CSCI 347, Data Mining

Celia Schahczenski

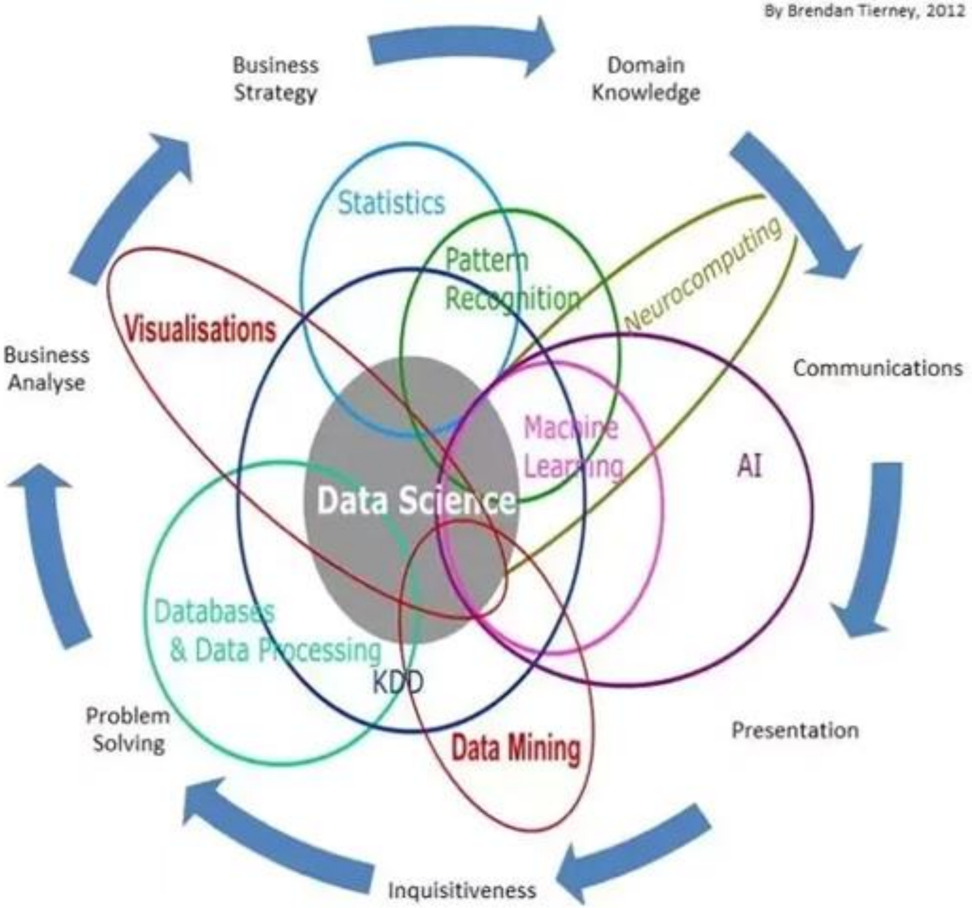
Data Mining

Term first came into widespread use in the mid to late 1990s

Originally - data analysis approach that sought to empower business people to explore and model data, without requiring extensive training in statistics.

Characterized by use of machine learning techniques, testing performance (testing is necessary because data mining methods do not adhere to the assumptions required by statistical theory) and statistics, and special tools that featured visual programming interfaces.

Data Science Is Multidisciplinary



https://www.simplilearn.com/data-science-vs-data-analytics-vs-machine-learning-article

Data Mining vs Data Science

Data mining process of discovering patterns in large datasets involving methods of machine learning, statistics, and database systems to identify future patterns

Data Science field of study that includes everything from Big Data, analytics, data mining, predictive modeling, data visualization, mathematics and statistics

Data Mining vs Data Science

Area	Data Mining	Data Science
What is it?	A technique	An area
Focus	Business process	Scientific study
Goal	Make data more usable	Building data-centric products for an organization
Output	Patterns	Varied
Purpose	Finding trends previously not known	Social analysis, building predictive models, unearthing unknown facts, and more
Deals with (type of data)	Mostly structured	All forms of data – structured, semi-structured and unstructured

Data Mining vs Other Areas

Statistics concerned with probabilistic models, specifically inference on these models using data

Machine Learning designing algorithms that can learn from and make predictions on the data, less concern about parameter estimates than statistics and focuses on computational efficiency and large datasets

Data Mining applied machine learning and statistics discovering hidden patterns or unknown knowledge which can be used for decision making

AI anything that is concerned with intelligence in computers, can be seen as a superset of machine learning and data mining

H1N1 Flu

- 2009 H1N1 flu virus outbreak
- Center for Disease Control and Prevention (CDC) requires doctors to report new flu cases
- Useful information but 2 week delay
- Google took 50 million most common search terms, compared with a CDC list of terms, and processed 450 million different mathematical models to test search terms, comparing their predictions against actual 2007 & 2008 flu cases.
- Found a combination of 45 search terms that had a strong correlation with incidences of the flu to identify the spread of flu **in real time.**

Colleges using Data Mining

Under a Watchful Eye: How colleges are tracking students to boost graduation APM Reports, Educate series of podcasts

How One University Used Big Data to Boost Graduation Rates

Amazon - different way

- Greg Linden suggests, don't compare customers, find associations between products using item-to-item collaborative filtering (finding correlations)
- Generated significantly more sales
- Don't know **why** people buy, just know **what** (don't know **causation**, just know **correlation**)

WalMart

- LOTS of data
- Can hypothesize that prior to a hurricane sales of flashlights increase
- Found also sales of Pop-Tarts increase

- As storm approaches, stock boxes of Pop-Tarts at the front

Aviva, insurance firm

- Want to identify those who might be at higher risk of illnesses such as high blood pressure, diabetes or depression
- Collecting blood and urine samples predicts fairly well and costs around $125/person
- Found can predict from lifestyle data including hobbies, websites visited, amount of TV watched, income level for around $5/person

So successful that other insurance agencies s let clients opt-in to sharing their lifestyle information