Principles of Database Systems (CS307)

Lecture 15: Advanced Topics

Ran Cheng

Department of Computer Science and Engineering
Southern University of Science and Technology

- Most contents are from slides made by Stéphane Faroult, Dr Yuxin Ma and the authors of Database System Concepts (7th Edition).
- Their original slides have been modified to adapt to the schedule of CS307 at SUSTech.

Beyond Tables: More Data Types

Semi-Structured Data

- Many applications require storage of complex data, whose schema changes often
- The relational model's requirement of atomic data types may be **an overkill**
  - E.g., storing set of interests as a set-valued attribute of a user profile may be simpler than normalizing it
- Data exchange can benefit greatly from semi-structured data
  - Exchange can be between applications, or between back-end and front-end of an application
  - Web-services are widely used today, with complex data fetched to the front-end and displayed using a mobile app or JavaScript
- **JSON** and **XML** are widely used semi-structured data models

Features of Semi-Structured Data Models

- Flexible schema
  - Wide column representation: allow each tuple to have a different set of attributes, can add new attributes at any time
  - Sparse column representation: schema has a fixed but large set of attributes, by each tuple may store only a subset

Features of Semi-Structured Data Models

- Multivalued data types
  - Sets, multisets
    - E.g.,: set of interests: {‘basketball, ‘cooking’, ‘anime’, ‘jazz’}
  - Key-value map (or just map for short)
    - Store a set of key-value pairs
    - E.g.,
      - {(brand, Apple), (ID, MacBook Air), (size, 13), (color, silver)}
    - Operations on maps
      - put(key, value)
      - get(key)
      - delete(key)

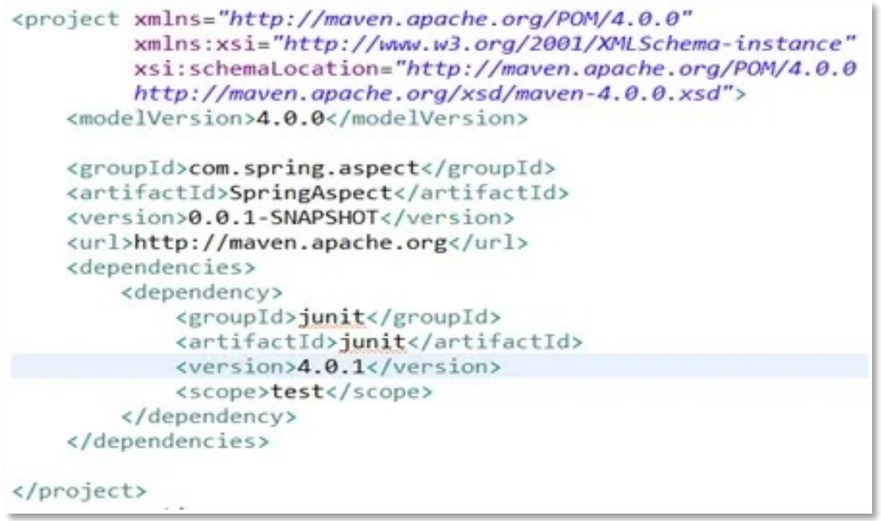
Features of Semi-Structured Data Models

- Arrays
  - Widely used for scientific and monitoring applications
  - E.g., readings taken at regular intervals can be represented as array of values instead of (time, value) pairs
    - [5, 8, 9, 11] instead of {(1,5), (2, 8), (3, 9), (4, 11)}
- Array database: a database that provides specialized support for arrays
  - E.g., compressed storage, query language extensions, etc.
  - Oracle GeoRaster, PostGIS, SciDB, etc

Nested Data Types

- Hierarchical data is common in many applications

- **JSON** (JavaScript Object Notation)
  - Widely used today
- **XML** (eXtensible Markup Language)
  - Earlier generation notation, still used extensively



JSON

- Textual representation widely used for data exchange

- Types: integer, real, string, and
  - Objects: key-value maps, i.e. sets of (attribute name, value) pairs
  - Arrays: also key-value maps (from offset to value)

```
{
    "ID": "22222",
    "name": {
        "firstname: "Albert",
        "lastname: "Einstein"
    },
    "deptname": "Physics",
    "children": [
        {"firstname": "Hans", "lastname": "Einstein" },
        {"firstname": "Eduard", "lastname": "Einstein" }
    ]
}
```

JSON

- JSON is ubiquitous in data exchange today
  - Widely used for web services
    - Most modern applications are architected around on web services
- PostgreSQL supports JSON format columns

```
create table json_test (
    id serial not null primary key,
    student json not null
);

insert into json_test (student) values ('{"name": "aaa", "age": 20, "major": {"primary": "cs", "minor":
"math"}}');
insert into json_test (student) values ('{"name": "bbb", "major": {"primary": "math", "minor": "physics"}}');
insert into json_test (student) values ('{"name": "ccc", "age": 19, "major": {"primary": "biology"}}');
```

JSON

- JSON is ubiquitous in data exchange today
  - Widely used for web services
    - Most modern applications are architected around on web services
- PostgreSQL supports JSON format columns

```
-- select all content from the column
select * from json_test;

-- select a value of a key with "->"
select student -> 'major' -> 'minor' from json_test;
```

	id	student
1	1	{"name": "aaa", "age": 20, "major": {"primary": "cs", "minor": "math"}}
2	2	{"name": "bbb", "major": {"primary": "math", "minor": "physics"}}
3	3	{"name": "ccc", "age": 19, "major": {"primary": "biology"}}

	?column?
1	"math"
2	"physics"
3	<null>

XML

- XML uses tags to mark up text
  - Tags make the data self-documenting
  - Tags can be hierarchical

```
<course>
    <course id>CS-101</course id>
    <title>Intro. to Computer Science</title>
    <dept name>Comp. Sci.</dept name>
    <credits>4</credits>
</course>
```

```
<info>
    <name>aaa</name>
    <age>20</age>
    <major>
        <primary>cs</primary>
        <minor>math</minor>
    </major>
</course>
```

Textual Data

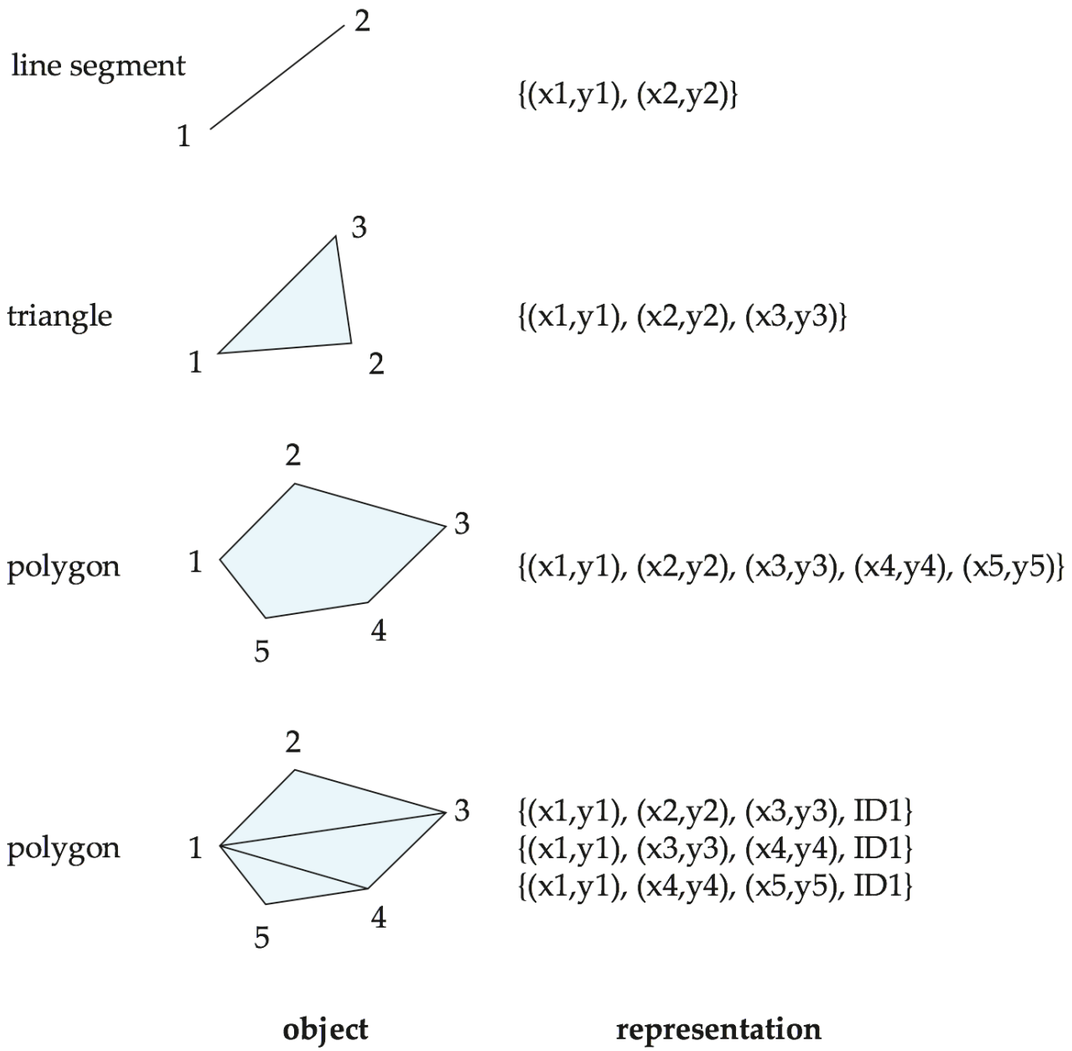
- Information retrieval: querying of unstructured data
  - Simple model of keyword queries: given query keywords, retrieve documents containing all the keywords
  - More advanced models rank relevance of documents
  - Today, keyword queries return many types of information as answers
    - E.g., a query “cricket” typically returns information about ongoing cricket matches
- Relevance ranking
  - Essential since there are usually many documents matching keywords

Spatial Data

- **Spatial databases** store information related to spatial locations, and support efficient storage, indexing and querying of spatial data.
  - **Geographic data**: road maps, land-usage maps, topographic elevation maps, political maps showing boundaries, land-ownership maps, and so on.
    - Geographic information systems (GIS) are special-purpose databases tailored for storing geographic data.
    - Round-earth coordinate system may be used
      - (Latitude, longitude, elevation)
  - **Geometric data**: design information about how objects are constructed
    - E.g., designs of buildings, aircraft, layouts of integrated-circuits.
    - 2 or 3 dimensional Euclidean space with (X, Y, Z) coordinates

Representation of Geometric Information

- Various geometric constructs can be represented in a database in a normalized fashion



Representation of Geometric Information

- Representation of points and line segment in 3-D similar to 2-D, except that points have an extra “z” component
  - Represent arbitrary polyhedra by dividing them into tetrahedrons
    - Similar to triangulating polygons
  - Alternative
    - List their faces, each of which is a polygon, along with an indication of which side of the face is inside the polyhedron

Representation of Geometric Information

- Geometry and geography data types supported by many databases
  - E.g. PostGIS
    - point, linestring, curve, polygons
    - Collections: multipoint, multilinestring, multicurve, multipolygon
    - LINESTRING(1 1, 2 3, 4 4)
    - POLYGON((1 1, 2 3, 4 4, 1 1))
    - Type conversions: ST GeometryFromText() and ST GeographyFromText()
    - Operations: ST Union(), ST Intersection(), …

NoSQL Database

- “Not Only SQL”
  - Useful when working with a huge quantity of data when nature of data does not require a relational model
  - Usually not built on tables and queried by SQL
- Examples
  - Document store – MongoDB
  - Graph structure – Neo4j
  - Key-value storage – Redis, LevelDB
  - Tabular – Apache Hbase (Hadoop-based)

Beyond Storage: Big Data Analytics

What is Data, By the Way?

data noun, plural in form but singular or plural in construction, often attributive

Save Word

da·ta | \ ˈdā-tə, ˈda- *also* ˈdä- \

Definition of *data*

1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

// the *data* is plentiful and easily available
— H. A. Gleason, Jr.

// comprehensive *data* on economic growth have been published
— N. H. Jacoby

2 : information in digital form that can be transmitted or processed

3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

https://www.merriam-webster.com/dictionary/data

Data Attribute

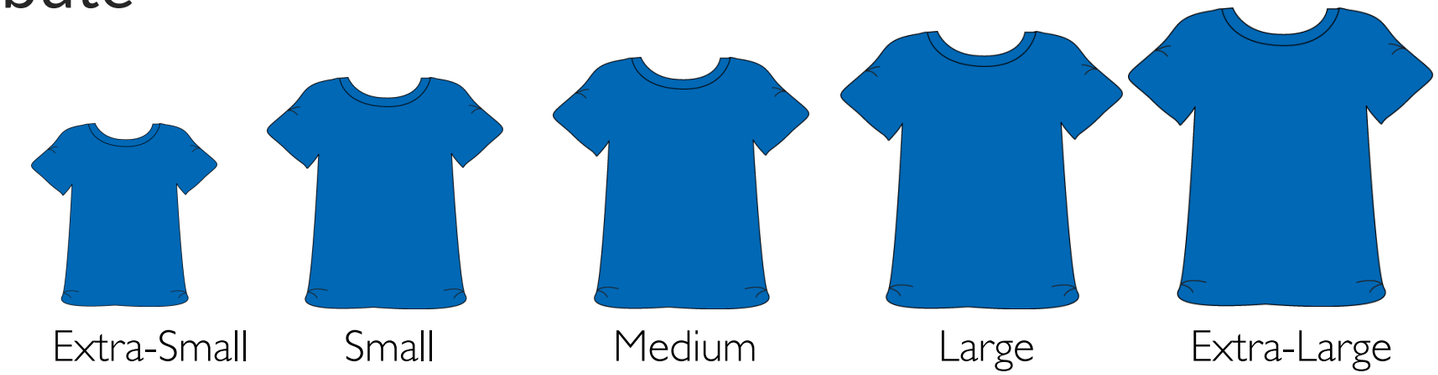
- Characteristics or feature of a data object
  - Other names:
    - Feature
    - Dimension
    - Variable

- Set of attributes: attribute vector

Attribute Types

- Nominal (Categorical) attribute

- Ordinal attribute



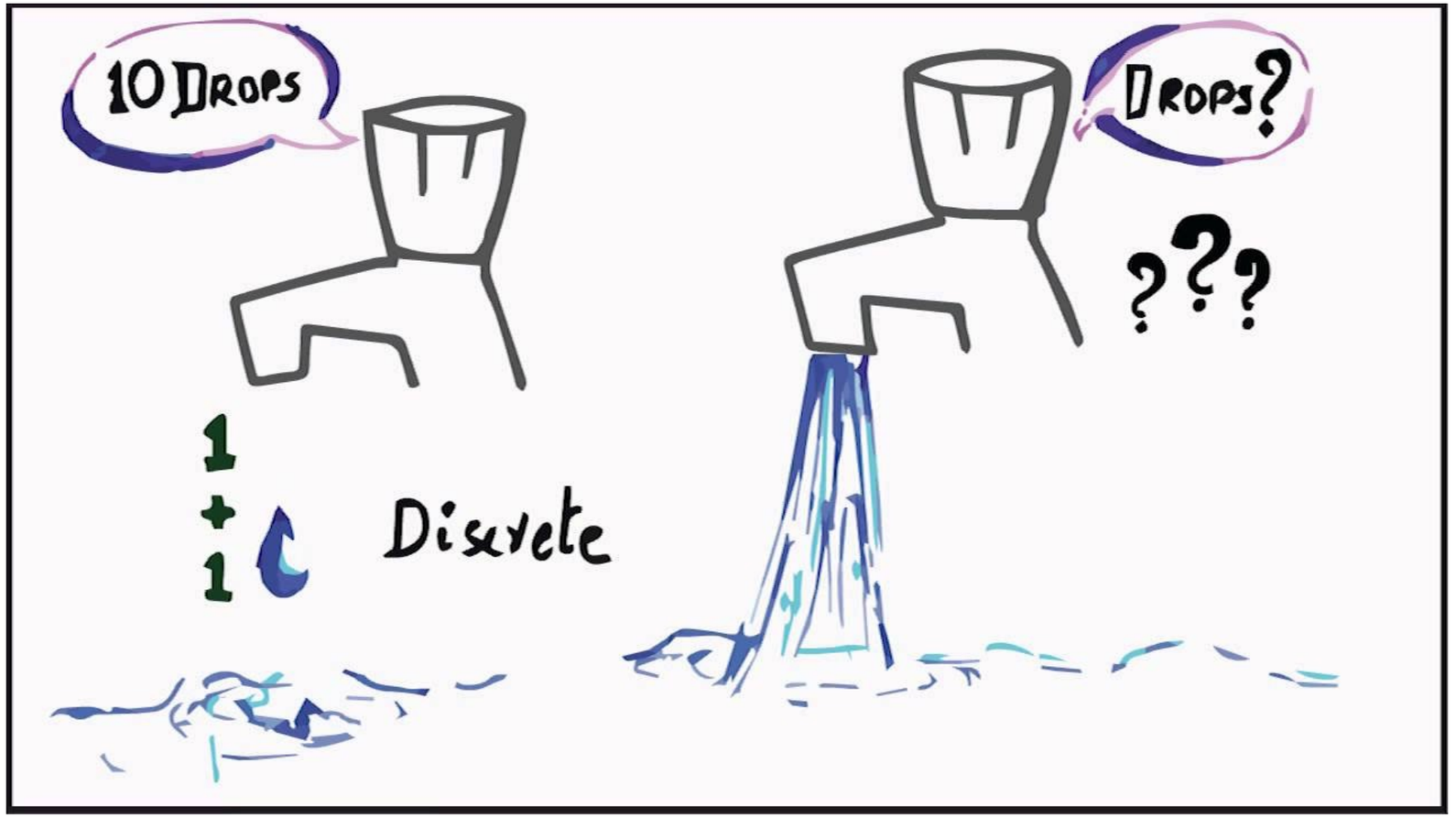
Attribute Types

- Numeric attribute

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa
5.4	3.4	1.7	0.2	setosa
5.1	3.7	1.5	0.4	setosa

https://archive.ics.uci.edu/ml/datasets/iris

Discrete vs. Continuous



Basic Statistical Descriptions

- Overall picture of your data
- Basis of exploratory data analysis

- Mean

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

- Median

$$\mathrm{Q}_{\frac{1}{2}}(x) = \begin{cases} x'_{\frac{n+1}{2}}, & \text{if } n \text{ is odd.} \\ \frac{1}{2}(x'_{\frac{n}{2}} + x'_{\frac{n}{2}+1}), & \text{if } n \text{ is even.} \end{cases}$$

- Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \bar{x}^2,$$

Relationship between Data Objects: Data (Dis)Similarity

- Measurement of relationships
  - Commonly used in many statistical methods and data mining algorithms

- Dissimilarity Matrix & Distance Measures

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Euclidean	$d(x,y) = \sqrt{\sum (x_i - y_i)^2}$
Squared Euclidean	$d(x,y) = \sum (x_i - y_i)^2$
Manhattan	$d(x,y) = \sum \lvert (x_i - y_i) \rvert$
Canberra	$d(x,y) = \sum \frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert}$
Chebychev	$d(x,y) = \max(\lvert x_i - y_i \rvert)$
Bray Curtis	$d(x,y) = \frac{\sum \lvert x_i - y_i \rvert}{\sum x_i + y_i}$
Cosine Correlation	$d(x,y) = \frac{\sum (x_i y_i)}{\sqrt{\sum (x_i)^2 \sum (y_i)^2}}$
Pearson Correlation	$d(x,y) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (y_i - \overline{y})^2} \sqrt{\sum (y_i - \overline{y})^2}}$
Uncentered Peason Correlation	$d(x,y) = \frac{\sum x_i y_i}{\sqrt{\sum (y_i - \overline{y})^2} \sqrt{\sum (y_i - \overline{y})^2}}$
Euclidean Nullweighted	Same as Euclidean, but only the indexes where both x and y have a value (not NULL) are used, and the result is weighted by the number of values calculated. Nulls must be replaced by the missing value calculator (in dataloader).

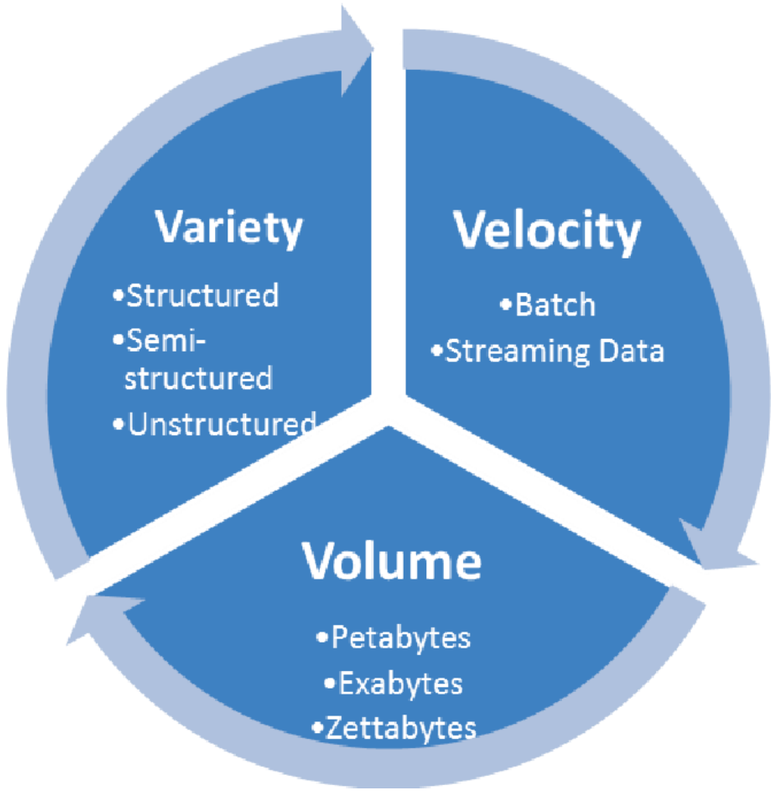
What is Big Data?

- A collection of data sets so large and complex

Three Dimensions of Big Data

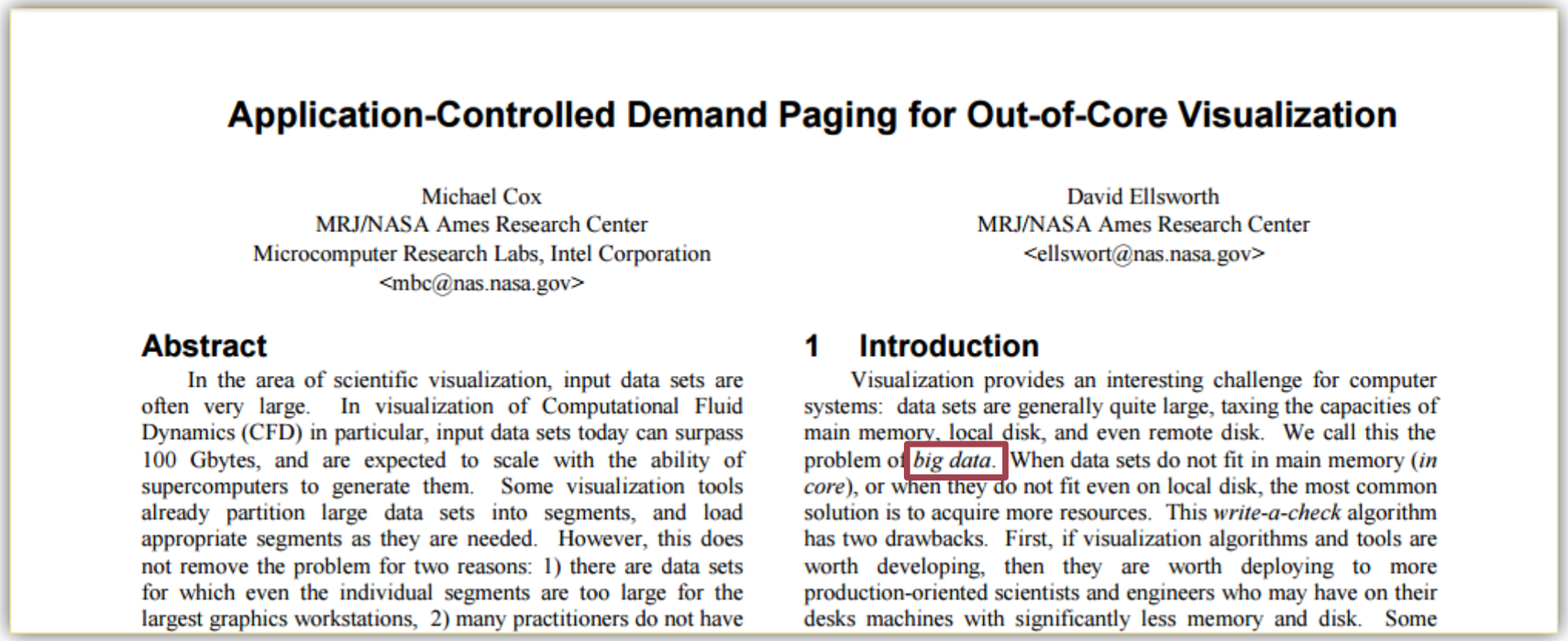
- **Volume**
  - From GB to TB, PB, or higher
- **Velocity**
  - Processing speed
- **Variety**
  - Text, sensor data, multimedia, …

- Other (new) aspects:
  - Veracity: Trustworthiness
  - Value: Worth of data



The Emergence of Data Science

- 1997: The first article to use the term “big data” in the ACM digital library.



Application-Controlled Demand Paging for Out-of-Core Visualization

Michael Cox
MRJ/NASA Ames Research Center
Microcomputer Research Labs, Intel Corporation
<mbc@nas.nasa.gov>

David Ellsworth
MRJ/NASA Ames Research Center
<ellswort@nas.nasa.gov>

Abstract

In the area of scientific visualization, input data sets are often very large. In visualization of Computational Fluid Dynamics (CFD) in particular, input data sets today can surpass 100 Gbytes, and are expected to scale with the ability of supercomputers to generate them. Some visualization tools already partition large data sets into segments, and load appropriate segments as they are needed. However, this does not remove the problem for two reasons: 1) there are data sets for which even the individual segments are too large for the largest graphics workstations, 2) many practitioners do not have

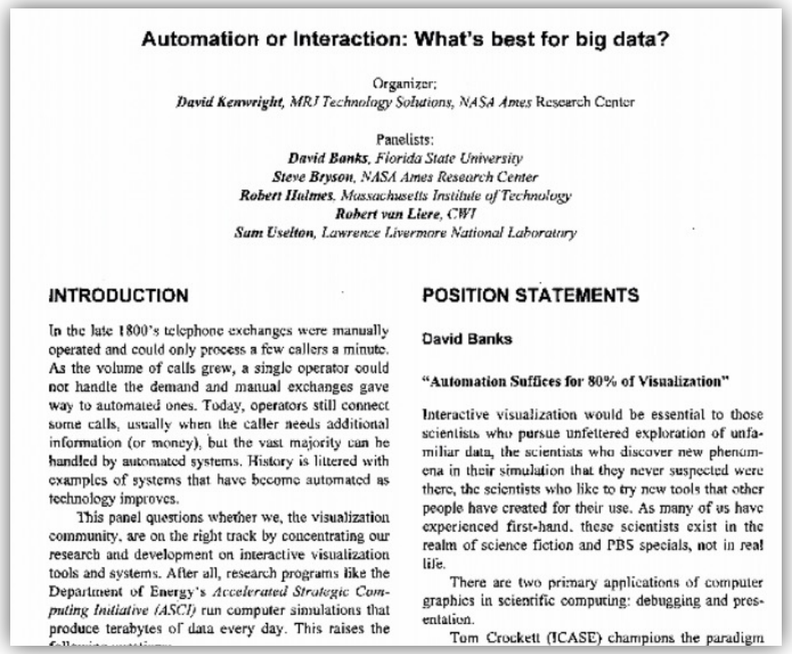
1 Introduction

Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of *big data*. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources. This *write-a-check* algorithm has two drawbacks. First, if visualization algorithms and tools are worth developing, then they are worth deploying to more production-oriented scientists and engineers who may have on their desks machines with significantly less memory and disk. Some

Proceedings of the Visualization ‘97

The Emergence of Data Science

- 1999: Bryson, Kenwright and Haimes join David Banks, Robert van Liere, and Sam Uselton on a panel titled “Automation or interaction: what’s best for big data?” at the IEEE Conference on Visualization in 1999



Automation or Interaction: What's best for big data?

Organizer:
David Kenwright, MRJ Technology Solutions, NASA Ames Research Center

Panelists:
***David Banks**, Florida State University*
***Steve Bryson**, NASA Ames Research Center*
***Robert Haimes**, Massachusetts Institute of Technology*
***Robert van Liere**, CWI*
***Sam Uselton**, Lawrence Livermore National Laboratory*

INTRODUCTION

In the late 1800's telephone exchanges were manually operated and could only process a few callers a minute. As the volume of calls grew, a single operator could not handle the demand and manual exchanges gave way to automated ones. Today, operators still connect some calls, usually when the caller needs additional information (or money), but the vast majority can be handled by automated systems. History is littered with examples of systems that have become automated as technology improves.

This panel questions whether we, the visualization community, are on the right track by concentrating our research and development on interactive visualization tools and systems. After all, research programs like the Department of Energy's *Accelerated Strategic Computing Initiative (ASCI)* run computer simulations that produce terabytes of data every day. This raises the

POSITION STATEMENTS

David Banks

"Automation Suffices for 80% of Visualization"

Interactive visualization would be essential to those scientists who pursue unfettered exploration of unfamiliar data, the scientists who discover new phenomena in their simulation that they never suspected were there, the scientists who like to try new tools that other people have created for their use. As many of us have experienced first-hand, these scientists exist in the realm of science fiction and PBS specials, not in real life.

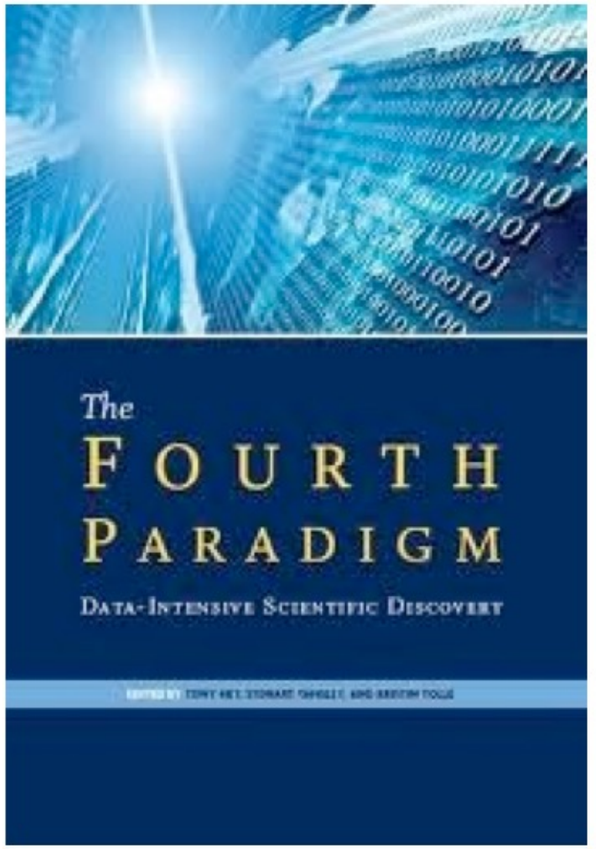
There are two primary applications of computer graphics in scientific computing: debugging and presentation.

Tom Crockett (ICASE) champions the paradigm

Proceedings of the IEEE Conference on Visualization in 1999

The Emergence of Data Science

- 2007: “The Fourth Paradigm”
  - “Data-Intensive Scientific Discovery”



The Emergence of Data Science

- 2012: Data Scientist: The Sexiest Job of the 21st Century

Harvard Business Review

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

The Emergence of Data Science

- 2013: The IEEE Task Force on Data Science and Advanced Analytics was launched.
- 2014: The first international conference: IEEE International Conference on Data Science and Advanced Analytics was launched.
- 2015: The International Journal on Data Science and Analytics was launched by Springer to publish original work on data science and big data analytics.
- 2016: The American Statistical Association section on Statistical Learning and Data Mining renamed to "Statistical Learning and Data Science".

The Emergence of Data Science

- 2016: “Trump vs. Clinton: How Big Data and scientists helped Trump win the election”

Digital campaigning

The role of technology in the presidential election

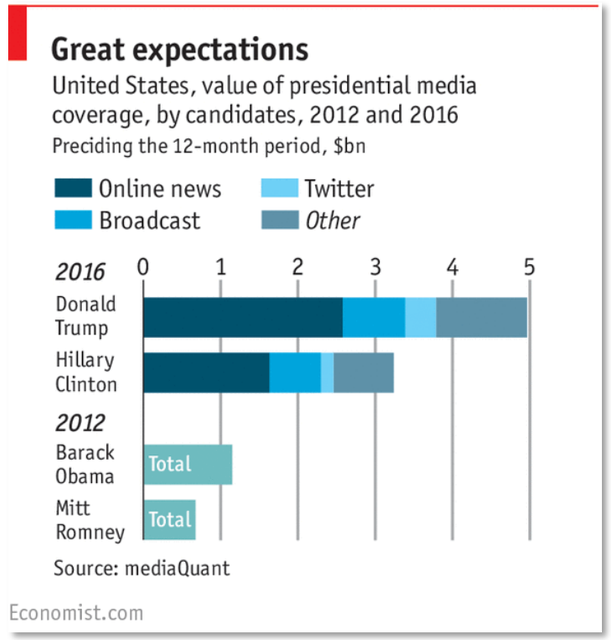
All latest updates

From fake news to big data, a post mortem is under way

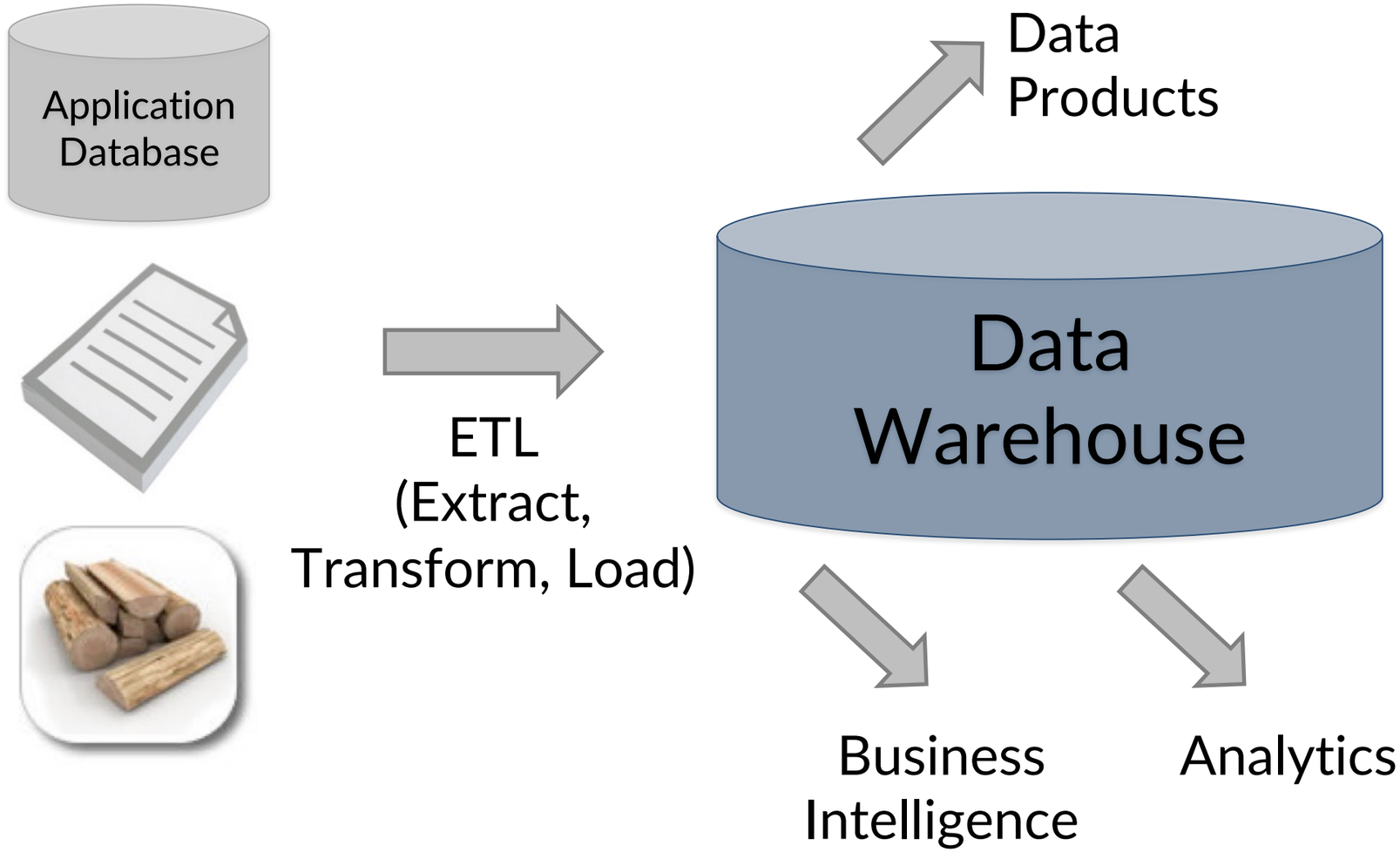
Nov 20th 2016 | United States

Timekeeper Like 916 Tweet

EARLY in America’s presidential campaign, pundits compared the contest between Hillary Clinton and Donald Trump to a fight between a large tanker and Somali pirates. This turned out to be particularly true of the digital campaigns: a massive data battleship lost to a chaotic flotilla of social-media speedboats. The big question now is what this means for future elections, both in America and abroad.



Standard Architecture



Instantiations(1) - Businesspersons

- Data Sources
  - Web pages
  - Excel
- Extract-Transform-Load (ETL)
  - Copy & paste
- Data Warehouse
  - Excel
- BI and Analytics
  - Excel functions
  - Excel charts
  - VB scripts?
  - Visualization tools: Power BI, Tableau

Instantiations(2) - Programmers

- Data Sources
  - Web scraping, web services API
  - CSV files
  - Database queries
- ETL
  - wget, curl, Beautiful Soup, lxml, …
- Data Warehouse
  - Files
- Analytics
  - Numpy, pandas, Matplotlib, R, Octave, …

Instantiations(3) - Enterprises

- Data Sources
  - Application databases(Oracle, IBM, ...)
  - Intranet files
  - Application log files
- ETL
  - Infomatica, IBM DataStage, ...
- Data Warehouse
  - Teradata, Oracle, IBM DB2, ...
- Business Intelligence & Analytics
  - SAS, SPSS, R, ...
  - Power BI, Tableau, Spotfire, ...

Instantiations(4) – Web Companies

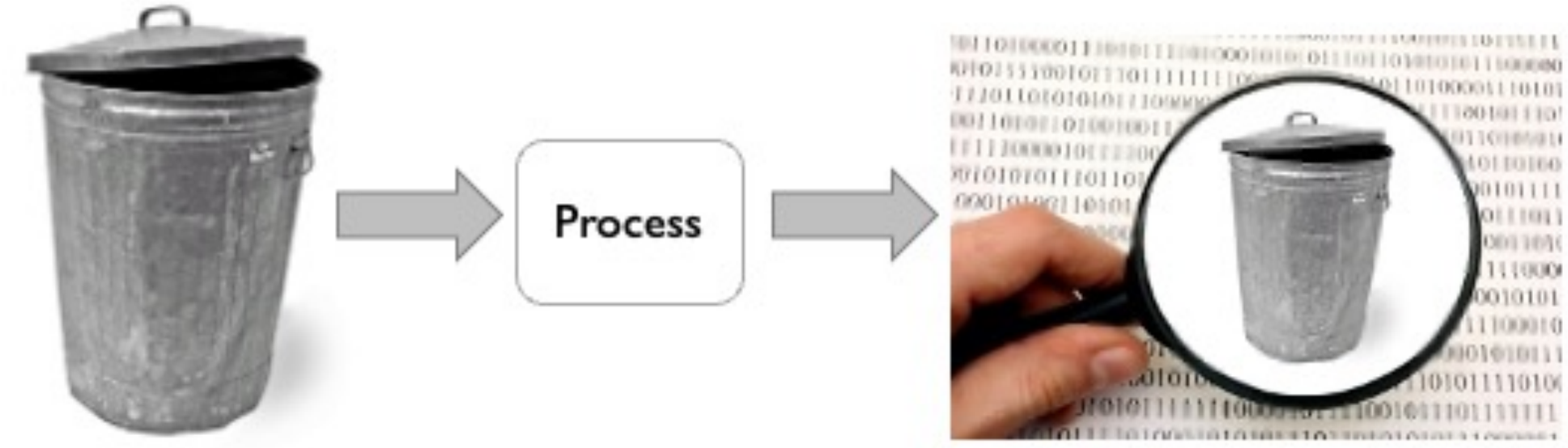
- Data Sources
  - Application databases
  - Logs
  - Web crawl data
- ETL
  - Apache Flume, Apache Sqoop, …
- Data Warehouse
  - Hadoop-based: Hive, Hbase
  - Microsoft Azure, Amazon Redshift
- Business Intelligence & Analytics
  - Argus, R, …

ETL: What is Inside

- Data Cleaning
- Data Integration

“Garbage in, garbage out.”

- Raw data can always be **DIRTY**!



Data Quality

- Data quality: data has quality if it satisfies the requirements of its intended use
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Interpretability

Data Cleaning

- Deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data

Data Cleaning

- Examples of data quality issues
  - Schema-level

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Illegal values	bdate=30.13.70	values outside of domain range
Record	Violated attribute dependencies	age=22, bdate=12.02.70	age = (current date – birth date) should hold
Record type	Uniqueness violation	emp_1=(name="John Smith", SSN="123456") emp_2=(name="Peter Miller", SSN="123456")	uniqueness for SSN (social security number) violated
Source	Referential integrity violation	emp=(name="John Smith", deptno=127)	referenced department (127) not defined

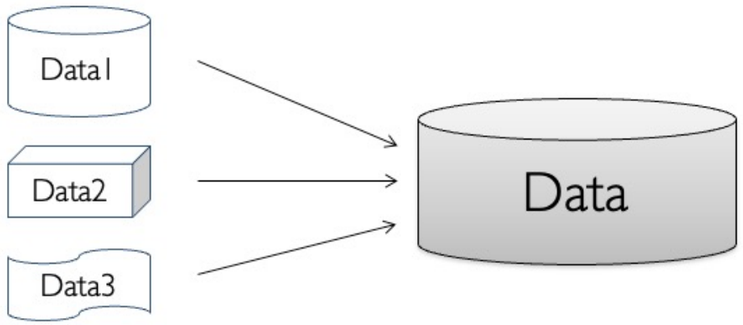
  - Instance-level

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Missing values	phone=9999-999999	unavailable values during data entry (dummy values or null)
	Misspellings	city="Liipzig"	usually typos, phonetic errors
	Cryptic values, Abbreviations	experience="B"; occupation="DB Prog."	
	Embedded values	name="J. Smith 12.02.70 New York"	multiple values entered in one attribute (e.g. in a free-form field)
	Misfielded values	city="Germany"	
Record	Violated attribute dependencies	city="Redmond", zip=77777	city and zip code should correspond
Record type	Word transpositions	$name_1$= "J. Smith", $name_2$="Miller P."	usually in a free-form field
	Duplicated records	emp_1=(name="John Smith",...); emp_2=(name="J. Smith",...)	same employee represented twice due to some data entry errors
	Contradicting records	emp_1=(name="John Smith", bdate=12.02.70); emp_2=(name="John Smith", bdate=12.12.70)	the same real world entity is described by different values
Source	Wrong references	emp=(name="John Smith", deptno=17)	referenced department (17) is defined but wrong

Data Integration

- Data integration involves combining data residing in different sources and providing users with a unified view of these data.
  - Remember “views” in DBMS?

- Management of data from multiple sources



Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Customers (integrated target with cleaned data)

No	*LName*	*FName*	*Gender*	*Street*	*City*	*State*	*ZIP*	*Phone*	*Fax*	*CID*	*Cno*
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

Typical Data Cleaning and Integration Workflow

- Data analysis
  - Detailed inspection before operations
- Conflicts resolution
  - Resolve data confliction between data sources to be integrated
- Definition of transformation workflow and mapping rules
  - Workflow methods for schema adaption and transformation
- Verification of Workflow
  - Verify each steps
- Transformation
  - start the process

Load and Store Data

- File-based Storage
  - Simplest way & easy to manage
  - Scalability is low

- **Database & DBMS**
  - What we have learned for 10+ weeks

- Data Warehouse

Data Warehouse

A data warehouse is a **subject-oriented**, **integrated**, **time-varient**, and **nonvolatile** collection of data in support of management's decision making process.

-- W. H. Inmon, "Building the Data Warehouse". 1996.

Loosely Speaking, a data warehouse refers to a data repository that is **maintained separately** from an organization's operational databases.

-- J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 3rd ed., 2011.

Differences between Databases and Data Warehouses

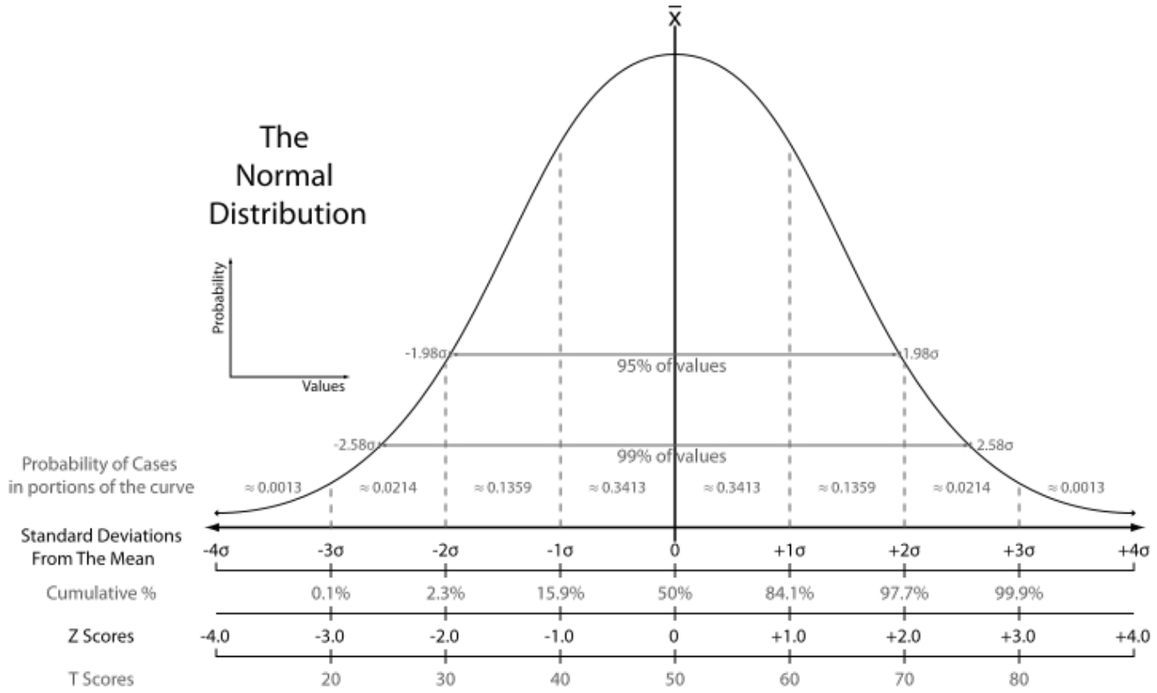
	DB	DW
Characteristics	operational processing	Informational processing
Orientation	transaction	analysis
User	terminal users: clerk, database administrator(DBA)	knowledge workers: manager, analyst, executive
Function	everyday operations	long-term informational requirements decision support
Data	current, up-to-date	historic, accuracy maintained over time
Access	read/write	mostly read
Focus	data in	information/knowledge out
Size	GB to high-order GB	>=TB

Data Analysis

- Exploratory Data Analysis
- Data Mining

Something Basic: Statistics

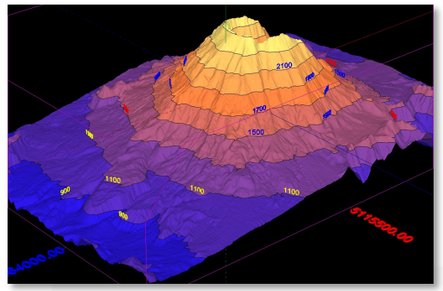
- (Probably) Foundation of modern data analysis
- (Also) Foundation of machine learning, data mining, etc.

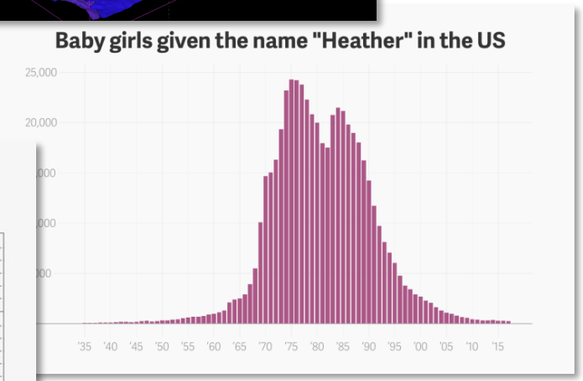


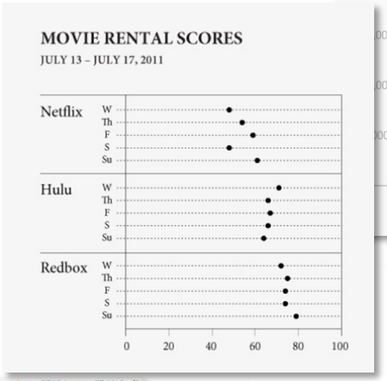
Exploratory Data Analysis (EDA)

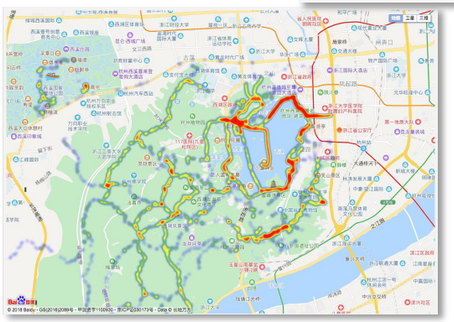
- Based on statistics
  - Data visualization-driven method
  - Summary of main characteristics in easy-to-understand form

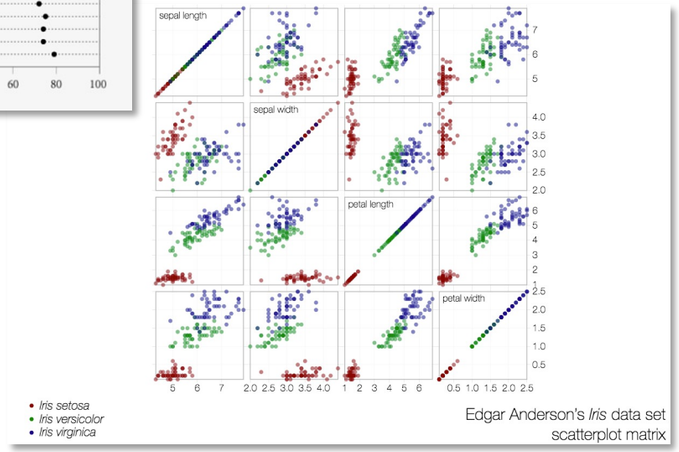
- Types of data visualization methods in EDA:
  - Plotting of raw data
  - Plotting of statistical values
  - Multiple coordinated views (Dashboard)











Exploratory Data Analysis (EDA)

"Some of my friends felt that I should be very explicit in warning you of how much time and money can be wasted on computing, how much clarity and insight can be lost in great stacks of computer output. In fact, I ask you to remember only two points:

1. The tool that is so dull that you cannot cut yourself on it is not likely to be sharp enough to be either useful or helpful.

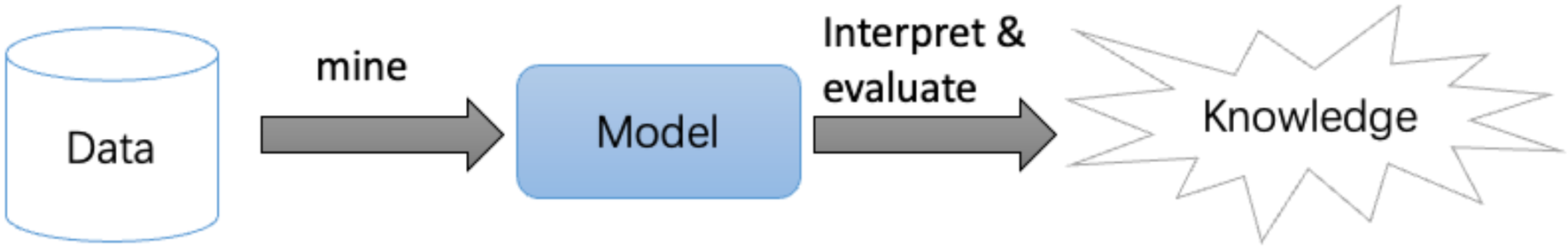
2. Most uses of the classical tools of statistics have been, are, and will be , made by those who know not what they do."

-- John W. Tukey, "The Technical Tools of Statistics",
at the 125th Anniversary Meeting of American Statistical Association, 1964

Data Mining

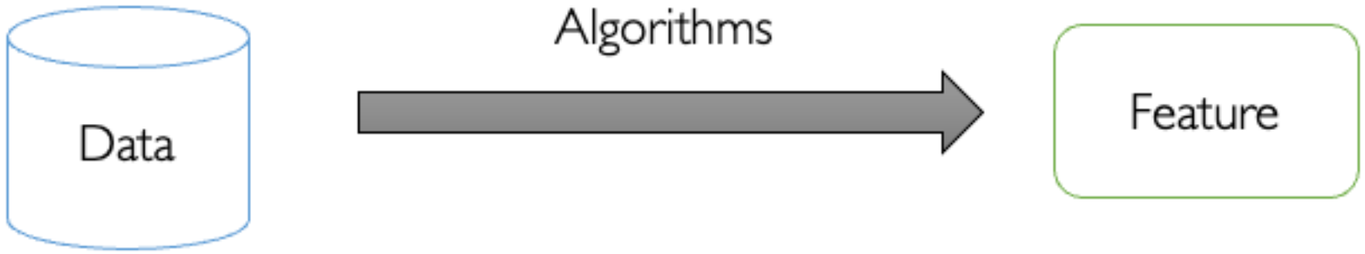
"**Data Mining**, also popularly referred to as **knowledge discovery** from data (KDD), is the automated or convenient **extraction of patterns representing knowledge** implicitly stored or captured in large databases, data warehouses, the Web, other massive repositories, or data streams."

-- H. Jiawei and M. Kamber, "Data Mining: Concepts and Techniques", 3rd ed., 2011.

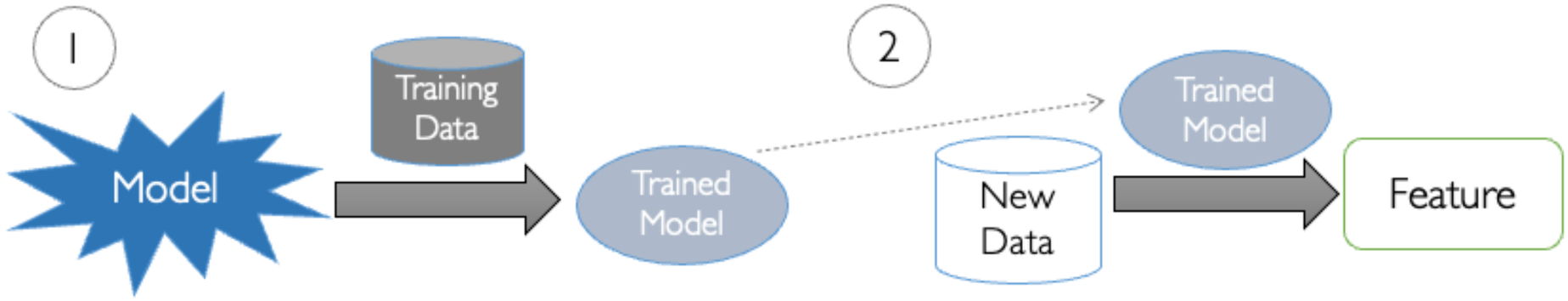


Tasks in Data Mining

- Descriptive Tasks



- Predictive Tasks



Descriptive Tasks

- Concept Description
  - Describe features of data directly
- Association Analysis
  - Analyze “feature-value” pairs that occur frequently in data
- Clustering
  - Group data on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity
- Outlier Detection
  - Analyze objects that do not comply with the general behavior or model of the data

Predictive Tasks

- Regression
  - Model the relationship between a scalar response and a number of variables
- Classification
  - Find a model/function that describes and distinguish data classes or concepts based on analysis of a set of training data
- Evolution Analysis
  - Analyze temporal and spatial patterns in dataset, model these patterns and predict data in unknown spatio-temporal positions

Data Visualization

- **Visualization is the creation and study of the visual representation of data**

Input: data

Output: visual form

Goal: insight

Why Do We Need Visualization?

- Sometimes, statistics may not work

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics
$u_X = 9.0$ $\sigma_X = 3.317$
$u_Y = 7.5$ $\sigma_Y = 2.03$

Linear Regression
$Y = 3 + 0.5\,X$
$R^2 = 0.67$

[Anscombe 73]