Data Science for Business
- Cambridge Analytica Case Demo

Jung PARK, PhD
Research Fellow in Data Science

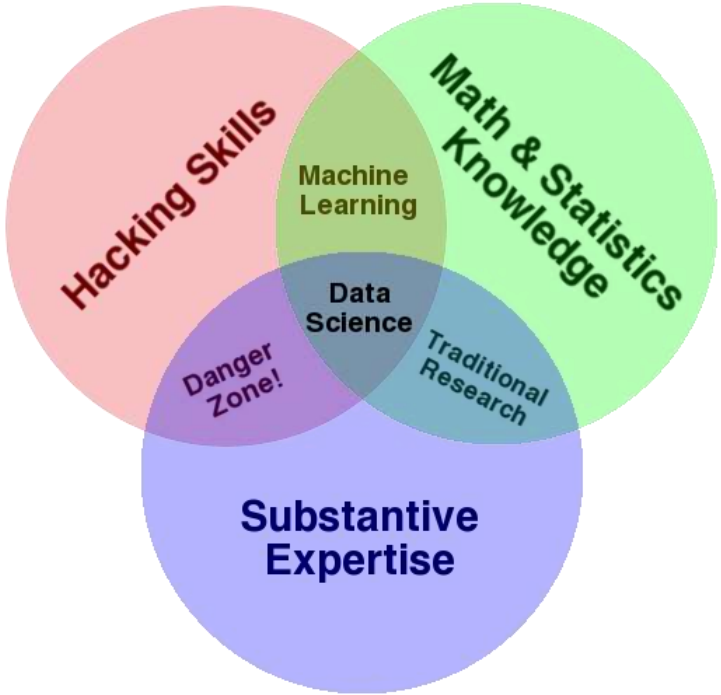
Last modified 2019 Dec 17

Distinct Research Fields in AI/ Data Science

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

…New technologies around AI, Data Science and Machine Learning

Multidisciplinary field

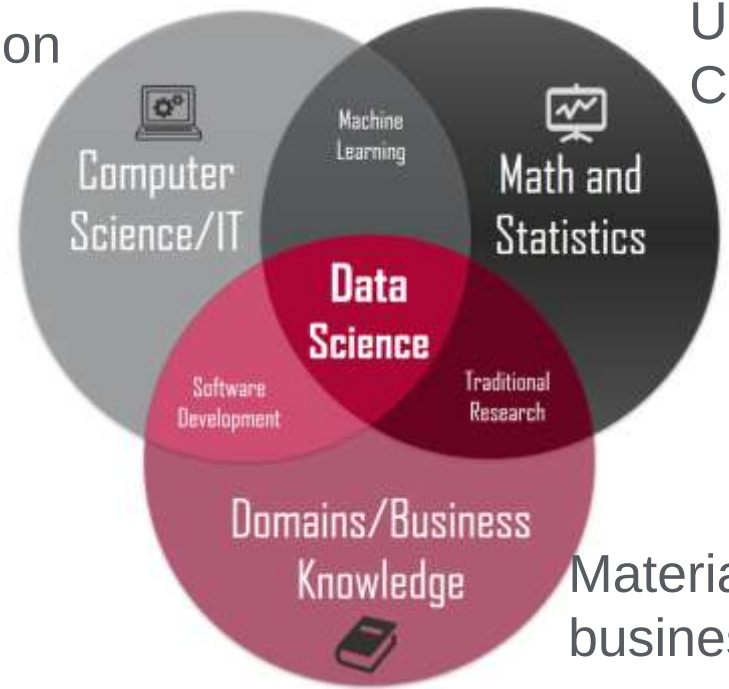


- Data Science is a multidisciplinary field that combines information technology (IT), statistics and management study.
- Due to the rapid advancement of IT, much more data and new analytic techniques became available.
- We need to balance it with sound statistical knowledge and business expertise to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

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

Multidisciplinary field – alternative description

R/ Python



Uncertainty
Causality

- This alternative diagram is less popular but shows more acceptable description

Materializing the business opportunity

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

Data Science: analogy to a flying vehicle

- The rapid advancement of IT forces us to adjust other areas accordingly

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

Five processes in Data Science

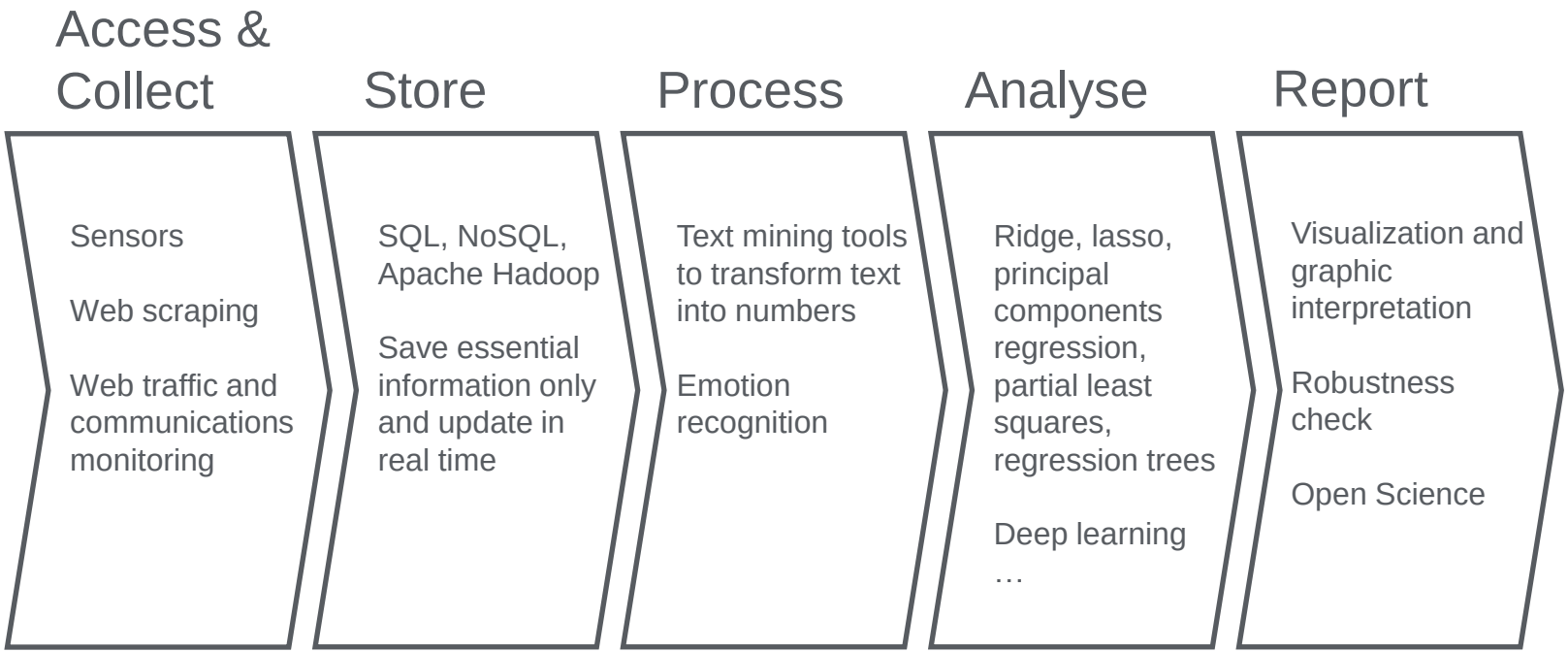
We will use these five processes as a framework

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

Big Data Challenges and Solutions

Process	Challenges	Solutions	Key references
Data access and collection	• Easy access to data offered in standardized formats. No practical limit to the size of these data offering unlimited scalability • Efficiently obtain detailed data for a large number of agents • Protocols on security, privacy, and data rights	• Sensors • Web scraping • Web traffic and communications monitoring	• Chaffin et al. (2015) • Sismeiro and Bucklin (2004)
Data storage	• Tools for data storage, matching and integration of different big datasets • Data reliability • Warehousing	• SQL, NoSQL, Apache Hadoop • Save essential information only and update in real time	• Varian (2014) • Prajapati (2013)
Data processing	• Use non-numeric data for quantitative analyses	• Text mining tools to transform text into numbers • Emotion recognition	• Manning, Raghavan, and Schütze (2009) • Teixeira, Wedel, and Pieters (2012)
Data analysis	• Large number of variables • Causality • Find latent topics and attach meaning • Data too large to process	• Ridge, lasso, principal components regression, partial least squares, regression trees • Topic modeling, latent Dirichlet allocation, entropy-based measures, and deep learning • Cross-validation and holdout samples • Field experiments • Parallelization, bags of little bootstrap, sequential analysis	• Hastie, Tibshirani, and Friedman (2009) • George and McCulloch (1993) • Archak, Ghose, and Ipeirotis (2011) • Tirunillai and Tellis (2012) • Blei, Ng, and Jordan (2003) • LeCun, Bengio, and Hinton (2015) • Lambrecht and Tucker (2013) • Wang, Chen, Schifano, Wu, and Yan (2015) • Wedel and Kannan (2016)
Reporting and visualization	• Facilitate interpretation, representation with external partners and knowledge users • Difficult to understand complex patterns	• Describe data sources • Describe methods and specifications • Bayesian analysis • Visualization and graphic interpretations	• Loughran and McDonald (2011) • Simonsohn, Simmons, and Nelson (2015)

Five processes in Data Science



Five processes in Experimental Research in Thermodynamics

Access & Collect

Sensors

Web scraping

Web traffic and communications monitoring

Store

SQL, NoSQL, Apache Hadoop

Save essential information only and update in real time

Process

Text mining tools to transform text into numbers

Emotion recognition

Analyse

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

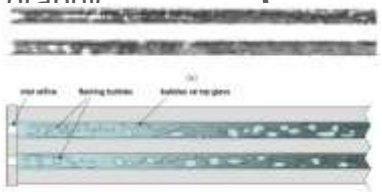
eep learning

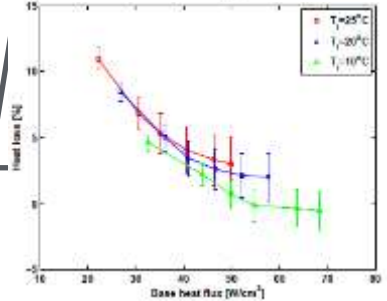
Excel

Matlab

Report

Visualization and graphic





Five processes in Qualitative Management Research

Access & Collect

Sensors

Web scraping

Interview or survey

Store

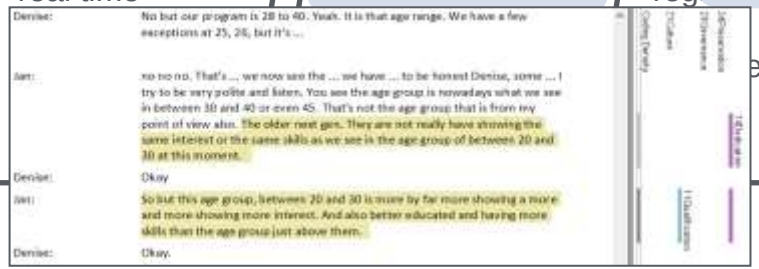
SQL, NoSQL, Apache Hadoop

Save essential information only and update in real time

Process

Text mining tools to transform text into numbers

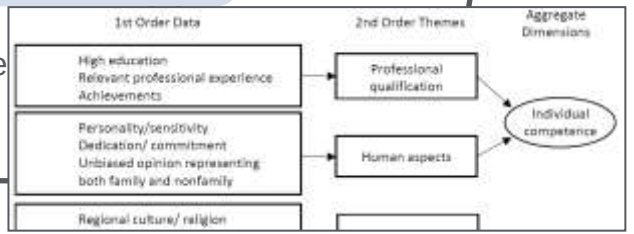
Text-coding



Analyse

Ridge, lasso, principal components reg
pa
squ
reg

Traditional analytics



Report

Visualization and graphic interpretation

obustness
eck

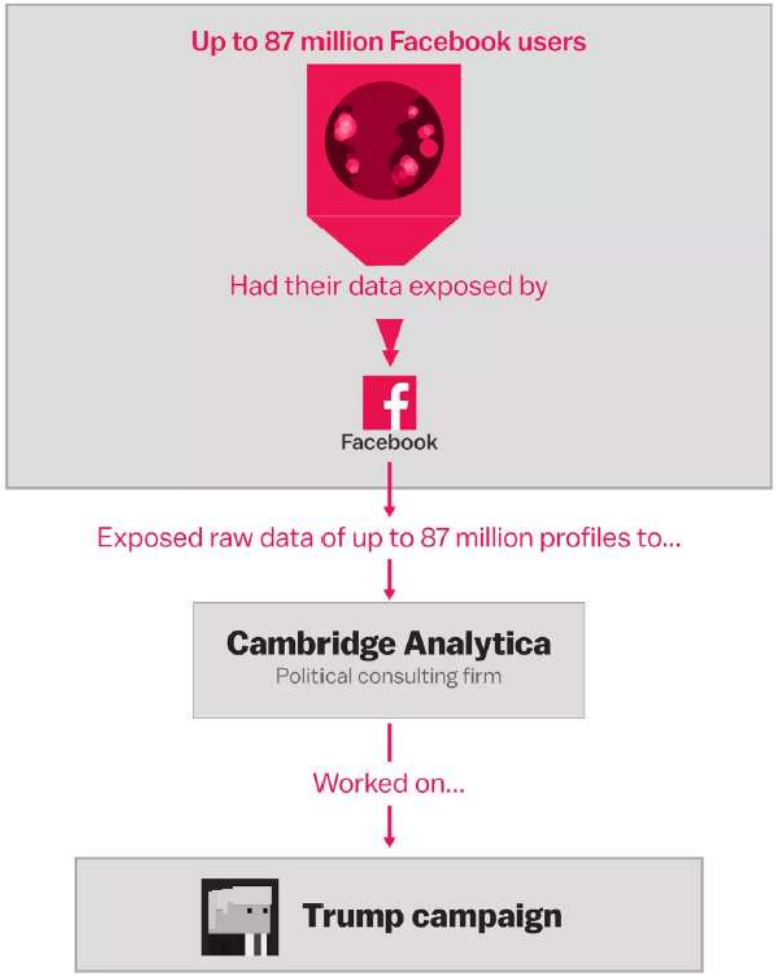
New Data is
New Opportunity

Case: Cambridge Analytica Scandal

Photographer: Chesnot/Getty Images & © Bloomberg

What happened?

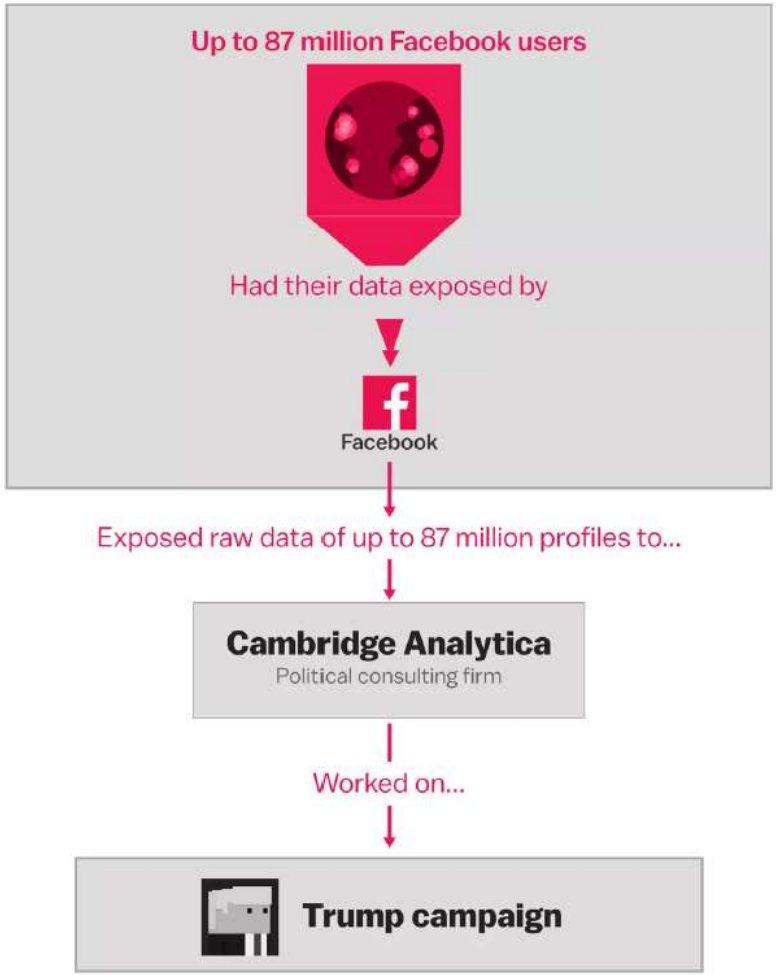
Facebook exposed data on up to 87 million Facebook users to a researcher who worked at Cambridge Analytica, which worked for the Trump campaign.



Chang, A. (2018). The Facebook and Cambridge Analytica scandal, explained with a simple diagram. *Vox, March*, *23*.

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.



Chang, A. (2018). The Facebook and Cambridge Analytica scandal, explained with a simple diagram. *Vox, March*, *23*.

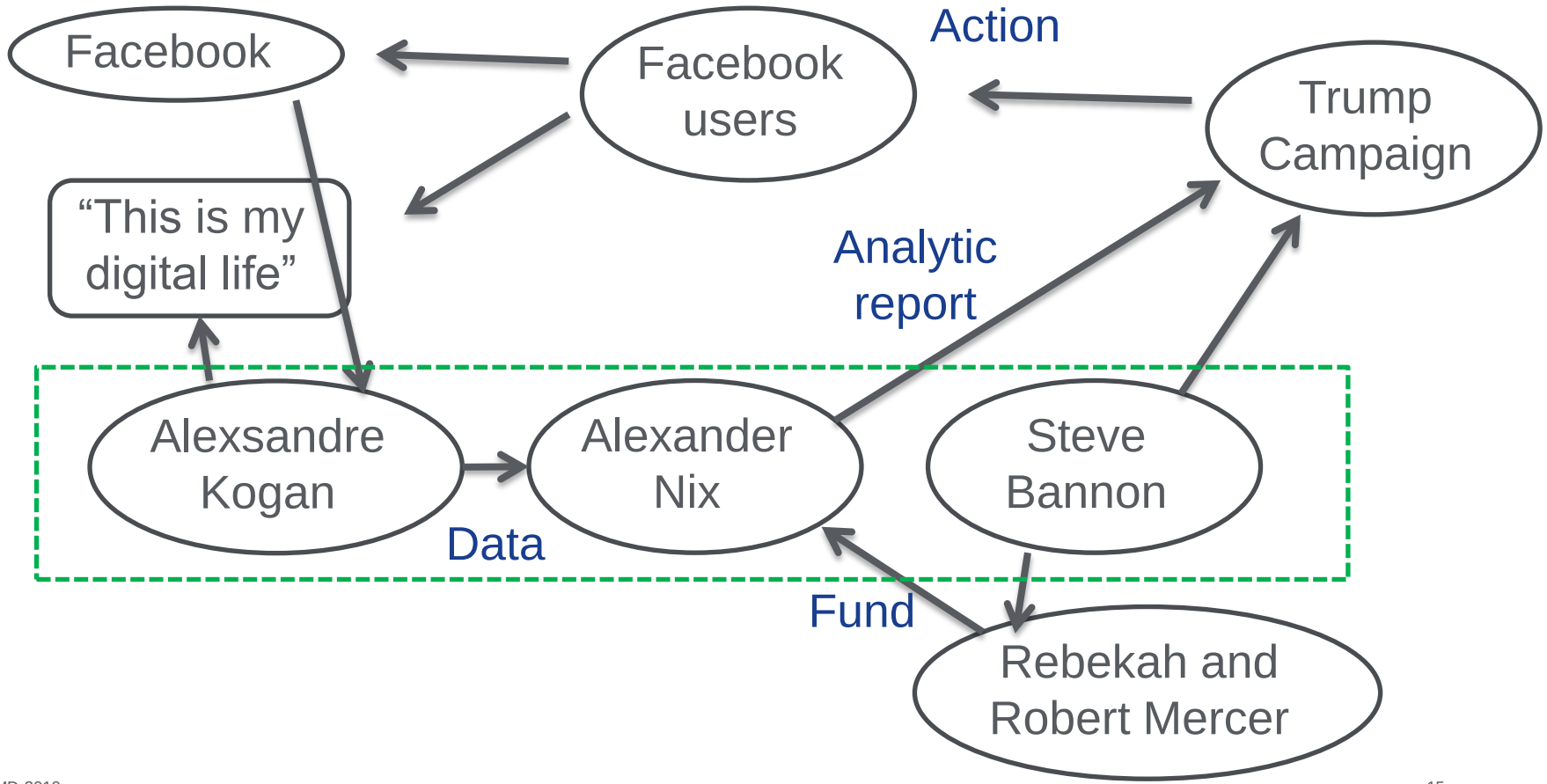
Cambridge Analytica

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

Complex relations



Complex relations

Facebook

Facebook users

Action

Trump Campaign

"This is my digital life"

Analytic report

Alexsandre Kogan

Alexander Nix

Steve Bannon

Data

new data access!

Fund

Rebekah and Robert Mercer

Complex relations

Facebook

Facebook users

Action

Trump Campaign

"This is my digital life"

Analytic report

Alexsandre Kogan

Alexander Nix

Steve Bannon

Data

Fund

New data is
New opportunity!

Rebekah and Robert Mercer

Complex relations

Facebook

Facebook users

Action

Trump Campaign

"This is my digital life"

Analytic report

Alexsandre Kogan

Alexander Nix

Steve Bannon

Data

Fund

I know how to materialize opportunities!

Rebekah and Robert Mercer

Complex relations

Sorry…

Thank you!

Facebook

Facebook users

Action

Tru… Campaign

"This is my digital life"

Analytic report

Alexsandre Kogan

Alexander Nix

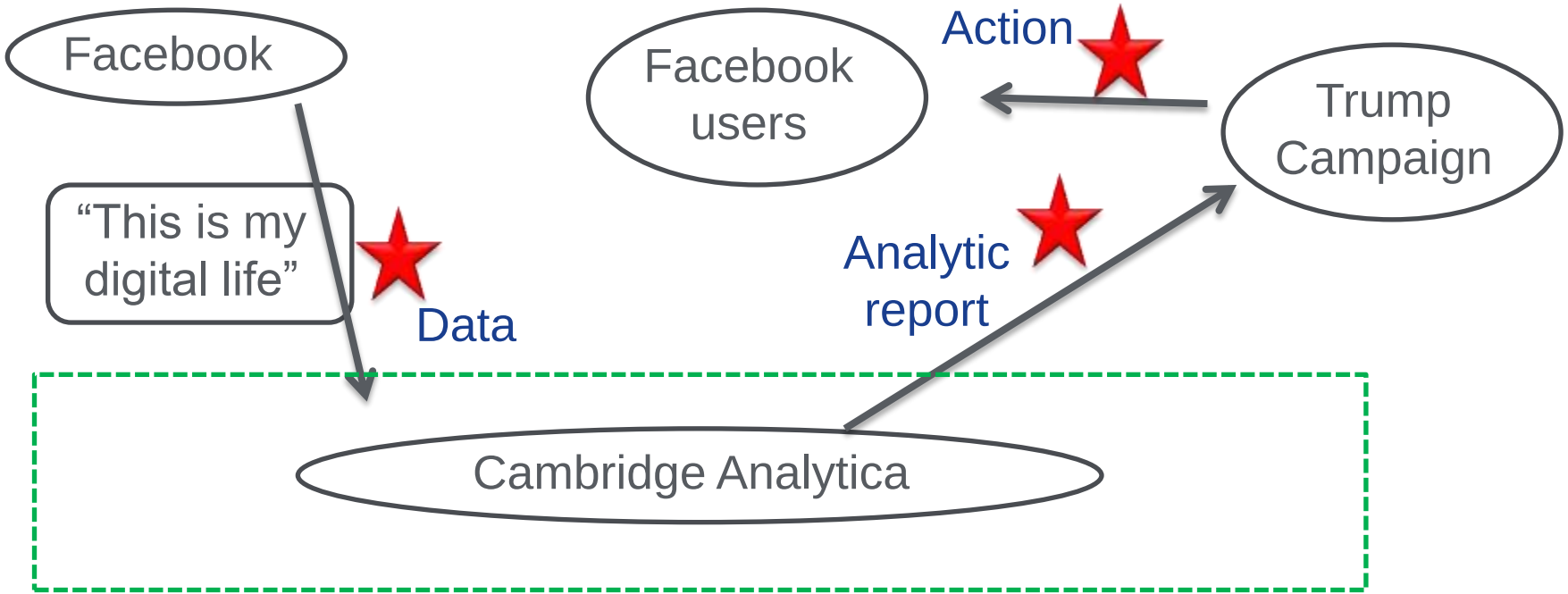
Steve Bannon

Data

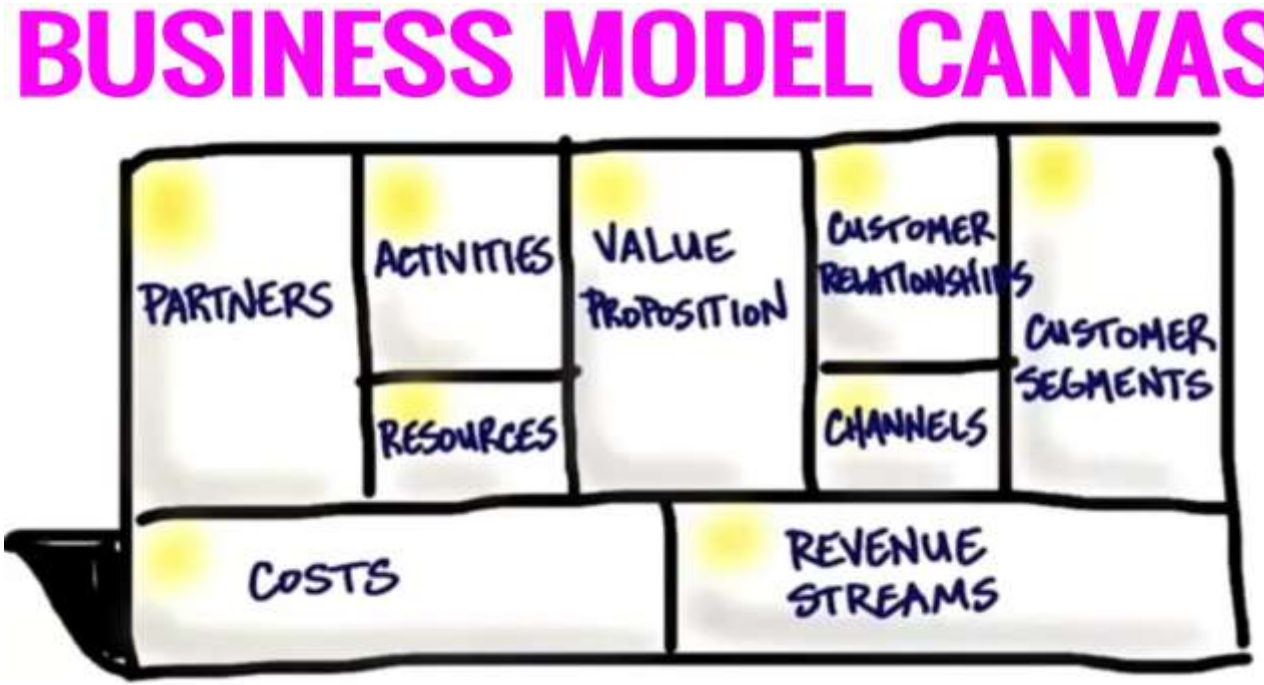
Fund

Rebekah and Robert Mercer

Our focus is on Data Science



If this were Entrepreneurship discussion…

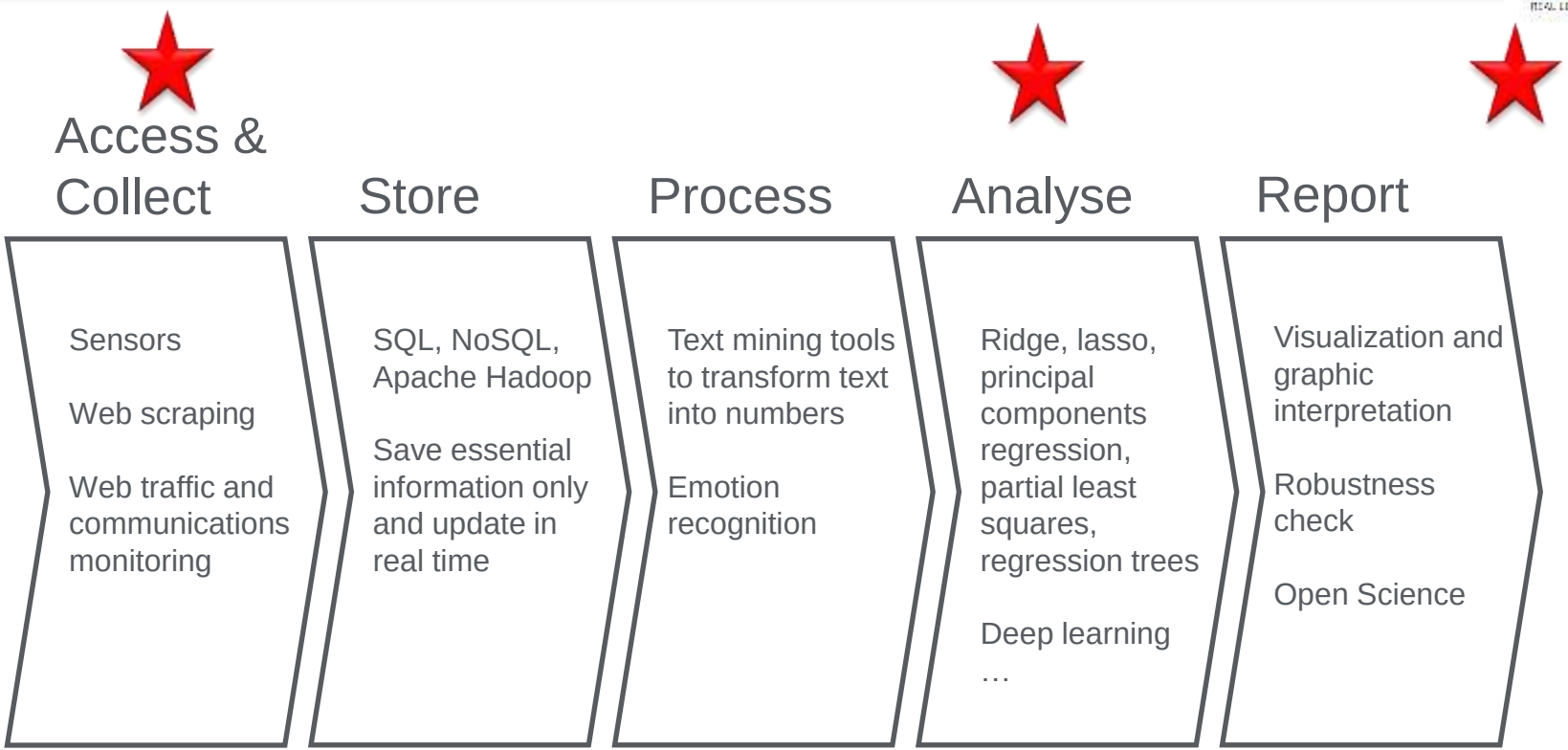


https://www.youtube.com/watch?v=IP0cUBWTgpY

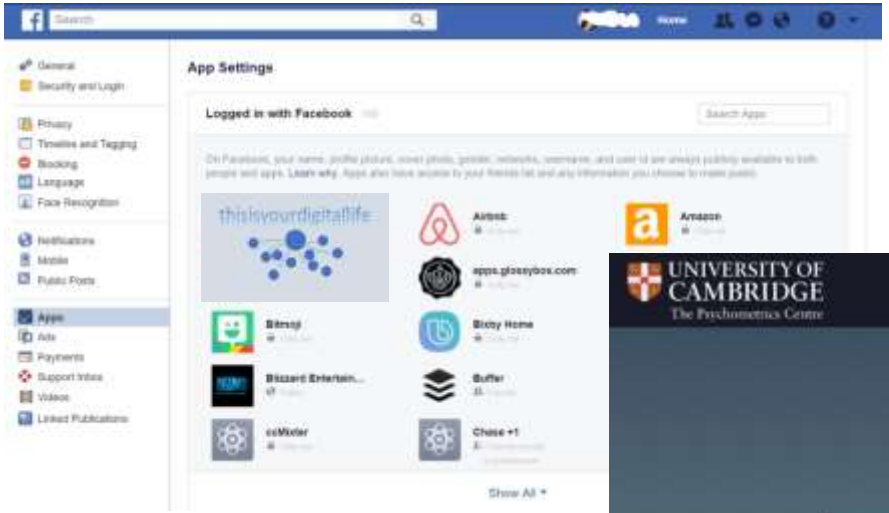
If this were Ethics discussion…

https://scrippsmediaethics.blogspot.com/2018/09/remaining-ethical-in-fake-news.html

Our focus is on Data Science

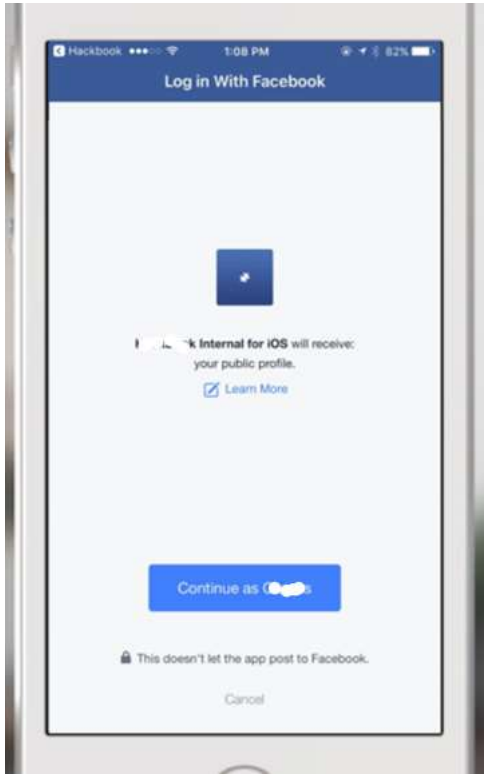


[Data collect] “This is my digital life” App



- Download the app
- Do your personality test
- And get paid!

[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)

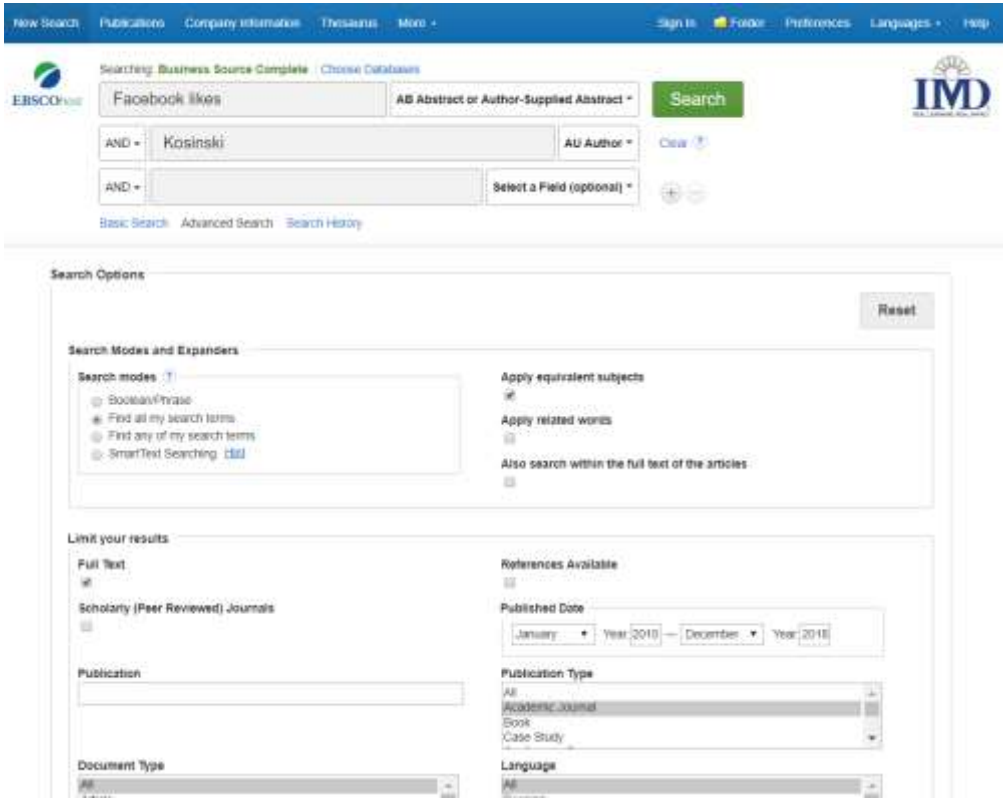
- (Over-)Simplified definition of API is: **a pre-defined way for external users to collect data from a website's database**

- 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	Type	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+...
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

```
# R code
library(httr)
library(rvest)
library(xml2)

get_infobox <- function(title){
  base_url <- "https://en.wikipedia.org/w/api.php"

  query_params <- list(action = "parse",
                       page = title,
                       format = "xml")

  resp <- GET(url = base_url, query = query_params)
  resp_xml <- content(resp)

  page_html <- read_html(xml_text(resp_xml))
  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)
  colnames(wiki_table) <- c("key", "value")
  cleaned_table <- subset(wiki_table, !wiki_table$key == "")
  name_df <- data.frame(key = "Full name", value = page_title)
  wiki_table <- rbind(name_df, cleaned_table)

  wiki_table
}

get_infobox(title = "International Institute for Management Development")
```

Call some packages
Define a function get_infobox
We will use wikipedia api

Send query parameters and receive data

Rearrange the data into a table form

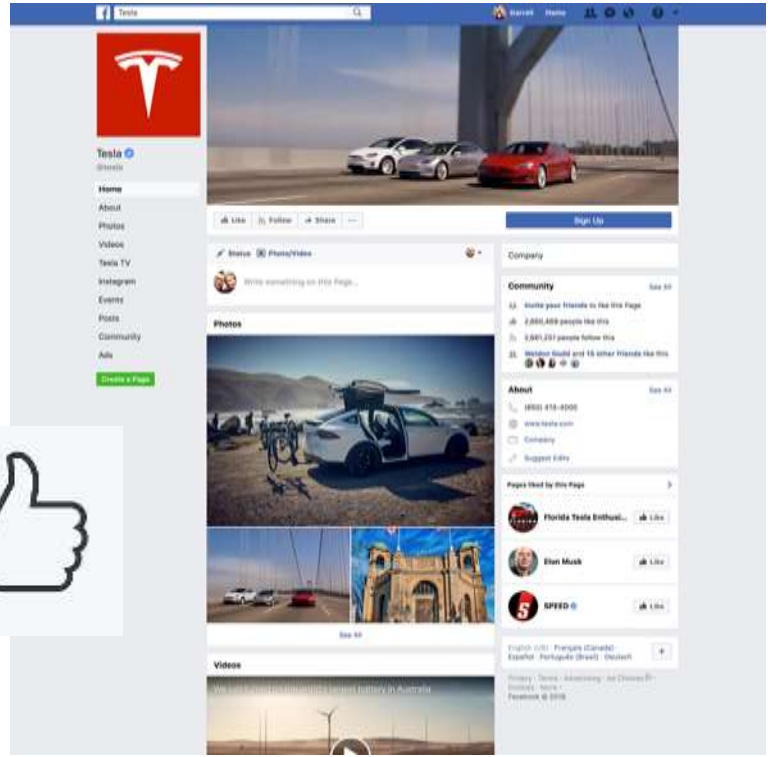
Call the function for “IMD” wikipage

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

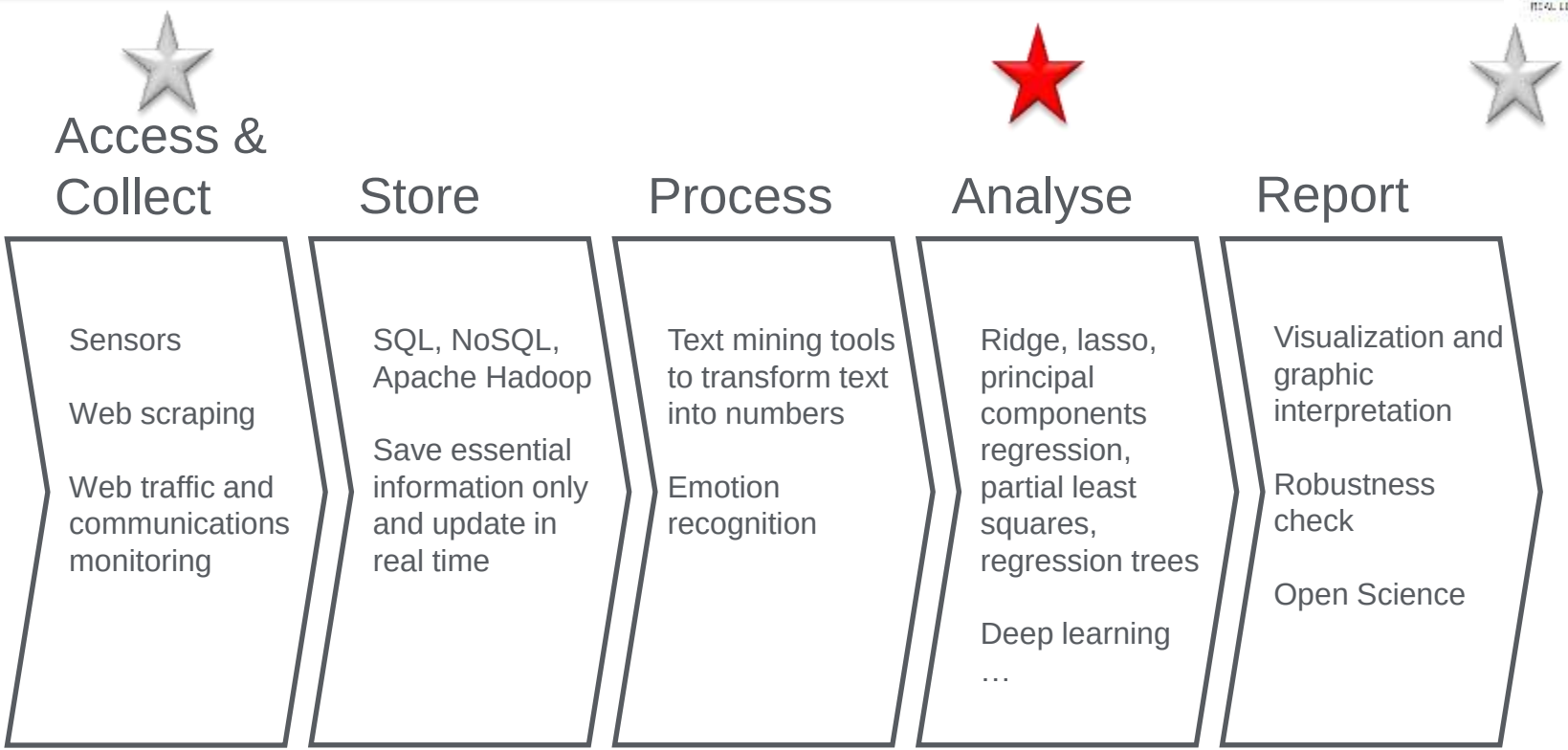
270k users’
Facebook likes
& Personality test results

87mil users’ (friends)
Facebook likes
& ...

+ Facebook profiles (names, address, etc)

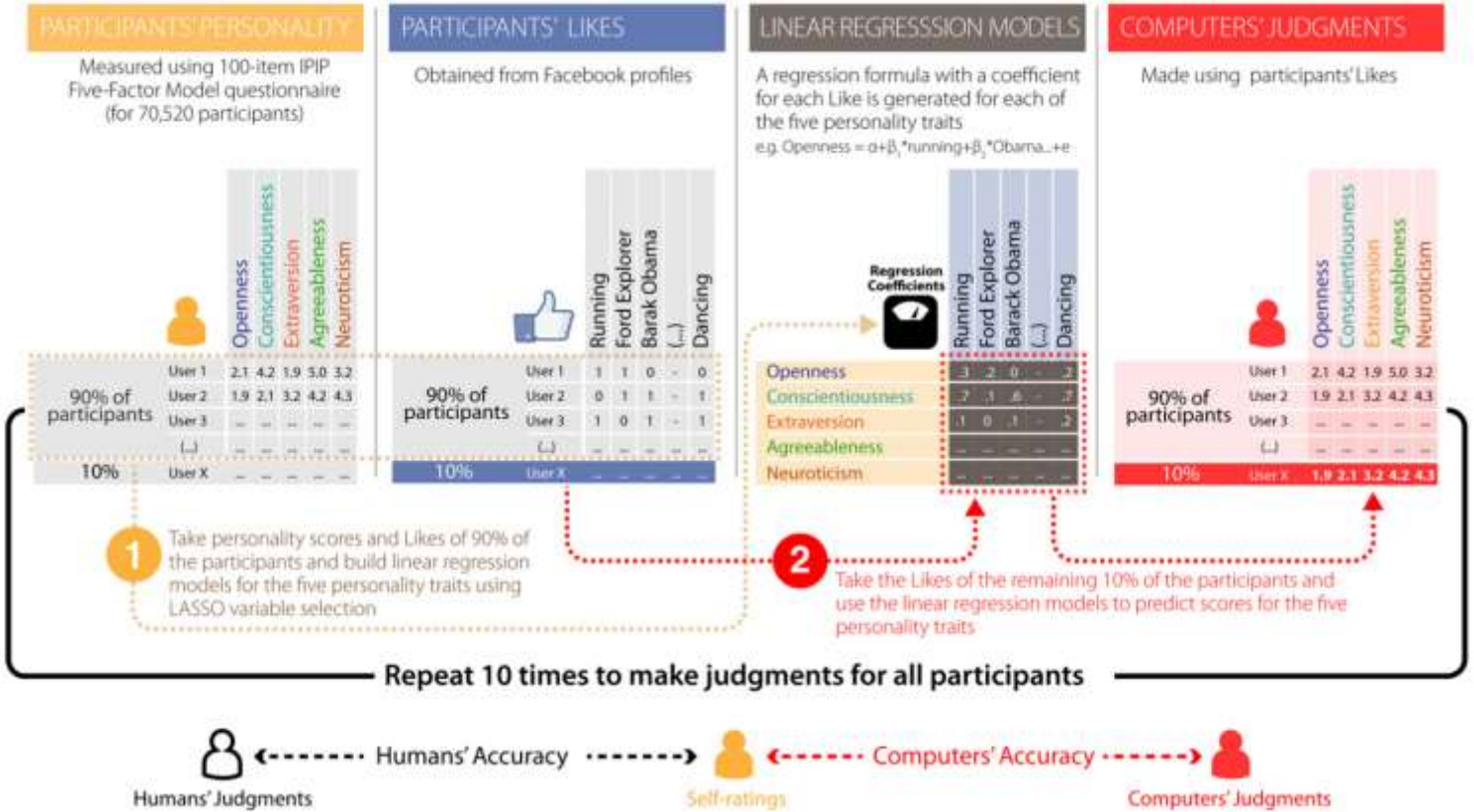


Our focus is on Data Science

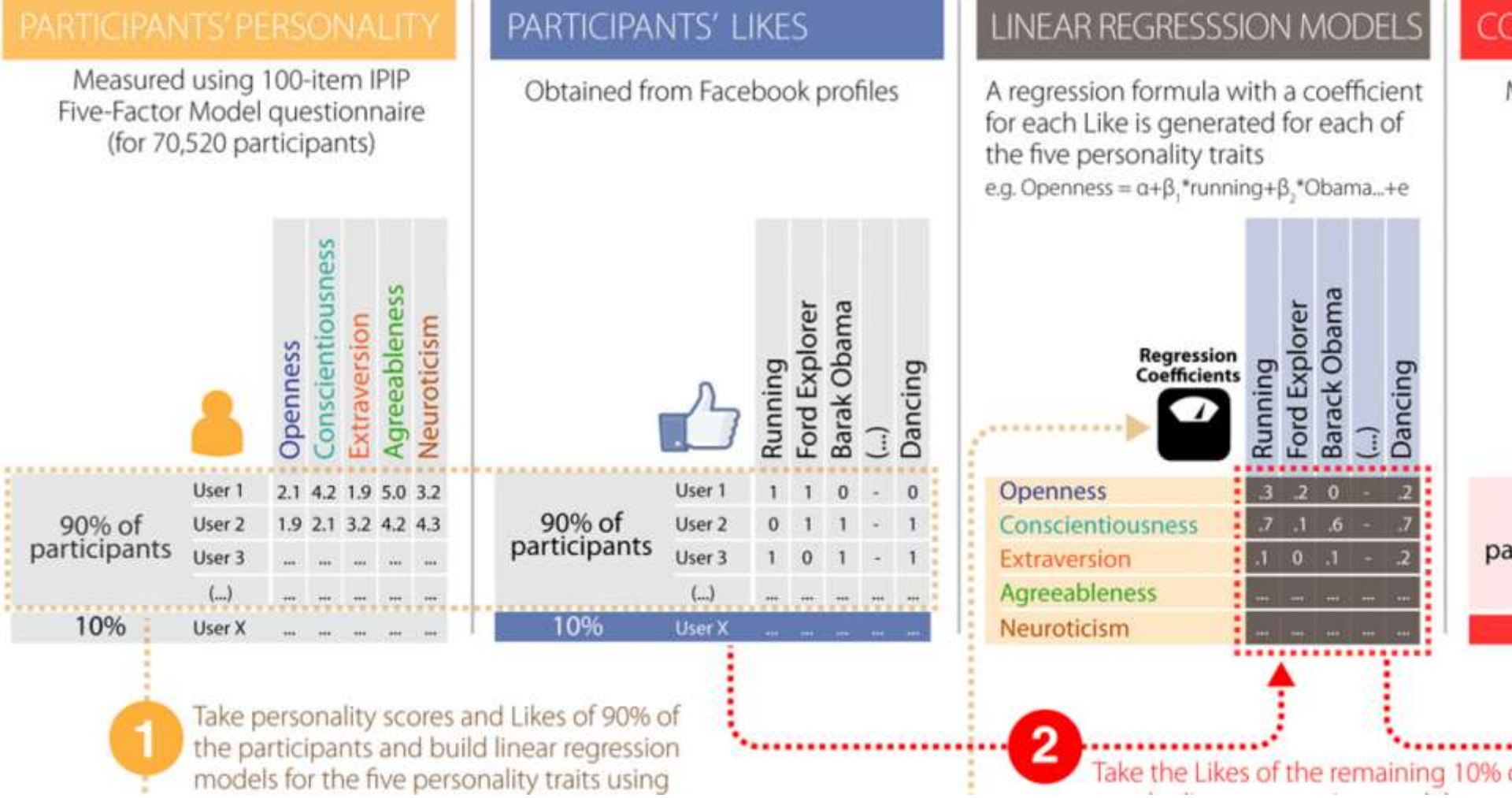


[Data Analysis] Psychographics – Personality Test Results

[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.


PARTICIPANTS' PERSONALITY
Measured using 100-item IPIP Five-Factor Model questionnaire (for 70,520 participants)
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism
90% of participants
User 1 2.1 4.2 1.9 5.0 3.2
User 2 1.9 2.1 3.2 4.2 4.3
User 3 ... ...
(...) ... ...
10%
User X ... ...
PARTICIPANTS' LIKES
Obtained from Facebook profiles
Running
Ford Explorer
Barak Obama
(...)
Dancing
90% of participants
User 1 1 1 0 - 0
User 2 0 1 1 - 1
User 3 1 0 1 - 1
(...) ... ...
10%
User X ... ...
LINEAR REGRESSSION MODELS
A regression formula with a coefficient for each Like is generated for each of the five personality traits
e.g. Openness = α+β1*running+β2*Obama...+e
Regression Coefficients
Running
Ford Explorer
Barack Obama
(...)
Dancing
Openness .3 .2 0 - .2
Conscientiousness .7 .1 .6 - .7
Extraversion .1 0 .1 - .2
Agreeableness ... ...
Neuroticism ... ...
1
Take personality scores and Likes of 90% of the participants and build linear regression models for the five personality traits using
2
Take the Likes of the remaining 10%

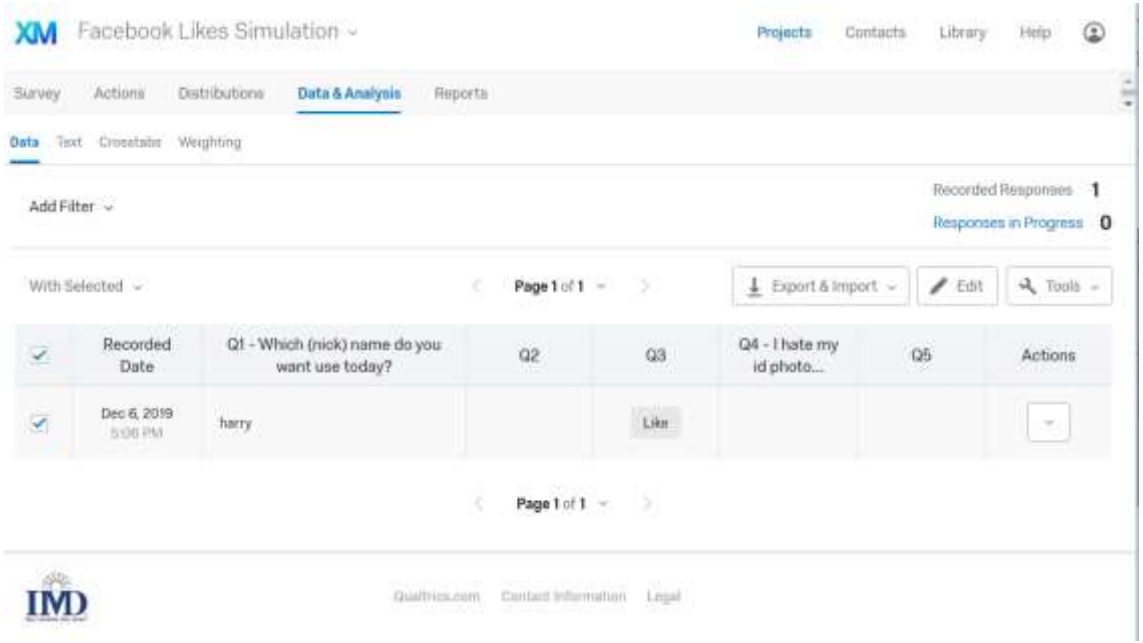
Demo: 1) Make you “pseudo-Facebook Likes”

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

imddemoca_20191209a.R | DataRaw

Filter

	Name	Q1	Q2	Q3	Q4	Q5
	Which (nick) name do you want use today?	Q1	Q2	I hate my id photo...	Q4	Q5
"}	{"ImportId":"QID1_TEXT"}	{"ImportId":"QID2"}	{"ImportId":"QID3"}	{"ImportId":"QID4"}	{"ImportId":"QID5"}	{"ImportId":"QID6"}
	harry		Like			Like
	hulk					Like
	hermione	Like		Like	Like	
	pooh	Like	Like	Like	Like	Like
	Phone	Like			Like	
	a		Like			Like

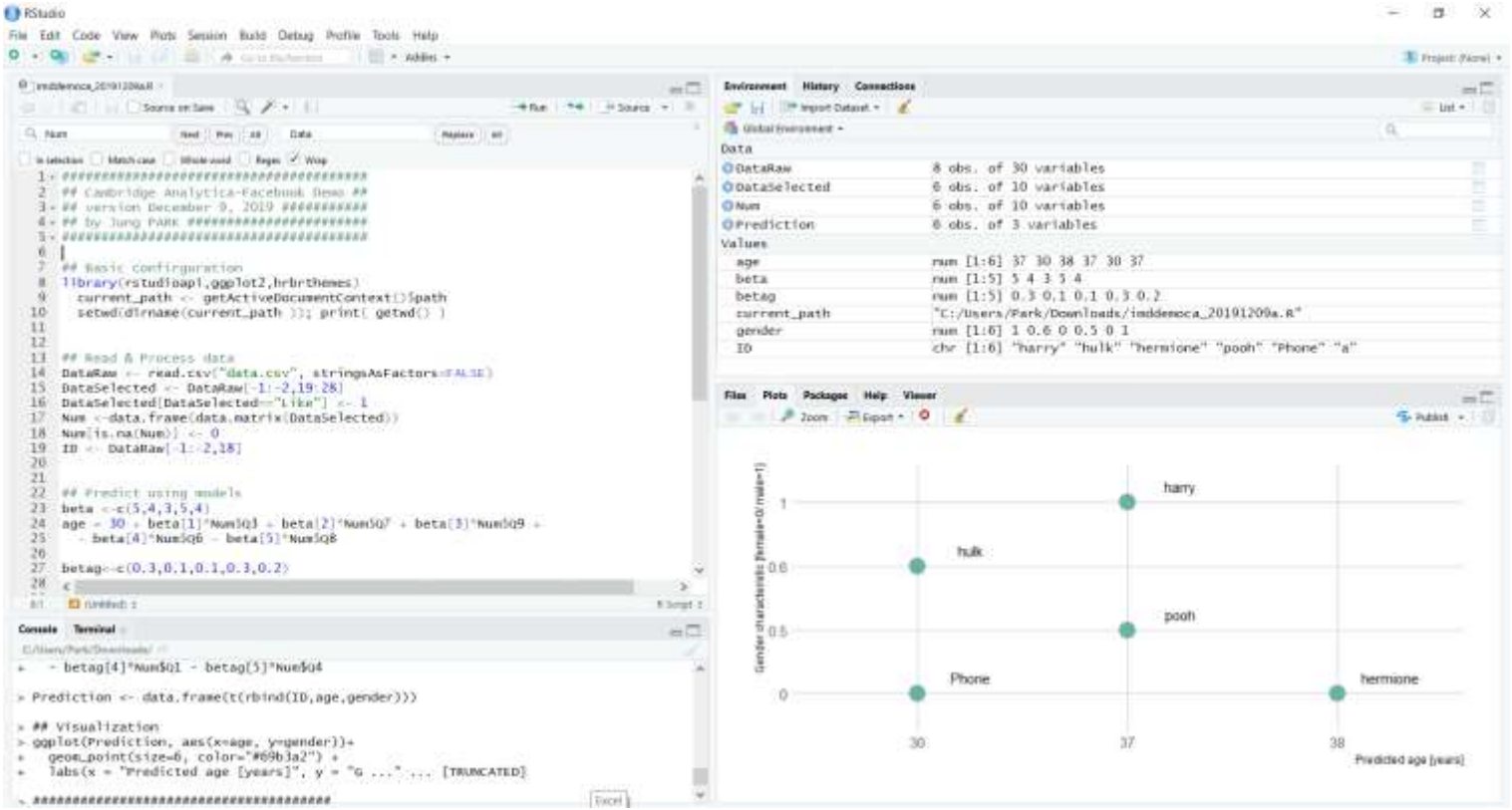
Demo: 4) Process the data, Do regression and Predict Age/ Gender

Q1	Q2	Q3	Q4	Q5	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

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

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



Demo: 4) Regression using Training data

```
## Create Linear Regression Models using the training data
TrainData <- read.csv("traindata.csv", stringsAsFactors=FALSE)

AgeLinearModel <- lm(age ~ Q6+Q8+Q3+Q7+Q9, data=TrainData)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.4859     0.5046   86.19   <2e-16 ***
Q6            7.1564     0.4031   17.75   <2e-16 ***
Q8            7.5213     0.4047   18.58   <2e-16 ***
Q3           -6.7378     0.4054  -16.62   <2e-16 ***
Q7           -6.6812     0.4049  -16.50   <2e-16 ***
Q9           -7.0746     0.4073  -17.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Demo: 4) Regression using Training data

```
## Predict using models
AgePred <- predict(AgeLinearModel, Num)
```

	ID	AgePred	GenderPred	AgeTrue	GenderTrue
3	harry	29.7	1.1		
4	hulk	44.3	0.6		
5	hermione	29.7	-0.1		
6	pooh	30.1	0.6	46	
7	Phone	44.3	-0.1	46	0
8	a	29.7	1.1	45	1
9	apple	43.9	0.6	37	0
10	imd	36.8	0.9	20	1

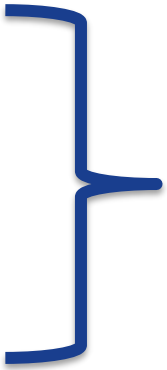
[Data Analysis] Regression models

Training data

270k users'
Facebook likes

&

Personality test results



Predictive model

New data

87mil users'
Facebook likes

Personality test results

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

Political views: selected most predictive Likes

Selected most predictive Likes

Politics	Republican	Democrat
	George W Bush	Joe Biden
	John McCain	Speaker Nancy Pelosi
	Conservative	Health Care Reform
	Rush Limbaugh	The White House
	Sean Hannity	Democrats
	Bill Oreilly	Barbara Boxer
	Positively Republican	Anthony Weiner
	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.

Openness: selected most predictive Likes

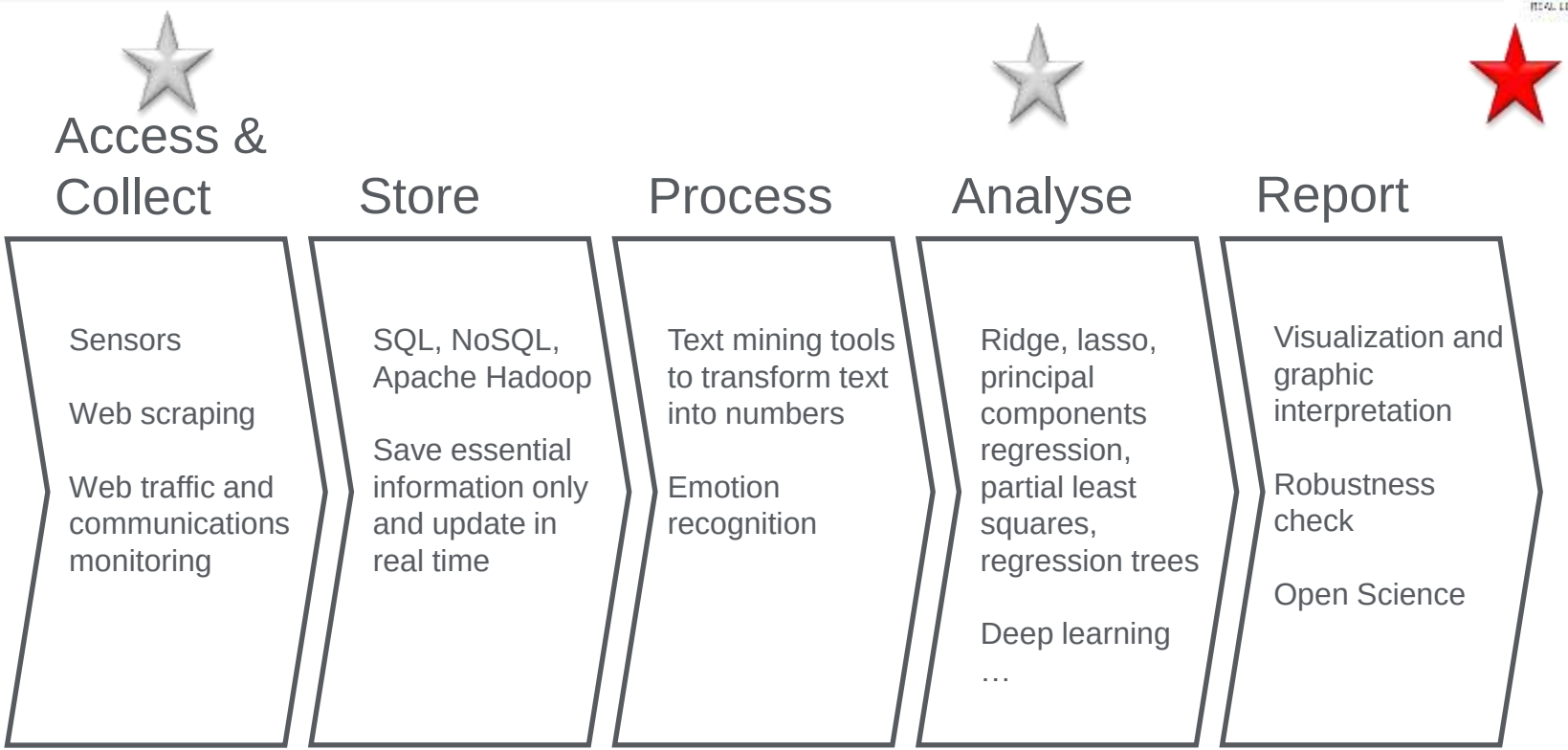
Selected most predictive Likes

Openness	*Liberal & Artistic*			*Conservative*
		Oscar Wilde	NASCAR	
		Charles Bukowski	Austin Collie	
		Sylvia Plath	Monster-In-Law	
		Leonardo Da Vinci	I don't read	
		Bauhaus	Justin Moore	
		Dmt The Spirit Molecule	ESPN2	
		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



[Use of the analytics] Psychographics – Personality Test Results

[Use of the analytics] choosing the most effective message

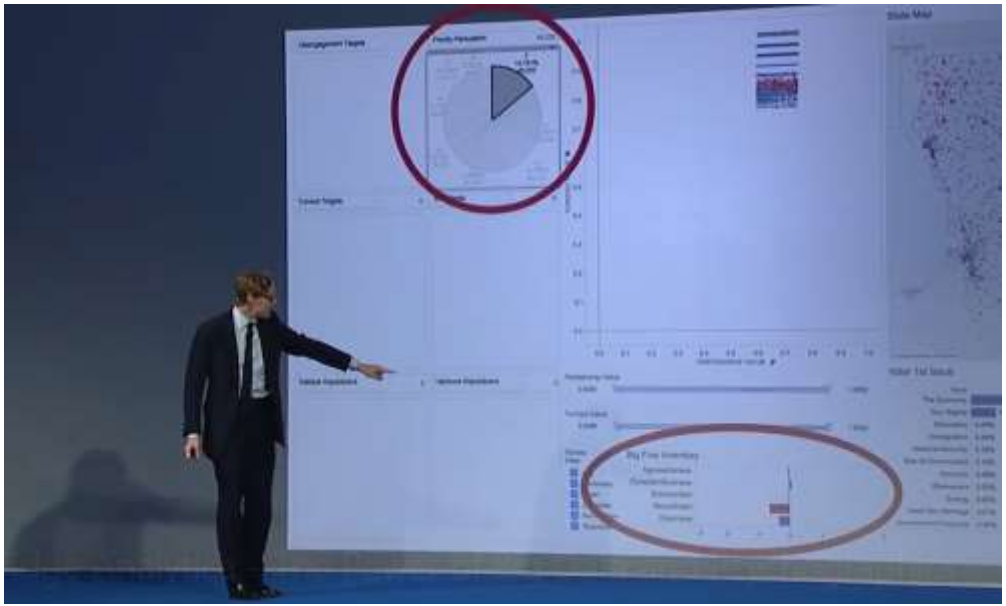
Anxious type
would respond to messages highlighting the threat of a break-in

On gun rights issue

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



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=n8Dd5aVXLCc&t=126s

[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 predictions actually work well? How much contribution did Cambridge Analytica make to the republican's win?

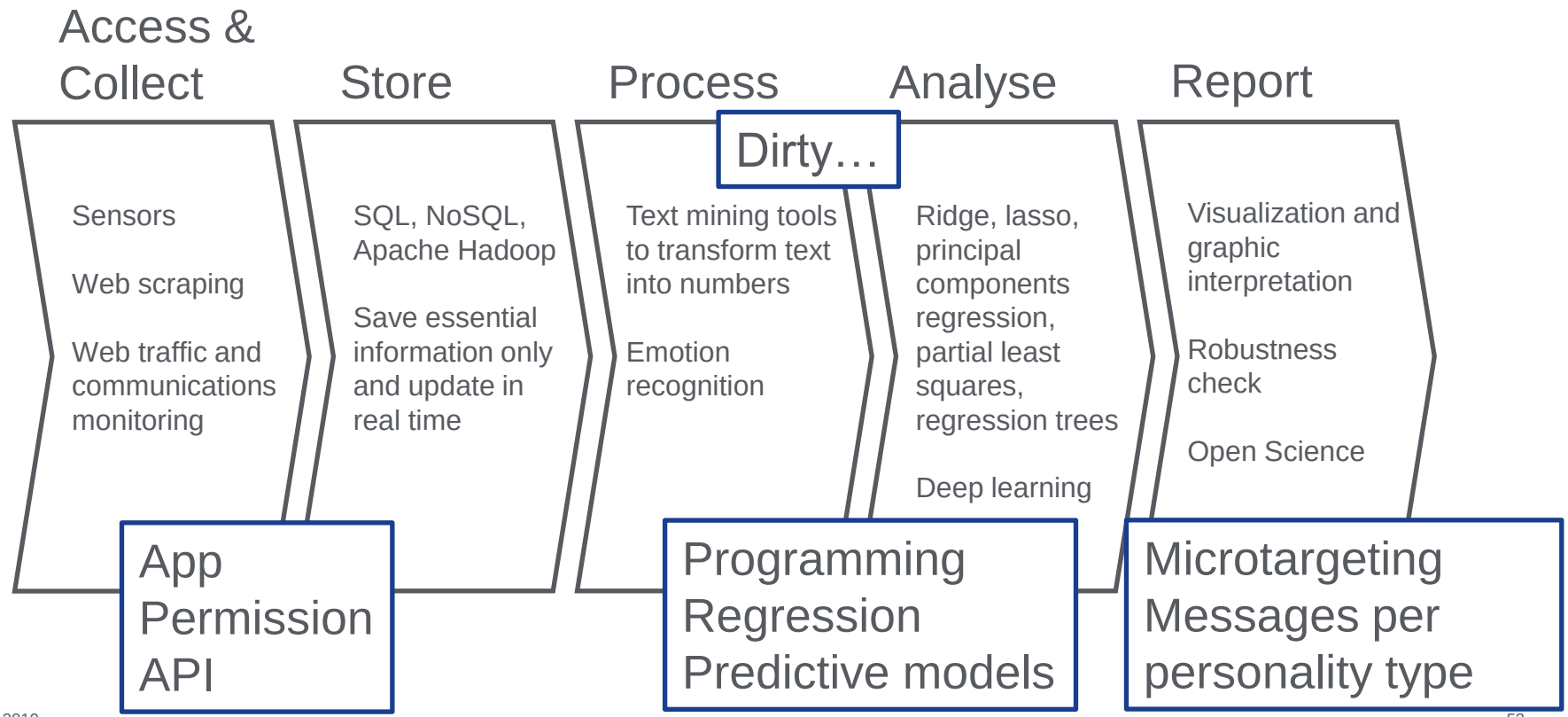
- Why did Alexander Nix (CA's CEO) present CA's involvement in the election at several conferences and sales meetings? Didn't he expect that he could be accused of it?

- Implication to your business?

New Data is
New Opportunity

New Data is
New Opportunity
and New Risk!

Any questions?



Appendix (in case of a technical problem)

Simulating Facebook Likes

Q1

https://www.bizjournals.com/phoenix/news/2017/01/30/wet-seal-closing-remaining-arizona-locations.html

Q2

https://www.ksfy.com/content/news/ESPN-Caesars-Entertainment-ink-sports-betting-deal-509912801.html

Simulating Facebook Likes

Q3

I hate my id photo…

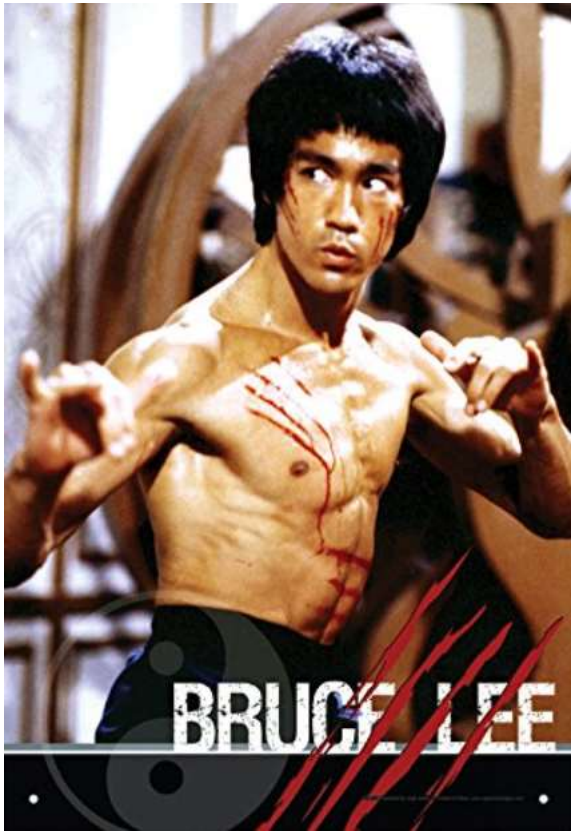
https://twitter.com/mariaforever12/status/365220589717311488

Q4

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

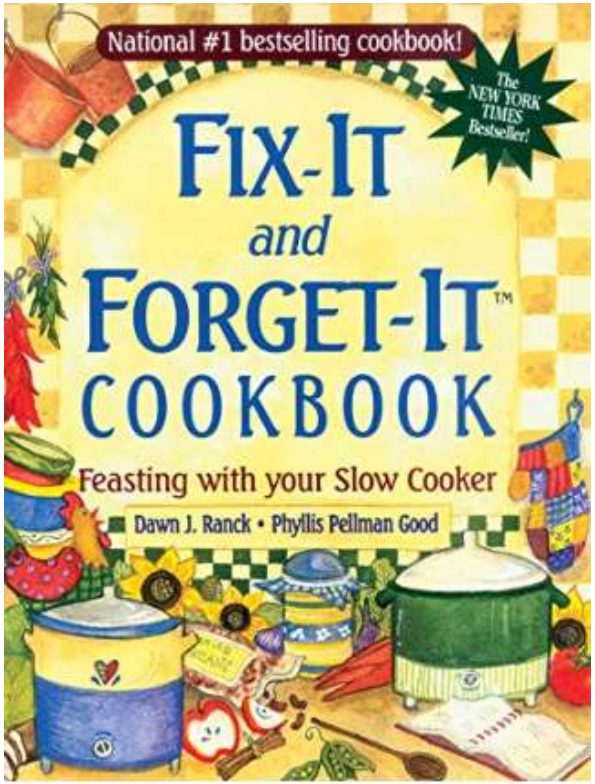
Simulating Facebook Likes

Q5



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

Q6



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

Simulating Facebook Likes

Q7

https://me.me/i/dude-wait-what-3243555

Q8

Small Business Saturday

https://eu.usatoday.com/story/money/2019/11/29/museum-store-sunday-small-business-saturday-2019-holiday-shopping/4331921002/

Simulating Facebook Likes

Q9

Because I am a girl movement

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

Q10

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

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

$$\text{Age} = 30 + 5*(Q6+Q8-Q3-Q7-Q9)$$

$$\text{Gender} = 0.5 + 0.2*(Q2+Q5+Q10-Q1-Q4)$$

Likes = 1, Skip = 0