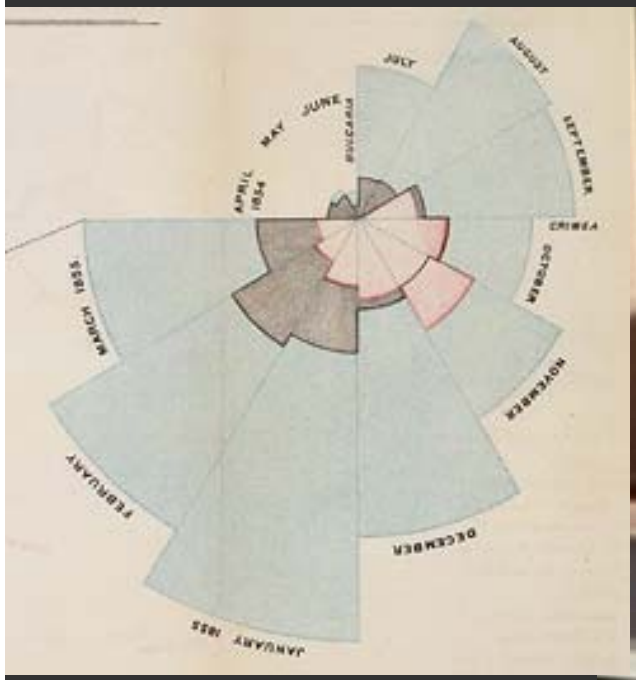
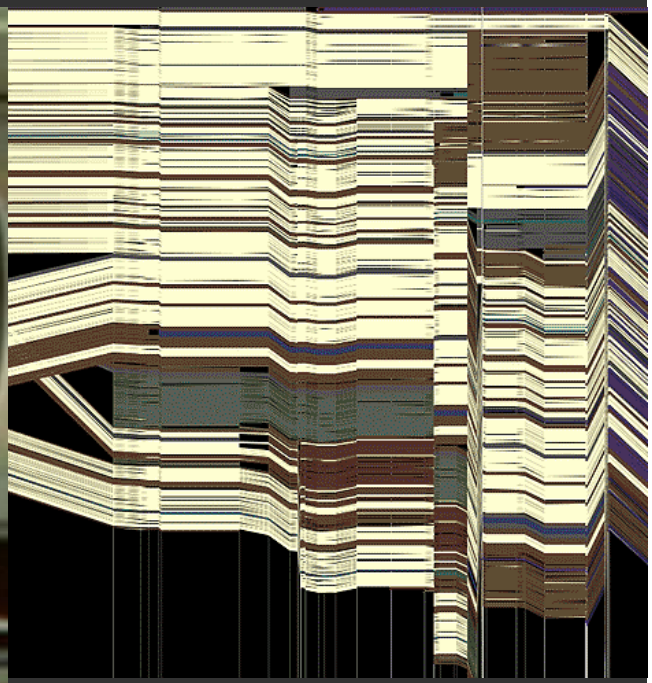
Interactive Tools for Data Transformation & Visualization





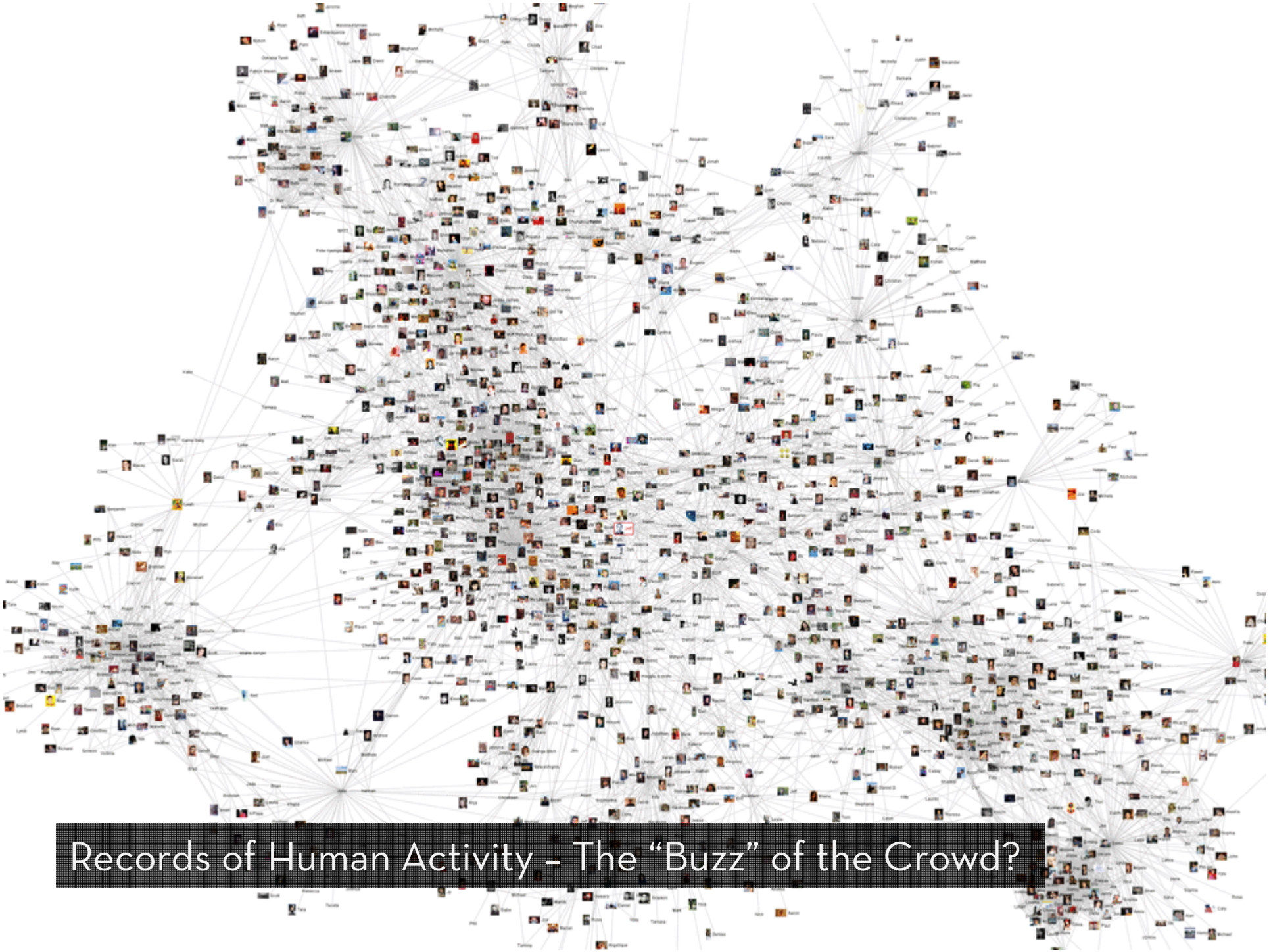
Jeffrey Heer Stanford University

How much data (bytes) will we produce in 2010?

2010: 1,200 exabytes

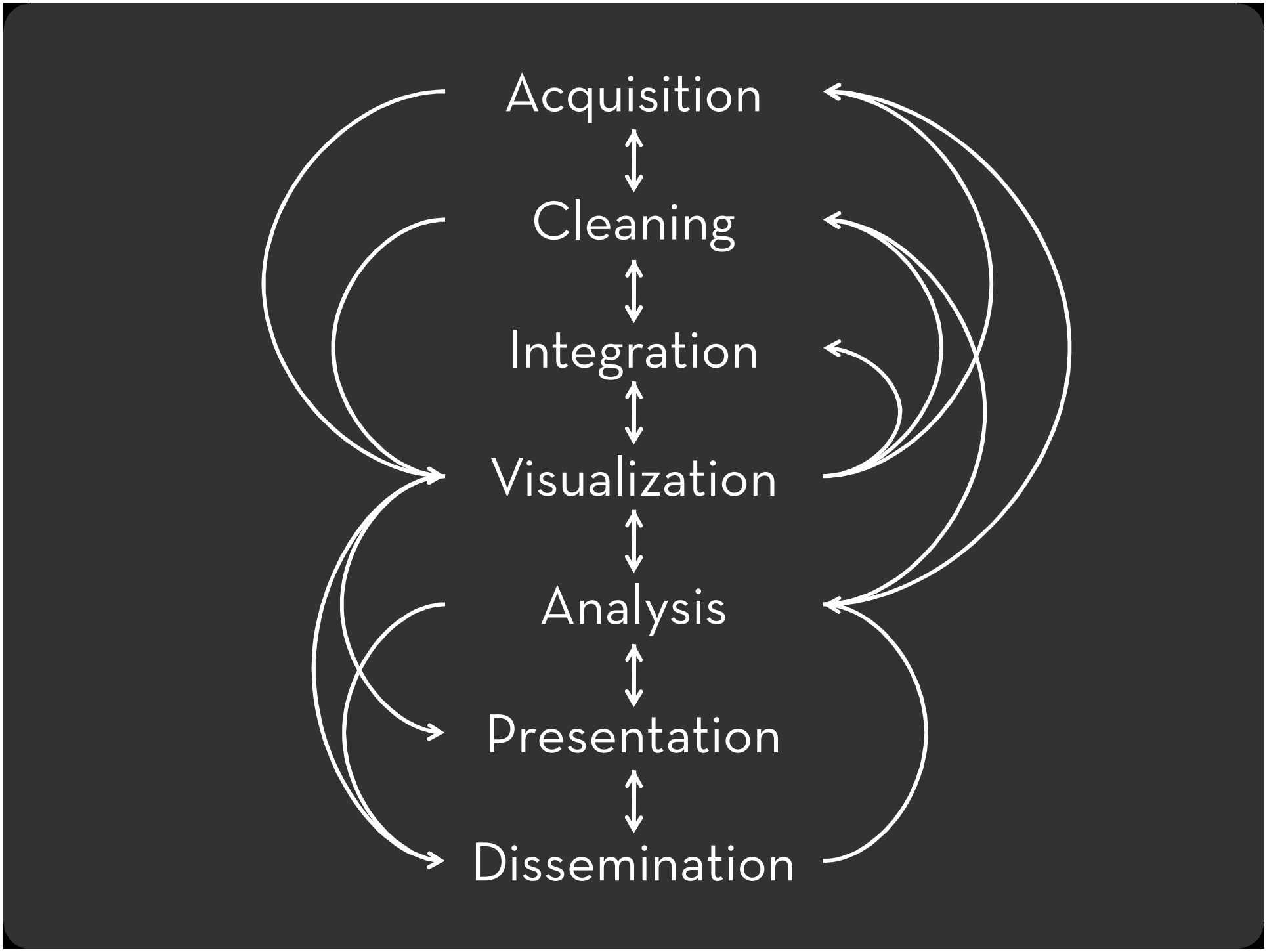
10x increase over 5 years

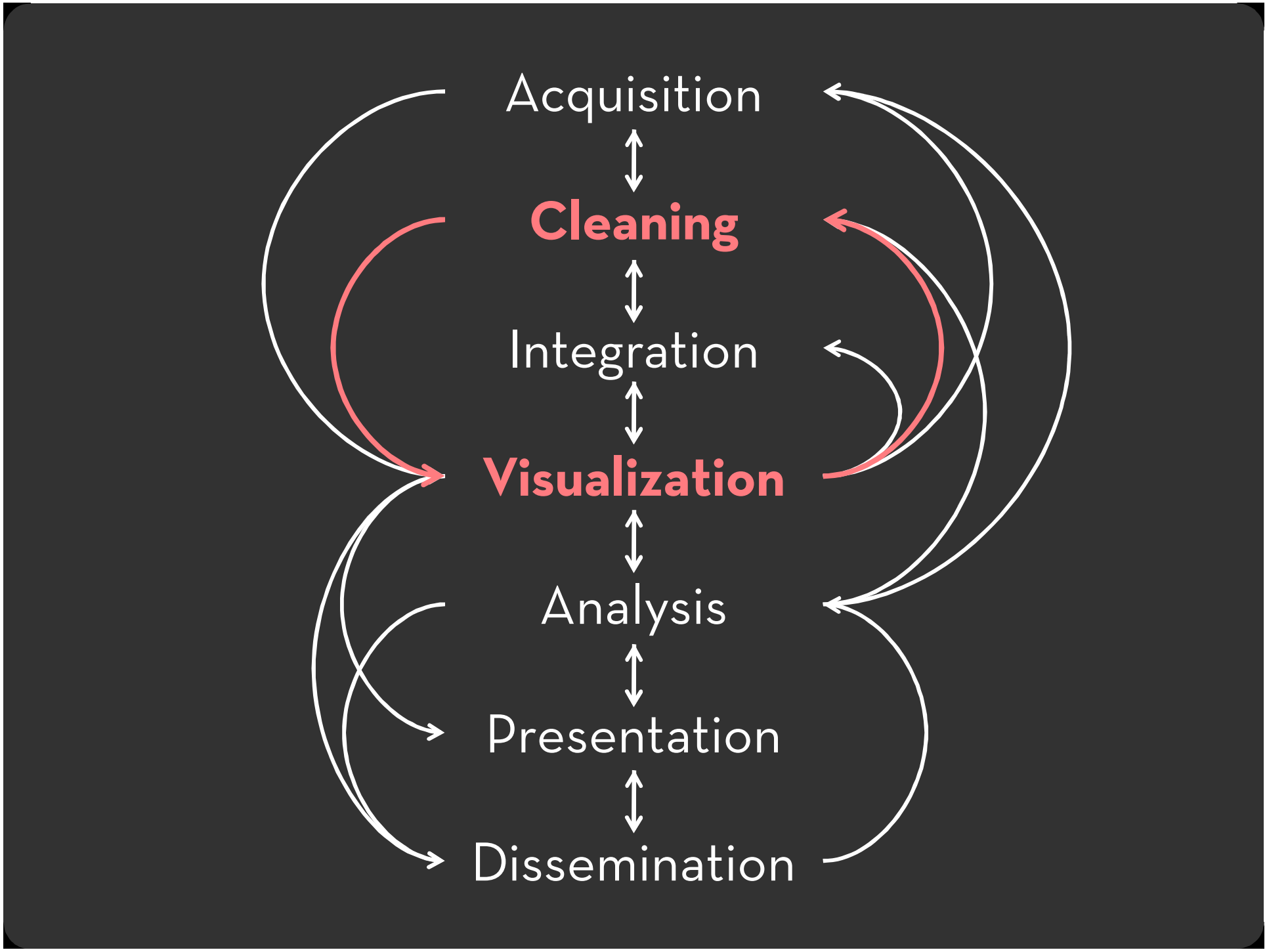
Gantz et al, 2008, 2010


Records of Human Activity – The “Buzz” of the Crowd?

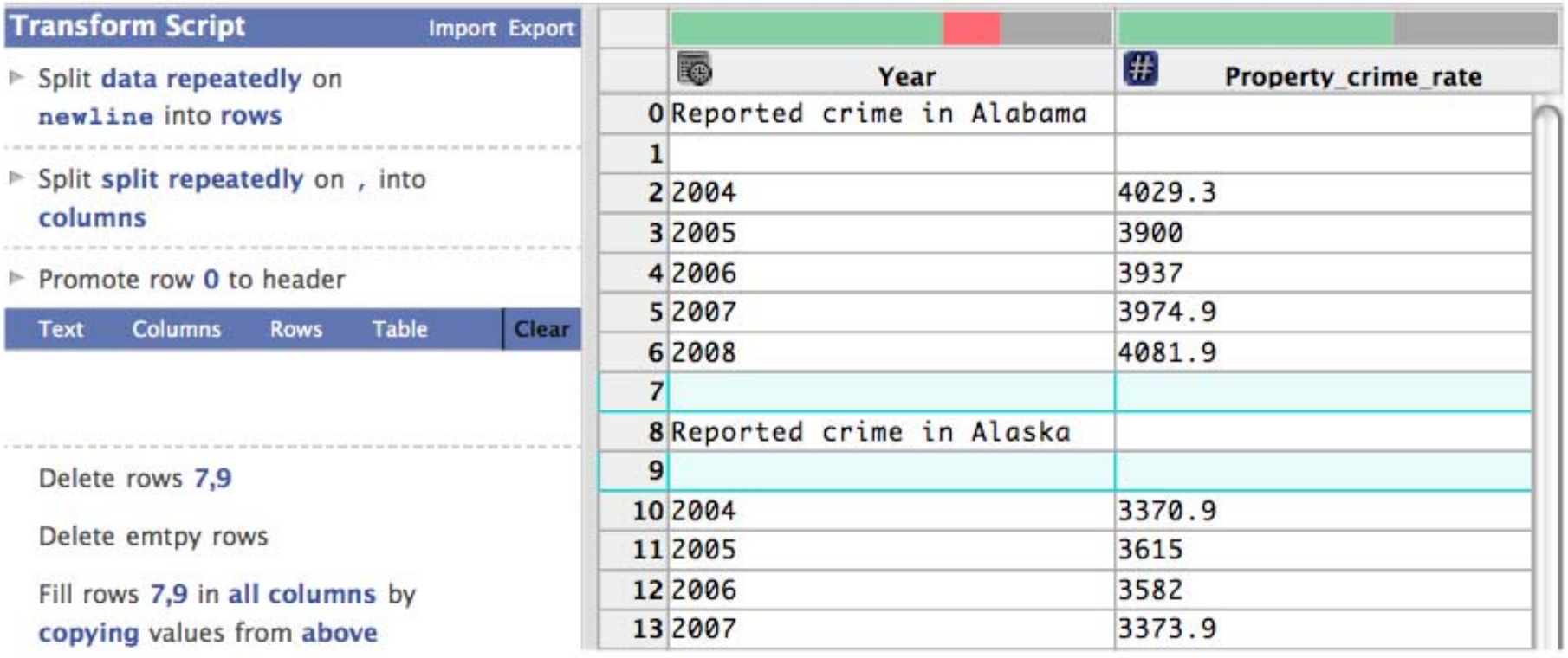
The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, … because now we really do have **essentially free and ubiquitous data**. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination

DataWrangler



with **Sean Kandel**, Andreas Paepcke & Joe Hellerstein

From UI to running code...

```
split('data').on(NEWLINE).max_splits(NO_MAX)
split('split').on(COMMA).max_splits(NO_MAX)
columnName().row(0)
delete(isEmpty())
extract('Year').on(/.*/).after(/in /)
fill('extract').method(COPY).direction(DOWN)
delete('Year starts with "Reported crime in"')
columnName('extract').to('State')
```

Data Wrangler

Declarative data transformation language

- **Tuple mapping** – split, merge, extract, delete
- **Lookups and joins** – e.g., FIPS code to US state
- **Reshaping** – e.g., cross-tabulation
- **Sorting, aggregation, etc.**

- Informed by prior work in databases, namely Potter's Wheel & SchemaSQL

Data Wrangler

Declarative data transformation language

+

Mixed-initiative interface for data transforms

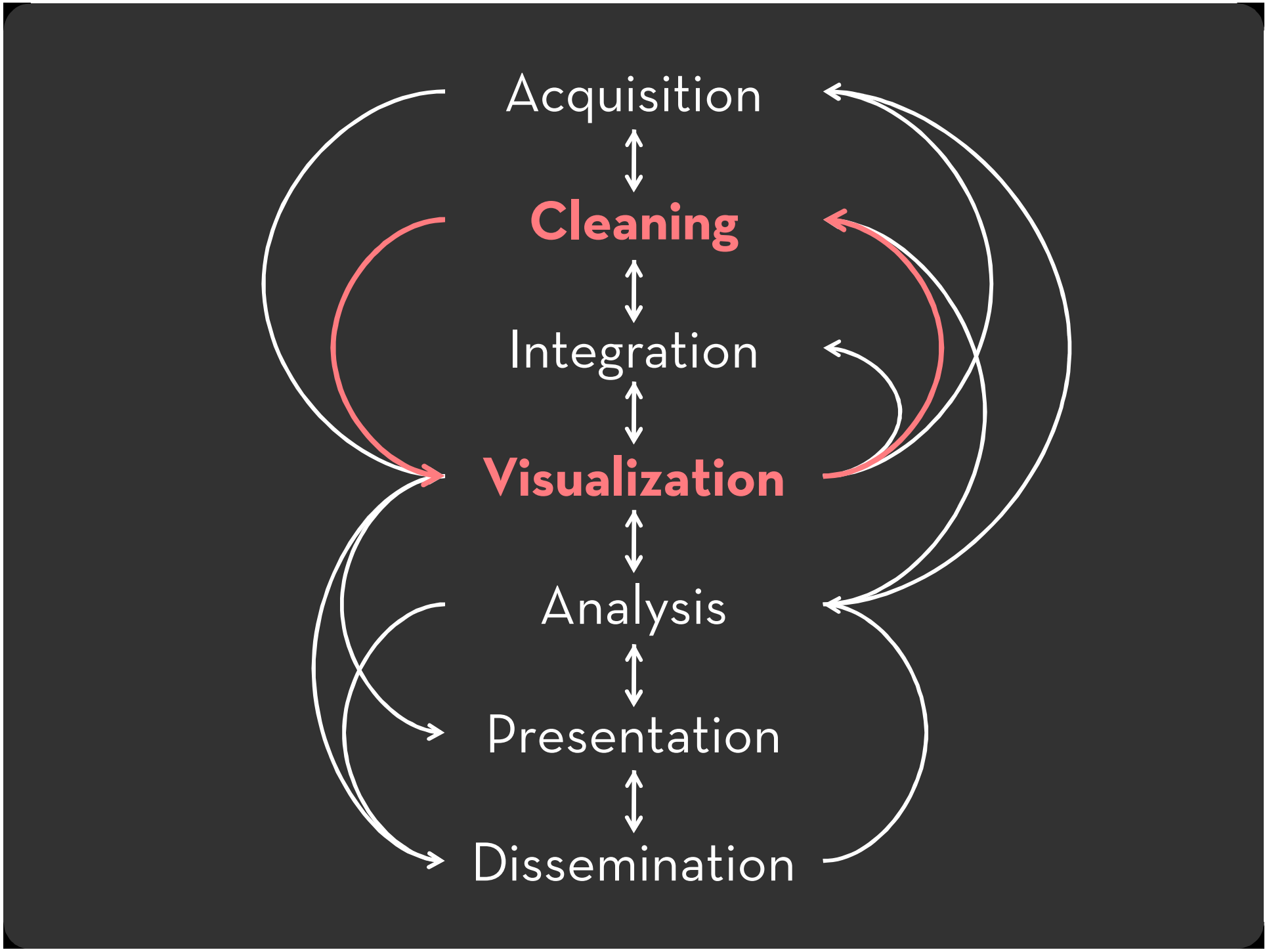
- **Select** data elements of interest
- **Suggest** applicable transforms
- Enable rapid **preview and refinement**

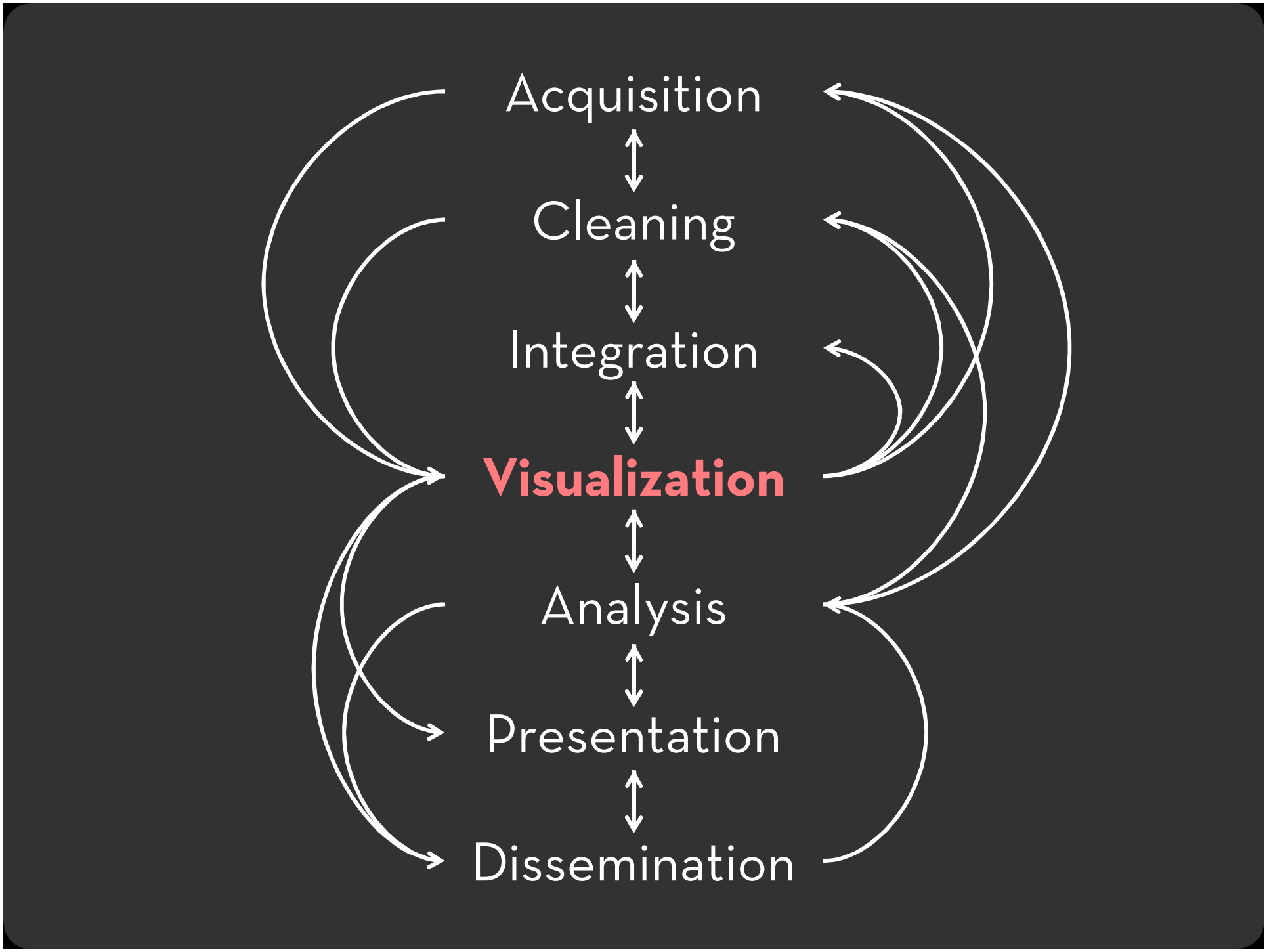
Comparative Evaluation

Compared Wrangler performance to Excel with 3 data cleaning tasks on small data sets.

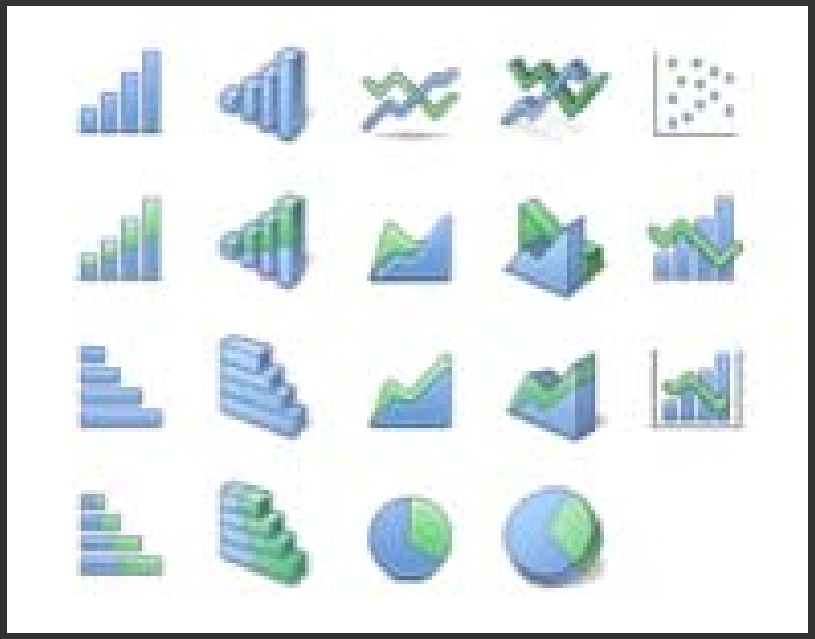
Median completion time for Wrangler at least twice as fact in all tasks.

Skilled Excel users benefit disproportionately!


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination

How do people create visualizations?



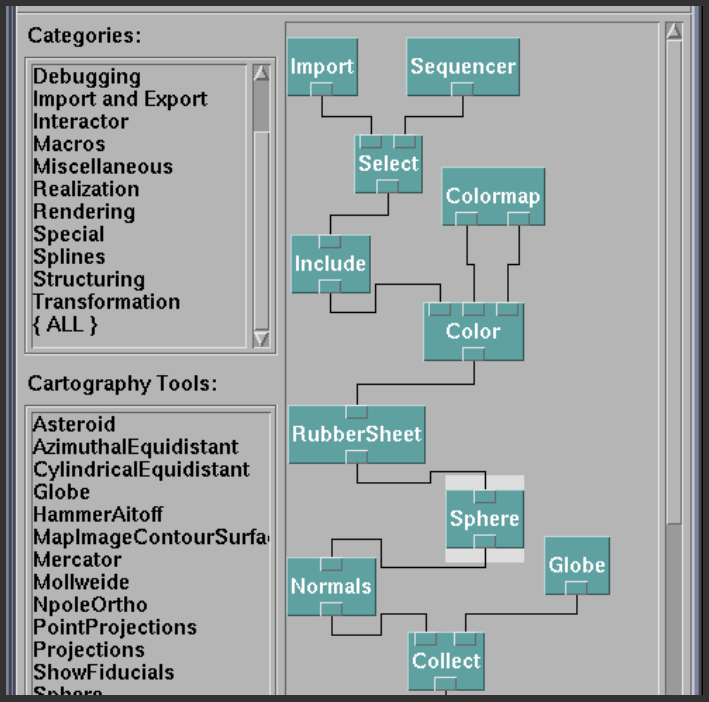


Chart Typology

Pick from a stock of templates

Easy-to-use but limited expressiveness

Prohibits novel designs, new data types

Component Model Architectures

Compose common high-level operations

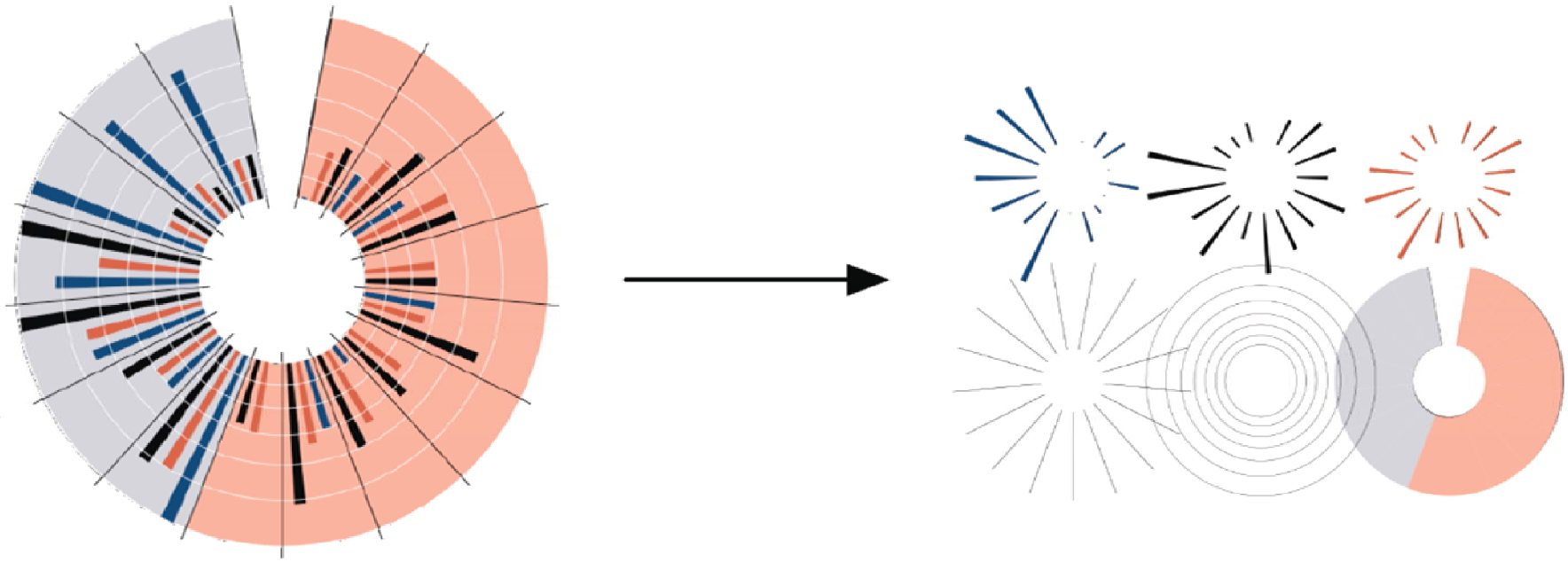
Permits more combinatorial possibilities

Novel views require new operators, in turn requiring software engineering.

Today's first task is not to invent wholly new [*graphical*] techniques, though these are needed. Rather we need most vitally to recognize and reorganize the essential of old techniques, to make easy their assembly in new ways, and to modify their external appearances to fit the new opportunities.

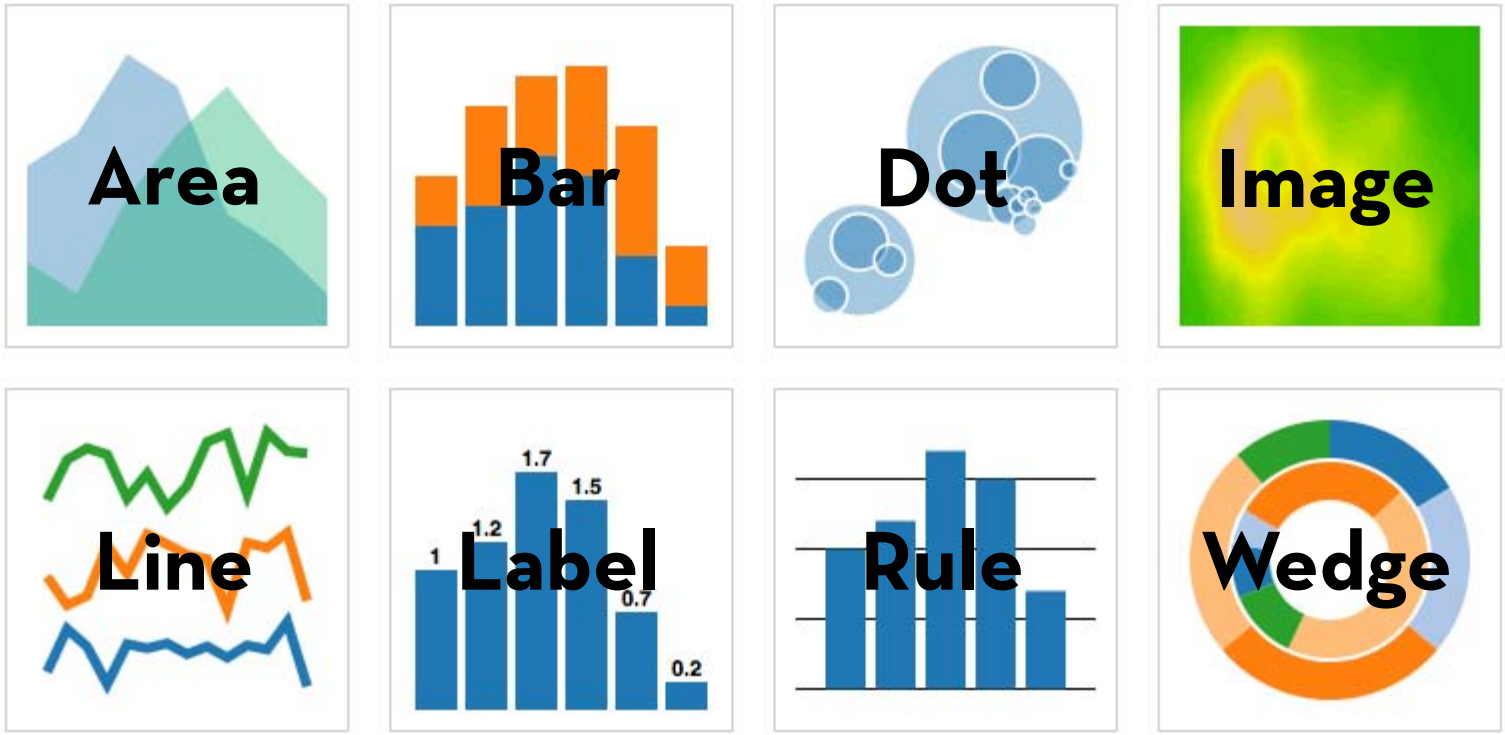
J. W. Tukey, *The Future of Data Analysis*, 1962.

Protovis: A Declarative Language for Visualization



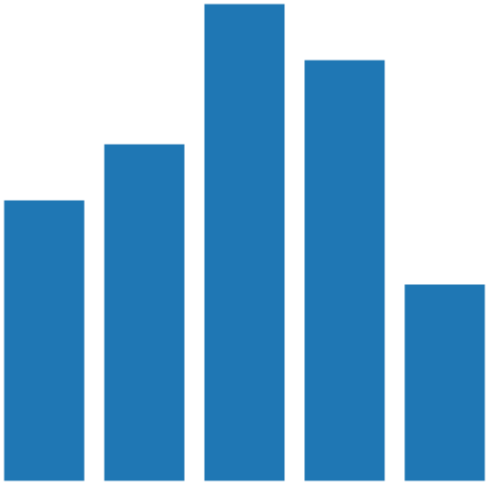
A graphic is a composition of data-representative marks.

with **Mike Bostock** & **Vadim Ogievetsky**


Area
Bar
Dot
Image
Line
1
1.2
1.7
1.5
0.7
0.2
Label
Rule
Wedge

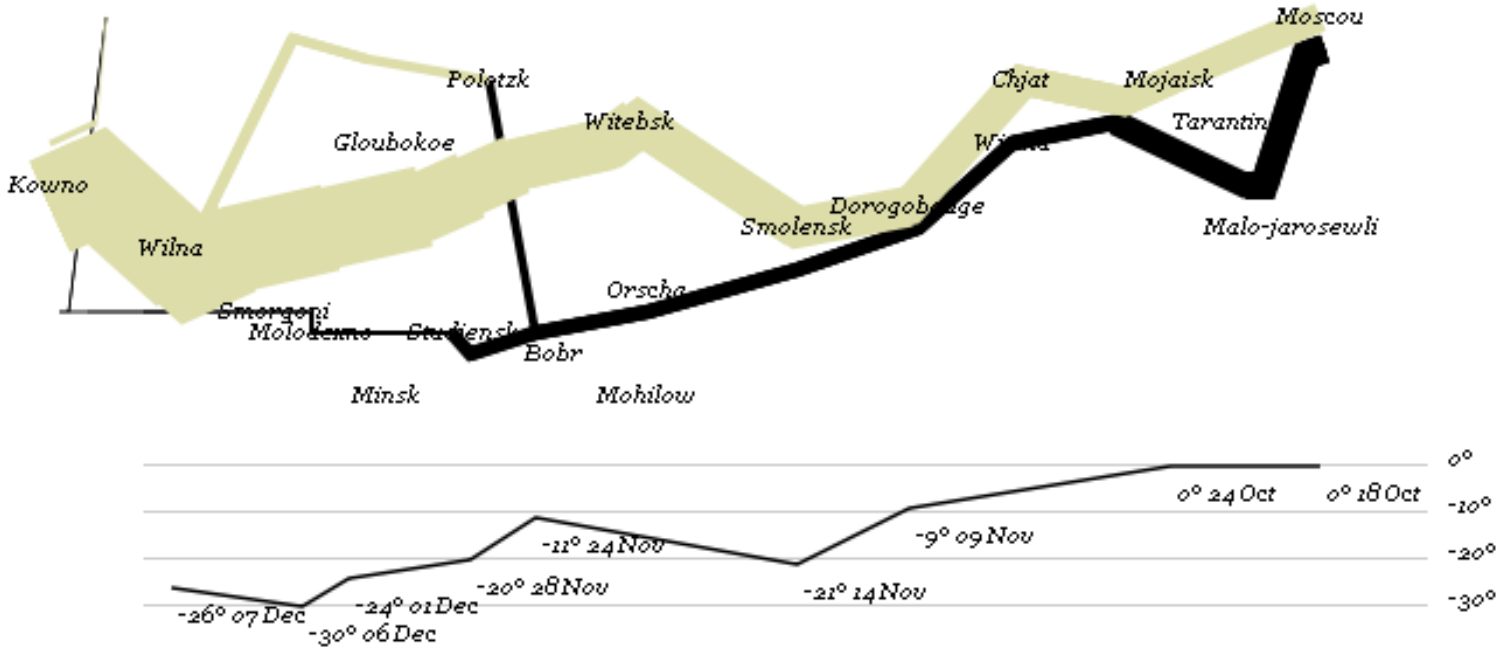
Protovis

Create customized visualizations using a declarative specification language.



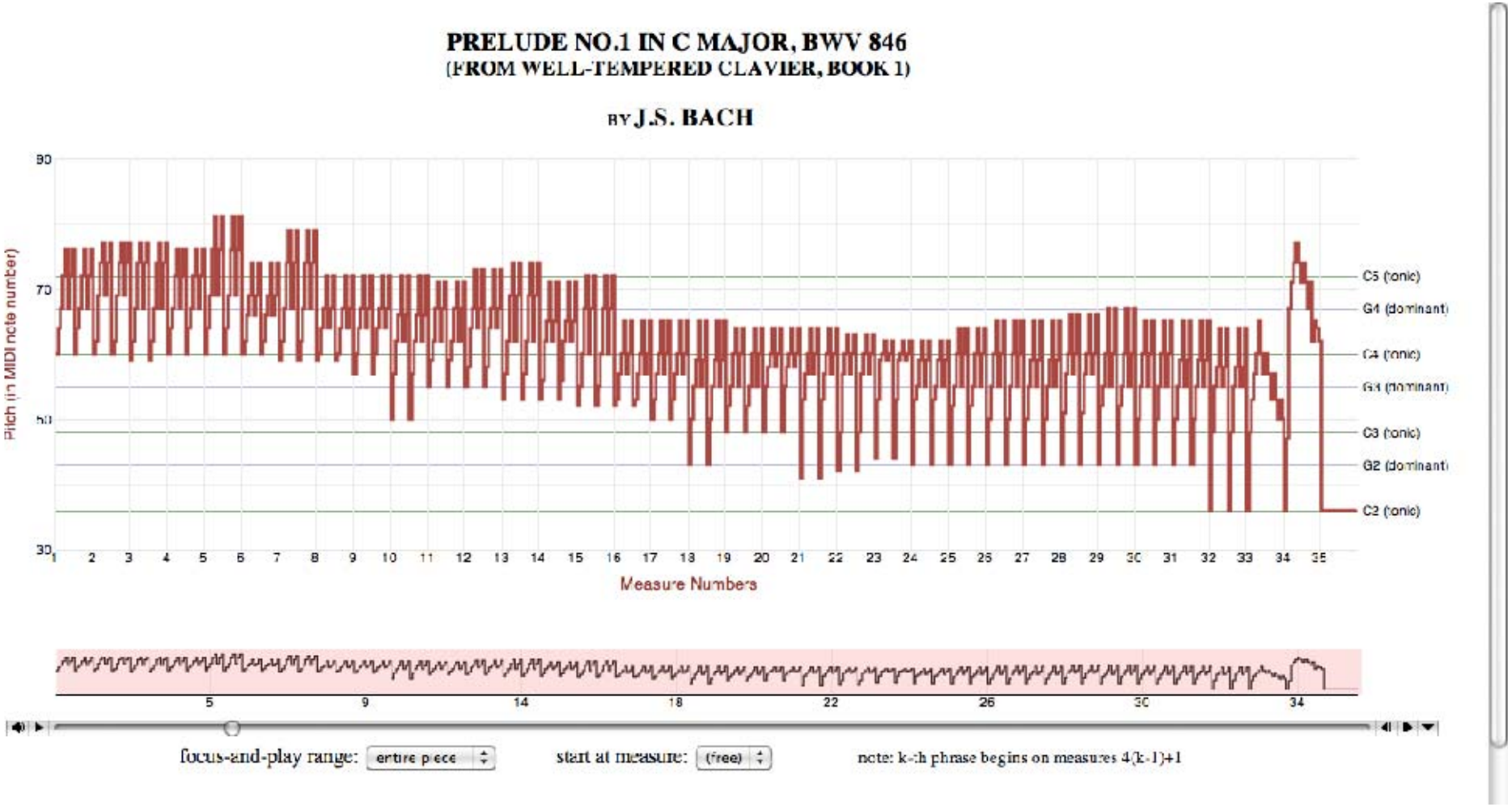
```
var vis = new pv.Panel();
vis.add(pv.Bar)
  .data([1, 1.2, 1.7, 1.5, .7])
  .bottom(10)
  .width(20)
  .height(function(d) d * 70)
  .left(function()  this.index * 25 + 20);
vis.render();
```

Protovis (http://protovis.org) – Declarative Visualization Specification

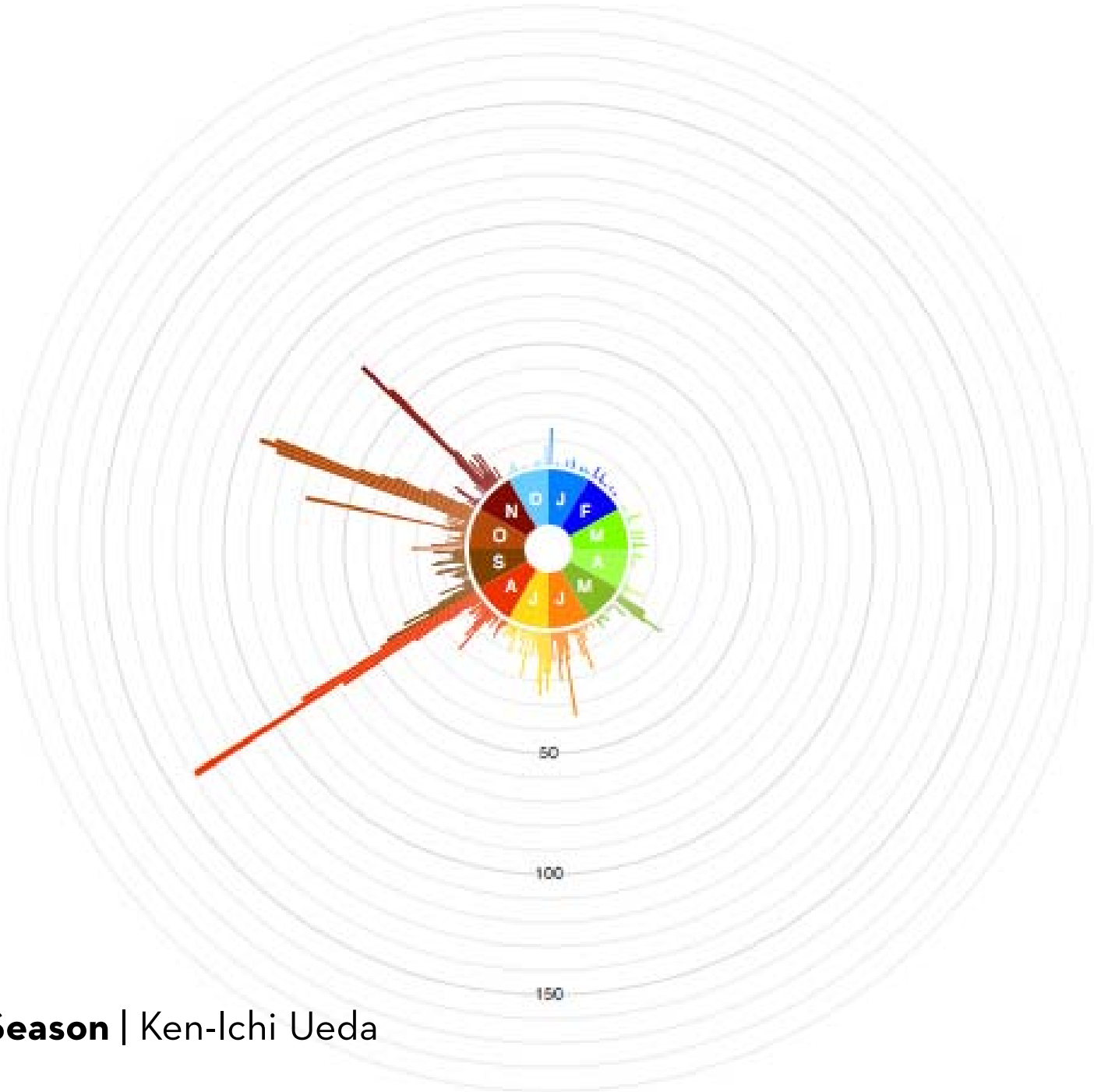



```
var army = pv.nest(napoleon.army, "dir", "group“);
var vis = new pv.Panel();

var lines = vis.add(pv.Panel).data(army);
lines.add(pv.Line)
  .data(function() army[this.idx])
  .left(lon).top(lat).size(function(d) d.size/8000)
  .strokeStyle(function() color[army[paneIndex][0].dir]);

vis.add(pv.Label).data(napoleon.cities)
  .left(lon).top(lat)
  .text(function(d) d.city).font("italic 10px Georgia")
  .textAlign("center").textBaseline("middle");
```

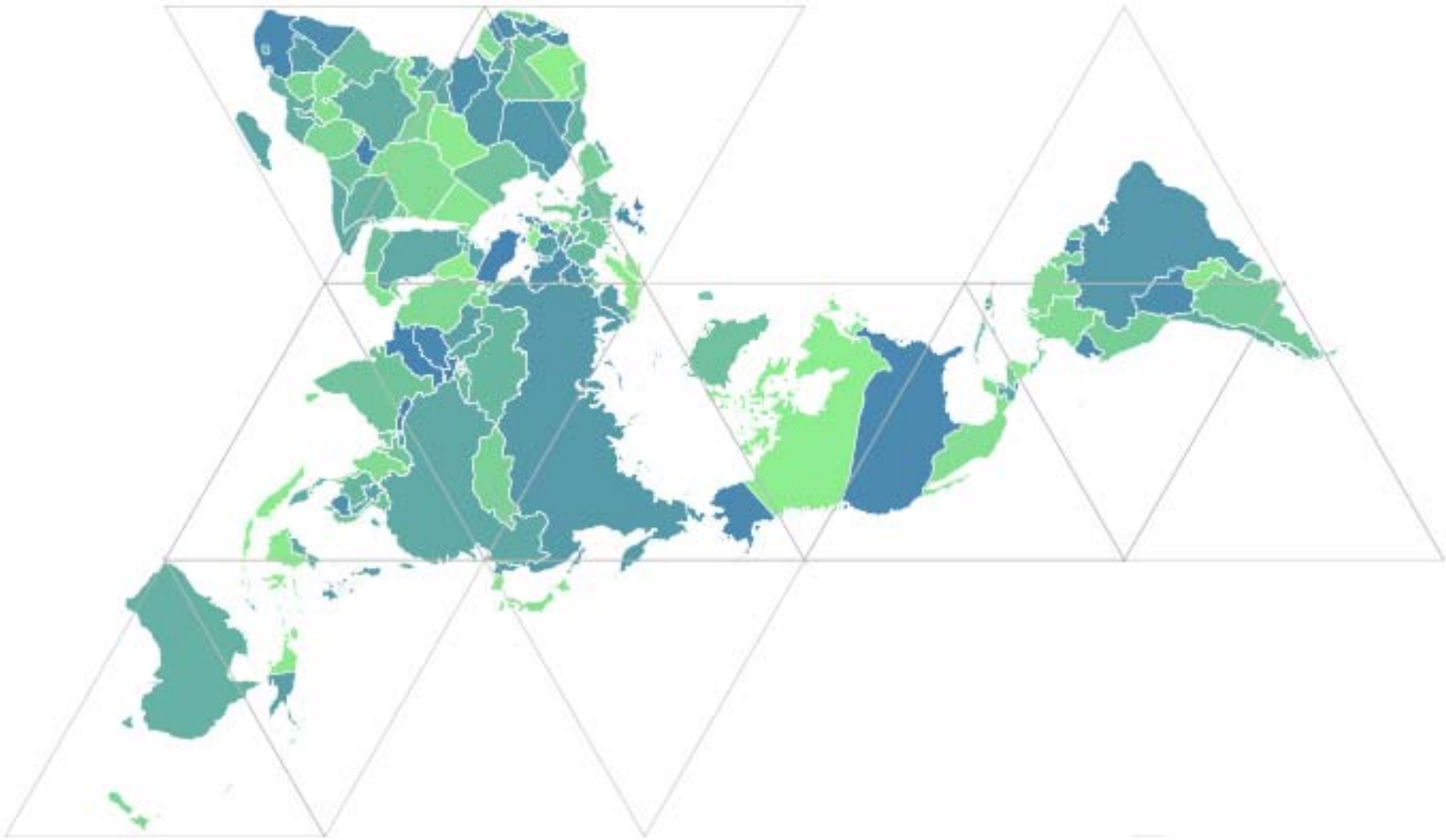
```
vis.add(pv.Rule).data([0,-10,-20,-30])
  .top(function(d) 300 - 2*d - 0.5).left(200).right(150)
  .lineWidth(1).strokeStyle("#ccc")
  .anchor("right").add(pv.Label)
    .font("italic 10px Georgia")
    .text(function(d) d+"°").textBaseline("center");

vis.add(pv.Line).data(napoleon.temp)
  .left(lon).top(tmp) .strokeStyle("#0")
 .add(pv.Label)
  .top(function(d) 5 + tmp(d))
  .text(function(d) d.temp+"° "+d.date.substr(0,6))
  .textBaseline("top").font("italic 10px Georgia");
```



Bach's Prelude #1 in C Major | Jieun Oh



FlickrSeason | Ken-Ichi Ueda



Dymaxion Maps | Vadim Ogievetsky

Exploiting Declarative Specification

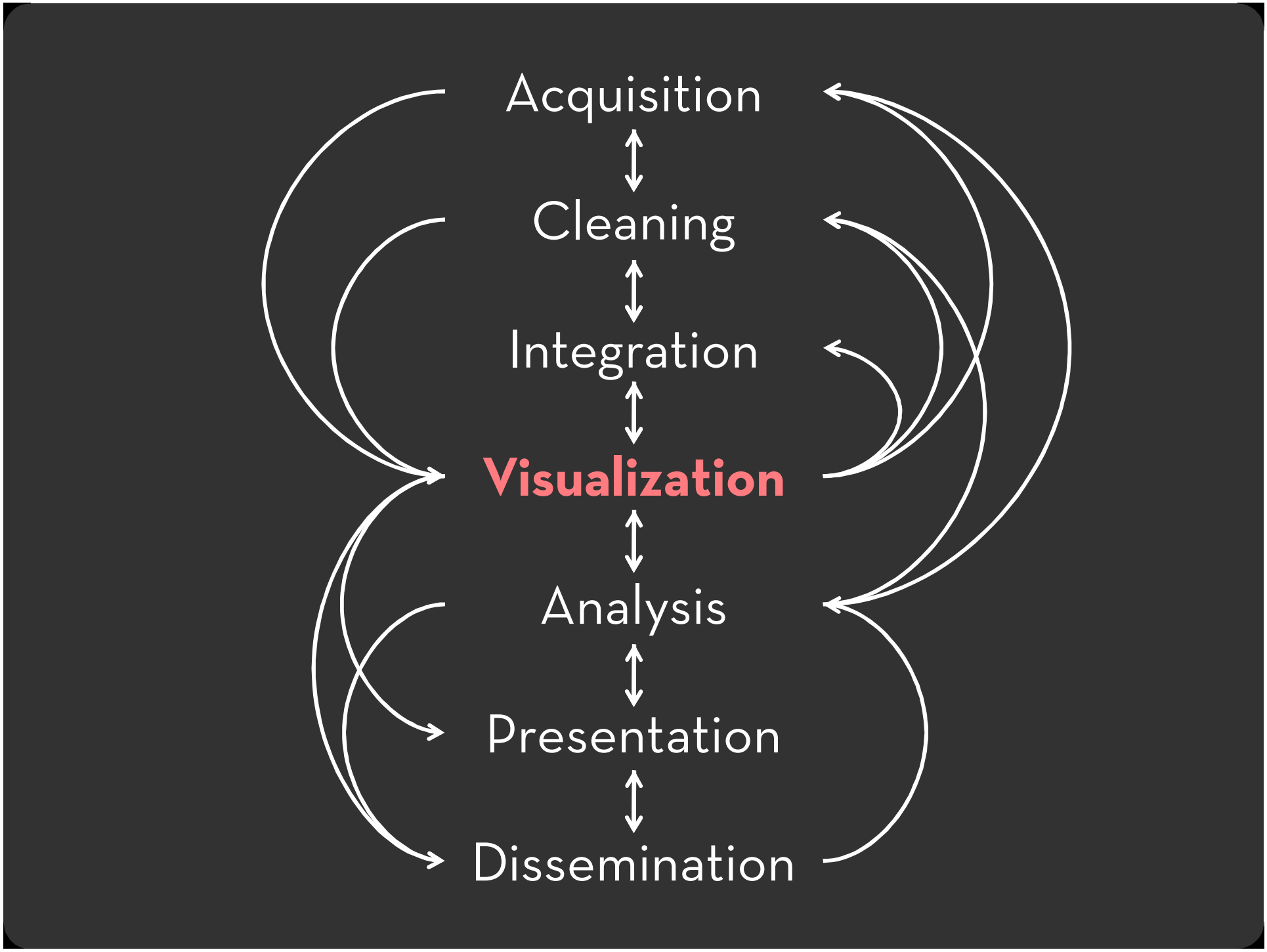
Protovis has led to faster designs, less code

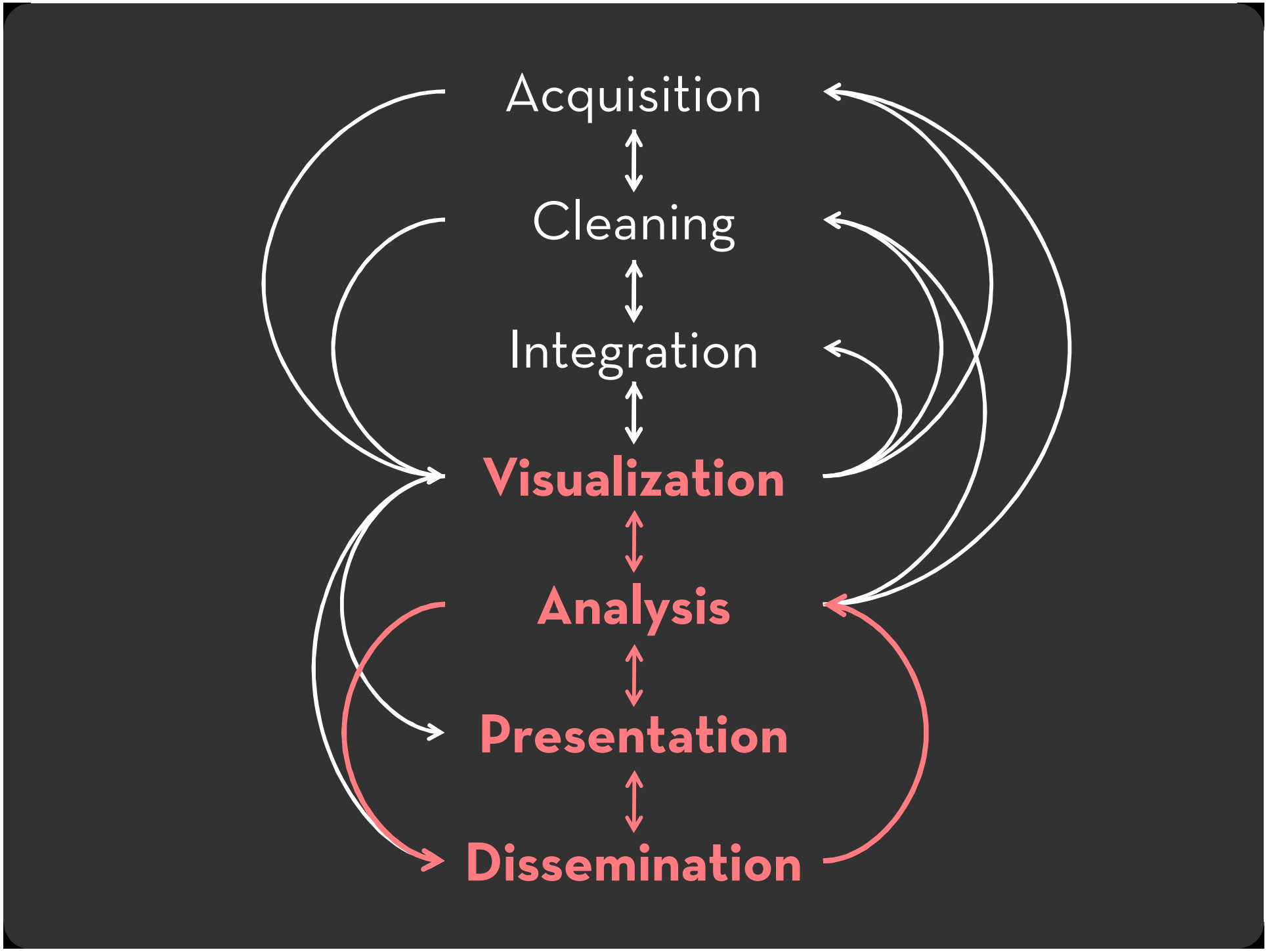
- Job Voyager: 5x less code, 10x less dev time
- Over 20,000 downloads and widely in use

Multiple implementations: JavaScript & Java

Behind-the-scenes optimization & parallelization

- 20x scalability over prior systems (in Java)


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination

sense.us

A Web Application for Collaborative Visualization of Demographic Data

with **Fernanda Viégas** and **Martin Wattenberg**

sense.us > social data visualization - Mozilla Firefox

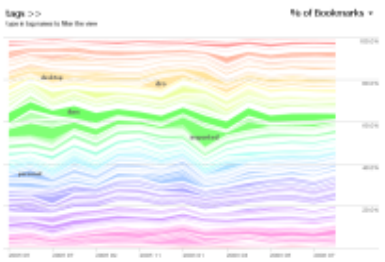
File Edit View History Bookmarks Tools Help

http://sense.us/

Google

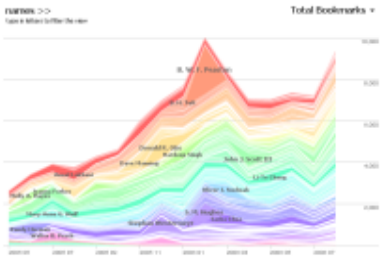
sense.us

Log in to add comments! ibm e-mail pass login help



Dogear Tags

Dogear Tag Usage, May 2005 to August 2006
source: IBM Dogear
10 comments



Dogear People

Dogear Bookmarking by Person, May 2005 to August 2006
source: IBM Dogear
1 comment

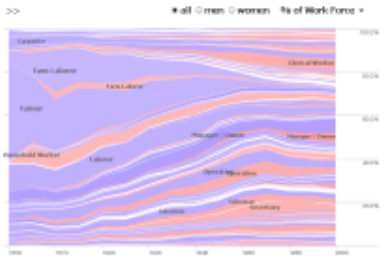
sense.us - social data visualization

sense.us is a prototype system for collaborative visualization.
- See the data. See what people have to say about it.
- Dive into the data and share your explorations.

The site requires **Java 1.5+** and either **Firefox** or **Internet Explorer**.
Use your **IBM W3/BluePages e-mail and password** to login.
Use at least **1024x768** resolution for the best experience.

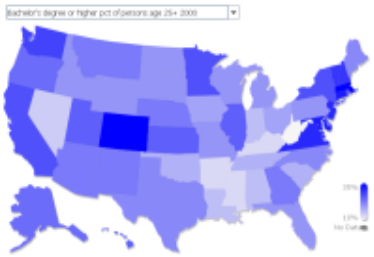
Check out the **user's guide** and **privacy policy** before getting started.

Having problems using Firefox with Java 1.4? Some users using Firefox and Java 1.4 have found that comments aren't loading properly. If you run into this problem, consider upgrading to Java 1.5 (**Windows**, **Linux**) or using Internet Explorer (on Windows systems). Sorry for any inconvenience!



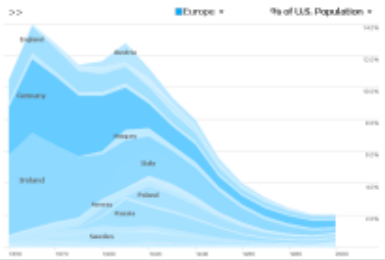
Job Voyager

Reported Occupations of U.S. Labor Force, 1850-2000
source: http://ipums.org
139 comments



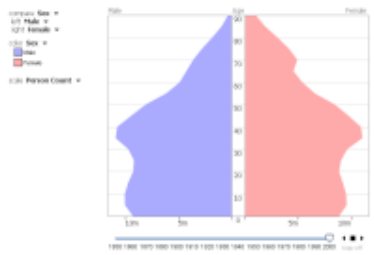
U.S. Census State Map

State Map of 2000-2005 Census Data
source: U.S. Census Bureau
16 comments



Birthplace Voyager

Reported Birthplace of U.S. Residents, 1850-2000
source: http://ipums.org
10 comments



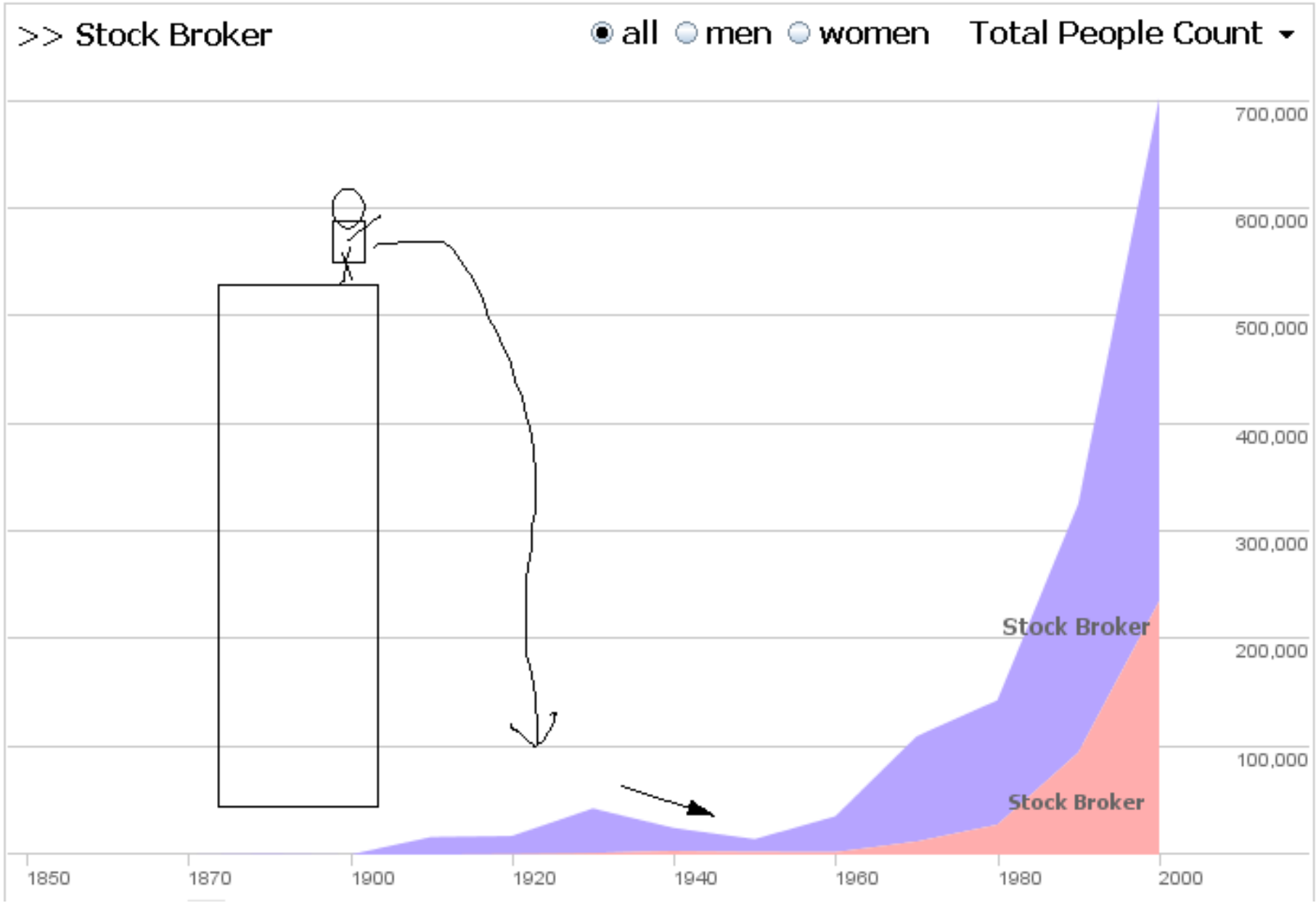
Population Pyramid

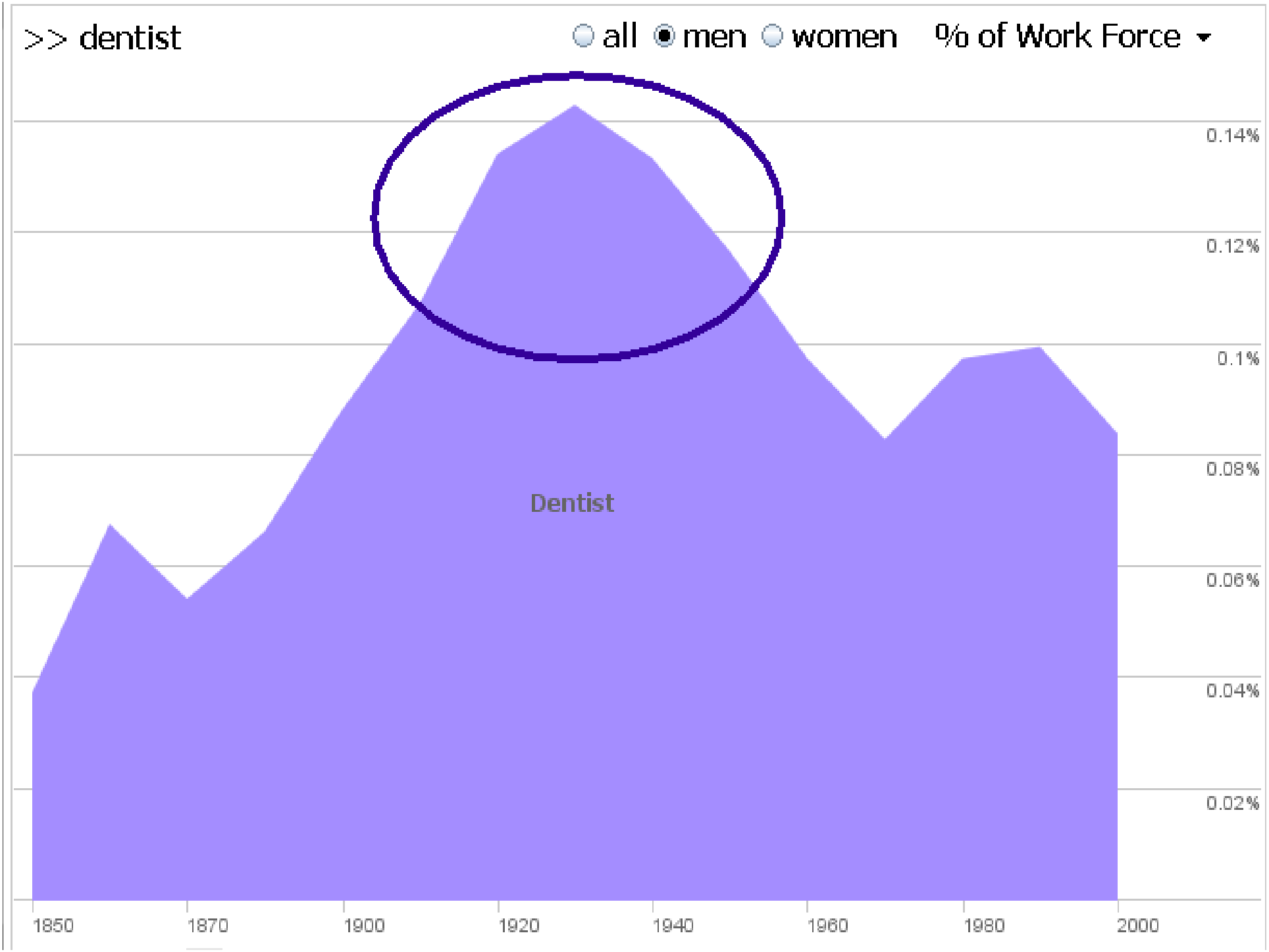
U.S. Population Demographics, 1850-2000
source: http://ipums.org
7 comments

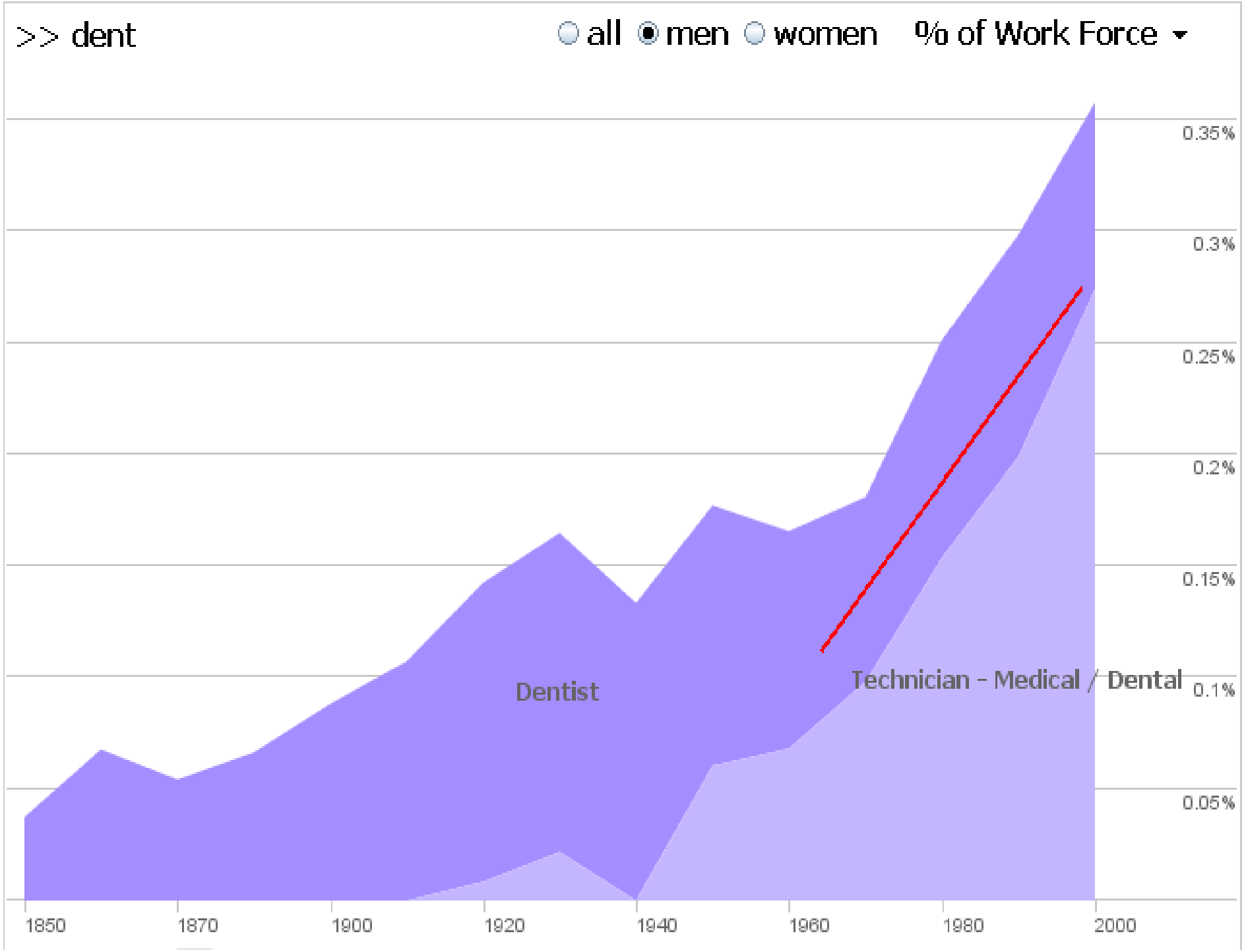
Done


>> mili
all
men
women
% of Work Force
2.0%
1.5%
1.0%
0.5%
WW I
Civil war
Military
Military
1850
1870
1900
1920
1940
1960
1980
2000

Great depression "killed" a lot of brokers



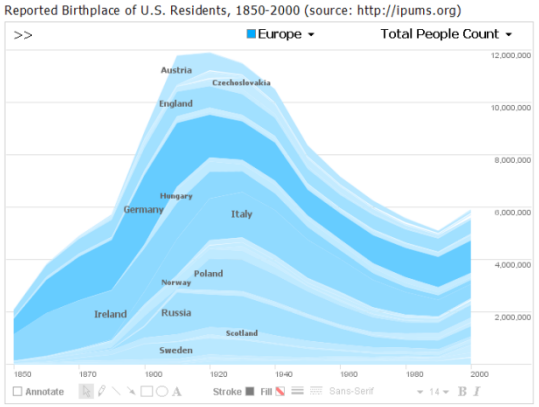

>> dentist
all
men
women
% of Work Force
0.14%
0.12%
0.1%
0.08%
0.06%
0.04%
0.02%
Dentist
1850
1870
1900
1920
1940
1960
1980
2000


>> dent
all
men
women
% of Work Force
0.35%
0.3%
0.25%
0.2%
0.15%
0.1%
0.05%
Dentist
Technician - Medical / Dental
1850
1870
1900
1920
1940
1960
1980
2000

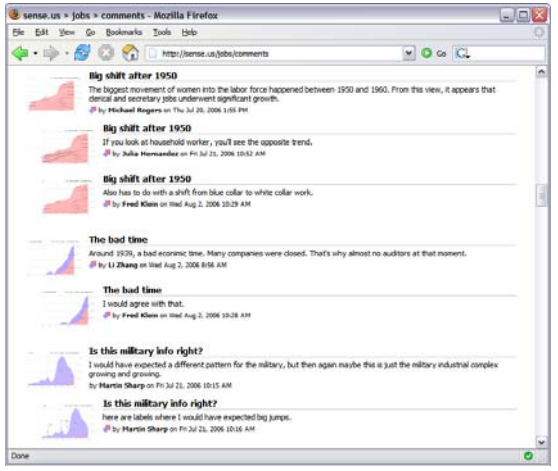
Voyagers and Voyeurs

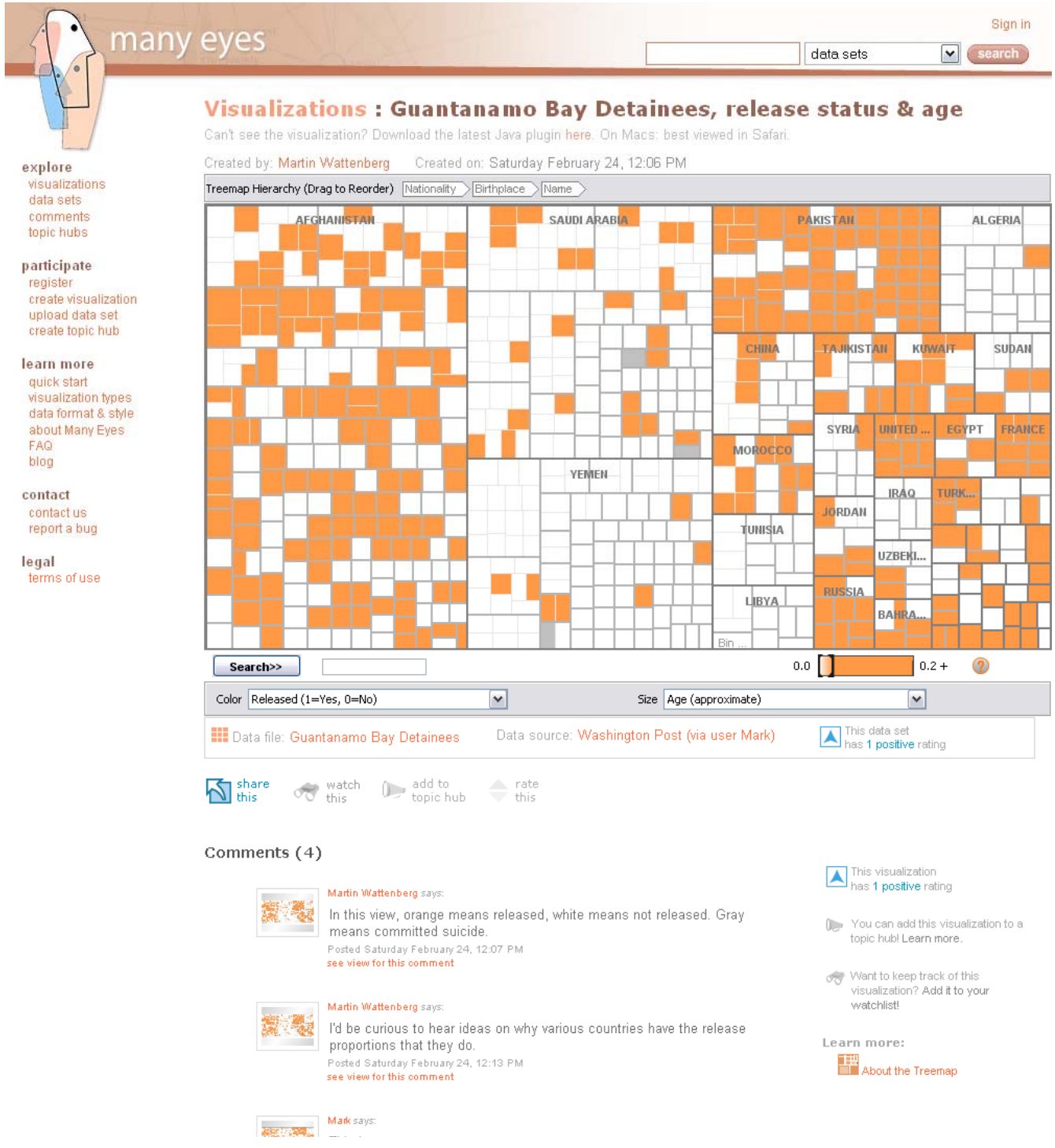
Complementary faces of analysis

Voyager – focus on visualized data
Active engagement with the data
Serendipitous comment discovery



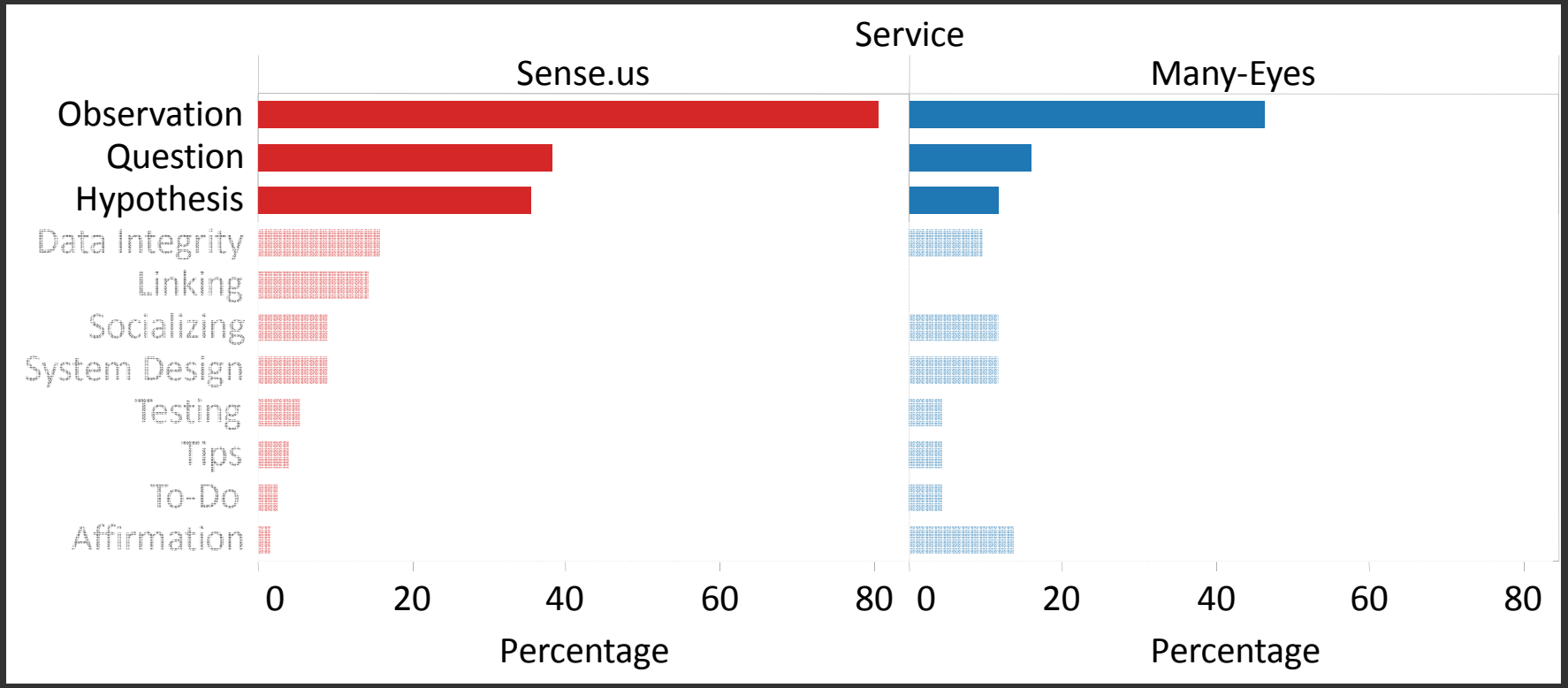
Voyeur – focus on comment listings
Investigate others' explorations
Find people and topics of interest
Catalyze new explorations




many eyes
Sign in
data sets
search
Visualizations : Guantanamo Bay Detainees, release status & age
Can't see the visualization? Download the latest Java plugin here. On Macs: best viewed in Safari.
Created by: Martin Wattenberg
Created on: Saturday February 24, 12:06 PM
explore
visualizations
data sets
comments
topic hubs
participate
register
create visualization
upload data set
create topic hub
learn more
quick start
visualization types
data format & style
about Many Eyes
FAQ
blog
contact
contact us
report a bug
legal
terms of use
Treemap Hierarchy (Drag to Reorder)
Nationality
Birthplace
Name
AFGHANISTAN
SAUDI ARABIA
PAKISTAN
ALGERIA
CHINA
TAJIKISTAN
KUWAIT
SUDAN
SYRIA
UNITED ...
EGYPT
FRANCE
MOROCCO
YEMEN
IRAQ
TURK...
JORDAN
TUNISIA
UZBEKI...
RUSSIA
LIBYA
BAHRA...
Bin ...
Search>>
0.0
0.2 +
Color
Released (1=Yes, 0=No)
Size
Age (approximate)
Data file: Guantanamo Bay Detainees
Data source: Washington Post (via user Mark)
This data set has 1 positive rating
share this
watch this
add to topic hub
rate this
Comments (4)
Martin Wattenberg says:
In this view, orange means released, white means not released. Gray means committed suicide.
Posted Saturday February 24, 12:07 PM
see view for this comment
Martin Wattenberg says:
I'd be curious to hear ideas on why various countries have the release proportions that they do.
Posted Saturday February 24, 12:13 PM
see view for this comment
Mark says:
This visualization has 1 positive rating
You can add this visualization to a topic hub! Learn more.
Want to keep track of this visualization? Add it to your watchlist!
Learn more:
About the Treemap

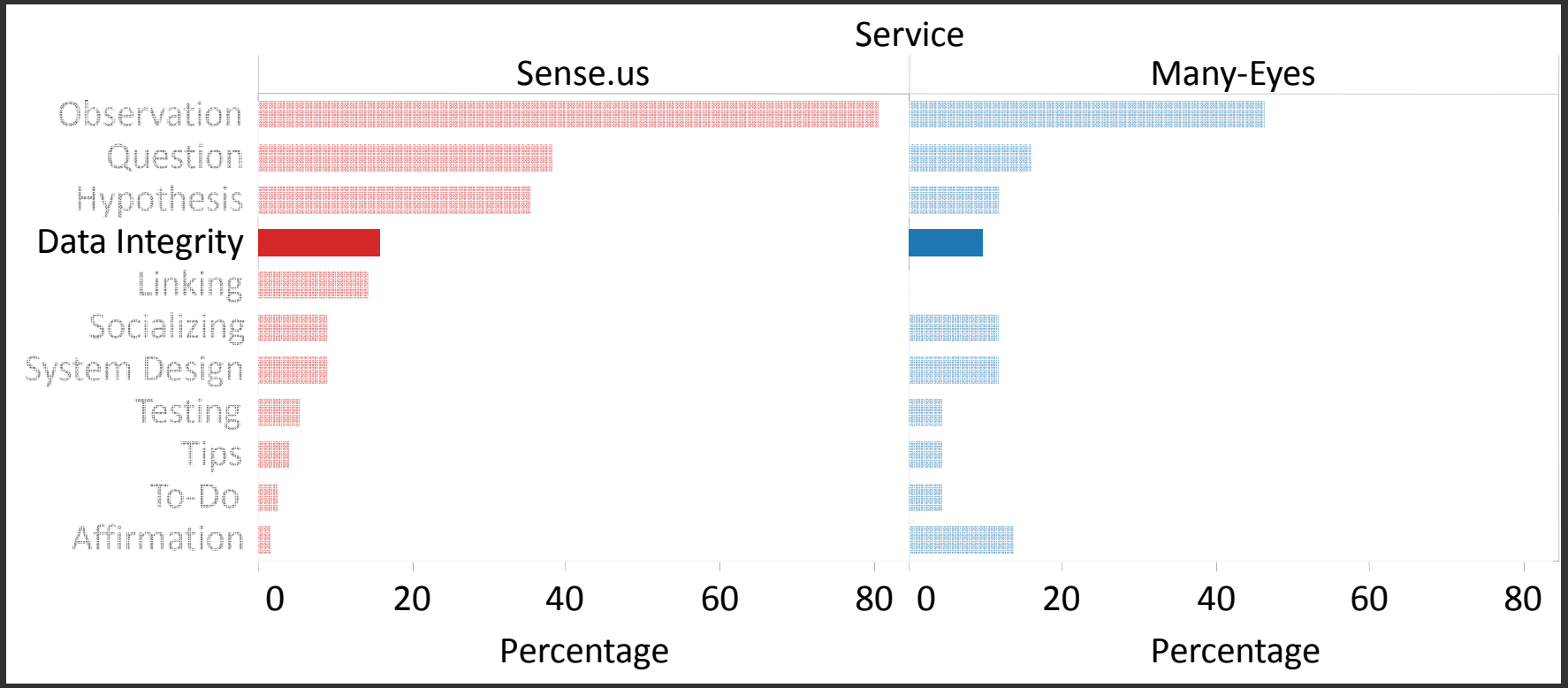
Many-Eyes

Content Analysis of Comments

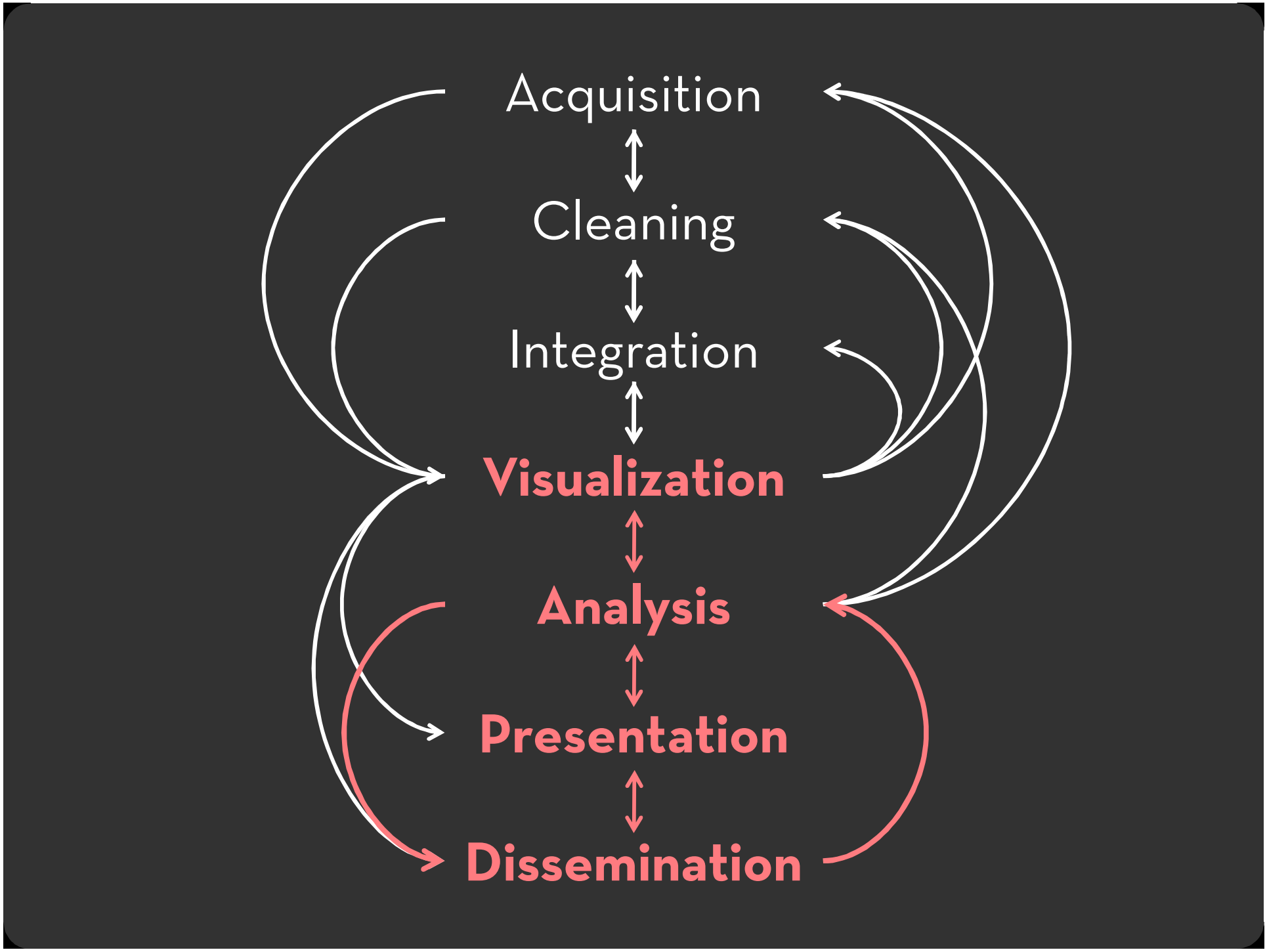


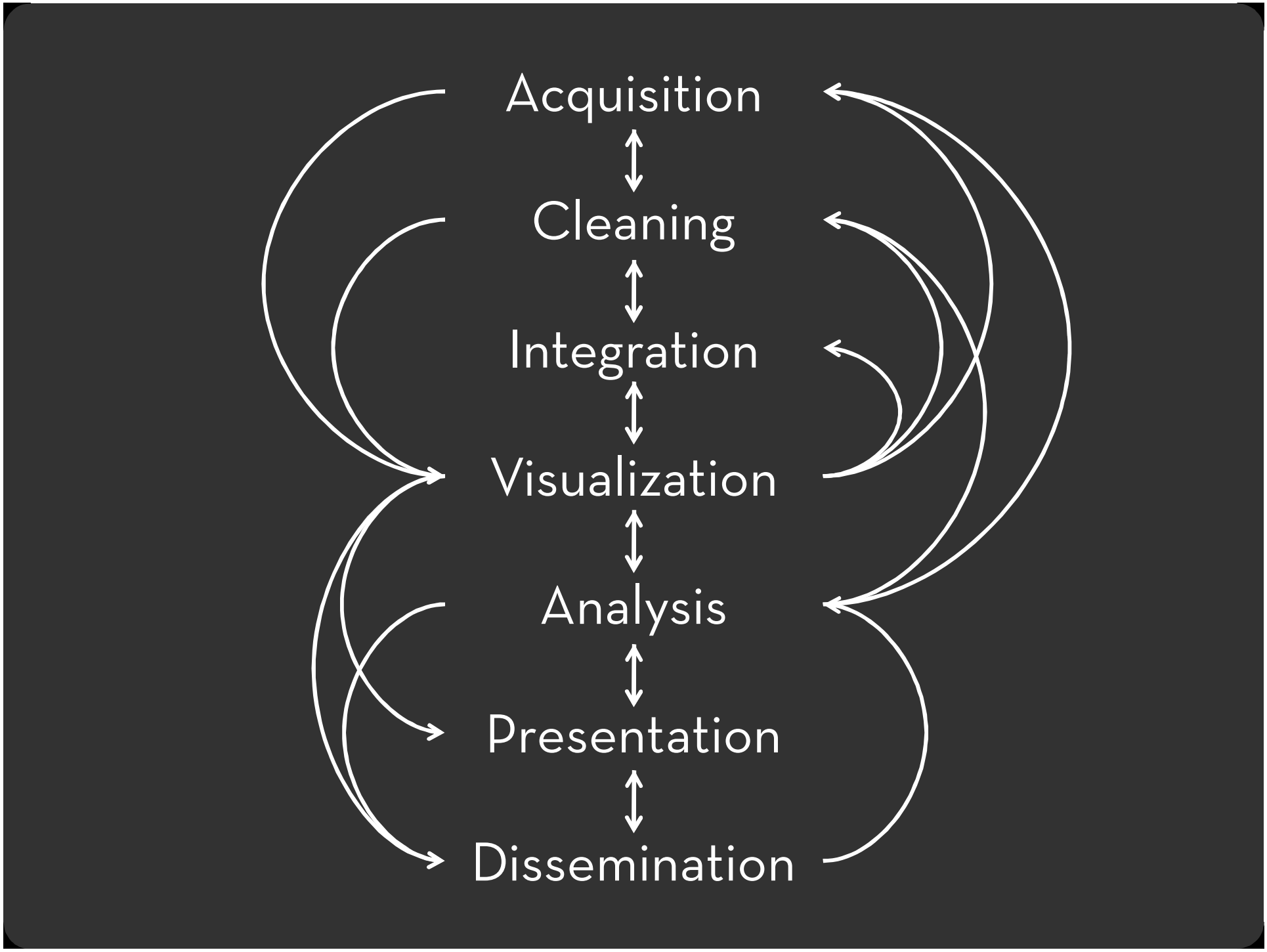
Feature prevalence from content analysis (min Cohen's κ = .74)
High co-occurrence of Observation, Question, and Hypothesis

Content Analysis of Comments



16% of sense.us comments and **10%** of Many-Eyes comments reference *data integrity* issues.


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination

Students & Collaborators

Mike Bostock

Jason Chuang

Sean Kandel

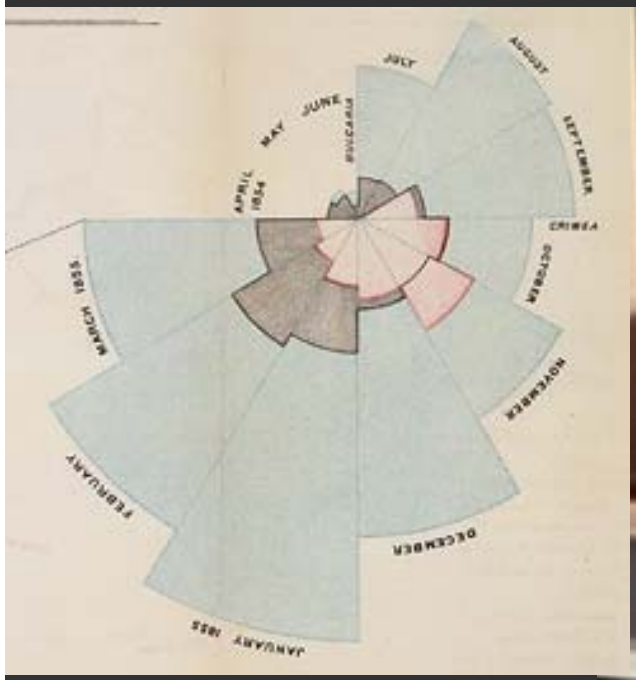
Diana MacLean

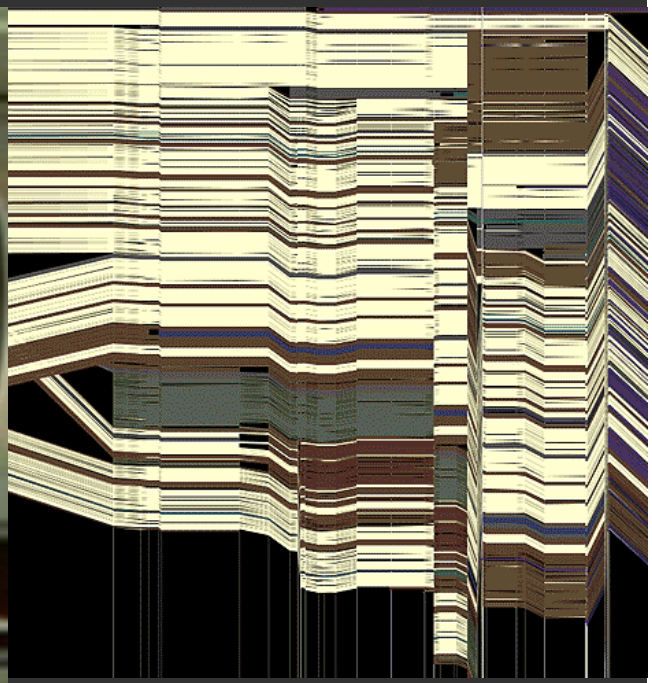
Vadim Ogievetsky

Joe Hellerstein, Andreas Paepcke

Fernanda Viégas, Martin Wattenberg

Interactive Tools for Data Transformation & Visualization





Jeffrey Heer http://vis.stanford.edu