CPT-S 415

Big Data

Yinghui Wu

EME B45

CPT_S 415
Big Data

Big data: conclusion

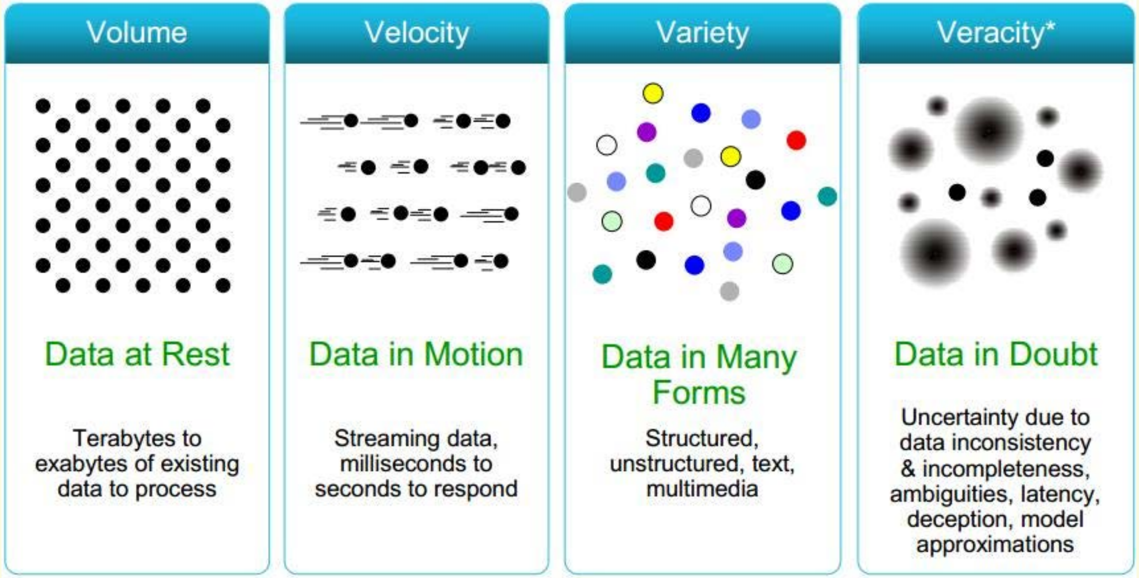
- Big Data: Summary & Vision
- Course project and presentation

Big picture

Big Data: Models	Big Data: Algorithms	Big Data: Systems	Big Data: Analytics & Privacy
•Big V's •Relational model •XMLs and RDFs	•Query models •Sequential search strategies: Make Big Data Small •Parallel and distributed query processing: Make Big Data Distributed •Theory and practice	•Relational DBMS •NoSQL DBMS •In-memory DBMS •NewSQL	•Classification & Clustering •Pattern mining •Data Quality •Data security and privacy
What's Big Data How to present and store Big Data?	Big Data system design	How to search Big Data?	How to analyze Big Data?

Big data models

Big data: the 4 big V's

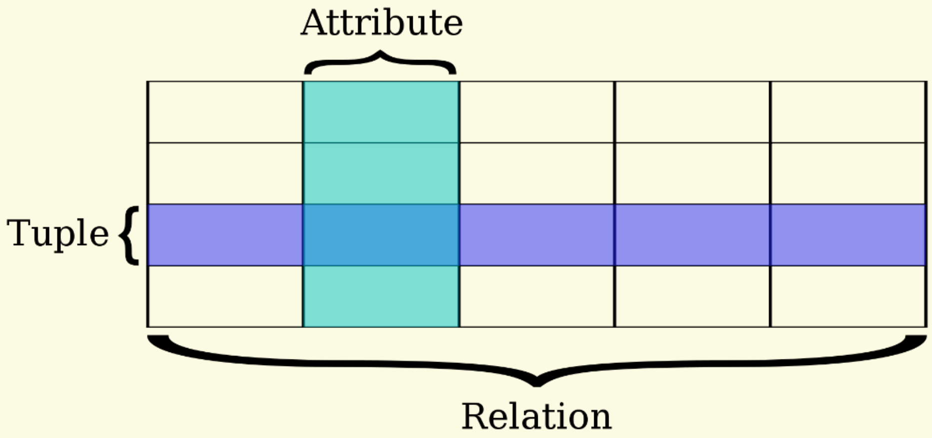


Big data models:

- the 4 big V's,
- data types,
- applications.,
- research trend /topics

- ✓ What is big data? a large, complex data set; a challenge; a trend; an approach of data analytics
- ✓ What is the volume of big data? Variety? Velocity? Veracity?
- ✓ Why do we care about big data?
- ✓ Is there any fundamental challenge introduced by querying big data?
- ✓ Why study Big Data?

Relational data models and DBMS

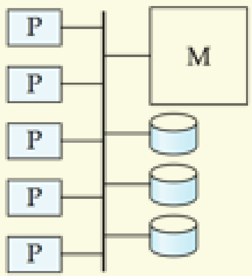


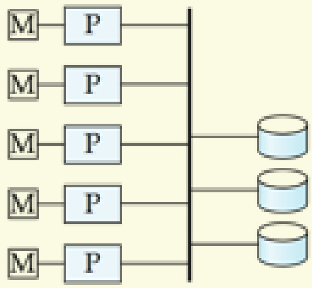
Relational DBMS

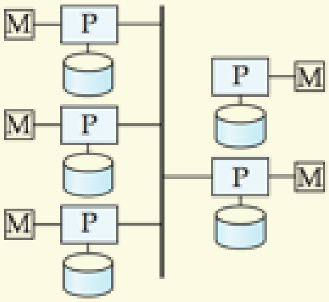
- Relational Model Concepts
- Relational Model Constraints and Schemas
- Update Operations and Dealing with Constraint Violations

- ✓ The relational model has rigorously defined query languages that are simple and powerful.
- ✓ Relational algebra is more operational; useful as internal representation for query evaluation plans.
- ✓ Several ways of expressing a given query; a query optimizer should choose the most efficient version.

DBMS architectures & design

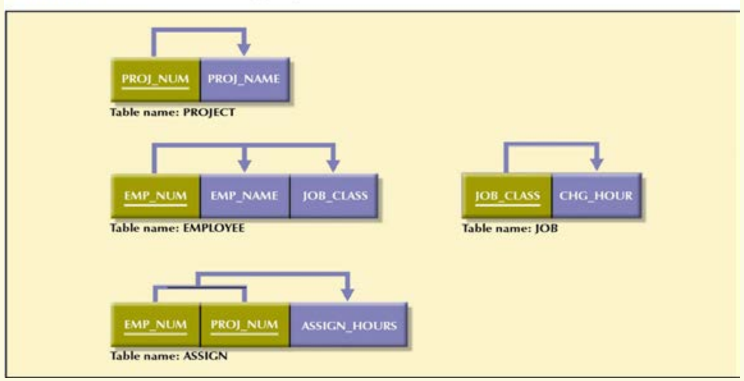






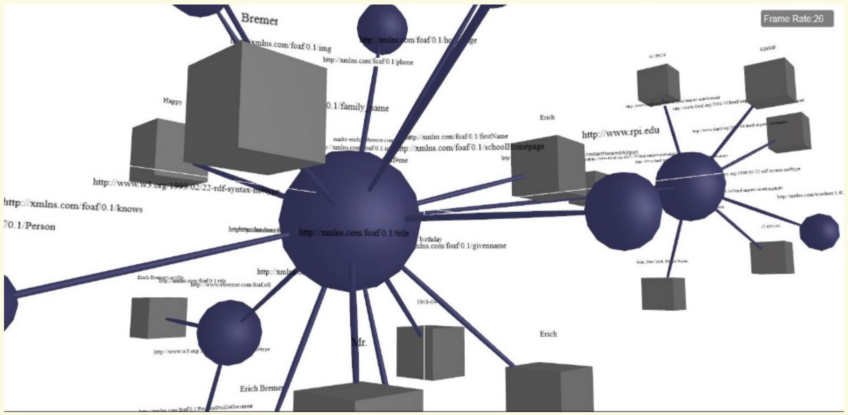
- DBMS: architecture
  - Centralized
  - Client-server
  - Parallel
  - Distributed

FIGURE 5.5 THIRD NORMAL FORM (3NF) CONVERSION RESULTS



- RDBMS design
  - the normal forms 1NF, 2NF, 3NF, BCNF
  - normal forms transformation

Beyond Relational Data



Introduction to XML

- XML basics
- DTD
- XML Schema
- XML Constraints

Introduction to RDF

- RDF data model and syntax
- RDF schemas
- RDF inferencing

XML is a prime data exchange format.

DTD provides useful syntactic constraints on documents.

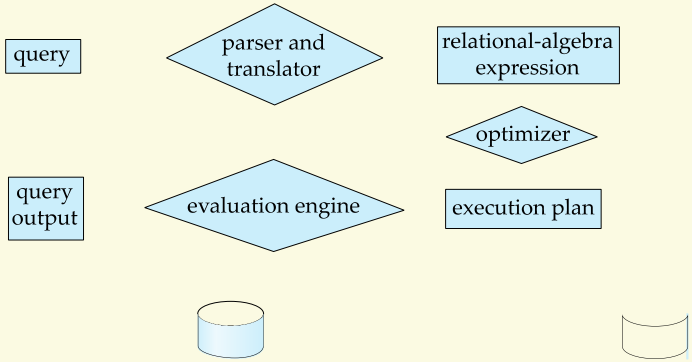
XML Schema extends DTD by supporting a rich type system

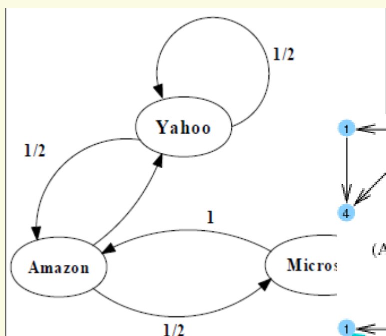
Integrity constraints are important for XML, yet are nontrivial

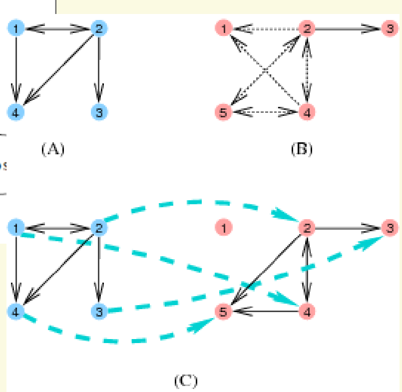
RDF provides a foundation for representing and processing metadata
RDF has a graph-based data model
RDF has an XML-based syntax to support syntactic interoperability
RDF has a decentralized philosophy and allows incremental building of knowledge, and its sharing and reuse

Big data search

Query Processing



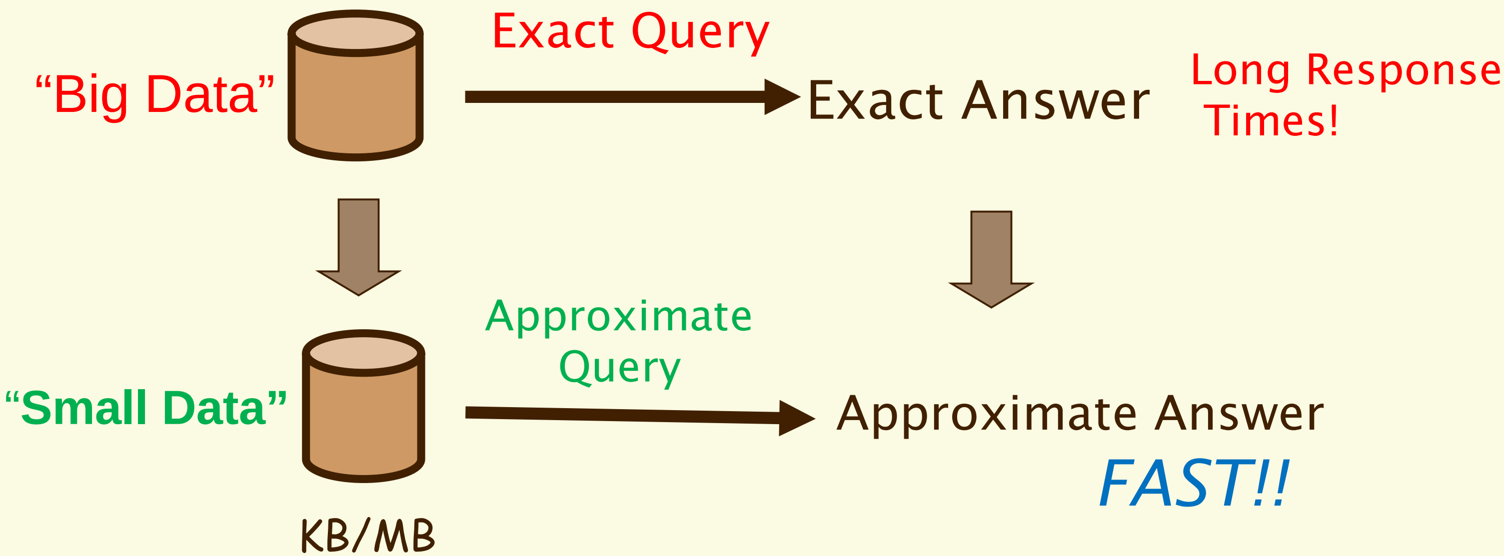




- Querying framework overview
- Measures of Query Cost
- Basic of Database operations
- Basics of Graph Queries

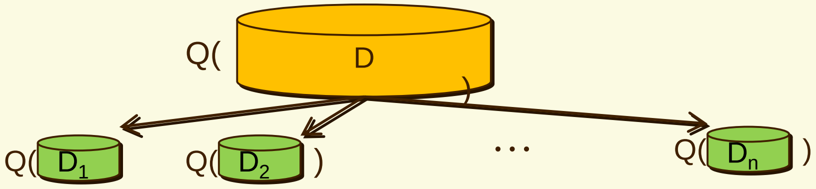
- Basics of Graph Algorithms
  - Graph search (traversal)
  - PageRank
  - Nearest neighbors
  - Keyword search
  - Graph pattern matching

Big data search: Make big data small

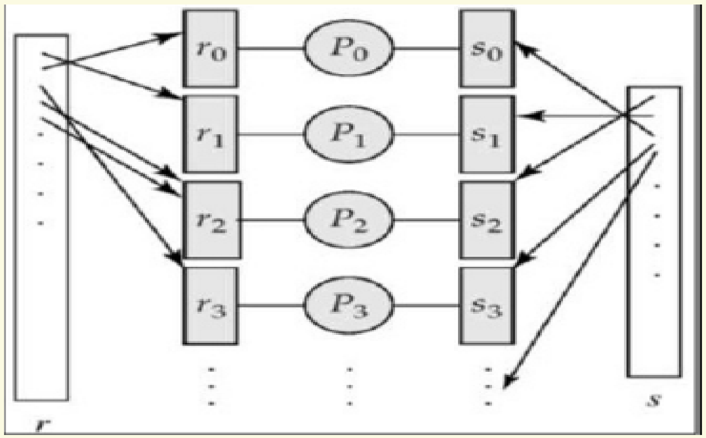


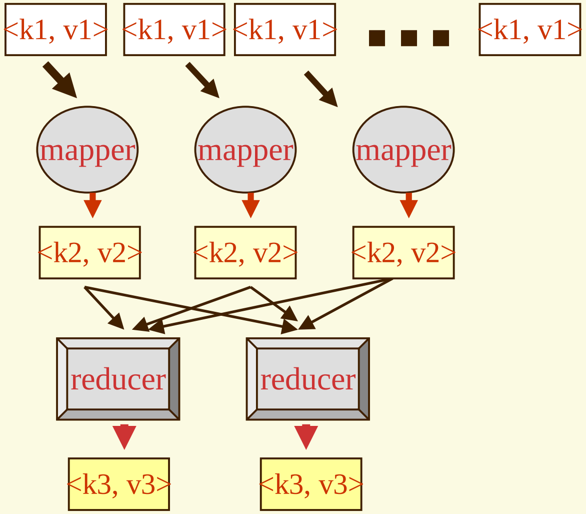
- ✓ Approximate query evaluation
  - query driven: approximate query models
  - data driven: synopses, histogram, sampling, sketches, spanners…
- ✓ View-based query evaluation
- ✓ Make big data small: indexing, sketch, sampling, spanners
- ✓ Cope with data streams: incremental query evaluation

Parallel data management



- parallel DBMS Architectures
- 4 Parallelism: Intraquery, Interquery Intraoperation,Interoperation
- MapReduce

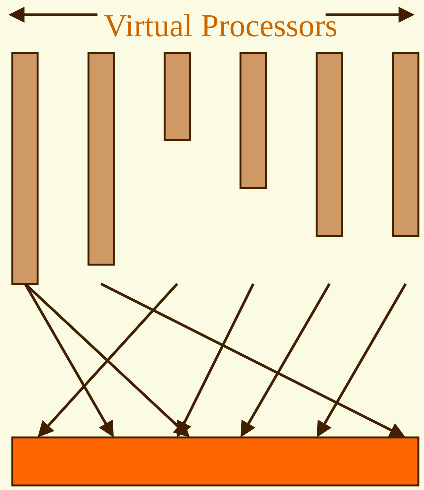


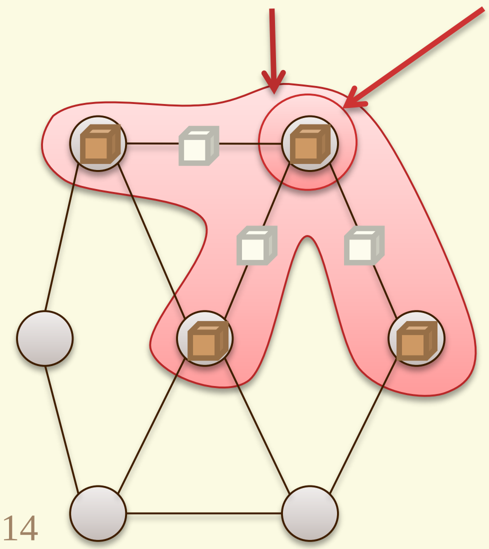


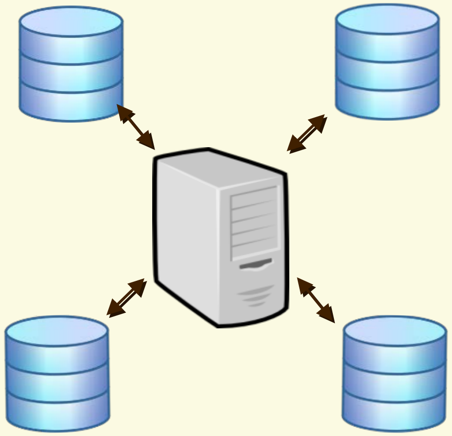
Scalable Big Data Search

Query processing: Make it distributed

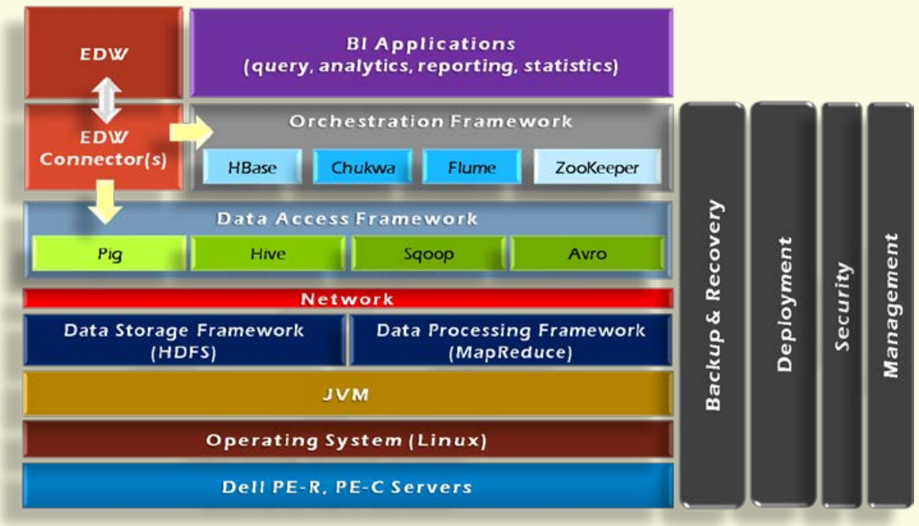
- ✓ Parallel programming models
  - MapReduce for BFS for distance queries, PageRank..
  - Vertex Centric Programming: GraphLab and Pregel
  - Graph Centric Programming: Giraph ++
  - GRAPE: Hybrid models







Hadoop



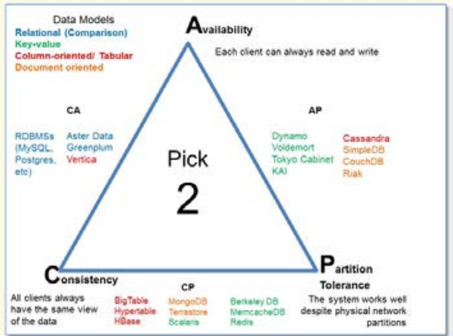
Hadoop: history, features and design

Hadoop ecosystem

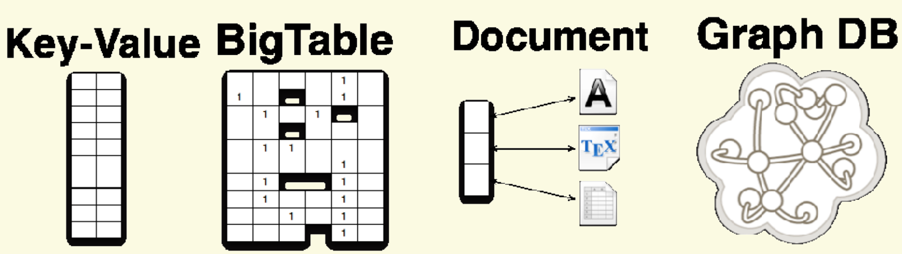
- HDFS
- Hive & Pig
- Hbase
- Zookeeper

oldSQL vs. noSQL

ACID



EASE

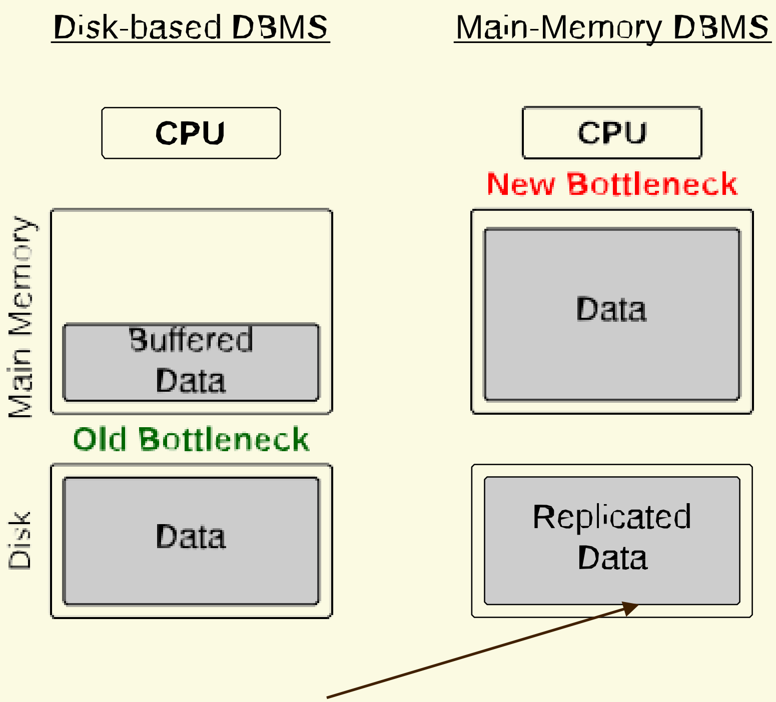


- noSQL: concept and theory
  - CAP theory
  - ACID vs EASE
  - noSQL vs RDBMS
- noSQL databases
  - Key-value stores
  - Document DBs
  - Column family
  - Graph databases

- Cheap, easy to implement (open source)
- Data are replicated to multiple nodes (fault-tolerant)
- Easy to distribute
- Don't require a schema
- Can scale up and down
- Relax the data consistency requirement (CAP)

- Joins, ACID transactions
- SQL as a sometimes frustrating but still powerful query language
- easy integration with other applications that support SQL

Disk-based vs. Main-Memory DBMS



Row-store or column store?

Disk bottleneck is removed as database is kept in main memory

→ Access to main memory becomes new bottleneck

tuple-at-a-time

vectorized execution

operator-at-a-time

oldSQL vs. noSQL vs. NewSQL

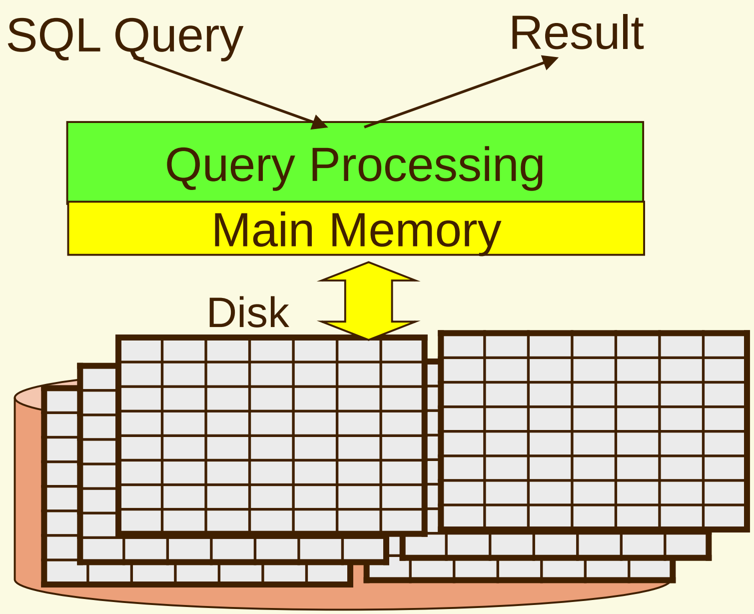
- “A DBMS that delivers the scalability and flexibility promised by NoSQL while retaining the support for SQL queries and/or ACID, or to improve performance for appropriate workloads.”

- SQL + ACID + performance and scalability through modern innovative software architecture

- *Principle 1: minimizing or stay away from locking*
- *Principle 2: rely on main memory*
- *Principle 3: try to avoid latching*
- *Principle 4: cheaper solutions for HA*

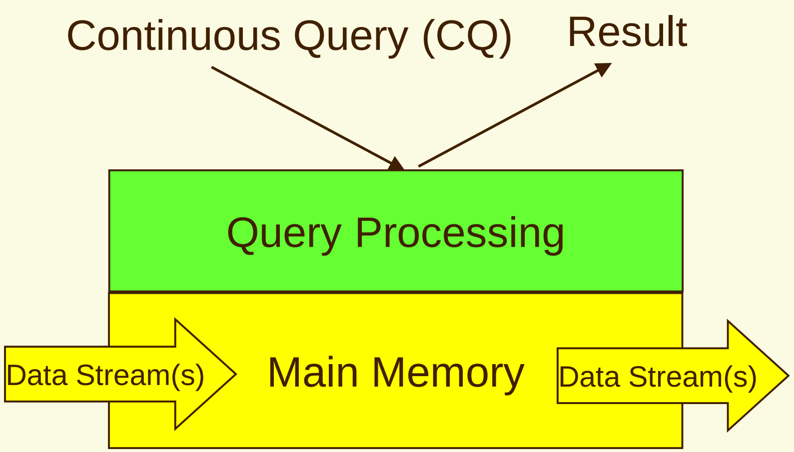
DBMS vs. DSMS

- Traditional DBMS:
  - static records with no pre-defined notion of time
  - persistent data storage and complex querying



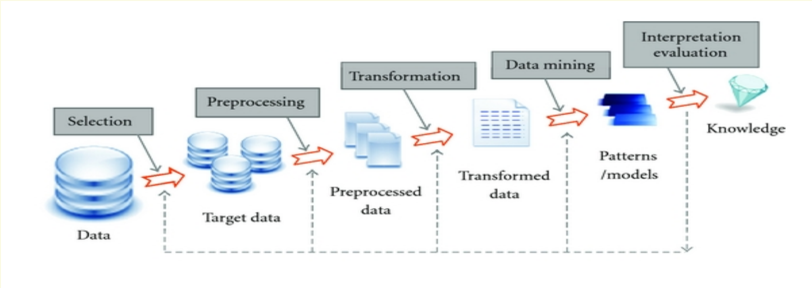
DSMS:

- on-line analysis of rapidly changing data streams
- *data stream*
- sequence of items, too large to store entirely, not ending
- continuous queries

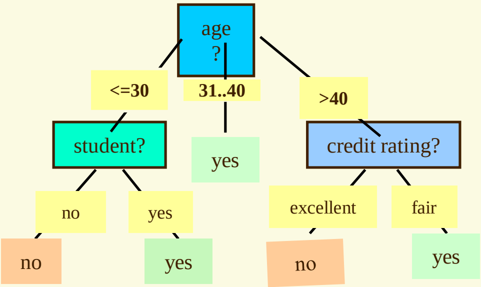


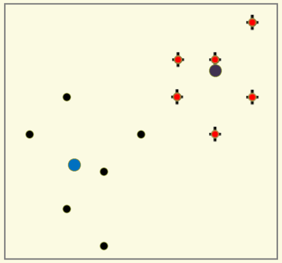
Big data: Special topics

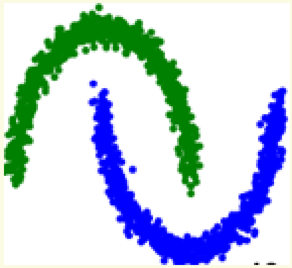
Data Mining and Graph Mining Basic

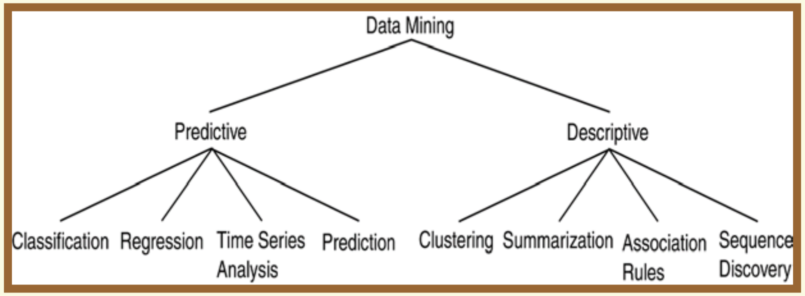


- Data mining: from data to knowledge
- Graph Mining: pattern mining
- Classification: decision tree
- Clustering: k-means, DBSCAN



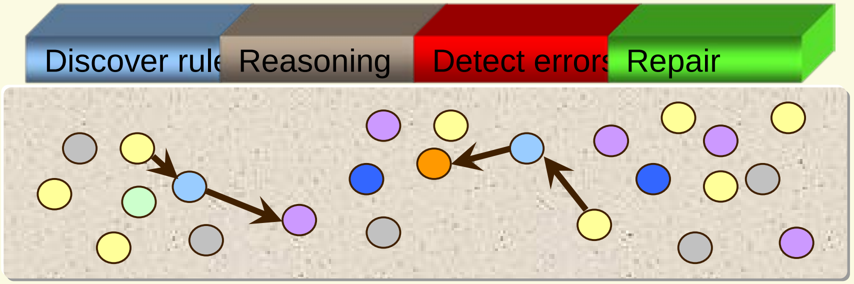






Data quality

- ✓ Data quality: The No.1 problem for data management
- ✓ Real life data are dirty, dirty data are costly
  - – The quest for a principled approach
  - – Critical issues:
    - Data consistency
    - Data accuracy
    - Entity resolution (record matching)
    - Information completeness
    - Data currency

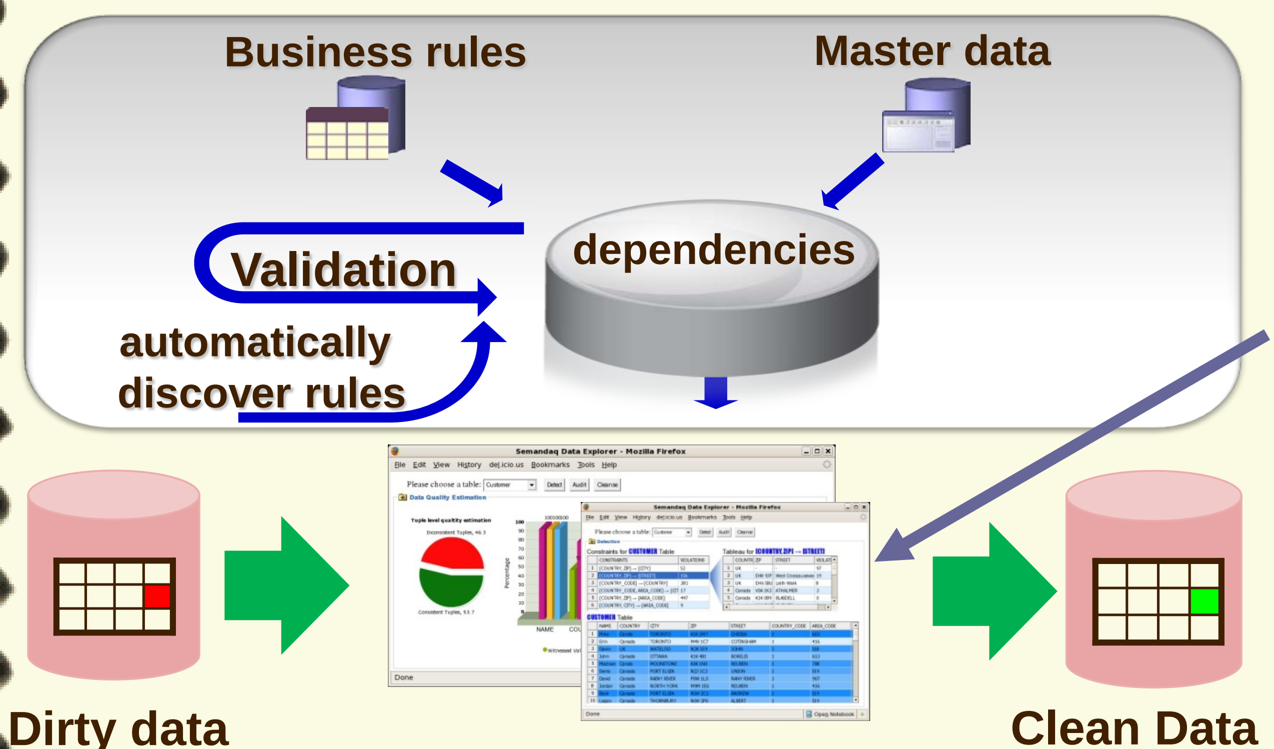


- ✓ Many challenges remain
  - – certain fixes (minimum user interaction), information completeness, data currency, Interaction between central issues of data quality

telecommunication, life sciences, finance, e-government, …

Data quality: A rich source of questions and vitality

A platform for improving data quality



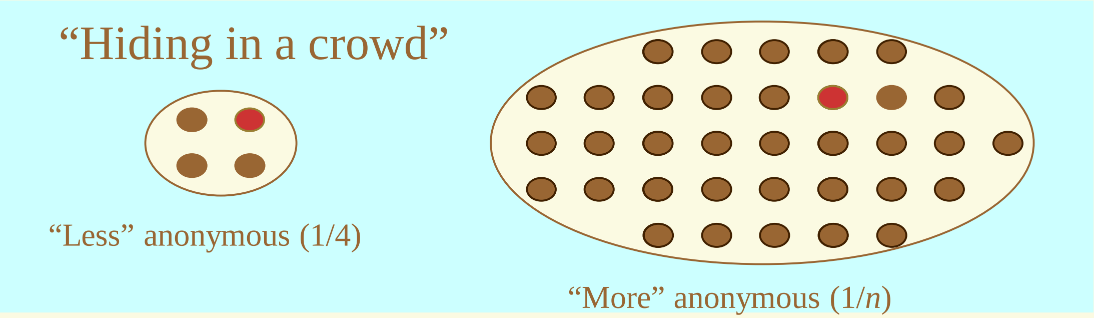
- profiling
- validating
- error detecting
- data repairing
- record matching
- certain fixes

- Standardization
- Auditing
- Enrichment
- Monitoring
- Data explorer

Develop practical data cleaning system

Data security and privacy

- Information Security: basic concepts
- Privacy: basic concepts and comparison with security
- K-anonymity, I-diversity & t-closeness



Future of Big Data and DBMS

The Beckman report

http://cacm.acm.org/magazines/2016/2/197411-the-beckman-report-on-database-research/fulltext

- Research challenges
  - Challenge 1：Scalable big/fast data infrastructures – parallel and distributed processing (volume)
    - Query processing and optimization (process monitoring)
      - Integrate data mining, sampling, machine learning
    - New hardware
    - Cost-efficient storage
    - High-speed data streams
    - Late-bound schemas
    - Consistency
    - Metrics and benchmarks

The Beckman report

http://cacm.acm.org/magazines/2016/2/197411-the-beckman-report-on-database-research/fulltext

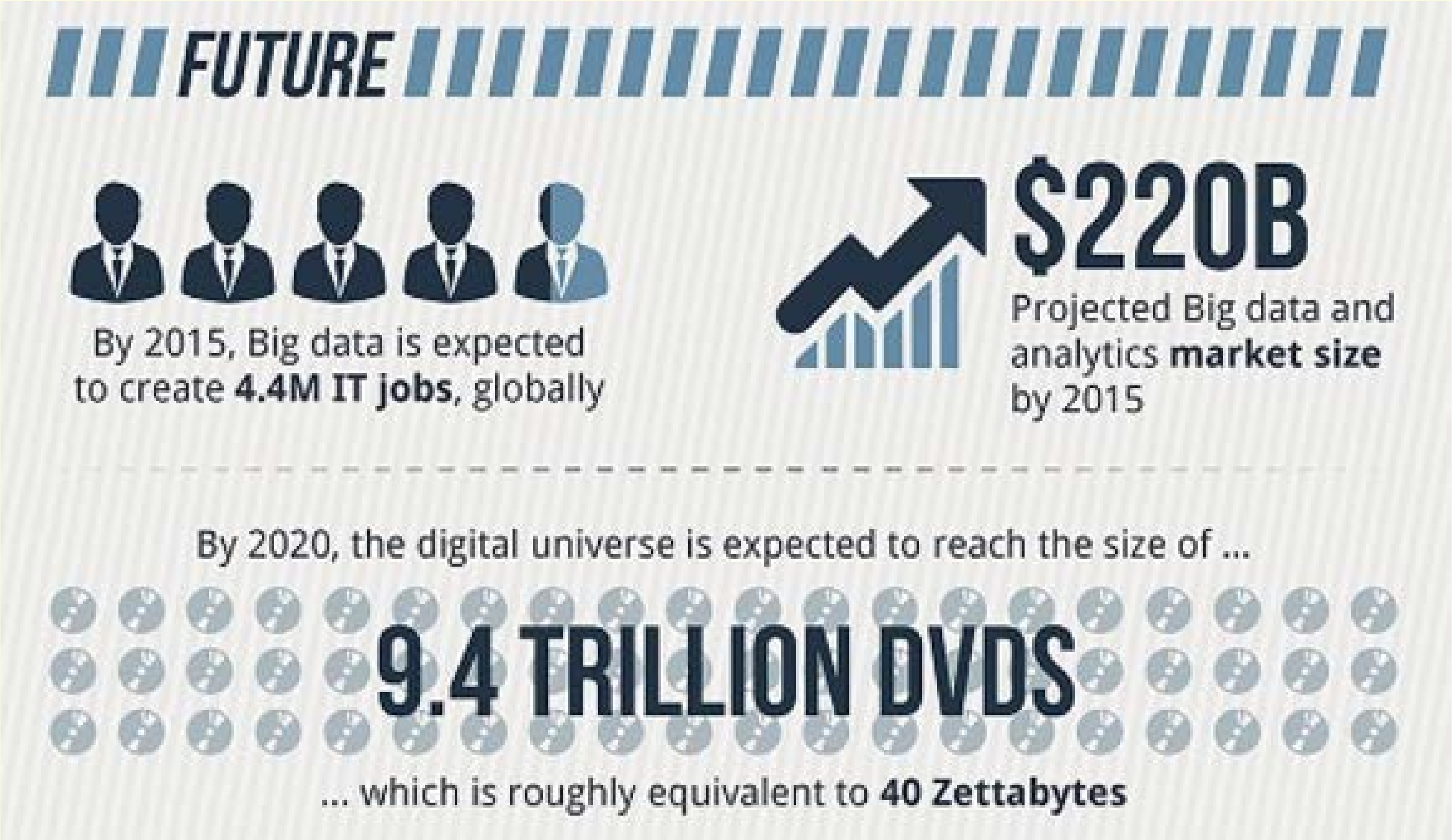
- Research challenges
  - Challenge 2: Diversity in data management
    - No-one-size-fits-all
    - Cross-platform integration
    - Programming models
    - Data processing workflows
  - Challenge 3: End-to-end processing of data
    - Data-to-knowledge pipeline
    - Tool-diversity and customizability
    - Open source
    - Understanding data/knowledge bases

The Beckman report

http://cacm.acm.org/magazines/2016/2/197411-the-beckman-report-on-database-research/fulltext

- Research challenges
  - Challenge 4: Cloud Service
    - Elasticity
    - Data replication
    - System administration and tuning
    - Multitenancy
    - Data sharing
    - Hybrid clouds (cyber-physical systems)
  - Challenge 5: Roles of humans in the data life cycle
    - Data producer (meta-data)
    - Data curators (crowdsourcing)
    - Data consumers (fuzzy queries)
    - Online communities (data community)

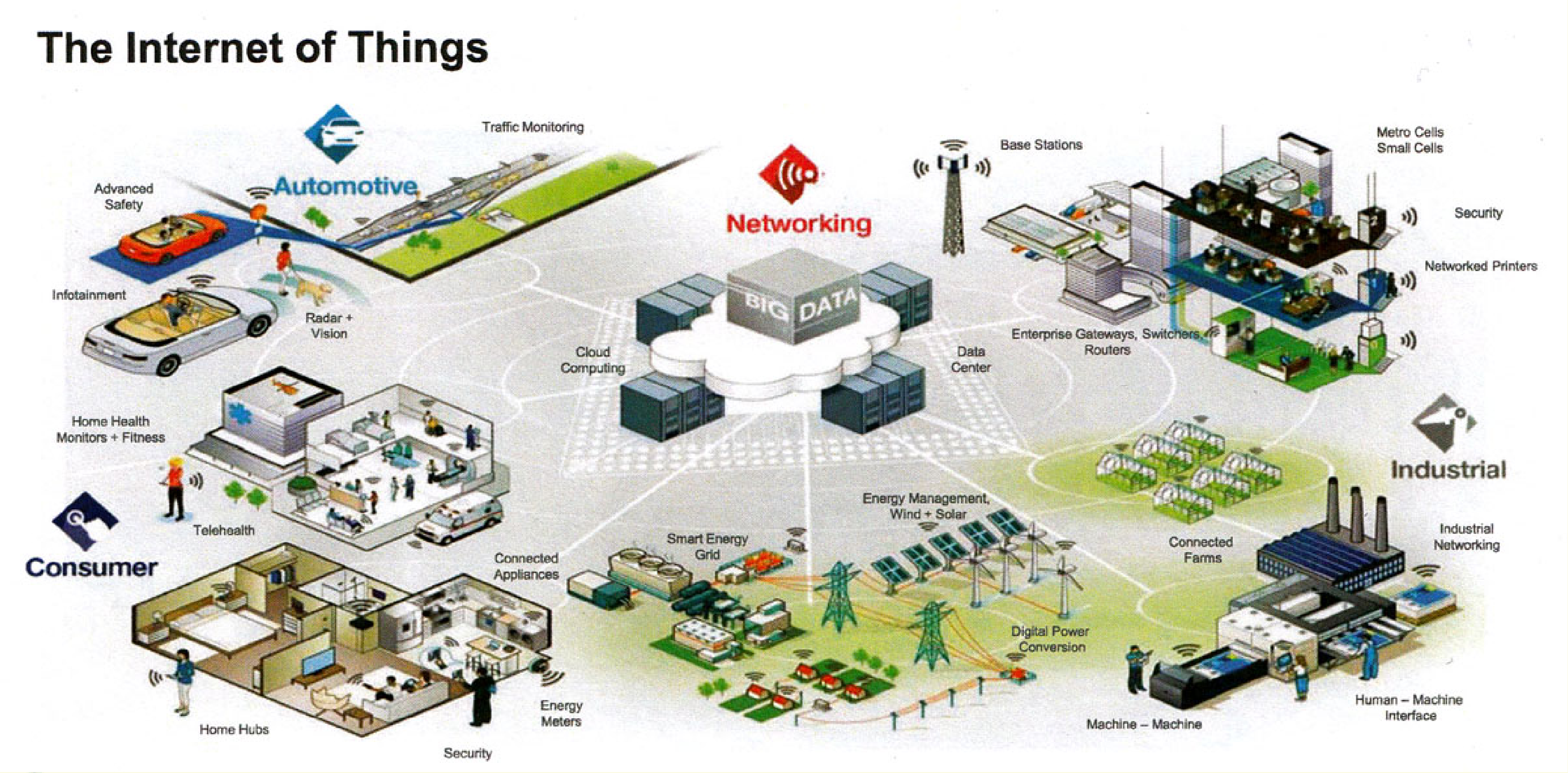
The future of Big Data



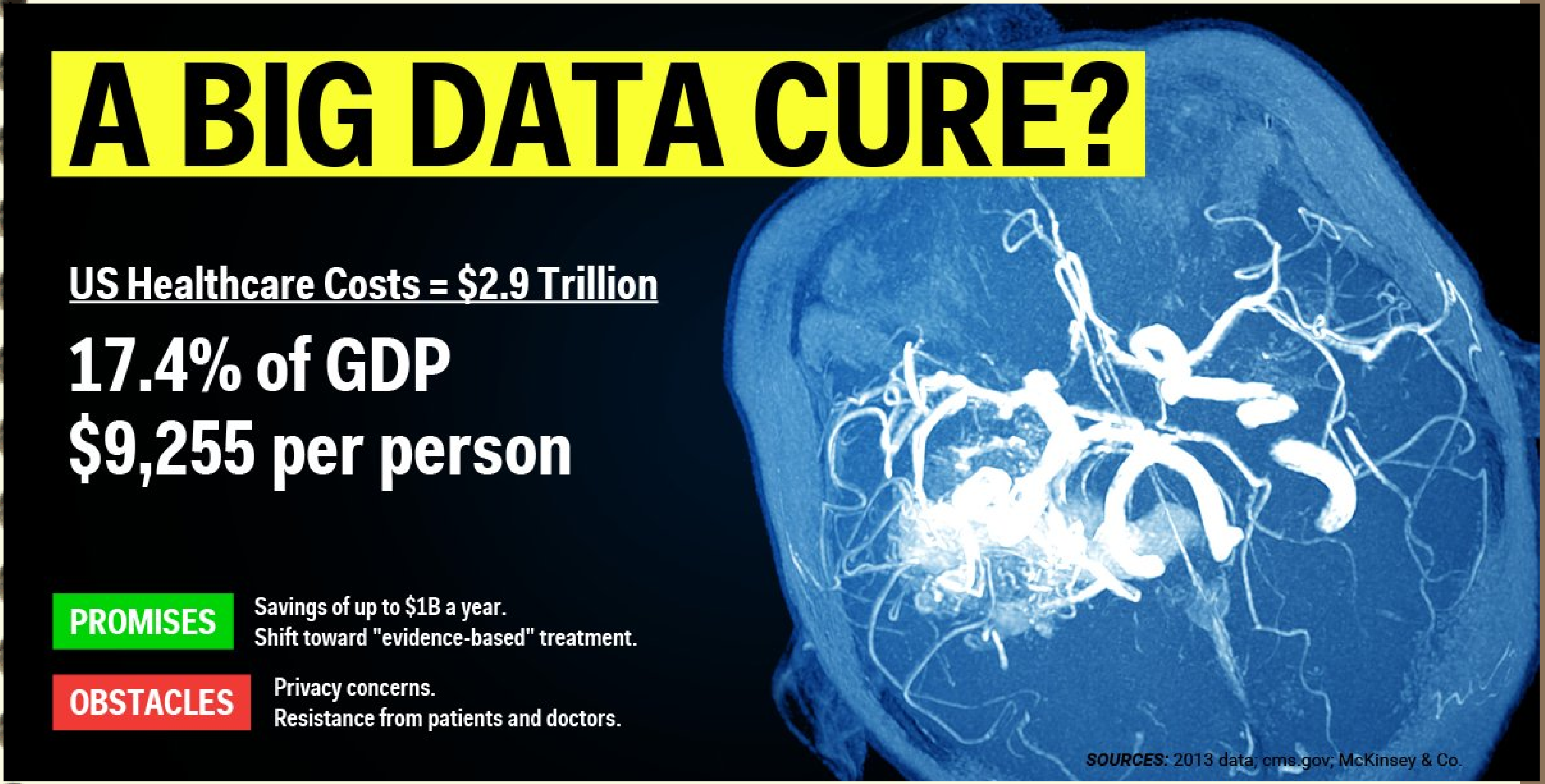
Machine BI

- Business Intelligence

Smart Living



Big data & Healthcare



http://www.healthcareitnews.com/category/resource-topic/government

Future of Big Data techs (NSF National Priorities)

NSF Big Ideas:
https://www.nsf.gov/news/mmg/mmg_disp.jsp?med_id=81537

Course project: presentation & report

Schedule of presentation
Dec 12.13: 4 – 5:30 pm, SLOA 161

Presentation Schedule (12th December)

Group1 : Srinivas Siddarth Vodnala
Project : ***Sentiment Analysis on Yelp dataset***

Group2: Hongyang Gao, Zhengyang Wang, Lei Cai.
Project: ***Node Classification using Graph Convolution***

Group3: Chin-Wei, Chang
Project: ***Multi-label Classification with the Column Subset Selection Problem Approach***

Group4: Aditi Deepak Thuse, Ankita Tanwa
Project: ***Flight Data Analysis***

Group5: Michael Antosz
Project: ***Airline Search Engine***

Group6: Ehdieh Khaledian
Project: ***Analyzing Relationship of Organisms and Proteins Using Graphs***

Group7: Nathan Scott, Joshua.R.Meyer
Project: ***Amazon Purchasing Recommendation Sysytem***

Group8: JingLin Tao
Project: ***Result matching of a food database***

Group9: Vishal Sonawane, Anirudh Rao
Project: ***IMDB dataset of reviews mining***

Presentation Schedule (13th December)

Group1 : Justin Jackson, Ryan Torelli
Project : ***Amazonco-purchase analysis***

Group2 : Hang Guo, Stefanie Watson, Jerdon Helgeson
Project : ***Mining the Association Rules of the Medical Costs in US Medicare Patients***

Group3: Arman Ahmed
Project: ***Cyber Physical Security Analytics for Transactive Energy Systems***

Group4: Kudart Kaur, Chih-Che Sun
Project: ***Intrusion Detection in Big Data using Machine Learning***

Group5: Xin Zhang
Project: ***Knowledge base search engine***

Group6: Insun Lee, Kim Nguyen, Chao Zheng
Project: ***Youtube Analyzer***

Group7: Jason Kramberger, Tyler Walker
Project: ***YouTube recommendation engine***

Group8: Matthew Green, Wyatt Fraley, and Kayl Coulston
Project: ***Flight Database***

Group9: Sheng Guang
Project: ***To be filled.***

Project presentation

- Presentation (8 minutes + 2-3 minutes Q&A)
  - Background and motivation
    - why the problem is important
    - application of the solutions
    - Challenges and difference with related work
  - Problem formulation:
    - Input and output
    - Object function, if any
  - Algorithm description
    - Correctness analysis
    - Complexity analysis
    - Properties/features/optimization techniques
  - Experimental study/demo
    - Data sets/generation of dataset
    - Algorithm implemented/baseline algorithms/platforms/test settings
    - Figures/trend/explain
    - Summary of experimental result
  - Conclusion and Future work
    - How your current work can be improved

General tips

- ✓ Talk is about idea
- ✓ Every talk motivates a single problem/solution
- ✓ Simple Slides are better
- ✓ A picture is worth a thousand words
- ✓ Keep logical flow
- ✓ Prepare for Questions
- ✓ Practice makes perfect

Course project report

- ✓ You have milestones 1-5. Combine your milestones and enrich with experimental results, references and future work to a complete, comprehensive report. Do not simply glue them together.
- ✓ Make title concise and right to the point
- ✓ Abstract: describe your problem, solution and experimental result.
- ✓ Section 1: Introduction:
  - Background
  - challenges (why your problem is hard)
  - Related work
- ✓ Section 2: Problem statement
  - Input/output
  - Object function
  - Hardness of the problem
- ✓ Section 3: Solution
  - Algorithm description
  - correctness and time complexity analysis (try big O notation)
  - Optimization techs (e.g. , applying Big data search strategy)

Course project report

- ✓ Section 4: Experimental study
  - Data sets/generation of dataset
  - Algorithm implemented/baseline algorithms/platforms/test settings
  - Figures/trend/explain
  - Summary of experimental result
- ✓ Section 5: Conclusion & Future work
  - What have you observed in your project
  - What problem remains to be unresolved? What are possible extension of your problem? What’s your plan to solve it in future?

Submit your course project report to your TA as a single pdf,
with name “CPTS415_”+your firstname+”_report”.
Due date: 11:59 pm, Dec 16.

CPT_S 415
Big Data