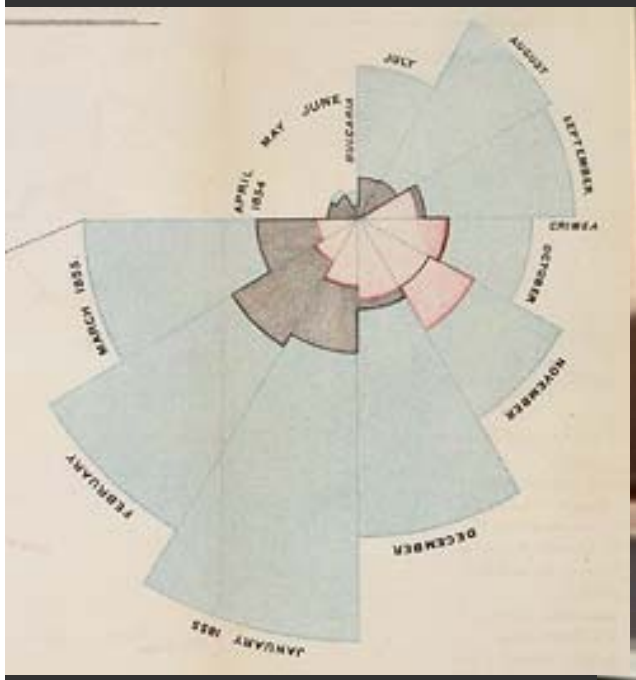
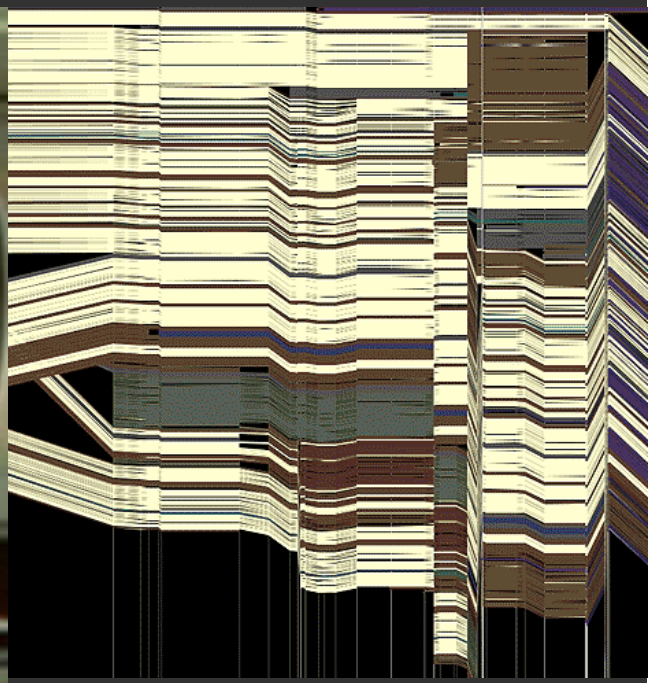
Interactive Tools for Data Transformation & Visualization





Jeffrey Heer Stanford University

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics

u_X = 9.0 σ_X = 3.317

u_Y = 7.5 σ_Y = 2.03

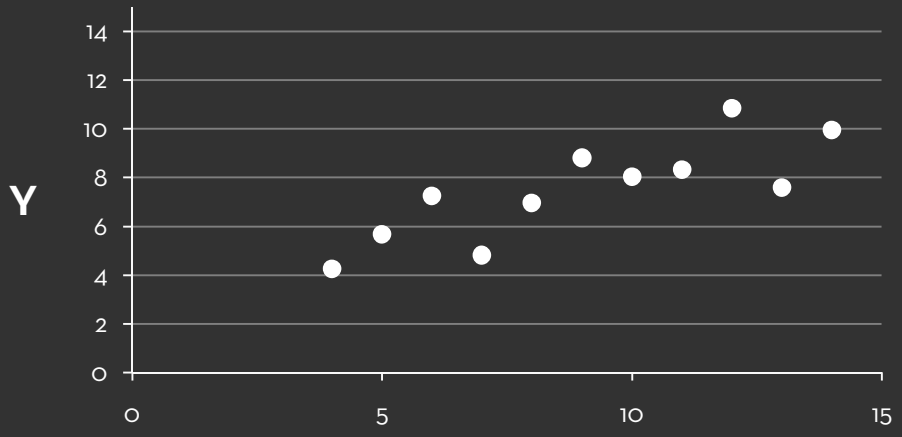
Linear Regression

Y^2 = 3 + 0.5 X

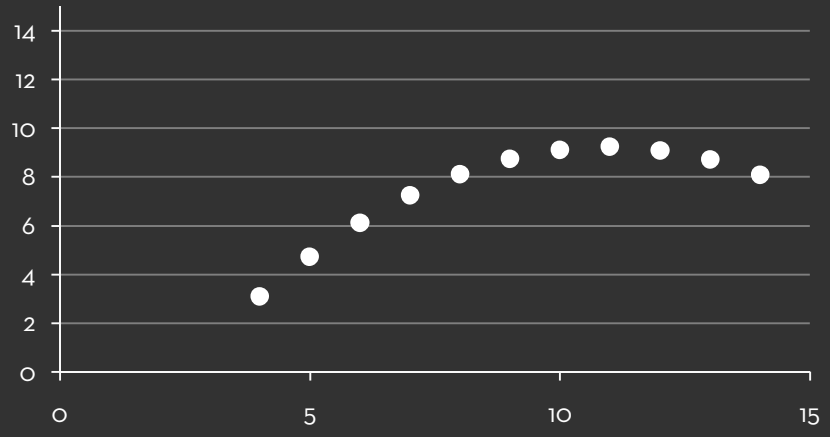
R^2 = 0.67

[Anscombe 73]

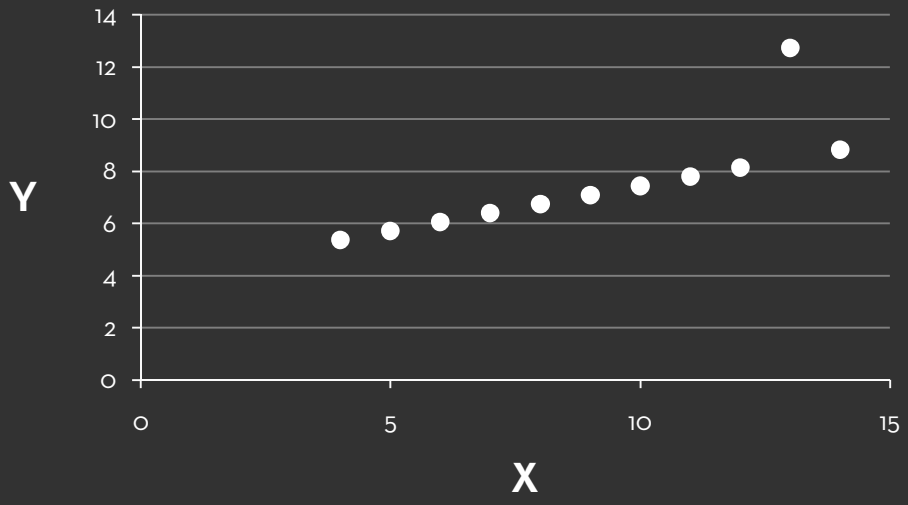
Set A



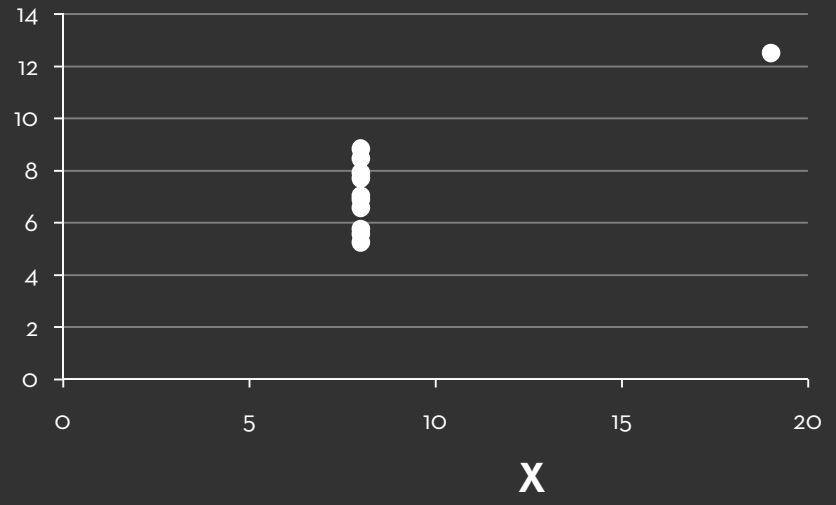
Set B

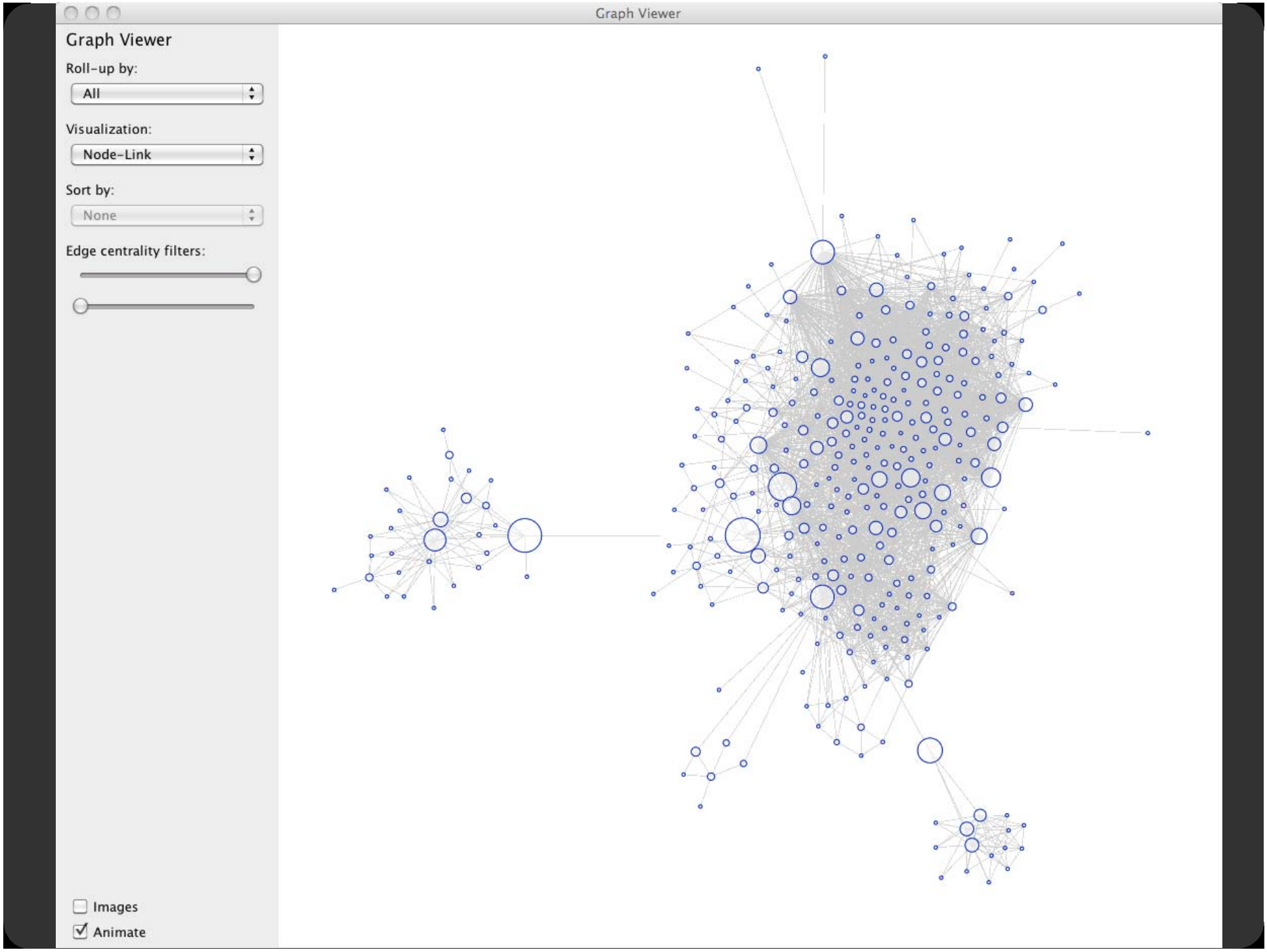


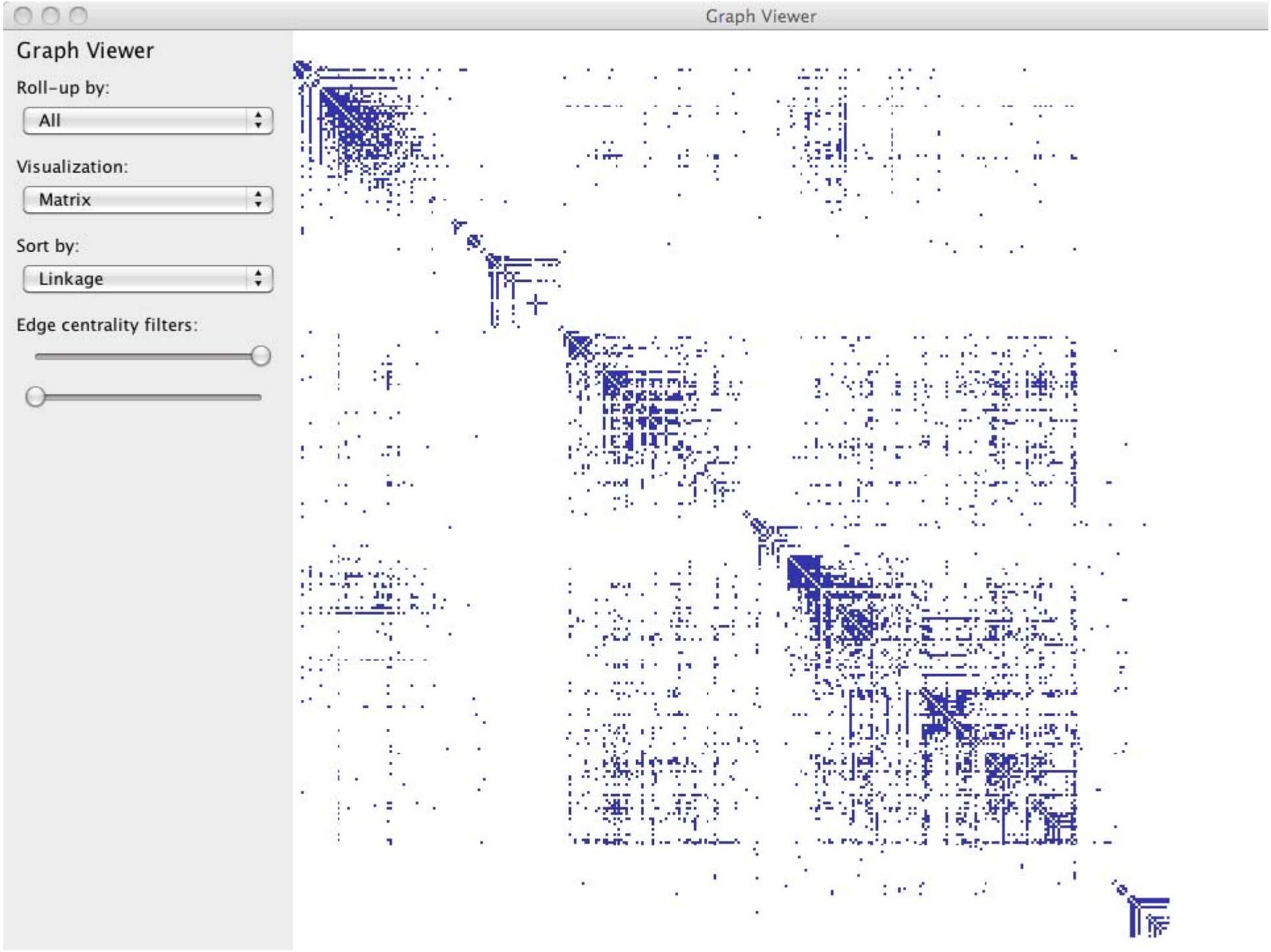
Set C

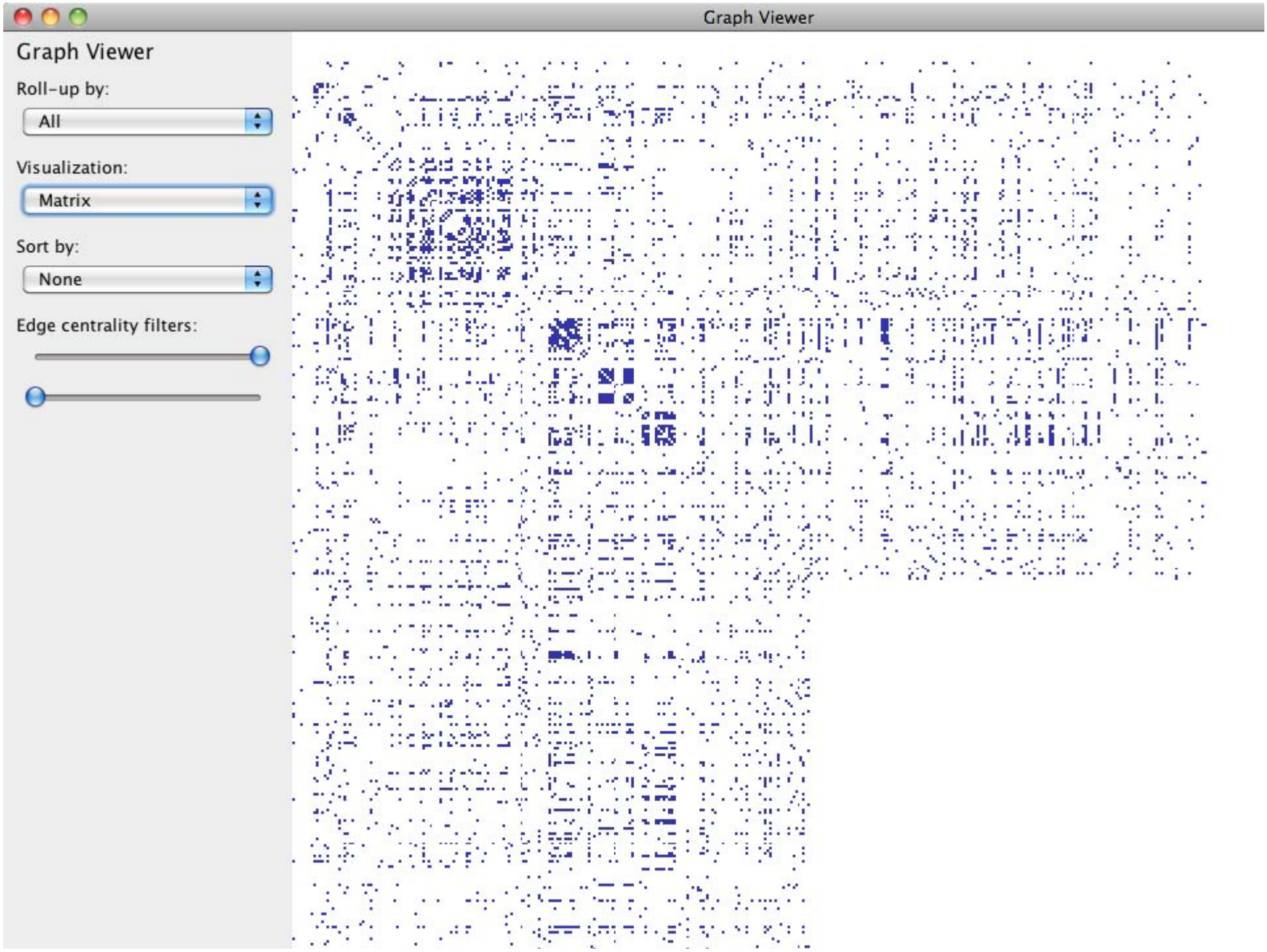


Set D




Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
None
Edge centrality filters:
Images
Animate


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
Linkage
Edge centrality filters:


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
None
Edge centrality filters:

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

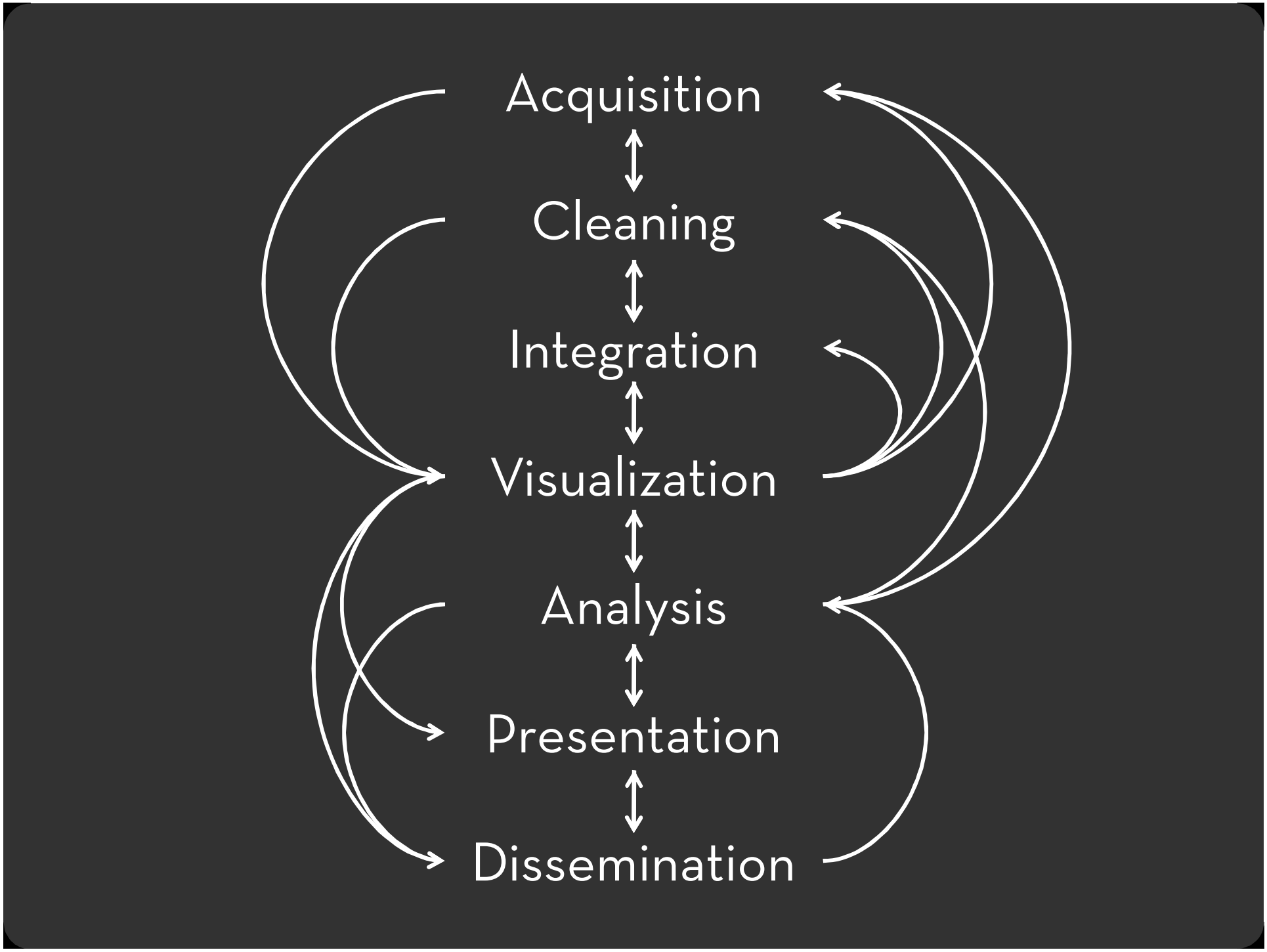
2010: 1,200 exabytes

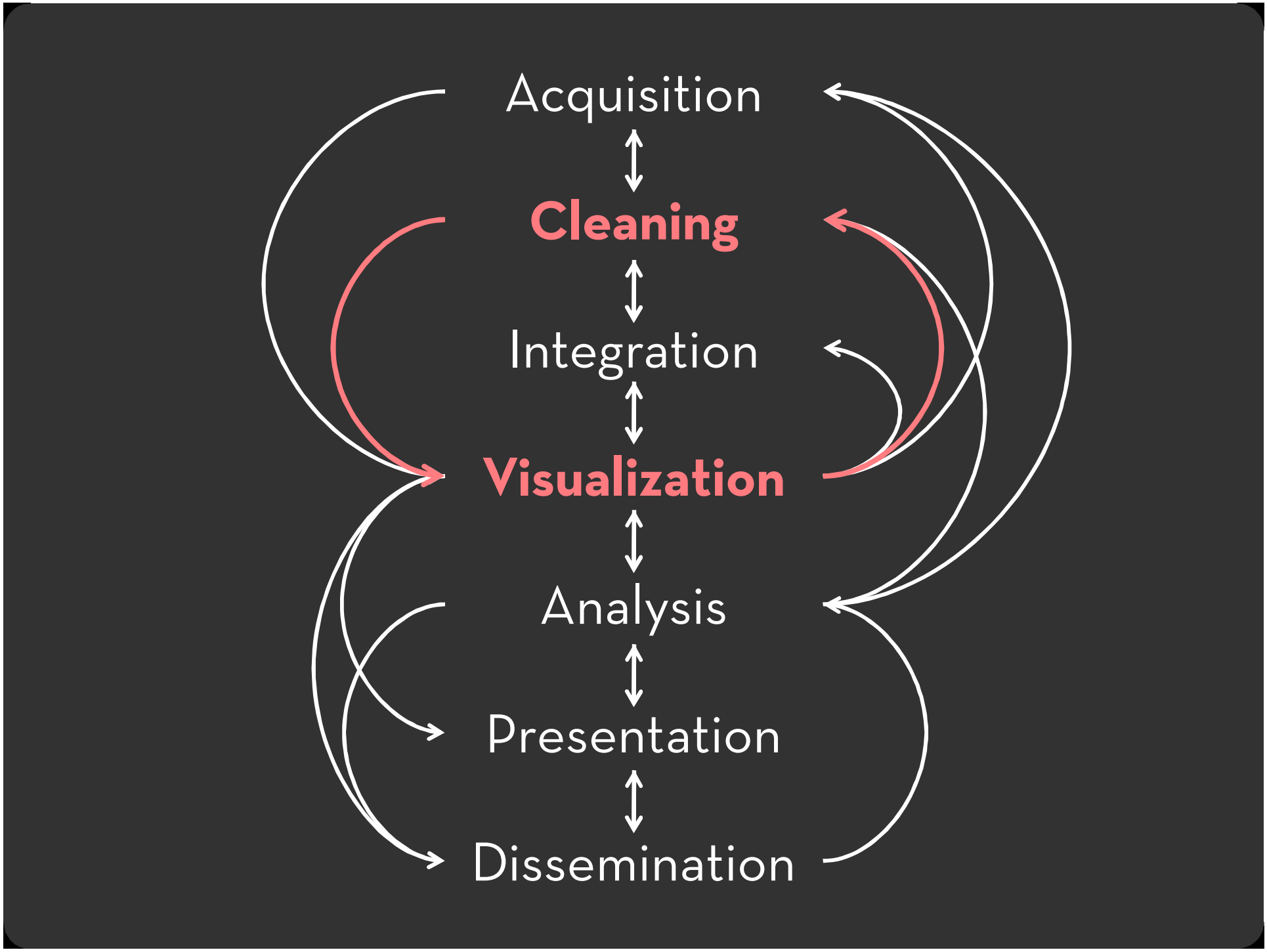
10x increase over 5 years

Gantz et al, 2008, 2010

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, *The McKinsey Quarterly*, Jan 2009


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination

Data Wrangler

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

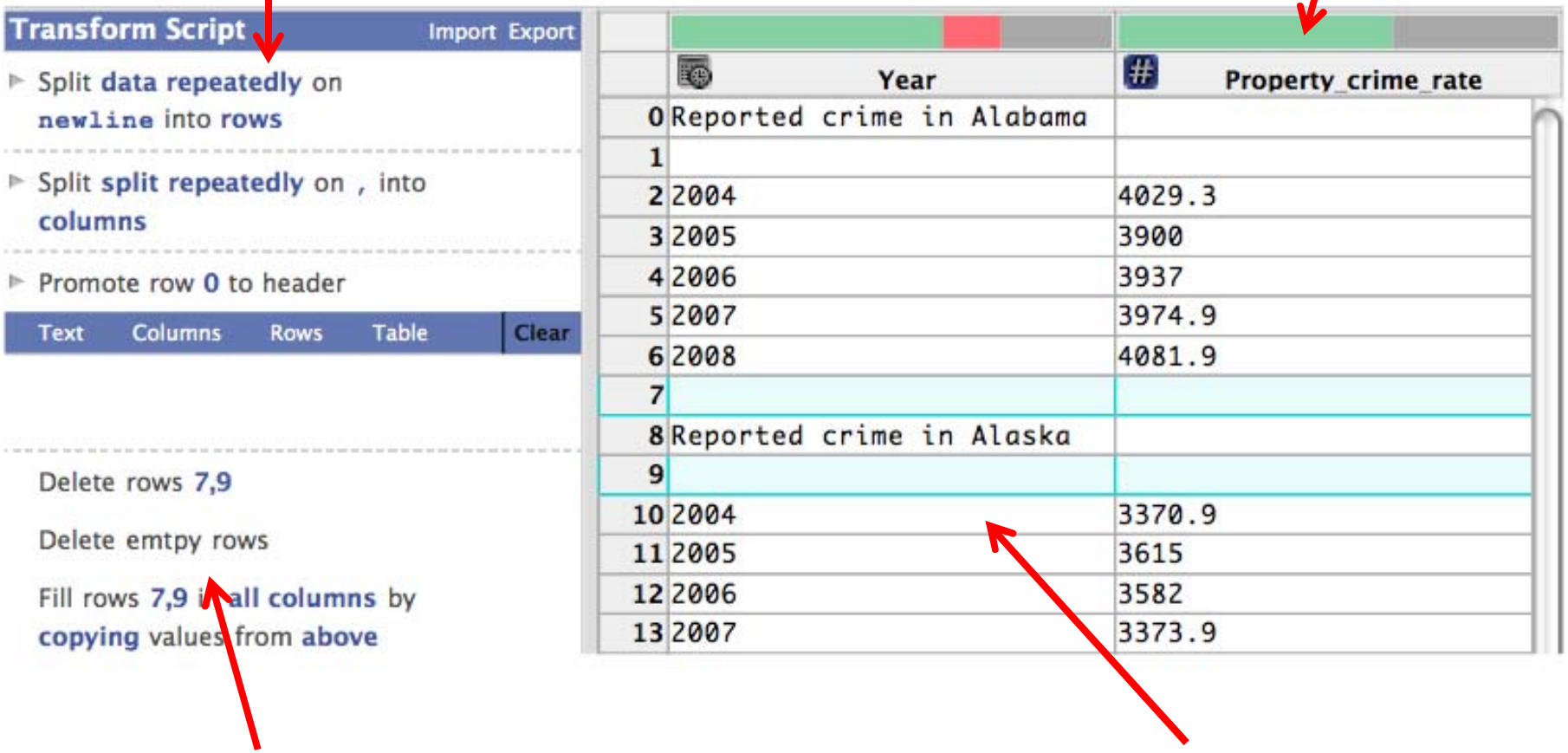
+

Mixed-initiative interface for data transforms

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

Transform History

Data Quality Meter



Suggested Transforms

Interactive Data Table

Generated Wrangler Script

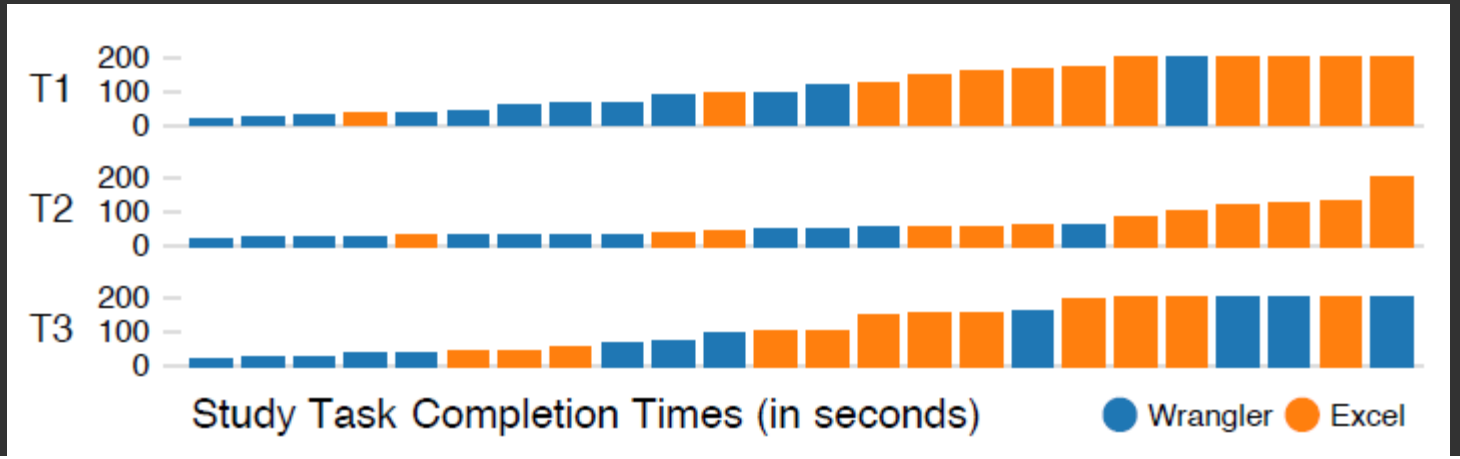
```
split('data').on(NEWLINE).max_splits(NO_MAX)
split('split').on(COMMA).max_splits(NO_MAX)
columnName().row(0)
delete(isEmpty())
extract('Year').on(/.*/).after(/in /)
fill('State').method(COPY).direction(DOWN)
delete('Year starts with "Reported crime in"')
columnName('extract').to('State')
```

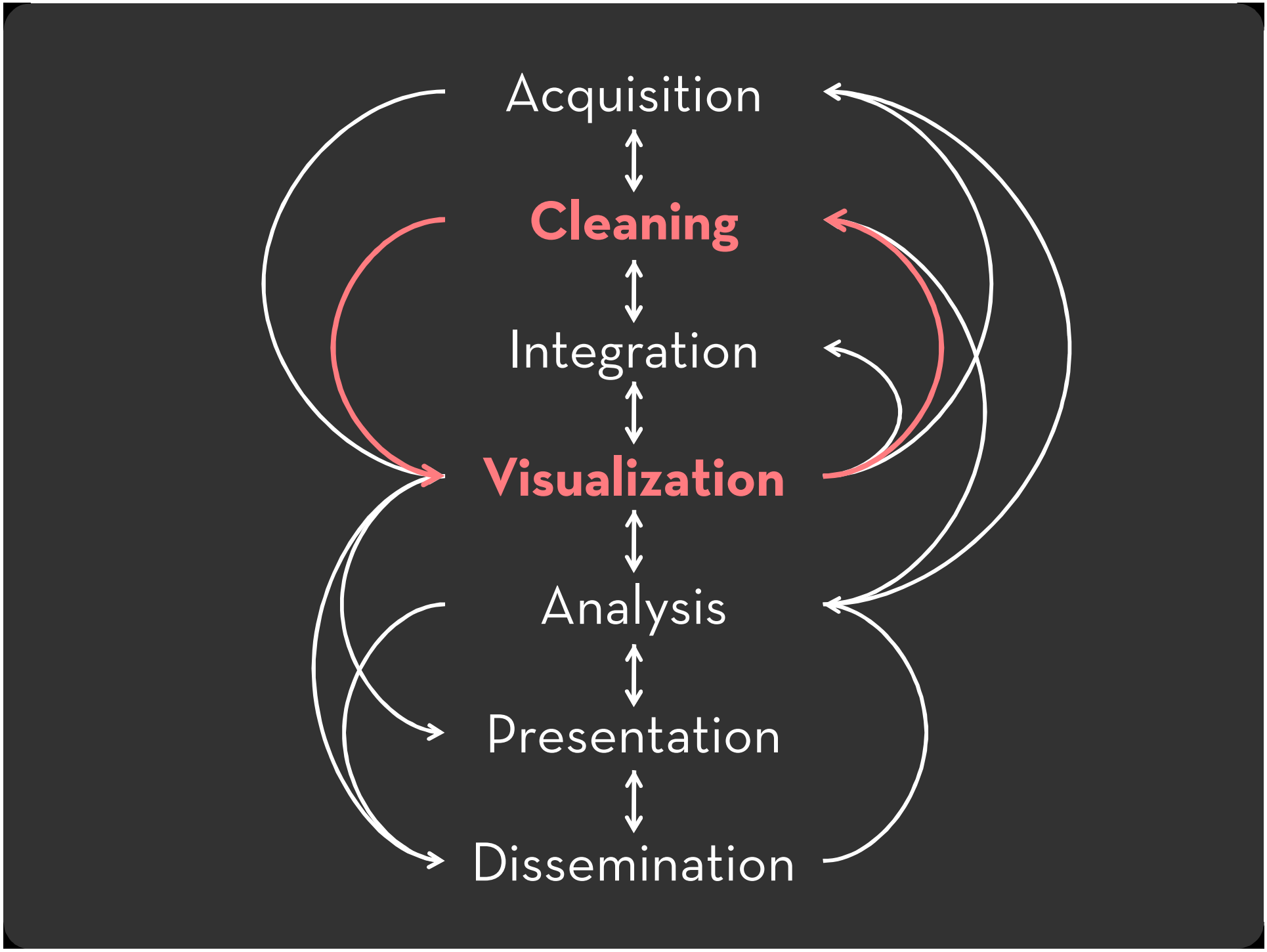
Evaluation

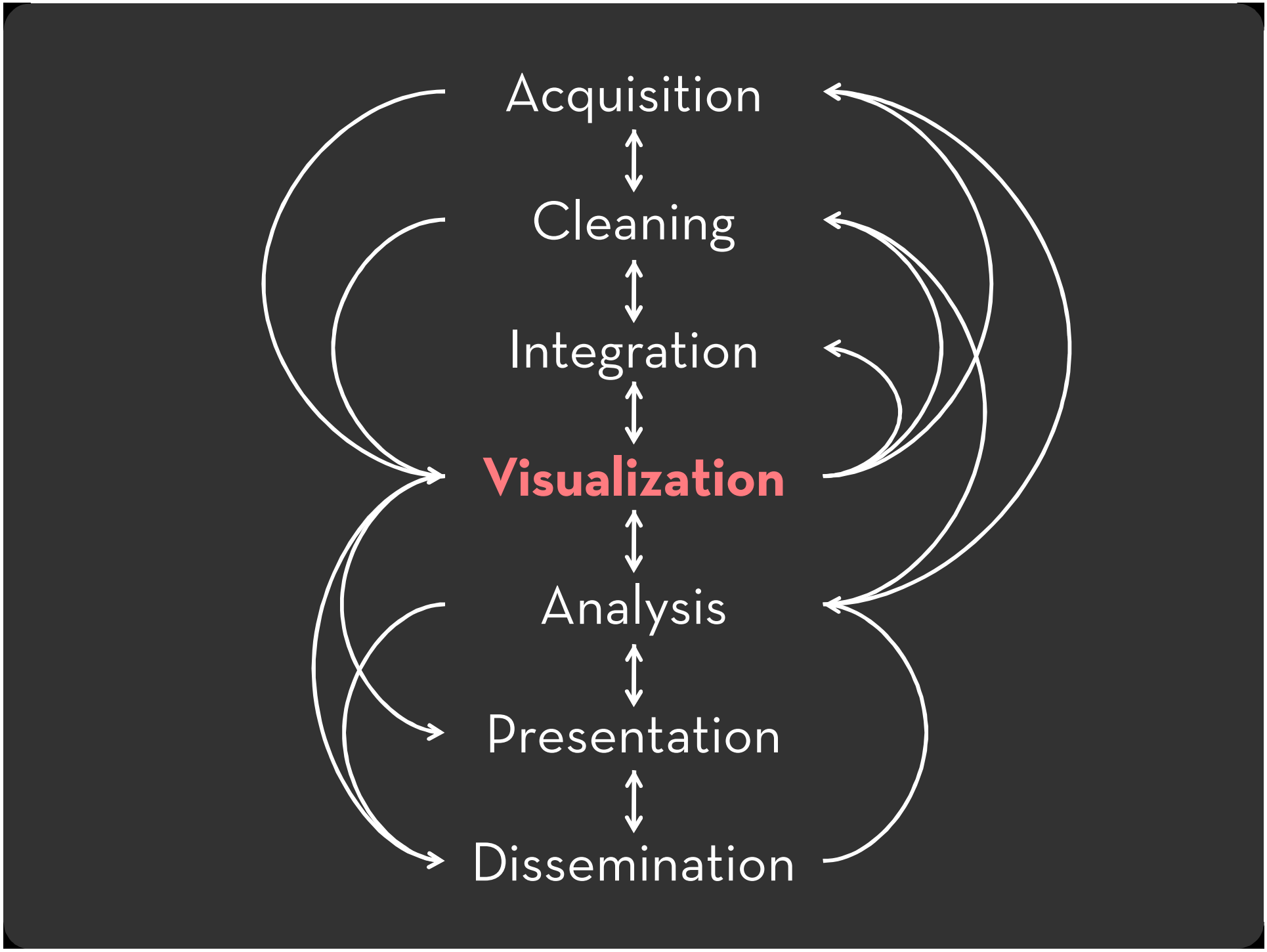
Compared Wrangler performance to Excel with 3 data cleaning tasks with a small data set.

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

Skilled Excel users benefitted the most!




Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination

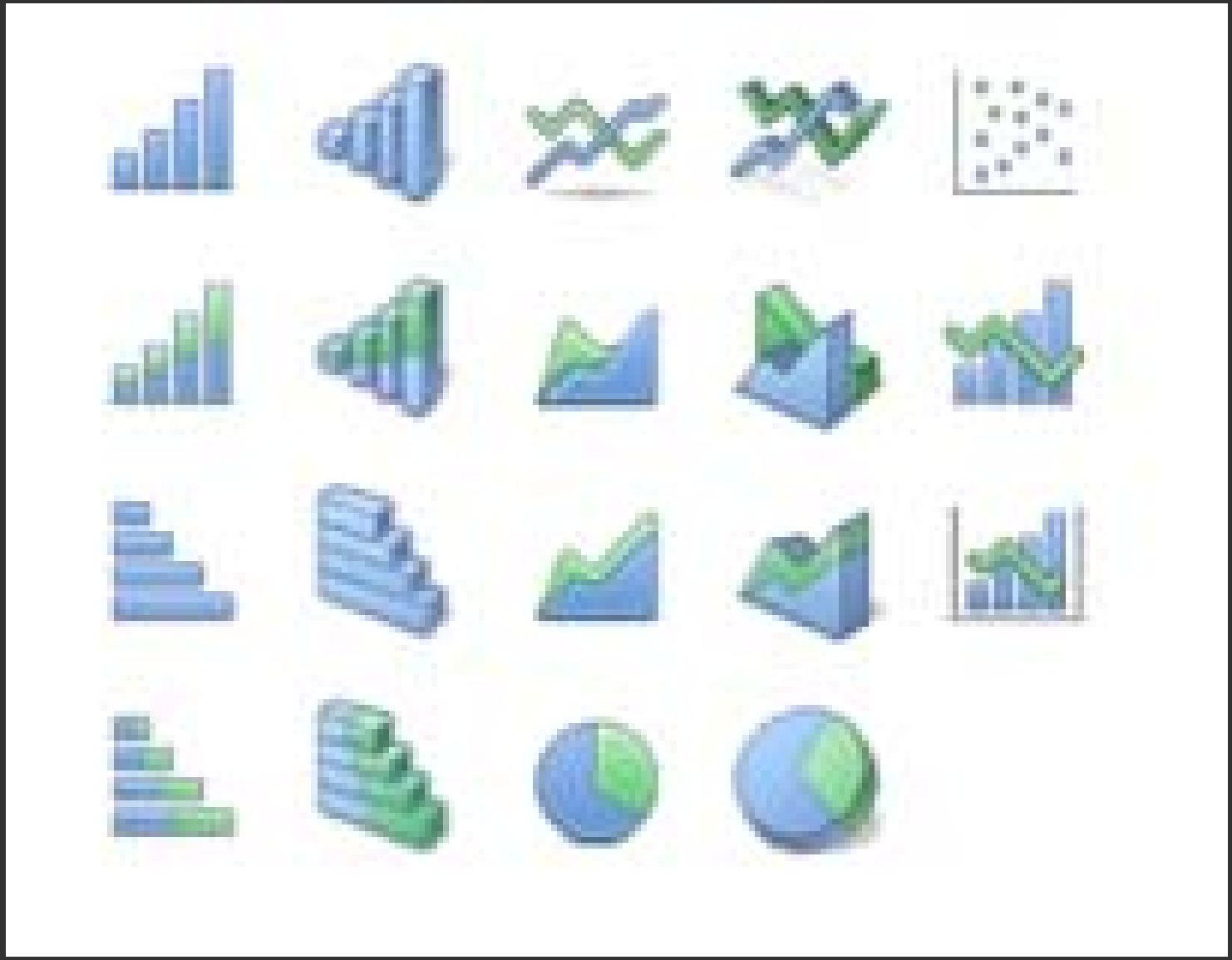


Chart Typology

Data Sets : State Quick Facts

Uploaded By: zinggoat
Data Source: US Census Bureau
Description:
Tags: people census

Created at: Friday May 18, 3:08 PM

view as text | edit data set

	People QuickFacts	Population 2005 estimate	Population percent change April 1 2000 to July 1 2005	Population 2000	Population percent change 1990 to 2000	Persons under 5 years old percent 2004	Persons under 18 years old percent 2004	Persons 65 years old and over percent 2004
1	Alabama	4557808	0.03	4447100	0.1	0.07	0.24	0.13
2	Alaska	663661	0.06	626932	0.14	0.08	0.29	0.06
3	Arizona	5939292	0.16	5130632	0.4	0.08	0.27	0.13
4	Arkansas	2779154	0.04	2673400	0.14	0.07	0.25	0.14
5	California	36132147	0.07	33871648	0.14	0.07	0.27	0.11
6	Colorado	4665177	0.08	4301261	0.31	0.07	0.26	0.1
7	Connecticut	3510297	0.03	3405565	0.04	0.06	0.24	0.14
8	Delaware	843524	0.08	783600	0.18	0.07	0.23	0.13
9	Florida	17789864	0.11	15982378	0.24	0.06	0.23	0.17
10	Georgia	9072576	0.11	8186453	0.26	0.08	0.26	0.1
11	Hawaii	1275194	0.05	1211537	0.09	0.07	0.24	0.14
12	Idaho	1429096	0.1	1293953	0.29	0.07	0.27	0.11
13	Illinois	12763371	0.03	12419293	0.09	0.07	0.26	0.12

Choosing a visualization type for State Quick Facts

Analyze a text

Tag Cloud

How are you using your words? This enhanced tag cloud will show you the words popularity in the given set of text.

Learn more

Wordle

Wordle is a toy for generating "word clouds" from text that you provide. The clouds give greater prominence to words that appear more frequently in the source text.

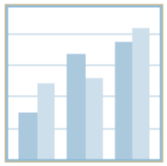
Learn more

Word Tree

See a branching view of how a word or phrase is used in a text. Navigate the text by zooming and clicking.

Learn more

Compare a set of values



Bar Chart

How do the items in your data set stack up? A bar chart is a simple and recognizable way to compare values. You can display several sets of bars for multivariate comparisons.

Learn more

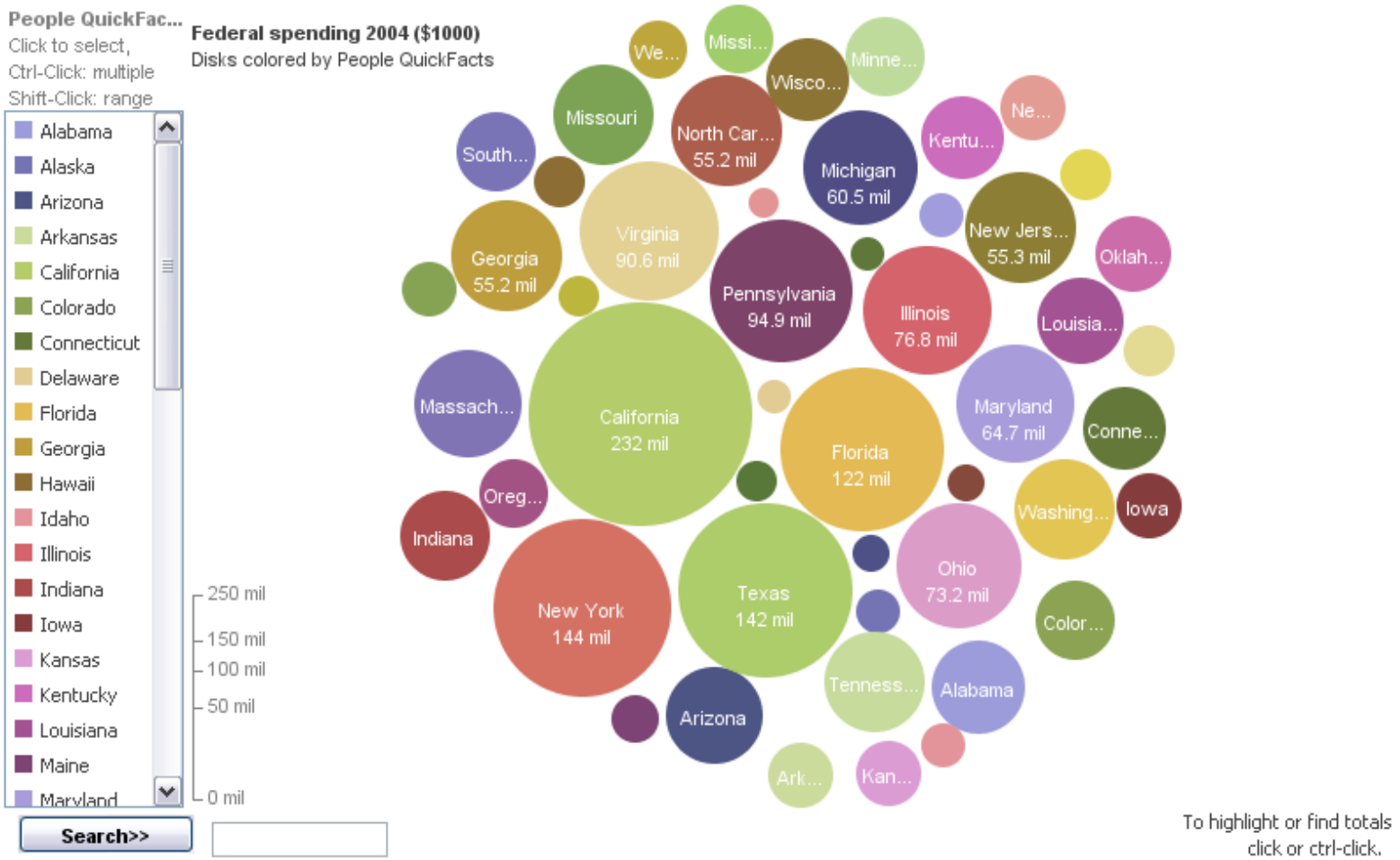
Block Histogram

This versatile chart lets you get a quick sense of how a single set of data is distributed. Each item in the data is an individually identifiable block.

Learn more

Visualizations : Federal Spending by State, 2004

Creator: Anonymous
Tags: census people



Search>>

To highlight or find totals click or ctrl-click.

Bubble Size: Federal spending 2004 ($1000) | Label: People QuickFacts | Color: People QuickFacts

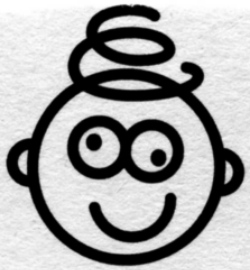
- Retail sales per capita 2002
- Minority-owned firms percent of total 1997
- Women-owned firms percent of total 1997
- Housing units authorized by building permits 2004
- Federal spending 2004 ($1000)
- Land area 2000 (square miles)
- Persons per square mile 2000
- FIPS Code

Data file ... Census Bureau

This data set has not yet been rated

full image

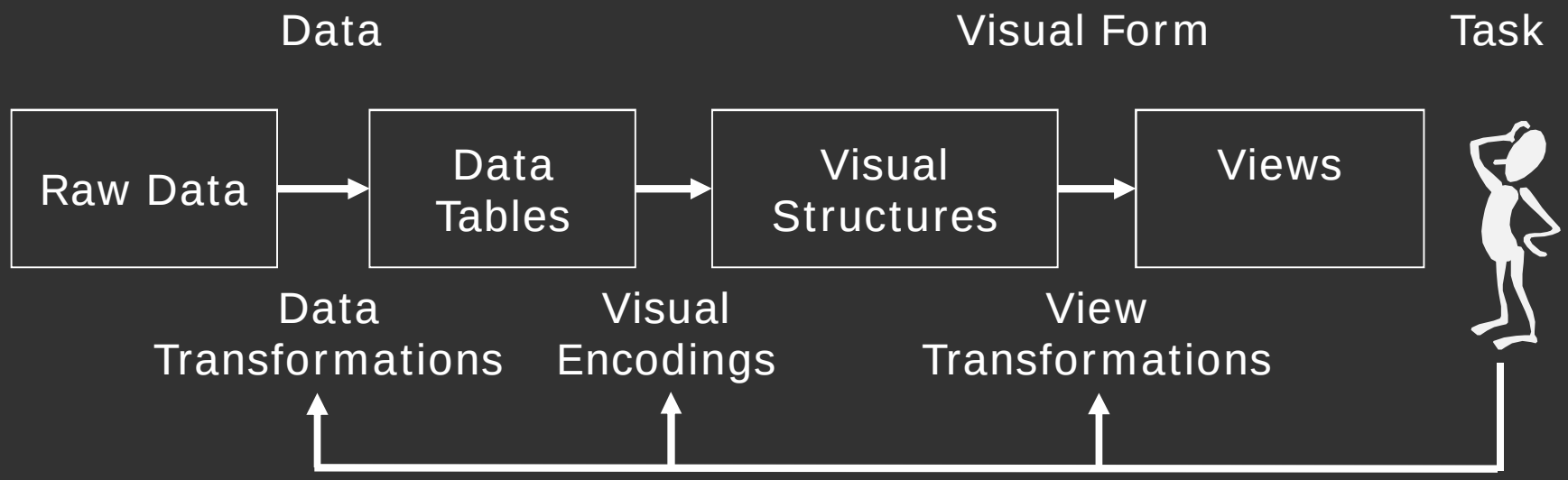
rate this



MAD LIBS®

MY MUSIC LESSON

Every Wednesday, when I get home from school, I have a piano lesson. My teacher is a very strict house (NOUN). Her name is Hillary Clinton (CELEBRITY (FEMALE)). Our piano is a Steinway Concert tree (NOUN) and it has 88 ~~[illegible]~~ cups (PLURAL NOUN). It also has a soft pedal and a/an Smily (ADJECTIVE) pedal. When I have a lesson, I sit down on the piano AIBERTO (NOUN) and play for 16 Minuts (PERIOD OF TIME). I do scales to exercise my cats (PLURAL NOUN), and then I usually play a minuet by Johann Sebastian Washington (CELEBRITY (LAST NAME)). Teacher says I am a natural Haunted House (NOUN) and have a good musical leg (PART OF THE BODY). Perhaps when I get better I will become a concert vet (PROFESSION) and give a recital at Carnegie hospital (TYPE OF BUILDING).


Data
Visual Form
Task
Raw Data
Data Tables
Visual Structures
Views
Data Transformations
Visual Encodings
View Transformations

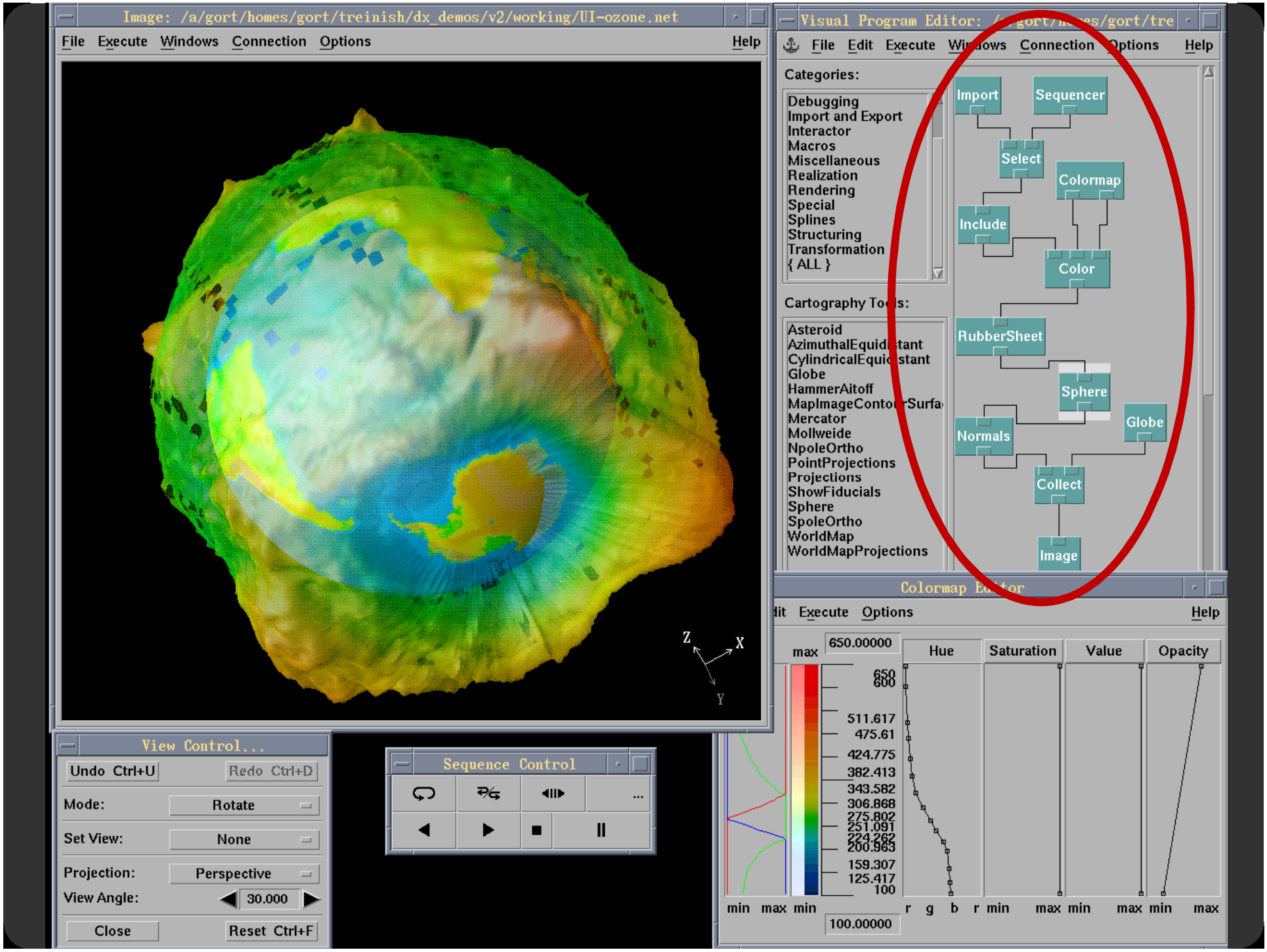
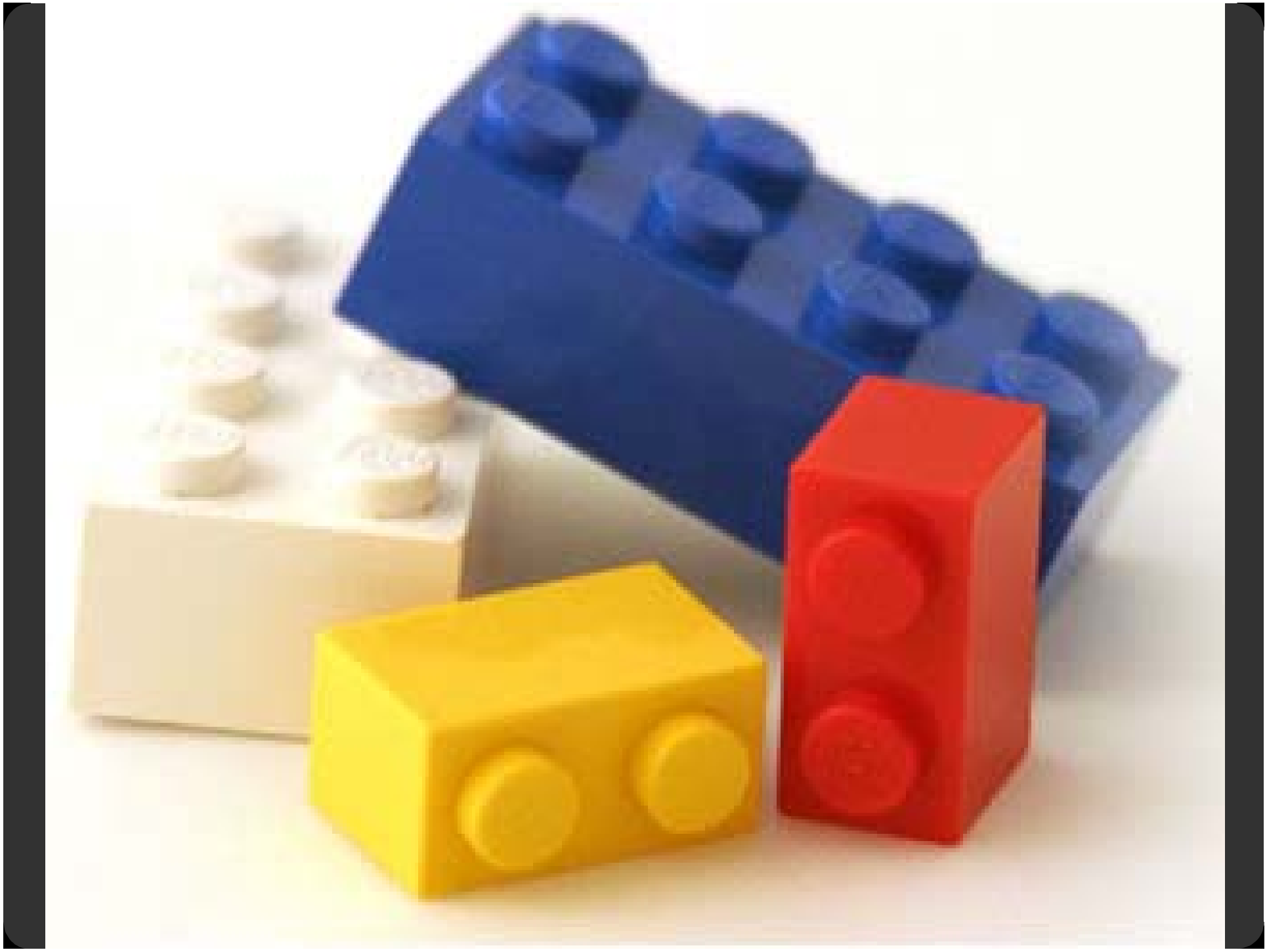


Image: /a/gort/homes/gort/treinish/dx_demos/v2/working/UI-ozone.net
File Execute Windows Connection Options Help
Z
X
Y
Visual Program Editor:
File Edit Execute Windows Connection Options Help
Categories:
Debugging
Import and Export
Interactor
Macros
Miscellaneous
Realization
Rendering
Special
Splines
Structuring
Transformation
{ ALL }
Cartography Tools:
Asteroid
AzimuthalEquidistant
CylindricalEquidistant
Globe
HammerAitoff
MapImageContourSurface
Mercator
Mollweide
NpoleOrtho
PointProjections
Projections
ShowFiducials
Sphere
SpoleOrtho
WorldMap
WorldMapProjections
Import
Sequencer
Select
Colormap
Include
Color
RubberSheet
Sphere
Globe
Normals
Collect
Image
Colormap Editor
Execute Options Help
max
650.00000
650
600
511.617
475.61
424.775
382.413
343.582
306.868
275.802
251.091
224.262
200.363
159.307
125.417
100
100.00000
Hue
Saturation
Value
Opacity
min max min r g b r min max min max min max
View Control...
Undo Ctrl+U
Redo Ctrl+D
Mode:
Rotate
Set View:
None
Projection:
Perspective
View Angle:
30.000
Close
Reset Ctrl+F
Sequence Control

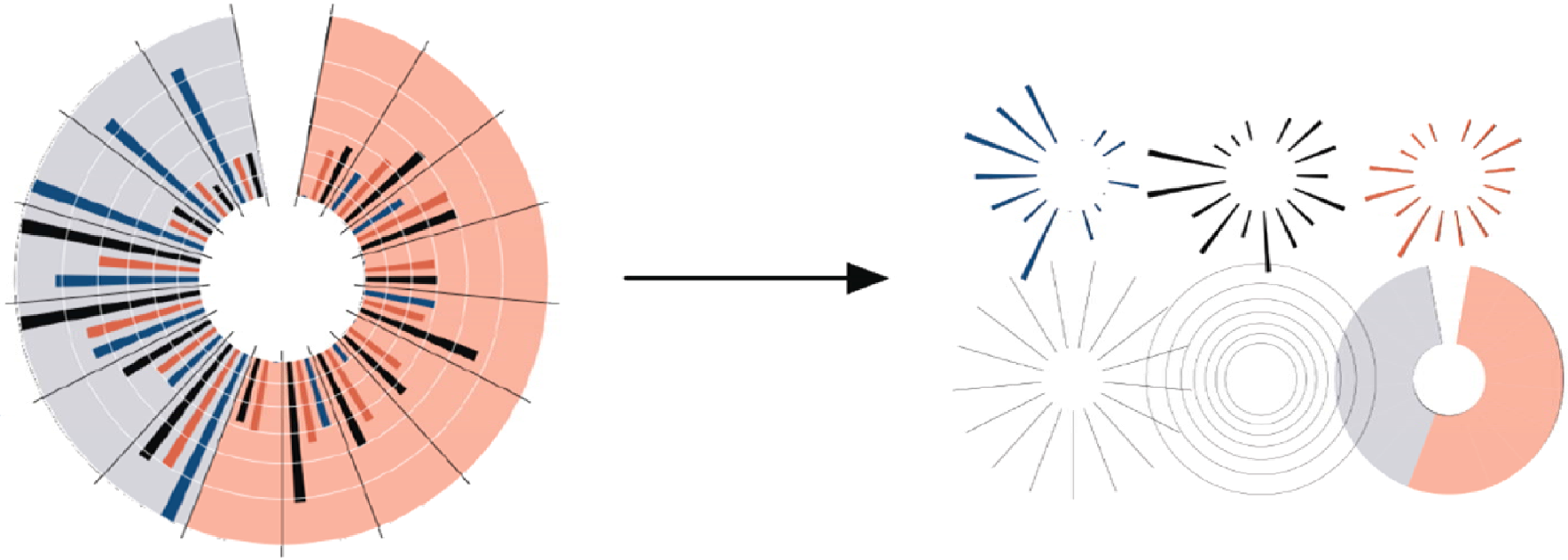


Today's first task is not to invent wholly new [*graphical*] techniques, though these are needed. Rather we need most vitally to recognize and reorganize the essential of old techniques, to make easy their assembly in new ways, and to modify their external appearances to fit the new opportunities.

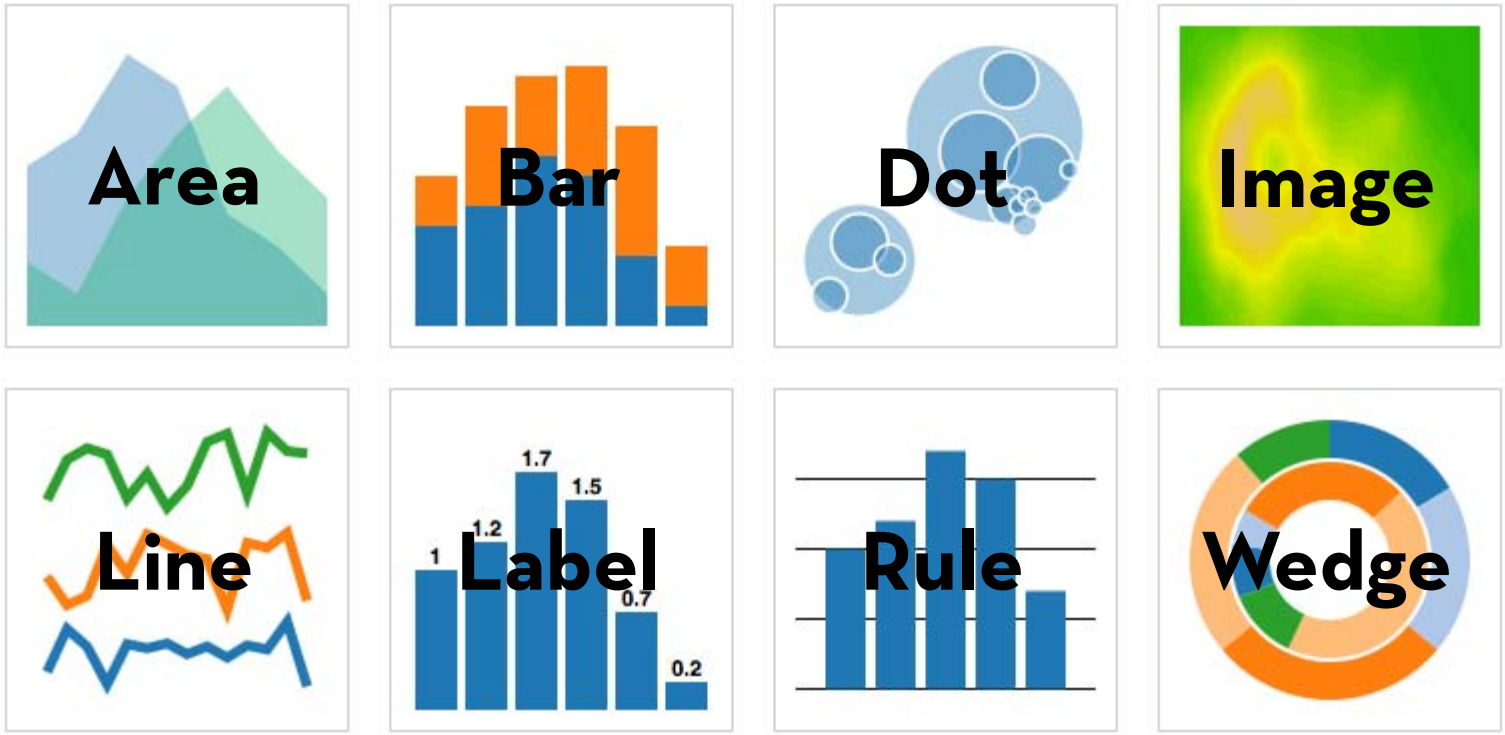
J. W. Tukey, *The Future of Data Analysis*, 1962.

Protovis: A Declarative Language for Visualization

http://protovis.org/

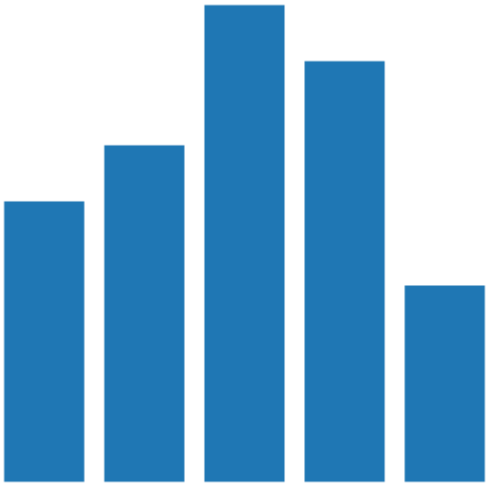


A graphic is a composition of data-representative marks.


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

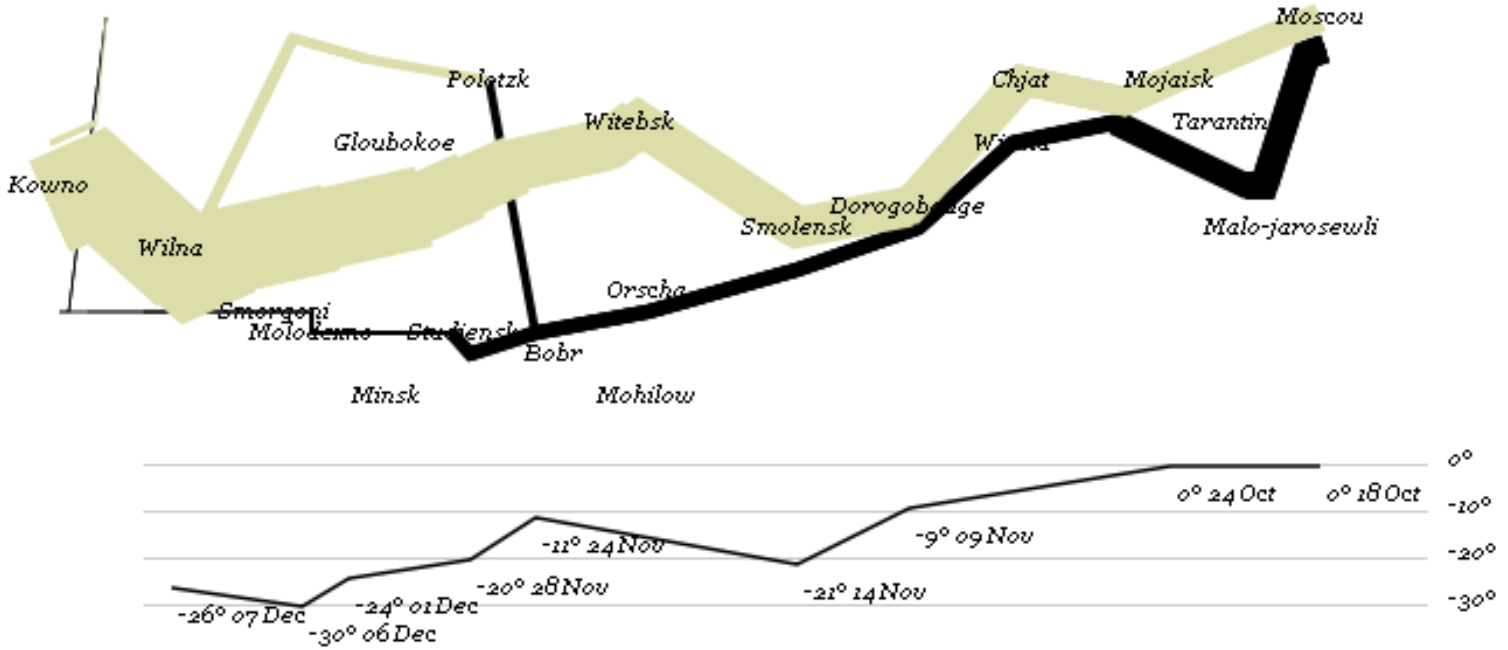
Protovis

Create customized visualizations using a declarative specification language.



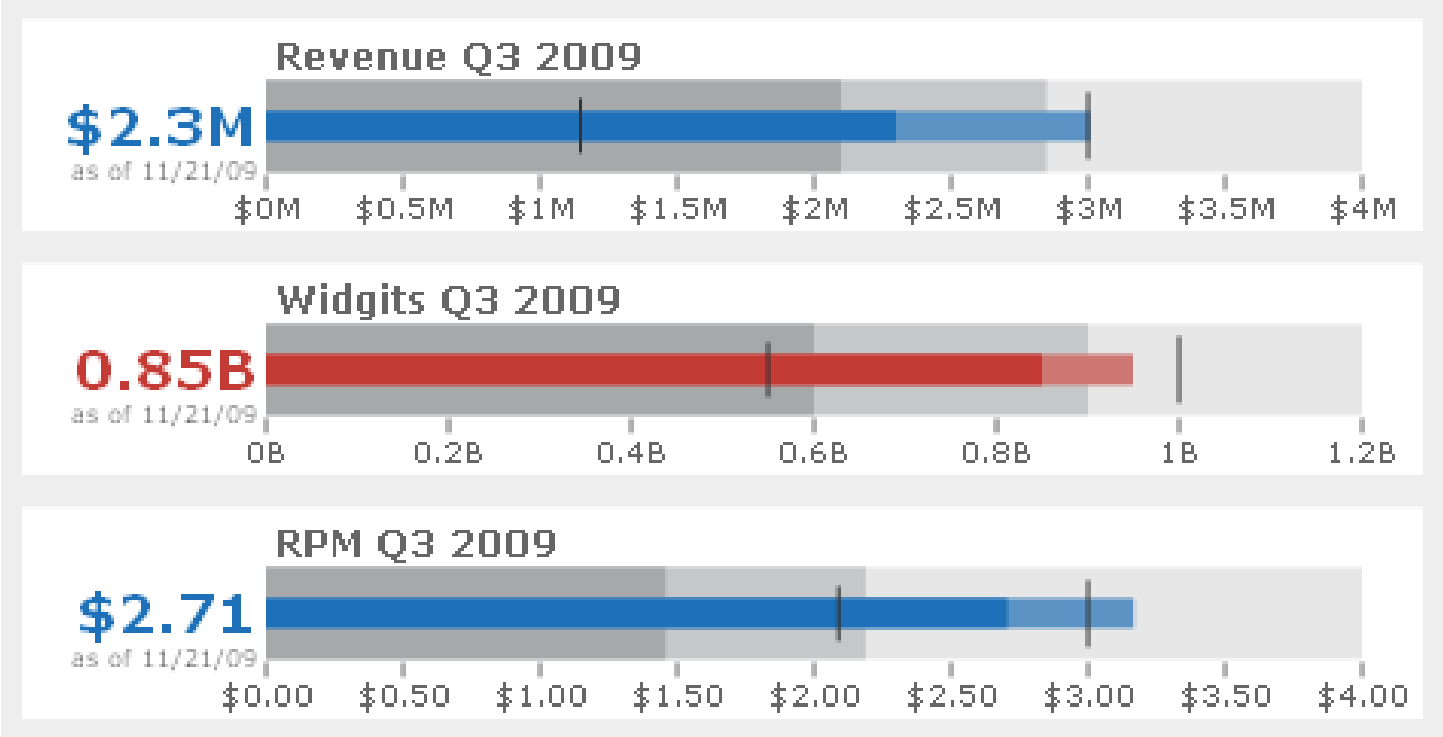
```
var vis = new pv.Panel();
vis.add(pv.Bar)
  .data([1, 1.2, 1.7, 1.5, .7])
  .bottom(10).width(20)
  .height(function(d) d * 70)
  .left(function()  this.index * 25 + 20);
vis.render();
```

Protovis (protovis.org) – JavaScript Visualization Tools

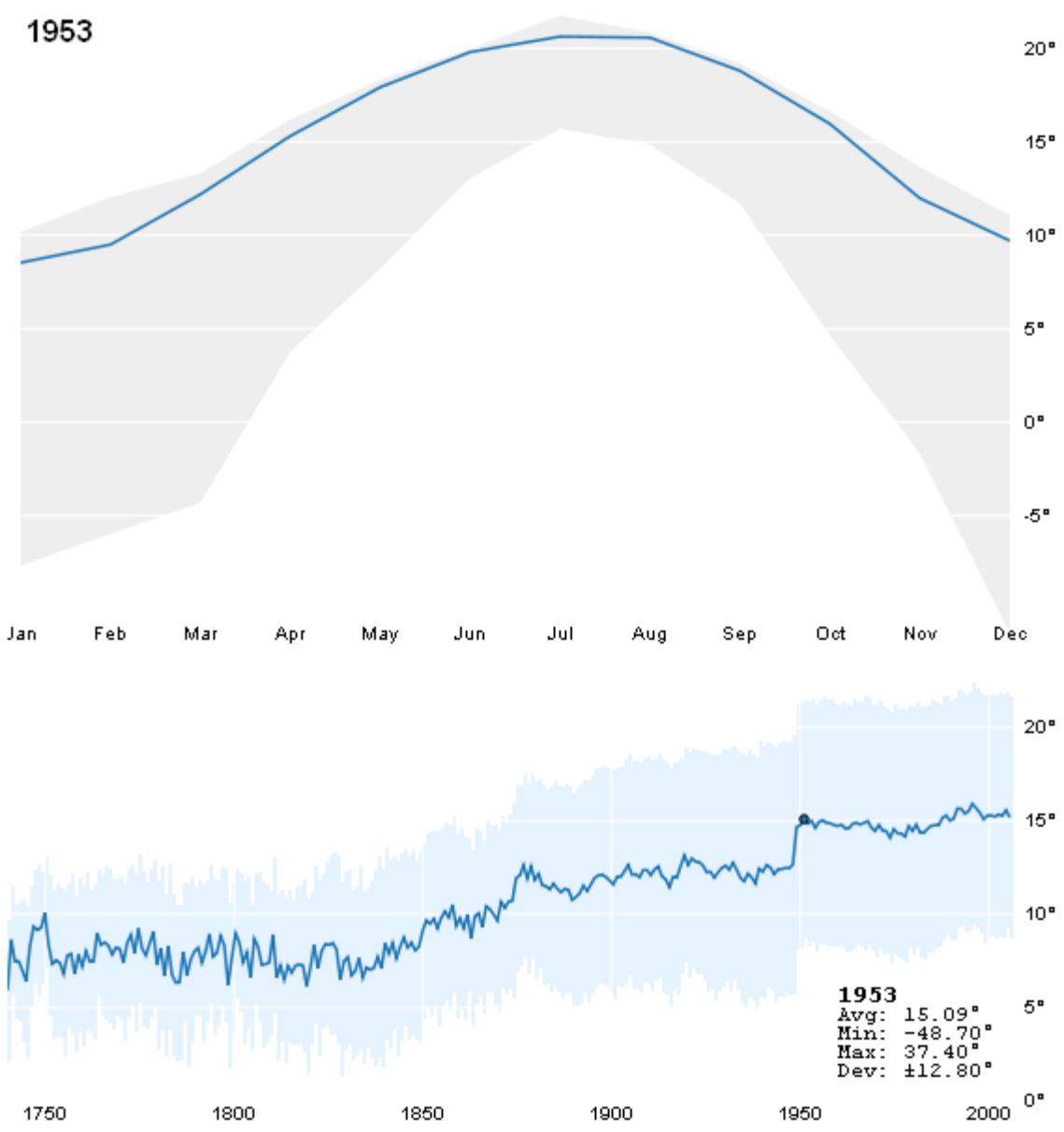



```
var army = pv.nest(napoleon.army, "dir", "group“);
var vis = new pv.Panel();

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");
```

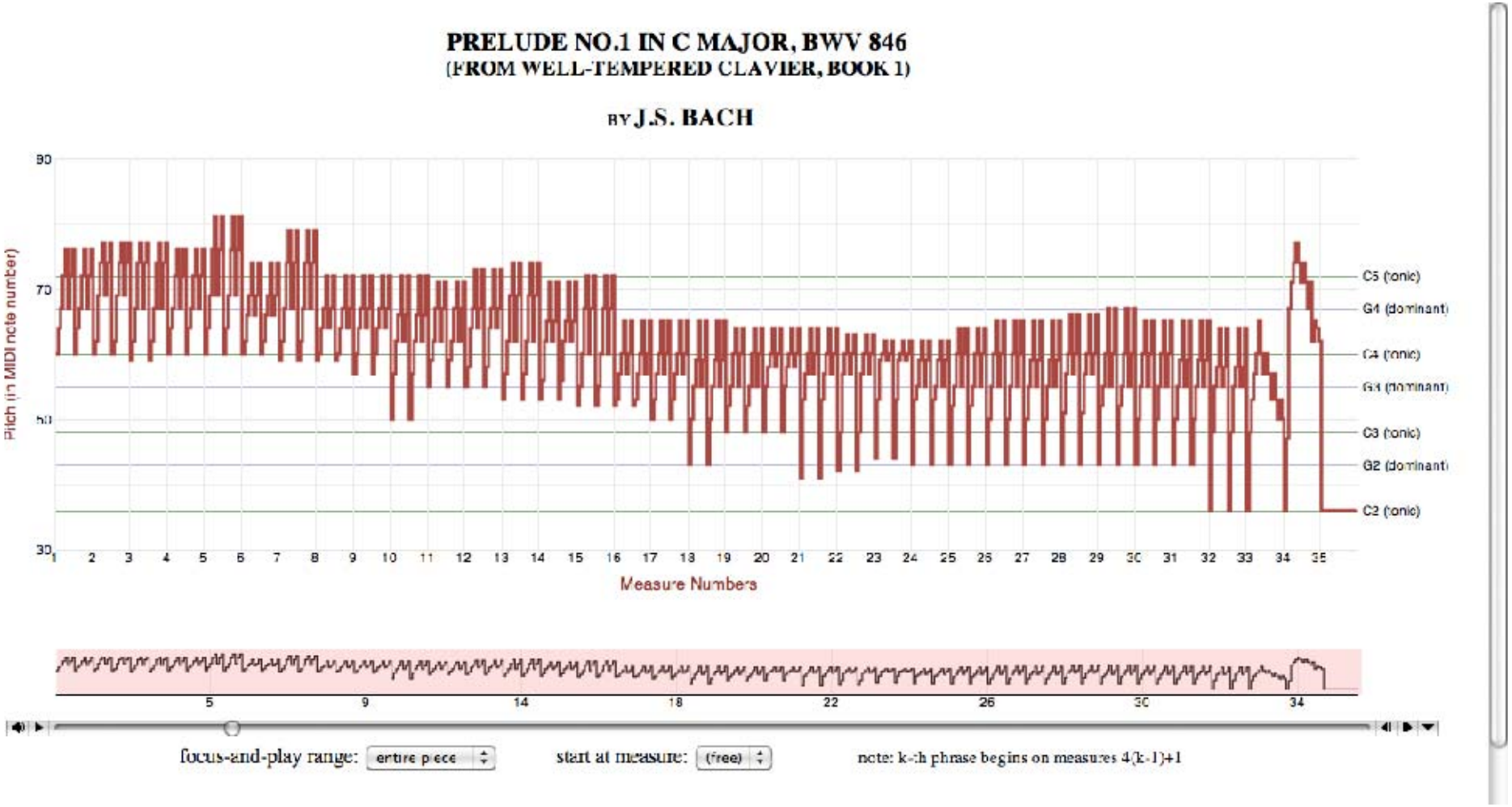
```
vis.add(pv.Rule).data([0,-10,-20,-30])
  .top(function(d) 300 - 2*d - 0.5).left(200).right(150)
  .lineWidth(1).strokeStyle("#ccc")
  .anchor("right").add(pv.Label)
    .font("italic 10px Georgia")
    .text(function(d) d+"°").textBaseline("center");

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");
```



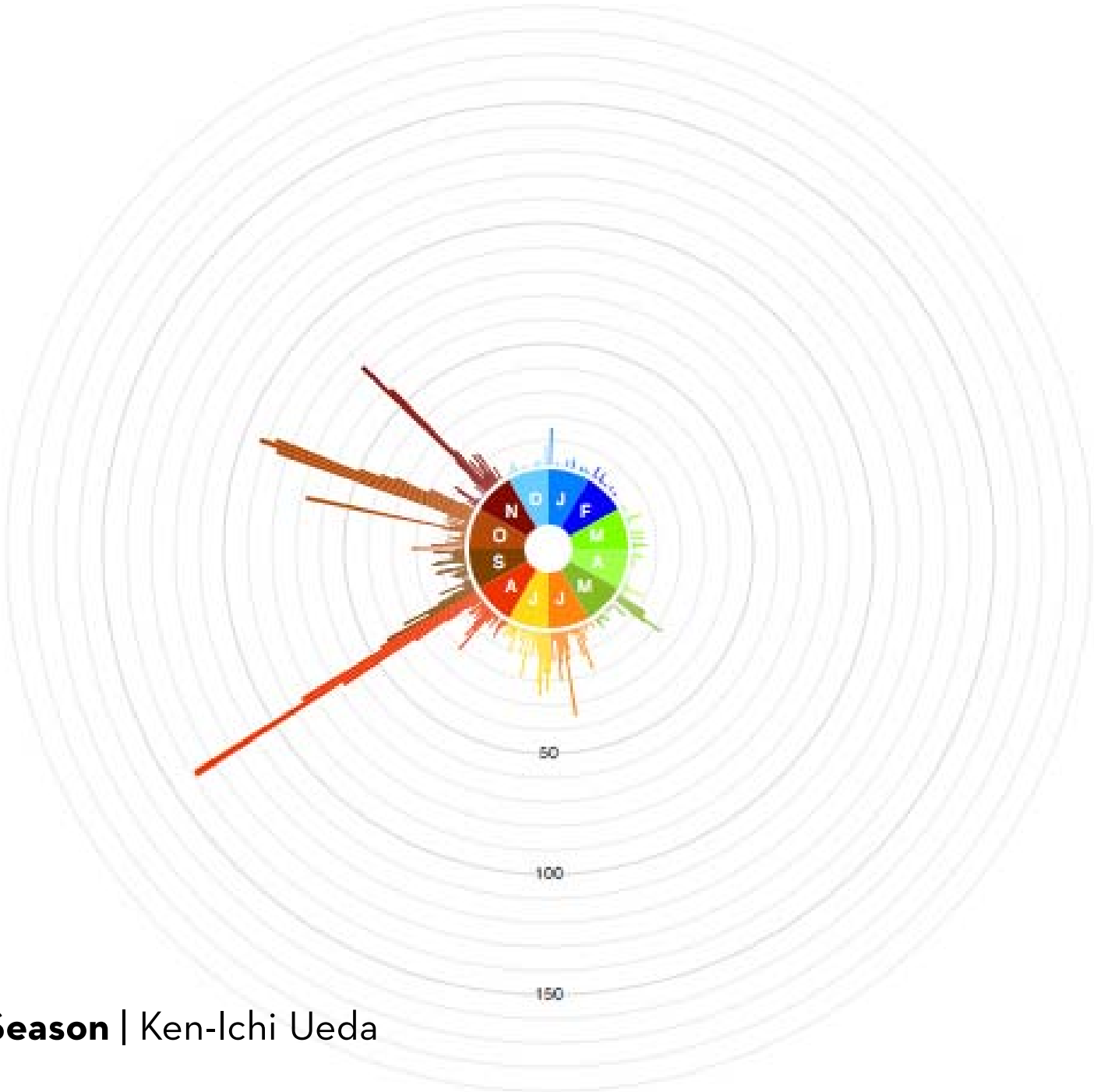
Bullet Charts | Clint Ivy



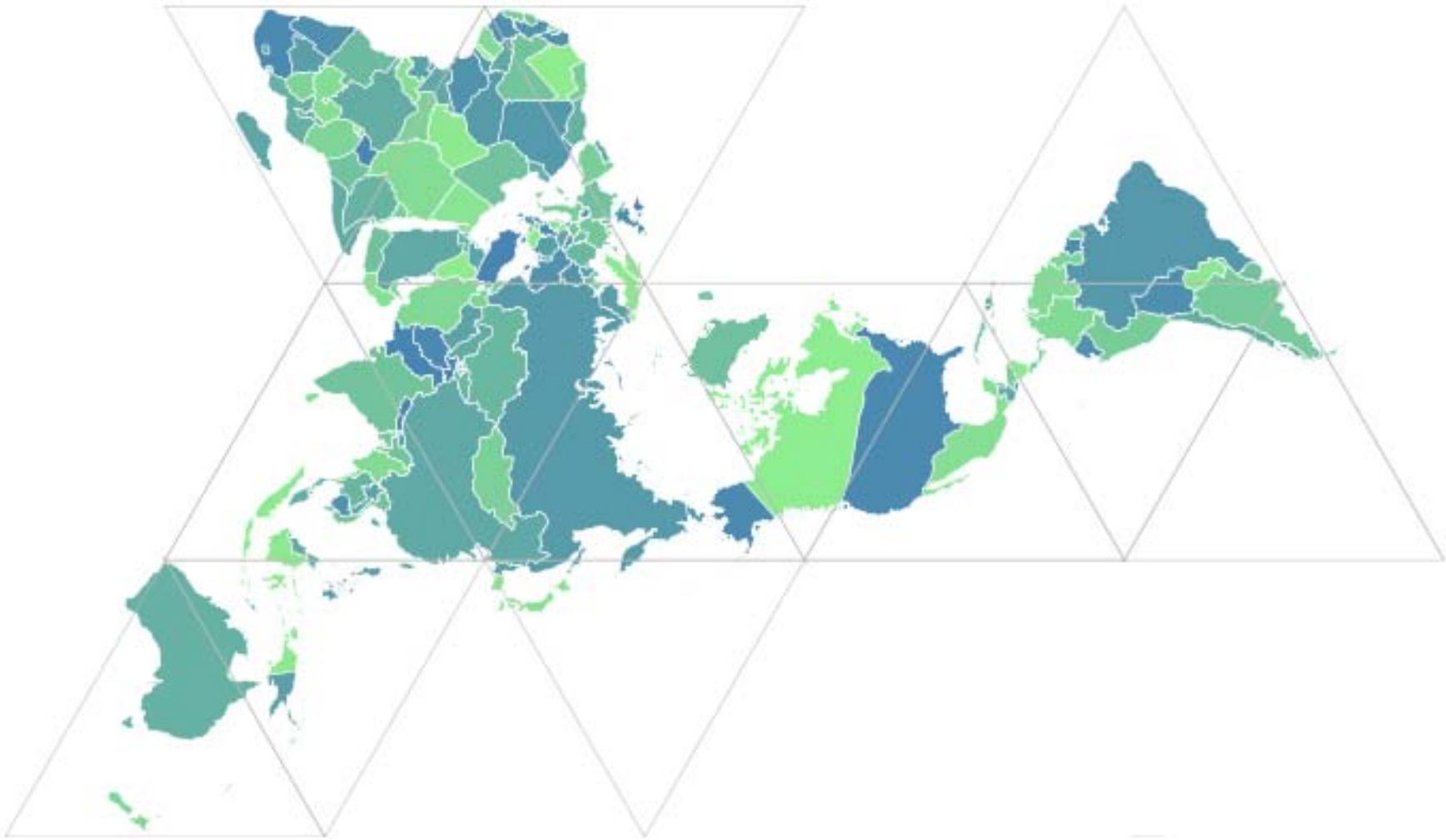
Climate Graph | Robert Kosara



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

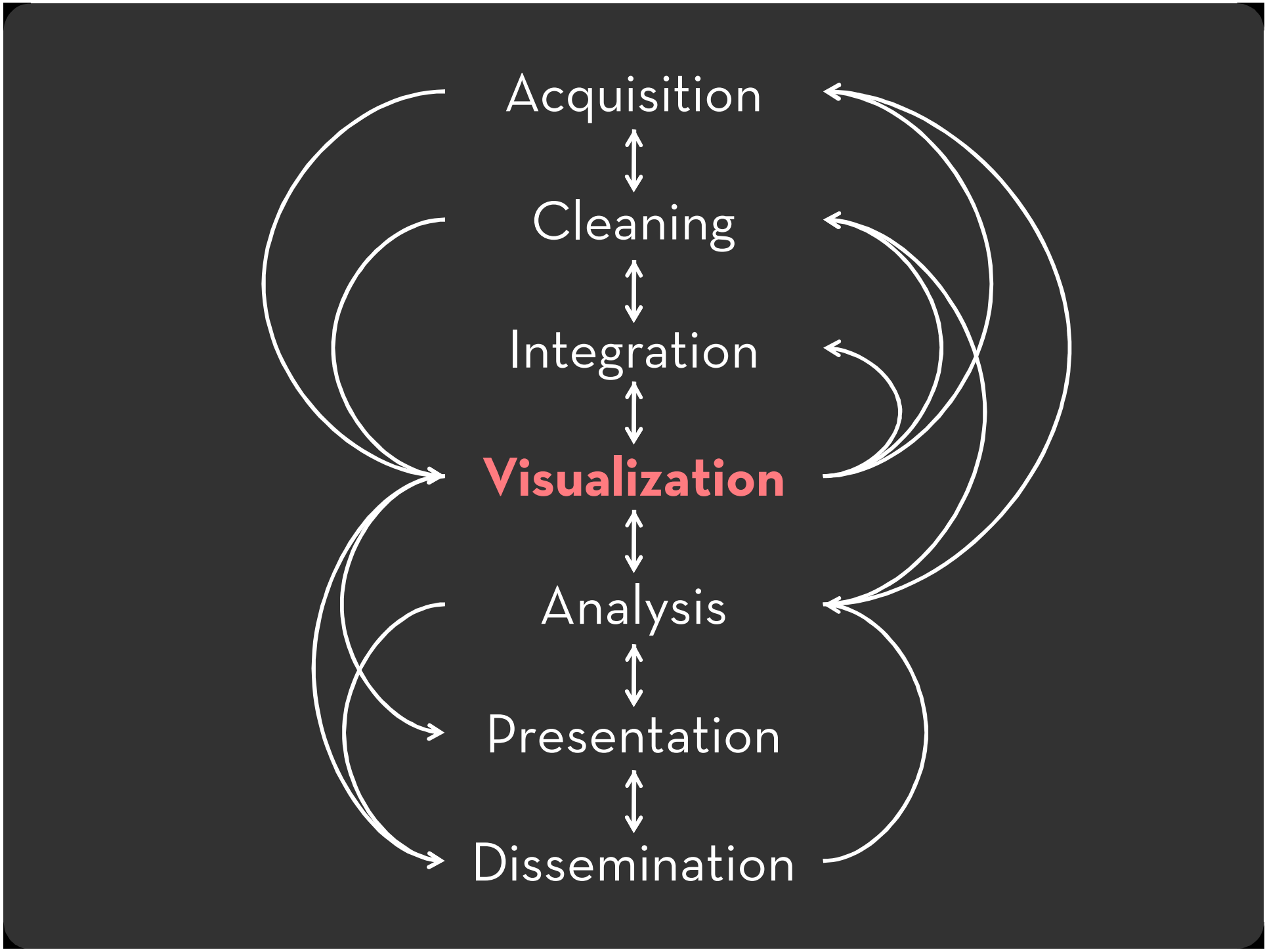
- Job Voyager: 5x less code, 10x less dev time

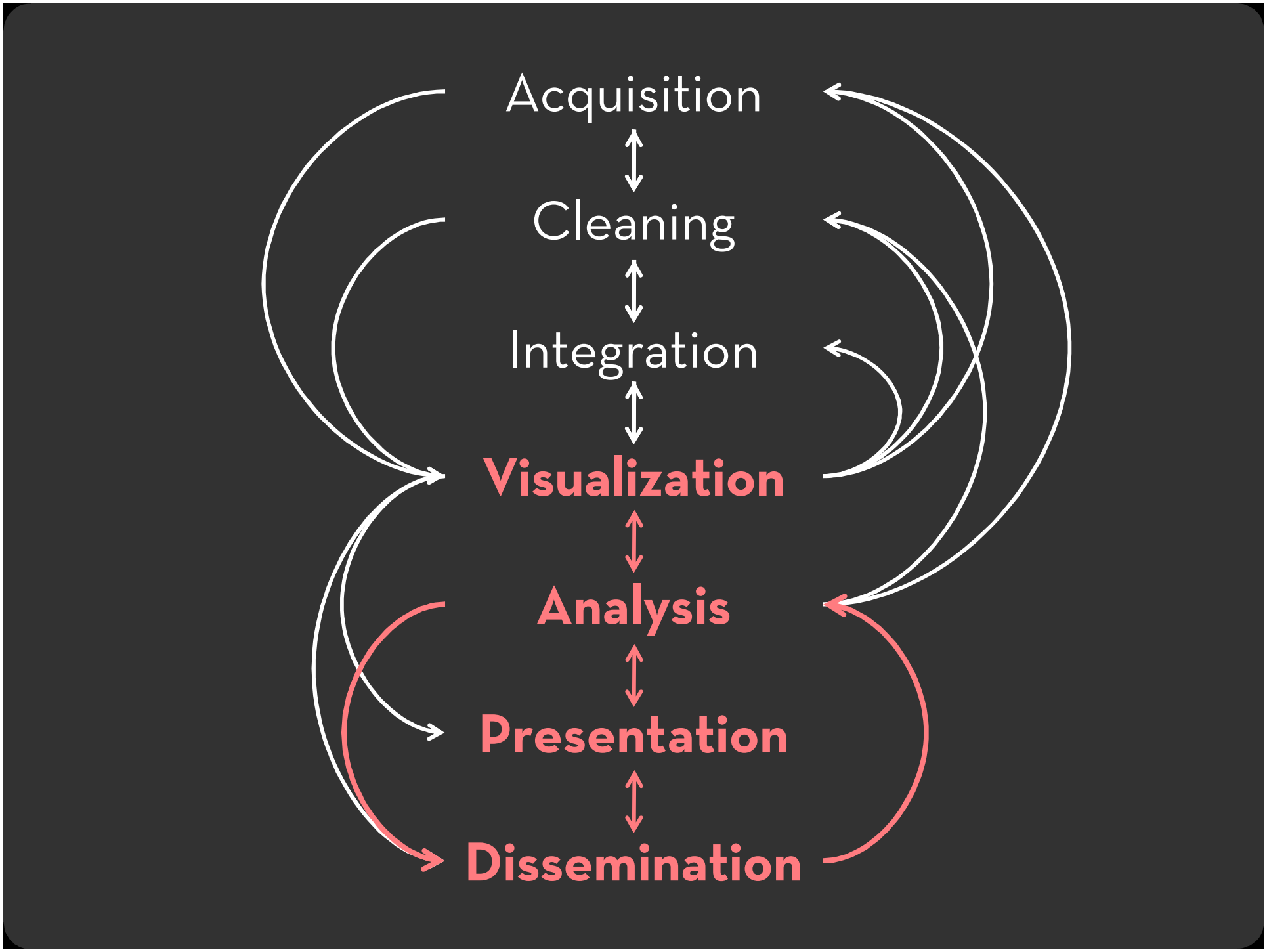
Retargeting across platforms / devices

Behind-the-scenes optimization

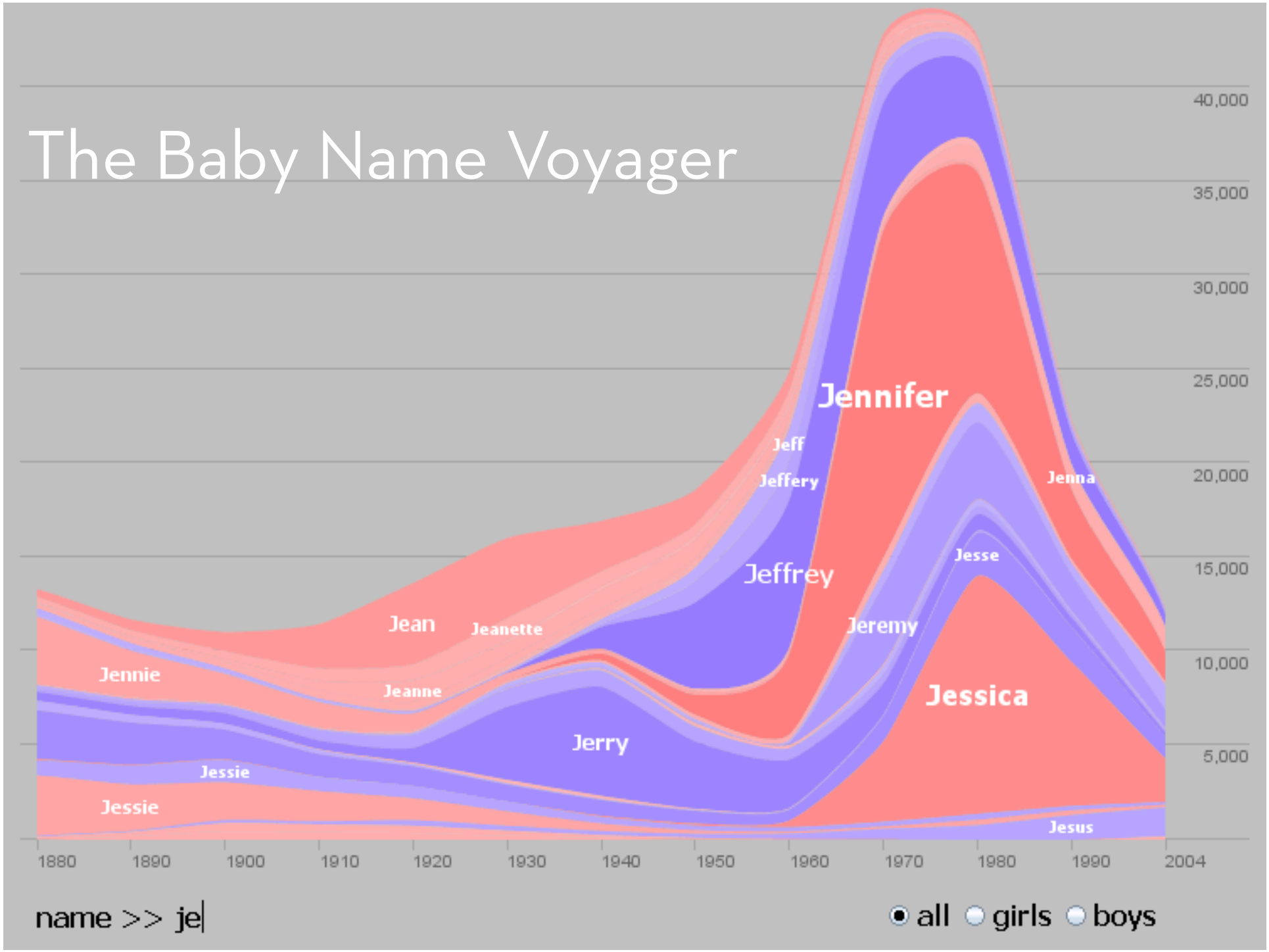
- (a) parallel execution
- (b) streamline code generation

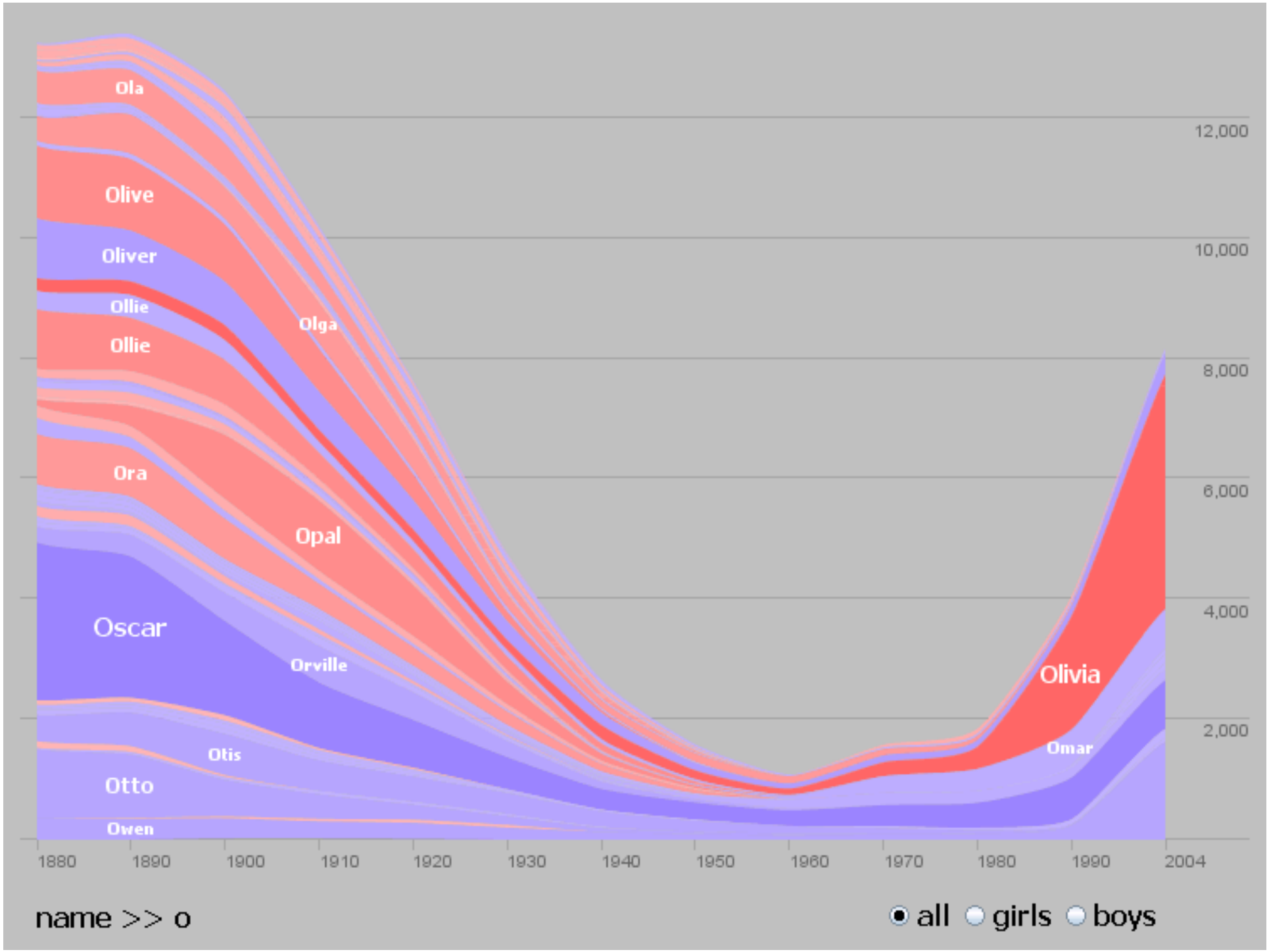
20x scalability over comparable systems (in Java)


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination

Social Data Analysis & sense.us


The Baby Name Voyager
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
Jennifer
Jeff
Jeffery
Jenna
Jeffrey
Jesse
Jean
Jeanette
Jeremy
Jennie
Jeanne
Jessica
Jerry
Jessie
Jessie
Jesus
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2004
name >> je
all
girls
boys


Ola
Olive
Oliver
Ollie
Ollie
Olga
Ora
Opal
Oscar
Orville
Otis
Otto
Owen
Olivia
Omar
12,000
10,000
8,000
6,000
4,000
2,000
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2004
name >> o
all
girls
boys

Social Data Analysis

Visual sensemaking is a **social** as well as cognitive process.

How can user interfaces support and encourage **collaborative analysis**?

Understand impact by building and deploying **real systems**.

sense.us

A Web Application for Collaborative Visualization of Demographic Data

sense.us > social data visualization - Mozilla Firefox

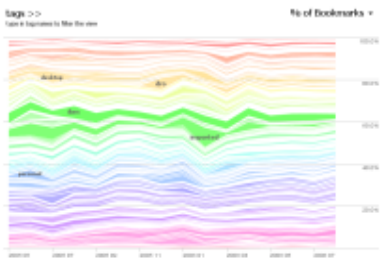
File Edit View History Bookmarks Tools Help

http://sense.us/

Google

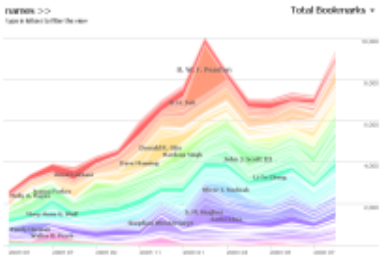
sense.us

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

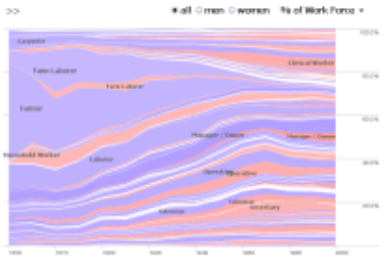
sense.us - social data visualization

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.

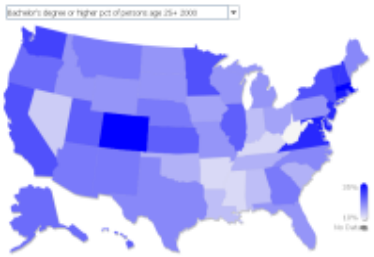
Check out the **user's guide** and **privacy policy** before getting started.

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!



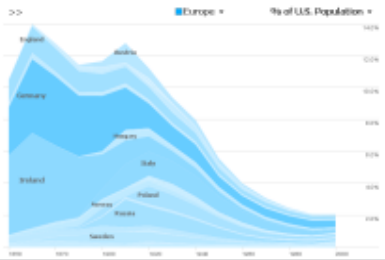
Job Voyager

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



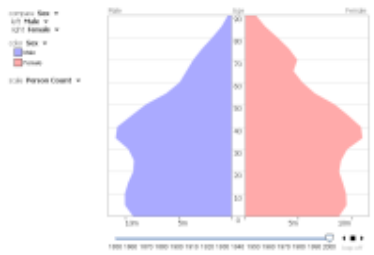
U.S. Census State Map

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



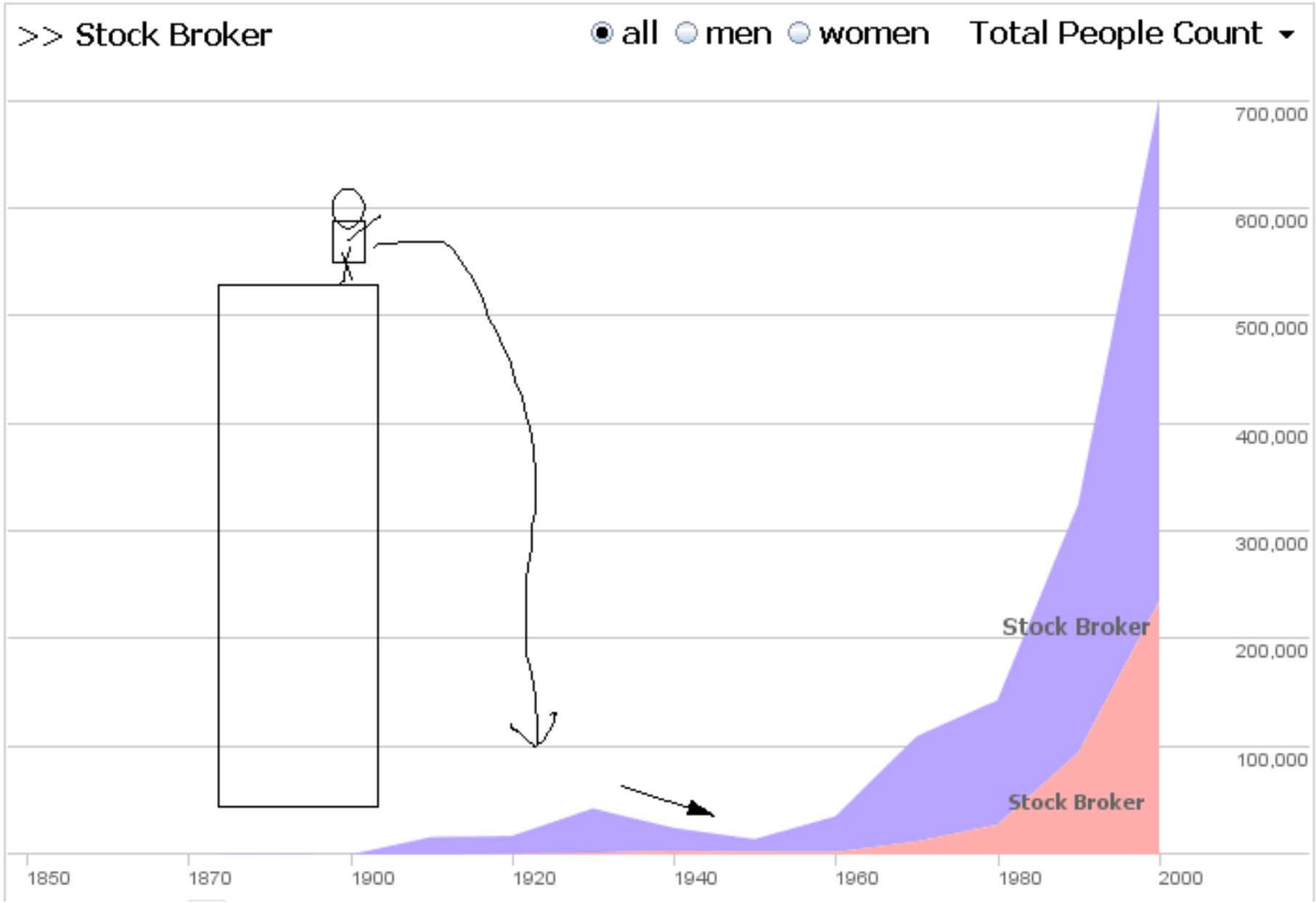
Population Pyramid

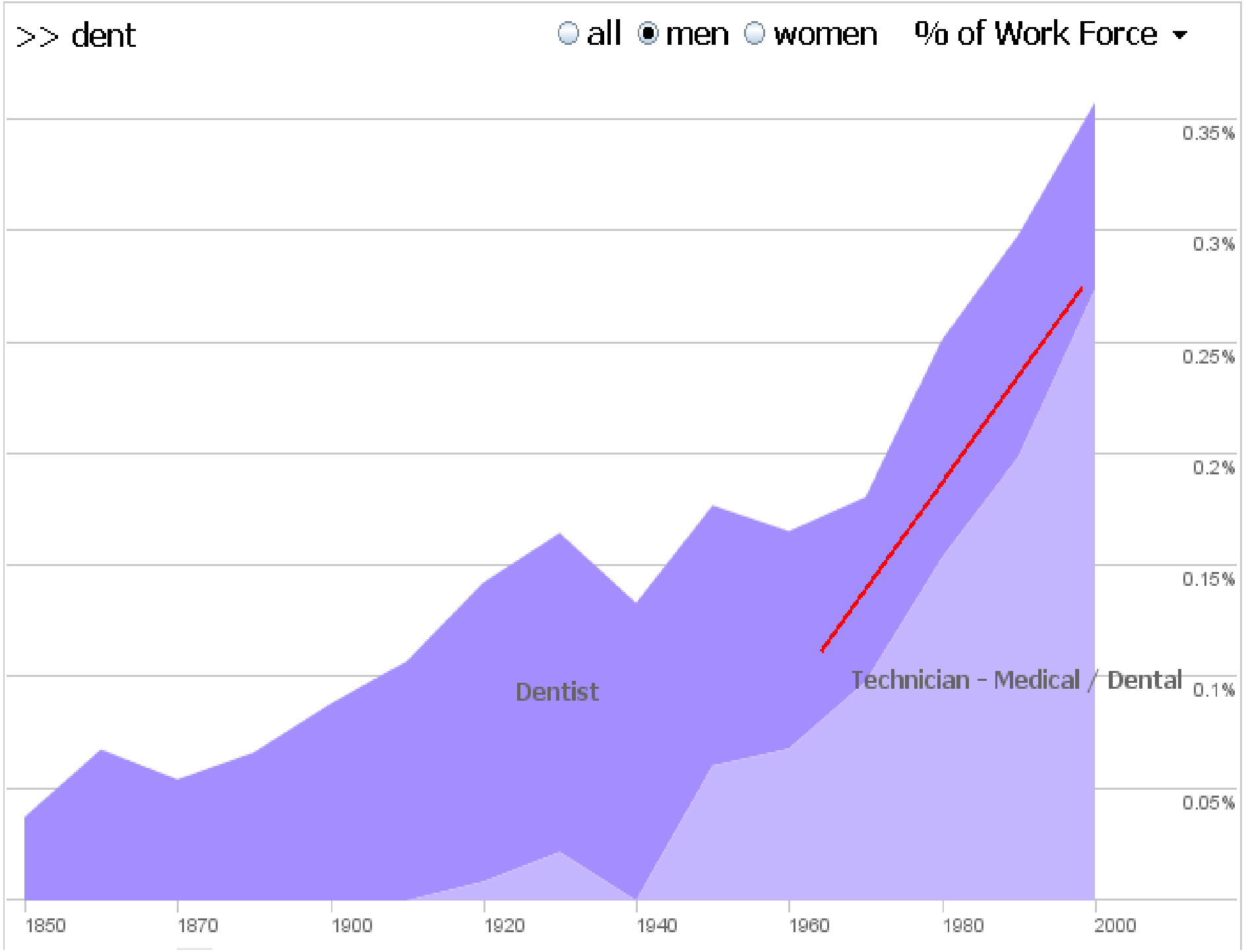
U.S. Population Demographics, 1850-2000
source: http://ipums.org
7 comments

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

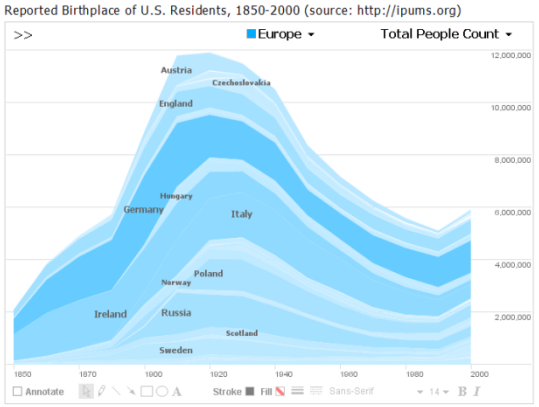



>> dent
all
men
women
% of Work Force
0.35%
0.3%
0.25%
0.2%
0.15%
0.1%
0.05%
Dentist
Technician - Medical / Dental
1850
1870
1900
1920
1940
1960
1980
2000

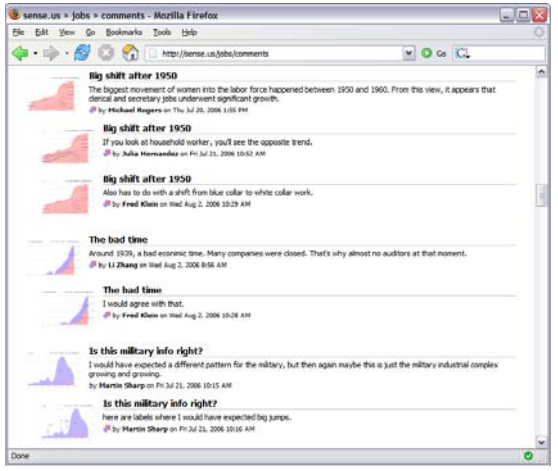
Voyagers and Voyeurs

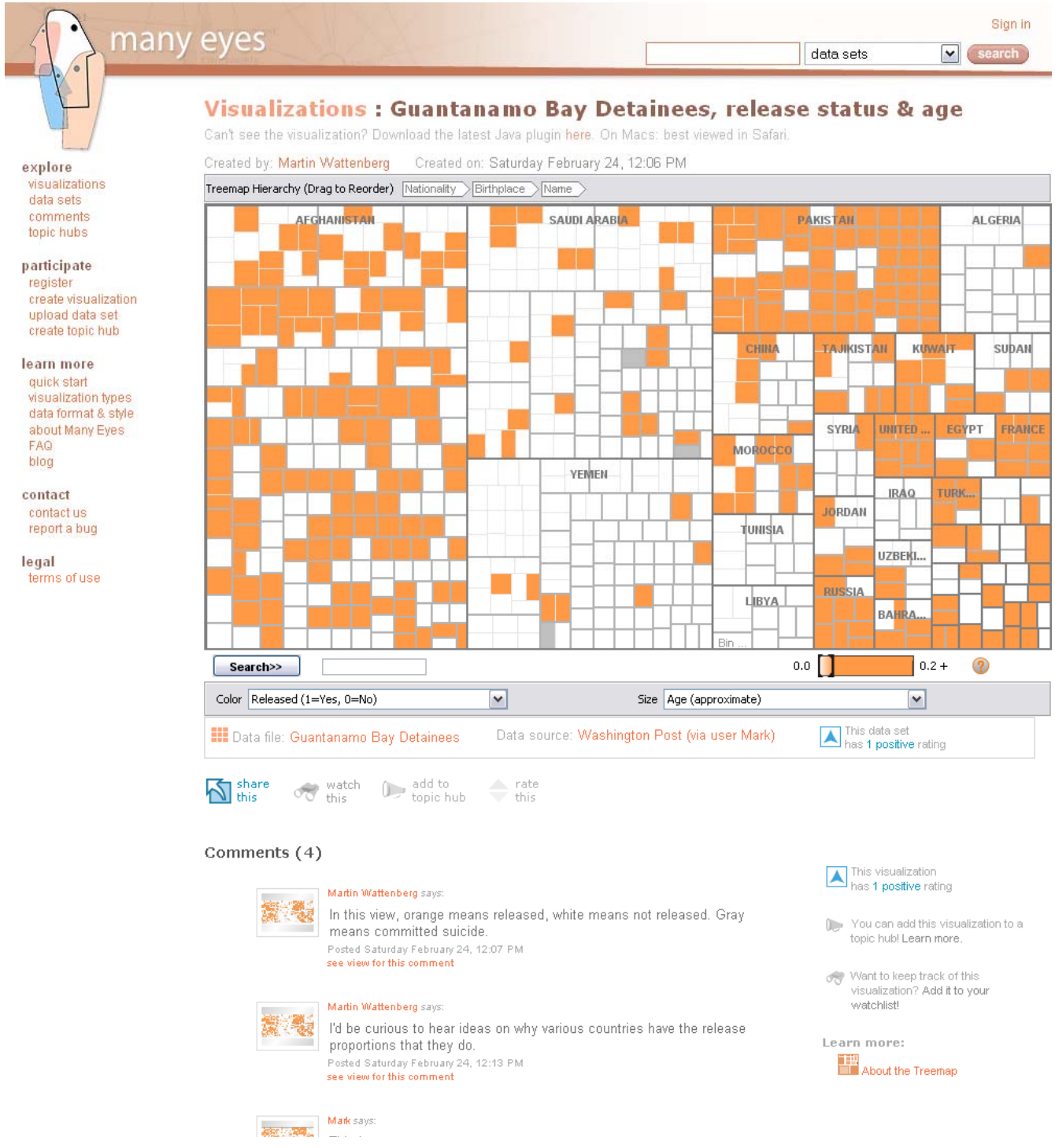
Complementary faces of analysis

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



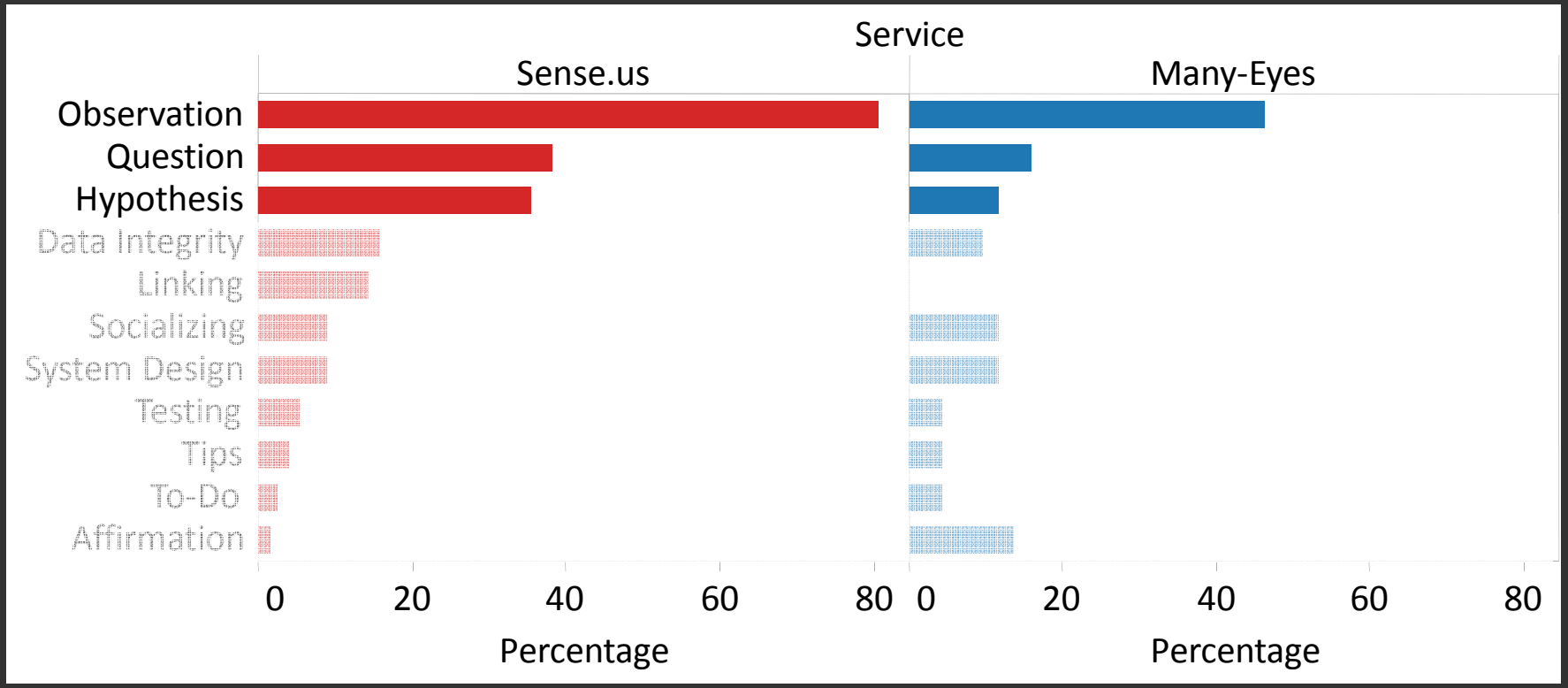
Voyeur – focus on comment listings
Investigate others' explorations
Find people and topics of interest
Catalyze new explorations




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

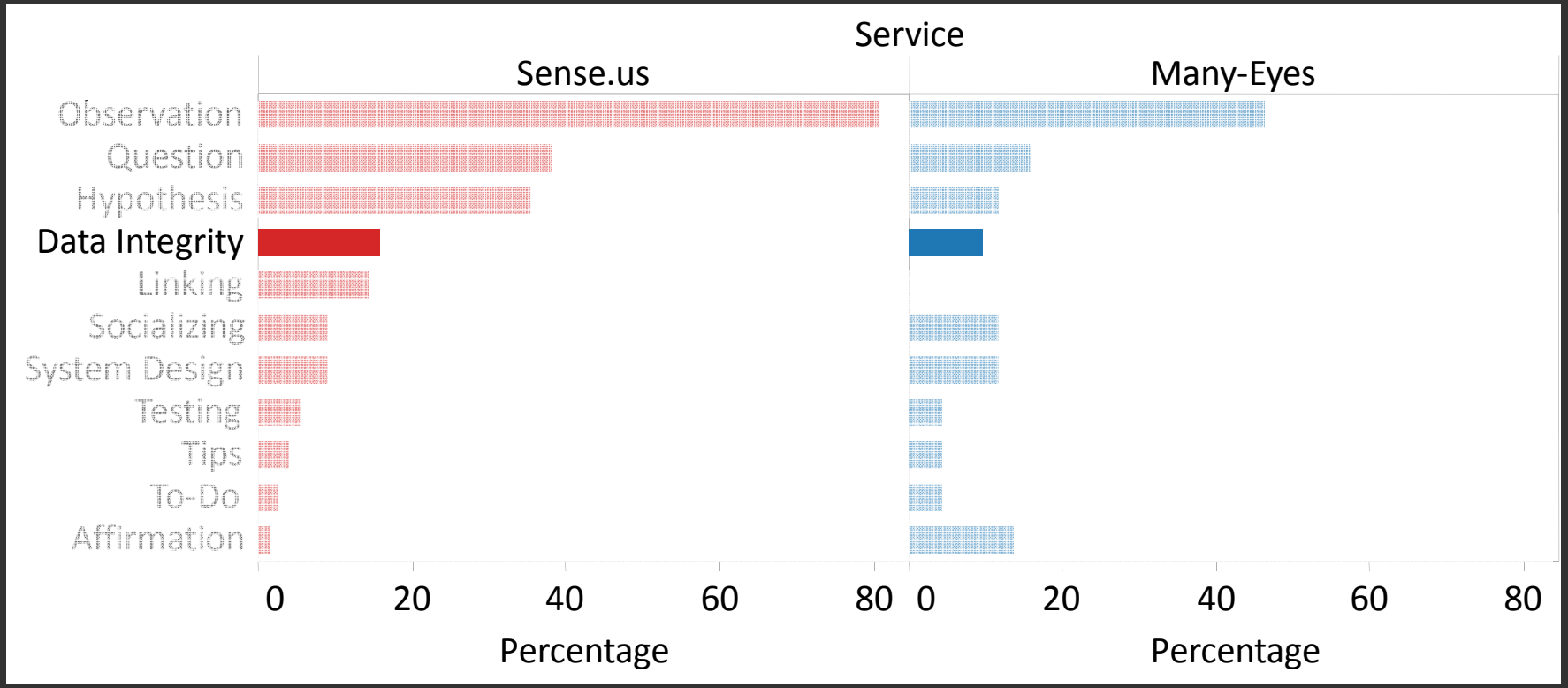
Many-Eyes

Content Analysis of Comments

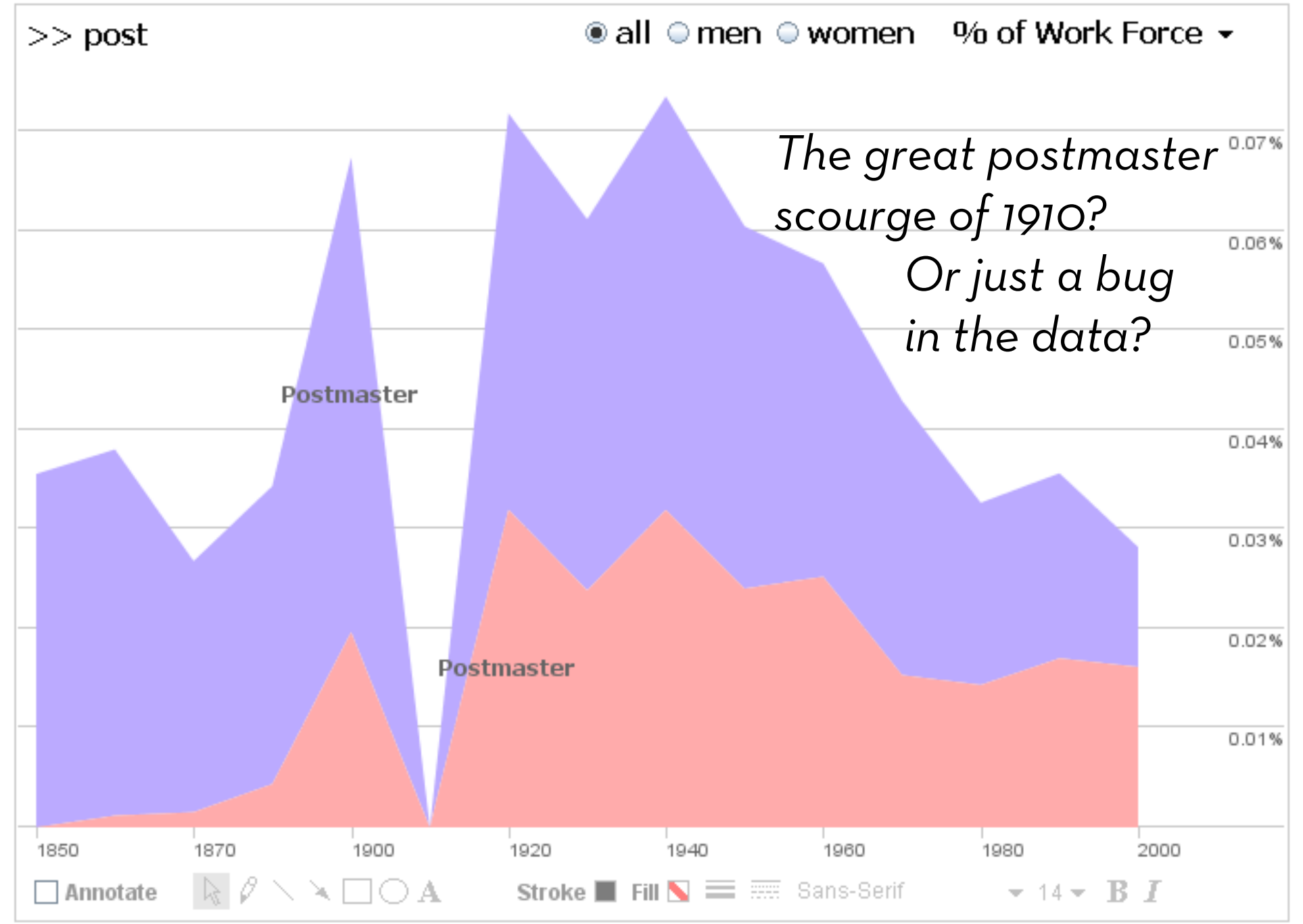


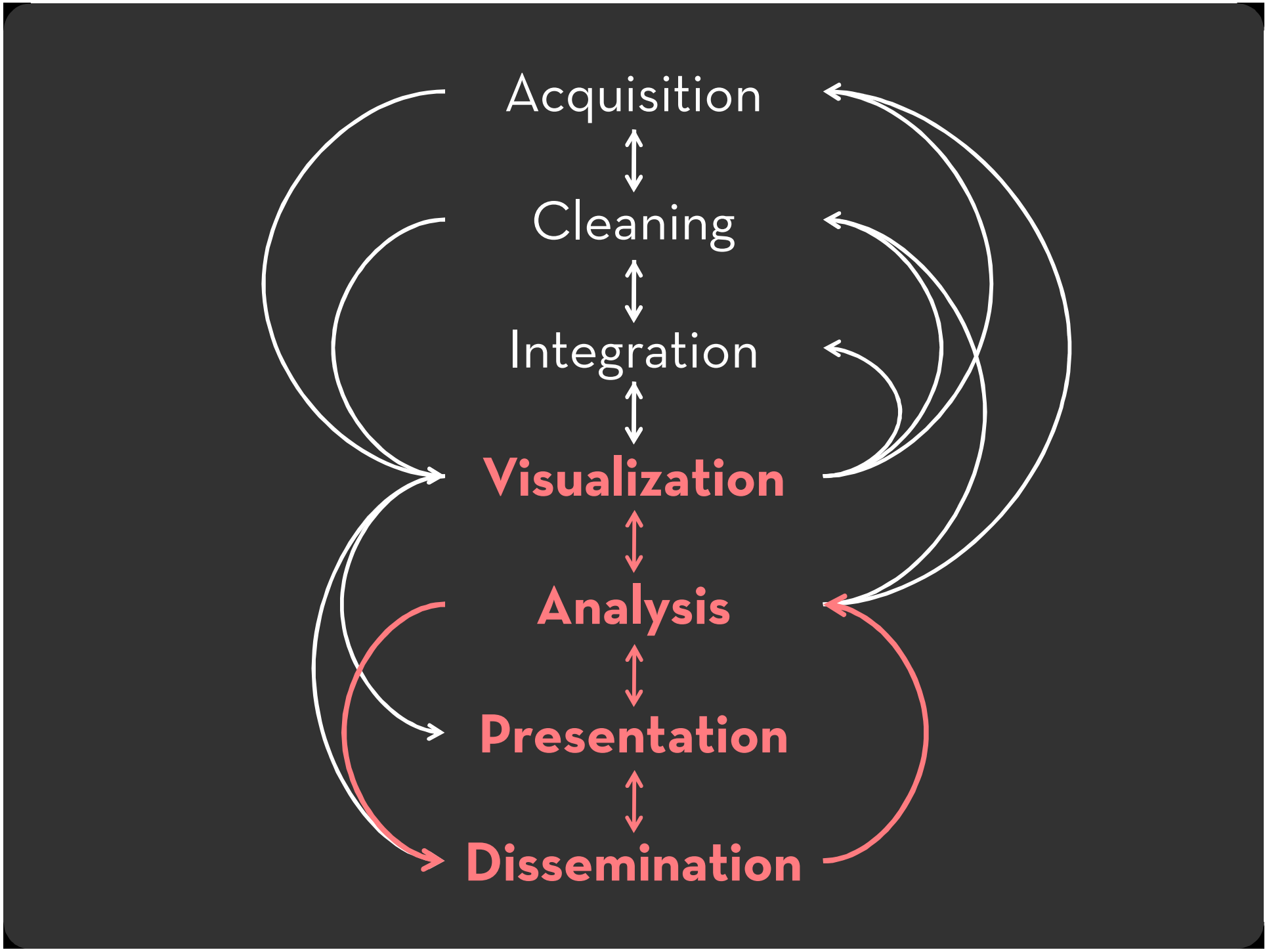
Feature prevalence from content analysis (min Cohen's κ = .74)
High co-occurrence of Observation, Question, and Hypothesis

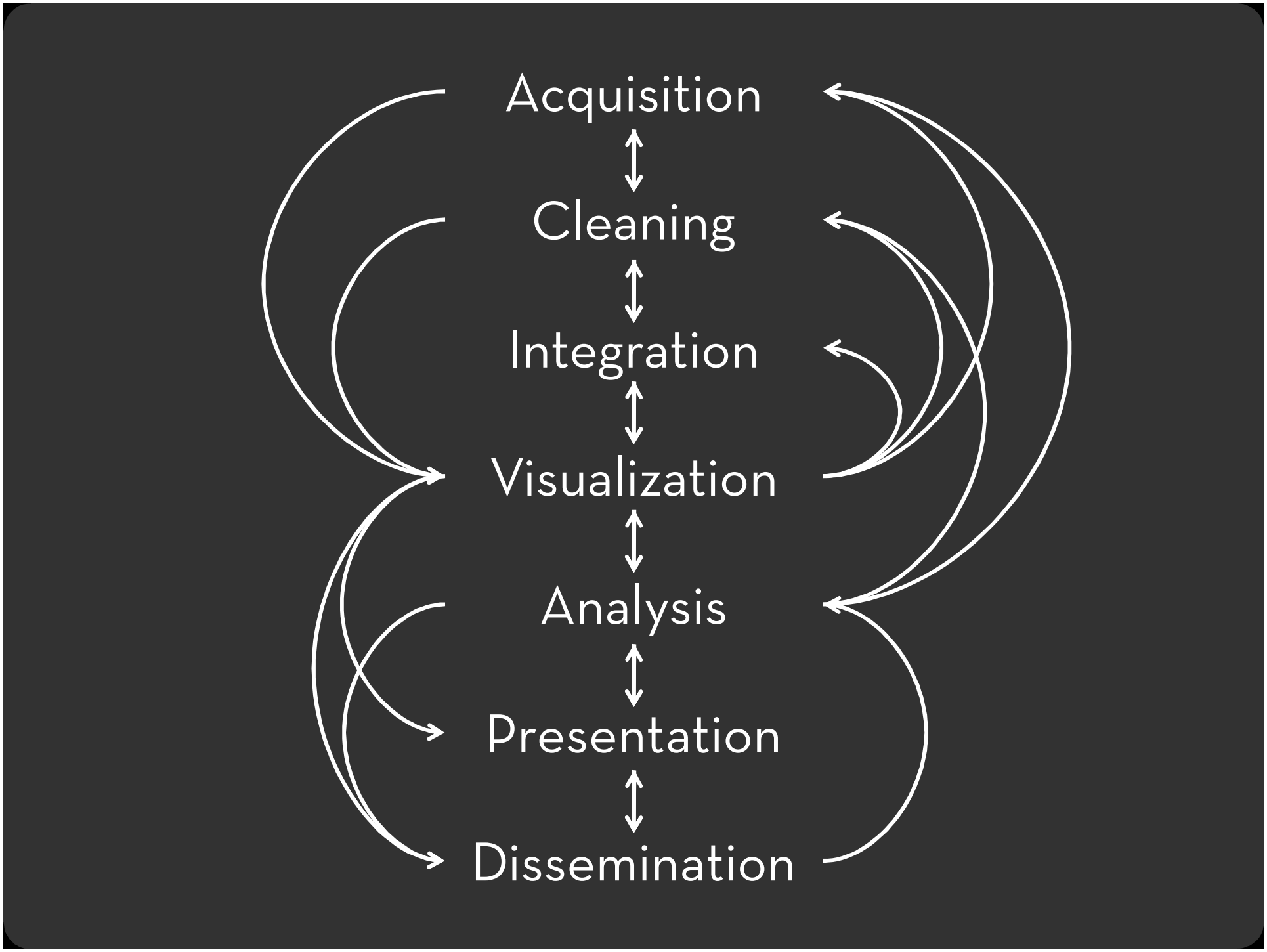
Content Analysis of Comments



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


>> post
all
men
women
% of Work Force
The great postmaster scourge of 1910?
Or just a bug in the data?
Postmaster
Postmaster
0.07%
0.06%
0.05%
0.04%
0.03%
0.02%
0.01%
1850
1870
1900
1920
1940
1960
1980
2000
Annotate
Stroke
Fill
Sans-Serif
14
B
I


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination


Acquisition
Cleaning
Integration
Visualization
Analysis
Presentation
Dissemination

Students & Collaborators

Mike Bostock

Jason Chuang

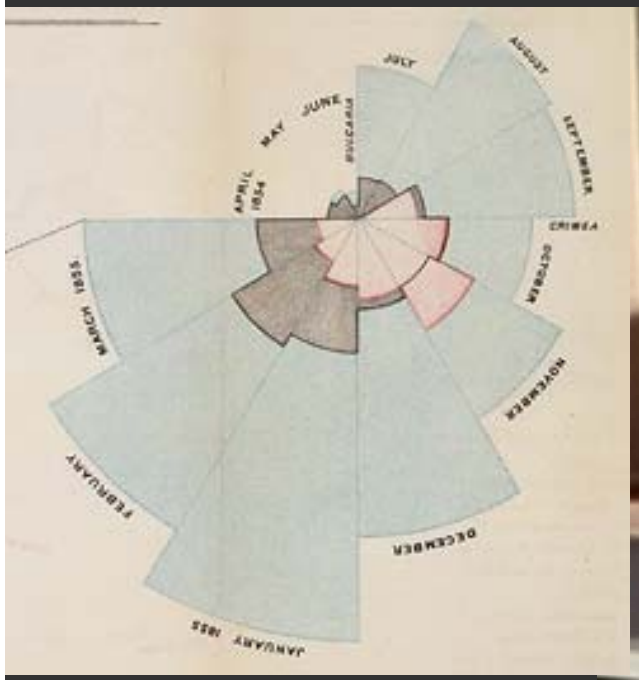
Sean Kandel

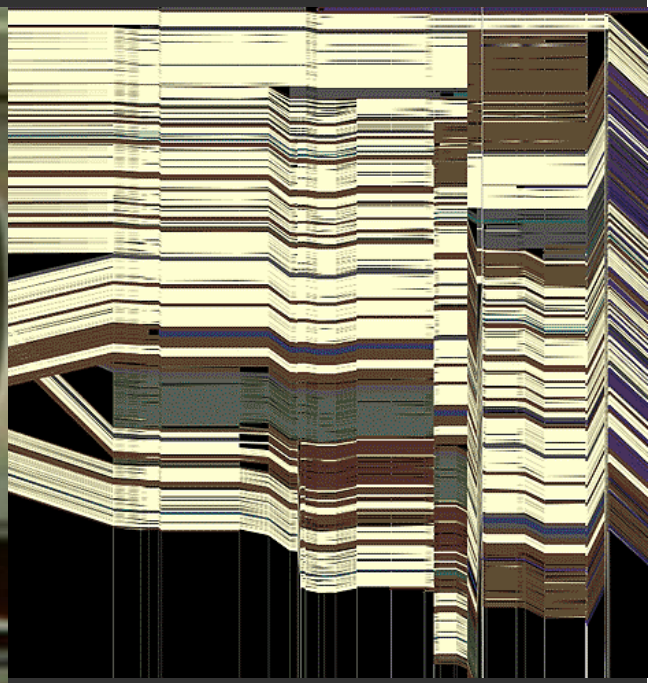
Diana MacLean

Vadim Ogievetsky

Joe Hellerstein, Andreas Paepcke

Interactive Tools for Data Transformation & Visualization





Jeffrey Heer vis.stanford.edu/jheer