Big Data and Data Mining

Big Data

- Massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques.

- Technology that an organization requires to store and analyze the large amounts of data.
- Data mining methods do the analysis.

Big Data and Data Mining

Big Data Applications

Facebook

- Captures more than 1.5PB weblog data daily.
- Recommends friend and items.
- Makes targeted ads.

Amazon

- Collects more than 200TB of weblog data daily.
- Recommends items.
- Makes targeted ads.

Big Data and Data Mining

Big Data Applications

Obama election campaign

- Voter's particular interests were recorded by door-to-door campaign members.
- Stored in campaign database.
- Designed emails from the local organizer to voters, each corresponding to a voter's favorite campaign issue.
- More effective and more economical ads targeting the precise demographic slices the Obama campaign was trying to reach.

Big Data Applications

Large Hadron Collider

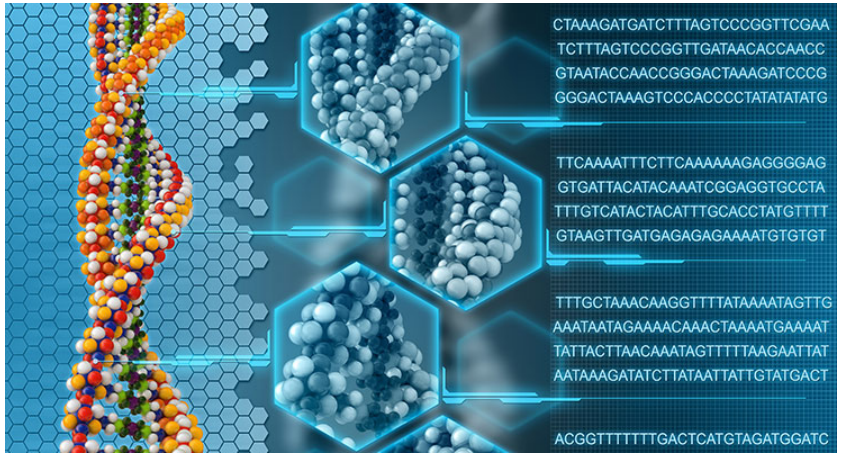
- For particle physics experiments, the largest experimental facility in the world.
- About 150 million sensors delivering data 40 million times per second.
- Nearly 600 million collisions per second.
- Detect the 100 collisions of interest per second.
- Detect and characterize new particles.

Big Data and Data Mining

Big Data Applications

Genomics

- Sequencing one human genome produces ~1TB of data.
- Precision Medicine Initiative: 1 million volunteers provide EHRs, healthcare claims, biological samples (e.g. for DNA collection).



- Goal: understand relationship between genotype and phenotype for more precise, personalized diagnostics and treatment.

Machine Learning

Definition

- Machine learning is a field of computer science that gives computers the ability to learn withput explicitly being programmed. (Arthur Samuel, 1959)

Applications

- Robotics
- Computer vision
- Natural language processing
- Recommender systems
- Data mining

Machine Learning

Tasks

- Clustering (unsupervised)
- Classification (supervised)
- Prediction
- Anomaly detection
- Dimensionality reduction

Precision Medicine

Definition

- Precision medicine is an approach for disease treatment and prevention that takes into account individual variability in genes, environment and lifestyle for each person.

  https://ghr.nlm.nih.gov/primer/precisionmedicine/

- In contrast to a one-size-fits-all approach.
- As a first step towards fully personalized medicine, determine groups of similar patients that can be treated similarly.

Precision Medicine

Medical tasks

- Patient stratification (clustering)
- Diagnosis (classification)
- Prognosis (prediction)
- Recommendation of treatment
- . . .

Public health tasks

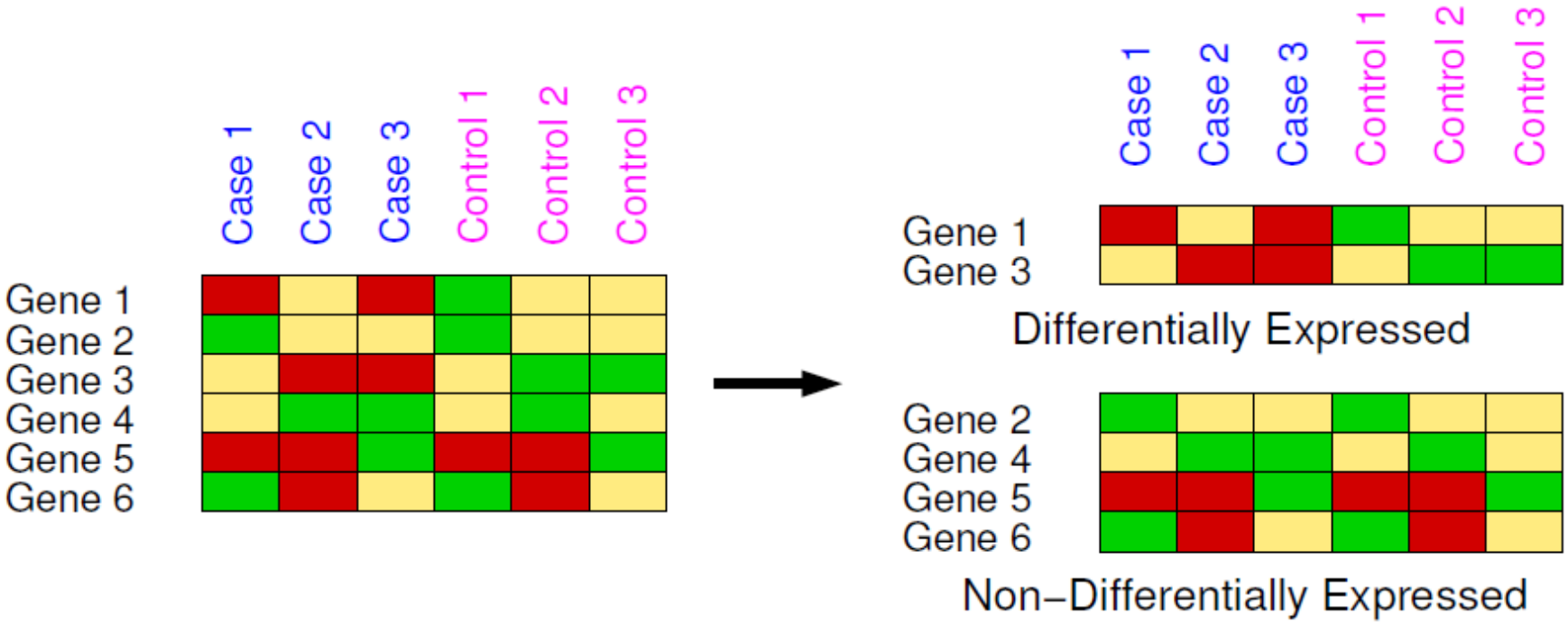
- Detection of outbreaks of infectious diseases
- Detection of phylogenetic trees of strains of bacteria
- . . .

Precision Medicine

Scientific tasks

- Construction of biological networks
- Protein function prediction
- Biomarker discovery
- Discovery of causal genes
- Drug-target interaction prediction
- . . .

Precision Medicine



Machine Learning Challenges

Small, high-dimensional datasets

- Patient data is expensive and private.

Noise

- High-throughput data collection.

Explainability

- Need to explain the reason for a prediction.

Causal relationships

- Understand the causes of effects in order to manipulate them.

Course Outline

Introductory tutorials

- Probabilistic graphical models
- Deep neural networks

Paper presentations

Course project presentations

Tentative Grading Scheme

Paper presentation	30%
Course project presentation	30%
Course project report	40%