DSO530 Statistical Learning

Lecture 1: What is statistical (machine) learning?

Dr. Xin Tong
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
xint@marshall.usc.edu

Today's agenda

- Introduction of professor
- Statistics, statistical learning and their definitions
- Examples and exercises
- Syllabus highlights
- Expectations

About the instructor

My research site: https://sites.google.com/site/xintonghomepage/

Let's start from "Statistics"

- Statistics is an old term.
- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data
- Statistics is much more than counting in agriculture, compiling baseball scores, creating life and death comparison data, etc!
- Canadian science philosopher Ian Hacking (1936-) captures the essence of Statistics:

"*The quiet statisticians have changed our world - not by discovering new facts or technical developments but by changing the ways we reason, experiment and form our opinions about it.* ''

Let's start from "Statistics"

- Statistics is an old term.
- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data
- Statistics is much more than counting in agriculture, compiling baseball scores, creating life and death comparison data, etc!
- Canadian science philosopher Ian Hacking (1936-) captures the essence of Statistics:

"*The quiet statisticians have changed our world - not by discovering new facts or technical developments but by changing the ways we reason, experiment and form our opinions about it.* ''

Let's start from "Statistics"

- Statistics is an old term.
- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data
- Statistics is much more than counting in agriculture, compiling baseball scores, creating life and death comparison data, etc!
- Canadian science philosopher Ian Hacking (1936-) captures the essence of Statistics:

"The quiet statisticians have changed our world - not by discovering new facts or technical developments but by changing the ways we reason, experiment and form our opinions about it. ''

Let's start from "Statistics"

- Statistics is an old term.
- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data
- Statistics is much more than counting in agriculture, compiling baseball scores, creating life and death comparison data, etc!
- Canadian science philosopher Ian Hacking (1936-) captures the essence of Statistics:

"*The quiet statisticians have changed our world - not by discovering new facts or technical developments but by changing the ways we reason, experiment and form our opinions about it*. ''

21st century: big data->machine learning, AI!

Figure 2: Big data everywhere

Statistical (machine) learning

- Machine learning constructs **algorithms** that can learn from data , especially for **prediction**
- Statistical learning is a branch of Statistics that was developed in response to Machine learning, emphasizing building statistical models, drawing **inferences** and assessing uncertainty
- Data Science is the extraction of knowledge from data, using ideas from mathematics, statistics, machine learning, computer programming, data engineering ...
- **In this course, we use statistical learning and machine learning interchangeably**.
- Other terms you encounter often: artificial intelligence (AI), deep learning, etc.

Statistical (machine) learning

- Machine learning constructs **algorithms** that can learn from data , especially for **prediction**
- Statistical learning is a branch of Statistics that was developed in response to Machine learning, emphasizing building statistical models, drawing **inferences** and assessing uncertainty
- Data Science is the extraction of knowledge from data, using ideas from mathematics, statistics, machine learning, computer programming, data engineering ...
- **In this course, we use statistical learning and machine learning interchangeably**.
- Other terms you encounter often: artificial intelligence (AI), deep learning, etc.

Definitions from CS community

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
- Tom Mitchell (1998). Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

An exercise

- Recall "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."
- Suppose your email program watches which emails you mark as spam, and based on that learns how to better filter spam. *What are the task T, performance measure P, and experience E in this setting?*

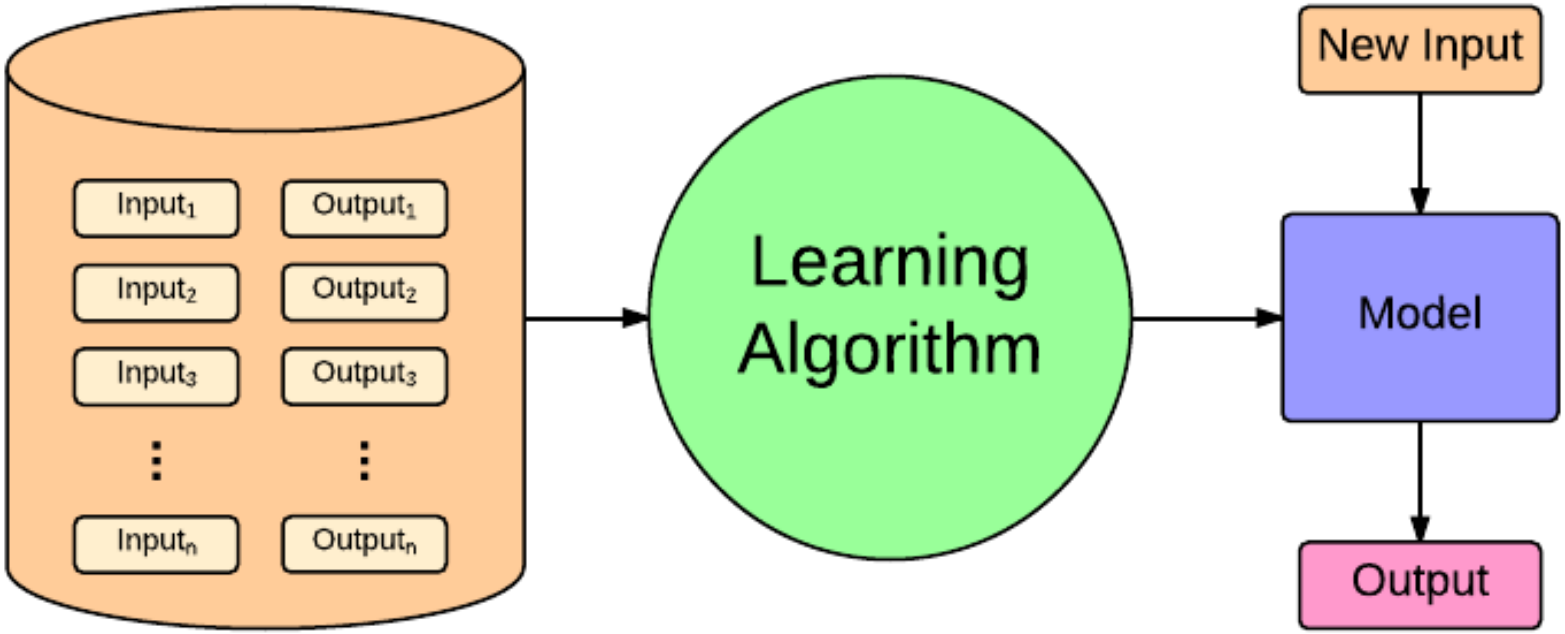
Structured vs. unstructured data

- Structured data: a flat file with a fixed number of measurements, e.g., patient response to a drug under consideration of certain conditions (such as age, weight, size, nutrition intake)
- Unstructured data: doctor's notes, Twitter feeds, broker reports
- We focus on structured data in this course.

21st century statistical learning examples

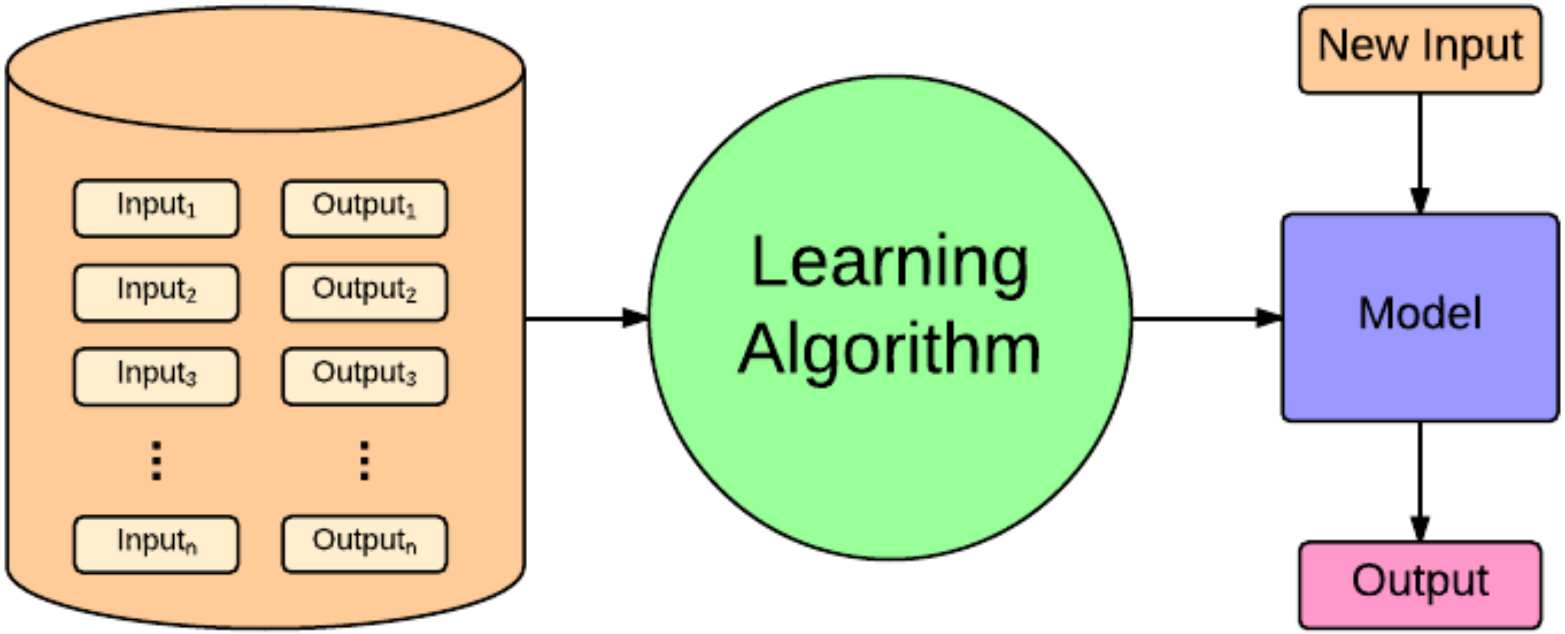
- Use **classification techniques** to classify which accounts are the most likely to upgrade their service contract (this helps the salesforce to know which leads / accounts to focus on to sell more)...
- Use **regression techniques** to improve how sales people pitch prospective clients and what features of the company's services they should highlight...
- **Regression and classification** are examples of **Supervised Learning** techniques (see next slide)
- Use **optimization techniques** to maximize the number of views of company promotional material a prospective customer sees for a given dollar amount of promotional spend

Supervised learning paradigm



Traditional statistics: domain experts work for many years to learn good **features**; they bring statisticians a small clean dataset
Today's scenario: Domain knowledge is limited in new fields and large data sets are readily available. we (are sometimes forced to) start with a large dataset with many features

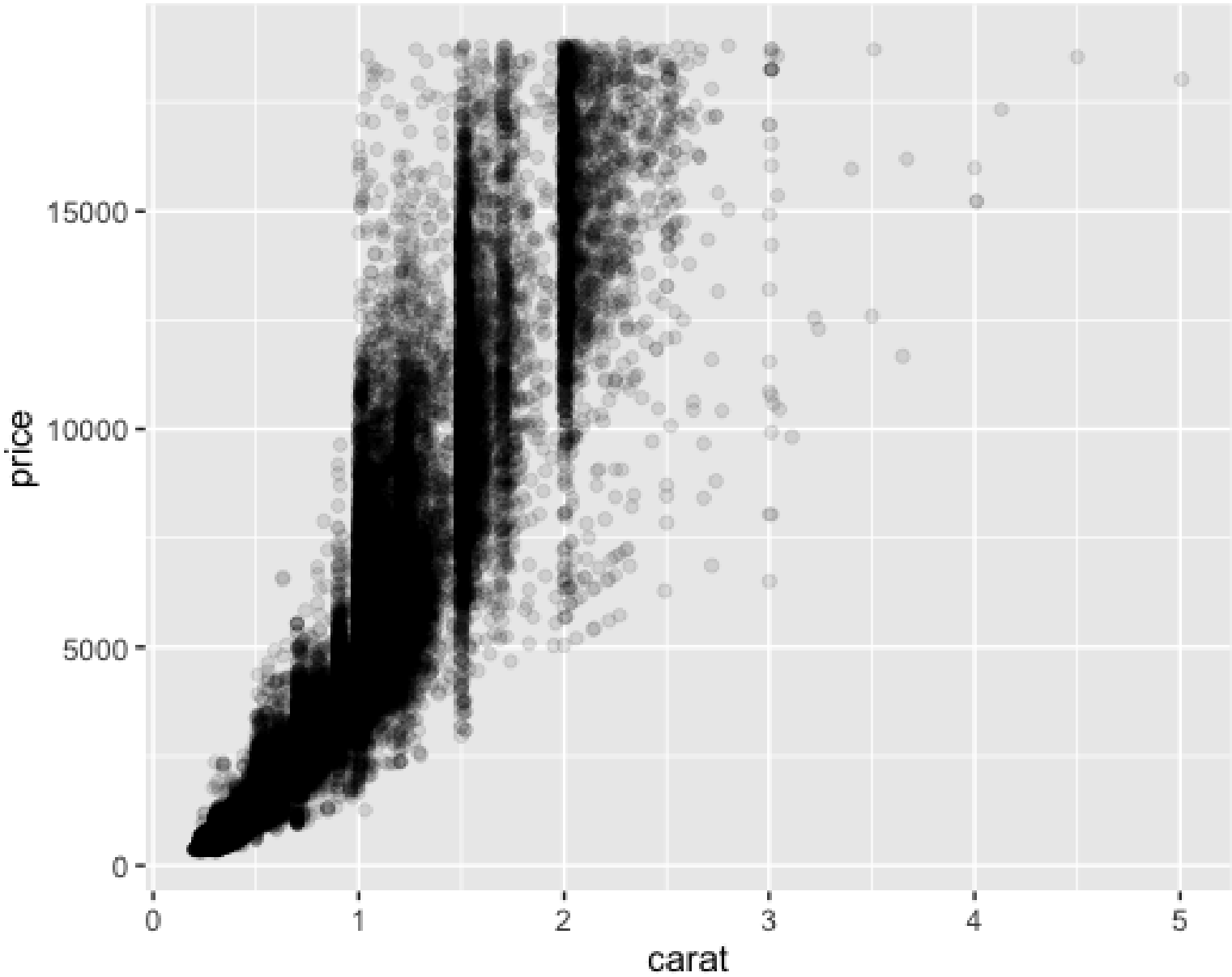
Supervised learning paradigm



Traditional statistics: domain experts work for many years to learn good **features**; they bring statisticians a small clean dataset
Today's scenario: Domain knowledge is limited in new fields and large data sets are readily available. we (are sometimes forced to) start with a large dataset with many features

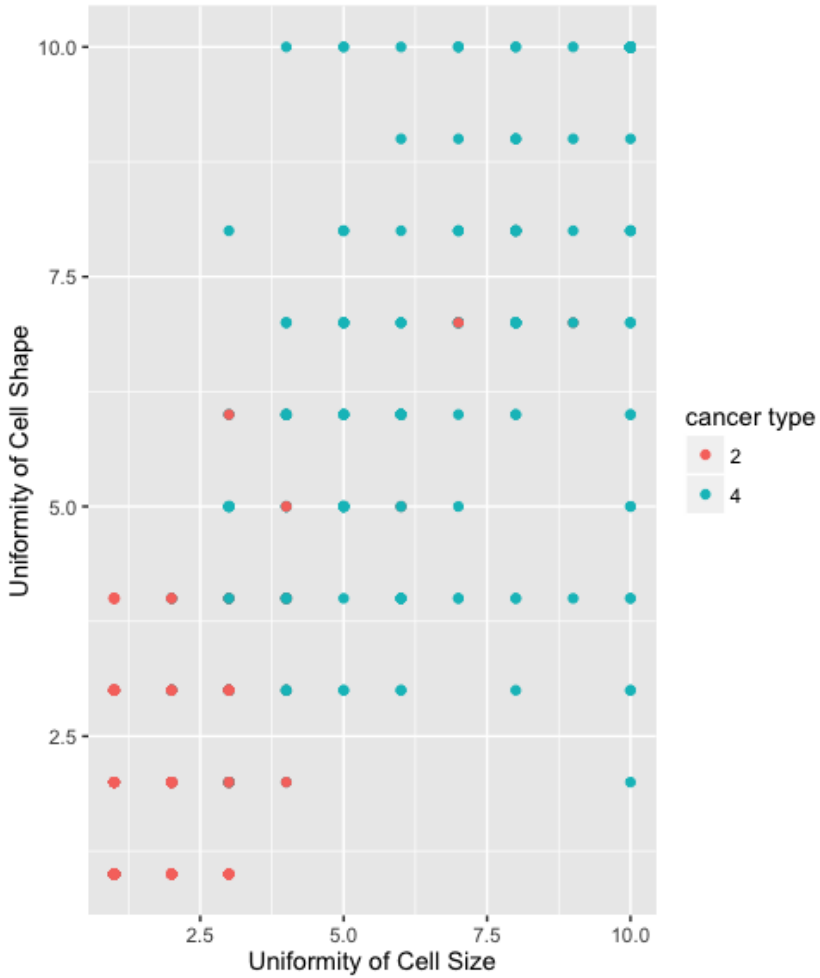
First supervised learning example: diamond price prediction

- Task: predict diamond price based on weight (regression)



Second example: cancer diagnosis (benign, malignant)

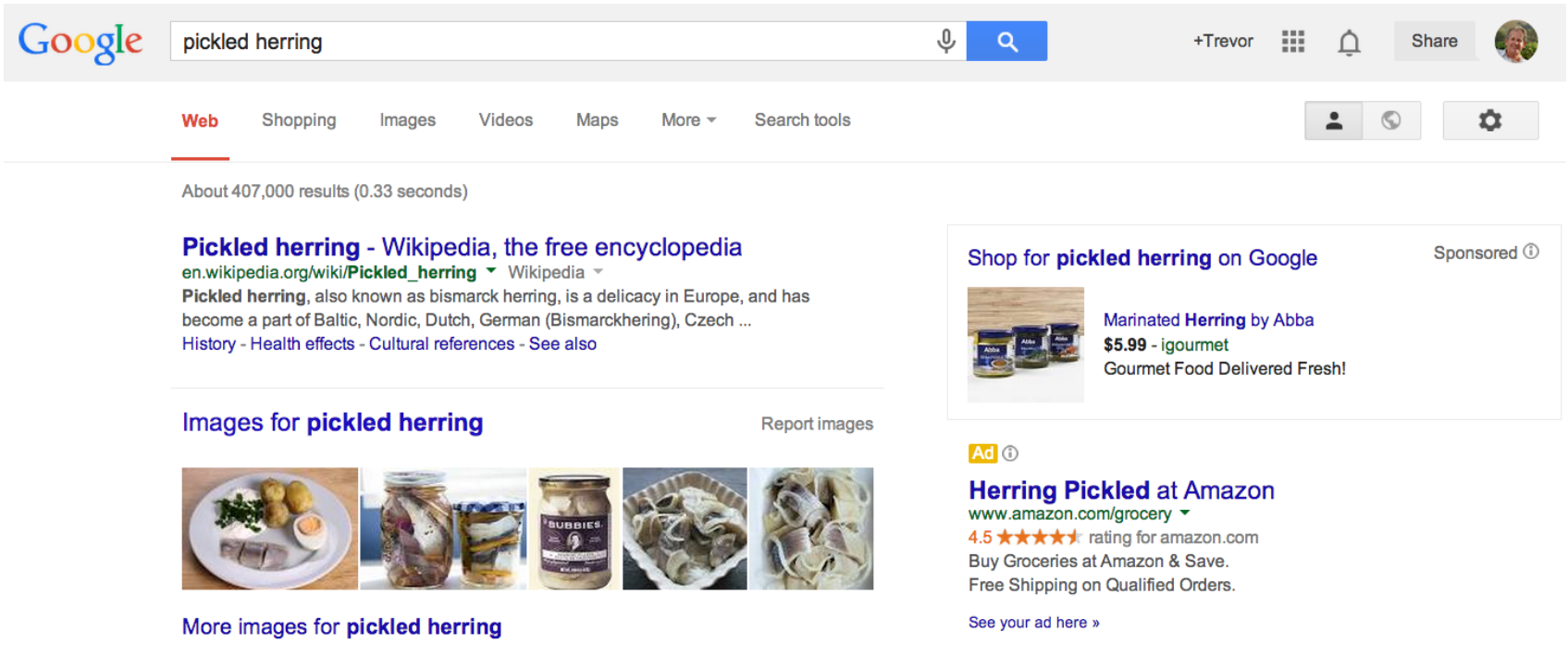
- This breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. (2 codes benign and 4 codes malignant). A classification example



More examples: Netflix challenge

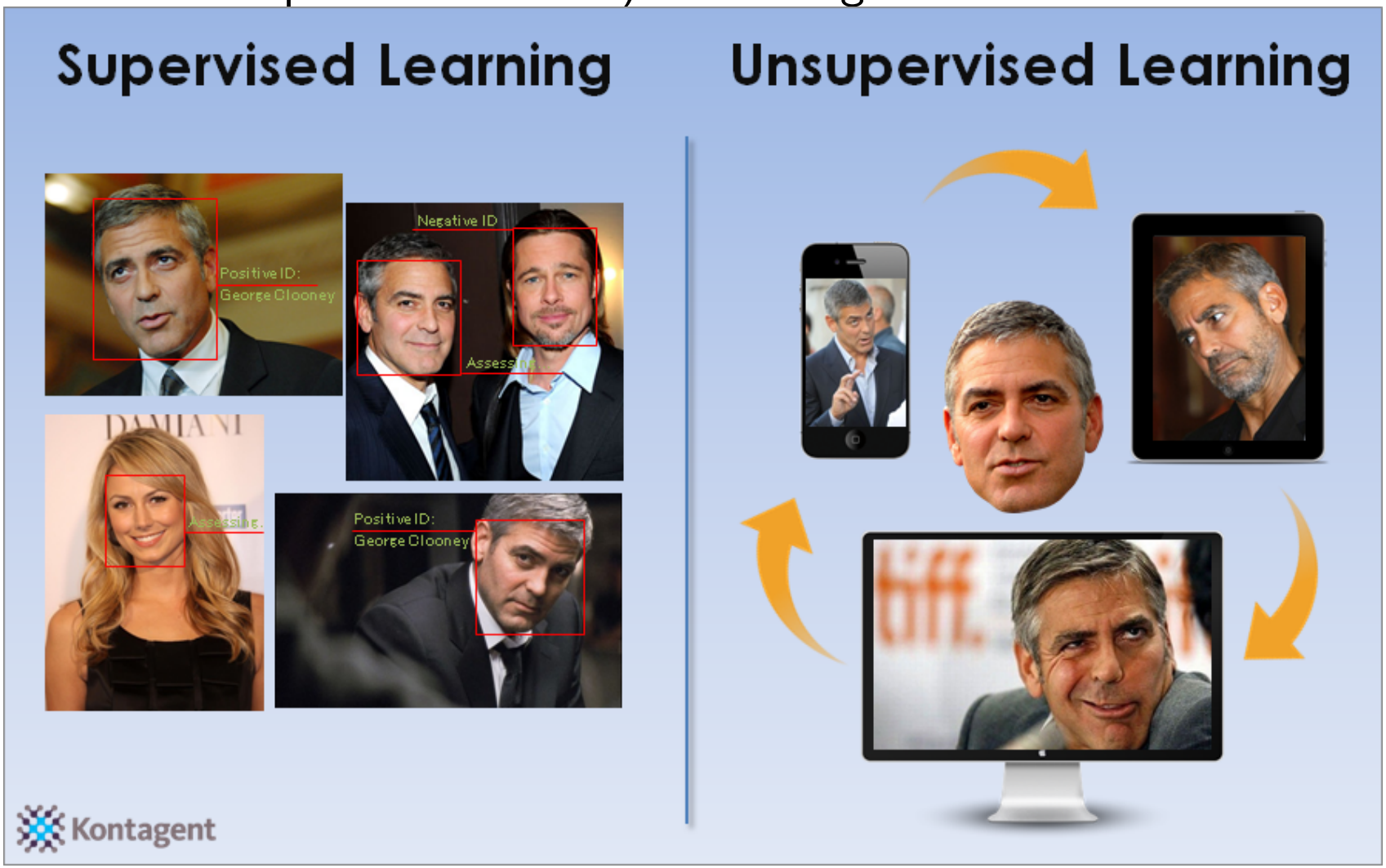
- October 2006: Netflix offers $1M for an improved recommender algorithm.
- **Training data**:
  - 100M ratings
  - 480K users
  - 17,770 movies
  - 6 years of data: 2000-2005
- **Test data**:
  - Last few ratings of each user (2.8M)
  - Evaluation via RMSE: root mean squared error
  - Netflix Cinematch RMSE: 0.9514
- Competition:
  - $1M grand prize for 10% improvement
  - If 10% not met, $50K annual "Progress Prize" for best improvement

More examples: how Google has changed advertising



Supervised vs Unsupervised Learning

- **Supervised**: Both inputs (features, a.k.a. covariates, a.k.a. independent variables) and outputs (labels, a.k.a. response, a.k.a. dependent variable) in training set.



An exercise

Are the following problems supervised or unsupervised problems? If they are supervised problems, are they regression (output is numerical) or classification (output is categorical) problems?

- Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
- Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.
- Problem 3: Given a database of customer data, automatically discover market segments and group customers into different market segments.

Recall the diamond and cancer examples, what are their types?

Syllabus highlights

- Our slides will mainly follow the book *An Introduction to Statistical Learning (ISLR)*, which is freely available at https://hastie.su.domains/ISLR2/ISLRv2_website.pdf. The lecture and slides are self-contained; you are not responsible for the contents in the book beyond the slides.
- We use the **Python** language. We will provide multiple Python tutorials.
- I assume that you know basic statistics.
- I will communicate through **Blackboard**.

Sullabus highlights (cont)

What this course is about?

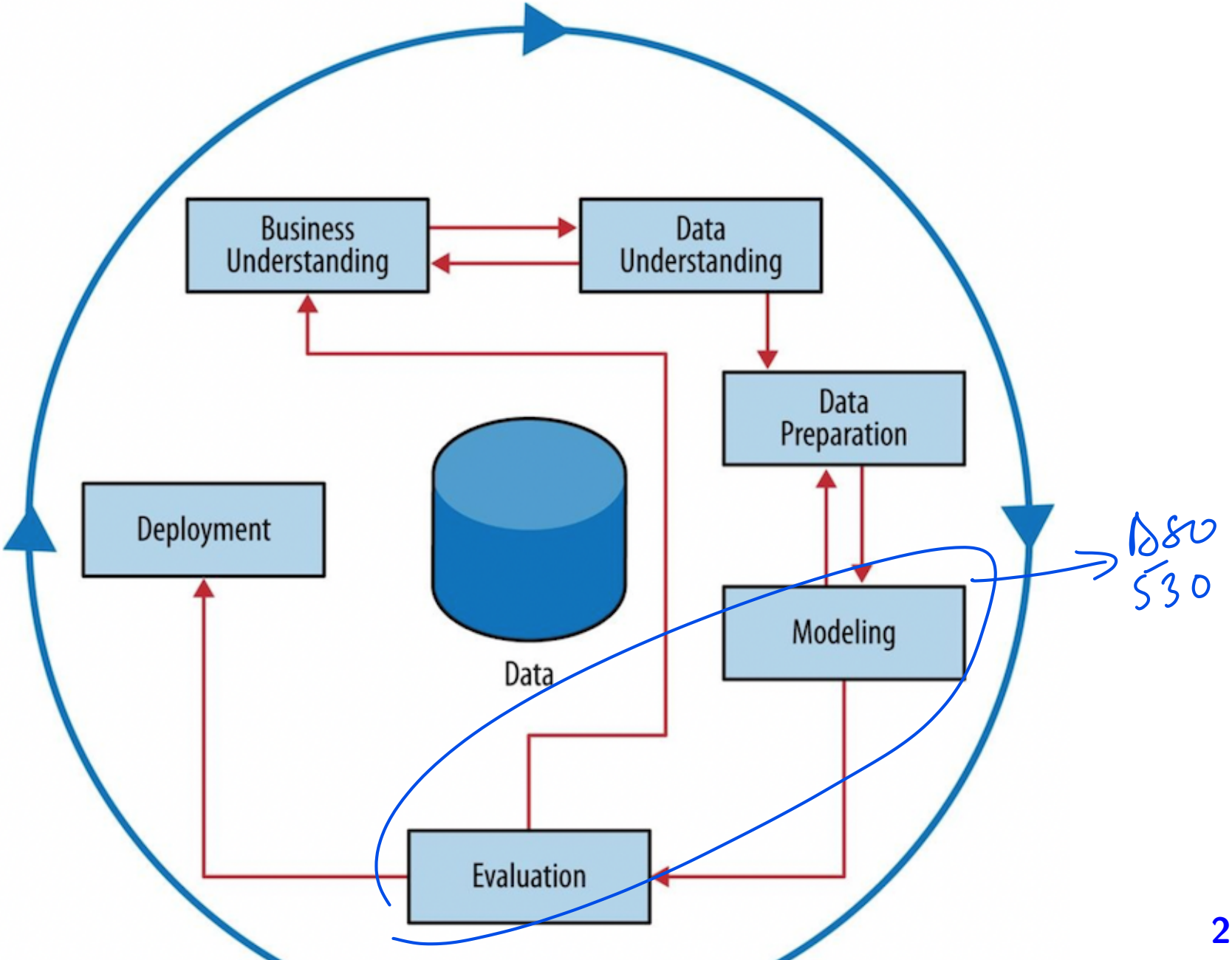
- We will cover a wide range of topics in statistical/machine learning, including LASSO, boosting, missing data, asymmetric group sizes in classification, etc.
- We will try to understand on an intuitive level on how these algorithms work.
- This course is not a theory course, but minimum math/prob notations are unavoidable.
- When we learn an algorithm, we pick up not only its execution in Python, but also why it is considered for a given problem, how the algorithm is trained behind the scenes, etc. In other words, we strive for a more comprehensive understanding of statistical learning algorithms.

Syllabus highlights (cont)

What will your course grade be based on?

- **in-class test I 29%** in the evening of **TBA**, open notes
- **in-class test II 40%** in the evening of **TBA**, open notes
- **quiz 2%** in class quizzes will be assigned
- **assignments 16%** there will be four assignments, and problems include both theoretical and coding questions. One submits assignments individually, although discussion is encouraged
- **group project 13%** each group will be of size (up to) 6. I will assign the project after Test 1. Each group will hand in a project report and deliver a 10-15 min presentation. Evaluation criterion and due dates will be sent out later. Students also have the freedom to come up with their own project.

Cross Industry Standard Process for Data Mining



Achievements after taking DSO530

- Understand the most popular machine learning algorithms: linear regression, logistic regression, LASSO, random forest, support vector machines, boosting, neural networks, etc.
- Work through examples in real estate, marketing, finance, etc.
- Be able to select proper algorithms for a given problem
- Understand the limits of machine learning algorithms
- Get (more) familiar with the language **Python**

Components of the Course

- Component 1: Lectures cover fundamentals of statistical learning. Think actively during class and review slides after class.
- Component 2: Python tutorials provide implementation of the algorithms and real case studies. Practice (instead of read) the tutorials.
- Component 3: Ungraded exercises and graded homework.
- Component 4: A group project.
- Component 5: Two tests. Opportunities to connect all dots.

Expectations

- Join (Zoom) sessions on time and be ready to learn.
- Review contents after each lecture. Memory fades quickly if you do not refresh the contents.
- Submit homework and project on time. You hear, you forget; you see, you remember; you do, you understand
- Don't expect to learn a universal pipeline that works for all problems. Develop a more mature mindset about statistical learning
- Ask in class/Come to office hours/make appointments if you have any questions
- Depending on your background, the course load might be on the **heavy** side
- Follow "the University Student Conduct Code"
- A tip: download the slides to your tablet and take notes on slides.
- Treat each other with respect

Expectations

- Join (Zoom) sessions on time and be ready to learn.
- Review contents after each lecture. Memory fades quickly if you do not refresh the contents.
- Submit homework and project on time. You hear, you forget; you see, you remember; you do, you understand
- Don't expect to learn a universal pipeline that works for all problems. Develop a more mature mindset about statistical learning
- Ask in class/Come to office hours/make appointments if you have any questions
- Depending on your background, the course load might be on the **heavy** side
- Follow "the University Student Conduct Code"
- A tip: download the slides to your tablet and take notes on slides.
- Treat each other with respect

To do list in the first week

- Review lecture 1 and syllabus
- Practice Python Tutorial 1 posted on Blackboard
- Go over the four technical notes for basic probability and statistics posted on Blackboard; come to office hours if you have any questions.
- If you forget Python, please review the contents of DSO 545