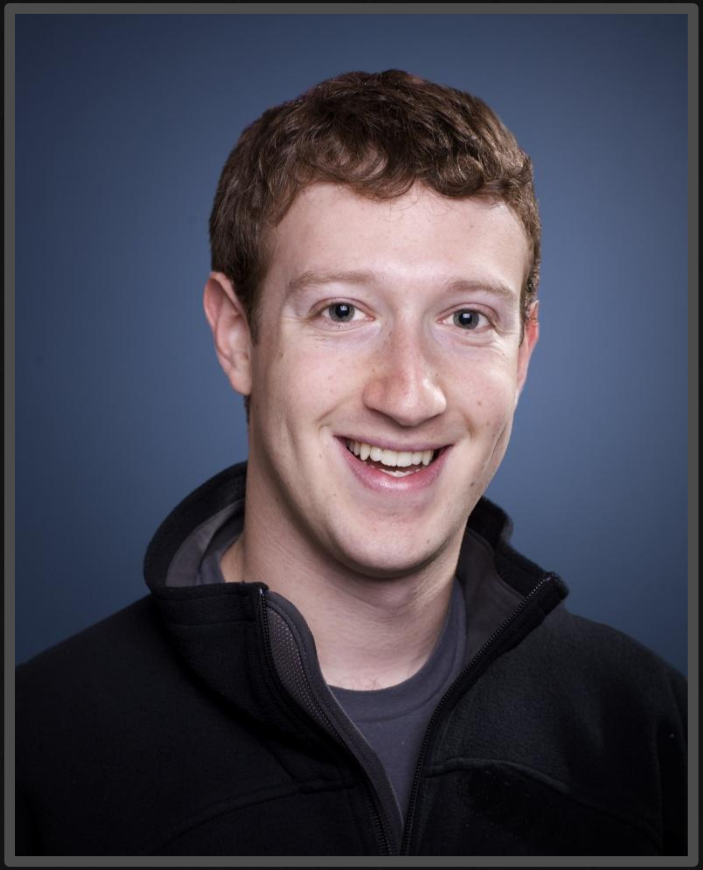
Course Motivation

Why Should You Take This Course?



Big Data Era

- We have more data now than ever before.
  - *Google: "Every 2 Days We Create As Much Info As We Did Up To 2003"*
  - *Facebook: "500+ TB of New Data Each Day"*
  - *Twitter: "500 million tweets per day."*

Source: TechCrunch, August 2010, http://techcrunch.com/2010/08/04/schmidt-data/

Source: Gigaom.com, August 2012, http://bit.ly/13wyPj7

Source: Twitter Blog, August 2013, http://bit.ly/14ySYrr

Facebook

- 2.5 billion items shared / day.
- 2.7 billion “Likes” / day.
- 300 million photos / day.
- 100+ PB Hadoop cluster.
- 70k database queries / day.

Source: Gigaom.com, August 2012, http://bit.ly/13wyPj7

Big Data for the Masses

- Large-scale data analysis is not new.
- It’s now just easier to generate a lot of data quickly:
  - *Web/Mobile Applications.*
  - *Sensors.*
  - *Public Data.*

Data-Driven Decision Making

Decision Making

- Use data to inform every aspect of our lives.
  - *Medicine.*
  - *Public Policy.*
  - *Science.*
  - *Business.*

The Three V's of Big Data

Source: Gartner, August 2001, http://gtnr.it/1etFI9e

Volume

Velocity

Variety

Crime Forecasting

- Used earthquake prediction algorithm to predict locations of future crimes.
- Deployed in Santa Cruz and Los Angeles.
    - *12% Decrease in Property Crime*
    - *26% Decrease in Burglaries*

Source: NY Times, August 2011, http://nyti.ms/13wDh1e

Internet Advertising

- Exchanges allow advertisers to bid on ad impressions in real-time.
- Each impression is supplemented by third-party historical data.

Source: NY Times, November 2012, http://nyti.ms/15bQltY

Language Translation

- Previous automatic translation techniques were difficult and inaccurate.
- Google scrapped the web to find matching translations and trained their models.

Source: A. Halevy, et al. March 2009, http://bit.ly/14yTDZT

Course Outline

- Background
- Data Ingestion
- Data Digestion
- Potpourri

Background (Sept)

- History of Databases
- Consensus Protocols
- Distributed Transactions

Ingestion (Sept — Oct)

- NoSQL
- NewSQL
- Distributed Data Stores

Digestion (Oct – Nov)

- Stream Processing Platforms
- Data Warehouses
- Machine Learning Systems

Potpourri (Nov – Dec)

- Non-Standard Techniques
- Hybrid Systems
- Crowdsourcing

For Next Class

- First paper reviews are due.
- Sign up for a lightning talk (date only).
- Sign up for two paper presentation dates.
- Links will be sent out on mailing list.