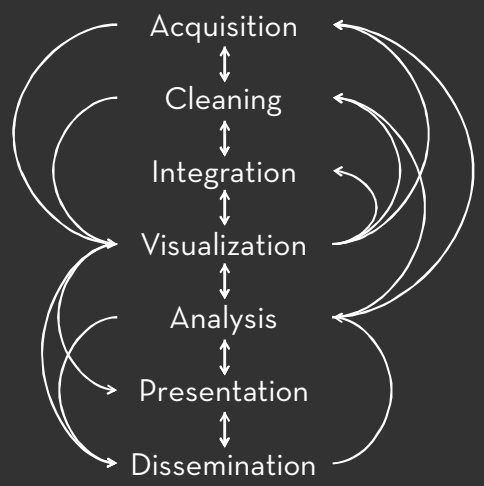
CS448G :: 2 May 2011

Visual Analysis

Jeffrey Heer Stanford University



Guiding Principles

Analysis is often **iterative** and **exploratory**
- Hypothesis formation vs. confirmation
- Ad hoc queries – difficult to know workload

Design interfaces to enable **tighter iteration**
- Visualization: rapid perception of patterns
- Interaction: specification of data & views

Support analysis *"at the speed of thought"*

Effects of Latency

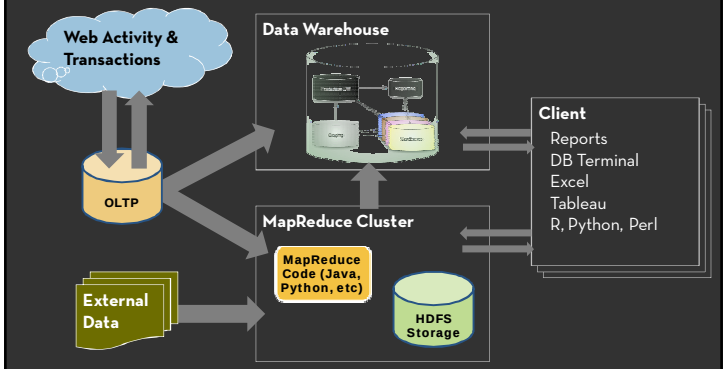
Milliseconds Matter
< 100ms: perception of animation, causation
Wait times > 1s may interrupt flow of thought
To reduce (perceived) latency, drop details

Example: Latency in Google search results
Latency (+300ms) reduces searches ~0.5%
Reduction persists after performance resumes

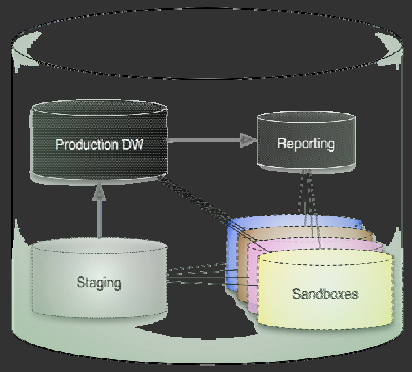
Goal: support analysis at **interactive rates** to enable fluid **conversations with data**.

Review of Analysis Infrastructure

Analysis Infrastructure

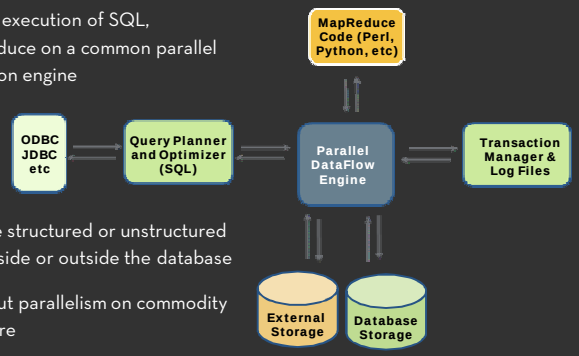


Data Modeling for Analysis



Greenplum: SQL & MapReduce

Unified execution of SQL, MapReduce on a common parallel execution engine



Analyze structured or unstructured data, inside or outside the database

Scale out parallelism on commodity hardware

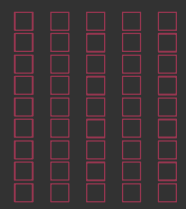
Greenplum

Relational Data Organizations

Transactions vs. **Analysis**

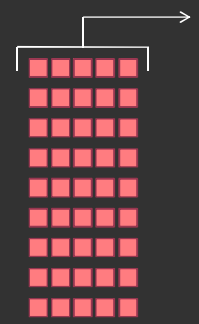
Row-oriented

Column-oriented

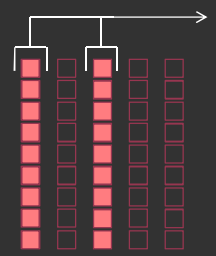


Relational Data Organizations

Row-oriented



Column-oriented



Relational Data Organizations

Speed-up Analysis

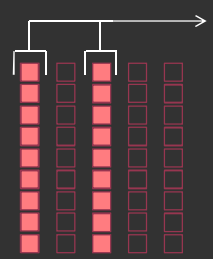
Reduce data transfer

Improved locality

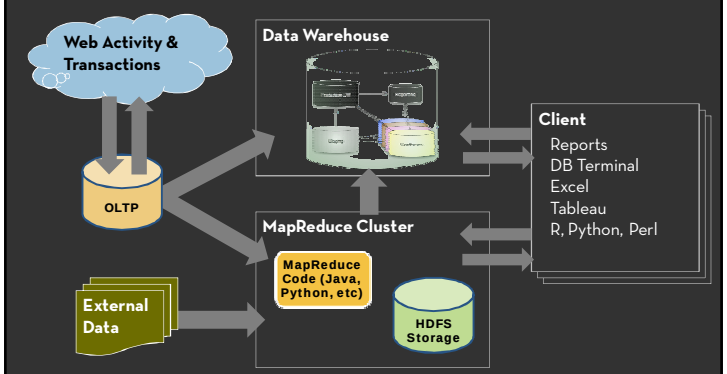
Better data compression

Query processing on compressed data

Column-oriented



Analysis Infrastructure



Other Thoughts

Much current attention is focused on scalable infrastructure, but don't forget the client

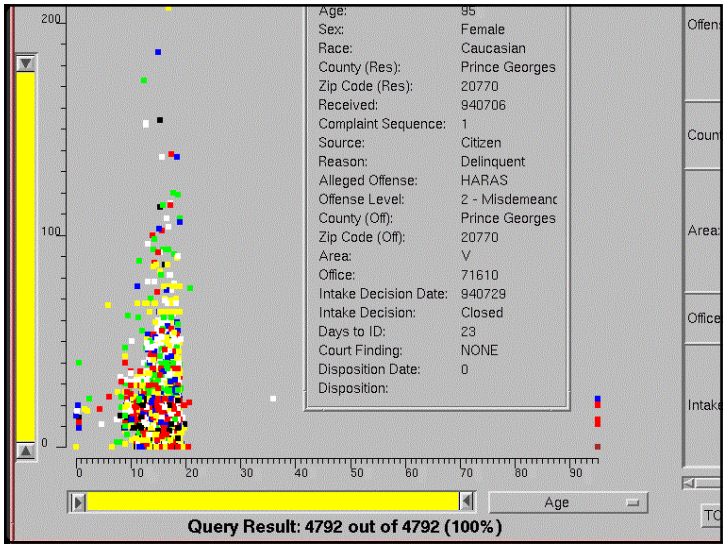
- Multi-core and GPU computing for analysis
- Monitoring, review on mobile devices?

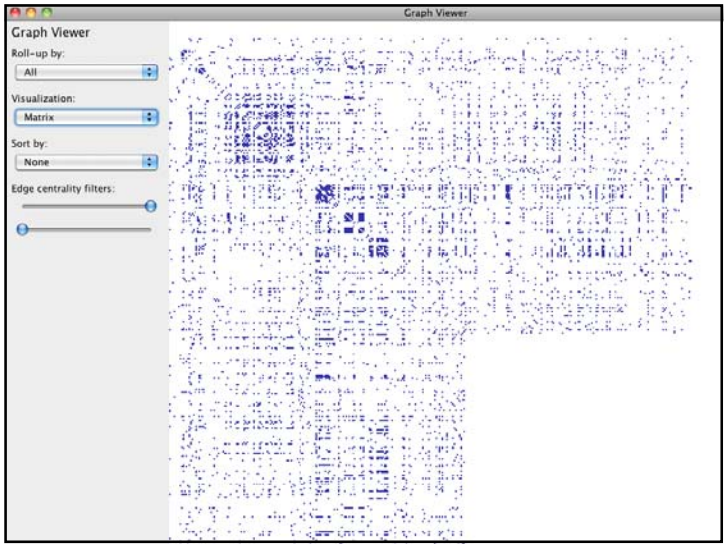
Increasing variety of cloud computing services, solid state storage – networking as limiter?

Visual Data Analysis

Visual Data Analysis

1 Data diagnostics





Visual Data Analysis

1 Data diagnostics

2 Exploratory data analysis

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

Summary Statistics

$u_X = 9.0$ $\sigma_X = 3.317$

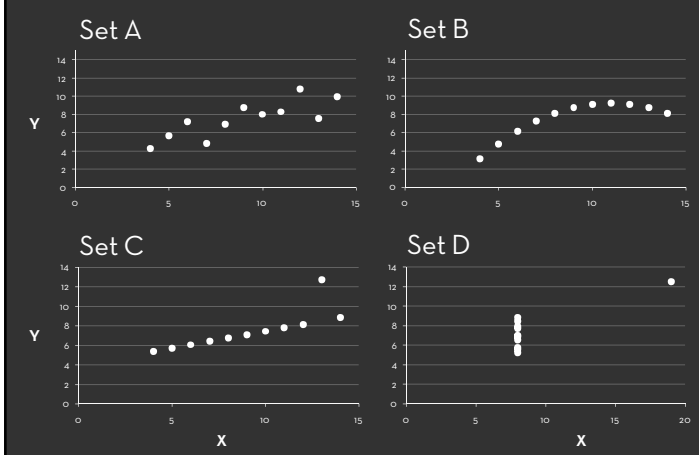
$u_Y = 7.5$ $\sigma_Y = 2.03$

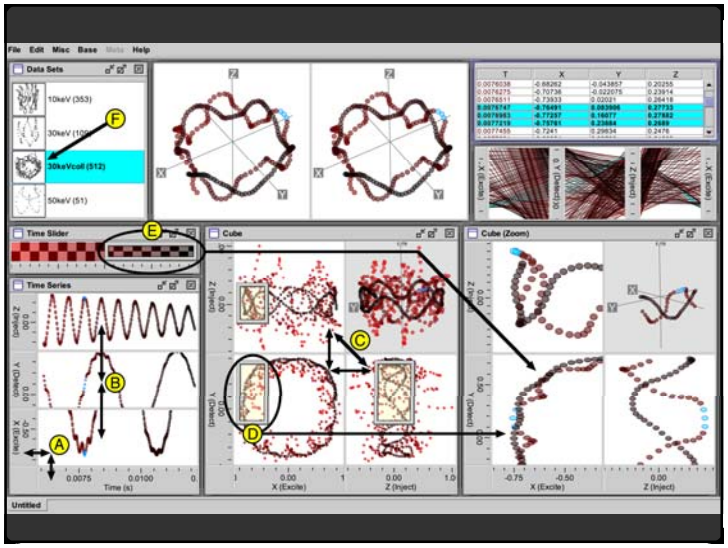
Linear Regression

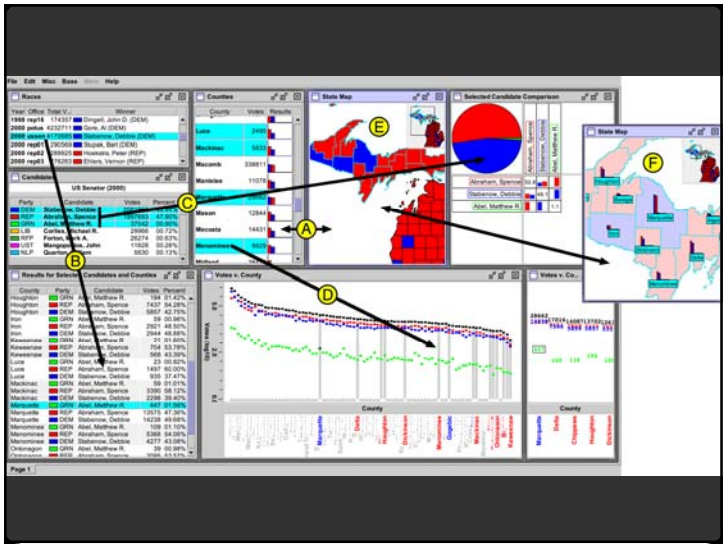
$Y^2 = 3 + 0.5\ X$

$R^2 = 0.67$

Anscombe 1973







Visual Data Analysis

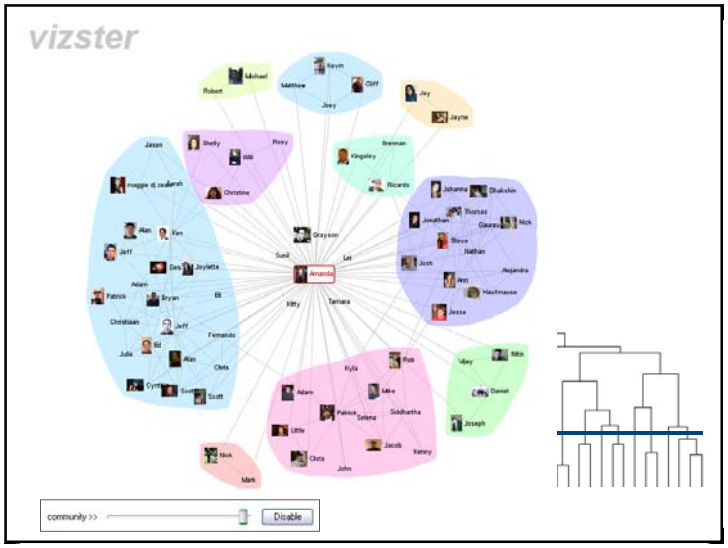
1 Data diagnostics

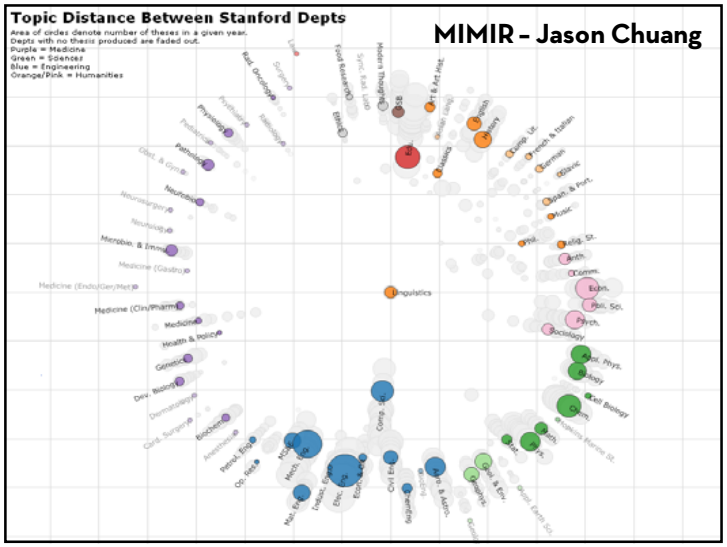
2 Exploratory data analysis

3 Assessing analytic results

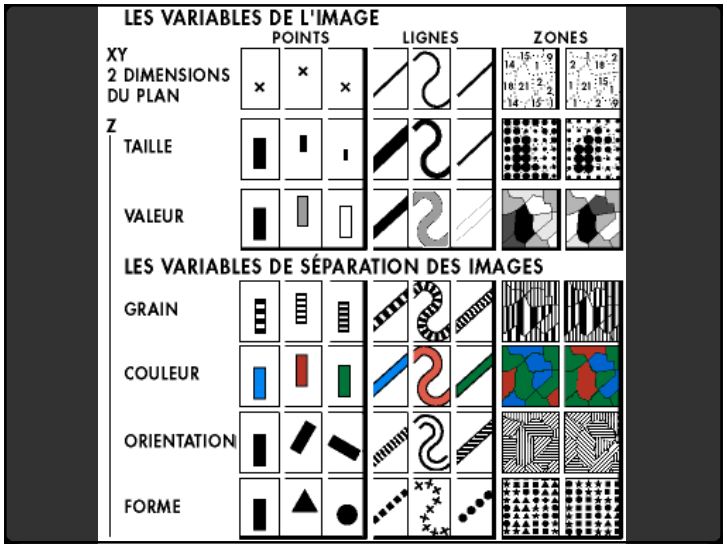
While it is often most helpful to "plot the data," this is rarely enough. We need also to "plot the results of analysis" as a routine matter. There is often more analysis than there was data.

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



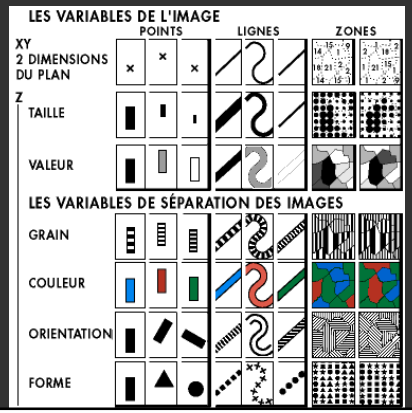


Graphical Perception
& Automated Design



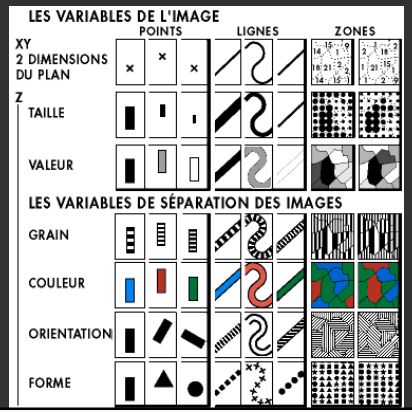
Visual encoding variables

Position (x 2)
Size
Value
Texture
Color
Orientation
Shape



Visual encoding variables

Position
Length
Area
Volume
Value
Texture
Color
Orientation
Shape
Transparency
Blur / Focus ...



Combinatorics of Encodings

Challenge:

Pick the best encoding from the exponential number of possibilities $(n+1)^8$

Principle of Consistency:

The properties of the image (visual variables) should match the properties of the data.

Principle of Importance Ordering:

Encode the most important information in the most effective way.

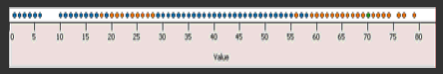
Design Criteria (Mackinlay)

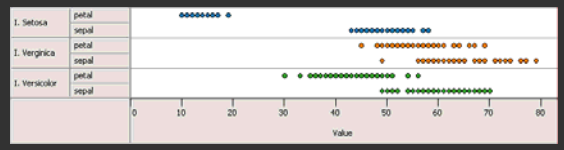
Expressiveness

A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express *all* the facts in the set of data, and *only* the facts in the data.

Cannot express the facts

A one-to-many (1 → N) relation cannot be expressed in a single horizontal dot plot because multiple tuples are mapped to the same position





Expresses facts not in the data

A length is interpreted as a quantitative value;

∴ Length of bar says something untrue about N data

Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

[Mackinlay, APT, 1986]

Design Criteria (Mackinlay)

Expressiveness

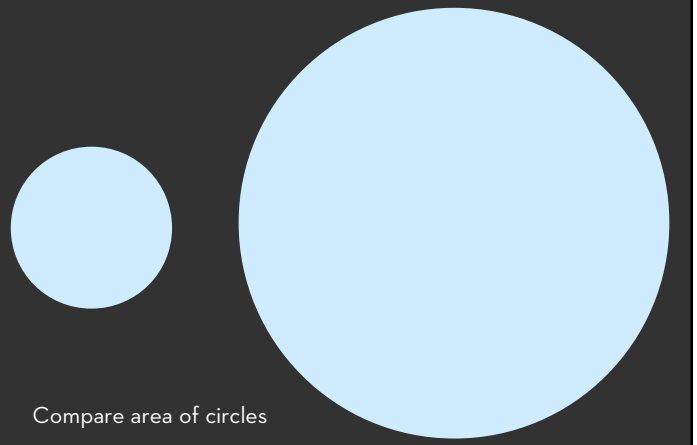
A set of facts is expressible in a visual language if the sentences (i.e. the visualizations) in the language express *all* the facts in the set of data, and *only* the facts in the data.

Effectiveness

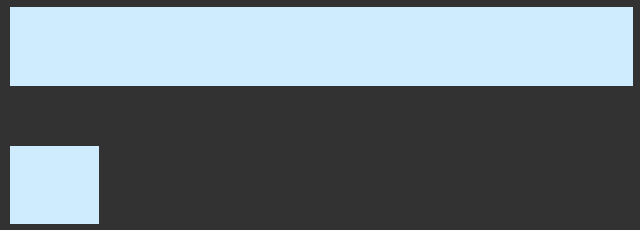
A visualization is more effective than another visualization if the information conveyed by one visualization is more readily *perceived* than the information in the other visualization.

Which best encodes quantities?

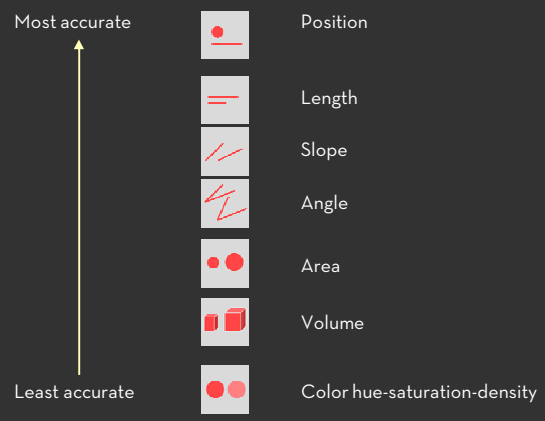
Position
Length
Area
Volume
Value (Brightness)
Color Hue
Orientation (Angle)
Shape



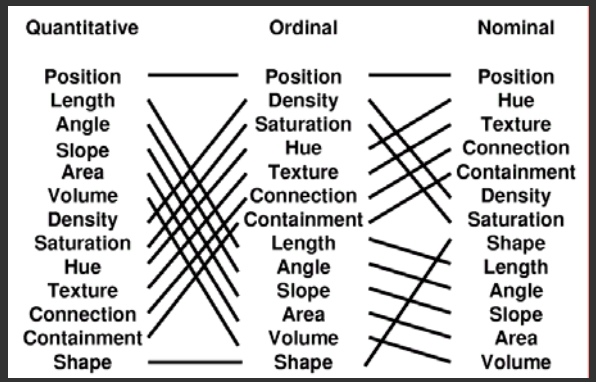
Compare area of circles



Compare length of bars



Rankings Used by Mackinlay



Conjectured *effectiveness* of the encoding

Mackinlay's Design Algorithm

User formally specifies data model and type
- Additional input: ordered list of data variables to show

APT searches over design space
- Tests expressiveness of each visual encoding
- Generates image for encodings that pass test
- Tests perceptual effectiveness of resulting image

Outputs the "most effective" visualization

Final Project Presentations

Project Presentations

Present your project plans to the class.

Keep it short! (≤ 3 minutes)

Very briefly introduce the problem domain.

Spend most of the time presenting your initial design ideas – flowcharts, sketches, prototypes

Your final slide should have 3 questions for the class – no need to read them out loud.

Send slides to *cs448g* by **8am Wed 5/4**.

Discussants

Amir Ghazvinian
Eli Marschner
Letitia Lew