上海科技大学
ShanghaiTech University

ARTS1422 Data Visualization

Lecture 5

Basics of Data Visualization (II)

Quan Li

Spring 2024

2024. 03.12

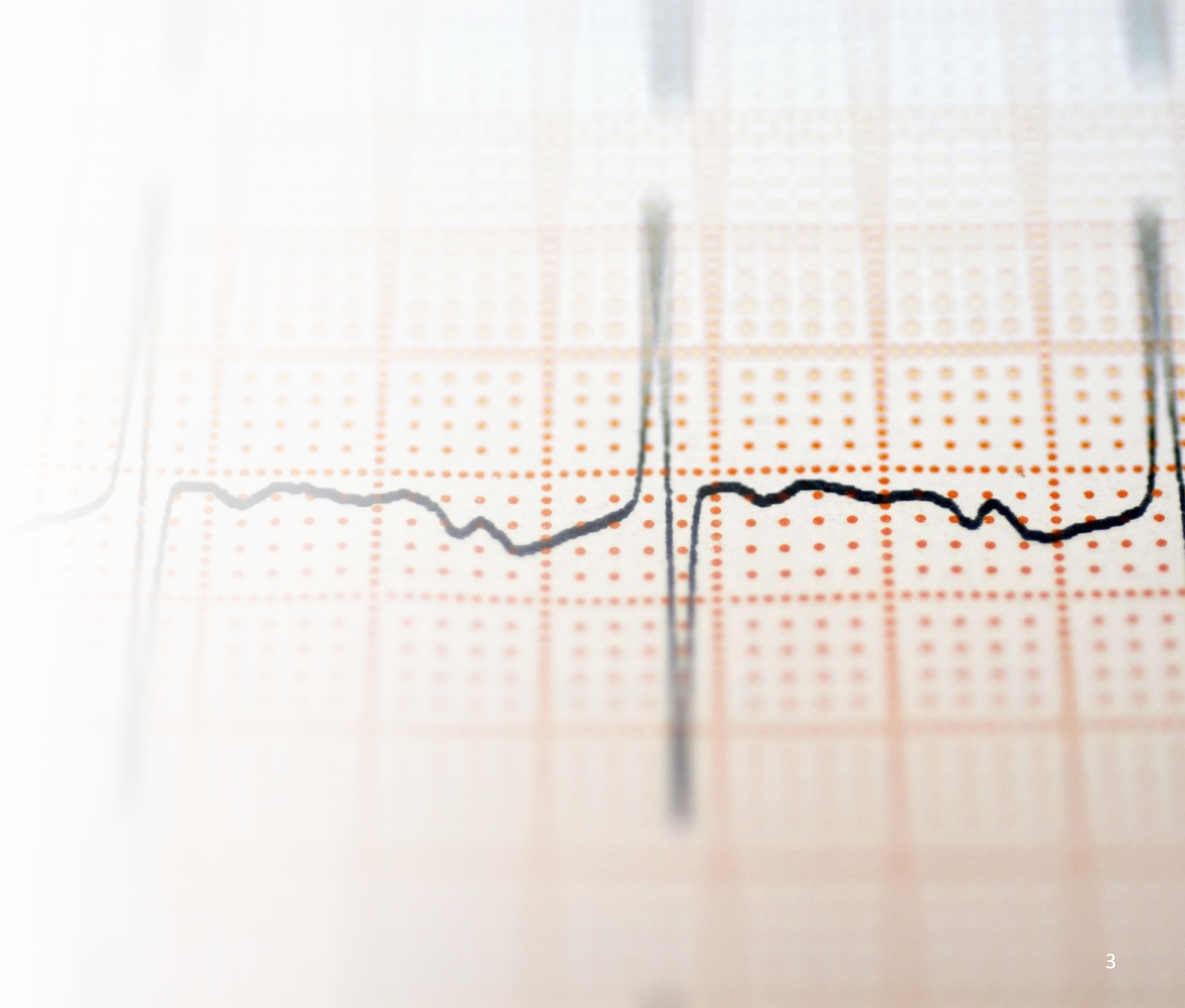
OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Normalization

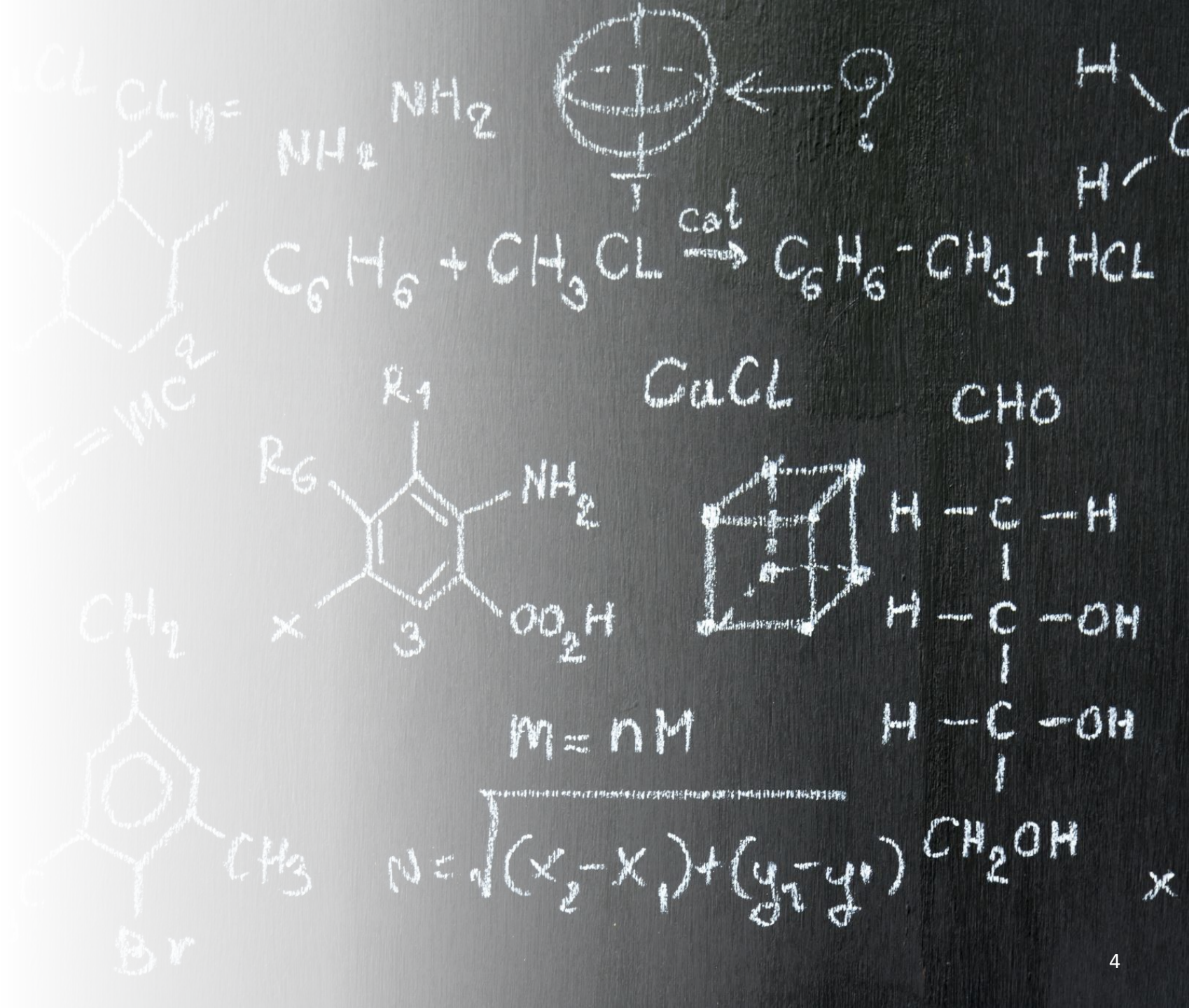
- Why normalize?
  - Mapping data according to its distributions
  - Color/size/coordinate encoding

- [0, 1] normalization
- [-1, 1] normalization



Normalization

- Linear transformation
  - y=(x-MinValue)/(MaxValue-MinValue)
- Arc tangent transformation
  - y=atan(x)*2/PI
- User-defined transformation

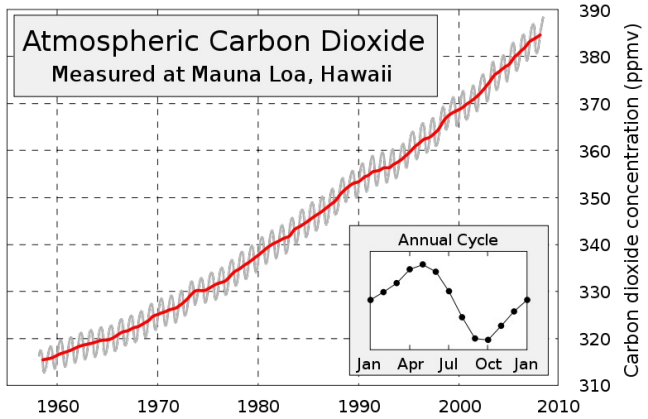


OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Curve Fitting/Smoothness

- Why curve fitting?
  - Finding the trends of data



- Fitting data points to a polynomial curve
  - PLSR: partial least squares regression
  - Locally weighted scatterplot smoothing (LOESS)

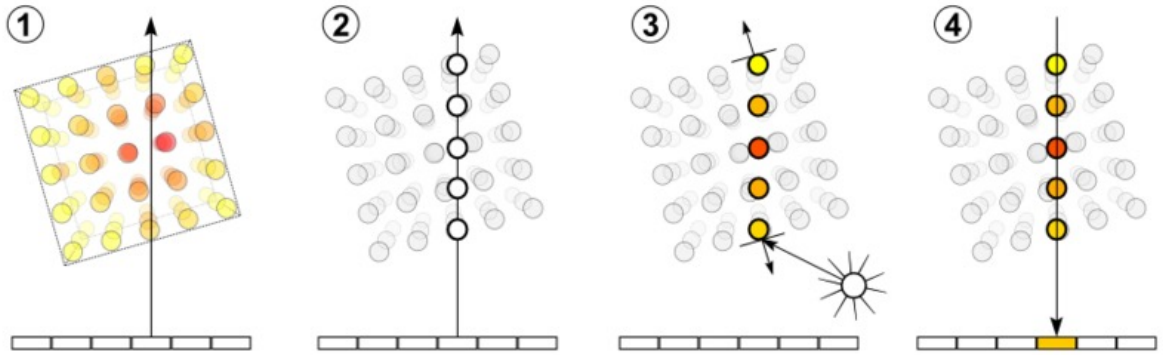
$$\min_{\vec{x}} \sum_{i=1}^{n} (y_m - y_i)^2.$$

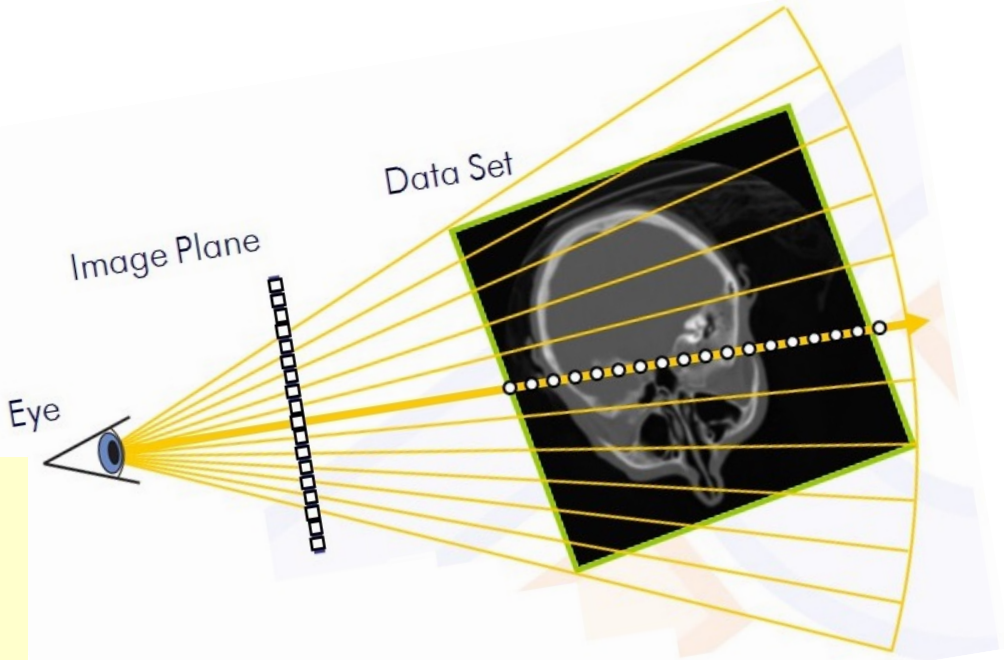
OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Sampling (Signal)

- What is sampling (signal)?
  - In signal field, sampling is the reduction of a continuous signal to a discrete signal.
- Sampling in Volume Rendering





1. Ray casting: For each pixel of the final image, a ray of sight is shot ("cast") through the volume
2. Sampling: Along the part of the ray of sight that lies within the volume, equidistant *sampling points* or *samples* are selected
3. Shading: For each sampling point, a gradient of illumination values is computed
4. Compositing: After all sampling points have been shaded, they are composited along the ray of sight, resulting in the final color value for the pixel that is currently being processed.

Sampling (Statistics)

- What is sampling (statistics)?
  - The selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population.

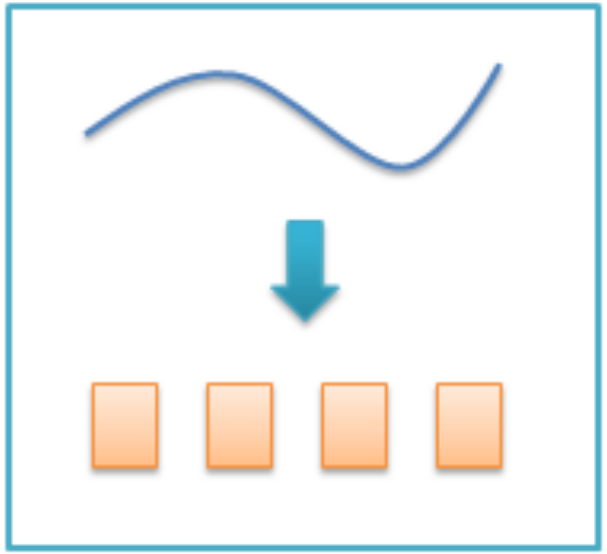
- Influencing factors of various sampling methods
  - Nature and quality of the frame
  - Availability of auxiliary information about units on the frame
  - Accuracy requirements, and the need to measure accuracy
  - Whether detailed analysis of the sample is expected
  - Cost/operational concerns

OUTLINE

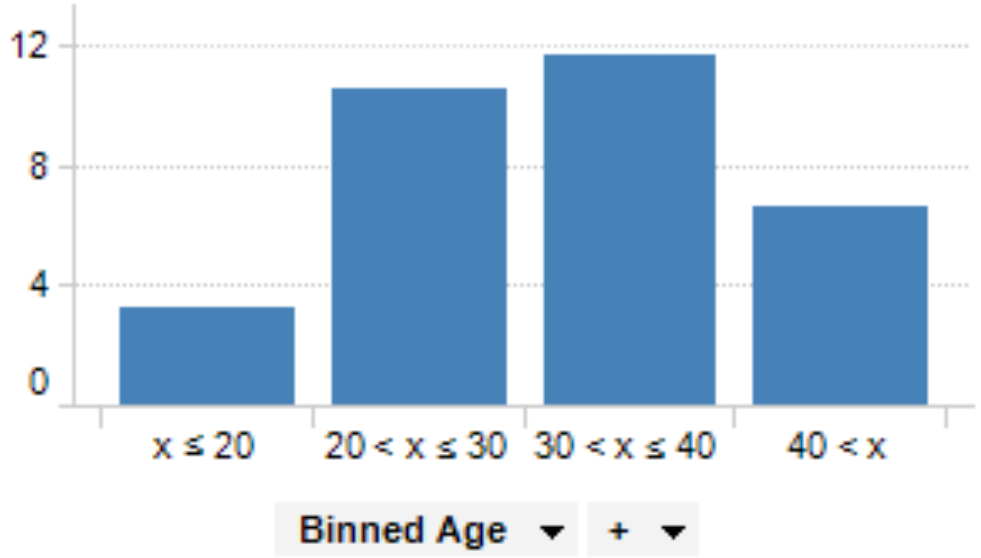
- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Binning/Discretization

- Binning is a way to group a number of more or less continuous values into a smaller number of "bins".



https://www.saedsayad.com/binning.htm



https://docs.tibco.com/pub/spotfire/7.0.1/doc/html/bin/bin_what_is_binning.htm

OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Dimensionality Reduction*

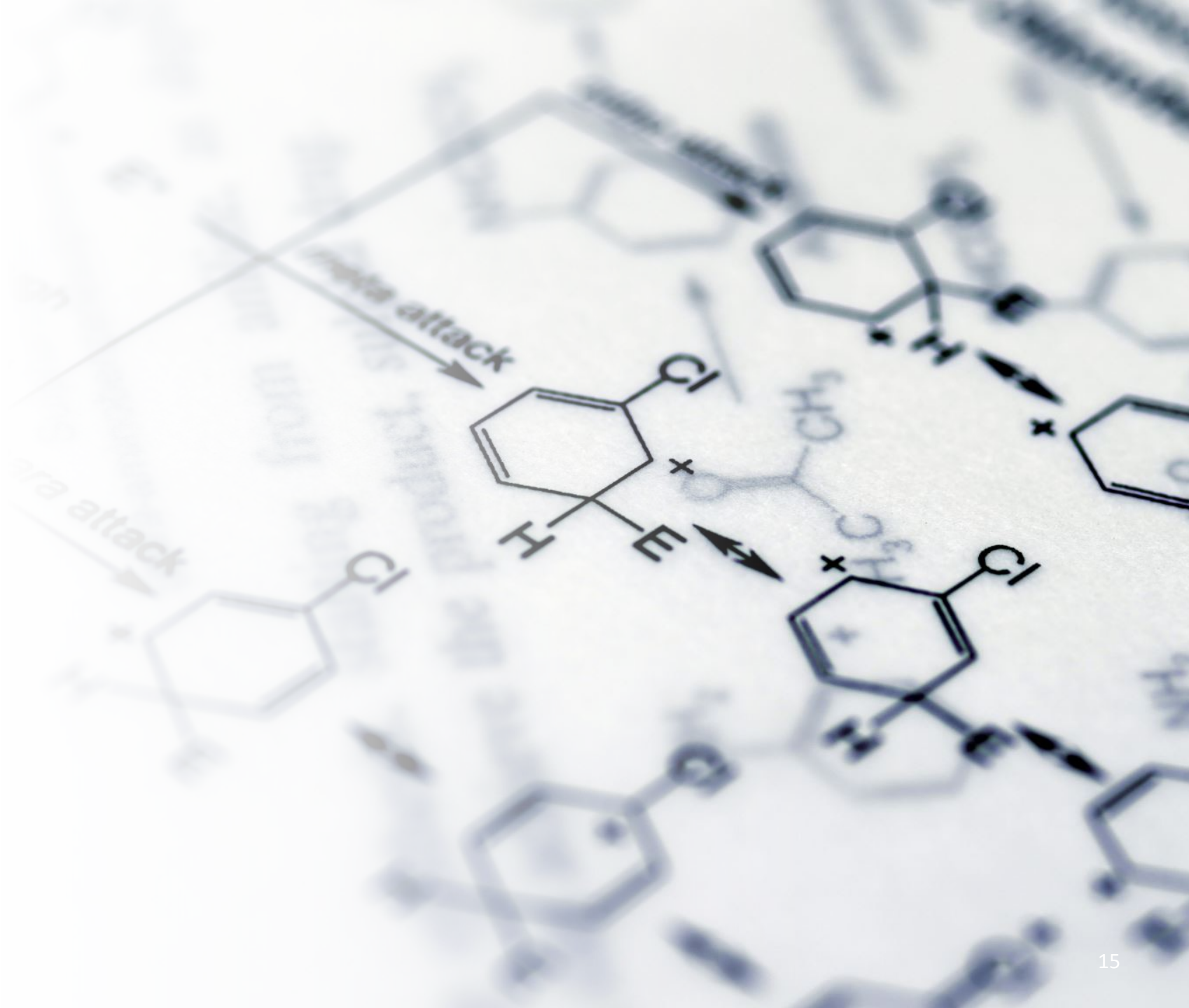
- Principal Components Analysis (PCA)
- Multidimensional Scaling (MDS)
- Self-Organizing Map (SOM)

OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Clustering

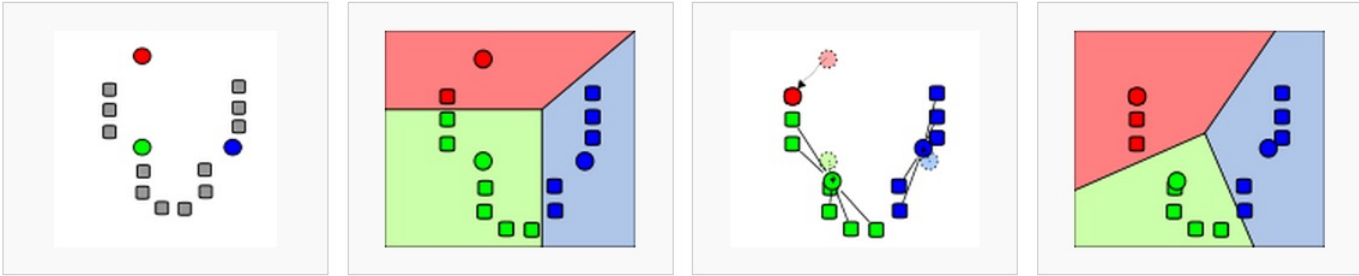
- K-means Clustering
- Expectation-Maximization Clustering (EM)*
- Gaussian Mixture Model (GMM)*
- Spectral Clustering*
- Hierarchical Clustering*



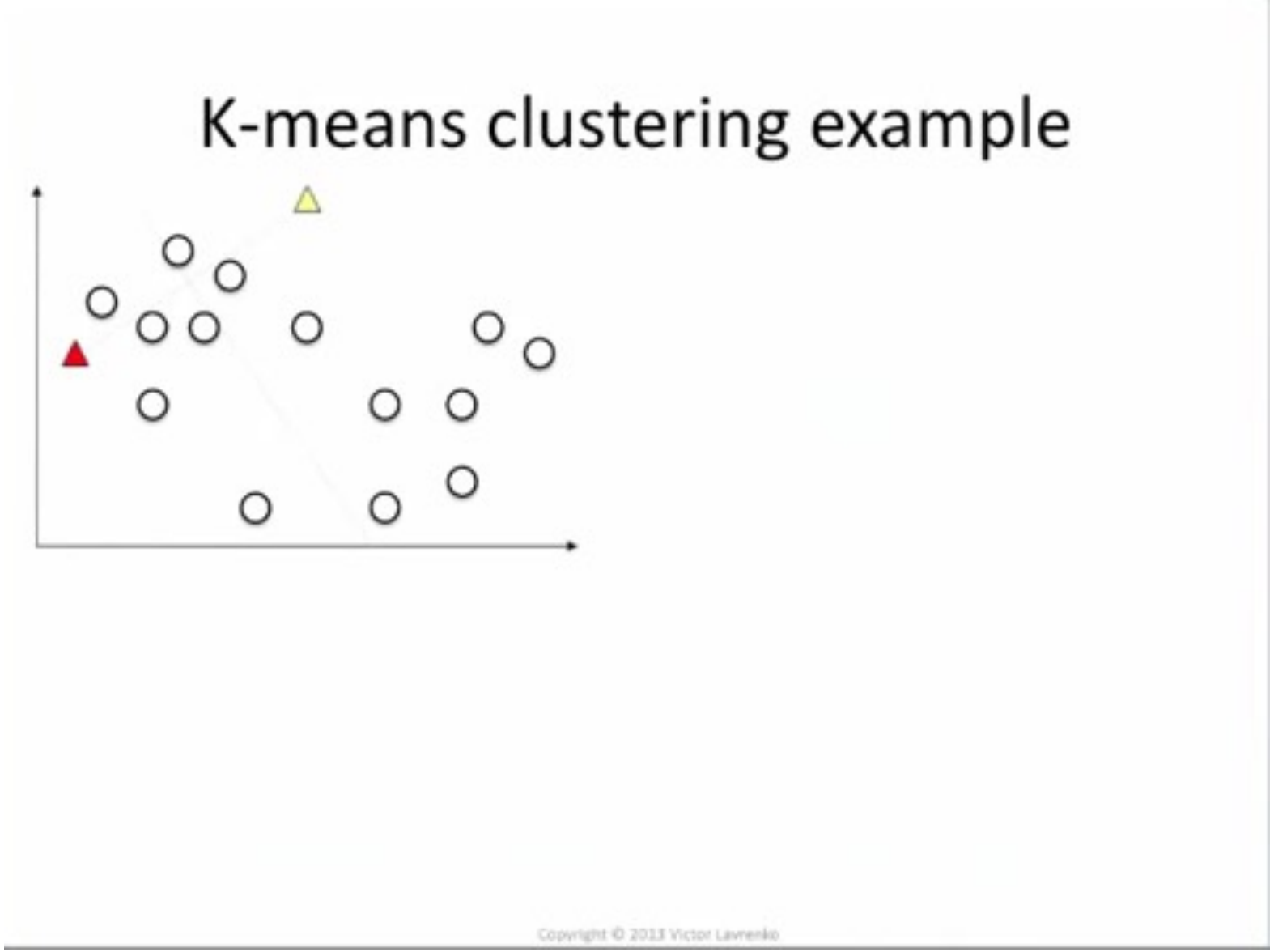
K-Means Clustering

- K-Means
  - k initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).
  - k clusters are created by associating every observation with the nearest mean (virtual points). The partitions here represent the Voronoi diagram generated by the means.
  - The centroid of each of the k clusters becomes the new mean.

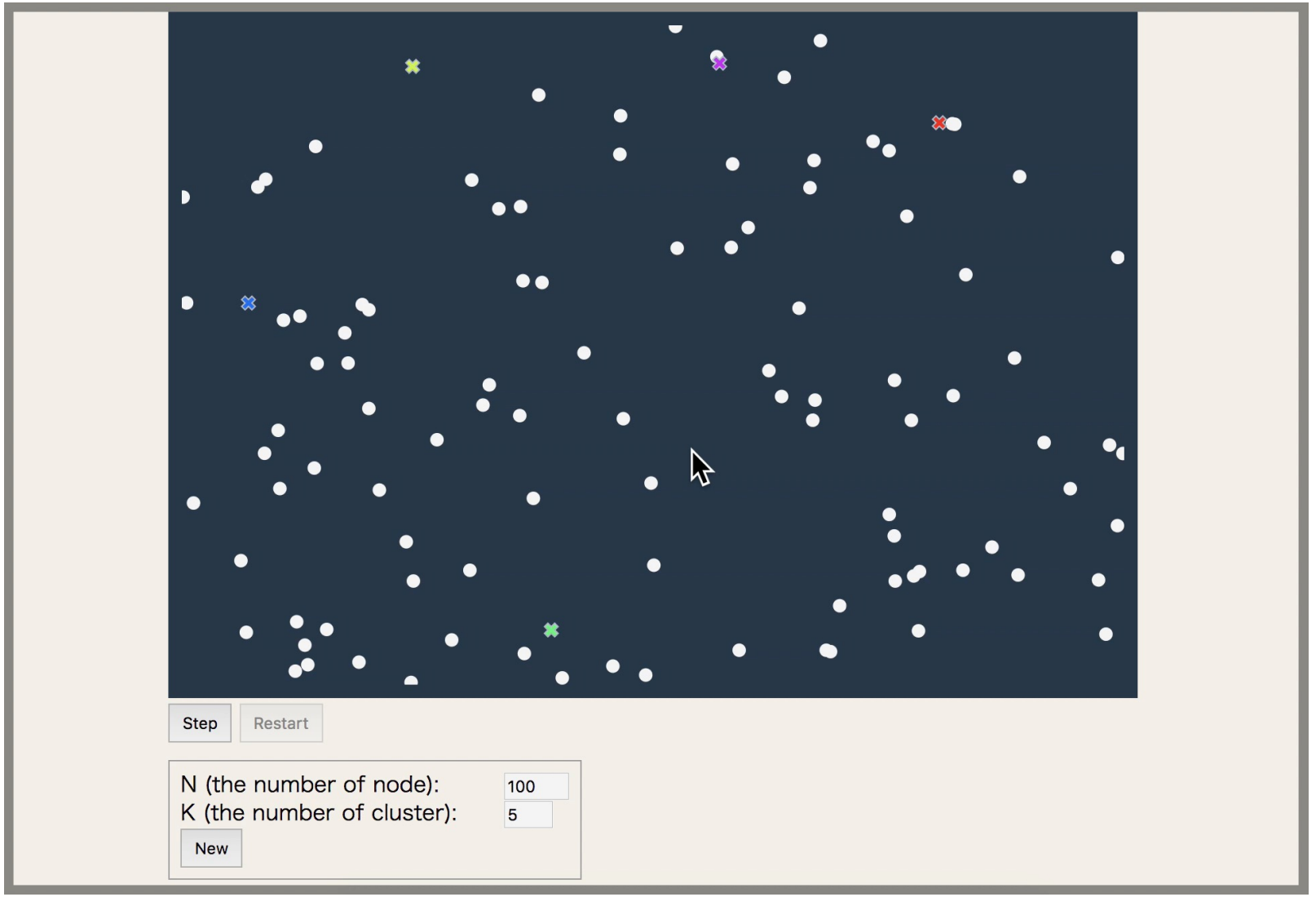
- K-Medoids —centroids are points in the dataset
  - Can be used in non-Euclidean space.



K-Means Clustering - Demo



Visualizing K-Means algorithm with D3.js

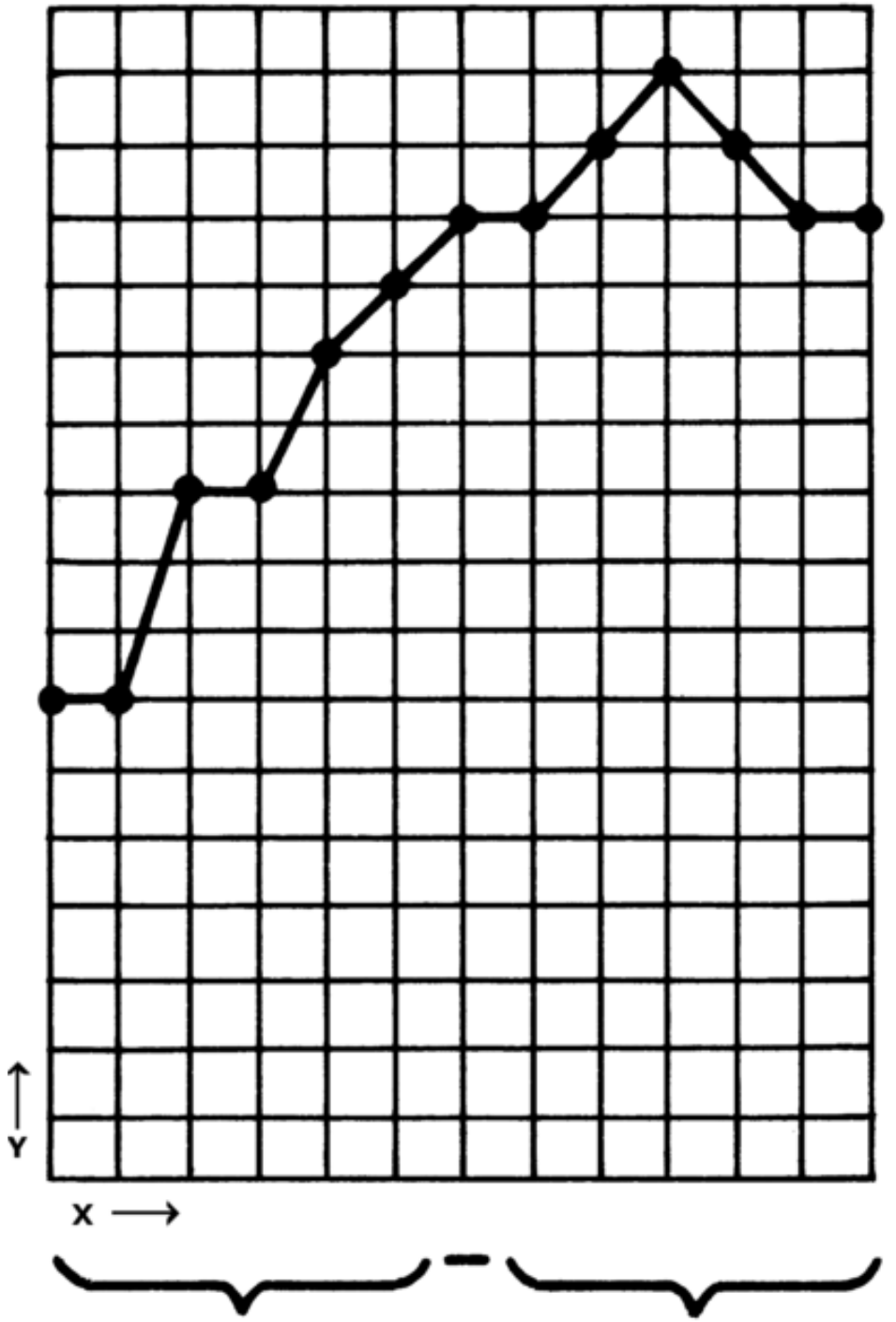


OUTLINE

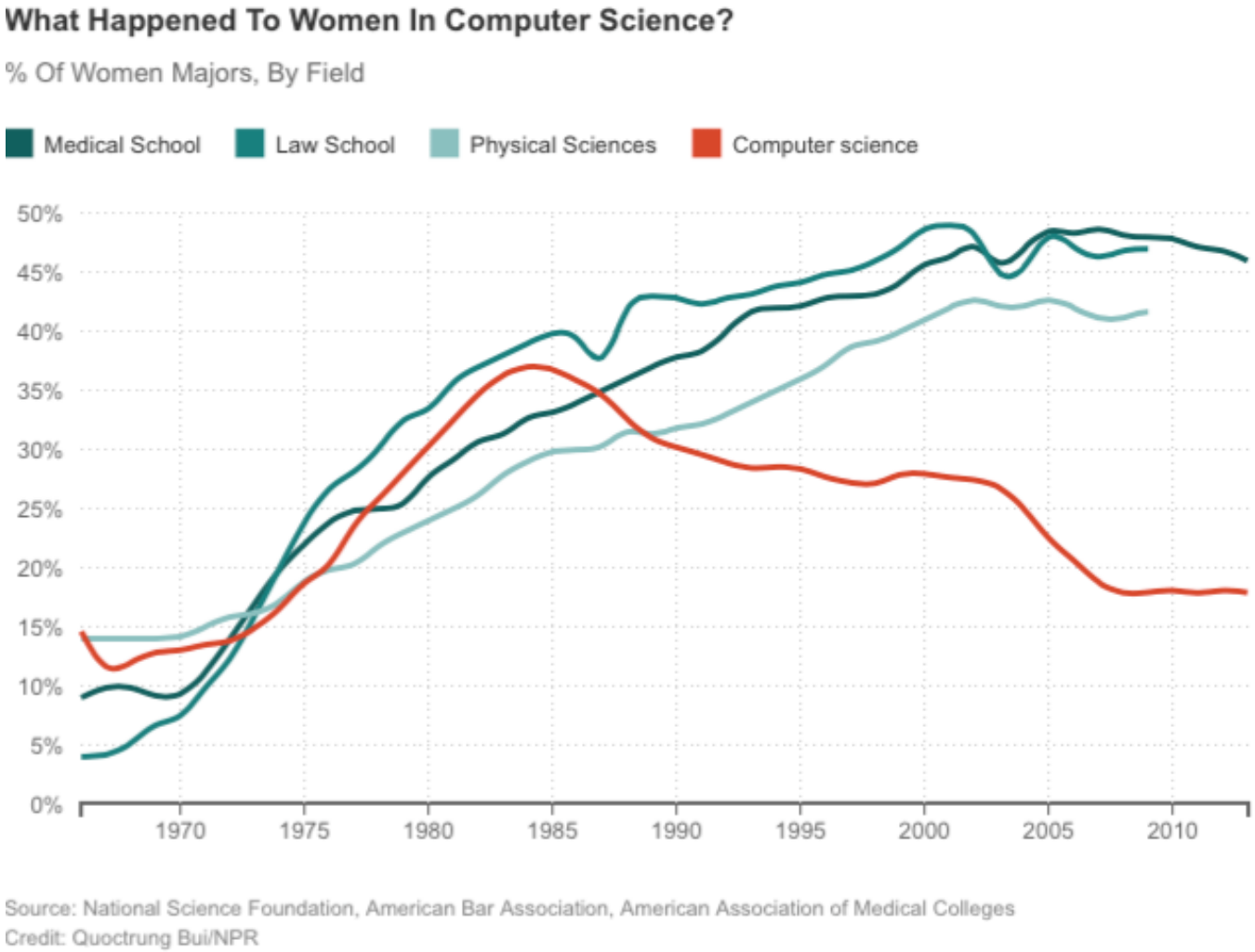
- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Line Chart

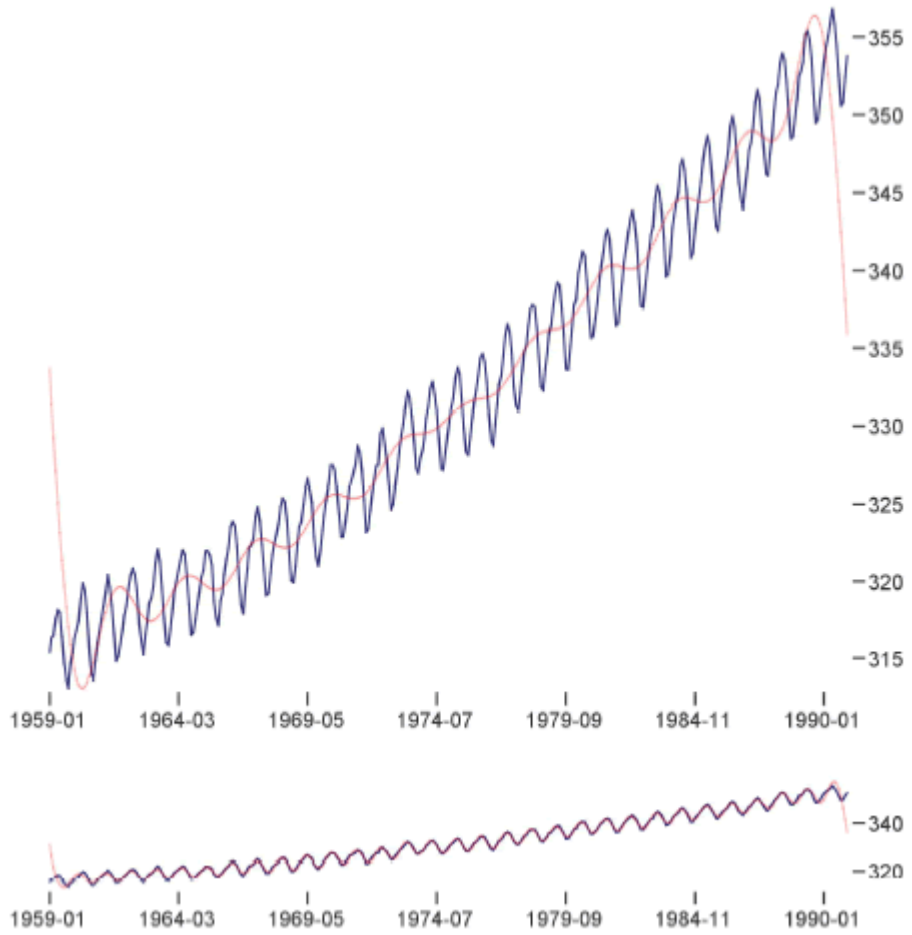
A line chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.



Line Chart


What Happened To Women In Computer Science?
% Of Women Majors, By Field
Medical School
Law School
Physical Sciences
Computer science
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
1970
1975
1980
1985
1990
1995
2000
2005
2010
Source: National Science Foundation, American Bar Association, American Association of Medical Colleges
Credit: Quoctrung Bui/NPR

Aspect Ratio



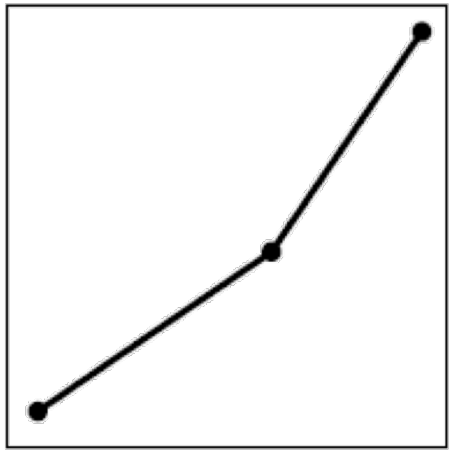
http://vis.stanford.edu/papers/banking

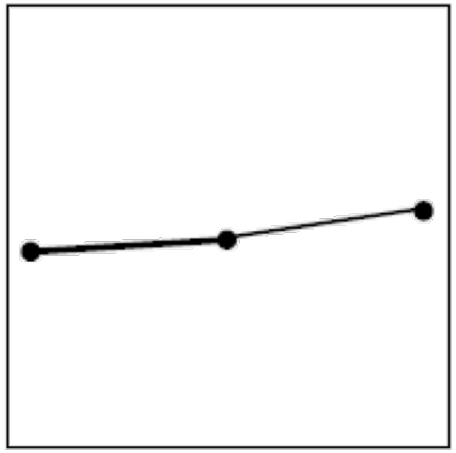
Two plots of monthly atmospheric carbon dioxide measurements, taken from 1959 to 1990. The first plot, with an aspect ratio of 1.17, reveals an accelerating increase in CO_2 levels. The second plot, with an aspect ratio of 7.87, facilitates closer inspection of seasonal fluctuations, revealing a gradual attack followed by a steeper decay. These aspect ratios were automatically determined using multi-scale banking.

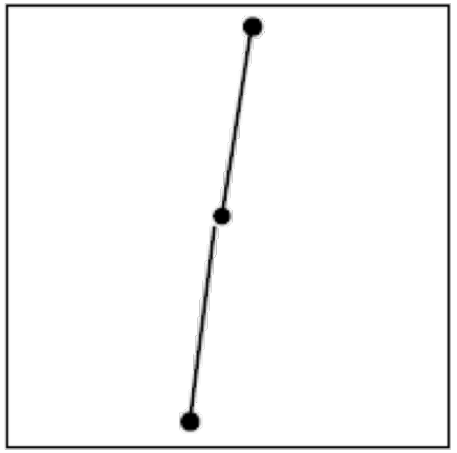
Aspect Ratio

Banking to 45°

The optimized aspect ratio is such that the average absolute orientation of line segments in the chart is equal to 45 degrees.

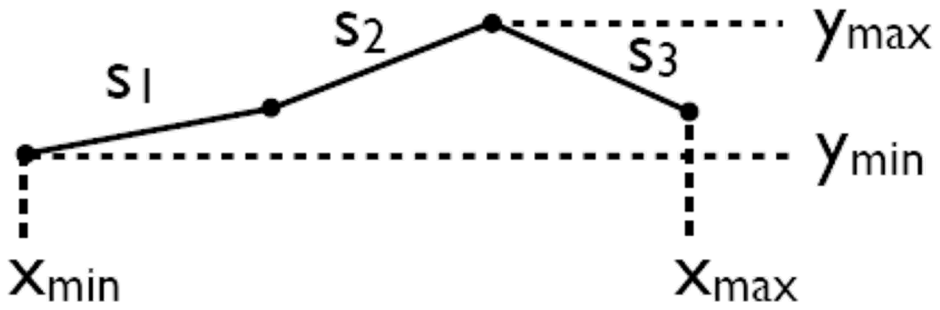






Median-Absolute-Slopes Banking

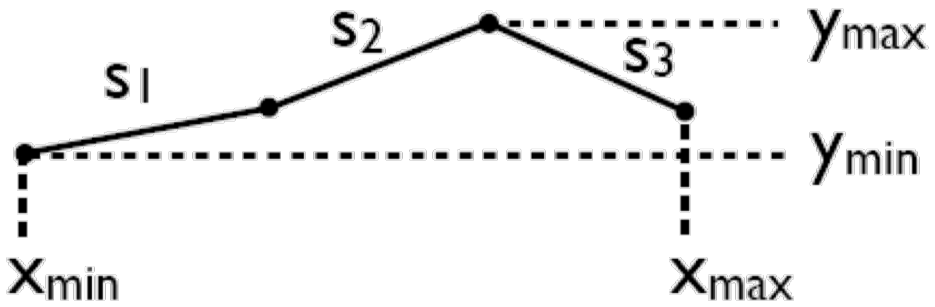
$$s_i = \frac{\Delta y}{\Delta x} \qquad R_x = x_{max} - x_{min} \qquad R_y = y_{max} - y_{min}$$



$$\alpha = \frac{w}{h} = median|s_i|\frac{R_x}{R_y}$$

Average-Absolute-Slopes Banking

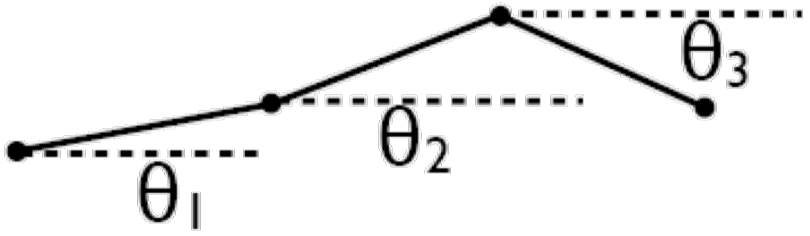
$$s_i = \frac{\Delta y}{\Delta x} \qquad R_x = x_{max} - x_{min} \qquad R_y = y_{max} - y_{min}$$



$$\alpha = \frac{w}{h} = mean|s_i|\frac{R_x}{R_y}$$

Average-Absolute-Orientation Banking

$$\theta_i(\alpha) = \tan^{-1}(s_i/\alpha)$$



$$\sum_i \frac{|\theta_i(\alpha)|}{n} = 45^o$$

Aspect Ratio Banking

- Average-Absolute-Slopes Banking

$$\alpha = mean|s_i|\frac{R_x}{R_y}$$

- Median-Absolute-Slopes Banking

$$\alpha = median|s_i|\frac{R_x}{R_y}$$

- Average-Absolute-Orientation Banking

$$\sum_i \frac{|\theta_i(\alpha)|}{n} = 45^o$$

Multi-Scale Banking to 45º

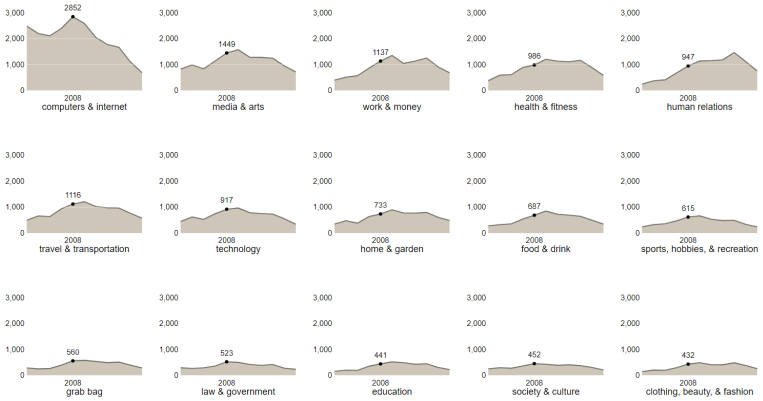
- Objective
  - maximize the discriminability of the orientations of the line segments in the chart
- Methods
  - Automatically identify frequency scales of interest
  - Generate trend curves for identified scales of interest
  - Bank the curves to 45º
  - cull similar aspect ratios

Multi-Scale Banking to 45º

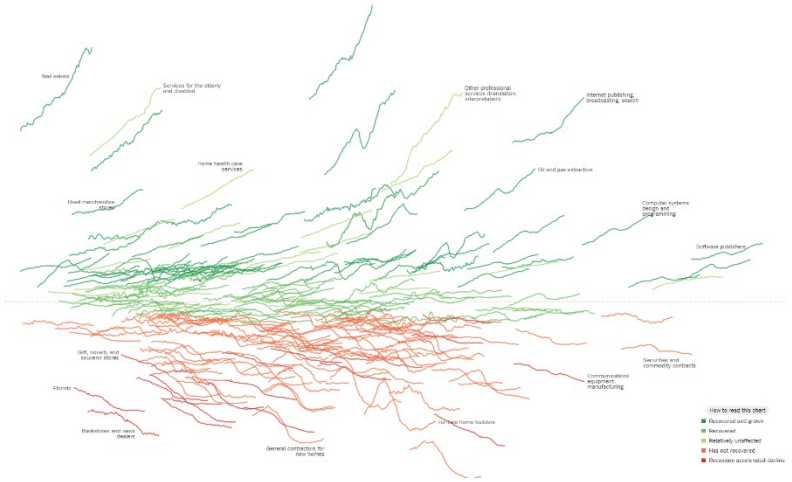
- Search for trends of interest in the power spectrum
  - retain only the highest-frequency value
  - Low-pass filter the data to form the trend curves

Applications

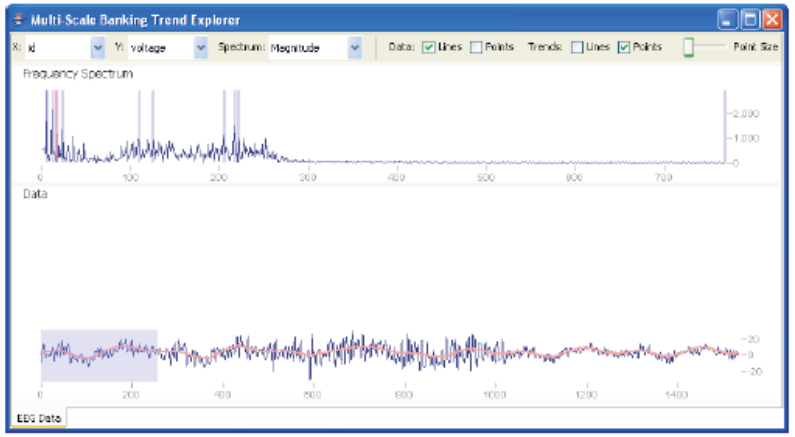
small multiples

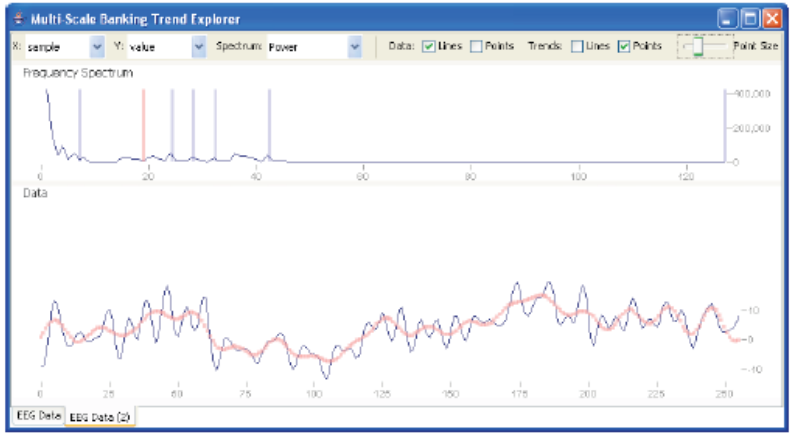


sparkline



Trend Explorer application





OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Sparklines

- A sparkline is a very small line chart, typically drawn without axes.

Using d3.js, we can fairly easily draw SVG-based sparklines. This is 2013 historical stock prices for **Google**  **$1084.75**. And this is for **Facebook**  **$55.57**. And this is for **Apple**  **$550.77**. Each sparkline has 244 data points, but it's condensed very nicely.

	A	B	C	D	E	F	G	H	I
1									
2	5	6	4	8	2	7			
3	2	5	7	8	9	0			
4	9	4	6	3	5	1			

- Sparklines are small enough to be embedded in text, or several sparklines may be grouped together as elements of a small multiple.

Sparklines in Small Multiple

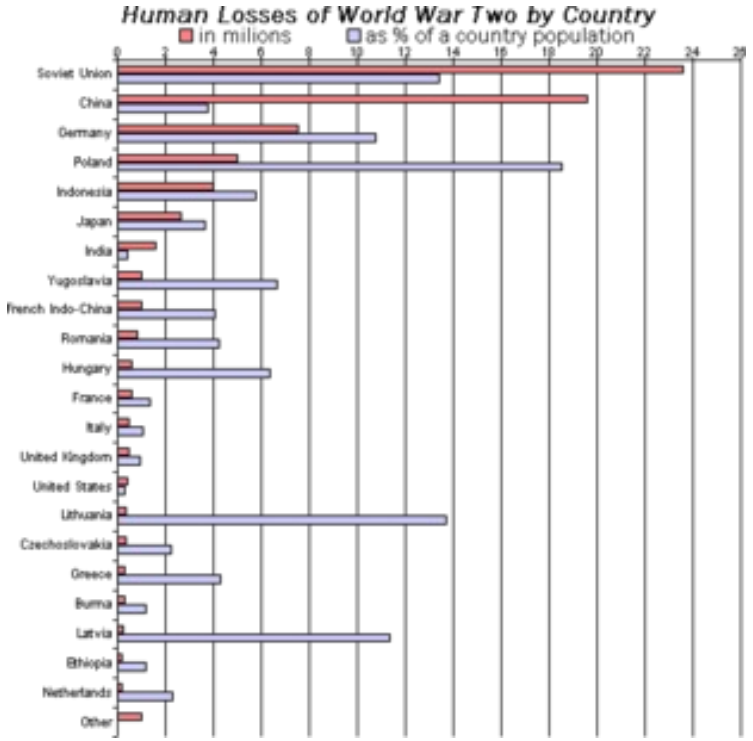
http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Bar Chart

A bar chart or bar graph is a chart that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

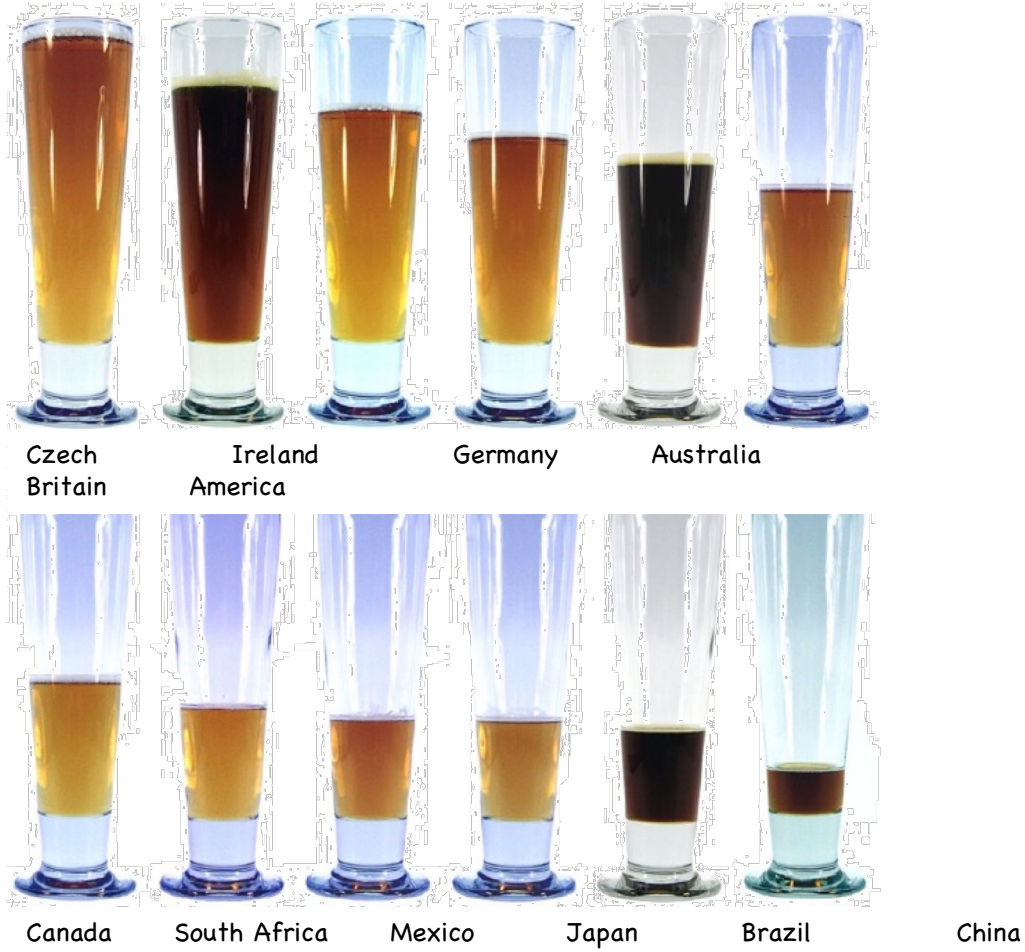


https://en.wikipedia.org/wiki/Bar_chart

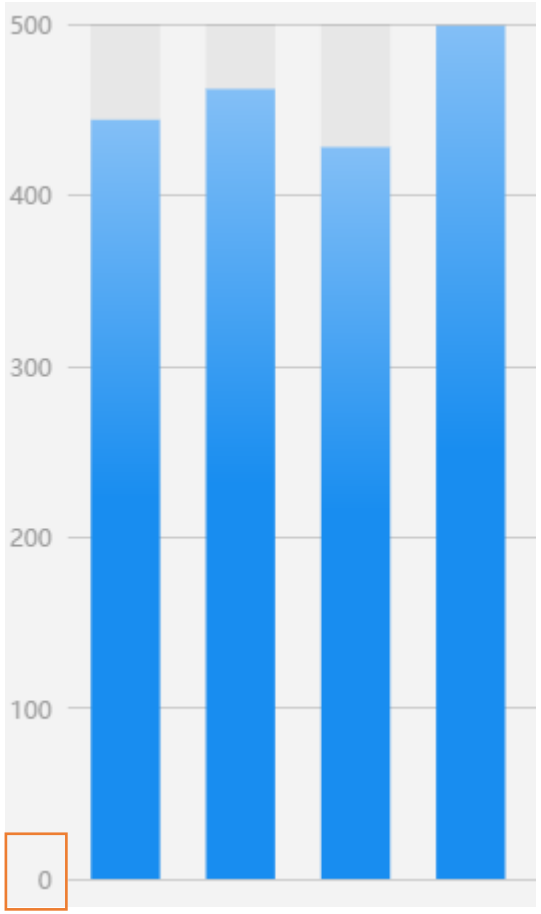
Bar Chart

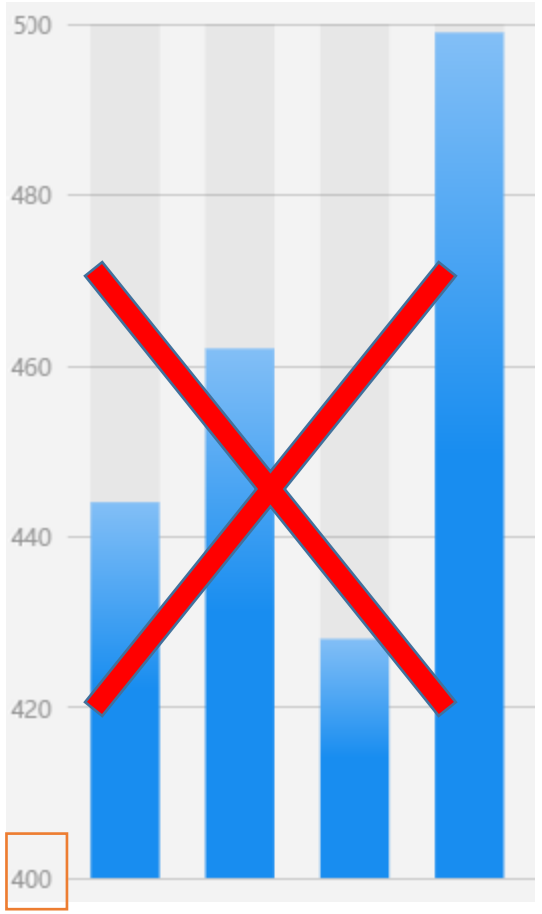
How much beer does each country consume?

Average beer bottles per one per week

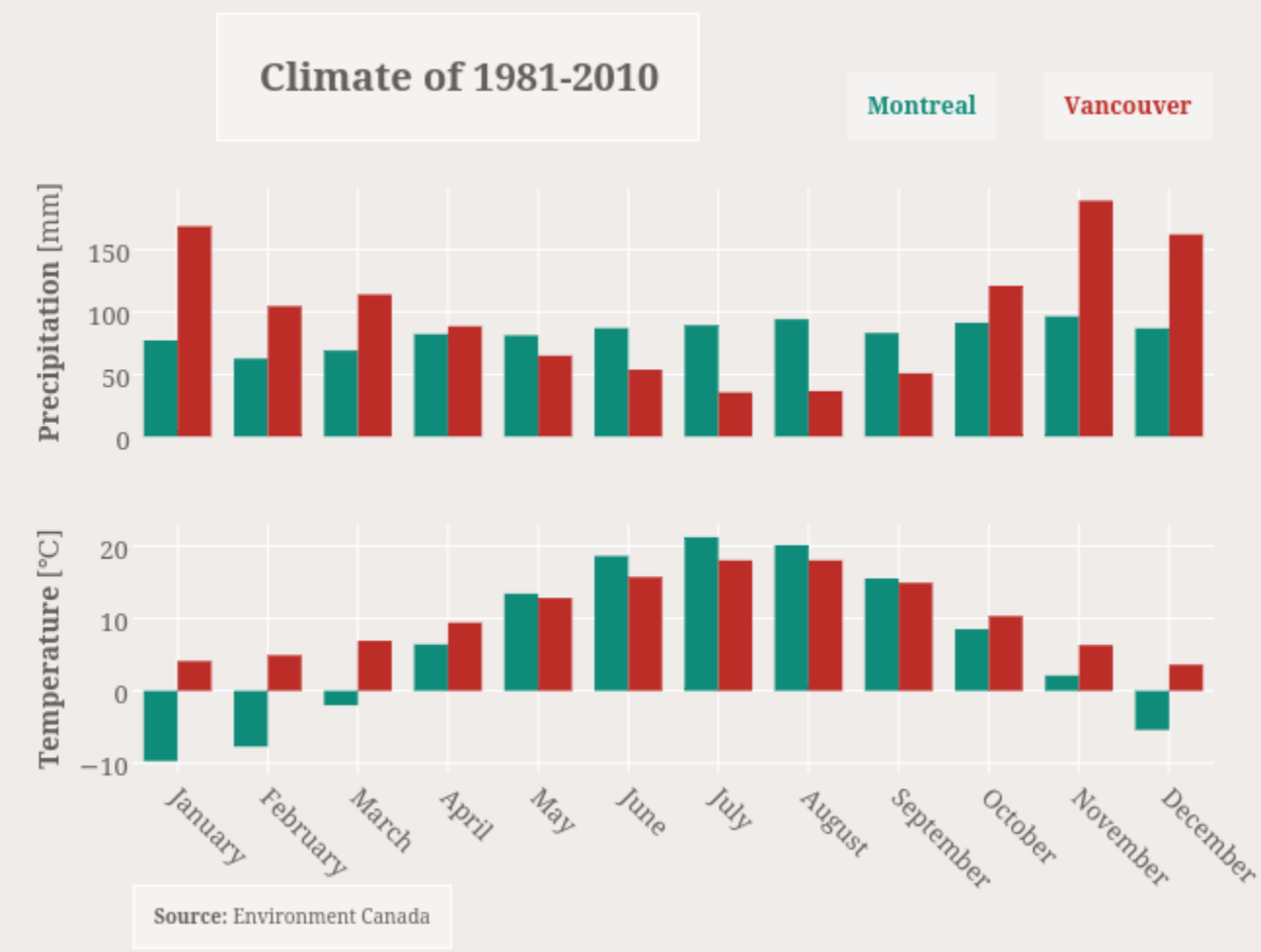


Scales

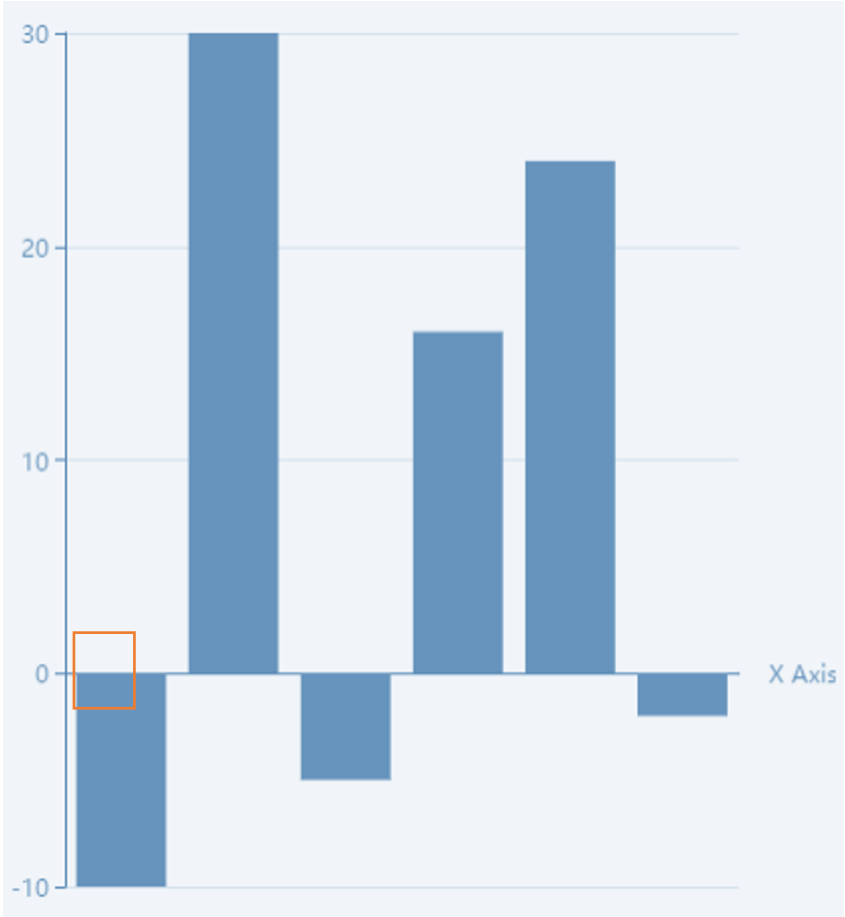


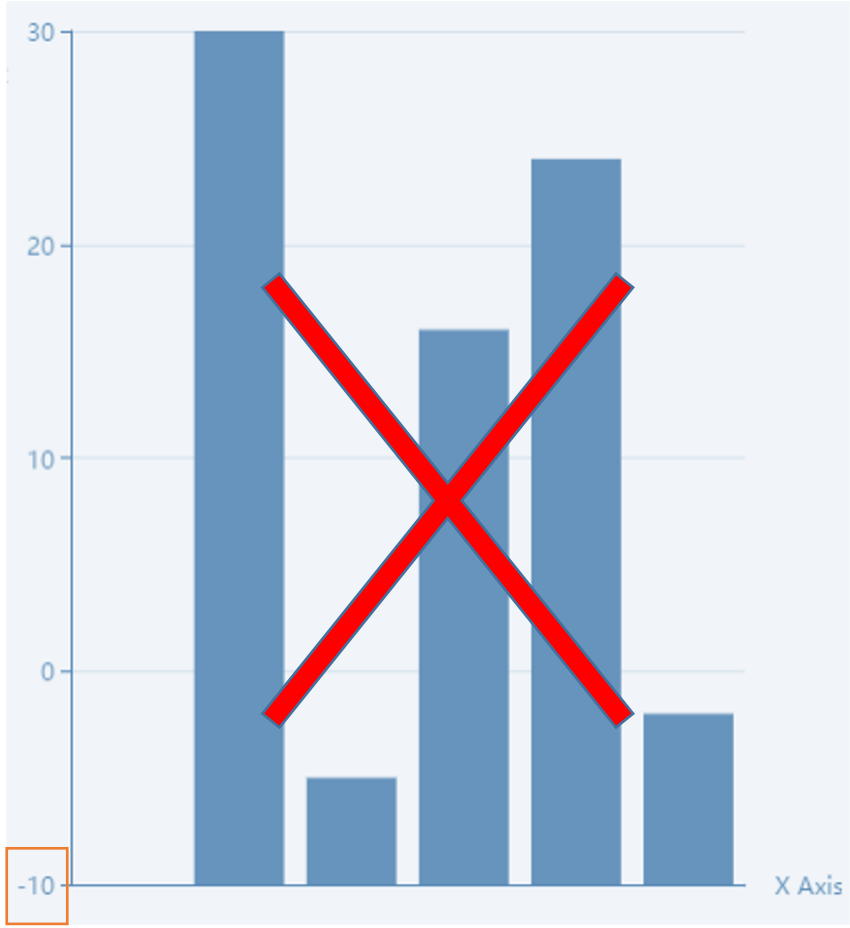


Deviation Design

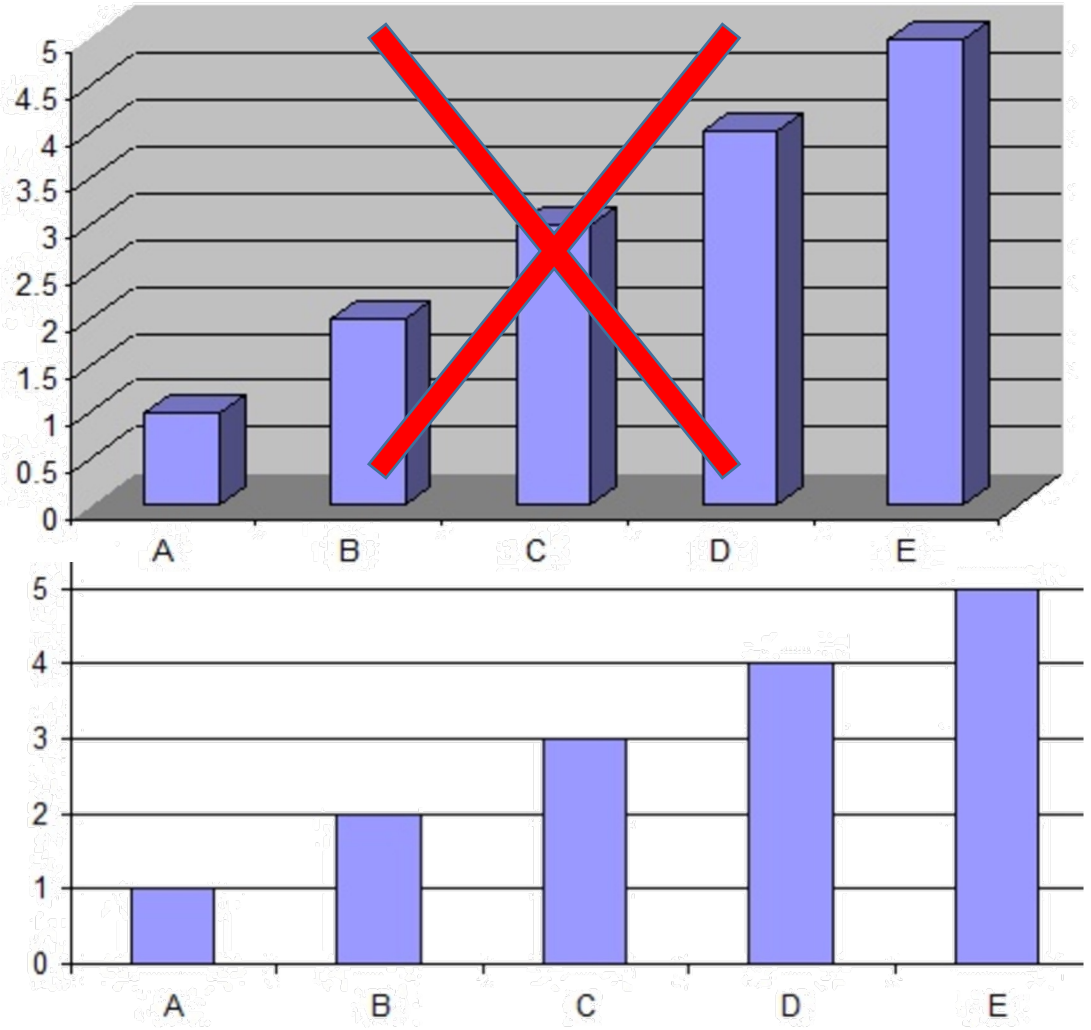


Use Zero as Base





3D is Usually not Preferred

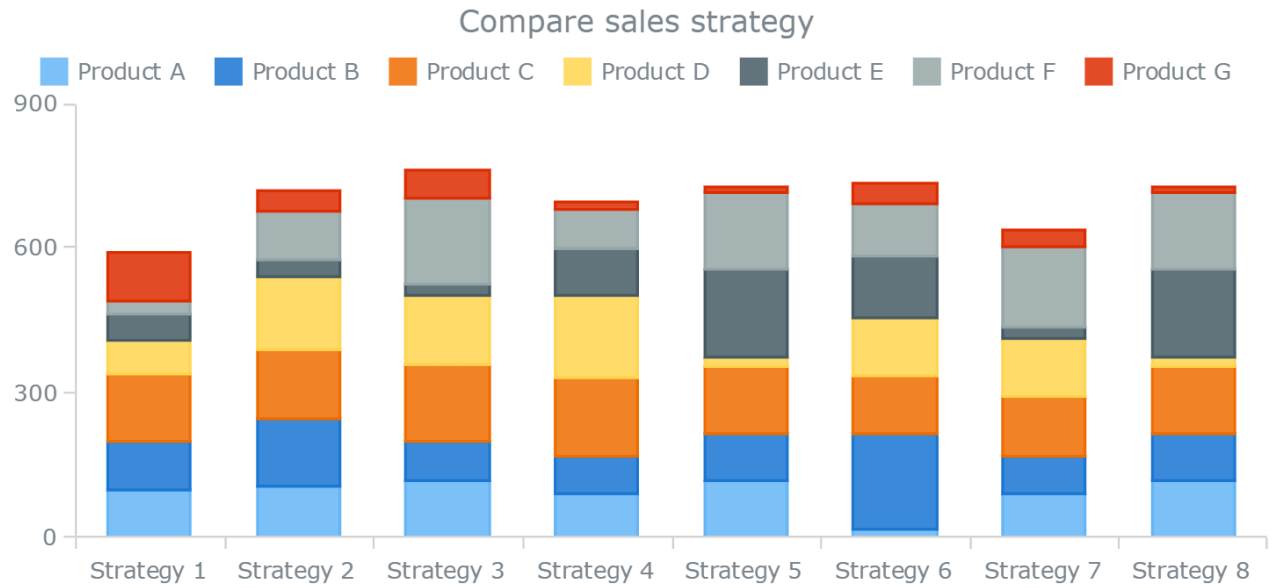


OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Stacked Bar Chart

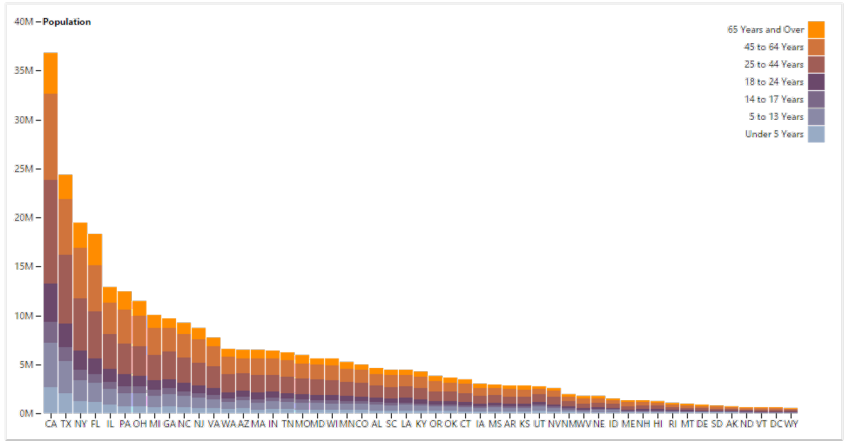
A stacked bar chart, also known as a stacked bar graph, is a graph that is used to break down and compare parts of a whole. Each bar in the chart represents a whole, and segments in the bar represent different parts or categories of that whole.



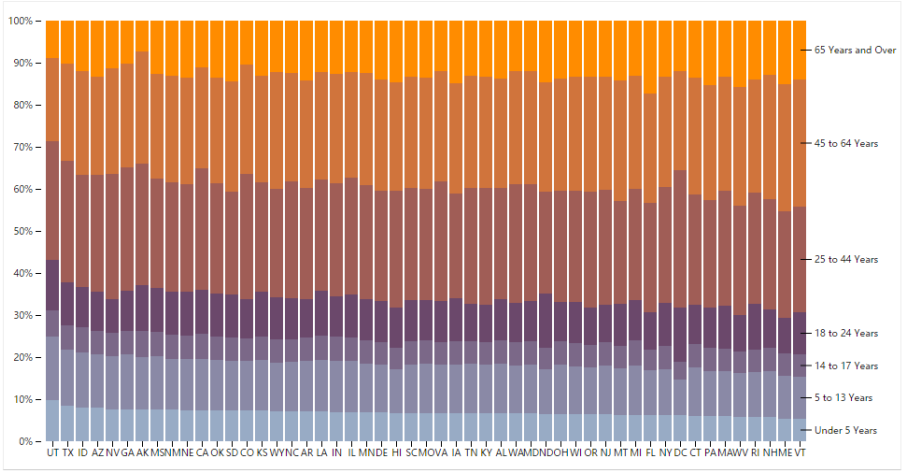
https://www.smashingmagazine.com/2017/03/understanding-stacked-bar-charts/

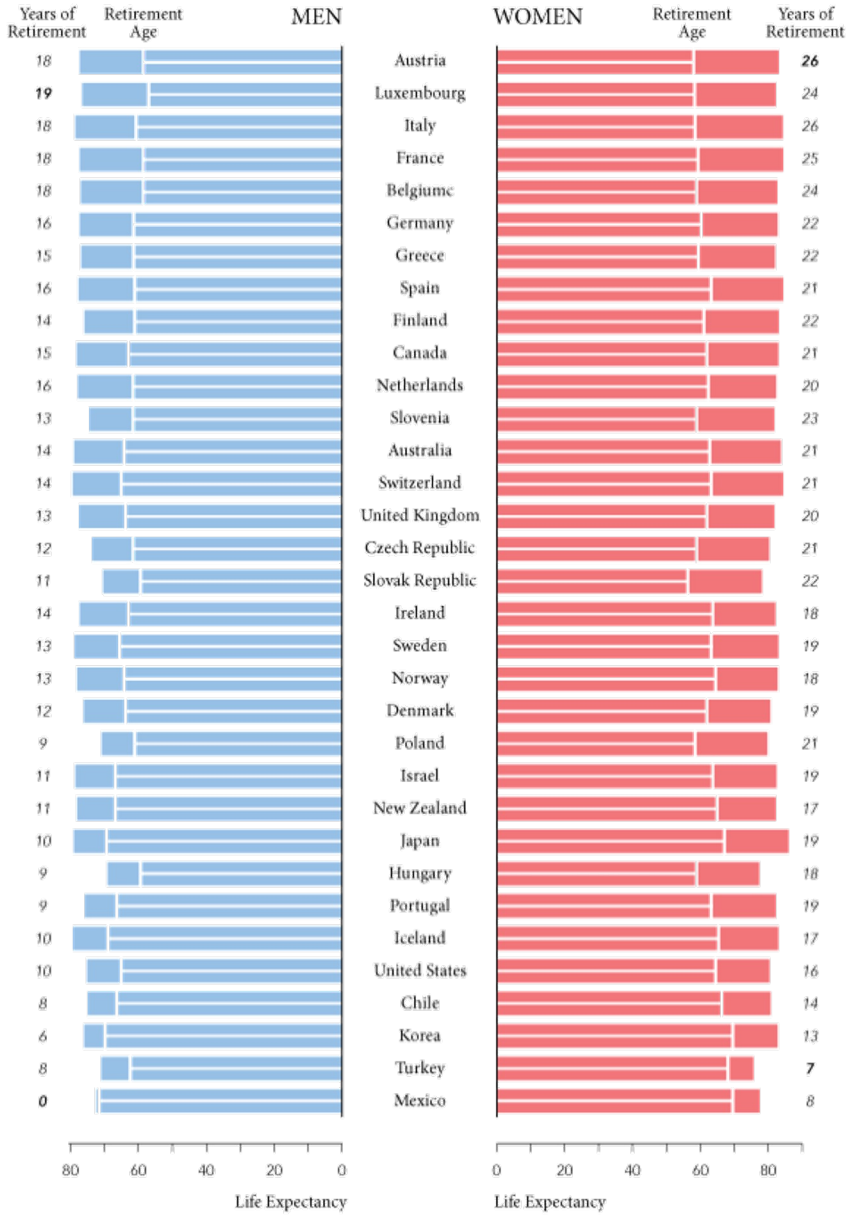
Stacked Bar Chart

Stacked Bar Chart



Normalized Stacked Bar Chart





Paired Bar Chart

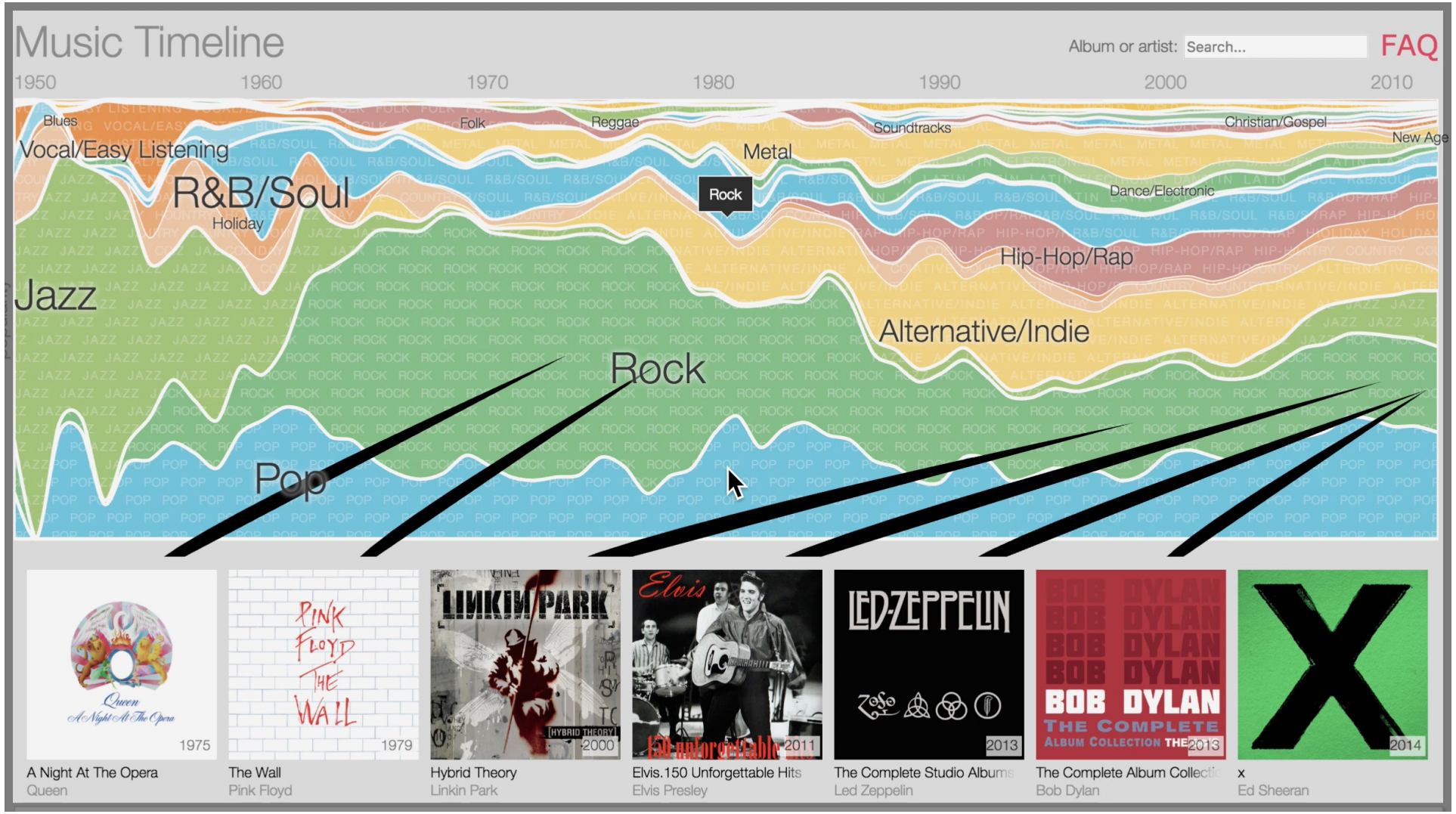
Life Expectancy
&
Retirement Years

http://Flowingdata.com

You call it
stacked bar chart?

http://www.lokeshdhakar.com/2007/08/20/an-illustrated-coffee-guide/

Stacked Chart



Music timeline of plays and history

http://research.google.com/bigpicture/music/#

OUTLINE

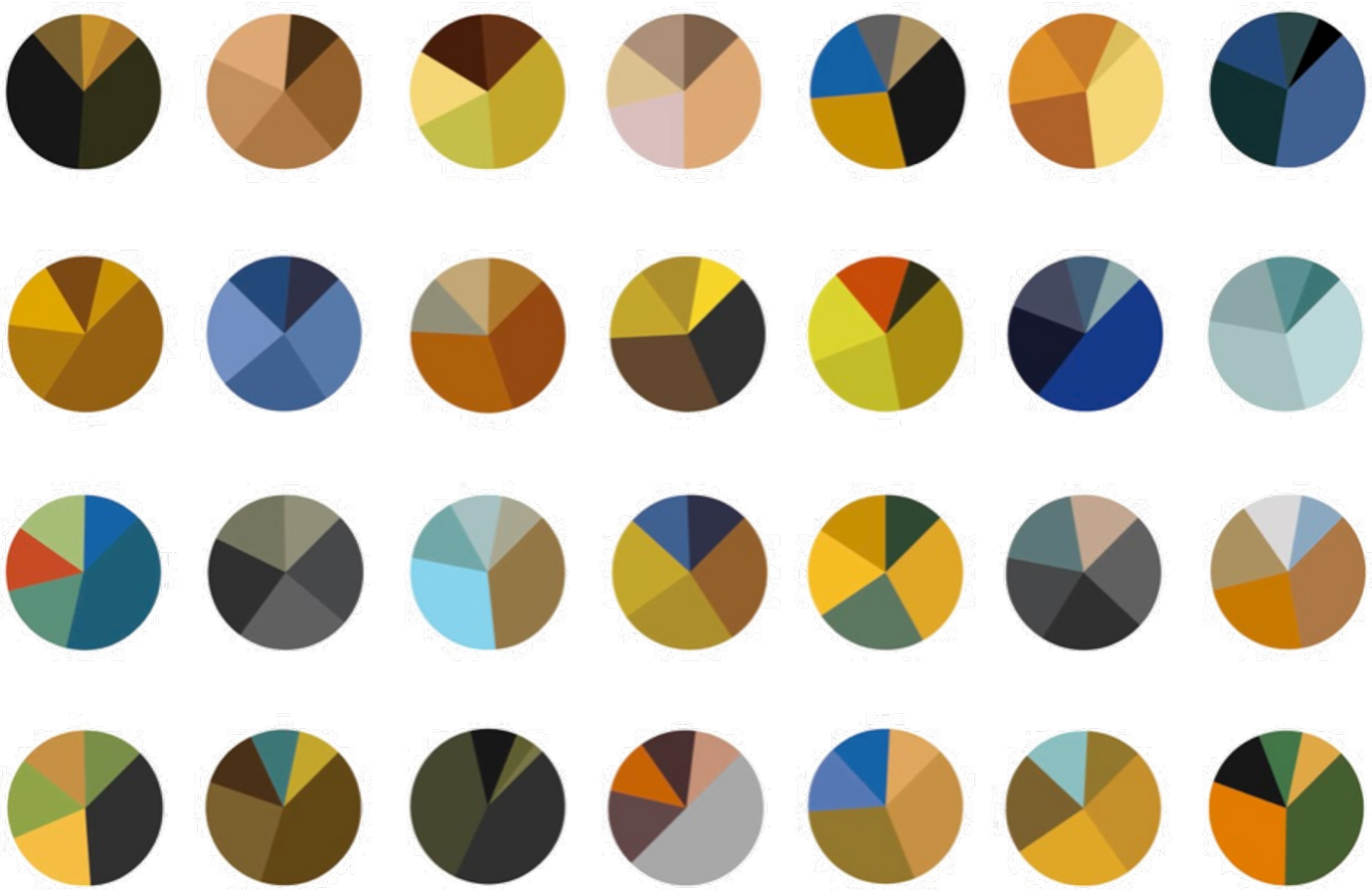
- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Pie Chart

Percentage data

Van Gogh's Paintings

Van Gogh Visualization

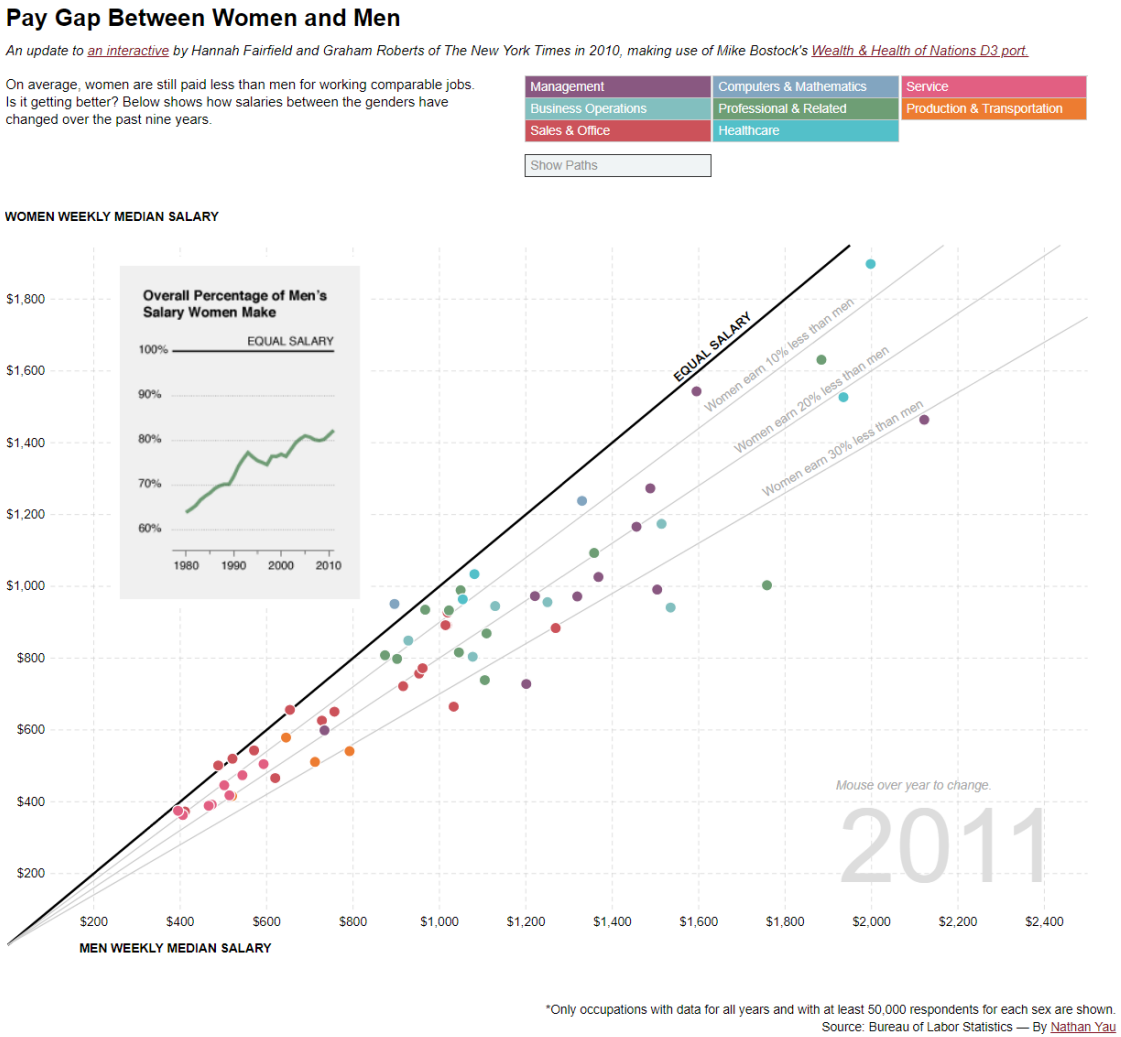


http://www.arthurbuxton.com/2010/11/van-gogh-visualisation.html

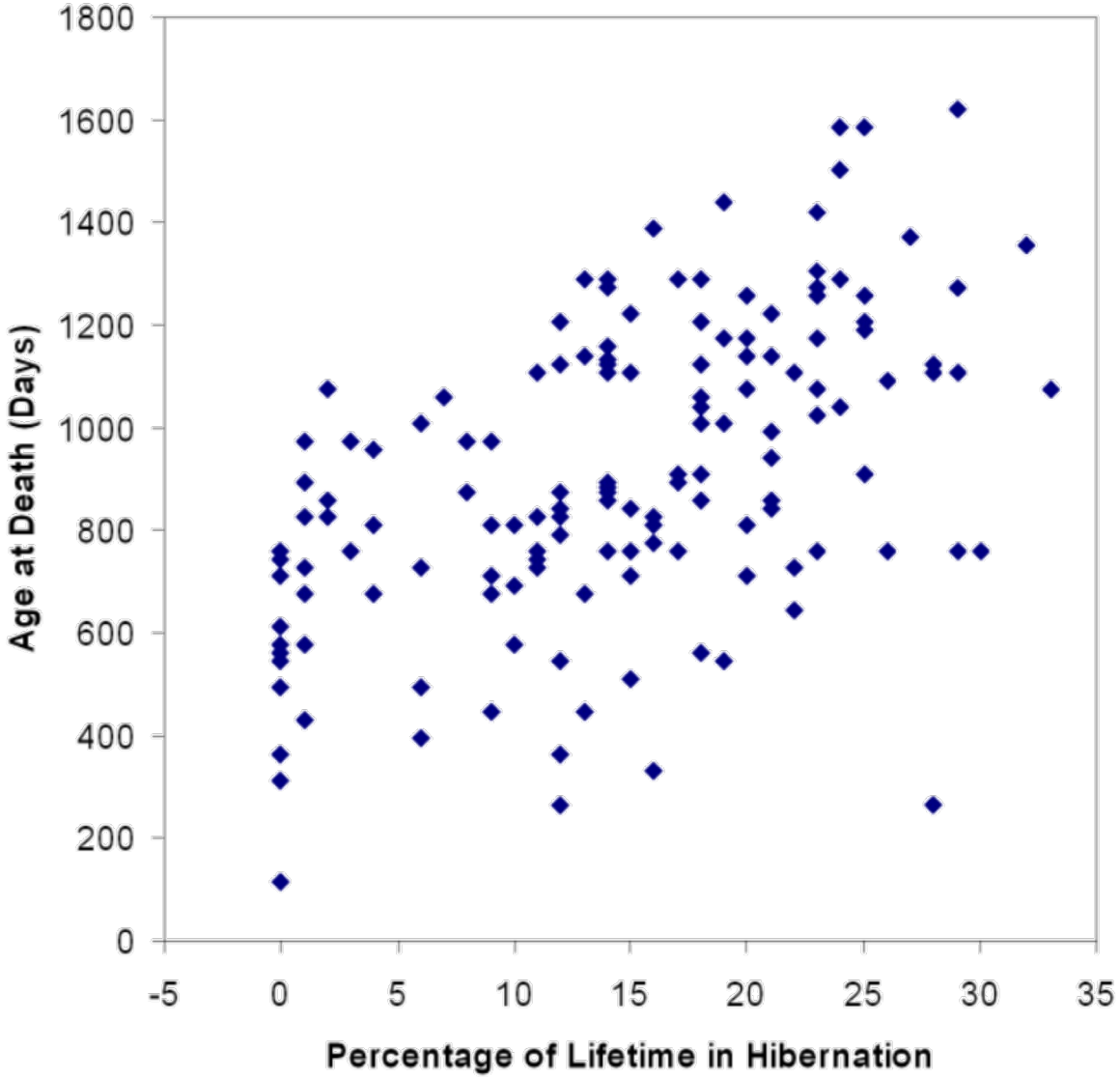
OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

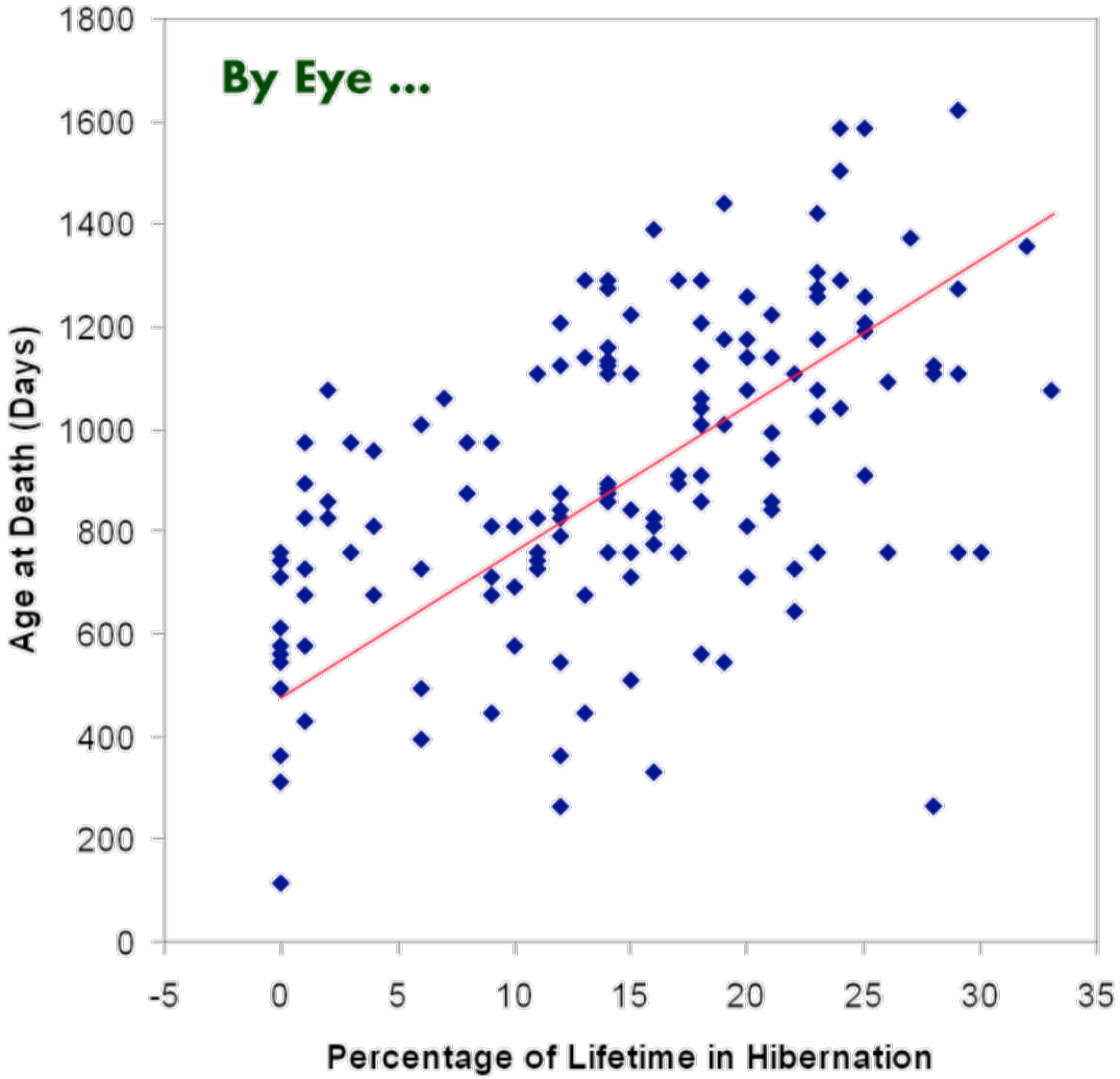
Scatter Plot



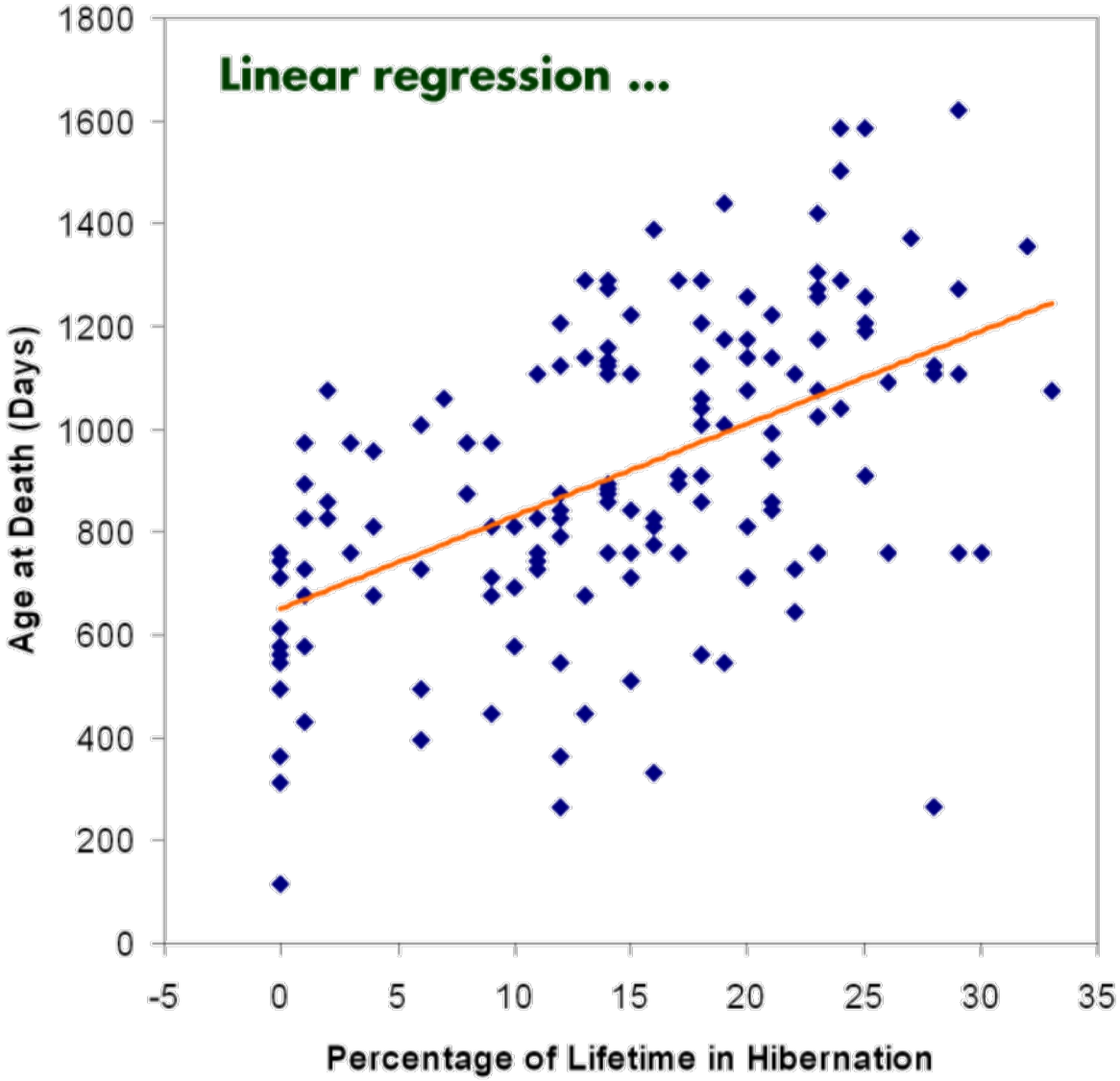
http://projects.flowingdata.com/salary/



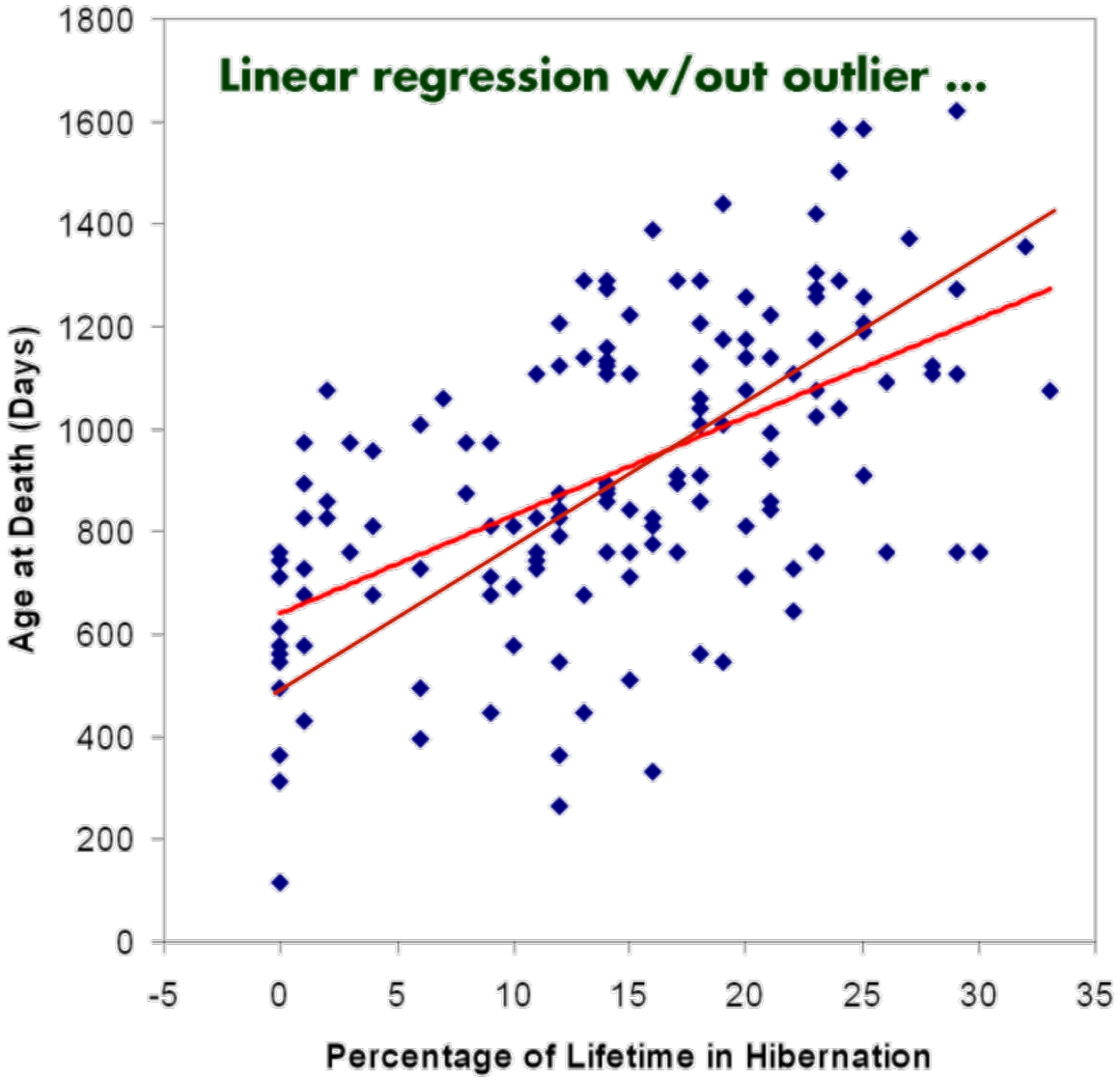
[The Elements of Graphing Data. Cleveland 94]



[The Elements of Graphing Data. Cleveland 94]



[The Elements of Graphing Data. Cleveland 94]

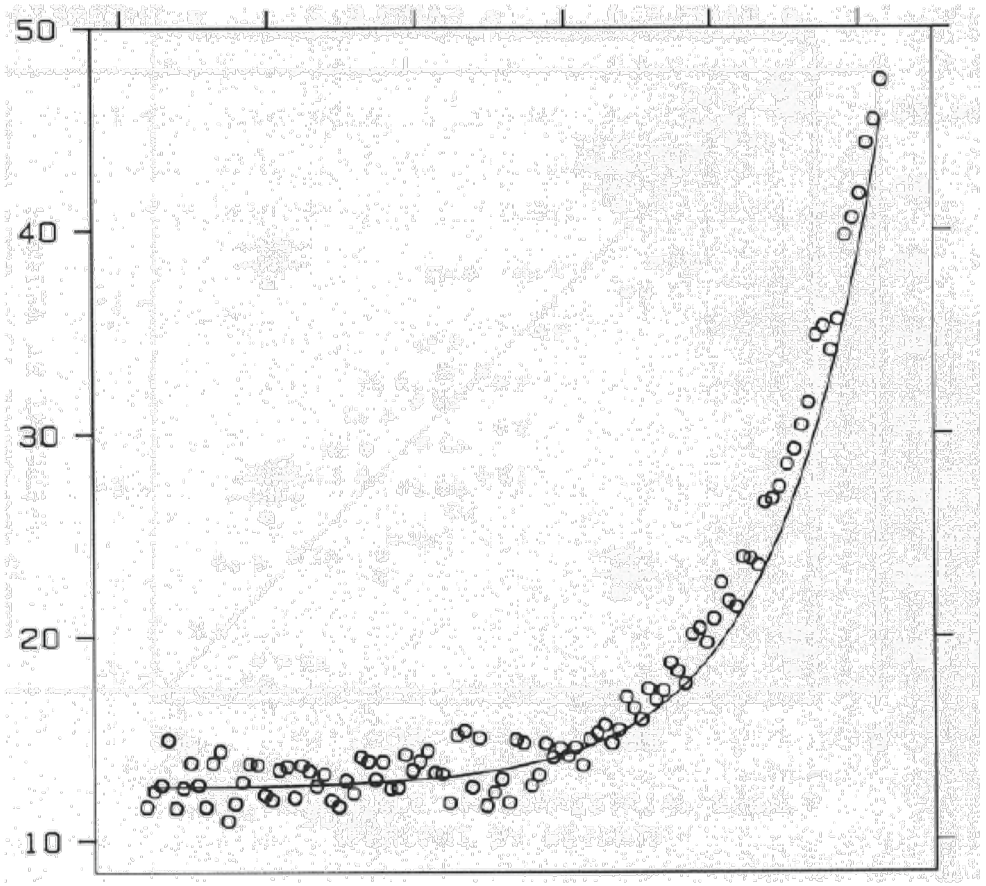


[The Elements of Graphing Data. Cleveland 94]

Fit Curve

How much it fits to the curve?

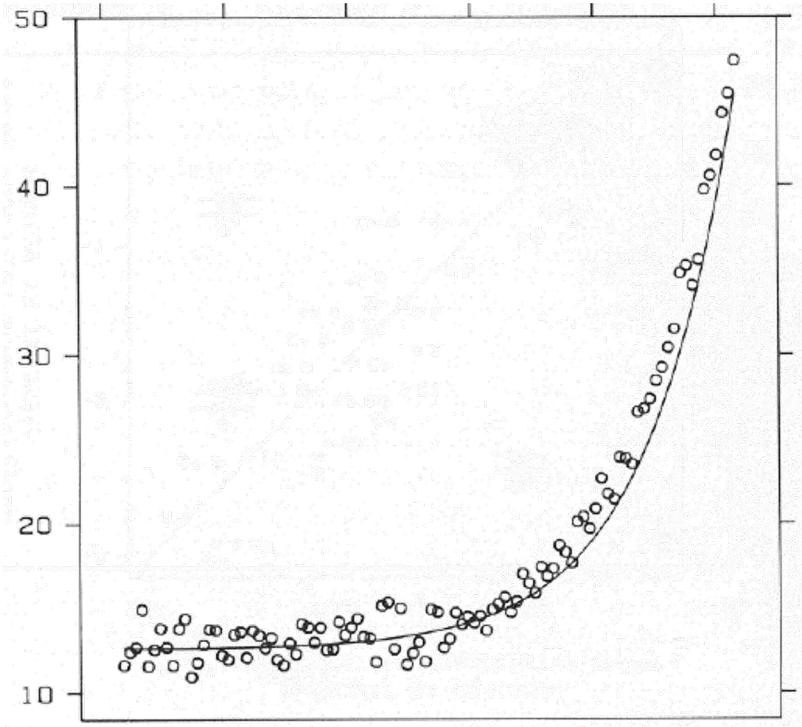
Y Value



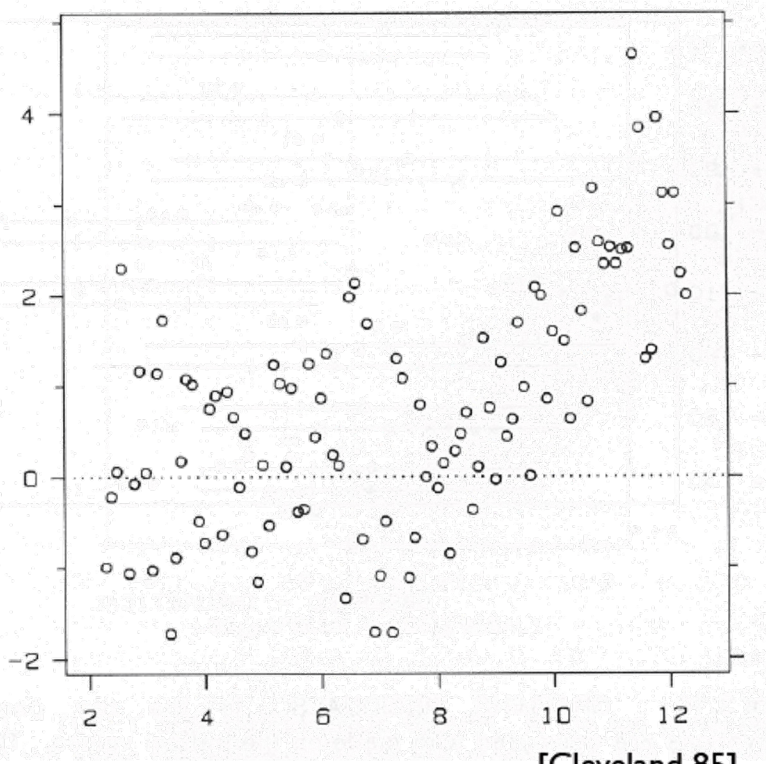
[Cleveland 85]

Remainder

Fit curve



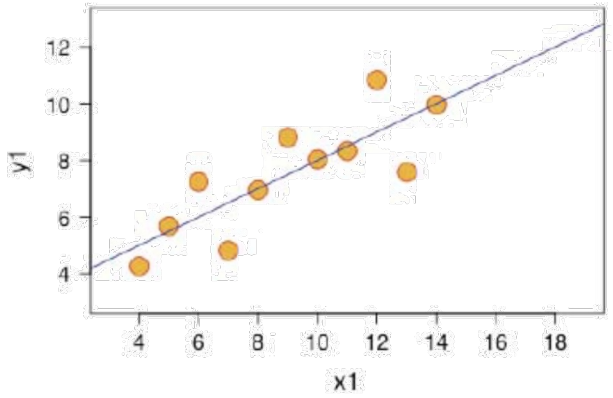
Remainder variation

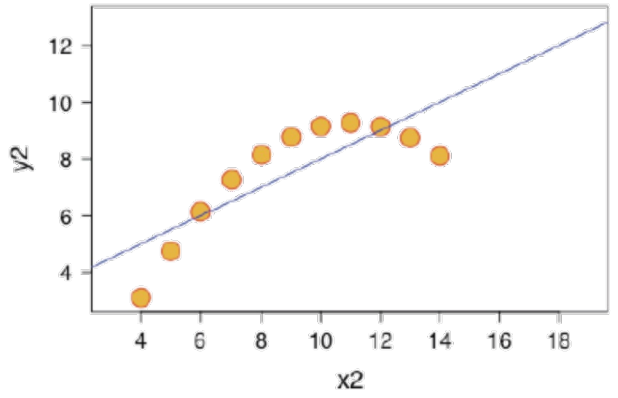


[Cleveland 85]

Show the Data

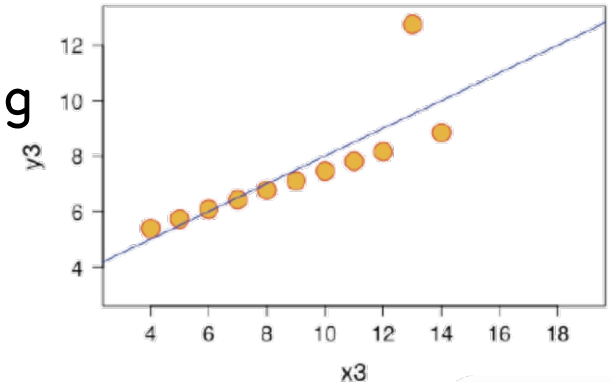
Mean

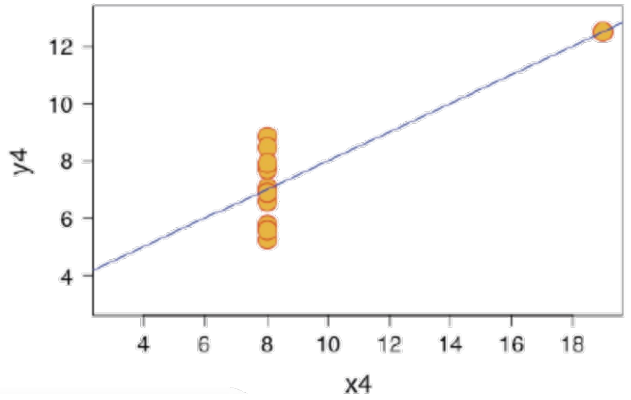




Deviation

Corresponding coefficient



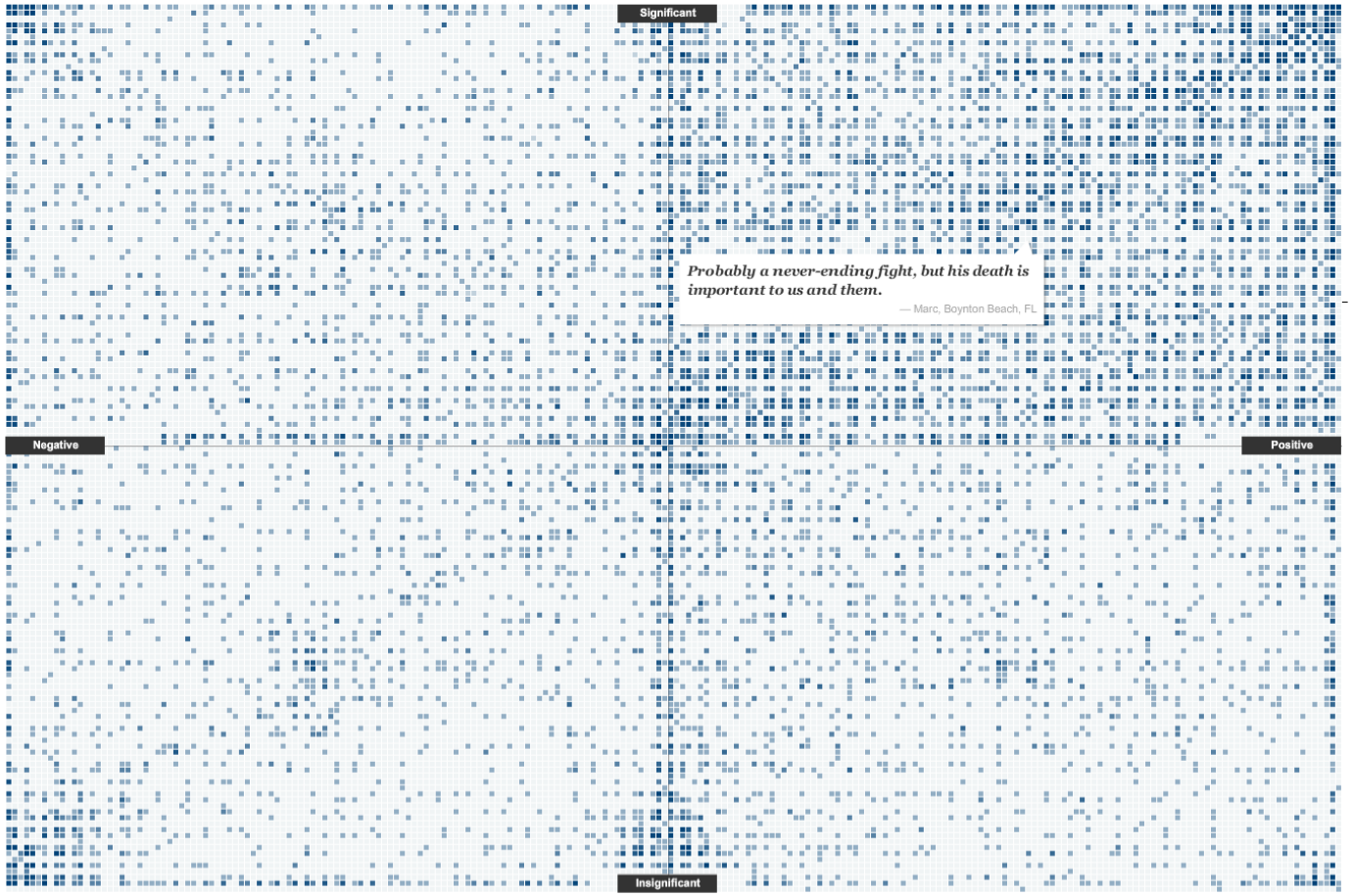


Linear regression

THE DEATH OF OSAMA BIN LADEN

Significant

Negative

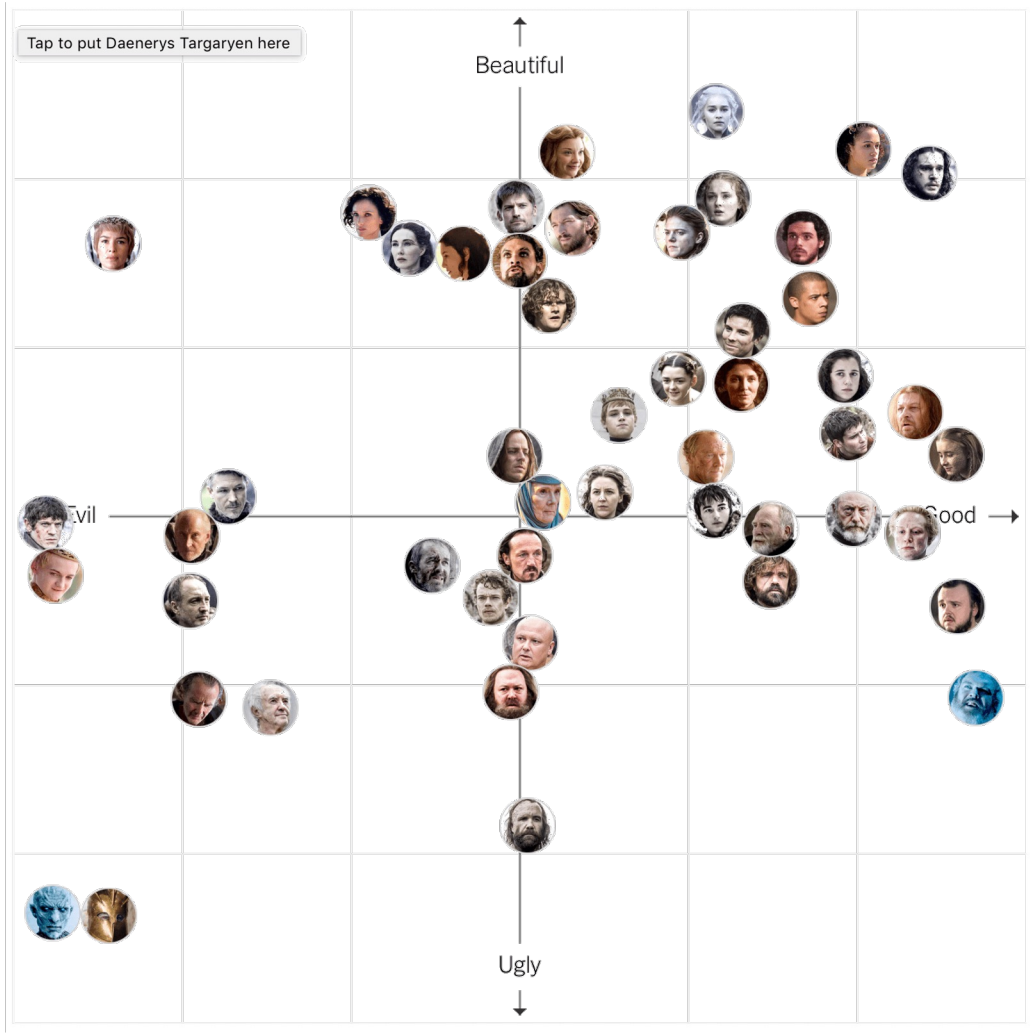


Positive

Insignificant

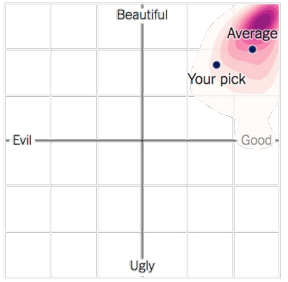
http://www.nytimes.com/interactive/2011/05/03/us/20110503-osama-response.html

Game of Thrones character chart, you decide



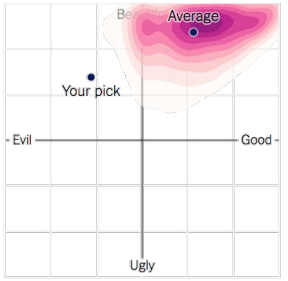
Jon Snow

The hero of the tale — so far. He's been the picture of fair and just and strong, though perhaps a little brooding at times.



Daenerys Targaryen

A stunning beauty who feels for the oppressed. But she has no problem sending people to their deaths as she conquers kingdoms.

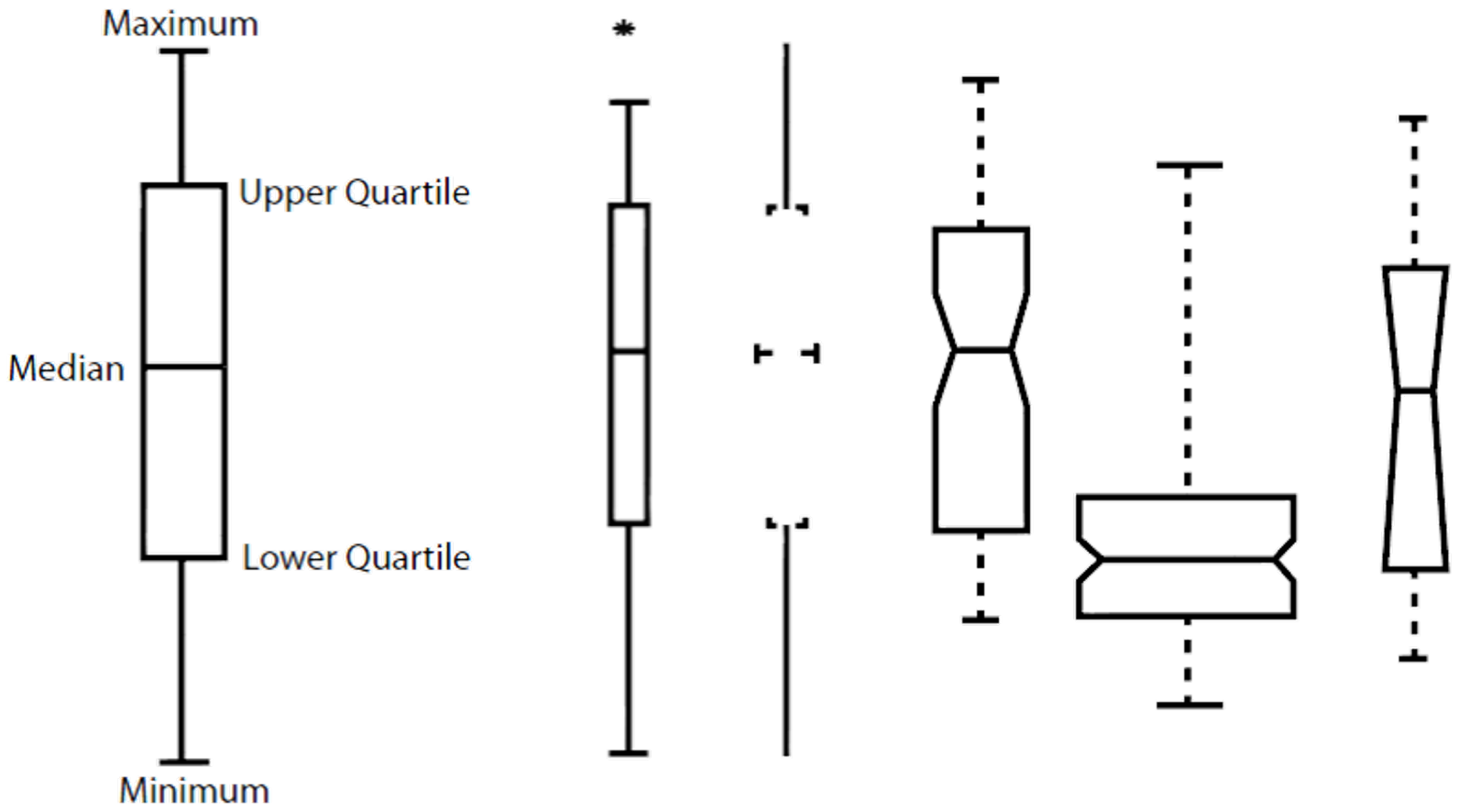


https://flowingdata.com/2017/08/13/game-of-thrones-character-chart-you-decide/

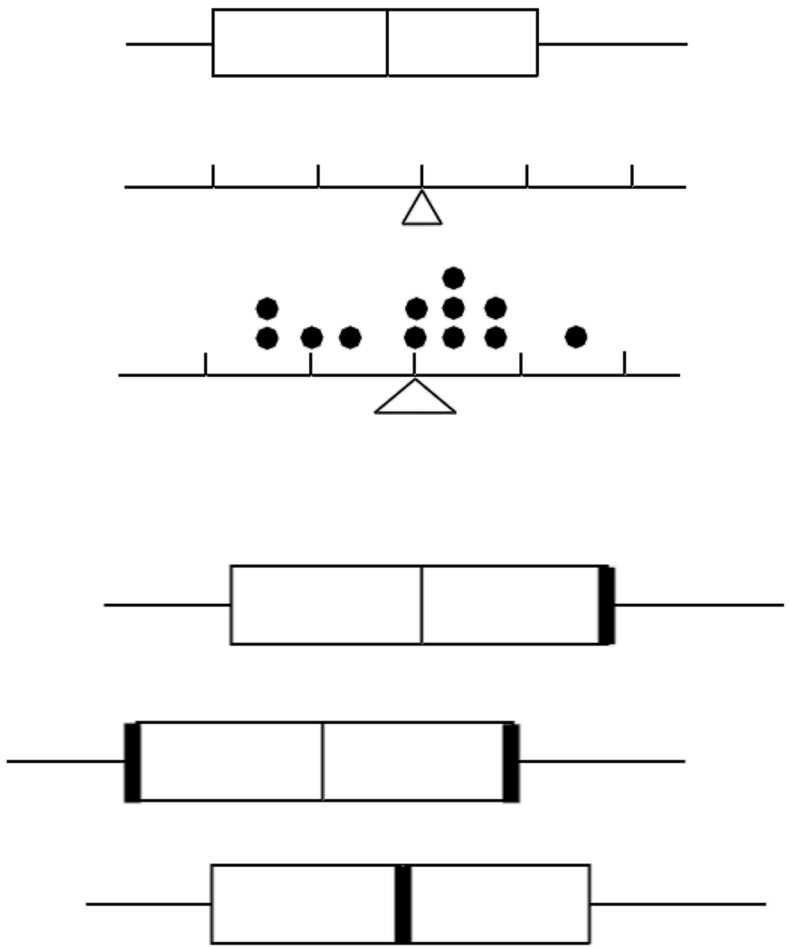
OUTLINE

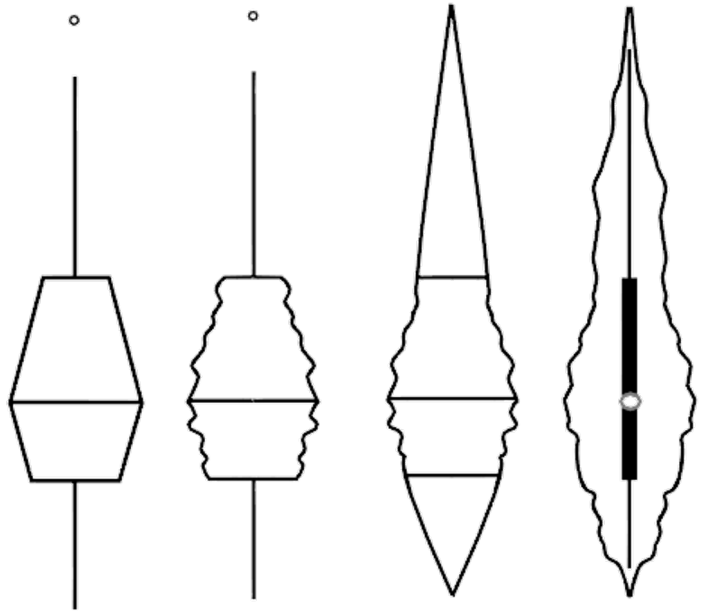
- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Box Plot

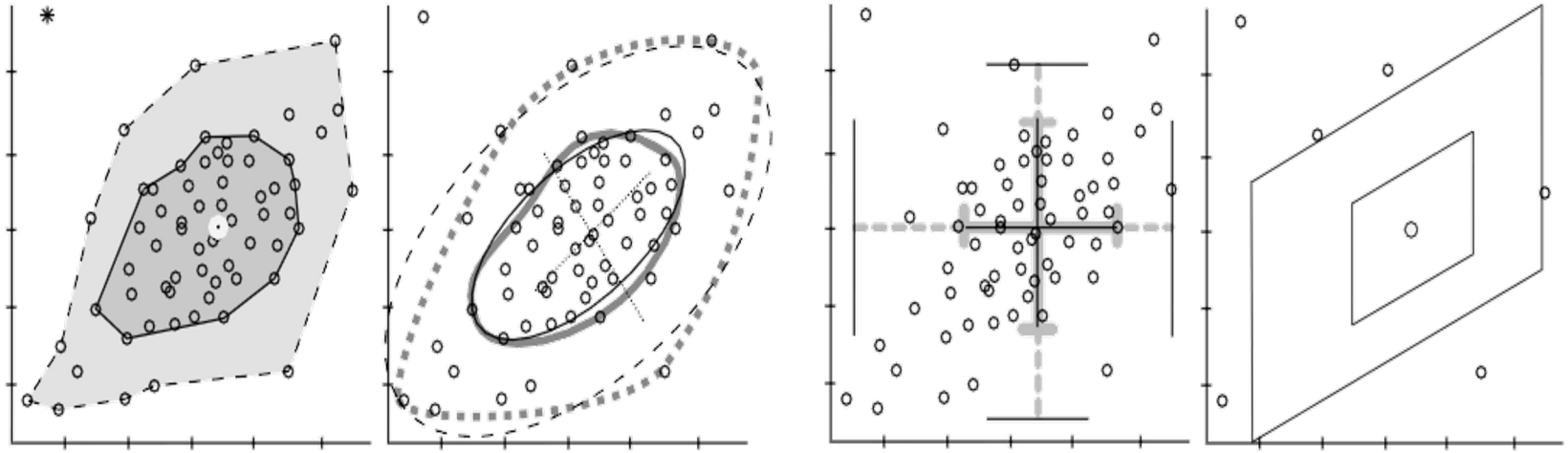


Box Plot Variations





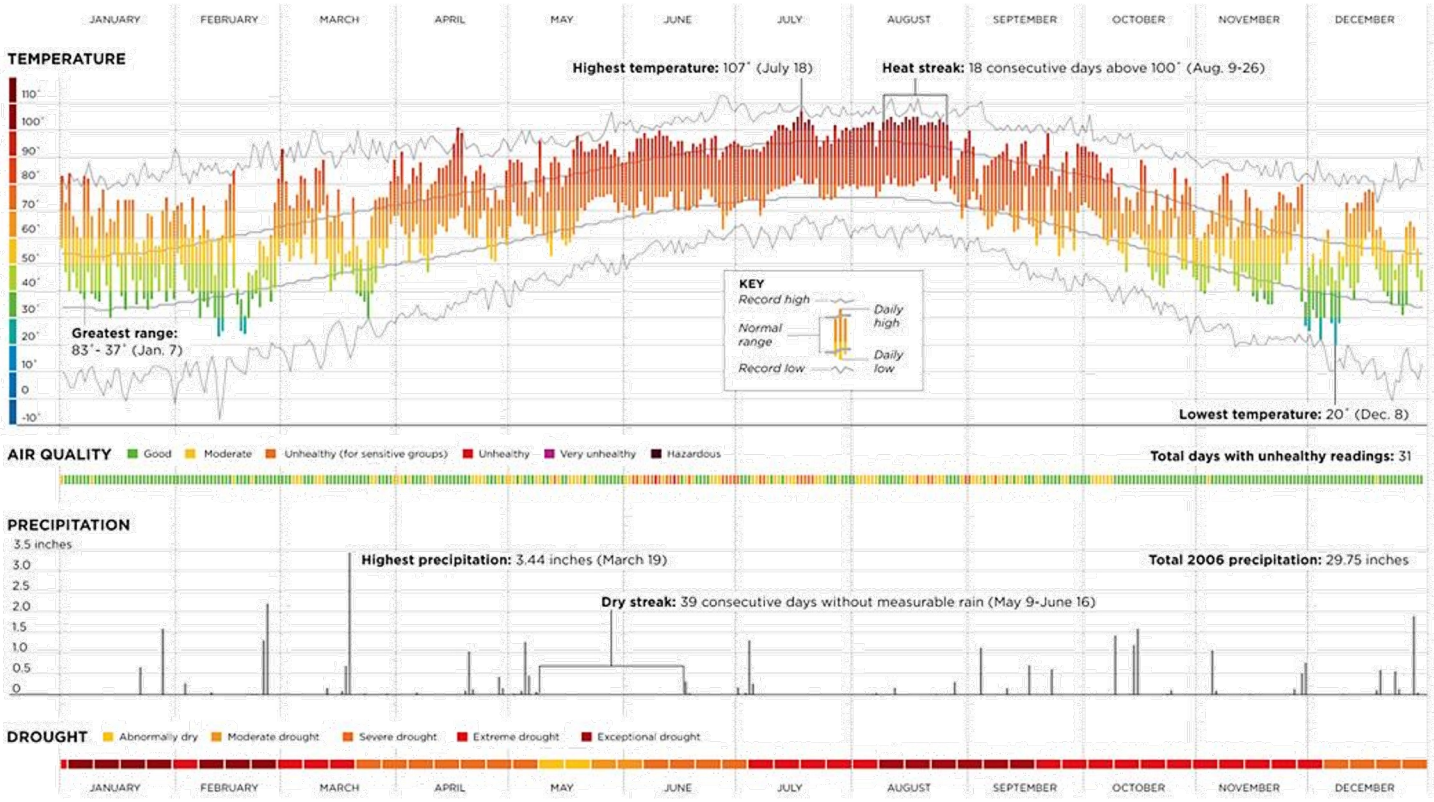
Box Plot Variations



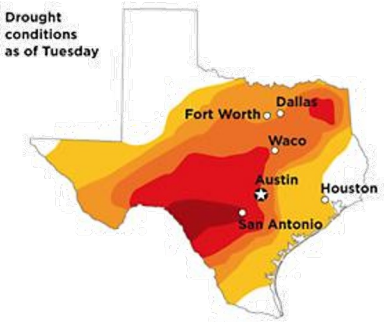
2D Box Plot　　Relplot　　Rangefinder Box Plot　Bag Plot

2006: The warmest year on record in the D-FW area

Drought persists as the annual average temperature increased for the fifth consecutive year with 43 days at 100° or above

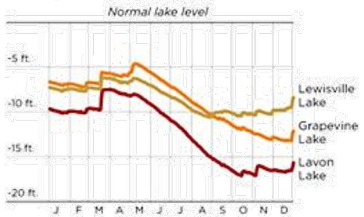


Drought conditions as of Tuesday



BELOW-NORMAL LAKE LEVELS

Area lake elevations have been consistently lower than the normal lake level — an indication of drought conditions.



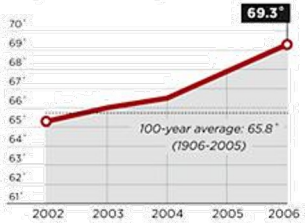
THE HOTTEST YEAR ON RECORD

Annual average temperature for the Dallas-Fort Worth area:

Rank	Temperature	Year	100° days
1	69.3°	2006	43
2	68.6°	1999	33
3	68.3°	1998	56
4	68.2°	1933	11
5	68.1°	1954	52
6	68.1°	1934	34
7	67.9°	2005	18
8	67.9°	1921	11
9	67.8°	1925	33
10	67.7°	1939	21

HOTTER AND HOTTER

Forecasts indicate that the warming trend in D-FW area average temperatures is likely to continue this year.



NOTES: Monthly temperatures for the D-FW area are recorded from different locations but are considered continuous. Air quality readings are daily summations for the D-FW area. Trace amounts of precipitation are marked as zero inches. Drought readings are weekly.

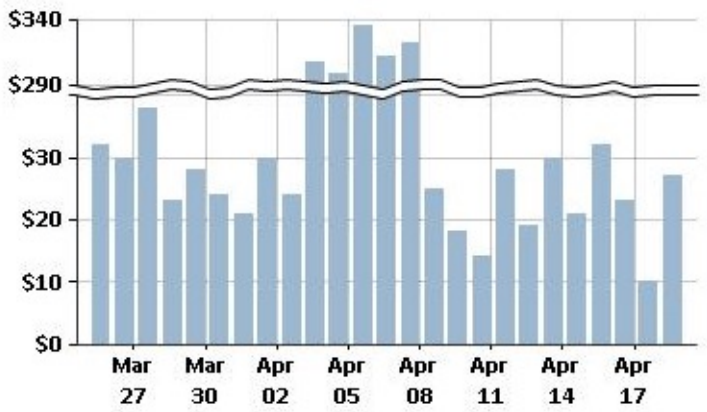
SOURCES: National Weather Service in Fort Worth, Texas Commission on Environmental Quality; U.S. Drought Monitor; U.S. Army Corps of Engineers-Fort Worth District; *Dallas Morning News* research

TROY OXFORD, TOM SETZER/Staff Artists. E-mail: Graphics@dallasnews.com

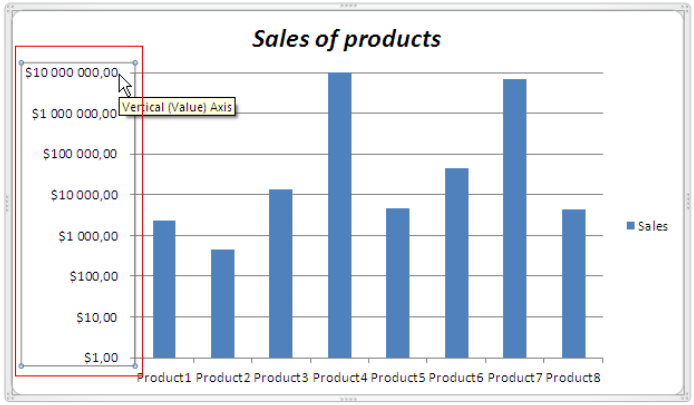
OUTLINE

- Transformation
  - Normalization
  - Smoothness
  - Sampling
  - Binning/Discretization
  - Dimensionality Reduction*
  - Clustering
- Statistical Charts
  - Line chart
  - Sparkline
  - Bar chart
  - Stacked bar chart
  - Pie chart
  - Scatter plot
  - Box plot
  - Scale

Scale Break and Log Scale



Scale Break

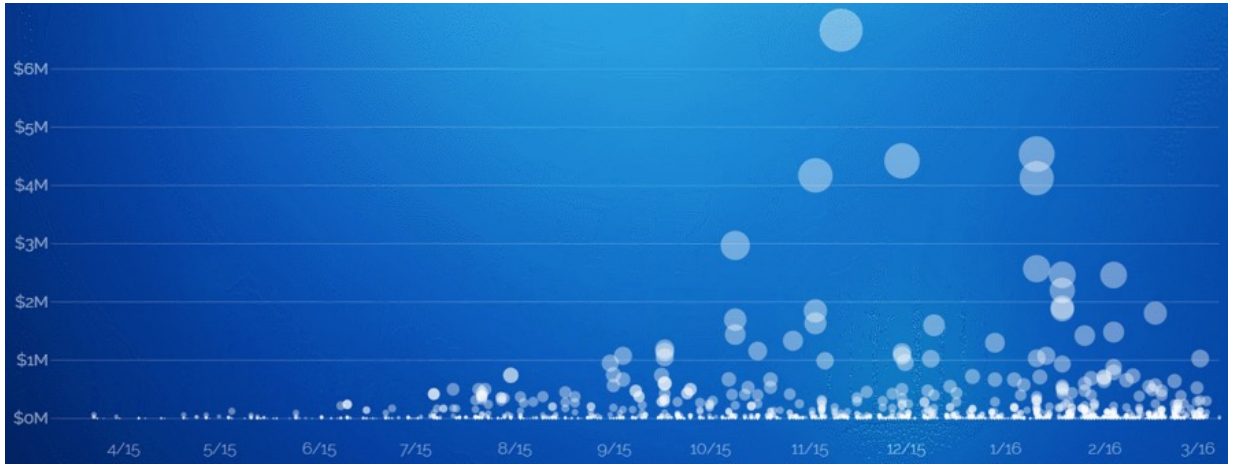


Logarithmic transformation
y=log10(x)

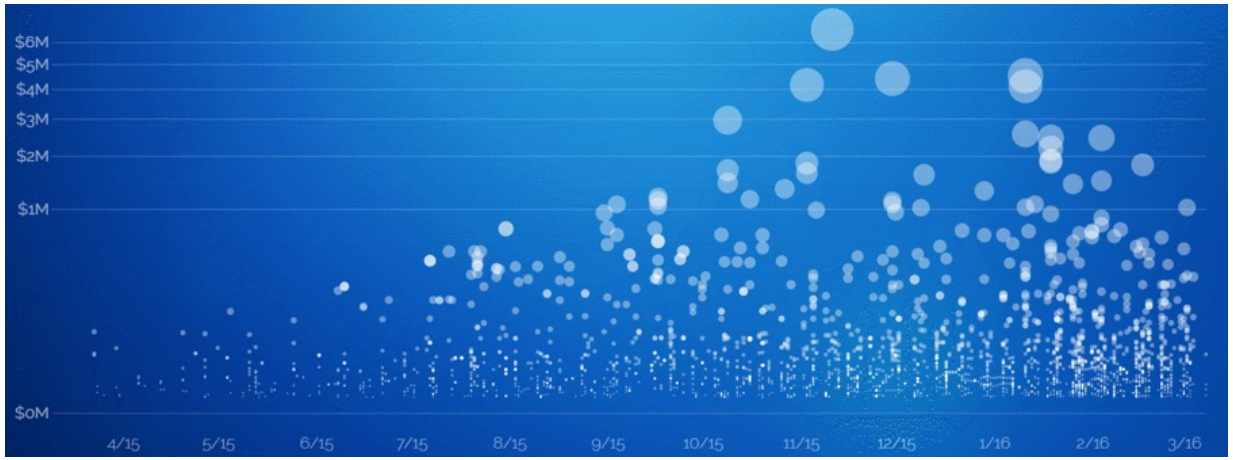
Both improve visual resolution
But it is hard to compare all the data in scale break

Linear Scale and Log Scale

Linear Scale

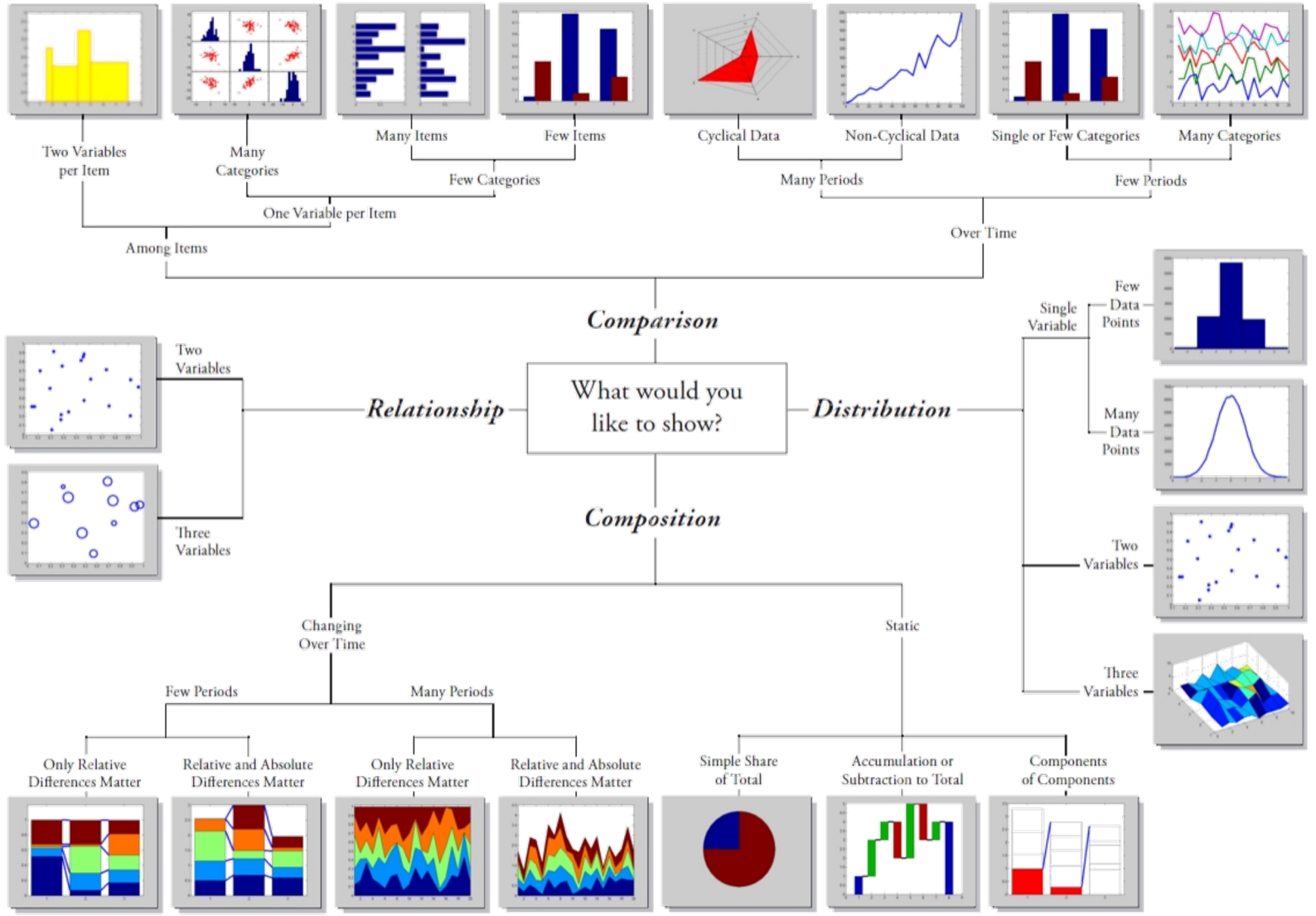


Log Scale



Variable Scale

Chart Suggestions—A Thought-Starter



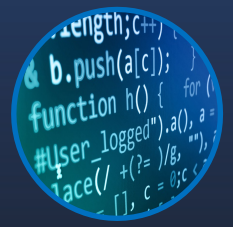
Modified with permision -Doug Hull
blogs.mathworks.com/videos
hull@mathworks.com 2009

www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com

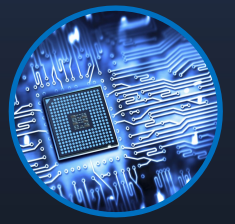
Big Data Era

The big problem: Scalability

Visualization



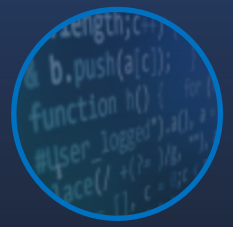
Algorithm



Hardware

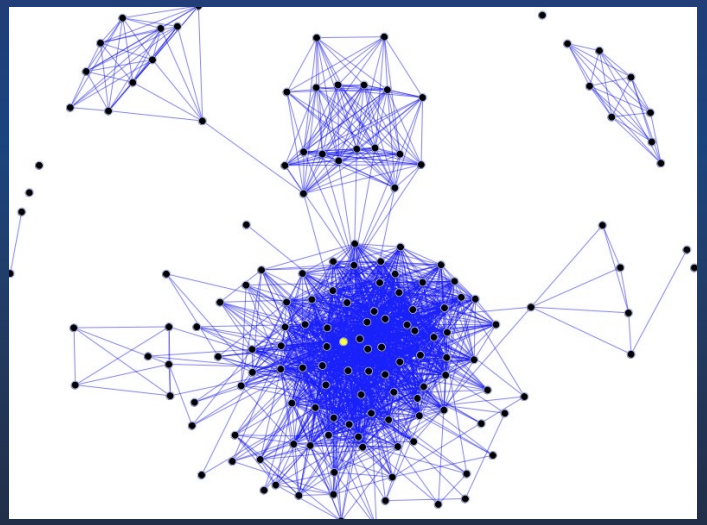
The big problem: Scalability

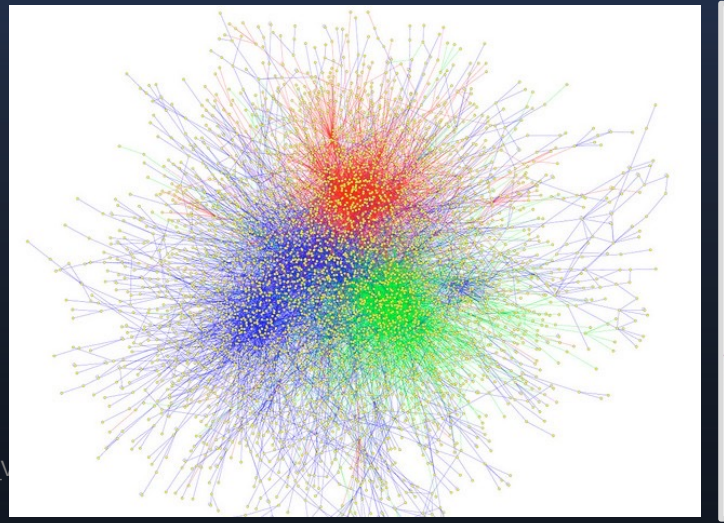
Visualization



Algorithm

Hardware



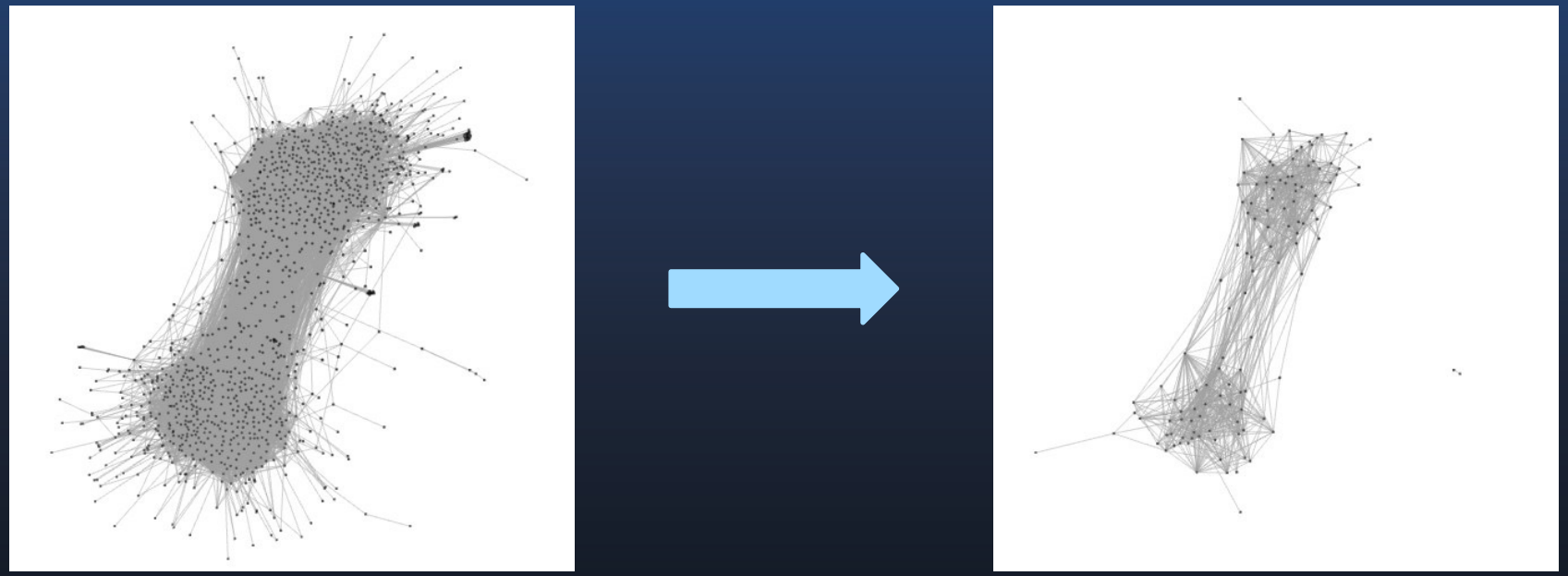


VIS 2016

https://upload.wikimedia.org/wikipedia/commons/0/05/Sna_large.png
https://upload.wikimedia.org/wikipedia/commons/9/9b/Social_Network_Analysis_V
https://c1.staticflickr.com/5/4033/4520018121_6dd39e8d7e_z.jpg
https://c1.staticflickr.com/1/1/916142_ddc2fd0140.jpg

Graph Sampling

- Randomly pick nodes /edges to construct a subgraph that represents the original unfiltered graph:



Which sampling strategy to use?

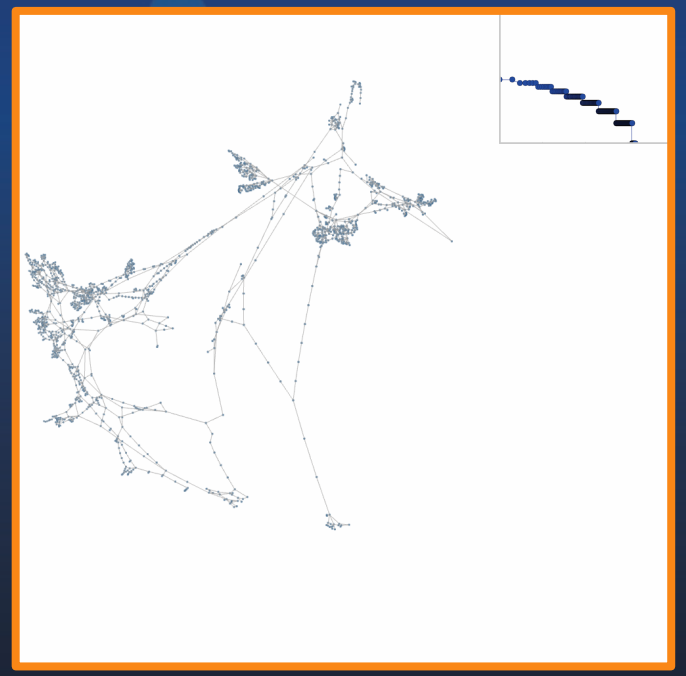
Graph Sampling Evaluation

	Static graph patterns								Temporal graph patterns				
	in-deg	out-deg	wcc	scc	hops	sng-val	sng-vec	clust	diam	cc-sz	sng-val	clust	**AVG**
RN	0.084	0.145	0.814	0.193	0.231	0.079	0.112	0.327	0.074	0.570	0.263	0.371	0.272
RPN	**0.062**	**0.097**	0.792	0.194	**0.200**	0.048	0.081	0.243	0.051	0.475	0.162	0.249	0.221
RDN	0.110	0.128	0.818	0.193	0.238	0.041	0.048	0.256	0.052	0.440	**0.097**	0.242	0.222
RE	0.216	0.305	**0.367**	0.206	0.509	0.169	0.192	0.525	0.164	0.659	0.355	0.729	0.366
RNE	0.277	0.404	0.390	0.224	0.702	0.255	0.273	0.709	0.370	0.771	0.215	0.733	0.444
HYB	0.273	0.394	0.386	0.224	0.683	0.240	0.251	0.670	0.331	0.748	0.256	0.765	0.435
RNN	0.179	0.014	0.581	0.206	0.252	0.060	0.255	0.398	0.058	0.463	0.200	0.433	0.258
RJ	0.132	0.151	0.771	0.215	0.264	0.076	0.143	**0.235**	0.122	0.492	0.161	**0.214**	0.248
RW	0.082	0.131	0.685	0.194	0.243	0.049	**0.033**	**0.243**	**0.036**	**0.423**	**0.086**	0.224	**0.202**
FF	0.082	0.105	0.664	0.194	**0.203**	**0.038**	0.092	**0.244**	0.053	0.434	0.140	**0.211**	**0.205**

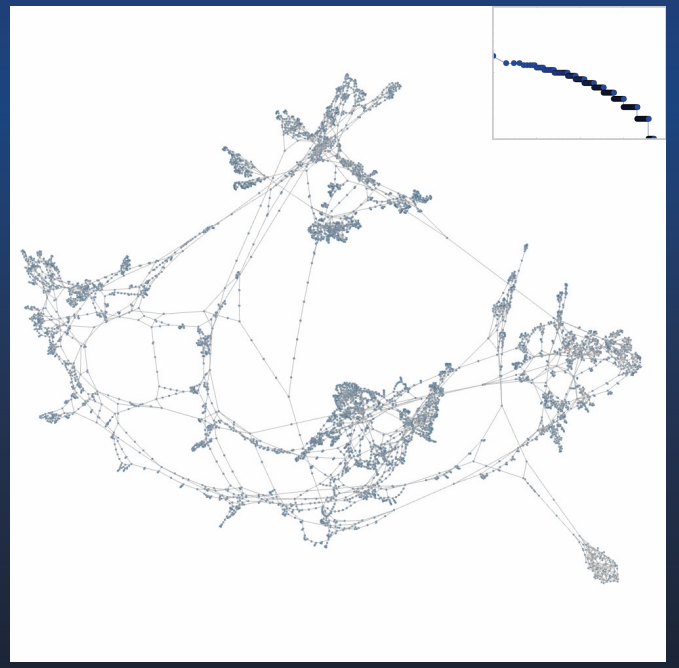
Random Walk (RW) v.s. Forest Fire (FF)

[Leskovec and Faloutsos, KDD 2006]

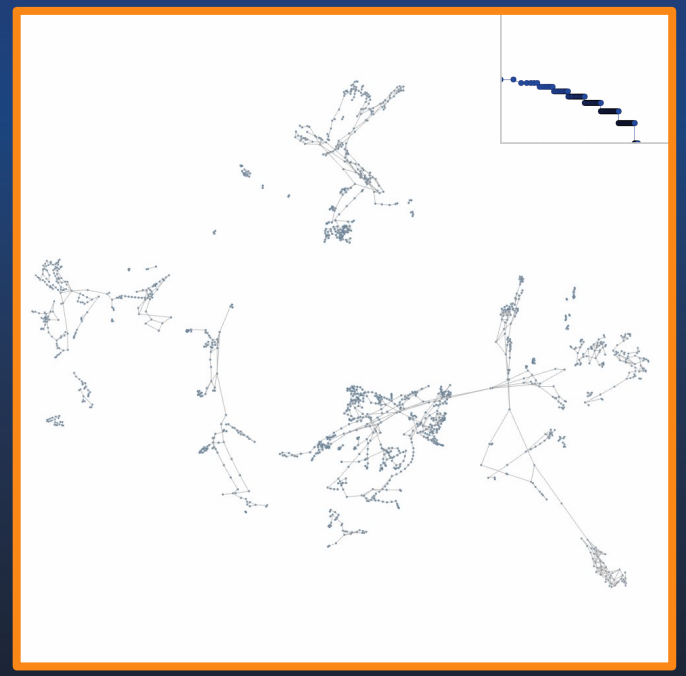
Graph Sampling Evaluation in Visualization



Random Walk (RW)
Avg. node degree: 2.4
Power-law degree distribution



Original Graph



Forest Fire (FF)
Avg. node degree: 2.4
Power-law degree distribution

Distinct Visual Result!

VIS 2016

Graph Sampling Evaluation in Visualization

Similarity Measurements

Statistical Features: Hub Inclusion Clustering Coeff. Discovery Quotient …	?
Data Mining	Visualization

Graph Sampling Evaluation in Visualization

Similarity Measurements

Statistical Features:

Hub Inclusion
Clustering Coeff.
Discovery Quotient
…

Data Mining

Visual Factors:

?

Visualization

Goals

G1: Identify the key visual factors that makes the sampled graphs representative

G2: Evaluate the performance of different sampling algorithms on these visual factors

Procedure

Pilot Study

Formal Studies

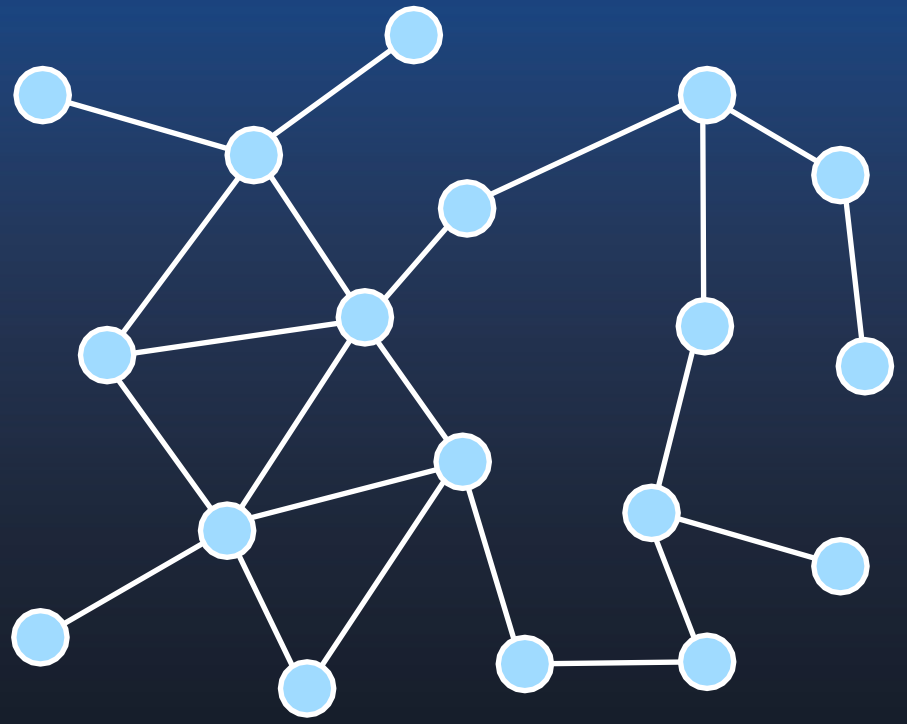
VIS 2016

Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
  - Perception of High Degree Nodes
  - Perception of Cluster Quality
  - Perception of Coverage Area

Node-Based Sampling

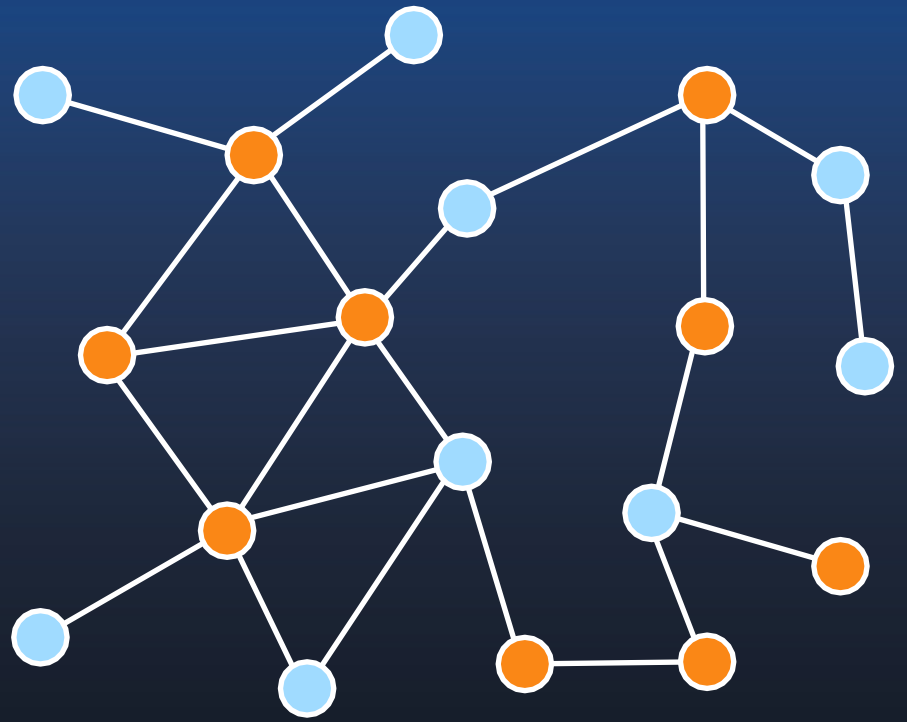
Original Graph



Random Node Sampling

Node-Based Sampling

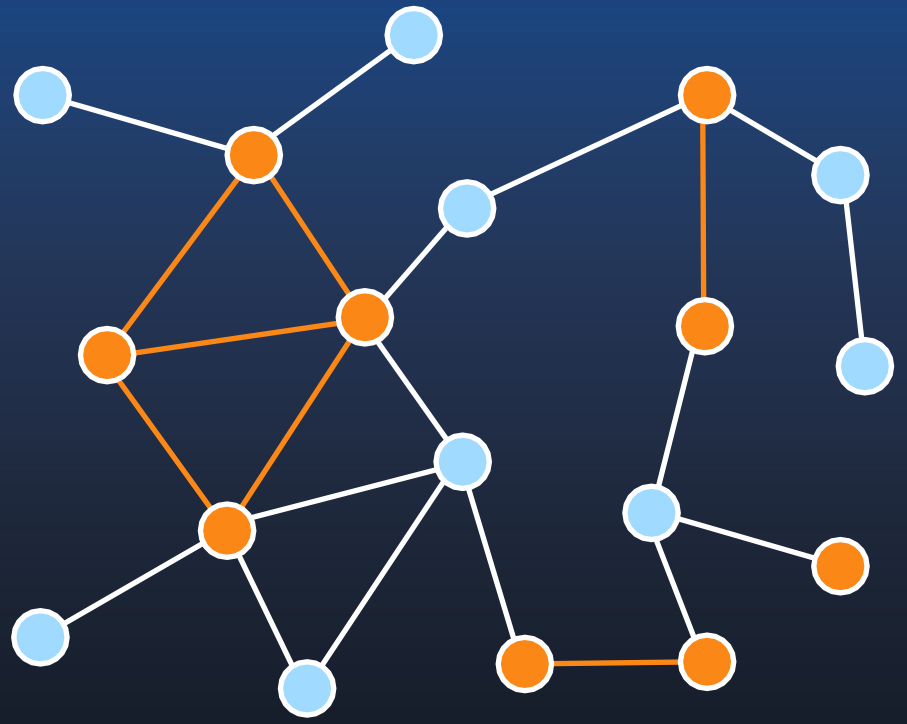
Original Graph



Random Node Sampling

Node-Based Sampling

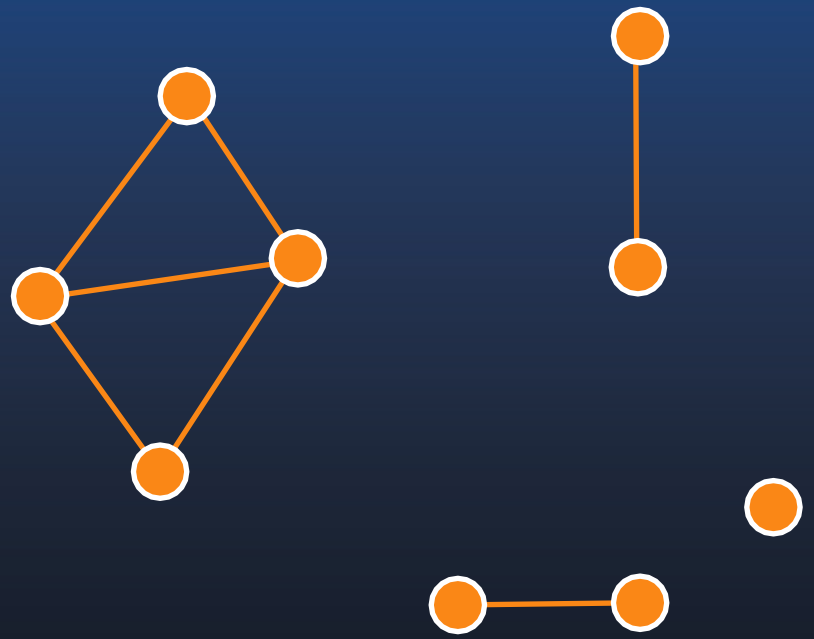
Original Graph



Random Node Sampling

Node-Based Sampling

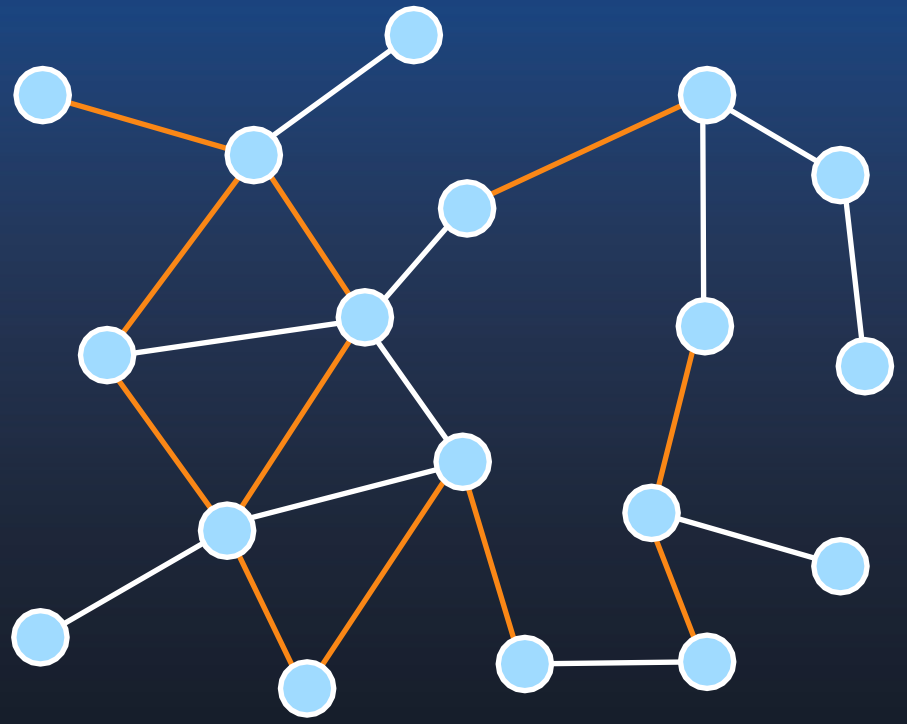
Original Graph



Random Node Sampling

Edge-Based Sampling

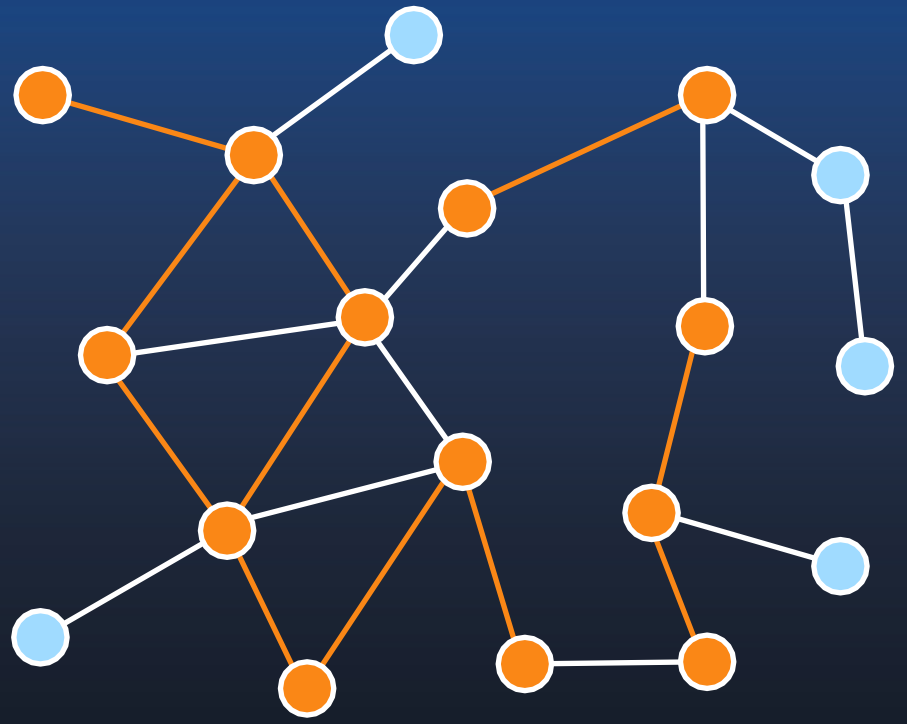
Original Graph



Random Edge Sampling

Edge-Based Sampling

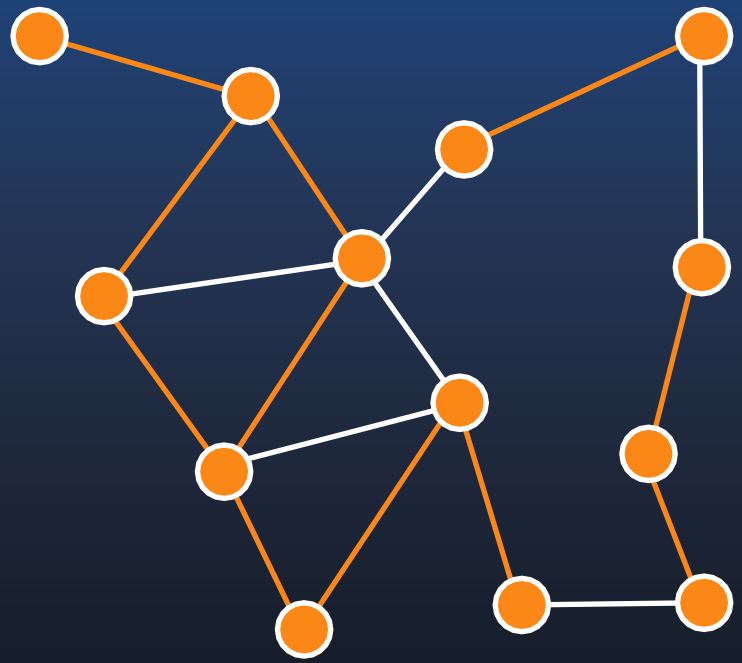
Original Graph



Random Edge Sampling

Edge-Based Sampling

Original Graph



Random Edge Sampling

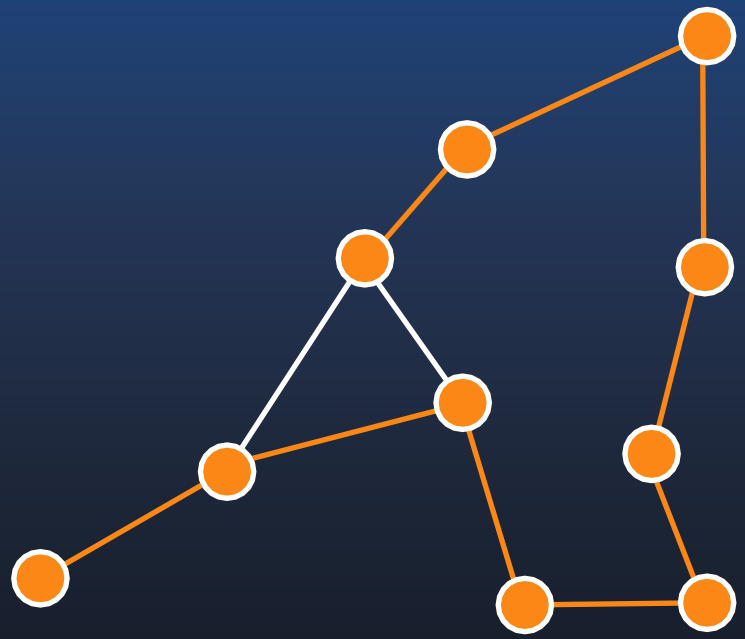
Traversal-Based Sampling: Random Walk

Original Graph

Random Walk

Traversal-Based Sampling: Random Walk

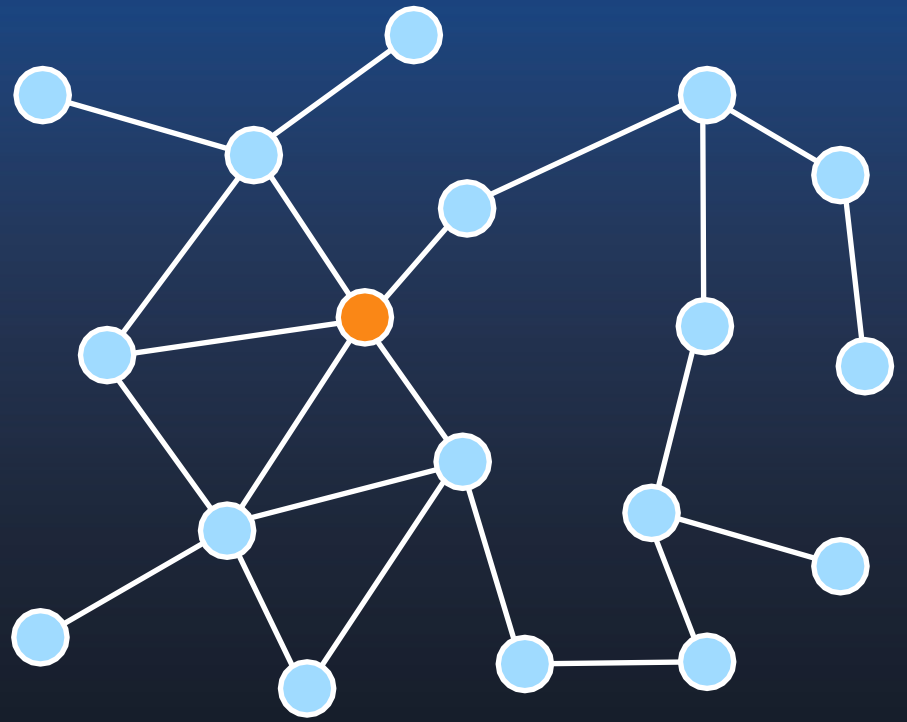
Original Graph



Random Walk

Traversal-Based Sampling: Random Jump

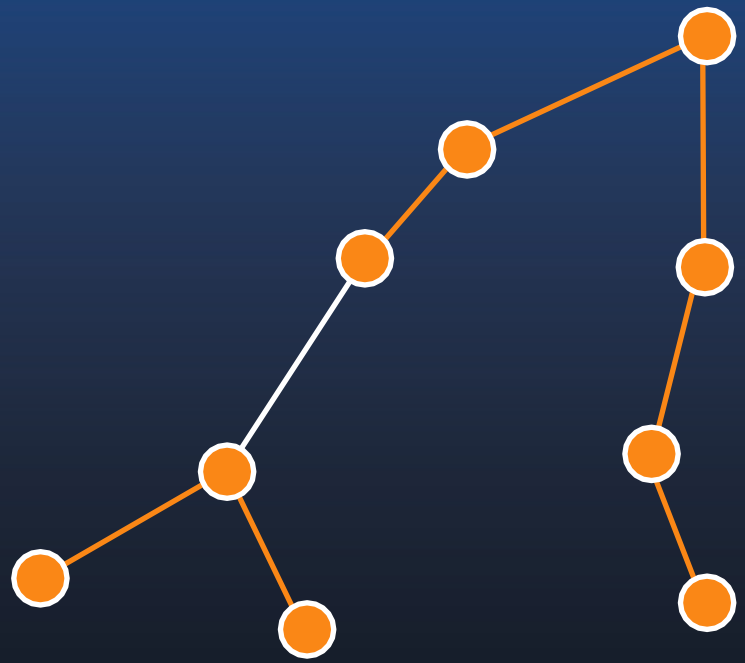
Original Graph



Random Jump

Traversal-Based Sampling: Random Jump

Original Graph



Random Jump

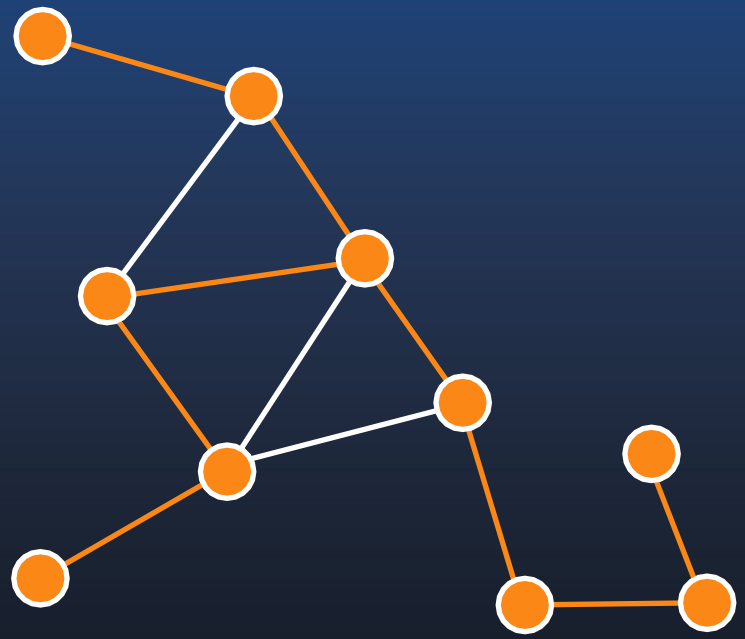
Traversal-Based Sampling: Forest Fire

Original Graph

Forest Fire

Traversal-Based Sampling: Forest Fire

Original Graph



Forest Fire

Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
  - Perception of High Degree Nodes
  - Perception of Cluster Quality
  - Perception of Coverage Area

Pilot Study

- Task:
  - Identify the visual factors that strongly influence the representativeness of sampled graphs
  - We also determine the sampling rate used in the formal studies.

Network	N	D	AD	CC	PL
ResidentRating (*RR*)	217	0.1002	21.6	0.50	1.9
PoliticalBlogs (*PB*)	1,222	0.0220	27.4	0.32	2.7
AdolescentHealth (*AH*)	2,539	0.0054	13.7	0.33	2.3
PowerGrid (*PG*)	4,941	0.0005	1.3	0.08	19.0
Google+ (*G+*)	23,613	0.0001	3.3	0.17	4.0

Dataset: 5 Real-World Graphs

Network Level	Node Level	Edge Level
Coverage Area (*CA*)	High Degree Nodes (*HN*)	Edges Linking *HN*
Cluster Quality (*CQ*)	Margin Nodes (*MN*)	Edges Linking *MN*
	Boundary Nodes (*BN*)	Edges Linking *BN*

Visual Factor Candidates

Pilot Study

- Task:
  - Identify the visual factors that strongly influence the representativeness of sampled graphs
  - We also determine the sampling rate used in the formal studies.

High Degree Nodes
Cluster Quality
Coverage Area

Results (key visual factors)

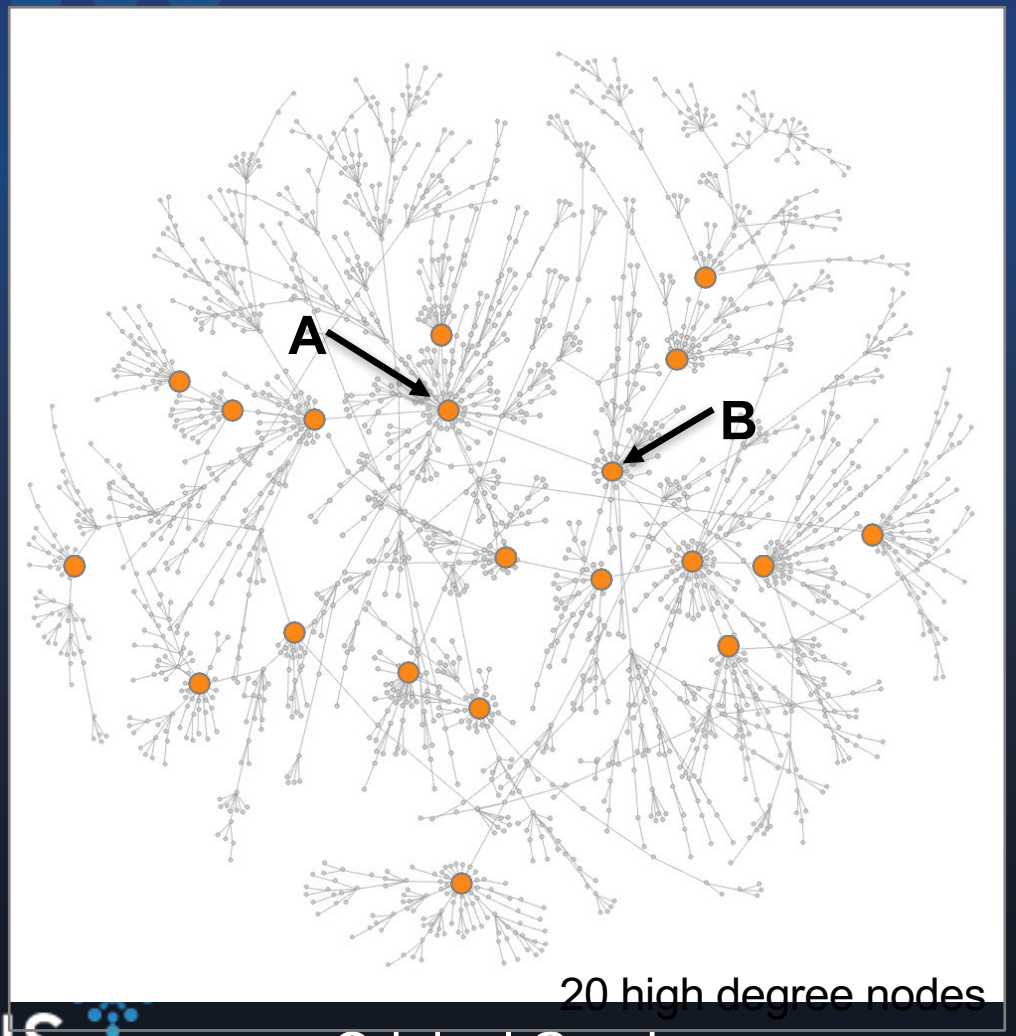
Network Level	Node Level	Edge Level
Coverage Area (*CA*)	High Degree Nodes (*HN*)	Edges Linking *HN*
Cluster Quality (*CQ*)	Margin Nodes (*MN*)	Edges Linking *MN*
	Boundary Nodes (*BN*)	Edges Linking *BN*

Visual Factor Candidates

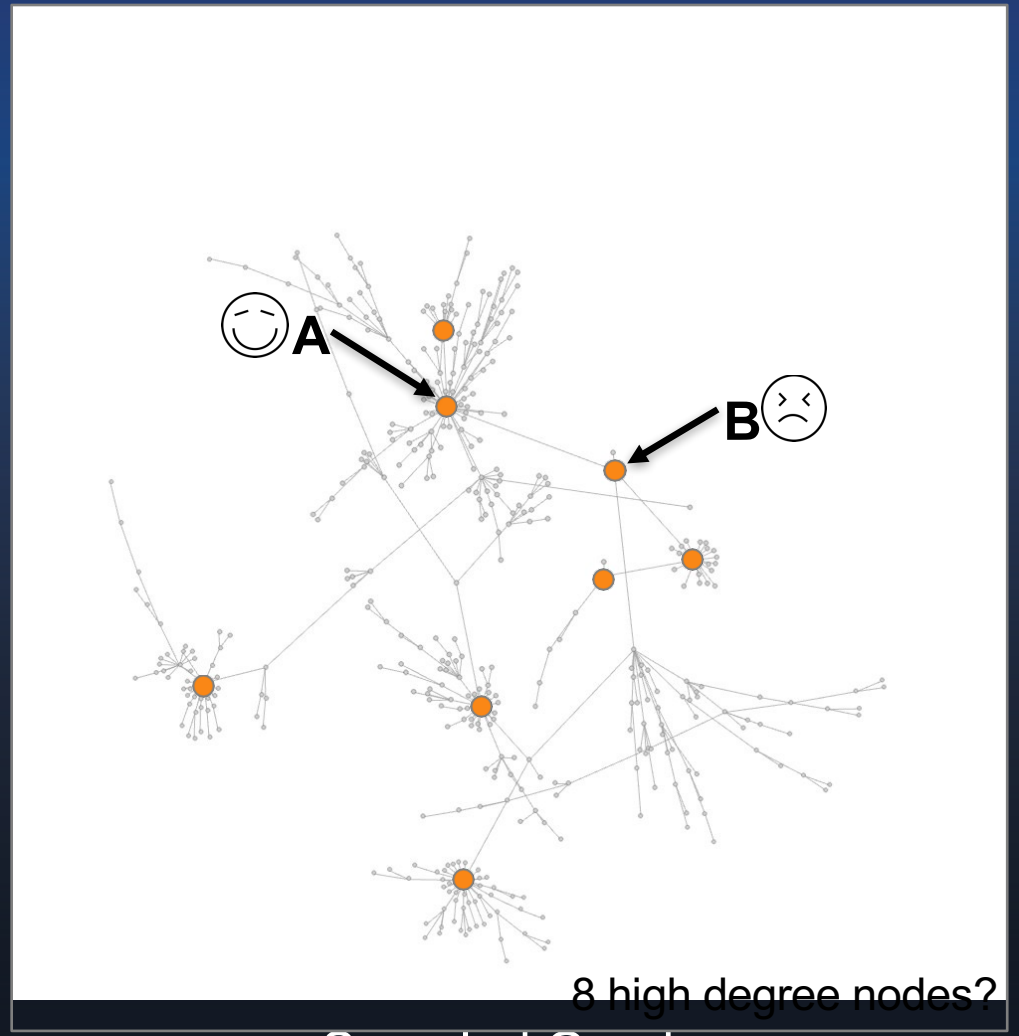
Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
  - Perception of High Degree Nodes
  - Perception of Cluster Quality
  - Perception of Coverage Area

Formal Study I: High Degree Nodes



Original Graph



Sampled Graph

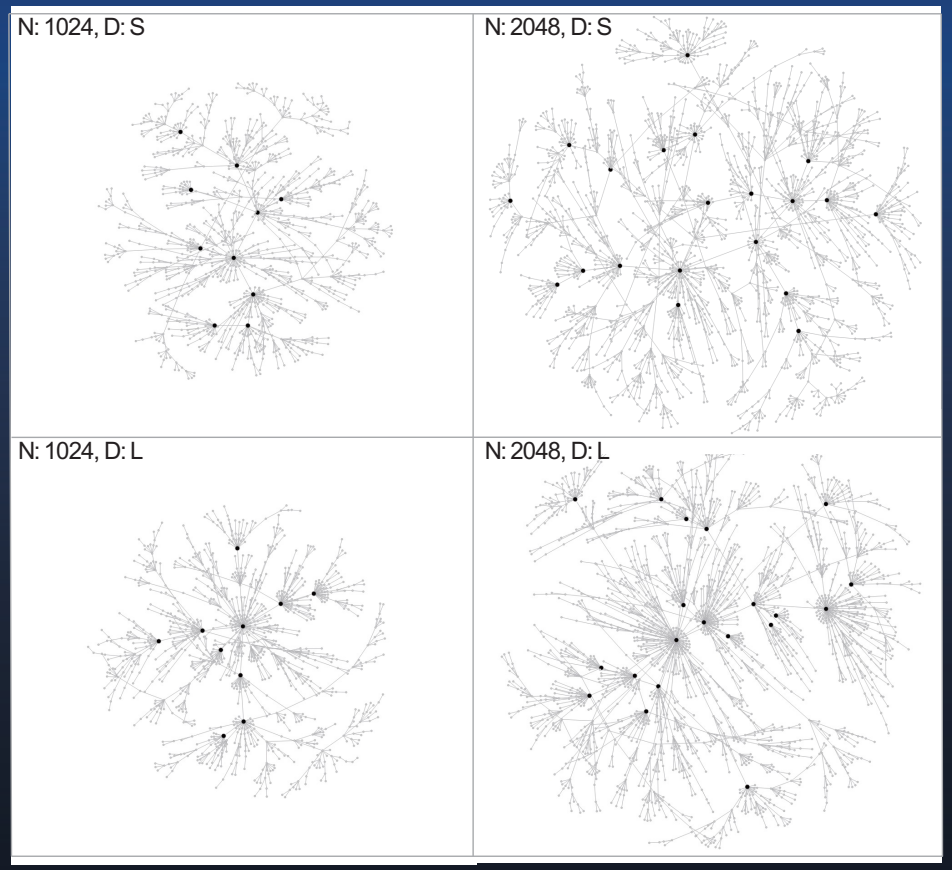
Formal Study I: High Degree Nodes

Formal Study I: High Degree Nodes

	2	graph sizes (small, large)
	2	average degrees of hub nodes (small, large)
	5	sampling strategies (*RN*, *REN*, *RW*, *RJ*, *FF*)
	3	random seeds (3 different seeds)
×	3	repetitions
	180	trials per participant
×	20	participants
	3,600	**trials in total**

Experiment Setting

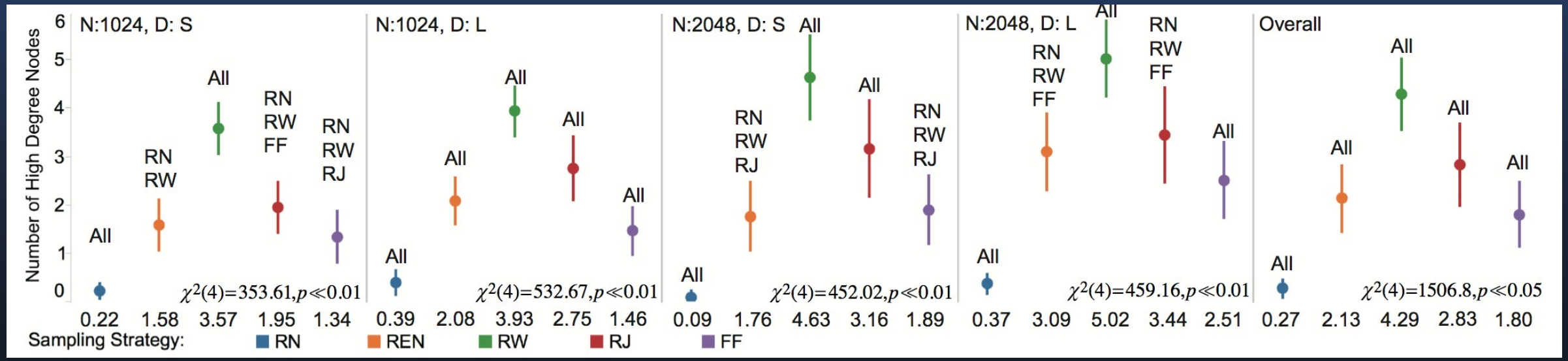
20 high degree nodes



Data Generation

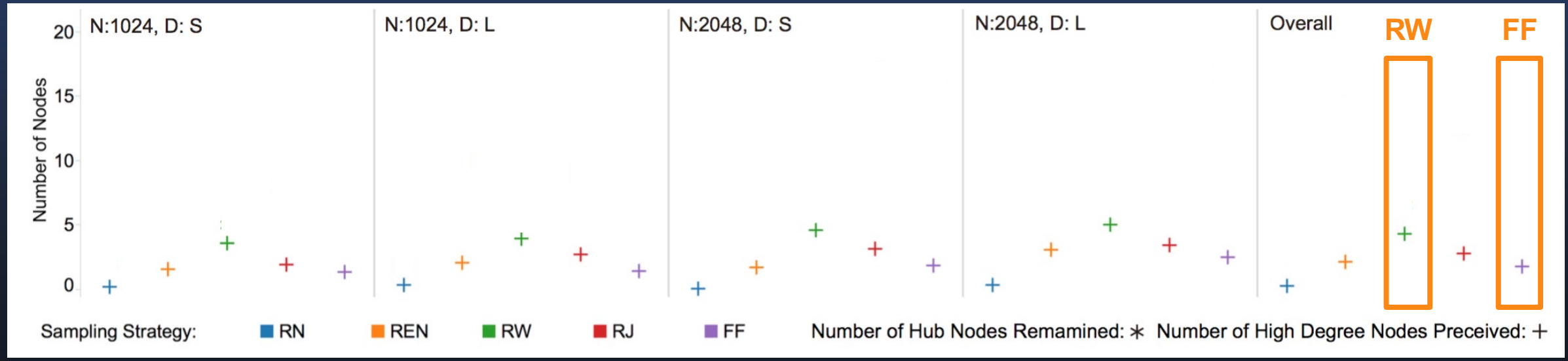
Formal Study I: High Degree Nodes Results

- Discussions:
  - It is easier to perceive high degree nodes in the *RW* Samples
  - It is more difficult to perceive high degree nodes in *RN* Samples
  - Above results hold across datasets



Formal Study I: High Degree Nodes Results

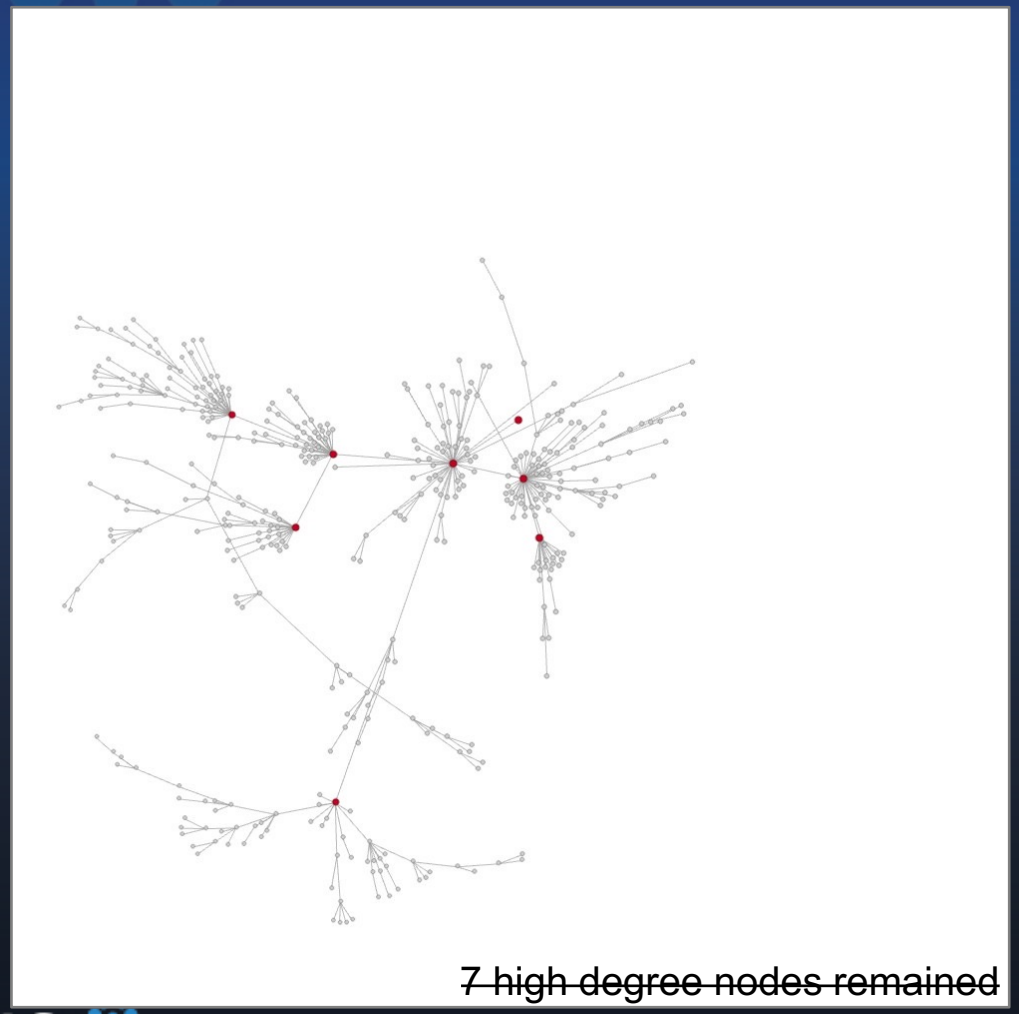
- Discussions:
  - It will be easier to perceive high degree nodes in the *RW* Samples
  - It will be more difficult to perceive high degree nodes in *RN* Samples.
  - Above results hold across datasets



Contradiction with metric-based results!

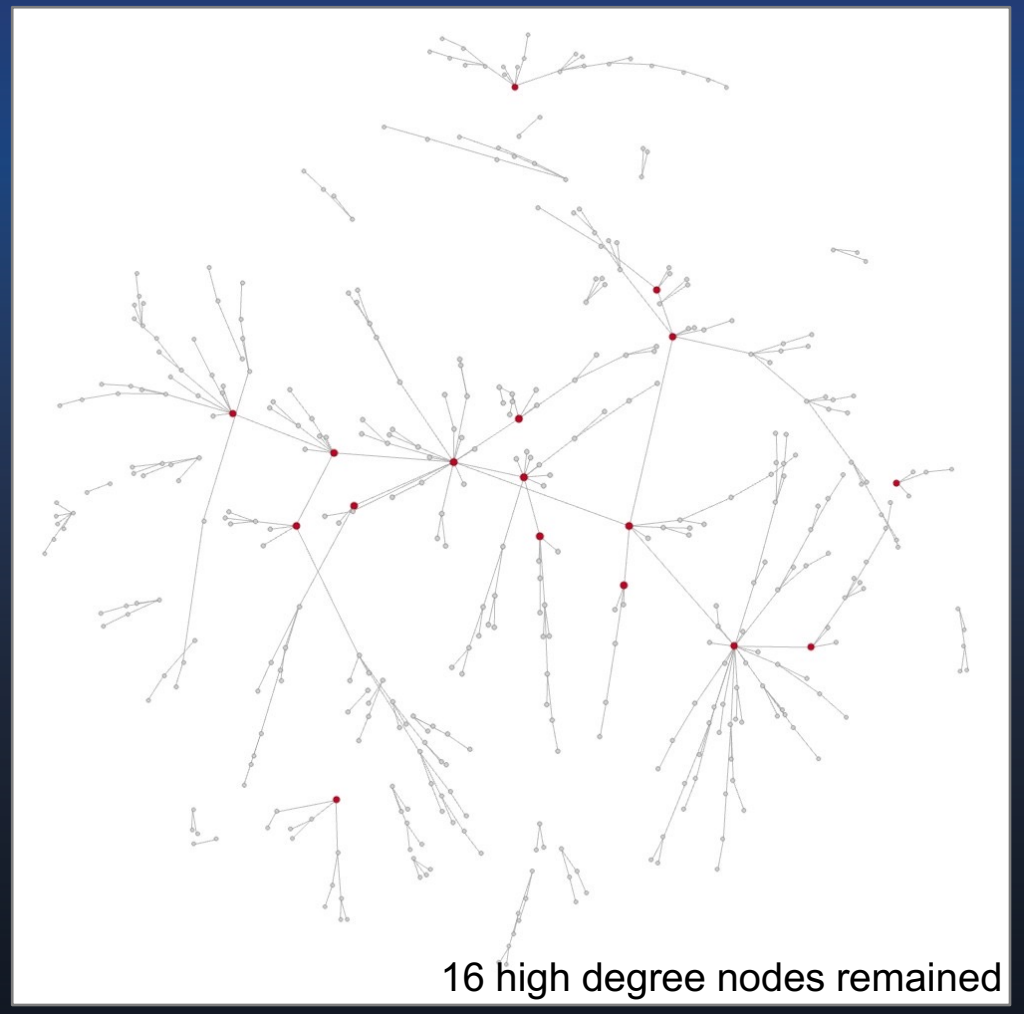
Number of high degree nodes perceived (Visualization): +
Number of high degree nodes remained (Data Mining): *

Formal Study I: High Degree Nodes Results



~~7 high degree nodes remained~~

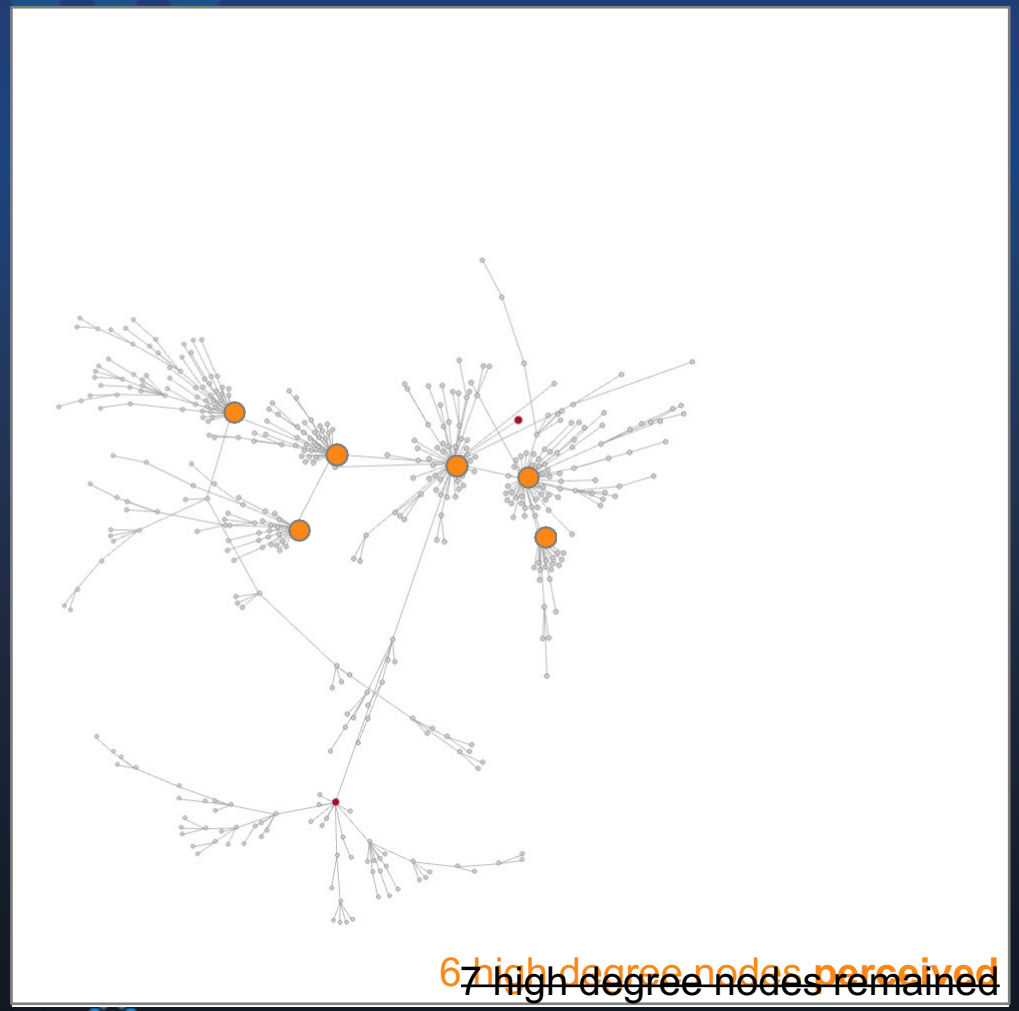
Random Walk (RW)



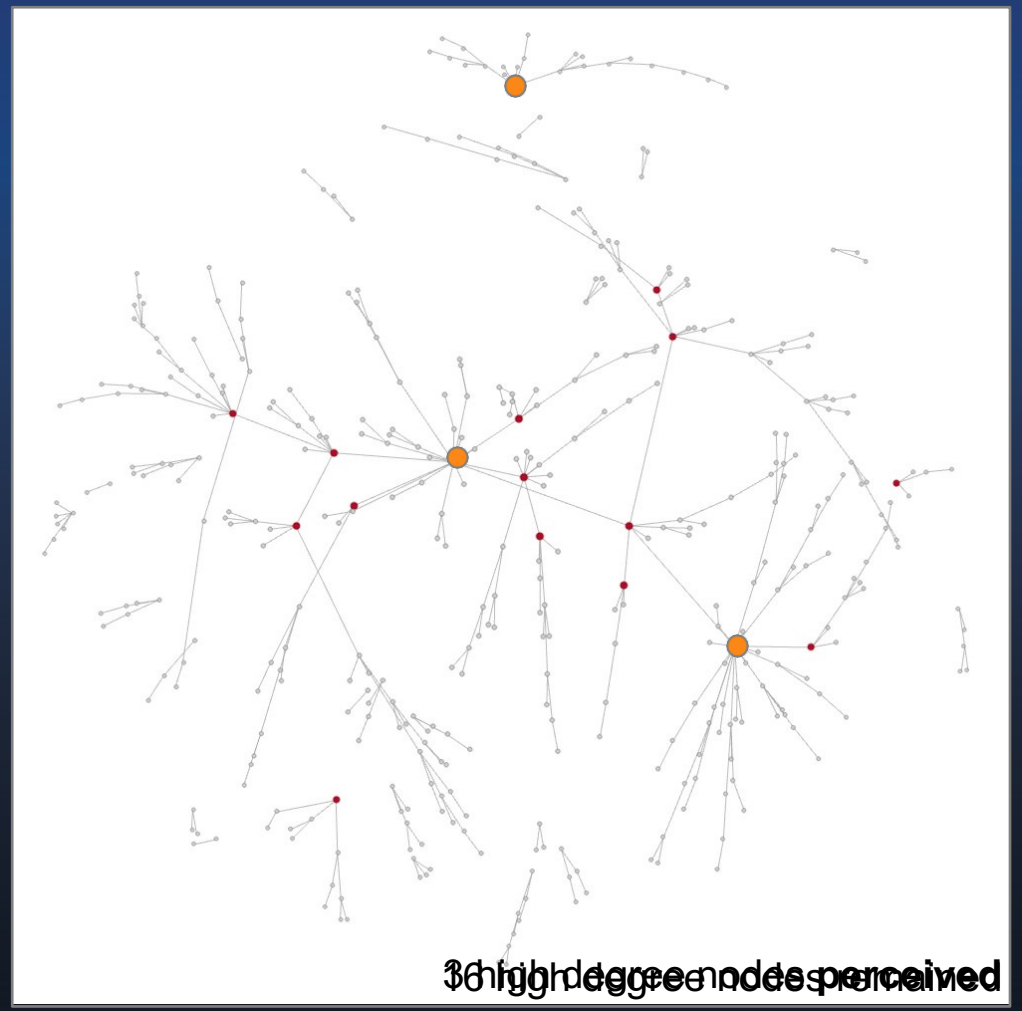
16 high degree nodes remained

Forest Fire (FF)

Formal Study I: High Degree Nodes Results



Random Walk (RW)

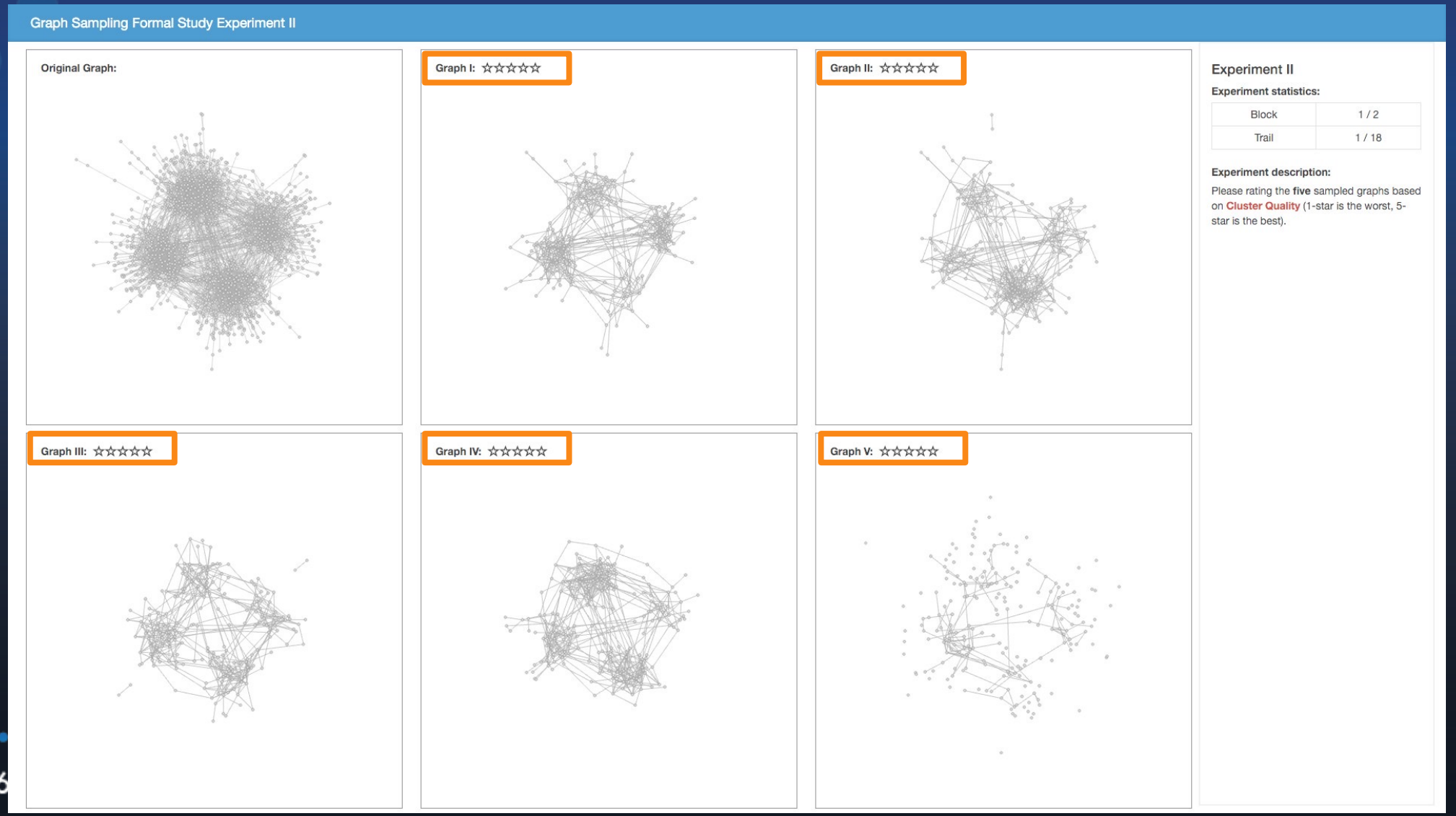


Forest Fire (FF)

Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
  - Perception of High Degree Nodes (more high degree nodes are perceived in *RW*)
  - Perception of Cluster Quality
  - Perception of Coverage Area

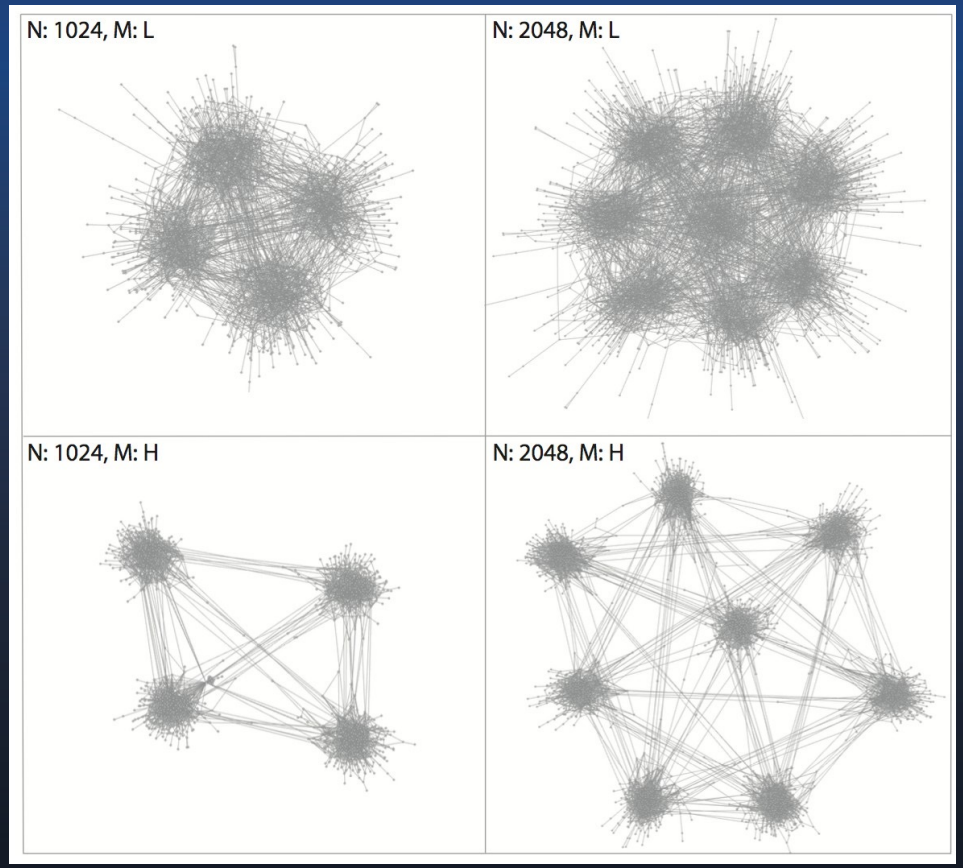
Formal Study II: Cluster Quality



Formal Study II: Cluster Quality

	2	graph sizes (small=1024 nodes, large=2048 nodes)
	2	graph modularities (low, high)
	3	random seeds (3 different seeds)
×	3	repetitions
	36	trials per participant
×	20	participants
	720	**trials in total**

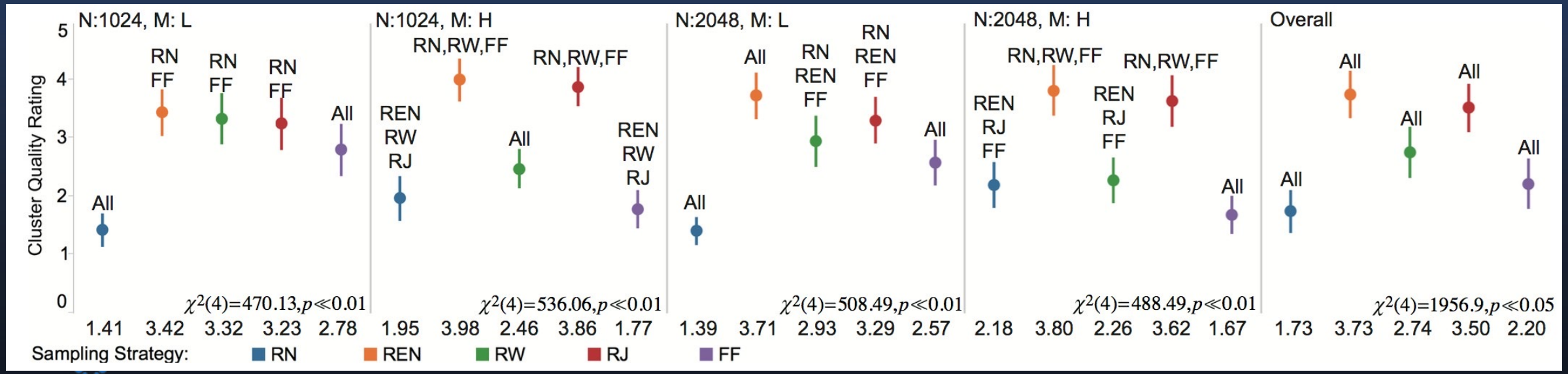
Experiment Setting



Data Generation

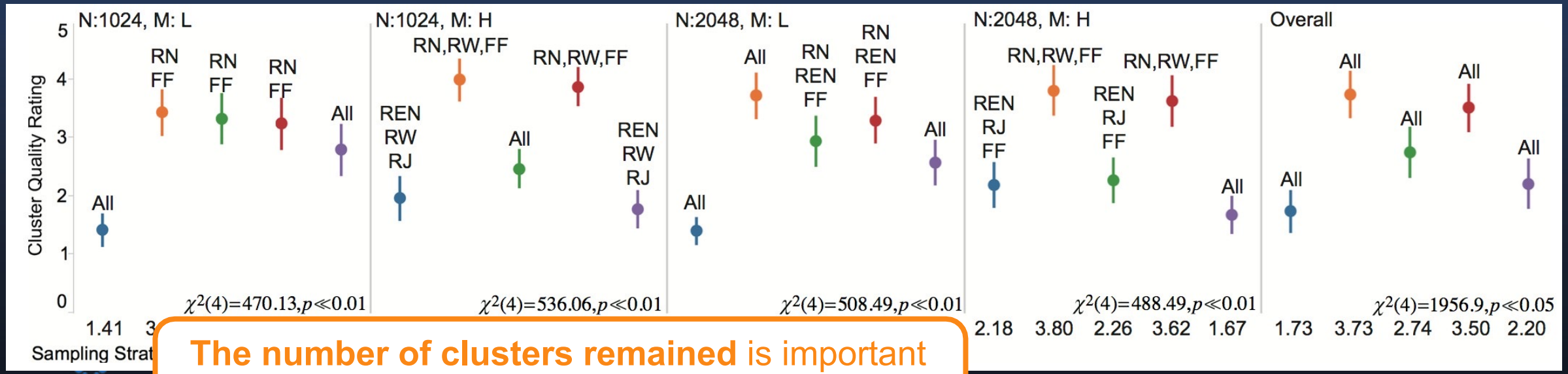
Formal Study II: Cluster Quality Results

- Discussions:
  - *RE* and *RJ* best preserve the perceived cluster quality in samples
  - *RN and FF* struggles in preserving the perceived cluster quality
  - The performance of *RW* and *FF* depends on graph modularity



Formal Study II: Cluster Quality Results

Graph	N: 1024, M: L				N: 1024, M: H				N: 2048, M: L				N: 2048, M: H				Overall			
	M	CN	CS	ER	M	CN	CS	ER	M	CN	CS	ER	M	CN	CS	ER	M	CN	CS	ER
Original	0.55	4	256	0.50	0.68	4	256	0.15	0.67	8	256	0.50	0.80	8	256	0.15	0.68	6	256	0.33
RN	0.77	4.6	14.0	0.15	0.80	4.3	15.9	0.07	0.84	2.4	21.7	0.08	0.88	4.1	26.4	0.02	0.82	3.8	19.5	0.08
REN	0.62	6	14.0	0.15	0.72	**4.0**	50.0	0.03	0.73	**8.0**	50.2	0.17	0.85	**8.0**	50.4	0.02	0.73	6.2	48.4	0.10
RW	0.59	**4.2**	**48.2**	0.20	0.57	4.4	48.0	**0.20**	0.70	**8.0**	51.5	**0.19**	0.74	6.0	68.2	**0.03**	**0.65**	5.6	**54.0**	0.16
RJ	0.60	4.9	41.5	0.22	**0.69**	**4.0**	**50.5**	0.03	0.72	**8.0**	51.0	0.16	**0.83**	**8.0**	51.0	0.02	**0.71**	6.2	48.5	0.11
FF	**0.56**	4.9	41.8	**0.27**	0.45	6.5	33.5	0.62	**0.69**	7.5	**53.9**	0.17	0.66	5.0	**80.8**	**0.03**	0.59	**6.0**	52.5	**0.27**

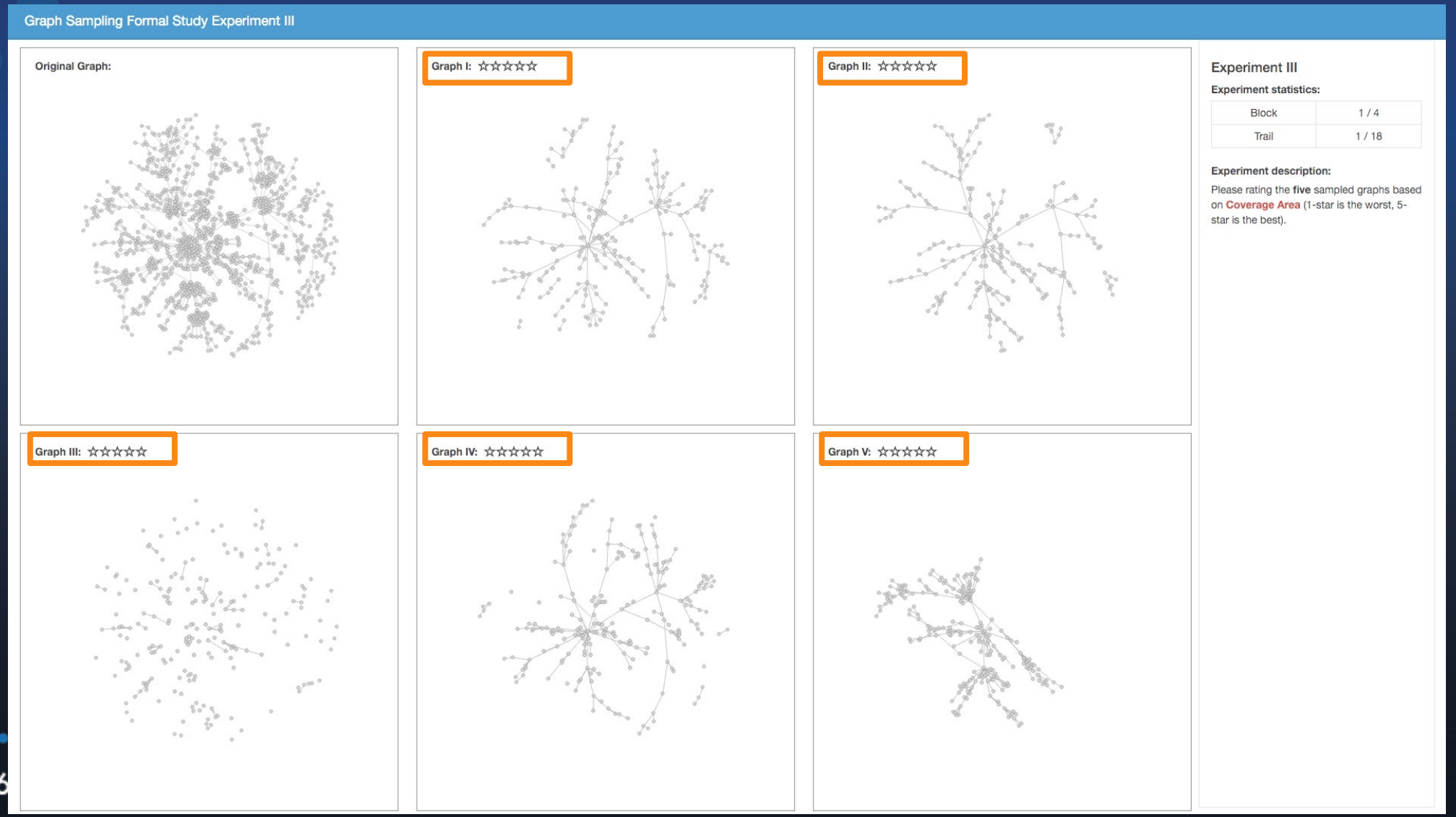


The number of clusters remained is important for perceiving the cluster quality in visualization!

Outline

- Selected Sampling Methods
- Pilot Study
- Formal Studies
  - Perception of High Degree Nodes (more high degree nodes are perceived in *RW*)
  - Perception of Cluster Quality (cluster number is important)
  - Perception of Coverage Area

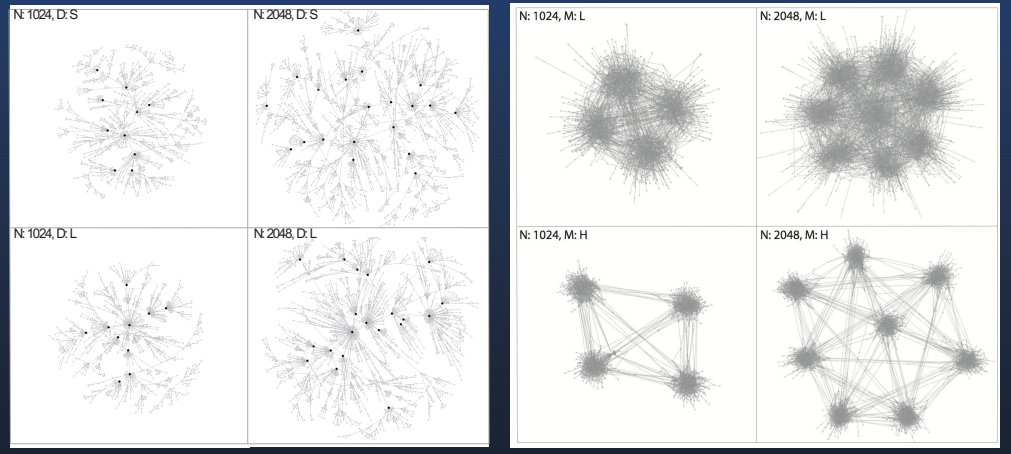
Formal Study III: Coverage Area



Formal Study III: Coverage Area

	2	graph models (Barabási-Albert model [7] and Sah et al.'s model [46])
	2	graph sizes (small=1024 nodes, large=2048 nodes)
	2	corresponding parameters for each graph model
	3	random seeds (3 different seeds)
×	3	repetitions
	72	trials per participant
×	24	participants
	1728	**trials in total**

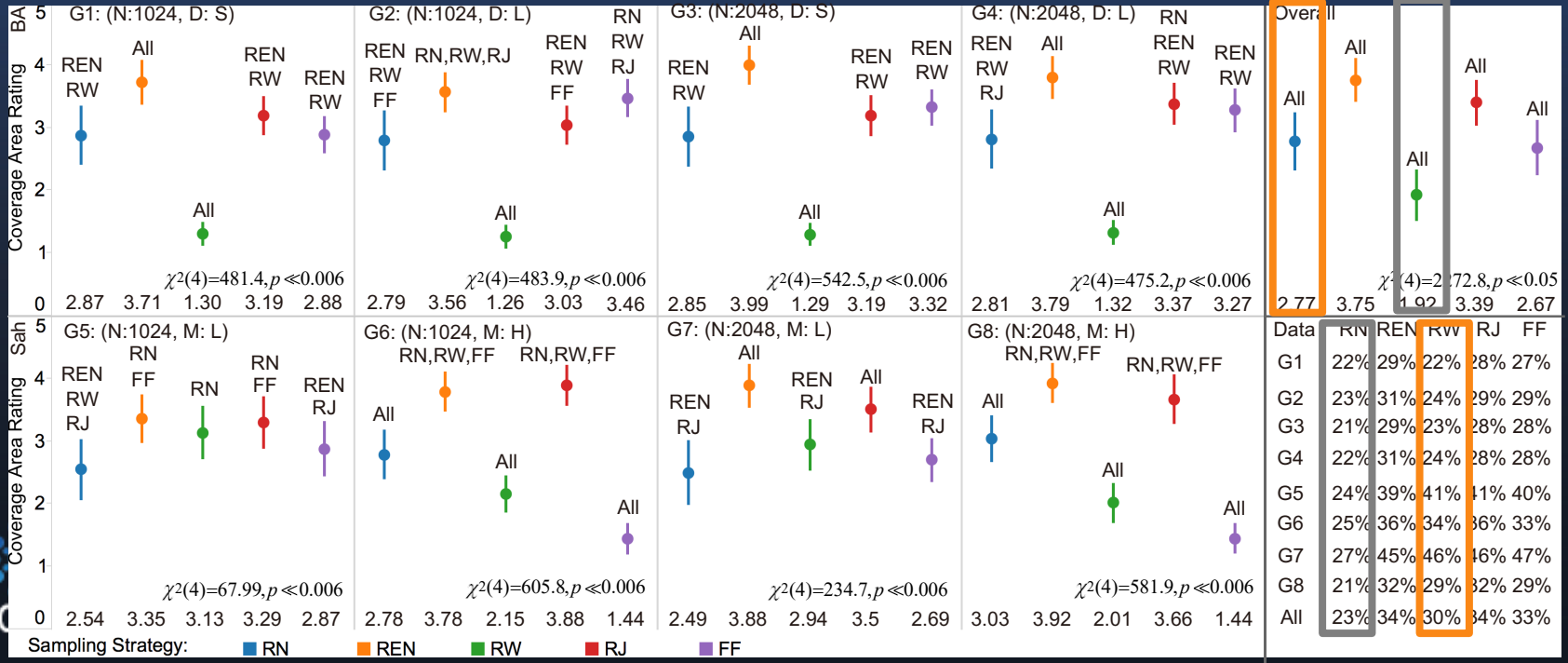
Experiment Setting



Data Generation

Formal Study III: Coverage Area Results

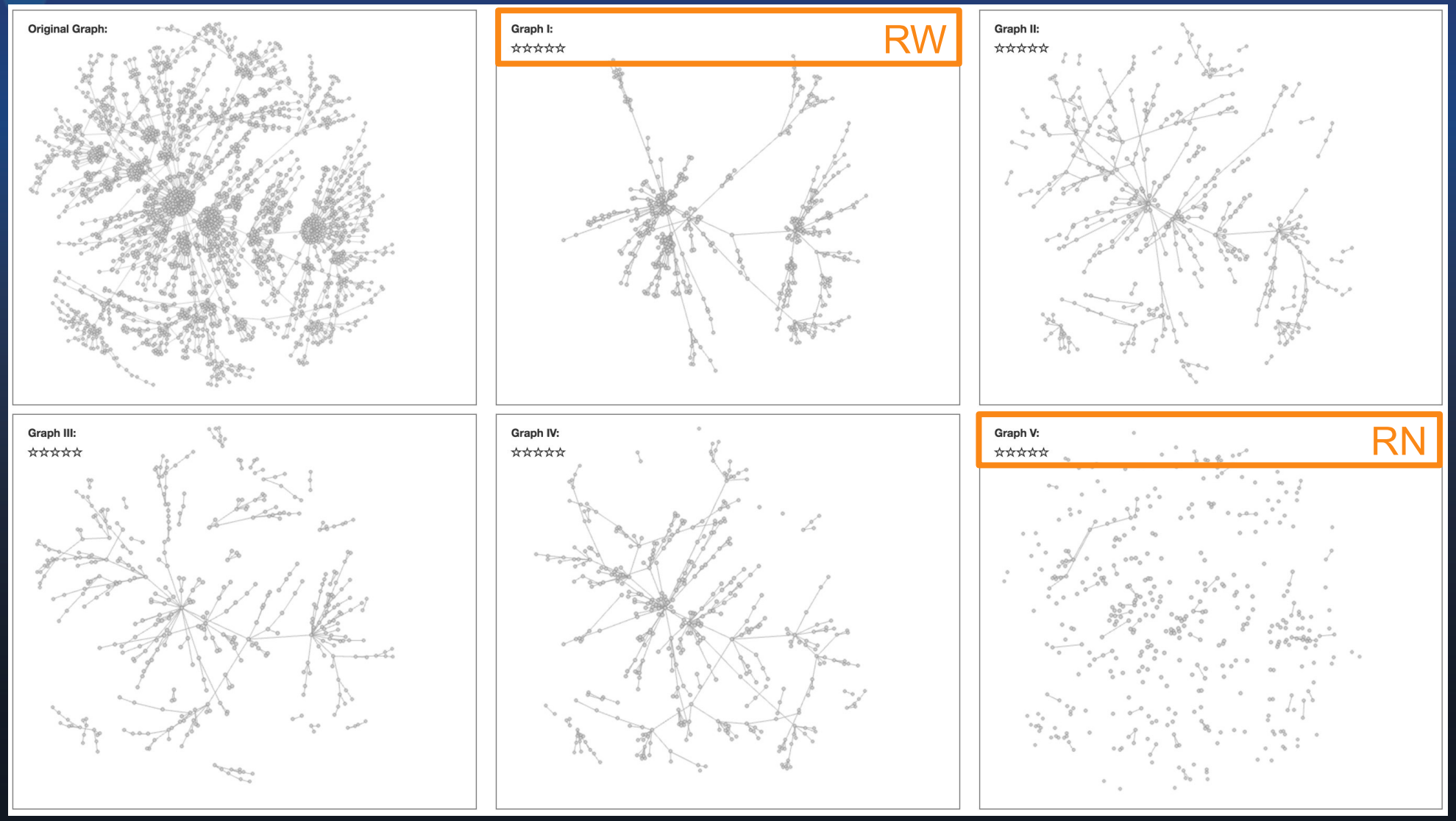
- Discussions:
  - *RE* and *RJ* have the largest perceived coverage area
  - *RW* has a smallest perceived coverage area in most cases
  - *RW* and *FF* 's performance vary depending on graph properties



Data	RN	REN	RW	RJ	FF
G1	22%	29%	22%	28%	27%
G2	23%	31%	24%	29%	29%
G3	21%	29%	23%	28%	28%
G4	22%	31%	24%	28%	28%
G5	24%	39%	41%	41%	40%
G6	25%	36%	34%	36%	33%
G7	27%	45%	46%	46%	47%
G8	21%	32%	29%	32%	29%
All	23%	34%	30%	34%	33%

Contradiction with metric-based results!

Formal Study III: Coverage Area Results



VIS 2016

Conclusion

- We provided the first study of how graph sampling strategies can influence the perception of node-link visualizations
  - Important visual factors: high degree nodes, cluster quality, and coverage area
  - Recommendations for sampling network visualizations:
    - Recommend *Random Edge* and *Random Jump* for global structure and cluster quality
    - Recommend *Random Walk* for perceived high degree nodes
    - Use *Random Node* unless for specific requirements
    - *Random Walk* and *Forest Fire* are modularity sensitive

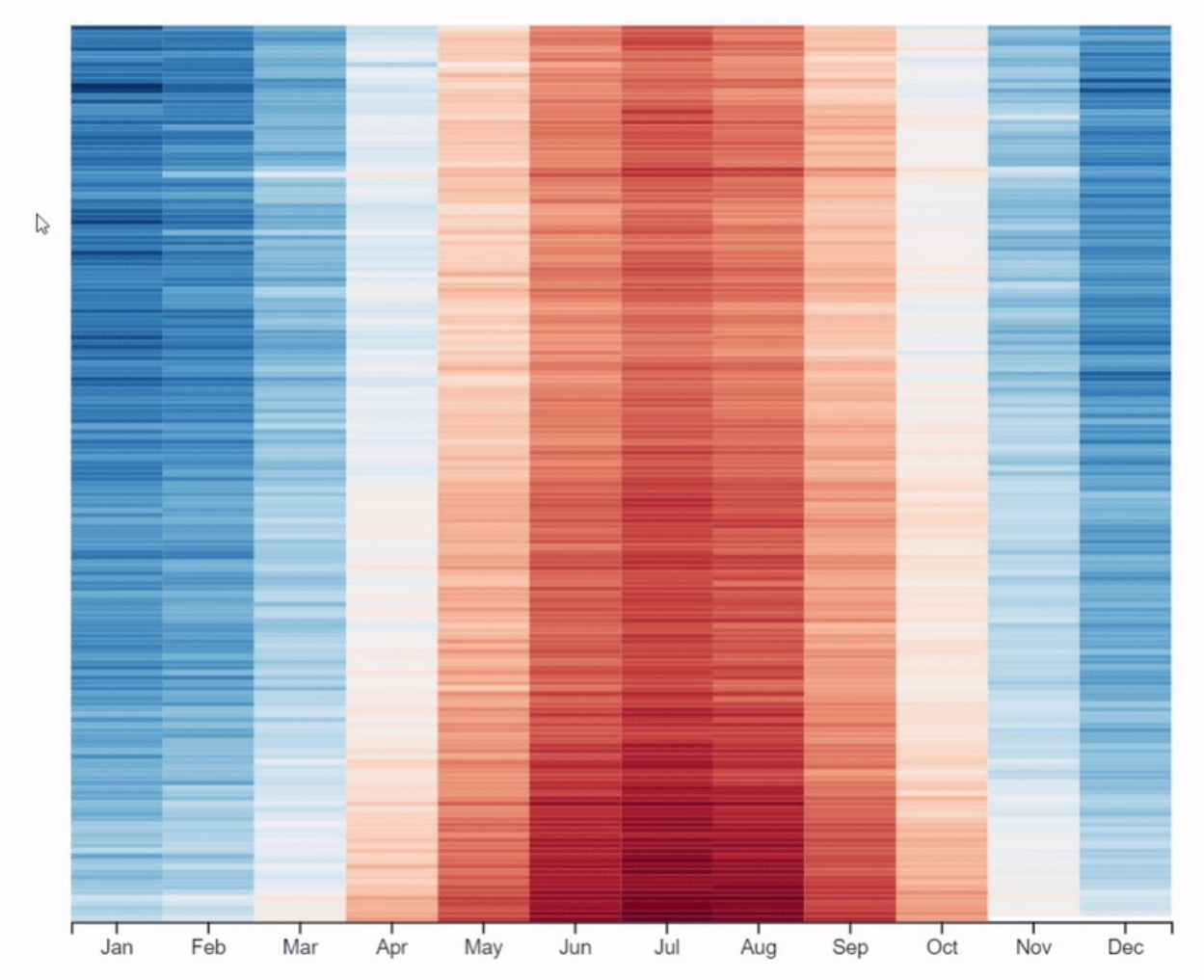
Graph sampling performance in visualization may **VARY** from previous metric-based results!

VIS 2016

Homework 1

- Task description:
  - Visualize the global monthly mean temperature data with JavaScript.
  - D3.js is the only library allowed to finish this assignment.
  - Provided data: temperature.csv
  - Not hard or time-consuming but help you quickly become familiar with D3.js.
- The detailed requirements will be provided later in Piazza.
- Deadline: Tuesday, Mar.26 at 23:59
- Submission: Through ShanghaiTech Pan

https://epan.shanghaitech.edu.cn/l/FkmC6q



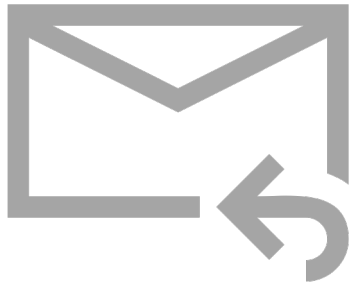
上海科技大学
ShanghaiTech University

立志成才报国裕民

In-class Training

- The 2011 Visualization Contest challenge was related to the spread of the epidemic. The task consisted of visualizing millions of tweets posted by residents of the virtual metropolis of Vastopolis during the epidemic to discover trends in the spread of the epidemic
  - Dataset 1: tweets collected from various GPS-enabled devices.
  - Dataset 2: Geographic information for the entire city: including satellite maps, demographics and weather conditions.
- Task objectives.
  - Identify the approximate geographic location of the epidemic outbreak, mark the infected area, and explain why you came to this conclusion?
  - What is the mode of transmission of the epidemic? For example, is it person-to-person transmission, airborne, waterborne, or other? Explain why.

Questions?
Thank you ☺



Quan Li

liquan@shanghaitech.edu.cn