Visualization in Data Science: What is it for?

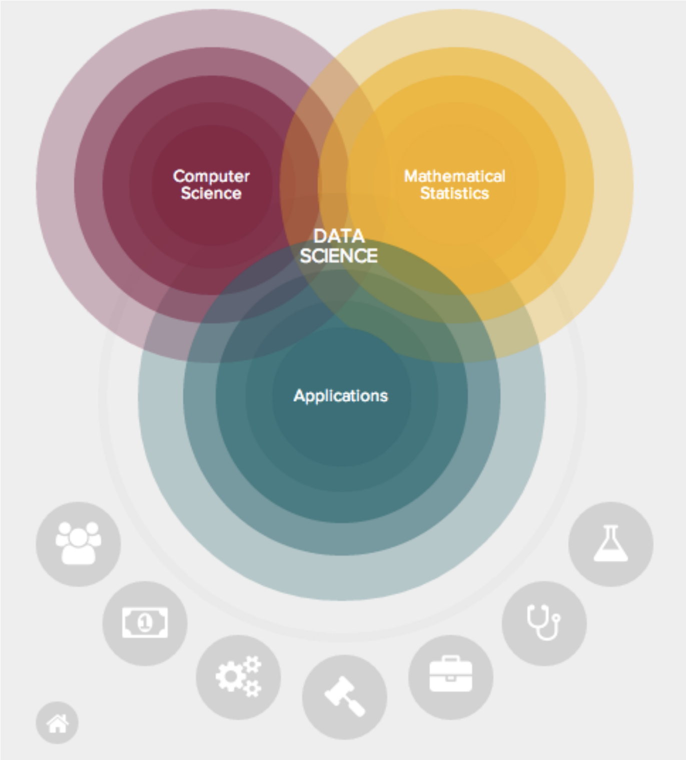
Enrico Bertini

NYU | POLYTECHNIC SCHOOL OF ENGINEERING

What is data science

What is Data Science?

There is much debate among scholars and practitioners about what data science is, and what it isn't. Does it deal only with big data? What constitutes big data? Is data science really that new? How is it different from statistics and analytics?



Data-Driven Discovery

WHAT DATA SCIENCE MEANS FOR RESEARCH

In virtually all areas of intellectual inquiry, data science offers a powerful new approach to making discoveries. By combining aspects of statistics, computer science, applied mathematics, and visualization, data science can turn the vast amounts of data the digital age generates into new insights and new knowledge.

Click on the icons to the left to see how social scientists, medical researchers, and many others are using data science to advance their fields.

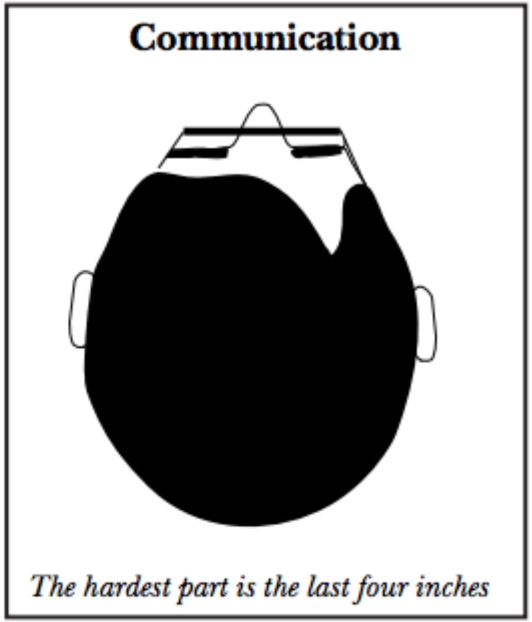
One way to consider data science is as an evolutionary step in interdisciplinary fields like business analysis that incorporate computer science, modeling, statistics, analytics, and mathematics.

At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them. With such automated methods turning up everywhere from genomics to high-energy physics, data science is helping to create new branches of science, and influencing areas of social science and the humanities. The trend is expected to accelerate in the coming years as data from mobile sensors, sophisticated instruments, the web, and more, grows. In academic research, we will see an increasingly large number of traditional disciplines spawning new sub-disciplines with the adjective "computational" or “quantitative” in front of them. In industry, we will see data science transforming everything from healthcare to media.

But knowledge is something that happens in our head.

How do we get information from the computer into our head?

… and from our head back into the computer?



Frederick P. Brooks, Jr.

Fred Brooks is the first recipient of the **ACM Allen Newell Award***—an honor to be presented annually to an individual whose career*
other disciplines. Brooks was honored for a breadth
and engineering and his interdisciplinary contribu
Here, we present his acceptance lecture delivered at

© Jerry Markatos

The Computer Scientist as Toolsmith II

It is a special honor to receive an award named for Allen Newell. Allen was one of the fathers of computer science. He was especially important as a visionary and a leader in developing artificial intelligence (AI) as a subdiscipline, and in enunciating a vision for it.

What a man *is* is more important than what he does professionally, however, and it is Allen's humble, honorable, and self-giving character that makes it a double honor to be a Newell awardee. I am profoundly grateful to the awards committee.

Rather than talking about one particular research area, I should like to stay in the spirit of the Newell Award by sharing some lifetime reflections on the computer science enterprise, reflections which naturally reflect my convictions about the universe. The title and opening section of this talk were first formulated for a 1977 speech [1]. Let me reiterate the points, since many of you were barely born then.

In some quarters and at some times, computer graphics has been seen as a left-handed stepchild of computer science. Another view of computer science sees it as a discipline focused on problem-solving systems, and in this view computer graphics is very near the center of the discipline.

A Discipline Misnamed

When our discipline was newborn, there was the usual perplexity as to its proper name. We at Chapel Hill, following, I believe, Allen Newell and Herb Simon, settled on "computer science" as our department's name. Now, with the benefit of three decades' hindsight, I believe that to have been a mistake. If we understand why, we will better understand our craft.

What Is a Science?

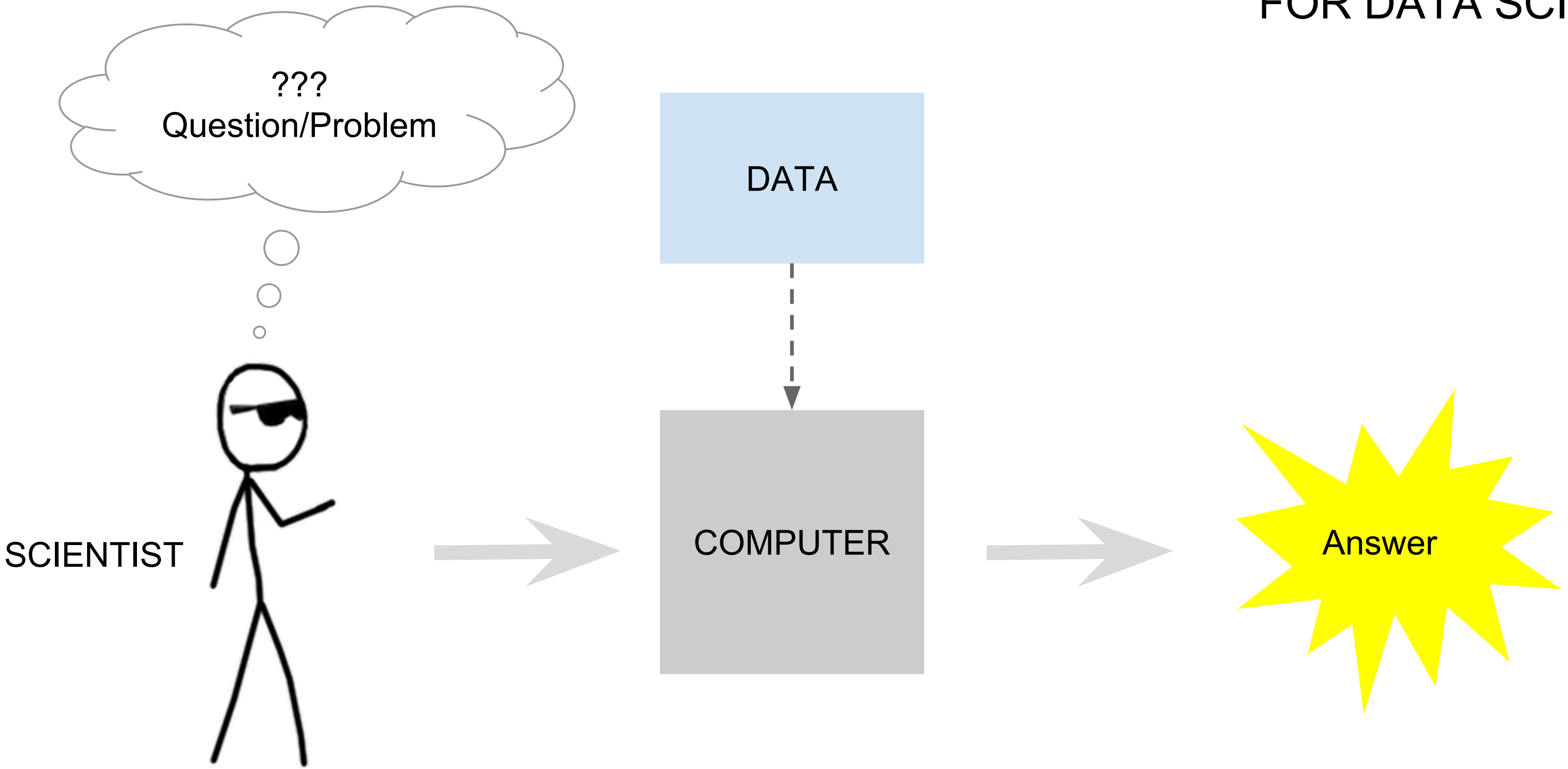
Webster says science is "a branch of study concerned with the observation and classification of facts, especially with the establishment and quantitative formulation of verifiable general laws." [2]

This puts it pretty well—a science is concerned with the *discovery* of facts and laws.

A folk adage of the academic profession says, "Any-

Brooks Jr, Frederick P. "The computer scientist as toolsmith II."*Communications of the ACM* 39.3 (1996): 61-68.

AN ORACLE MODEL FOR DATA SCIENCE



Problems

1) Formulating the question.
2) Understanding the answer.
3) Getting an explanation for the answer.

Note: all three happen at the interface between the human and the machine.

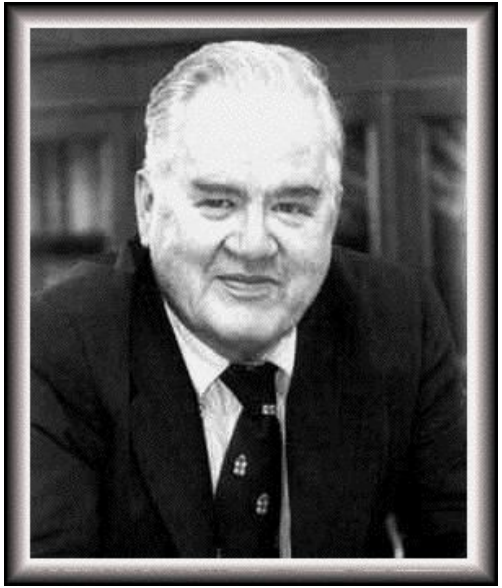
more (practical) problems ...

Is that the right data?

Is that the right model/algorithm?

Is that the right parameter set?

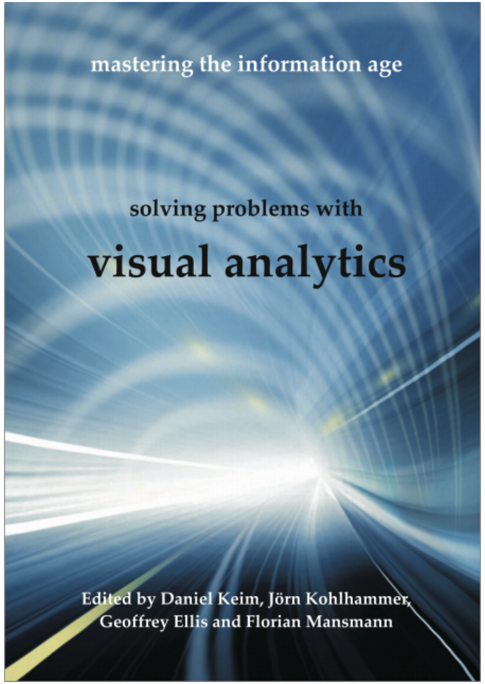
… and by the way do these data contain an answer for my problem at all?

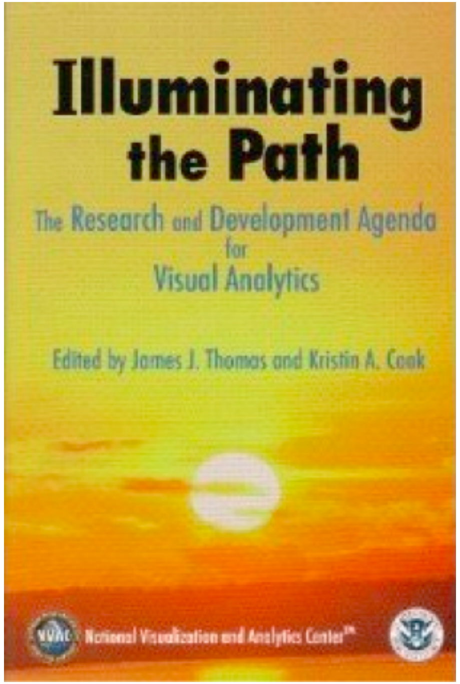


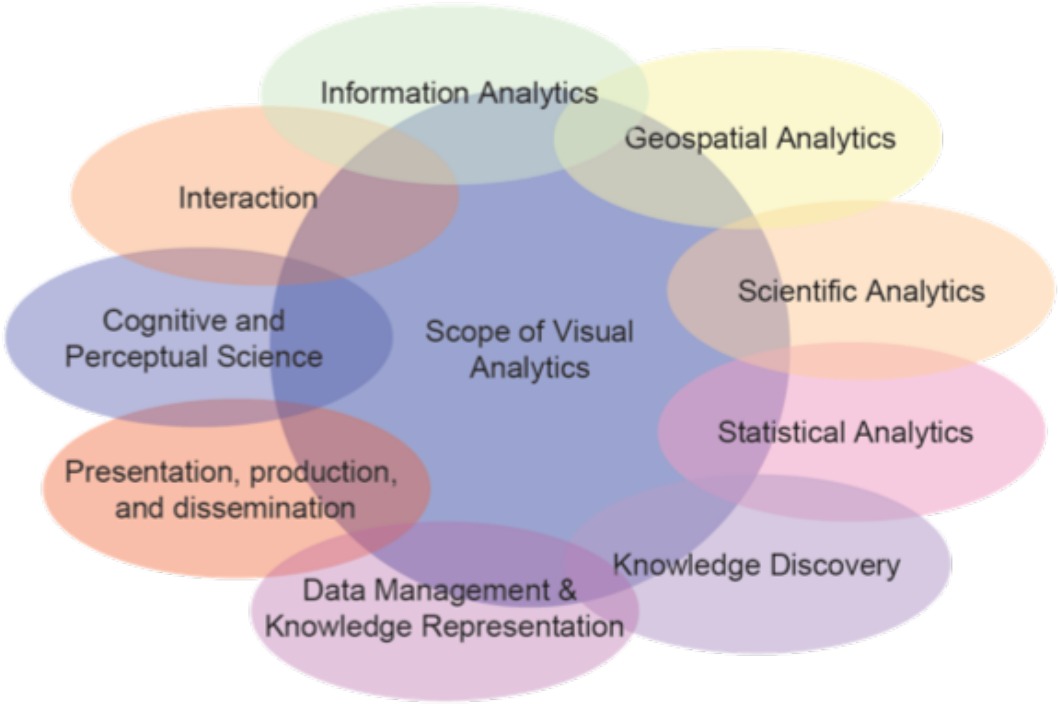
JOHN W. TUKEY*

We often forget how science and engineering function. Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. Broad general inquiries are also important. Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught. Confirmatory data analysis, by contrast, is easier to teach and easier to computerize. We need to teach both; to think about science and engineering more broadly; to be prepared to randomize and avoid multiplicity.

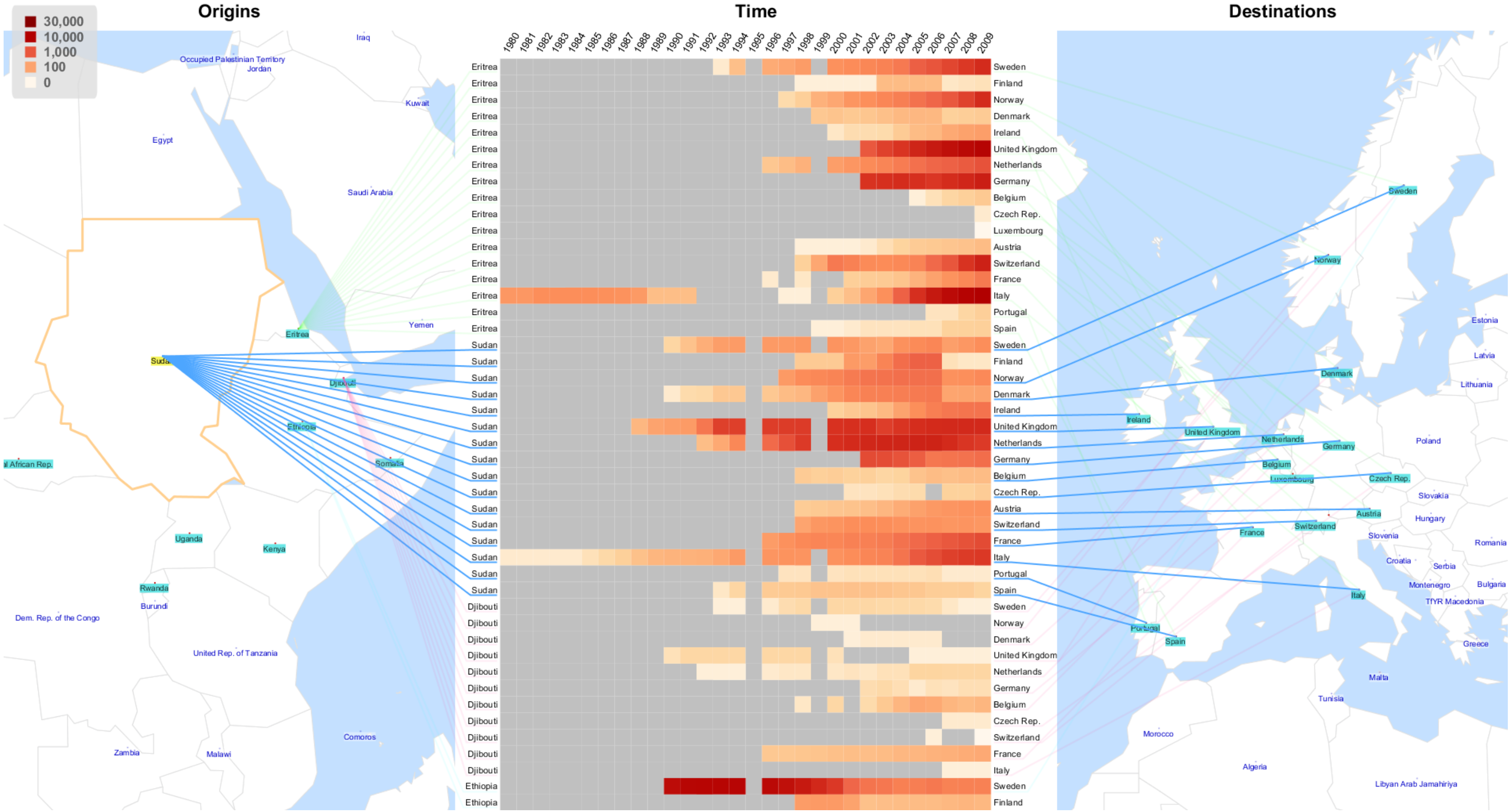
Visual Analytics: "The science of analytical reasoning facilitated by interactive visual interfaces"






Information Analytics
Geospatial Analytics
Interaction
Scientific Analytics
Cognitive and Perceptual Science
Scope of Visual Analytics
Statistical Analytics
Presentation, production, and dissemination
Knowledge Discovery
Data Management & Knowledge Representation

Example: Flowstrates



What is Visualization for?

Sensemaking

Discovery

Communication

(monitoring / situational awareness)

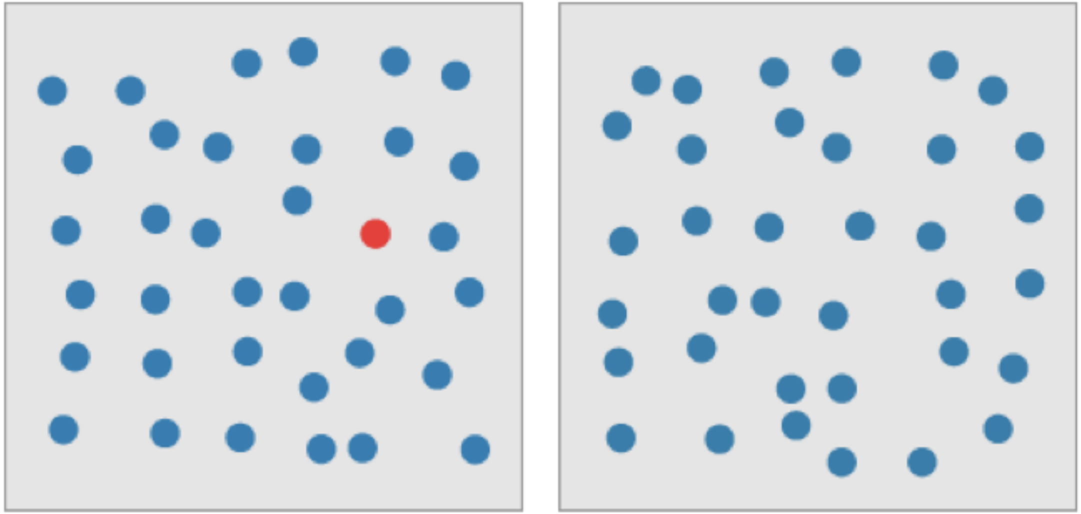
Why Use Visualization?

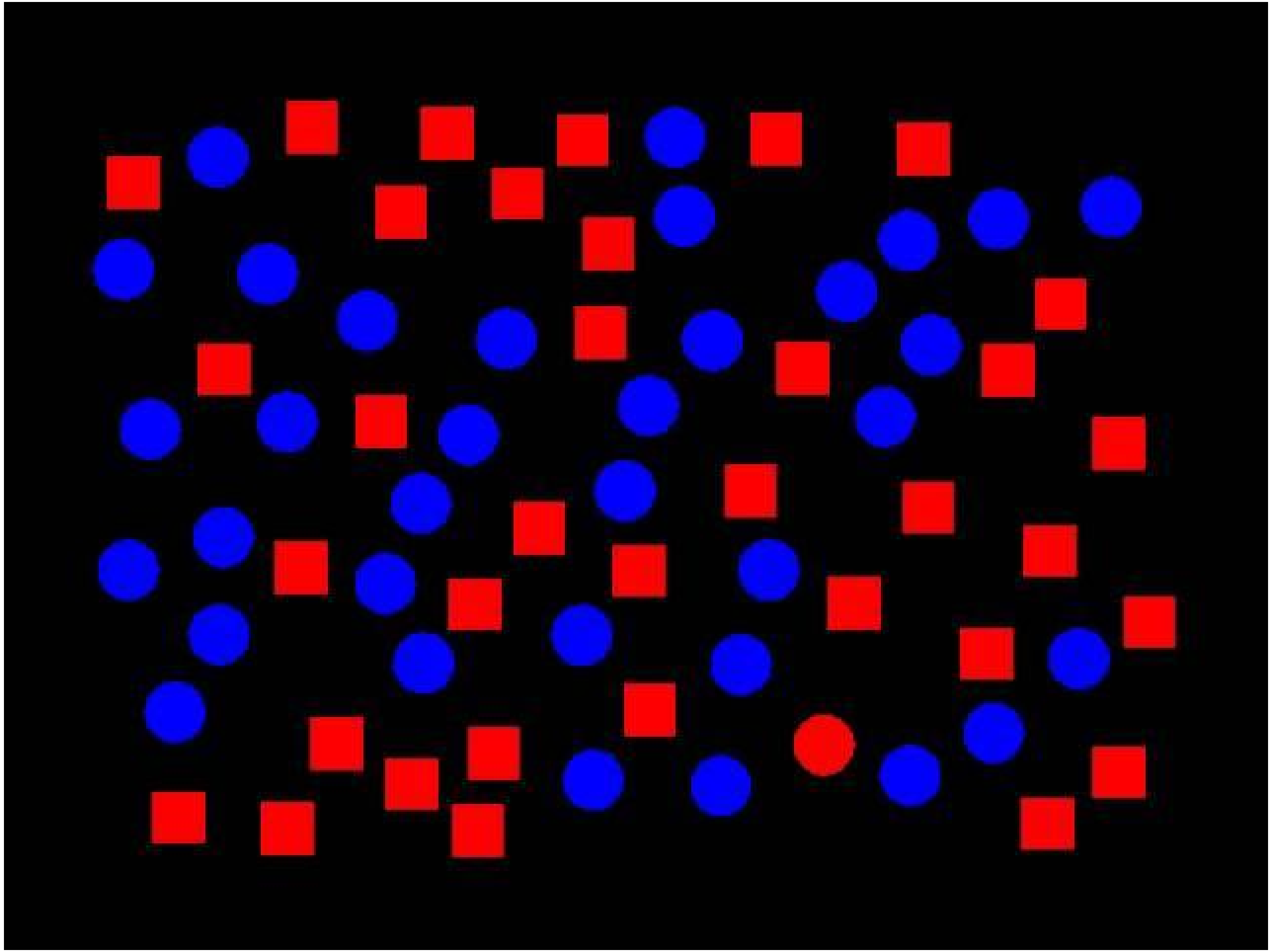
1) Vision is the most powerful communication channel humans posses.

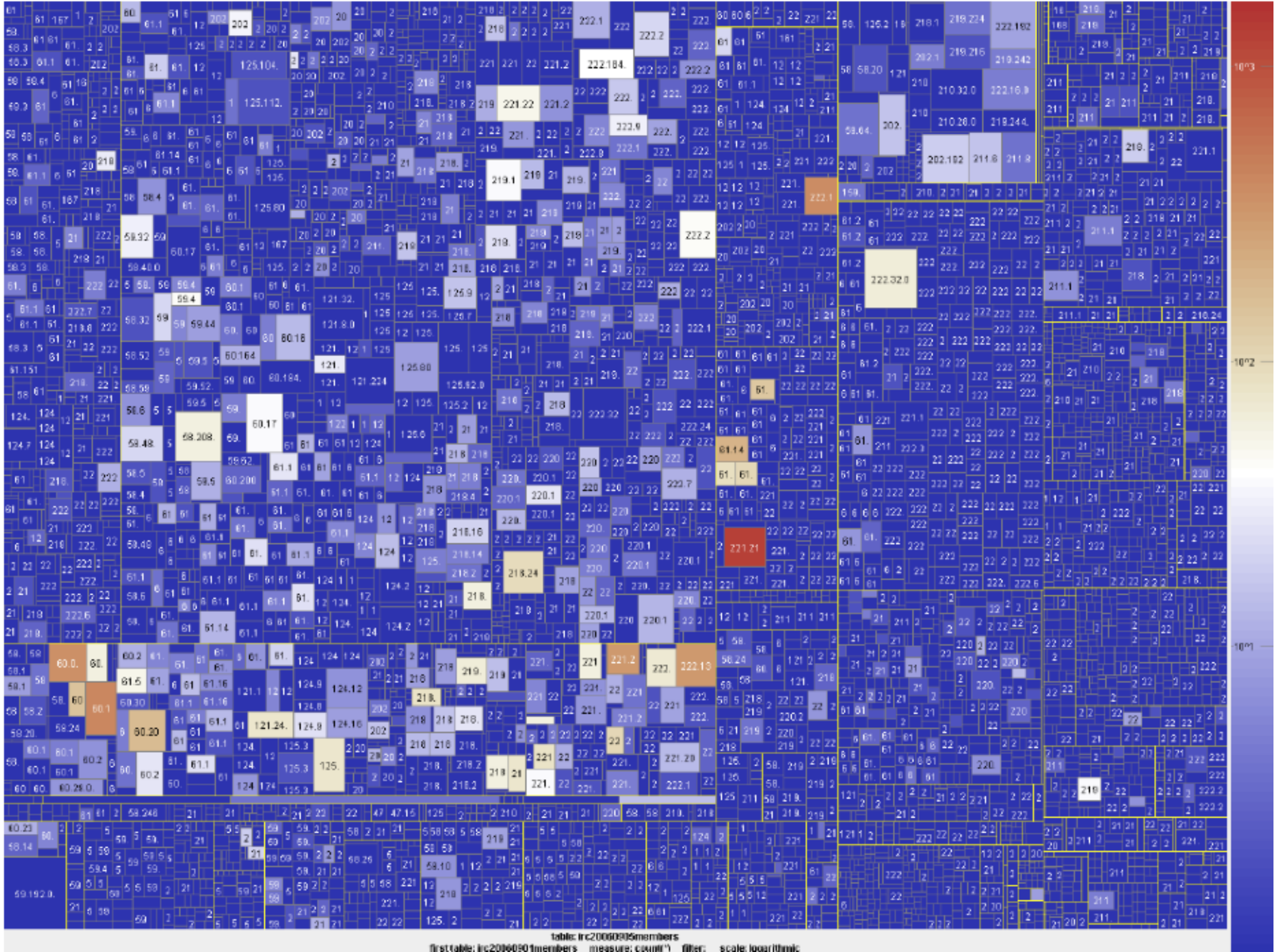
For instance, We can detect information faster than our eye can move ...

Preattentive Processing

Preattentive features can be detected faster than eye movement (200 msec).







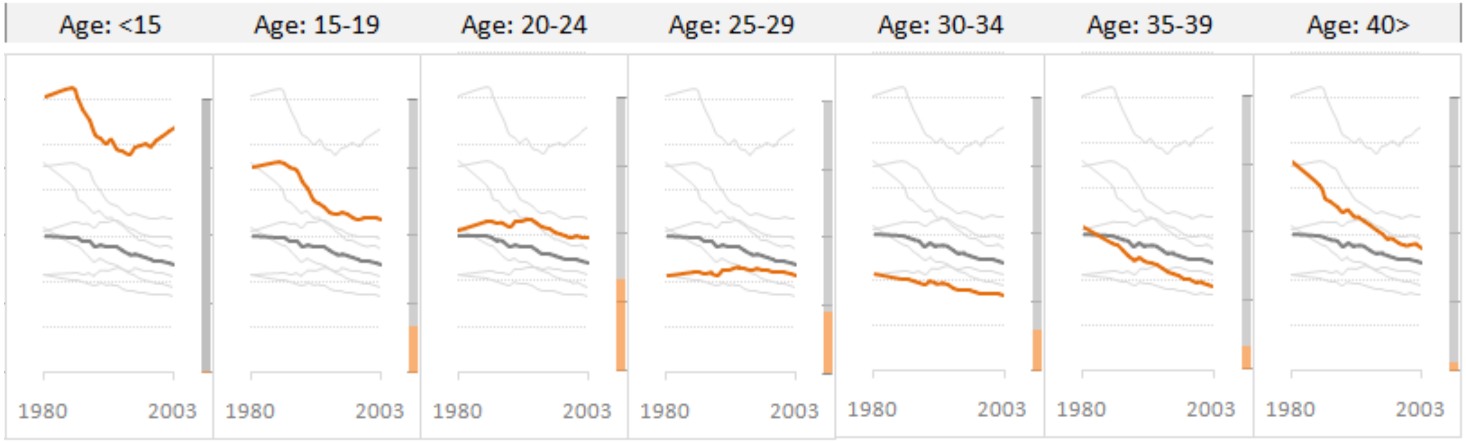
10^3
10^2
10^1
table: irc20060905members
first table: irc20060901members measure: count(*) filter: scale: logarithmic

2) Humans are not very good at detecting patterns from numbers.

	1980	1985	1986	1987	1988	1989	1990	1991	1992	1993
Less than 15 years old	607	624	605	578	553	523	515	502	511	492
15 to 19 years old	451	462	457	449	444	418	403	379	370	364
20 to 24 years old	310	328	328	327	327	318	328	330	333	334
25 to 29 years old	213	219	219	216	218	213	224	224	228	230
30 to 34 years old	213	203	201	197	194	189	196	192	192	189
35 to 39 years old	317	280	277	265	254	244	249	241	239	234
40 years old and over	461	409	381	374	361	350	354	339	338	329

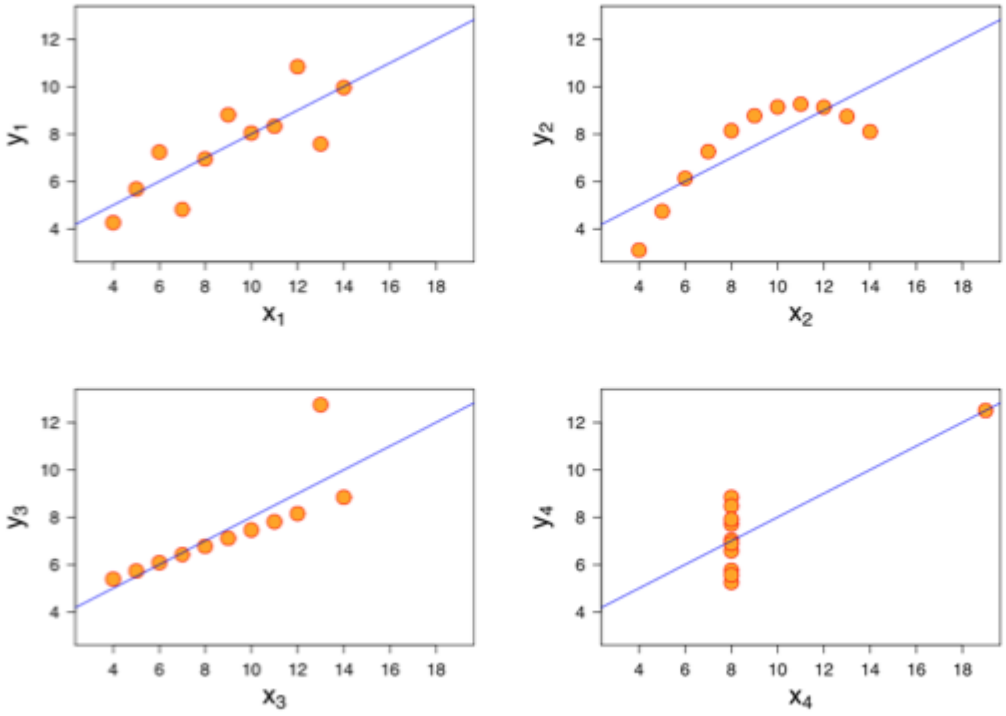
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Less than 15 years old	488	479	493	498	504	497	512	519	529	537
15 to 19 years old	353	347	350	346	341	337	339	341	339	337
20 to 24 years old	326	317	314	307	301	297	296	298	296	293
25 to 29 years old	227	224	228	226	224	221	220	219	215	211
30 to 34 years old	183	179	178	176	174	171	169	171	169	167
35 to 39 years old	226	219	215	208	203	200	195	195	190	186
40 years old and over	320	309	301	291	290	283	276	276	278	268

Which group has the highest/lowest rates? When?
Which group has an increasing/decreasing temporal trend?
Which group has a faster/slower rate of change?



3) Summary statistics can hide important information.

Anscombe's Quartet



Property	Value
Mean of x in each case	9 (exact)
Variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

4) The “right” representation can turn complex problems into simple ones.

(Experiential Vs. Reflective Thinking)

Let’s Play a Game! The Game of “15”

RULES

1) There are 2 players

2) Each player takes a digit in turn

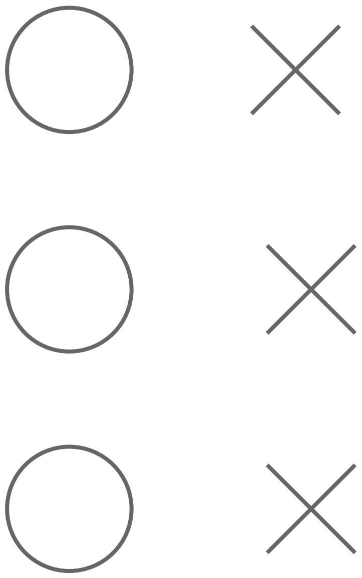
3) Once a digit is taken, it cannot be used by any of the players again

4) The first player to get three digits that sum to 15 wins

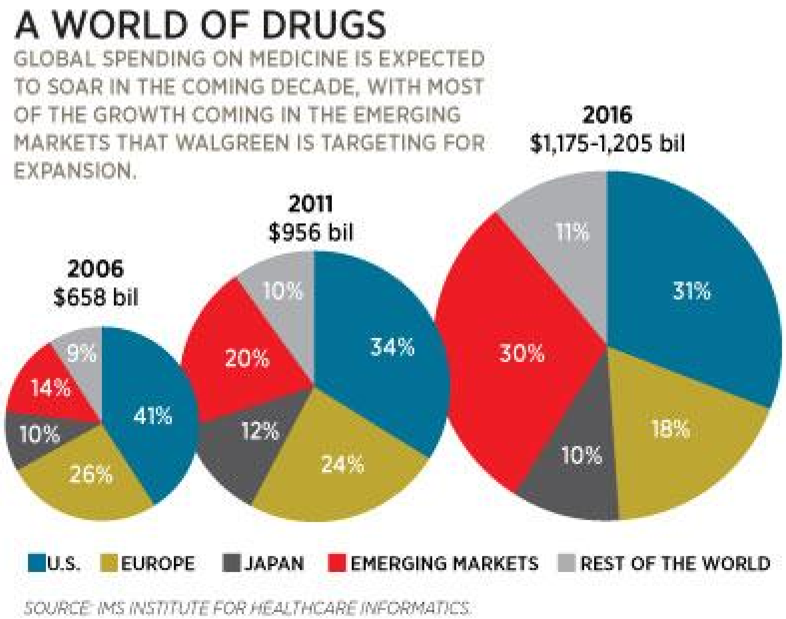
$$\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

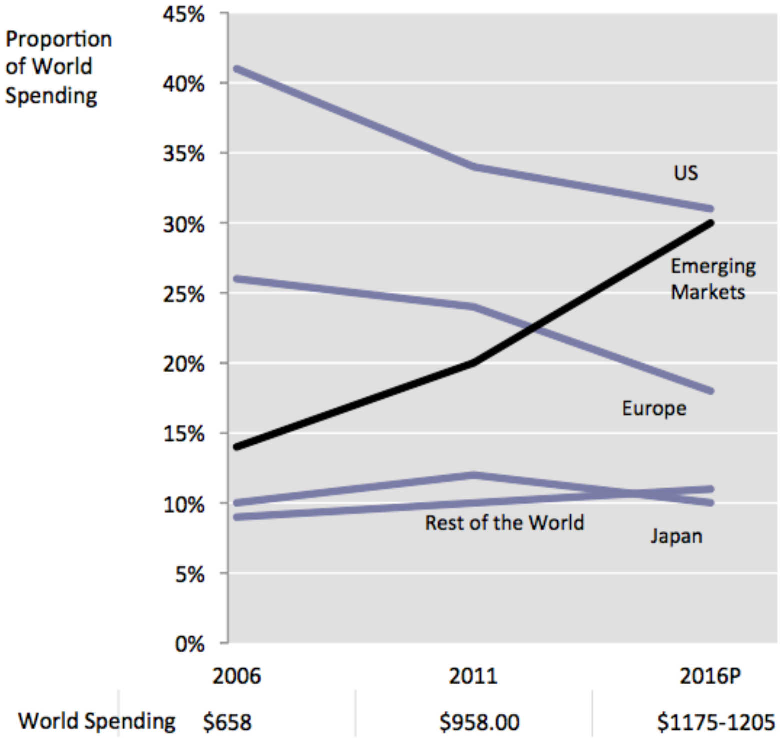
Tic-Tac-Toe: Herbert Simon’s Problem Isomorph

4	9	2
3	5	7
8	1	6



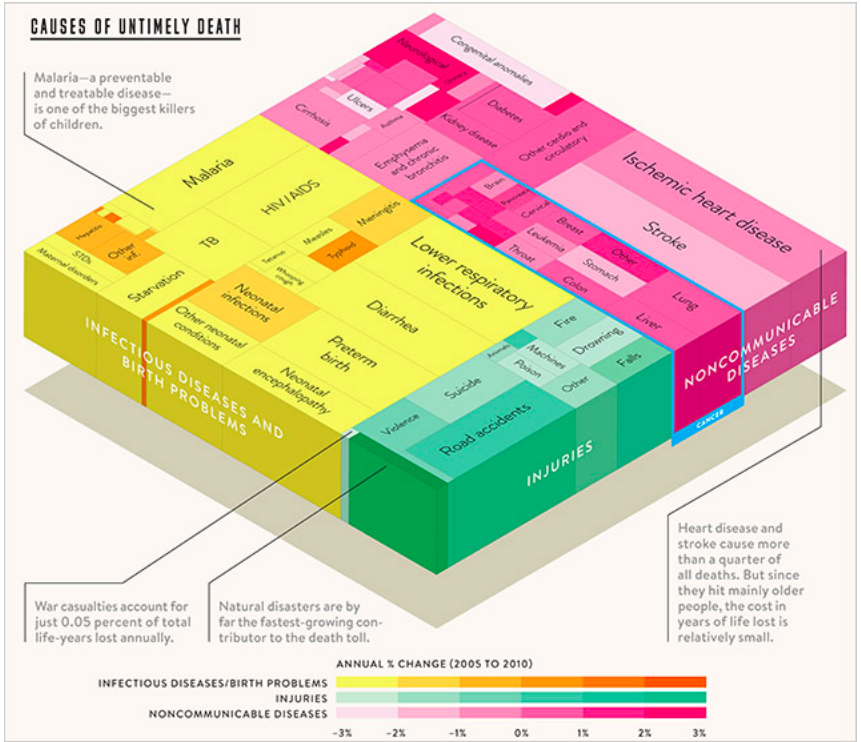
But not all visualizations are equally effective!

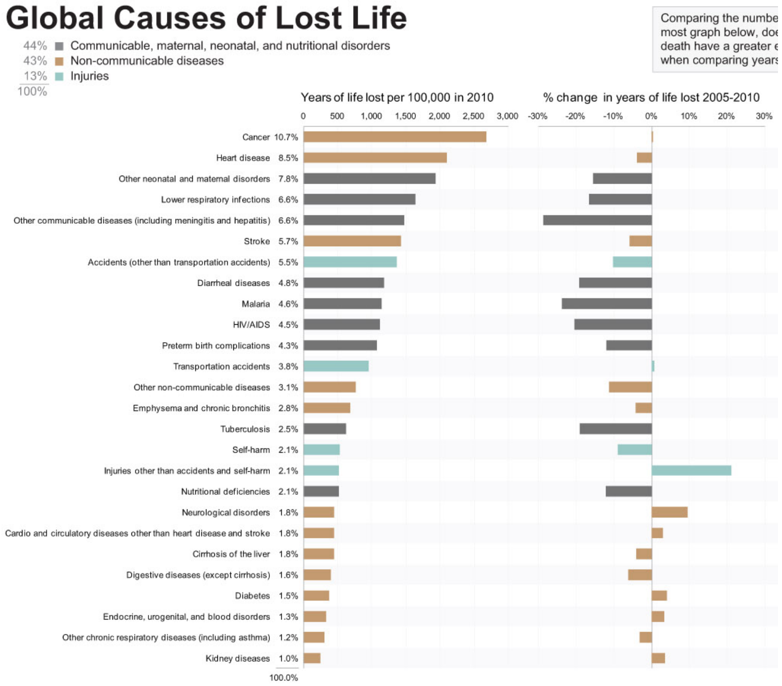




Junk Charts: [Expanding circles of error]

Change of years for life lost between 2005-2010.

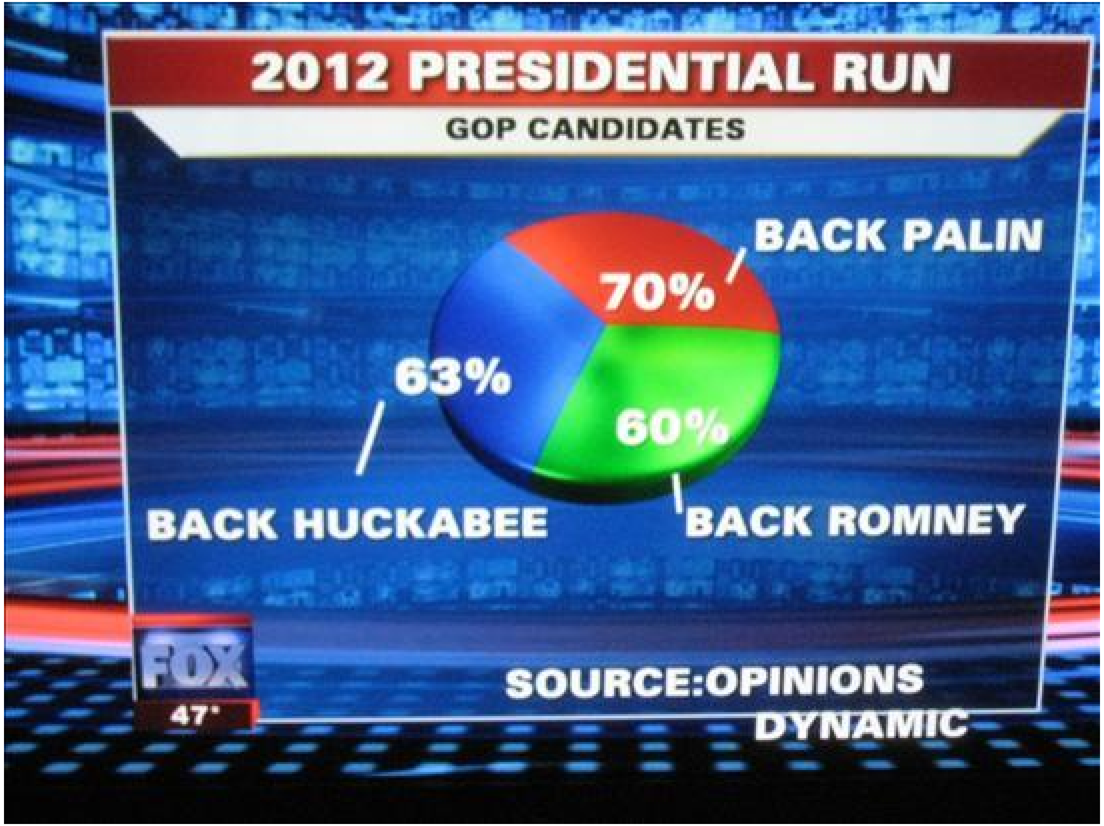




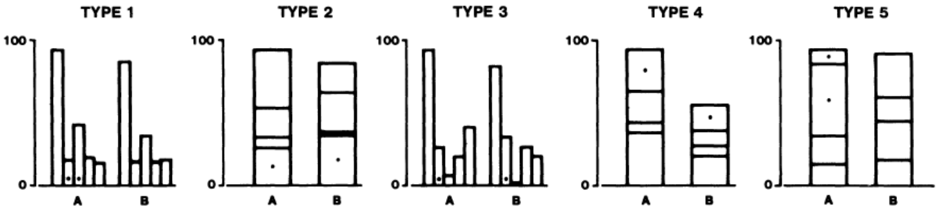
Example taken from Stephen Few's Perceptual Edge: http://www.perceptualedge.com/blog/?p=1829

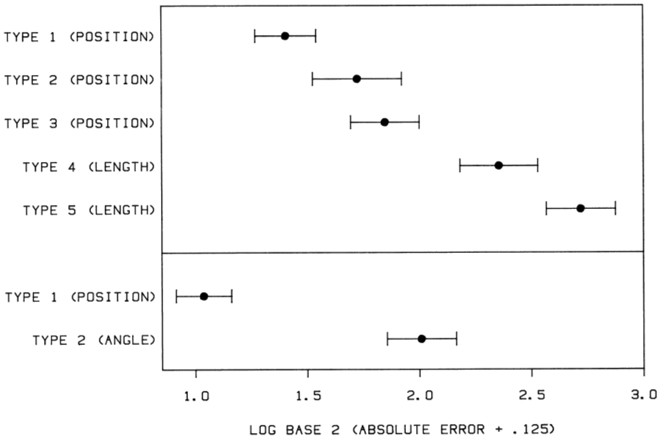
This visualization violates many visualization principles.

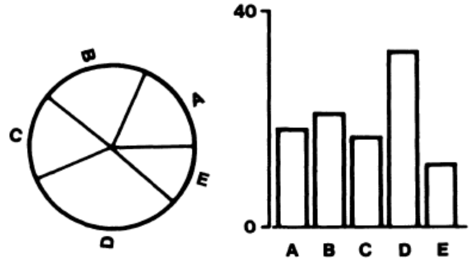
Some are plain wrong!



Ranking of Visual Variables







More Accurate

Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Hue
Angle	Saturation	Density
Slope	Hue	Saturation
Area	Length	Shape
Density	Angle	Length
Saturation	Slope	Angle
Hue	Area	Slope
Shape	Shape	Area

Less Accurate

Cleveland, William S., and Robert McGill. "Graphical perception: Theory, experimentation, and application to the development of graphical methods." *Journal of the American Statistical Association* 79.387 (1984): 531-554.

Automated or Interactive Data Analysis?

Golden Rule of Visual Analytics

Don’t use Visual Analytics if an automated solution exists.

Yet, many problems not well-understood or specified in the first place!

More exploration needed.

a. Exploration to understand the data.

b. Exploration to find good models.

And Visualization does not scale!

There's simply too much data.

Too many dimensions/features.

Too many options and possible views.

Many Data Science problems require a synergistic integration of *automated* **AND** *interactive* methods.

We need methods that support humans in their thinking process.

High-Throughput Screening for Drug Discovery

Assay Plates

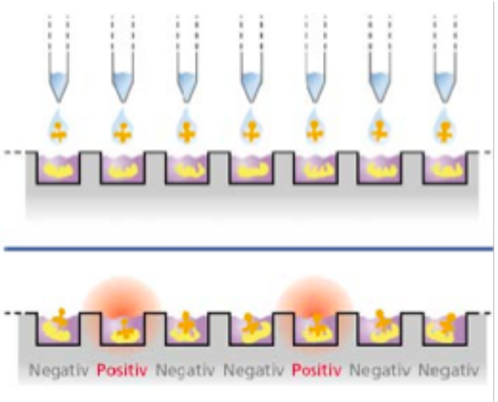
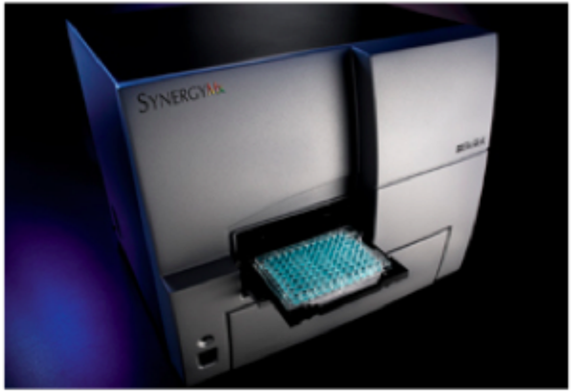
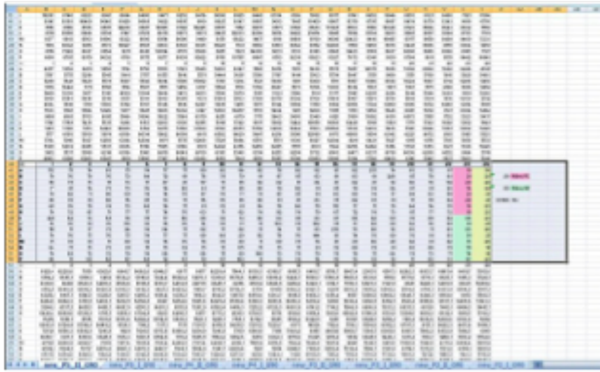


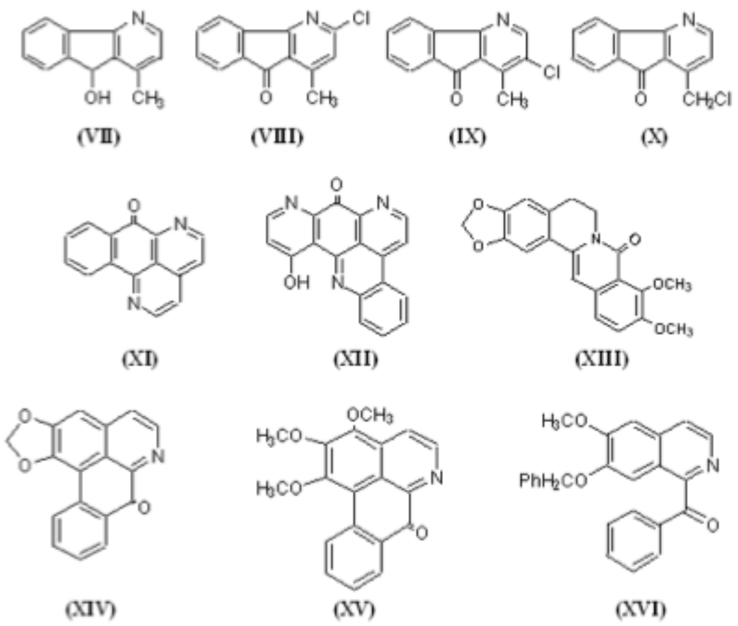
Plate Reader



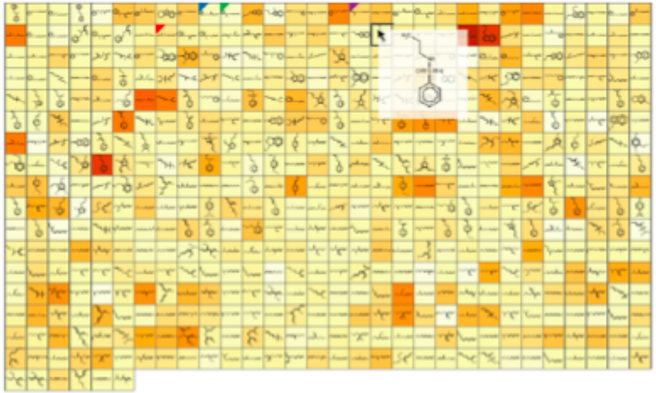
Data



Structure-Activity Relationship (SAR) Analysis

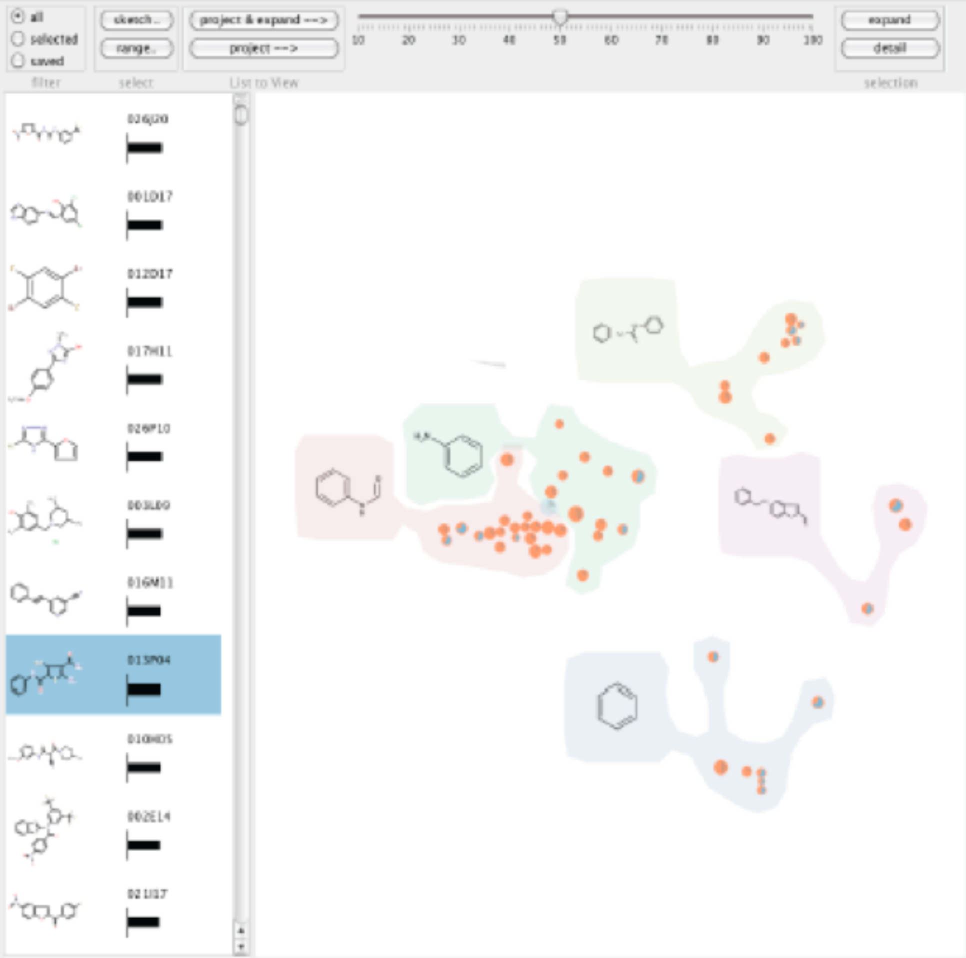


	O-S-C(=O)	S-C(=O)-N	S-C-N(-C)	S-C-N	C-N-C
O-S-C(=O)-N	1	1	0	1	0
S-C(=O)-N-C	0	1	1	1	1



Mining algorithms necessary to extract meaningful molecular fragments.

HiTSEE



Bertini, Enrico, et al. "**HiTSEE: a visualization tool for hit selection and analysis in high-throughput screening experiments.**" Biological Data Visualization (BioVis), 2011 IEEE Symposium on. IEEE, 2011.

High-Dimensional Data Visualization

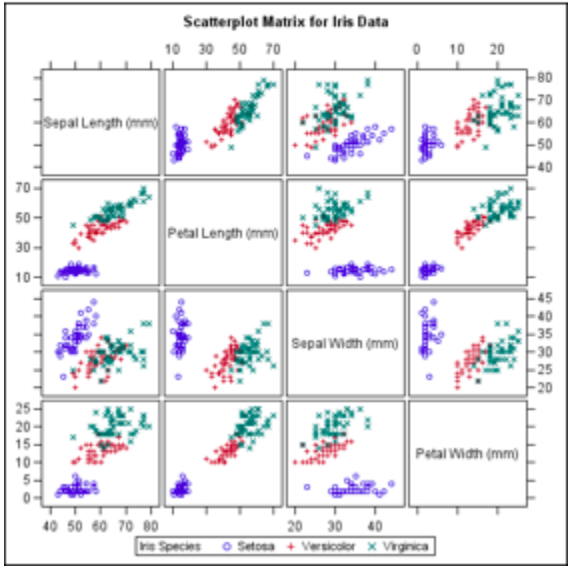
10s, 100s, 1000s, ...

D1	D2	D3	...	...	...	...	...	...	...	...	...	Dk

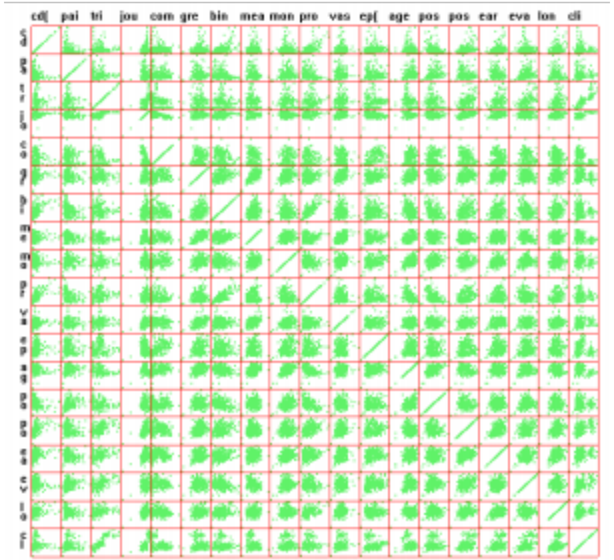
- Set of visual features very limited
- Resolution very limited
- Ability to make sense of it very limited!

Example: Scatter Plot Matrix

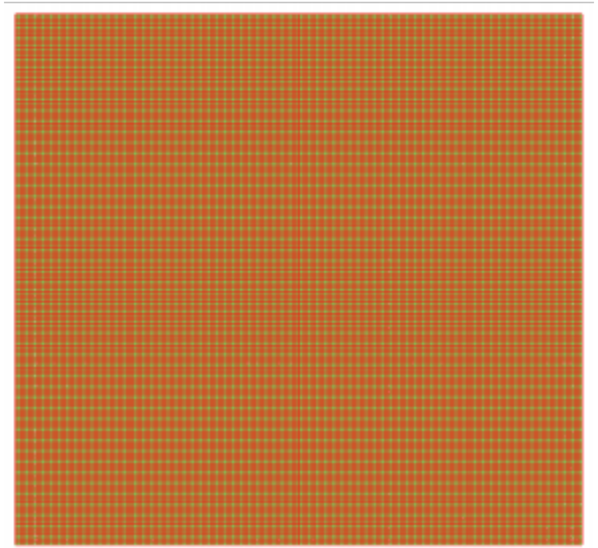
4 dimensions



20 dimensions



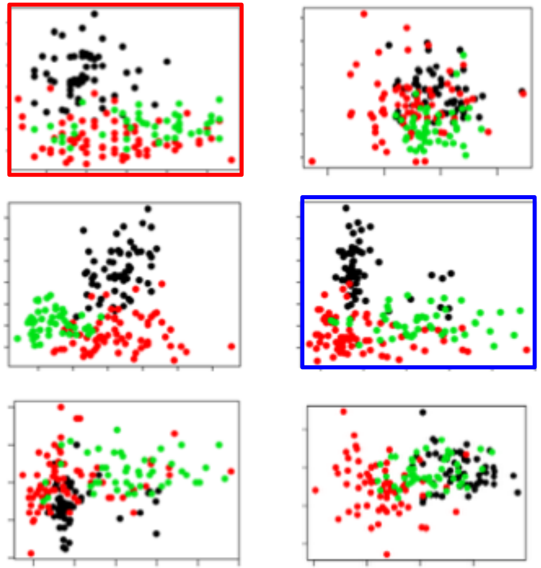
100 dimensions



* Taken from: Jing Yang's [Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets.]

Quality Metrics-Driven Visualization

KA	f2	f3	f4	f5	f6	f7	...	fn
2	60	733	476	72	73	46	...	63
2	422	37	319	4635	422	37	...	463
1	62	201	519	33	612	282	...	39
3	154	22	405	59	15	224	...	45
3	611	320	461	545	61	32	...	54
1	431	67	22	173	431	67	...	173
2	165	0	254	112	244	535	...	12
1	40	16	307	352	43	61	...	52
3	877	327	596	54	77	27	...	55
2	92	12	244	535	192	48	...	35
...	...	...	...	...	...	...	...	...

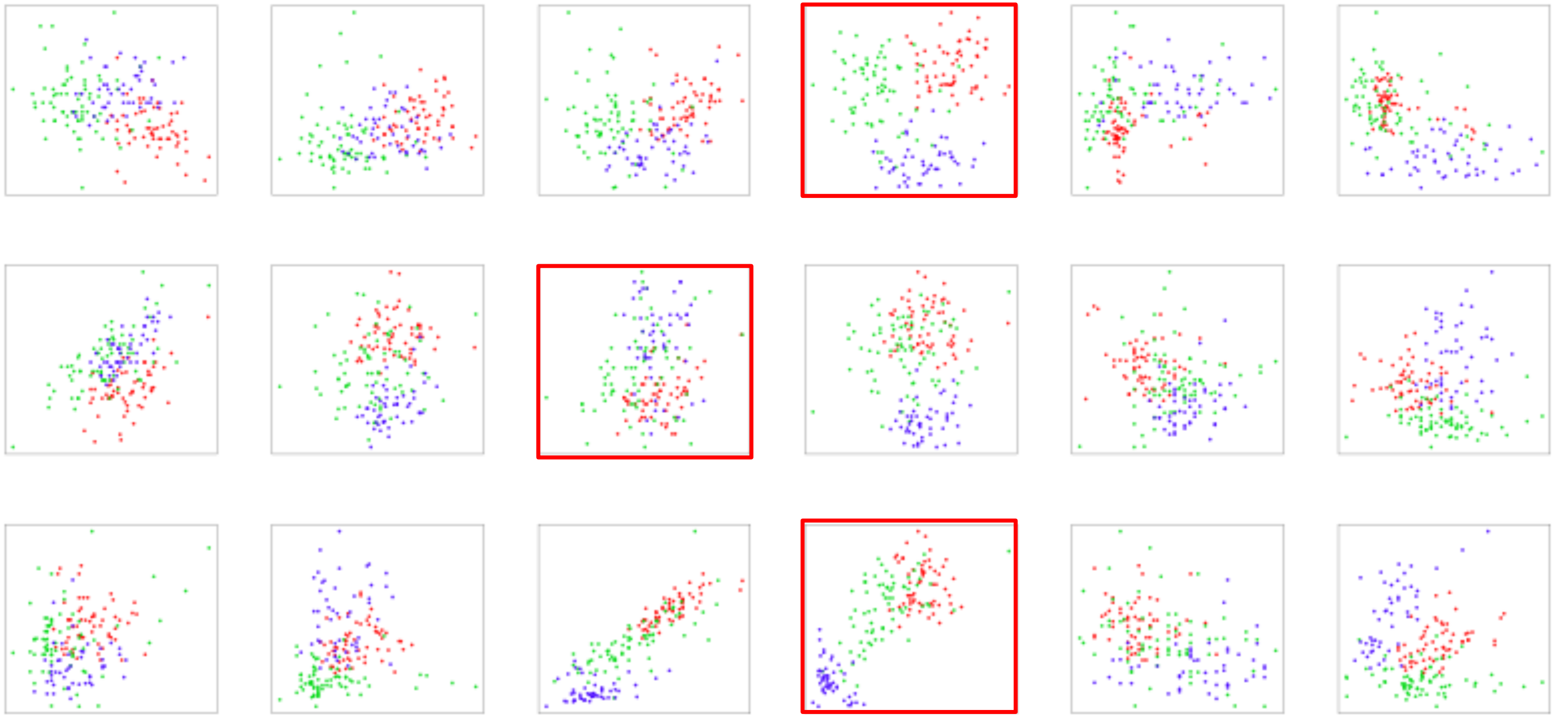


human perception

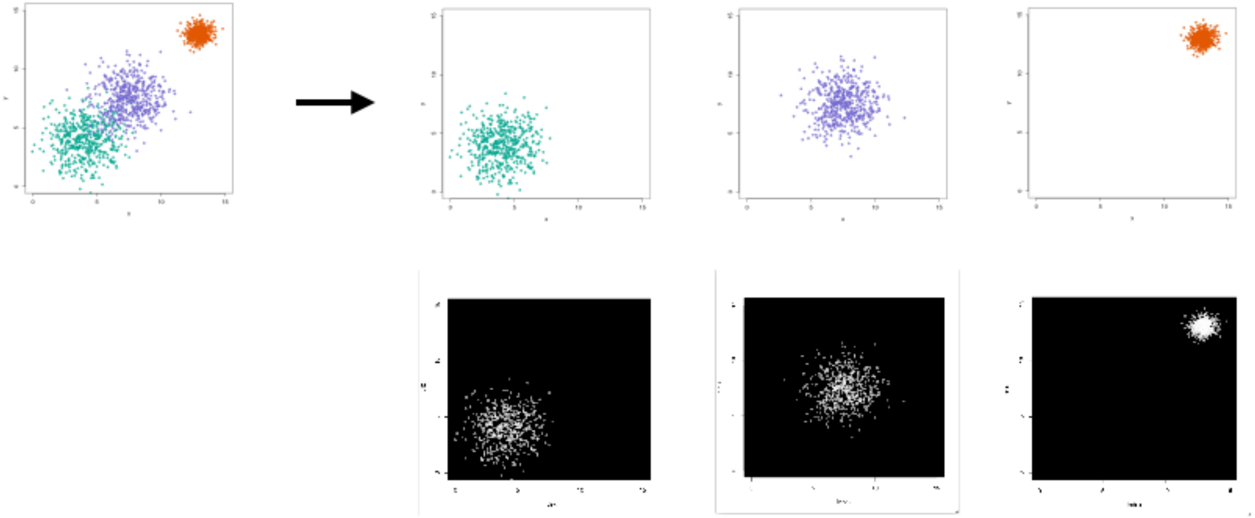
How can we measure the interestingness of a projection?

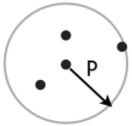
Does it correlate with human perception?

Automatic Ranking of Class Separation



Example: Class Density Measure



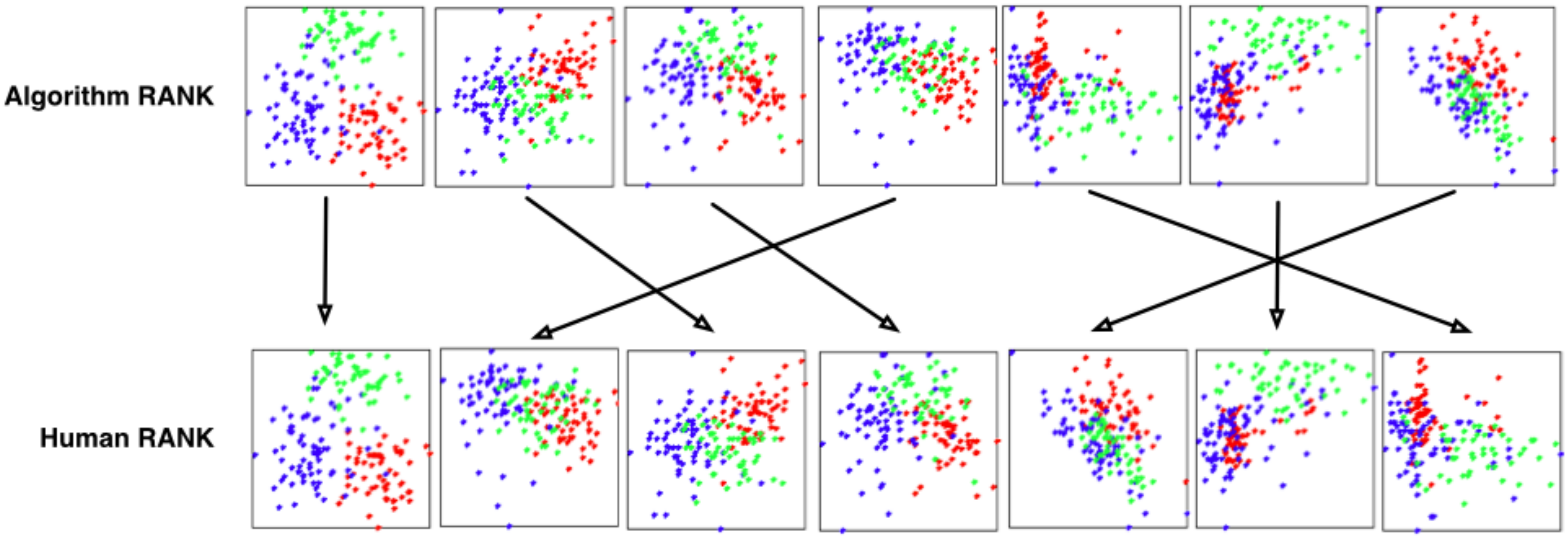


$$CDM = \sum_{k=1}^{m-1} \sum_{l=k+1}^{m} \sum_{i=1}^{P} ||\mathbf{p}_k^i - \mathbf{p}_l^i||$$

m = number of classes
p = local density

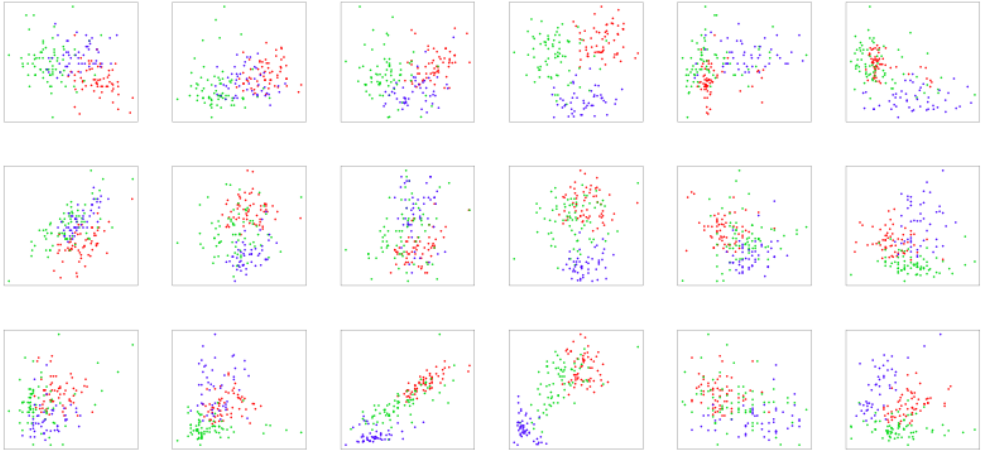
Work of Andrada Tatu, et al. "Combining automated analysis and visualization techniques for effective exploration of high-dimensional data." Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on. IEEE, 2009.

Do the metrics reflect human perception?



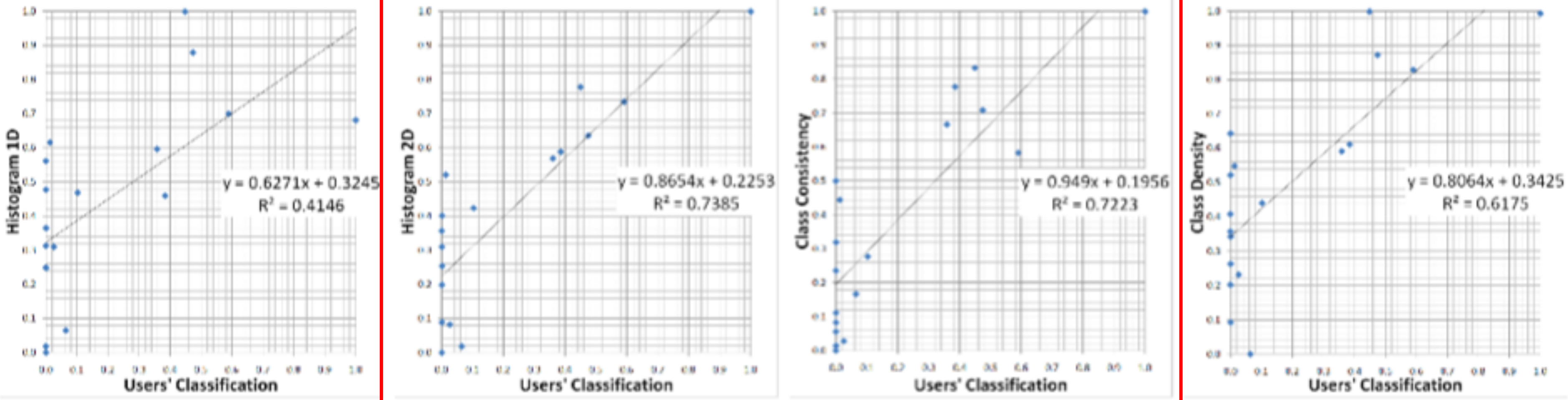
User Study

Data: Wine data set (from UCI), 178 samples, 13 attributes, 3 classes
Participants: 18 undergraduate students natural science
Visualization: scatter plots made of every possible pair of axes (18 selected)
Question: “*select projections suited to classify the three types of wine*”



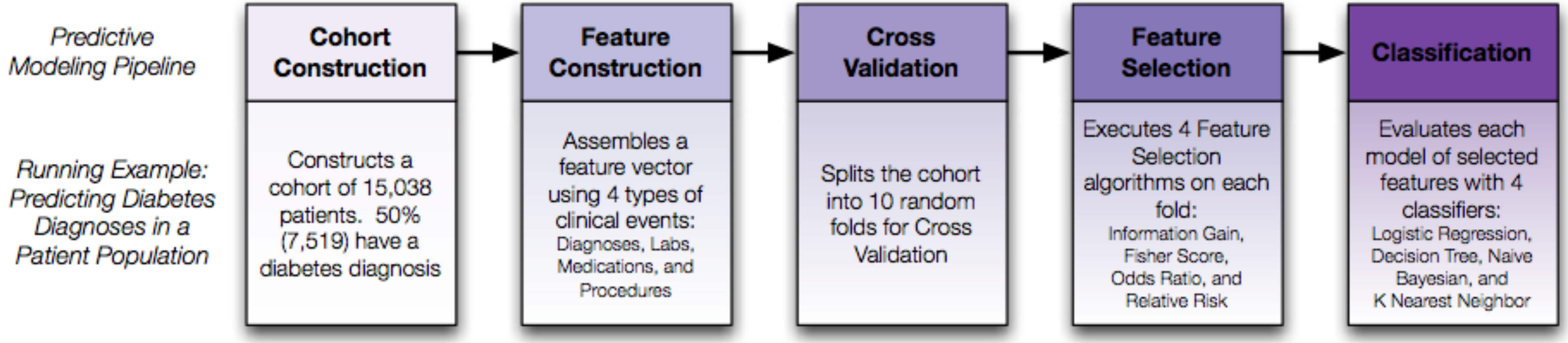
A. Tatu, P. Bak, E. Bertini, D. Keim, and J. Schneidewind.[Visual Quality Metrics and Human Perception: An Initial Study on 2D Projections of Large Multidimensional Data].In Proc. of ACM SIGCHI Working Conference on Advanced Visual Interfaces (AVI), 2010.

Results



(a) 1D-HDM (b) 2D-HDM (c) CCM (d) CDM

Making Sense of Feature Selection Algorithms



(Work in collaboration with Adam Perer @ IBM Watson)

Parallel computation of multiple models ...

Feature Selection
(Information Gain, Fisher Score, Odds Ratio, Relative Risk, …)

X

Classification
(Logistic Regression, Decision Trees, Naive Bayes, kNN, …)

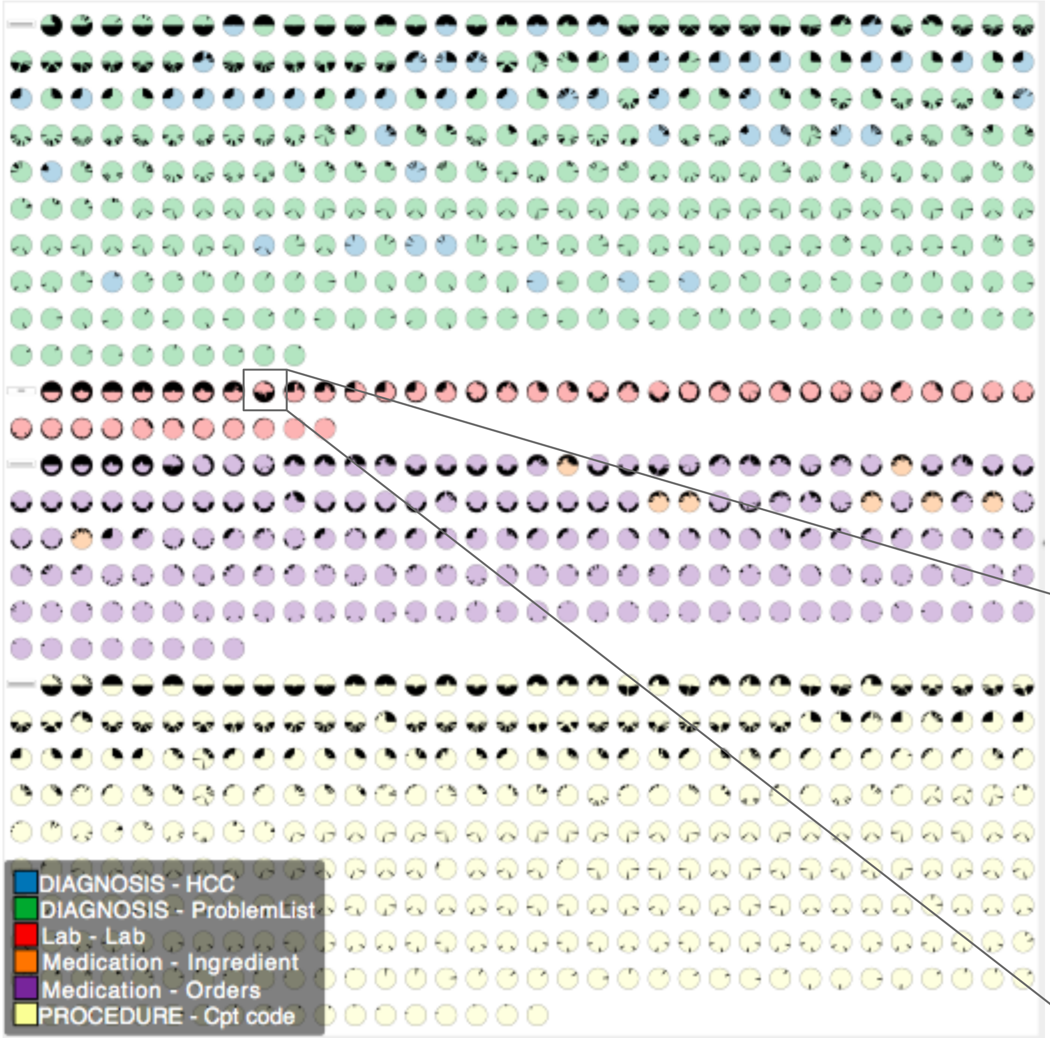
X

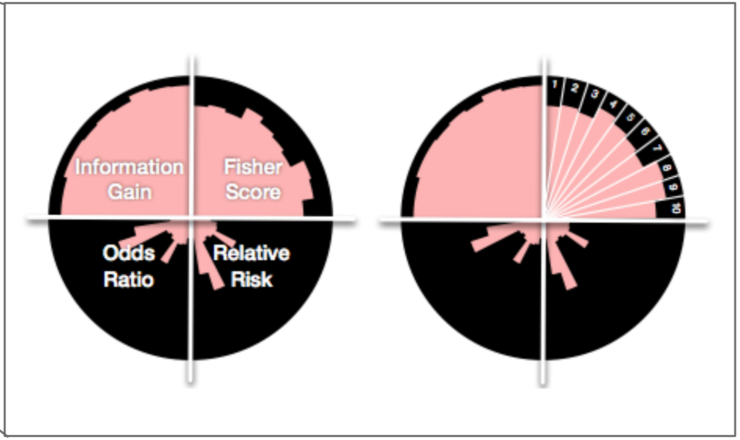
Folds (Samples)
10-folds validation

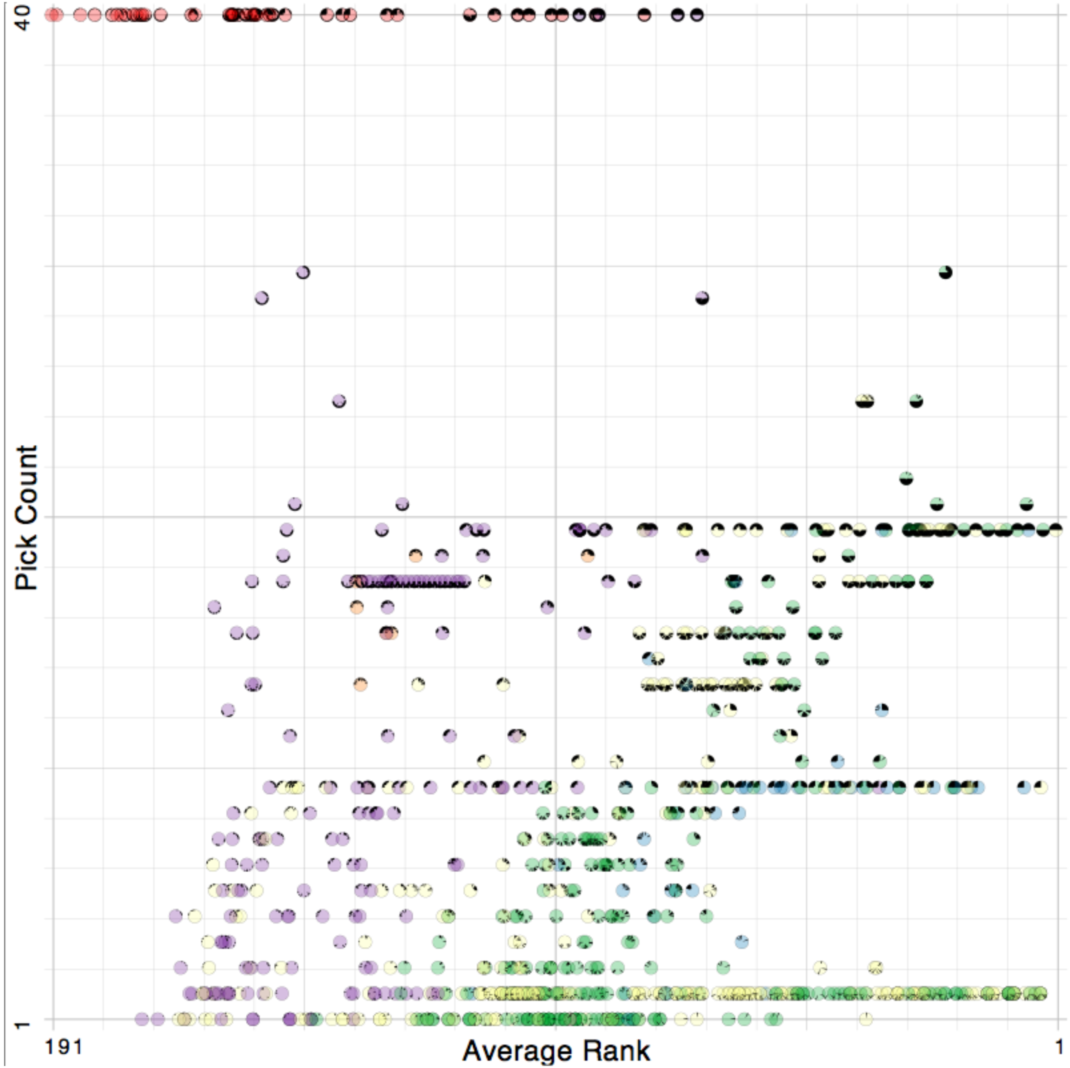
How to make sense of the results? What can we learn out of this? Can we better understand how features should be selected?

INFUSE

- Each dot is a feature (e.g., lab test)
- Each quadrant represents a feature selection algorithm
- Each segment represents a fold (sample)
- Length of the bar represents the ranking







40
Pick Count
1
191
Average Rank
1

Interesting open challenges for vis in data science and data science in vis.

1) From Data Visualization to Model Visualization

Many data mining methods for knowledge discovery exist.

- Results sometimes more complex (larger) than the data itself
- Parameters and techniques: Which one to use? How do they compare?
- A good grasp of the data is needed to feed mining algorithms with the "right" data

How do we make sense of them? And how do we build trust in them?

A Few Ideas for Model Visualization ...

Visual Diagnostics

Visual Ensembles

Computational Steering

Explanatory Visualization
(work on causality?)

We need your help!

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the many differences that arise in the process of modeling for an explanatory versus a predictive goal. The purpose of this article is to clarify the distinction between explanatory and predictive modeling, to discuss its sources, and to reveal the practical implications of the distinction to each step in the modeling process.

Key words and phrases: Explanatory modeling, causality, predictive modeling, predictive power, statistical strategy, data mining, scientific research.

1. INTRODUCTION

Looking at how statistical models are used in different scientific disciplines for the purpose of theory building and testing, one finds a range of perceptions regarding the relationship between causal explanation and empirical prediction. In many scientific fields such as economics, psychology, education, and environmental science, statistical models are used almost exclusively for causal explanation, and models that possess high explanatory power are often assumed to inherently possess predictive power. In fields such as natural language processing and bioinformatics, the focus is on empirical prediction with only a slight and indirect relation to causal explanation. And yet in other research fields, such as epidemiology, the emphasis on causal explanation versus empirical prediction is more mixed. Statistical modeling for description, where the purpose is to capture the data structure parsimoniously, and which is the most commonly developed within the field of statistics, is not commonly used for theory building and testing in other disciplines. Hence, in this article I focus on the use of statistical modeling for causal explanation and for prediction. My main premise is that the two are often conflated, yet the causal versus predictive distinction has a large impact on each step of the statistical modeling process and on its consequences. Although not explicitly stated in the statistics methodology literature, applied statisticians instinctively sense that predicting and explaining are different. This article aims to fill a critical void: to tackle the distinction between explanatory modeling and predictive modeling.

Clearing the current ambiguity between the two is critical not only for proper statistical modeling, but more importantly, for proper scientific usage. Both explanation and prediction are necessary for generating and testing theories, yet each plays a different role in doing so. The lack of a clear distinction

Galit Shmueli is Associate Professor of Statistics, Department of Decision, Operations and Information Technologies, Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742, USA (e-mail: gshmueli@umd.edu).

2) “Intelligent” Visual Data Exploration

We need
your help!

There are simply too many ways to encode and ways to interact with data.

“Where are the interesting views?”

Computational approaches can (and should) help users navigate through this huge design/decision space.

3) Visual Analytics Evaluation

We need your help!

Hard! Hard! Hard!

Problem 1: Visualization happens in our brain not in the computer.

Problem 2: Sensemaking, discovery, learning … hard to measure and often happen on a long time span.

Interesting Approach: Synthetic Benchmark Datasets

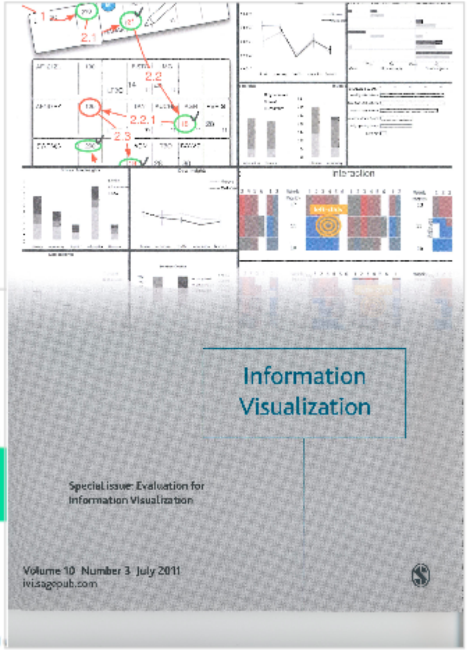
Synthetic data that look very real.

Ground truth known and manually hidden in the data.

VAST Challenge has new data sets each year.

http://vacommunity.org/VAST+Challenge+2014

Special Issue on Evaluation in InfoVis Journal (Jul 2011)



Empirical Studies in Information Visualization: Seven Scenarios

Heidi Lam   Enrico Bertini   Petra Isenberg   Catherine Plaisant   Sheelagh Carpendale

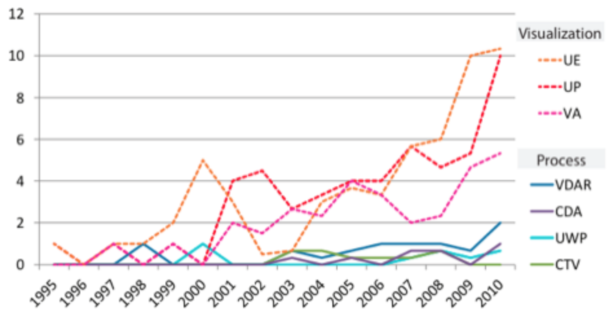
Abstract—We take a new, scenario based look at evaluation in information visualization. Our seven scenarios, evaluating visual data analysis and reasoning, evaluating user performance, evaluating user experience, evaluating environments and work practices, evaluating communication through visualization, evaluating visualization algorithms, and evaluating collaborative data analysis were derived through an extensive literature review of over 800 visualization publications. These scenarios distinguish different study goals and types of research questions and are illustrated through example studies. Through this broad survey and the distillation of these scenarios we make two contributions. One, we encapsulate the current practices in the information visualization research community and, two, we provide a different approach to reaching decisions about what might be the most effective evaluation of a given information visualization. Scenarios can be used to choose appropriate research questions and goals and the provided examples can be consulted for guidance on how to design one's own study.

Index Terms—Information visualization, evaluation

Venue	Year	Papers	Papers with Eval
EuroVis*	2002–2011	151	66
InfoVis	1995–2010	381	178
IVS	2002–2010	183	86
VAST	2006–2010	123	43
Total		850	361

Process
UWP: Understanding Environments and Work Practices
VDAR: Evaluating Visual Data Analysis and Reasoning
CTV: Evaluating Communication Through Visualization
CDA: Evaluating Collaborative Data Analysis

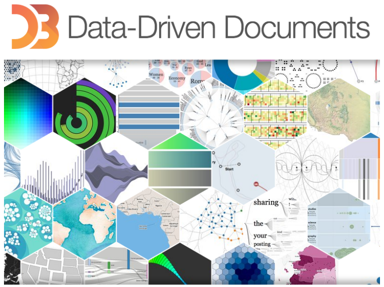
Visualization
UP: Evaluating User Performance
UE: Evaluating User Experience
VA: Evaluating Visualization Algorithms



Interesting research I am NOT doing …

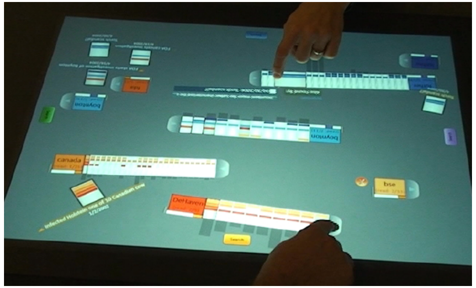
Visualization Programming Frameworks (D3.js).

(See Jeff Heer's IDL Lab @ UW)



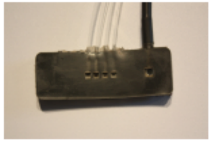
Collaborative Visualization and Large Displays.

(See Petra Isenberg and others at INRIA)



Brain Sensing for Evaluation.

(See Claudio Silva's work on EEG and Remco Chang @ Tufts)

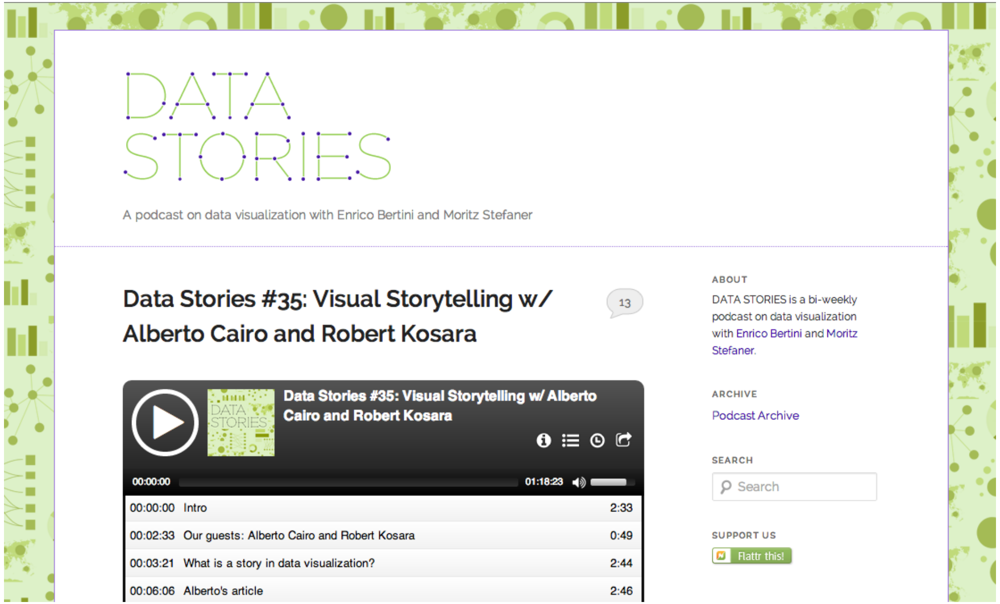


Crowdsourced Data Analysis.

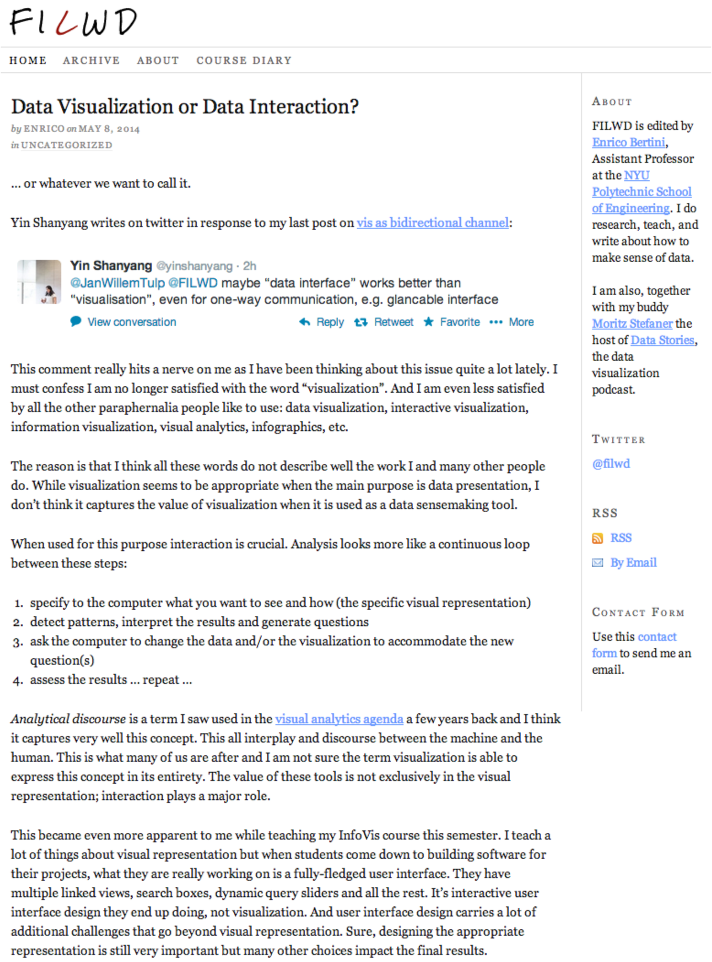
(See Maneesh Agrawala @ Berkeley)

http://datastori.es

http://fellinlovewithdata.com/



w/ Moritz Stefaner



FILWD

HOME ARCHIVE ABOUT COURSE DIARY

Data Visualization or Data Interaction?

by ENRICO on MAY 8, 2014
in UNCATEGORIZED

... or whatever we want to call it.

Yin Shanyang writes on twitter in response to my last post on vis as bidirectional channel:

Yin Shanyang @yinshanyang · 2h
@JanWillemTulp @FILWD maybe "data interface" works better than "visualisation", even for one-way communication, e.g. glancable interface
View conversation Reply Retweet Favorite More

This comment really hits a nerve on me as I have been thinking about this issue quite a lot lately. I must confess I am no longer satisfied with the word "visualization". And I am even less satisfied by all the other paraphernalia people like to use: data visualization, interactive visualization, information visualization, visual analytics, infographics, etc.

The reason is that I think all these words do not describe well the work I and many other people do. While visualization seems to be appropriate when the main purpose is data presentation, I don't think it captures the value of visualization when it is used as a data sensemaking tool.

When used for this purpose interaction is crucial. Analysis looks more like a continuous loop between these steps:

1. specify to the computer what you want to see and how (the specific visual representation)
2. detect patterns, interpret the results and generate questions
3. ask the computer to change the data and/or the visualization to accommodate the new question(s)
4. assess the results ... repeat ...

Analytical discourse is a term I saw used in the visual analytics agenda a few years back and I think it captures very well this concept. This all interplay and discourse between the machine and the human. This is what many of us are after and I am not sure the term visualization is able to express this concept in its entirety. The value of these tools is not exclusively in the visual representation; interaction plays a major role.

This became even more apparent to me while teaching my InfoVis course this semester. I teach a lot of things about visual representation but when students come down to building software for their projects, what they are really working on is a fully-fledged user interface. They have multiple linked views, search boxes, dynamic query sliders and all the rest. It's interactive user interface design they end up doing, not visualization. And user interface design carries a lot of additional challenges that go beyond visual representation. Sure, designing the appropriate representation is still very important but many other choices impact the final results.

ABOUT

FILWD is edited by Enrico Bertini, Assistant Professor at the NYU Polytechnic School of Engineering. I do research, teach, and write about how to make sense of data.

I am also, together with my buddy Moritz Stefaner the host of Data Stories, the data visualization podcast.

TWITTER

@filwd

RSS

RSS

By Email

CONTACT FORM

Use this contact form to send me an email.

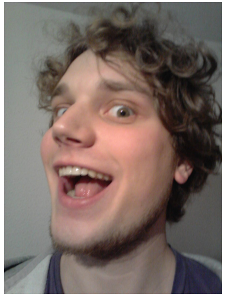
Acknowledgements

Andrada Tatu

Hendrik Strobelt

Ilya Boyandin

Josua Kruse



Thanks!

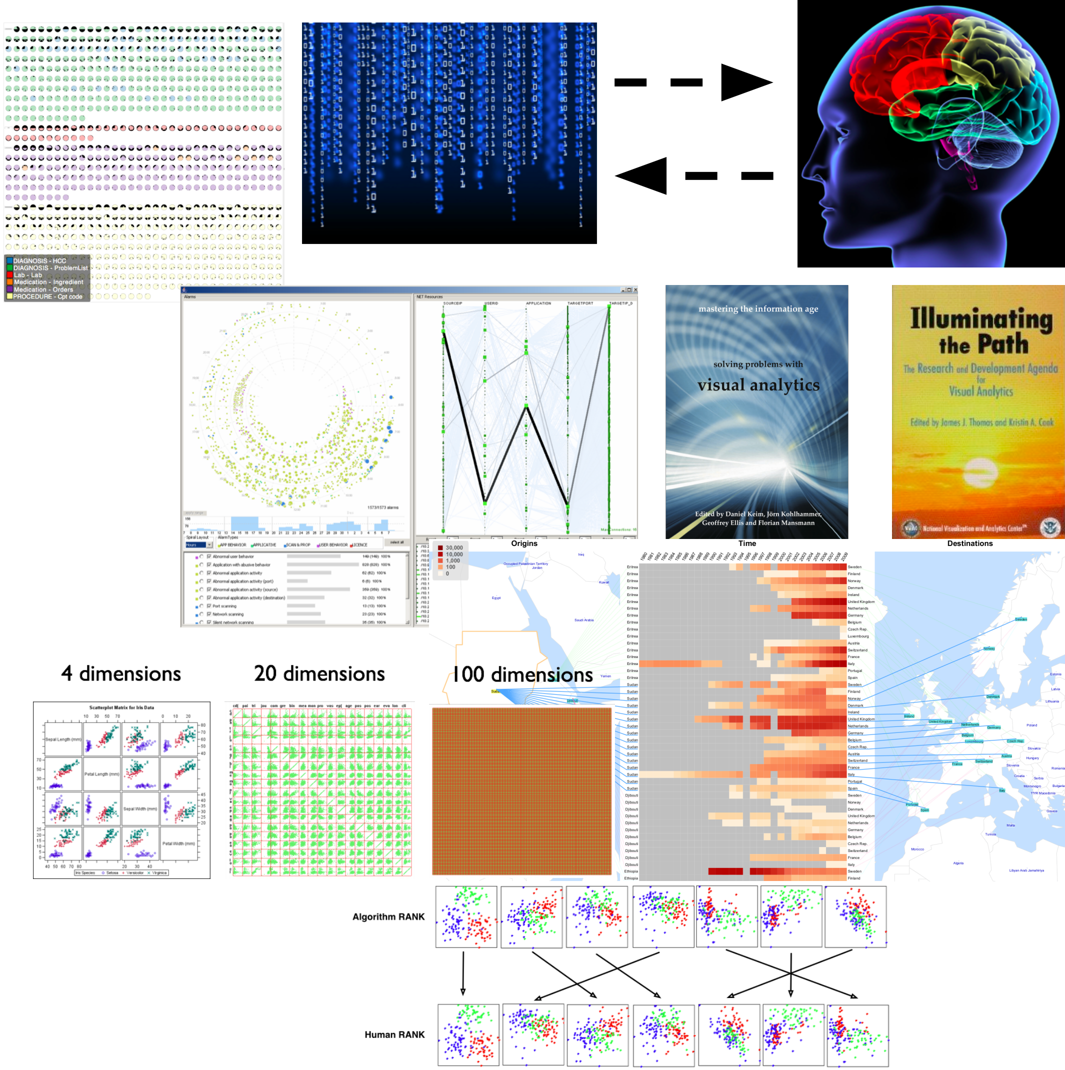
Enrico Bertini

enrico.bertini@nyu.edu

enrico.bertini.me

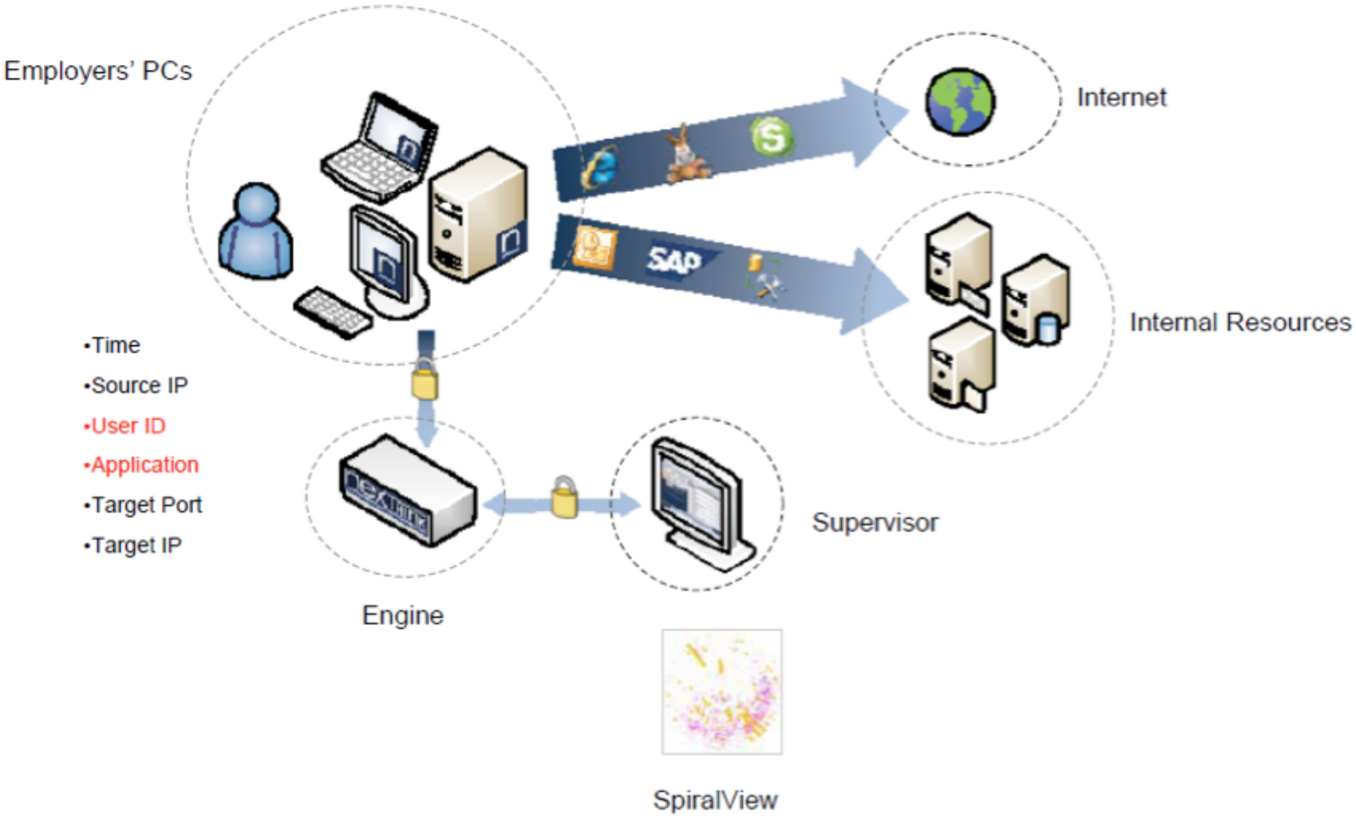
@FILWD

NYU POLYTECHNIC SCHOOL OF ENGINEERING

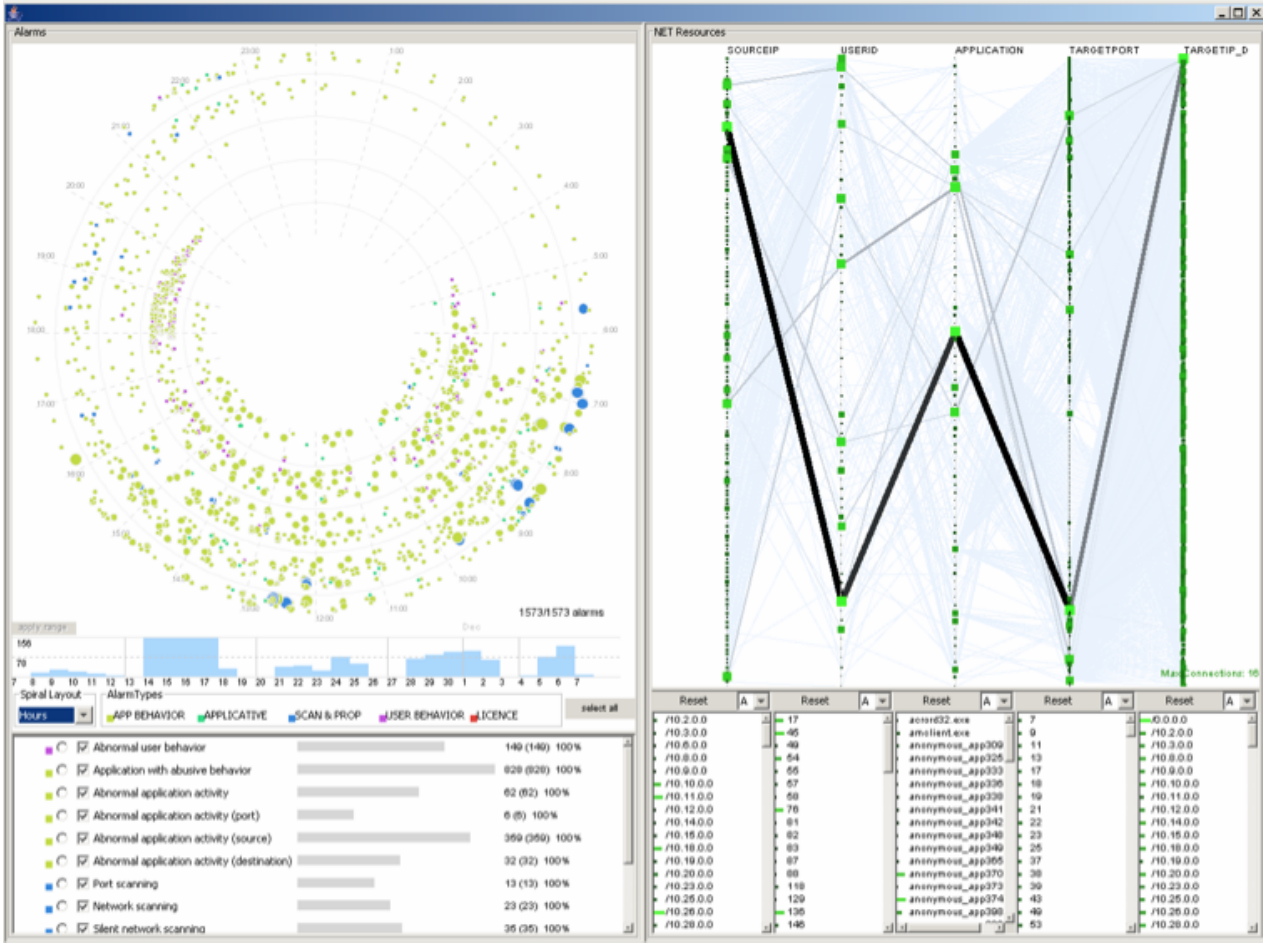


Backup slides ...

Visual Diagnostics for Intrusion Detection

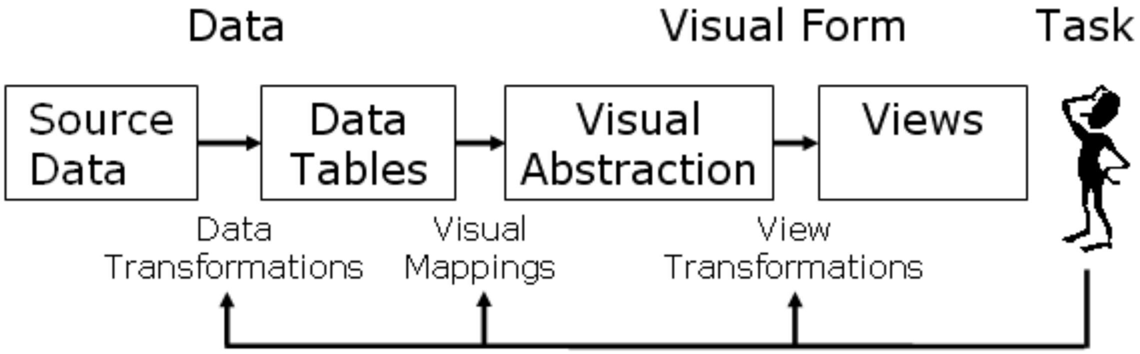


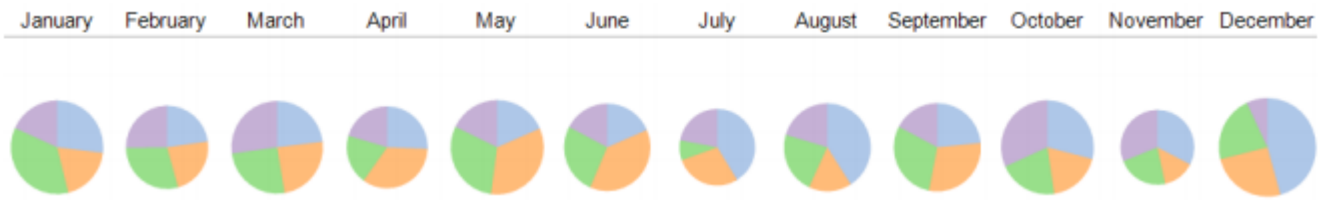
SpiralView

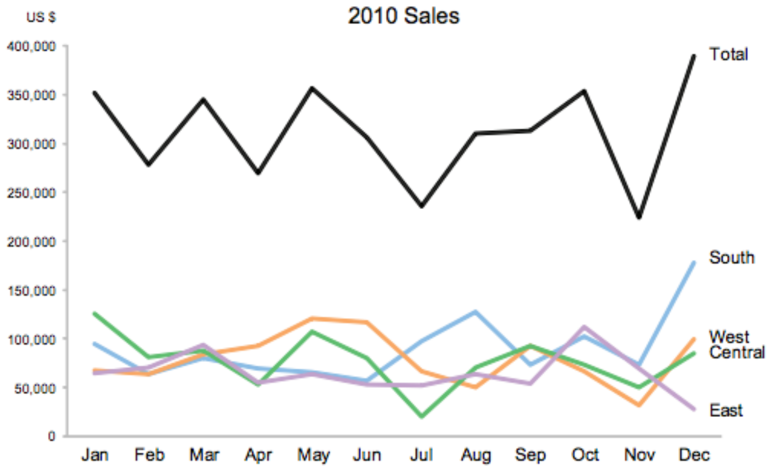


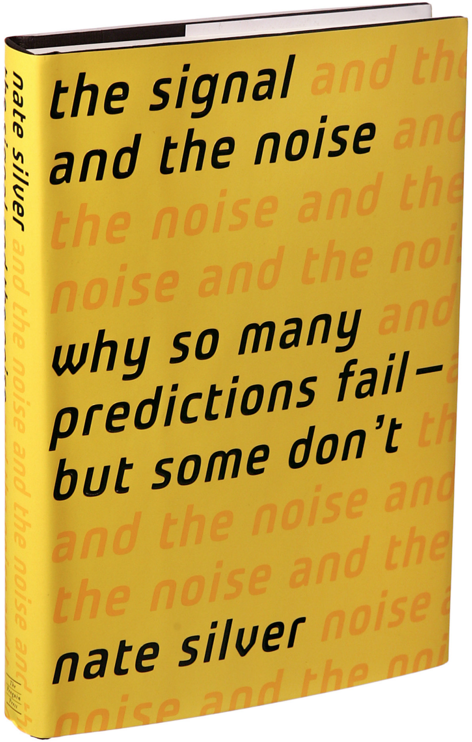
Enrico Bertini, Patrick Hertzog, and Denis Lalanne.**SpiralView: Towards Security Policies Assessment through Visual Correlation of Network Resources with Evolution of Alarms.**In Proc. of IEEE Symposium on Visual Analytics Science and Technology (VAST), 2007.

Information Visualization Pipeline




January
February
March
April
May
June
July
August
September
October
November
December


2010 Sales
US $
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Total
South
West
Central
East



"The numbers have no way of speaking for themselves. We speak for them.

Data-driven predictions can succeed—and they can fail. **It is when we deny our role in the process that the odds of failure rise.**

Before we demand more of our data, we need to demand more of ourselves."