What is Data Analytics?
Data contains value and knowledge
Data Analytics

- But to extract the knowledge, data needs to be
  - Stored
  - Managed
  - Analyzed \(\leftarrow\) emphasis on this class
  - Visualized \(\leftarrow\) emphasis on this class

Data Analytics \(\approx\) Data Mining \(\approx\) Big Data \(\approx\) Predictive Analytics \(\approx\) Data Science
what is data analytics?
Objective of Data Analysis

- **Input:** lots of data
- **Output:** patterns and models that are:
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern
Types of Data Analysis

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering (e.g., find communities of interest)

- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommendations (e.g., suggest new friends in a social network)
Data Analytics: Cultures

- **Data analysis overlaps with:**
  - **Databases:** Large data, simple queries
  - **Machine learning:** Large data, complex models
  - **CS Theory:** (Randomized) Algorithms

- **Different cultures:**
  - To a DB person, data analysis is an extreme form of **analytic processing** – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data analysis is the **inference of models**
    - Result is the parameters of the model
Growing market revenue of Big Data Analytics in billion U.S. dollars from the year 2011 to 2027

https://www.edureka.co/blog/what-is-big-data/
What Type of Data?

Text Data

Network Data

Multivariate Data
The “Age of Networks”

Technological  Social  Biological
why should we care about networks?
Why Networks? Why Now?

- **Universal language for describing complex data**
  - Networks from science, nature, and technology are more similar than one would expect
- **Shared vocabulary between fields**
  - Computer Science, Social science, Physics, Economics, Statistics, Biology
- **Data availability (/computational challenges)**
  - Web/mobile, bio, health, and medical
- **Impact!**
  - Social networking, Social media, Brain, Drug design
  - *We will never understand these systems unless we understand the networks behind them!*
how do we reason about networks?
How do we reason about networks?

- **Empirical**: Study network data to find organizational principles
- **Mathematical models**: Probabilistic, graph theory
- **Algorithms**: Methods for analyzing graphs
What do we study in networks?

- **Structure and evolution**
  - What is the structure of a network?
  - Why and how did it become to have such structure?

- **Processes and dynamics**
  - Networks provide “skeleton” for spreading of information, behavior, diseases
What We Have Covered?

- **basic graph theory**
  - graphs, networks
  - bow-tie structure

- **network measurements**
  - degree distributions, power-laws
  - shortest paths, clustering coefficient

- **network models**
  - Erdos-Renyi model
  - small-world model
  - configuration model
  - scale-free networks

- **models of evolving graphs**
  - preferential attachment model
  - microscopic/macroscopic evolution of networks
  - forest-fire model

- **community structure in networks**
  - Strength of weak ties, structural holes
  - community detection, Girvan-Newman algorithm
  - graph partitioning, graph cuts, conductance
  - spectral graph theory, spectral graph clustering
  - overlapping communities in networks

- **link analysis**
  - web search
  - hubs and authorities (HITS)
  - PageRank, topic-sensitive PageRank

- **link prediction**
  - neighborhood-based methods
  - node proximity based methods
  - supervised learning models, FB's "PYMK", Twitter's "WtF"

- **cascading behavior in networks**
  - Granovetter’s model, threshold model
  - game theoretic model
  - epidemic model on trees
  - disease spreading models (SIR, SIS, SIRS)
  - independent cascade model
  - influence maximization
  - outbreak detection

- **data visualization**
  - visual variables (Jacques Bertin’s)
  - perception & cognition
  - pre-attentive vs attentive processing
  - gestalt principles
  - principles of graphical excellence (Tufte’s)
  - a taxonomy of representation
  - visual elements intro (charts, graphs, maps)
How It All Fits Together

**Properties**
- Small diameter, Edge clustering
- Scale-free
- Strength of weak ties, Core-periphery
- Densification power law, Shrinking diameters
- Information virality, reproductive number

**Models**
- Small-world model, Erdös-Renyi model
- Preferential attachment, Copying model
- Community-affiliation Graph Model
- Microscopic model of evolving networks
- Independent cascade model, Game theoretic model, SIR

**Algorithms**
- Decentralized search
- PageRank, Hubs and authorities
- Community detection: Girvan-Newman, Modularity
- Link prediction, Supervised random walks
- Influence maximization, Outbreak detection, LIM
Small-World Phenomena

**Properties:**
- Six degrees of separation
  - Networks have small diameters
- Edges in the networks cluster
  - Large clustering coefficient

**Models:**
- Erdös-Renyi model
  - Baseline model for networks
- The Small-World model
  - Small diameter and clustered edges

**Algorithms:**
- Link analysis in networks
  - PageRank algorithm; link prediction
Scale-Free Networks

- **Properties:**
  - Power-law degrees
    - Degrees are heavily skewed
  - Network resilience
    - Networks are resilient to random attacks

- **Models:**
  - Preferential attachment
    - Rich get richer

- **Algorithms:**
  - Hubs and Authorities
    - Recursive: \( a_i = \sum_{j \rightarrow i} h_j, \ h_i = \sum_{i \rightarrow j} a_j \)
  - PageRank
    - Recursive formulation, Random jumps
Community Detection

- **Properties:**
  - Strength of weak ties
  - Core-periphery structure

- **Models:**
  - Community-affinity model

- **Algorithms:**
  - Spectral Clustering
  - Girvan-Newman (Betweenness centrality)
  - **Modularity:** \( \text{#edges within group} - E[\text{#edges within group}] \)
  - Clique Percolation Method
    - Overlapping communities
Network Diffusion

- **Properties:**
  - Node-to-node influence
  - Node threshold
  - Cascade spread

- **Models:**
  - Game theoretic model:
    - Payoffs, Competing products
  - Independent Cascade Model
    - Each node infects a neighbor with some probability
Map of Superpowers

Properties
- Small diameter, Edge clustering
- Scale-free
- Strength of weak ties, Core-periphery
- Densification power law, Shrinking diameters
- Information virality, reproductive number

Models
- Small-world model, Erdös-Renyi model
- Preferential attachment, Copying model
- Community-affiliation Graph Model
- Microscopic model of evolving networks
- Independent cascade model, Game theoretic model, SIR

Algorithms
- Decentralized search
- PageRank, Hubs and authorities
- Community detection: Girvan-Newman, Modularity
- Link prediction, Supervised random walks
- Influence maximization, Outbreak detection, LIM
Applying Your Superpowers
Applying Your Superpowers

- Social media analytics
- Viral marketing
Applying Your Superpowers

- Predicting epidemics: Ebola
Applying Your Superpowers

- Interactions of human diseases
- Drug design
Data Visualization – Review
Why Visualize Data?

Summary statistics for all four datasets
- \( \text{avg}(x) = 9 \)
- \( \text{avg}(y) = 7.50 \)
- \( \text{Var}(x) = 11 \)
- \( \text{Var}(y) = 4.12 \)
- \( \text{Correlation}(x,y) = 0.816 \)
- A linear regression line: \( y = 0.5x + 3 \)

Always plot your data!

Anscombe’s Quartet
Jacques Bertin’s Visual Variables

Jacques Bertin proposed an original set of “retinal variables” in Semiology of Graphics (1967)

- **Position**: changes in x, y, (z) location
- **Size**: change in length/area
- **Shape**: changes in shape
- **Value**: changes in light value
- **Hue**: changes in hue value
- **Orientation**: changes in alignment
- **Texture**: changes in pattern
perception is fragmented
eyes are constantly scanning and constructing reality

The “Door Study”*
https://www.youtube.com/embed/FWSxSOsspiQ

Pre-attentive vs Attentive Processing

**PRE-ATTENTIVE PROCESSING**
- bottom-up
- fast, automatic
- instinctive
- efficient
- multitasks

**ATTENTIVE PROCESSING**
- top-down
- slow, deliberate
- focused
- singe-task

**goal of information design**
- help humans process information as efficiently as possible
- make as much use of pre-attentive processing as possible
Gestalt Principles

- Figure/Ground
- Proximity
- Similarity
- Symmetry
- Continuity
- Closure

**Gestalt Principles**

**Good Figure**
Objects grouped together tend to be perceived as a single figure. Tendency to simplify.

**Proximity**
Objects tend to be grouped together if they are close to each other.

**Similarity**
Objects tend to be grouped together if they are similar.

**Continuation**
When there is an intersection between two or more objects, people tend to perceive each object as a single uninterrupted object.

**Closure**
Visual connection or continuity between sets of elements which do not actually touch each other in a composition.

**Symmetry**
The object tend to be perceived as symmetrical shapes that form around their center.
What Makes a Good Visualization?

https://informationisbeautiful.net/visualizations/what-makes-a-good-data-visualization/
Data Types

Data

Qualitative (Descriptive)
- Nominal
  - Data has no natural order
  - examples: gender, race, religion, sport
- Ordinal
  - Data can be arranged in order or rank
  - examples: sizes (s/m/l), attitudes (disagree, neutral, agree), house number.

Quantitative (Numerical)
- Continuous
  - Data is measured on a continuous scale
  - examples: temperature, length, height
- Discrete
  - Data is countable, and exists only in whole numbers
  - examples: number of people taking this class
Use of **visual elements** like **charts**, **graphs**, and **maps** to see and understand **trends**, **outliers**, and **patterns** in data.
What’s Next?
What’s Next?

- **Project presentation**
  - Mon, Mar 23rd and Mar 30th online (Zoom)
    - 12 minutes + 3 min QA
    - See course website for more info

- **Project final report**
  - Fri, Apr 17th Midnight (11:59PM) Pacific Time
    - Email and submit electronically your PDF report
    - see course website for more info
What Next?

- **Related conferences / Journals:**
  - **Conferences**
    - **DSAA:** IEEE Data Science and Advanced Analytics
    - **KDD:** ACM Conf. on Knowledge Discovery & Data Mining
    - **WWW:** ACM World Wide Web Conference
    - **WSDM:** ACM Web search and Data Mining
    - **ICDM:** IEEE International Conference on Data Mining
    - **ICWSM:** AAAI Int. Conf. on Web-blogs & Social Media
    - **Complex Networks:** Int. Conf. on Complex Networks
  - **Journals**
    - **Complex Networks:** Journal of Complex Networks
    - **TKDD:** ACM Transactions on Knowledge Discovery from Data
    - **TKDE:** IEEE Transactions on Knowledge and Data Engineering
In Closing...

You have worked a lot...

...and (hopefully) learned a lot!
thank you & happy holidays