Introduction to Data Science

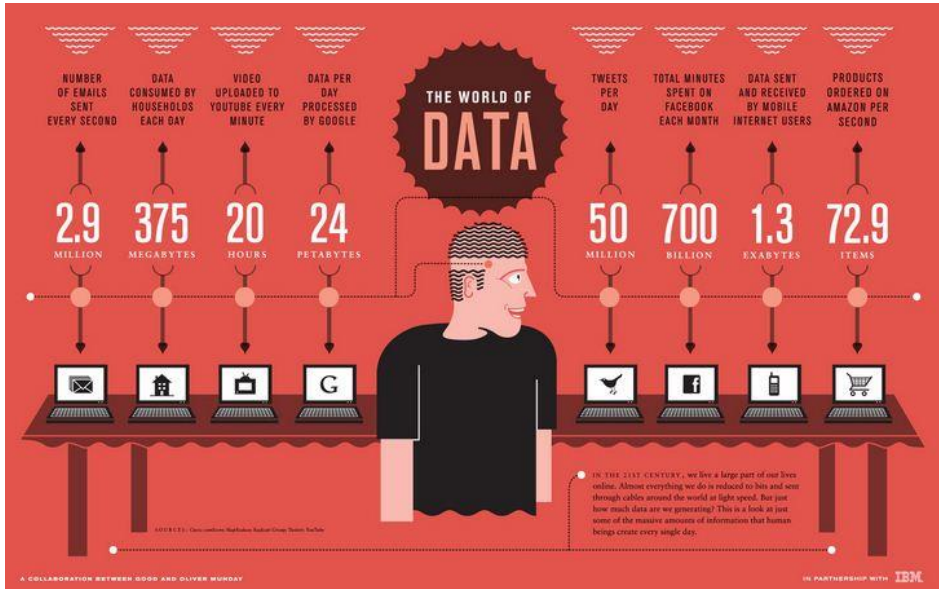
Data All Around

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network

How Much Data Do We have?

- Google processes 20 PB a day (2008)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 1000 genomes project: 200 TB

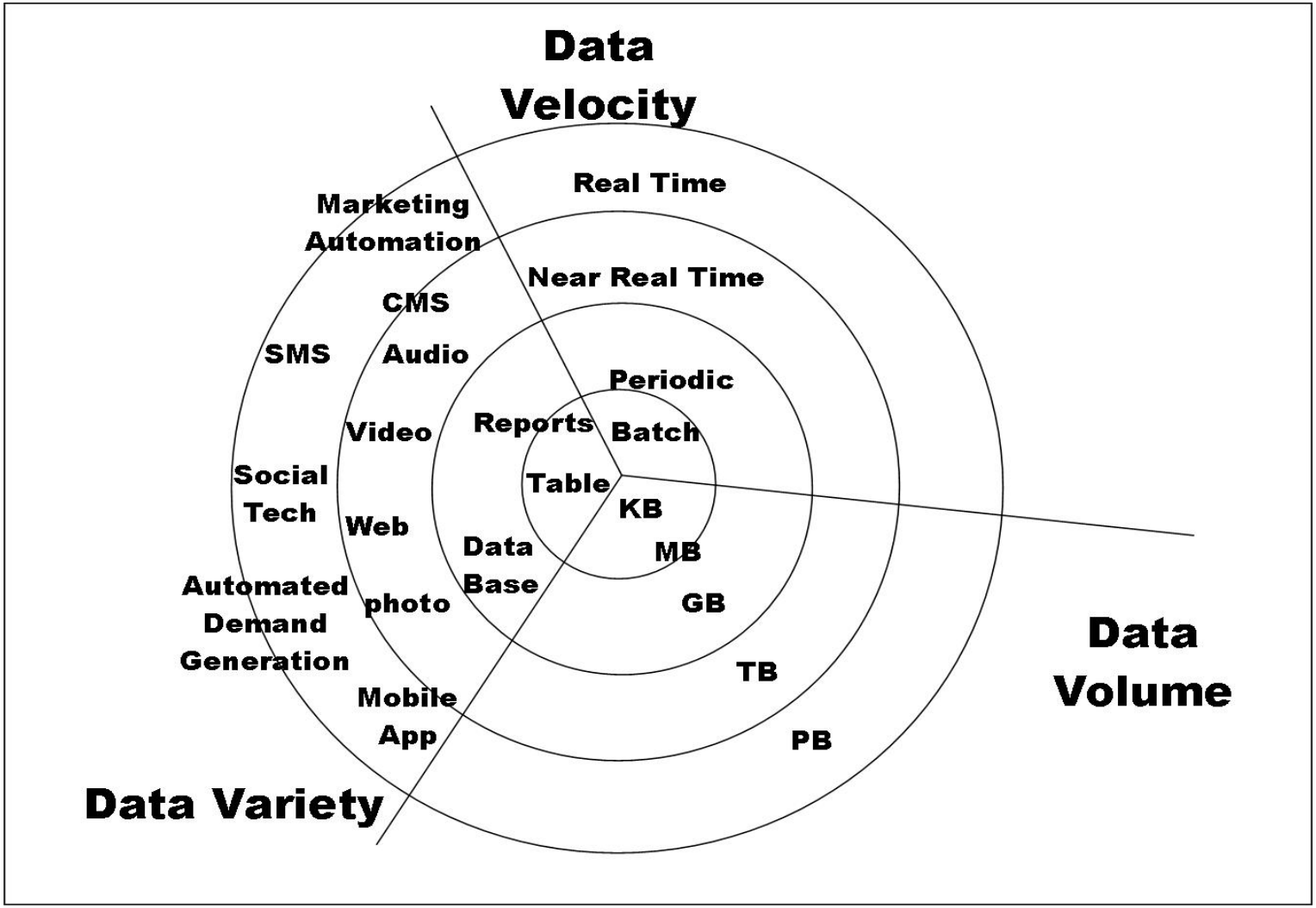
- Cost of 1 TB of disk: $35
- Time to read 1 TB disk: 3 hrs (100 MB/s)



Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity
  - Variety and Complexity
    - the diversity of sources, formats, quality, structures.

Big Data


Data
Velocity
Real Time
Near Real Time
Periodic
Batch
Marketing
Automation
CMS
SMS
Audio
Video
Reports
Social
Tech
Table
KB
Web
Data
Base
MB
Automated
Demand
Generation
photo
GB
TB
Mobile
App
PB
Data
Volume
Data Variety

Types of Data We Have

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- You can afford to scan the data once

What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Big Data and Data Science

- "… the highly demand job in the next 10 years will be statisticians," Hal Varian, Google Chief Economist
- The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute's June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - One proposal (elsewhere) for an MS in "Big Data Science"

What is Data Science?

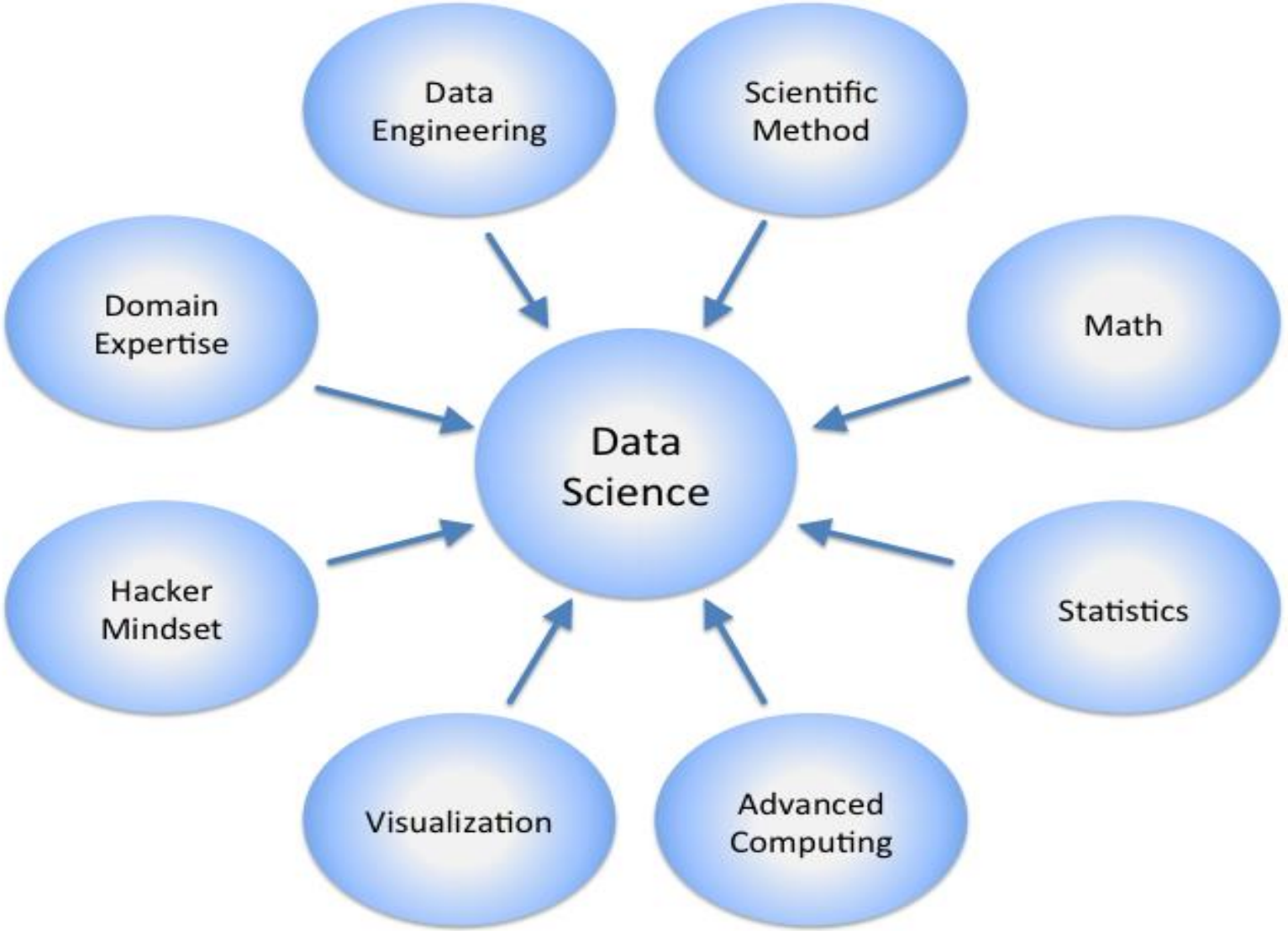
- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

What is Data Science?

- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

Data Science



Data Scientists

- Data Scientist
  - The highest paid Job of the 21st Century
- They find stories, extract knowledge. They are not reporters

Data Scientists

- Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions

What do Data Scientists do?

- National Security
- Cyber Security
- Business Analytics
- Engineering
- Healthcare
- And more ….

Concentration in Data Science

- Mathematics and Applied Mathematics
- Applied Statistics/Data Analysis
- Solid Programming Skills (R, Python, Julia, SQL)
- Data Mining
- Data Base Storage and Management
- Machine Learning and discovery

What is data analytics?

- The term data analytics refers to the process of examining datasets to draw conclusions about the information they contain.
- Data analytic techniques enable you to take raw data and uncover patterns to extract valuable insights from it.
- Data Scientists and Analysts use data analytics techniques in their research, and businesses also use it to inform their decisions.
- Data analysis can help companies better understand their customers, evaluate their ad campaigns, personalize content, create content strategies and develop products. Ultimately, businesses can use data analytics to boost business

4 Ways to Use Data Analytics

1. **Improved Decision Making**

- Data analytics eliminates much of the guesswork from planning marketing campaigns, choosing what content to create, developing products and more.

2.More Effective Marketing

- When you understand your audience better, you can market to them more effectively.

3.Better Customer Service

Data analytics provide you with more insights into your customers, allowing you to tailor customer service to their needs, provide more personalization and build stronger relationships with them.

4.More Efficient Operations you.

Data analytics can help you streamline your processes, save money and boost your bottom line. When you have an improved understanding of what your audience wants, you waste less time on creating ads and content that don't match your audience's interests.

- *A successful data analytics practice can—should—provide a better strategy for where your business can go. When done well, data analytics can help*
- *Find trends*
- *Uncover opportunities*
- *Predict actions, triggers, or events*
- *Make decisions*

Processes in data analytics

- **Collecting and ingesting the data**
- **Categorizing the data** into structured/unstructured forms, which might also define next actions
- **Managing the data**, usually in databases, data lakes, and/or data warehouses
- **Storing the data** in hot, warm, or cold storage
- **Performing ETL** (extract, transform, load)
- **Analyzing the data** to extract patterns, trends, and insights
- **Sharing the data** to business users or consumers, often in a dashboard or via specific storage

What is data analysis?

- Data analysis consists of cleaning, transforming, modeling, and questioning data to find useful information.
- When you're done analyzing a dataset, you'll turn to other data analytics activities to:
- Give others access to the data
- Present the data (ideally with data visualization or storytelling)
- Suggest actions to take based on the data

Type of data analysis

- **Text analysis.** This is also referred to as Data Mining. This method discovers a pattern in large form data sets using databases or other data mining tools.
- **Statistical analysis.** This analysis answers “What happened?” by utilizing past data in dashboard form. Statistic analysis involves the collection, analysis, interpretation, presentation, and modeling of data.
- **Diagnostic analysis.** This analysis answers “Why did it happen?” by seeking the cause from the insights discovered during statistical analysis. This type of analysis is beneficial for identifying behavior patterns of data.
- **Predictive analysis.** This analysis suggests what is likely to happen by utilizing previous data. The predictive analysis makes predictions about future outcomes based on the data.

Application of Data Science

- Banking
- Finance
- Manufacturing
- Transport
- Healthcare
- E-Commerce