Data Science Training

November 2017

Introduction to Data Science

Xavier Bresson

Data Science and AI Research Centre
NTU, Singapore

http://data-science-optum17.tk

About Me

Prof. Xavier Bresson

xavier.bresson@gmail.com

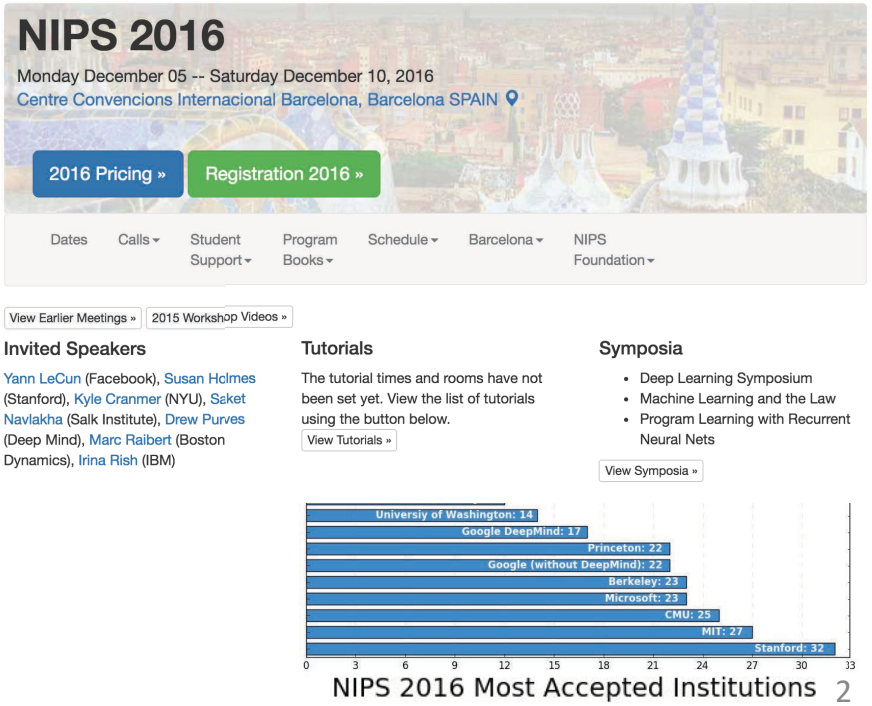
http://www.ntu.edu.sg/home/xbresson

https://github.com/xbresson

https://twitter.com/xbresson

https://www.linkedin.com/in/xavier-bresson-738585b

- *Prof of Data Science, NTU, Singapore*
- *Data Science and AI Research Centre*
- *Received 2.5MUS$ for DL research*
- *Teach Master and PhD courses in DS*
- *Consulting (Nestle, Cisco, Philips)*
- *Publications at NIPS, ICML, JMLR*



About Me

Neural Information Processing Systems (NIPS)

Geometric deep learning on graphs and manifolds

M. Bronstein, J. Bruna, A. Szlam, X. Bresson, Y. LeCun

Dec 4, 2017

New Deep Learning Techniques

Feb 5-9, 2018

Organizers

Xavier Bresson, NTU

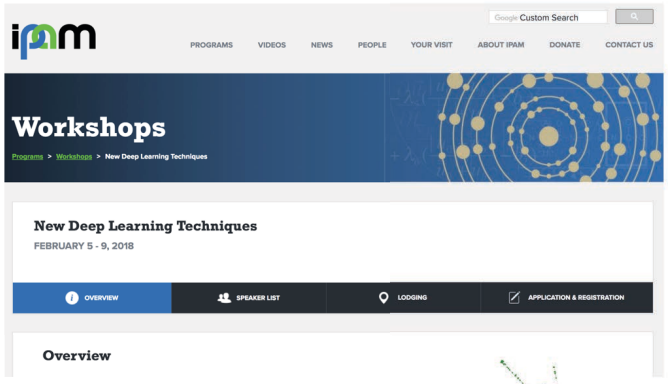
Michael Bronstein, USI/Intel

Joan Bruna, NYU/Berkeley

Yann LeCun, NYU/Facebook

Stanley Osher, UCLA

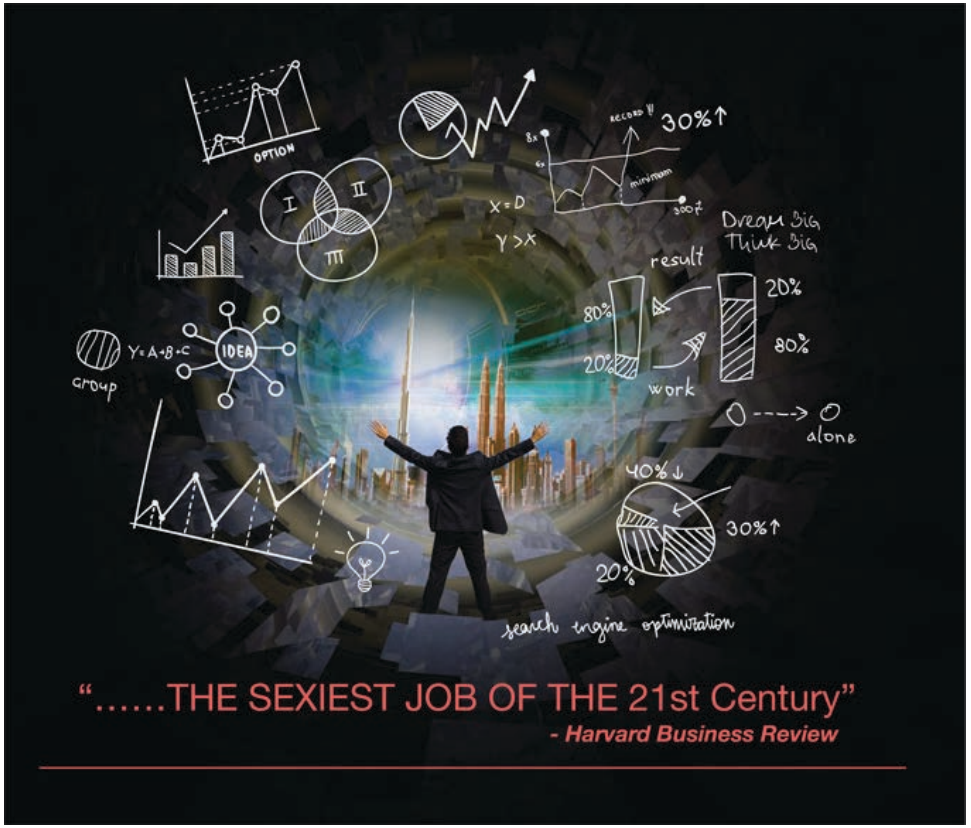
Arthur Szlam, Facebook

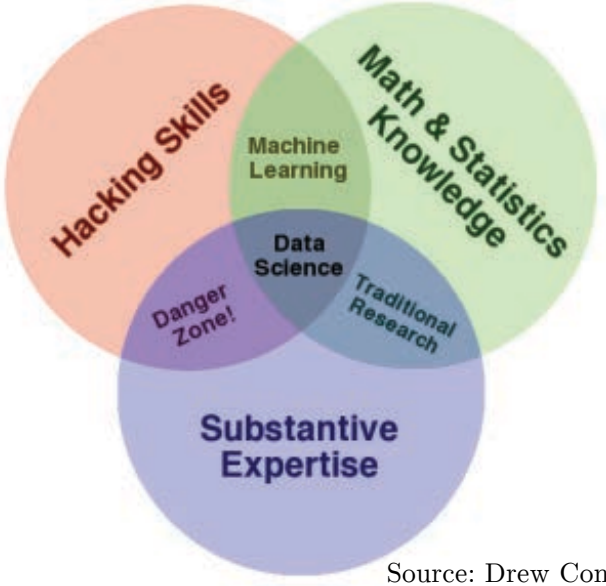


UCLA

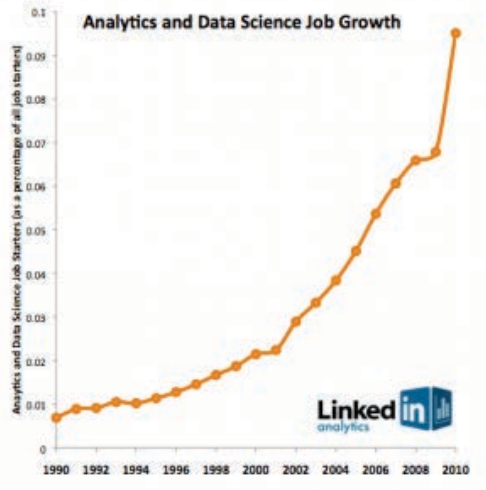
Participants

Data Scientist





Source: Drew Conway



Best job in the U.S in 2015 [Forbes, LinkedIn].

Salary has jumped from \$125,000 to \$200,000+ [Glassdoor].

McKinsey projects that "*by 2018, the U.S. alone may face a 50 percent to 60 percent gap between supply and requisite demand of deep analytic talent.*"

In the News


The New York Times

Violent Protests in Egypt As Leader Expands Power

Morsi Declares 'God's Will and Elections Made Me the Captain of This Ship'

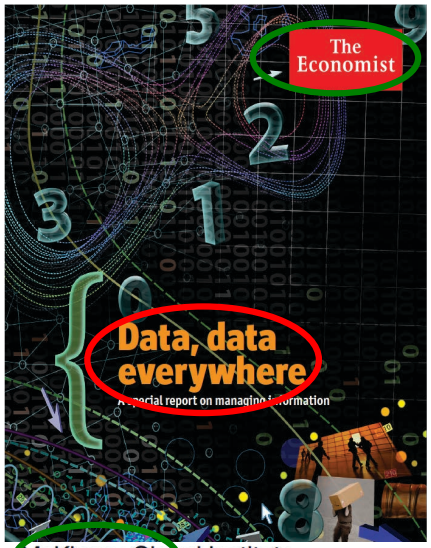
Optimistic Consumers Go on Shopping Spree

Learning Curve: No Longer Just A Human Trait

New Zealand Wants a Hollywood Put on Its Map

Tension and Confusion Linger In Gaza Strip After Cease-Fire

In Storm's Debris, the Macabre



McKinsey Global Institute

People

Contact us

Big data: The next frontier for innovation, competition, and productivity

May 2011

Download | Executive Summary PDF-922KB | Full Report PDF-6MB | Kindle MOBI-4MB | eBook EPUB-3MB

The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus, according to research by MGI and McKinsey's Business Technology Office. Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future.

Interactive

Deep analytical talent: Where are they now?

Research by MGI and McKinsey's Business Technology Office

MGI studied big data in five domains—healthcare in the United States, the public sector in Europe, retail in the United States, and manufacturing and personal-location data globally. Big data can generate value in each. For example, a retailer using big data to the full could increase its operating margin by more than 60 percent. Harnessing big data in the public sector has enormous potential, too. If US healthcare were to use big data creatively and effectively to drive efficiency and quality, the sector could create more than $300 billion in value every year. Two-thirds of that would be in the form of reducing US healthcare expenditure by about 8 percent.

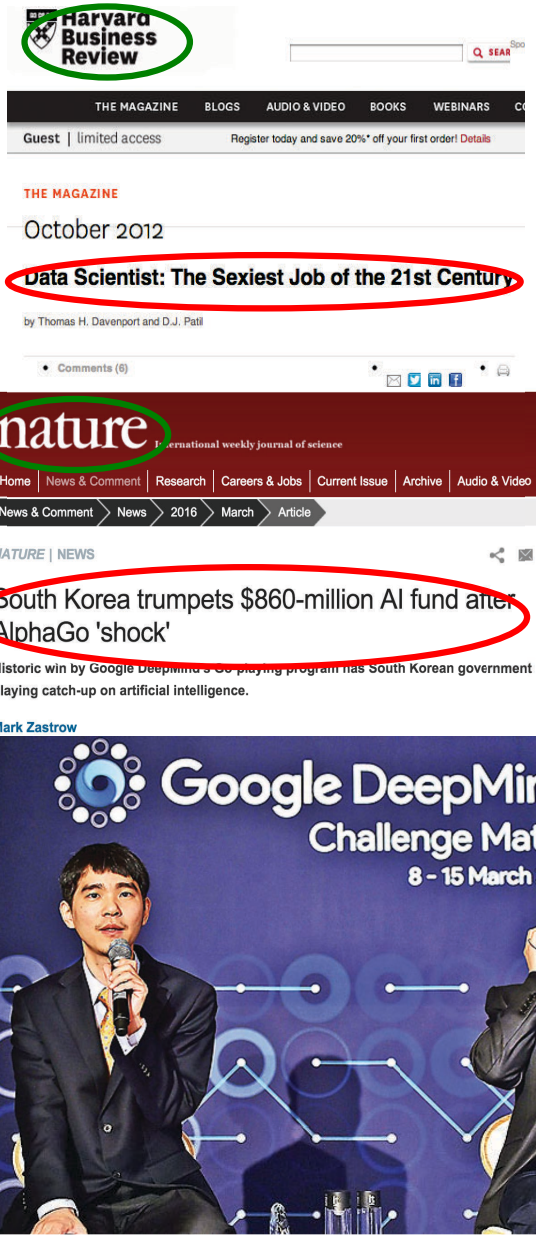
Related content

The social economy: Unlocking value and productivity through social technologies

Are you ready for the era of 'big data'?

Internet matters: The net's sweeping impact on growth, jobs, and prosperity

The great transformer: The impact of the Internet on economic growth and prosperity


Harvard Business Review

THE MAGAZINE | BLOGS | AUDIO & VIDEO | BOOKS | WEBINARS

Guest | limited access

Register today and save 20%* off your first order! Details

THE MAGAZINE

October 2012

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (6)

nature

International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For

News & Comment | News | 2016 | March | Article

NATURE | NEWS

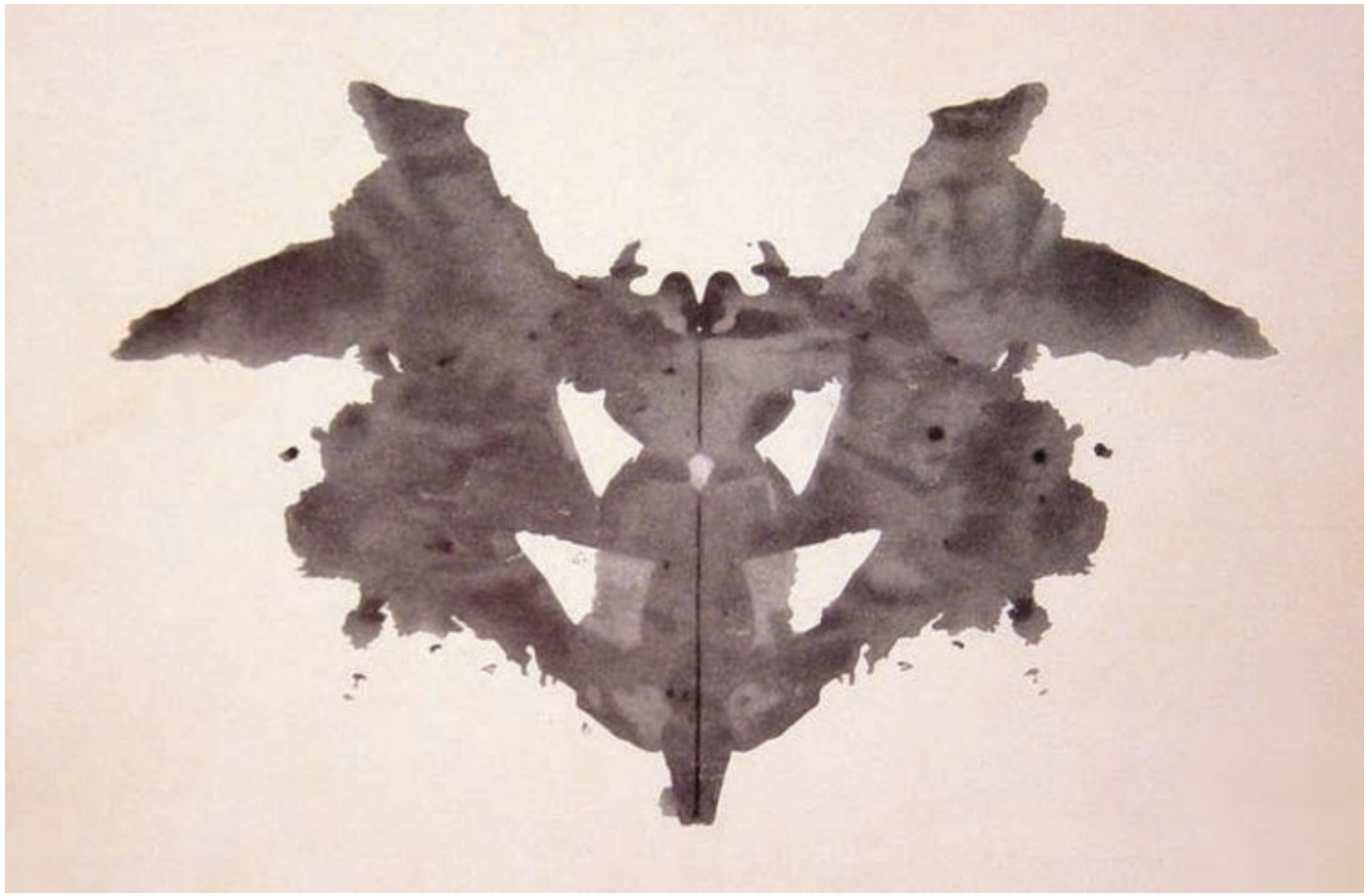
South Korea trumpets $860-million AI fund after AlphaGo 'shock'

Historic win by Google DeepMind's Go-playing program has South Korean government playing catch-up on artificial intelligence.

Mark Zastrow

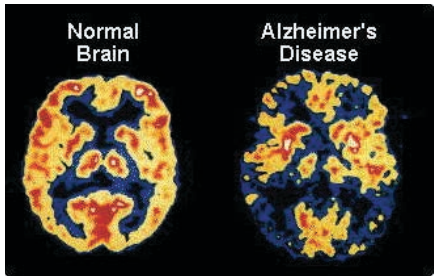
Data Science

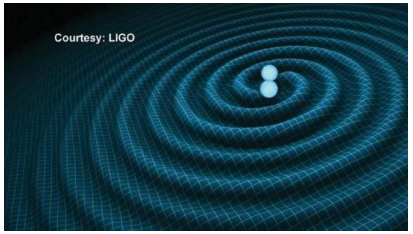
Q: What is Data Science?

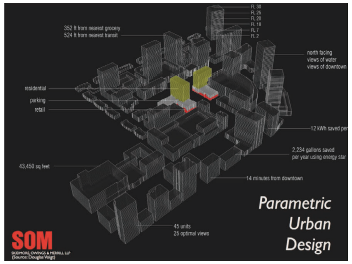


What is Data Science? - Short Answer

Science of transforming raw data into meaningful knowledge to provide smart decisions to real-world problems.







What is Data Science? - Long Answer

Q: What are the fields of Data Science?

Computer Science

Scalable databases for storing, accessing data.
E.g. Cloud computing, Amazon EC2, Hadoop.
Distributed and parallel frameworks for data processing.
E.g. MapReduce, GraphLab.

Mathematical Modeling

Design algorithms that transform data into knowledge.
Use Linear algebra, optimization, graph theory, statistics.

Data

Collection of massive amounts of data at increasing rate.
E.g. Social networks, sensor networks, mobile devices, biological networks, administrative, economics data

Issues of privacy, security, ownership

Data Science

Multidisciplinary field: "1+1=3"

Domain Expertise

Sciences
E.g. Economy, Biology, Physics, Neuroscience, sociology.
Government
E.g. Healthcare, Defense, Education, Transportation.
Industry
E.g. E-commerce, Telecommunications, Finance.

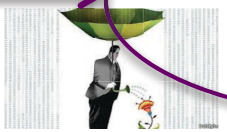
Q: What are the applications?

Personalized Services

E.g. Healthcare (enhanced diagnostics)
Commerce (products)

Knowledge Discovery

E.g. Physics, genomics, social sciences.



Intelligent Systems

E.g. Autonomous cars, chatbots, security, interactive tools for data organization and exploration.

Q: What are the main challenges?

Major challenges: Multidisciplinary integration, large-scale databases, scalable computational infrastructures, design math algorithms for massive datasets, trade-off speed and accuracy for real-time decisions, interactive visualization tools.

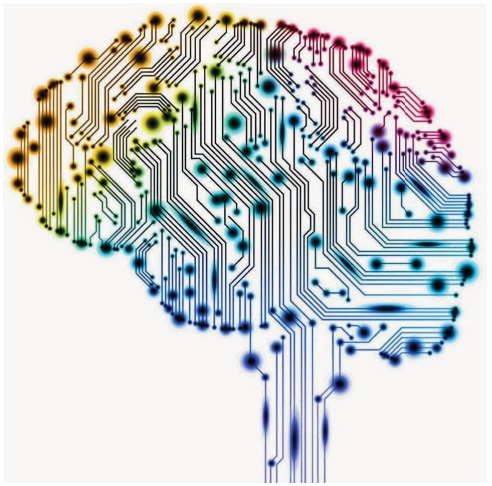
What is Data Science? - Medium Answer

Q: What is big data?

Q: Is AI new?

$$\textit{Data Science} = \underbrace{\textit{Big Data}}_{\text{3rd industrial revolution}} + \underbrace{\textit{Computational Infrastructure}}_{\text{Cloud computing, GPU}} + \textit{Artificial Intelligence}$$

Not new!

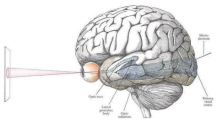


A Brief History of AI

Q: Have you heard about the 4th industrial revolution?

Hardware
GPU speed 2x/1.5 year

Big Data
Volume 2x/1.5 year



Visual primary cortex
Hubel-Wiesel
1959

First
NIPS

RNN/LSTM
Schmidhuber

CNN
LeCun

First
KDD
1989

1997 1998

First Amazon
Cloud Center

Kaggle
Platform

GAN, Goodfellow
ResNet, He
Deep RL, Mnih

Google AI
TensorFlow
Facebook AI
Torch

1958 1962 1982 1987 1995 1999 2006 2010 2012 2014 2015

Perceptron
Rosenblatt

Backprop
Werbos

Neocognitron
Fukushima

Birth of
Data Science
Split from Statistics
Tukey

First NVIDIA
GPU

SVM/Kernel techniques
Vapnik

amazon web services

IMAGENET

Deep Learning
Breakthough
Hinton

Facebook Center
OpenAI Center

OpenAI DeepMind facebook

Data scientist
1st Job in US

4th industrial revolution?
Digital intelligence revolution or new AI bubble?

AI Birth

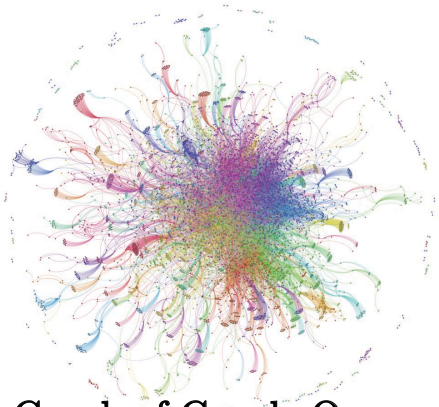
AI Winter [1960's-2012]

AI Renaissance

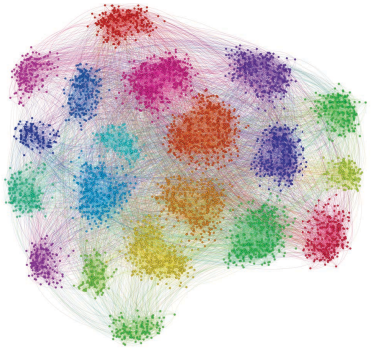
Kernel techniques
Graphical models
Wavelets, numerical PDEs
Handcrafted Features

Networks/Graphs

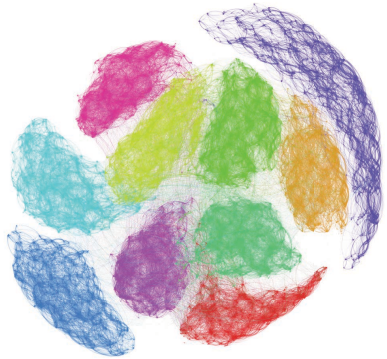
- Graphs encode complex data structures. They are everywhere: WWW, Facebook, Amazon, etc



Graph of Google Query "California"



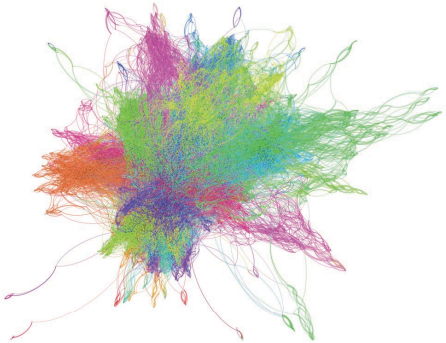
Social Network



MNIST Image Network

GTZAN Music Network

- "Graphs are the most important discrete models in the world!" - G. Strang (MIT)

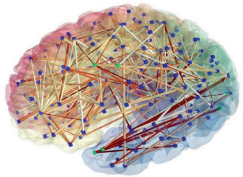


Network of Text Documents 20newsgroups

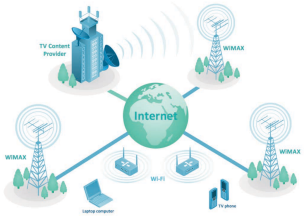
Why Networks are important?

- Networks improve all data science tasks.
- Essential data lie on networks:
  (1) Social networks (Facebook, Twitter)
  (2) Biological networks (genes, brain connectivity)
  (3) Communication networks (Internet, wireless, traffic)

Social networks



Brain structure



Telecommunication networks

=

Graph/network-structured data

Program Schedule

Monday Nov 20, 2017

Lecture 1 : Introduction to Data Science
Lecture 2 : Python
Lecture 3 : SVM Techniques
Lecture 4 : Feature Extraction
Lecture 5 : Data Visualization (if time permits)
Discussion with Optum team

Thursday Nov 21, 2017

Lecture 6 : Introduction to Deep Learning
Lecture 7 : Classification with Neural Networks
Lecture 8 : Training Neural Networks
Lecture 9 : TensorFlow
Lecture 10 : Reinforcement Learning (if time permits)
Discussion with Optum team

Wednesday Nov 22, 2017

Lecture 11 : Convolutional Neural Networks
Lecture 12 : Recurrent Neural Networks
Lecture 13 : Graph Science (if time permits)
Lecture 14 : Deep Learning on Graphs (if time permits)
Lecture 15 : Conclusion
Discussion with Optum team

http://data-science-optum17.tk

Data Science Training

Optum Services Ireland Ltd

November 20-22, 2017

Overview

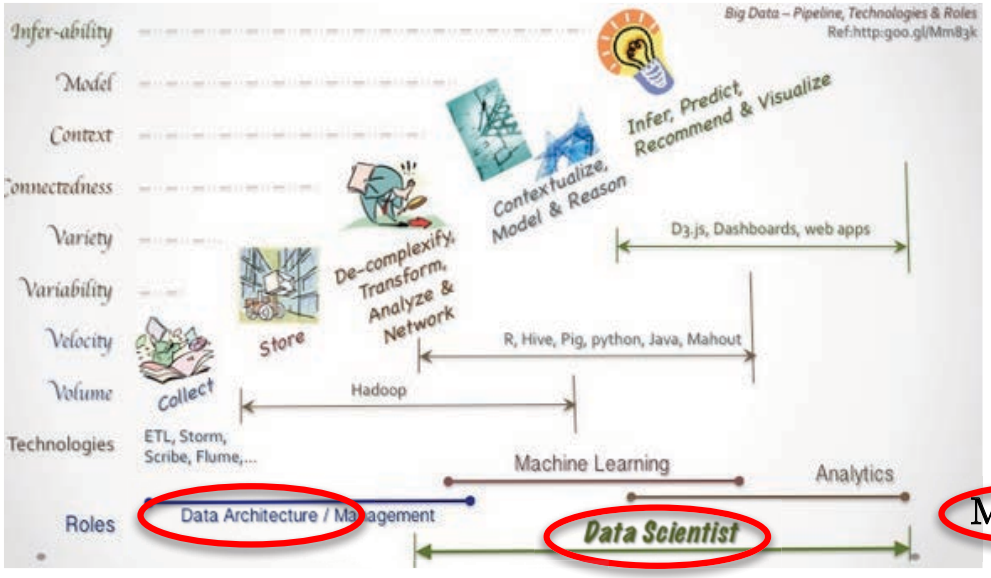
In 2012, the Harvard Business School called data scientist the sexiest job of the
a 50-60 percent gap between supply and requisite demand of deep analytic tale
Today massive amounts of data are available in all areas of science, governr
efficiency of research, services and industries in as many fields as healthcare, er
How are powerful companies like Google, Facebook, IBM or Apple using data
apply them to your projects ?

Coffee/croissants from 8:30
9:00 – 11:00 Session 1
Break: Coffee/Tea/Refreshments
11:15 – 13:15 Session 2
Lunch:
14:15 – 15:45 Session 3
Break: Coffee/Tea/Refreshments
16:00 – 17:30 Session 4
Finish

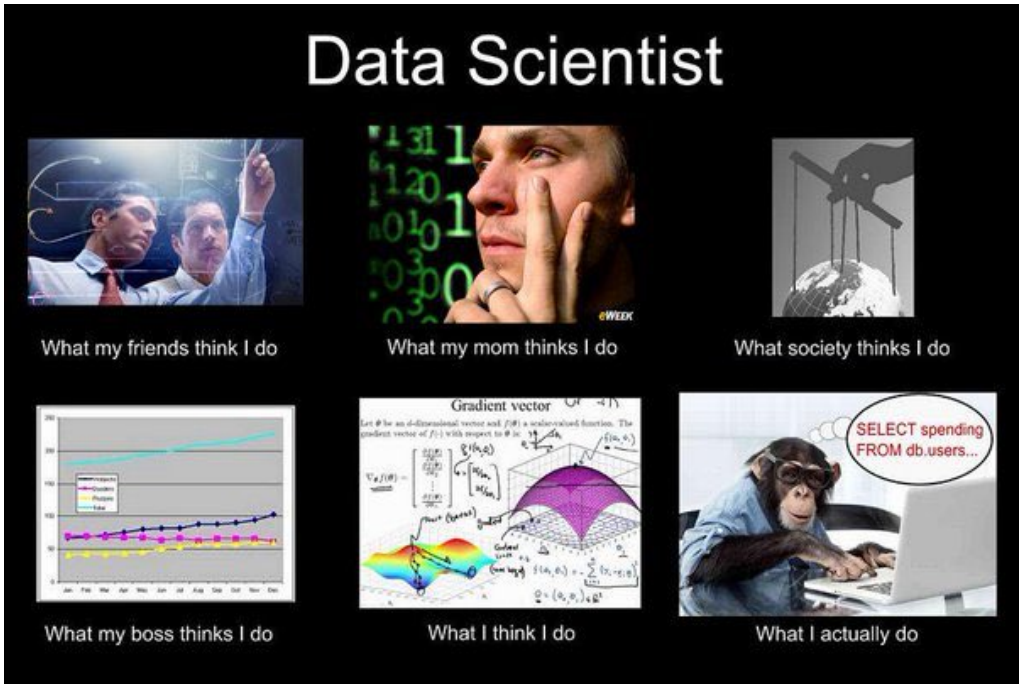
Structure of the Course

- Introduction of main concepts/ideas, technical details
- Coding illustration on real-world data
- Coding exercise at the end of each lecture
- Please, share your own data science problem for discussion
- Please, ask questions!

Goal of the Course



Data science training



Pipeline of Data Science

Data World
Ex: social, medical, business data

Collect raw data
Ex: API crawler ☺
Robot scrapper
(hacking skills)

Storage
Ex: Harddrive
Databases: SQL
Hybrids: hdf5

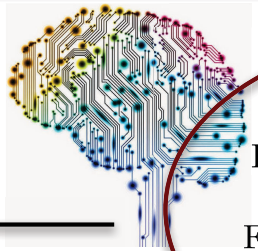
Clean data

Domain expertise

Common Sense

Cleaning
Ex: Remove corrupted, incomplete data

☹: Start again



Data Analysis
Ex: Unsupervised learning
Supervised learning
Feature/Pattern extraction
Visualization
Deep learning

Solve real-world challenges

Report results (html, pdf, excel)

Management
Decision level

☺

Production
Ex: Softwares
Devices

Domain Expertise
Ex: science, government, business

Coding
Python dev

Maths
Statistics, optimization, graphs, linear algebra

Data Science Core:
Design powerful data analysis algorithms

Computer Science
Large-scale computing platforms
Ex: Hadoop, Spark, Graphlab

Case Study: Mining Star Wars Universe

Data World

Wookieepedia

The Star Wars encyclopedia

+130,000 webpages

http://starwars.wikia.com/wiki/Main_Page

Collect raw data

Robot web scrapper

Extract data from every character of Wookieepedia

Storage

Neo4j Graph Database

Clean data

Cleaning

A character A is linked to another one B if its name appears on A's associated wiki page.

Data Analysis

Supervised learning

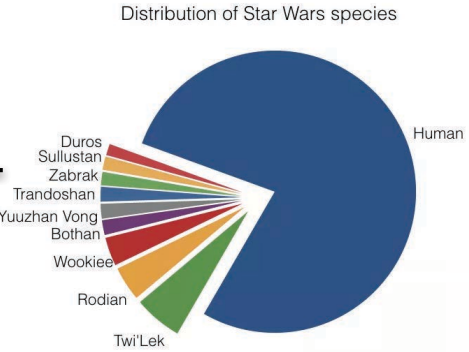
Visualization

Feature/Pattern extraction

20,000 characters spread among 640 communities over a period of 36,000 years

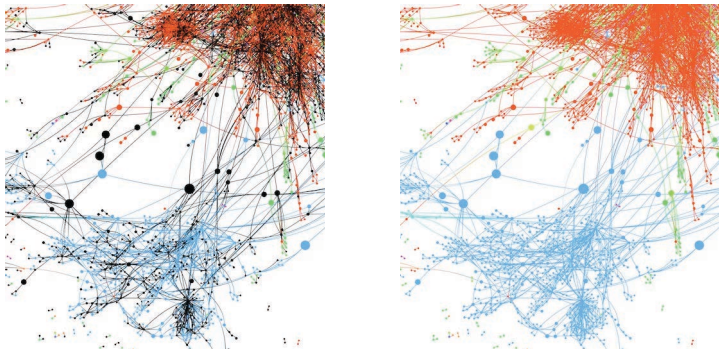
Results:

Analysis of Star Wars Universe



EPFL PR Mediacom Feb 2016

Visualization of Star Wars Universe



Filling missing information of characters

Historical application: find the time period of all characters using their connections?

Domain Expertise

Star Wars

Computer Science

Python dev

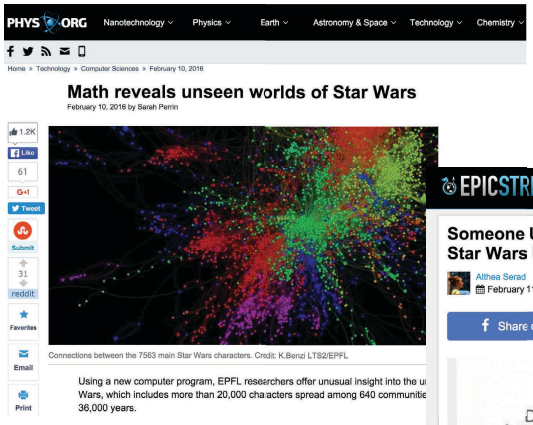
Maths

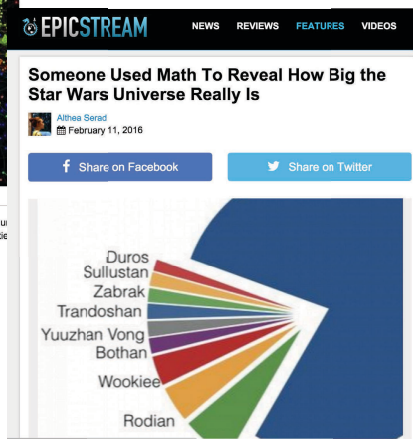
Statistics, optimization, graphs, linear algebra

More info: http://kirellbenzi.com/blog/exploring-the-star-wars-expanded-universe

Case Study: Mining Star Wars Universe

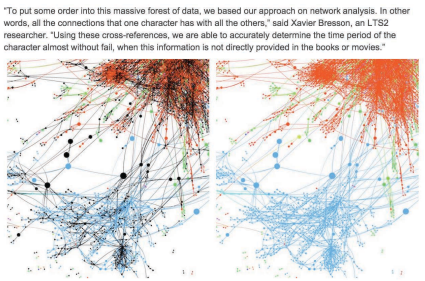
100+ international articles *including popular* ***Dailymail*** *(1.66M Twitter followers),* ***Gizmodo*** *(1.48M),* ***Engadget*** *(1.57M),* ***MUY Interesante*** *(7.1M).*





The mythical world of Star Wars can feel almost as vast as the universe itself, with characters and events spanning not just seven films, but also a plethora of books, comics and video games.

Now, thanks to data collected from the Ecole Polytechnique Fédérale de Lausanne (EPFL), the extended Star Wars universe is easier to take in for nerd and newcomer alike.



"To put some order into this massive forest of data, we based our approach on network analysis. In other words, all the connections that one character has with all the others," said Xavier Bresson, an LTS2 researcher. "Using these cross-references, we are able to accurately determine the time period of the character almost without fail, when this information is not directly provided in the books or movies."

Part of the Star Wars character graph colored by era. Black nodes represent missing values. Red nodes: the Rise of the Empire Era (episodes 1,2,3). Blue nodes: the Rebellion era (episodes 4,5,6). Green ones: both eras.Figure 2: Result of the label propagation algorithm. Black nodes have been replaced by the best compromise using their neighbors.

Mapping out connections

The researchers want to use this study to demonstrate the program's ability to extract and analyze digital data. "The program maps out connections in the mass of unorganized data available on the net," said Benzi.

Given a huge amount of information, the algorithms developed by the LTS2 researchers offer a service that cannot be matched by human beings. In addition to extracting data according to extremely precise criteria, the algorithms can also create links among data points, sort them, quantify them, interpret them and find missing information. All this in very little time. The results are then presented in the form of interactive charts

References

- **Reviews of data science (w/ the famous discussion statistics vs. data science):**
  - J.W. Tukey, The Future of Data Analysis, 1962
  - L. Breiman, Statistical modeling: The two cultures, 2003
  - Frontiers in Massive Data Analysis, National Academy of Sciences, 2013
  - D. Donoho, 50 years of Data Science, 2015
  - Study Panel, Stanford University, One Hundred Year Study on Artificial Intelligence, 2016

Questions?