Scalable Data Analysis (CIS 602-01)

Introduction

Dr. David Koop
NYC Taxi Data

[Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance, T. W. Schneider]
What are your questions about this data?
NYC Taxi Data

[Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance, T. W. Schneider]
NYC Taxi Data: Day analysis

Number of Trips for the years of 2011, and 2012

[Figure showing the number of trips for the years of 2011 and 2012, with multiple graphs depicting the trend over time.]
NYC Taxi Data: Region analysis

[Image of maps and graphs showing taxi data over different neighborhoods in NYC]

- Number of trips per time
- Fare/mile per time

For example, by examining the average cost of trip per mile, we can see that it is higher within Manhattan. This provides evidence of the higher density of taxi services in this area.

Another interesting observation is that the number of trips originating at airports and major train stations (i.e., Penn Station and Grand Central) is significantly higher than that of other regions.

This one-week overview provides an accurate overview of city life, reflecting the behavior of many commuters who go to the City during the week, and starts to decrease on Friday, hitting a low on Saturday. This is accentuated during rush hour on weekdays, when trips take much longer (see Fig. 10). The nighttime peak for pickups on Friday (May 6) in the evening—this indicates that it is the Friday night of the Labour Day weekend when most people go out.

Peak for pickups on Friday (May 6) in the evening—this indicates that it is the Friday night of the Labour Day weekend when most people go out. The figure also shows that most people go to Midtown (the darkest region), followed by the Lower East Side and the Upper East Side. But over the weekend, there is an increased number of dropoffs in Downtown. The figure also highlights the fact that Harlem is underserved by taxis.

Studying how taxi demand varies over time provides valuable insights for decision making, both to schedule driver shifts and maximize profits. To simplify the process of comparing multiple information sets, TaxiVis provides a mechanism.

Using the summary view, we can further explore features of the data. For example, by selecting 2011, May and Sunday, 5 times, we can simply group them using the Group/Ungroup button (note the two green outlines); we also group the parameters associated with the trips we found one surprising fact: the fare per mile is lower for Harlem, and thus, there is less economic incentive for drivers that go to Harlem.

[Image of a map showing taxi data by neighborhood]

A significant drop is observed in the number of trips assigned to its time range. This is illustrated in Fig. 11. Here, we select the regions of interest and maximize profits. To compare the number of trips originating at the airports with trips starting at the train stations, Penn Station and Grand Central, we can simply select the regions (by double-clicking on them), and dropoffs. Starting from the map view shown in Fig. 8, we can group them (Fig. 9). In the regular selection mode, the list of time ranges is already expressed and generated by the widget. For example, by selecting 2011, May and Sunday, 5 times, we can simply group them using the Group/Ungroup button (note the two green outlines); we also group the parameters associated with the trips we found one surprising fact: the fare per mile is lower for Harlem, and thus, there is less economic incentive for drivers that go to Harlem.

Note the increase in the number of trips that starts to happen on Thursday (May 5), with big peaks on Friday (May 6). This is a common practice when the destination is JFK. Another interesting question is where passengers go. The choropleth (Fig. 12) shows that most people go to Midtown (the darkest region), followed by the Lower East Side and the Upper East Side.

The Upper West Side.

This one-week overview provides an accurate overview of city life, reflecting the behavior of many commuters who go to the City during the week, and starts to decrease on Friday, hitting a low on Saturday. This is accentuated during rush hour on weekdays, when trips take much longer (see Fig. 10). The nighttime peak for pickups on Friday (May 6) in the evening—this indicates that it is the Friday night of the Labour Day weekend when most people go out.

Peak for pickups on Friday (May 6) in the evening—this indicates that it is the Friday night of the Labour Day weekend when most people go out. The figure also shows that most people go to Midtown (the darkest region), followed by the Lower East Side and the Upper East Side. But over the weekend, there is an increased number of dropoffs in Downtown. The figure also highlights the fact that Harlem is underserved by taxis.

Studying how taxi demand varies over time provides valuable insights for decision making, both to schedule driver shifts and maximize profits. To simplify the process of comparing multiple information sets, TaxiVis provides a mechanism.

Using the summary view, we can further explore features of the data. For example, by selecting 2011, May and Sunday, 5 times, we can simply group them using the Group/Ungroup button (note the two green outlines); we also group the parameters associated with the trips we found one surprising fact: the fare per mile is lower for Harlem, and thus, there is less economic incentive for drivers that go to Harlem.

Note the increase in the number of trips that starts to happen on Thursday (May 5), with big peaks on Friday (May 6). This is a common practice when the destination is JFK. Another interesting question is where passengers go. The choropleth (Fig. 12) shows that most people go to Midtown (the darkest region), followed by the Lower East Side and the Upper East Side.

The Upper West Side.
Marine Traffic Data
Marine Traffic Data
Baseball Data

[Deitrich et al., 2014]
**Baseball Data**

[Deitrich et al., 2014]
Baseball Data

[Deitrich et al., 2014]
Mobile Data Growth

Overall mobile data traffic is expected to grow to 49 exabytes per month by 2021, a sevenfold increase over 2016. Mobile data traffic will grow at a CAGR of 47 percent from 2016 to 2021 (Figure 2).

Figure 2. Cisco Forecasts 49 Exabytes per Month of Mobile Data Traffic by 2021

Source: Cisco VNI Mobile, 2017

Asia Pacific will account for 47 percent of global mobile traffic by 2021, the largest share of traffic by any region by a substantial margin, as shown in Figure 3. North America, which had the second-largest traffic share in 2016, will have only the fourth-largest share by 2021, having been surpassed by Central and Eastern Europe and Middle East and Africa. Middle East and Africa will experience the highest CAGR of 65 percent, increasing 12-fold over the forecast period. Asia Pacific will have the second-highest CAGR of 49 percent, increasing 7-fold over the forecast period (Figure 3).

Figure 3. Global Mobile Data Traffic Forecasts by Region

Source: Cisco VNI Mobile, 2017

[Cisco Visual Networking Index Mobile, 2017]
Mobile Video Keeps Growing

Figure 23. IP Traffic by Access Technology

Note: Fixed/Wi-Fi from Mobile Devices may include a small amount of Fixed/Wired from Mobile Devices

Source: Cisco VNI Mobile, 2017

Trend 5: Identifying New Mobile Applications and Requirements

Because mobile video content has much higher bit rates than other mobile content types, mobile video will generate much of the mobile traffic growth through 2021. Mobile video will grow at a CAGR of 54 percent between 2016 and 2021, higher than the overall average mobile traffic CAGR of 47 percent. Of the 49 exabytes per month crossing the mobile network by 2021, 38 exabytes will be due to video (Figure 24). Mobile video represented more than half of global mobile data traffic beginning in 2012.

Figure 24. Mobile Video Will Generate More Than Three-Quarters of Mobile Data Traffic by 2021

Note: Figures in parentheses refer to 2016 and 2021 traffic share.

Source: Cisco VNI Mobile, 2017

Mobile Video Keeps Growing

Mobile File Sharing (2%,2%)
Mobile Audio (8%,5%)
Mobile Web/Data/VoIP (30%,14%)
Mobile Video (60%,78%)

Exabytes per Month


Note: Figures in parentheses refer to 2016 and 2021 traffic share.

[Cisco Visual Networking Index Mobile, 2017]
Data Science Venn Diagram

[D. Conway, The Data Science Venn Diagram, 2013]
Questions are important!

• Having data is great, but most of the time it just sits waiting for someone to analyze it
• The reason data analysis is not completely automated is that there are so many potential questions
• Humans need to stay involved in the loop
• Interaction and visualization can be important, especially early in data analysis
Scalability

• “Big Data”
  - What is “big”? For whom is it “big”?
  - variety, velocity, volume, …

• Lots of data that was big is not an issue now

• Understanding the scalability of techniques is important

• There will always be larger datasets, want to understand
  - how methods scale
  - performance bounds
  - storage constraints
Real-time Analysis

• Want to have results now

• How?
  - Faster machines
  - Clusters
  - Progressive techniques
About Me

• Research Interests
  - Visualization
  - Computational Provenance
  - Geospatial Analysis

• Research Projects
  - VisTrails: www.vistrails.org
  - Dataflow Notebooks
  - Meta-versioning
  - Marine Traffic Data

• See my web page for more information
  - http://www.cis.umassd.edu/~dkoop/
About You

- Previous topics course (CIS 602)?
- Research Papers?
- Data Science?
- Python?
- Database Experience?
- Analytics Experience?
- Cloud Computing Experience?
- Anything you want to see covered?
About this course

• Course web page is authoritative:
  - Schedule, Readings, Assignments will be posted online
  - Check the web site before emailing me

• Topics course
  - A current research area the professor works in
  - A chance to be on the “cutting edge” of research

• Requires student participation
  - Reading responses
  - Project presentations
About this course

• Balance of techniques and research ideas
• Some background (Python) followed by topic areas and readings
• Assignments at the beginning of course, project at end
• Two tests
• Topic areas:
  - Exploratory Data Analysis and Visualization
  - Data Acquisition
  - Data Storage and Access
  - Cloud Computing and Scalable Computation
  - Applications and specific data considerations
Project

- Do scalable data analysis of a large dataset
  - Questions
  - Analysis
  - Visualizations
  - Cloud/Cluster Computing
- Another option: research-related topic
- Waypoints:
  - Proposal
  - Progress Report
  - Final Presentation
About this course

• Course Registration:
  - Make sure you have registered in COIN for the course
  - Email me if you are not registered but are interested in taking the course

• Review of course policies:
  - Plagiarism and academic honesty
  - If you have any concerns or questions, please email me as soon as possible

• If you are not sure if this course is a good fit, please email me or talk to me
Data

• What is this data?

<table>
<thead>
<tr>
<th></th>
<th>2ND STREET &amp; 8TH AVENUE</th>
<th>00228985</th>
<th>00008471</th>
<th>00000441</th>
<th>00001455</th>
<th>00000134</th>
<th>00033341</th>
<th>00071255</th>
</tr>
</thead>
<tbody>
<tr>
<td>R011</td>
<td>14TH STREET-UNION SQUARE</td>
<td>00224603</td>
<td>00111051</td>
<td>00000827</td>
<td>00003026</td>
<td>00000660</td>
<td>00089367</td>
<td>00199841</td>
</tr>
<tr>
<td>R046</td>
<td>2ND STREET &amp; GRAND CENTRAL</td>
<td>00207758</td>
<td>00007908</td>
<td>00000323</td>
<td>0001183</td>
<td>0003001</td>
<td>00040759</td>
<td>00096613</td>
</tr>
</tbody>
</table>

• Semantics: real-world meaning of the data
• Type: structural or mathematical interpretation
• Both often require metadata
  - Sometimes we can infer some of this information
  - Line between data and metadata isn’t always clear
## Data

<table>
<thead>
<tr>
<th>REMOTE</th>
<th>STATION</th>
<th>FF</th>
<th>SEN/DIS</th>
<th>7-D AFAS UNL</th>
<th>D AFAS/RFM</th>
<th>JOINT RR TKT</th>
<th>7-D UNL</th>
<th>30-D UNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R011 42ND STREET &amp; 8TH AVENUE</td>
<td>00228985</td>
<td>00008471</td>
<td>00000441</td>
<td>0001455</td>
<td>0000134</td>
<td>00033341</td>
<td>00071255</td>
</tr>
<tr>
<td>2</td>
<td>R170 14TH STREET-UNION SQUARE</td>
<td>00224603</td>
<td>0011051</td>
<td>00000827</td>
<td>0003026</td>
<td>00000660</td>
<td>00089367</td>
<td>00199841</td>
</tr>
<tr>
<td>3</td>
<td>R046 42ND STREET &amp; GRAND CENTRAL</td>
<td>00207758</td>
<td>00007908</td>
<td>00000323</td>
<td>0001183</td>
<td>00003001</td>
<td>00040759</td>
<td>00096613</td>
</tr>
<tr>
<td>4</td>
<td>R012 34TH STREET &amp; 8TH AVENUE</td>
<td>00188311</td>
<td>00006490</td>
<td>00000498</td>
<td>0001279</td>
<td>00003622</td>
<td>00035527</td>
<td>00067483</td>
</tr>
<tr>
<td>5</td>
<td>R293 34TH STREET – PENN STATION</td>
<td>00168768</td>
<td>00006155</td>
<td>00000523</td>
<td>0001065</td>
<td>00005031</td>
<td>00030645</td>
<td>00054376</td>
</tr>
<tr>
<td>6</td>
<td>R033 42ND STREET/TIMES SQUARE</td>
<td>00159382</td>
<td>00005945</td>
<td>00000378</td>
<td>0001205</td>
<td>00000690</td>
<td>00058931</td>
<td>00078644</td>
</tr>
<tr>
<td>7</td>
<td>R022 34TH STREET &amp; 6TH AVENUE</td>
<td>00156008</td>
<td>00006276</td>
<td>00000487</td>
<td>0001543</td>
<td>00000712</td>
<td>00058910</td>
<td>00110466</td>
</tr>
<tr>
<td>8</td>
<td>R084 59TH STREET/COLUMBUS CIRCLE</td>
<td>00155262</td>
<td>00009484</td>
<td>00000589</td>
<td>0002071</td>
<td>00000542</td>
<td>00053397</td>
<td>00113966</td>
</tr>
<tr>
<td>9</td>
<td>R020 47-50 STREETS/ROCKEFELLER</td>
<td>00143500</td>
<td>00006402</td>
<td>00000384</td>
<td>0001159</td>
<td>00000723</td>
<td>00037978</td>
<td>00090745</td>
</tr>
<tr>
<td>10</td>
<td>R179 86TH STREET-LEXINGTON AVE</td>
<td>00142169</td>
<td>00010367</td>
<td>00000470</td>
<td>0001839</td>
<td>00000271</td>
<td>00050328</td>
<td>00125250</td>
</tr>
<tr>
<td>11</td>
<td>R023 34TH STREET &amp; 6TH AVENUE</td>
<td>00134052</td>
<td>00005005</td>
<td>00000348</td>
<td>0001112</td>
<td>00000649</td>
<td>00031531</td>
<td>00075040</td>
</tr>
<tr>
<td>12</td>
<td>R029 PARK PLACE</td>
<td>00121614</td>
<td>00004311</td>
<td>00000287</td>
<td>0000931</td>
<td>00000792</td>
<td>00025404</td>
<td>00065362</td>
</tr>
<tr>
<td>13</td>
<td>R047 42ND STREET &amp; GRAND CENTRAL</td>
<td>00100742</td>
<td>00004273</td>
<td>00000185</td>
<td>0000704</td>
<td>00001241</td>
<td>00022808</td>
<td>00068216</td>
</tr>
<tr>
<td>14</td>
<td>R031 34TH STREET &amp; 7TH AVENUE</td>
<td>00095076</td>
<td>00003990</td>
<td>00000232</td>
<td>0000727</td>
<td>00001459</td>
<td>00024284</td>
<td>00038671</td>
</tr>
<tr>
<td>15</td>
<td>R017 LEXINGTON AVENUE</td>
<td>00094655</td>
<td>00004688</td>
<td>00000190</td>
<td>0000833</td>
<td>00000754</td>
<td>00020018</td>
<td>00055066</td>
</tr>
<tr>
<td>16</td>
<td>R175 8TH AVENUE-14TH STREET</td>
<td>00094313</td>
<td>00003907</td>
<td>00000286</td>
<td>0001144</td>
<td>00000256</td>
<td>00038272</td>
<td>00074661</td>
</tr>
<tr>
<td>17</td>
<td>R057 BARCLAYS CENTER</td>
<td>00093804</td>
<td>00004204</td>
<td>00000454</td>
<td>0001386</td>
<td>00001491</td>
<td>00039113</td>
<td>00068119</td>
</tr>
<tr>
<td>18</td>
<td>R138 WEST 4TH ST-WASHINGTON SQ</td>
<td>00093562</td>
<td>00004677</td>
<td>00000251</td>
<td>0000965</td>
<td>00000127</td>
<td>00031628</td>
<td>00074458</td>
</tr>
</tbody>
</table>
Data Types

- **Items**
  - An item is an individual discrete entity
  - e.g. row in a table, node in a network

- **Attributes**
  - An attribute is some specific property that can be measured, observed, or logged
  - a.k.a. variable, (data) dimension
## Items & Attributes

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>S</th>
<th>T</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order ID</td>
<td>Order Date</td>
<td>Order Priority</td>
<td>Product Container</td>
<td>Product Base Margin</td>
<td>Ship Date</td>
</tr>
<tr>
<td>3</td>
<td>10/14/06</td>
<td>5-Low</td>
<td>Large Box</td>
<td>0.8</td>
<td>10/21/06</td>
</tr>
<tr>
<td>6</td>
<td>2/21/08</td>
<td>4-Not Specified</td>
<td>Small Pack</td>
<td>0.55</td>
<td>2/22/08</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Small Pack</td>
<td>0.79</td>
<td>7/17/07</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Jumbo Box</td>
<td>0.69</td>
<td>7/17/07</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Medium Box</td>
<td>0.69</td>
<td>7/18/07</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Medium Box</td>
<td>0.69</td>
<td>7/18/07</td>
</tr>
<tr>
<td>35</td>
<td>10/23/07</td>
<td>4-Not Specified</td>
<td>Wrap Bag</td>
<td>0.52</td>
<td>10/24/07</td>
</tr>
<tr>
<td>35</td>
<td>10/23/07</td>
<td>4-Not Specified</td>
<td>Small Box</td>
<td>0.58</td>
<td>10/25/07</td>
</tr>
<tr>
<td>36</td>
<td>11/3/07</td>
<td>1-Urgent</td>
<td>Small Box</td>
<td>0.55</td>
<td>11/3/07</td>
</tr>
<tr>
<td>65</td>
<td>3/18/07</td>
<td>1-Urgent</td>
<td>Small Pack</td>
<td>0.49</td>
<td>3/19/07</td>
</tr>
<tr>
<td>66</td>
<td>1/20/05</td>
<td>5-Low</td>
<td>Wrap Bag</td>
<td>0.56</td>
<td>1/20/05</td>
</tr>
<tr>
<td>69</td>
<td>5</td>
<td>4-Not Specified</td>
<td>Small Pack</td>
<td>0.44</td>
<td>6/6/05</td>
</tr>
<tr>
<td>69</td>
<td>5</td>
<td>4-Not Specified</td>
<td>Wrap Bag</td>
<td>0.6</td>
<td>6/6/05</td>
</tr>
<tr>
<td>70</td>
<td>12/18/06</td>
<td>5-Low</td>
<td>Small Box</td>
<td>0.59</td>
<td>12/23/06</td>
</tr>
<tr>
<td>70</td>
<td>12/18/06</td>
<td>5-Low</td>
<td>Wrap Bag</td>
<td>0.82</td>
<td>12/23/06</td>
</tr>
<tr>
<td>96</td>
<td>4/17/05</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.55</td>
<td>4/19/05</td>
</tr>
<tr>
<td>97</td>
<td>1/29/06</td>
<td>3-Medium</td>
<td>Small Box</td>
<td>0.38</td>
<td>1/30/06</td>
</tr>
<tr>
<td>129</td>
<td>11/19/08</td>
<td>5-Low</td>
<td>Small Box</td>
<td>0.37</td>
<td>11/28/08</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.37</td>
<td>5/9/08</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Medium Box</td>
<td>0.38</td>
<td>5/10/08</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.6</td>
<td>5/11/08</td>
</tr>
<tr>
<td>132</td>
<td>6/11/06</td>
<td>3-Medium</td>
<td>Medium Box</td>
<td>0.6</td>
<td>6/12/06</td>
</tr>
<tr>
<td>132</td>
<td>6/11/06</td>
<td>3-Medium</td>
<td>Jumbo Box</td>
<td>0.69</td>
<td>6/14/06</td>
</tr>
<tr>
<td>134</td>
<td>5/1/08</td>
<td>4-Not Specified</td>
<td>Large Box</td>
<td>0.82</td>
<td>5/3/08</td>
</tr>
<tr>
<td>135</td>
<td>10/21/07</td>
<td>4-Not Specified</td>
<td>Small Pack</td>
<td>0.64</td>
<td>10/23/07</td>
</tr>
<tr>
<td>166</td>
<td>9/12/07</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.55</td>
<td>9/14/07</td>
</tr>
<tr>
<td>193</td>
<td>8/8/06</td>
<td>1-Urgent</td>
<td>Medium Box</td>
<td>0.57</td>
<td>8/10/06</td>
</tr>
<tr>
<td>194</td>
<td>4/5/08</td>
<td>3-Medium</td>
<td>Wrap Bag</td>
<td>0.42</td>
<td>4/7/08</td>
</tr>
</tbody>
</table>
Data Types

• Nodes
  - Synonym for item but in the context of networks (graphs)

• Links
  - A link is a relation between two items
  - e.g. social network friends, computer network links
Items & Links

[Bostock, 2011]
Dataset Types

- **Tables**
  - Attributes (columns)
  - Items (rows)
  - Cell containing value

- **Networks**
  - Link
  - Node (item)

- **Fields (Continuous)**
  - Grid of positions
  - Attributes (columns)
  - Value in cell

- **Geometry (Spatial)**
  - Position

- **Multidimensional Table**
  - Key 1
  - Key 2
  - Attributes
  - Value in cell

- **Trees**

[Munzner (ill. Maguire), 2014]
Attribute Types

- **Attribute Types**
  - Categorical
  - Ordered
  - Ordinal
  - Quantitative

- **Ordering Direction**
  - Sequential
  - Diverging
  - Cyclic

[Munzner (ill. Maguire), 2014]
Categorial, Ordinal, and Quantitative

<table>
<thead>
<tr>
<th>Date</th>
<th>Description</th>
<th>Box Type</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2/21/08 4-Not Spec</td>
<td>Small Pack</td>
<td>0.55</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07 2-High</td>
<td>Small Pack</td>
<td>0.79</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07 2-High</td>
<td>Jumbo Box</td>
<td>0.72</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07 2-High</td>
<td>Medium Box</td>
<td>0.6</td>
</tr>
<tr>
<td>32</td>
<td>7/16/07 2-High</td>
<td>Medium Box</td>
<td>0.65</td>
</tr>
<tr>
<td>35</td>
<td>10/23/07 4-Not Spec</td>
<td>Wrap Bag</td>
<td>0.52</td>
</tr>
<tr>
<td>35</td>
<td>10/23/07 4-Not Spec</td>
<td>Small Box</td>
<td>0.58</td>
</tr>
<tr>
<td>36</td>
<td>11/3/07 1-Urgent</td>
<td>Small Box</td>
<td>0.55</td>
</tr>
<tr>
<td>65</td>
<td>3/18/07 1-Urgent</td>
<td>Small Pack</td>
<td>0.49</td>
</tr>
<tr>
<td>66</td>
<td>1/20/05 5-Low</td>
<td>Wrap Bag</td>
<td>0.56</td>
</tr>
<tr>
<td>69</td>
<td>6/4/05 4-Not Spec</td>
<td>Small Pack</td>
<td>0.44</td>
</tr>
<tr>
<td>69</td>
<td>6/4/05 4-Not Spec</td>
<td>Small Pack</td>
<td>0.6</td>
</tr>
<tr>
<td>70</td>
<td>12/18/06 5-Low</td>
<td></td>
<td>0.59</td>
</tr>
<tr>
<td>70</td>
<td>12/18/06 5-Low</td>
<td></td>
<td>0.82</td>
</tr>
<tr>
<td>96</td>
<td>4/17/05 2-High</td>
<td></td>
<td>0.55</td>
</tr>
<tr>
<td>97</td>
<td>1/29/06 3-Medium</td>
<td></td>
<td>0.38</td>
</tr>
<tr>
<td>129</td>
<td>11/19/08 5-Low</td>
<td></td>
<td>0.37</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08 2-High</td>
<td>Small Box</td>
<td>0.37</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08 2-High</td>
<td>Medium Box</td>
<td>0.38</td>
</tr>
<tr>
<td>130</td>
<td>5/8/08 2-High</td>
<td>Small Box</td>
<td>0.6</td>
</tr>
<tr>
<td>132</td>
<td>6/11/06 3-Medium</td>
<td>Medium Box</td>
<td>0.6</td>
</tr>
<tr>
<td>132</td>
<td>6/11/06 3-Medium</td>
<td>Jumbo Box</td>
<td>0.69</td>
</tr>
<tr>
<td>134</td>
<td>5/1/08 4-Not Spec</td>
<td>Large Box</td>
<td>0.82</td>
</tr>
<tr>
<td>135</td>
<td>10/21/07 4-Not Spec</td>
<td>Small Pack</td>
<td>0.64</td>
</tr>
<tr>
<td>166</td>
<td>9/12/07 2-High</td>
<td>Small Box</td>
<td>0.55</td>
</tr>
</tbody>
</table>
## Categorial, Ordinal, and Quantitative

<table>
<thead>
<tr>
<th></th>
<th>10/14/08</th>
<th>4-Low</th>
<th>Large Box</th>
<th></th>
<th></th>
<th>10/21/08</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2/21/08</td>
<td>4-Not Spec.</td>
<td>Small Pack</td>
<td>0.55</td>
<td>2/22/08</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Small Pack</td>
<td>0.79</td>
<td>7/17/07</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Jumbo Box</td>
<td>0.72</td>
<td>7/17/07</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Medium Box</td>
<td>0.6</td>
<td>7/18/07</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>7/16/07</td>
<td>2-High</td>
<td>Medium Box</td>
<td>0.65</td>
<td>7/18/07</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>10/23/07</td>
<td>4-Not Spec.</td>
<td>Wrap Bag</td>
<td>0.52</td>
<td>10/24/07</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>10/23/07</td>
<td>4-Not Spec.</td>
<td>Small Box</td>
<td>0.58</td>
<td>10/25/07</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>11/3/07</td>
<td>1-Urgent</td>
<td>Small Box</td>
<td>0.55</td>
<td>11/3/07</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>3/18/07</td>
<td>1-Urgent</td>
<td>Small Pack</td>
<td>0.49</td>
<td>3/19/07</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>1/20/05</td>
<td>5-Low</td>
<td>Wrap Bag</td>
<td>0.56</td>
<td>1/20/05</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>6/4/05</td>
<td>4-Not Spec.</td>
<td>Small Pack</td>
<td>0.44</td>
<td>6/6/05</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>6/4/05</td>
<td>4-Not Spec.</td>
<td></td>
<td>0.6</td>
<td>6/6/05</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>12/18/06</td>
<td>5-Low</td>
<td></td>
<td>0.59</td>
<td>12/23/06</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>12/18/06</td>
<td>5-Low</td>
<td></td>
<td>0.82</td>
<td>12/23/06</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>4/17/05</td>
<td>2-High</td>
<td></td>
<td>0.55</td>
<td>4/19/05</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>1/29/06</td>
<td>3-Medium</td>
<td></td>
<td>0.38</td>
<td>1/30/06</td>
<td></td>
</tr>
<tr>
<td>129</td>
<td>11/19/08</td>
<td>5-Low</td>
<td></td>
<td>0.37</td>
<td>11/28/08</td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.37</td>
<td>5/9/08</td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Medium Box</td>
<td>0.37</td>
<td>5/10/08</td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>5/8/08</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.6</td>
<td>5/11/08</td>
<td></td>
</tr>
<tr>
<td>132</td>
<td>6/11/06</td>
<td>3-Medium</td>
<td>Medium Box</td>
<td>0.6</td>
<td>6/12/06</td>
<td></td>
</tr>
<tr>
<td>132</td>
<td>6/11/06</td>
<td>3-Medium</td>
<td>Jumbo Box</td>
<td>0.69</td>
<td>6/14/06</td>
<td></td>
</tr>
<tr>
<td>134</td>
<td>5/1/08</td>
<td>4-Not Spec.</td>
<td>Large Box</td>
<td>0.82</td>
<td>5/3/08</td>
<td></td>
</tr>
<tr>
<td>135</td>
<td>10/21/07</td>
<td>4-Not Spec.</td>
<td>Small Pack</td>
<td>0.64</td>
<td>10/23/07</td>
<td></td>
</tr>
<tr>
<td>166</td>
<td>9/12/07</td>
<td>2-High</td>
<td>Small Box</td>
<td>0.55</td>
<td>9/14/07</td>
<td></td>
</tr>
</tbody>
</table>

- **Quantitative**
- **Ordinal**
- **Categorical**
Semantics

- The type of data does not tell us what the data means or how it should be interpreted
- Tables have keys/values, fields have independent/dependent vars

[Munzner (ill. Maguire), 2014]
Analysis

**What?**

- Analyze
  - Consume
    - Discover
    - Present
    - Enjoy
  - Produce
    - Annotate
    - Record
    - Derive

**Why?**

- All Data
  - Trends
  - Outliers
  - Features

- Attributes
  - One
    - Distribution
    - Extremes
  - Many
    - Dependency
    - Correlation
    - Similarity

- Network Data
  - Topology
    - Paths
  - Spatial Data
    - Shape

- Search

<table>
<thead>
<tr>
<th>Target known</th>
<th>Target unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location known</td>
<td>· · ·</td>
</tr>
<tr>
<td>Location unknown</td>
<td>·</td>
</tr>
</tbody>
</table>

- Query
  - Identify
  - Compare
  - Summarize

[Munzner (ill. Maguire), 2014]
Analysis: Consume & Produce

• Consume
  - Exploration
  - Explanation
  - Enjoyment

• Produce
  - Annotation
  - Record
  - Derivation
    • Leads to new directions/ideas

[Munzner (ill. Maguire), 2014]
Analysis: Search and Query

• Search based on what a user knows
  - Target
  - Location

• Query depends on what data matters
  - One
  - Some (Often Two)
  - All

[Table]

<table>
<thead>
<tr>
<th></th>
<th>Target known</th>
<th>Target unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location known</td>
<td>Lookup</td>
<td>Browse</td>
</tr>
<tr>
<td>Location unknown</td>
<td>Locate</td>
<td>Explore</td>
</tr>
</tbody>
</table>

[Diagram]

[Source: Munzner (ill. Maguire), 2014]
Targets

ALL DATA
- Trends
- Outliers
- Features

ATTRIBUTES
- One
  - Distribution
  - Extremes
- Many
  - Dependency
  - Correlation
  - Similarity

NETWORK DATA
- Topology
- Paths

SPATIAL DATA
- Shape

[Munzner (ill. Maguire), 2014]
More Reading

• Listed on course schedule:
  - Challenges and Opportunities with Big Data, D. Agrawal et al.
  - Toward Scalable Systems for Big Data Analysis: A Technology Tutorial, H. Hu et al.
  - Big Data computing and clouds: Trends and future directions, M. D. Assuncao et al.
Next Class

• Introduction to/review of Python
• Download anaconda distribution:
  - https://www.anaconda.com/download/
  - I am planning to use Python 3 (3.6)