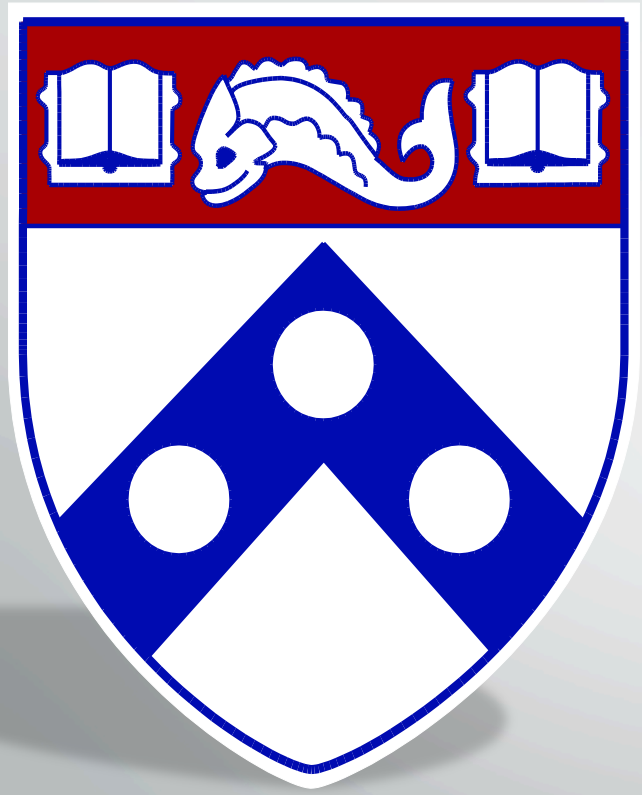


Advanced Topics in Databases: the "Big Data" Revolution

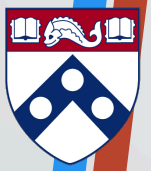
Susan B. Davidson

CIS 700: Advanced Topics in Databases

MW 1:30-3

Towne 309

http://www.cis.upenn.edu/~susan/cis700/homepage.html



The evolution of data models

- Hierarchical (IBM IMS) – 60's-70's
- Network, CODASYL (Backman, IDS) – 60's
- Relational – 70's
- Object-relational (Stonebraker, et al) – 90's
- OODBMS (Atkinson, et al) – 90's
- Array databases (MonetDB, SciDB, et al) – 90's
- XML (document-oriented) – 2000's
- **NoSQL – 2010's**

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

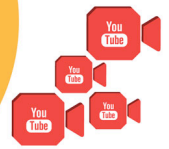
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month



30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

- almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

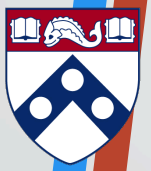
Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

in one survey were unsure of how much of their data was inaccurate

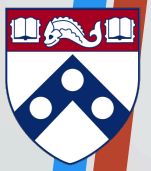
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM



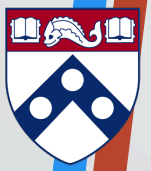
"Big Data" is two problems

- The **storage** problem
  - How to store and manipulate huge amounts of data to facilitate fast queries and analysis
- The **analysis** problem
  - How to extract useful info, using modeling, ML and stats.
- Problems with traditional (relational) storage
  - Not flexible
  - Hard to partition, i.e. place different segments on different machines



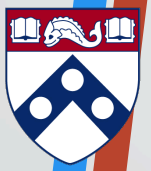
Dimensions of the revolution

- "Big data" has driven the revolution of database technology in several dimensions, including
  - more flexible models
  - streaming and time-varying data
  - different notions of updates and consistency
  - need for parallelism
- Due to the tight interaction with complex analysis and inference pipelines, it has also increased the need for more *accountability* and the careful consideration of *ethical issues* surrounding the use of the data.



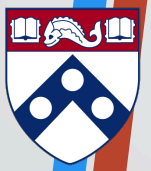
Course format

- Lectures on introductory material
  - Foundations of relational databases: relational algebra, relational calculus, Datalog
  - NoSQL "foundations": JSON-based solutions, graph-based solutions
  - Time varying/ streaming databases
  - Provenance
  - Transactions/consistency
- Research papers on related topics (to be posted)
  - Students are expected to present 2-3 papers during the semester, and write a summary of papers presented by others.



Intended audience

- Students who have taken a basic course in databases, e.g. CIS550
- Students who are interested in research topics in databases



Grading

- Class participation and attendance: 20%
- Paper presentation: 30%
- Paper reviews: 20%
- Project: 30%