Guiding Principles

Analysis is often **iterative** and **exploratory**
- Hypothesis formation vs. confirmation
- Ad hoc queries – difficult to know workload

Design interfaces to enable **tightly iterative**
- Visualization: rapid perception of patterns
- Interaction: specification of data & views

Support analysis “at the speed of thought”

Effects of Latency

**Milliseconds Matter**
- < 100ms: perception of animation, causation
- Wait times > 1s may interrupt flow of thought

To reduce (perceived) latency, drop details

**Example: Latency in Google search results**
- Latency (+300ms) reduces searches ~0.5%
- Reduction persists after performance resumes

**Goal**: support analysis at **interactive rates** to enable fluid **conversations with data.**
Review of Analysis Infrastructure

Analysis Infrastructure

- Web Activity & Transactions
- MapReduce Cluster
- External Data
- OLTP
- Datanbase Storage
- HDFS
- MapReduce Code (Java, Python, etc)
- Client
  - Reports
  - DB Terminal
  - Excel
  - Tableau
  - R, Python, Perl

Data Modeling for Analysis

- Production DW
- Staging
- Sandbox
- Reporting

Greenplum: SQL & MapReduce

Unified execution of SQL, MapReduce on a common parallel execution engine

- Query Planner and Optimizer (SQL)
- Parallel Dataflow Engine
- Transaction Manager & Log Files
- External Storage
- Database Storage
- MapReduce Code (Perl, Python, etc)

Analyze structured or unstructured data, inside or outside the database
Scale out parallelism on commodity hardware
Relational Data Organizations

<table>
<thead>
<tr>
<th>Transactions</th>
<th>vs.</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row-oriented</td>
<td></td>
<td>Column-oriented</td>
</tr>
</tbody>
</table>

Speed-up Analysis
- Reduce data transfer
- Improved locality
- Better data compression
- Query processing on compressed data

Analysis Infrastructure
- Web Activity & Transactions
- Data Warehouse
- OLTP
- External Data
- MapReduce Cluster
- HDFS Storage
- Client (Reports, DB Terminal, Excel, Tableau, R, Python, Perl)
Other Thoughts

Much current attention is focused on scalable infrastructure, but don’t forget the client!
Multi-core and GPU computing for analysis
Monitoring, review on mobile devices?
Increasing variety of cloud computing services, solid state storage – networking as limiter?

Visual Data Analysis

1 Data diagnostics
Visual Data Analysis

1. Data diagnostics
2. Exploratory data analysis

<table>
<thead>
<tr>
<th>Set A</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.04</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8.11</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td></td>
</tr>
</tbody>
</table>

Summary Statistics
- $\bar{x}_A = 9.0$, $\sigma_x = 3.317$
- $\bar{y}_A = 7.5$, $\sigma_y = 2.03$

<table>
<thead>
<tr>
<th>Set B</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9.14</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8.14</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>8.74</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8.77</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>9.26</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>8.1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6.13</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>9.11</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7.26</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4.74</td>
<td></td>
</tr>
</tbody>
</table>

Linear Regression
- $Y = 3 + 0.5X$
- $R^2 = 0.67$

<table>
<thead>
<tr>
<th>Set C</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7.46</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6.77</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>12.74</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>7.11</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>7.81</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>8.84</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6.08</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5.39</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>8.15</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6.42</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5.73</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set D</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>6.58</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5.76</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7.11</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8.84</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8.47</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7.04</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5.25</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>12.5</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5.56</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7.91</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6.89</td>
<td></td>
</tr>
</tbody>
</table>

Anscombe 1973
Visual Data Analysis

1. Data diagnostics
2. Exploratory data analysis
3. Assessing analytic results

While it is often most helpful to “plot the data,” this is rarely enough. We need also to “plot the results of analysis” as a routine matter. There is often more analysis than there was data.

Graphical Perception & Automated Design
Visual encoding variables

Position (x 2)
Size
Value
Texture
Color
Orientation
Shape

Combinatorics of Encodings

Challenge:
Pick the best encoding from the exponential number of possibilities \((n+1)^8\)

Principle of Consistency:
The properties of the image (visual variables) should match the properties of the data.

Principle of Importance Ordering:
Encode the most important information in the most effective way.

Design Criteria (Mackinlay)

Expressiveness
A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.
Cannot express the facts

A one-to-many (1 → N) relation cannot be expressed in a single horizontal dot plot because multiple tuples are mapped to the same position.

Expresses facts not in the data

A length is interpreted as a quantitative value; \( \therefore \) Length of bar says something untrue about N data.

Design Criteria (Mackinlay)

Expressiveness

A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

A visualization is more effective than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Which best encodes quantities?

Position
Length
Area
Volume
Value (Brightness)
Color Hue
Orientation (Angle)
Shape
Compare area of circles

Compare length of bars

Rankings Used by Mackinlay

Conjectured effectiveness of the encoding
**Mackinlay’s Design Algorithm**

User formally specifies data model and type
- Additional input: ordered list of data variables to show

APT searches over design space
- Tests expressiveness of each visual encoding
- Generates image for encodings that pass test
- Tests perceptual effectiveness of resulting image

Outputs the “most effective” visualization

---

**Project Presentations**

Present your project plans to the class.
Keep it short! (≤ 3 minutes)
Very briefly introduce the problem domain.
Spend most of the time presenting your initial design ideas – flowcharts, sketches, prototypes
Your final slide should have 3 questions for the class – no need to read them out loud.
Send slides to cs448g by 8am Wed 5/4.

---

**Final Project Presentations**

---

**Discussants**

Amir Ghazvinian
Eli Marschner
Letitia Lew