

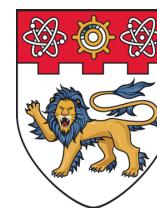
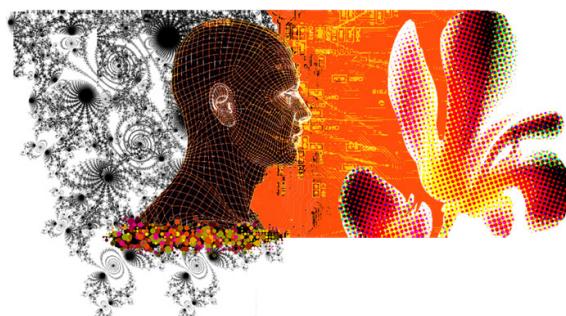
CE9010: Introduction to Data Science

Lecture 1: Introduction

Semester 2 2017/18

Xavier Bresson

School of Computer Science and Engineering
Data Science and AI Research Centre
Nanyang Technological University (NTU), Singapore



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

About Me



Prof. Xavier Bresson
xbresson@ntu.edu.sg

Office: N4-02b-40

Tel: 6790 6300

- *Prof of Computer Science*
- *Undergraduate in France, PhD in Switzerland, Postdoctorate in the US*
- *Member of NTU Data Science and AI Research Centre*
- *NRF Fellowship for DS research*
- *Teach Master, PhD courses in DS since 2015*
- *Consulting/training in DS for companies*
- *Publications at NIPS, ICML, JMLR*

- <http://www.ntu.edu.sg/home/xbresson>
- <https://github.com/xbresson>
- <https://twitter.com/xbresson>
- <https://www.facebook.com/xavier.bresson.1>
- <https://www.linkedin.com/in/xavier-bresson-738585b>

Students



CE9010 Class Logistics

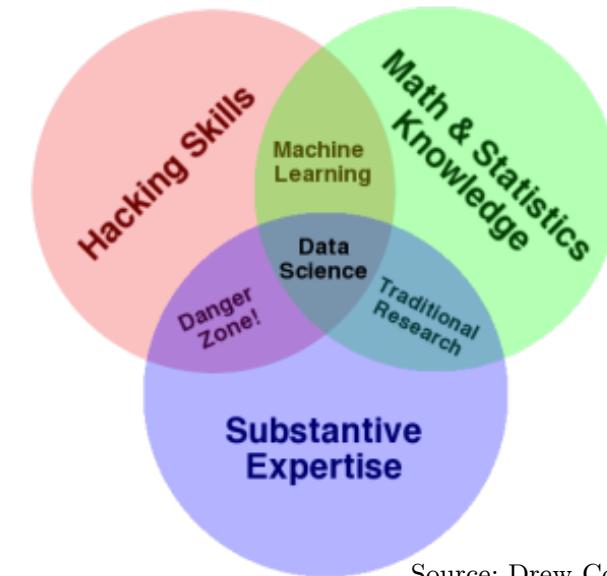
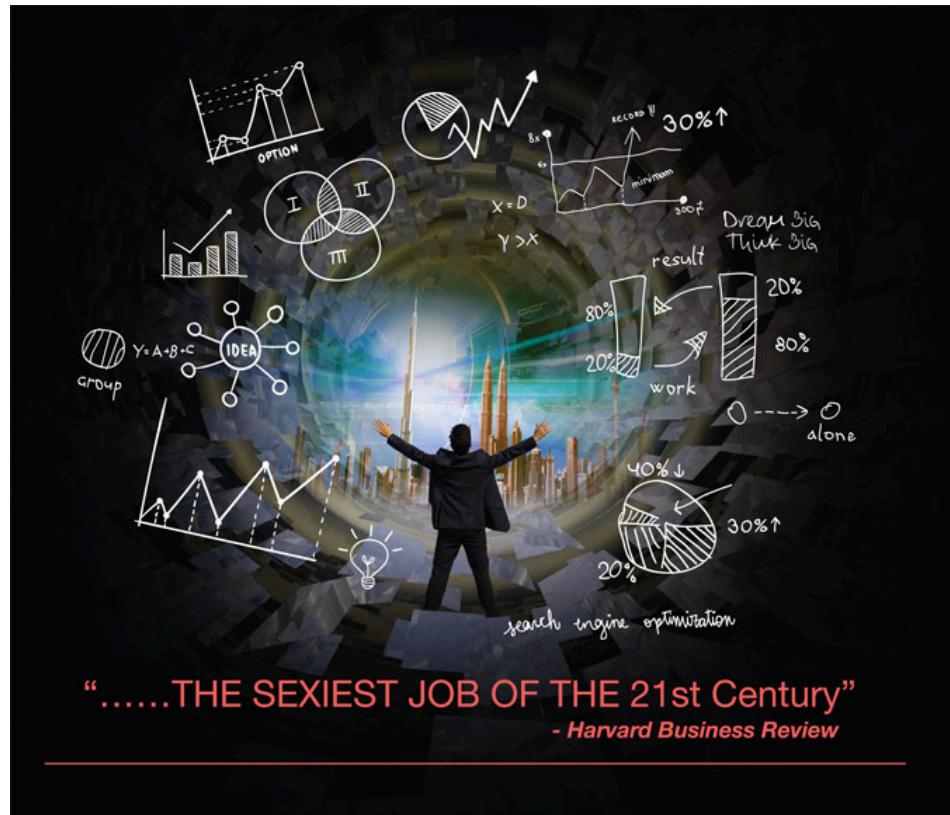
- Class Schedule
 - Lectures: 2 hours on Wednesdays 12:30-14:30, LT10
 - Tutorials: 1 hour on Wednesdays 16:30-17:30, LT10, starting Week 2
 - Laboratories: 5 x 2 hours on Wednesdays 14:30-16:30, SWLAB1/N4-01a-02, Weeks 5,7,9,11,13
- Happy for extra discussion on Wednesdays 17:30-18:30 at my office N4-02b-40.
- Use your own laptops for the coding exercises (tutorials, labs)
⇒ It is important to learn to code (e.g. 15min/day).
- Use NTULearn for course material and
https://github.com/xbresson/CE9010_2018
- Evaluation
 - Laboratory: 30%
 - Project: 30%
 - Final Examination: 40%

Course Schedule

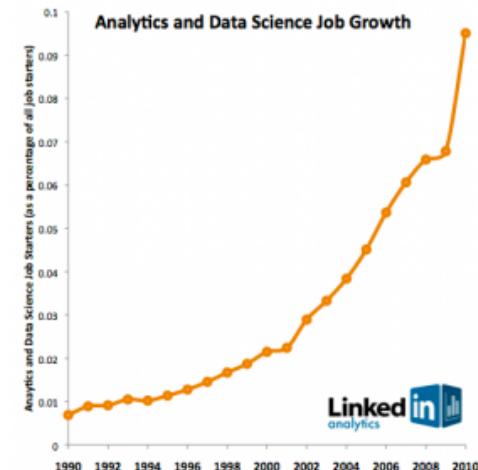
- The course structure follows the successful Andrew Ng Coursera course on Machine learning:

Week	Topic
1	Introduction to data science
	learning techniques
2-4	Supervised regression
5	Supervised classification
6	Gradient descent tricks
7	Generalization and regularization
8-9	Developing data science projects
10-11	Unsupervised learning
12	Recommender systems
13	Neural Networks

Data Scientist



Source: Drew Conway



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

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

McKinsey projected in 2015 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

SATURDAY, NOVEMBER 24, 2012

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 Shopping Spree

Black Friday in New York and elsewhere shows signs of returning to normal, increasingly optimistic about the economy and its impact on sales.

Tension and Confusion Linger In Gaza Strip After Cease-Fire

After a week of fighting, Israel and the Gaza Strip's militant rulers, Hamas, are still at odds over who controls the strip.

Learning Curve No Longer Just A Human Trait

Computers can learn to play chess, recognize faces, and even identify individual people.

In Storm's Debris, the Macabre

A family working along a beach in Queens found the body of a man who had drowned in a bathtub.

A New Sojourn for Europe

A meeting of the European Union's top leaders in Brussels will focus on how to build a bridge for the era and prepare for the challenges ahead.

South Korean Confidante Quits

She is not the only one to leave. The South Korean president has been under fire for her role in a scandal.

As Outbreak of Suntanning Sweeps Europe, Sunscreen Sales Surge

More than 100 million Americans are using some form of sun protection every day.

SPORTS SATURDAY-BEACH

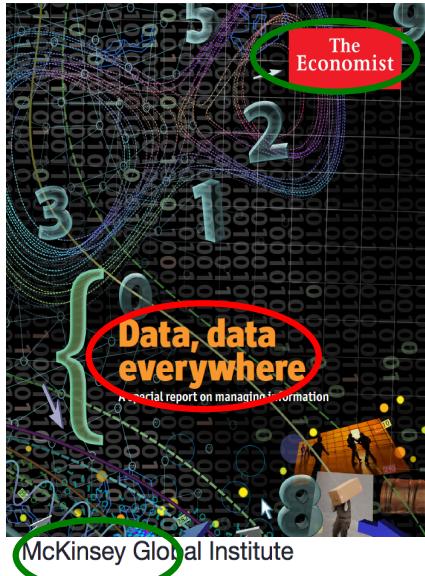
Spots on the Waves

Surfing is a sport that requires a lot of skill and practice to master.

NEW YORK CITY

Capturing Big Ideas

Charles M. Blow



McKinsey Global Institute

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

May 2011 | by James Manyika, Michael Chui, Richard Dobbs, Charles Roxburgh, Angela Hung Byers

Download Executive Summary PDF-622KB 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 and innovation. 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

MGIMI

Harvard Business Review

THE MAGAZINE BLOGS AUDIO & VIDEO BOOKS WEBINARS CO

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 | Forum

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.



What is Data Science?

- Data science is an emerging field that develops mathematical and algorithmic techniques to analyze (big) data and solve real-world problems.
- Data science has different names: machine learning, AI.
- Data science solves tasks by learning from data (like the brain).



Example

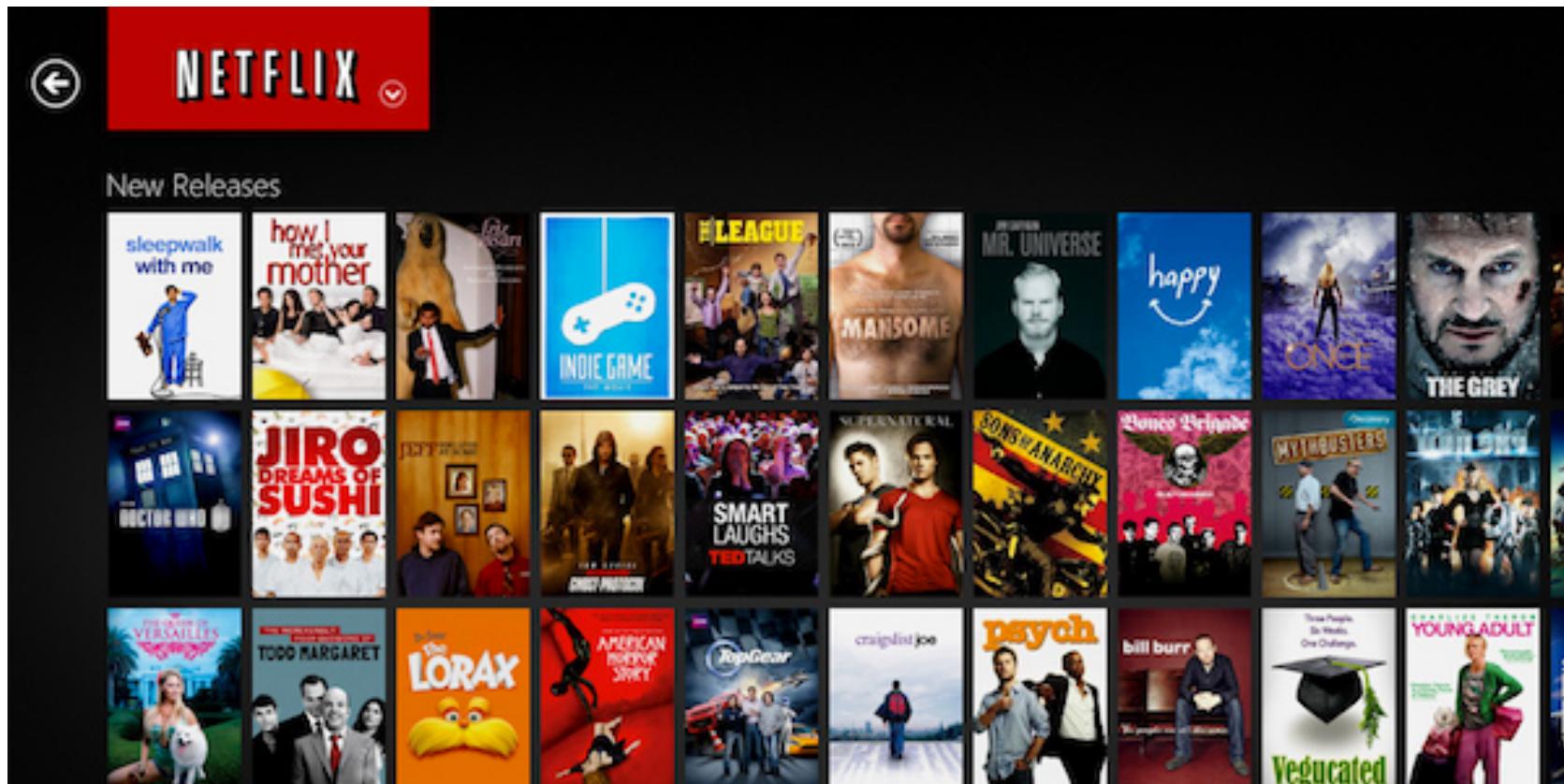
➤ Google search engine

The screenshot shows a Google search results page for the query "data science". The search bar at the top contains "data science". Below the search bar are navigation links for "All", "Images", "News", "Videos", "Books", and "More", along with "Settings" and "Tools" buttons. A status message indicates "About 64,100,000 results (0.91 seconds)". The search results are listed below:

- Data science - Wikipedia**
https://en.wikipedia.org/wiki/Data_science ▾
Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.
Business analytics · Information science · Information explosion · Jim Gray
- Data Science vs. Big Data vs. Data Analytics - Simplilearn**
<https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article> ▾
Dealing with unstructured and structured data, Data Science is a field that encompasses anything related to data cleansing, preparation, and analysis. Put simply, Data Science is an umbrella term for techniques used when trying to extract insights and information from data.
20 Most Popular Data Science ... · Data Analyst vs. Data Scientist
- Data Science I Coursera**
<https://www.coursera.org/specializations/jhu-data-science> ▾
Data Science from Johns Hopkins University. Ask the right questions, manipulate data sets, and create visualizations to communicate results. This Specialization covers the concepts and tools you'll need throughout the entire data science ...
- Data Science I edX**
<https://www.edx.org/micromasters/data-science> ▾
Excel in Data Science, one of the hottest fields in tech today. Learn how to gain new insights from big data by asking the right questions, manipulating data sets and visualizing your findings in compelling ways. In this MicroMasters program, you will develop a well-rounded understanding of the mathematical and ...
- Data Science & Artificial Intelligence Research Centre@ NTU (DSAIR)**
<https://dsair.ntu.edu.sg/> ▾
Data Science & Artificial Intelligence Research Centre@ NTU (DSAIR)
You've visited this page 5 times. Last visit: 17/12/17
- Top Data Science Courses Online - Updated January 2018 - Udemy**
<https://www.udemy.com/topic/data-science/> ▾
Learn Data Science today: find your Data Science online course on Udemy.

Example

- Netflix: Movie recommendations



Example

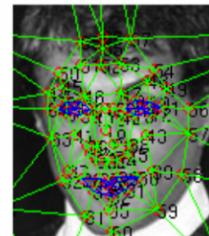
➤ Facebook face recognition



(a)



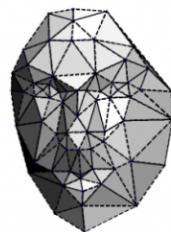
(b)



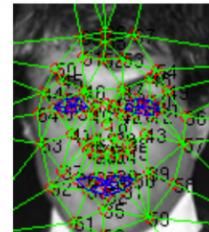
(c)



(d)



(e)



(f)



(g)



(h)



Gary Chavez added a photo you might ...
be in.

about a minute ago · 



Example

➤ Google translate machine

<i>Input sentence:</i>	<i>Translation (PBMT):</i>	<i>Translation (GNMT):</i>	<i>Translation (human):</i>
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

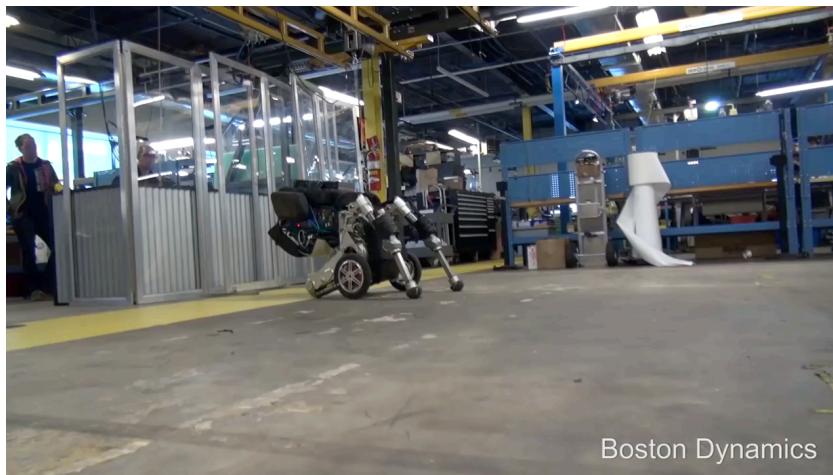
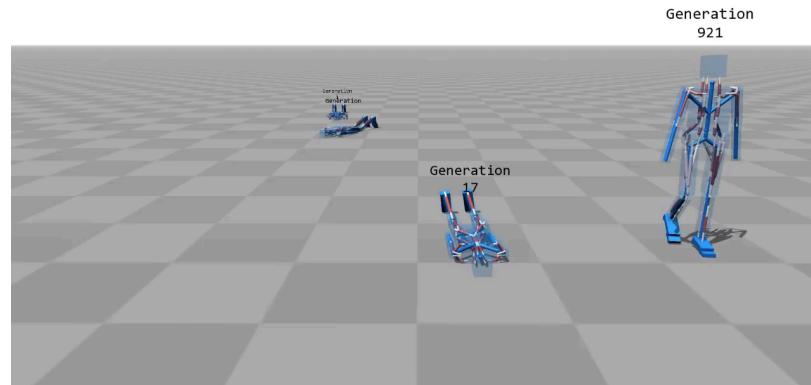
Example

- Self-driving cars

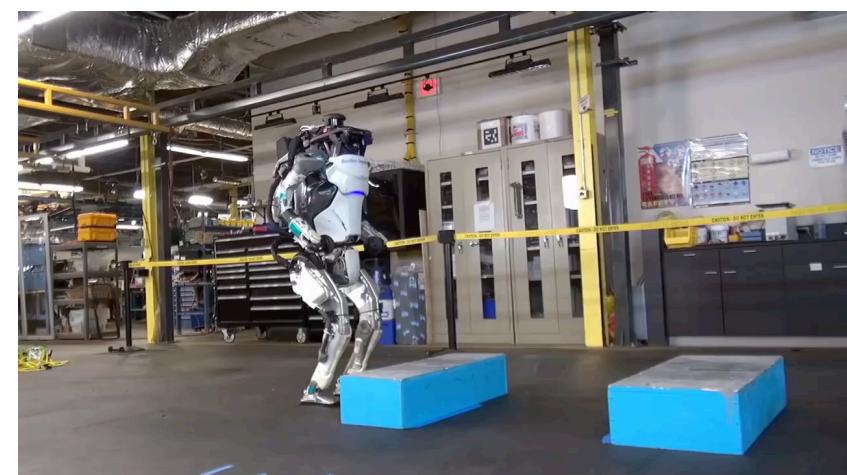


Example

➤ Robots

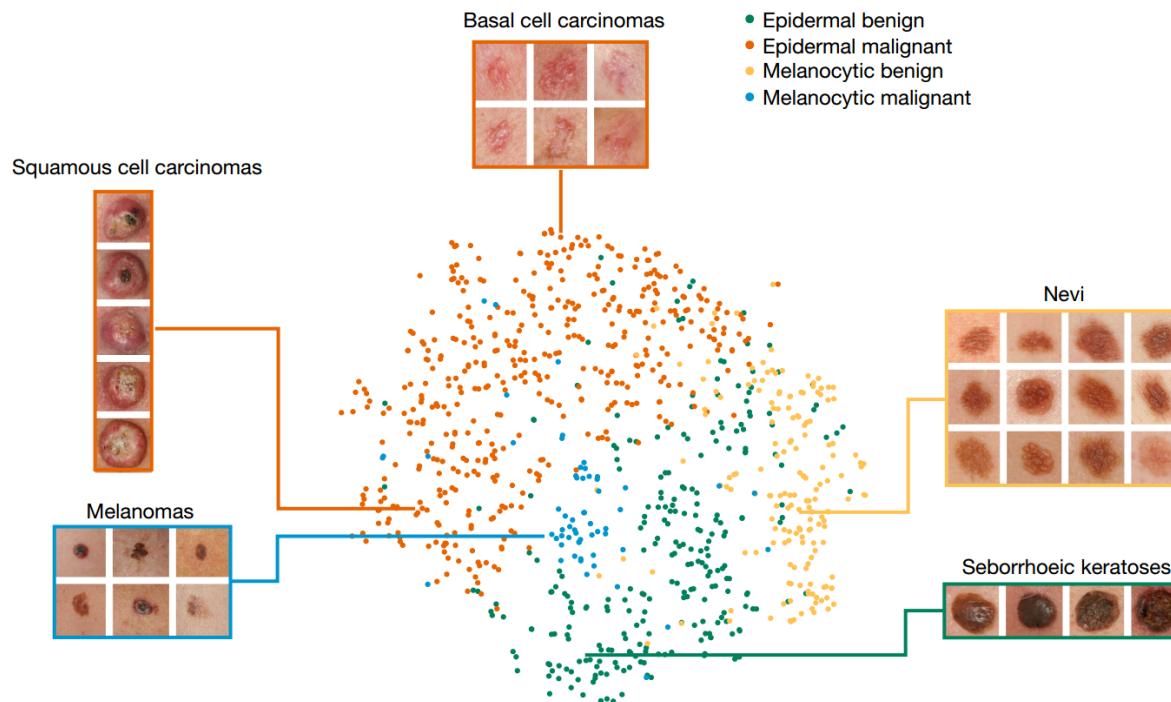


Boston Dynamics



Example

- Medical diagnostic for skin cancer



Follow

A deep learning algorithm made by Stanford researchers outperformed expert radiologists at diagnosing **#pneumonia**: stan.md/2AJ6SIY



4:38 PM - 15 Nov 2017

Example

- Playing games



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



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.



Q: What are the applications?

Personalized Services

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



Knowledge Discovery

E.g. Physics, genomics, social sciences.



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.



Intelligent Systems

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

What is Data Science? – Short Answer

?

Q: What is big data?

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

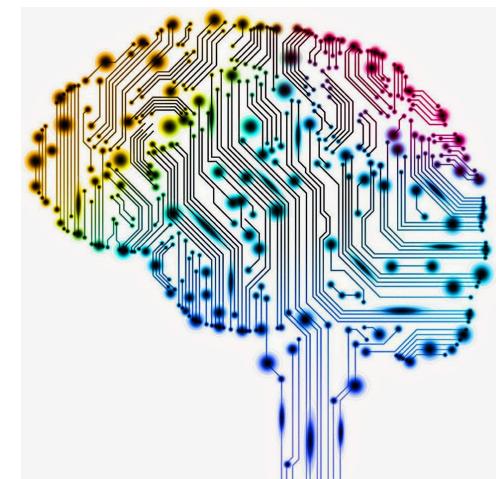


?

Q: Is AI new?

Artificial Intelligence

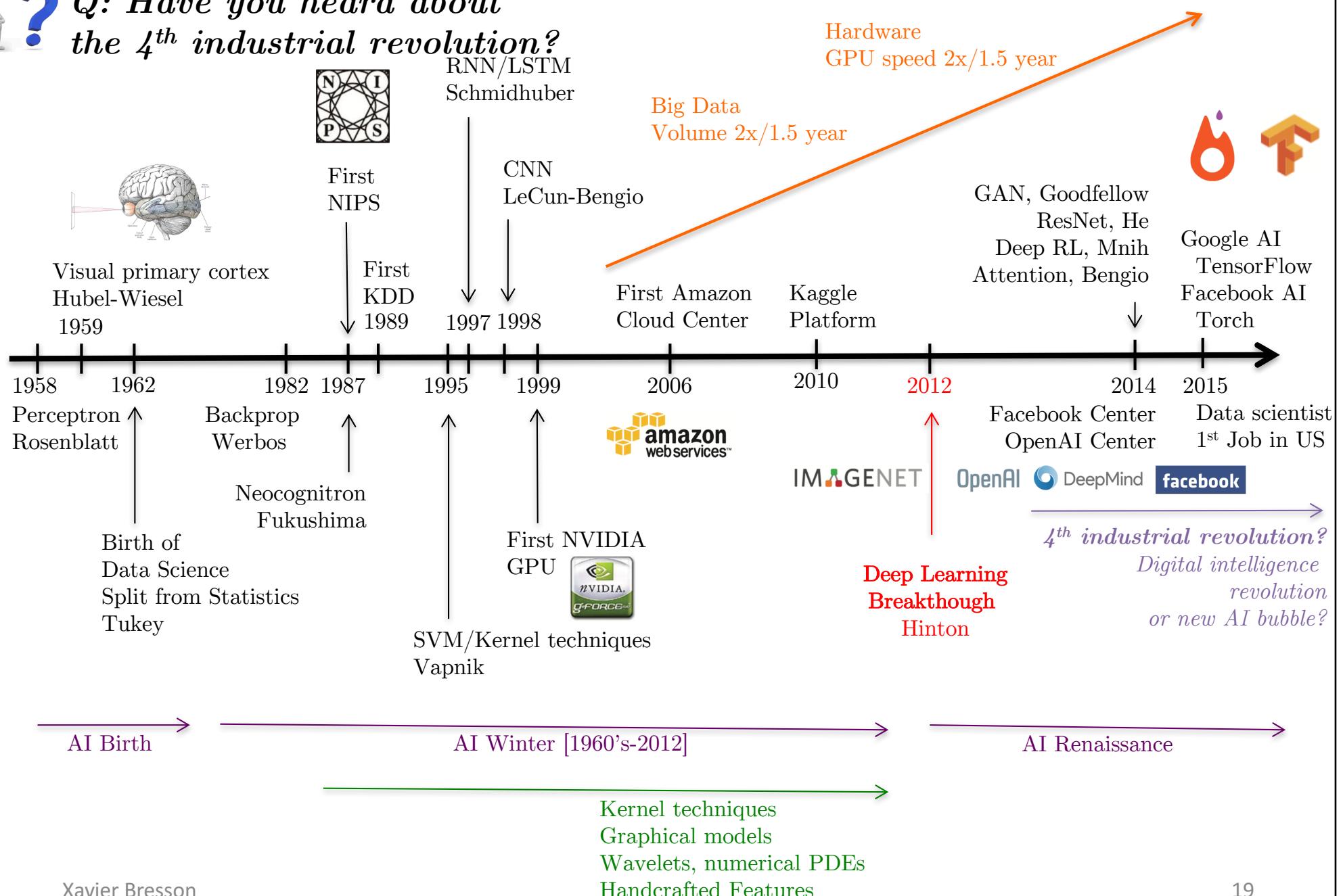
Not new!



A brief history of artificial neural networks



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



Intended Learning Outcomes (ILO)

- By the end of this course, you (as a student) would be expected to be able to:
 - Identify and apply data analysis techniques to different data problems.
 - Implement data science techniques with Python.
 - Analyze and solve real-world data science projects.
 - Engage more professionally in written and oral scientific communication.
 - Work cohesively and effectively in teams.

Why you should consider this course?

1. *Short-term goal:* Learn how to apply the most essential data analysis tools.
2. *Long-term goal:* Data scientist has become the most-wanted job in all industries.



*Data Science is coming.
Better be ready!*

Reference book

- Pattern Recognition and Machine Learning

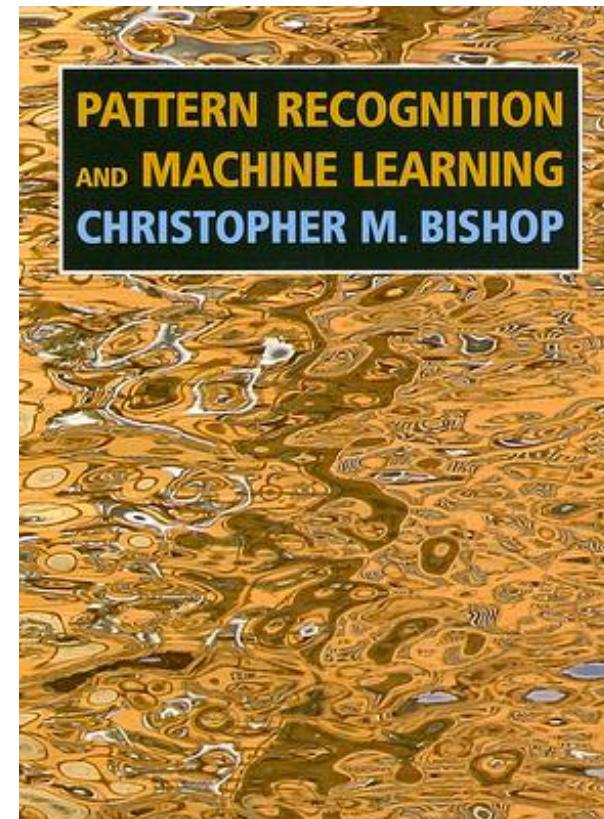
Christopher Bishop

Director at Microsoft Research
Cambridge

Springer-Verlag New York

2006

ISBN 978-0-387-31073-2





Questions are welcome!