



Data Science for Business - Cambridge Analytica Case Demo

Jung PARK, PhD Research Fellow in Data Science

[Draft] Last modified 2019 Dec 10

Distinct Research Fields in Al/ Data Science

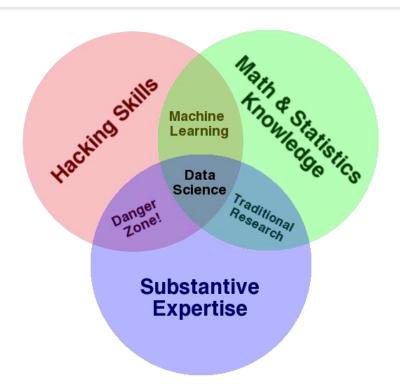


Attitude to Al	Al as a goal	Al as a tool	Al as a phenomenon		
Expertise	Computer science	Statistics	Social science		
Roles	Develop AI and machine learning algorithms	Use big data and machine learning techniques to solve research questions	Discuss the impact of AI on the society and organizations		
	Advance IT technologies for higher computing power and connectivity	Add values/ create insights using data	Understand the potential changes in ethics and policies		

... New technologies around AI, Data Science and Machine Learning

Multidisciplinary field





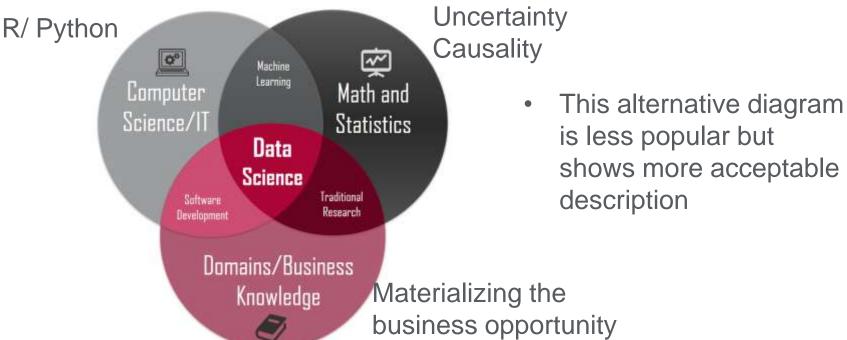
- Data Science is a multidisciplinary field that combines information technology (IT), statistics and management study.
- Due to the rapid advancement of IT, much more data and new analytic techniques became available.
- We need to balance it with sound statistical knowledge and business expertise to create useful insights.

Data science venn diagram

Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Multidisciplinary field – alternative description

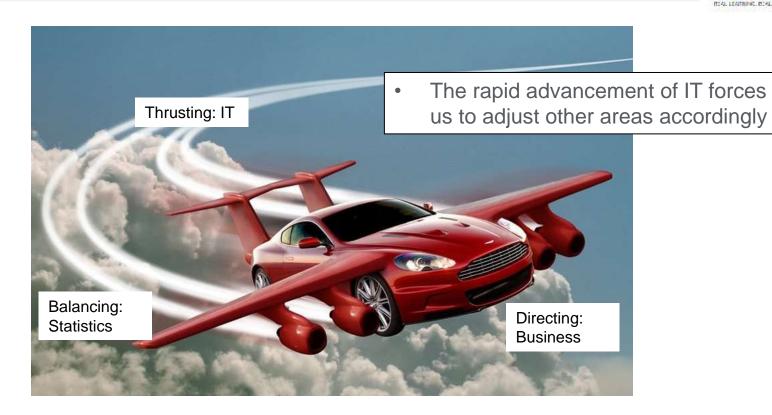




https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Data Science: analogy to a flying vehicle





Source: https://sourceable.net/degree-in-flying-cars-coming-soon/

Five processes in Data Science



We will use these five processes as a framework

Source: George et al. (2016). Editorial. *Academy of Management Journal*

Process	Challenges	Solutions	Key references
Data account and collection	Easy access to data offered in standardized formats. No practi- cal limit to the size of these data	Sensors Web scraping	Chaffin et al. (2015) Sameiro and Bucklin (2004)
	offering unlimited scalability • Efficiently obtain detailed data for a large number of agents	Web traffic and communications munitoring	
	 Protocols on security, privacy, and data rights 		
Data etorage	Tools for data storage, matching	• SQL, NoSQL, Apache Hadoop	• Varian (2014)
	and integration of different big datasets	Save essential information only and update in real time	Prajapati (2013)
	Data reliability		
	 Warehousing 		
Data processing	 Use non-numeric data for quanti- tative analyses 	 Text mining tools to transform text into numbers 	 Manning, Raghavan, and Schiltze (2009)
		• Emotion recognition	 Teixeira, Wedel, and Pieters (2012)
Data analysis			
	 Large number of variables Consulity 	 Ridge, Iasso, principal compo- nents regression, partial least aquares, regression troce 	 Hastie, Tibshirani, and Friedman (2009).
	 Find latent topics and attach 	Topic modeling, latent	George and McCulloch (1993) Asshed, Observer of Socienting
	Data too large to process	Dirichlet allocation, entropy- based measures, and deep	 Archak, Ghose, and speirotis (2011)
		learning	 Tirunillai and Tellis (2012)
		 Cross-validation and holdout samples 	 Blei, Ng, and Jordan (2003) LeCan, Beegin, and Hinton
		Field experiments	(2015) • Lambrecht and Tucker (2013)
		 Parallelization, bags of little boot- strup, sequential analysis 	Wang, Chen, Schifano, Wu, and Yan (2015)
			• Wodel and Kannan (2016)
Reporting and	TO SHARE THE STATE OF	1940 (1944) 1.4 (1945) 1.5 (1945)	november of the second
visualization	 Facilitate interpretation, repre- sentation with external partners 	Describe data sources Describe methods and	 Loughran and McDonald (2011)
	and knowledge users	specifications	· Simonsohn, Simmons, and
	 Difficult to understand complex patterns 	Bayesian analysis	Nelsem (2015)
		 Visualization and graphic interpretations 	

Five processes in Data Science



Access & Collect

Store

Process

Analyse

Report

Sensors

Web scraping

Web traffic and communications monitoring

SQL, NoSQL, Apache Hadoop

Save essential information only and update in real time

Text mining tools to transform text into numbers

Emotion recognition

Ridge, lasso, principal components regression, partial least squares, regression trees

Deep learning

. . .

Visualization and graphic interpretation

Robustness check

Open Science

Five processes in Experimental Research in Thermodynamics



Access & Collect

Store

Process

Analyse

Report

aranhic

Visualization and

Sensors

Web scraping

Web traffic and communications monitoring

SQL, NoSQL, Apache Hadoop

Save essential information only and update in real time

Text mining tools to transform text into numbers

Emotion recognition

Ridge, lasso, principal components regression, partial least squares, regression trees

ep learning



Excel

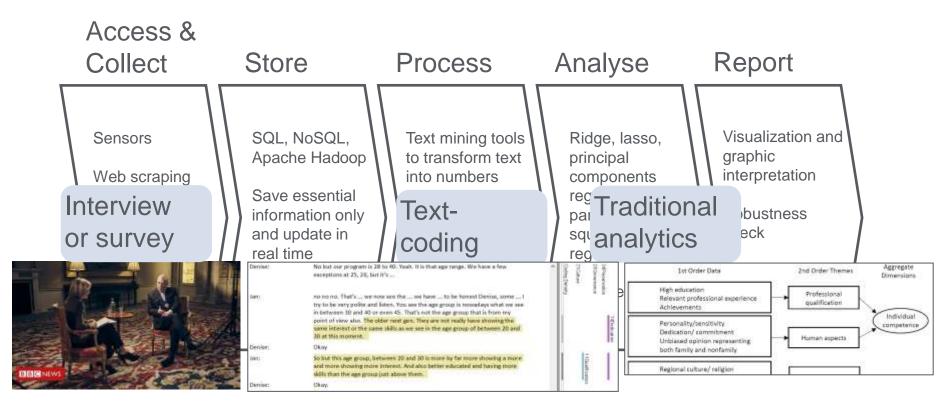
Matlab



17 - 25°C | 1, 20°C | 1, 2

Five processes in Qualitative Management Research







New Data is New Opportunity

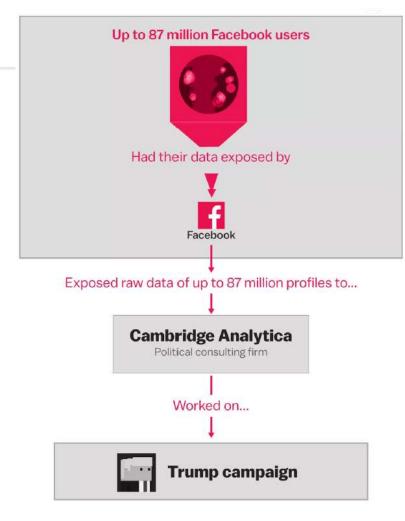
Case: Cambridge Analytica Scandal





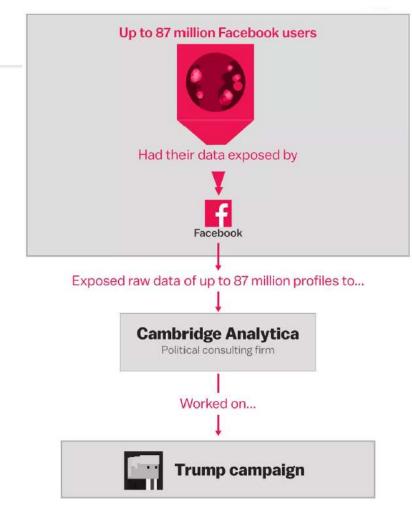
What happened?

Facebook exposed data on up to 87 million Facebook users to a researcher who worked at Cambridge Analytica, which worked for the Trump campaign.



What happened?

Facebook exposed data on up to 87 million Facebook users to a researcher who worked at Cambridge Analytica, which worked for the Trump campaign.





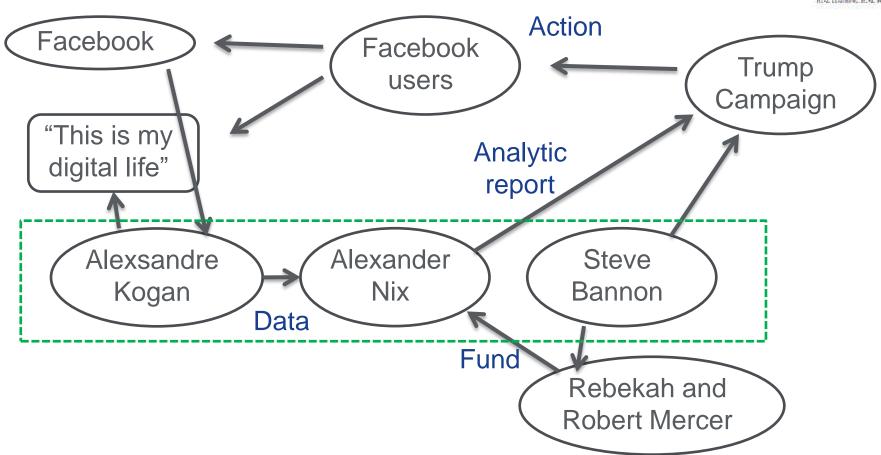
Cambridge Analytica

A scientist used to work at Cambridge University

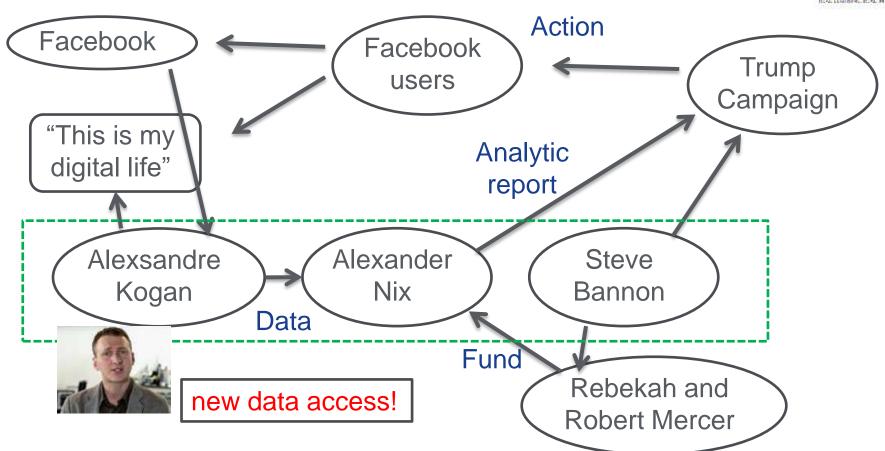
Latin/ Greek word for "Analytics" - more "sciency" cool sound to it

- A subsidiary of SCL (Strategic Communication Laboratories)
 which uses the study of mass behaviour and how to change it
 for commercial, military and political purposes
- SCL formed CA to participate in the election process in the United States

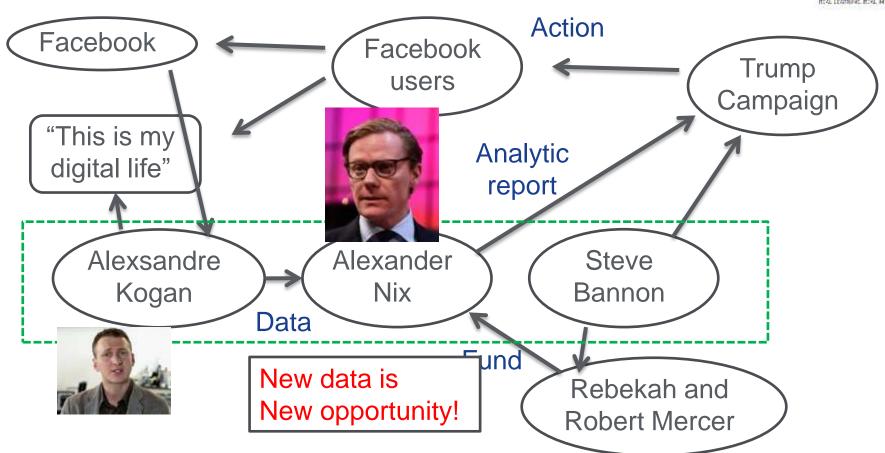




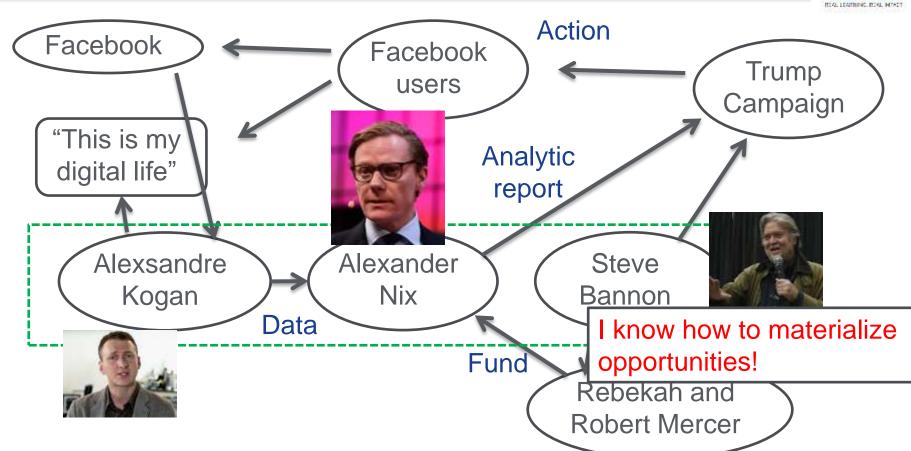


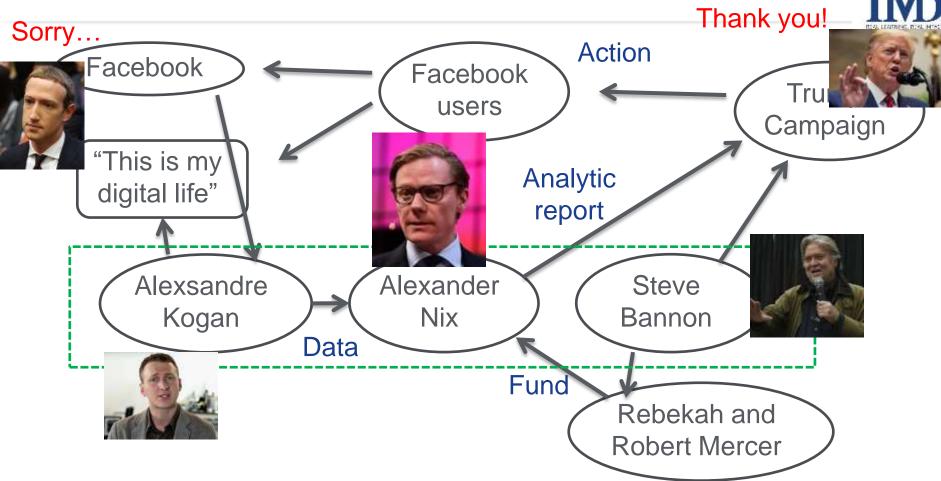






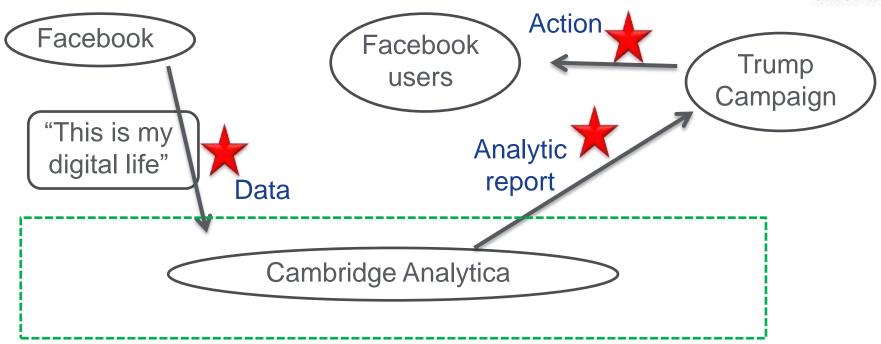






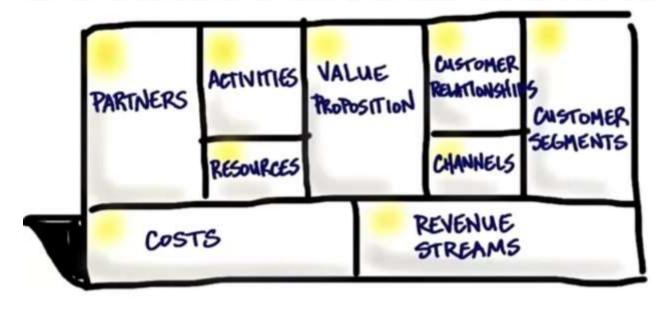
Our focus is on Data Science







BUSINESS MODEL CANVAS



If this were Ethics discussion...





Our focus is on Data Science









Store

Process

Analyse

Report

Sensors

Web scraping

Web traffic and communications monitoring

SQL, NoSQL, Apache Hadoop

Save essential information only and update in real time

Text mining tools to transform text into numbers

Emotion recognition

Ridge, lasso, principal components regression, partial least squares, regression trees

Deep learning

- - -

Visualization and graphic interpretation

Robustness check

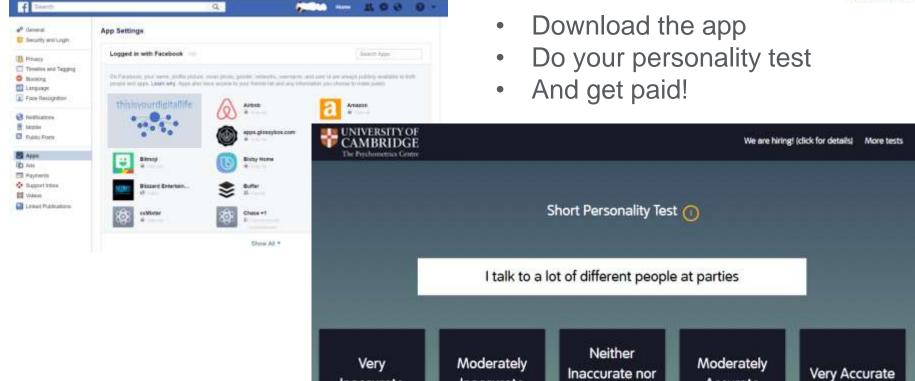
Open Science

[Data collect] "This is my digital life" App



Suggest a Question | Report

Accurate



© IMD 2019

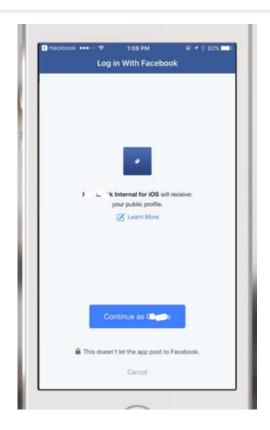
Inaccurate

Inaccurate

Accurate

[Data collect] "This is my digital life" App





An example of what Facebook Login looks like. Facebook

To be paid, login with Facebook



Facebook has allowed third-party app developers to access some private user data since May 2007, when it first opened the Facebook platform.

[Data collect] Using API (Application Programming Interface)



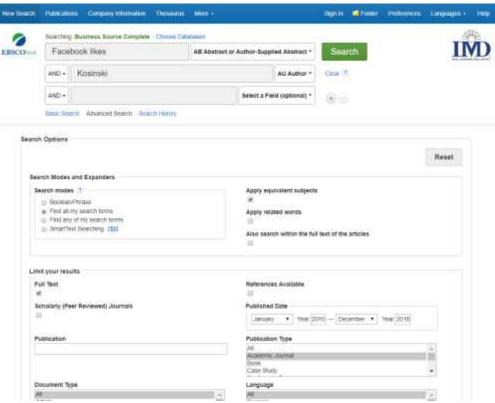
 (Over-)Simplified definition of API is: a pre-defined way for external users to collect data from a website's database

- Websites like Facebook and Wikipedia provide API to encourages developing a third-party software providing additional services based on the company's product
- Also, providing API can avoid the abuse of data access using a risky (hacking or overloading) method to the server

[Data collect] Example of E-Library API



We need to fill in the form to request the relevant information from the library database



[Data collect] Example of using API ex. R code for collecting table information from Wikipedia





Output

•	key	value
1	Full name	IMD
2	Logo of IMD	Logo of IMD
3	Туре	Private Business School
4	Established	1990
5	President	Jean-François Manzoni
6	Academic staff	50
7	Administrative staff	300
8	Students	8,900
9	Other students	90
10	Location	Lausanne, Vaud, Switzerland46°31'N 6°37'E <u+feff> / <u+< th=""></u+<></u+feff>
11	Campus	Urban
12	Website	http://www.imd.org/
13	Alumni: 100,000	Alumni: 100,000

[Data collect] Example of using API ex. R code for collecting table information from Wikipedia

get infobox(title = "International Institute for Management Development")



```
# R code
library(httr)
library(rvest)
library(xml2)
                                                                         Call some packages

Define a function get_infobox
get infobox <- function(title){
 base url <- "https://en.wikipedia.org/w/api.php"
 query params <- list(action = "parse",
                                                                         We will use wikipedia api
            page = title,
            format = "xml")
 resp <- GET(url = base_url, query = query_params)
 resp xml <- content(resp)
                                                                         Send query parameters and receive
 page html <- read html(xml text(resp xml))
                                                                         data
 infobox element <- html node(x = page html, css =".infobox")
 page name <- html node(x = infobox element, css = ".fn")
 page title <- html text(page name)
 wiki table <- html table(infobox element)
                                                                         Rearrange the data into a table form
 colnames(wiki_table) <- c("key", "value")
 cleaned table <- subset(wiki table.!wiki table$kev == "")
 name df <- data.frame(key = "Full name", value = page title)
 wiki table <- rbind(name df, cleaned table)
                                                                         Call the function for "IMD" wikipage
 wiki table
```

[Data collect] Data collected from "This is my digital life" App



270k users'

Facebook likes

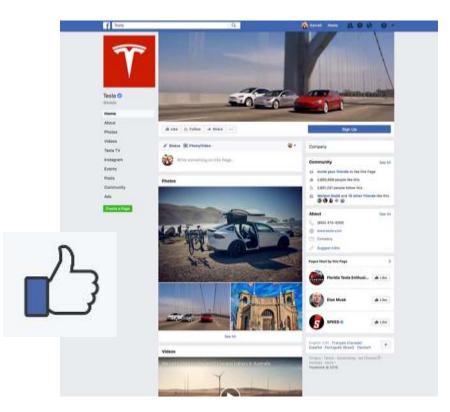
& Personality test results

87mil users' (friends)

Facebook likes

& ...

+ Facebook profiles (names, address, etc)



Our focus is on Data Science







Access & Collect



Process Analyse

Report

Sensors

Web scraping

Web traffic and communications monitoring

SQL, NoSQL, Apache Hadoop

Store

Save essential information only and update in real time

Text mining tools to transform text into numbers

Emotion recognition

Ridge, lasso, principal components regression, partial least squares, regression trees

Deep learning

- -

Visualization and graphic interpretation

Robustness check

Open Science

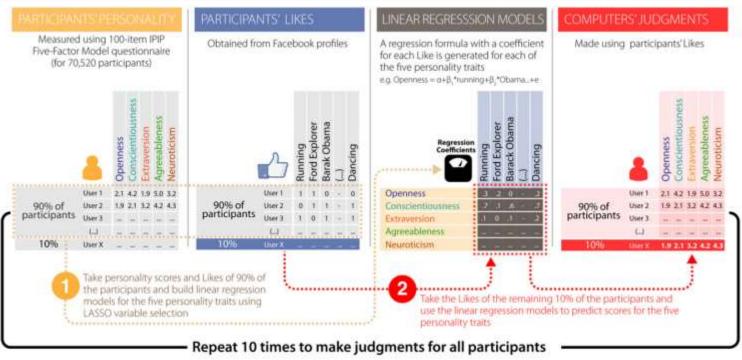
[Data Analysis] Psychographics – Personality Test Results





[Data Analysis] Regression models







Youyou, W., **Kosinski**, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4), 1036-1040.

Demo: 1) Make you "pseudo-Facebook Likes"



- 1. Go to bit.ly/imddemoca
- 2. Enter any nick name: hulk, hermione, wolverine,...
- 3. Click "Likes" or skip for the 10 postings as you feel like
- 4. Enter age and gender



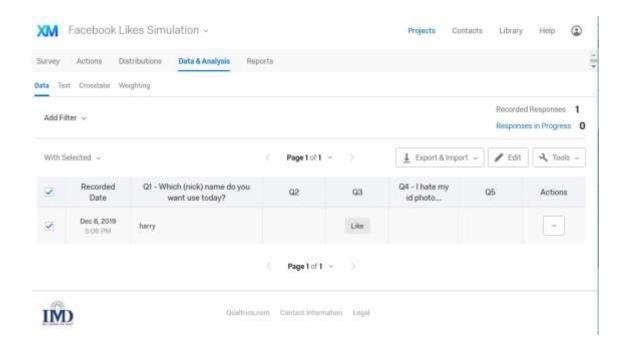
Interesting and relevant to you?



or SKIP

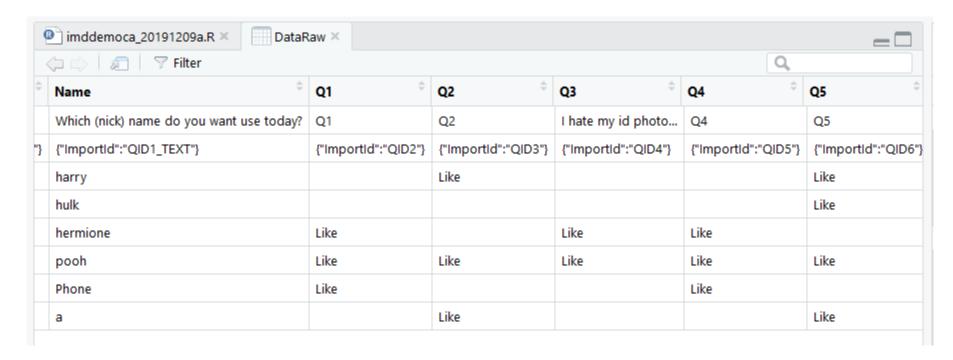
Demo: 2) Downloading the data





Demo: 3) Import Data





Demo: 4) Process the data, Do regression and Predict Age/ Gender



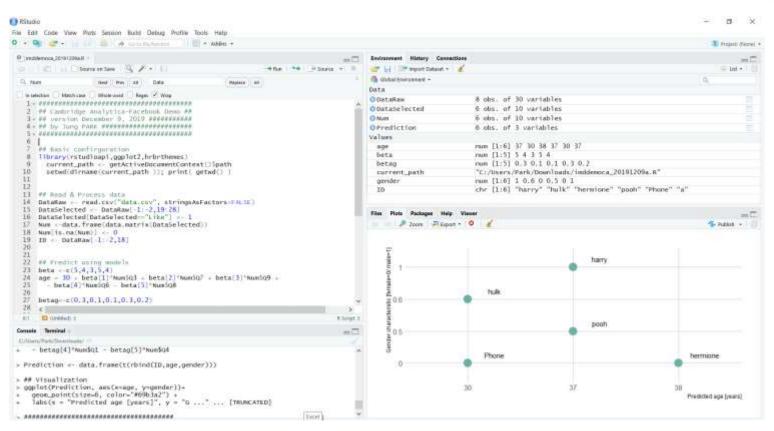
Q1 [‡]	Q2 [‡]	Q3 [‡]	Q4 [‡]	Q 5 [‡]	Q6 [‡]	Q7 [‡]	Q8 [‡]	Q9 [‡]	Q10 [‡]
0	1	0	0	1	0	1	0	1	1
0	0	0	0	1	0	1	1	0	0
1	0	1	1	0	0	0	0	1	0
1	1	1	1	1	1	1	0	1	1
1	0	0	1	0	0	1	1	0	0
0	1	0	0	1	0	1	0	1	1

Age =
$$\alpha + \beta_1 Q_1 + \beta_2 Q_2 + ... + \beta_{10} Q_{10}$$

Ex. Age = 30 + (Fix-it and forget-it) + (Small Business Sat)
- (I hate my id photo) - (Dude... wait) - (Because I am a girl)

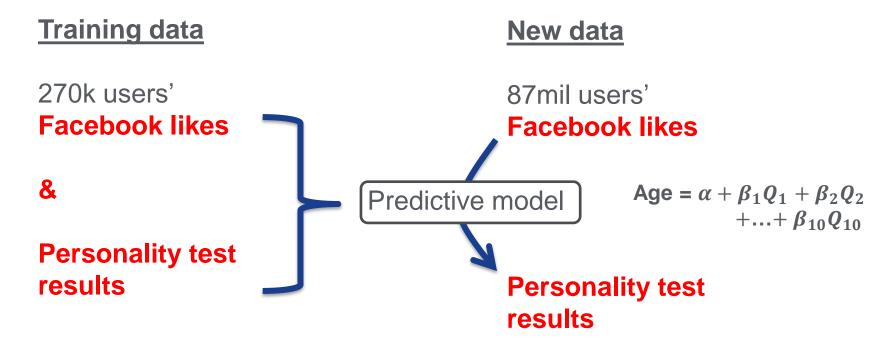
Demo: 4) Plot graphs





[Data Analysis] Regression models





Selected most predictive Likes

Politics Republican		George W Bush	Joe Biden Speaker Nancy Pelosi		
		John McCain			
		Conservative	Health Care Reform		
	u	Rush Limbaugh	The White House		
	ica	Sean Hannity	Democrats		
	Iqi	Bill Oreilly	Barbara Boxer		
	abı	Positively Republican	Anthony Weiner		
	R	Sarah Palin	Being Liberal		
		Ronald Reagan	Left Action		
		Glenn Beck	Barack Obama2012		
			Ted Kennedy		

Some obvious words

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802-5805.

Selected most predictive Likes

Openness Liberal & Artistic	Oscar Wilde	NASCAR		
		Charles Bukowski	Austin Collie	
	stic	Sylvia Plath	Monster-In-Law	
	rii	Leonardo Da Vinc	I don't read	
	4	Bauhaus	Justin Moore	
		Dmt The Spirit Molecule	ESPN2	
	era	American Gods	Farmlandia	
	John Waters	The Bachelor		
	Plato	Oklahoma State University		
	Leonard Cohen	Teen Mom 2		

Some less obvious words based on US data

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802-5805.

Our focus is on Data Science









Access & Collect

Store

Process

Analyse

Report

Sensors

Web scraping

Web traffic and communications monitoring

SQL, NoSQL, Apache Hadoop

Save essential information only and update in real time

Text mining tools to transform text into numbers

Emotion recognition

Ridge, lasso, principal components regression, partial least squares, regression trees

Deep learning

- -

Visualization and graphic interpretation

Robustness check

Open Science

© IMD 2019

[Use of the analytics] Psychographics – Personality Test Results





[Use of the analytics] choosing the most effective message

END THE RIGHT TO BEAR ARMS



Anxious type

would respond to messages highlighting the threat of a break-in

High Neuroticism

On gun rights issue

Psychographic Messaging



Traditional person

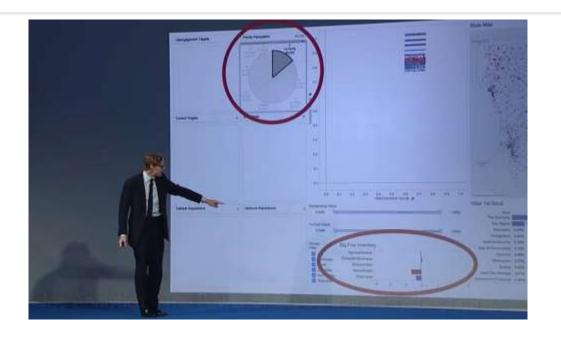
could be targeted with messages about a grandfather teaching subsequent generations to hunt

> https://www.youtube.com/watch?v= n8Dd5aVXLCc&t=126s

[Use of the analytics] finding the campaign targets



n8Dd5aVXLCc&t=126s



At the 2016 Concordia Annual Summit in New York, Mr. Alexander Nix discusses the power of big data in... behavioral microtargeting for election processes around the world.

https://www.youtube.com/watch?v=

© IMD 2019

[Use of the analytics] finding the campaign targets





https://www.toledoblade.com/Politics/2016/05/17/Campaigns-looking-to-technologyto-harvest-voters-habits-target-ads.html

Visit targets who

- 1) are most likely to vote
- 2) may swing

With the most effective messages for their personality types

Additional food of thoughts



 Do the psychographic prediction actually work well?
 How much contribution did Cambridge Analytica make to the republican's win?

• Why did Alexander Nix (CA's CEO) present CA's involvement in the election at several conferences and sales meetings? Didn't he expect that he could be accused of it?



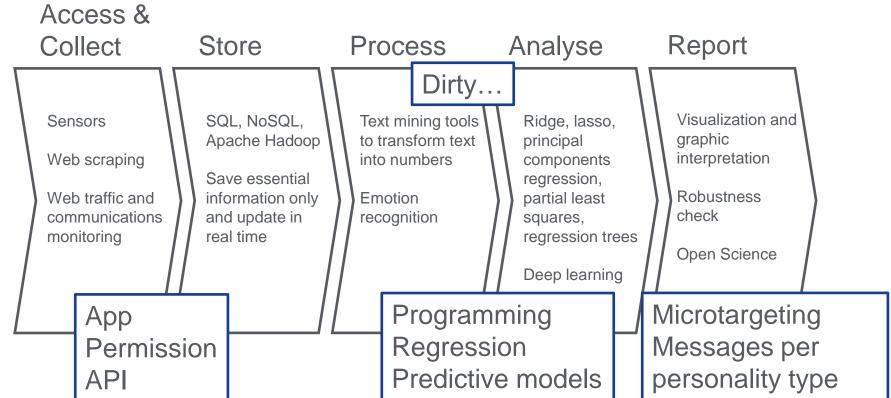
New Data is New Opportunity



New Data is New Opportunity and New Risk!

Any questions?





© IMD 2019



Appendix (in case of a technical problem)



 Q'



Q2





Q3



I hate my id photo...

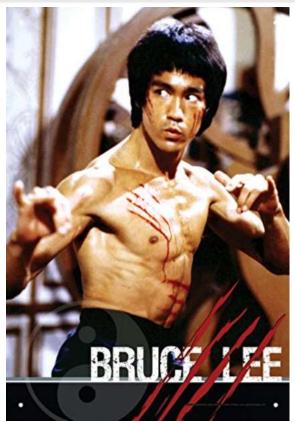
Q4



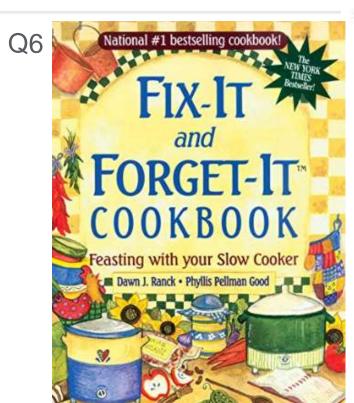
https://www.pinterest.ch/pin/349451252316384401

IND

Q5



https://www.amazon.in/Bruce-Lee-Fight-Tin-Sign/dp/B00MHRBKIA



https://www.amazon.com/Fix-Forget-Cookbook-Ranck-Hower-ebook/dp/B00RW2UVPS



Q7



Q8



Small Business Saturday

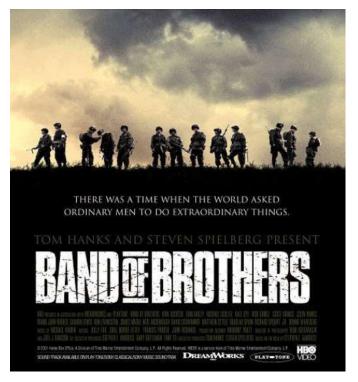


Q9



Because I am a girl movement

Q10



https://plan-international.org/because-i-am-a-girl

https://www.imdb.com/title/tt0185906/mediaviewer/rm1833146624

Simulating Facebook Likes: (over-)simplified version for hand calculation



Age =
$$30 + 5*(Q6+Q8 -Q3-Q7-Q9)$$

Gender =
$$0.5+ 0.2*(Q2+Q5+Q10$$

-Q1-Q4)

Likes = 1, Skip = 0