

# DATA MINING INTRODUCTION

---

What is data mining?

Applications and techniques

“

Data is the new oil”

Clive Humby



**“Data is the new oil. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.”**

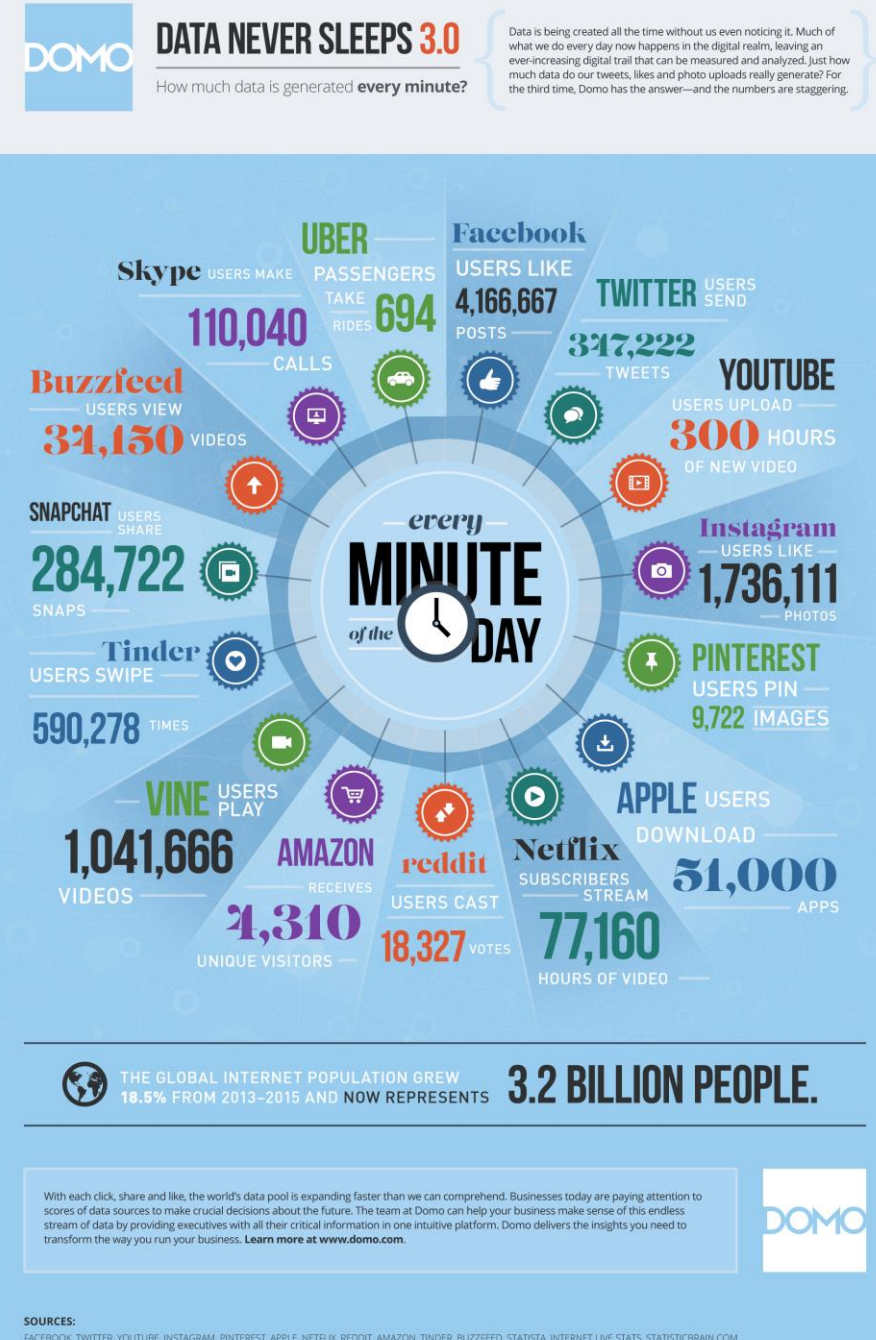
# Data Mining

- In simple terms:



# There is a lot of data

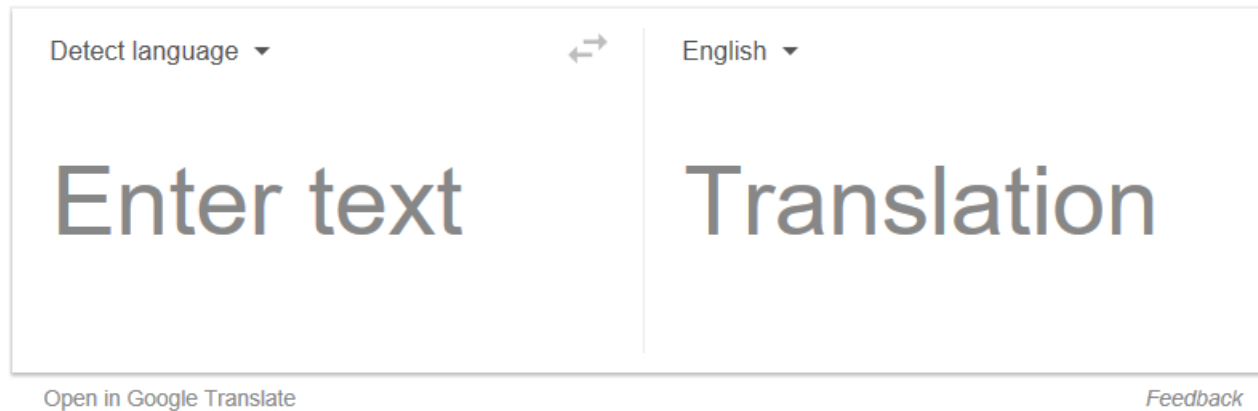
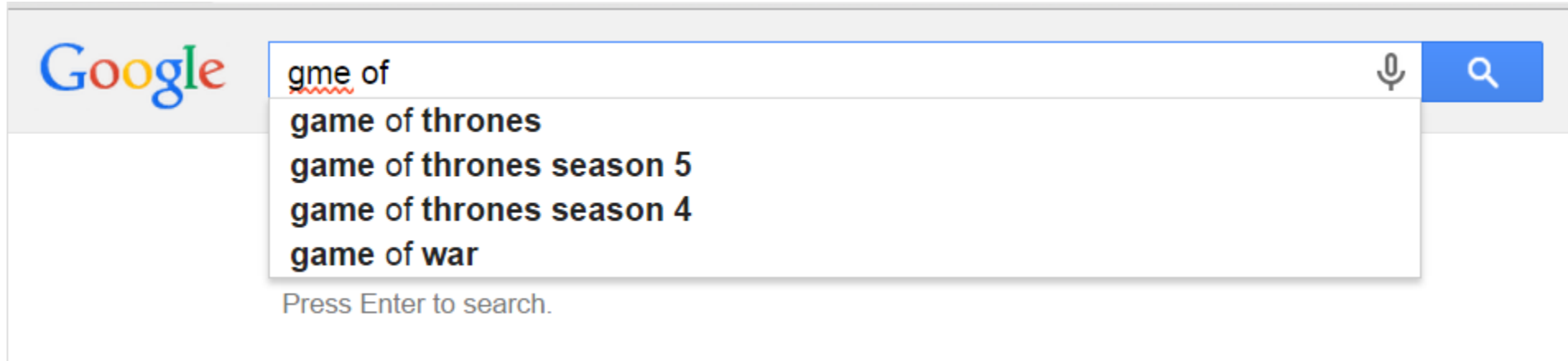
- Every human, physical, or machine activity generates data.
  - Transaction data in stores, credit cards
  - Scientific measurements
  - DNA sequences, gene coexpression
  - Health records, brain images, daily measurements
  - The Web, Wikipedia, Facebook posts, Tweets, Online Reviews
  - Queries to Google, Clicks, Browsing behavior, Ads
  - Facebook likes and comments, Twitter retweets
  - The Web graph, Facebook friends, Twitter followers
  - Movement data, Trajectories,
  - Mobile use, telephone calls
  - Wearable devices
  - Machine and workflow monitoring
- **Everybody** collects data!



# The data is complex and interconnected

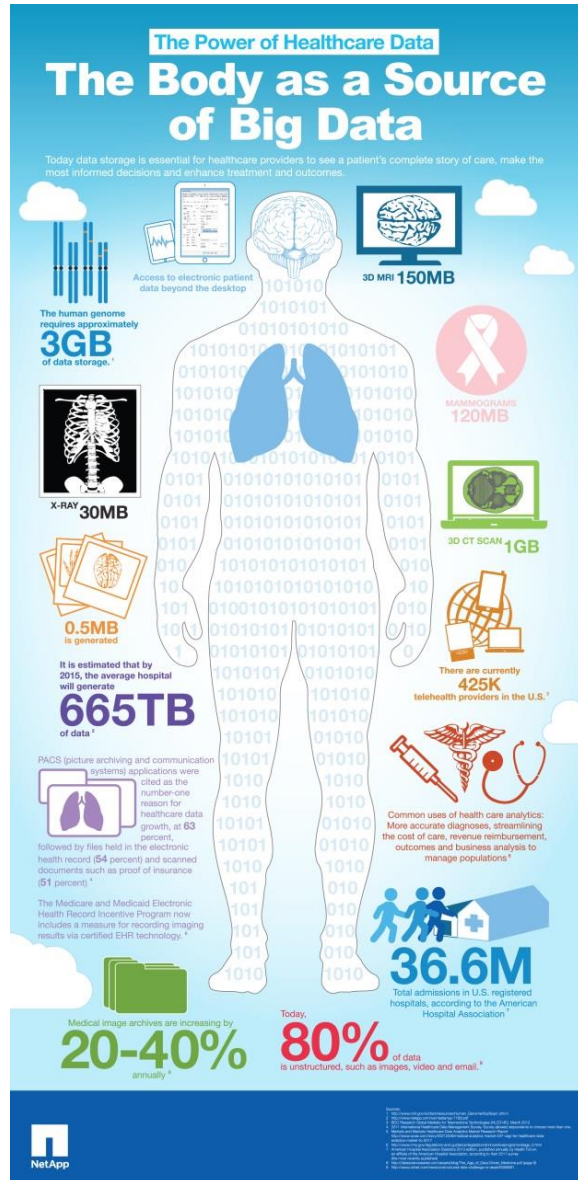
- Multiple types of data: database tables, text, time series, images, videos, graphs, etc
- Spatial and temporal aspect
- Interconnected data of different types:
  - From the mobile phone we can collect, location of the user, friendship information, check-ins to venues, opinions through twitter, status updates in FB, images through cameras, queries to search engines

# Data creates value



Natural language understanding is driven by data

# Data creates value



**Precision/Personalized medicine:**  
Find the best treatment for patients using their genotype and all data that are related to them

Also: understanding drug side-effects through google queries

# Data creates value

## The self-driving car's sensors

*Just like a person has five senses, Google's self-driving car has a variety of gadgets that detect nearby objects so it can avoid them.*

**Global Positioning System software**  
Helps car determine its location.

**Position sensor**  
Located in the wheel hub, this sensor helps determine car's location from wheel rotations.

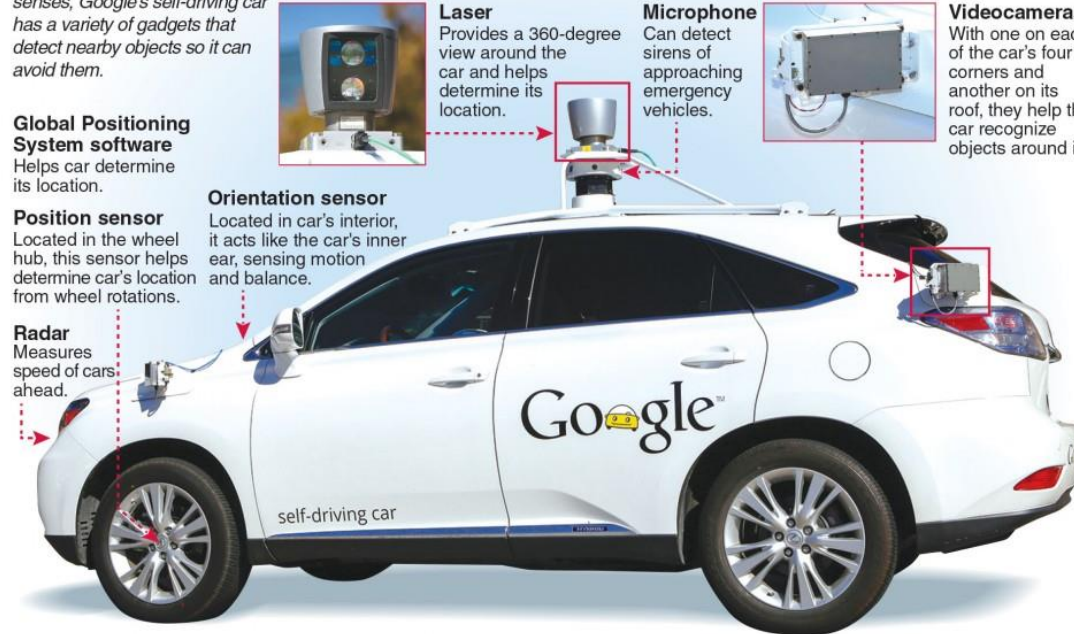
**Radar**  
Measures speed of cars ahead.

**Orientation sensor**  
Located in car's interior, it acts like the car's inner ear, sensing motion and balance.

**Laser**  
Provides a 360-degree view around the car and helps determine its location.

**Microphone**  
Can detect sirens of approaching emergency vehicles.

**Videocameras**  
With one on each of the car's four corners and another on its roof, they help the car recognize objects around it.



### How the car operates

- 1 Any object the vehicle's sensors spot is interpreted by software to determine if it's a pedestrian, cyclist, vehicle or something else.
- 2 Using what it's learned from previous driving, the software makes predictions about what objects will do next.
- 3 The software analyzes the information to decide whether it is safe to accelerate, turn or hit the brakes.



### How the car sees the world

This computerized image is what Google researchers monitoring sensor data see as they ride in the vehicle.

- Other vehicle
- Pedestrian
- Cyclist
- Objects that warrant caution
- A crosswalk, indicating the car needs to stop
- A traffic signal, warning of upcoming railroad tracks
- Path where Google's car intends to go

Source: Google  
Graphic: Tribune News Service

## Self-Driving Cars:

Car is the next computer. A future of smart cars that can drive themselves and learn from data

Also: smart cities – urban computing

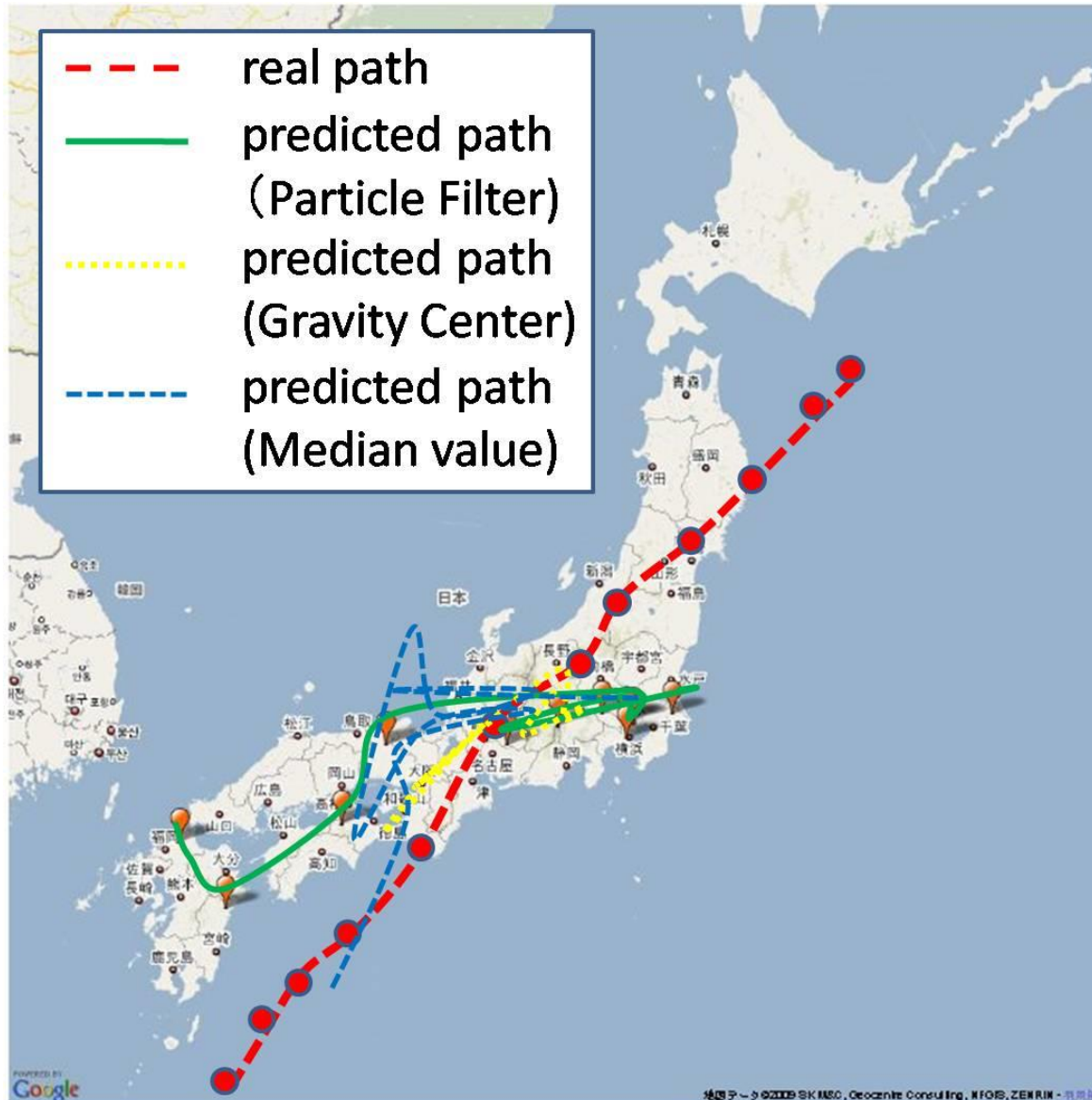
# Data creates value



Computers learn to play games by observing data



# Data creates value



Use of data for crisis management

# Data creates value

- All major soccer and basketball teams use data mining to make decisions.

The national team of Germany had a special software for the analysis of video.

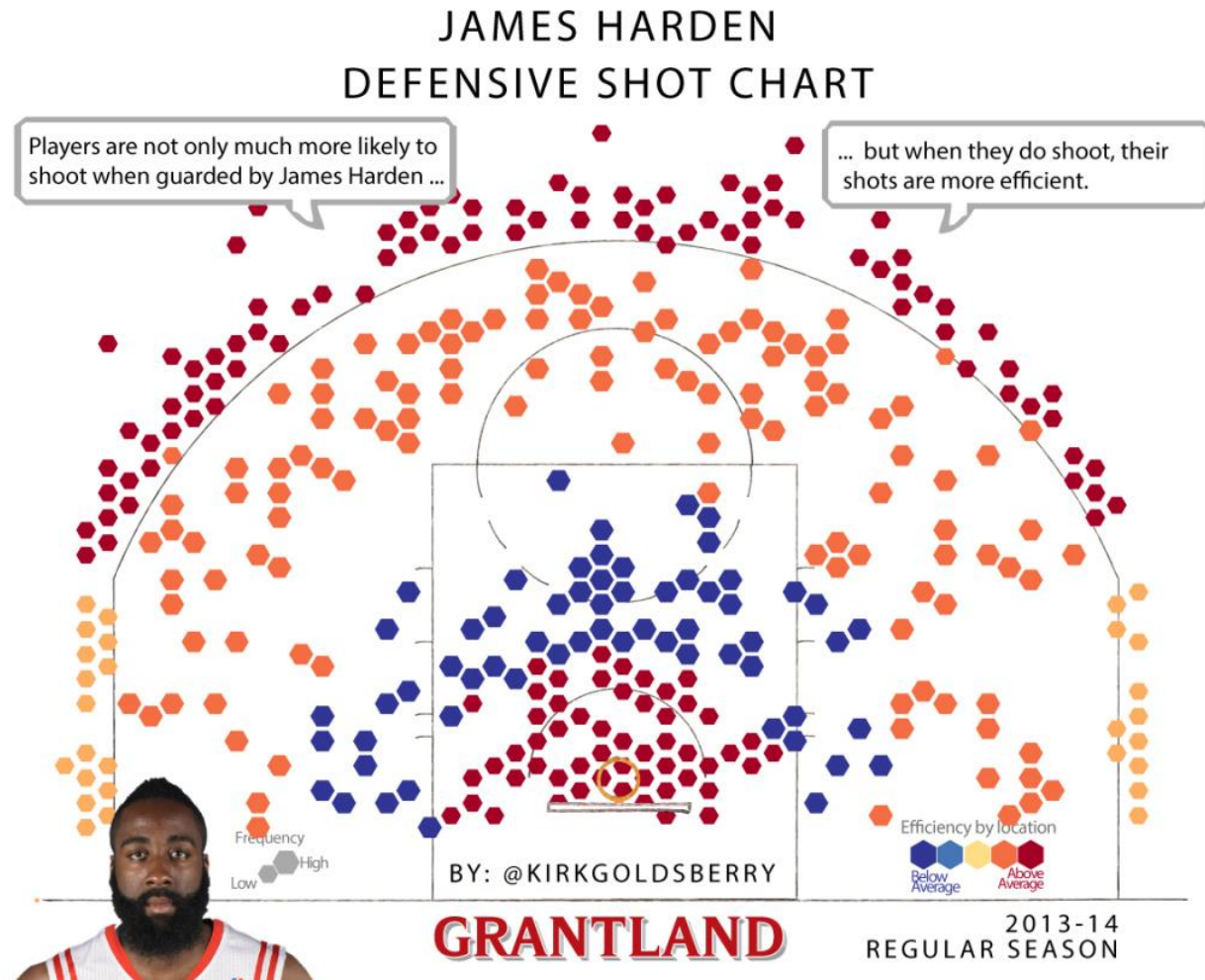
They concluded that the possession time per player should be reduced.



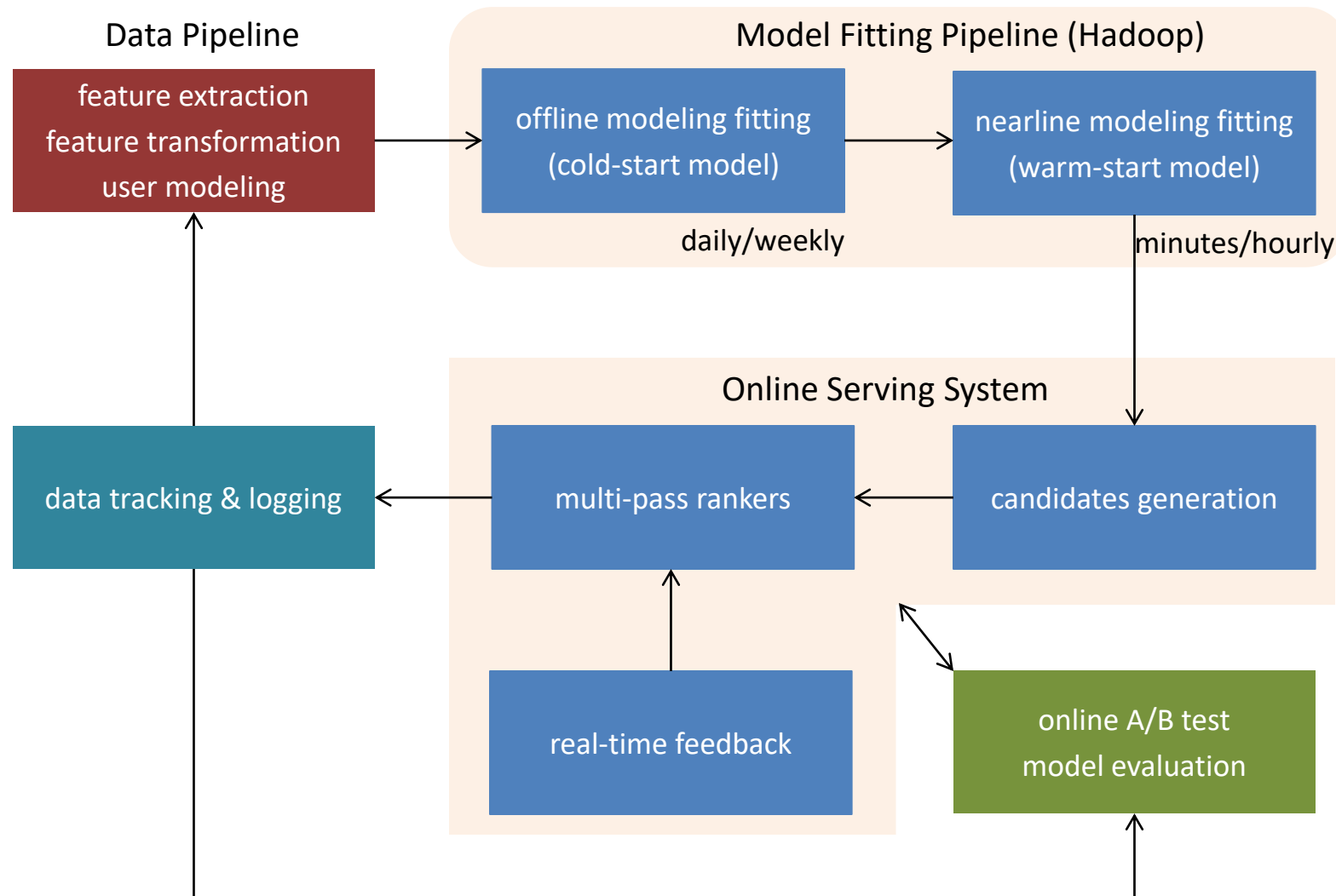
Germany won the 2014 word cup

# Data creates value

## James Harden defence



# Putting it all together: The Data Mining Pipeline (LinkedIn)



# Data Mining Example

- Suppose that you were creating the Greek Facebook.
- What kind of data would you collect and store?

Social network contacts

Posts, content of posts

Interactions with feed: Clicks, Likes, Comments, Shares

Interaction with contacts: messages, likes, replies, shares

Photos

Videos uploaded videos consumed

Demographics: Age, City, etc

Ads seen, ads clicked

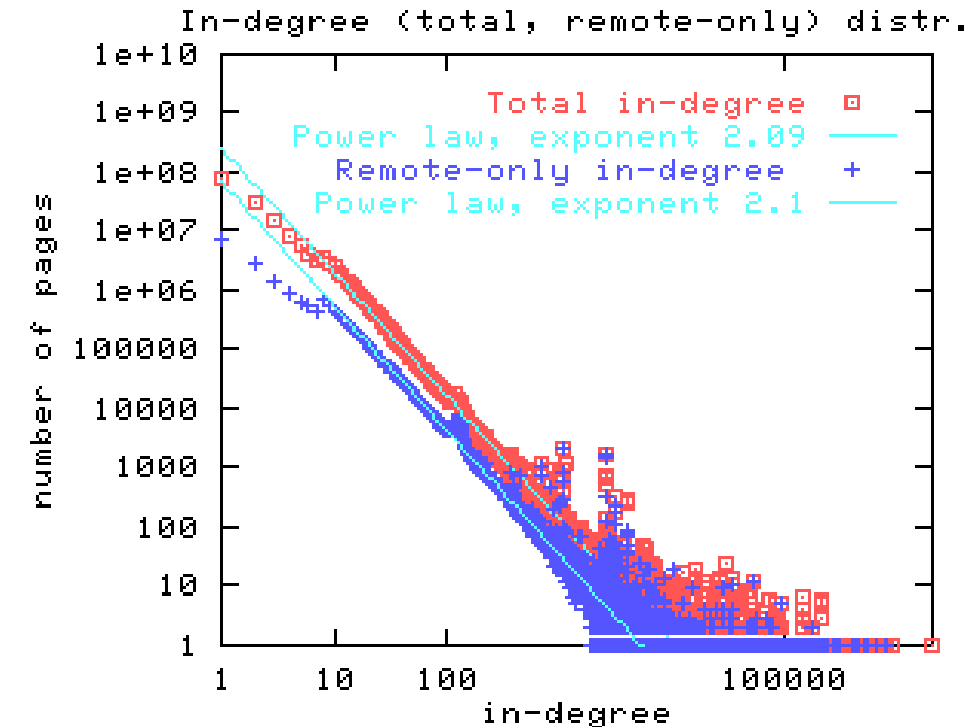
Products bought

and many more!

What would you do with this data?

# Exploratory Analysis

- Make **measurements** to understand what the data looks like
- Example: Posts
  - How often do users post, how many posts per user, when do they post, is there a correlation between number of posts and number of friends, etc
- This is one of the first steps when collecting data.
  - **Metrics**: Deciding **what to measure** is important
- The example of the Web graph



# Exploiting similarities

- Consider the following data for six users:
  - Number of times they have clicked on posts from these pages

|   | NBA | ESPN | Sports.com | MSNBC | NY Times | Wall Street | Politico |
|---|-----|------|------------|-------|----------|-------------|----------|
| A | 100 | 50   | 73         | 10    | 1        | 1           | 4        |
| B | 500 | 200  | 400        | 20    | 10       | 4           | 1        |
| C | 80  | 100  | 60         | 1     | 3        | 1           | 1        |
| D | 4   | 2    | 1          | 12    | 90       | 100         | 80       |
| E | 9   | 3    | 4          | 9     | 100      | 80          | 70       |
| F | 3   | 4    | 5          | 30    | 300      | 200         | 500      |

- What conclusion can we draw?

# Exploiting similarities

- Two types of users and two types of pages
  - Sports and politics

|   | NBA | ESPN | Sports.com | MSNBC | NY Times | Wall Street | Politico |
|---|-----|------|------------|-------|----------|-------------|----------|
| A | 100 | 50   | 73         | 10    | 1        | 1           | 4        |
| B | 500 | 200  | 400        | 20    | 10       | 4           | 1        |
| C | 80  | 100  | 60         | 1     | 3        | 1           | 1        |
| D | 4   | 2    | 1          | 12    | 90       | 100         | 80       |
| E | 9   | 3    | 4          | 9     | 100      | 80          | 70       |
| F | 3   | 4    | 5          | 30    | 300      | 200         | 500      |

- Questions:
  - How do we compute **similarity**?
  - How do we group similar users? **Clustering**

# Exploiting similarities

- What if we were missing this entry?

|   | NBA | ESPN | Sports.com | MSNBC | NY Times | Wall Street | Politico |
|---|-----|------|------------|-------|----------|-------------|----------|
| A | 100 | 50   | 73         | 10    | 1        | 1           | 4        |
| B | 500 | 200  | 400        | 20    | 10       | 4           | 1        |
| C | 80  | 100  | ???        | 1     | 3        | 1           | 1        |
| D | 4   | 2    | 1          | 12    | 90       | 100         | 80       |
| E | 9   | 3    | 4          | 9     | 100      | 80          | 70       |
| F | 3   | 4    | 5          | 30    | 300      | 200         | 500      |

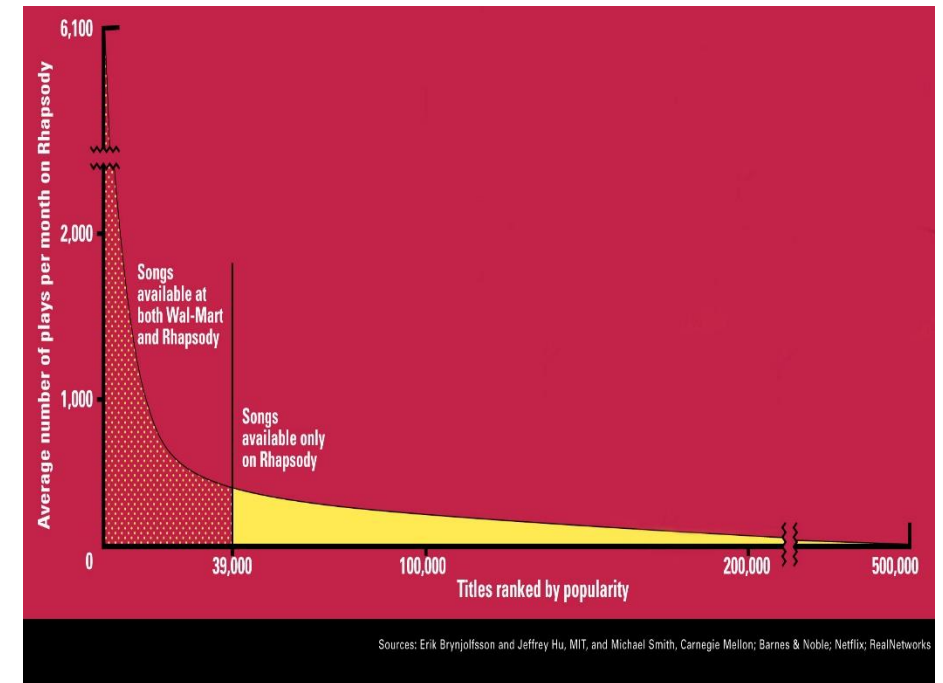
- Can we fill this value?
- Similar users like items similarly: **Recommendation** systems

# Amazon Recommendations

- “People who have bought this also bought...”



- A huge breakthrough for amazon
  - Took advantage of the long tail
- A big breakthrough for data mining in general



# Making predictions

- Filling the missing value can also be viewed as a **prediction** task
- Types of prediction tasks:
  - Predicting a real value (e.g. number of clicks): **Regression**
  - Predicting a YES/NO value (e.g., will the user click?): **Binary classification**
  - Predicting over multiple classes (e.g., what is the topic of a post): **Classification**
- Can you think of prediction/classification tasks for your social network?

Ad click prediction

Ad clickthrough prediction

Like prediction

Predict if a user will like a post over another:  
Learning to rank

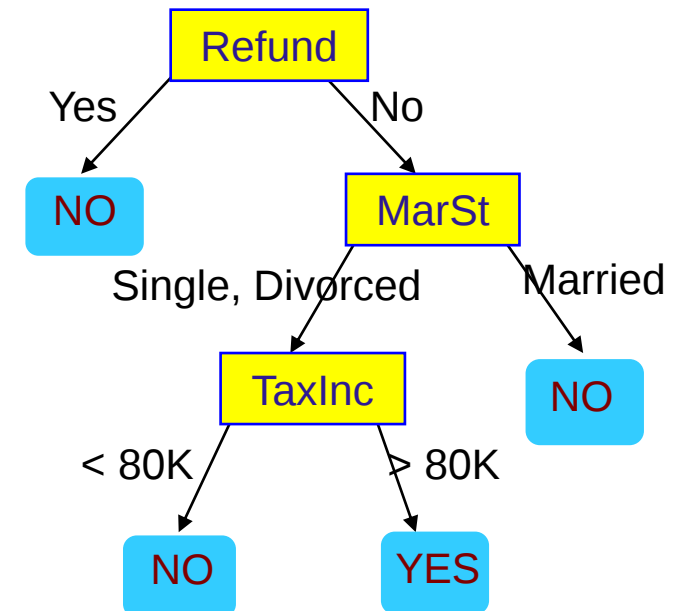
Predict if a post is offensive

Predict if a photo contains nudity

# Classification

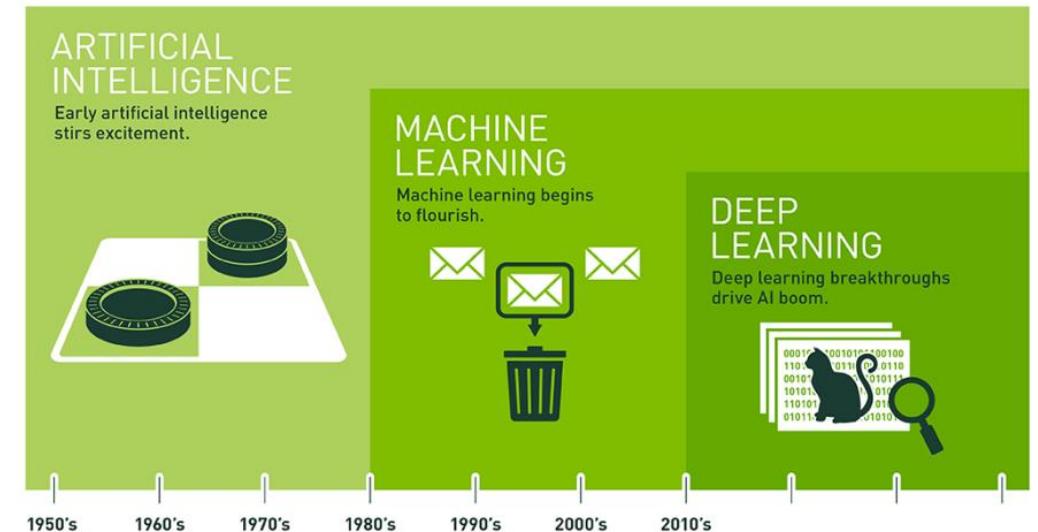
- Classification process:
  - Find **features** that describe an entity.
  - Use **examples** of the classes you want to predict.
  - **Learn** a model (function) that predicts
- Classification is the engine behind the AI revolution
  - Used in all systems that make decisions
  - Became very powerful with **Deep Learning**
  - Huge applications in vision

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1   | Yes    | Single         | 125K           | No    |
| 2   | No     | Married        | 100K           | No    |
| 3   | No     | Single         | 70K            | No    |
| 4   | Yes    | Married        | 120K           | No    |
| 5   | No     | Divorced       | 95K            | Yes   |
| 6   | No     | Married        | 60K            | No    |
| 7   | Yes    | Divorced       | 220K           | No    |
| 8   | No     | Single         | 85K            | Yes   |
| 9   | No     | Married        | 75K            | No    |
| 10  | No     | Single         | 90K            | Yes   |



# Deep learning

- Machine learning systems that use neural networks with multiple layers and are trained on **very large quantities of data**
  - Able to learn **complex representations** and **powerful models**.
  - Applications in **recommendations, network analysis, text analysis, image recognition, car driving, playing games...**
  - Require less feature engineering



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

# The social graph

- Your Greek Facebook also has a **social graph**. What can you do with this data?

Who is important and influential in the graph?

What is the shortest path between two nodes?

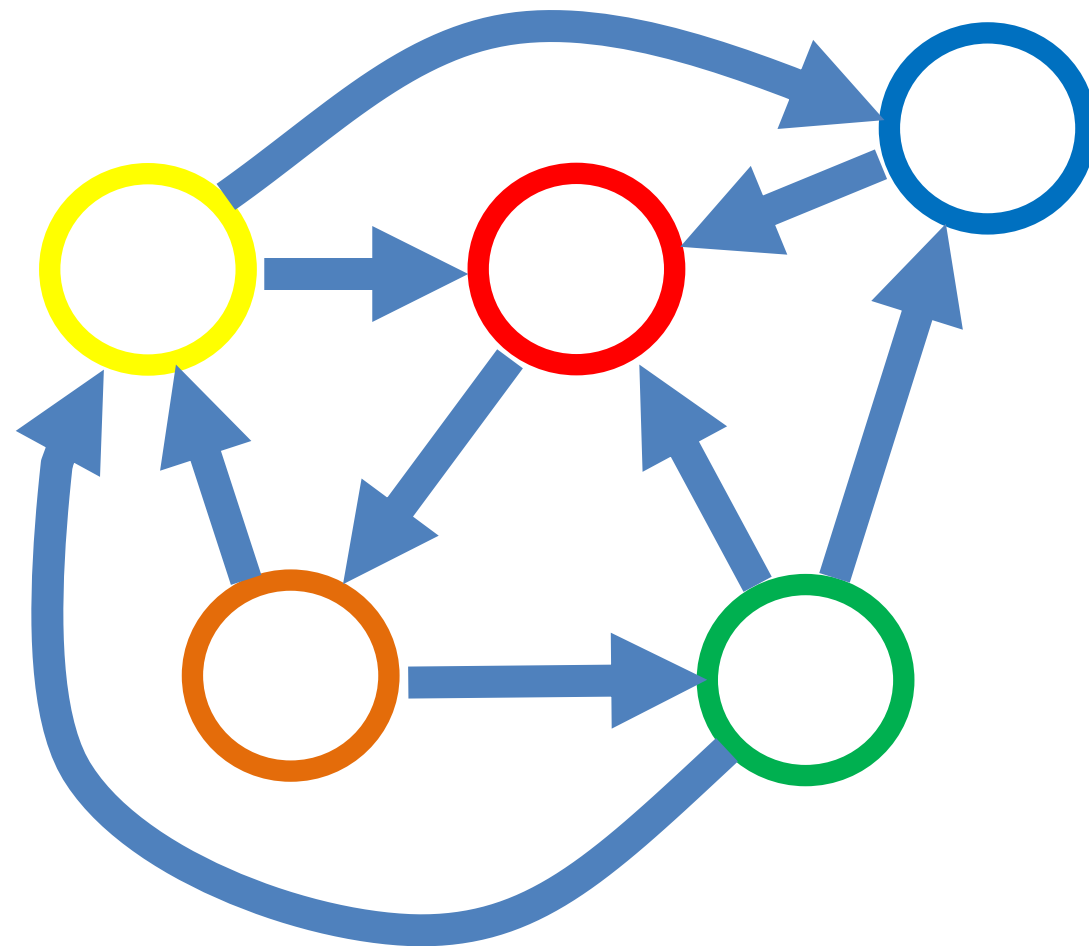
How does information spread in the network?

What becomes viral?

Will two users become friends in the future?

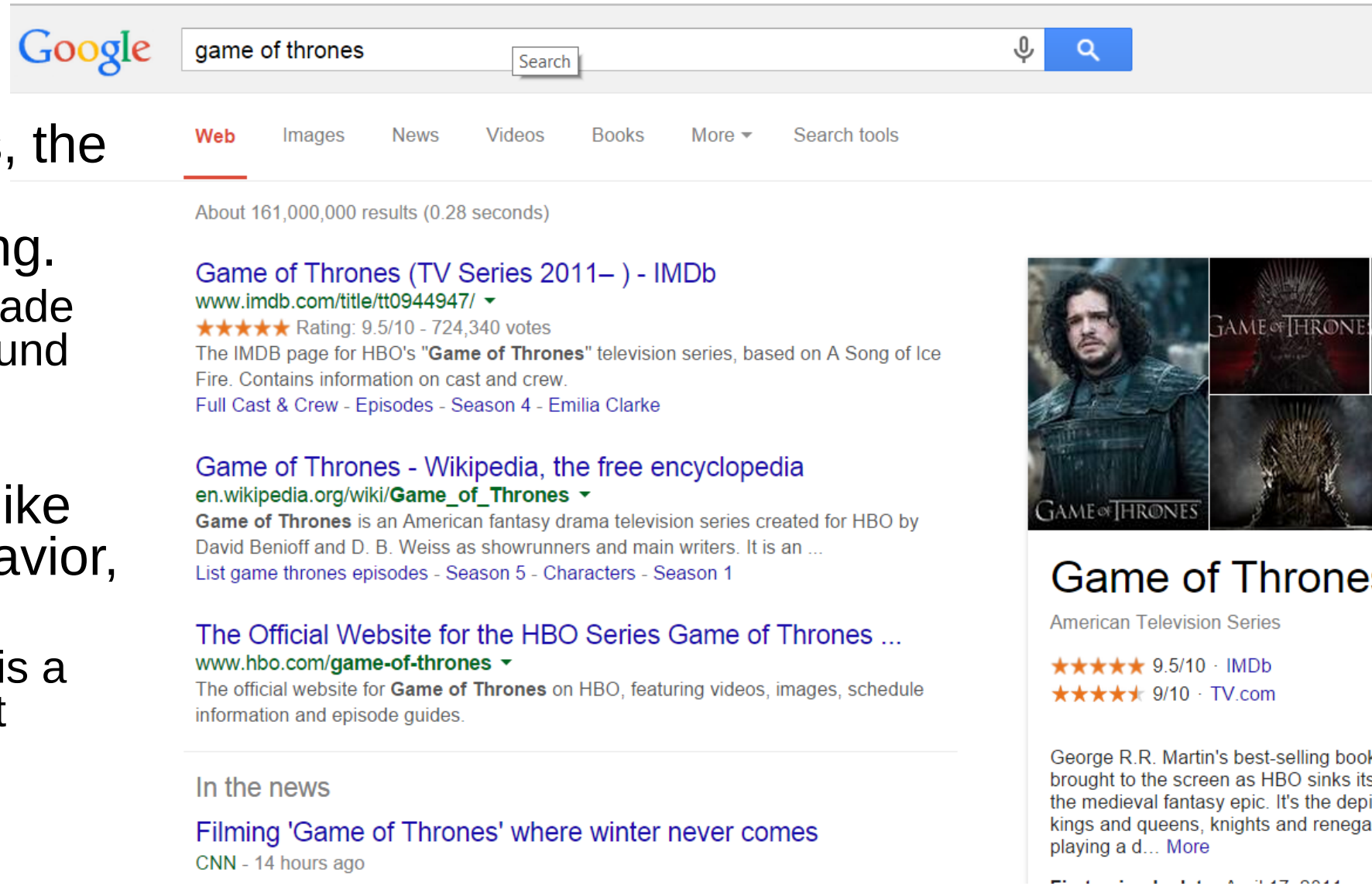
# Node importance

- What is the most important node in this graph?
- The PageRank algorithm: A node is important if it is pointed to by other important nodes



# The Web as a graph

- When ranking pages, the authoritativeness is factored in the ranking.
  - This is the idea that made **Google** a success around 2000
- Today a lot more information is used, like clicks, browsing behavior, etc
  - Ranking of the pages is a very complex task that requires sophisticated techniques



The screenshot shows a Google search interface with the query "game of thrones". The search results are displayed under the "Web" tab, showing approximately 161,000,000 results in 0.28 seconds. The top results include the IMDb page for the TV series, the Wikipedia page, and the official HBO website. A knowledge panel on the right side of the page provides additional information about the series, including its rating on IMDb and TV.com, and a brief description of its origin from George R.R. Martin's books.

Google

game of thrones

Search

Web Images News Videos Books More Search tools

About 161,000,000 results (0.28 seconds)

**Game of Thrones (TV Series 2011– ) - IMDb**  
[www.imdb.com/title/tt0944947/](http://www.imdb.com/title/tt0944947/)  
★★★★★ Rating: 9.5/10 - 724,340 votes  
The IMDB page for HBO's "**Game of Thrones**" television series, based on A Song of Ice Fire. Contains information on cast and crew.  
[Full Cast & Crew](#) - [Episodes](#) - [Season 4](#) - [Emilia Clarke](#)

**Game of Thrones - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Game\\_of\\_Thrones](http://en.wikipedia.org/wiki/Game_of_Thrones)  
**Game of Thrones** is an American fantasy drama television series created for HBO by David Benioff and D. B. Weiss as showrunners and main writers. It is an ...  
[List game thrones episodes](#) - [Season 5](#) - [Characters](#) - [Season 1](#)

**The Official Website for the HBO Series Game of Thrones ...**  
[www.hbo.com/game-of-thrones](http://www.hbo.com/game-of-thrones)  
The official website for **Game of Thrones** on HBO, featuring videos, images, schedule information and episode guides.

**In the news**  
**Filming 'Game of Thrones' where winter never comes**  
[CNN](#) - 14 hours ago

**Game of Thrones**  
American Television Series  
★★★★★ 9.5/10 · [IMDb](#)  
★★★★★ 9/10 · [TV.com](#)  
George R.R. Martin's best-selling book brought to the screen as HBO sinks its the medieval fantasy epic. It's the depiction of kings and queens, knights and renegades playing a deadly game of thrones. [More](#)



game of thrones

Search



Web

Images

News

Videos

Books

More ▾

Search tools

Page 10 of about 159,000,000 results (0.45 seconds)

## Game of Thrones Show Summary and Episode Schedule ...

[www.pogdesign.co.uk/cat/Game-of-Thrones-summary](http://www.pogdesign.co.uk/cat/Game-of-Thrones-summary) ▾

**Game of Thrones**. Seven noble families fight for control of the mythical land of Westeros. Political and sexual intrigue abound. The primary families are the Stark, ...

## Will Bibi's Doomsday Speech Matter? - The Daily Beast

[www.thedailybeast.com/.../bibi-israel-in-deadly-game-of-thrones-with-ir...](http://www.thedailybeast.com/.../bibi-israel-in-deadly-game-of-thrones-with-ir...) ▾

2 days ago - "In this deadly **game of thrones**, there's no place for America or for Israel, no peace for Christians, Jews or Muslims who don't share the Islamist ...

## Is 'Winds of Winter' finished? 'Game of Thrones' Nikolaj ...

[www.zap2it.com/.../is\\_winds\\_of\\_winter\\_finished\\_game\\_of\\_thrones\\_nik...](http://www.zap2it.com/.../is_winds_of_winter_finished_game_of_thrones_nik...) ▾

6 hours ago - Nikolaj Coster-Waldau of **Game of Thrones** Is "**Game of Thrones**" fans' impatient wait for George R.R. Martin's next book, "The Winds of Winter," ...

## Sand Snakes or Snow Snakes? Not Everyone Is Happy With ...

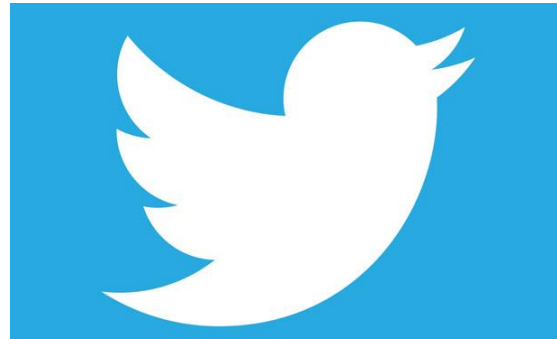
[www.styleite.com/.../sand-snakes-or-snow-snakes-new-game-of-thrones-...](http://www.styleite.com/.../sand-snakes-or-snow-snakes-new-game-of-thrones-...) ▾

2 days ago - **Game of Thrones** is getting a trio of badass new female characters next season. Obara (Keisha Castle-Hughes), Tyene (Rosabell Laurenti ...

## OMG The 'Game Of Thrones' Sand Snakes Look Amazing

# Friendship suggestions

- LinkedIn, Twitter, Facebook **friendship suggestions**
  - Useful for the users to discover their friends, but also useful for the network in order to **grow**, and increase **engagement**
    - LinkedIn success story



- **Triadic closure principle**: Links are created in a way that usually closes a triangle
  - If both Bob and Charlie know Alice, then they are likely to meet at some point.

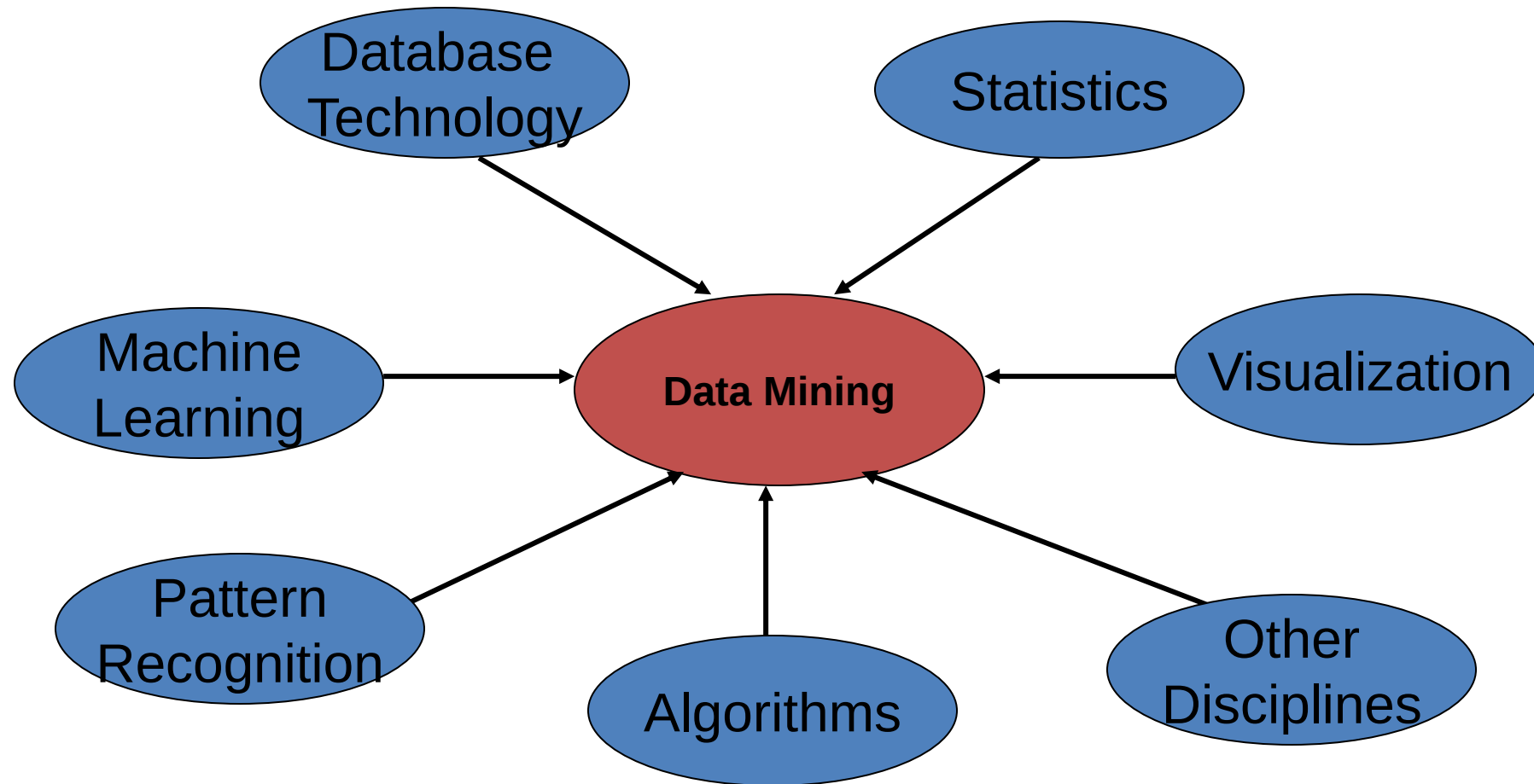
# What is Data Mining again?

- “Data mining is the analysis of (often large) observational data sets to find **unsuspected relationships** and to **summarize** the data in novel ways that are both **understandable and useful** to the data analyst” (Hand, Mannila, Smyth)
- “Data mining is the discovery of **models** for data” (Rajaraman, Ullman)
  - We can have the following types of models
    - Models that **explain** the data (e.g., a single function)
    - Models that **predict** the future data instances.
    - Models that **summarize** the data
    - Models the **extract** the most prominent **features** of the data.
- “Data Mining is the study of **collecting, processing, analyzing, and gaining useful insights** from data” – Charu Aggarwal

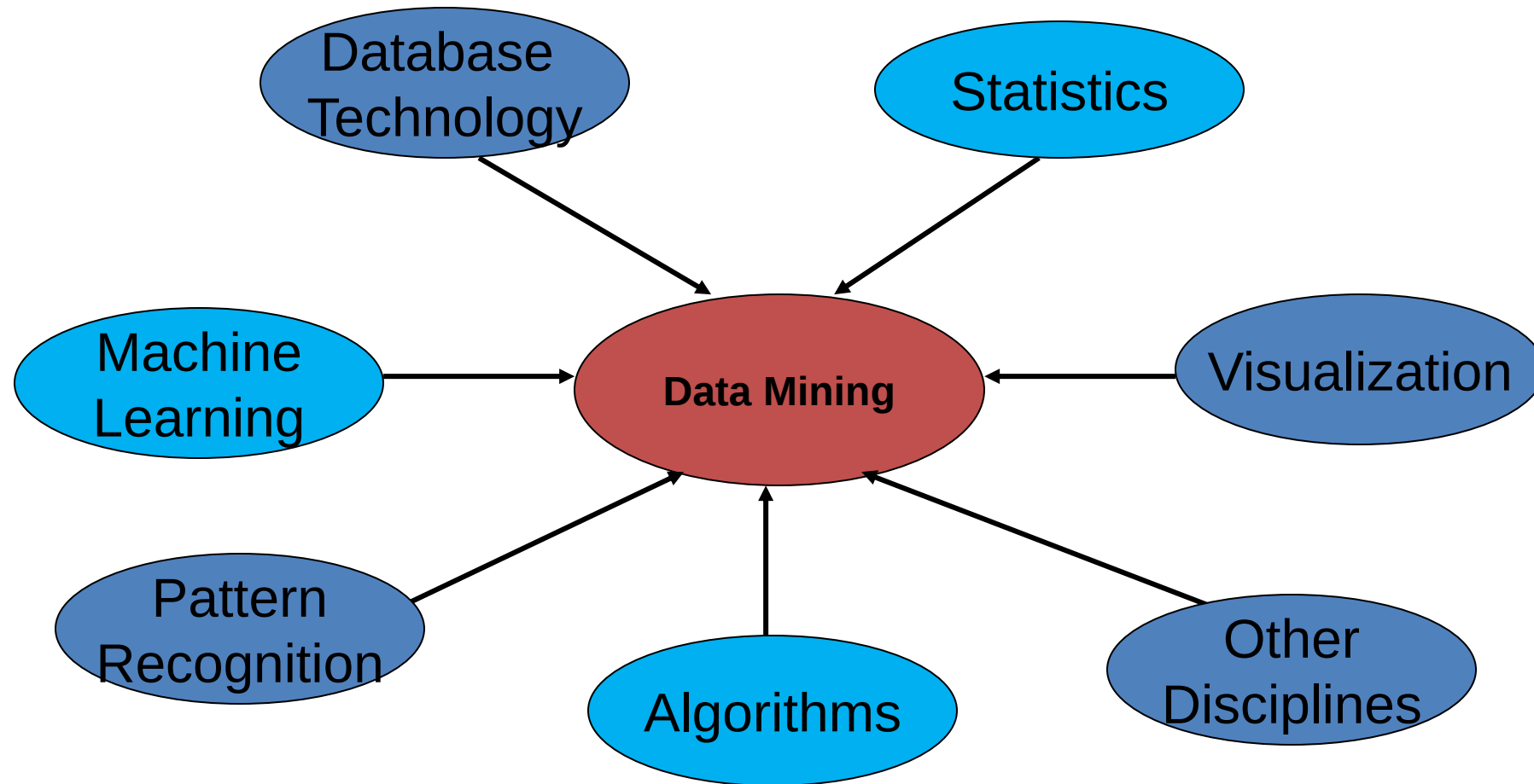
# Why data mining?

- **Scientific** point of view
  - Scientists are at an unprecedented position where they can collect TB of information
    - Examples: Sensor data, astronomy data, social network data, gene data
  - We need the tools to analyze such data to get a better understanding of the world and advance science and help people
- **Commercial** point of view
  - Data has become the key competitive advantage of companies
    - Examples: Facebook, Google, Amazon
  - Being able to extract useful information out of the data is key for exploiting them commercially.
- **Scale** (in data **size** and feature **dimension**)
  - Why not use traditional analytic methods?
  - Enormity of data, **curse of dimensionality**
  - The amount and the complexity of data does not allow for manual processing of the data. We need automated techniques.

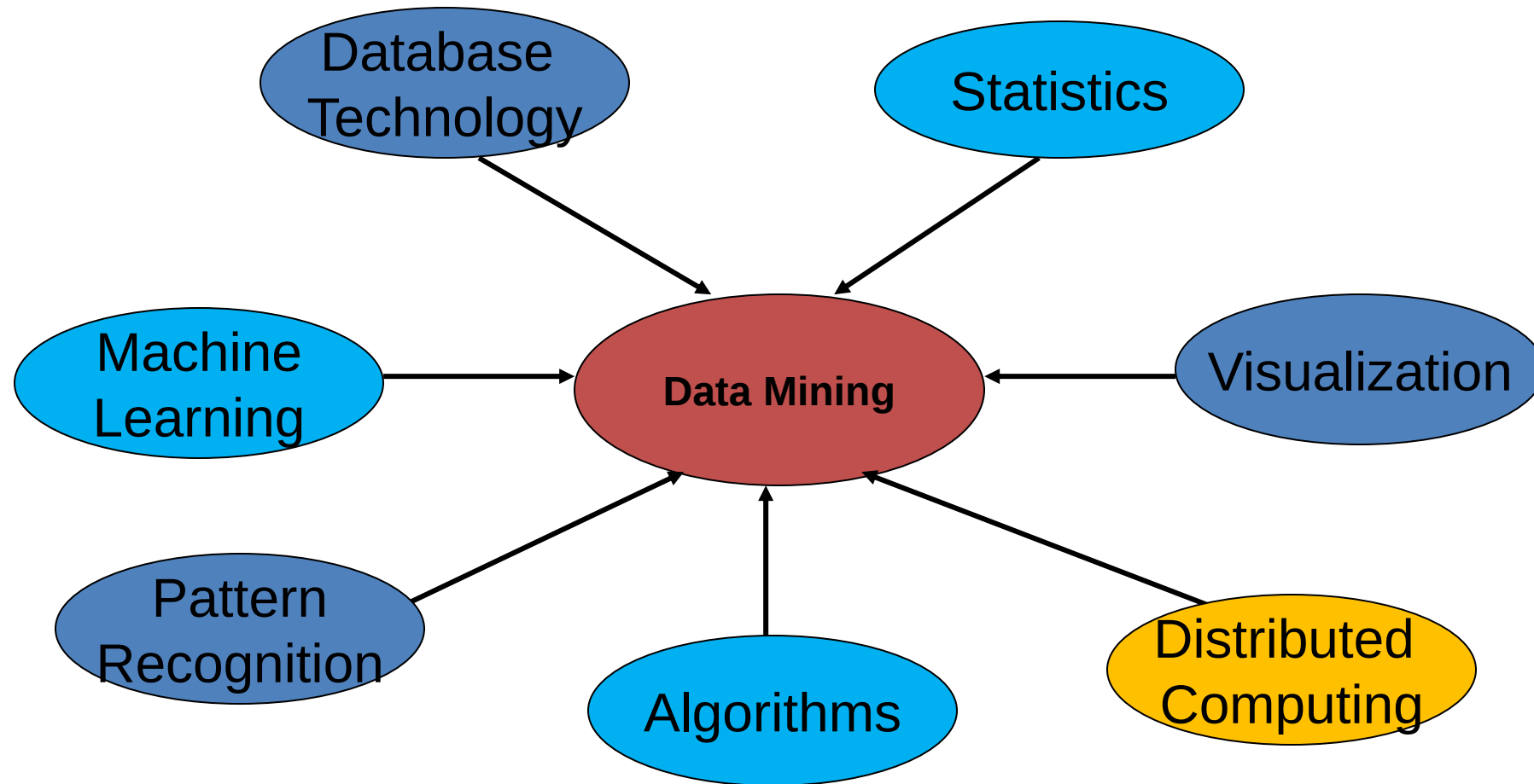
# Data Mining: Confluence of Multiple Disciplines



# Data Mining: Confluence of Multiple Disciplines



# Data Mining: Confluence of Multiple Disciplines



# The buzz around data

- **Data Science**: Data is useful to understand a process and improve it. All organizations should have a data science team that analyses their data and proposes improvements
  - Focuses on more immediate applications and insights
- **Big Data**: Data appear everywhere. We should process it collectively and interconnect them. We need infrastructure (cloud computing, cloud storage) to do this
  - More systems oriented
- **AI/Machine Learning/Deep Learning**: These have been around for a while but now we have the data to learn more complex models that are significantly more powerful
  - More emphasis on scientific breakthroughs

# New era of data mining

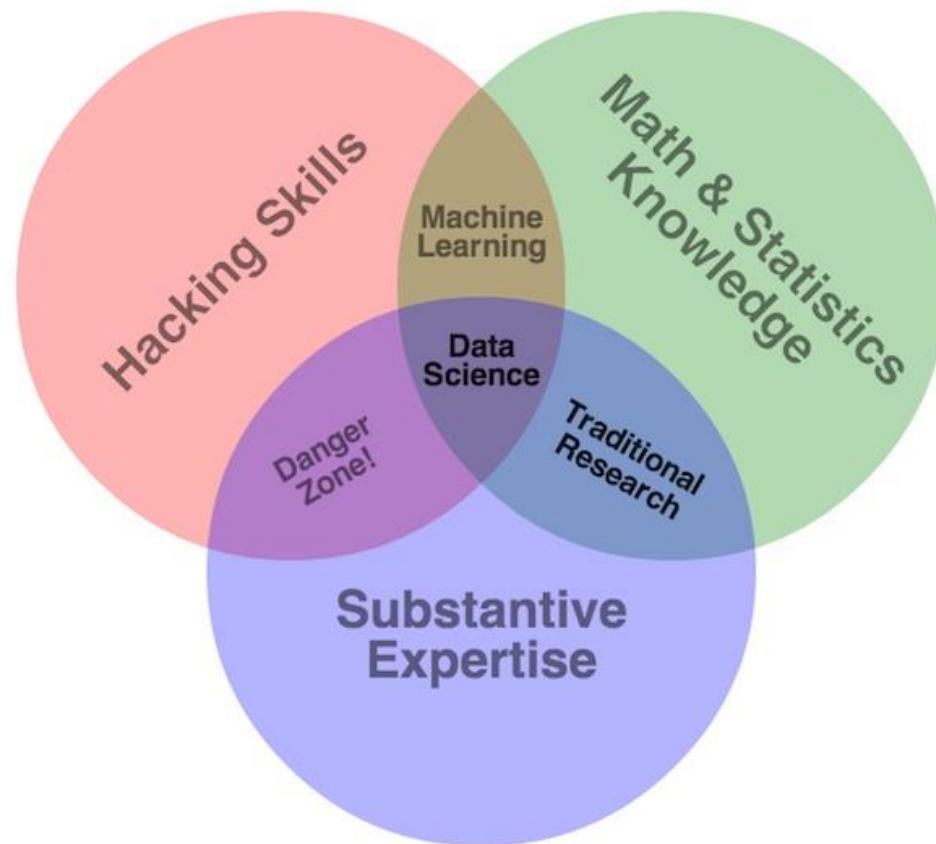
- Boundaries are becoming less clear
  - Today data mining, machine learning, and AI are synonymous. It is assumed that the algorithms should scale. It is clear that statistical inference is used for building the models.
  - Data is the engine for AI
  - Data Mining touches everything related to data.

# Which also has a dark side

- Are the algorithms making fair and correct decisions?
- Do algorithms create filter bubbles, echo chambers, and promote misinformation? Are they a threat to democracy?
- Surveillance capitalism
- Is AI a threat?



# The Skills of a Data Miner – Data Scientist



It is a hard job

# But also a rewarding one

*"The success of companies like Google, Facebook, Amazon, and Netflix, not to mention Wall Street firms and industries from manufacturing and retail to healthcare, is increasingly driven by better tools for extracting meaning from very large quantities of data. 'Data Scientist' is now the hottest job title in Silicon Valley."* — Tim O'Reilly

## Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (87)



Artwork: Tamar Cohen, Andrew J Buboltz, 2011, silk screen on a page from a high school yearbook. 9.5" x 12"

### RELATED

**Executive Summary**

### ALSO AVAILABLE

- [Buy PDF](#)