Introduction to Data Science

Session 1

10 March 2018

Agenda

- Introduction to Data Science
- Introduction to R
- R Exercises
- Q&A
- Use Cases

About Us

William Yap
Data Scientist
Digital Strategy

Nadav Rindler
Data Scientist
Marketing Analytics

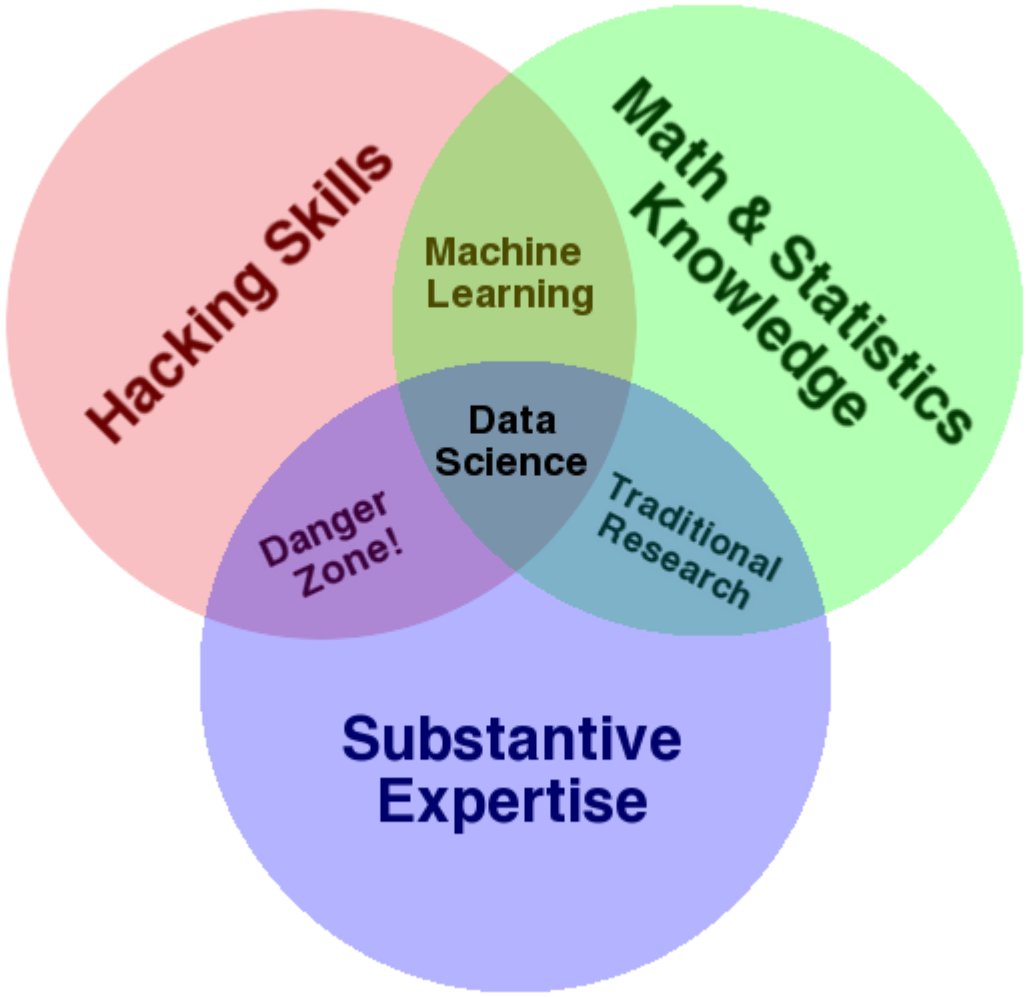
Series Map

- 10 March - Cover first half of “The Data Scientist’s Toolbox”
- 24 March - Cover second half of “The Data Scientist’s Toolbox”

- 7 April - Cover first half of “R Programming”
- 28 April - Cover second half of “R Programming”

- 12 May - Cover first half of “Getting and Cleaning Data”
- 26 May - Cover second half of “Getting and Cleaning Data”

Drew Conway's Data Science Venn diagram



Series Details

- Location: Universiti Tunku Abdul Rahman, Sungai Long Campus
- Time: 2-4pm
- Start each session with an overview of the online syllables material
- Add value through first-hand experience in the local environment

Agenda

- **Introduction to Data Science**
- Introduction to R
- R Exercises
- Q&A
- Use Cases

Digital Economy Evolution

2000 **2005** **2013** **2020**

Web & E-Business → Digital Marketing → Digital Businesses → Autonomous Businesses

+ Smart Machines

Welcome to the New World

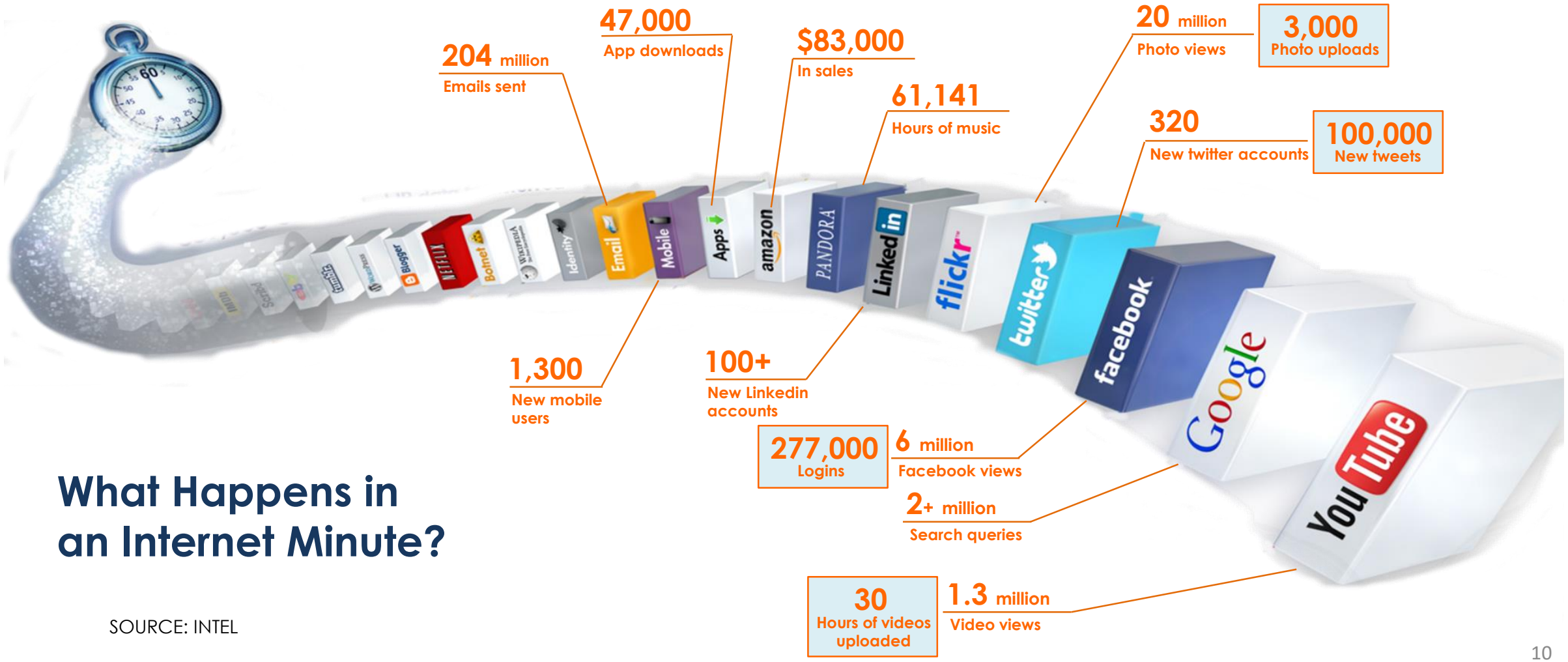
The world's largest taxi company, owns no vehicles.

The world's most popular media owner, creates no content.

The world's largest accommodation provider, owns no real estate.

The most valuable retailer, has no inventory.

Big Data Explosion



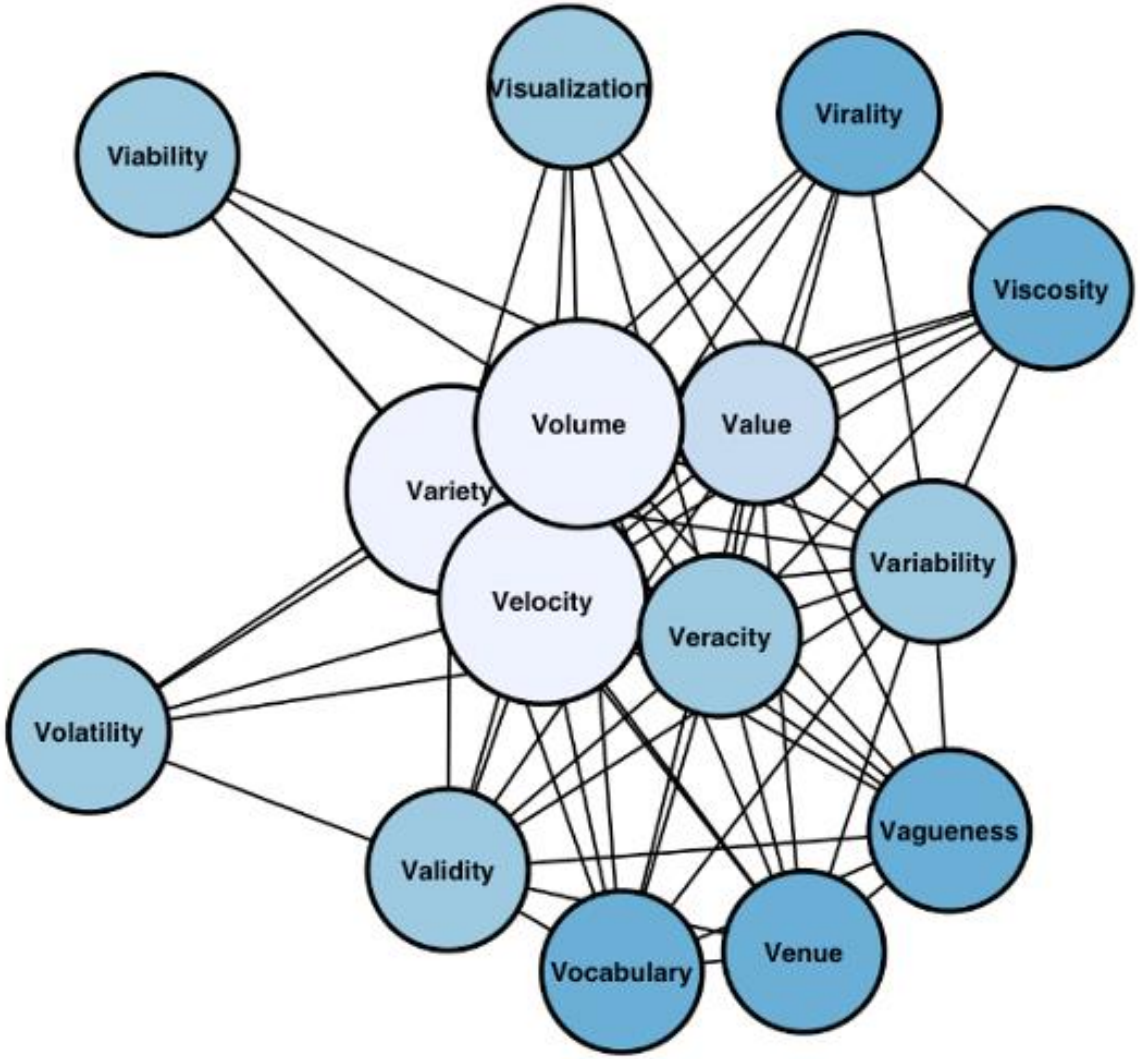
What Happens in an Internet Minute?

SOURCE: INTEL

The Four V of Big Data

- Volume – Amount of Data
- Velocity – Growth of Data
- Variety – Different Forms of Data
- Veracity – Uncertainty of Data

Now 42 V of Big Data



Data Science Definition

- Data Science is an interdisciplinary field of scientific methods to provide meaningful information from large amounts of complex data

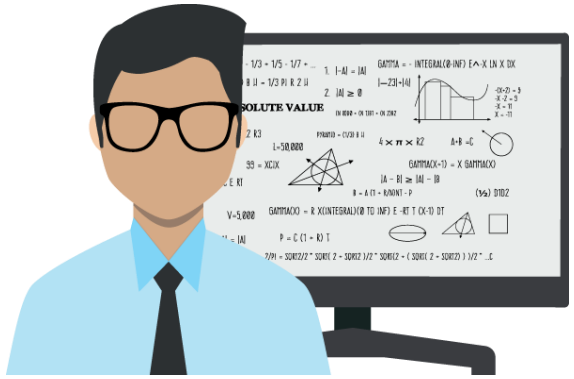
- It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science

Data Analytics vs Data Science

- Data Analytics = Finding **Right** Answers

- Data Science = Finding **Right** Questions

Data Scientists: What People Say

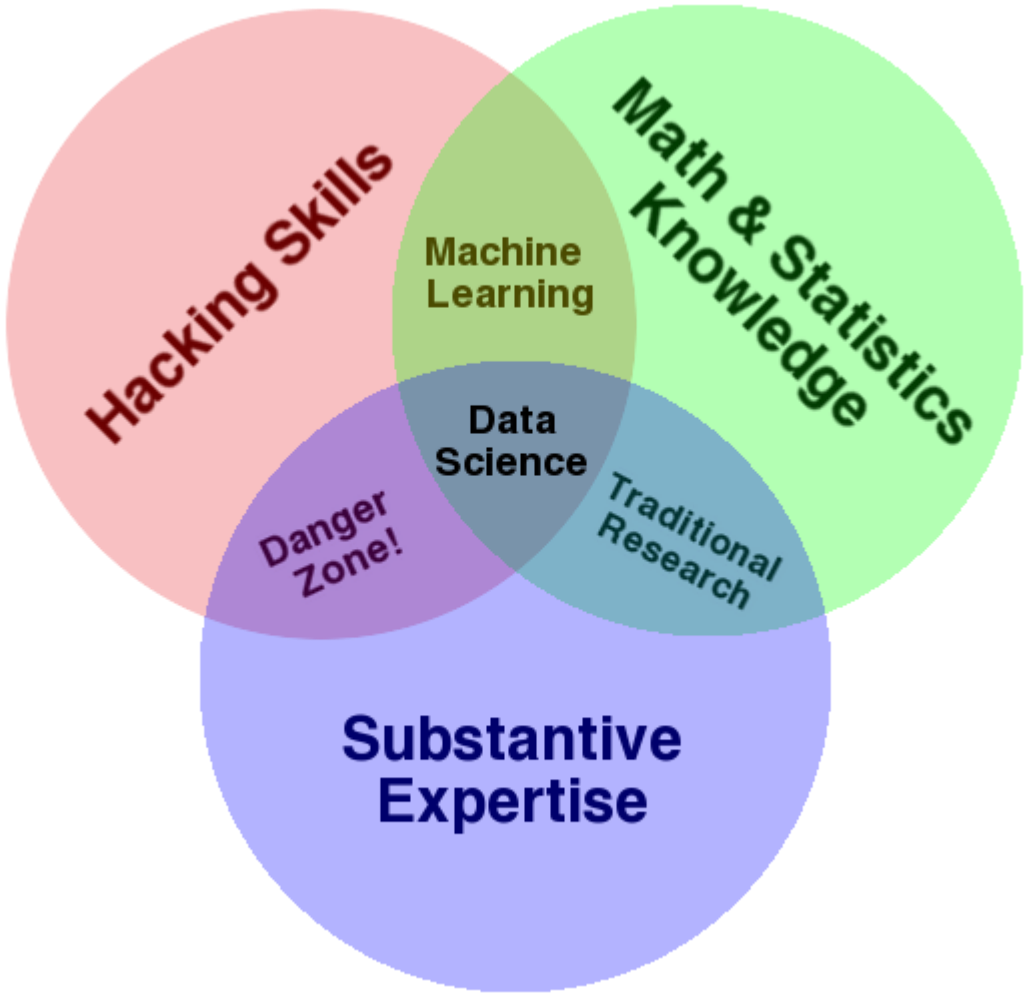


- *"The sexiest job of the 21st century"* Harvard Business Review, Oct 2012

- *"Person who is better at statistics than any software engineer and better at software engineering than any statistician"* Josh Wills, May 2012

- *"The data scientist was called, only half-jokingly, 'a caped superhero'"* Wall Street Journal, Apr 2012

Drew Conway's Data Science Venn diagram



Top Jobs of the Future

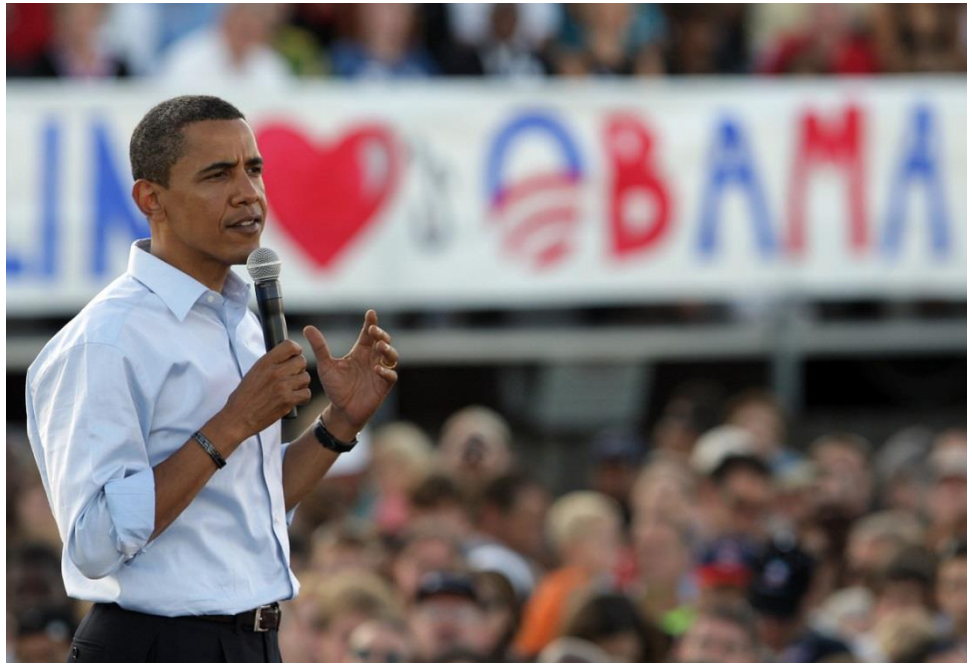
Job	Job openings	Median base salary	Career opportunity	Job score
1. Data Scientist	**1,736**	**$116,840**	**4.1**	**4.7**
2. Tax Manager	1,574	$108,000	3.9	4.7
3. Solutions Architect	2,906	$119,500	3.5	4.6
4. Engagement Manager	1,356	$125,000	3.8	4.6
5. Mobile Developer	2,251	$90,000	3.8	4.6
6. HR Manager	3,468	$85,000	3.7	4.6
7. Physician Assistant	3,364	$97,000	3.5	4.6
8. Project Manager	6,607	$106,680	3.3	4.5
9. Software Engineer	49,270	$95,000	3.3	4.5
10. Audit Manager	1,001	$95,000	3.9	4.5

Source : Glassdoor, 2016

Applications of Data Science

USA Election 2012

- The Obama campaign recruited a team of data scientists
- Allowed the Obama campaign to create a portrait of shifting voter allegiances
- The power of this operation stunned the Romney team as they saw voters they never even knew existed



Applications of Data Science

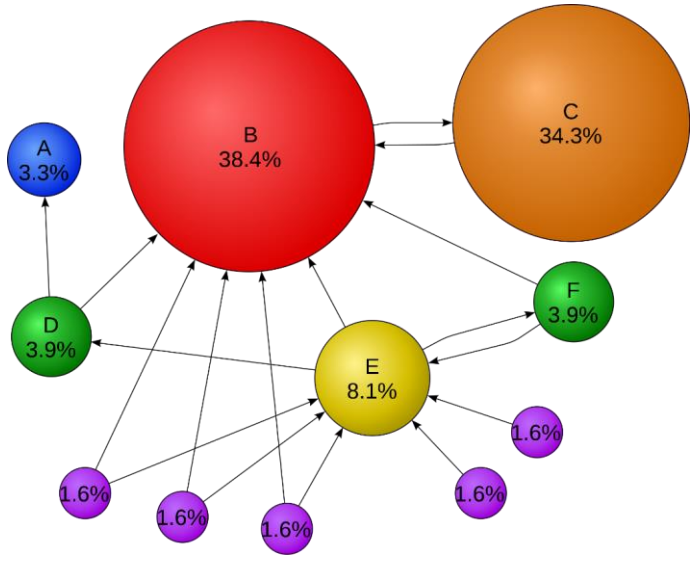
- US Government’s first Chief Data Scientist
- Dr. Dhanurjay "DJ" Patil
- Harvard Business Review “Data Scientist: The Sexiest Job of the 21st Century”
- Credited with coining the term “Data Science”

What Happened?

Applications of Data Science

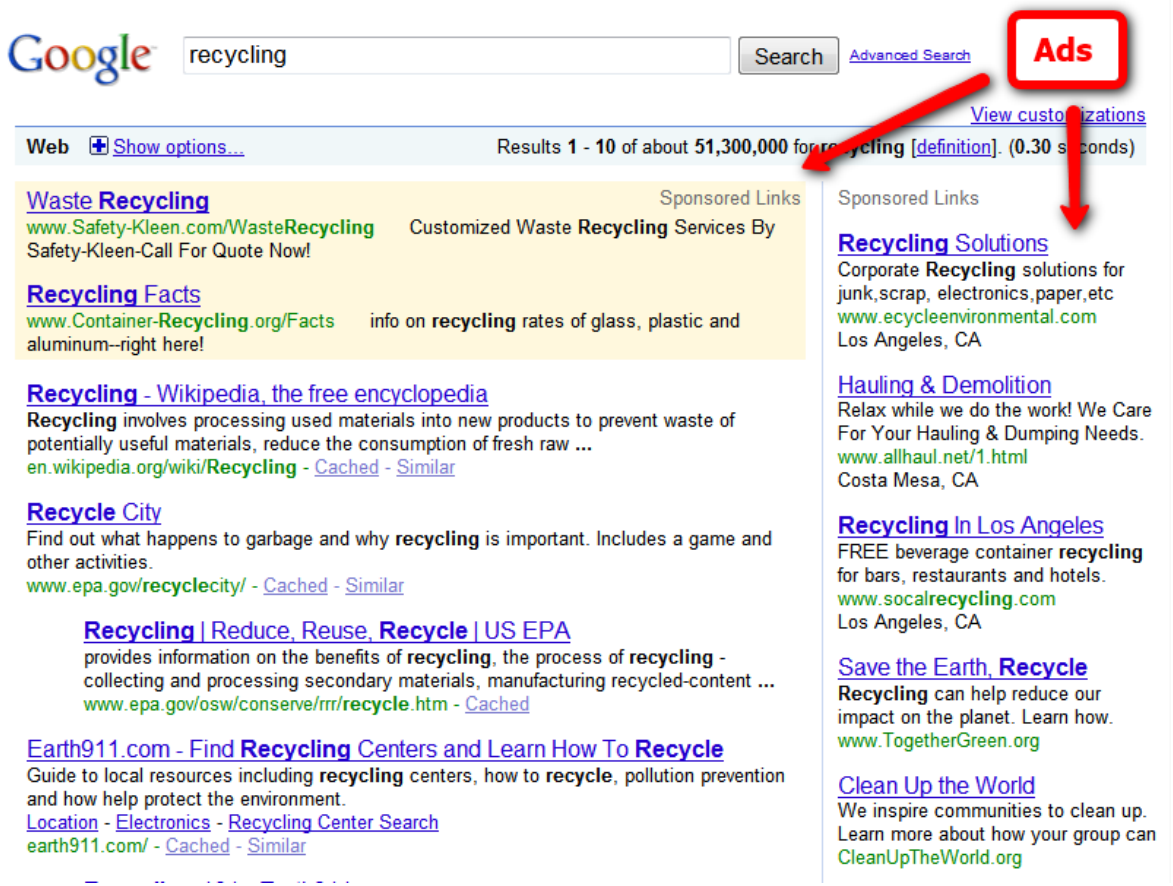
GOOGLE

- PageRank is an algorithm used by Google Search to rank websites in their search engine results



Applications of Data Science

GOOGLE ADS



Applications of Data Science

- There are around 30 billion search requests a month.

- Google revenue around $50 billion per year from marketing, 97% of the companies revenue.

Application in Insurance Industry

- In 2017, Fukoku Mutual Life Insurance Company replaced 34 employees with IBM Watson
- Used AI to calculate insurance payouts to policyholders
- Target to increase productivity by 30% and save $1.2m per year

REKA Self Driving Cars

- Since 2016, REKA has been developing its own self-driving tech
- Tested the self-driving car from KL to Melaka with no driver
- Speed 60 - 80 km/h with navigation based on Google Maps

How to become a Data Scientist

Actuarial Society of Malaysia
Persatuan Aktuari Malaysia

coursera

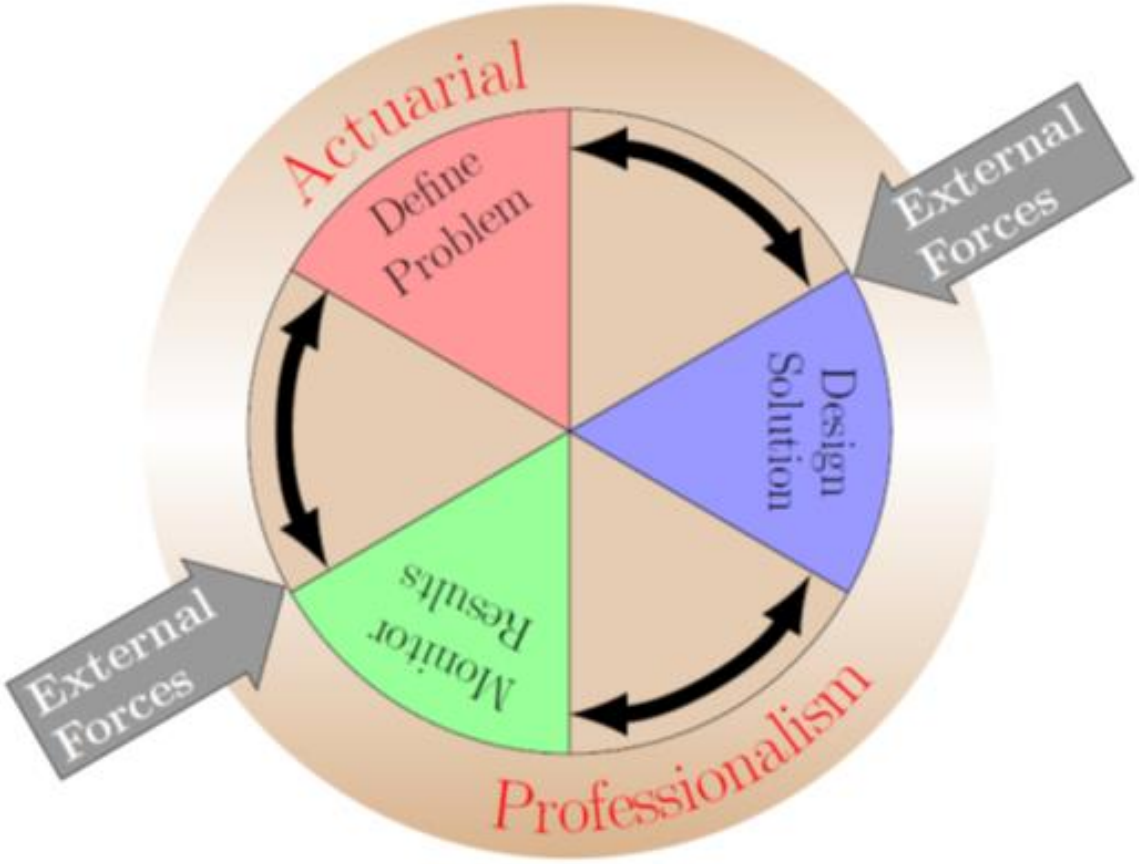
MDEC®

kaggle™

Actuary and Data Scientist

- Similar objectives
- Similar skill-sets
- Similar techniques
- Both are Data Dependent

Actuarial Control Cycle



5

Data Scientist's Workflow

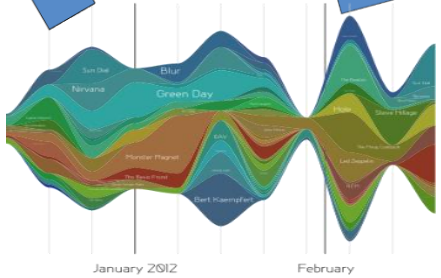
Digging Around in Data

Clean, prep

$$\begin{bmatrix} \cos 90^\circ & \sin 90^\circ \\ -\sin 90^\circ & \cos 90^\circ \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

Hypothesize Model

Large Scale Exploitation



Evaluate Interpret

Actuary VS Data Scientist

Actuary	Data Scientist
Primarily in insurance industry	Applicable to any industry
Formal qualifications	Still no formal qualification
SAS, Excel, VBA, SQL, MoSes, Prophet	R, Python and NoSQL databases (Hadoop, etc)
Directly impacts company financials	Impact in the long term

The Future

- Which profession has a better future? Actuary VS Data Scientist

- Regardless of what the future holds, predictive analytics & big data is the place to be nowadays!

R studio

- R Studio is a free and open-source integrated development environment (IDE) for R.
- R is a programming language for statistical computing and graphics.
- Commonly used for data exploration.