Global Big Data Conference

BIG DATA BOOTCAMP
September 30th, Oct 1st & 2nd 2016

Denver

Colorado Convention Center, 700 14th St, Denver, CO 80202

www.globalbigdataconference.com
Twitter: @bigdataconf
Enabling a dialog between People & Data
Lessons in Designing for Big Data

© D.A. Szafir, 2016
LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
More Analysts (& Fewer Experts)
More Heterogeneity
More Questions

Why don’t we just compute the answer?
Data Sample 1:
Mean(x) = 9
Variance(x) = 11
Correlation(x, y) = 0.816
Regression: y = 3 + 0.5x

Data Sample 2:
Mean(x) = 9
Variance(x) = 11
Correlation(x, y) = 0.816
Regression: y = 3 + 0.5x

Data Sample 3:
Mean(x) = 9
Variance(x) = 11
Correlation(x, y) = 0.816
Regression: y = 3 + 0.5x

Data Sample 4:
Mean(x) = 9
Variance(x) = 11
Correlation(x, y) = 0.816
Regression: y = 3 + 0.5x
Data Sample 1:

\[
\begin{align*}
\text{Mean}(x) &= 9 \\
\text{Variance}(x) &= 11 \\
\text{Correlation}(x, y) &= 0.816 \\
\text{Regression}: y &= 3 + 0.5x
\end{align*}
\]

Data Sample 2:

\[
\begin{align*}
\text{Mean}(x) &= 9 \\
\text{Variance}(x) &= 11 \\
\text{Correlation}(x, y) &= 0.816 \\
\text{Regression}: y &= 3 + 0.5x
\end{align*}
\]

Data Sample 3:

\[
\begin{align*}
\text{Mean}(x) &= 9 \\
\text{Variance}(x) &= 11 \\
\text{Correlation}(x, y) &= 0.816 \\
\text{Regression}: y &= 3 + 0.5x
\end{align*}
\]

Data Sample 4:

\[
\begin{align*}
\text{Mean}(x) &= 9 \\
\text{Variance}(x) &= 11 \\
\text{Correlation}(x, y) &= 0.816 \\
\text{Regression}: y &= 3 + 0.5x
\end{align*}
\]
Statistical tools are powerful, but the human brain understands patterns
Apple Earnings Dashboard: September 2012 Quarter

How Apple’s Revenue Stacks Up (Billions)

- iPhone
- iPad
- iPod
- Mac
- Software
- iTunes
- Peripherals

iPhone shipments (Millions)
- Sep. ’11: 10
- Mar. ’12: 30
- Sep. ’12: 40

iPad shipments (Millions)
- Sep. ’11: 10
- Mar. ’12: 20
- Sep. ’12: 25

Mac shipments (Millions)
- Sep. ’11: 2
- Mar. ’12: 3
- Sep. ’12: 5

Revenue by Product (Billions)
- iPhone
- iPad
- iPod
- Mac
- Software
- iTunes
- Peripherals

© D.A. Szafir, 2016
Cleveland & McGill, 1985

More Effective
- Position
- Length
- Orientation
- Area
- Value/Lightness

Less Effective
- Color

© D.A. Szafir, 2016
What happens when our tools don’t suit our data?
Low-Level Tasks $\rightarrow$ Individual Values

- Los Angeles
- Phoenix

High-Level Tasks $\rightarrow$ Combine Many Values

- Midwest
- Southeast

© D.A. Szafir, 2016
Four types of ensemble coding in data visualizations

Danielle Albers Szafir
Department of Computer Sciences,
University of Wisconsin–Madison, Madison, WI, USA

Steve Haroz
Department of Psychology,
Northwestern University, Evanston, IL, USA

Michael Gleicher
Department of Computer Sciences,
University of Wisconsin–Madison, Madison, WI, USA

Steven Franconeri
Department of Psychology,
Northwestern University, Evanston, IL, USA

Ensemble coding supports rapid extraction of visual statistics about distributed visual information. Researchers typically study this ability with the goal of drawing conclusions about how such coding extracts

Kahn, 2012). Other types of information can be extracted and combined in parallel from large numbers of objects at once, such as the average object size (Arielv. 2001). A growing body of work seeks to

Binary Comparisons don’t scale!
<table>
<thead>
<tr>
<th>Visual Feature</th>
<th>Identification (Outlier)</th>
<th>Summary (Mean)</th>
<th>Segmentation (Clustering)</th>
<th>Structure Estimation (Trends)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orientation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color &amp; Luminance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Big Picture Analyses

Computational Aggregation: Compute the answer then visualize it

Visual Aggregation: Use the visual system to estimate the answer

© D.A. Szafir, 2016
Big Picture Analyses

Computational Aggregation: Compute the answer then visualize it

Visual Aggregation: Use the visual system to estimate the answer

© D.A. Szafir, 2016
Encodings

Tasks

Maxima

Minima

Range

Average

Variance

Outliers

© D.A. Szafir, 2016
What month has the highest sales day?
What month has the highest sales on average?
What month has the **highest sales on average**?
What month has the highest sales day?
How you map the data impacts what information is readily extracted.
Can we design better visualization systems that do support these analyses?
Two Challenges for Visualization

**Scalability**

How can we support insight across larger numbers and higher complexity?

**Comprehensibility**

How can we ensure estimates from a visualization are accurate?
Visualization in the Age of Big Data

Understand limits in current tools
  Large Scale Sequence Alignment

Derive inspiration across domains
  Literary Patterns

Link big and small
  Machine Learning & Molecules
Visualization in the Age of Big Data

Understand limits in current tools

What does the data look like?

Derive inspiration across domains

Literary Patterns

Link big and small

Machine Learning & Molecules
Limited Number of Sequences
Limited Length of Sequences
Only Reference-Dependent Analysis
Difficult to Analyze High-Level Relationships

Darling et al, 2004
Visualization in the Age of Big Data

Understand limits in current tools

What does the data look like?
The Fix: Aligning patterns with tasks

Derive inspiration across domains

 Literary Patterns

Link big and small

 Machine Learning & Molecules

© D.A. Szafir, 2016
Color better supports visual processing at scale

© D.A. Szafir, 2016
Color better supports visual processing at scale

© D.A. Szafir, 2016
Color better supports visual processing at scale

© D.A. Szafir, 2016
Color better supports visual processing at scale
Mapping America: Every City, Every Block

Browse local data from the Census Bureau's American Community Survey, based on samples from 2005 to 2009.

Distribution of racial and ethnic groups

Census tract 66
- Whites: 26%
- Blacks: 23%
- Hispanics: 68%
- Asians: 2%
- Other groups: 5%

© D.A. Szafir, 2016

Register Now.

Note: Due to limitations in the Census data, foreign-born populations are not available in all areas for all years.
Often too many genes to display on the monitor
Sequence Block

Average
Sequence Block

Robust Average
Sequence Block

Event Striping

© D.A. Szafir, 2016
Average

Event Striping

Robust Average

Color Weaving

© D.A. Szafir, 2016
Visualization in the Age of Big Data

Understand limits in current tools

What does the data look like?
The Fix: Aligning patterns with tasks
Building The System

Derive inspiration across domains

Literary Patterns

Link big and small

Machine Learning & Molecules
Sequence Surveyor

Task-Driven Sequence Aggregation
Length of Sequences

Perceptually-Driven Encoding
Number of Sequences

Dynamic Compression, Color and Position Encoding Choices
Variety of Tasks

© D.A. Szafir, 2016
10x More Sequences

100x Longer Sequences

Reference-Dependent, Independent, and Metadata-Based Analyses

Explicit Support for High-Level and Low-Level Relationships
Explore Evolutionary Patterns in Organisms
Explore Phylogenetic Relationships
Explore Phylogenetic Relationships
Explore Phylogenetic Relationships
"At a Glance" Algorithm Debugging
Color-based aggregation better supports analyses at scale
Visualization in the Age of Big Data

Understand limits in current tools
  Large Scale Sequence Alignment

Derive inspiration across domains
  Literary Patterns

Link big and small
  Machine Learning & Molecules

© D.A. Szafir, 2016
All the world's a stage,
And all the men and women merely players:
They have their exits and their entrances;
And one man in his time plays many parts,
His acts being seven ages. At first the infant,
Mewling and puking in the nurse's arms.
And then the whining school-boy, with his satchel
And shining morning face, creeping like snail
Unwillingly to school. And then the lover,
Sighing like furnace, with a woeful ballad
Made to his mistress' eyebrow. Then a soldier,
Full of strange oaths and bearded like the pard,
Jealous in honour, sudden and quick in quarrel,
Seeking the bubble reputation
Even in the cannon's mouth.
Large Digitized Collections

Google N-Grams: 5,195,769 books

© D.A. Szafir, 2016
Position

Length/Height

Orientation

Area

Value/Lightness

Color

More Effective

Less Effective

© D.A. Szafir, 2016

Cleveland & McGill, 1985
⚠️ Please don’t use wordclouds
More Effective

Position

Length/Height

Orientatio

Area

Value/Lightnes

Color

Less Effective

© D.A. Szafir, 2016

Cleveland & McGill, 1985
Word Usage Analysis Tasks

Characterize and compare authors

Measure shifts in an author's writing over time

Evolution of language over time

Evolution of cultural influences over time

Indicate recurring themes and topics

Characterize typographic conventions

© D.A. Szafir, 2016
Word Usage Analysis Tasks

Characterize and compare organisms

Measure shifts in organisms over species

Evolution of organisms over time

Evolution of cultural influences over time

Indicate recurring genetic material

Characterize typographic conventions

© D.A. Szafir, 2016
All the world's a stage,
And all the men and women merely players:
They have their exits and their entrances,

all the world a stage
and all the men and women merely players
they have their exits and their entrances

Turning texts into sequences

© D.A. Szafir, 2016
Text Sequence:
Present words in their original reading order

Highlight word locations

Precise analysis for single texts

Ranked Count:
Order words by how often they occur in a text collection

Highlight word frequency

Aggregate multiple texts

A Midsummer Night’s Dream

Text Sequence: now fair Hippolyta our nuptial hour draws on apace four

Ranked Count: the and to l you of a in my is

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
</table>
© D.A. Szafir, 2016
A Midsummer Night’s Dream

Text Sequence: now fair Hippolyta our nuptial hour draws on apace four

Ranked Count: the and to l you of a in my is

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.2 million books from 1660-2009
175,000 words over 35 decades

© D.A. Szafir, 2016

Mitchel et al, 2011
Explore Evolutionary Patterns in Writing
Confirm Prior Hypotheses
<table>
<thead>
<tr>
<th>1660</th>
<th>1670</th>
<th>1680</th>
<th>1690</th>
<th>1700</th>
<th>1710</th>
<th>1720</th>
<th>1730</th>
<th>1740</th>
<th>1750</th>
<th>1760</th>
<th>1770</th>
<th>1780</th>
<th>1790</th>
<th>1800</th>
<th>1810</th>
<th>1820</th>
<th>1830</th>
<th>1840</th>
</tr>
</thead>
</table>
Identify Cultural Shifts
Women

Wife

1910 - 1919

© D.A. Szafir, 2016
Author Attribution
Look for inspiration in other data domains
Visualization in the Age of Big Data

Understand limits in current tools
  Large Scale Sequence Alignment

Derive inspiration across domains
  Literary Patterns

Link big and small
  Machine Learning & Molecules
Hundreds of proteins with binding site predictions computed over hundreds of ligands

© D.A. Szafir, 2016
Task-driven overviews of large-scale machine learning performance data
DNA binding predictions over 216 proteins with 40 to 800 residues per protein

© D.A. Szafir, 2016
Binding big data with details gives findings more meaning
Visualization in the Age of Big Data

Understand limits in current tools
  Large Scale Sequence Alignment

Derive inspiration across domains
  Literary Patterns

Link big and small
  Machine Learning & Molecules

© D.A. Szafir, 2016
Designing for Big Data

Consider how the ways we communicate data support high-level tasks.

Look at parallels in the data structure and tasks associated with your data.

Don’t lose sight of the details.
Thank You!

Danielle Albers Szafir
danielle.szafir@colorado.edu
@dalbersszafir

Funding by BACTER and the NSF

Demos & Papers at: http://danielleszafir.com
Extra Slides