DESIGNING WITH DATA

Jeffrey Heer  @jeffrey_heer
Trifacta Inc. + U. Washington
Data Sets I Have *Known* and *Loved*
Enron 'Mastermind' Pledges Guilty

SAN FRANCISCO, Oct. 17, 2002

(AP) A former top energy trader, considered the mastermind of Enron Corp.'s scheme to drive up California's energy prices, pleaded guilty Thursday to a federal conspiracy charge.

Timothy Belden, the former head of trading at Enron's Portland, Ore., office, admitted to one count of conspiracy to commit wire fraud and promised to cooperate with state and federal prosecutors as well as any non-criminal effort to investigate the energy industry.

"I did it because I was trying to maximize profit for Enron," Belden told U.S. District Judge Martin Jenkins.
Having real data as part of your project is as important as having real users look at your project. I have seen people ... create mockups with fake data, show them to real users, get all sorts of feedback, and in the end it's not worth anything.

It's almost as if the data is one of the stakeholders in the project and you need its input from the beginning.

Martin Wattenberg  [ACM Queue ’09]

Big shift after 1950

The biggest movement of women into the labor force happened between 1950 and 1960. From this view, it appears that clerical and secretarial jobs underwent significant growth.

Michael Rogers on Thu Jul 20, 2006 1:05 PM

If you look at household worker, you'll see the opposite trend.

Julia Hernandez on Fri Jul 21, 2006 10:52 AM

Also has to do with a shift from blue collar to white collar work.

Fred Klein on Wed Aug 2, 2006 10:29 AM

...and then women move behind the bar.

by Adam Kleinberg on Tue Aug 1, 2006 8:02 AM

reply

Disappearing professors???

Note how a bunch of different kinds of professors "disappear" in the year 2000, whereas the "professor" category take off. Is this just a change in the way the census aggregates data?

by Julia Hernandez on Fri Jul 21, 2006 11:07 AM
reply
SENSE.US DESIGN PROCESS
SENSE.US DESIGN PROCESS

Collect data
SENSE.US DESIGN PROCESS

Collect data
Collect more data
SENSE.US DESIGN PROCESS

Collect data
Collect more data
Learn data is piecemeal & incompatible
SENSE.US DESIGN PROCESS

Collect data
Collect more data
Learn data is piecemeal & incompatible
Get frustrated!
SENSE.US DESIGN PROCESS

Collect data
Collect more data
Learn data is piecemeal & incompatible
Get frustrated!
Seek out subject matter experts
SENSE.US DESIGN PROCESS

Collect data
Collect more data
Learn data is piecemeal & incompatible
Get frustrated!
Seek out subject matter experts
Find better data source
SENSE.US DESIGN PROCESS (2)
Collect data
SENSE.US DESIGN PROCESS (2)

Collect data
Find data is too big for spreadsheets
Sense.Us Design Process (2)

Collect data
Find data is too big for spreadsheets
Peek at data through text editors
Collect data
Find data is too big for spreadsheets
Peek at data through text editors
Write code to parse data
Collect data
Find data is too big for spreadsheets
Peek at data through text editors
Write code to parse data
Run code; find errors; fix code; repeat
SENSE.US DESIGN PROCESS (2)

Collect data
Find data is too big for spreadsheets
Peek at data through text editors
Write code to parse data
Run code; find errors; fix code; repeat
Get frustrated!
SENSE.US DESIGN PROCESS (3)
SENSE.US DESIGN PROCESS (3)
Load data into relational database
SENSE.US DESIGN PROCESS (3)

Load data into relational database
Discover errors on import
Load data into relational database
Discover errors on import
More frustration!
SENSE.US DESIGN PROCESS (3)
Load data into relational database
Discover errors on import
More frustration!
Fix import script; load into database
SENSE.US DESIGN PROCESS (3)
Load data into relational database
Discover errors on import
More frustration!
Fix import script; load into database
Write queries against data; learn a little
SENSE.US DESIGN PROCESS (3)

Load data into relational database
Discover errors on import
More frustration!
Fix import script; load into database
Write queries against data; learn a little
Point Tableau at database to explore
SENSE.US DESIGN PROCESS (4)
SENSE.US DESIGN PROCESS (4)

Explore and profile data
SENSE.US DESIGN PROCESS (4)

Explore and profile data; *FUN!!*
SENSE.US DESIGN PROCESS (4)

Explore and profile data; \textit{FUN!!}

Prototype visualizations (various tools)
SENSE.US DESIGN PROCESS (4)

Explore and profile data; **FUN!!**
Prototype visualizations (various tools)
Get crits & feedback from team
Sense.us Design Process (4)

Explore and profile data; FUN!!
Prototype visualizations (various tools)
Get crits & feedback from team
Write custom visualization code
SENSE.US DESIGN PROCESS (4)

Explore and profile data; FUN!!
Prototype visualizations (various tools)
Get crits & feedback from team
Write custom visualization code
More team critique; user test
SENSE.US DESIGN PROCESS (4)

Explore and profile data; FUN!!
Prototype visualizations (various tools)
Get crits & feedback from team
Write custom visualization code
More team critique; user test
Deploy!
SENSE.US DESIGN PROCESS (4)

Explore and profile data; FUN!!
Prototype visualizations (various tools)
Get crits & feedback from team
Write custom visualization code
More team critique; user test
Deploy!
DATA
DATA

...is a stakeholder.
DATA

...is a stakeholder.

...comes in all shapes & sizes.
DATA

...is a stakeholder.
...comes in all shapes & sizes.

How to design tools for data unseen?
Trifacta v1:
Enable Interactive Data Transformation

Trifacta v2:
Accelerate Data Discovery & Profiling
**Trifacta v1:**
Enable Interactive Data Transformation

**Trifacta v2:**
Accelerate Data Discovery & Profiling
I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any “analysis” at all.

Anonymous Data Scientist

*from our interview study [IEEE VAST ’12]*
Demo: Predictive Interaction
Traditional Burden of Specification

1. User authors a draft transformation script

2. User tests the script on a small amount of data

3. User inspects output data to assess effects
Visualizations and Interactions

Data Transformations Code

1. User highlights features of a data visualization

2. ML methods predict distribution over DSL statements

3. Data previews allow user to choose, adjust and confirm

Trifacta: Predictive Interaction™
DESIGN DESIDERATA
DESIGN DESIDERATA

Co-evolution of language & interface
Co-evolution of language & interface
Embrace ambiguity
DESIGN DESIDERATA

Co-evolution of language & interface

Embrace ambiguity

Visualize transformation effects
DESIGN DESIDERATA

Co-evolution of language & interface
Embrace ambiguity
Visualize transformation effects
Interpretable transformations
SUGGESTED TRANSFORMS

- extract col: Screen_Detail on: /(?<=adtam_source\=)[^\&]*(?=\&)/
- extract col: Screen_Detail on: /(?<=\-)[^\&]*(?=\&)/ limit: 2
- extract col: Screen_Detail on: /(?<=\=)[a-z]+/ limit: 2
- extract col: Screen_Detail on: /[a-z]+/ limit: 4
- countpattern col: Screen_Detail on: /[a-z]+/
Designing for Interpretability

**SUGGESTED TRANSFORMS**

extract col: Screen_Detail on: /(?<=adtm_source=)[^\&]*(?=\&)/
extract col: Screen_Detail on: /(?<==)[^\&]*(?=\&)/ limit: 2
extract col: Screen_Detail on: /(?<=a-z)+/ limit: 2
extract col: Screen_Detail on: /[a-z]+/ limit: 4
countpattern col: Screen_Detail on: /[a-z]+/

**SUGGESTED TRANSFORMS**

extract col: Screen_Detail after: `adtm_source=` before: `&`
extract col: Screen_Detail limit: 2 after: `=` before: `&`
extract col: Screen_Detail on: `\{lower}+` limit: 2
extract col: Screen_Detail on: `\{lower}+` limit: 4
countpattern col: Screen_Detail on: `\{lower}+`
Co-evolution of language & interface
Embrace ambiguity
Visualize transformation effects
Interpretable transformations
…and lots of testing (data + users)!

DESIGN DESIDERATA
Trifacta v1:
Enable Interactive Data Transformation

Trifacta v2:
Accelerate Data Discovery & Profiling
Trifacta v1:
Enable Interactive Data Transformation

Trifacta v2:
Accelerate Data Discovery & Profiling
What’s in your data?
What's in your data?
The first sign that a visualization is good is that it shows you a problem in your data. Every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.

Martin Wattenberg  [ACM Queue '09]
Demo: Visual Data Profiling
DESIGN DESIDERATA
DESIGN DESIDERATA

Automatic visualization presentation
<table>
<thead>
<tr>
<th>Field</th>
<th>Summary</th>
<th>Frequent Values</th>
<th>Invalid Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans_Date</td>
<td>Unique</td>
<td>469 (19.6%)</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Valid</td>
<td>2,391 (100.0%)</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Mismatched</td>
<td>0 (0.0%)</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Missing</td>
<td>0 (0.0%)</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Extreme</td>
<td>0 (0.0%)</td>
<td>21</td>
</tr>
</tbody>
</table>

**Value Histogram**

**Month**

**Day of Week**

**Day of Month**
Exposure of all variables
Show relevant perspectives
DESIGN DESIDERATA

Automatic visualization presentation
DESIGN DESIDERATA

Automatic visualization presentation
Perceptually effective encodings
Compare area of circles
Compare length of bars
Perception of Quantitative Values

Most accurate
- Position (common) scale
- Position (non-aligned) scale
- Length
- Slope
- Angle
- Area
- Volume

Least accurate
- Color hue-saturation-density
Area encoding >> Color encoding
DESIGN DESIDERATA

Automatic visualization presentation
Perceptually effective encodings
DESIGN DESIDERATA

Automatic visualization presentation
Perceptually effective encodings
Visualize all data, big or small
### Summary of Trans_Amount

<table>
<thead>
<tr>
<th>SUMMARY</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique</td>
<td>212</td>
<td>8.9%</td>
</tr>
<tr>
<td>Valid</td>
<td>2,391</td>
<td>100.0%</td>
</tr>
<tr>
<td>Mismatched</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Extreme</td>
<td>212</td>
<td>8.9%</td>
</tr>
</tbody>
</table>

### Frequent Values

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>524</td>
</tr>
<tr>
<td>500</td>
<td>452</td>
</tr>
<tr>
<td>1,000</td>
<td>316</td>
</tr>
<tr>
<td>200</td>
<td>130</td>
</tr>
<tr>
<td>300</td>
<td>106</td>
</tr>
</tbody>
</table>

### Value Histogram

Histogram showing frequency distribution of Trans_Amount values. The histogram is skewed towards lower values with a peak around 250 and 500.

### Frequent Values

Bar chart showing individual values and their frequency. The highest frequency is seen around 250 and 500.
# Trans_Amount

<table>
<thead>
<tr>
<th>Summary</th>
<th>Value</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique</td>
<td>4,657</td>
<td>0.3%</td>
</tr>
<tr>
<td>Valid</td>
<td>1,539,514</td>
<td>100.0%</td>
</tr>
<tr>
<td>Mismatched</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Extreme</td>
<td>148,056</td>
<td>9.6%</td>
</tr>
</tbody>
</table>

**Frequent values**

<table>
<thead>
<tr>
<th>Value</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>339,126</td>
</tr>
<tr>
<td>500</td>
<td>278,466</td>
</tr>
<tr>
<td>1,000</td>
<td>204,290</td>
</tr>
<tr>
<td>200</td>
<td>100,182</td>
</tr>
<tr>
<td>2,600</td>
<td>82,535</td>
</tr>
</tbody>
</table>

**Invalid values**

None

---

**Value histogram**

![Value histogram](image)

**Frequent values**

![Frequent values](image)

Summarize full data set
Immediate brushing & linking
Immediate brushing & linking
Immediate brushing & linking
Immediate brushing & linking
DESIGN DESIDERATA

Automatic visualization presentation
Perceptually effective encodings
Visualize all data, big or small
DESIGN DESIDERATA

Automatic visualization presentation
Perceptually effective encodings
Visualize all data, big or small
Springboard correction & refinement
DESIGN DESIDERATA

Automatic visualization presentation
Perceptually effective encodings
Visualize all data, big or small
Springboard correction & refinement
Trifacta v1:
Enable Interactive Data Transformation

Trifacta v2:
Accelerate Data Discovery & Profiling
DESIGNING WITH DATA

Jeffrey Heer  @jeffrey_heer
http://www.trifacta.com
Visit us at our booth or at http://www.trifacta.com