Introduction to Data Visualization

CPS 563 – Data Visualization

Dr. Tam Nguyen

tamnguyen@udayton.edu

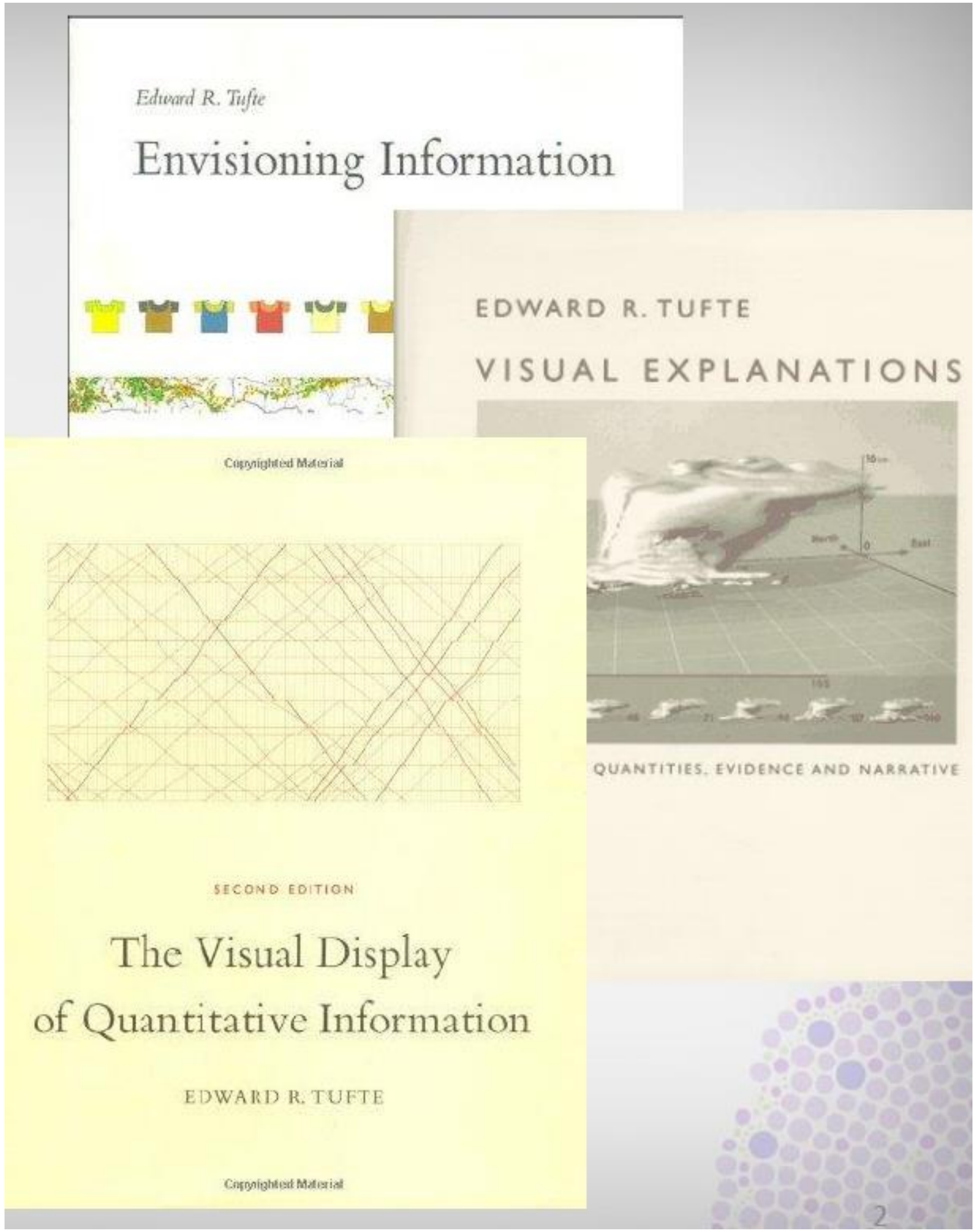
University of Dayton

Outline

- Books about Data Visualization
- What is Data?
- What is Big Data?
- Examples of Data Visualization
- Modes of Visualization

Books

- The Visual Display of Quantitative Information
- Visual Explanations
- Envisioning Information
  - Edward Tufte, Yale



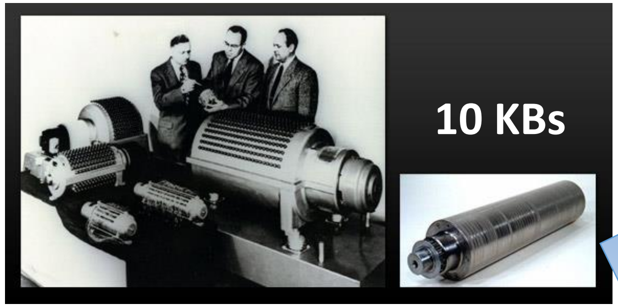
What is Data?

- Data is everywhere
  - Web data, e-commerce
  - Purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - Social Network (photos, texts, etc.)

Type of Data

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data

Data size



- Bit (1/0)
- Byte = 8 bits
- KB = 1000 bytes
- MB = 1000 KBs
- GB = 1000 MBs
- TB = 1000 GBs
- PB = 1000 TBs
- EB = 1000 PBs
- ZB = 1000 EBs
- YB = 1000 ZBs

What is “Big Data”?

- Big Data is also **data** but with a **huge size**.
- Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.

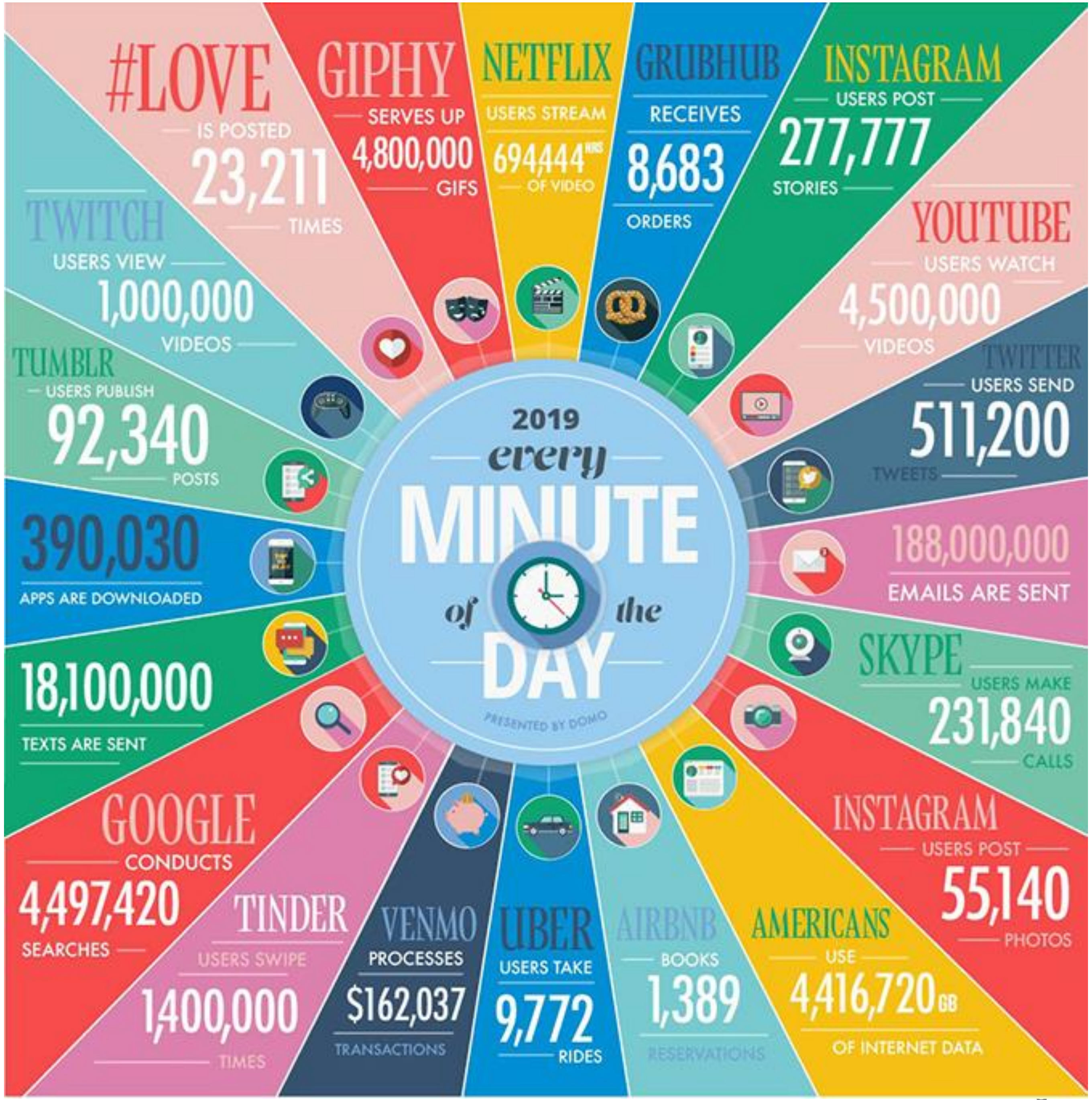
How much data then?

- Google processes 20 PB a day (2008)
- Facebook has 2.5 PB of user data/day (4/2009)
- eBay has 6.5 PB of user data/day (5/2009)

640K ought to be enough for anybody.

1981

How much data now?



Source: https://www.visualcapitalist.com/big-data-keeps-getting-bigger/

Big Data and Data Visualization

Again, what is Data Visualization?

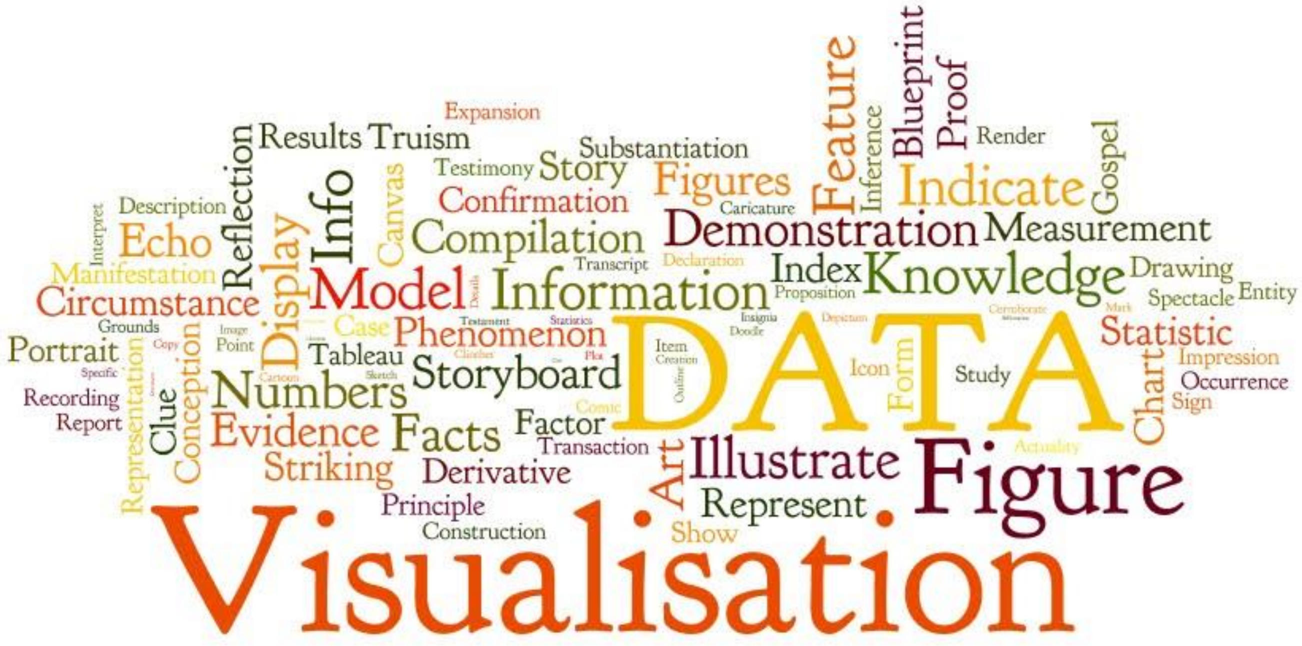
- "... finding the artificial memory that best supports our natural means of perception." [Bertin 1967]

- "Transformation of the symbolic into the geometric" [McCormick et al. 1987]

- "The use of computer-generated, interactive, visual representations of data to amplify cognition." [Card, Mackinlay, & Shneiderman 1999]

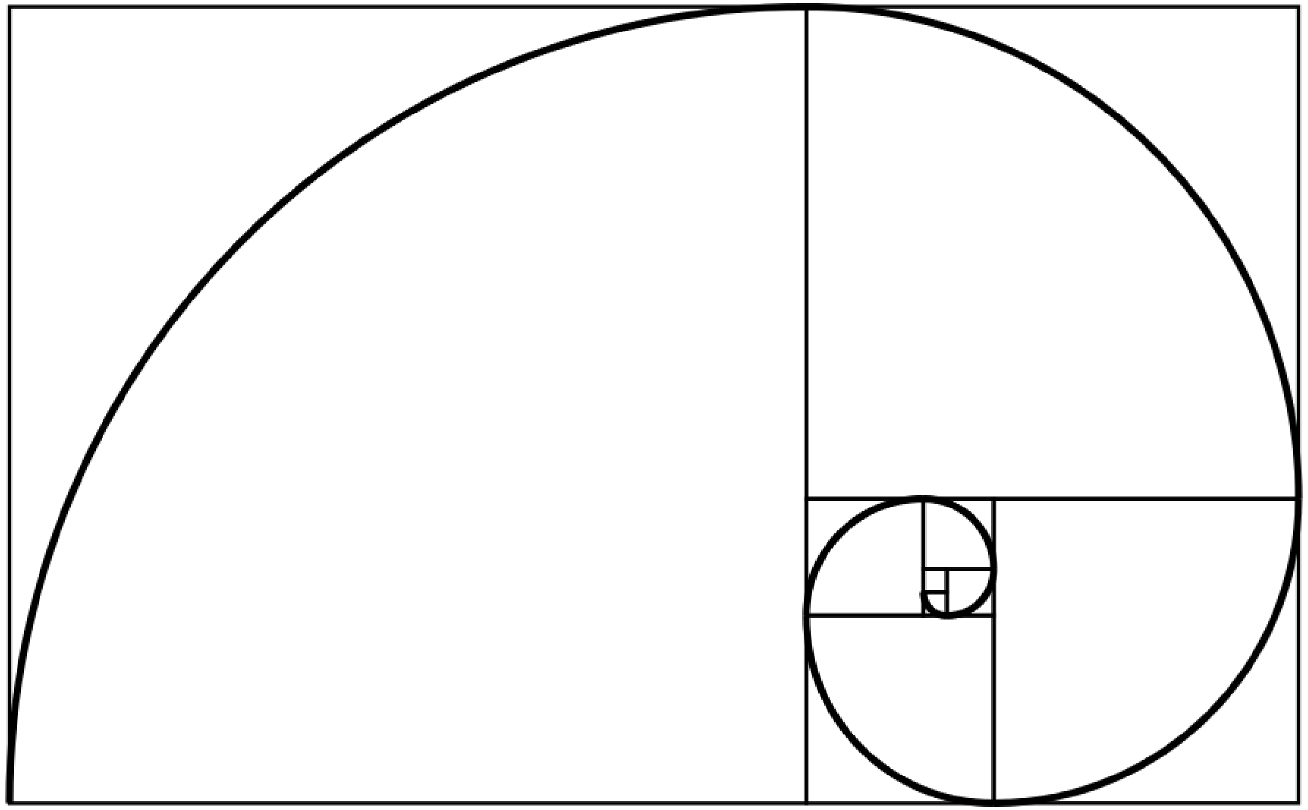
Again, what is Data Visualization?

- Visualization is the representation of the processed data graphically as a means of gaining understanding and insight into the data. It is sometimes referred to as visual data analysis.



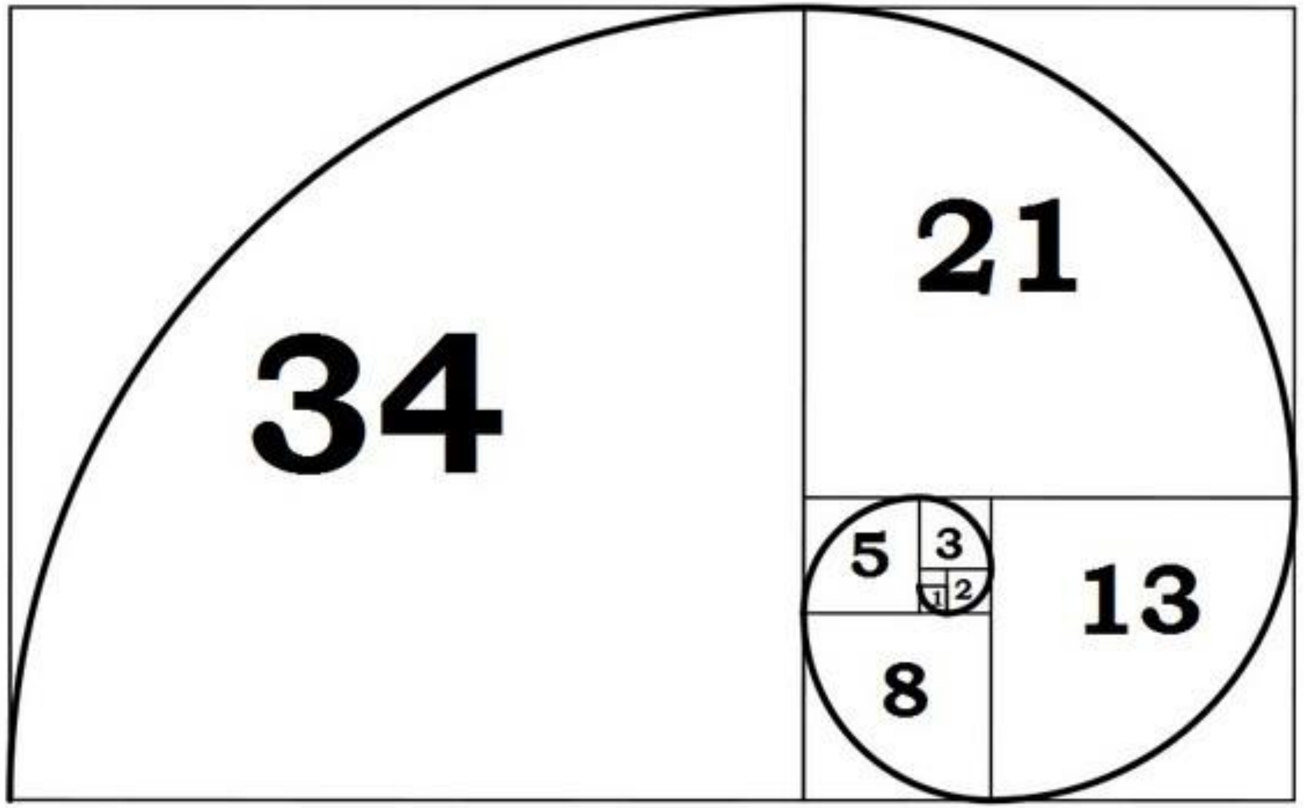
Mathematical Visualization

- Data results from mathematics
- Missing data can be readily generated by the computer



Mathematical Visualization

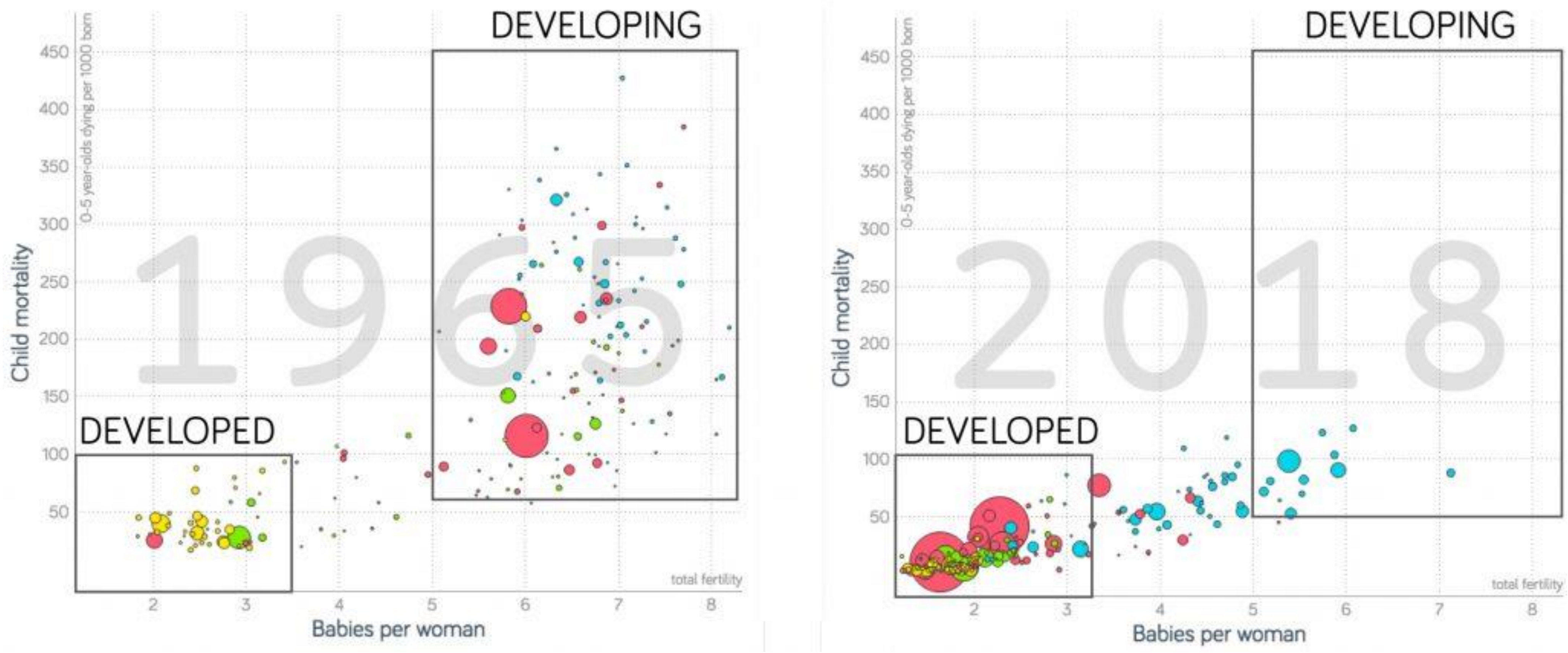
- Data results from mathematics
- Missing data can be readily generated by the computer



Scientific Visualization

- Visualization of scientific data
- Coordinate data
  - spatial coordinates
  - temperature, pressure, etc.
  - time
  - precision, recall
  - error rate

Example



Life expectancy vs. children per woman

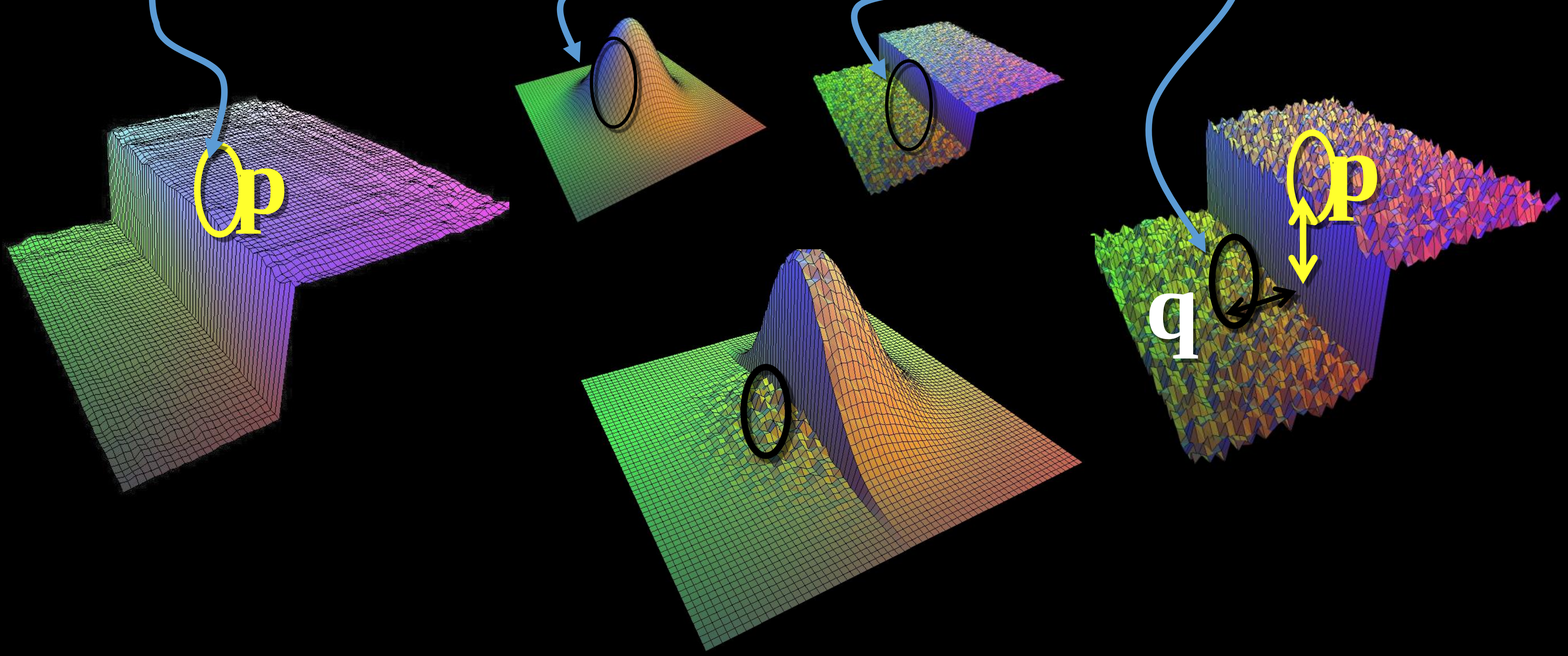
Another Example: Bilateral Filter on a Height Field

Formula: $$BF[I]_{\mathbf{p}} = \frac{1}{W_{\mathbf{p}}} \sum_{\mathbf{q} \in S} G_{\sigma_{\mathrm{s}}}\left(\| \mathbf{p} - \mathbf{q} \|\right) \; G_{\sigma_{\mathrm{r}}}\left(| I_{\mathbf{p}} - I_{\mathbf{q}} |\right) \; I_{\mathbf{q}}$$

Definition: A **bilateral filter** is a non-linear, edge-preserving, and noise-reducing smoothing **filter** for images. It replaces the intensity of each pixel with a weighted average of intensity values from nearby pixels.

Example: Bilateral Filter on a Height Field

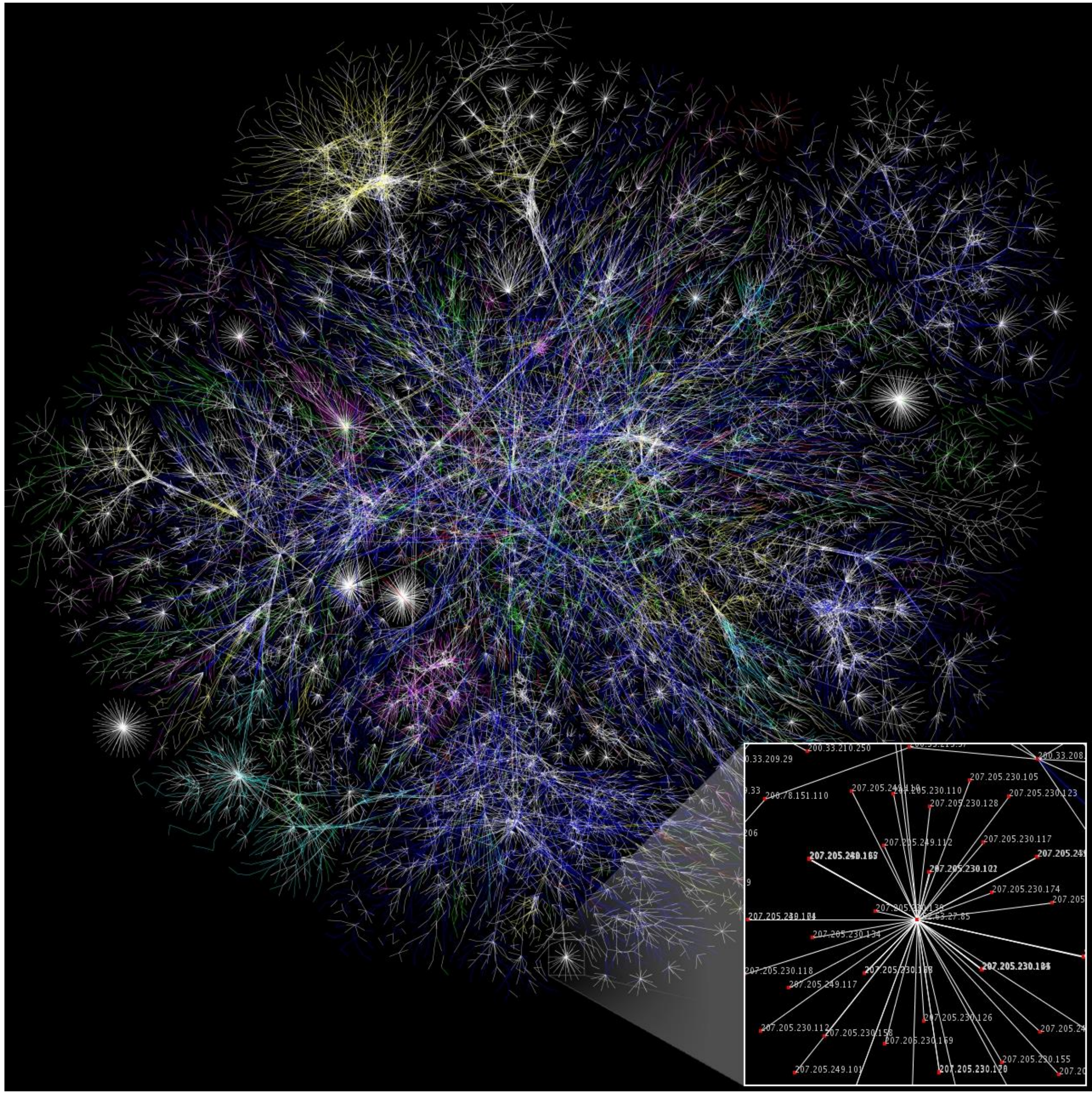
$$BF[I]_{\mathbf{p}} = \frac{1}{W_{\mathbf{p}}} \sum_{\mathbf{q} \in S} G_{\sigma_s}(\|\mathbf{p}-\mathbf{q}\|) \; G_{\sigma_r}(|I_{\mathbf{p}} - I_{\mathbf{q}}|) \; I_{\mathbf{q}}$$



reproduced from [Durand 02]

Information Visualization

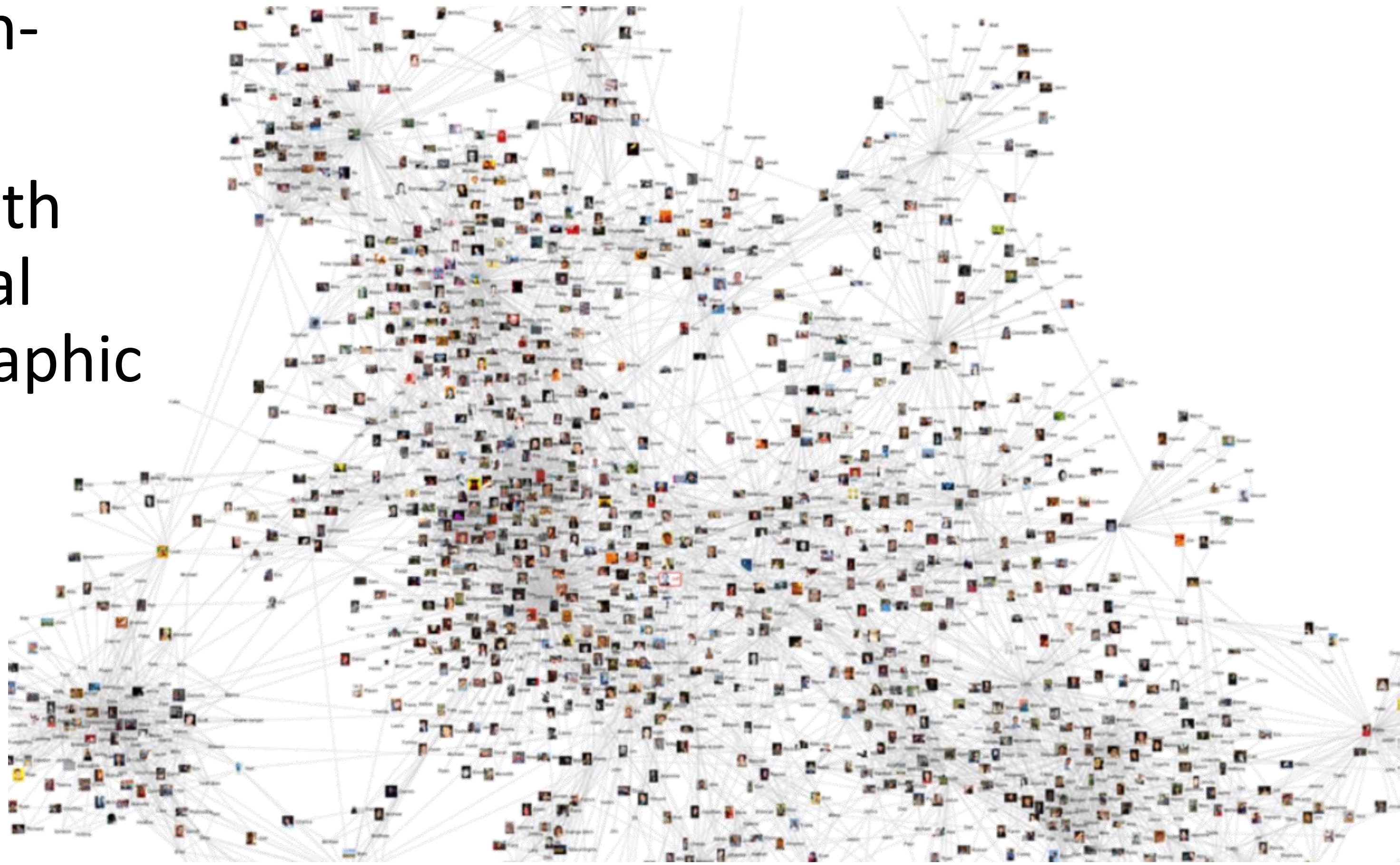
- Visualization of abstract, non-coordinate data
- The abstract data include both numerical and non-numerical data, such as text and geographic information.



Partial map of the Internet early 2005, each line represents two IP addresses, and some delay between those two nodes.

Information Visualization

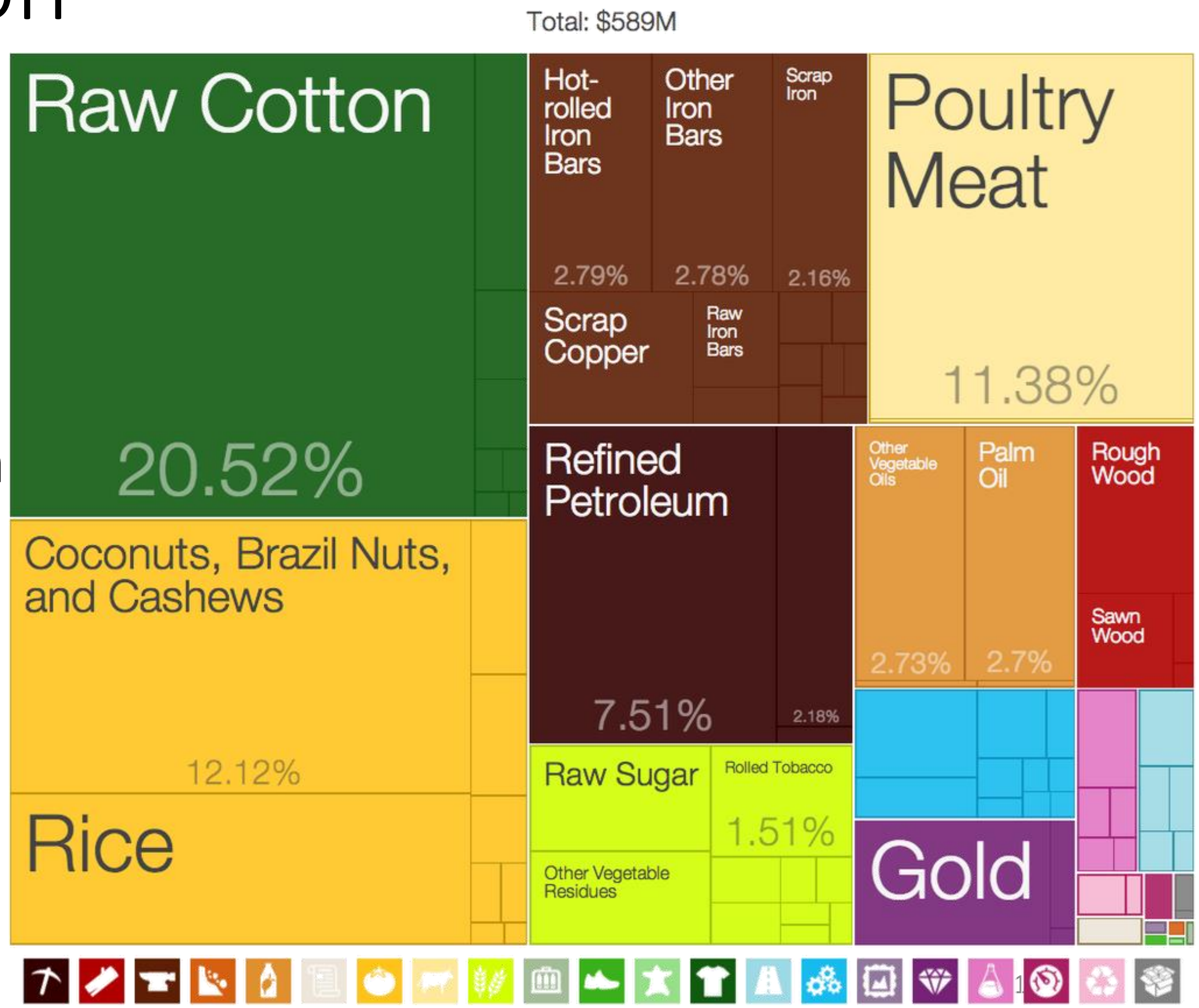
- Visualization of abstract, non-coordinate data
- The abstract data include both numerical and non-numerical data, such as text and geographic information.



"Making friends" on social media

Information Visualization

- Visualization of abstract, non-coordinate data
- The abstract data include both numerical and non-numerical data, such as text and geograph information.



Tree Map of Benin Exports (2009) by product category.

Data Visualization for Spotting Patterns

Paper Gestalt

Carven von Bearnensquash
Department of Computer Science
University of Phoenix
bearensquash@live.com

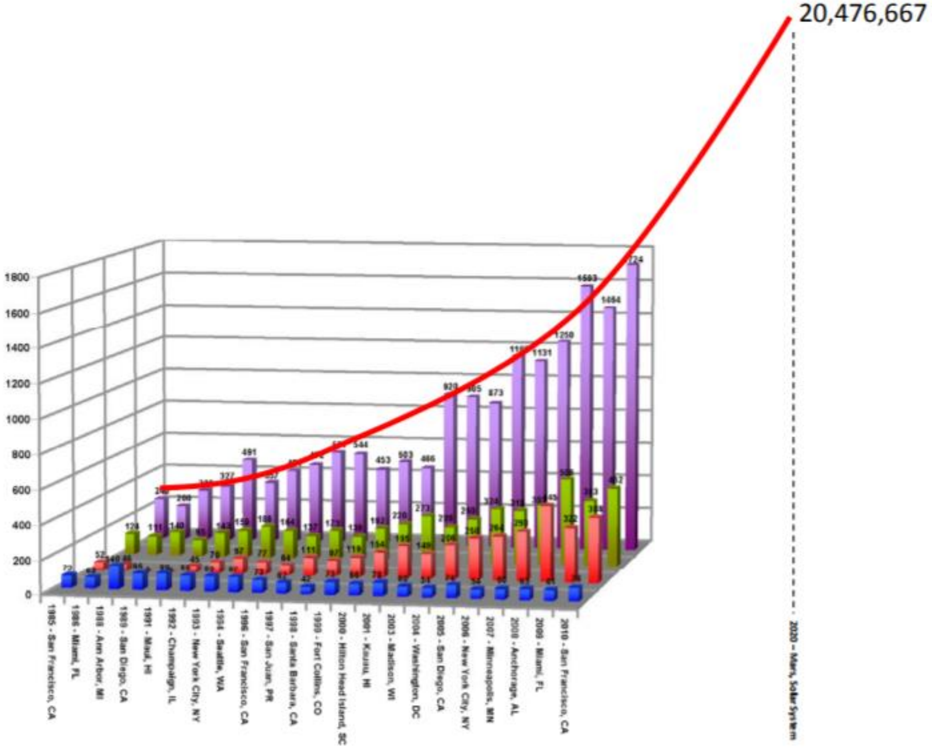


Figure 1. **Paper submission trends.** The number of submitted papers to CVPR, and other top tier computer vision conferences, is growing at an alarming rate. In this paper we propose an automated method of rejected sub-par papers, thereby reducing the burden on reviewers.

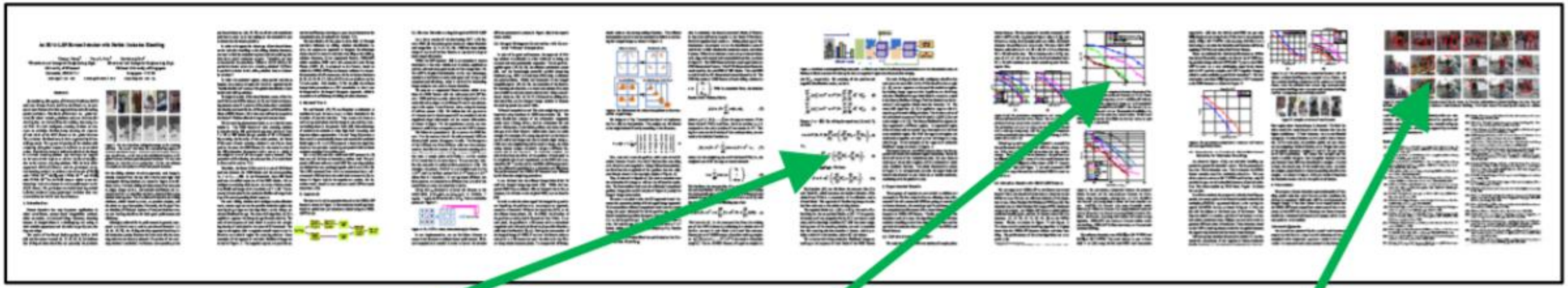
Why do we need to read research papers?

- Most ideas coming from research papers
- Most lectures coming from research papers
- Help us develop critical thinking skills
- Good for paper/report writing
- Good for future PhD study

Data Visualization for Spotting Patterns

Paper Gestalt

Carven von Bearnensquash
Department of Computer Science
University of Phoenix
bearensquash@live.com



Math: Sophisticated mathematical expressions make a paper look technical and make the authors appear knowledgeable and "smart".

Plots: ROC, PR, and other performance plots convey a sense of thoroughness. Standard deviation bars are particularly pleasing to a scientific eye.

Figures/Screenshots: Illustrative figures that express complex algorithms in terms of 3rd grade visuals are always a must. Screenshots of anecdotal results are also very effective.

Figure 6. **Characteristics of a "Good" paper.**

Patterns

Paper Gestalt

Carven von Bearnensquash
Department of Computer Science
University of Phoenix
bearensquash@live.com

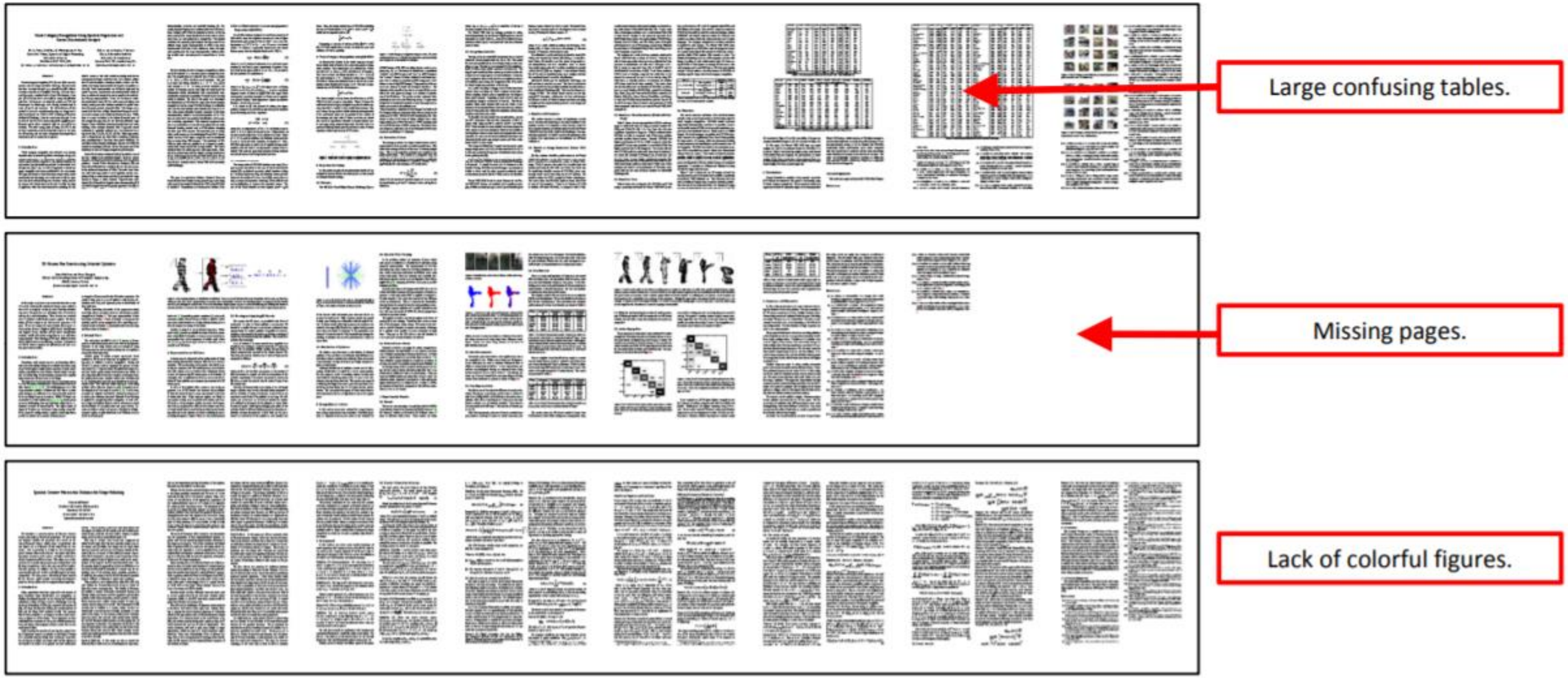


Figure 7. **Characteristics of a “Bad” paper.**

Patterns

An HOG-LBP Human Detector with Partial Occlusion Handling

Xiaoyu Wang* Tony X. Han* Shuicheng Yan†

*Electrical and Computer Engineering Dept.
University of Missouri
Columbia, MO 65211

†Electrical and Computer Engineering Dept.
National University of Singapore
Singapore 117576

xw9x9@mizzou.edu hantx@missouri.edu eleyans@nus.edu.sg

Abstract

By combining Histograms of Oriented Gradients (HOG) and Local Binary Pattern (LBP) as the feature set, we propose a novel human detection approach capable of handling partial occlusion. Two kinds of detectors, i.e., *global detector for whole scanning windows and part detectors for local regions, are learned from the training data using linear SVM. For each ambiguous scanning window, we construct an occlusion likelihood map by using the response of each block of the HOG feature to the global detector. The occlusion likelihood map is then segmented by Mean-shift approach. The segmented portion of the window with a majority of negative response is inferred as an occluded region. If partial occlusion is indicated with high likelihood in a certain scanning window, part detectors are applied on the unoccluded regions to achieve the final classification on the current scanning window. With the help of the augmented HOG-LBP feature and the global-part occlusion handling method, we achieve a detection rate of* 91.3% *with FPPW=* 10^{-6}, 94.7% *with FPPW=* 10^{-5}, *and* 97.9% *with FPPW=* 10^{-4} *on the INRIA dataset, which, to our best knowledge, is the best human detection performance on the INRIA dataset. The global-part occlusion handling method is further validated using synthesized occlusion data constructed from the INRIA and Pascal dataset.*

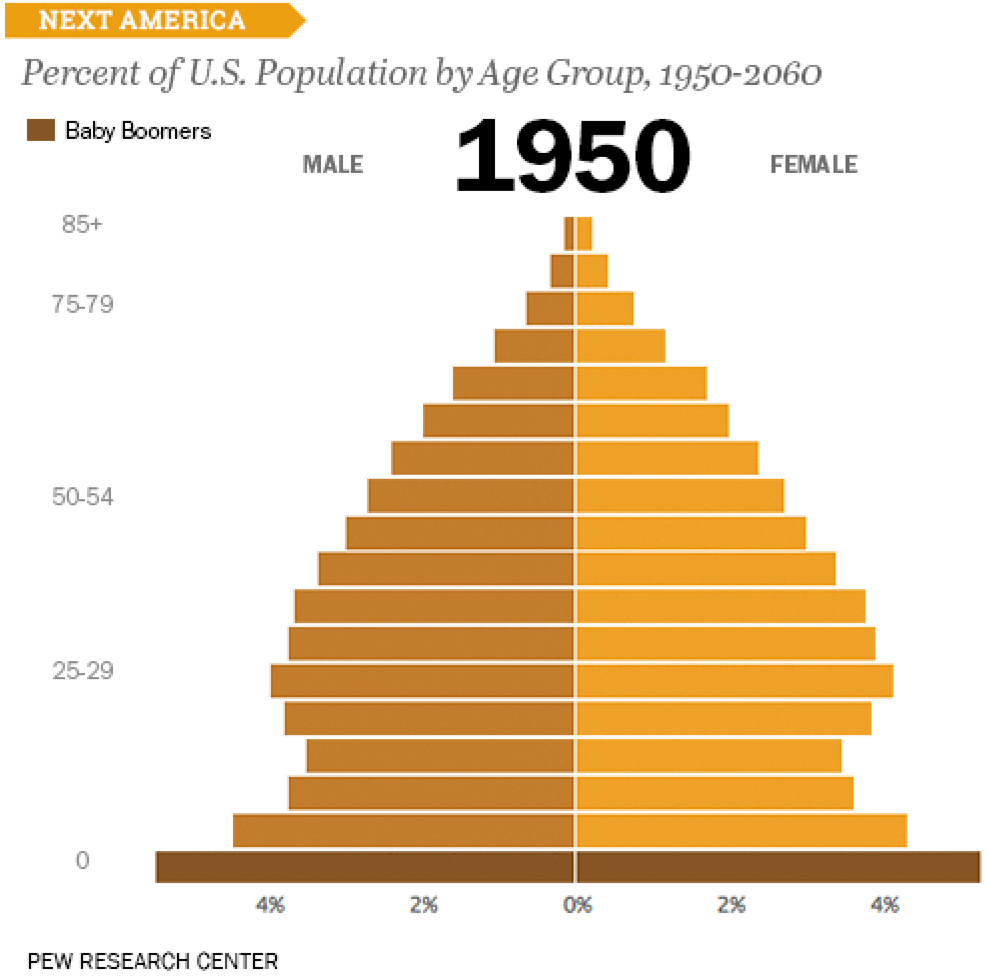
1. Introduction

Figure 1. The first row shows ambiguous images in the scanning windows. The second row shows the corresponding segmented occlusion likelihood images. For each segmented region, the negative overall score, *i.e.* the sum of the HOG block responses to the global detector, indicates possible partial occlusion. The first four columns are from the INRIA testing data. The last two columns are samples of our synthesized data with partial occlusion.

For the sliding window detection approach, each image is densely scanned from the top left to the bottom right with rectangular sliding windows (as shown in Figure 1) in different scales. For each sliding window, certain features such as edges, image patches, and wavelet coefficients are extracted and fed to a classifier, which is trained offline using labeled training data. The classifier will classify the sliding windows, which bound a person, as positive samples, and

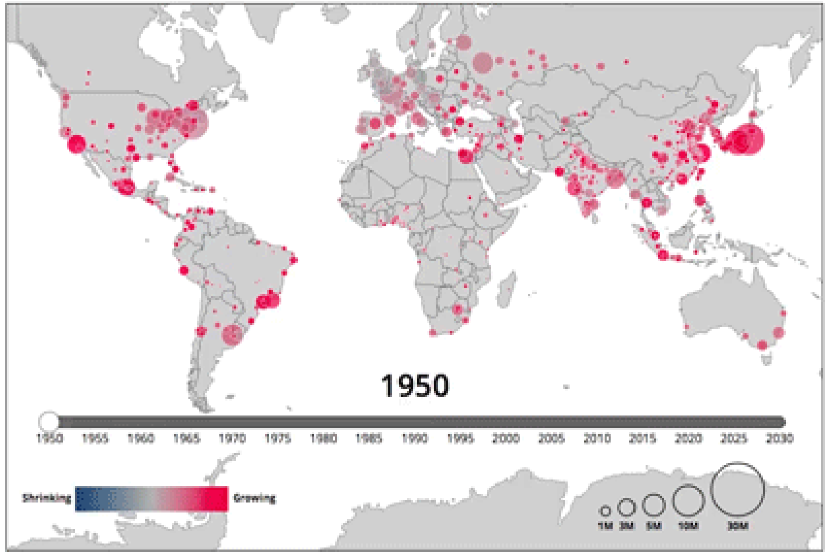
Why do we need Data Visualizations?

- Make decisions
- Enhance memory
- Find patterns
- Inspire
- Present argument or tell a story



Modes of Visualization

- **Interactive Visualization**
  - Used for discovery
  - Intended for a single investigator or collaborators
  - Rerenders based on input
  - Prototype quality
- **Presentation Visualization**
  - Used for communication
  - Intended for large group or mass audience
  - Does not support user input
  - Highly polished



Q&A