CS639: Data Management for Data Science

Lecture 1: Intro to Data Science and Course Overview

Theodoros Rekatsinas


data is the new oil
BIG DATA
CLOUD
SECURITY
DEVELOPERS
AZURE
APPS
PLATFORM
WEB
SOCIAL
PEOPLE
memory
DEVICES
MAGIC

NATIONAL CANCER INSTITUTE
GENOMIC DATA COMMONS

ELIXIR Innovation and SME Forum: Data-driven innovation in food, nutrition and the microbiome

Big science is data driven.

Increasingly many companies see themselves as **data driven**.

Even more “traditional” companies…

DIGITAL INDUSTRIAL COMPANY

Bloomberg Businessweek: How GE Became A 124-Year-Old Startup

The digital power plant is one of Industrial Internet applications. Image credit: GE Power

The story tracks GE’s digital transformation from its inception after the financial crisis in 2008. It started with a broad idea. “I said, ‘Look, we need to start building analytic capability, big data capability, and let’s do it in California,’” Immelt told the magazine. “That was as sophisticated as my original thinking was.”

The world is increasingly **driven by data…**

This class teaches **the basics** of how to use & manage data to obtain useful insights.

Today's Lecture

1. Motivation, Admin, & Setup

2. Introduction to Data Science

1. Motivation, Admin & Setup

What you will learn about in this section

1. Motivation for studying Data Science

2. Administrative structure

3. Course logistics

Data Analysis has always been around

1935: "The Design of Experiments"

R.A. Fisher

"Correlation does not imply causation"

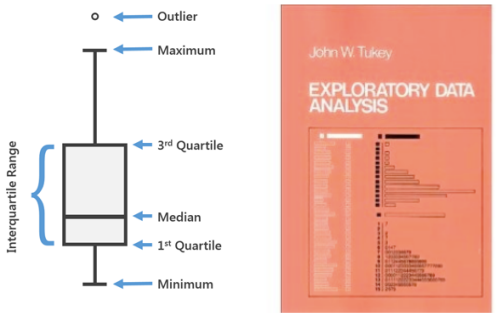
1958: "A Business Intelligence System"

Peter Luhn

A pioneer in hash coding and full text processing
Coined the term "business intelligence"

Data Analysis has always been around

1977: "Exploratory Data Analysis"



John Tukey

Introduced the box-plot
Also introduced the term "bit" (later used by Shannon)

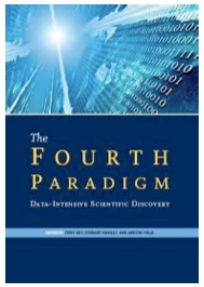
1997: "Machine Learning"

Tom Mitchell

The time of data-driven AI

Data Analysis has always been around

2007: "The Fourth Paradigm"



By Jim Gray

Essays on Scientific Discovery based on data-intensive science
Father of ACID
(requirements for reliable transaction processing)

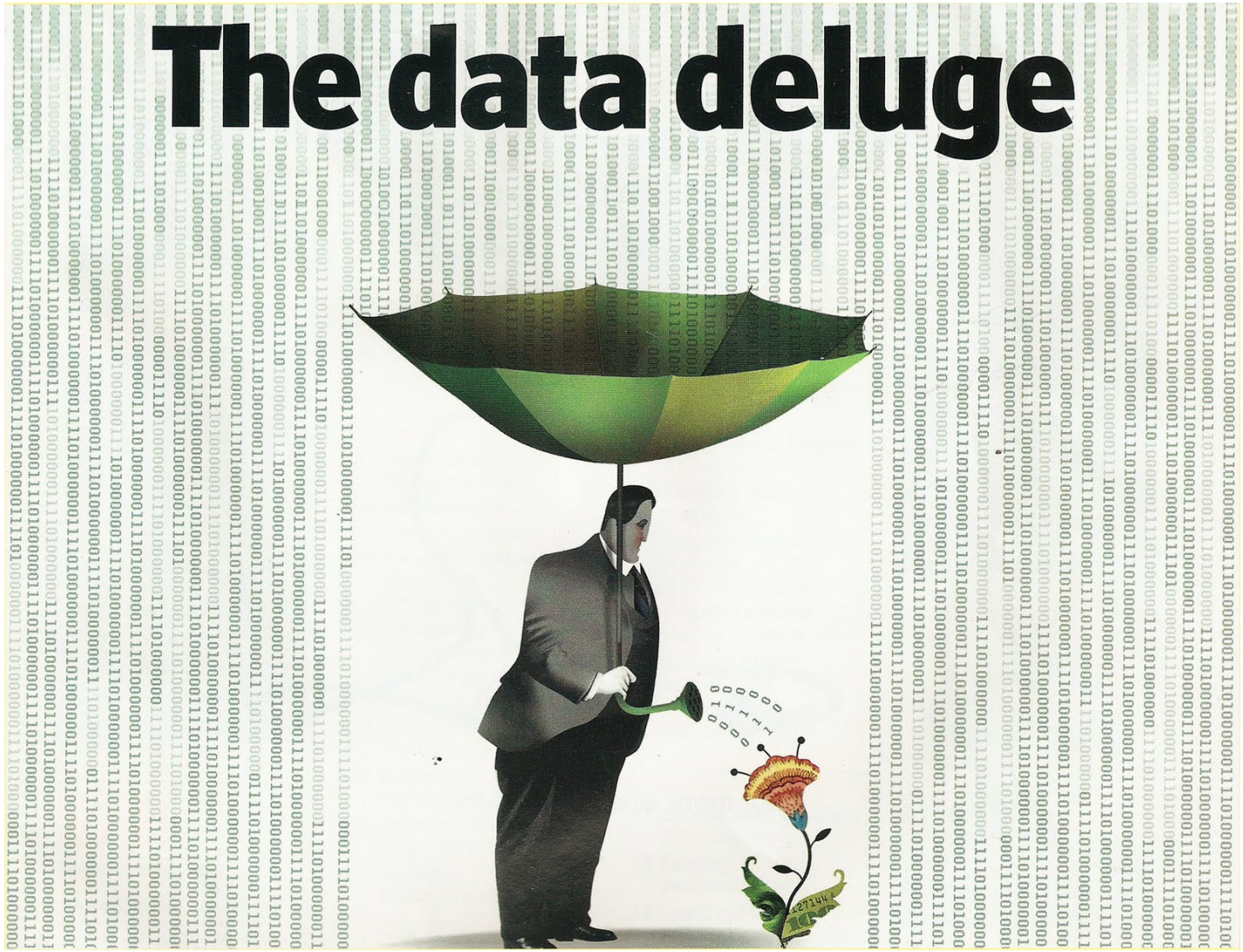
2009: "The Unreasonable Effectiveness of Data"

Peter Norvig

Known for AI programming
(at Google)

Data Analysis is popular

The Economist



Why should you study data science?

- **Mercenary-make more $$$:**
  - Startups need data science talent right away = low employee #
  - Massive industry…

- **Intellectual**:
  - Science: data poor to data rich
    - No idea how to handle the data!
  - Fundamental ideas to/from all of CS:
    - Systems, theory, AI, logic, stats, analysis….

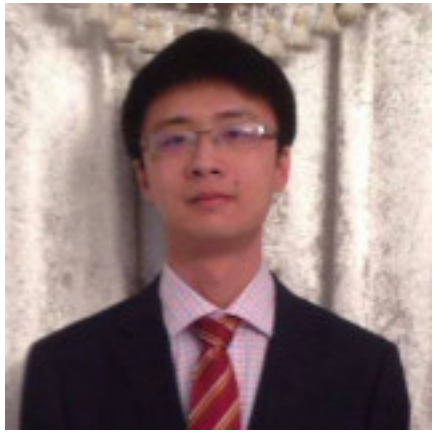
What this course is (and is not)

- Discuss **the fundamentals data management** for data science workflows
  - How to represent and store data
  - How to extract and prepare data for analysis
  - How to analyze data and how to visualize and communicate insights
- You will learn **how to design data science pipelines**
- This is not a databases, systems, or machine learning class. We will touch many topics covered in these classes but we will not go into details.

Who we are…

Instructor (me) Theo Rekatsinas

- Faculty in the Computer Sciences and part of the UW-Database Group
- **Research**: data integration and cleaning, statistical analytics, and machine learning.
- thodrek@cs.wisc.edu
- Office hours: MWF after class @CS 4361



Frank Zou

Huawei
Wang

Communication w/ Course Staff

- Piazza https://piazza.com/wisc/spring2019/cs639
- Class mailing list: compsci639-4-s19@lists.wisc.edu
- Office hours: Listed on the website
- *Also By appointment!*

The goal is to get you to answer each other's questions so you can benefit and learn from each other.

Course Website:
https://thodrek.github.io/cs639_spring19/

Course Github:
https://github.com/thodrek/cs639_spring19

Lectures

- Lecture slides cover **essential material**
  - This is your best reference.
  - We will have pointers as needed
  - Recommended textbooks listed on website

- Try to cover same thing in **many ways**: Lecture, lecture slides, homework, exams (no shock)
  - Attendance makes your life easier…

Graded Elements

- **Five** Programming assignments (45%)
  - Focus on different aspects of data science
- Midterm (20%)
- Final exam (35%)

Dates are posted on the website!!!

What is expected from you

- **Attend lectures**
  - If you don't, it's at your own peril
- **Be active and think critically**
  - Ask questions, post comments on forums
- **Do programming projects**
  - Start early and be honest
- **Study for tests and exams**

Programming Assignments

- **Five programming assignments**
  - Python plus Jupiter notebooks
- **These are individual assignments**
  - Submission via Canvas
- **~1 week per programming assignments**
  - Ask questions, post comments on forums
  - Start early!
- **You have late days. The policy is described on the website**

Programming setup for class

1. For all assignments we will use the provided Virtual Machine.
   - Ubuntu + necessary python libraries
   - Link provided on the website
   - We will not provide support for any other platform (you can still use your own machine)

2. To deploy and run the provided VM:
   1. **Download and Install Virtual Box: https://www.virtualbox.org/wiki/Downloads**
   2. **Download the Class VM from: https://www.dropbox.com/s/xjvj3jlaurzjfas/cs639_vm.ova.zip?dl=0**
   3. **Import the VM by following the instructions here: https://blogs.oracle.com/oswald/importing-a-vdi-in-virtualbox**
   4. **Run the VM and login with the following credentials:**
      1. Username: CS639_DS_USER
      2. Password: cs639_ds_user

3. Come to office hours if you need help with installation!

Please help out your peers by posting issues / solutions on Piazza!

2. Introduction to Data Science

What you will learn about in this section

1. What is Data Science?
2. Data Science workflows
3. What should a Data Scientist know?
4. Overview of lecture coverage

Data Science is an emerging field

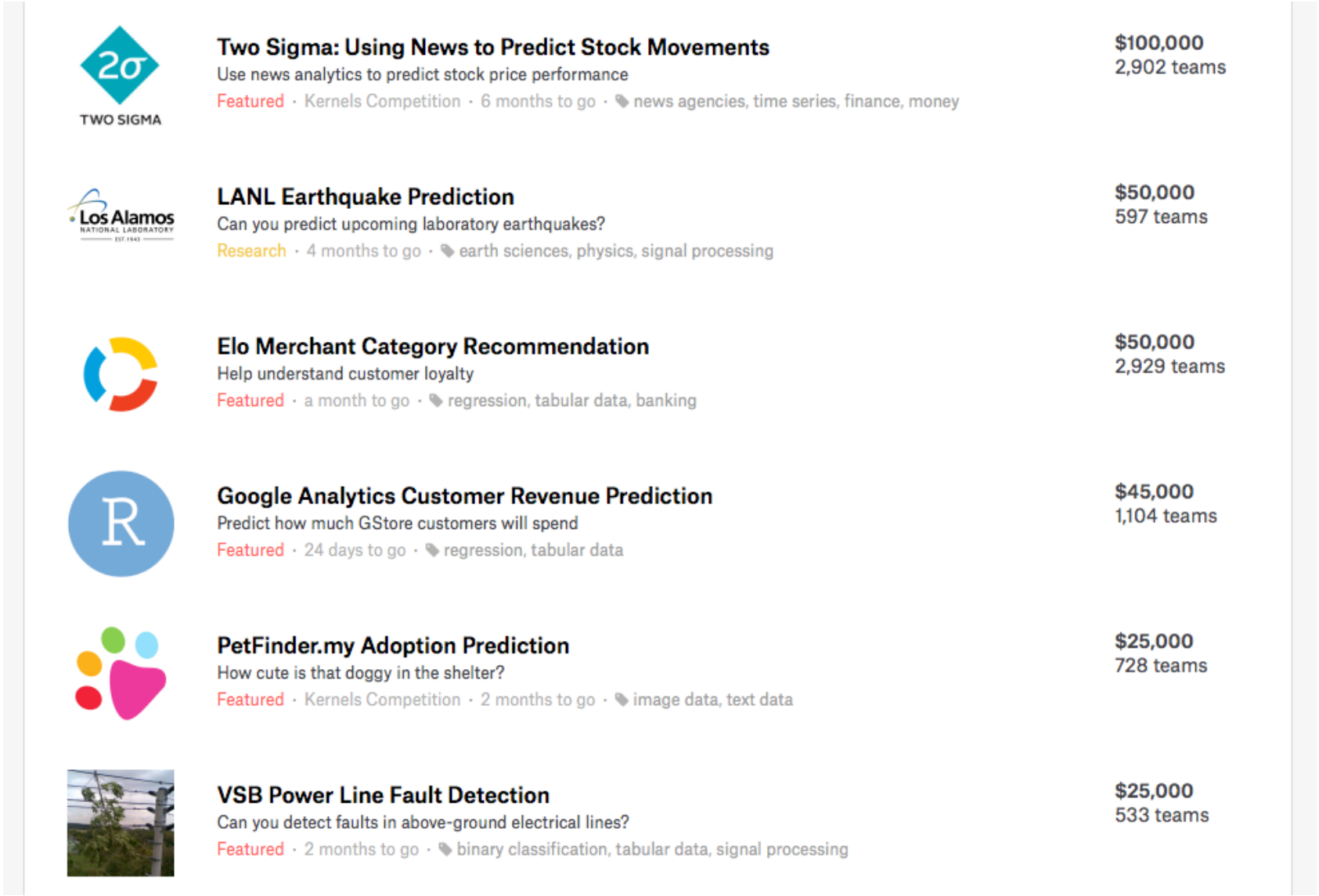
What is data science?

The future belongs to the companies and people that turn data into products.

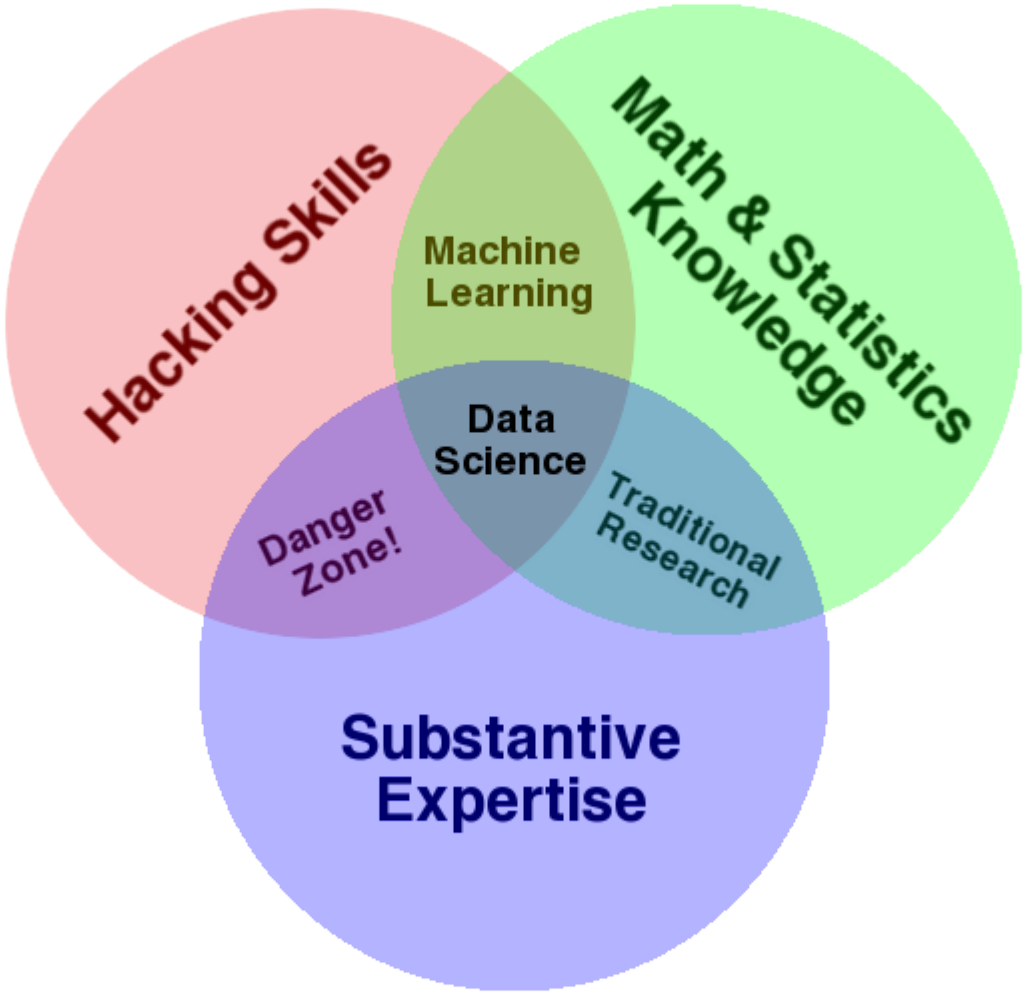
By Mike Loukides. June 2, 2010

- https://www.oreilly.com/ideas/what-is-data-science

Data Science products



One definition of data science



Data science is a broad field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data.

Source: Techopedia

Data science is not databases

	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: Riak, Memcached, MongoDB, CouchDB, Hbase, Cassandra,…

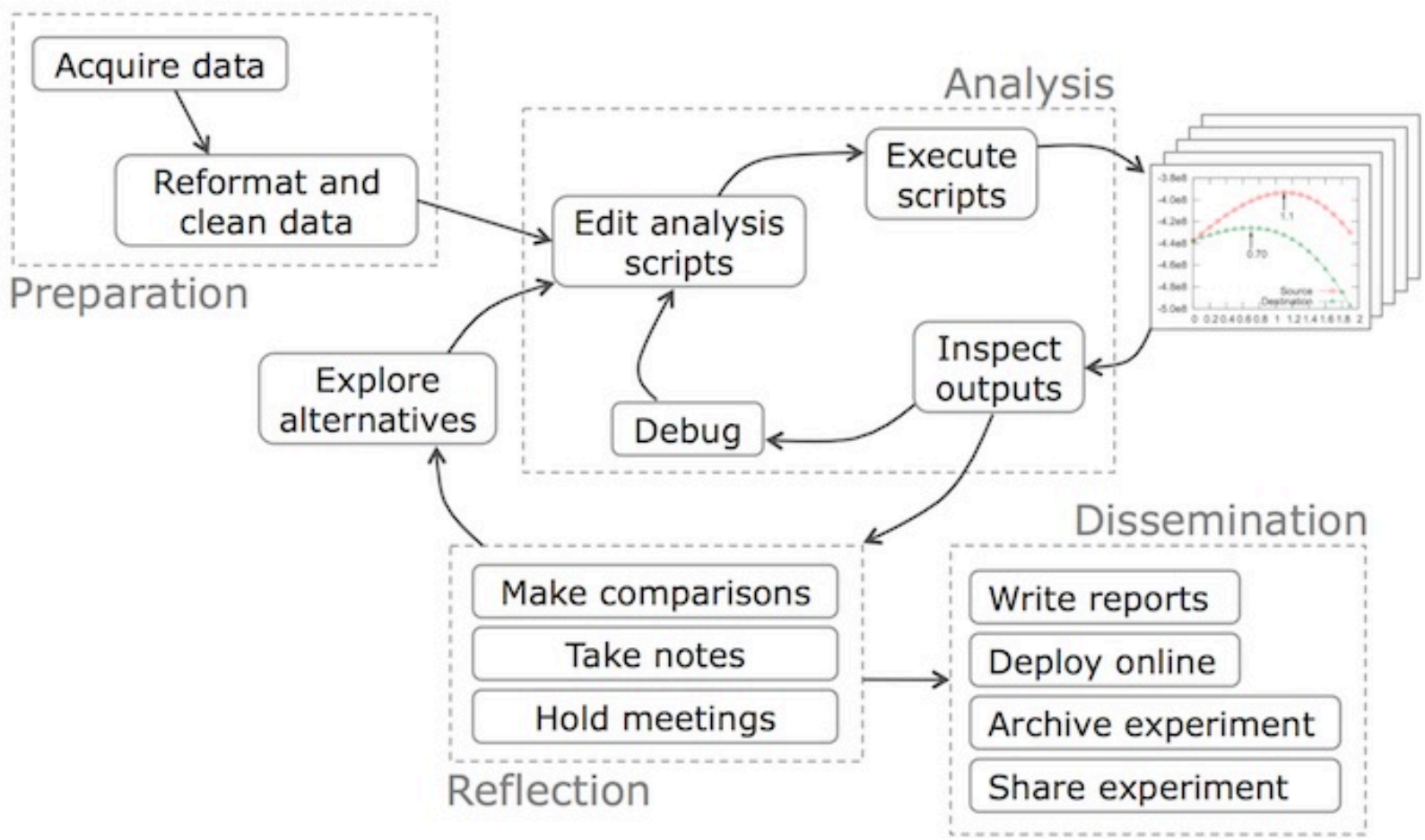
ACID = Atomicity, Consistency, Isolation and Durability CAP = Consistency, Availability, Partition Tolerance

Data science is not databases

Databases	Data Science
Querying the past	Querying the future

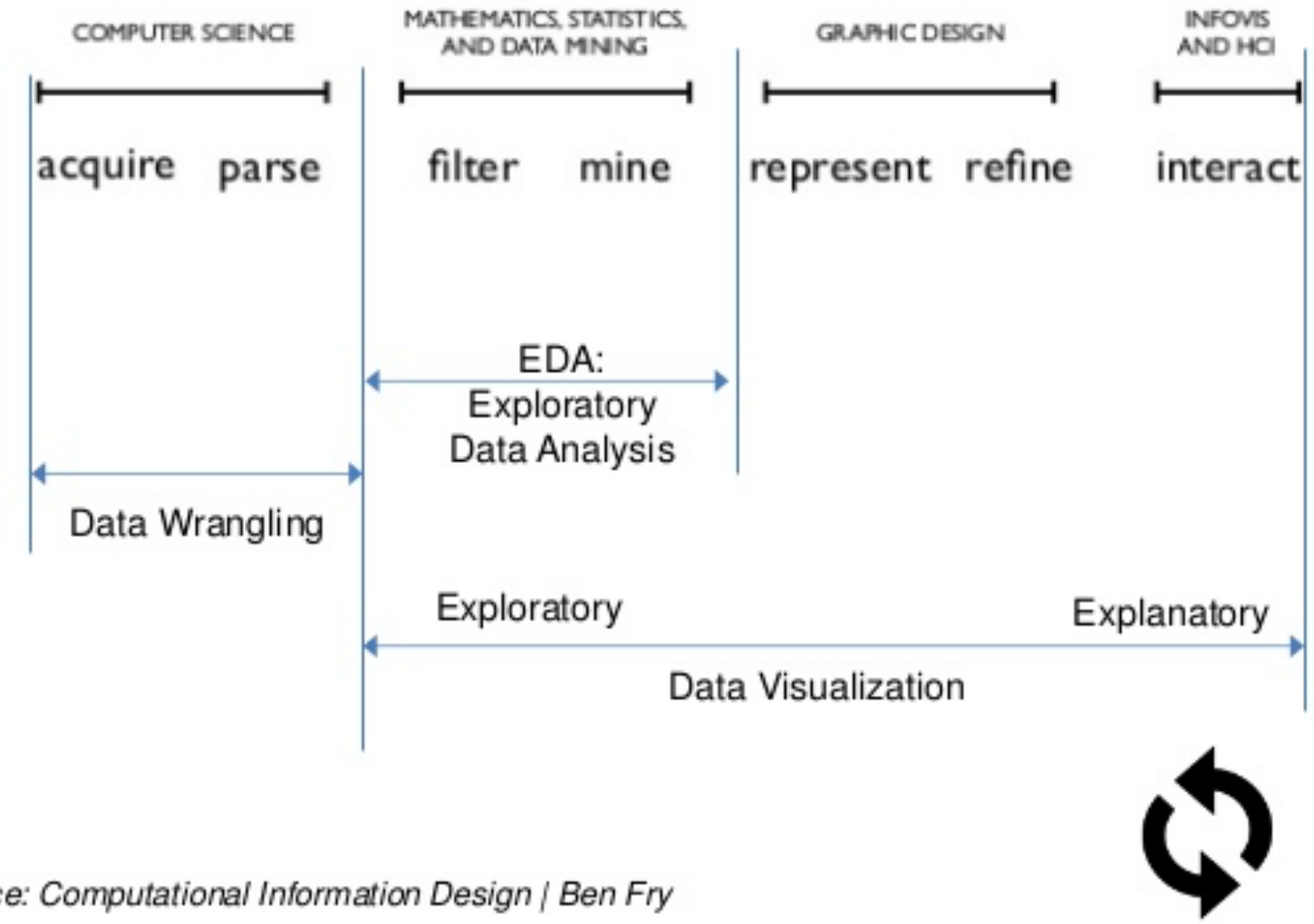
Business intelligence (**BI**) is the transformation of raw data into meaningful and useful information for business analysis purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia

Data science workflow



https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

Data science workflow

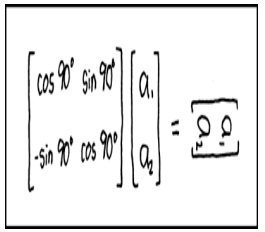


Source: Computational Information Design | Ben Fry

Data science workflow

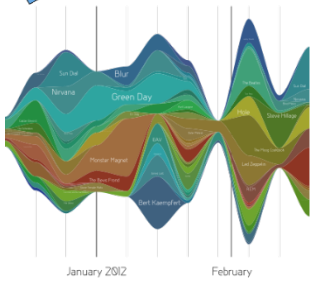
Digging Around in Data

Clean, prep



Hypothesize Model

Large Scale Exploitation

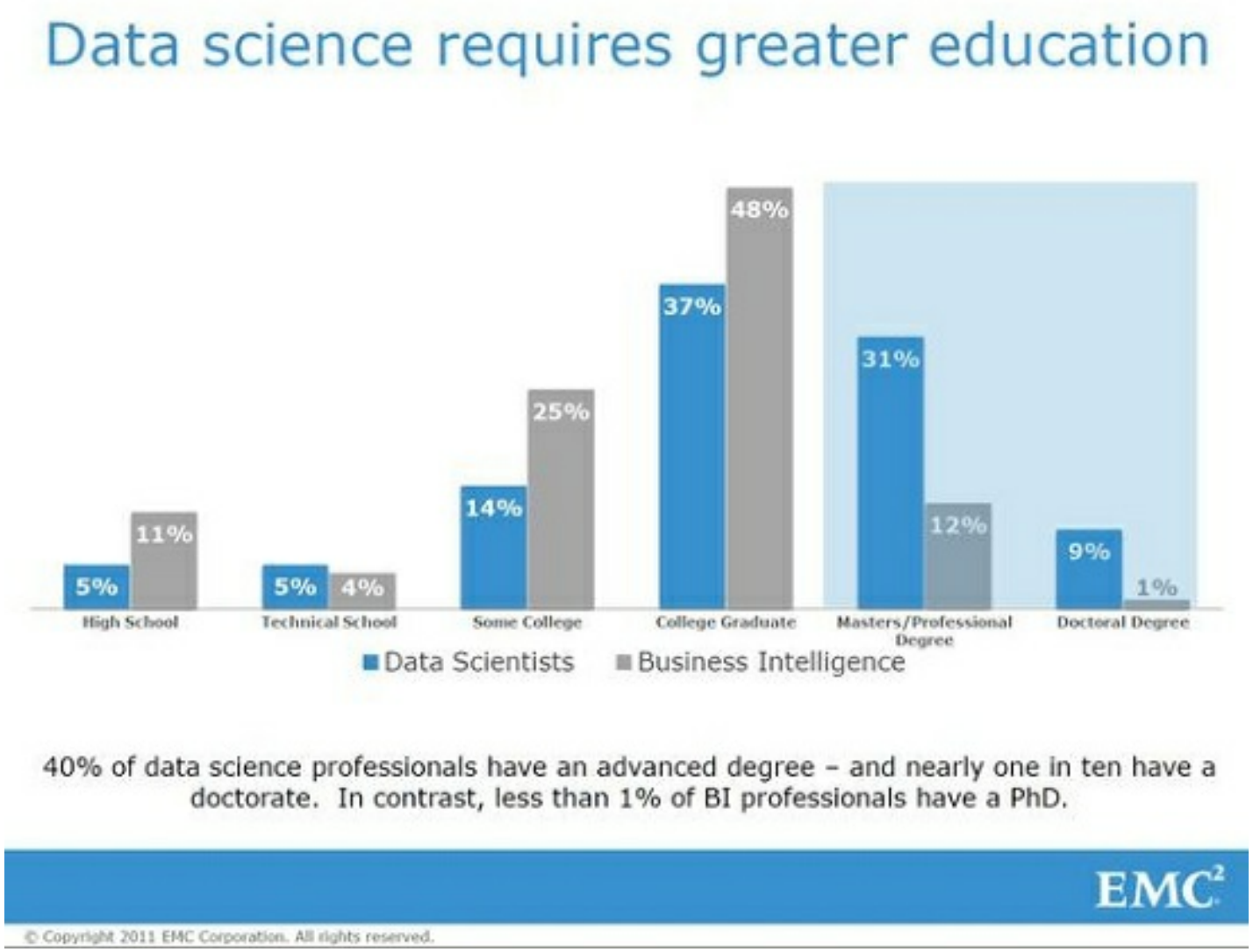


Evaluate Interpret

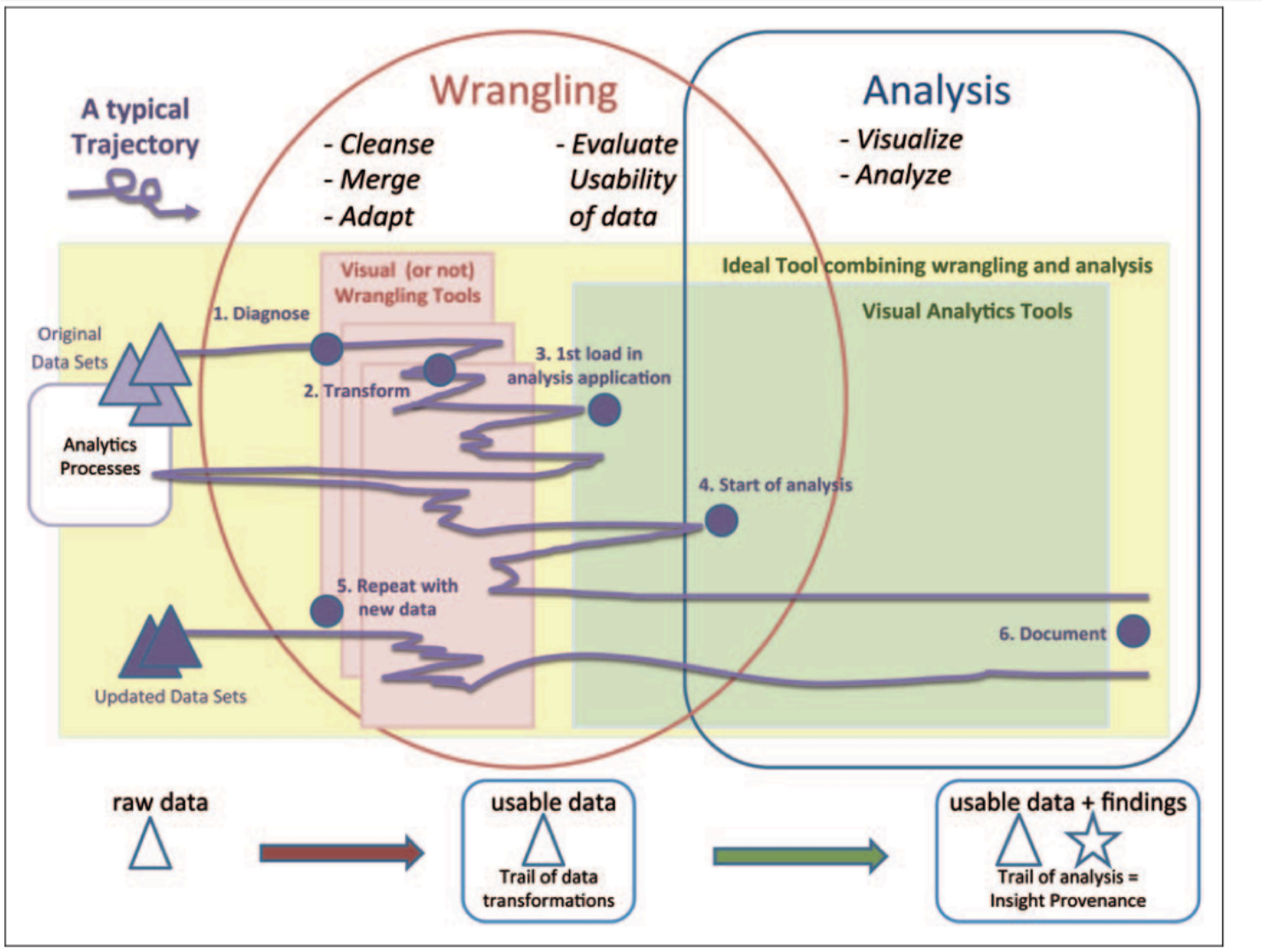
What is hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

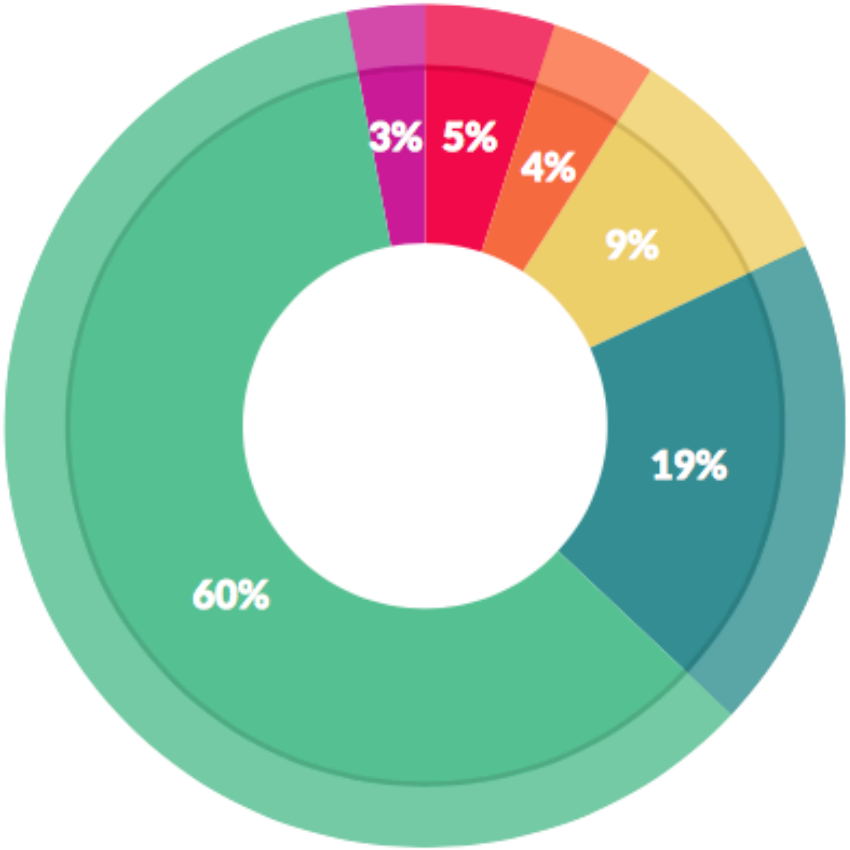
What is hard about Data Science



What is hard about Data Science



What are Data Scientists really doing?



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Lectures: 1st part – Data Storage

1. **Overview:** What is data science
   - Lectures 1-3
2. **Foundations:** Relational data models & SQL
   - Lectures 4-7
   - How to manipulate data with SQL, a declarative language
     - *reduced expressive power but the system can do more for you*
   - Query optimization
3. **MapReduce and NoSQL systems**: MapReduce, KeyValue Stores, Graph DBs
   - Lectures 8-14
   - Dealing with massive amounts of data and non-relational data

Lectures: 2nd part – Predictive analytics

4. Statistical Reasoning: Inference, Sampling Bayesian Methods

- Lectures 15-17
- How to reason about patterns in data

5. Machine Learning: Decision Trees, Evaluation of ML models, Ensembles

- Lectures 18-22
- Overview of different ML paradigms

6. Optimization: How to train ML models (efficiently)?

- Lecture 23
- Loss Functions, Optimization via Gradient Descent
- Stochastic Gradient Descent (SGD), Parallel SGD

Lectures: 3rd part – Data Integration

7. Information Extraction: Named entity recognition and relation extraction
- Lecture 24
- How to identify entities of interest in unstructured data?
- How to find relationships between them?

8. Data Integration: Combine information from different data sources
- Lecture 25
- How to find if a real-world entity is mentioned in different sources?
- How to align data from different sources?

9. Data Cleaning: Remove errors and noise from data
- Lecture 26
- How can we detect errors in data to be used for analytics?
- How can we fix these errors automatically?

Lectures: 4th part – Communicating Insights

10. Data Visualization: Creating data charts that convey interesting findings

- Lectures 27-29
- How to convey insights most effectively?
- How to explore raw data?

11. Data Privacy: Sharing sensitive information

- Lecture 30-32
- How can we share sensitive data?
- How can we perform analytics on sensitive data?