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 of 2011, the global size of data in healthcare was estimated to be 150 exabytes (1,000 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.

1 in 3 business leaders don't trust the information they use to make decisions.

27% of respondents in one survey were unsure of how much of their data was inaccurate.

By 2015, 4.4 million IT jobs will be created globally to support big data, with 1.9 million in the United States.

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

**Volume**

- **Scale of Data**
  - 40 zettabytes (43 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005.
  - 6 billion people have cell phones.
  - World population: 7 billion.

**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.

**Variety**

- Different forms of data
  - As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (1,000 billion gigabytes).

**Veracity**

- Uncertainty of data
  - 27% of respondents in one survey were unsure of how much of their data was inaccurate.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, WEPTEC, QAS
“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%