IST 3420: Introduction to Data Science and Management

Langtao Chen, Fall 2017

1. Introduction to Data Science

Agenda

- What is data science?
- Big data
- Work as a data scientist
- Ethical conduct in data science

Walmart Case: What They Know About You

- Event
  - Hurricane Frances was threatening Florida's Atlantic coast
- Intuition
  - The demand for flashlights would increase in local stores
- Data Science Approach
  - Use predictive technology to see what's going to happen based on historical data
- Findings (Hidden Patterns)

"We didn't know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane," Ms. Dillman said in a recent interview. "And the pre-hurricane top-selling item was beer."

What is Data Science?

- https://en.wikipedia.org/wiki/Data_science

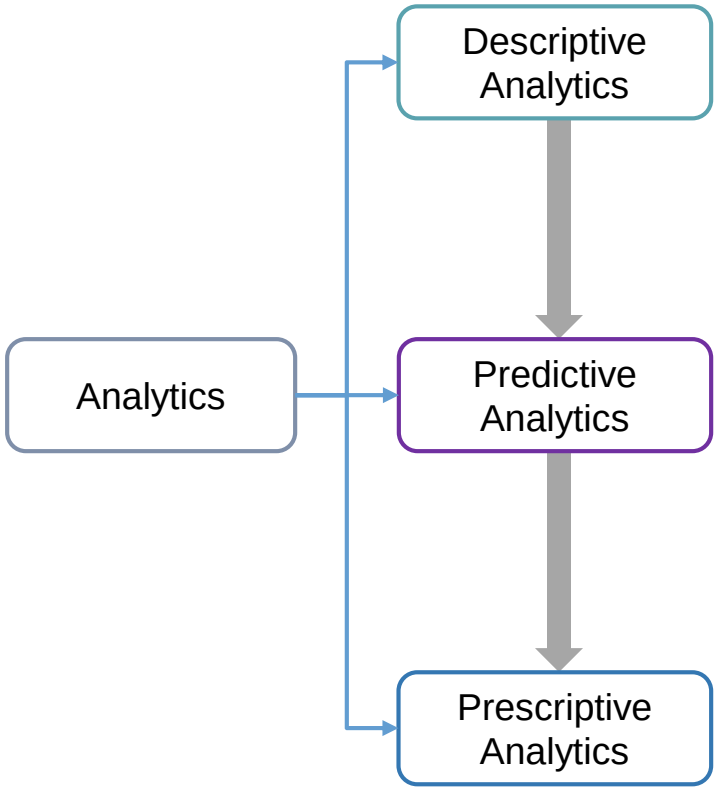
> Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).

- "Data" means dealing with every aspects of data such as collecting data, cleansing data, transforming data, and analyzing data.
- "Science" means extracting knowledge through systematic approach
- A briefer definition: "Data science is the study of the generalizable extraction of knowledge from data." (Dhar 2013)

Data Science vs. Business Analytics

- The boundary between Data Science and Business Analytics is blurry.
- For a comparison, refer to https://onlinebusiness.american.edu/blog/comparing-analytics-data-science/
- In this course, we use the two terms interchangeably.

Overview of Analytics (INFORMS taxonomy)



What happened?

- Prepares and analyzes ***historical*** data
- Identifies patterns from samples for reporting of trends

- Visualization
- Report
- Dashboard
- …

What will happen?

- Predicts future probabilities and trends
- Finds relationships in data that may not be readily apparent with descriptive analysis

- Data mining
- Forecasting
- Statistical analysis
- …

How can we make it happen?

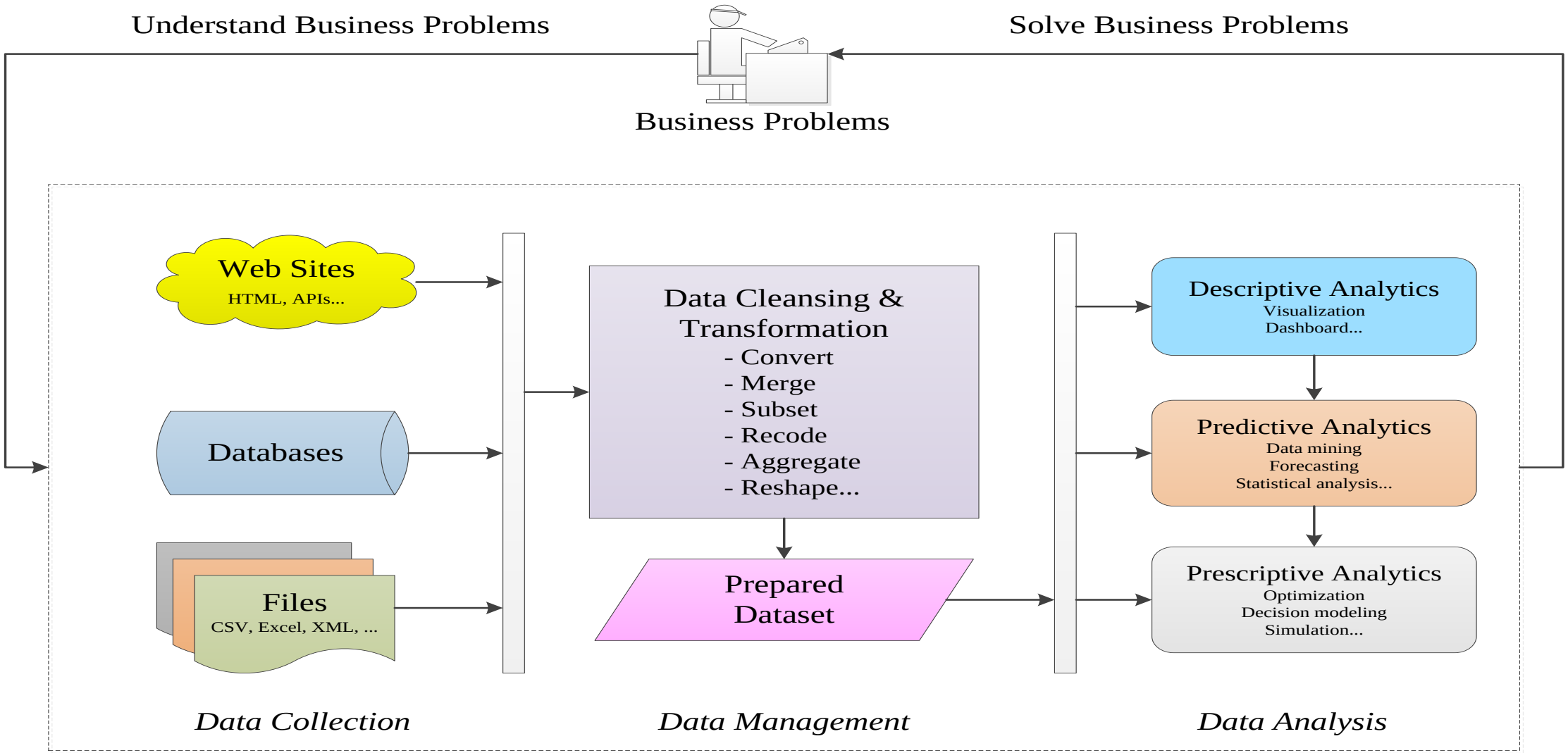
- Evaluates and determines new ways to operate
- Targets business objectives
- Balances all constraints

- Optimization
- Simulation
- Decision modeling
- …

The Institute for Operations Research and the Management Sciences (INFORMS) is the largest society in the world for professionals in the field of operations research (O.R.), management science, and analytics.

Source: https://www.informs.org/Community/Analytics

Data Science/Analytics Procedures



Data Science/Analytics Tools

- Open source

- Commercial

Agenda

- What is data science?
- Big data
- Work as a data scientist
- Ethical conduct in data science

The Era of Big Data

- Big data
  - Data sets with characteristics beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

- 4Vs Model

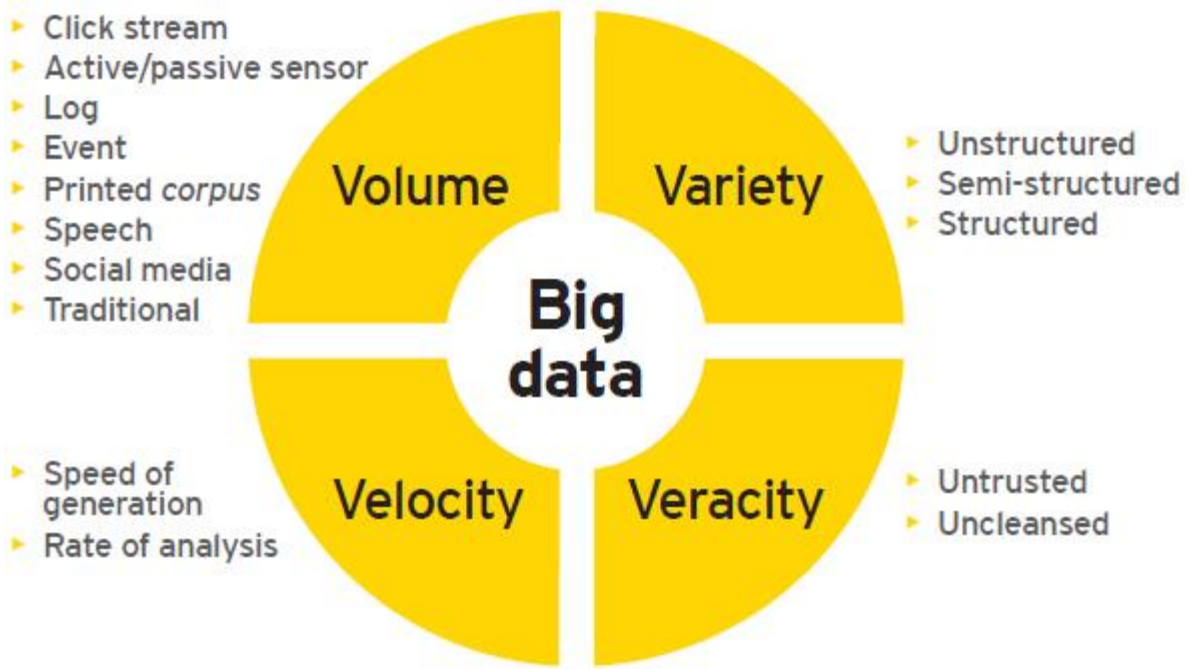
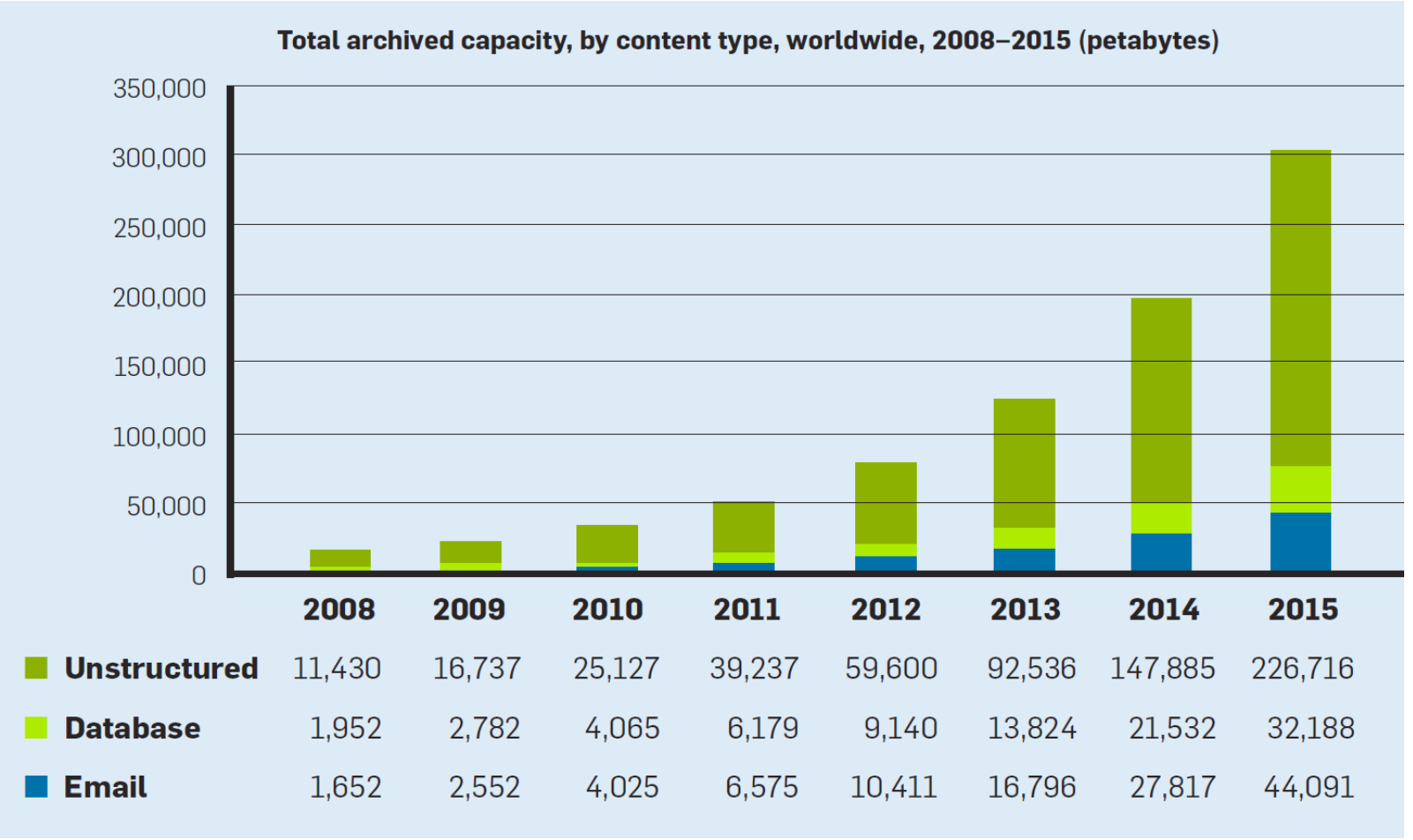


Image source: http://www.ey.com/GL/en/Services/Advisory/EY-big-data-big-opportunities-big-challenges

Projected growth of unstructured & structured data (Dhar 2013)



	2008	2009	2010	2011	2012	2013	2014	2015
Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
Email	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

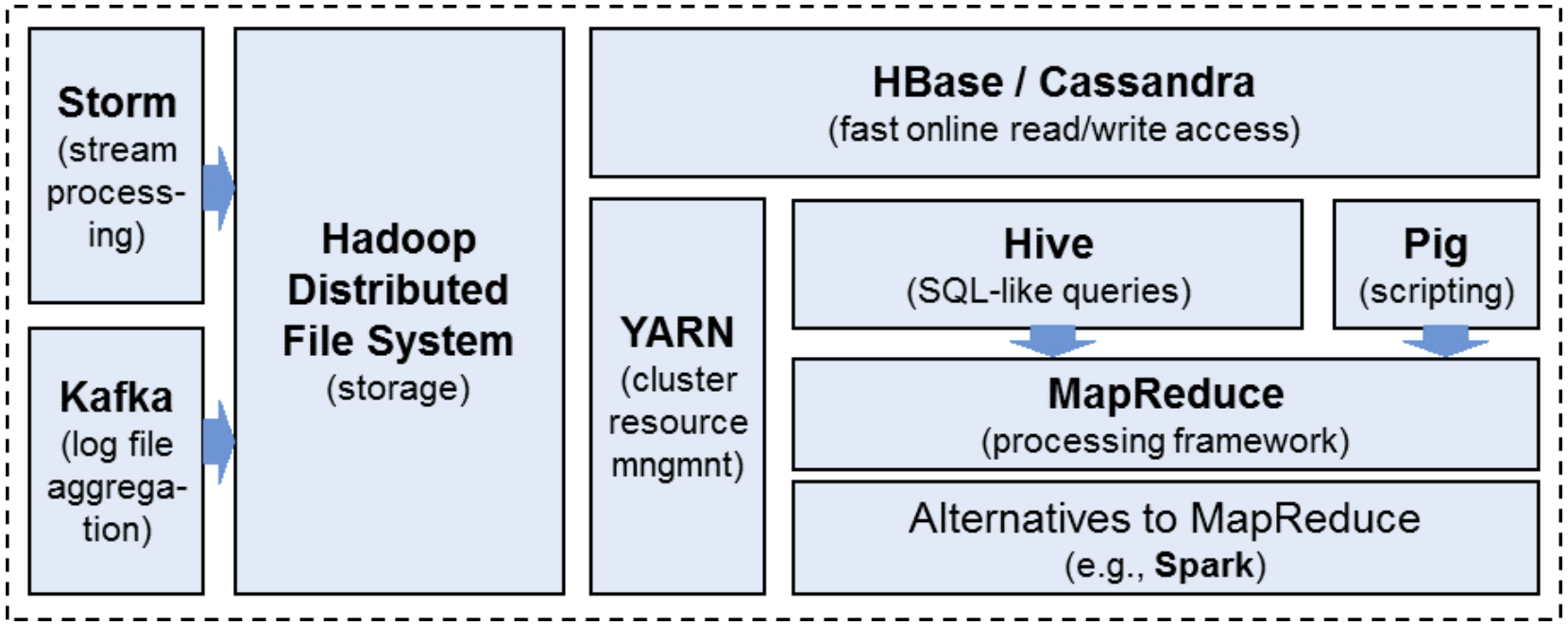
Limitations of Relational DBMS

- RDBMS
  - Bases on E.F. Codd model
  - Provides good support for ACID (Atomicity, Consistency, Isolation and Durability)

- Limitations to deal with big data
  - Scalability: expensive and hard to scale
  - Complexity: data fit into a predefined schema (table structure)
  - Performance/Speed: data join is very time consuming

Big Data Ecosystem

Processing <=> Storage <=> Analytics

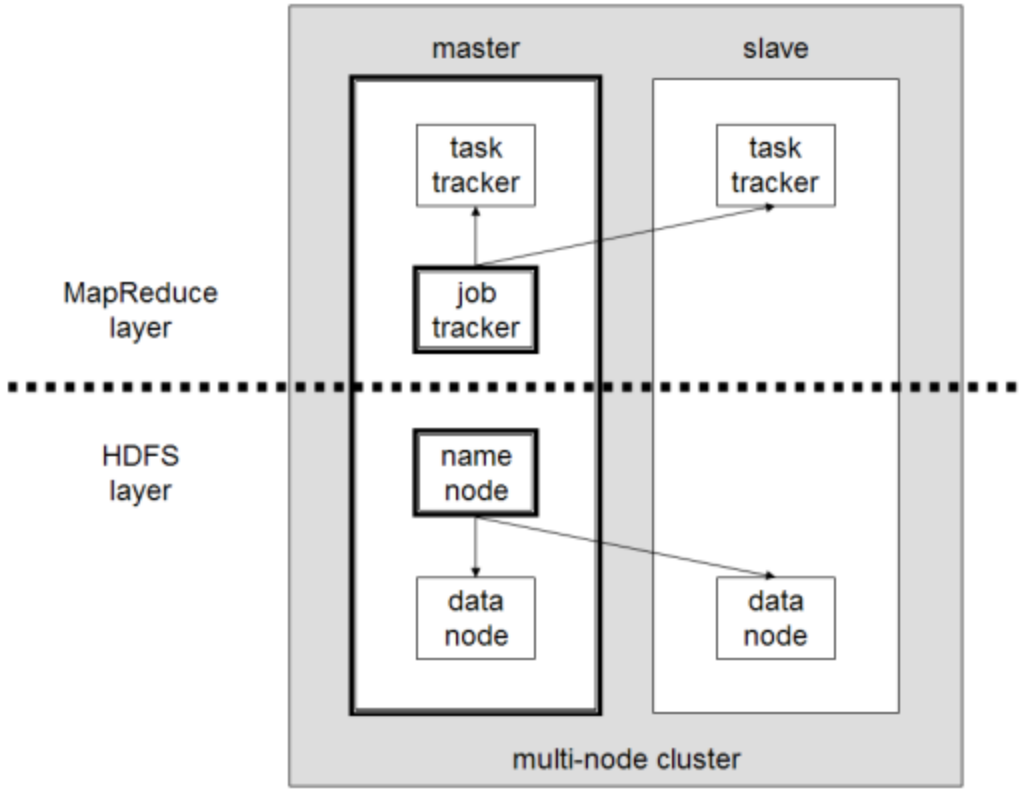


For a complete list of technologies, refer to
http://bigdata.andreamostosi.name/

Source: http://blogs.gartner.com/martin-kihn/what-marketers-should-know-about-big-data/

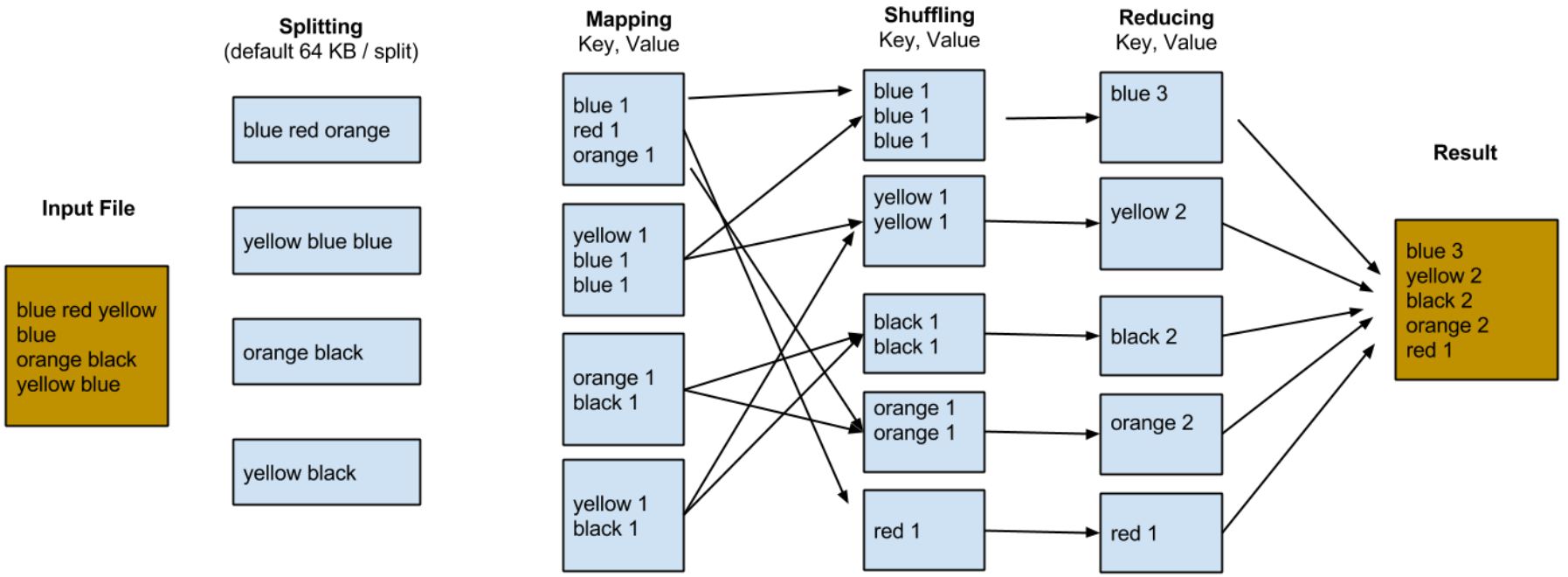
Hadoop

- An open-source framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware
- Based on the idea of Google File System
- Consists of two main layers
  - Hadoop distributed file system (HDFS)
  - MapReduce: processing engine



Hadoop Word Count Example

- MapReduce = Map + Reduce



For the programming detail of the word count example, refer to
http://wiki.apache.org/hadoop/WordCount

Image source: http://louaialarabi.blogspot.com/

Hadoop Design Principles

- Process big data in parallel
- Run on cheap commodity hardware
- Be fault-tolerant and robust (self-healing)

Doug Cutting and Hadoop the elephant

The name, on the other hand, is a homey story going back 10 years - into the realm of a toddler's experimentation with old-fashioned human language. Cutting's son, then 2, was just beginning to talk and called his beloved stuffed yellow elephant "Hadoop" (with the stress on the first syllable).

"Hadoop: Toddler Talk Provides Big Data Name"
http://www.cnbc.com/id/100769719

Video: What is Hadoop?

https://www.youtube.com/watch?v=RQr0qd8gxW8

Agenda

- What is data science?
- Big data
- Work as a data scientist
- Ethical conduct in data science

Data Scientist: the "Sexist" Job of the 21st Century

- Top 1 in 25 best jobs in America (2016)

United States | 2016

1

Data Scientist

Job Openings	1,736
Median Base Salary	$116,840
Career Opportunity	4.1
Job Score	4.7

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

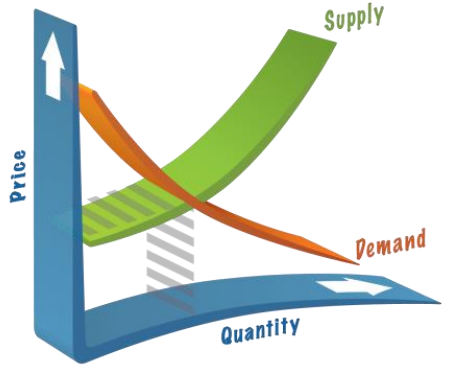
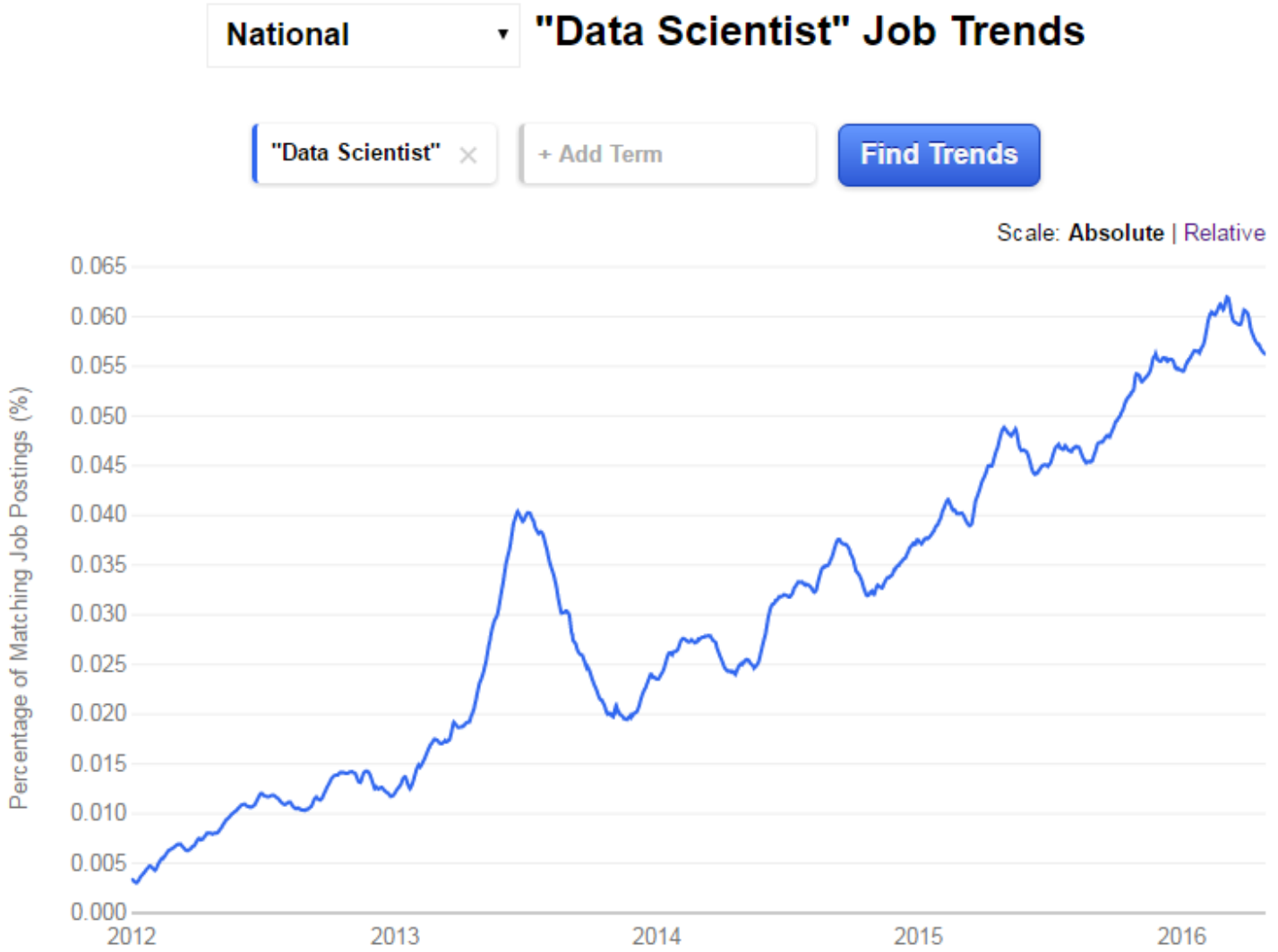


Image source
http://precision-market-research.com/supply-demand/

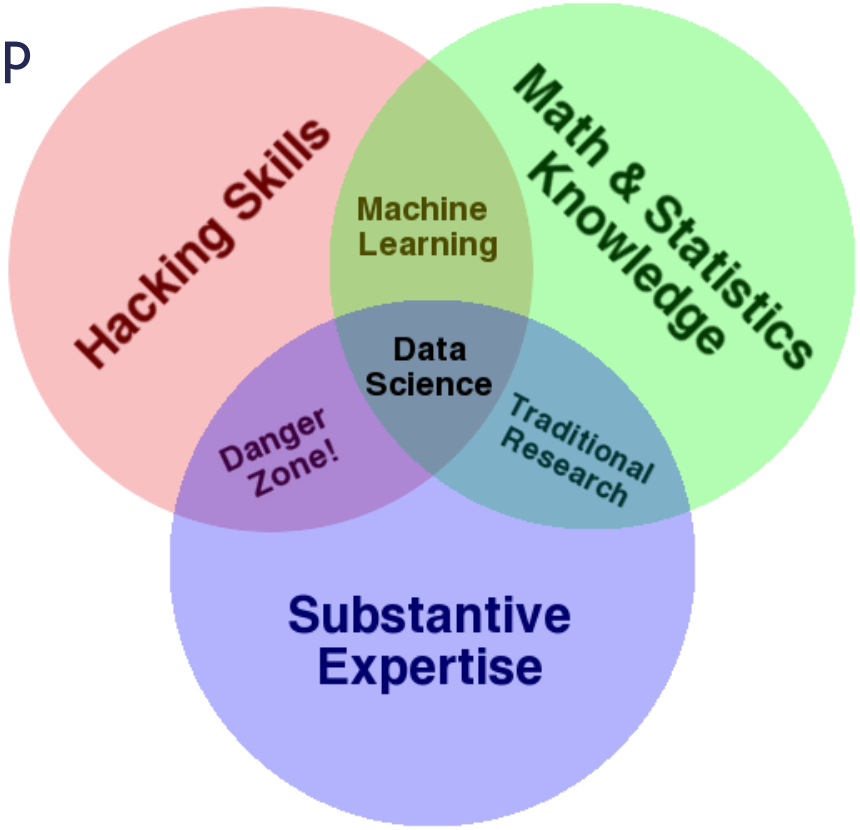
Data Scientist Job Trend



Source: http://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html

Skill Set for BA/Data Science

- Hacking Skills
  - Programming (R and Python), data structures, markup languages, algorithms, database, cloud computing, big data etc.
- Math and Statistics Knowledge
  - Probabilities, distributions, hypothesis testing, multivariate analyses, econometrics etc.
  - Prediction, classification, clustering, text mining etc.
- Curiosity and Expertise
  - Academic curiosity, domain knowledge, storytelling, visualization, product sense etc.



A Sample Job Description

Data Scientist & Analytics Developer

Lancer Insurance Company - Long Beach, NY

Lancer Insurance Company is looking for a talented and motivated data scientist and analytics developer to help provide business insight from our data assets and provide new algorithms to solve data problems. The ideal candidate will demonstrate skills in R, Python, SQL, ODBC, JDBC, REST, JSON, and Excel. This individual will be expected to learn and understand the information needs of an insurance company and develop for the current and future data integrations and intelligence infrastructure. The right candidate will also be a successful communicator and self-starter with strong analytical skills who exhibits creativity, hustle, integrity, and teamwork.

Duties and Responsibilities :

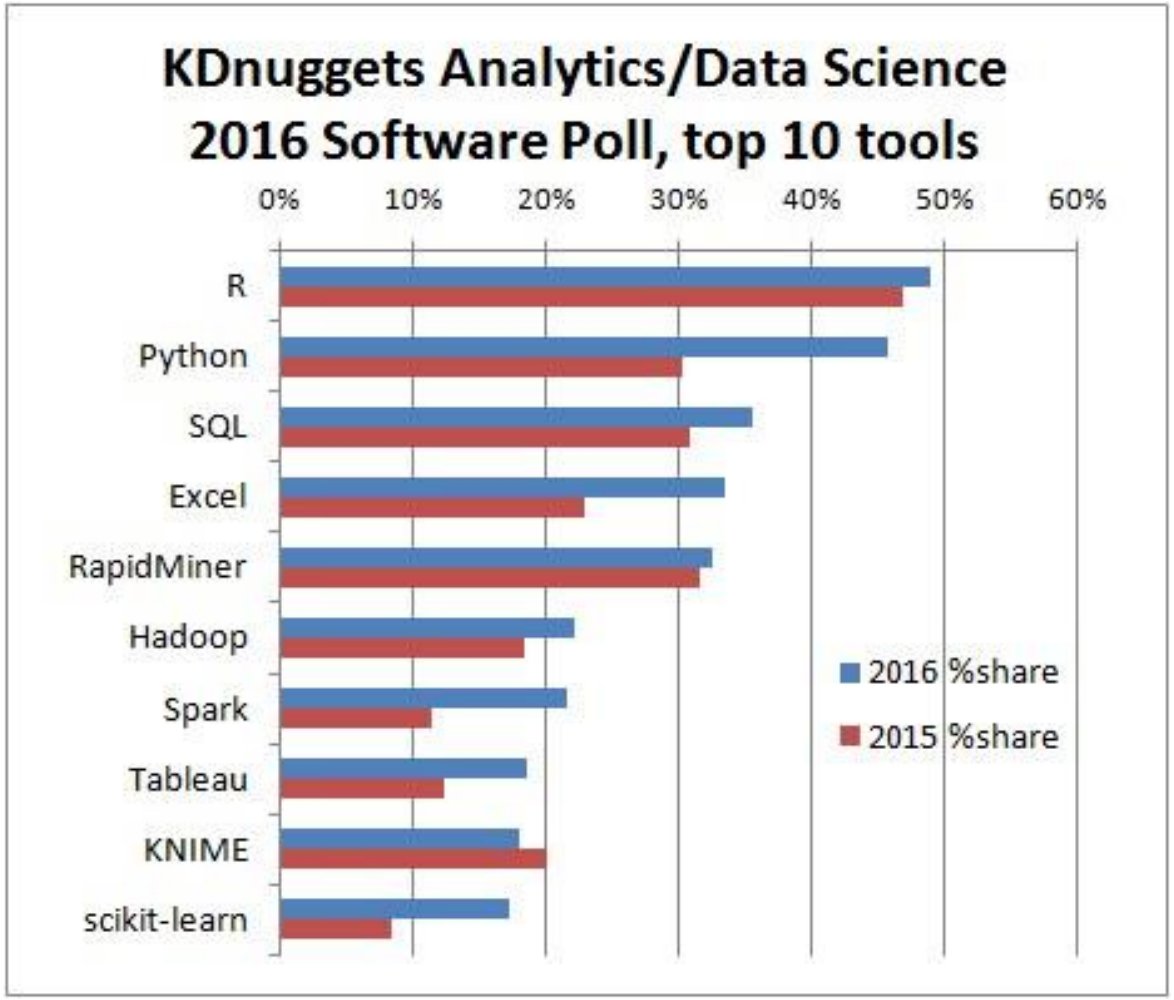
- Develop required analytic projects in response to business needs
- Contribute to data mining architectures, modeling standards, reporting and data analysis methodologies.
- Collaborate with stakeholders to integrate data mining results with existing systems.
- Enforce data integrity, security, and stewardship.
- Provide data structure, integration and intelligence solutions
- Monitor data mining system performance and implement efficiency improvements.
- Craft quality code and solutions for maintainability and extensibility

Skills & Experience:

- A Bachelor's degree (or equivalent experience) in a math or computer science program is required. Recent graduates are encouraged to apply.
- Ability to effectively manage multiple competing priorities at any given time
- Detail-oriented and ability to work collaboratively in a deadline-driven environment
- Excellent communication skills
- Strong relational data modeling experience
- Mathematics, statistics, correlation, data mining, predictive analysis skills
- Scripting and/or high level programming language experience
- R or other statistical package experience
- Data visualization skills
- Insurance experience a major plus

http://www.indeed.com/cmp/Lancer-Insurance-Company/jobs/Data-Scientist-Analytic-Developer-60daa4cc2519aa23 [Access on 6/26/2016]

Top Tools for Data Science/Analytics



http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

Agenda

- What is data science?
- Big data
- Work as a data scientist
- Ethical conduct in data science

Ethical Issues in Data Science

- Business case: Target's customer tracking technology
  - "How Companies Learn Your Secrets". Refer to http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=2&hp=&pagewanted=all

Standards of Professional Conduct

- Legal, policy, and ethical concerns can arise in any steps of data science including data collection, storage, and analysis

- Some organizations such as Data Science Association start to create standards of conduct
  - http://www.datascienceassn.org/code-of-conduct.html

Reference

- Davenport, T.H., and Patil, D.J. 2012. "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review* (90:10), pp. 70-76.
- Dhar, V. 2013. "Data Science and Prediction," *Communications of the ACM* (56:12), pp. 64-73.
- Provost, F., and Fawcett, T. 2013. "Data Science and Its Relationship to Big Data and Data-Driven Decision Making," *Big Data* (1:1), pp. 51-59.

Q & A