FROM VISION SCIENCE TO DATA SCIENCE: APPLYING PERCEPTION TO PROBLEMS IN BIG DATA

Remco Chang
Associate Professor, Computer Science
Tufts University
ROLE OF VISUALIZATION IN BIG DATA

- Big Data
  - Machine Learning
  - Databases
  - Visualization

- Visualization allows the user to interactively and visually explore large amounts of data
ROLE OF VISUALIZATION IN BIG DATA

- Big Data
  - Machine Learning
  - Databases
  - Visualization

Visualization allows the user to interactively and visually explore large amounts of data

Tableau
ROLE OF VISUALIZATION IN BIG DATA

- Big Data
  - Machine Learning
  - Databases
  - Visualization

Visualization allows the user to interactively and visually explore large amounts of data.
ROLE OF VISUALIZATION IN BIG DATA

- **Big Data**
  - Machine Learning
  - Databases
  - Visualization

- Visualization allows the user to **interactively** and **visually** explore large amounts of data

SAS Visual Analytics
DOMAIN-SPECIFIC VISUAL ANALYTICS SYSTEMS

- Financial Fraud Detection
  - Bank of America
- Political Simulation
  - DARPA
- Global Terrorism Database
  - DHS
- Bridge Maintenance
  - US DOT
- Biomechanical Motion
  - Brown University
APPLICATIONS OF VISUAL ANALYTICS

- Financial Fraud Detection
  - Bank of America
- Political Simulation
  - DARPA
- Global Terrorism Database
  - DHS
- Bridge Maintenance
  - US DOT
- Biomechanical Motion
  - Brown University
APPLICATIONS OF VISUAL ANALYTICS

- Financial Fraud Detection
  - Bank of America
- Political Simulation
  - DARPA
- Global Terrorism Database
  - DHS
- Bridge Maintenance
  - US DOT
- Biomechanical Motion
  - Brown University

APPLICATIONS OF VISUAL ANALYTICS

- Financial Fraud Detection
  - Bank of America
- Political Simulation
  - DARPA
- Global Terrorism Database
  - DHS
- Bridge Maintenance
  - US DOT
- Biomechanical Motion
  - Brown University
APPLICATIONS OF VISUAL ANALYTICS

- Financial Fraud Detection
  - Bank of America
- Political Simulation
  - DARPA
- Global Terrorism Database
  - DHS
- Bridge Maintenance
  - US DOT
- Biomechanical Motion
  - Brown University
INSUFFICIENT FOR BIG DATA?

- **Scalability**
  - Data growing faster than compute power
  - With Big Data, every user’s interaction can take seconds, minutes, hours, or days to complete.

- For *Library-of-Congress-scale* of IC data, existing tools do not scale to meet the needs (Dr. Fingar).
DATA VISUALIZATION PIPELINE
INTERACTIVE SYSTEMS FOR BIG DATA

- Effects of Latency
  - Interrupts analysis flow
  - Cannot meet “time sensitive” needs (Dr. Wilson)
  - Affects short-term memory
    - 500ms affects exploration
    - 10 minutes affects recall

- Conversely, highly interactive systems allow:
  - Higher “throughput” for analysts
  - Better “communication” between human and computer
  - Increase efficiency and accuracy of decision
Taking a Step Back: Big Data Visualization

- Visualize 10 million data points
  - “Information theoretic” upper-bound
  - 10 million data -> 1 million pixels
  - 10:1 reduction in information

- Problem is in fact worse, because...
  - There is a perceptual and cognitive limit
  - Adding more pixels doesn’t solve the problem

1000 pixels

1000x1000 = 1 million

1000 pixels
TURNING OBSTACLES INTO OPPORTUNITIES!

- Fitting billions of data points into million of pixels can be an obstacle...
  - Or it can be an opportunity!!
OPPORTUNITIES IN BIG DATA VISUALIZATION

1. **We should never fetch more than 1 million rows of data from the database**
   - How to pick the 1 million points?

2. **Since the user can’t see the details anyway, we can “simplify” the visualization a little for as long as the user can’t perceive the difference**
   - How to model the perceptual limits in data visualization?
A SIMPLE EXAMPLE OF PERCEPTION-DRIVEN COMPUTATION

- Simple Box-and-Whisker plots (or Box Plots)

- Spot the difference?
A Simple Example of Perception-Driven Computation

- This difference, although perceptually small, can mean the difference of scanning all data points or using a smaller sampled data set.

- If the user cannot perceive the difference, we can save computation time by not rendering the perfect visualization!
WHY VISION SCIENCE MATTERS

- Spot the difference?

- Although the difference between the two bars is the same as the previous example, we know from vision science that the difference here is easier to spot.

- But these are “rules”, not “models”
  - Does not indicate how many pixels for how much sampling
PERCEPTION-DRIVEN COMPUTATION

Perception & Cognitive Model

Visualization

Approximate Data & Computation
PERCEPTUAL MODELING IN DATA VISUALIZATION?

- Requires collaboration between the Perceptual Psychologists with Computer Scientists.

- One promising direction is to consider modeling the JND (just noticeable differences) in different visualization designs.

- Example of JND: Imagine yourself in a dark room...
PERCEPTUAL MODELING

- Weber’s Law (mid 1800s)
  - Low-level perceptual discrimination (sound, touch, taste, brightness, etc.)

\[
doP = k \frac{dS}{S}
\]

- Perceived Difference
- Change in Intensity
- Weber Fraction (via experiments)
- Intensity of the Stimulus
PERCEPTUAL MODELING

- Weber’s Law (mid 1800s)
  - Low-level perceptual discrimination (sound, touch, taste, brightness, etc.)

\[ dP = k \frac{dS}{S} \]

Given a fixed stimulus \( S \), the smallest of \( dS \) that can be perceived by humans is known as the “Just Noticeable Difference”, or JND
BALDRIDGE & RENSINK, 2010: PERCEPTION OF CORRELATION

$$\Delta r = 0.05$$
BALDRIDGE & RENSINK, 2010: PERCEPTION OF CORRELATION

$$\Delta r = 0.05$$
PERCEPTION OF CORRELATION

- In 2010, Ron Rensink and colleague found that the perception of correlation in a scatterplot follows Weber’s Law.

- In 2014, our team extended this work to determine if perception of correlation in all bivariate visualizations follows Weber’s Law.
The diagram shows three types of stacked plots:

1. **Stacked Area**
   - For $r = -1$, the area is uniformly filled.
   - As $r$ decreases to $-0.8$, $-0.3$, and $0.3$, the area becomes more sparse.
   - For $r = 0.8$, the area is almost empty.
   - At $r = 1$, the area is completely empty.

2. **Stacked Line**
   - The lines for $r = -1$ are dense and closely packed.
   - As $r$ decreases to $-0.8$, $-0.3$, and $0.3$, the lines become sparser and more spread out.
   - For $r = 0.8$, the lines are quite sparse.
   - At $r = 1$, the lines are almost nonexistent.

3. **Stacked Bar**
   - The bars for $r = -1$ are densely packed.
   - As $r$ decreases to $-0.8$, $-0.3$, and $0.3$, the bars become more spread out.
   - For $r = 0.8$, the bars are less dense.
   - At $r = 1$, the bars are almost nonexistent.

Each plot type demonstrates how the parameter $r$ affects the density and distribution of the data.
less precise

more precise
The perception of correlation in every tested chart can be modeled using Weber’s law.
Why Is This So Cool?

Positives:
1. These models have both descriptive and prescriptive powers
   - It “describes” what people see (and don’t see) from the data
   - It “predicts” what people will see given a new data
2. They can be used to quantitatively compare the effectiveness of visualizations
3. They can be used to drive computation (sampling)

Useful, but limited to correlation...
**Beyond Correlation: Visualization Metamers**

- **Metamers:**
  - Stimuli that are *different but appear the same* (under some condition)

- **Visualization Metamers**
  - Visualizations that *appear the same* (under some condition)
  - Uses much less data
  - Are much faster to render

- **Collaboration with Steve Franconeri (Northwestern) and Eugene Wu (Columbia)**
Visualization Metamers

- Focus: 1D visualizations
  - Bar charts, line graphs, heatmaps, etc.

- Properties of the model for visualization metamers
  - Recursive
  - Peaks and valleys
  - Slope
  - Concavity
  - Texture
  - ...

- Progressive Refinement
In 10 Years...

- **Science** of Perception (and Cognition)
  - What do people see (and not see) when perceiving data?
  - Can we model these properties and anticipate what mistakes will be made?

- **Engineering** of Data Systems based on Perception and Cognitive Principles
  - "Compute only what you can see"
  - Conversely, "Make more clear what an analyst cannot see"

- Two sides of the same coin. Collaboration between the fields (Psych, CogSci, CS, IC) is essential
QUESTIONS?
REMCO@CS.TUFTS.EDU
SUMMARY & CONCLUSION

- Need more cross-disciplinary research between Vision Science and Computer Science
- In 10 Years: need more systems to help analyst combat the data deluge
  - More CPUs is not the answer!
- Perception- and Cognitive- based approaches enable “building machines that make analysts smarter” (Dr. Klein)
  - Data analysis for the Masses (DARPA)
  - Support Bayesian Reasoning
  - Interaction Inferencing for Predictive Analysis
HYPOTHESIS

- Perception of Correlation follows Weber’s Law because the Visual Features used to discriminate Correlation follows Weber’s Law.
  - Regression (Logistic) Analysis of Visual Features vs. Judgment
  - Mathematical Confirmation
**VISUAL FEATURES**

- a. A 95% prediction ellipse
- b. A 95% confidence bounding box
- c. A convex hull
- d. The maximum distance
- e. A Minimum Spanning Tree
- f. k-Nearest Neighbors (each point connected to 3NN)
- g. All distances to the regression line
- h. The 75% percentile distance
HYPOTHESIS

- Perception of Correlation follows Weber’s Law because the Visual Features used to discriminate Correlation follows Weber’s Law.
  - Regression (Logistic) Analysis of Visual Features vs. Judgment
**HYPOTHESIS**

- Perception of Correlation follows Weber’s Law **because** the Visual Features used to discriminate Correlation follows Weber’s Law.
  - Mathematical Confirmation

Model for the Perception of Correlation (from our experiment):

\[ JND_r = k_r \cdot I_r + b_r \]

Model for the Perception of the Visual Feature (from Psych):

\[ JND_{vf} = k_{vf} \cdot I_{vf} + b_{vf} \]
HYPOTHESIS

- Perception of Correlation follows Weber’s Law because the Visual Features used to discriminate Correlation follows Weber’s Law.
  - Mathematical Confirmation

- If a Visual Feature (VF) is used to judge correlation, then the two equations should equate:

\[ JND_r = k_r I_r + b_r \iff JND_{vf} = k_{vf} I_{vf} + b_{vf} \]
Visualization Metamers: Beyond Correlation

Original

Immediate memory reproductions

Participant 1  Participant 2  Participant 3