EECS6414: Data Analytics & Visualization

Thanks to Jure Leskovec, Anand Rajaraman, Jeff Ullman Stanford University - **http://www.mmds.org**

What is Data Analytics?

Data contains value and knowledge

Data Analytics

- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - Analyzed ← emphasis on this class
  - Visualized ← emphasis on this class

Data Analytics ≈ Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

what is data analytics?

Objective of Data Analysis

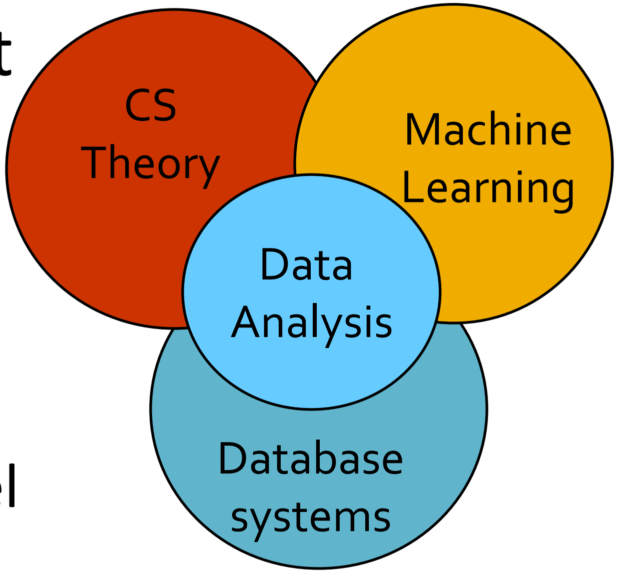
- **Input: lots of data**
- **Output: patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

Types of Data Analysis

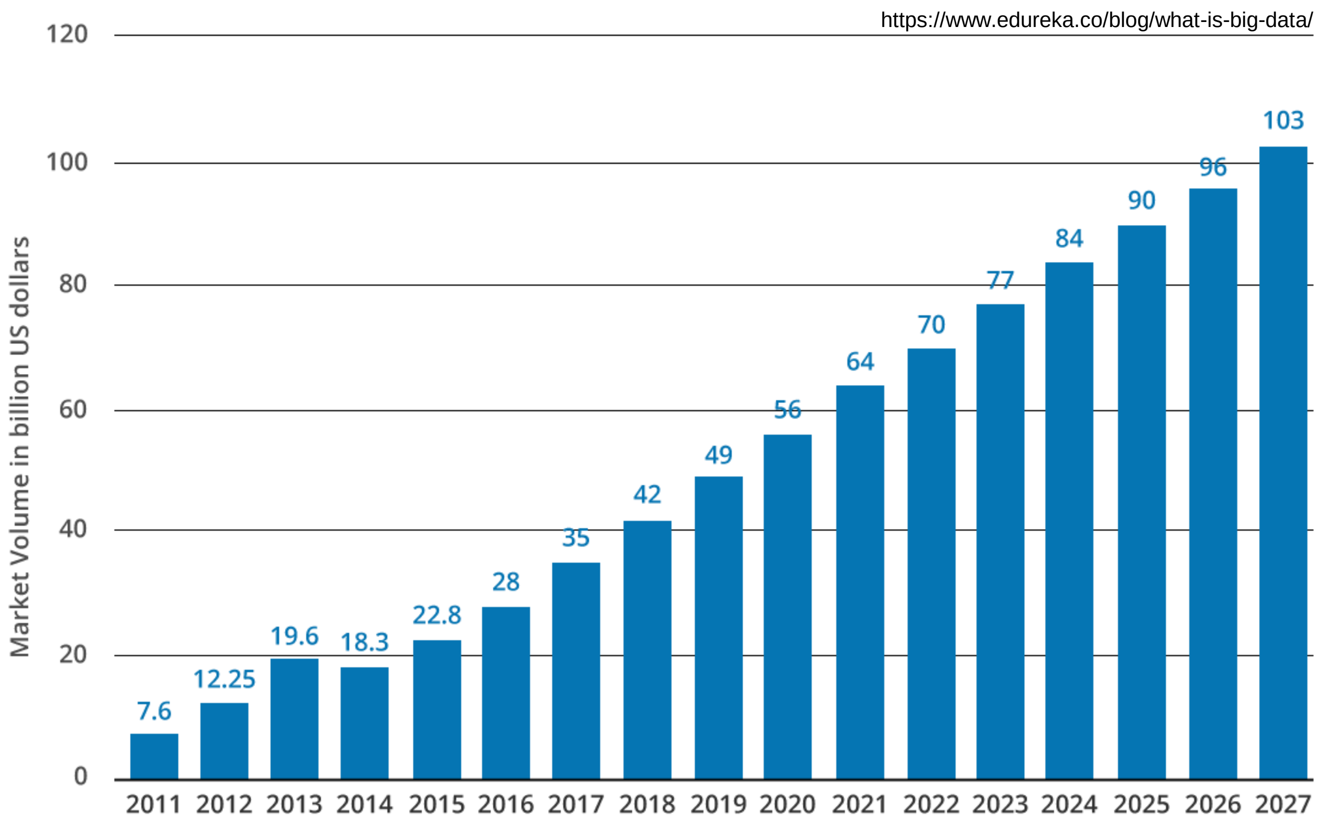
- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering (e.g., find communities of interest)
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommendations (e.g., suggest 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



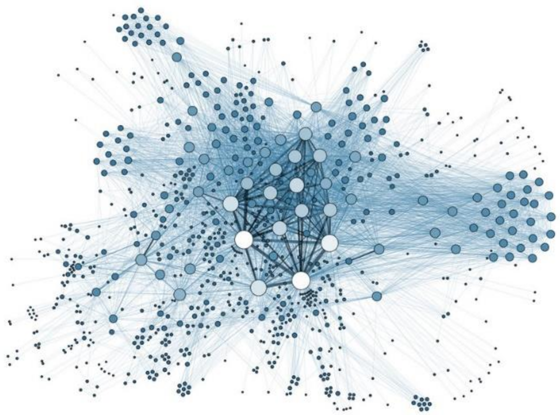
Demand for Data Analytics



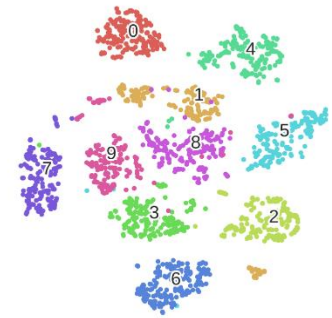
Growing market revenue of Big Data Analytics in billion U.S. dollars from the year 2011 to 2027

What Type of Data?

Text Data



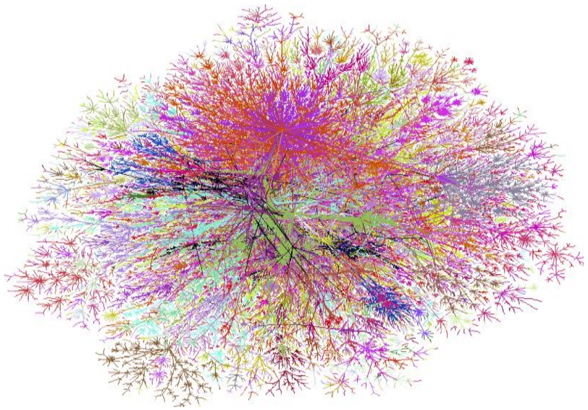
Network Data



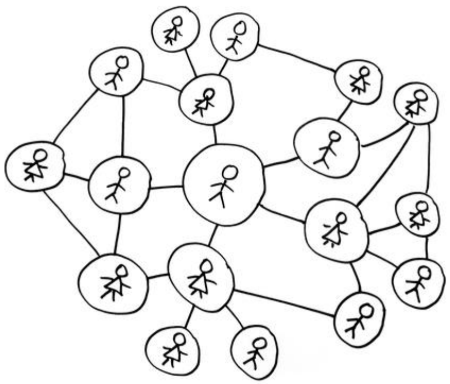
Multivariate Data

Networks Analytics – Review

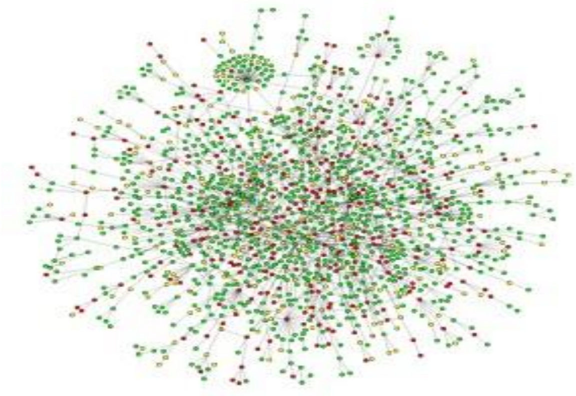
The "Age of Networks"



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?

Reasoning About Networks

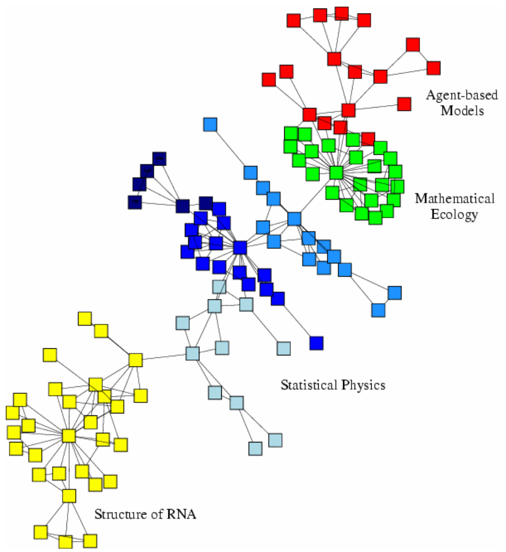
How do we reason about networks?

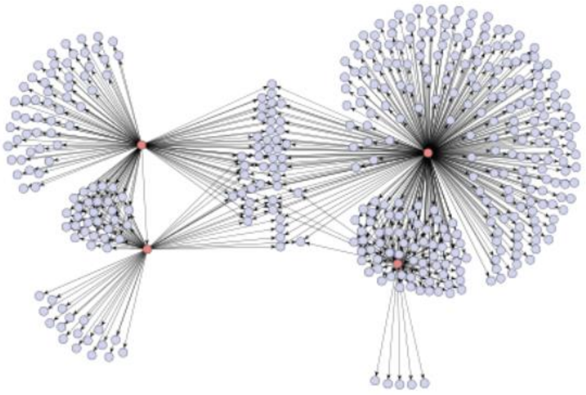
- **Empirical**: Study network data to find organizational principles
- **Mathematical models**: Probabilistic, graph theory
- **Algorithms**: Methods for analyzing graphs

Networks: Structure & Process

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?

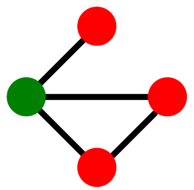
- basic graph theory
  - graphs, networks
  - bow-tie structure
- network measurements
  - degree distributions, power-laws
  - shortest paths, clustering coefficient
- network models
  - Erdos-Renyi model
  - small-world model
  - configuration model
  - scale-free networks
- models of evolving graphs
  - 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

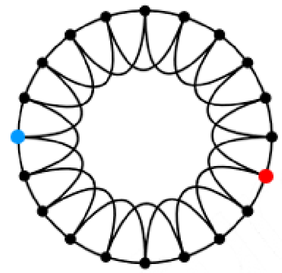
Properties	Models	Algorithms
Small diameter, Edge clustering	Small-world model, Erdös-Renyi model	Decentralized search
Scale-free	Preferential attachment, Copying model	PageRank, Hubs and authorities
Strength of weak ties, Core-periphery	Community-affiliation Graph Model	Community detection: Girvan-Newman, Modularity
Densification power law, Shrinking diameters	Microscopic model of evolving networks	Link prediction, Supervised random walks
Information virality, reproductive number	Independent cascade model, Game theoretic model, SIR	Influence maximization, Outbreak detection, LIM

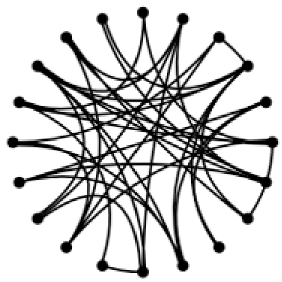
Small-World Phenomena

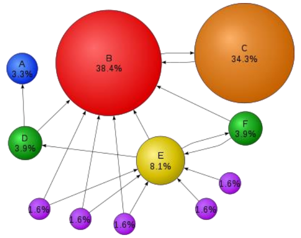
- **Properties:**
  - **Six degrees of separation**
    - Networks have small diameters
  - **Edges in the networks cluster**
    - Large clustering coefficient
- **Models:**
  - **Erdös-Renyi model**
    - Baseline model for networks
  - **The Small-World model**
    - Small diameter and clustered edges
- **Algorithms:**
  - **Link analysis in networks**
    - PageRank algorithm; link prediction

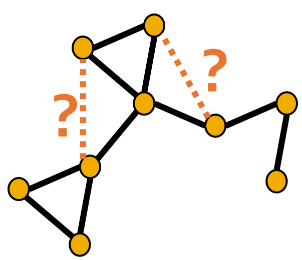


$C_i=1/3$

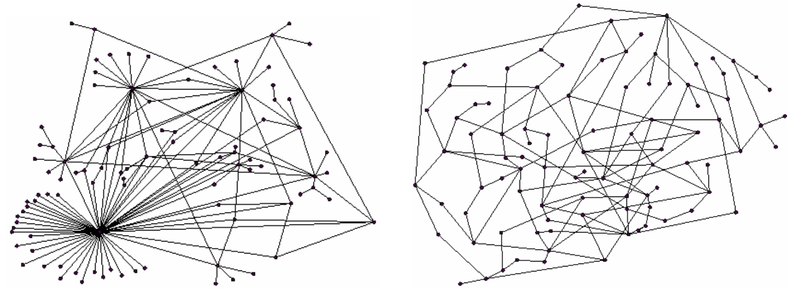




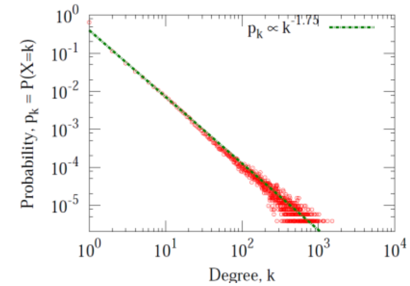




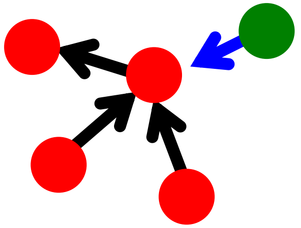
Scale-Free Networks



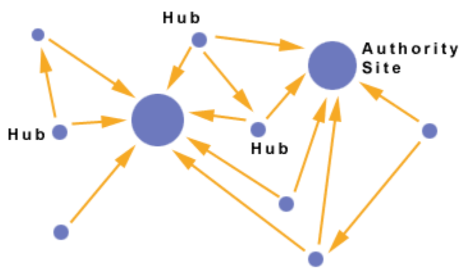
- **Properties:**
  - **Power-law degrees**
    - Degrees are heavily skewed



  - **Network resilience**
    - Networks are resilient to random attacks
- **Models:**
  - **Preferential attachment**
    - Rich get richer

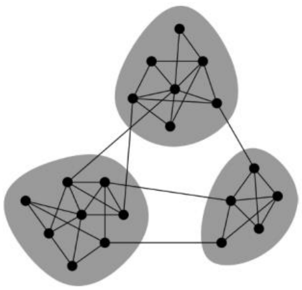


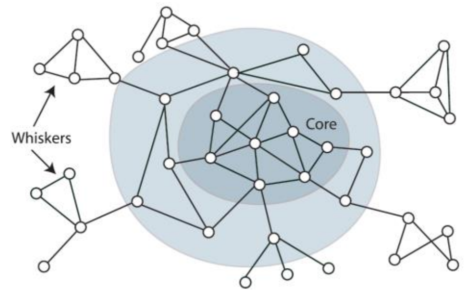
- **Algorithms:**
  - **Hubs and Authorities**
    - Recursive: $a_i = \sum_{j \to i} h_j,\ h_i = \sum_{i \to j} a_j$
  - **PageRank**
    - Recursive formulation, Random jumps



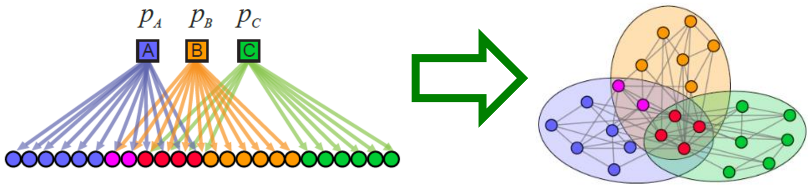
Community Detection

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

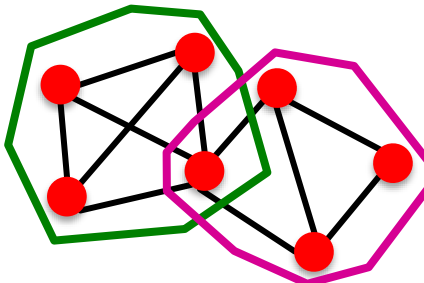




- **Models:**
  - **Community-affinity model**

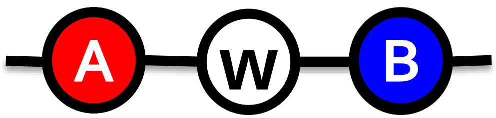


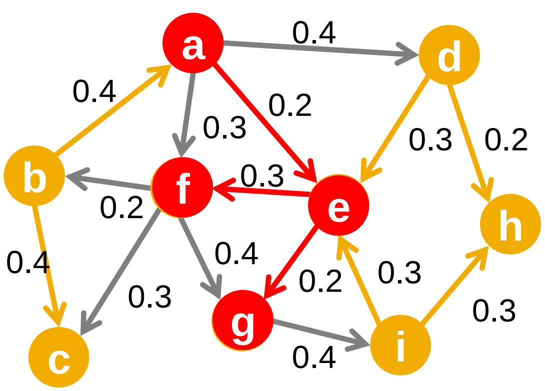
- **Algorithms:**
  - **Spectral Clustering**
  - **Girvan-Newman** (Betweeness centrality)
  - **Modularity:** #edges within group – E[#edges within group]
  - **Clique Percolation Method**
    - Overlapping communities



Network Diffusion

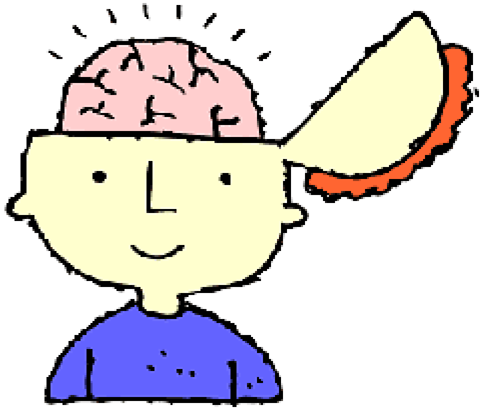
- **Properties:**
  - **Node-to-node influence**
  - **Node threshold**
  - **Cascade spread**
- **Models:**
  - **Game theoretic model:**
    - Payoffs, Competing products
  - **Independent Cascade Model**
    - Each node infects a neighbor with some probability





Map of Superpowers

Properties	Models	Algorithms
Small diameter, Edge clustering	Small-world model, Erdös-Renyi model	Decentralized search
Scale-free	Preferential attachment, Copying model	PageRank, Hubs and authorities
Strength of weak ties, Core-periphery	Community-affiliation Graph Model	Community detection: Girvan-Newman, Modularity
Densification power law, Shrinking diameters	Microscopic model of evolving networks	Link prediction, Supervised random walks
Information virality, reproductive number	Independent cascade model, Game theoretic model, SIR	Influence maximization, Outbreak detection, LIM

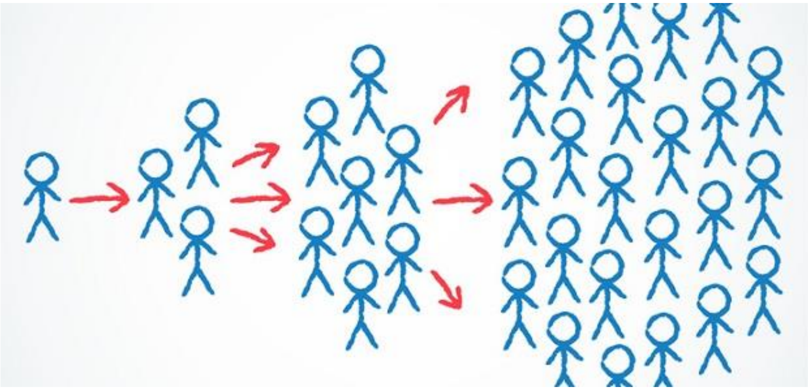


Applying Your Superpowers

Applying Your Superpowers

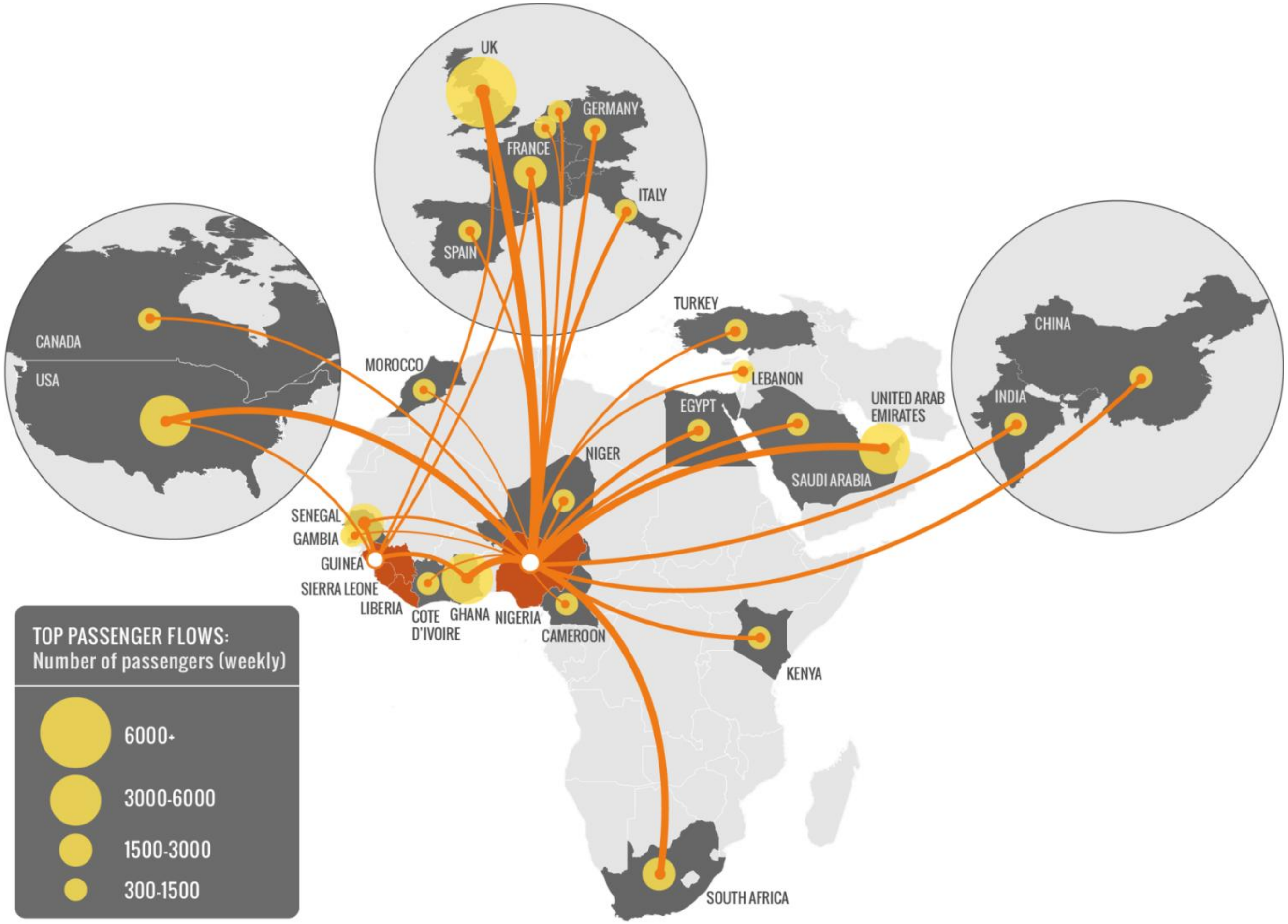
- **Social media analytics**

- **Viral marketing**



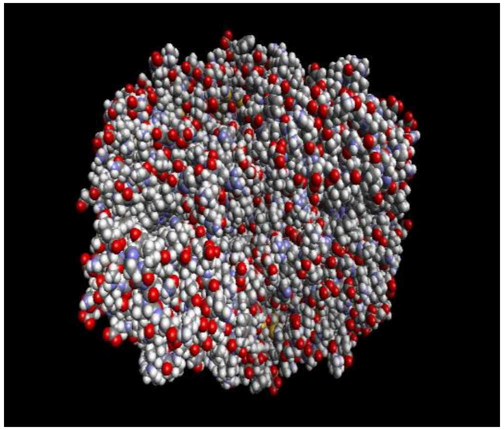
Applying Your Superpowers

- **Predicting epidemics: Ebola**

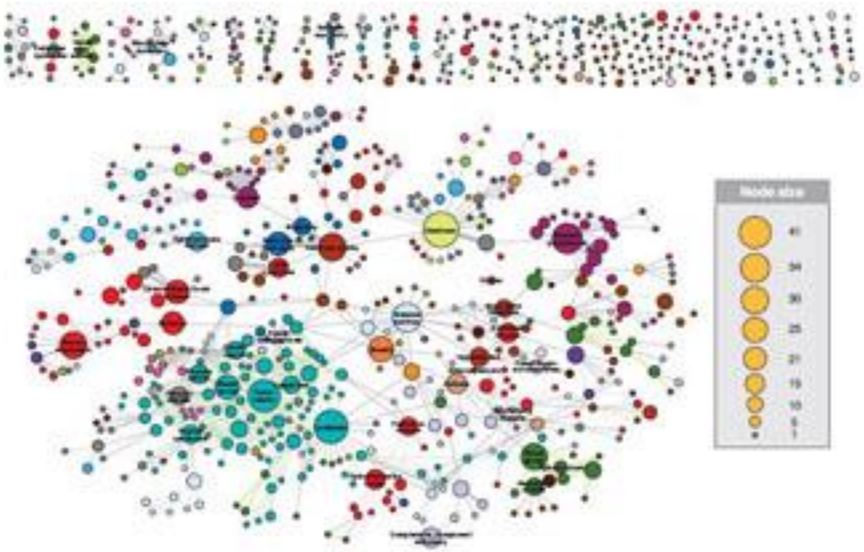


Applying Your Superpowers

- **Interactions of human diseases**

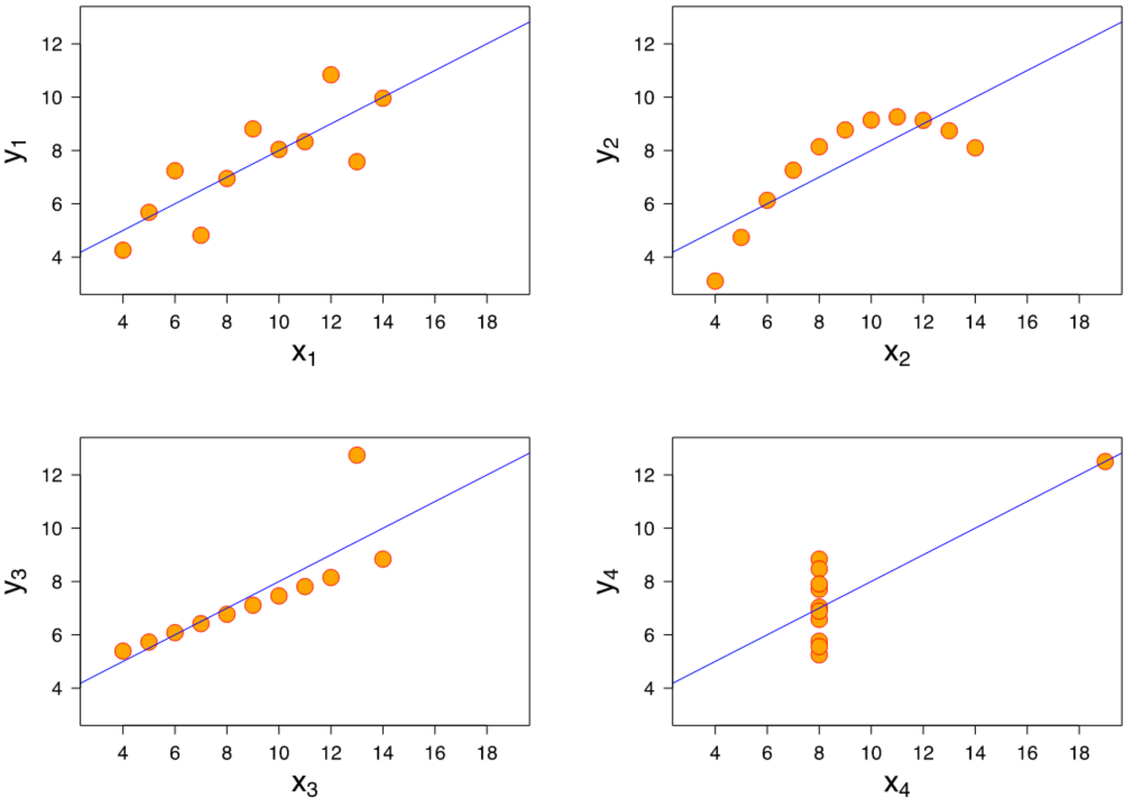


- **Drug design**



Data Visualization – Review

Why Visualize Data?



Summary statistics for all four datasets

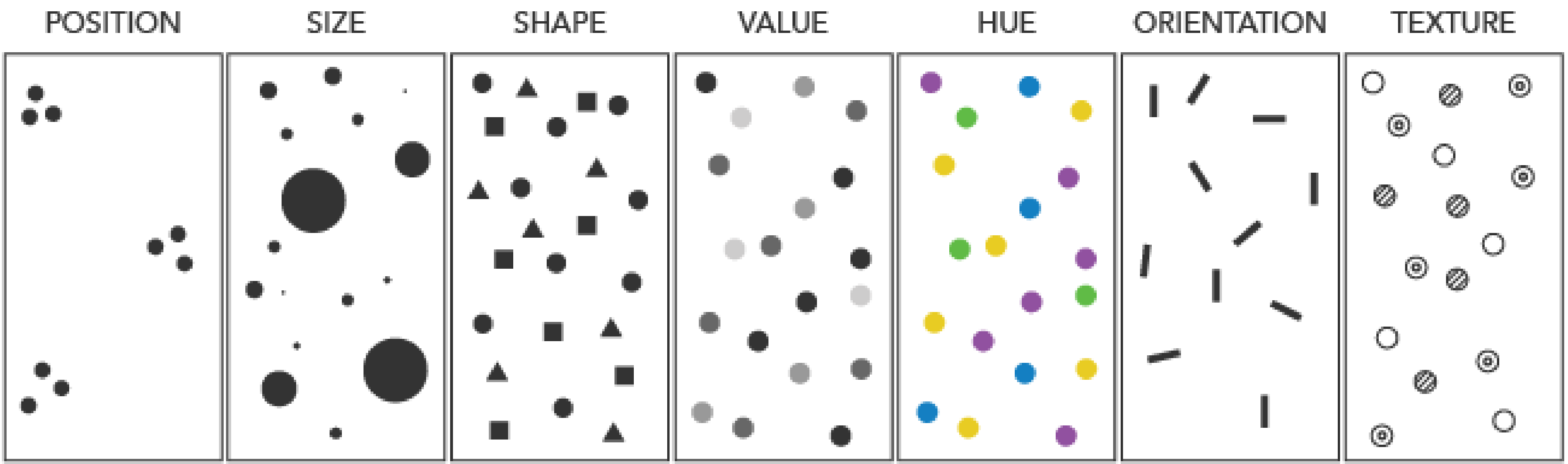
- avg(x) = 9
- avg(y) = 7.50
- Var(x) = 11
- Var(y) = 4.12
- Correlation(x,y) = 0.816
- A linear regression line: y = 0.5x + 3

Always plot your data!

Anscombe's Quartet

Anscombe, F. (1973). Graphs in statistical analysis. American Statistician, 27:17--21.

Jacques Bertin's Visual Variables



changes in x, y, (z) location

change in length/area

changes in shape

changes in light value

changes in hue value

changes in alignment

changes in pattern

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

Perception & Cognition

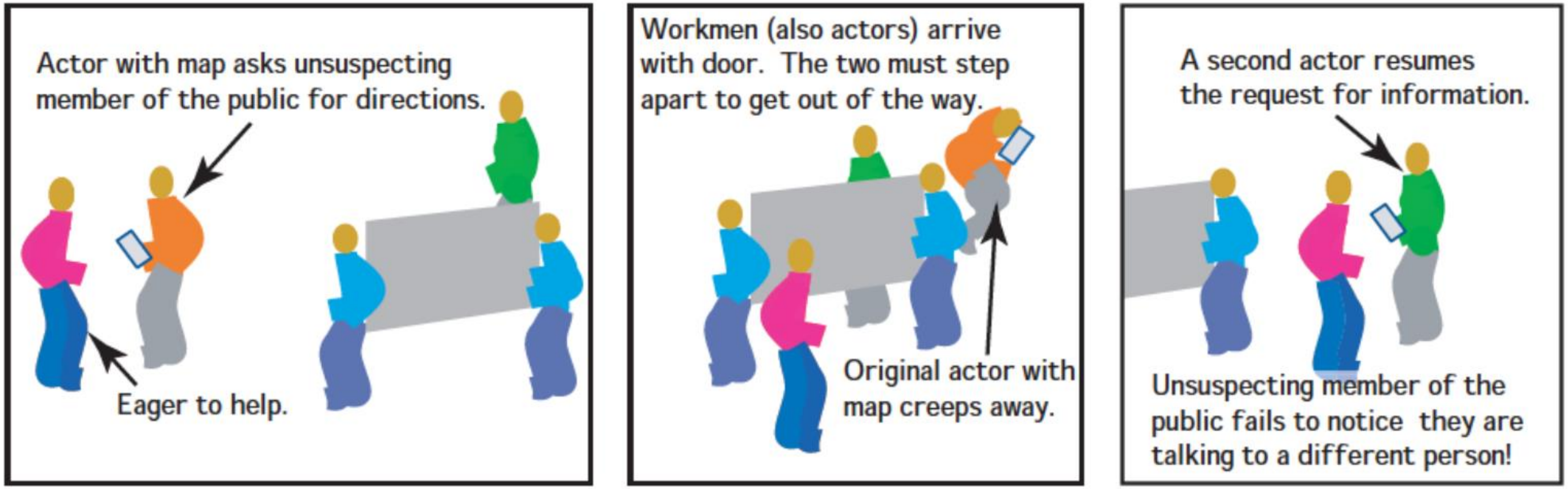


Image: Ware, Colin. Visual thinking: For design. Morgan Kaufmann, 2010

- perception is fragmented
- eyes are constantly scanning and constructing reality

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

* Daniel J. Simons and Daniel T. Levin. 1998. “Failure to detect changes to people during a real world interaction.” Psychonomic Bulletin and Review. 5: 644–669.

Pre-attentive vs Attentive Processing

PRE-ATTENTIVE PROCESSING

- bottom-up
- fast, automatic
- instinctive
- efficient
- multitasks

ATTENTIVE PROCESSING

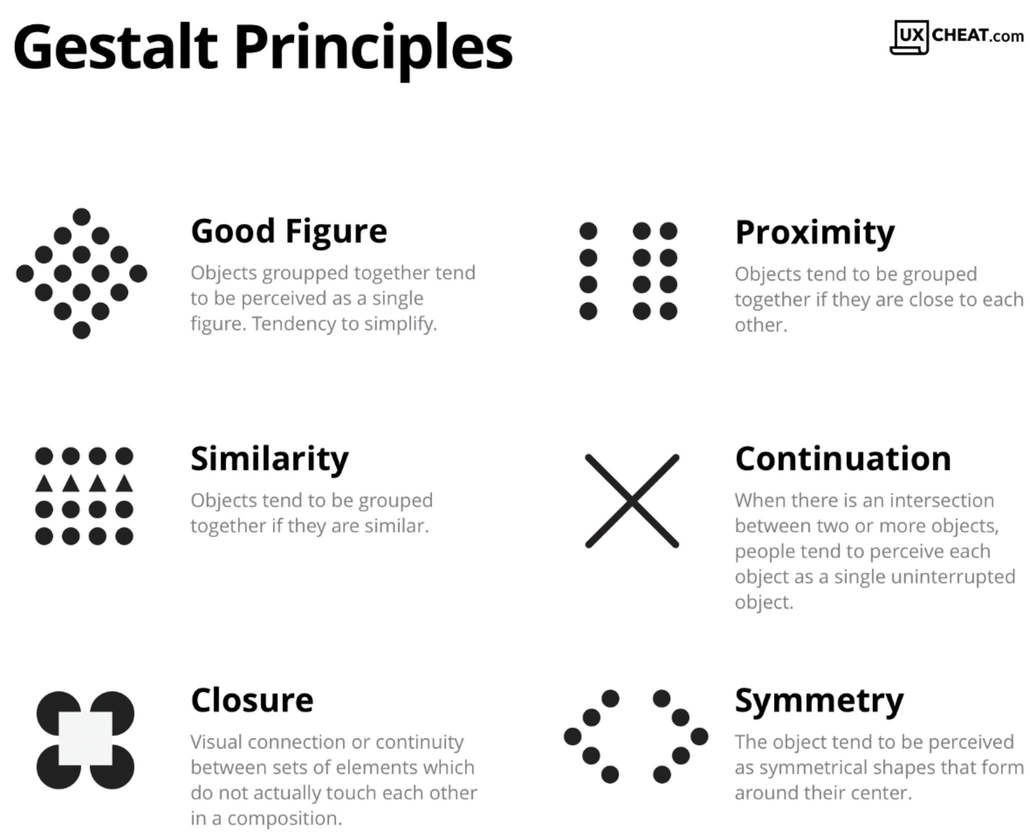
- top-down
- slow, deliberate
- focused
- singe-task

goal of information design

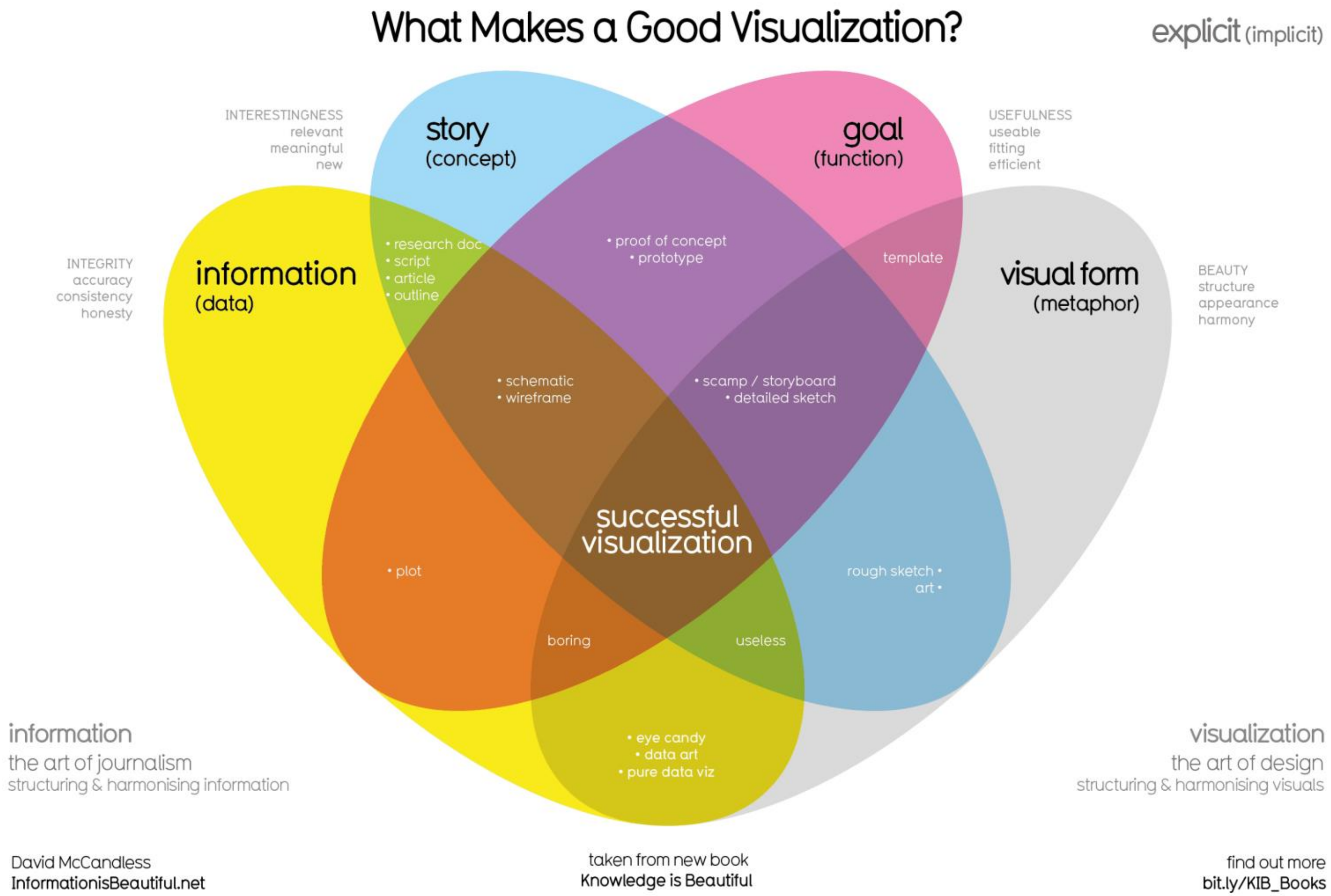
- help humans process information as efficiently as possible
- make as much use of pre-attentive processing as possible

Gestalt Principles

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

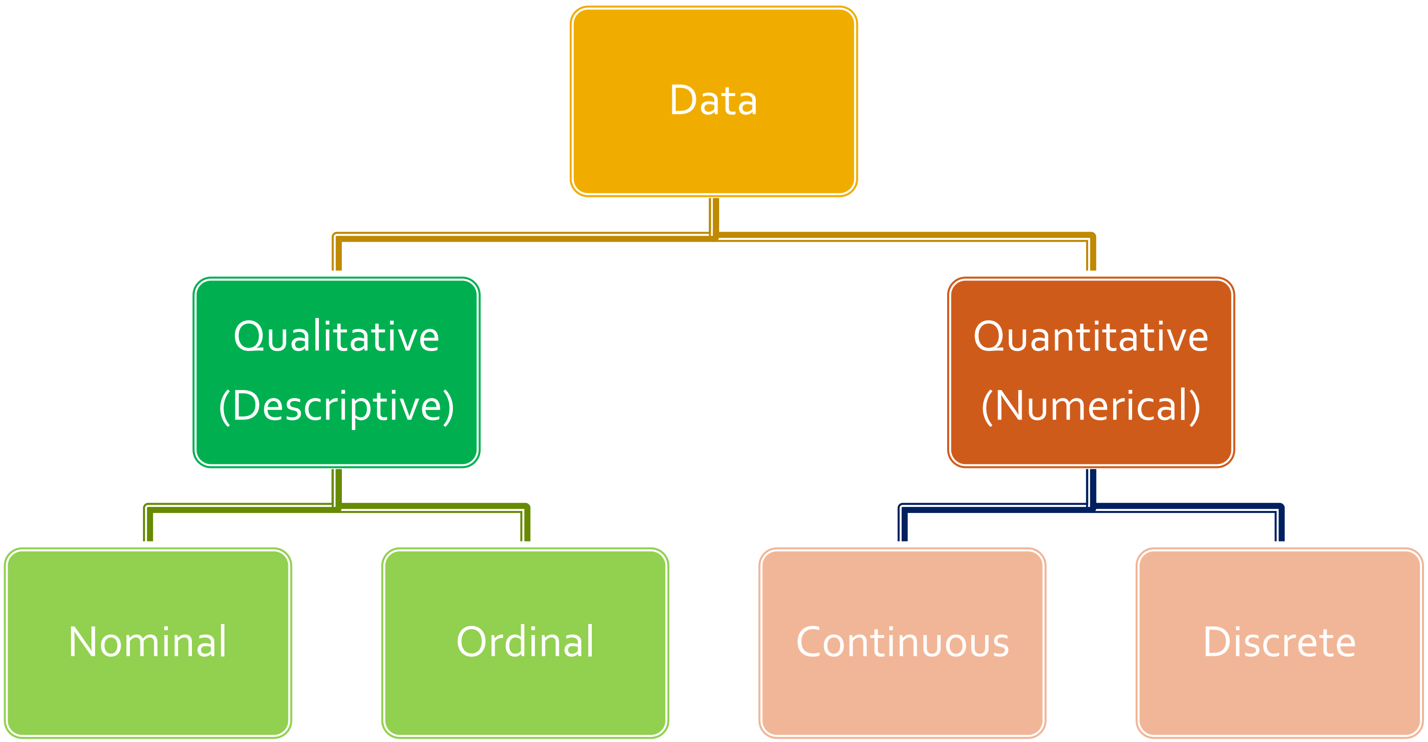


What Makes a Good Visualization?



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

Data Types



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.

Continuous

Data is measured on a continuous scale

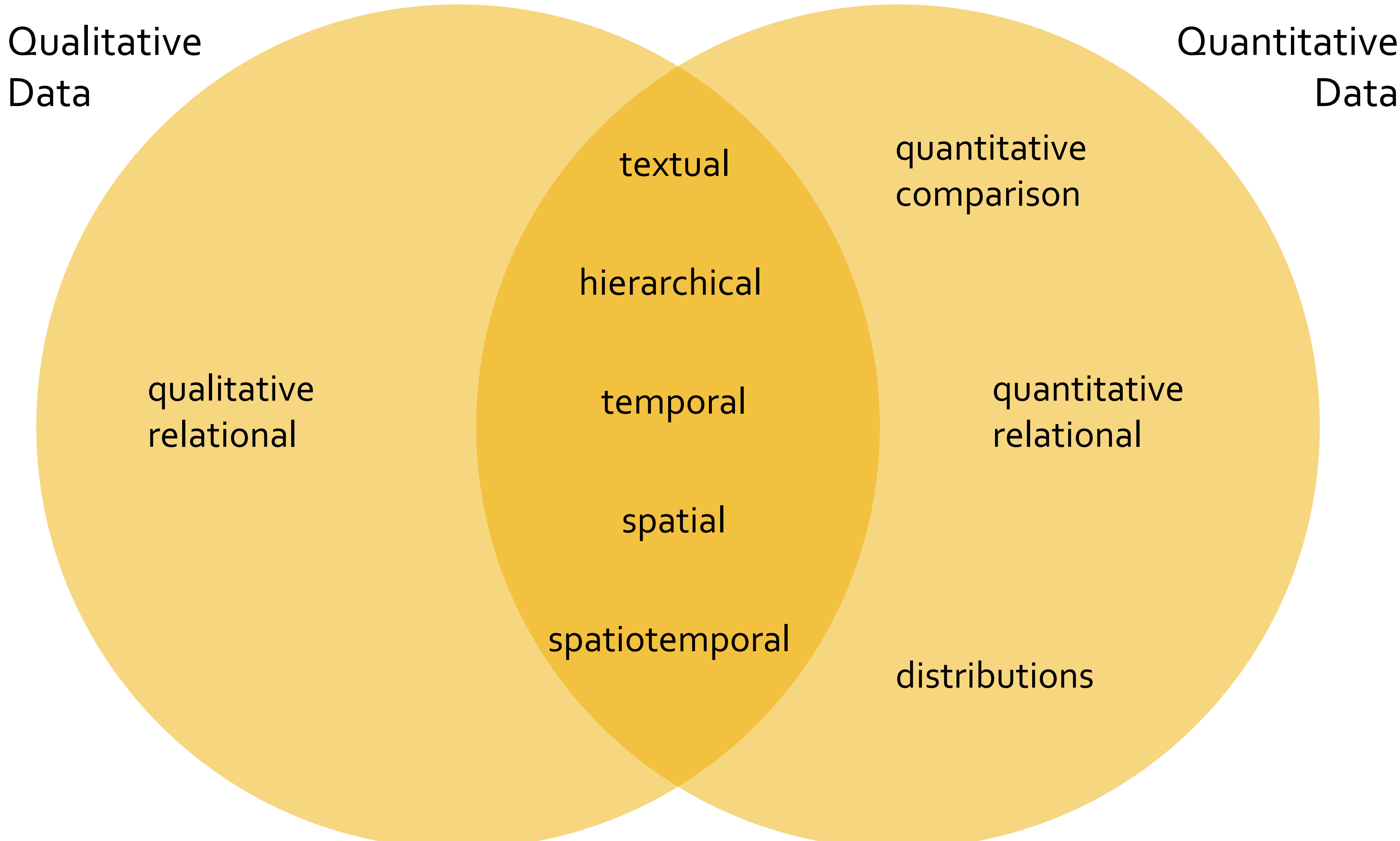
examples
temperature, length, height

Discrete

Data is countable, and exists only in whole numbers

examples
number of people taking this class

Information Visualization Taxonomy



Visual Elements

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**
  - **Wed, Apr 3rd**
    - 10 minutes (7 min presentation + 3 min QA)
    - Submit electronically your presentation (PPTX and PDF)
    - See course website for more info
- **Project final report**
  - **Sun, Apr 7th Midnight (11:59PM) Pacific Time**
    - Submit electronically your report (PDF) and code (zip)
    - 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

course evaluations ☺