Data Science in the Wild

Giri Iyengar

Cornell University

giyengar@gmail.com

Jan 24, 2018

Overview

1. Introduction
   - About your Instructor
   - What is Data Science?
   - What we will cover in this course?

2. Class Mechanics
   - Software tools

Overview

1. Introduction
   - About your Instructor
   - What is Data Science?
   - What we will cover in this course?

2. Class Mechanics
   - Software tools

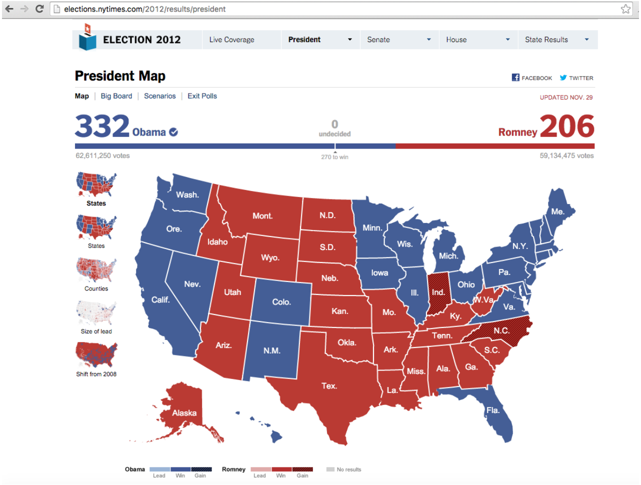
About me

- EE from IIT Mumbai, India. PhD from MIT (Media Lab)
- Researcher at IBM Research doing Audio-Visual Speech Recognition and Multimedia Mining
- Startup No. 1 - Mobile and Social Media apps
- Startup No. 2 - Big Data Machine Learning as a Service. Acquired by AOL/Verizon in 2015
- Engineering Director of Merchandising, currently Head of Computer Vision, eBay

What is the excitement all about?

- Harvard Business Review called it the **sexiest** job of the 21st century (https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/)
- Mashable called it the **best job** in America (http://mashable.com/2016/01/20/the-best-jobs-in-america-2016/), based on a recently concluded Glassdoor annual survey ▸ Mashable

Elections Projections



elections.nytimes.com/2012/results/president

ELECTION 2012 | Live Coverage | President | Senate | House | State Results

President Map

FACEBOOK TWITTER

Map | Big Board | Scenarios | Exit Polls

UPDATED NOV. 29

332 Obama

0 undecided

Romney 206

62,611,250 votes

270 to win

59,134,475 votes

States

States

Counties

Size of lead

Shift from 2008

Obama Lead Win Gain

Romney Lead Win Gain

No results

Figure: Nate Silver Elections projections

A Definition of Data Science

Wikipedia

Data Science is an **interdisciplinary** field about processes and systems to extract knowledge or insights from large volumes of data in various forms, either **structured or unstructured**, which is a continuation of some of the data analysis fields such as statistics, data mining and predictive analytics, as well as Knowledge Discovery in Databases (KDD).
Data scientists use their data and analytical ability to find and interpret **rich data sources**; manage **large amounts of data** despite hardware, software, and bandwidth constraints; merge data sources; ensure consistency of datasets; create visualizations to aid in understanding data; build **mathematical models** using the data; and present and **communicate** the data insights and findings.

Data Science

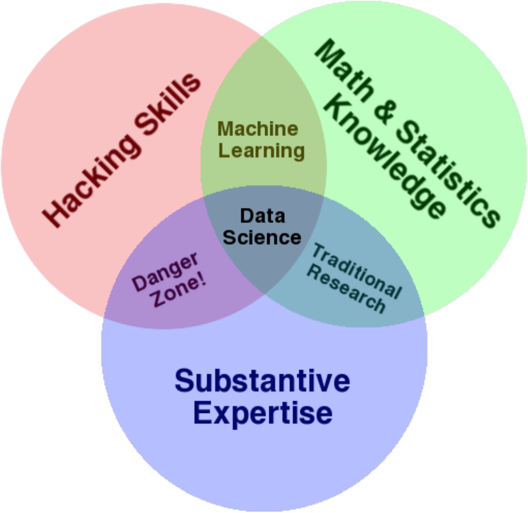


Figure: Drew Conway Venn Diagram

Who is a Data Scientist?

Josh Wills, Slack Data Scientist, Open Source Committer

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

IBM Developer Works

Part Scientist, Part Artist

A broader perspective

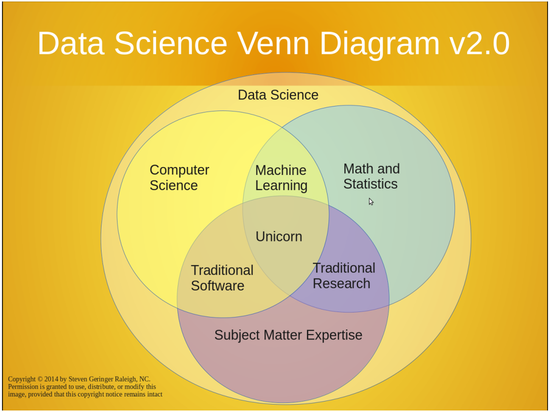

Data Science Venn Diagram v2.0
Data Science
Computer Science
Machine Learning
Math and Statistics
Unicorn
Traditional Software
Traditional Research
Subject Matter Expertise
Copyright © 2014 by Steven Geringer Raleigh, NC. Permission is granted to use, distribute, or modify this image, provided that this copyright notice remains intact

Figure: Rob Hyndman Venn Diagram

The goal of this course

Teach skills beyond Machine Learning and Database management systems

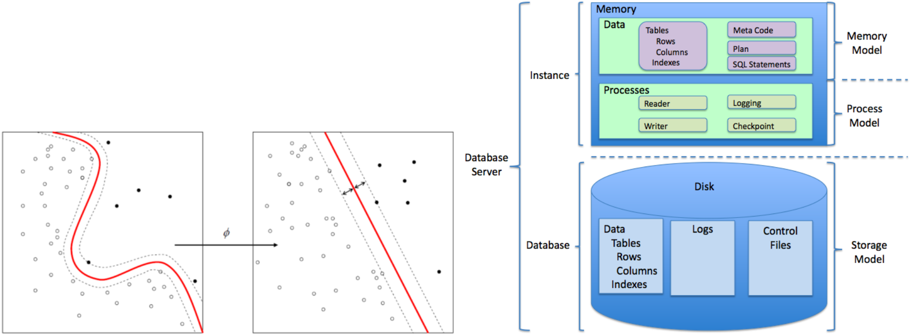


Figure: Kernel Machine by Alisneaky, RDBMS by Scifipete

What is Machine Learning?

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.

Machine Learning as per HBR

How Machines Learn (and you win) ▸ HBR

ML compared with DS

Machine Learning

1. Develop new models
2. Prove mathematical properties
3. Validate on relatively clean (possibly small) datasets
4. Publish paper

Data Science

1. Explore many models, focus on tuning
2. Understand empirical properties of models
3. Handle messy, massive datasets
4. Actionable systems

DBMS compared with DS

Database Systems

1. Individual records valuable
2. Modest data volumes
3. Structured, Consistent, Auditable
4. ACID compliance

Data Science

1. Individual rows "cheap"
2. Massive data volumes
3. Structured, Unstructured, and everything in between
4. Lots of ad-hoc querying/transformations

The Data Science Process

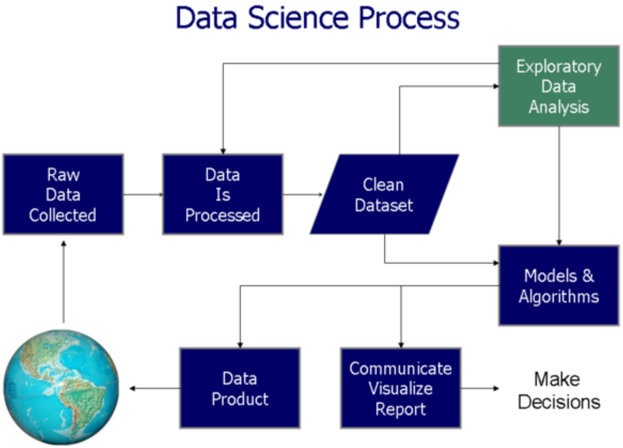


Figure: Data Science Process by Farcaster

Data provides valuable insights

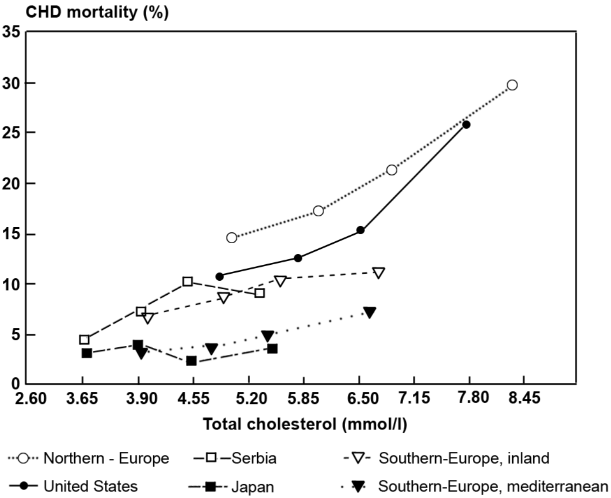


Figure: Seven Countries Study - Cholesterol vs Mortality ▸ Go

Good Data Visualization is Invaluable

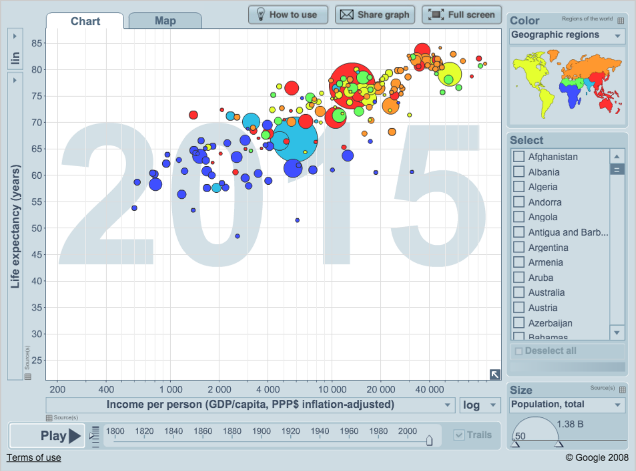


Figure: Gapminder - Wealth vs Health ▸ Go

Good Data Visualization is Invaluable

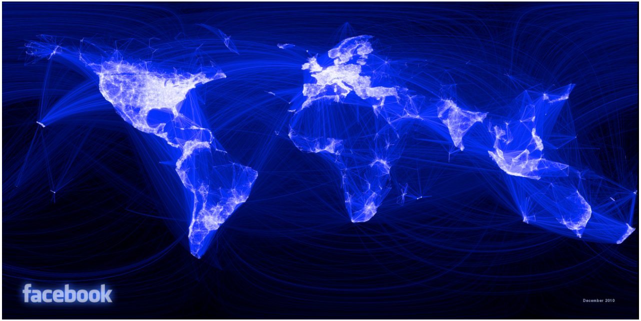


Figure: Facebook World Connections

Good Data Visualization is Invaluable

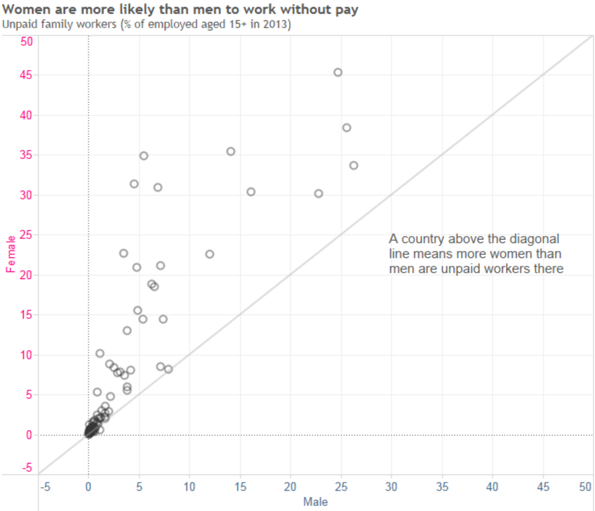
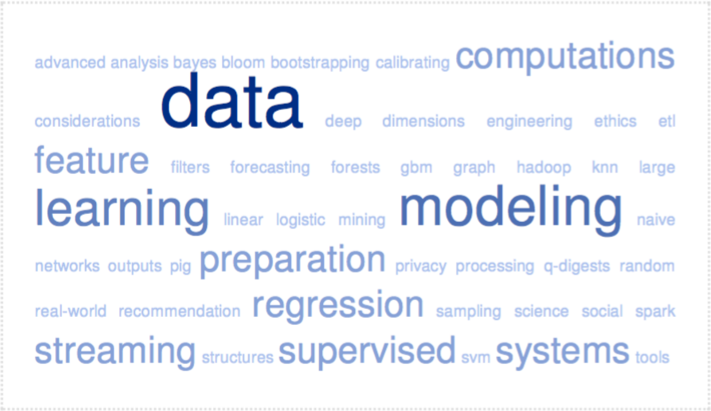


Figure: World Bank OpenData ▸ World Bank

What makes Data Science hard?

- Insufficient domain knowledge
- Incorrect assumptions
- Ad-hoc explanations of data patterns
- Overreach
- Validation/Data integrity
- Complex data and modeling pipelines
- Going from prototype to production
- Communicating the implications

Topics Covered


advanced analysis bayes bloom bootstrapping calibrating computations
considerations data deep dimensions engineering ethics etl
feature filters forecasting forests gbm graph hadoop knn large
learning linear logistic mining modeling naive
networks outputs pig preparation privacy processing q-digests random
real-world recommendation regression sampling science social spark
streaming structures supervised svm systems tools

Overview

1. Introduction
   - About your Instructor
   - What is Data Science?
   - What we will cover in this course?

2. Class Mechanics
   - Software tools

Class Mechanics

- Meet twice a week. Mondays, Wednesdays 4:45-6:00 PM
- 6 Programming assignments
- 1 Course Project

Software tools we'll be using

- PyTorch ▸ PyTorch

Weekly Reading

- Forrester Analyst Video ▸ Play
- Hilary Mason Video ▸ Play
- Hans Rosling TED Talk ▸ TED
- Short History of Data Science ▸ Blog
- O'Reilly Definition of Data Science ▸ OReilly

рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
mersi
kia ora
barka
welalin
spas
tack
mahalo
tapadh leat
kiitos
dankie
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
gracias
хвала
asante
manana
obrigada
murakoze
tenki
nanni
nandri
dhanyavad
bedankt
bayarlalaa
gracie
hvala
mauruuru
köszönöm
thank you
dankon
aciū
akun
enkosi
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
go raibh maith agat
sobodi
děkuji
obrigado
mèsi
didi madloba
sagolun
sukriya
kop khun krap
taiku
ありがとう
najis tuke
kam sah hamnida
rahmat
terima kasih
tanemirt
grazie
rahmet
arigatô
takk
dakujem
trugarez
ευχαριστώ
diolch
dhanyavadagalu
shukriya
merce
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci