College of Information Studies

University of Maryland Hornbake Library Building College Park, MD 20742-4345

Data as an Asset

Session 10

INST 301

Introduction to Information Science

Data as a Model

- Data represents some aspect(s) of reality
  - Reality itself is way too complex

- All models are wrong
  - Some models are useful

- Things that are useful are useful for a purpose
  - Which need not be the original intended purpose

Some Examples

- Bank account
- Airline ticket
- Email

Find the Data

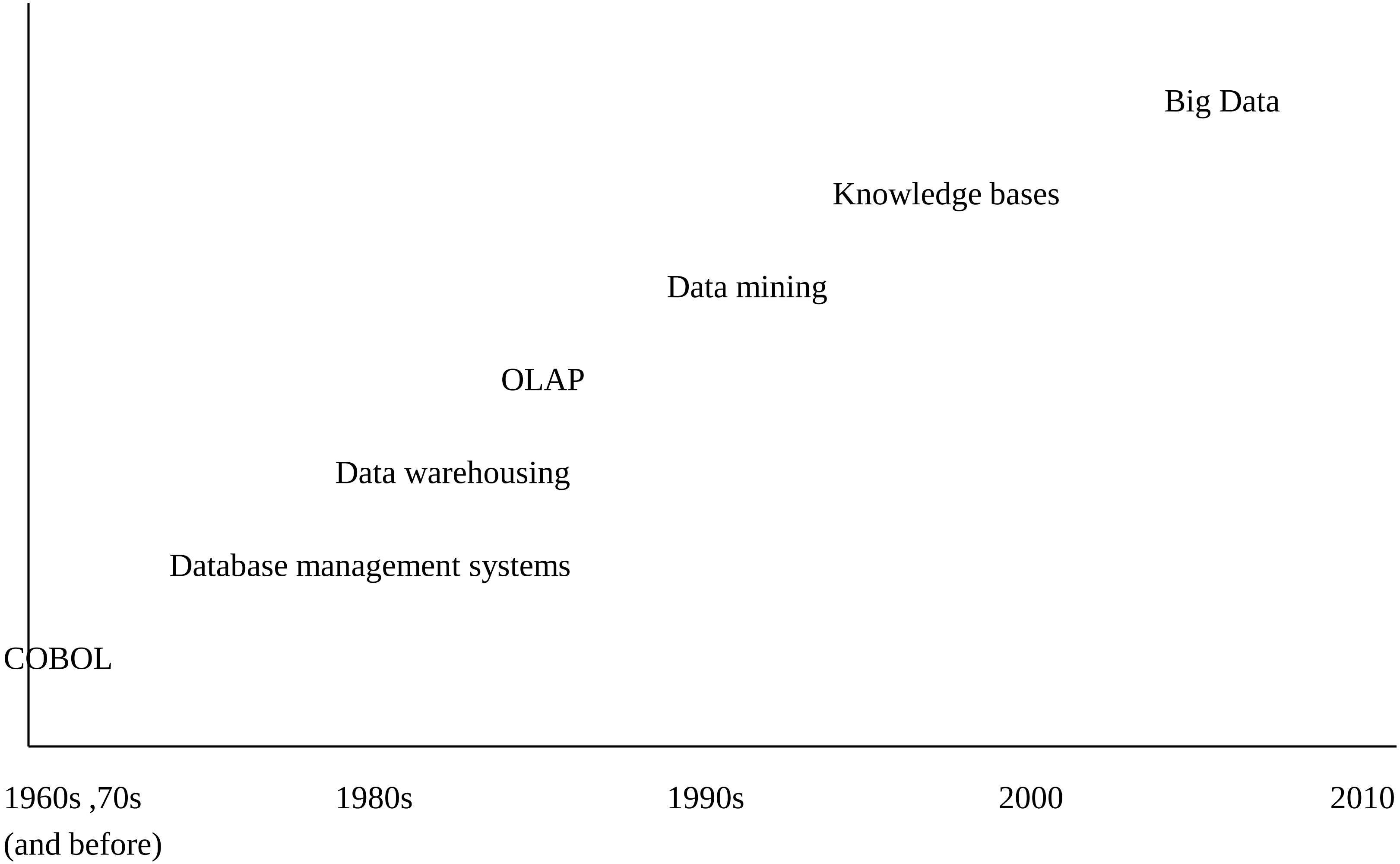
***Date*: Wed Dec 20 08:57:00 EST 2000**
***From*: Kay Mann <kay.mann@enron.com>**
***To*: Suzanne Adams <suzanne.adams@enron.com>**
***Subject*: Re: GE Conference Call has be rescheduled**

Did Sheila want Scott to participate? Looks like the call will be too late for him.

Some Aspects of Reality

- People
- Places
- Organizations
- Events
- Objects
- Concepts
- Amounts
- Times
- Time periods
- Statements
- Attitudes

Historical Development



Big Data

Knowledge bases

Data mining

OLAP

Data warehousing

Database management systems

COBOL

1960s ,70s
(and before)

1980s

1990s

2000

2010

Data abstraction

- moving from the nuts and bolts to the big picture
- hiding the complexity of data storage and operations from the user
- levels of abstraction - from the trees to the forest
  - the physical level
  - the conceptual level
  - the view level

Database Management System

- Special-purpose programming language for:
  - Defining a database
    - Data
    - Relationships
  - Populating the database
    - Initialization, update, deletion, …
  - Using the database
    - Queries
    - Reports

Database Lifecycle

- Identify a need
- Analyze specific goals
- Create a model
- Implement the database
- Initialize the data
- Use it for the intended purpose(s)
  - Support the business process(es)
  - Interoperate with other data systems
- Optionally, repurpose the data
- Migrate or retire the data

Data Warehouse

- Collection of technologies designed to convert heaps of data to usable information
  - an environment, not a process
- Three applications
  - improve traditional information presentation technologies
  - support online analytical processing
  - enable use of data mining techniques

Data Warehouse

- Aggregates data from different systems
  - Can require reconciliation that no system needed
- Updated mostly by addition
  - Permits current and historical analysis
- Read-intensive offline analysis
  - Different access pattern than operational systems
- Small number of expert users

Online Analytical Processing (OLAP)

- Exploration tool
  - Slice and dice to “preprocess” the data
  - Visualization to explore relationships

- Example:
  - What percentage of employees in the southeast region who had been covered under health plan A have switched to health plan B since January, broken down by employee family demographics and by office, and how does that compare with our projections?

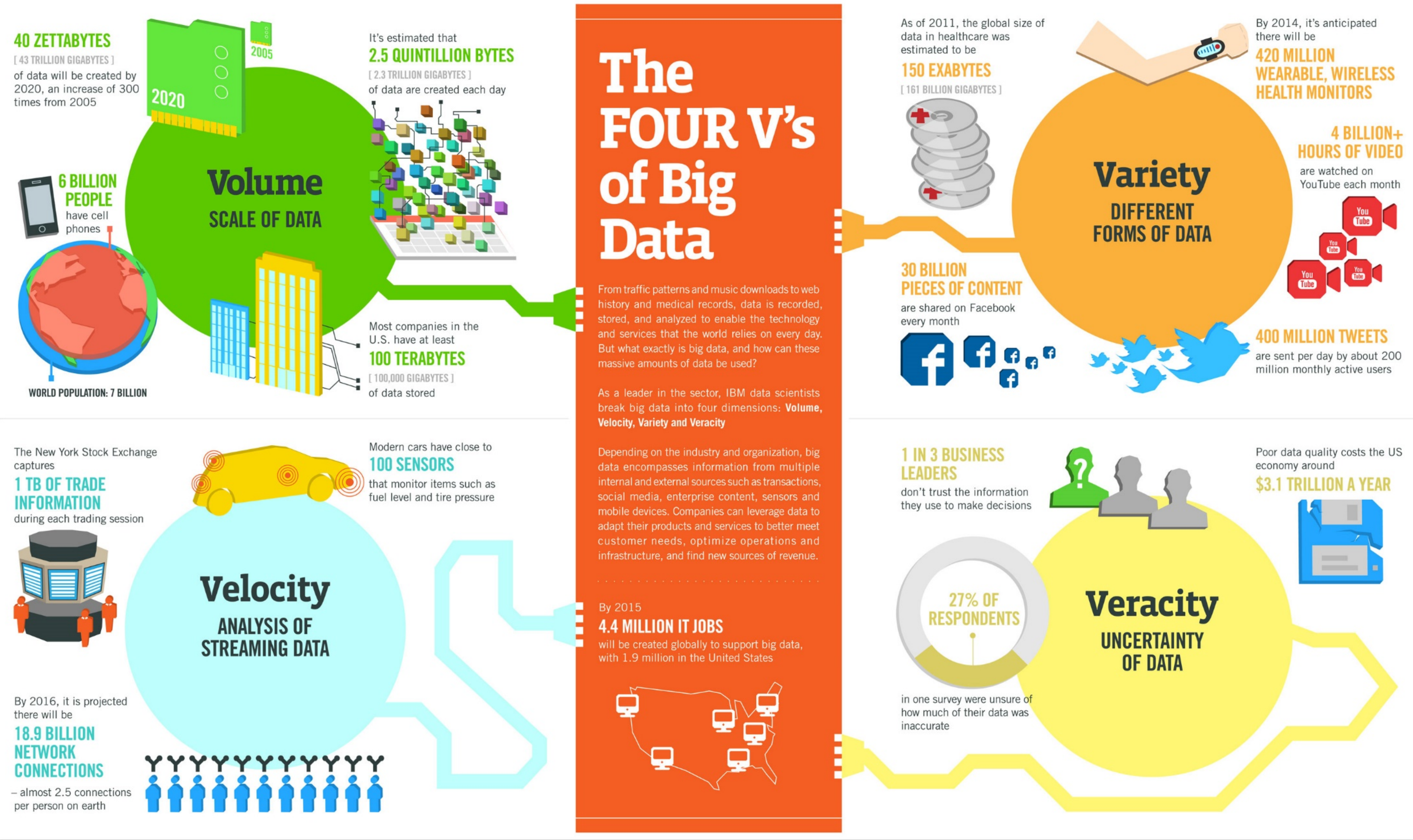
Data Mining

- Statistical analysis to uncover patterns

- Rule-based:
  - We know what pattern we want

- Supervised:
  - We have examples of patterns like what we want

- Unsupervised:
  - We know what kind of a pattern we want



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

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

[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

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

27% OF RESPONDENTS

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

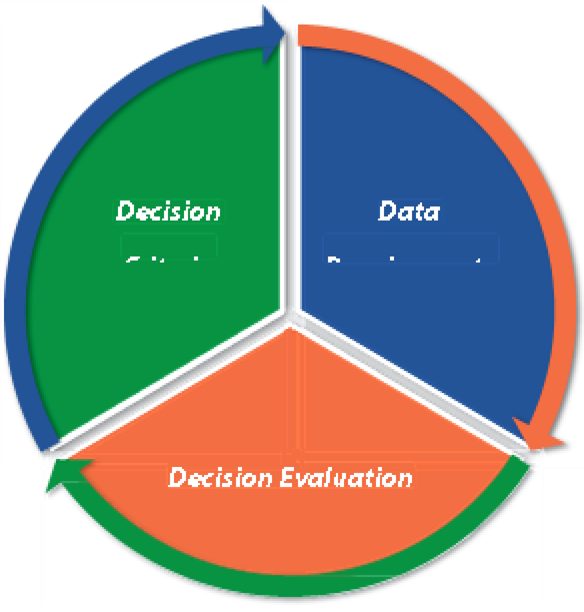
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

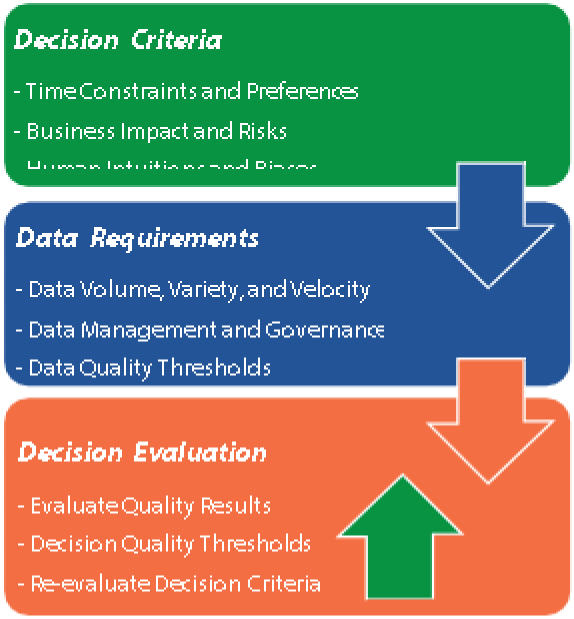
IBM

Data Quality

- Valid
- Accurate
- Precise
- Consistent
- Complete
- Current
- …

Data-Driven Decisions





Decision Criteria

- Time Constraints and Preferences
- Business Impact and Risks
- Human Intuitions and Biases

Data Requirements

- Data Volume, Variety, and Velocity
- Data Management and Governance
- Data Quality Thresholds

Decision Evaluation

- Evaluate Quality Results
- Decision Quality Thresholds
- Re-evaluate Decision Criteria