CSE 332
Introduction to Visualization

Visual Analytics & The Visual Sense Making Process

Klaus Mueller

Computer Science Department
Stony Brook University
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intro, schedule, and logistics</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Applications of visual analytics, data types</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Basic tasks</td>
<td>Project 1 out</td>
</tr>
<tr>
<td>4</td>
<td>Data preparation and representation</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Data reduction, notion of similarity and distance</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Dimension reduction</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Introduction to D3</td>
<td>Project 2 out</td>
</tr>
<tr>
<td>8</td>
<td>Visual perception and cognition</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Visual design and aesthetic</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Visual analytics tasks</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Cluster analysis</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>High-dimensional data, dimensionality reduction</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Visualization of spatial data: volume visualization intro</td>
<td>Project 3 out</td>
</tr>
<tr>
<td>14</td>
<td>Introduction to GPU programming</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Visualization of spatial data: raycasting, transfer functions</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Illumination and isosurface rendering</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Midterm</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Scientific visualization</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Non-photorealistic and illustrative rendering</td>
<td>Project 4 out</td>
</tr>
<tr>
<td>20</td>
<td>Midterm discussion</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Principles of interaction</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td><strong>Visual analytics and the visual sense making process</strong></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Predictive and causal analysis</td>
<td>Project 5 out</td>
</tr>
<tr>
<td>24</td>
<td>Visualization of graphs and hierarchies</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Visualization of time-varying and time-series data</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Memorable visualizations, visual embellishments</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Evaluation and user studies</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Narrative visualization, storytelling, and data journalism</td>
<td></td>
</tr>
</tbody>
</table>
RECAP: BRUSHING THE BASEBALL DATA

- How long in majors
- Avg. assists (x) vs. avg. putouts (y) (fielding ability)
- Distribution of positions played
- Select high salaries
- Avg. career home runs (y) vs. avg. career hits (x) (batting ability)
- Anything interesting?
Word Cloud of Your Responses

CAREER
HIGH
HITS
RUNS
HOMERUNS
PERFORMANCE
HIGHER
NUMBER
SALARIES
PLAYER
POSITION
PLAYER
HITS/yr
HOMESTAND
HOME
• Higher salaried players get more hits
• Home runs are less frequent than regular hits
• high salary is correlated with home runs
• the bipolt graphs may be combine into one complete graph (looks like the right one match the left one)
• the more home runs a person has in their career the higher their salary
• it seems the points from left graph is split along the 2 axes, nothing with both high values in both axes
• the position you play determines the salary
• "high salary players get lots of hits and homerunssecond base players are poor"
• Players with high salaries have been playing for a variable number of years, from 4 to around 18. Can say salary doesn't depend on the number of years played
• salary is tied to hitting but not fielding performance
yellow
highly paid players have higher hits per year
high salary players tend to take position of 1B
more salary, more years of experience and they have more homeruns hit with a higher battering accuracy
high salary for high number of average career hits & home runs
Players make less money as they get older
Almost all of the high paid athletes seem to fall in the upper end of all performance metrics
The highest salaries are most frequent for 2-3 positions.
Players need # of hits to earn high salary. # of homerun does not influence it much
More putouts/home runs/hits... higher salary
more runs, more money, but not necessary more school
the high paying players are outliers in terms of career hits
people who make high salaries have more hits/yr
higher the salary the more years the player plays the more home runs he hits
baseball players tend to make a higher salary in the middle of their career
the higher number of homeruns and hits correlate with higher salary
interesting trend on the scatter plot
The higher putons the lower assists
people who have higher salaries doing better in games
if you are making dem stacks boi, you are most likely a 4-6 year guy or a 11-13 year guy. Theres a big gap!
the more salary the player has, the more possibility that the player would hit the ball.
high salaries pays off
• Seems impossible to earn a high salary in the first three years
• High salaried players have a bimodal distribution (peaking around 7 & 13 yrs)
• Hits/Year a better indicator of salary than HR/Year
• High paid outlier with low HR and medium hits/year.
  (Reason: person is player-coach)
• There seem to be two differentiated groups in the put-outs/assists category (but not correlated with salary)
Why Visual Analytics?

Big Data

12+ TBs of tweet data every day

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones worldwide

25+ TBs of log data every day

100s of millions of GPS enabled devices sold annually

76 million smart meters in 2009... 200M by 2014

2+ billion people on the Web by end 2011
Visual Analytics
PROBLEMS WITH SCALABILITY

Must be scalable to

- number of data points
- number of dimensions
- data sources
- diversity of data sources
- number of users
- diversity of users and tasks
- quality of the data

Visual Analytics comes to the rescue...
Ease understanding of the data by providing an effective visual representation

*Amplify Perception*

*Detect the Expected, Discover the Unexpected™*
Visualization plus...

- interaction (HCI)
- data processing (analytics)
- story telling
- scientific approach

but also...

- intelligent computing (AI, machine learning)
- behavioral psychology (cognitive science, human factors)

**Visual Analytics is the science of analytical reasoning supported by a highly interactive visual interface**

Agenda setting book:
http://nvac.pnl.gov/agenda.stm
The Daniel Keim Mantra of Visual Analytics

“Analyze First - Show the Important – Zoom, Filter and Analyze Further - Details on Demand"

The triangle of Visual Analytics (VA)
Intelligence analysis is challenging
Huge amounts of data
Low signal vs. noise (SNR)
Many data types
  - text, images, video, sensor data, etc.
Uncertainty
Contradictions
Omissions
Use of Visualization

Visual perception
- high bandwidth
- fast screening of a lot of data
- pattern recognition
- higher-level cognition

Interaction
- direct manipulation
- two-way communication

*Recall intro lecture on the human visual system...*
Use of Visualization

Visual perception
- high bandwidth
- fast screening of a lot of data
- pattern recognition
- higher-level cognition

Interaction
- direct manipulation
- two-way communication

Recall intro lecture on the human visual system...
But... humans are imperfect
Humans tend to overlook/ignore non-focus (and unexpected) objects even when very close and obvious
- note the Visual Analytics slogan: *Detect the Unexpected*

Humans also have limited working memory
- fine details are quickly forgotten when focus changes
- big effect in animated or interactive visualizations
- need to preserve temporal context
EXAMPLE #1

Spot a difference?

This is called change blindness
In this video you will do some counting.

It is very important that you get the right number!

Ready?

[watch video](https://www.youtube.com) (YouTube)

Video by Dan Simons (U Illinois)
Another distraction experiment

watch video (YouTube)

Video by Dan Simons (U Illinois)
Thoroughly studied by Dan Simons (U Illinois)
- see http://www.dansimons.com/index.html

Visual Analytics tools
- help human analysts cope with insufficient memory
  → visualizations externalize memory
  → allow humans to perform *visual queries* (see C. Ware book)
- help human analysts deal with change blindness
  → analytics can detect changes
  → visualization can highlight/emphasize these changes
- we have seen many visual tools this semester
  → this lecture is more about strategy building
Another deficiency of humans

- humans tend to stick with an “opinion” for a long time
- how long does it take you to switch?

man/woman

Young/old woman
The Magic Number Seven

- $\pm 2$: the number of things most people can keep in working memory at one time
- causes problems for complicated analysis

An excellent book that has more on this topic
- discussed next
Strategies for Dealing with Complexity

Decomposition
- decompose a complex problem into simpler problems
- get one’s thinking straight in these simpler problems

Externalization
- get the decomposed problem out of one’s head and down on paper or on a computer screen in some simplified form
- shows the main variables, parameters, or elements of the problem and how they relate to each other

Recall principles of information visualization
- overview and detail
- focus and context
- analyze, filter, zoom,...
Mentioned his method of solving decision problems

Why is the decision problem so difficult?
  - folks cannot keep all pros and cons in mind at the same time

Solution?
  - write down all the pros and cons onto paper in some visible, shorthand form
  - allows you make a global judgment effectively
You want to choose the best car among various cars

What is the best car?
- lowest maintenance cost?
- highest resale value?
- slickest styling?
- best gas mileage?
- largest trunk space?

How to make a decision?
- create the car purchase matrix
#1 List the important attributes you want to maximize

- Price
- Maintenance Cost
- Styling
- Gas Mileage
- Comfort
- Handling
#2 Quantify the Relative Importance of Each Attribute

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>30%</td>
</tr>
<tr>
<td>Operating Cost</td>
<td>10%</td>
</tr>
<tr>
<td>Styling</td>
<td>20%</td>
</tr>
<tr>
<td>Comfort</td>
<td>20%</td>
</tr>
<tr>
<td>Handling</td>
<td>15%</td>
</tr>
<tr>
<td>Safety</td>
<td>5%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>
#3 Judge Each Car How it Values on Each of these Attributes

<table>
<thead>
<tr>
<th></th>
<th>%Value</th>
<th>Car 1</th>
<th>Car 2</th>
<th>Car 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>30%</td>
<td>3.5%</td>
<td>3.0%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Operating Cost</td>
<td>10%</td>
<td>3.5%</td>
<td>2.0%</td>
<td>4.5%</td>
</tr>
<tr>
<td>Styling</td>
<td>20%</td>
<td>2.5%</td>
<td>4.5%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Comfort</td>
<td>20%</td>
<td>4.0%</td>
<td>2.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Handling</td>
<td>15%</td>
<td>3.0%</td>
<td>4.0%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Safety</td>
<td>5%</td>
<td>3.5%</td>
<td>2.5%</td>
<td>4%</td>
</tr>
</tbody>
</table>
#4 Multiply the Overall Attribute Value by the Car’s Attribute Value

<table>
<thead>
<tr>
<th></th>
<th>% Value</th>
<th>Car 1</th>
<th>Car 2</th>
<th>Car 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>30%</td>
<td>105</td>
<td>90</td>
<td>105</td>
</tr>
<tr>
<td>Operating Cost</td>
<td>10%</td>
<td>35</td>
<td>20</td>
<td>45</td>
</tr>
<tr>
<td>Styling</td>
<td>20%</td>
<td>50</td>
<td>90</td>
<td>60</td>
</tr>
<tr>
<td>Comfort</td>
<td>20%</td>
<td>80</td>
<td>50</td>
<td>70</td>
</tr>
<tr>
<td>Handling</td>
<td>15%</td>
<td>45</td>
<td>60</td>
<td>45</td>
</tr>
<tr>
<td>Safety</td>
<td>5%</td>
<td>17.5</td>
<td>12.5</td>
<td>20</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>332.5</td>
<td>322.5</td>
<td>345</td>
</tr>
</tbody>
</table>
When working on difficult intelligence issues

- which is the correct explanation?
- which is the most likely outcome?

Alternative hypothesis
Analysis of Competing Hypotheses = ACH

Used to
- aid judgment on important issues
- minimize cognitive limitations

Basic insights from
- cognitive psychology
- decision analysis
- scientific method
1. Identify hypothesis
2. List evidence
3. Prepare matrix
4. Refine matrix
5. Draw conclusions
6. Analyze conclusions
7. Report conclusions
8. Identify milestones

Check evidence source
**Step 1: Identify Hypotheses**

Hypothesis generation vs. hypothesis evaluation
- **generation**: bring together all possibilities
- **evaluation**: focus on each of them and rule out from weak to strong

Disproved vs. unproven
- for a disproved hypothesis there is positive evidence that it is wrong
- for an unproven hypothesis, there is no evidence that it is correct
Step 2: List Evidence

Don’t limit to the evidences currently available

For each hypothesis, list supporting and contradicting factors

Absence and presence of evidence

- for example: Did the dog bark in the night?
  no. nobody heard it barked (absence)
## Step 3: Prepare Matrix

<table>
<thead>
<tr>
<th>Event (E)</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1. Saddam public statement of intent not to retaliate.</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E2. Absence of terrorist offensive during the 1991 Gulf War.</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>E3. Assumption that Iraq would not want to provoke another US attack.</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>E4. Increase in frequency/length of monitored Iraqi agent radio broadcasts.</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E5. Iraqi embassies instructed to take increased security precautions.</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E6. Assumption that failure to retaliate would be unacceptable loss of face for Saddam.</td>
<td>- -</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
**Question:**

**Will Iraq Retaliate for a US Bombing?**

H1: Iraq will not retaliate
H2: It will sponsor some minor terrorist actions.
H3: Iraq is planning a major terrorist attack, perhaps against one or more CIA installations

<table>
<thead>
<tr>
<th>Evidence</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1. Saddam public statement of intent not to retaliate.</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E2. Absence of terrorist offensive during the 1991 Gulf War.</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>E3. Assumption that Iraq would not want to provoke another US attack.</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>E4. Increase in frequency/length of monitored Iraqi agent radio broadcasts.</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E5. Iraqi embassies instructed to take increased security precautions.</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E6. Assumption that failure to retaliate would be unacceptable loss of face for Saddam.</td>
<td>---</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Step 4: Refine Matrix

Diagnostic value – likeliness of hypothesis
- high-temp indicates sickness, but can’t determine which illness

Reconsider the hypotheses
- add, or need finer distinction
- combine

Reconsider the evidences
- put in missing factors
- delete evidence that have no diagnostic value
Work down the matrix, looking at each hypothesis
Proceed by trying to disprove the hypotheses rather than prove them

<table>
<thead>
<tr>
<th></th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1. Saddam public statement of intent not to retaliate.</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E2. Absence of terrorist offensive during the 1991 Gulf War.</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>E3. Assumption that Iraq would not want to provoke another US attack.</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>E4. Increase in frequency/length of monitored Iraqi agent radio broadcasts.</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E5. Iraqi embassies instructed to take increased security precautions.</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>E6. Assumption that failure to retaliate would be unacceptable loss of face for Saddam.</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>
Analyze how sensitive your conclusion is to a few critical items of evidence

- the consequences if the evidence were wrong
- check the original source
Decision-maker needs to make decisions on the basis of a full set of alternative possibilities

The importance is on eliminating not confirming!

Discuss the relative likelihood of all the hypotheses
Analytical conclusion should always be regarded as tentative. Specify in advance things will change possibly.
### Summary and Conclusion

Key differences b/t competing hypotheses from conventional intuitive analysis

<table>
<thead>
<tr>
<th></th>
<th>ACH</th>
<th>Conventional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of possibilities</td>
<td>Full set</td>
<td>Most likely one</td>
</tr>
<tr>
<td>Diagnostic value</td>
<td>Greatest</td>
<td>Maybe no</td>
</tr>
<tr>
<td>Use of evidence</td>
<td>Refute</td>
<td>Confirm</td>
</tr>
</tbody>
</table>
Support visualization with computations for data processing
Form a loop: visualize - refine
Gather (forage) information
Re-represent
  ▪ choose form that aids analysis
Develop insight
  ▪ through manipulation of representation
Produce results
  ▪ “product”
Nominal Sense-Making Process
Reasoning Artifacts

Elemental artifacts
  ▪ source intelligence, evidence, assumptions

Pattern artifacts
  ▪ relationships, temporal and spatial structure

Higher-order knowledge constructs
  ▪ arguments, causality, models

Complex reasoning constructs
  ▪ hypotheses, scenarios

All these become part of the Visual Analytics sense-making (reasoning) process
Standard Information Displays

Showing about 50 - 300 data values; 10-20 dimensions
Cross-filtering

- a method for interactively expressing sequences of multidimensional set queries by selecting and filtering unique data values across pairs of views
- the next 2 slides are due to Chris Weaver (U Oklahoma) – check out his *IEEE Trans Visualization and Computer Graphics* (16(2): 192-204) paper
Cross-Filtering

A general pattern for constructing an interdependent set of data transformation operations that support the method:

- Group (\(\gamma\)) data records into sets for each unique attribute value.
- Filter (\(\phi\)) each set, keeping records whose attribute values match those selected in other views.
- Project/visually encode (\(\pi\)) each value and its filtered set.
- Select (\(\sigma\)) values/sets corresponding to brushed glyphs in the view.
## Design Variations

<table>
<thead>
<tr>
<th>Attributes</th>
<th>KEDS</th>
<th>Hotels</th>
<th>Retrosheet</th>
<th>Cinegraph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>code (event)</td>
<td>name (guest)</td>
<td>name (home team, away team)</td>
<td>name (movie, genre, oscar, person, role)</td>
</tr>
<tr>
<td>Temporal</td>
<td>date (event)</td>
<td>date (visit)</td>
<td>date &amp; time (game)</td>
<td>date (release)</td>
</tr>
<tr>
<td>Spatial</td>
<td>region (countries)</td>
<td>location (hotel, residence)</td>
<td>location (stadium)</td>
<td>-</td>
</tr>
<tr>
<td>Numerical</td>
<td>cooperative/conflictual weight</td>
<td>-</td>
<td>capacity, attendance, temperature, wind speed</td>
<td>box office, rating average, rating count</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Auxiliary Views</th>
<th>Pre-filter</th>
<th>Post-filter</th>
<th>Detail</th>
<th>Nested</th>
</tr>
</thead>
<tbody>
<tr>
<td>list (data sources)</td>
<td>-</td>
<td>map (world)</td>
<td>drill-down table, split time series</td>
<td>scatter plot (date vs. weight)</td>
</tr>
<tr>
<td>-</td>
<td>map (Pennsylvania)</td>
<td>drill-down table</td>
<td>1-D heatmap (visit count by date)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>map (North America), rich drill-down table</td>
<td>-</td>
<td>1-D heatmap (game count by date)</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>histogram (rating distribution)</td>
<td></td>
</tr>
<tr>
<td>sliders (ratings &amp; roles thresholds)</td>
<td>attribute relationship graph</td>
<td>movie viewer</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Use Visualizations to Evoke The Right Thoughts
How Many 9s Do You See?
How Many 9s Do You See?
**Who has the best profit and who has the worst sales?**

<table>
<thead>
<tr>
<th>Product Type</th>
<th>Product</th>
<th>Central</th>
<th>East</th>
<th>South</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sum of Profit</td>
<td>Sum of Sales</td>
<td>Sum of Profit</td>
<td>Sum of Sales</td>
<td>Sum of Profit</td>
</tr>
<tr>
<td>Coffee</td>
<td>Amaretto</td>
<td>$5,105</td>
<td>$14,011</td>
<td>$1,009</td>
<td>$2,993</td>
</tr>
<tr>
<td></td>
<td>Columbian</td>
<td>$8,528</td>
<td>$28,913</td>
<td>$27,253</td>
<td>$47,386</td>
</tr>
<tr>
<td></td>
<td>Decaf Irish Cream</td>
<td>$9,632</td>
<td>$26,155</td>
<td>$2,727</td>
<td>$6,261</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>$23,265</td>
<td>$69,080</td>
<td>$30,989</td>
<td>$56,640</td>
</tr>
<tr>
<td>Espresso</td>
<td>Caffe Latte</td>
<td>$14,640</td>
<td>$35,218</td>
<td>($6,230)</td>
<td>$16,646</td>
</tr>
<tr>
<td></td>
<td>Caffe Mocha</td>
<td>$8,860</td>
<td>$24,485</td>
<td>$2,410</td>
<td>$7,722</td>
</tr>
<tr>
<td></td>
<td>Decaf Espresso</td>
<td>$10,062</td>
<td>$24,036</td>
<td>$5,930</td>
<td>$15,384</td>
</tr>
<tr>
<td></td>
<td>Regular Espresso</td>
<td>$23,500</td>
<td>$59,703</td>
<td>$6,242</td>
<td>$48,405</td>
</tr>
<tr>
<td>Herbal Tea</td>
<td>Chamomile</td>
<td>$14,434</td>
<td>$36,570</td>
<td>$765</td>
<td>$2,194</td>
</tr>
<tr>
<td></td>
<td>Lemon</td>
<td>$6,251</td>
<td>$21,978</td>
<td>$7,901</td>
<td>$27,176</td>
</tr>
<tr>
<td></td>
<td>Mint</td>
<td>$4,069</td>
<td>$9,337</td>
<td>$4,069</td>
<td>$9,337</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>$24,754</td>
<td>$67,885</td>
<td>$6,424</td>
<td>$41,362</td>
</tr>
<tr>
<td>Tea</td>
<td>Darjeeling</td>
<td>$10,772</td>
<td>$30,289</td>
<td>$6,497</td>
<td>$14,096</td>
</tr>
<tr>
<td></td>
<td>Earl Grey</td>
<td>$10,331</td>
<td>$32,881</td>
<td>$3,405</td>
<td>$6,505</td>
</tr>
<tr>
<td></td>
<td>Green Tea</td>
<td>$1,227</td>
<td>$5,211</td>
<td>$5,654</td>
<td>$11,571</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>$22,330</td>
<td>$68,380</td>
<td>$15,557</td>
<td>$32,172</td>
</tr>
</tbody>
</table>
Who has the best profit and who has the worst sales?

<table>
<thead>
<tr>
<th>Product Type</th>
<th>Product</th>
<th>Central</th>
<th>East</th>
<th>South</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffee</td>
<td>Amaretto</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Columbian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decaf Irish Cream</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Espresso</td>
<td>Caffe Latte</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Caffe Mocha</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decaf Espresso</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Regular Espresso</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Herbal Tea</td>
<td>Chamomile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lemon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mint</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tea</td>
<td>Darjeeling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Earl Grey</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Green Tea</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>0K 20K 40K 60K</td>
<td>0K 20K 40K 60K</td>
<td>0K 20K 40K 60K</td>
<td>0K 20K 40K 60K</td>
</tr>
</tbody>
</table>
Do The Right Analytics, Don’t Just Visualize Data
Doubling down on states for strong growth

Maria
Senior Sales Analyst
March 15th, 2012
Today’s question

In which states should we invest additional marketing spend during the upcoming campaign?

Based upon sales growth potential...
2011 sales by state

2011 sales per state (top/bottom 3 labeled)

ID at $697 0.0%

Top- TX at $467,644 11.6%
2011 sales by state

The top 3 states generate 30% of sales
The top 8 states 50%
And the top 21 80%
Potential sales by state???

Is there a better metric?
The emphasis is on potential

Average sale per capita for top states multiplied by
Current population of top sales states
What are the top states based on sales per capita?

2011 sales per million residents by state (top/bottom 3 labeled)

Top sales states are quite low in sales per million people! Great potential!
### Highest growth potential in top 8

#### Top 8 states by 2011 sales, 3 year growth potential

<table>
<thead>
<tr>
<th>State</th>
<th>Growth Potential Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>1558%</td>
</tr>
<tr>
<td>TX</td>
<td>1006%</td>
</tr>
<tr>
<td>NY</td>
<td>822%</td>
</tr>
<tr>
<td>FL</td>
<td>810%</td>
</tr>
<tr>
<td>PA</td>
<td>620%</td>
</tr>
<tr>
<td>IL</td>
<td>514%</td>
</tr>
<tr>
<td>OH</td>
<td>414%</td>
</tr>
<tr>
<td>VA</td>
<td>336%</td>
</tr>
</tbody>
</table>

+ If we were to pick just one state, California has the greatest potential

+ The next tier is Texas, New York & Florida
Useful metrics

1. Total sales per state was OK

2. Better: Total sales per million residents per capita is better than looking at existing customers, because we want new customers

3. Top five states to target: 90\textsuperscript{th} percentile +
Dashboards should pass the 5-second test
Important rules:

- most important view goes on top or top left
- legends go near their views
- avoid using multiple color schemes
- use 5 views or fewer
- provide interactivity