Graph Viewer

Roll-up by:
- All

Visualization:
- Matrix

Sort by:
- Linkage

Edge centrality filters:
d3.js Data-Driven Documents

with Mike Bostock & Vadim Ogievetsky
Reported crime in Alabama

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>4525375</td>
<td>4029.3</td>
<td>987</td>
<td>2732.4</td>
<td>309.9</td>
</tr>
<tr>
<td>2005</td>
<td>4548327</td>
<td>3900</td>
<td>955.8</td>
<td>2656</td>
<td>289</td>
</tr>
<tr>
<td>2006</td>
<td>4599030</td>
<td>3937</td>
<td>968.9</td>
<td>2645.1</td>
<td>322.9</td>
</tr>
<tr>
<td>2007</td>
<td>4627851</td>
<td>3974.9</td>
<td>980.2</td>
<td>2687</td>
<td>307.7</td>
</tr>
<tr>
<td>2008</td>
<td>4661900</td>
<td>4081.9</td>
<td>1080.7</td>
<td>2712.6</td>
<td>288.6</td>
</tr>
</tbody>
</table>

Reported crime in Alaska

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>657755</td>
<td>3370.9</td>
<td>573.6</td>
<td>2456.7</td>
<td>340.6</td>
</tr>
<tr>
<td>2005</td>
<td>663253</td>
<td>3615</td>
<td>622.8</td>
<td>2601</td>
<td>391</td>
</tr>
<tr>
<td>2006</td>
<td>670053</td>
<td>3582</td>
<td>615.2</td>
<td>2588.5</td>
<td>378.3</td>
</tr>
<tr>
<td>2007</td>
<td>683478</td>
<td>3373.9</td>
<td>538.9</td>
<td>2480</td>
<td>355.1</td>
</tr>
<tr>
<td>2008</td>
<td>686293</td>
<td>2928.3</td>
<td>470.9</td>
<td>2219.9</td>
<td>237.5</td>
</tr>
</tbody>
</table>

Reported crime in Arizona

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>5739879</td>
<td>5073.3</td>
<td>991</td>
<td>3118.7</td>
<td>963.5</td>
</tr>
<tr>
<td>2005</td>
<td>5953007</td>
<td>4827</td>
<td>946.2</td>
<td>2958</td>
<td>922</td>
</tr>
<tr>
<td>2006</td>
<td>6166318</td>
<td>4741.6</td>
<td>953</td>
<td>2874.1</td>
<td>914.4</td>
</tr>
<tr>
<td>2007</td>
<td>6338755</td>
<td>4502.6</td>
<td>935.4</td>
<td>2780.5</td>
<td>786.7</td>
</tr>
<tr>
<td>2008</td>
<td>6500180</td>
<td>4087.3</td>
<td>894.2</td>
<td>2605.3</td>
<td>587.8</td>
</tr>
</tbody>
</table>

Reported crime in Arkansas

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>2750000</td>
<td>4033.1</td>
<td>1096.4</td>
<td>2699.7</td>
<td>237</td>
</tr>
<tr>
<td>2005</td>
<td>2775708</td>
<td>4068</td>
<td>1085.1</td>
<td>2720</td>
<td>262</td>
</tr>
<tr>
<td>2006</td>
<td>2810872</td>
<td>4021.6</td>
<td>1154.4</td>
<td>2596.7</td>
<td>270.4</td>
</tr>
<tr>
<td>2007</td>
<td>2834797</td>
<td>3945.5</td>
<td>1124.4</td>
<td>2574.6</td>
<td>246.5</td>
</tr>
<tr>
<td>2008</td>
<td>2855390</td>
<td>3843.7</td>
<td>1182.7</td>
<td>2433.4</td>
<td>227.6</td>
</tr>
</tbody>
</table>

Reported crime in California

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>35842038</td>
<td>3423.9</td>
<td>686.1</td>
<td>2033.1</td>
<td>704.8</td>
</tr>
<tr>
<td>2005</td>
<td>36154147</td>
<td>3321</td>
<td>692.9</td>
<td>1915</td>
<td>712</td>
</tr>
<tr>
<td>2006</td>
<td>36457549</td>
<td>3175.2</td>
<td>676.9</td>
<td>1831.5</td>
<td>666.8</td>
</tr>
<tr>
<td>2007</td>
<td>36553215</td>
<td>3032.6</td>
<td>648.4</td>
<td>1784.1</td>
<td>600.2</td>
</tr>
<tr>
<td>2008</td>
<td>36756666</td>
<td>2940.3</td>
<td>646.8</td>
<td>1769.8</td>
<td>523.8</td>
</tr>
</tbody>
</table>

Reported crime in Colorado

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>4601821</td>
<td>3918.5</td>
<td>717.3</td>
<td>2679.5</td>
<td>521.6</td>
</tr>
<tr>
<td>2005</td>
<td>4617214</td>
<td>3933.5</td>
<td>723.9</td>
<td>2698.2</td>
<td>527.2</td>
</tr>
<tr>
<td>2006</td>
<td>4634705</td>
<td>3926.1</td>
<td>717.3</td>
<td>2714.3</td>
<td>528.3</td>
</tr>
<tr>
<td>2007</td>
<td>4653225</td>
<td>3919.7</td>
<td>717.3</td>
<td>2729.9</td>
<td>529.4</td>
</tr>
<tr>
<td>2008</td>
<td>4672867</td>
<td>3913.3</td>
<td>717.3</td>
<td>2745.5</td>
<td>530.5</td>
</tr>
</tbody>
</table>
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, 2012
### DataWrangler

<table>
<thead>
<tr>
<th>#</th>
<th>Year</th>
<th>Property_crime_rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Reported crime in Alabama</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2004</td>
<td>4029.3</td>
</tr>
<tr>
<td>3</td>
<td>2005</td>
<td>3900</td>
</tr>
<tr>
<td>4</td>
<td>2006</td>
<td>3937</td>
</tr>
<tr>
<td>5</td>
<td>2007</td>
<td>3974.9</td>
</tr>
<tr>
<td>6</td>
<td>2008</td>
<td>4081.9</td>
</tr>
<tr>
<td>7</td>
<td>Reported crime in Alaska</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2004</td>
<td>3370.9</td>
</tr>
<tr>
<td>9</td>
<td>2005</td>
<td>3615</td>
</tr>
<tr>
<td>10</td>
<td>2006</td>
<td>3582</td>
</tr>
<tr>
<td>11</td>
<td>2007</td>
<td>3373.9</td>
</tr>
</tbody>
</table>

with **Sean Kandel**, Philip Guo, Andreas Paepcke & Joe Hellerstein
Wrangler in 2 Parts...

1. Declarative data transformation language
   - **Tuple mapping** – split, merge, extract, delete
   - **Reshaping** – fold, unfold (cross-tabulation)
   - **Lookups & joins** – e.g., FIPS code to US state

   Sorting, aggregation, etc.

Informed by prior work in databases:
Potter’s Wheel, SchemaSQL, AJAX
Wrangler in 2 Parts...

1. Declarative data transformation language

+ 

2. Mixed-initiative interface for data transforms

User: **Selects** data elements of interest

System: **Suggests** applicable transforms via search over the space of viable transforms

Enable rapid **preview and refinement**
Transform Suggestion

Interaction

\downarrow

Infer Operands

\downarrow

Generate Transforms

\downarrow

Rank Transforms

\downarrow

Present Top-N
Transform Suggestion

**Interaction**

↓

Infer Operands

↓

Generate Transforms

↓

Rank Transforms

↓

Present Top-N

---

Text Selection

Text Editing

Row Selection

Column Selection

Transform Menu

Click Quality Meter
Transform Suggestion

Interaction ↓
Infer Operands ↓
Generate Transforms ↓
Rank Transforms ↓
Present Top-N

Map user input to transform operands.
Example: text highlight maps to row, column, and text selections.
Inferred text selections include string indices and regular expressions.
Text Selection Inference

Series Id: LNU020000000
Series Id: LNU02000000
-> ^ STR WS STR SYM WS STR NUM $
Text Selection Inference

Series Id: LNU02000000

-> ^ STR WS STR SYM WS STR NUM $
Text Selection Inference

Series Id: LNU02000000

MATCH Indices 11–22
Text Selection Inference

Series Id: LNU02000000

MATCH Indices 11–22

MATCH LNU02000000
Text Selection Inference

Series Id: LNU02000000
-> ^ STR WS STR SYM WS STR NUM $ 

Series Id: LNU02000000
MATCH Indices 11–22
MATCH LNU02000000
MATCH LNU NUM
MATCH STR NUM
Text Selection Inference

Series Id: LNU02000000
-> ^ STR WS STR SYM WS STR STR NUM $

Series Id: LNU02000000
MATCH Indices 11–22
MATCH LNU02000000
MATCH LNU NUM
MATCH STR NUM
AFTER : WS
Map user input to transform operands.

Example: text highlight maps to row, column, and text selections.

Inferred text selections include string indices and regular expressions.
Enumerate transforms that accept inferred operands as input.
Set unmatched params to default values.

Transform Suggestion

Generate Transforms

Rake Transforms

Infer Operands

Present Top-N

Interaction

Apply filter heuristics: No-ops, delete-all, and overly sparse outputs.
Enumenrate transforms that accept inferred operands as input.
Set unmatched params to default values.
Transform Suggestion

Interaction

↓

Infer Operands

↓

Generate Transforms

↓

Rank Transforms

↓

Present Top-N

Sort transforms by:

- Toolbar selection
- Specification difficulty
- Frequency in corpus
Transform Suggestion

Interaction
↓
Infer Operands
↓
Generate Transforms
↓
Rank Transforms
↓
Present Top-N

Extract from `unnamed_1` once between positions 17, 25

Extract from `unnamed_1` once on whitespace Alabama

Cut from `unnamed_1` once between positions 17, 25

Cut from `unnamed_1` once on whitespace Alabama

Split `unnamed_1` once between positions 17, 25 into columns

Split `unnamed_1` once on whitespace Alabama into columns
Comparative Evaluation with Excel

Median completion time for Wrangler at least twice as fast in all tasks ($p < 0.001$).

Suggestions and visual previews used heavily.
Difficult Transforms: Table Reshaping

### Fold

<table>
<thead>
<tr>
<th>Country</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Austria</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Belgium</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>China</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

### Pivot

<table>
<thead>
<tr>
<th>Country</th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>Boys</td>
<td>1</td>
</tr>
<tr>
<td>Australia</td>
<td>Girls</td>
<td>2</td>
</tr>
<tr>
<td>Austria</td>
<td>Boys</td>
<td>3</td>
</tr>
<tr>
<td>Austria</td>
<td>Girls</td>
<td>4</td>
</tr>
<tr>
<td>Belgium</td>
<td>Boys</td>
<td>5</td>
</tr>
<tr>
<td>Belgium</td>
<td>Girls</td>
<td>6</td>
</tr>
<tr>
<td>China</td>
<td>Boys</td>
<td>7</td>
</tr>
<tr>
<td>China</td>
<td>Girls</td>
<td>8</td>
</tr>
</tbody>
</table>
Proactive Wrangling

**Proactive transform suggestion** [UIST’11]
Guide users to a proper relational table

\[
S(T) = \left(1 - \frac{\sum_{c \in C} H_c(T)}{|C|}\right) + \frac{E(T) + D(T)}{|R||C|}
\]

Type homogeneity

\[
H_c = \sum_{Type} \left(\frac{|i \in R : c_i \in Type|}{|R|}\right)^2
\]
Proactive Wrangling

Proactive transform suggestion [UIST'11]
Guide users to a proper relational table

EVALUATION:
Compare automatic vs. manual transformation
53% of transforms automatically suggested
In those cases, the top-ranked suggestion is preferred 77% of the time (mean rank: 1.6).
Variables (with induced data types)

Results of Automatic Anomaly Detection

Data Profiler [AVI’12]
with Sean Kandel, Ravi Parikh & Joe Hellerstein
Schema Browser
- IMDB Rating
- IMDB Votes
- MPAA Rating
- Major Genre
- Production Budget
- Release Date
- Release Location
- Rotten Tomatoes Rating
- Running Time (min)
- Source

Anomaly Browser
- Missing (6)
  - MPAA Rating
  - Creative Type
  - Source
  - Major Genre
  - Distributor
  - Release Location
- Error (2)
- Extreme (7)
- Inconsistent (3)
- Schema (1)

Data Profiler [AVI’12]
with Sean Kandel, Ravi Parikh & Joe Hellerstein
Data Profiler [AVI’12]

with Sean Kandel, Ravi Parikh & Joe Hellerstein
Data Profiler [AVI'12]

with Sean Kandel, Ravi Parikh & Joe Hellerstein
Data Profiler [AVI’12]

with Sean Kandel, Ravi Parikh & Joe Hellerstein
Acquisition

Cleaning

Integration

Visualization

Modeling

Presentation

Dissemination
imMens: Real-Time Visual Querying of Big Data

with Zhicheng (Leo) Liu & Biye Jiang
Perceptual and interactive scalability should be limited by the chosen resolution of the visualized data, not the number of records.
Data
5-D Data Cube
Month, Hour, Day, X, Y
~2.3B bins
5-D Data Cube
Month, Hour, Day, X, Y
~2.3B bins
Full 5-D Cube
For any pair of 1D or 2D binned plots, the maximum number of dimensions needed to support brushing & linking is four.
Full 5-D Cube

13 3-D Data Tiles
Full 5-D Cube

- ~2.3B bins

13 3-D Data Tiles

- ~17.6M bins (in 352KB!)
<table>
<thead>
<tr>
<th>Index</th>
<th>X</th>
<th>Y</th>
<th>Day</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>256</td>
<td>512</td>
<td>0</td>
<td>378</td>
</tr>
<tr>
<td>1</td>
<td>256</td>
<td>512</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>30</td>
<td>256</td>
<td>512</td>
<td>30</td>
<td>1209</td>
</tr>
<tr>
<td>31</td>
<td>256</td>
<td>513</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>7935</td>
<td>256</td>
<td>767</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>7936</td>
<td>257</td>
<td>512</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2031615</td>
<td>511</td>
<td>767</td>
<td>30</td>
<td>466</td>
</tr>
<tr>
<td>index</td>
<td>X</td>
<td>Y</td>
<td>Day</td>
<td>Count</td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>0</td>
<td>256</td>
<td>512</td>
<td>0</td>
<td>378</td>
</tr>
<tr>
<td>1</td>
<td>256</td>
<td>512</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>30</td>
<td>256</td>
<td>512</td>
<td>30</td>
<td>1209</td>
</tr>
<tr>
<td>31</td>
<td>256</td>
<td>513</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>7935</td>
<td>256</td>
<td>767</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>7936</td>
<td>257</td>
<td>512</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2031615</td>
<td>511</td>
<td>767</td>
<td>30</td>
<td>466</td>
</tr>
</tbody>
</table>

The data is sparse and only includes specific entries.
<table>
<thead>
<tr>
<th>index</th>
<th>X</th>
<th>Y</th>
<th>Day</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>256</td>
<td>512</td>
<td>0</td>
<td>378</td>
</tr>
<tr>
<td>1</td>
<td>256</td>
<td>512</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>30</td>
<td>256</td>
<td>512</td>
<td>30</td>
<td>1209</td>
</tr>
<tr>
<td>31</td>
<td>256</td>
<td>513</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>7935</td>
<td>256</td>
<td>767</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>7936</td>
<td>257</td>
<td>512</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Sparse conversion to dense:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Day</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>512</td>
<td>0</td>
<td>378</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>256</td>
<td>512</td>
<td>30</td>
<td>1209</td>
</tr>
<tr>
<td>256</td>
<td>513</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>511</td>
<td>767</td>
<td>30</td>
<td>466</td>
</tr>
</tbody>
</table>

Dense conversion to sparse:

| 378 | 0   | ... | 1209 | 76   | ... | 0   | 0   | ... | 466 |

Diagram:

- The left table shows the sparse representation of data with indices and values.
- The right table demonstrates the conversion to dense representation, with repeated values indicating missing data.
- The middle table is the conversion process shown with arrows highlighting the transformation from sparse to dense and vice versa.
Dense packing more efficient if:

density > 25% in 3D tiles

density > 20% in 4D tiles
Query & Render on GPU via WebGL

Pack data tiles as PNG image files, bind to WebGL as image textures.
Query & Render on GPU via WebGL

Invoke program for each output bin.
Executes in parallel on GPU.
Query & Render on GPU via WebGL
Performance Benchmarks

Simulate interaction: brushing & linking across binned plots.

- imMens vs. Profiler
- 4x4 and 5x5 plots
- 10 to 50 bins

Measure time from selection to render.

Test setup:
2.3 GHz MacBook Pro (4-core)
NVIDIA GeForce GT 650M
Google Chrome v.23.0
~50fps querying of visual summaries of 1B data points.
Future Work

• Visualization specification interface
• Optimization considering resource constraints
• Integration with backend databases
• Server-side tile generation policies
• Activity modeling & prefetching schemes
Orion - Network Modeling & Analysis
with Adam Perer  [VAST’11]
GraphPrism

with Sanjay Kairam, Diana MacLean & Manolis Savva [AVI’12]
Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning


Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks—sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.
Termite Topic Model Viewer

with Jason Chuang & Chris Manning [AVI’12]
Interactive Data Analysis

http://vis.stanford.edu