FIT1043 Introduction to Data Science

Week 2: Data Scientist Roles and Skills, Impact of Data Science & Business Models with Data

Ian K T Tan

School of Information Technology
Monash University Malaysia

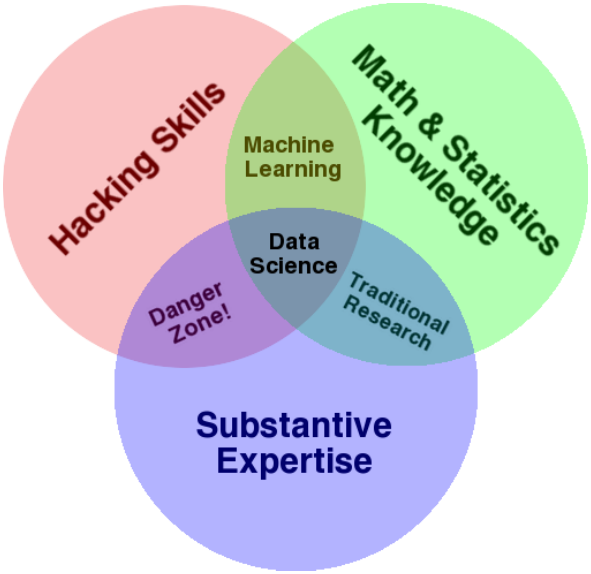
Week 1 Coverage
Python for Data Science
Overview of Data Science

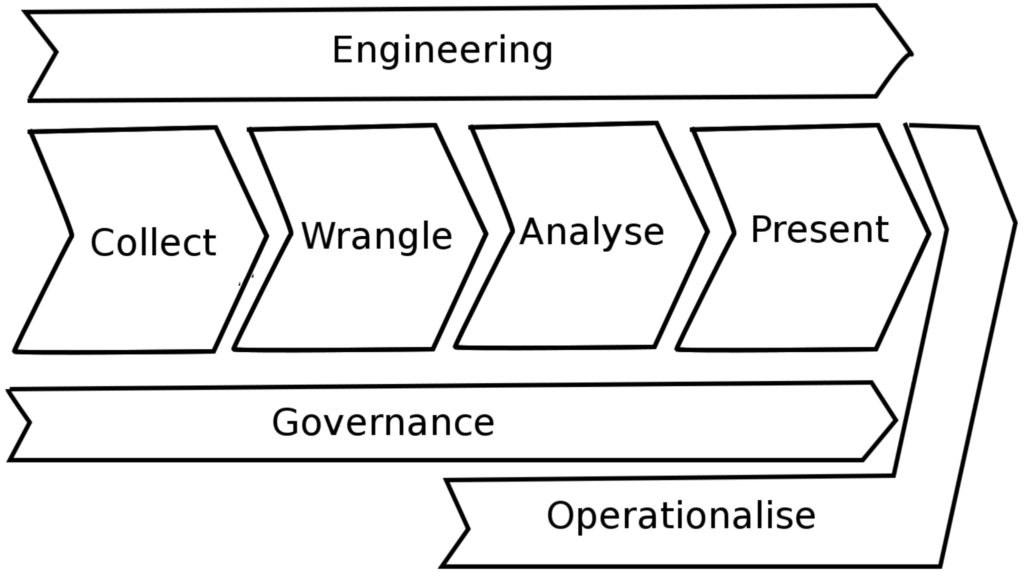
Week 1 Coverage

- Why study data science?
    - We had a look at “Data Scientist” as a job last week. How about searching for:
        - Data Analyst
        - Machine Leaning

- Drew Conway’s Venn Diagram
- Usefulness of Machine Learning
- Data Science Process and Our Standard Value Chain

flux.qa/GMBXRB





We call this the Standard Value Chain.

We will refer to this throughout the semester!

Collection

- Getting the data

Engineering

- Storage and computational resources across full lifecycle

Governance

- Overall management of data across full lifecycle

Wrangling

- Data pre-processing, cleaning

Analysis

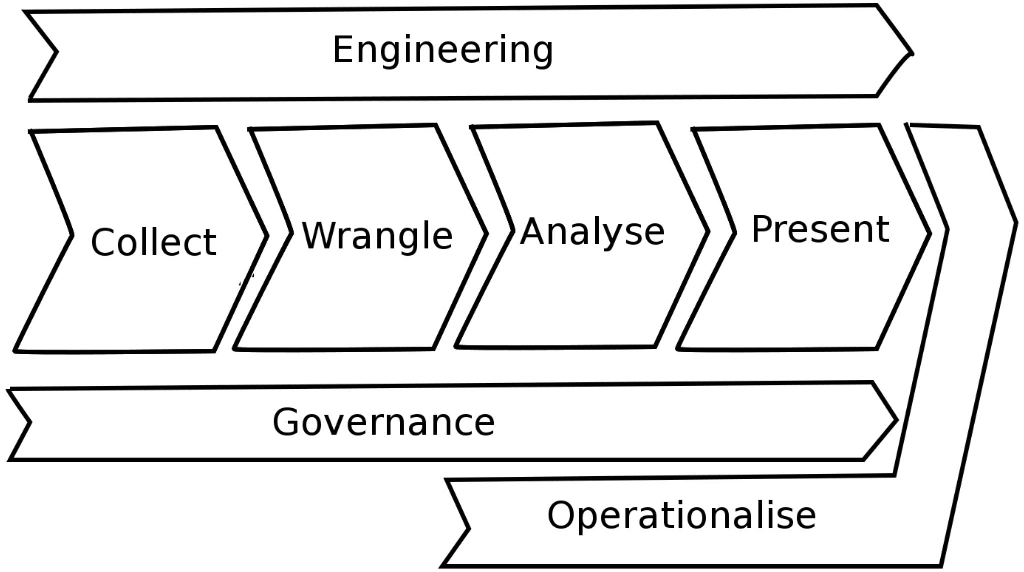
- Discovery (learning, visualisation, etc.)

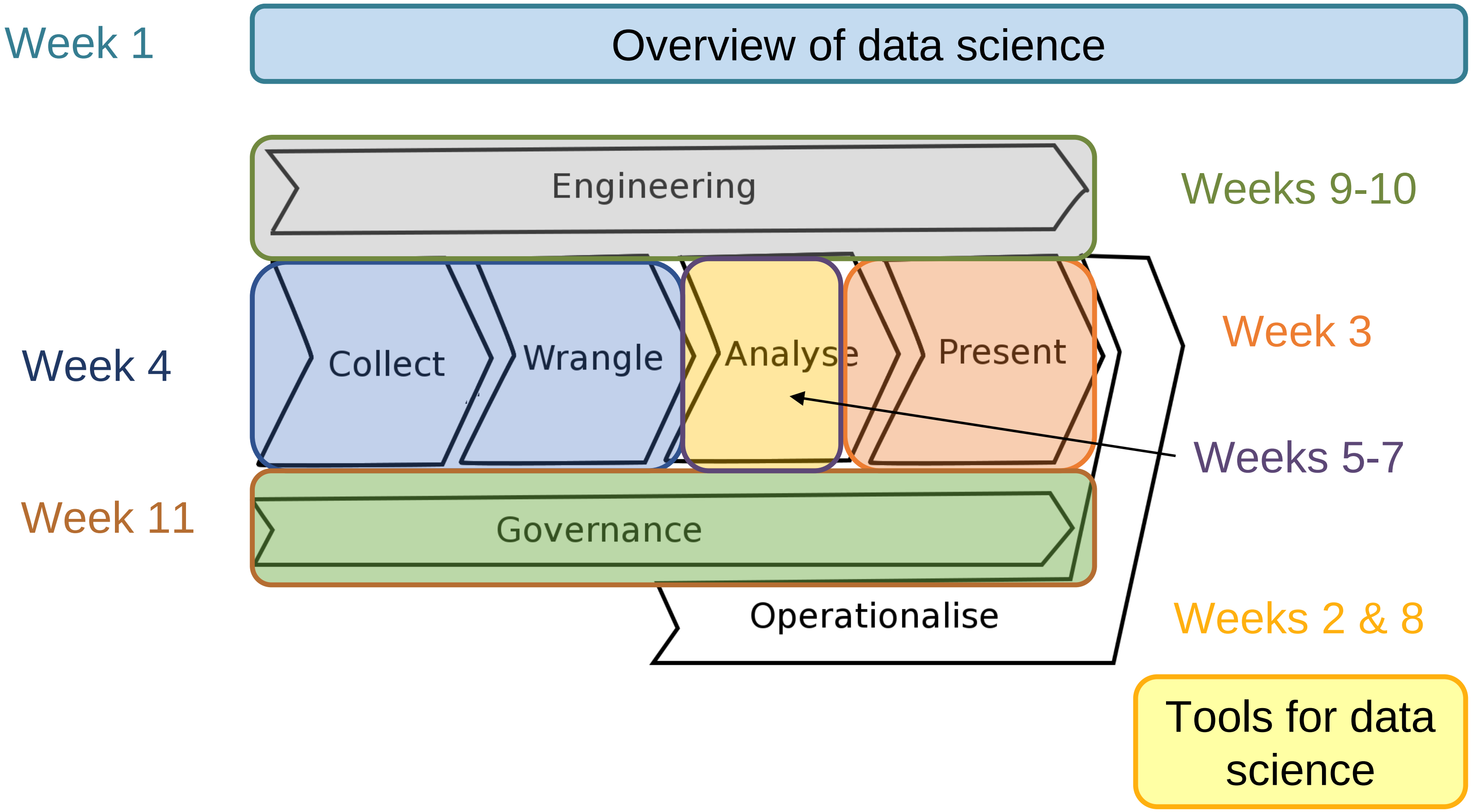
Visualization

- Arguing the case that the results are significant and useful

Operationalize

- Putting the results to work, so as to gain benefits or value




Week 1
Overview of data science
Engineering
Weeks 9-10
Week 4
Collect
Wrangle
Analyse
Present
Week 3
Weeks 5-7
Week 11
Governance
Operationalise
Weeks 2 & 8
Tools for data science

Week	Activities	Assignments
1	Overview of data science	
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	
9	Characterising data and "big" data	Assignment 2
10	Big data processing	
11	Issues in data management	
12	Industry guest lecture (tentative)	Assignment 3

Week 2 Outline

Introduction to Python for Data Science

- Motivation to studying Python
- Python data types (Video Lectures)
- Essential libraries

Overview of data science (con’t)

- Data science roles and skills
- Impact of data science
- Business models with data

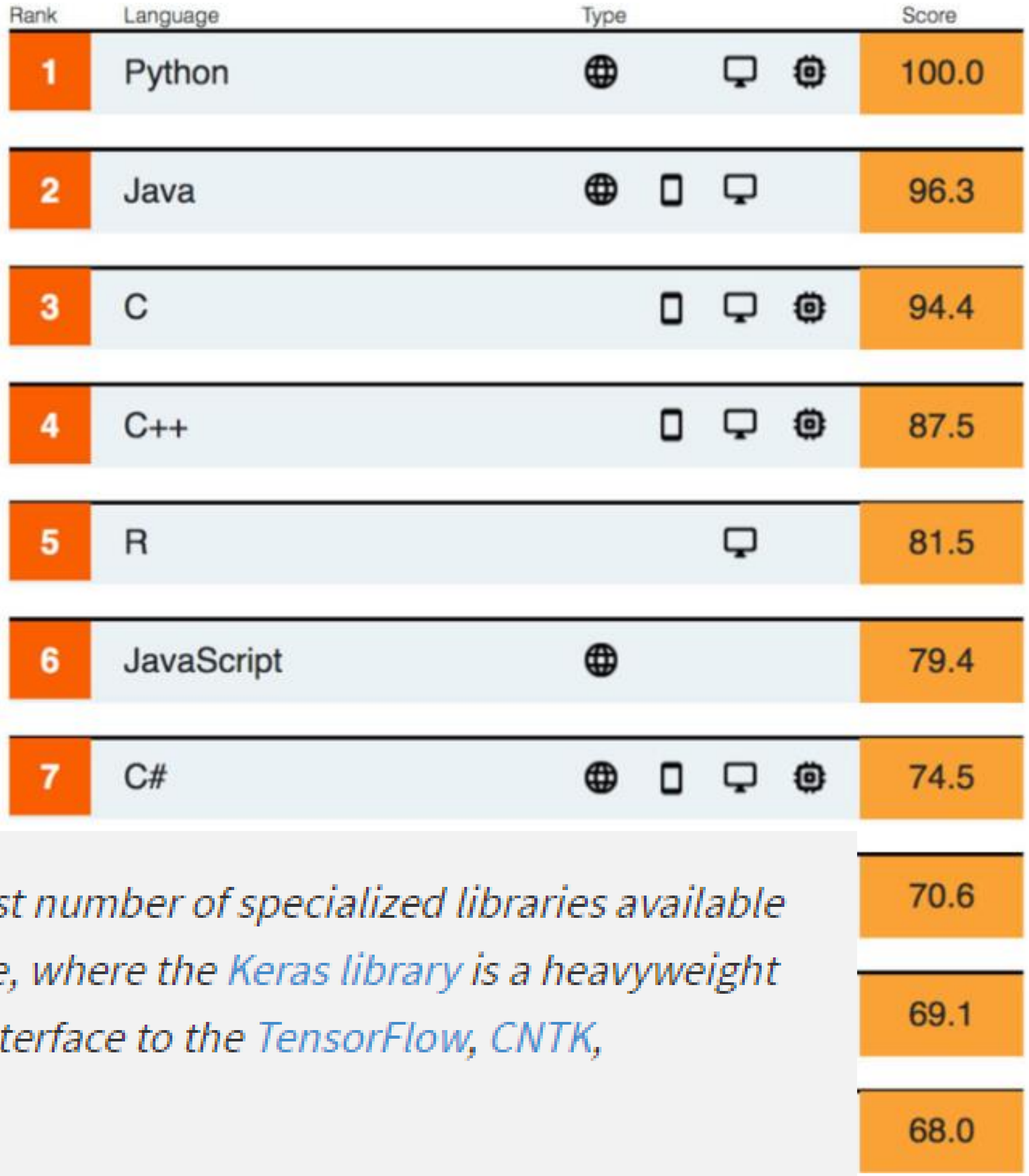
Learning Outcomes

Week 2

By the end of this week you should be able to:

- Comprehend essentials for coding in Python for data science
- Explain and interpret given **Python** codes
- Explain **different data science roles** and skills and comprehend the differences between them
- Explain **Impact** of data science
- Explain the **data business models** for organizations

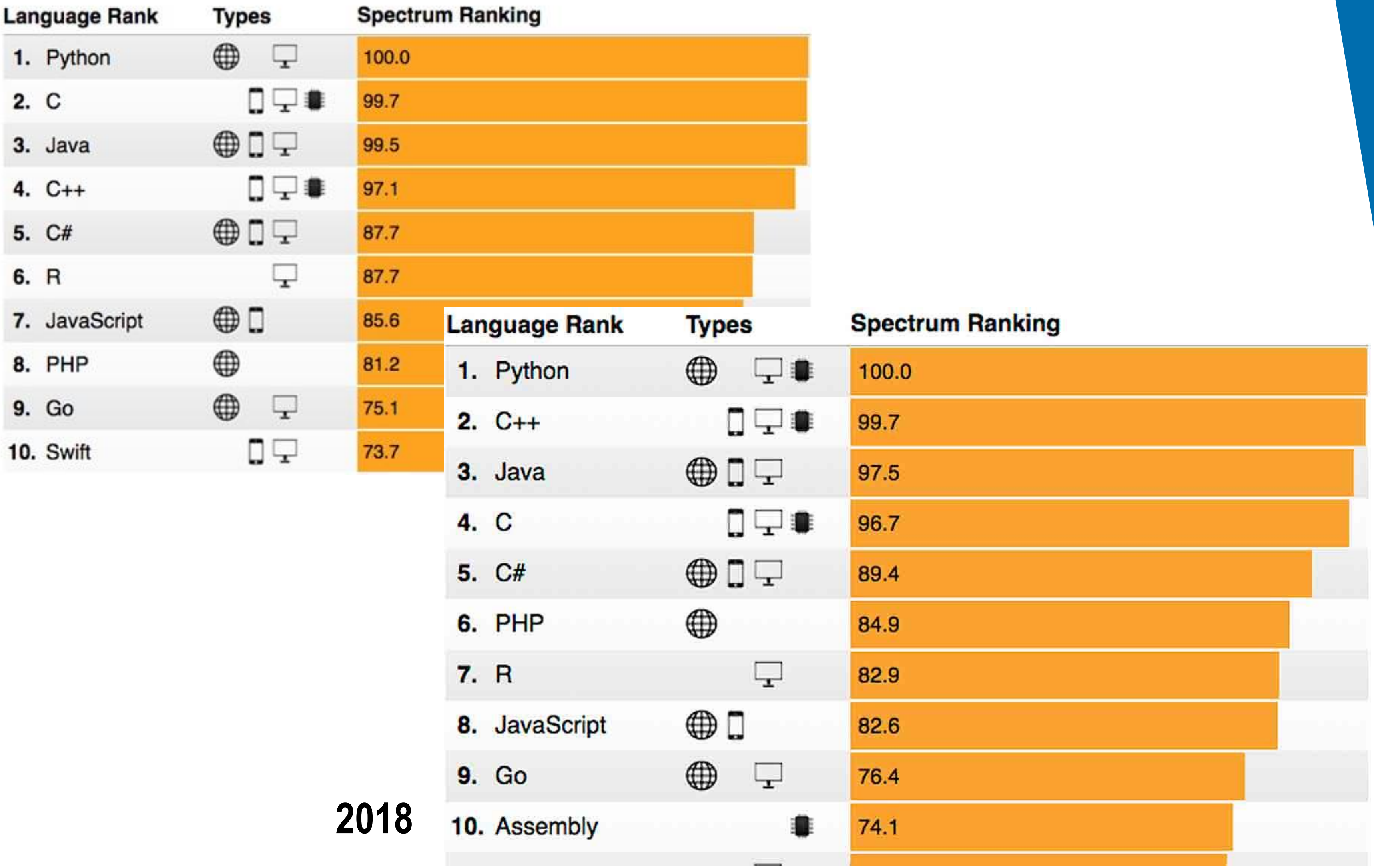
IEEE Top Programming Languages in 2019



Rank	Language	Type	Score
1	Python	🌐 🖥 ⚙	100.0
2	Java	🌐 📱 🖥	96.3
3	C	📱 🖥 ⚙	94.4
4	C++	📱 🖥 ⚙	87.5
5	R	🖥	81.5
6	JavaScript	🌐	79.4
7	C#	🌐 📱 🖥 ⚙	74.5
			70.6
			69.1
			68.0

> *Python's popularity is driven in no small part by the vast number of specialized libraries available for it, particularly in the domain of artificial intelligence, where the Keras library is a heavyweight among deep-learning developers: Keras provides an interface to the TensorFlow, CNTK, and Theano deep-learning frameworks and tool kits.*
>
> *IEEE Spectrum*

3 Years Running



2017

Language Rank	Types	Spectrum Ranking
1. Python	🌐 🖥	100.0
2. C	📱 🖥 ▪	99.7
3. Java	🌐 📱 🖥	99.5
4. C++	📱 🖥 ▪	97.1
5. C#	🌐 📱 🖥	87.7
6. R	🖥	87.7
7. JavaScript	🌐 📱	85.6
8. PHP	🌐	81.2
9. Go	🌐 🖥	75.1
10. Swift	📱 🖥	73.7

2018

Language Rank	Types	Spectrum Ranking
1. Python	🌐 🖥 ▪	100.0
2. C++	📱 🖥 ▪	99.7
3. Java	🌐 📱 🖥	97.5
4. C	📱 🖥 ▪	96.7
5. C#	🌐 📱 🖥	89.4
6. PHP	🌐	84.9
7. R	🖥	82.9
8. JavaScript	🌐 📱	82.6
9. Go	🌐 🖥	76.4
10. Assembly	▪	74.1

Source: IEEE

Python’s Role in Data Science

Many tools out there for data science.

Python has gained popularity over the last few years.

- easy to learn
- flexible and multi-purpose
- great libraries
- well designed computer language
- good visualization for basic analysis

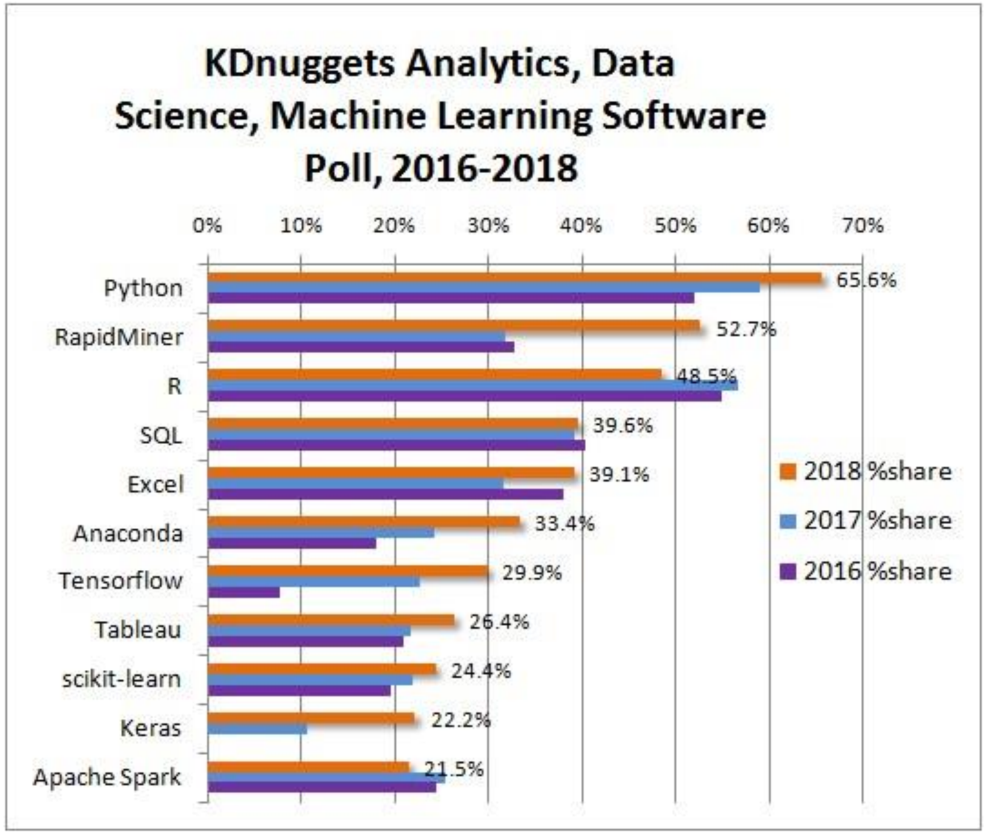
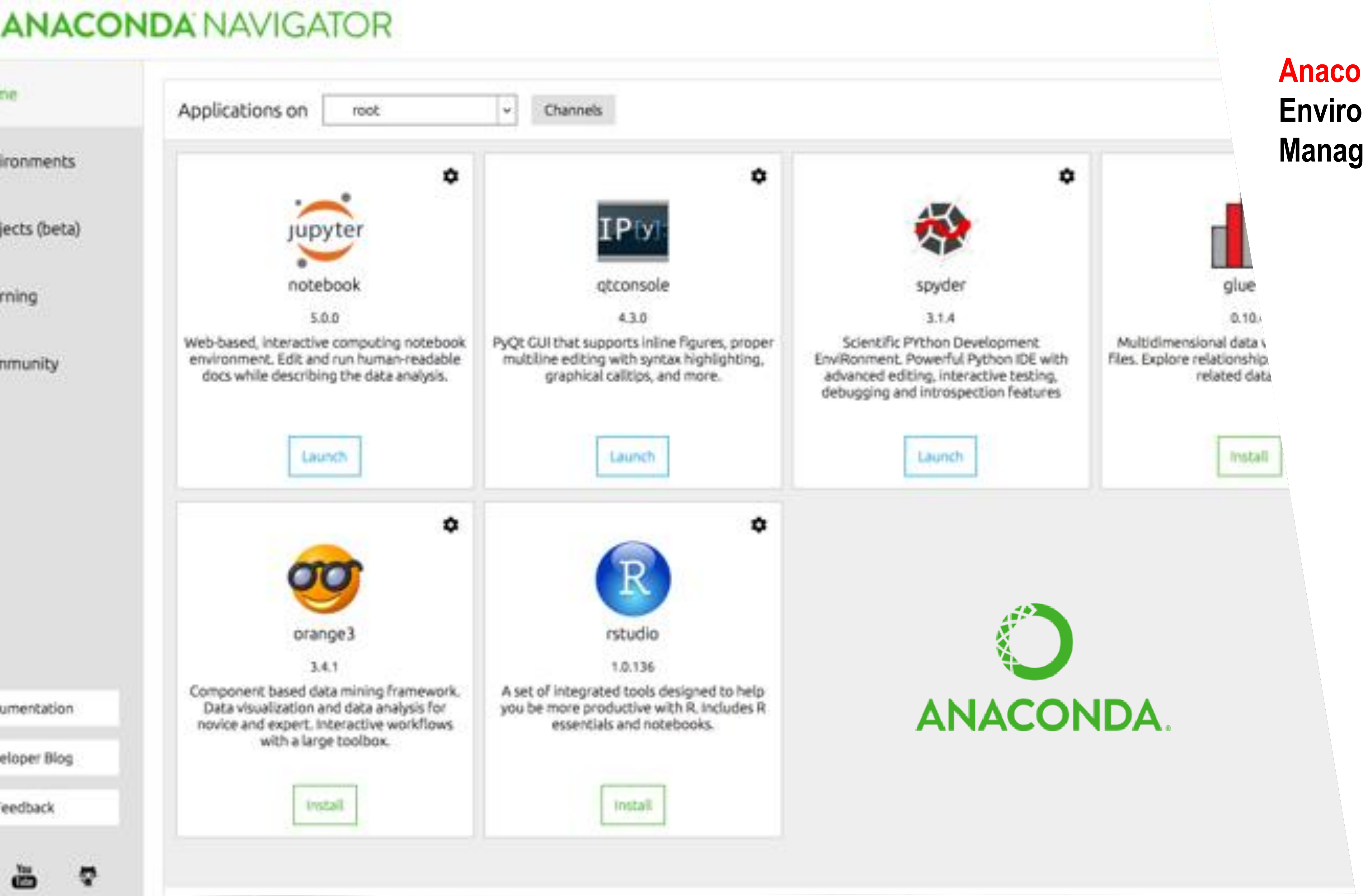


image source: kdnuggets.com

Anaconda: Environment Manager.



notebook 5.0.0: Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

qtconsole 4.3.0: PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

spyder 3.1.4: Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

glue 0.10.: Multidimensional data v Files. Explore relationship related data

orange3 3.4.1: Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

rstudio 1.0.136: A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Data Scientist Roles

ePub Section 1.4

For better understanding the different kinds of data scientists:

- Reviewing:

  Analyzing the Analyzers from Harris, Murphy and Vaisman

- Interviews:

  From *Data Analytics Handbook*

MONASH University

Roles of a Data Scientist

Analyzing the Analyzers http://www.oreilly.com/data/free/files/analyzing-the-analyzers.pdf

A quote from Quora from Jason Widjaja:

- Data analysts are primarily people who develop insights with data,
- Data scientists are primarily people who develop data models and products, that in turn produce insights, and
- Data engineers are primarily people who manage data infrastructure, automate data processing and deploy models at scale.

(Note the use of the word “primarily”!)

see also Job Comparison – Data Scientist vs Data Engineer vs Statistician

Skills of Data Scientists

Analyzing the Analyzers, Harris, Murphy and Vaisman, 2013

NB. typical data scientist doesn't have to know all of these!

Business:

Product development, business

Machine learning/Big data:

Unstructured data, structured data, machine learning, big and distributed data

Mathematics/Operations research:

Optimisation, mathematics, graphical models, algorithms

Programming:

Systems administration, back end programming, front end programming

Statistics:

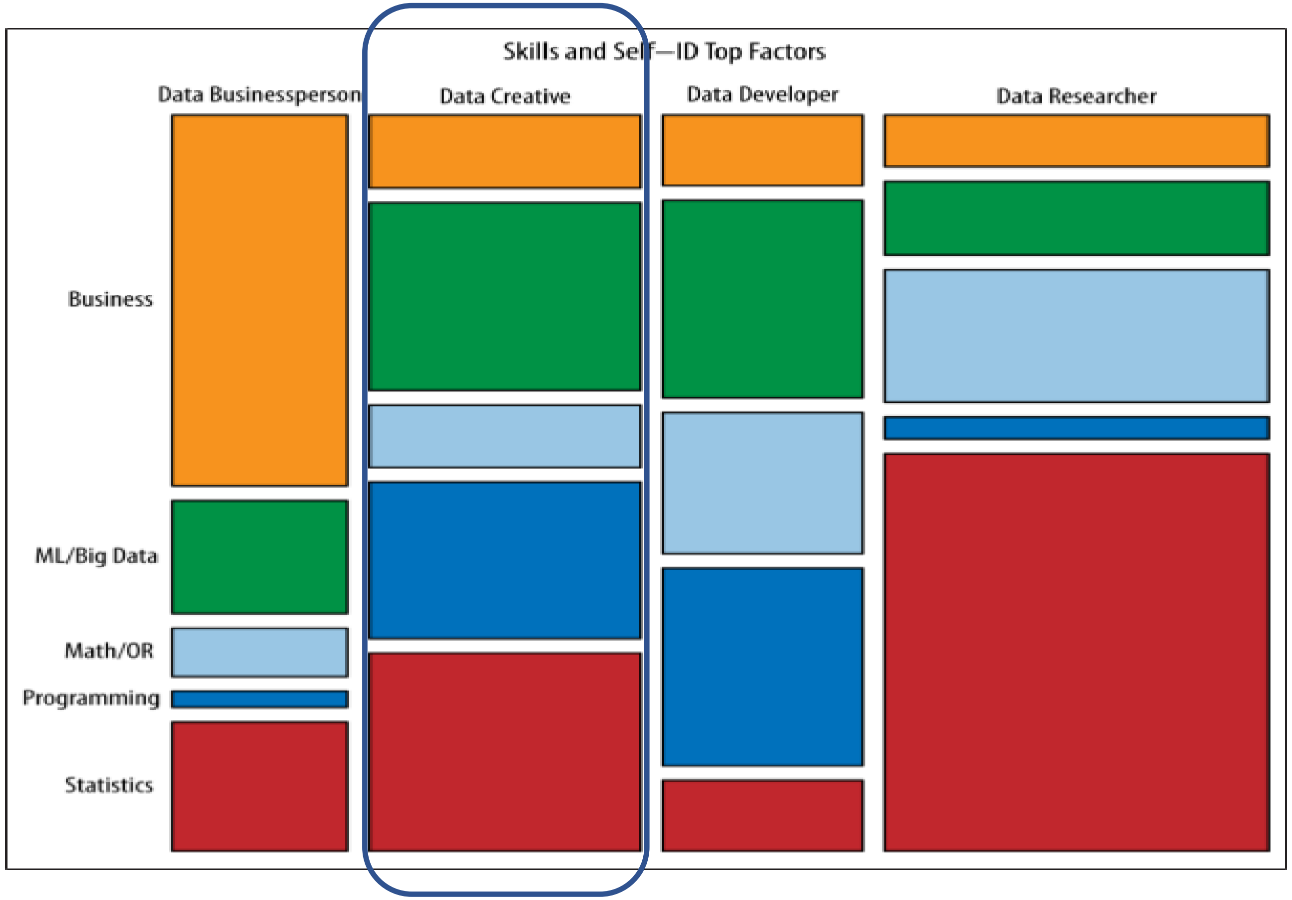
Visualisation, temporal statistics, surveys and marketing, spatial statistics, science, data manipulation

Mapping Styles to Skills

Analyzing the Analyzers, Harris, Murphy and Vaisman, 2013

The Variety of Data Scientist (pages 14 – 16)

- Data Businesspeople
- Data Creatives
- Data Developers
- Data Researchers



Roles of a Data Scientist 2

Interviews from Data Analytics Handbook (https://www.teamleada.com/handbook)

From Data Analytics Handbook

The Data Analytics Handbook is a four volume set of long interviews from industry and academic professionals in the field.

Volume 1 deals with practitioners:

- What exactly do the sexy “data scientists” do?
- What other professions are there in big data?
- What tools do they use to accomplish their tasks?
- How can I enter the industry if I don’t have a Ph.D. in Statistics?

Lessons from the DA Handbook

Summary (important bits)

- **Communication** skills are underrated.
- The biggest challenge for a data analyst is the **Collection** and **Wrangling** steps.
- A data scientist is better at statistics than a software engineer and better at software engineering than a statistician.
- The data industry is still nascent (growing) and the roles less well defined so you get to interact with many parts of the company from engineering to business intelligence to product managers.
- Keep a **curiosity** about working with data, a quality as important as your technical abilities.

Career as a Data Scientist

Your CV

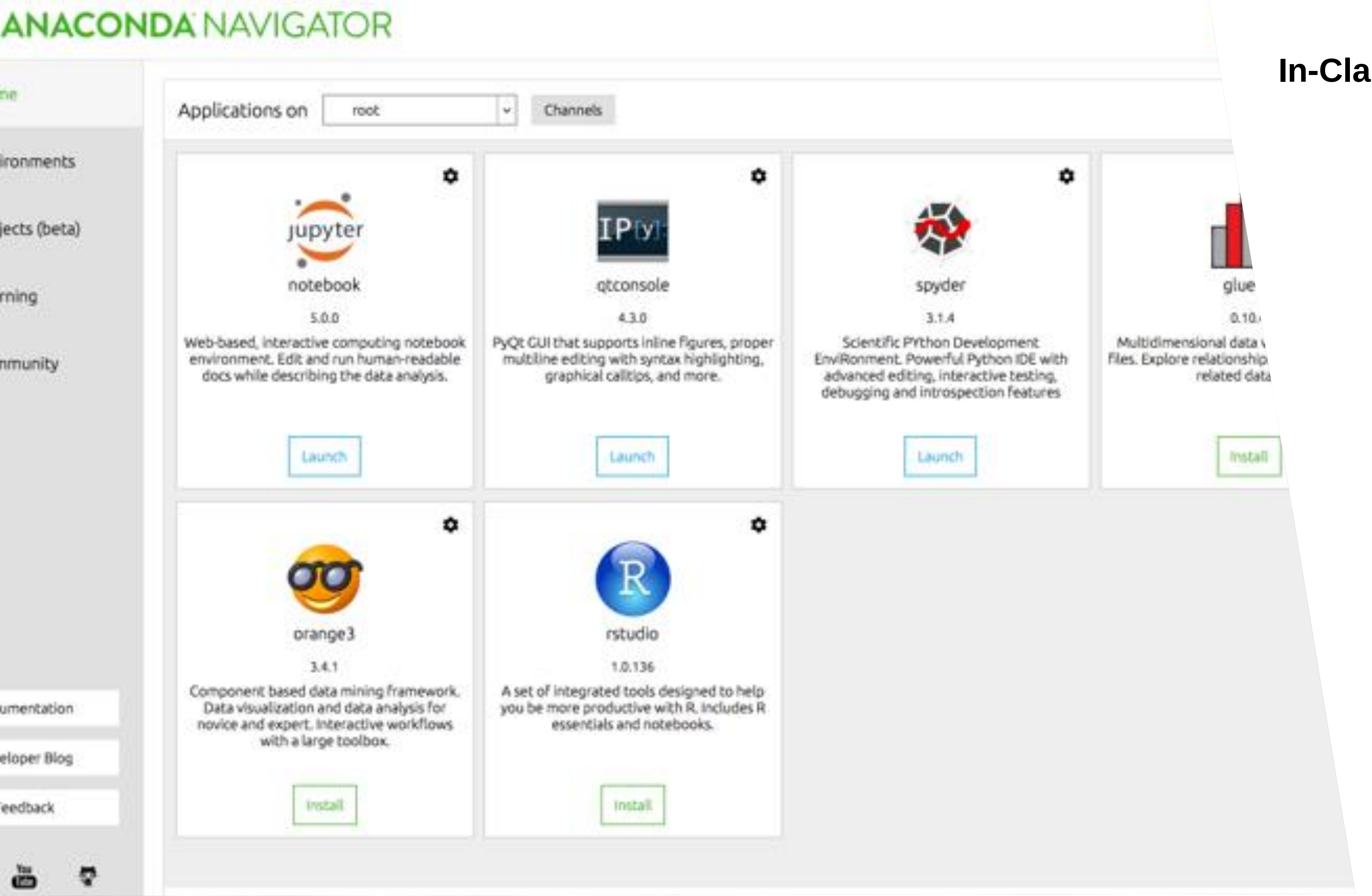
To become a specialist you need:

- Solid **machine learning** and **statistics**
- Related mathematics (1st+2nd year in many degrees)
- Solid prototyping (proof of concepts)
  - R, **Python**, Java
  - **Github**
  - Competitions, e.g. **Kaggle**
- **Unix experience** (Linux, Mac OSX). This unit provides an introduction and background only

Python for Data Science

Basic Python in Pre-Class Activity

In-Class Coding



ANACONDA NAVIGATOR

Applications on root

jupyter notebook
5.0.0
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

IP[y]: qtconsole
4.3.0
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

spyder
3.1.4
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

glue
0.10.
Multidimensional data v
Files. Explore relationship
related data

orange3
3.4.1
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

rstudio
1.0.136
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Impact of Data Science

ePub Section 1.6

MONASH University

Impact of Data Science

Some examples of how data science is impacting others:

- Your life in the cloud
  - Datafication of you

- Social good
  - Numerous examples and very rewarding

- Futurology
  - Healthcare and automobiles

Your Life on the Cloud

From Year Zero: Our life timelines begin

Our personal information is increasingly stored in the cloud:

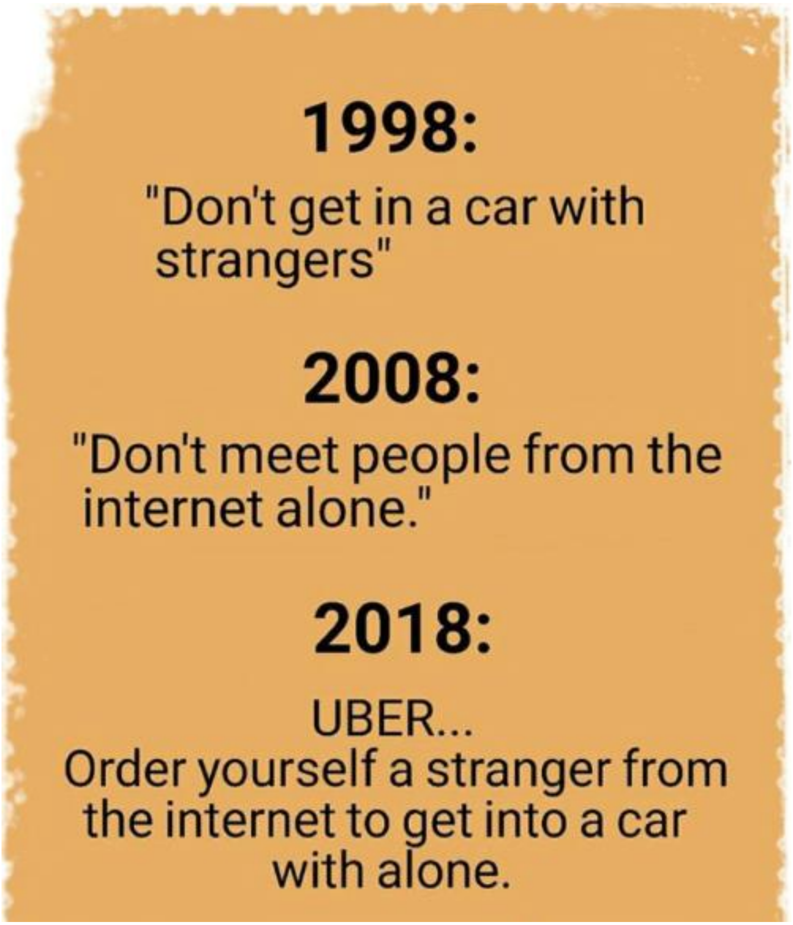
- Social life (Facebook, etc.),
- Career (LinkedIn),
- Search history (Google, etc.),
- Health and medical (Fitbit, etc.),
- Music (Apple, Spotify, etc.).

This provides many, many advantages:

- Personal agents, computerised support for health.

But also some disadvantages:

- Security and privacy breaches.



1998:

"Don't get in a car with strangers"

2008:

"Don't meet people from the internet alone."

2018:

UBER...
Order yourself a stranger from the internet to get into a car with alone.

Image source: me.me

Your Life on the Cloud

But

- Corporate leakage to government (security, tax, etc.)
- What if you don't have rights to access/delete our own data?
- Security and privacy breaches
- What if we've changed our ways?
- The department of pre-crime
- Corporate mergers

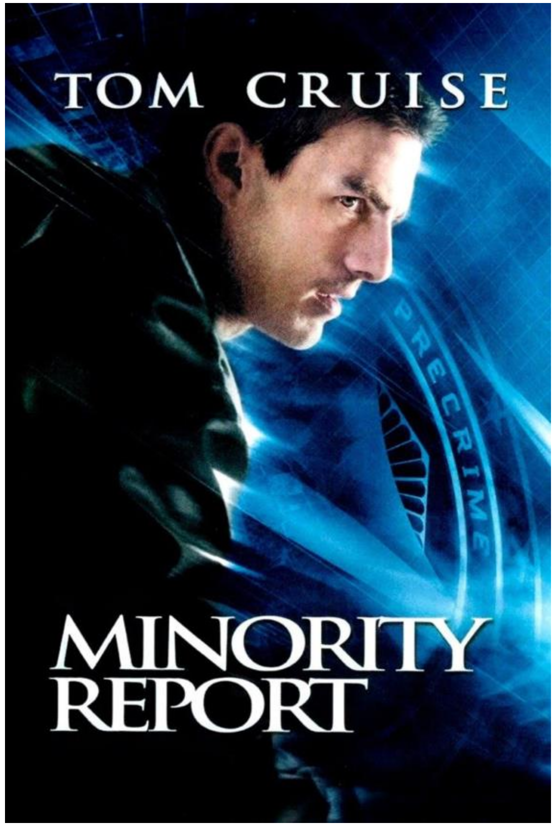


Image source: imdb

Your Life on the Cloud

Social Scoring (https://www.youtube.com/watch?v=xuqbx8tyW1Y)

Social Good

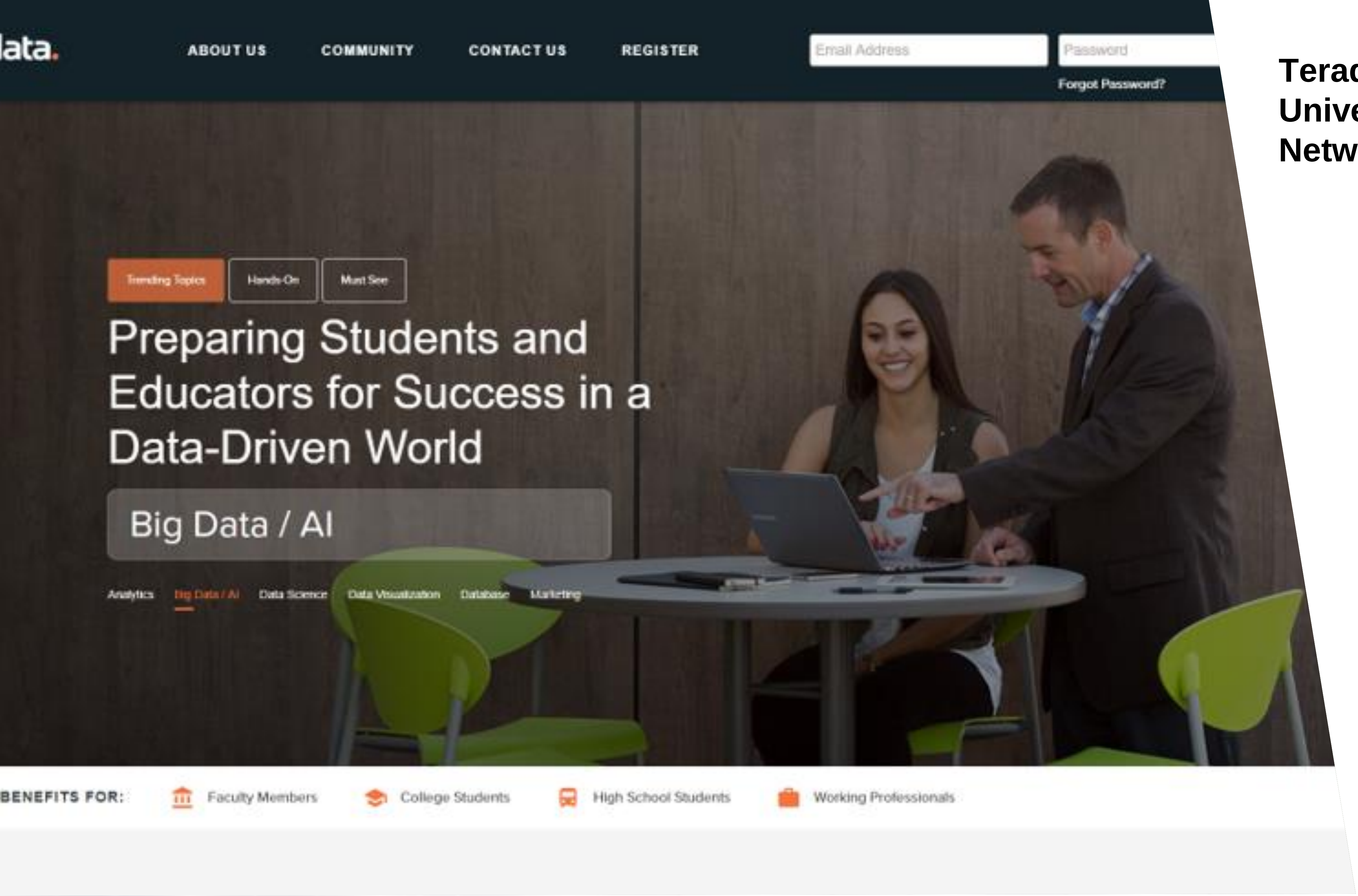
Data Science for Social Good movement training data scientists to support community and charity.

Fight accidents

Fight disease

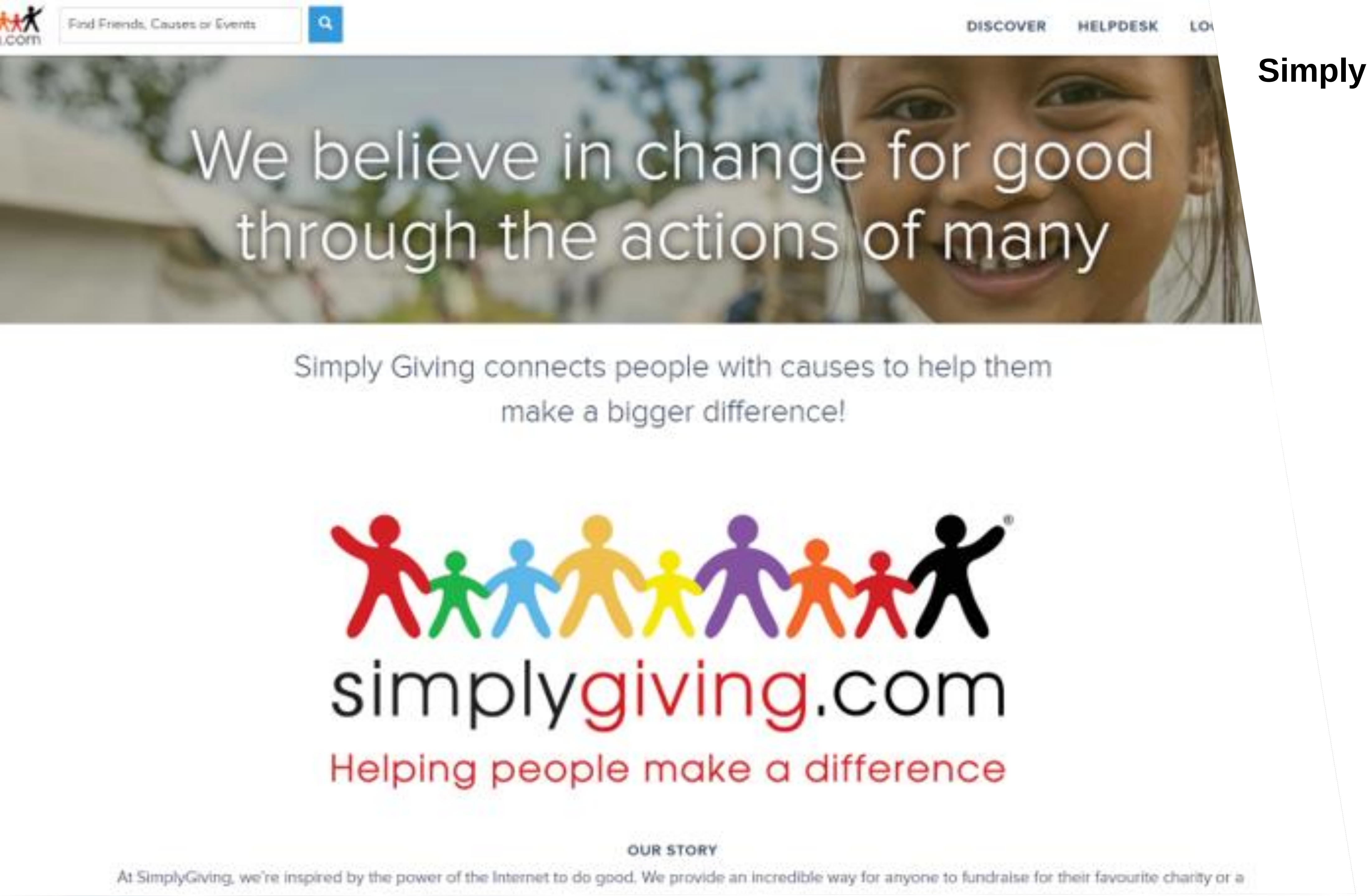
image src: emotiv.com, thehealthsite.com

Teradata University Network



Preparing Students and Educators for Success in a Data-Driven World

SimplyGiving.com



We believe in change for good through the actions of many

Simply Giving connects people with causes to help them make a bigger difference!

simplygiving.com

Helping people make a difference

OUR STORY

At SimplyGiving, we're inspired by the power of the Internet to do good. We provide an incredible way for anyone to fundraise for their favourite charity or a

Health Care Futurology

Some areas where significant impact is to be made in the future

- Your stomach can be instrumented to assess contents, nutrients, etc.
- Your bloodstream can be instrumented too assess insulin levels, etc.
- Your “health” dashboard can be online and shared by your GP
- Health management organisations (HMO) tying funding levels to patient care performance
- GP/HMO will know about your ice cream/beer binge last night and you missing your morning run

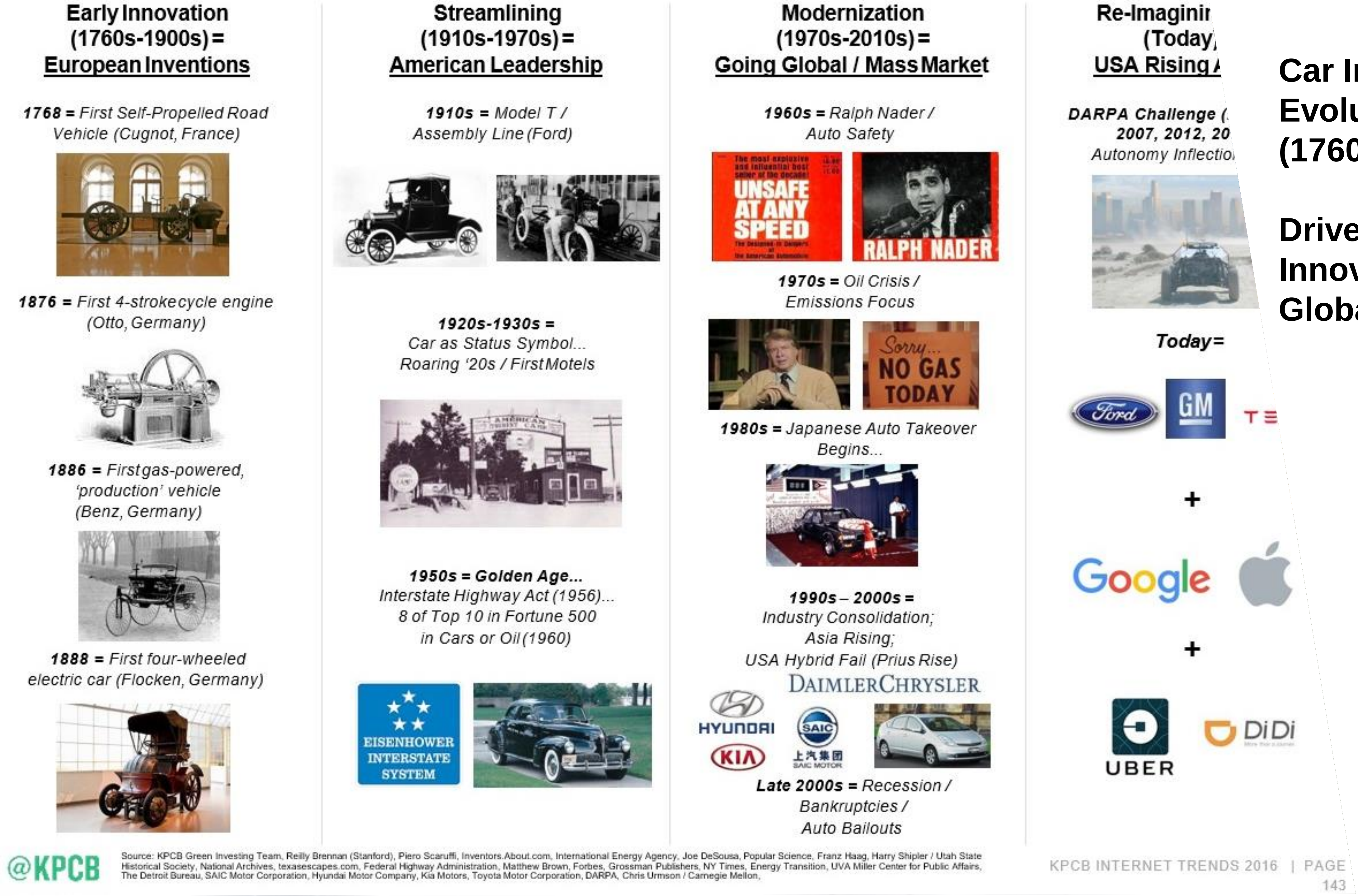
Longitudinal studies feasible

- Longitudinal studies is a method in which data is gathered for the same subjects repeatedly over a period of time

"Big Data - 2020 Vision"

John Schitka in Strata + Hadoop 2014

https://www.youtube.com/watch?v=kgue3iHbkcI



Early Innovation (1760s-1900s) = European Inventions

***1768** = First Self-Propelled Road Vehicle (Cugnot, France)*

***1876** = First 4-stroke cycle engine (Otto, Germany)*

***1886** = First gas-powered, 'production' vehicle (Benz, Germany)*

***1888** = First four-wheeled electric car (Flocken, Germany)*

Streamlining (1910s-1970s) = American Leadership

***1910s** = Model T / Assembly Line (Ford)*

***1920s-1930s** = Car as Status Symbol... Roaring '20s / First Motels*

***1950s = Golden Age...** Interstate Highway Act (1956)... 8 of Top 10 in Fortune 500 in Cars or Oil (1960)*

Modernization (1970s-2010s) = Going Global / Mass Market

***1960s** = Ralph Nader / Auto Safety*

***1970s** = Oil Crisis / Emissions Focus*

***1980s** = Japanese Auto Takeover Begins...*

***1990s – 2000s** = Industry Consolidation; Asia Rising; USA Hybrid Fail (Prius Rise)*

DAIMLERCHRYSLER

HYUNDAI

KIA

SAIC

上汽集团 SAIC MOTOR

***Late 2000s** = Recession / Bankruptcies / Auto Bailouts*

Re-Imaginir (Today) USA Rising A

***DARPA Challenge** (. **2007, 2012, 20** Autonomy Inflectio*

***Today**=*

Ford GM T≡

+

Google

+

UBER DiDi

@KPCB

Source: KPCB Green Investing Team, Reilly Brennan (Stanford), Piero Scaruffi, Inventors.About.com, International Energy Agency, Joe DeSousa, Popular Science, Franz Haag, Harry Shipler / Utah State Historical Society, National Archives, texasescapes.com, Federal Highway Administration, Matthew Brown, Forbes, Grossman Publishers, NY Times, Energy Transition, UVA Miller Center for Public Affairs, The Detroit Bureau, SAIC Motor Corporation, Hyundai Motor Company, Kia Motors, Toyota Motor Corporation, DARPA, Chris Urmson / Carnegie Mellon.

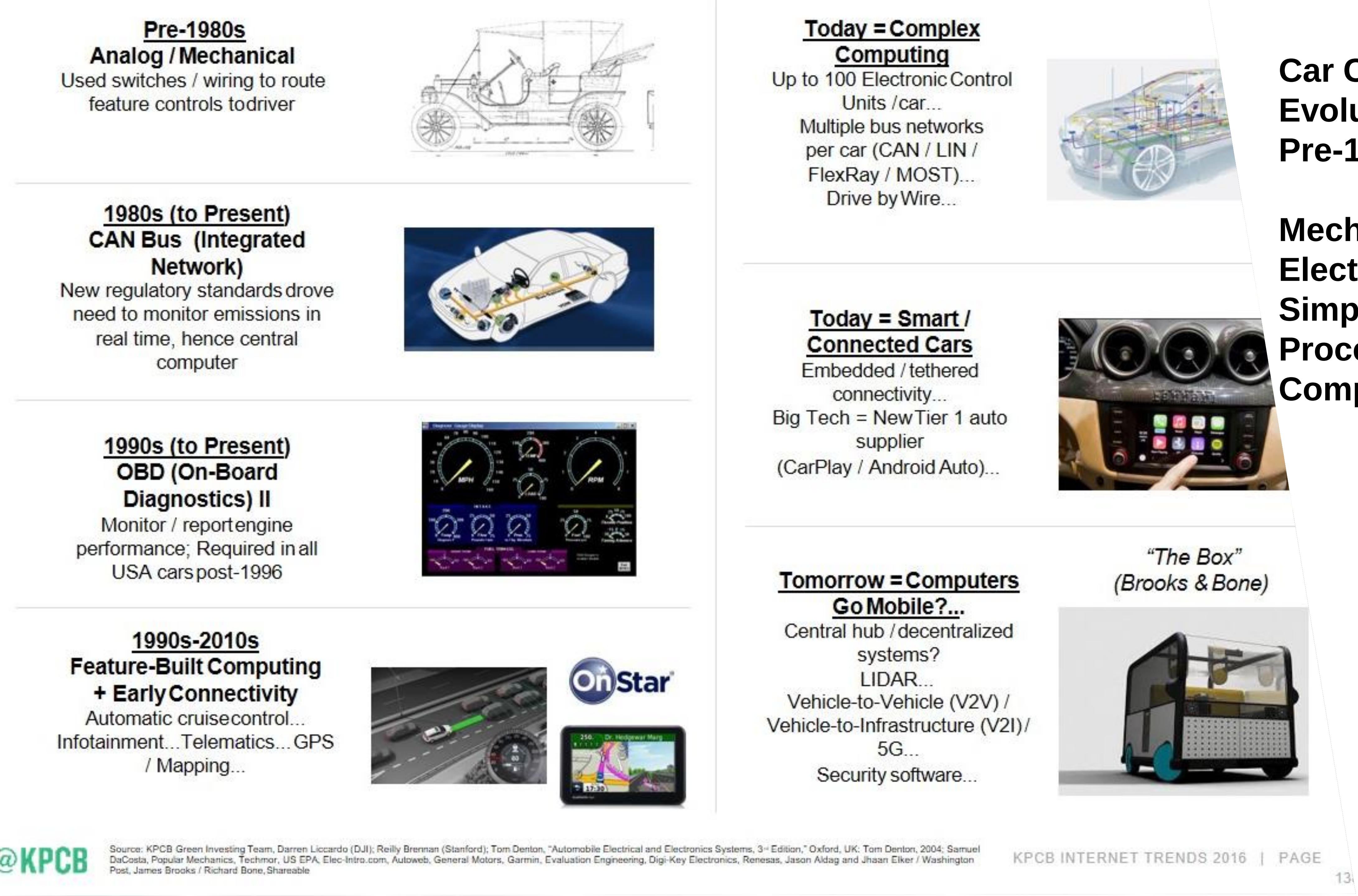
Car Industry Evolution (1760s – Today):

Driven by Innovation & Globalization

MONASH University

Car Computing Evolution Since Pre-1980s:

Mechanical / Electrical → Simple Processors → Computers



Pre-1980s
Analog / Mechanical
Used switches / wiring to route feature controls todriver

1980s (to Present)
CAN Bus (Integrated Network)
New regulatory standards drove need to monitor emissions in real time, hence central computer

1990s (to Present)
OBD (On-Board Diagnostics) II
Monitor / reportengine performance; Required in all USA cars post-1996

1990s-2010s
Feature-Built Computing + Early Connectivity
Automatic cruisecontrol... Infotainment...Telematics...GPS / Mapping...

Today = Complex Computing
Up to 100 Electronic Control Units /car...
Multiple bus networks per car (CAN / LIN / FlexRay / MOST)...
Drive by Wire...

Today = Smart / Connected Cars
Embedded / tethered connectivity...
Big Tech = New Tier 1 auto supplier (CarPlay / Android Auto)...

Tomorrow = Computers Go Mobile?...
Central hub / decentralized systems?
LIDAR...
Vehicle-to-Vehicle (V2V) / Vehicle-to-Infrastructure (V2I) /
5G...
Security software...

"The Box" (Brooks & Bone)

@KPCB

Source: KPCB Green Investing Team, Darren Liccardo (DJI); Reilly Brennan (Stanford); Tom Denton, "Automobile Electrical and Electronics Systems, 3rd Edition," Oxford, UK: Tom Denton, 2004; Samuel DaCosta, Popular Mechanics, Techmor, US EPA, Elec-Intro.com, Autoweb, General Motors, Garmin, Evaluation Engineering, Digi-Key Electronics, Renesas, Jason Aldag and Jhaan Elker / Washington Post, James Brooks / Richard Bone, Shareable

Automobile Futurology

“Big data – 2020 vision” talk by SAP manager John Schitka

Self driving cars:

- How does the city replace traffic fine revenue?
- Can you drink and drive if the car is automatic?
- What happens to the taxi industry?
- What happens to the auto insurance industry?
- What happens to people still “self” driving, and their insurance?
- For the Ultimate Driving Machine, how will self-driving cars impact it?

Business Models with Data

ePub Section 2.3

What kinds of businesses do we have operating in the Data Science world?

Business Models

From Wikipedia:

- A business model describes the rationale of how an organization creates, delivers, and captures value, in economic, social, cultural or other contexts.

Examples of general classes:

- Retailer versus wholesaler
- Luxury consumer products
- Software vendor
- Service provider

What kinds of businesses do we have operating in the Data Science world?

Business Models for Data Science

Many Data Science companies fit into traditional IT business models.

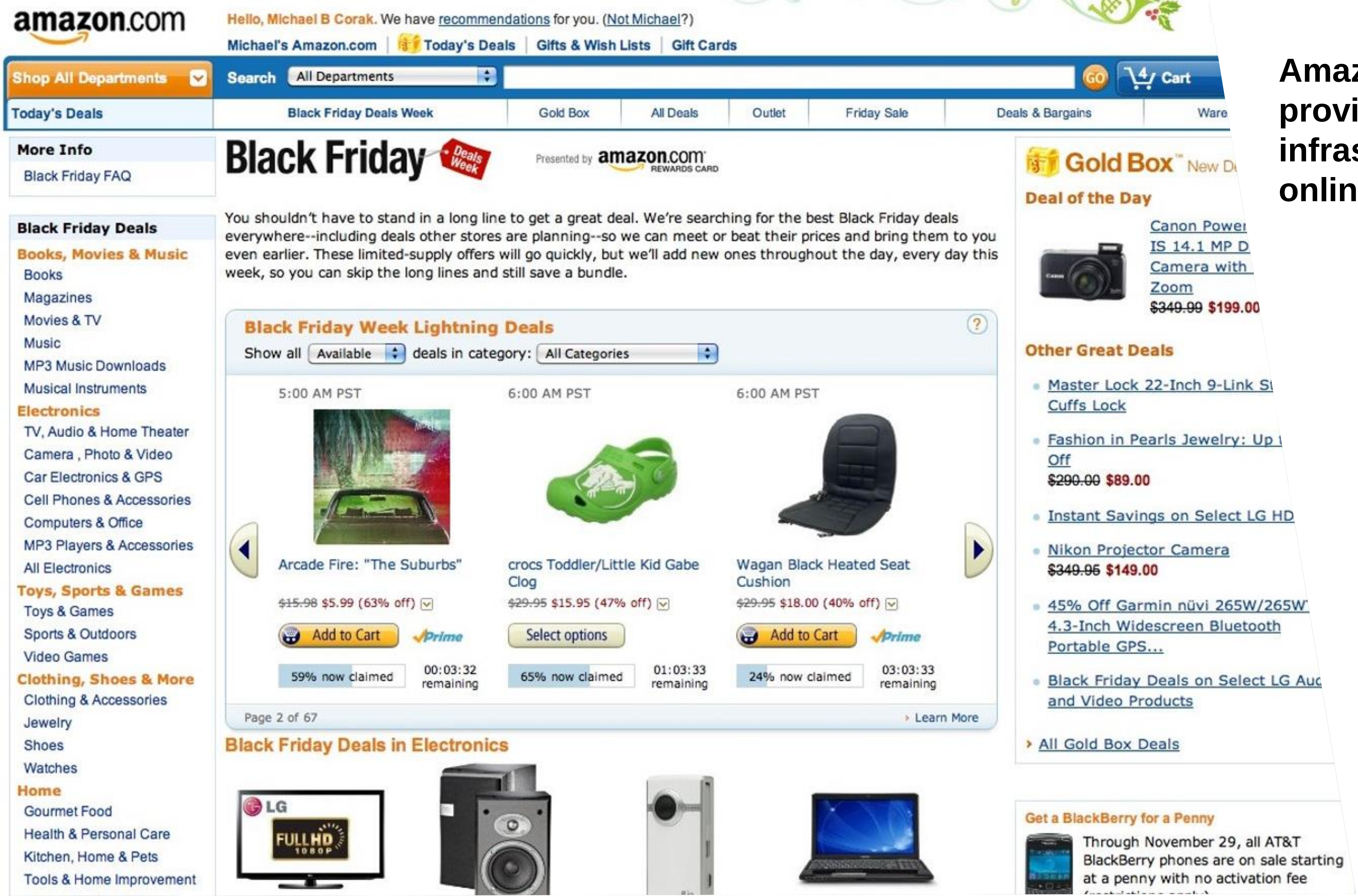
- Software as a service (SaaS)
- Consulting
- Customer relationship management

What are some business models specific to data science?

For example:

- SAS is both a software vendor and a consultancy, both traditional IT business models
- But there are business models somewhat unique to data-based businesses like data science.

sas THE POWER TO KNOW.



You shouldn't have to stand in a long line to get a great deal. We're searching for the best Black Friday deals everywhere--including deals other stores are planning--so we can meet or beat their prices and bring them to you even earlier. These limited-supply offers will go quickly, but we'll add new ones throughout the day, every day this week, so you can skip the long lines and still save a bundle.

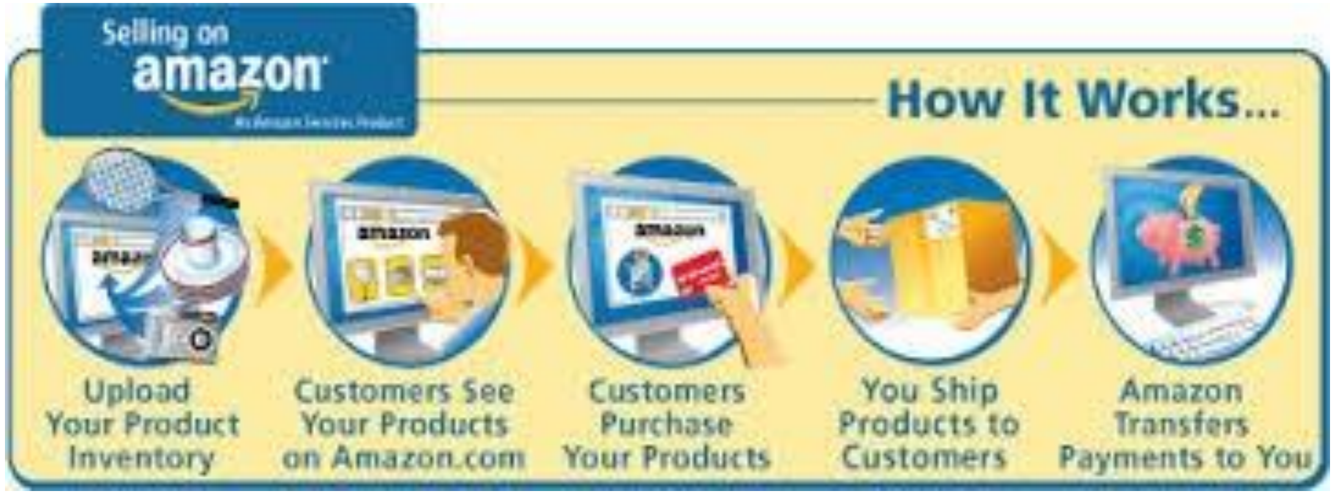
Through November 29, all AT&T BlackBerry phones are on sale starting at a penny with no activation fee

Amazon is providing online infrastructure for online retailers.

Amazon’s infrastructure, which includes packing.

And shipping (logistics)

Amazon.com


Selling on
amazon
How It Works...
Upload Your Product Inventory
Customers See Your Products on Amazon.com
Customers Purchase Your Products
You Ship Products to Customers
Amazon Transfers Payments to You

- An assembly line for the retail industry, with support for embedded online retailers.
- Huge stock of books, DVDs, CDs, etc. *easily searchable*
- Extensive customer *reviews*

Amazon.com

Information-based differentiation:

Satisfies customers by providing a differentiated service:

- Superior information including ***reviews*** about products
- Superior ***range***

Information-based delivery network:

They deliver information for others; retailers in the Amazon marketplace get:

- Customers ***directed*** to them
- Other retailers' support

Data Business Models

- Information brokering service:
  - Buys and sells data/information for others.
- Information-based differentiation:
  - Satisfies customers by providing a differentiated service built on the data/information. (www.amazon.com)
- Information-based delivery network:
  - Deliver data/information for others. (www.reuters.com) (www.plentisoft.com)
- Information provider:
  - Business selling the data/information it collects. (www.Nielsen.com)

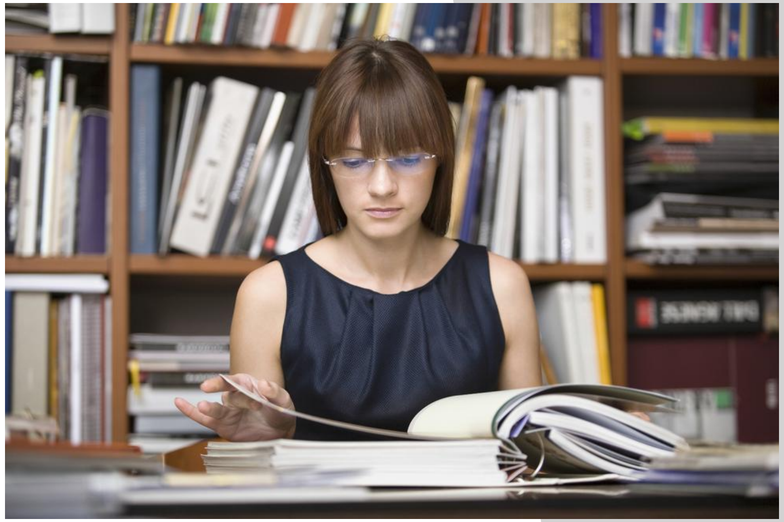
"What a Big-Data Business Model Looks Like" by Ray Wang in the Harvard Business Review claims these are unique in the data world.

Home Activities

Suggested Activities for the week

Videos

Watch John Schitka "Big Data – 2020 Vision"



From Data Analytics Handbook (Pt1) read the interviews of

- Abraham Cabangbang (2 pp) (pp 5 - 7)
- Ben Bregman (2 pp) (pp 13 - 15)
- Leon Rudyak (3 pp) (pp 16 - 19)

MONASH University

Recap: Learning Outcomes

Week 2

By the end of this week you should be able to:

- Comprehend essentials for coding in Python for data science
- Explain and interpret given **Python** codes
- Explain **different data science roles** and skills and comprehend the differences between them. (Read Analyzing the Analyzers, Harris, Murphy and Vaisman, 2013)
- Explain **Impact** of data science. (Data in cloud, Social good, Futurology)
- Explain the **data business models** for organizations, e.g. how amazon.com uses data to be more competitive.