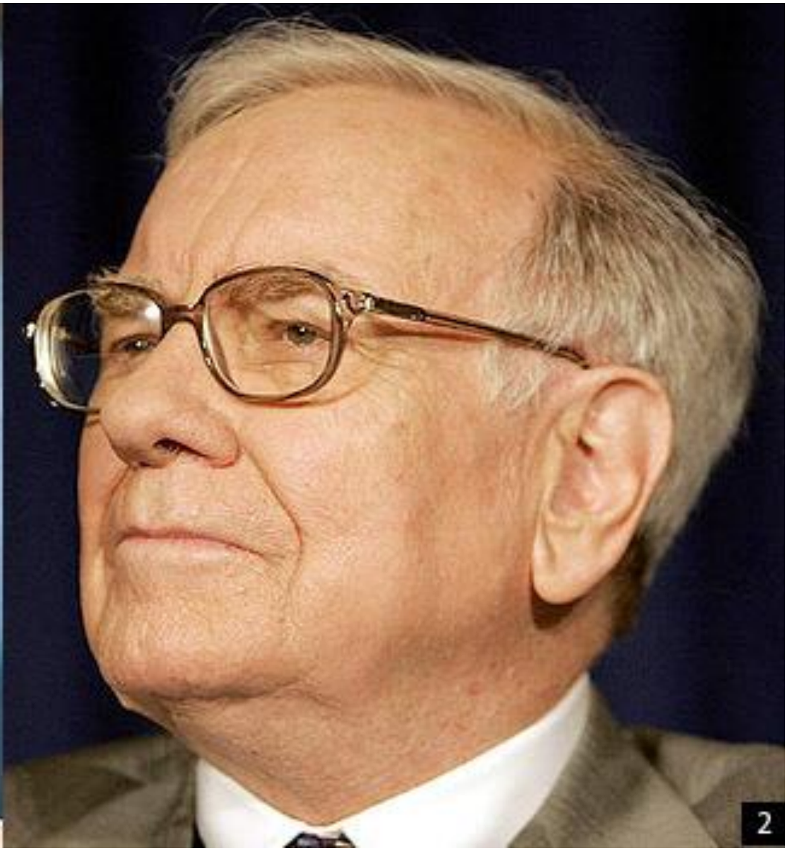
Rendered in GitHub

Can be extended to include section folding, run time logging, interactive widgets, Reveal.JS presentations and more

DATA NUGGETS

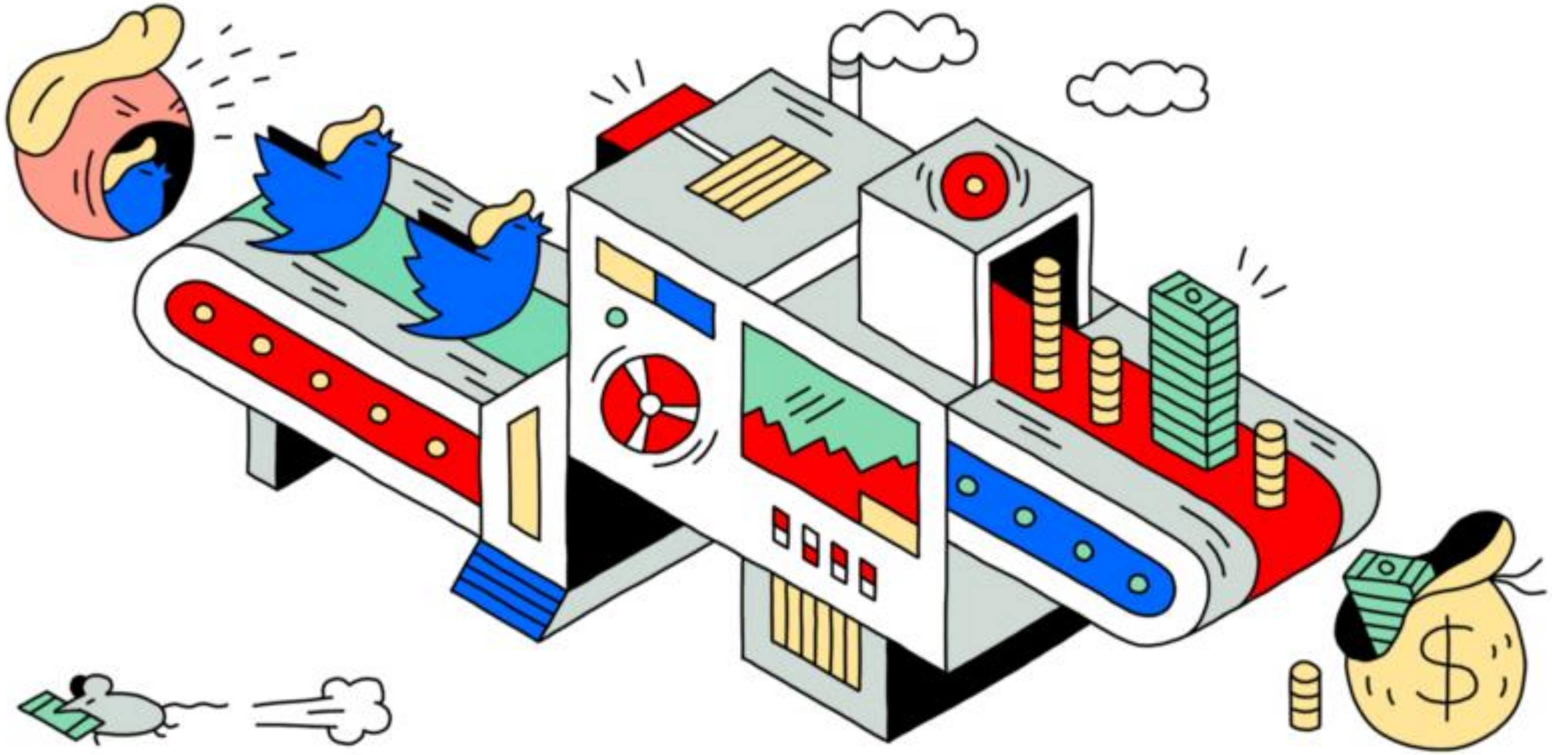
HOW ARE THESE TWO RELATED?

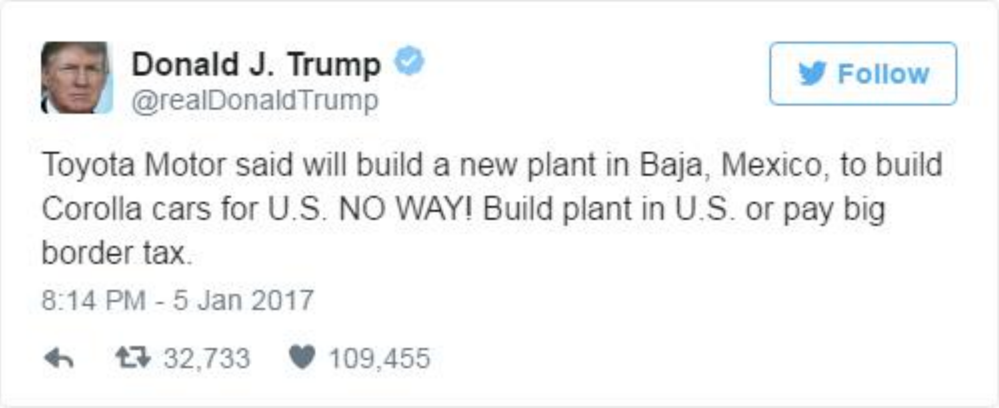




Source: The Atlantic

“THIS MACHINE TURNS TRUMP TWEETS INTO PLANNED PARENTHOOD DONATIONS” (BY MAX BRAUN)





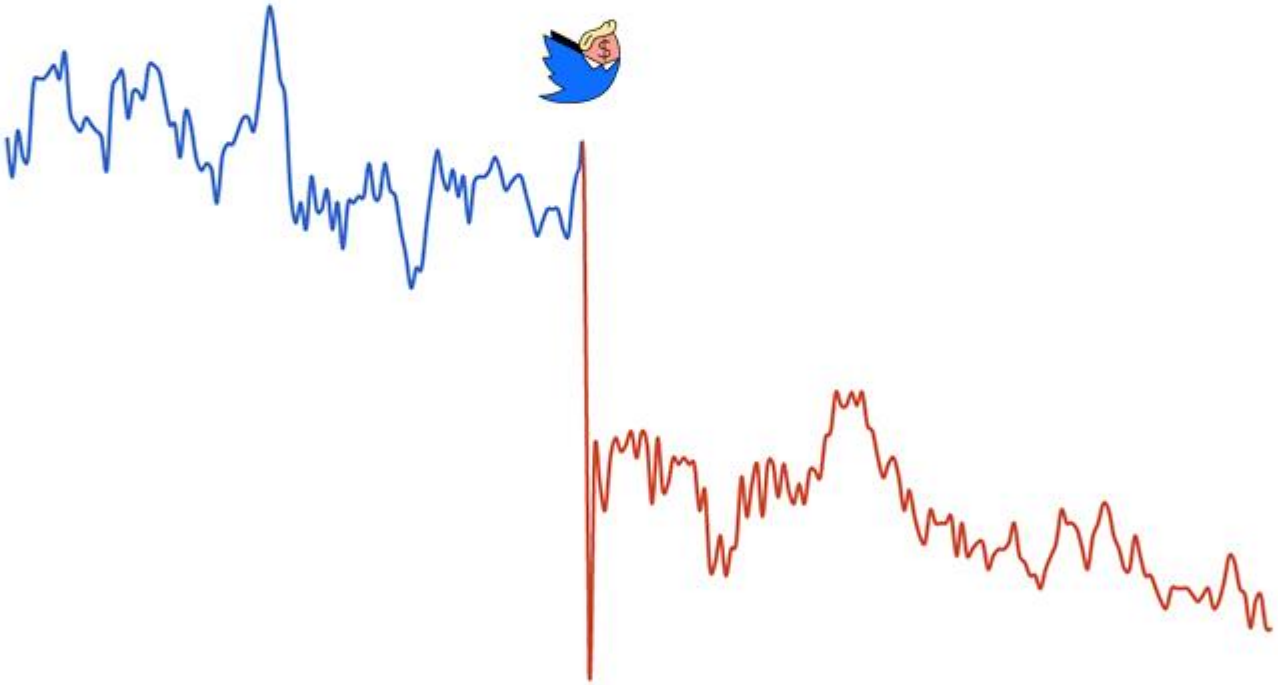
Donald J. Trump
@realDonaldTrump

Toyota Motor said will build a new plant in Baja, Mexico, to build Corolla cars for U.S. NO WAY! Build plant in U.S. or pay big border tax.

8:14 PM - 5 Jan 2017

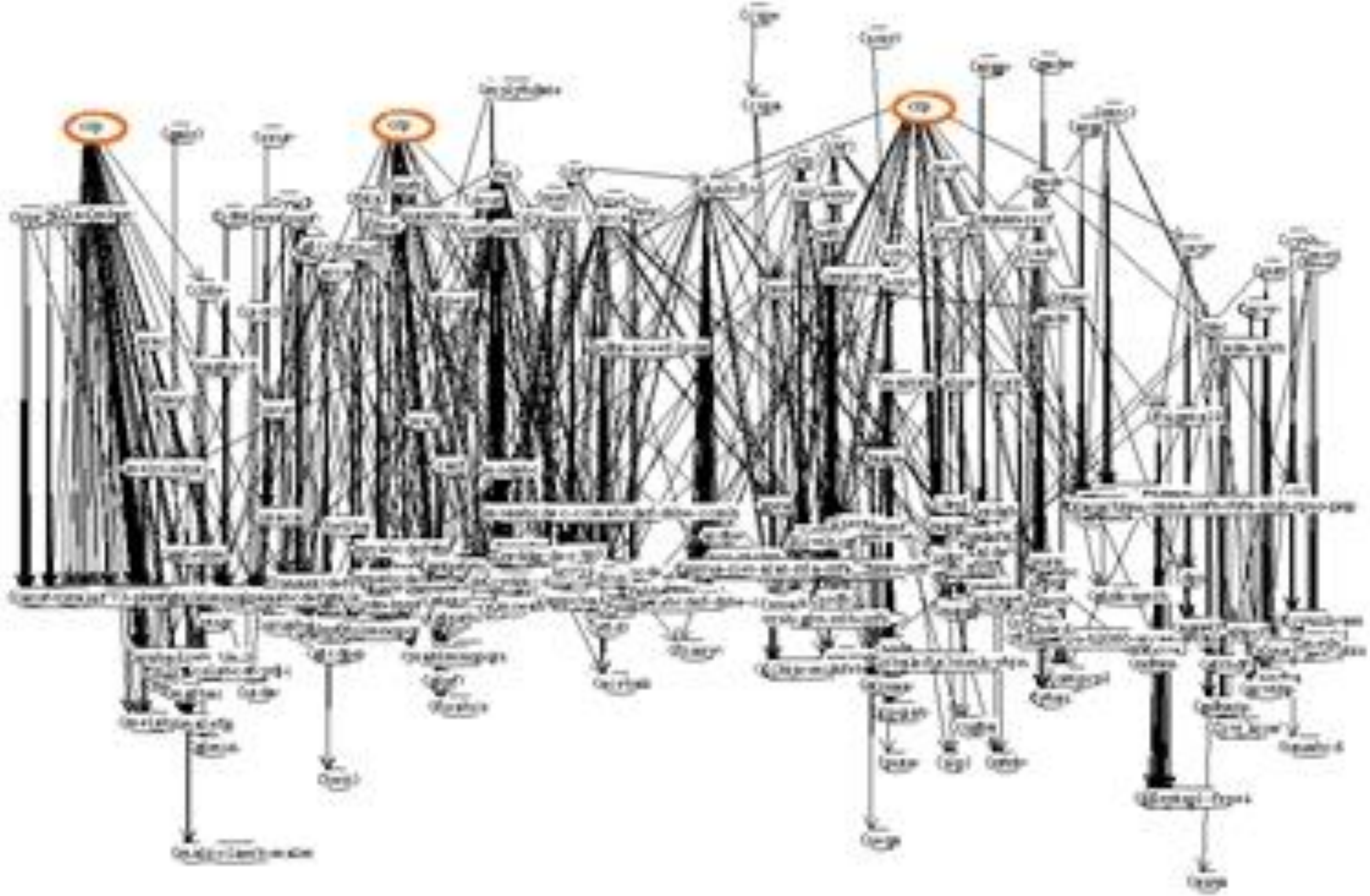
32,733 109,455

Mr. Trump's nuanced fiscal policy proposal. Sad!



Toyota's NYSE:TM stock price on January 5th 2017

DATA NUGGETS



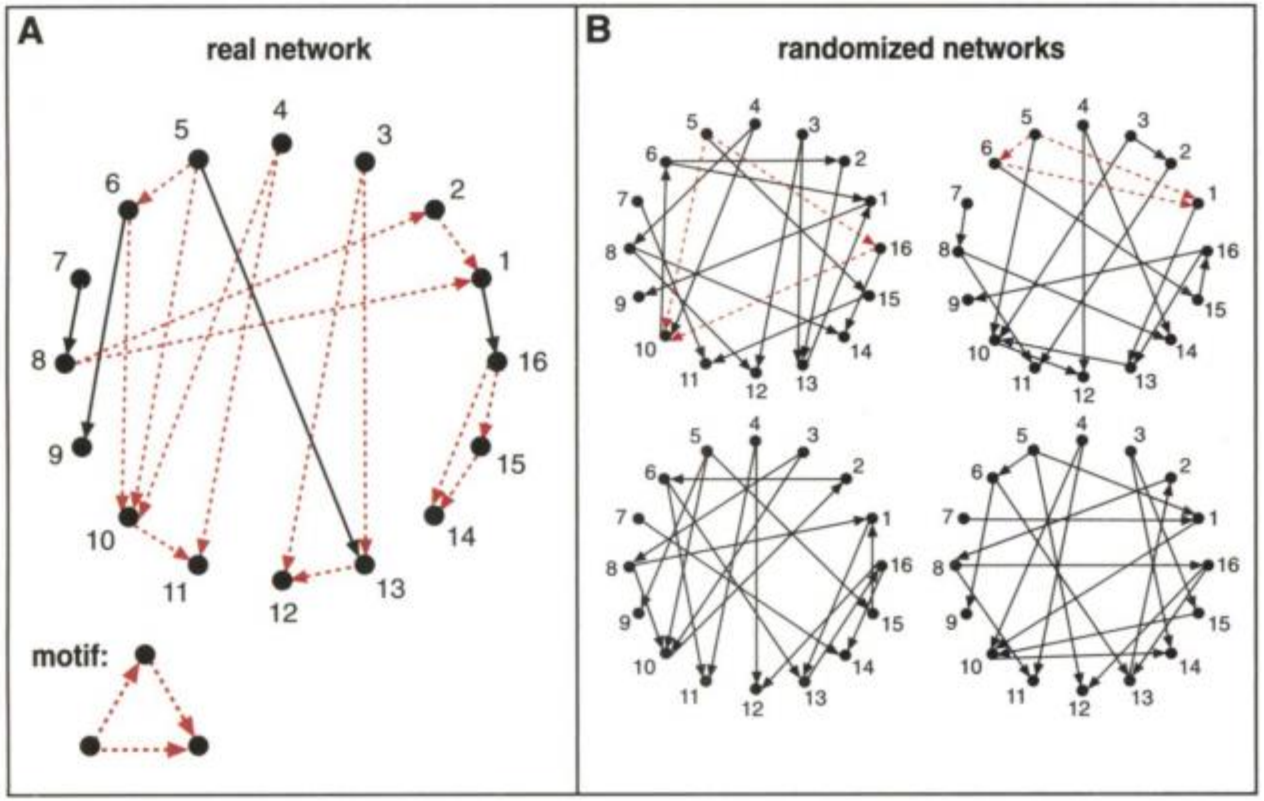


Fig. 2. Schematic view of network motif detection. Network motifs are patterns that recur much more frequently (**A**) in the real network than (**B**) in an ensemble of randomized networks. Each node in the randomized networks has the same number of incoming and outgoing edges as does the corresponding node in the real network. Red dashed lines indicate edges that participate in the feedforward loop motif, which occurs five times in the real network.

FINAL NOTES

HOMEWORK – PART 1

1. Familiarize yourself with Jupyter Notebooks. Get to know:
    1. Keyboard Shortcuts
    2. Markdown syntax
2. Follow Scikit Learn's basic Introduction to Machine Learning (make sure you run the code yourself, using Jupyter Notebook!)
3. Watch 2 Project Videos submitted for Harvard's CS109 (search YouTube for "CS109"). Email omri.allouche@gmail.com with a list of terms from each video, each marked with:

    1 – don't know anything about it

    2 – have a vague idea what it is

    3 – I know it pretty well

HOMEWORK – PART 2

1. Install an RSS feed app on your mobile phone (I recommend Feedly), and subscribe to the RSS feed of Analytics Vidhya
2. Subscribe to the weekly newsletter of Data Machina
3. Sign up to meetup.com and join a few groups related to data science
4. Read "A Few Useful Things to Know about Machine Learning"

REQUESTED TOPICS?

Databases and SQL

Natural Language Processing

Neural Networks and Deep Learning

Big Data Tools (Hadoop, Spark)

RECOMMENDED MATERIALS

ONLINE COURSES

Google's Python Class - https://developers.google.com/edu/python/

DataCamp's Interactive Tutorials - https://www.datacamp.com/courses/intro-to-python-for-data-science

EdX - https://www.edx.org/course/introduction-python-data-science-microsoft-dat208x-4#!

https://classroom.udacity.com/courses/ud120/lessons/2410328539/concepts/24185385370923#

CodeAcademy - https://www.codecademy.com/learn/python

https://www.udacity.com/course/intro-to-descriptive-statistics--ud827

READING

An elaborate guide to Python with examples

https://github.com/jakevdp/PythonDataScienceHandbook

http://www3.canisius.edu/~yany/python/Python4DataAnalysis.pdf

https://datajobs.com/what-is-data-science