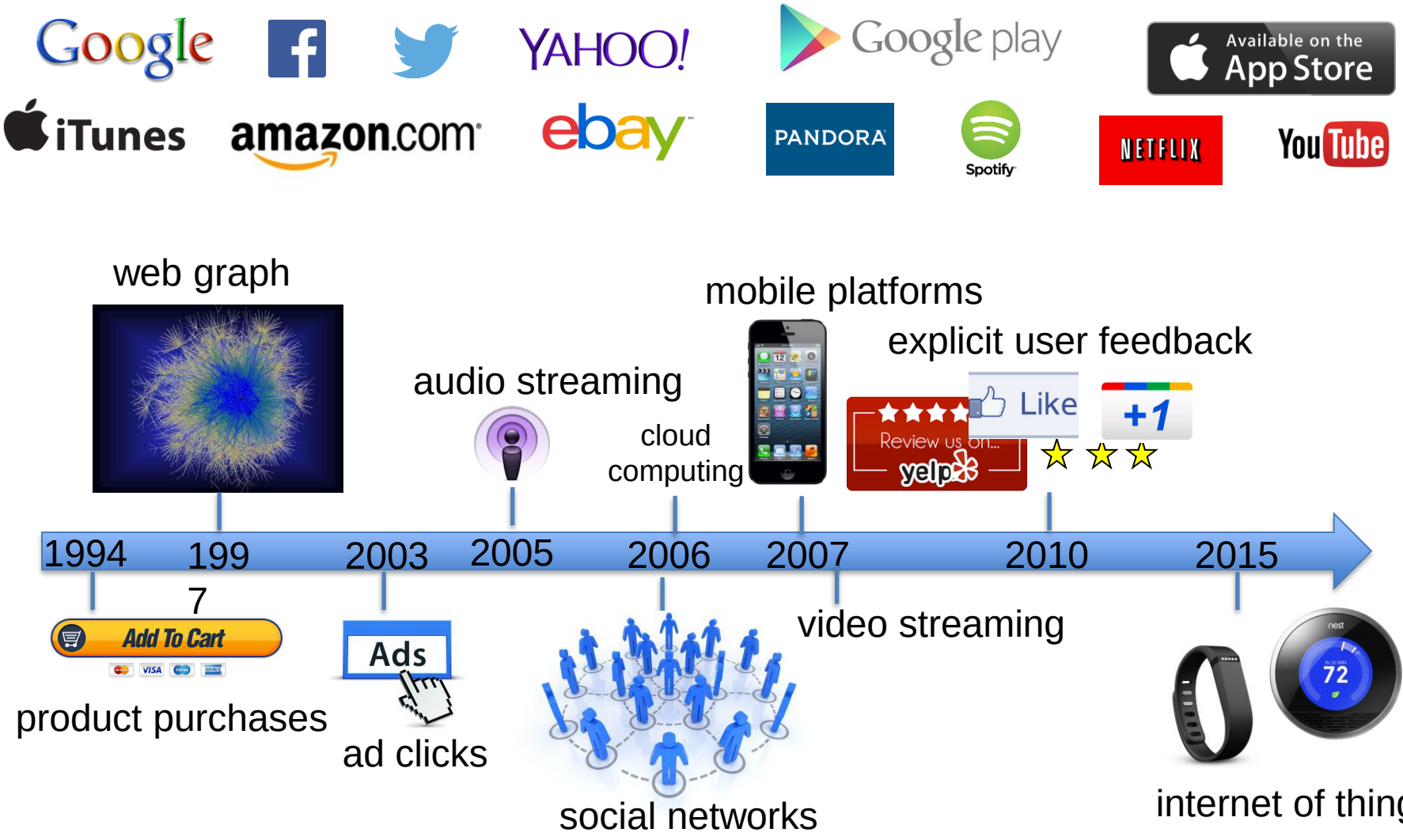
Northeastern

EECE5645
Parallel Processing for Data Analytics

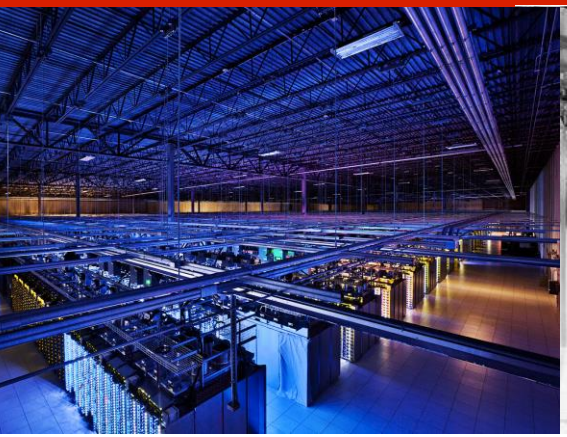
Lecture 0: Course Outline & Syllabus

Mining User Data



Northeastern

Dealing with Massive Computational Tasks

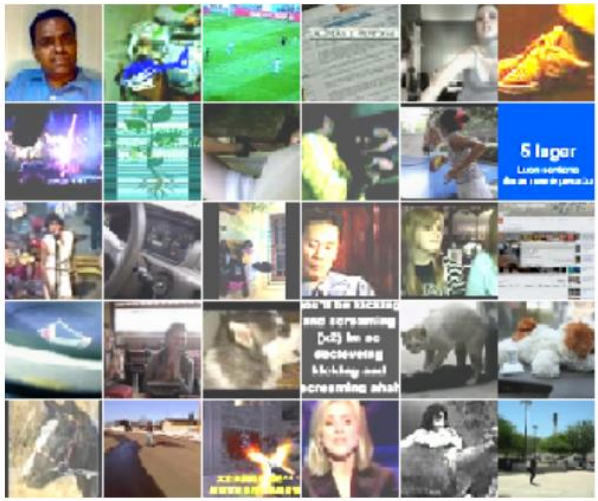


- Built using commodity, low-cost hardware
- Execute "embarrassingly parallel" operations
- Automated tools to access, process, and analyze collected data

Massive Data + Massive Computational Power…

…has led to breakthroughs in a broad variety of fields

- ❑ Image processing
- ❑ Speech recognition
- ❑ Natural language processing
- ❑ …



How Many Computers to Identify a Cat? 16,000

By JOHN MARKOFF JUNE 25, 2012 The New York Times

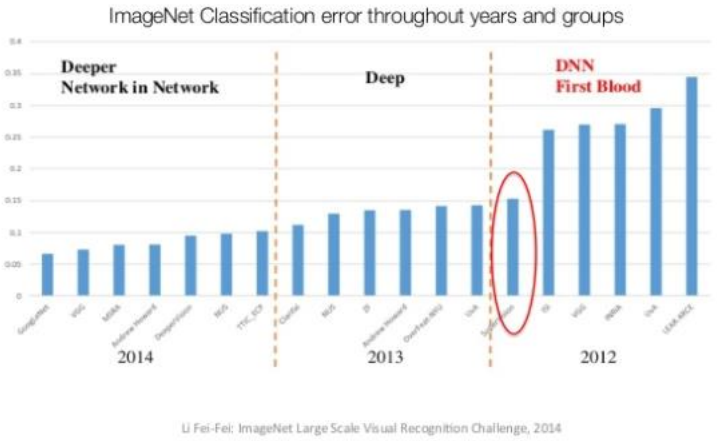
- ➢ Images sampled randomly from 10 million Youtube videos
- ➢ DNN trained over 1000 machines of 16 cores each

AlexNet

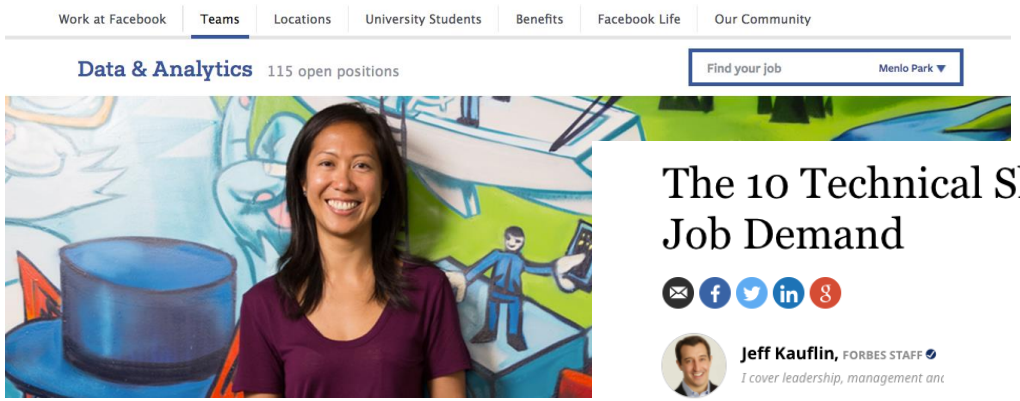
"ImageNet Classification with Deep Convolutional Neural Networks", Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton, NIPS 2012.

- Over 15M labeled high resolution images
- Roughly 22K categories
- Collected from web and labeled by Amazon Mechanical Turk

- Deeper, ReLU, dropout…
- Used **GPUs**



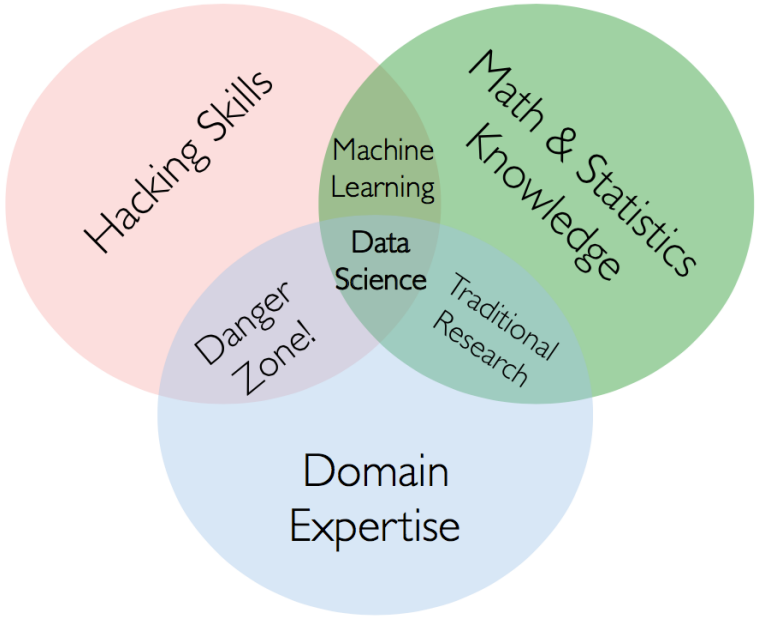
Data Analytics as a Profession



The 10 Technical Skills With Explosive Growth In Job Demand

Jeff Kauflin, FORBES STAFF

I cover leadership, management and



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

What does EECE5645 cover?

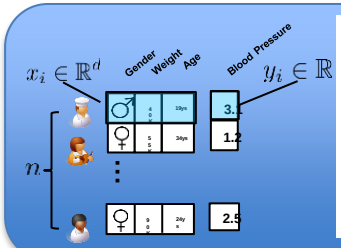
- Parallel Programming with Spark
- Convex Optimization fundamentals
  - How do I minimize a function?
- Regression and Classification
  - How do I predict whether a user will click on an ad?
  - How can I be confident that my prediction is correct?
- Advanced Topics
  - Matrix Factorization
  - DNNs
  - …

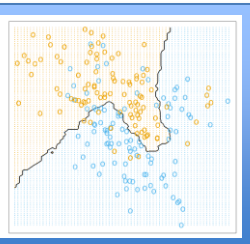
IN PARALLEL

See also the **Course Syllabus** on Canvas.

Overall Course Structure

Statistics &
Machine Learning

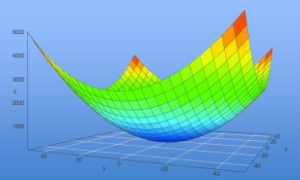


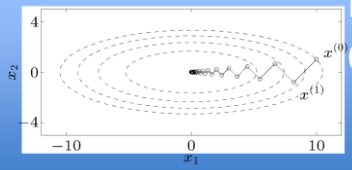


Regression, Classification
Regularization, Cross Validation
…

Convex Optimization

$$\arg\min_{x \in \mathbb{R}^d} f(x)$$





Descent Methods
Gradient Descent
Newton Method
…

Parallel Processing

APACHE Spark + python

```
map reduce reduceByKey join …
```


Northeastern

What is Spark?

- A general engine for **large-scale data processing**
- **Extensively** used in the industry
- **Seamlessly integrates** with Python, itself a powerful **general-purpose** programming language
- Very **easy to parallelize** your programs, and learn map reduce concepts
- Class will be **self-contained**: you can do with Spark+Python everything you learn here, and much more!

By the end of the class you will…

- ❑ …be confident in **programming in Python and Spark**
- ❑ …have **mastered map/reduce** concepts
- ❑ …know how to use a **computing cluster**
- ❑ …understand the basic **ML/statistics pipeline**, i.e., how to
  - ❑ **Train** a model
  - ❑ Identify **important features**
  - ❑ **Validate** your model's performance
- ❑ …understand **why** you are doing these steps
- ❑ …learn how do these steps **in parallel** over **many machines**, processing **massive datasets**!

This Course is Not…

- *…a Machine Learning class:*
  - There will be a math+statistics component, but we will focus on **parallelism** and **simple methods**

- *…a programming-only class:*
  - There will be math + stats: we should understand what we are doing, and why.

  See **Math Background** on Canvas.

- *…a "learn how to use toolboxes and libraries" class:*
  - You will program everything from scratch (i.e., learn how to build your own toolbox)

Reference Textbooks

- Karau, H., Konwinski, A., Wendell, P. and Zaharia, M., 2015. *Learning Spark.* Available online at NEU library.
- Friedman, J., Hastie, T., and Tibshirani, R.. *The Elements of Statistical Learning*. Springer. Available online: https://statweb.stanford.edu/~tibs/ElemStatLearn/
- Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press. Available online: http://stanford.edu/~boyd/cvxbook/

We will mostly rely on course slides + lecture notes.

Logistics

Grading Scheme

- ❑ Homework: 50%
- ❑ 2 Quizzes (**10/30, 11/24**) : 25%
- ❑ Course project: 25%

Consult the **Course Syllabus and/ or Course Calendar** on Canvas for dates. Dates are subject to change.

Course Summary:

Date	Details	
Thu Sep 10, 2020	First Class	2:50pm to 4:30pm
Fri Oct 2, 2020	Homework 1	due by 11:59pm
Sun Oct 4, 2020	Team Formation	due by 11:59pm
Mon Oct 5, 2020	Discovery Maintenance	8am to 12pm
Mon Oct 12, 2020	Columbus Day (USA), no classes	12am
Fri Oct 23, 2020	Homework 2	due by 11:59pm
Fri Oct 30, 2020	Quiz 1	due by 11:59pm
Sun Nov 1, 2020	Project Proposal	due by 11:59pm
Mon Nov 2, 2020	Discovery Maintenance	8am to 12pm
Wed Nov 11, 2020	Veterans' Day (USA), no classes	12am

Homework

- ❑ 4 programming assignments
- ❑ all will involve Spark
- ❑ Submit **code** and **typewritten** report
- ❑ Should be executed on the Discovery cluster

Instructor

- ❑ Stratis Ioannidis, ioannidis@ece.neu.edu
- ❑ TA: Gözde Özcan, gozcan@ece.neu.edu
- ❑ See also **Office Hours** on Canvas
- ❑ Use **Piazza** to ask questions!!

On Campus Instruction

- All sessions (Lectures+Office Hours) will be conducted online via Zoom.
- Classroom Technician (Nikhitha Sindhiya) will be onsite to set up lecture room
- If you choose to come to the lecture hall, practice **healthy distancing**, **and wear mask or face covering**. Use NUFlex to reserve seats.
- Bring laptop and connect to zoom silently, so that you can type questions in chat.
- **Do not be late**. If no-one is in classroom **10' after class starts**, Nikhitha will **terminate the session** and leave

Zoom Etiquette

- Connect using **NU credentials** (see Canvas for instructions). Do not share link with anyone.
- **All class lectures will be recorded and uploaded on Canvas**. Recordings include slides+voice but no video.
- **Mute** while not asking questions.
- **Videos**: feel free to turn on your video (or not) for lectures, but please turn on for office hours (not recorded).
- **Ask questions**! Unmute and interrupt me! Use chat and/or raise hand as well.
- On a **different timezone**? Please **email instructor** ASAP.

Class Project

- 25% of final grade
- In teams of **3-4 people**
- Need to:
  - Form teams by **10/4**
  - Submit 1-page proposal by **11/1** (not graded)
  - Submit report by **12/12 (last day of classes)**
  - Present results during exam week (**12/14,** exact time TBD)

Project Details

- Does not need to be in Spark (though it certainly can be)
- Can use existing libraries/tools (e.g. Mlib), though it does not have to
- Must involve parallelism
- Must involve analyzing an interesting, large dataset
- Must include a validation analysis
- Must demonstrate that parallelism helped
- Must clearly state what each participant did

Quizzes

- Both will be held at a morning time, off-class hours (**Friday 10/30**,**Tuesday 11/24,** just before TG break).
- Details/exact time TBD, but exam will be synchronous.
- 100 minutes needed but will give extra time.

Discovery Cluster

- Managed by NU Information Technology Services
- Housed at the Massachusetts Green High Performance Computing Center (MGHPCC) in Holyoke, MA
- >300 compute nodes, 20K cores
- 30 machines (20 CPUs each) reserved for class use

- Getting started: see **Programming Resources** on Canvas to find out more.

URGENT: Discovery Cluster Checklist (see Canvas)

1. **Get an account** ASAP!

1. Confirm you have access to **machines** reserved for class

1. Create a **.bashrc** file loading all modules you will need.

1. Familiarize yourself with **how to use the cluster**

Instructions for 1-4 can be found in **Programming Resources** on Canvas, we will do a demo of 3. and 4. in class

Academic Integrity

- Please read Canvas Syllabus and http://www.northeastern.edu/osccr/academic-integrity-policy/

- Do not use code taken from the internet or from other students to complete assignments.

Be Safe, Take Care of Yourselves, and Do Not Hesitate to Reach Out

- We are all not just working remotely, we are trying to work during a pandemic.

- Please reach out if you have any concerns whatsoever about the class, these are truly exceptional circumstances, and I will do my best to help.

Coming Up Next:

- Python Review
- Introduction to Spark
- Deep-dive into Spark Fundamentals
- Demos on how to connect to the Discovery cluster and run python and spark