MATH 4323 Data Science and Statistical Learning

W. Wang[1]

[1] Department of Mathematics
University of Houston

Aug. 22, 2023

Course Information

- Instructor: Dr. Wenshuang "Wendy" Wang
- Office: Fleming 11D
- Office Hours: Monday and Wednesday 9:30 am - 10:30 am, or by appointment
- Email: wwang51@uh.edu
- Course page: CANVAS

What Will Be Taught In This Course?

- Course will deal with applications for such statistical learning techniques as K-nearest neighbors, maximal margin classifiers, support vector machines, principle components analysis, K-means and hierarchical clustering.
- Other topics might include: algorithm performance evaluation, cluster validation, data scaling, re-sampling methods.
- R Statistical programming will be used throughout the course.
- Pre-requisite: MATH3339 Statistics for the Sciences

Learning Objectives

By the end of the course a successful student should:

- Have a solid conceptual grasp on the described statistical learning methods.
- Be able to correctly identify the appropriate techniques to deal with particular data sets.
- Have a working knowledge of *R* programming software in order to apply those techniques and subsequently assess the quality of fitted models.
- Demonstrate the ability to clearly communicate the results of applying selected statistical learning methods to the data.

References

While lecture notes will serve as the main source of material for the course, the following books constitute great references:

- "An Introduction to Statistical Learning (with applications in R)" by James, Witten et al. ISBN: 978-1-4614-7137-0
- "Machine Learning with R" by Brett Lantz. ISBN: 978-1-78829-586-4
- **Rstudio:** Make sure to download R and RStudio (which can't be installed without R) before the course starts. Use the link R download to download R first, then RStudio download to download it from the mirror appropriate for your platform.
- **New: Rstudio is in the cloud: RStudio.cloud.

Assessments

Class Participation & Labs	10%
Homework	15%
Midterm 1	25%
Midterm 2	25%
Final Group Project	25%

Class Participation and Labs:

Participation is required:

- During the lab, I will show you how to implement some of these in R as a "lab." There will be a series of questions that you will need to answer for each lecture based on these "labs".
- Remember to bring your laptop when we have labs.
- Each lecture lab is worth 10 points.
- The lecture lab will open at the time of the lecture and close one hour after the lab. Specific dates will be posted in Canvas.
- One lab assignment will be dropped (The one with the lowest grade).

Written Homework

- Written Assignments will be given several times during the semester. Tentatively due every two weeks.
- Students will submit their written homework by scanning and uploading their work in CANVAS.
- Instructions and due dates will be announced during lecture and on CANVAS.
- The lowest written assignment will be dropped. Work not submitted will be a zero at the end of the semester.
- You may discuss the problems with other classmates as you figure out how to do the problem or establish its truth, but the write-up should be done by you alone and in your own words. Otherwise, this will affect your grade.

Exams

- The exams will be on campus.
- Instructions of the exams will be given a week before the scheduled exam.
- Tentative dates of exams are as follows:

Test	**Date**
Midterm 1	Oct. 5
Midterm 2	Nov. 16

Final Group Project

- Form a group of 3 - 5 people consisting of students teaming up in the groups in BlackBoard. No later than **Sep. 8**. No more than 5 group members are allowed.
- This is a semester long group project, consisting of students teaming up (1 points), deciding on the data set of interest (3 points), posing research questions (6 points) and applying statistical learning techniques discussed in this course to address those questions (40 points).
- Detailed requirements and due dates for deciding the dataset of interest, forming research questions and submitting final report of your findings will be posted in CANVAS.
- For each part of the group project assignments, **EACH student** must submit the corresponding report in CANVAS. Work not submitted will be a zero at the end of the semester.

Tentative Grading Scale:

93% and above	A	at least 73% and below 77%	C
at least 90% and below 93%	A-	at least 70% and below 73%	C-
at least 87% and below 90%	B+	at least 67% and below 70%	D+
at least 83% and below 87%	B	at least 63% and below 67%	D
at least 80% and below 83%	B-	at least 60% and below 63%	D-
at least 77% and below 80%	C+	below 60%	F

Late Assignment Policy:

This is a course in which you need to keep up with the materials and assignments as best as possible.

If you turn in an assignment late, the grade will be reduced by 5% every day past the due date.

Let the grader know that you turned in the assignment late. Otherwise it may be overlooked and will not be graded.

Tentative Course Outline:

- **Review: Task of Statistical Learning.** Supervised and unsupervised learning.
- **Support Vector Classifier.** Maximal margin classifier: separating hyperplane, support vectors. Non-separable case: support vector classifier.
- **Support Vector Machines.** Non-linear decision boundaries. Kernels. One-versus-one and one-vs-all classification for $K > 2$ classes. Evaluating quality of classification.
- **Clustering Methods: K-Means.** Within-cluster variation. Computing centroids. Multiple starts. Selecting K.
- **Clustering Methods: Hierarchical.** Agglomerative clustering. Linkage. Interpreting dendrogram. Choice of dissimilarity measure. Data scaling.
- **Evaluation of Clustering Solution.** Is this a good clustering? Variance explained. Between- and within-cluster variation. Silhouette coefficient.

Excused Absence Policy:

- Regular class attendance, participation, and engagement in coursework are important contributors to student success.
- Absences may be excused as provided in the University of Houston Undergraduate Excused Absence Policy for reasons including: medical illness of student or close relative, death of a close family member, legal or government proceeding that a student is obligated to attend, recognized professional and educational activities where the student is presenting, and University-sponsored activity or athletic competition.
- Additional policies address absences related to military service, religious holy days, pregnancy and related conditions, and disability.

Other Information

Recording of Class (University Policy)

- Students may not record all or part of class, livestream all or part of class, or make/distribute screen captures, without advanced written consent of the instructor. If you have or think you may have a disability such that you need to record class-related activities, please contact the Center for Students with DisABILITIES. If you have an accommodation to record class-related activities, those recordings may not be shared with any other student, whether in this course or not, or with any other person or on any other platform. Classes may be recorded by the instructor. Students may use instructor's recordings for their own studying and notetaking. Instructor's recordings are not authorized to be shared with anyone without the prior written approval of the instructor. Failure to comply with requirements regarding recordings will result in a disciplinary referral to the Dean of Students Office and may result in disciplinary action.

Other Information

Syllabus Changes

- Please note that the instructor may need to make modifications to the course syllabus. Notice of such changes will be announced as quickly as possible through UH email.

Other Information

Resources for Online Learning

- Check the Power-On website for a comprehensive set of online learning resources, tools, and tips including: obtaining access to the internet, AccessUH, and Blackboard; Requesting a laptop through the Laptop Loaner Program; and downloading Microsoft Office 365 at no cost.
- For questions or assistance contact UHOnline@uh.edu.

Other Information

Honor Code Statement (University Policy)

- Students may be asked to sign an honor code statement as part of their submission of any graded work including but not limited to projects, quizzes, and exams: “I understand and agree to abide by the provisions in the University of Houston Undergraduate Academic Honesty Policy. I understand that academic honesty is taken very seriously and, in the cases of violations, penalties may include suspension or expulsion from the University of Houston."

Other Information

Helpful Information:

- COVID-19 Updates: https://uh.edu/covid-19/
- Coogs Care: https://www.uh.edu/dsaes/coogscare/
- Laptop Checkout Requests: https://www.uh.edu/infotech/about/planning/off-campus/index.phpdo-you-need-a-laptop
- Health FAQs: https://uh.edu/covid-19/faq/health-wellness-prevention-faqs/
- Student Health Center: https://uh.edu/class/english/lcc/current-students/student-health-center/index.php

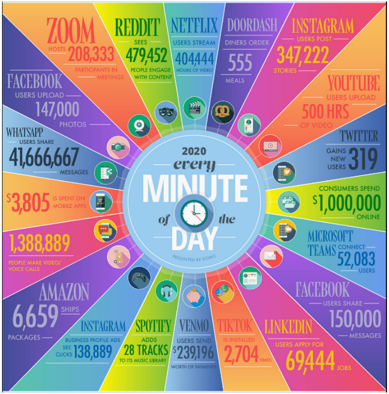
Email policy

(Required!) Include “MATH 4323” as well as a searchable description of the issue in the subject line for ALL course-related email correspondence.

Send a follow-up email if I do not respond to your email within two working days.

Why Data Science?

- Companies become digital with internet and cloud computing serving as the backbone of their establishment.
- Huge amount of data become available.
- It became mandatory for those companies to handle their data carefully.

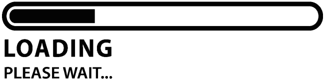


Why Data Science?

- The world has changed dramatically over the past decades and the economy has changed with it.
- Data is now the "fuel" that drivers business

Big Data: How Big is "Big"?

- **"Big"** is a **moving target.** (both over **time** and **application field**).
- Constructing thresholds such as "Big Data is $\geq$ 1 petabyte" or "Expensive campaign is 100$M\$$" is meaningless.
- **"Big"** - when **size becomes a challenge**.



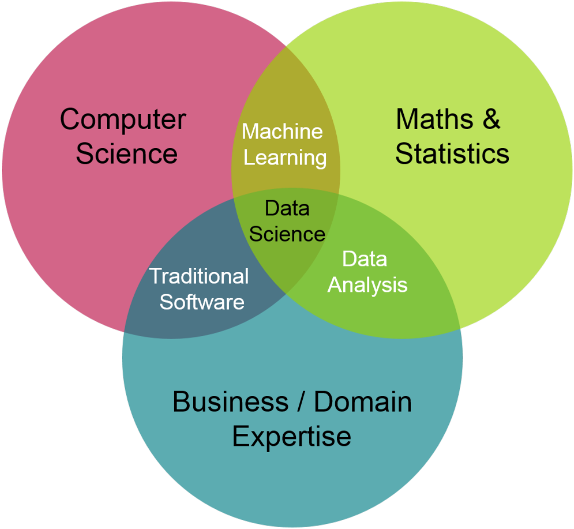
$$\Longrightarrow$$

have to learn a new host of tools.

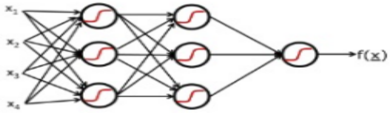
What is Data Science?

- "Data science is the application of statistical and computational techniques to solve problems in the real world."
- "A multi-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data."
- "A concept to unify statistics, data analysis, machine learning, and their related methods in order to understand and analyze actual phenomena with data."

What is Data Science?



What Data Scientists Do? Tech Companies.

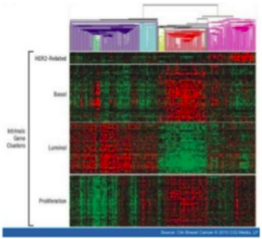


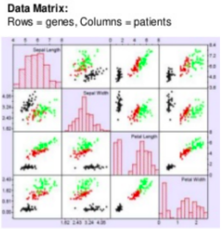
From Le Cun and Ranzato 2013

- **Face Recognition**: How does Facebook automatically tag people in photos?

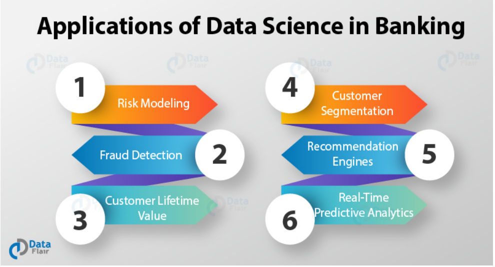
- **Lyft**: How do Lyft, Uber decide on surge pricing, promos?

What Data Scientists Do? Besides Tech Companies.




Data Matrix:
Rows = genes, Columns = patients

- **Personalized Medicine**: How does MD Anderson Cancer Center use genomics data to personalize treatment?


Applications of Data Science in Banking

1. Risk Modeling
2. Fraud Detection
3. Customer Lifetime Value
4. Customer Segmentation
5. Recommendation Engines
6. Real-Time Predictive Analytics

- **Banking**: How does bank manage their resources efficiently?

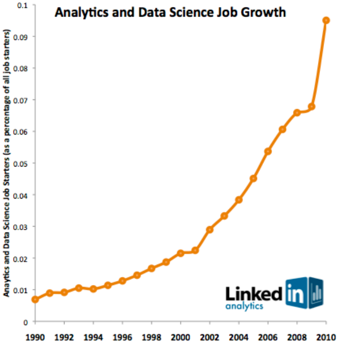
What Data Scientists Do? Energy and Utilities.

- **Outage detection and prediction**: Predict the influence of weather conditions on the power grid; Detect outages in specified areas; Real-time filtering of outage inputs and recognition of the outage type, etc.

- **Real-time customer billing**: How does utility company improve their customer service and satisfaction rates?

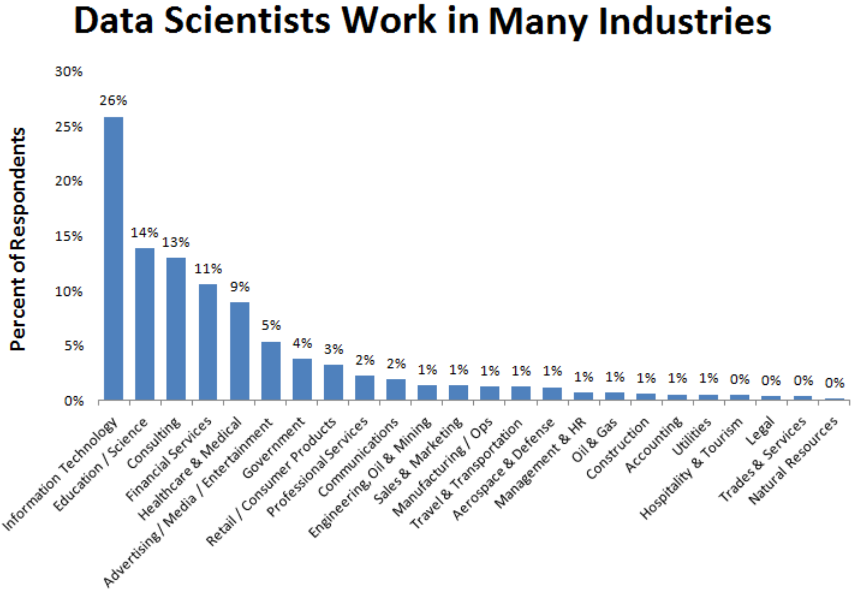
Might not be a real field, but has REAL JOBS.

"So even if data science isn't a 'real field', it has REAL JOBS."
(R. Schutt, C. O'Neil, "Doing Data Science")



Linkedin's 2019 report on the most promising jobs (`https://blog.linkedin.com/2019/january/10/linkedins-most-promising-jobs-of-2019/`)

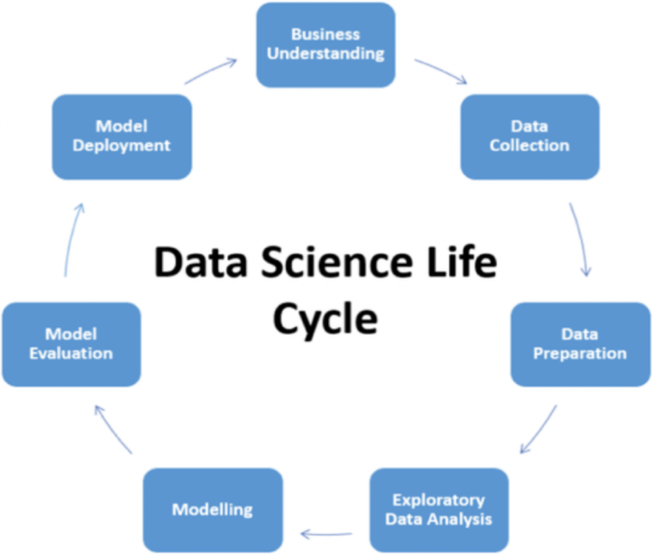
Might not be a real field, but has REAL JOBS.



Life cycle of Data Science

- Data Acquisition: There is no Data Science without data; Data may come from difference resources.
- Data pre-processing: How to deal with missing values; What if suspicious outliers or extreme values?
- Exploratory data analysis: Data visualization/summarization/basic reporting.
- Model building: Depending on the task, choose from supervised learning, unsupervised learning, reinforcement learning; Various models available for each type of learning methods.
- Evaluation of the models: They all do the job, which one to choose?
- Deployment of model: Integrate the final model to make practical business decisions; Visualization of the findings; Storytelling.

Life cycle of Data Science



Skills needed to do data science

Data Analyst:

- Data analysis and forecasting using Excel
- Data querying using SQL
- Communication: Creating dashboards using business intelligence software
- Some knowledge of mathematical statistics

Data Scientist (on TOP of Data Analyst skills of duties):

- Programming: R or Python
- Statistical analysis using machine learning algorithms such as natural language processing, support vector machine, random forest, or deep learning
- Automation techniques to simplify daily processes and for other members of their organization

Data Science: **Skills/responsibilities**.

Besides **expert mastery** of **Data Analyst** qualifications, **Data Scientist**:

1. **Formulates questions**, translates them into **math problems**.
2. Possesses **data intuition, curiosity, critical thinking**.
3. **Develops, implements** and **tests** a **ML algorithm/Stat. models**, used for:
   - **prediction**, or
   - **explanation** task.

$\Longrightarrow$ **DATA PRODUCTS**.

Who will fit Data Science?

You are if you

- hold a degree in n Mathematics, Statistics, Computer Science, Management or related area
- have a problem-solving attitude
- can understand business and customer needs
- be able to communicate in business language to visualize your data and results

Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working. simpler methods often perform as well as fancier ones!
- This is an exciting research area, having important applications in science, industry and finance.

Final Thoughts.

1. **Data Science** can be used **everywhere**. It doesn't take a fully-fledged data scientist to use data science.
2. Not a specialization, but a **state of mind** which can be applied to **any discipline**..
3. Adding a data-related project on your resume rarely hurts.