(In search of..)
Frictionless Data

Adam Shepherd
Technical Director
Biological & Chemical Oceanography Data Management Office

Funding provided by NSF Award #1435578
80% of time is spent on data massaging

What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Courtesy of Forbes.com
80% say data massaging is the least enjoyable

What's the least enjoyable part of data science?

- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Courtesy of Forbes.com
FRICTION
How can data repositories improve "time-to-insight" for data consumers?
Data Containerization
Tabular Data Package

Data Package | JSON Table Schema | CSV Dialect Description

http://specs.frictionlessdata.io/
Focus on Extensibility

- Fiscal Data Package: Budgets and finances
- Neuro Data Package: e.g. neuroscience
- Tabular Data Package
  - Data Package
  - JSON Table Schema
  - CSV
Distributed Data Support

Data Package

Data Repo 1
Data Repo 2
Data Repo 3

Local Store
Tools
Platform Integrations

Tools: http://frictionlessdata.io/tools/
## Web Visualization Support

### Redline.js - relax with your data

<table>
<thead>
<tr>
<th>Grid</th>
<th>Graph</th>
<th>Map</th>
<th>Timeline</th>
<th>Transform</th>
<th>Results found 71</th>
</tr>
</thead>
<tbody>
<tr>
<td>LU growth</td>
<td>noname</td>
<td>Millions</td>
<td>Millions</td>
<td>LU Average</td>
<td>LU plus...</td>
</tr>
<tr>
<td>-5.4%</td>
<td>2006/200...</td>
<td>151.3</td>
<td>23.5</td>
<td>73.9</td>
<td>158.8</td>
</tr>
<tr>
<td>-2.5%</td>
<td>2006/200...</td>
<td>159.9</td>
<td>23.5</td>
<td>73.8</td>
<td>159.1</td>
</tr>
<tr>
<td>-2.7%</td>
<td>2006/200...</td>
<td>158.4</td>
<td>23.2</td>
<td>73.6</td>
<td>159.6</td>
</tr>
<tr>
<td>6.2%</td>
<td>2006/200...</td>
<td>161.3</td>
<td>23.6</td>
<td>74.1</td>
<td>140.1</td>
</tr>
<tr>
<td>15.0%</td>
<td>2006/200...</td>
<td>153.5</td>
<td>227.2</td>
<td>74.6</td>
<td>141.1</td>
</tr>
<tr>
<td>7.9%</td>
<td>2006/200...</td>
<td>153.2</td>
<td>227.3</td>
<td>75.2</td>
<td>141.7</td>
</tr>
<tr>
<td>6.1%</td>
<td>2006/200...</td>
<td>160.1</td>
<td>246.2</td>
<td>75.4</td>
<td>142.1</td>
</tr>
<tr>
<td>5.8%</td>
<td>2006/200...</td>
<td>166.5</td>
<td>250.0</td>
<td>76.0</td>
<td>142.6</td>
</tr>
<tr>
<td>4.9%</td>
<td>2006/200...</td>
<td>166.5</td>
<td>231.1</td>
<td>76.3</td>
<td>142.8</td>
</tr>
<tr>
<td>8.4%</td>
<td>2006/200...</td>
<td>138.9</td>
<td>210.0</td>
<td>76.7</td>
<td>142.9</td>
</tr>
<tr>
<td>6.4%</td>
<td>2006/200...</td>
<td>160.5</td>
<td>236.1</td>
<td>77.0</td>
<td>143.4</td>
</tr>
<tr>
<td>5.4%</td>
<td>2006/200...</td>
<td>163.0</td>
<td>244.2</td>
<td>77.4</td>
<td>144.3</td>
</tr>
<tr>
<td>7.4%</td>
<td>2007/200...</td>
<td>164.5</td>
<td>242.1</td>
<td>78.1</td>
<td>159.9</td>
</tr>
<tr>
<td>5.3%</td>
<td>2007/200...</td>
<td>168.7</td>
<td>248.3</td>
<td>78.4</td>
<td>160.6</td>
</tr>
<tr>
<td>7.7%</td>
<td>2007/200...</td>
<td>166.4</td>
<td>246.5</td>
<td>78.9</td>
<td>161.2</td>
</tr>
<tr>
<td>8.5%</td>
<td>2007/200...</td>
<td>167.9</td>
<td>251.6</td>
<td>79.4</td>
<td>161.7</td>
</tr>
<tr>
<td>9.9%</td>
<td>2007/200...</td>
<td>163.1</td>
<td>244.1</td>
<td>79.9</td>
<td>162.4</td>
</tr>
<tr>
<td>5.2%</td>
<td>2007/200...</td>
<td>163.3</td>
<td>241.4</td>
<td>80.2</td>
<td>163.2</td>
</tr>
<tr>
<td>9.0%</td>
<td>2007/200...</td>
<td>174.7</td>
<td>258.9</td>
<td>80.6</td>
<td>163.9</td>
</tr>
</tbody>
</table>

**Graph Type:**
- Lines and Points

**Group Column (x-axis):**
- Year

**Series A (y-axis):**
- % of labor force

**Add Series:**
- Bus growth
- LU plus bus growth
Data Package Pipelines

https://github.com/frictionlessdata/datapackage-pipelines

- data processing, cleaning, massaging
- facilitates reproducibility & automatic provenance generation
- for tabular data: joins, add/remove columns, create columns, filter, dump to ZIP, SQL, etc.
## Frictionless Data Package

<table>
<thead>
<tr>
<th>What of it?</th>
<th>Benefit to Data Repositories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple, extendible schema</td>
<td>● Automated validation</td>
</tr>
<tr>
<td>Describing data and related information</td>
<td>● Tightly coupled complex, distributed data relations</td>
</tr>
<tr>
<td>For data transport</td>
<td>● Submission from producers</td>
</tr>
<tr>
<td></td>
<td>● Data Massaging w. PROV</td>
</tr>
<tr>
<td></td>
<td>● Tool-ready for consumers</td>
</tr>
</tbody>
</table>
Frictionless be like…
Questions?

http://frictionlessdata.io

Tools: http://frictionlessdata.io/tools/

Case Studies: http://frictionlessdata.io/case-studies/

Specifications: http://specs.frictionlessdata.io/

Thank you to Rufus Pollock and Paul Walsh of the Open Knowledge Foundation.