Learning from data

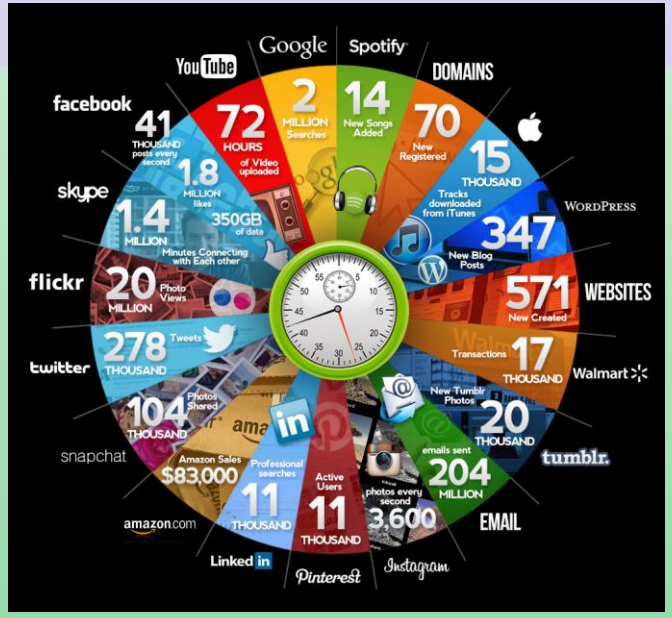
Course: Analytics, Machine Learning, and the Digital Economy

- **Lecturer Radjabova Dilnora**

Lesson Goals:

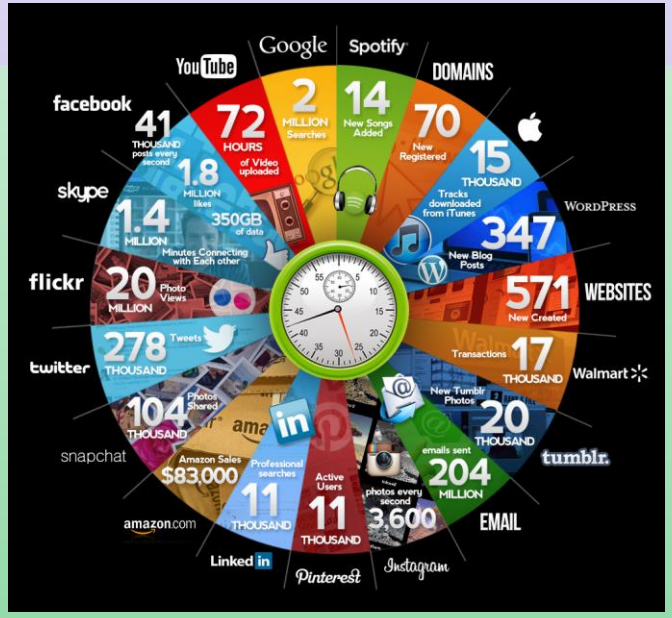
- Understand the basic concepts of the learning problem and why/how machine learning methods are used to learn from data to find underlying patterns for prediction and decision-making.
- Understand the learning algorithm trade-offs, balancing performance within training data and robustness on unobserved test data.
- Differentiate between supervised and unsupervised learning methods as well as regression versus classification methods.
- Understand the basic concepts of assessing model accuracy and the bias-variance trade-off.
- Become familiar with using the R statistical programming language and practice by exploring data using basic statistical analysis.

Big Data is Everywhere



- We are in the era of **big data**!
  - 40 billion indexed web pages
  - 100 hours of video are uploaded to YouTube every minute

Big Data is Everywhere



- The deluge of data calls for automated methods of data analysis, which is what **machine learning** provides!

What is Machine Learning?

- **Machine learning** is a set of methods that can *automatically* detect patterns in data.

What is Machine Learning?

- These uncovered patterns are then used to predict future data, or to perform other kinds of decision-making under uncertainty.

What is Machine Learning?

- The key premise is *learning* from data!!

What is Machine Learning?

- Addresses the problem of analyzing huge bodies of data so that they can be understood.

What is Machine Learning?

- Providing techniques to automate the analysis and exploration of large, complex data sets.

What is Machine Learning?

- Tools, methodologies, and theories for revealing patterns in data – critical step in knowledge discovery.

What is Machine Learning?

- Driving Forces:
  - Explosive growth of data in a great variety of fields
    - Cheaper storage devices with higher capacity
    - Faster communication
    - Better database management systems
  - Rapidly increasing computing power

- We want to make the data work for us!!

Examples of Learning Problems

- Machine learning plays a key role in many areas of science, finance and industry:
  - Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.
  - Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.
  - Identify the numbers in a handwritten ZIP code, from a digitized image.
  - Estimate the amount of glucose in the blood of a diabetic person, from the infrared absorption spectrum of that person's blood.
  - Identify the risk factors for prostate cancer, based on clinical and demographic variables.

Research Fields

- Statistics / Statistical Learning
- Data Mining
- Pattern Recognition
- Artificial Intelligence
- Databases
- Signal Processing

Applications

- Business
  - Walmart data warehouse mined for advertising and logistics
  - Credit card companies mined for fraudulent use of your card based on purchase patterns
  - Netflix developed movie recommender system

Applications (cont.)

- Genomics
  - Human genome project: collection of DNA sequences, microarray data

Applications (cont.)

- Information Retrieval
  - Terrabytes of data on internet, multimedia information (video/audio files)

Applications (cont.)

- Communication Systems
  - Speech recognition, image analysis

The Learning Problem

- Learning from data is used in situations where we don't have any analytic solution, but we do have data that we can use to construct an empirical solution

The Learning Problem

- The basic premise of learning from data is the use of a set of observations to uncover an underlying process.

The Learning Problem (cont.)

- Suppose we observe the output space Y_i and the input space $X_i = (X_{i1}, \ldots, X_{ip}) \; \forall \; i = 1, \ldots, n$

The Learning Problem (cont.)

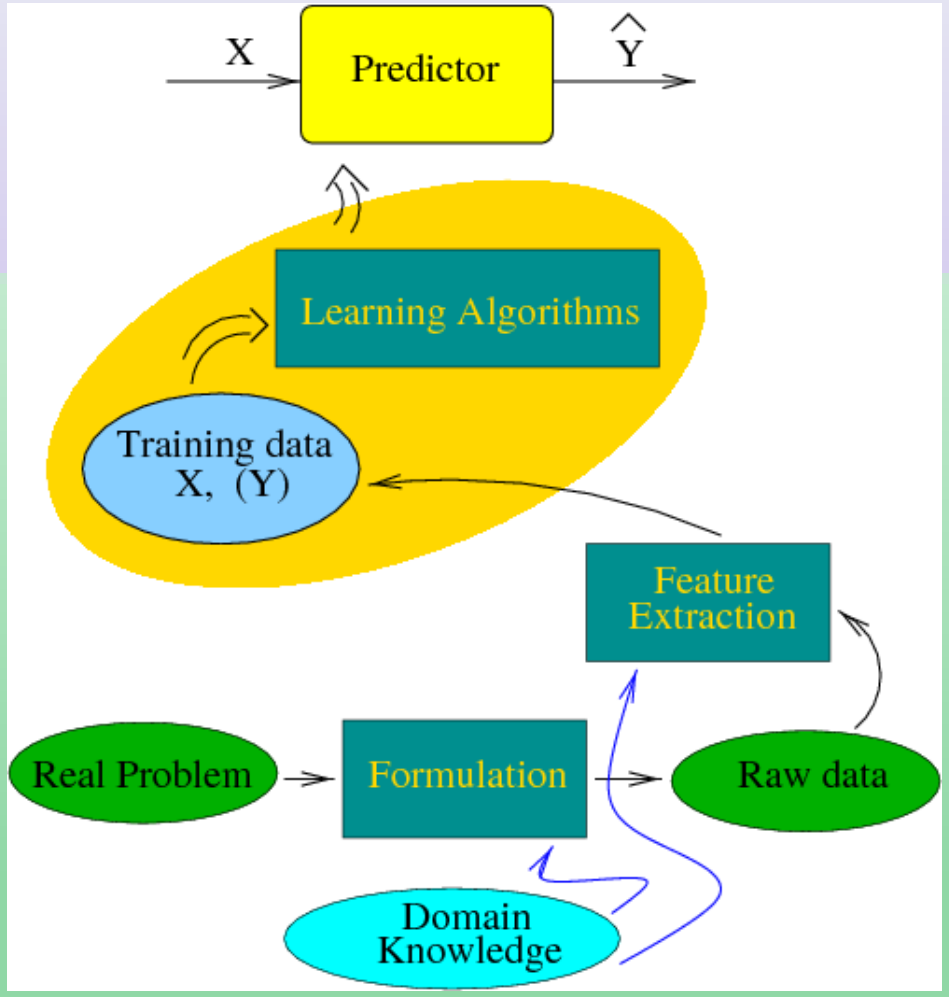
- We believe that there is a *relationship* between Y and at least one of the X's.

The Learning Problem (cont.)

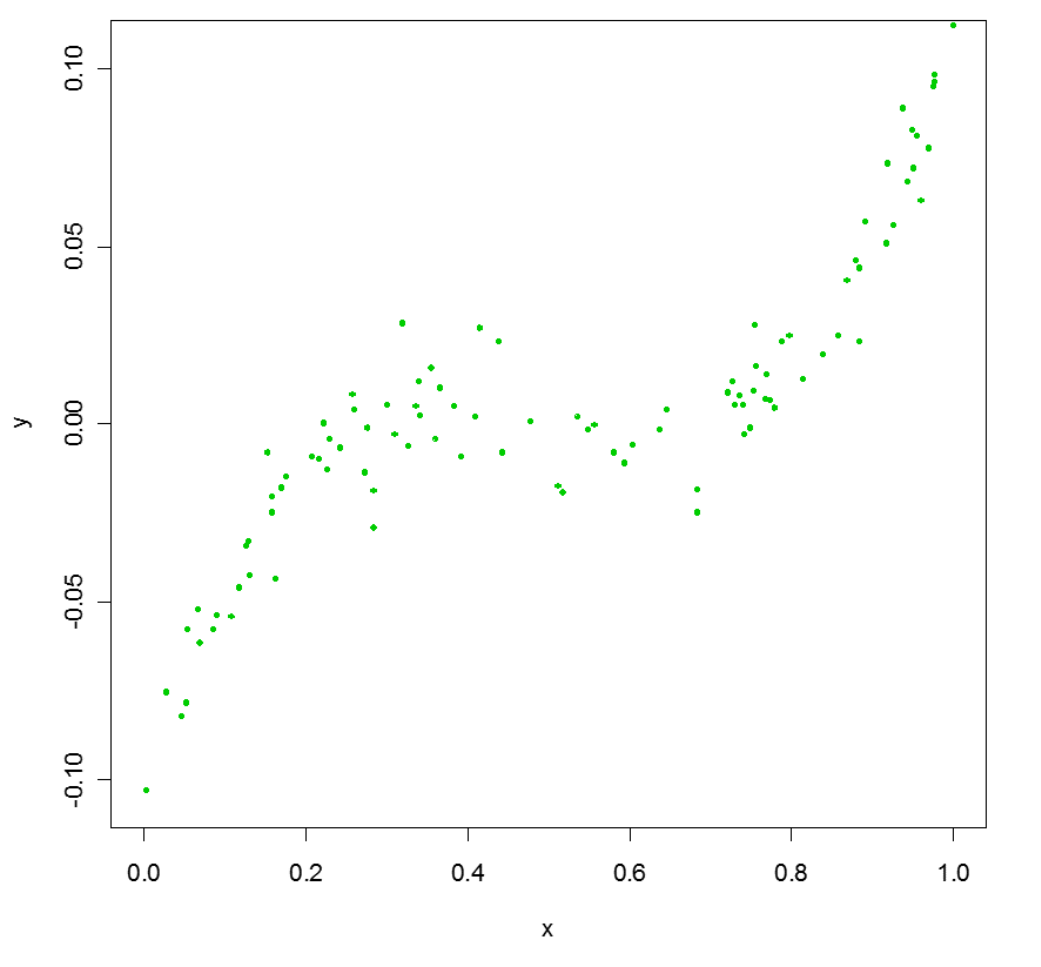
- We can model the relationship as: $Y_i = f(\boldsymbol{X_i}) + \varepsilon_i$

where f is an unknown function and ε is a random error (noise) term, independent of $\boldsymbol{X}$ with mean zero.

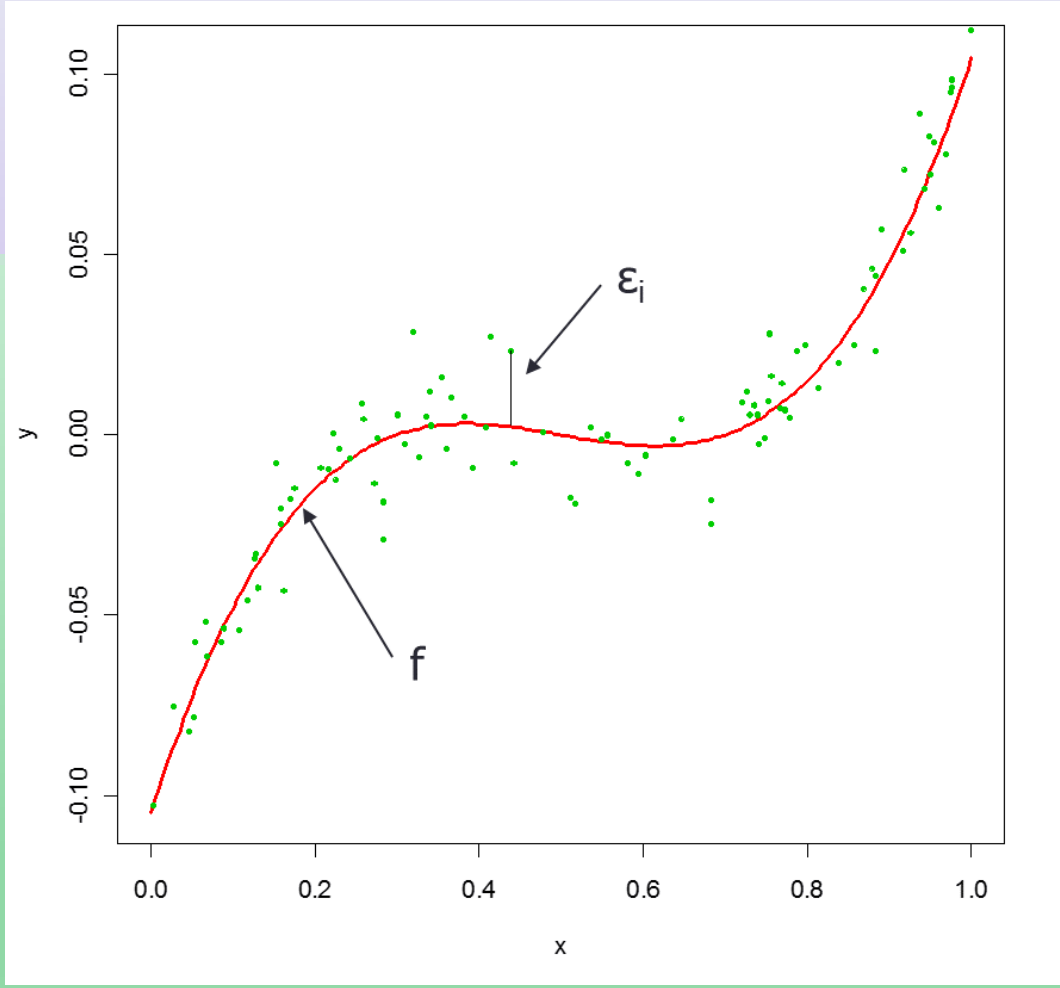
The Learning Problem (cont.)



The Learning Problem: Example

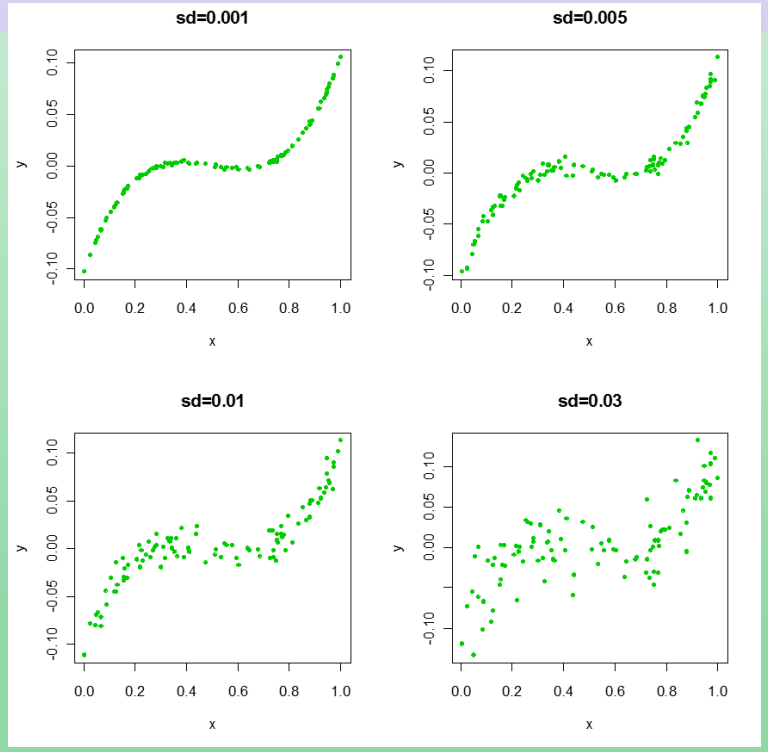


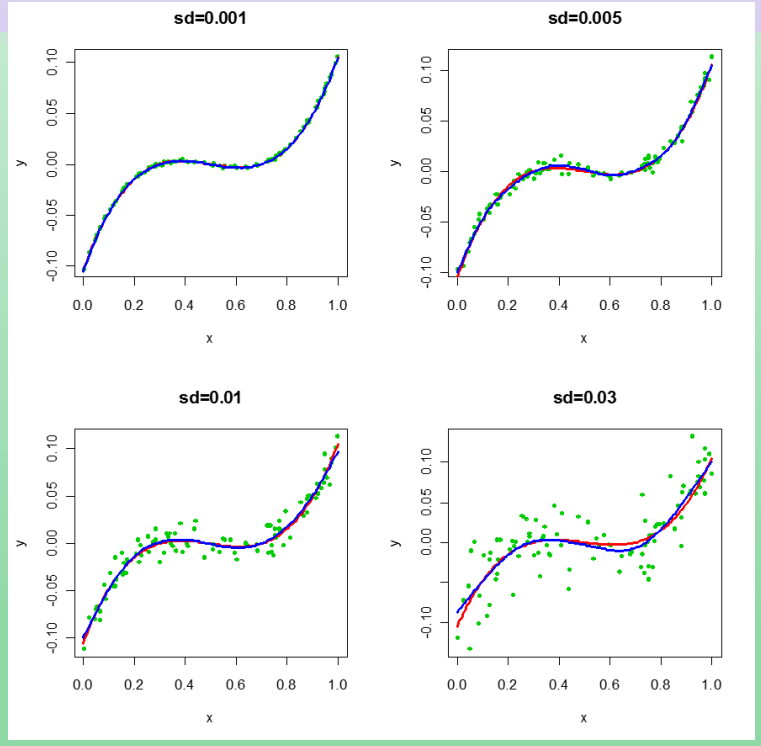
The Learning Problem: Example (cont.)



The Learning Problem: Example (cont.)

- Different estimates for the target function *f* that depend on the standard deviation of the ε's





Why do we estimate *f*?

- We use modern machine learning methods to estimate *f* by *learning* from the data.

Why do we estimate *f*?

- The target function f is unknown.

Why do we estimate *f*?

- We estimate *f* for two key purposes:
  - Prediction
  - Inference

Reference

- 1. Shalev-Shwartz and Ben-David. Understanding Machine Learning: From Theory to Algorithms (Cambridge University Press, 2014)
- 2. Daum´e. A Course in Machine Learning.
- 3. The Art of Statistics: How to Learn from Data by David Shpigelter
- 4. Learning From Data – January 1, 2012 by Yaser S. Abu-Mostafa (Author), Malik Magdon-Ismail (Author), Hsuan-Tien Lin (Author)
- 5. Statistics: The Art and Science of Learning from Data by Alan Agresti
- 6. Learning From Data: An Introduction To Statistical Reasoning by M.Glenber.
- 7. Statistics: Learning from Data (with JMP Printed Access Card) by Rocky Pek
- 8. The Elements of Statistical Learning by Gerim Garold
- 9. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition
- by Aurélien Géron (Author)