Intro to Data Mining Lecture 1a

Intro to Data Science and Syllabus

Created by Jon Witkowski on 12/28/2023

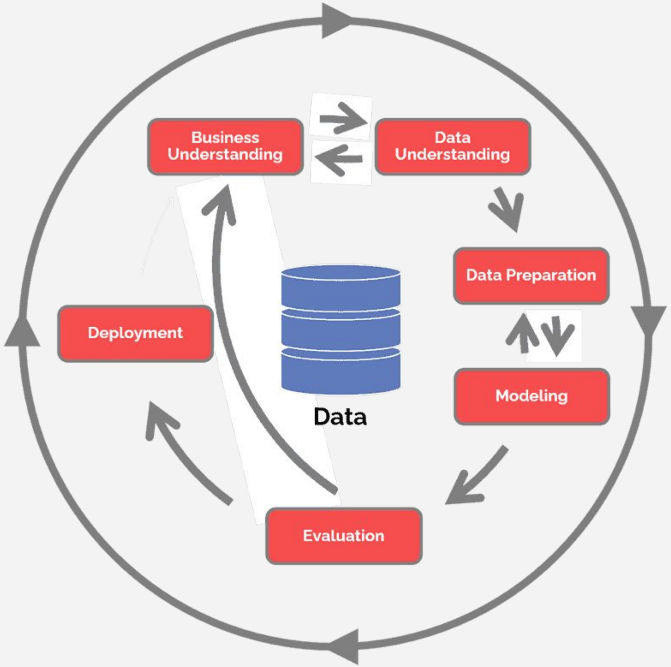
Why are we here?

- Data is being collected at an unprecedented rate. We are living in the age of data
- Every beep you hear at a supermarket register is a barcode being read and a purchase being entered into a database
- Every web site you enter or tweet you send is being recorded in a database somewhere
- The problem today is not lack of data, (we have too much in many cases), but the lack of human trained data analysts that can make sense of the data and turn it into knowledge
- With all of the tools out there, data mining is easy to do badly

What is data mining?

- The process of learning from data by observing and discovering patterns in large sets of data through the utilization of machine learning, statistics, linear algebra, calculus, etc.
- According to the Cross-Industry Standard Process (CRISP-DM), developed in 1996, Data Mining has 6 phases:

CRISP-DM



- The purpose of this order is so that data mining is iterative and adaptive
- Human intervention and understanding of the process is necessary in order to yield desired results.

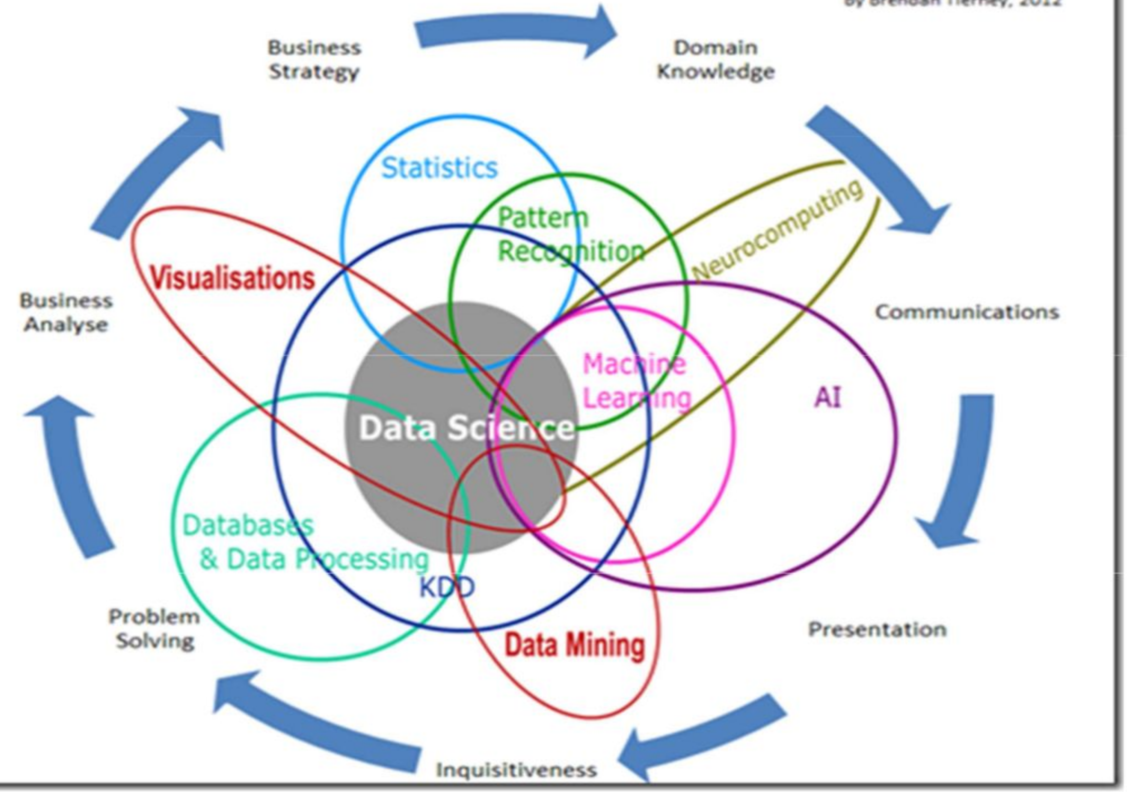
How is data mining used?

- How does Amazon suggest things you might like?
  - Data Mining
- How can credit card fraud be detected?
  - Data Mining
- How do lenders know whether to accept applications?
  - Data Mining

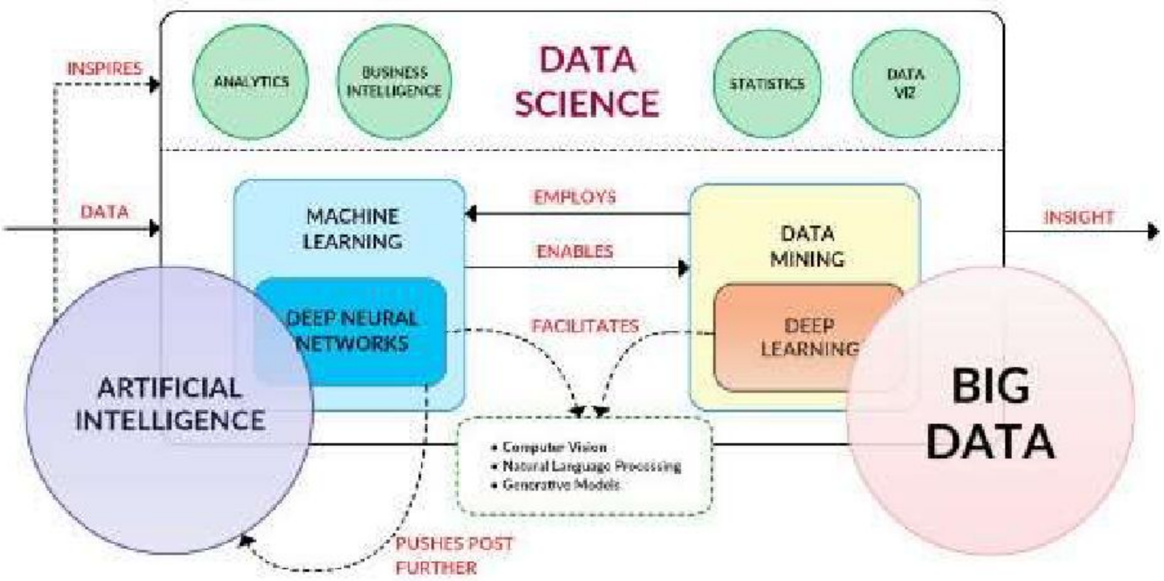
You get the point. Basically anything that involves data also probably involves data mining. If there are patterns to find, they will be found and used.

Data Science Is Multidisciplinary

By Brendan Tierney, 2012



Interactions Between Data Science, Machine Learning, Data Mining



Titles

- Data Mining – take a data set and try to answer a question. Uses tools from machine learning like neural networks, clustering, etc.
- Data Scientist does Data Mining to solve problems
- Data Analyst creates charts, graphs, reports. (Usually doesn’t program)
- Data Engineer – sets up data streams and maintains data warehouse
- Machine Learning Engineer – operationalizes models that get used over and over. Also does data scientist stuff.
- Machine Learning Scientist – Invents new algorithms or enhances existing algorithms (e.g. Neural Nets invented in 1958, but Deep Learning Neural Nets invented in 2013)

Titles

- 20 years ago there was no such thing as a Data Scientist
- There was data mining which belonged to Computer Science
- Statistics which belonged to well, statistics/math
- Machine learning which belonged to CS or Computer Engineering
- 10 years ago, Data Scientists emerged which were some combination of the above.
- More recently the roles have been more well defined and now we have data engineers, machine learning engineers, etc.
- Many companies still use Data Scientist as a catchall

Python

- In 2020 among Data Science/Data Engineer job listings, Python outnumbered R about 10 to 1.
- In 2016 R was the dominant language.
- As we will see, Sci-Kit learn is a much better ecosystem for running data models, and has contributed to Python overtaking and passing R.
- Last semester I tried to show Python and R to the class and nobody really cared about R, so we will just be using Python for everything

About this course

- Data Mining is a huge field.
- Rowan has 3 courses on it (DM 1, DM 2, and Text Mining) and there’s still so much that isn’t covered.

- In this course, we will cover some fundamentals with basic types of classification models, regression, neural networks, clustering, how to clean data, and some dimensionality reduction amongst other topics.
- By the end of the semester, you should have a good enough foundation of data mining that you could learn on your own and understand the reasoning behind why certain methods are used.

About this course

- We will have 10 assignments and 2 exams throughout the semester, with each homework being worth 3.5% of your grade, each programming assignment being worth 7.25% and each exam being worth 20% and the other 10% being attendance and participation.
  - The 10 assignments are split up into 6 homework assignments and 4 larger programming assignments
  - Each homework assignment will have several smaller questions, while the programming assignments will often be implementations of methods we discuss in class
- For the midterm and final, you will have a week to start the exam and then 2 hours to complete once you start. **Do not start the exams less than 2 hours before the submission deadline.**

Class Policies

- I don't accept late work. It's better to turn in an incomplete assignment than to not turn something in at all.
- Follow the university's academic integrity policy. Last semester I caught several students using ChatGPT on their assignments. It's a lot easier to notice than you think. Please do your own work.
- We have class from 6:30 - 9:15. I know that's a good time to eat dinner. You can eat in my class; just don't be distracting.
- Show up to class. I take attendance and it would suck to not get the grade you want because you didn't show up enough.
- Submit your programming assignments as .ipynb files. I grade everything the day after it's due and keeping everything in the same format makes it easier to grade everything quicker.
- If you have questions about your grade, ask. I can explain my reasoning so you understand, or I can see that I made a mistake and can adjust your grade.
- Don't use your phone during class. I can see it. It sucks. Do you like when the person you're talking to isn't paying attention to you? I don't.

Grading

Category	**Description**	**Percentage/Points**
Homework	6 homework assignments, 3.5% each	21%
Attendance and Participation	Show up each week and participate	10%
Programming Assignments	4 of them, 7.25% each	29%
Midterm	Will be online	20%
Final	Will be online	20%
Total		**100%**

Grading

Letter Grade	Range
A	[100-93]
A-	(93-90]
B+	(90-87]
B	(87-84]
B-	(84-80]
C+	(80-77]
C	(77-74]
C-	(74-70]
D+	(70-67]
D	(67-64]
D-	(64-60]
F	(60-0]