Introduction to Big Data

University of California, Berkeley

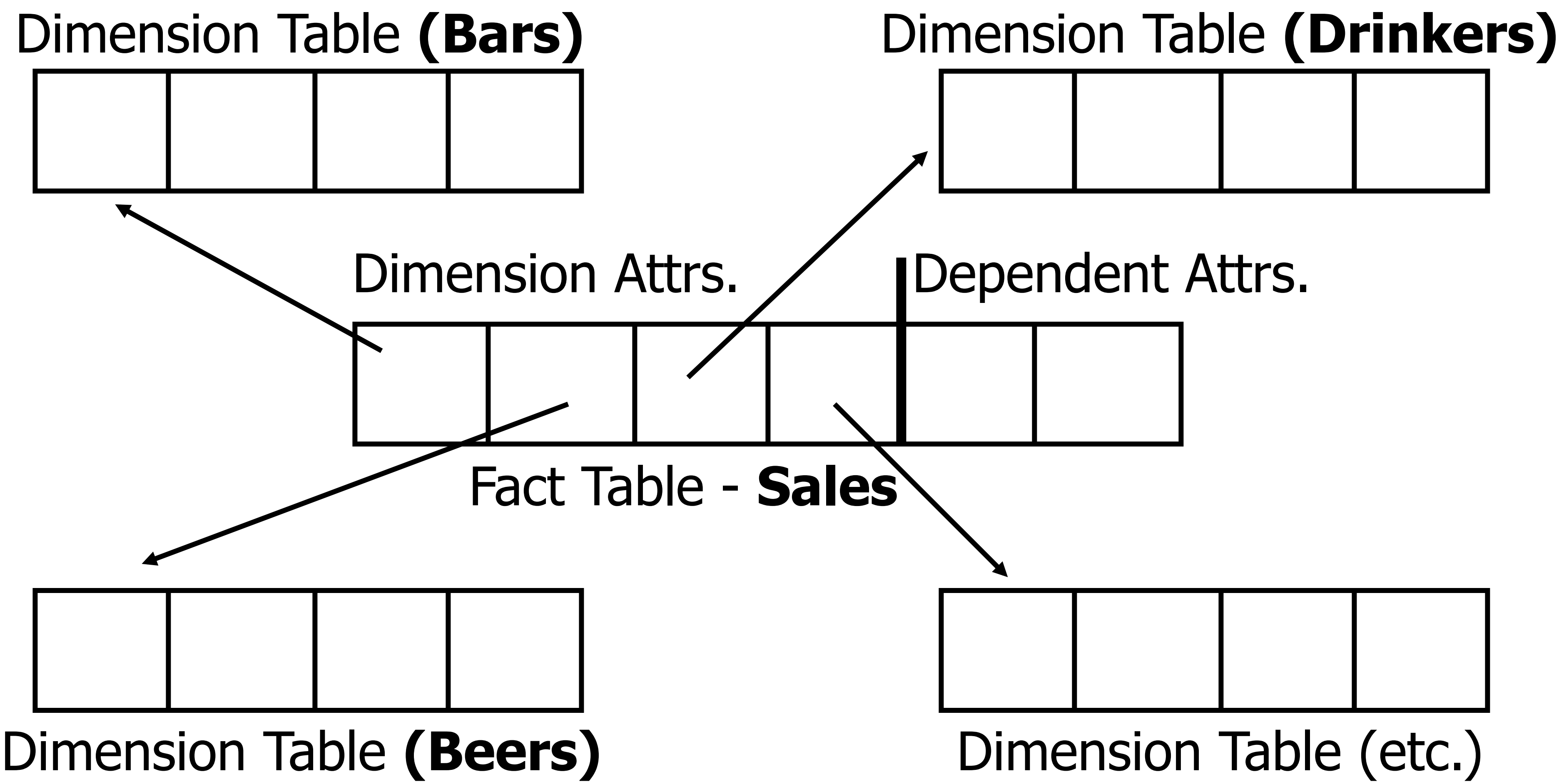
School of Information

IS 257: Database Management

Lecture Outline

- Review
  - OLAP with SQL
- Big Data (introduction) - Continued

Visualization – Star Schema



From anonymous "olap.ppt" found on Google

Typical OLAP Queries

- Often, OLAP queries begin with a "star join": the natural join of the fact table with all or most of the dimension tables.
- Example:

```
SELECT *
FROM Sales, Bars, Beers, Drinkers
WHERE Sales.bar = Bars.bar AND
  Sales.beer = Beers.beer AND
  Sales.drinker = Drinkers.drinker;
```

From anonymous "olap.ppt" found on Google

Example: OLAP Query

- For each bar in Palo Alto, find the total sale of each beer manufactured by Anheuser-Busch.
- Filter: addr = “Palo Alto” and manf = “Anheuser-Busch”.
- Grouping: by bar and beer.
- Aggregation: Sum of price.

Example: In SQL

```
SELECT bar, beer, SUM(price)
FROM Sales NATURAL JOIN Bars
  NATURAL JOIN Beers
WHERE addr = ’Palo Alto’ AND
  manf = ’Anheuser-Busch’
GROUP BY bar, beer;
```

From anonymous “olap.ppt” found on Google

Using Materialized Views

- A direct execution of this query from Sales and the dimension tables could take too long.
- If we create a materialized view that contains enough information, we may be able to answer our query much faster.

Example: Materialized View

- Which views could help with our query?
- Key issues:
  1. It must join Sales, Bars, and Beers, at least.
  2. It must group by at least bar and beer.
  3. It must not select out Palo-Alto bars or Anheuser-Busch beers.
  4. It must not project out addr or manf.

From anonymous "olap.ppt" found on Google

Example --- Continued

- Here is a materialized view that could help:

```
CREATE VIEW BABMS(bar, addr,
     beer, manf, sales) AS
SELECT bar, addr, beer, manf,
     SUM(price) sales
FROM Sales NATURAL JOIN Bars
     NATURAL JOIN Beers
GROUP BY bar, addr, beer, manf;
```

Since bar -> addr and beer -> manf, there is no real grouping. We need addr and manf in the SELECT.

From anonymous "olap.ppt" found on Google

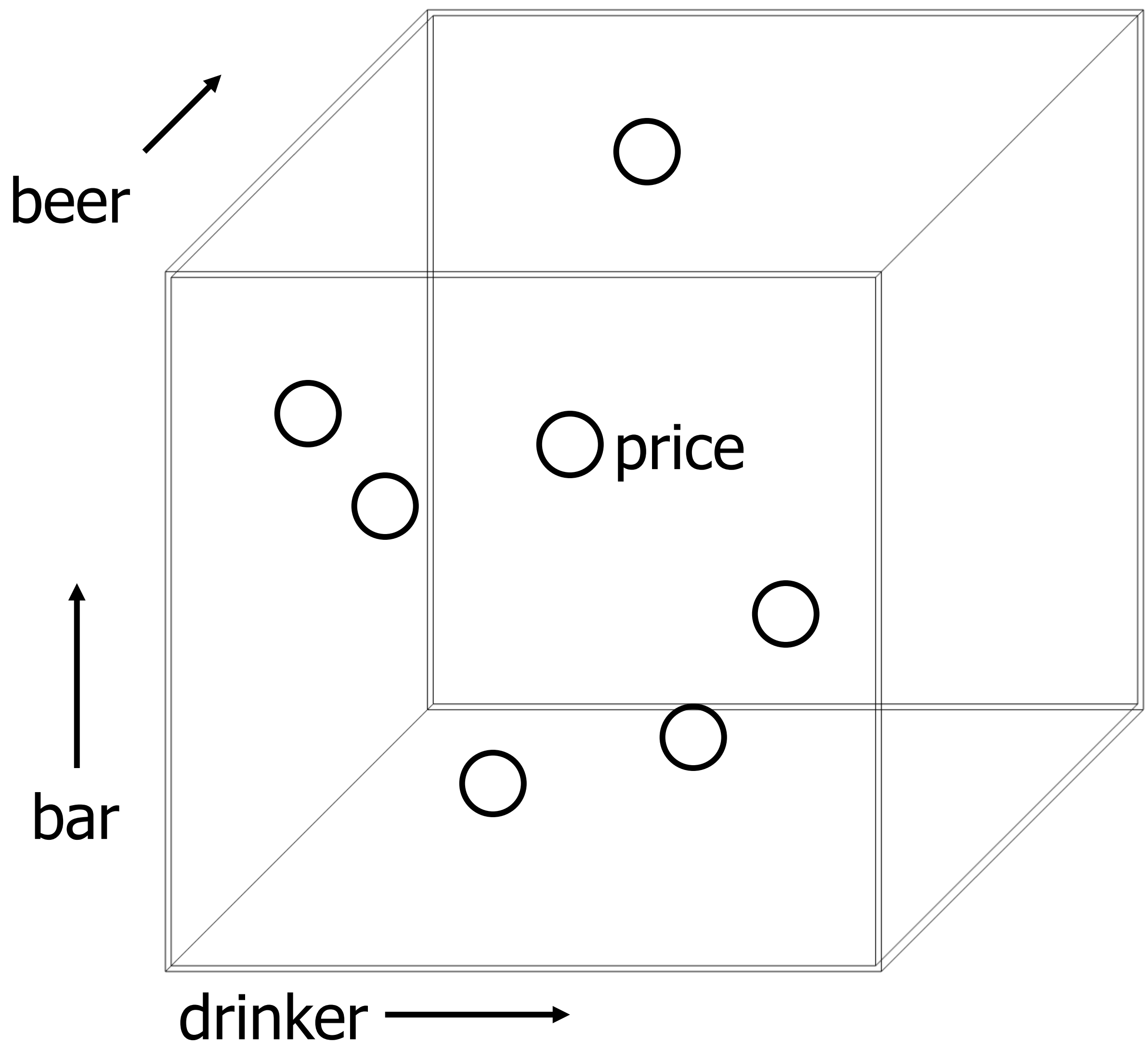
Example --- Concluded

- Here's our query using the materialized view BABMS:

```
SELECT bar, beer, sales
FROM BABMS
WHERE addr = 'Palo Alto' AND
    manf = 'Anheuser-Busch';
```

From anonymous "olap.ppt" found on Google

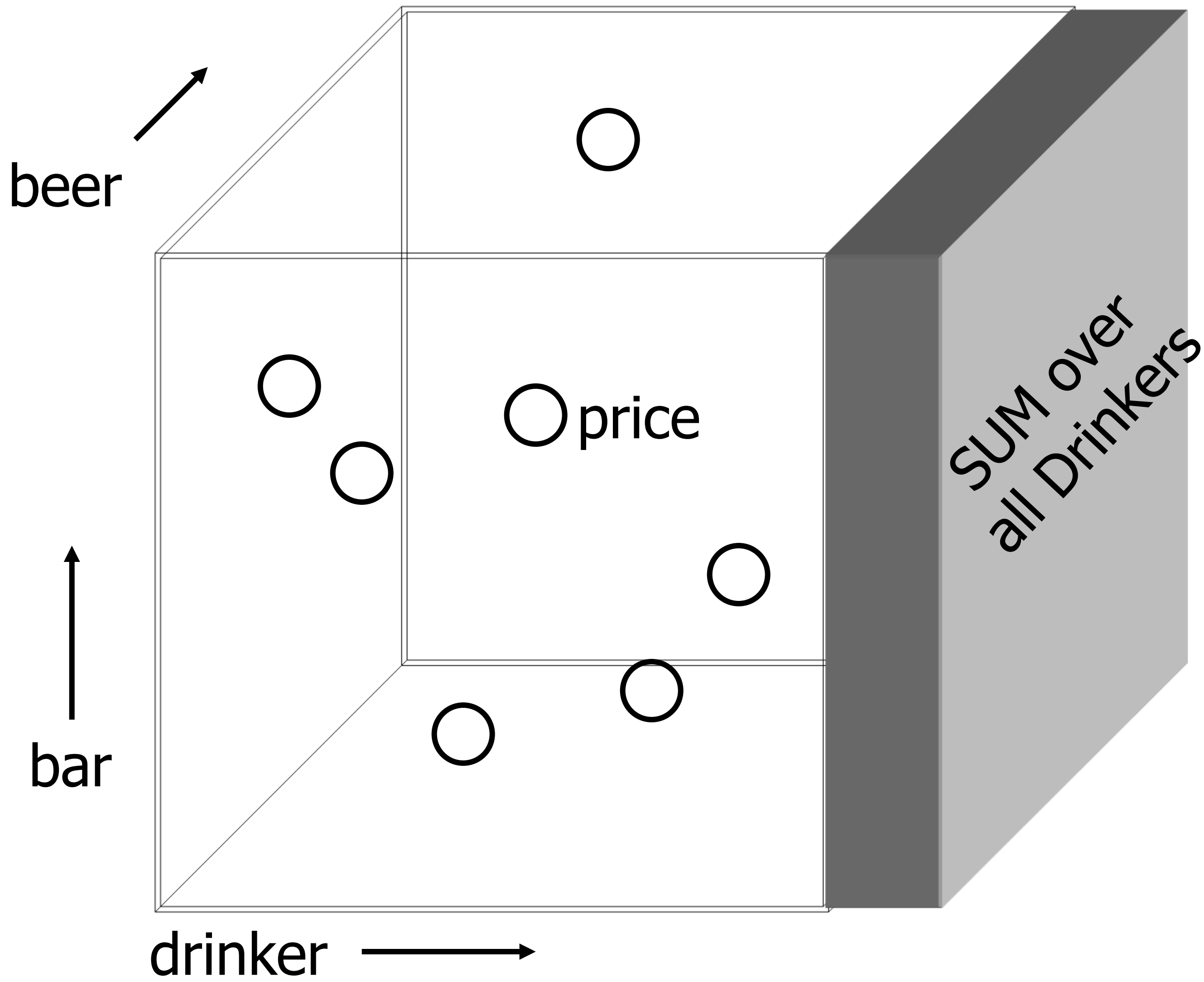
Visualization - Data Cubes



Marginals

- The data cube also includes aggregation (typically SUM) along the margins of the cube.
- The *marginals* include aggregations over one dimension, two dimensions,…

Visualization - Data Cube w/ Aggregation



Structure of the Cube

- Think of each dimension as having an additional value *.
- A point with one or more *'s in its coordinates aggregates over the dimensions with the *'s.
- Example: Sales("Joe's Bar", "Bud", *, *) holds the sum over all drinkers and all time of the Bud consumed at Joe's.

Roll Up and Drill Down

$ of Anheuser-Busch by drinker/bar

	Jim	Bob	Mary
Joe's Bar	45	33	30
Nut-House	50	36	42
Blue Chalk	38	31	40

Roll up by Bar

$ of A-B / drinker

Jim	Bob	Mary
133	100	112

Drill down by Beer

$ of A-B Beers / drinker

	Jim	Bob	Mary
Bud	40	29	40
M'lob	45	31	37
Bud Light	48	40	35

Materialized Data-Cube Views

- Data cubes invite materialized views that are aggregations in one or more dimensions.
- Dimensions may not be completely aggregated --- an option is to group by an attribute of the dimension table.

Data Mining

- *Data mining* is a popular term for queries that summarize big data sets in useful ways.
- Examples:
  1. Clustering all Web pages by topic.
  2. Finding characteristics of fraudulent credit-card use.

Market-Basket Data

- An important form of mining from relational data involves *market baskets* = sets of “items” that are purchased together as a customer leaves a store.
- Summary of basket data is *frequent itemsets* = sets of items that often appear together in baskets.

Finding Frequent Pairs

- The simplest case is when we only want to find "frequent pairs" of items.
- Assume data is in a relation Baskets(basket, item).
- The *support threshold* *s* is the minimum number of baskets in which a pair appears before we are interested.

From anonymous "olap.ppt" found on Google

Frequent Pairs in SQL

```
SELECT b1.item, b2.item
FROM Baskets b1, Baskets b2
WHERE b1.basket = b2.basket
  AND b1.item < b2.item
GROUP BY b1.item, b2.item
HAVING COUNT(*) >= s;
```

Look for two Basket tuples with the same basket and different items. First item must precede second, so we don't count the same pair twice.

Create a group for each pair of items that appears in at least one basket.

Throw away pairs of items that do not appear at least *s* times.

From anonymous "olap.ppt" found on Google

Lecture Outline

- Review
  - OLAP with SQL
- Big Data (introduction) - Continued

Big Data and Databases

- “640K ought to be enough for anybody.”
  - – Attributed to Bill Gates, 1981

Big Data and Databases

- We have already mentioned some Big Data
  - The Walmart Data Warehouse
  - Information collected by Amazon on users and sales and used to make recommendations
- Most modern web-based companies capture EVERYTHING that their customers do
  - Does that go into a Warehouse or someplace else?

Other Examples

- NASA EOSDIS
  - Estimated 10^{18} Bytes (Exabyte)
- Computer-Aided design
- The Human Genome
- Department Store tracking
  - Mining non-transactional data (e.g. Scientific data, text data?)
- Insurance Company
  - Multimedia DBMS support

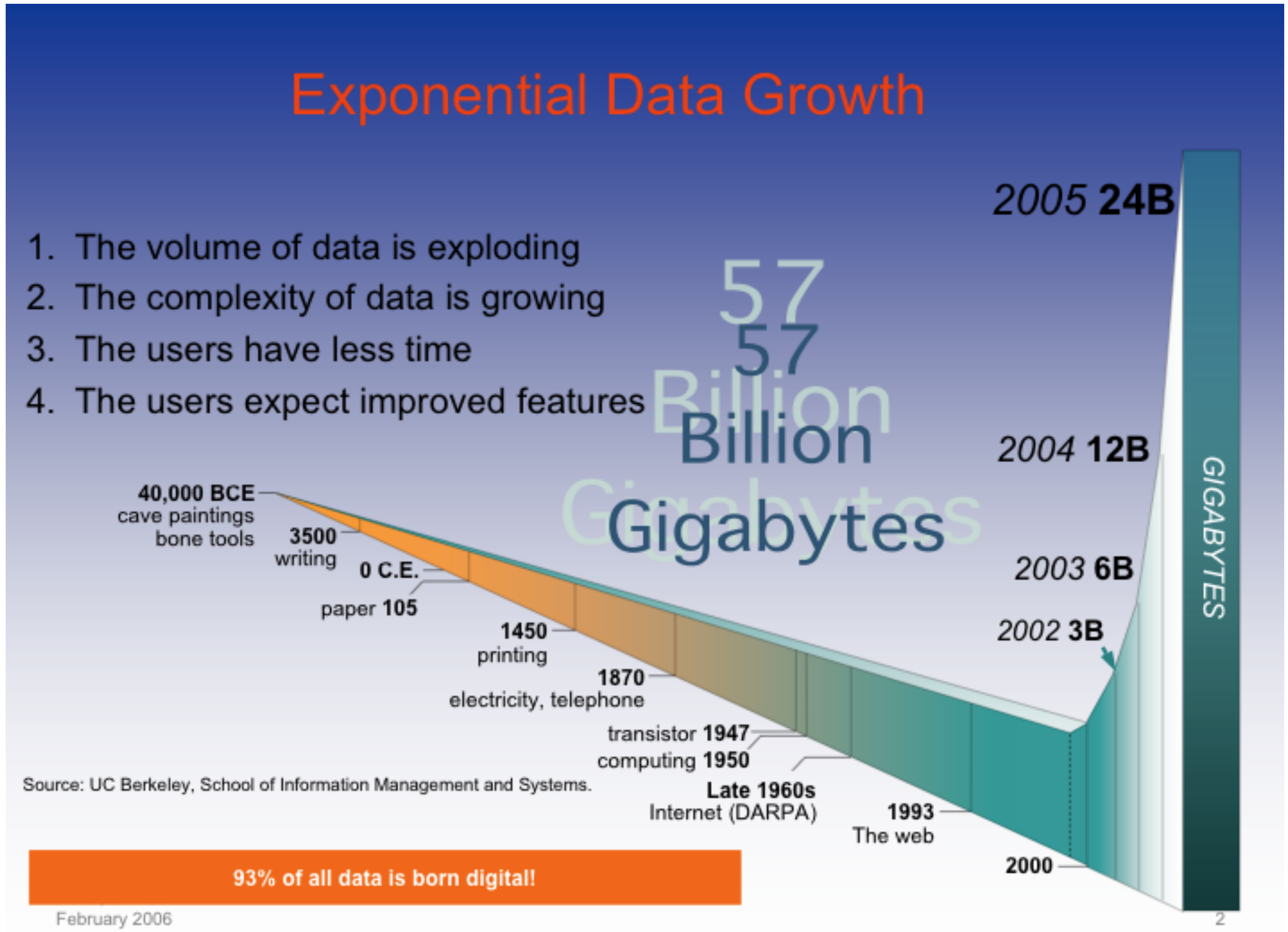
Table 1.1: How Big is an Exabyte?

Kilobyte (KB)	*1,000 bytes OR 10^3 bytes* 2 Kilobytes: A Typewritten page. 100 Kilobytes: A low-resolution photograph.
Megabyte (MB)	*1,000,000 bytes OR 10^6 bytes* 1 Megabyte: A small novel OR a 3.5 inch floppy disk. 2 Megabytes: A high-resolution photograph. 5 Megabytes: The complete works of Shakespeare. 10 Megabytes: A minute of high-fidelity sound. 100 Megabytes: 1 meter of shelved books. 500 Megabytes: A CD-ROM.
Gigabyte (GB)	*1,000,000,000 bytes OR 10^9 bytes* 1 Gigabyte: a pickup truck filled with books. 20 Gigabytes: A good collection of the works of Beethoven. 100 Gigabytes: A library floor of academic journals.
Terabyte (TB)	*1,000,000,000,000 bytes OR 10^{12} bytes* 1 Terabyte: 50000 trees made into paper and printed. 2 Terabytes: An academic research library. 10 Terabytes: The print collections of the U.S. Library of Congress. 400 Terabytes: National Climactic Data Center (NOAA) database.
Petabyte (PB)	*1,000,000,000,000,000 bytes OR 10^{15} bytes* 1 Petabyte: 3 years of EOS data (2001). 2 Petabytes: All U.S. academic research libraries. 20 Petabytes: Production of hard-disk drives in 1995. 200 Petabytes: All printed material.
Exabyte (EB)	*1,000,000,000,000,000,000 bytes OR 10^{18} bytes* 2 Exabytes: Total volume of information generated in 1999. 5 Exabytes: All words ever spoken by human beings.

Source: Many of these examples were taken from Roy Williams "Data Powers of Ten" web page at Caltech.

Digitization of Everything: the Zettabytes are coming

- **Soon most everything will be recorded and indexed**
- **Much will remain local**
- **Most bytes will never be seen by humans.**
- **Search, data summarization, trend detection, information and knowledge extraction and discovery are key technologies**
- **So will be infrastructure to manage this.**

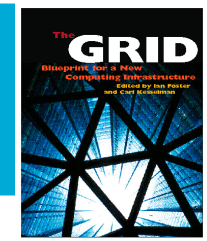


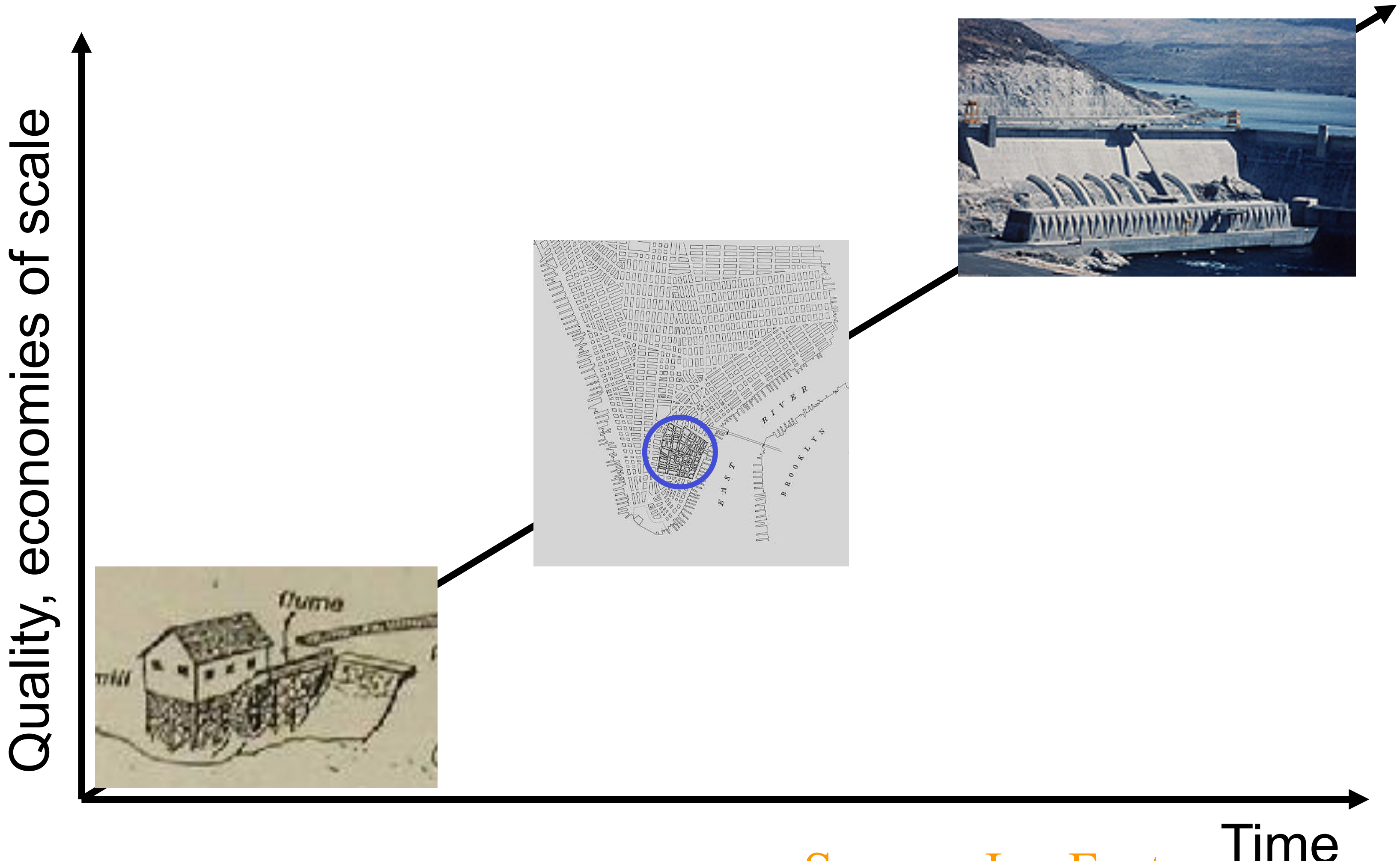
Before the Cloud there was the Grid

- So what's this Grid thing anyhow?
- Data Grids and Distributed Storage
- Grid vs "Cloud"

The following borrows heavily from presentations by Ian Foster (Argonne National Laboratory & University of Chicago), Reagan Moore and others from San Diego Supercomputer Center

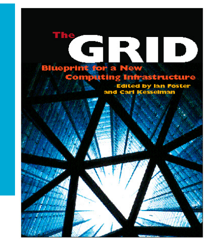
The Grid: On-Demand Access to Electricity





Source: Ian Foster

By Analogy, A Computing Grid



- Decouples production and consumption
  – Enable on-demand access
  – Achieve economies of scale
  – Enhance consumer flexibility
  – Enable new devices
- On a variety of scales
  – Department
  – Campus
  – Enterprise
  – Internet

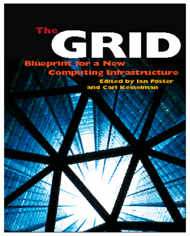
Source: Ian Foster

What is the Grid?

“The short answer is that, whereas the Web is a service for sharing information over the Internet, the Grid is a service for sharing computer power and data storage capacity over the Internet. The Grid goes well beyond simple communication between computers, and aims ultimately to turn the global network of computers into one vast computational resource.”

Source: The Global Grid Forum

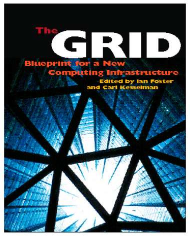
Not Exactly a New Idea …



- "The time-sharing computer system can unite a group of investigators …. one can conceive of such a facility as an … intellectual public utility."
  - Fernando Corbato and Robert Fano , 1966
- "We will perhaps see the spread of 'computer utilities', which, like present electric and telephone utilities, will service individual homes and offices across the country." Len Kleinrock, 1967

Source: Ian Foster

But, Things are Different Now



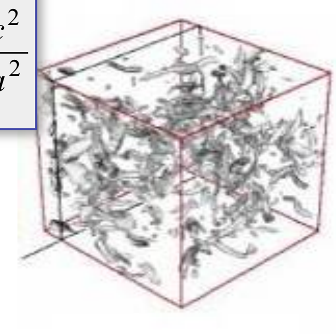
- Networks are far faster (and cheaper)
  - Faster than computer backplanes
- "Computing" is very different than pre-Net
  - Our "computers" have already disintegrated
  - E-commerce increases size of demand peaks
  - Entirely new applications & social structures
- We've learned a few things about software
- But, the needs are changing too…

Source: Ian Foster

Progress of Science

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

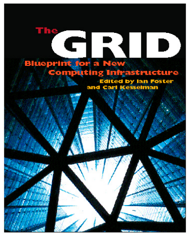
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



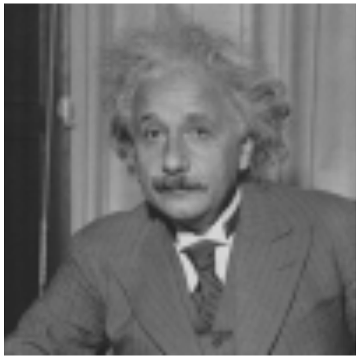
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **(big data/information)**
  data and information exploration (eScience)
  unify theory, experiment, and simulation - information driven
  - Data captured by sensors, instruments
    or generated by simulator
  - Processed/searched by software
  - Information/Knowledge stored in computer
  - Scientist analyzes database / files
    using data management and statistics
  - Network Science
  - Cyberinfrastructure

Source: Jim Gray

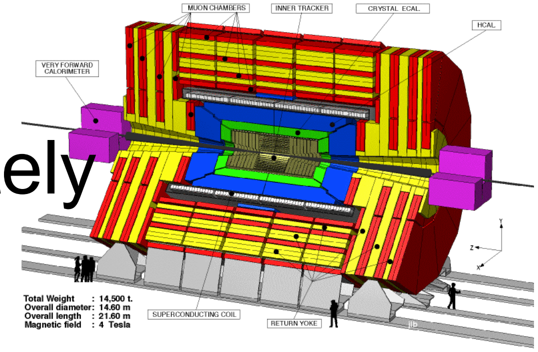
Why the Grid?
(1) Revolution in Science



- Pre-Internet
  - Theorize &/or experiment, alone or in small teams; publish paper



- Post-Internet
  - Construct and mine large databases of observational or simulation data
  - Develop simulations & analyses
  - Access specialized devices remotely
  - Exchange information within distributed multidisciplinary teams

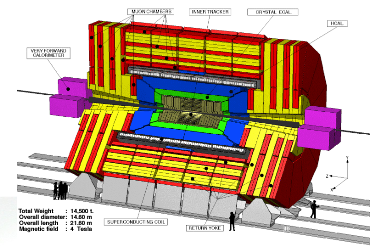


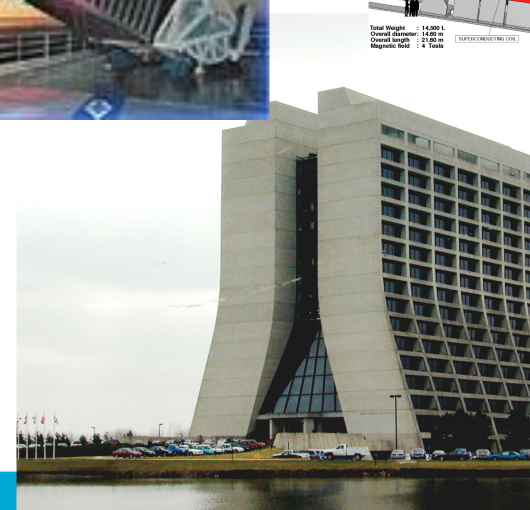
Source: Ian Foster

Computational Science

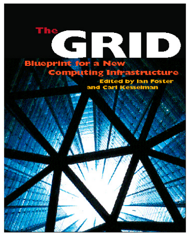
- **Traditional Empirical Science**
  – Scientist gathers data by direct observation
  – Scientist analyzes data

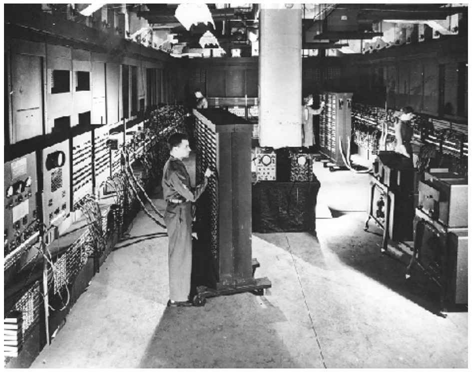
- **Computational Science**
  – Data captured by instruments
  Or data generated by simulator
  – Processed by software
  – Placed in a database
  – Scientist analyzes database
  – **`tcl scripts`**
    - **`or C programs`**
      – **`on ASCII files`**





Why the Grid?
(2) Revolution in Business





- Pre-Internet
  - Central data processing facility
- Post-Internet
  - Enterprise computing is highly distributed, heterogeneous, inter-enterprise (B2B)
  - Business processes increasingly computing- & data-rich
  - Outsourcing becomes feasible => service providers of various sorts

Source: Ian Foster

The Information Grid

Imagine a web of data

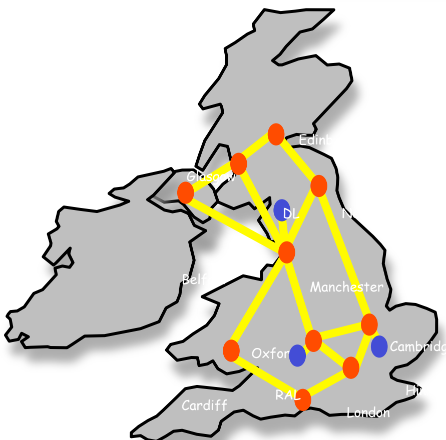
- Machine Readable
  – Search, Aggregate, Transform, Report On, Mine Data – using more computers, and less humans
- Scalable
  – Machines are cheap – can buy 50 machines with 100Gb or memory and 100 TB disk for under $100K, and dropping
  – Network is now *faster* than disk
- Flexible
  – Move data around without breaking the apps

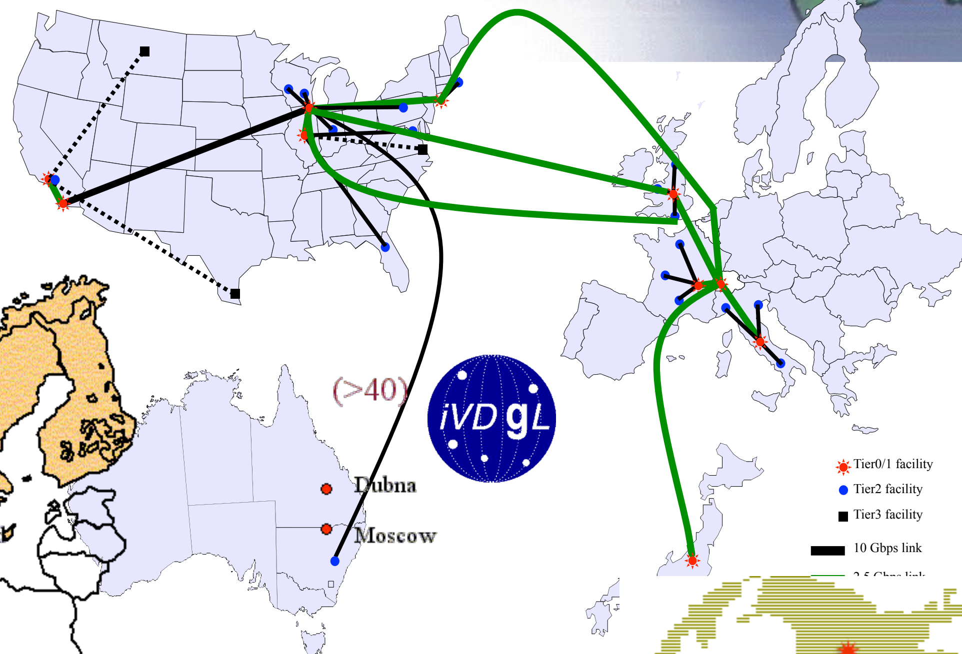
Source: S. Banerjee, O. Alonso, M. Drake - ORACLE

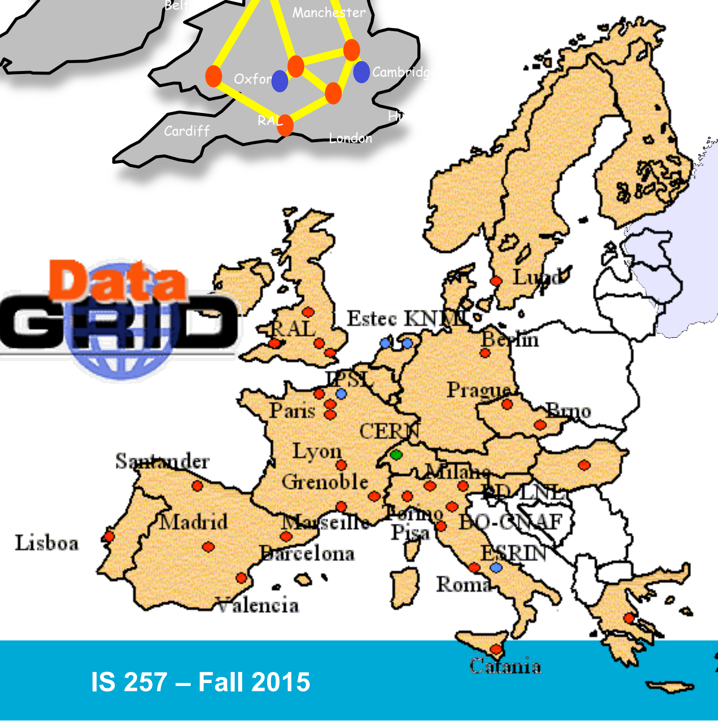
The Foundations are Being Laid

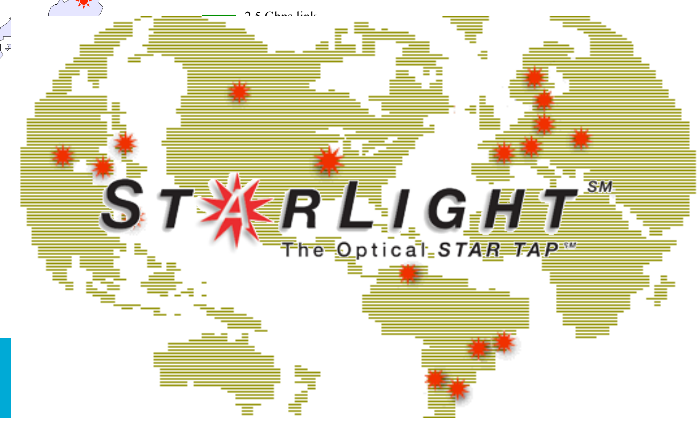
Building the National Virtual Collaboratory for Earthquake Engineering Research

NEESgrid









Current Environment

- "Big Data" is becoming ubiquitous in many fields
  - enterprise applications
  - Web tasks
  - E-Science
  - Digital entertainment
  - Natural Language Processing (esp. for Humanities applications)
  - Social Network analysis
  - Etc.
- Berkeley Institute for Data Science (BIDS)

Current Environment

- Data Analysis as a profit center
  - – No longer just a cost – **may be the entire business** as in Business Intelligence

Current Environment

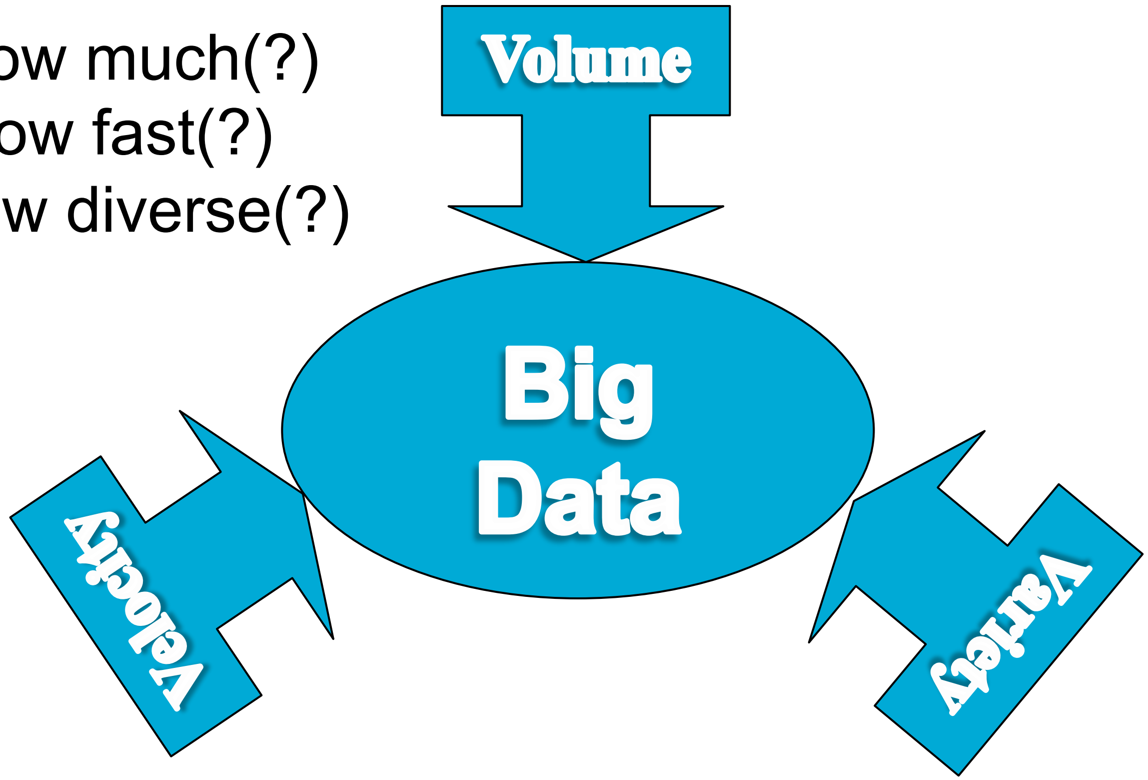
- Ubiquity of Structured and Unstructured data
  – Text
  – XML
  – Web Data
  – Crawling the Deep Web
- How to extract useful information from "noisy" text and structured corpora?

Current Environment

- Expanded developer demands
  - Wider use means broader requirements, and less interest from developers in the details of traditional DBMS interactions
- Architectural Shifts in Computing
  - The move to parallel architectures both internally (on individual chips)
  - And externally – Cloud Computing

The 3V's of Big Data

Volume – how much(?)
Velocity – how fast(?)
Variety – how diverse(?)



High Velocity Data

- Examples:
  - Harvesting hot topics from the Twitter "firehose"
  - Collecting "clickstream" data from websites
  - System logs and Web logs
  - High frequency stock trading (HFT)
  - Real-time credit card fraud detection
  - Text-in voting for TV competitions
  - Sensor data
  - Adwords auctions for ad pricing
    - http://www.youtube.com/watch?v=a8qQXLby4PY

High Velocity Requirements

- Ingest at very high speeds and rates
  - E.g. Millions of read/write operations per second
- Scale easily to meet growth and demand peaks
- Support integrated fault tolerance
- Support a wide range of real-time (or "near-time") analytics
- Integrate easily with high volume analytic datastores (Data Warehouses)

Put Differently

High velocity and you

You need to ingest a firehose in real time

You need to process, validate, enrich and respond in real-time (i.e. update)

You often need real-time analytics (i.e. query)

High Volume Data

- "Big Data" in the sense of large volume is becoming ubiquitous in many fields
  - enterprise applications
  - Web tasks
  - E-Science
  - Digital entertainment
  - Natural Language Processing (esp. for Humanities applications – e.g. Hathi Trust)
  - Social Network analysis
  - Etc.

High Volume Data Examples

- The Walmart Data Warehouse
  - Often cited as one of, if not the largest data warehouse
- The Google Web database
  - Current web
- The Internet Archive
  - Historic web
- Flickr and YouTube
- Social Networks (E.g.: Facebook)
- NASA EOSDIS
  - Estimated 10^{16} Bytes (Exabyte)
- Other E-Science databases
  - E.g. Large Hadron Collider, Sloan Digital Sky Survey, Large Synoptic Survey Telescope (2016)

How Big is Big Data

- How big is big?

1 Kilobyte	1,000 bits/byte
1 megabyte	1,000,000
1 gigabyte	1,000,000,000
1 terabyte	1,000,000,000,000
1 petabyte	1,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000

What is Big Data?

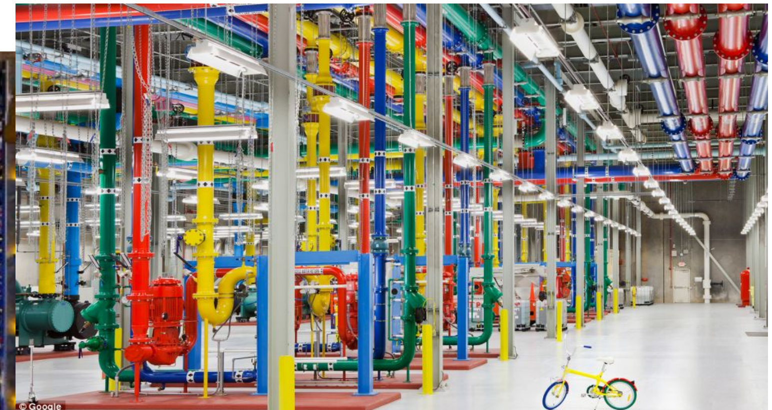
- Ran across some interesting slides from a decade ago that already frame the problem and did a fair job of predicting where we are today
  - Slides by Jim Gray and Tony Hey : “In Search of Petabyte Databases” ca. 2001

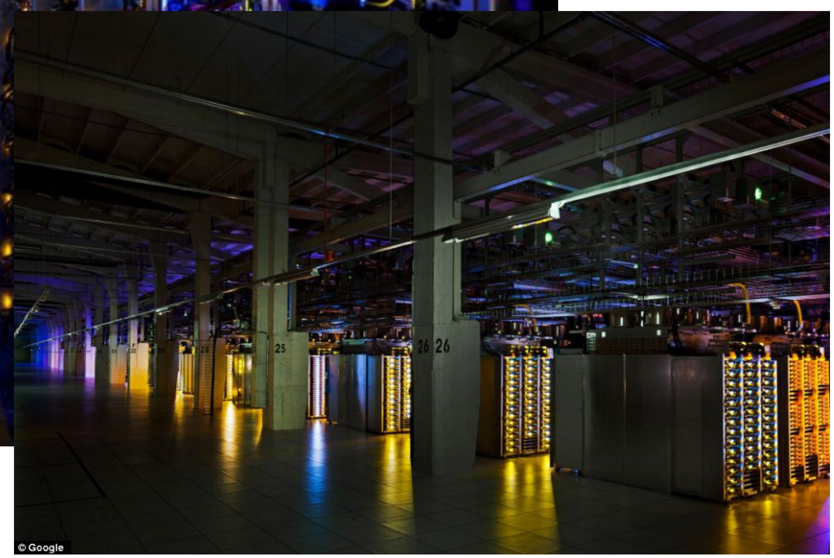
Summary from Gray & Hey

- DBs own the sweet-spot:
  - 1GB to 100TB
- Big data is ***not*** in databases
- HPTS (high performance transaction systems) crowd is not really high performance storage (BIG DATA)
- Cost of storage is people:
  - Performance goal:
    1 Admin per PB

From Jim Gray and Tony Hey : "In Search of Petabyte Databases" ca. 2001

Why People?





One row of one of Google's data centers

Also – the plumbing need for cooling, and the many rows of the data center

Difficulties with High Volume Data

- Browsibility
- Very long running analyses
- Steering Long processes
- Federated/Distributed Databases
- IR and item search capabilities
- Updating and normalizing data
- Changing requirements and structure

High Variety

- Big data can come from a variety of sources, for example:
  - Equipment sensors: Medical, manufacturing, transportation, and other machine sensor transmissions
  - Machine generated: Call detail records, web logs, smart meter readings, Global Positioning System (GPS) transmissions, and trading systems records
  - Social media: Data streams from social media sites like Facebook and miniblog sites like Twitter

High Variety

- The problem of high variety comes when these different sources must be combined and integrated to provide the information of interest
- Problems of:
  - Different structures
  - Different identifiers
  - Different scales for variables
- Often need to combine unstructured or semi-structured text (XML/JSON) with structured data

Various data sources

What Does Machine Data Look Like?

Sources

Order Processing

ORDER,2012-05-21T14:04:12.484,10098213,569281734,67.17.10.12,43CD1A7B8322,SA-2100

Middleware Error

May 21 14:04:12.996 wl-01.acme.com Order 569281734 failed for customer 10098213. Exception follows: weblogic.jdbc.extensions.ConnectionDeadSQLException: weblogic.common.resourcepool.ResourceDeadException: Could not create pool connection. The DBMS driver exception was: [BEA][Oracle JDBC Driver]Error establishing socket to host and port: ACMEDB-01:1521. Reason: Connection refused

Care IVR

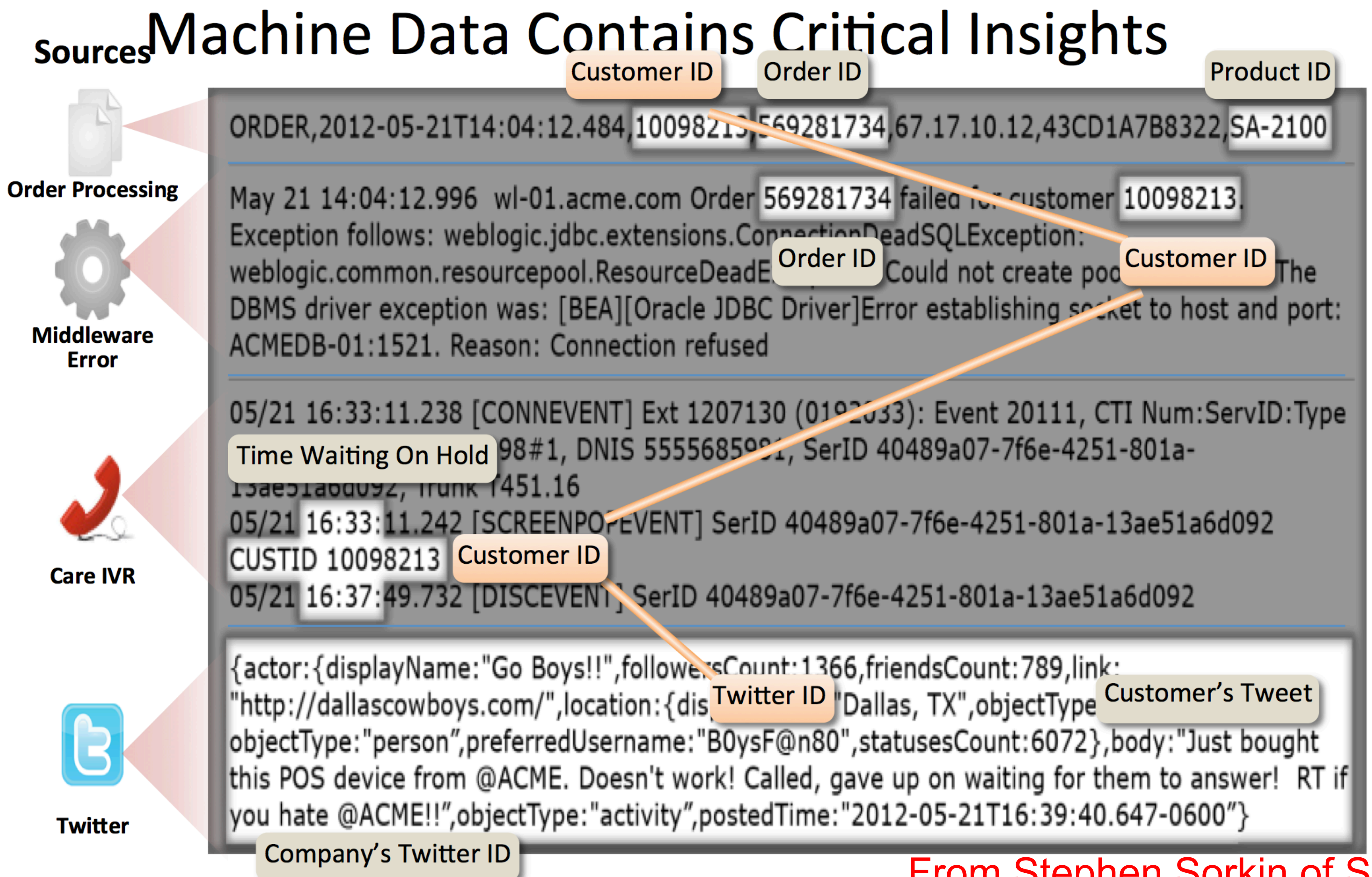
05/21 16:33:11.238 [CONNEVENT] Ext 1207130 (0192033): Event 20111, CTI Num:ServID:Type 0:19:9, App 0, ANI T7998#1, DNIS 5555685981, SerID 40489a07-7f6e-4251-801a-13ae51a6d092, Trunk T451.16
05/21 16:33:11.242 [SCREENPOPEVENT] SerID 40489a07-7f6e-4251-801a-13ae51a6d092 CUSTID 10098213
05/21 16:37:49.732 [DISCEVENT] SerID 40489a07-7f6e-4251-801a-13ae51a6d092

Twitter

{actor:{displayName:"Go Boys!!",followersCount:1366,friendsCount:789,link: "http://dallascowboys.com/",location:{displayName:"Dallas, TX",objectType:"place"}, objectType:"person",preferredUsername:"B0ysF@n80",statusesCount:6072},body:"Just bought this POS device from @ACME. Doesn't work! Called, gave up on waiting for them to answer! RT if you hate @ACME!!",objectType:"activity",postedTime:"2012-05-21T16:39:40.647-0600"}

From Stephen Sorkin of Splunk

Integration of Variety



From Stephen Sorkin of Splunk

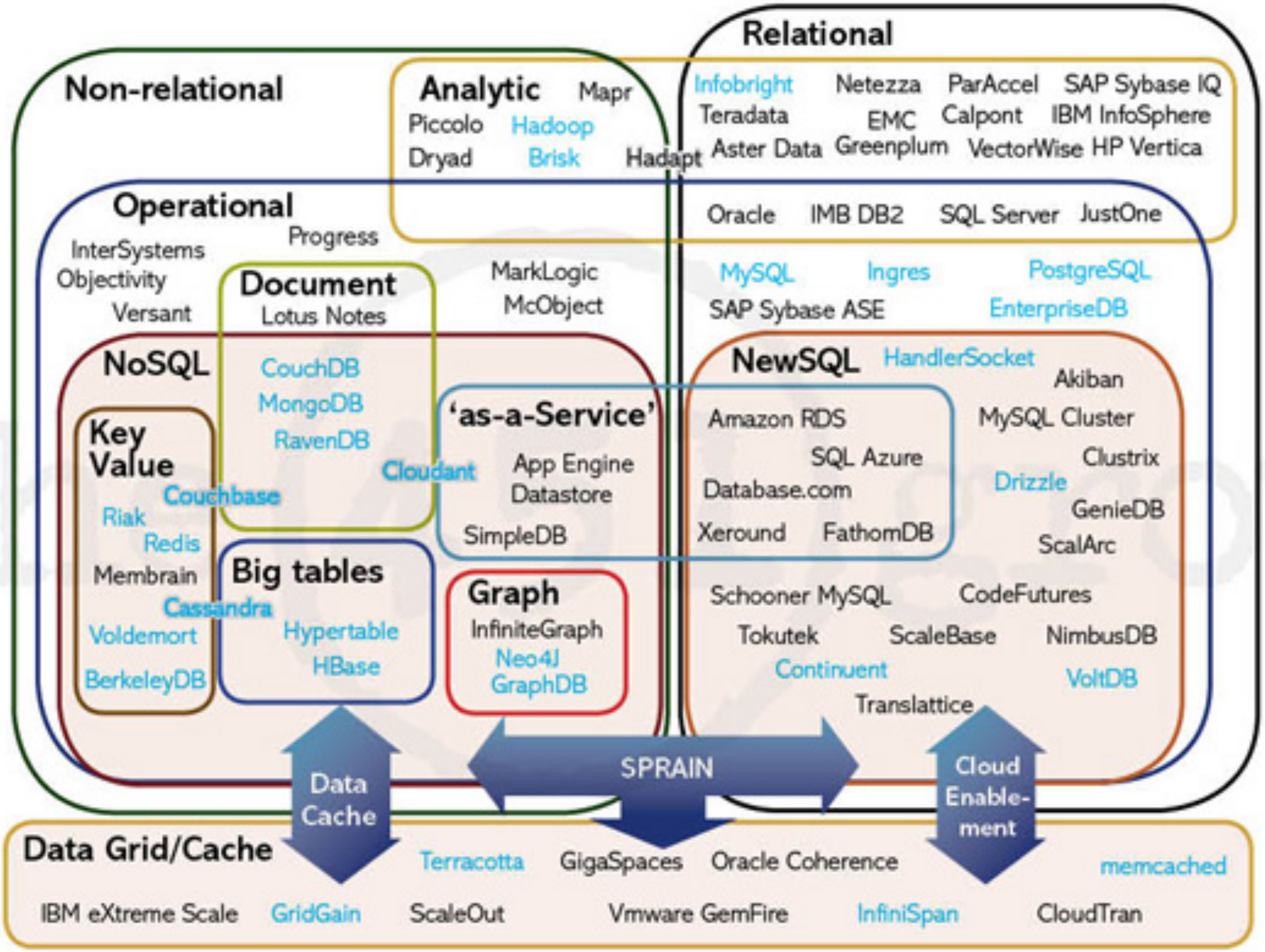
Current Environment

- Data Analysis as a profit center
  - – No longer just a cost – may be the entire business as in Business Intelligence

Current Environment

- Expanded developer demands
  - Wider use means broader requirements, and less interest from developers in the details of traditional DBMS interactions
- Architectural Shifts in Computing
  - The move to parallel architectures both internally (on individual chips)
  - And externally – Cloud Computing/Grid Computing

The Database Universe 201x



The Semantic Web

- The basic structure of the Semantic Web is based on RDF triples (as XML or some other form)
- Conventional DBMS are very bad at doing some of the things that the Semantic Web is supposed to do… (.e.g., spreading activation searching)
- “Triple Stores” are being developed that are intended to optimize for the types of search and access needed for the Semantic Web
- What if it really takes off?

Preview: Massively Parallel Processing

- MPP used to mean that you had to write a lot of code to partition tasks and data, run them on different machines, and combine the results back together
- That has now largely been replaced due to the MapReduce paradigm

MapReduce and Hadoop

- MapReduce developed at Google
  - To run the web crawlers and search engine
- MapReduce implemented in Nutch
  - Doug Cutting at Yahoo!
  - Became Hadoop (named for Doug's child's stuffed elephant toy)

Motivation

- Large-Scale Data Processing
  - Want to use 1000s of CPUs
    - But don’t want hassle of ***managing*** things
- MapReduce provides
  - Automatic parallelization & distribution
  - Fault tolerance
  - I/O scheduling
  - Monitoring & status updates

From “MapReduce…” by Dan Weld