Big Data Analytics using Spark

CSE255 / DSE230
What is “Big Data”?

• 1GB?
  • 1TB?
    • 1PB?
    • ....

• We need a definition that does not change over time.
  • More data than can fit on a single work-station.
  • Communication dominates computation.
“Data Science” vs. “Computer science”

• Computer science focuses on the **algorithm**
  • Requirements specify **input to output relationship** (find shortest path)
  • Algorithm should be **correct** and **efficient**
  • Input (data) can be anything that conforms to input format.

• Data Science focuses on the data.
  • The goal is to understand/ model / control the physical process generating the data.
  • Algorithms are used by the data scientist to identify patterns in the data.
  • Data is assumed to conform to a statistical model.
What is a data scientist?

From: Doing Data Science: Straight Talk from the Frontline
Rachel Schutt & Cathy O’Neil
There are many good jobs in data science

- **Data Scientist**: One of the ten top jobs in 2016 according to Forbes and glass-door.
- There are currently 8446 data science openings in the US (LinkedIn).
- 7000 openings in India (naukuri.com),
- Median base salary is around $116,000 per year (Glassdoor).
Alumnus Taner Halicioglu Kicks off Campaign for UC San Diego with $75 Million Gift

Facebook pioneer will establish the Halicioglu Institute for Data Science at UC San Diego

Halicioglu graduated with a bachelor’s degree in computer science in 1996
Nick Woodman, Founder of Go-Pro

Woodman graduated from UCSD in June 1997 with a B.A in visual arts and a minor in creative writing.
The output of a single goPro

- GoPro Hero Black 5: $400.
- 120 FPS 1080p 1920X1080
  - = 250Mpixel/sec each pixel 3*8 bits = 6Gbit / sec
- Max compressed output bitrate 60Mbit/sec
- Compression by a factor of 100.
- 2:14 minutes = 1GB compressed.
- Image processing requires uncompressed
Processing at the source

• Suppose you wanted to use GoPro to monitor your front door.
• The GoPro uses sophisticated lossy compression to reduce data by a factor of 100.
• However, to perform analysis, your PC would have to uncompress the data and then process >40GB per minute.
• You would need a beefy computer.
• But most of the time there is very little change from frame to frame, so if change detector is implemented on the camera, there is, most of the time, nothing to communicate.
Scaling up: Sensor networks & Smart cities

Libelium Smart World

- **Air Pollution**: Control of CO₂ emissions of factories, pollution emitted by cars and toxic gases generated in farms.

- **Forest Fire Detection**: Monitoring of combustion gases and preemptive fire conditions to define alert zones.

- **Wine Quality Enhancing**: Monitoring of soil moisture and their diameter in vineyards to control the amount of sugar in grapes and grapevine health.

- **Offspring Care**: Control of growing conditions of the offspring in animal farms to ensure its survival and health.

- **Sportsmen Care**: Vital signs monitoring in high performance centers and fields.

- **Structural Health**: Monitoring of vibrations and material conditions in buildings, bridges and historical monuments.

- **Quality of Shipment Conditions**: Monitoring of temperature, pressure, container openings or cold chain maintenance for insurance purposes.

- **Smartphones Detection**: Detect iPhone and Android devices and in general any device which works with Wi-Fi or Bluetooth interfaces.

- **Perimeter Access Control**: Access control to restricted areas and detection of people in non-sanctioned areas.

- **Radiation Levels**: Distributed measurement of radiation levels in nuclear power stations surrounding to generate leakage alerts.

- **Traffic Congestion**: Monitoring of vehicles and pedestrian influence to optimize driving and walking routes.

- **Electromagnetic Levels**: Measurement of the energy radiated by cell stations and Wi-Fi routers.

- **Waste Management**: Optimization of trash collection routes.

- **Vehicle Auto-diagnosis**: Information collection from CarBus to send real-time alarms to emergencies or provide advice to drivers.

- **Item Location**: Search of individual items in big surfaces like warehouses or harbors.

- **Water Quality**: Study of water suitability in rivers and the sea for fauna and eligibility for drinkable use.

- **Water Leakages**: Detect breaks in agricultural, outside tanks and pressure variations among pipes.

- **Smart Roads**: Warning messages and diversions according to climate conditions and unexpected events like accidents or traffic jams.

- **Smart Lighting**: Intelligent and weather adaptive lighting (street lights).

- **Intelligent Shopping**: Setting prices in the entire store according to customer habits, preferences, presence of alarming components for them or expiration dates.

- **Noise Urban Maps**: Sound monitoring in bar areas and public zones in real time.
<table>
<thead>
<tr>
<th>Project</th>
<th>Short Description</th>
<th>Status</th>
<th>Domain Expert</th>
<th>Domain Expert Department</th>
<th>Methods Expert</th>
<th>Methods Student Funding</th>
<th>Methods Student Openings</th>
<th>Last Date Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical NLP and Medical Note Analysis</td>
<td>Clinical Natural Language Processing (cNLP) to large corpora of medical notes.</td>
<td>Proposed</td>
<td>Chun-Nan Hsu</td>
<td>Biomedical Informatics</td>
<td>Julian McAuley</td>
<td>no</td>
<td>0.00</td>
<td>2017-03-27</td>
</tr>
<tr>
<td>Histology Browser</td>
<td>A web portal for viewing and annotating stacks of histology images.</td>
<td>Active</td>
<td>David Kleinfeld</td>
<td>Physics</td>
<td>Yoav Freund</td>
<td>yes</td>
<td>0.50</td>
<td>2017-03-16</td>
</tr>
<tr>
<td>Honey Bee Waggle Dance</td>
<td>Automate the analysis of videos capturing the honey bee waggle dances.</td>
<td>Active</td>
<td>James Nieh</td>
<td>Behavioral Ecology</td>
<td>Yoav Freund</td>
<td>no</td>
<td>0.00</td>
<td>2017-03-27</td>
</tr>
<tr>
<td>Whale classification from echo-location clicks</td>
<td>Develop a classifier of whale species using underwater recording of echo-location clicks</td>
<td>Proposed</td>
<td>Yoav Freund</td>
<td>Computer Science and Engineering</td>
<td>Yoav Freund</td>
<td>1.00</td>
<td>2017-03-28</td>
<td></td>
</tr>
</tbody>
</table>
CSE255 / DSE230

- A fun course
- Not an easy course.
- Weekly HW, from Friday to Friday expect to spend ~10 hours on each HW.
- You are expected to figure out things on your own.
  - Consult documentation of python, spark etc.
  - Brush up on your linear algebra, eigen-vectors, eigen-values, eigen-decomposition.
  - See linear algebra material on web site.
  - Wikipedia
- You are expected to participate in class and on Piazza.
What will you learn?

From: Doing Data Science: Straight Talk from the Frontline
Rachel Schutt & Cathy O’Neil

Python
Spark

Hacking Skills

Math & Statistics Knowledge
Linear Algebra
PCA
Regression
Classification

Machine Learning

Data Science

Traditional Research

Substantive & Communication skills
Expertise

Jupyter Notebooks
Visualization
Interpretation
Breakdown Problems

Danger Zone!
Jupyter Notebooks

• Pull them from the github repository.
• They are your main resource:
  • Class Slides are derived from the notebooks
  • Code
  • Explanations
  • Pointers to additional resources
  • Exercises
Grading

• HW: 50%
  • There will be 9 HW assignments, the one with the lowest grade will be dropped from the average.

• Quiz: 10%
  • Each Thursday. Lowest grade dropped from average.

• Breakdown Problems: 10%
  • Explained on class web page.

• Final: 30%
  • Yet do decide whether in-class or take home.
More details on the web site

• Go to
  • https://mas-dse.github.io/DSE230/