# Visualization

Seven dirty secrets of data visualization
By Nate Agrin and Nick Rabinowitz
http://www.netmagazine.com/features/seven-dirty-secrets-data-visualisation

And

Important Tools for Visualizing and Communicating Data
http://www.visualisingdata.com/index.php/resources/

# Data Visualization

- Data visualization - and in particular, web-based data visualization - is having its moment.
- JavaScript libraries like D3.js, Raphaël, and Paper.js, building on modern browser support for Canvas and SVG, have made it easier than ever to produce complex visualizations.
- Data visualization is a wonderful way of exploring data, finding new insights, and telling a compelling story.
- But what are the real challenges visualization developers face

# Real Data is Ugly

- Most data visualization tutorials start with a pleasant fantasy: a pristine data set.

- Whether you're learning to build a basic bar chart or a force-directed network graph, you're presented with clean, normalized, well-formatted base data.

- In practice, when dealing with most real-world data sets, expect to spend up to 80 per cent of your time finding, acquiring, loading, cleaning and transforming your data.

# Real Data is Ugly

- Some of this process can be done with automated tools, but almost any data cleaning involving two or more data sets will require some level of manual work.

- A wide variety of tools can convert XLS to XML or timestamps to other date formats, but nothing can automatically map one company's internal sales categories to those of its competitors, or deal reliably with data entry typos, incompatible character encodings, or poor OCR.

# Real Data is Ugly

**Tools and strategies**

- Google Refine (https://code.google.com/p/google-refine/ ) is a great data cleanup workhorse
  - It has limitations, particularly for non-tabular data.
- Other cleanup-specific tools include:
  - Data Wrangler (http://vis.stanford.edu/wrangler/ )
  - Mr. Data Converter (http://shancarter.github.io/mr-data-converter/).
- However, many tasks still require basic proficiency in a scripting language like Python or manual work in Excel.
  - Save your scripts - you'll use them again
- Visualization is a great tool for identifying data problems. Use scatter plots and histograms to find and fix suspicious outliers

# A Bar Chart is Usually Better

- One of the first questions to ask when considering a potential visualization design is:

  **"Why is this better than a bar chart?"**

- If you're visualizing a single quantitative measure over a single categorical dimension, there is rarely a better option.
- Time-based data is usually best displayed on a line chart
- Scatterplots are often best for exploring correlations between two linear measures
- Bar charts are one of the best tools available for facilitating visual comparisons, leveraging our innate ability to precisely compare side-by-side lengths.

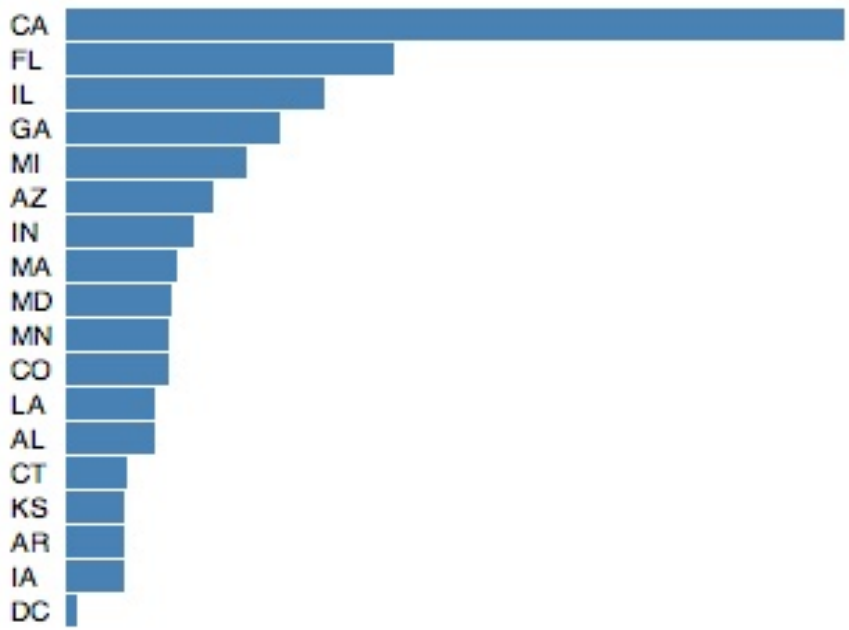
# A Bar Chart is Usually Better

- The corollary to bar chart superiority is that the coolest-looking visualizations are often the least useful
  - The novelty and aesthetic appeal of custom visualizations comes at a cost: **the clarity of the data**.
- Most bar chart alternatives ask the viewer to compare differences we have a harder time discerning: areas, angles, hues, or opacities.
- At best, such visualizations make comparison difficult; at worst, they distort the data entirely, leading viewers to false conclusions.

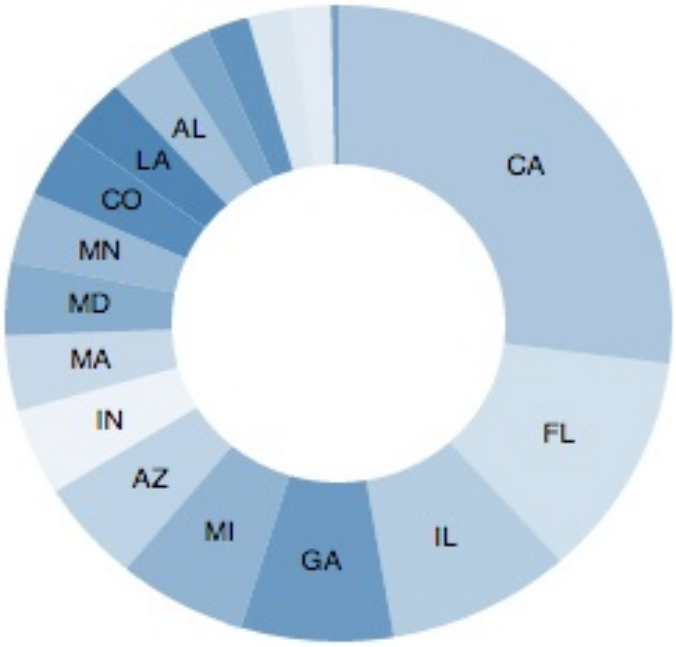
# A Bar Chart is Usually Better

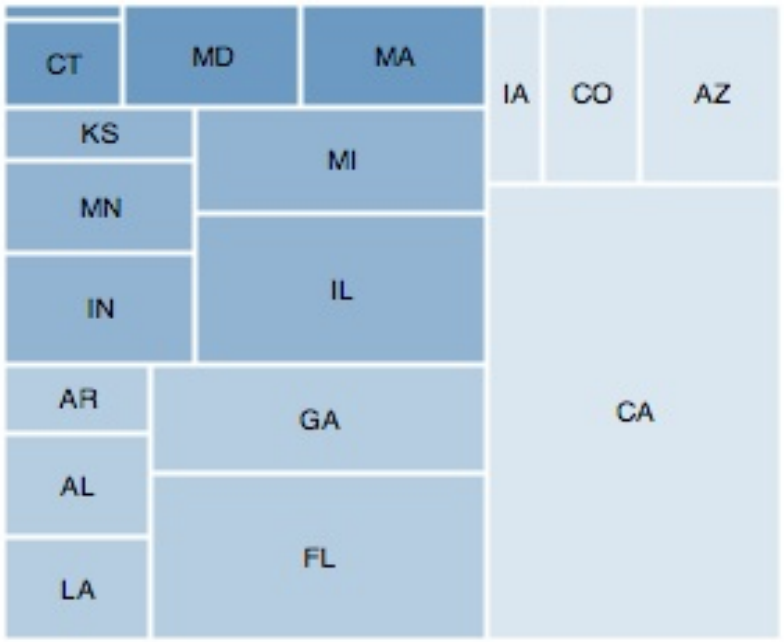
**Tools and strategies**

- Don't dismiss traditional visualization choices if they represent the best option for your data.
  - Start with bar and line charts, and look further only when the data requires it
- Have a good rationale for choosing other options. Compared to bar charts:
  - Bubble charts support more data points with a wider range of values
  - Pies and doughnuts clearly indicate part-whole relationships
  - Treemaps support hierarchical categories


CA
FL
IL
GA
MI
AZ
IN
MA
MD
MN
CO
LA
AL
CT
KS
AR
IA
DC

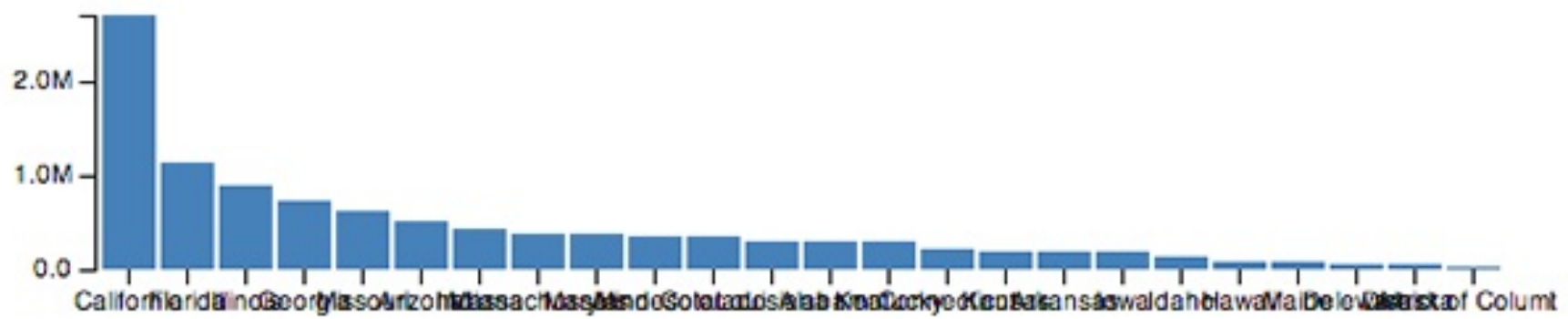

MI
MD
MA
AZ
KS
IN
AR
IA
FL
CT
CO
GA
CA
IL
MN
AL
LA

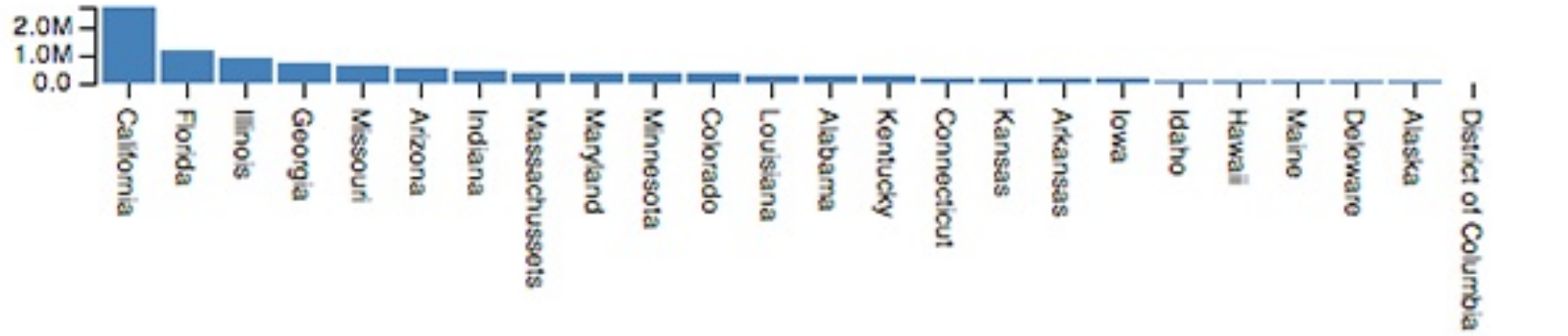

CA
FL
IL
GA
MI
AZ
IN
MA
MD
MN
CO
LA
AL

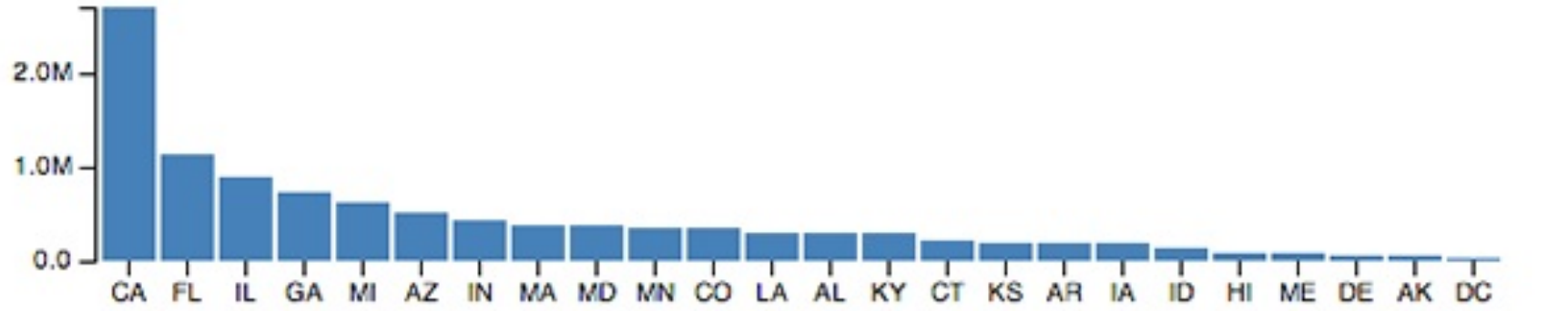

CT
MD
MA
IA
CO
AZ
KS
MI
MN
IL
IN
AR
GA
CA
AL
FL
LA

# The Devil is in the Details

- Designing the labels, legends and axes for your visualization is often an afterthought to the initial visualization.
- These elements are crucially important to the visualization, and can be difficult and time-consuming to get right, especially when you can't predict the data ahead of time.


2.0M
1.0M
0.0


2.0M
1.0M
0.0
California
Florida
Illinois
Georgia
Missouri
Arizona
Indiana
Massachussets
Maryland
Minnesota
Colorado
Louisiana
Alabama
Kentucky
Connecticut
Kansas
Arkansas
Iowa
Idaho
Hawaii
Maine
Delaware
Alaska
District of Columbia


2.0M
1.0M
0.0
CA
FL
IL
GA
MI
AZ
IN
MA
MD
MN
CO
LA
AL
KY
CT
KS
AR
IA
ID
HI
ME
DE
AK
DC

# The Devil is in the Details

**Tools and strategies**

- Plan space around your graphic for labels, axes and legends
- Designate a maximum character length for labels, truncating if needed to prevent crowding. Group nearby labels together, revealing them in response to user actions
- Consider scrolling or accordion-style expansion for long legends
- Whatever you do, don't leave these elements out. Labels may seem like a secondary concern when you're focused on the graphic elements, but they are incredibly important to your viewers

# Visualization is not Analysis

- It's a central tenet of the field that data visualization can yield meaningful insight.
  - It’s important to remember that visualization is a tool to aid analysis, not a substitute for analytical skill.
- It’s also not a substitute for statistics:
  - Your chart may highlight differences or correlations between data points, but to reliably draw conclusions from these insights often requires a more rigorous statistical approach.
  - The reverse can also be true - as Anscombe’s Quartet demonstrates, visualizations can reveal differences statistics hide.
- Really understanding your data generally requires a combination of analytical skills, domain expertise, and effort.

# Data Visualization Takes More than Code

- The range of libraries and tutorials now available make it easier than ever to produce production-quality web-based visualizations without specialized expertise.
- Creating visualizations that offer real insight or tell a compelling story still requires a particularly wide range of real skills in addition to coding
  - Including graphic design, data analysis, and an understanding of interaction design and human perception.
- No library or technology can substitute for knowing what you're doing.

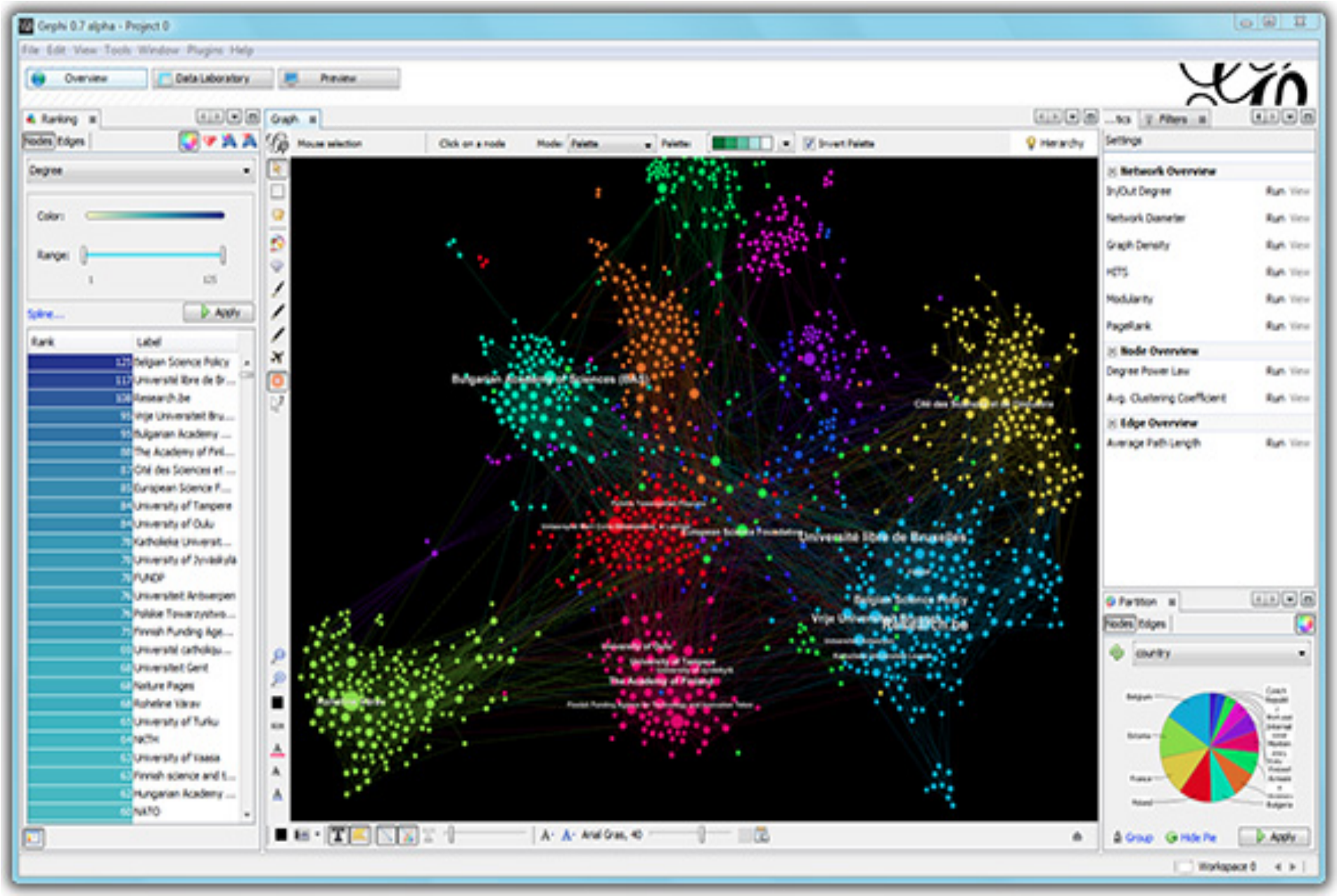
# Data Visualization Takes More than Code

- But the flip side of this secret is that you don't need to know that much.
- Especially if you use well-established visualizations and interaction principles.
- Learn enough about the field to avoid newbie mistakes (always zero-base your bar charts and never set a circle radius with a linear scale)
- Keep things simple (no 3D, limited animation, no drop shadows), base your work on solid examples and you can create great visualizations.

# Important Tools for Visualizing and Communicating Data

# Gephi

- Gephi is an open-source, free interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. It claims to be “like Photoshop but for data”, allowing the user to interact with the data representation, manipulate structures, shapes and colors to reveal hidden properties.
- http://gephi.org/
- http://gephi.org/features/
- Cost: Free


Gephi 0.7 alpha - Project 0
File Edit View Tools Window Plugins Help
Overview
Data Laboratory
Preview
Ranking
Nodes Edges
Degree
Color:
Range:
Spline...
Apply
Rank
Label
Belgian Science Policy
Universite libre de Br...
Research.be
Vrije Universiteit Bru...
Bulgarian Academy ...
The Academy of Fin...
Cnil des Sciences et ...
European Science F...
University of Tampere
University of Oulu
Katholieke Universit...
University of Jyväskylä
FUNDP
Universiteit Antwerpen
Polskie Towarzystwo...
Finnish Funding Age...
Universite catholiqu...
Universiteit Gent
Nature Pages
Roheline Varav
University of Turku
nachi
University of Vaasa
Finnish science and t...
Hungarian Academy ...
NATO
Graph
Mouse selection
Click on a node
Mode:
Palette:
Invert Palette
Hierarchy
Bulgarian Academy of Sciences (BAS)
Cnil des Sciences et Industrie
Universite libre de Bruxelles
Belgian Science Policy
Vrije Universiteit Brussel
Research.be
University of Oulu
University of Tampere
Roheline Varav
Arial Gras, 40
Settings
Network Overview
In/Out Degree
Network Diameter
Graph Density
HITS
Modularity
PageRank
Node Overview
Degree Power Law
Avg. Clustering Coefficient
Edge Overview
Average Path Length
Run
View
Partition
Nodes Edges
country
Group
Hide Pie
Apply
Workspace 0

# Processing

- Processing is an open source programming language and environment for people who want to create images, animations, and interactions. Today, there are tens of thousands of students, artists, designers, researchers, and hobbyists who use Processing for learning, prototyping, and production.

- http://processing.org/

- http://processing.org/exhibition/

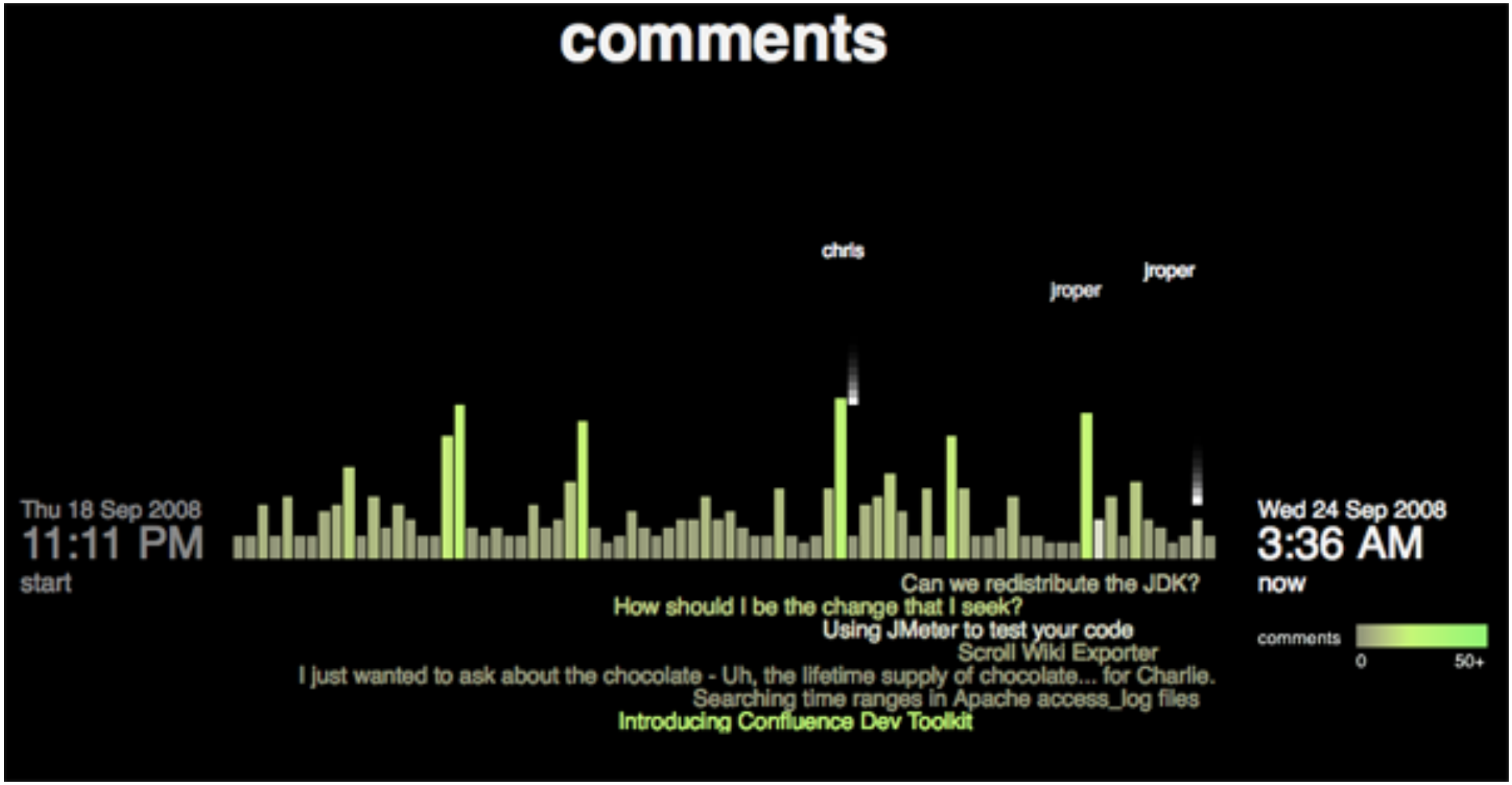
- For GNU/Linux, Mac OS X, and Windows

# Processing.js

- Processing.js is a 'port' of Processing, a sister project as it were, designed to make data visualizations, digital art, interactive animations, educational graphs, video games, etc. work using web standards and without any plug-ins.
- You write code using the Processing language and include it in a web page

- Interactive & Animations

- http://processingjs.org/

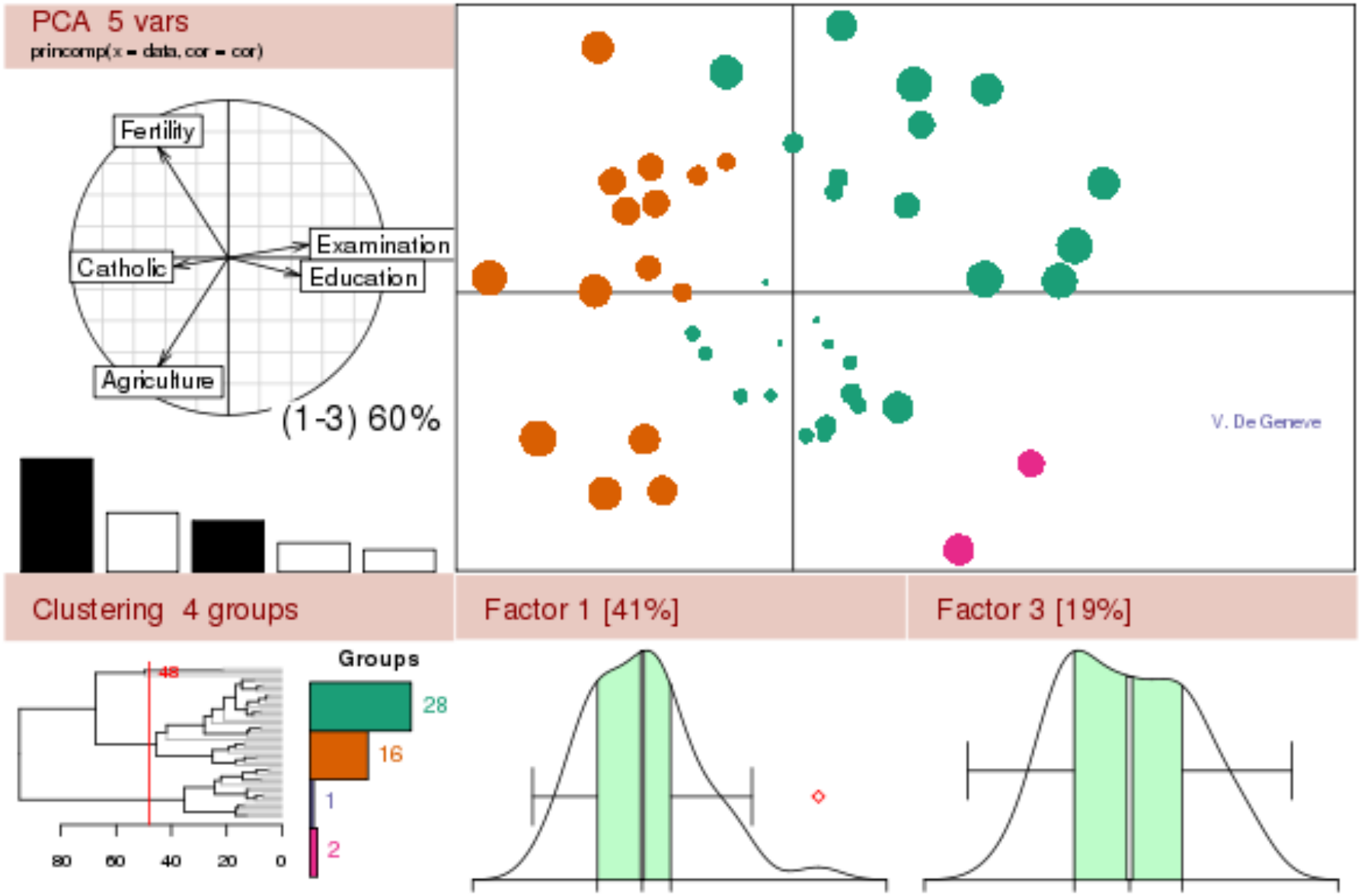

comments
chris
jroper
jroper
Thu 18 Sep 2008
11:11 PM
start
Wed 24 Sep 2008
3:36 AM
now
Can we redistribute the JDK?
How should I be the change that I seek?
Using JMeter to test your code
Scroll Wiki Exporter
I just wanted to ask about the chocolate - Uh, the lifetime supply of chocolate... for Charlie.
Searching time ranges in Apache access_log files
Introducing Confluence Dev Toolkit
comments
0
50+

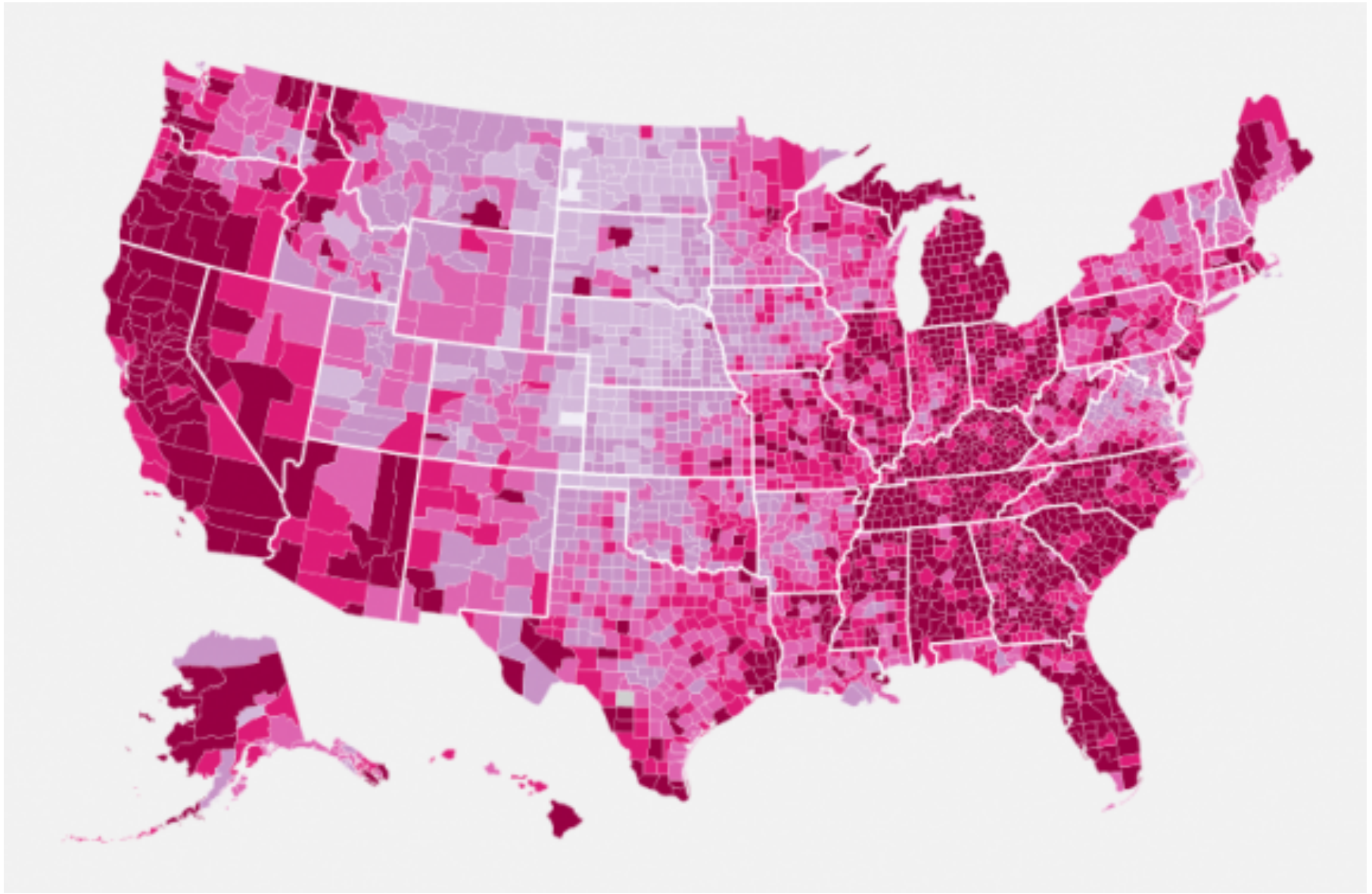
# R

- R is a highly extensible, open source language and environment for data handling, statistical computing and graphical techniques.
- One of R's key strengths is the ease with which well-designed publication-quality graphical plots can be produced.

- http://www.r-project.org/

- http://www.rstudio.com/
-


PCA 5 vars
princomp(x = data, cor = cor)
Fertility
Examination
Catholic
Education
Agriculture
(1-3) 60%
V. De Geneve
Clustering 4 groups
48
80
60
40
20
0
Groups
28
16
1
2
Factor 1 [41%]
Factor 3 [19%]

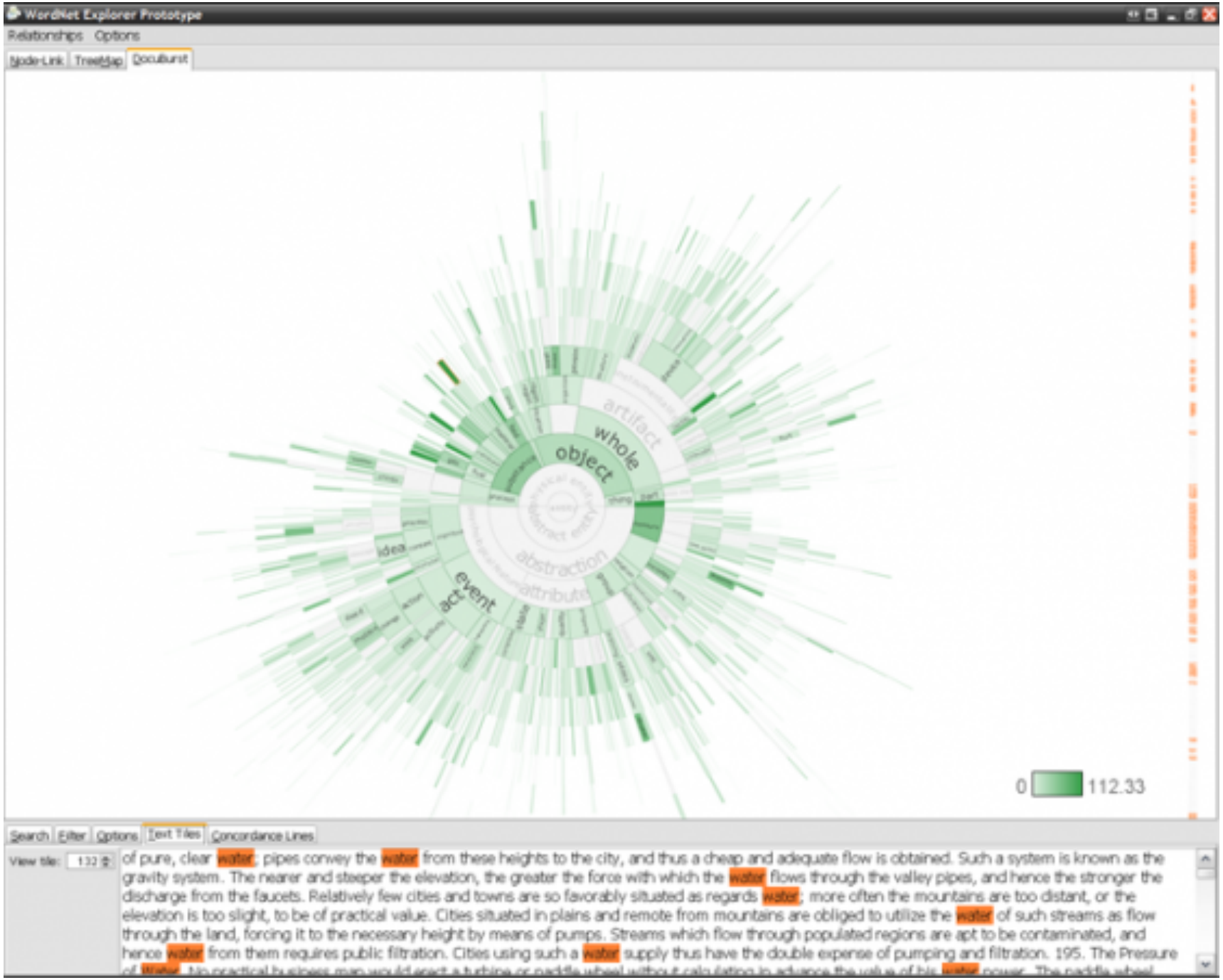


# Prefuse

- Prefuse is a Java programming framework for creating rich interactive data visualizations supporting a rich set of features for data modeling, visualization, and interaction.
- Provides optimized data structures for tables, graphs, and trees, a host of layout and visual encoding techniques, and support for animation, dynamic queries, integrated search, and database connectivity.

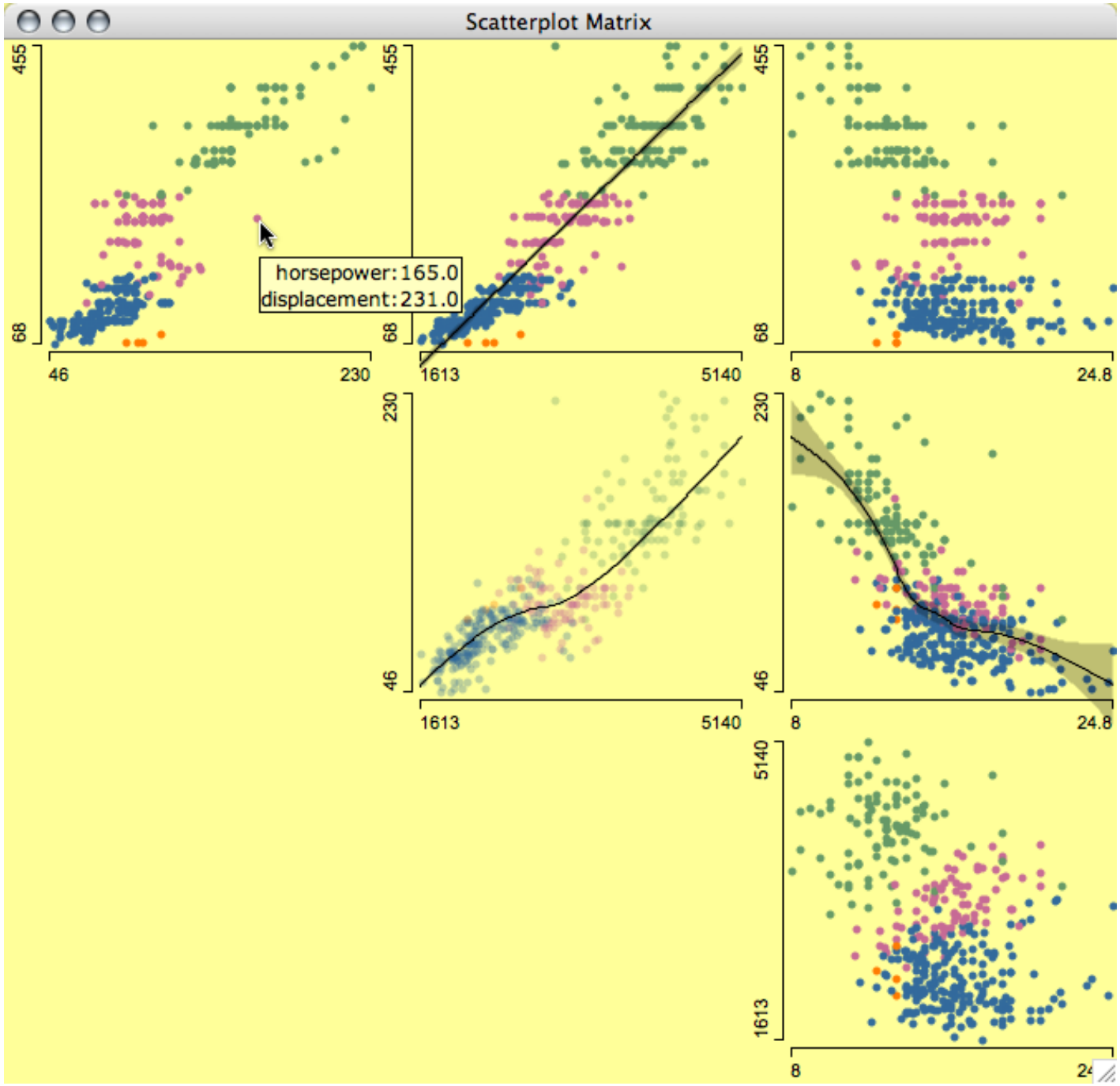
- Visual Programming Language, Java, Open Source

- http://prefuse.org/


WordNet Explorer Prototype
Relationships
Options
Node-Link
Treemap
Docuburst
artifact
whole
object
abstraction
idea
act
event
0
112.33
Search
Filter
Options
Text Files
Concordance Lines
View file:
of pure, clear water; pipes convey the water from these heights to the city, and thus a cheap and adequate flow is obtained. Such a system is known as the gravity system. The nearer and steeper the elevation, the greater the force with which the water flows through the valley pipes, and hence the stronger the discharge from the faucets. Relatively few cities and towns are so favorably situated as regards water; more often the mountains are too distant, or the elevation is too slight, to be of practical value. Cities situated in plains and remote from mountains are obliged to utilize the water of such streams as flow through the land, forcing it to the necessary height by means of pumps. Streams which flow through populated regions are apt to be contaminated, and hence water from them requires public filtration. Cities using such a water supply thus have the double expense of pumping and filtration. 195. The Pressure

# Mondarian

- Mondrian is a general purpose statistical data-visualization system written in Java.
- Has particular strengths, compared to other tools, for working with categorical, geographical and large data sets.
- Currently implemented plots comprise Histograms, Boxplots, Scatterplots, Barcharts, Mosaicplots, Missing Value Plots, Parallel Coordinates/Boxplots, SPLOMs and Maps.
- Mondrian works with data in standard tab-delimited or comma-separated ASCII files and can load data from R workspaces.
- http://www.theusrus.de/Mondrian/
- http://www.theusrus.de/Mondrian/Mondrian.html#Hist

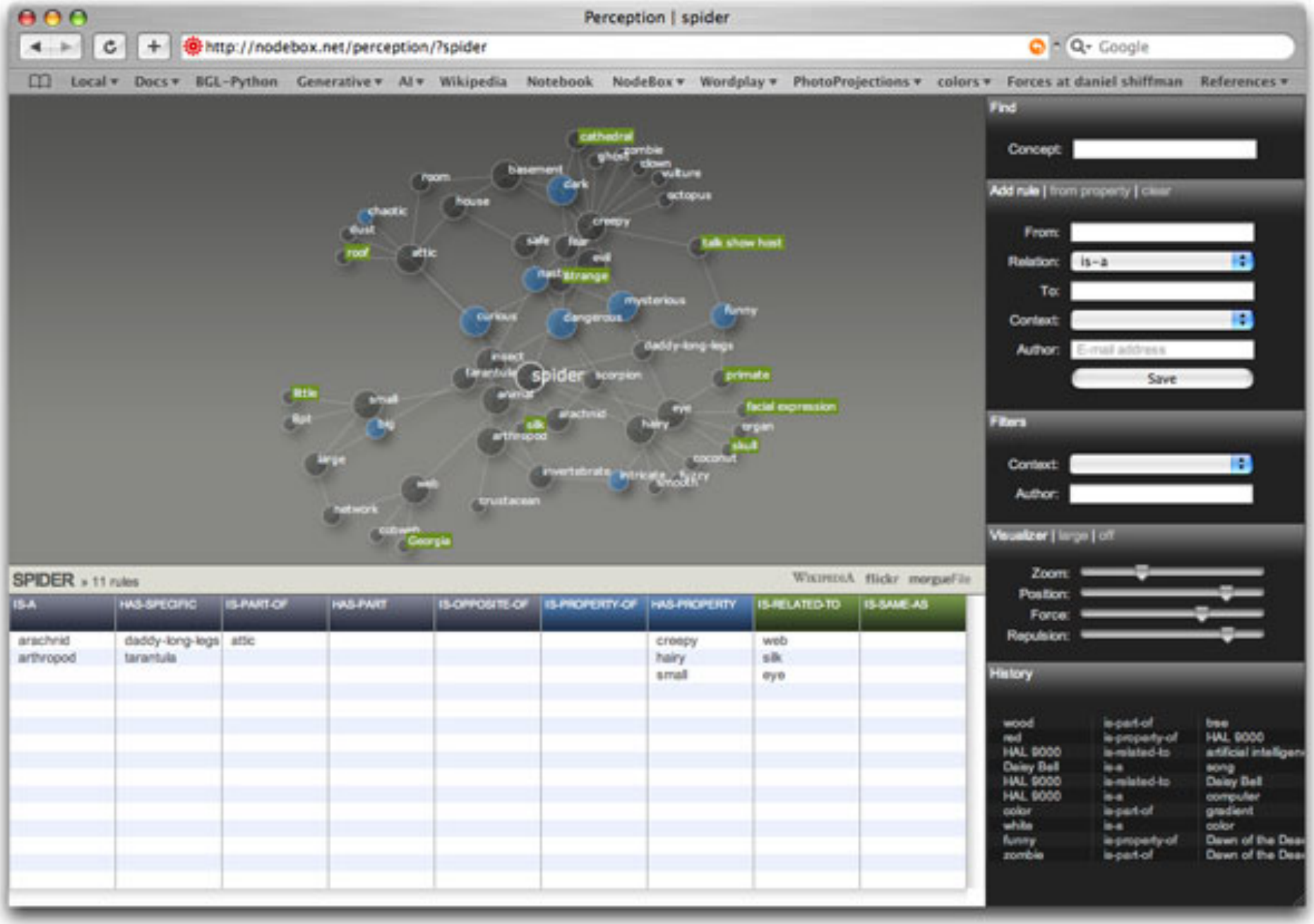

Scatterplot Matrix
horsepower: 165.0
displacement: 231.0
455
68
46
230
1613
5140
8
24.8

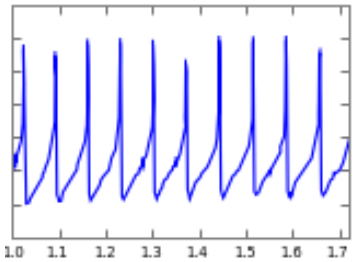
# Python (Matplotlib)

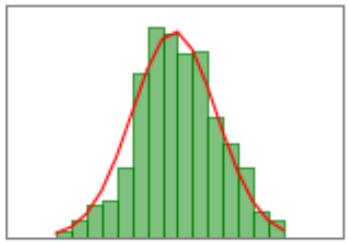
- Python is a powerful, versatile and increasingly common programming language usually deployed as an automation tool on the data handling side of visualization projects (eg. scraping data, parsing it, formatting it) but it is also used as the basis for graphing and visualization libraries too.

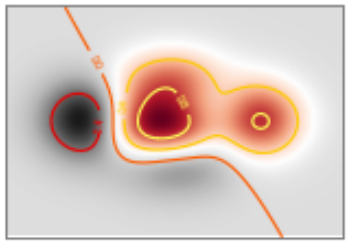
- http://matplotlib.org/

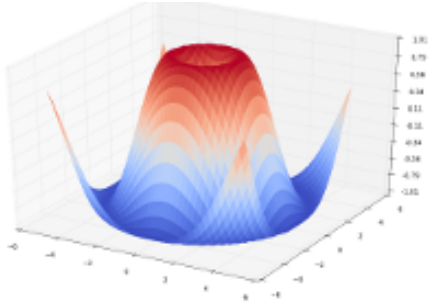
- http://www.youtube.com/watch?v=3Fp1zn5ao2M&feature=plcp

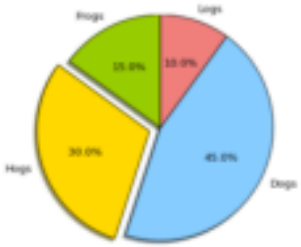

Perception | spider
http://nodebox.net/perception/?spider
Google
Local
Docs
BGL-Python
Generative
AI
Wikipedia
Notebook
NodeBox
Wordplay
PhotoProjections
colors
Forces at daniel shiffman
References
cathedral
zombie
ghost
clown
vulture
octopus
basement
dark
room
house
chaotic
dust
roof
attic
creepy
safe
fear
evil
talk show host
nest
strange
mysterious
funny
curious
dangerous
daddy-long-legs
insect
tarantula
spider
scorpion
primate
arthropod
arachnid
eye
facial expression
hairy
organ
skull
coconut
silk
web
crustacean
network
Georgia
SPIDER » 11 rules
WIKIPEDIA flickr morgueFile
IS-A
HAS-SPECIFIC
IS-PART-OF
HAS-PART
IS-OPPOSITE-OF
IS-PROPERTY-OF
HAS-PROPERTY
IS-RELATED-TO
IS-SAME-AS
arachnid
arthropod
daddy-long-legs
tarantula
attic
creepy
hairy
small
web
silk
eye
Find
Concept:
Add rule | from property | clear
From:
Relation:
is-a
To:
Context:
Author:
E-mail address
Save
Filters
Context:
Author:
Visualizer | large | off
Zoom:
Position:
Force:
Repulsion:
History
wood
red
HAL 9000
Daisy Bell
HAL 9000
HAL 9000
color
white
funny
zombie
is-part-of
is-property-of
is-related-to
is-a
is-related-to
is-a
is-part-of
is-a
is-property-of
is-part-of
tree
HAL 9000
song
Daisy Bell
computer
gradient
color
Dawn of the Dead
Dawn of the Dead

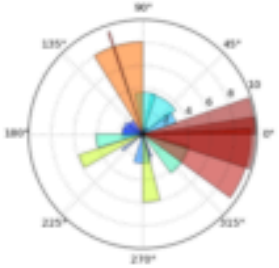

1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7

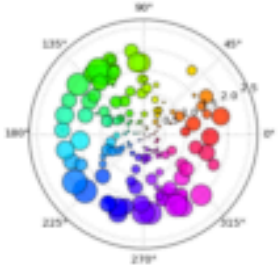




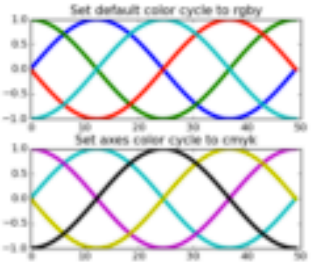


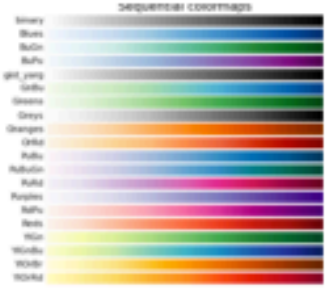

Frogs
Logs
Hogs
Dogs
15.0%
10.0%
30.0%
45.0%

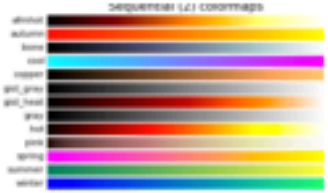


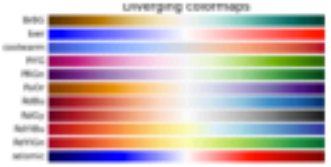


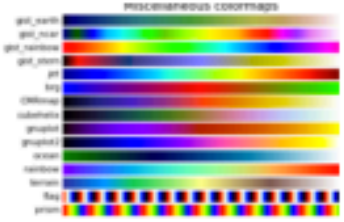
## Color


Set default color cycle to rgby
Set axes color cycle to cmyk


Sequential colormaps


Sequential (2) colormaps

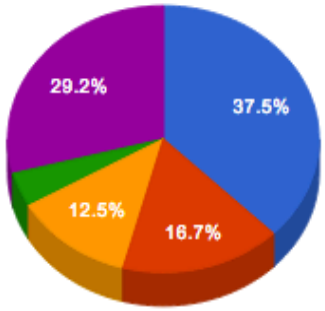

Diverging colormaps


Miscellaneous colormaps

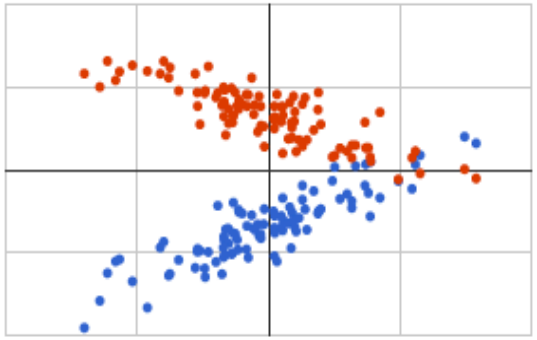
# Google Charts

- Google Charts provides a way to visualize data on your website. From simple line charts to complex hierarchical tree maps, the chart gallery provides a large number of ready-to-use chart types.
- The most common way to use Google Charts is with simple JavaScript that you embed in your web page.
- You load some Google Chart libraries, list the data to be charted, select options to customize your chart, and finally create a chart object with an id that you choose.
- Then, later in the web page, you create a  with that id to display the Google Chart.
- https://developers.google.com/chart/interactive/docs/index

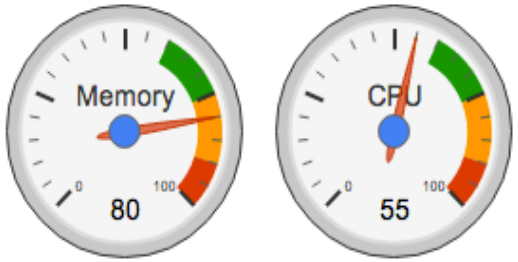
## Pie Chart




## Scatter Chart



## Gauge

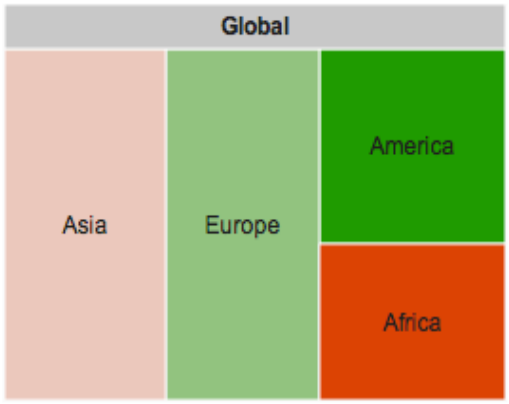



## Geo Chart

## Table

| | Name | Salary | Full Time |
|---|---|---|---|
| 1 | Marie | $24,700 | ✓ |
| 2 | Albert | $25,200 | ✗ |
| 3 | Enrico | $25,700 | ✓ |
| 4 | Lise | $26,600 | ✓ |

## Treemap



# Next class:

- Matplotlib – python
  - http://www.youtube.com/watch?v=3Fp1zn5ao2M&feature=plcp
- Google charts – web development
  - https://developers.google.com/maps/tutorials/visualizing/earthquakes
- Processing – java
  - http://www.youtube.com/watch?v=9UcL8B0GQuE