CS448B :: 11 Oct 2011

Multi-Dimensional Vis

Jeffrey Heer Stanford University

Last Time: Exploratory Data Analysis

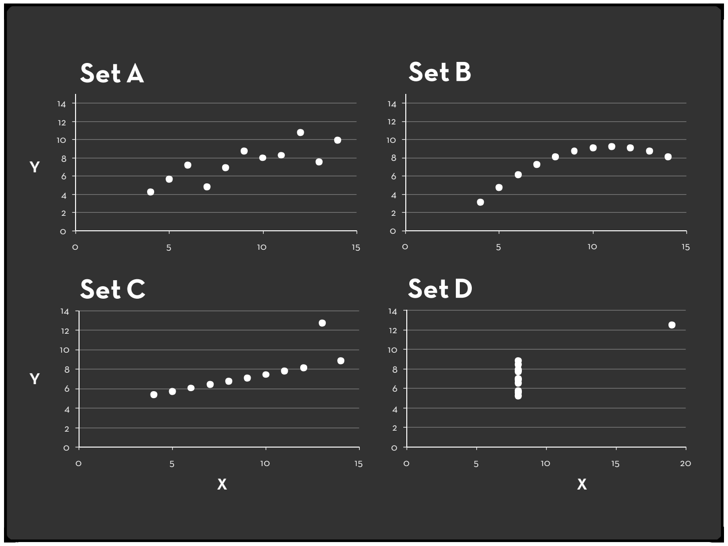
Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.

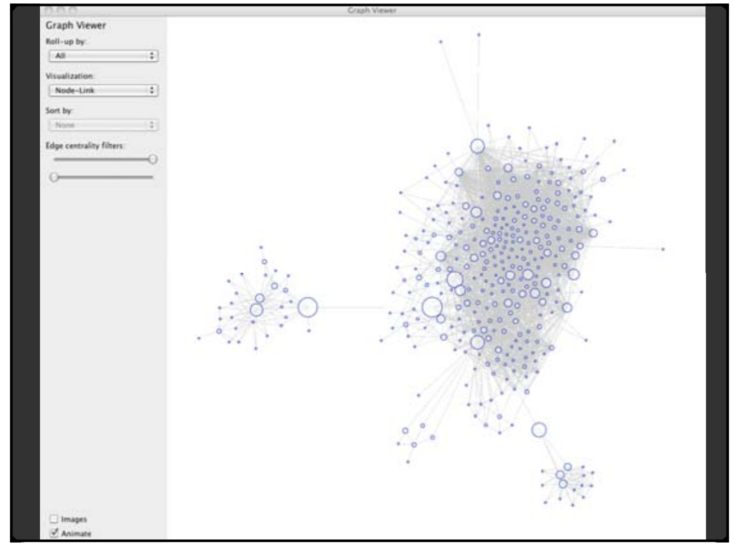
Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15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

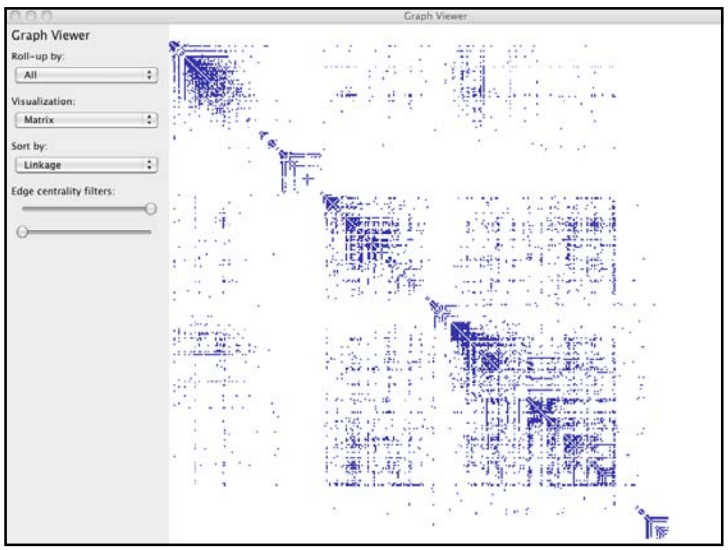
Summary Statistics
$u_X = 9.0$ $\sigma_X = 3.317$
$u_Y = 7.5$ $\sigma_Y = 2.03$

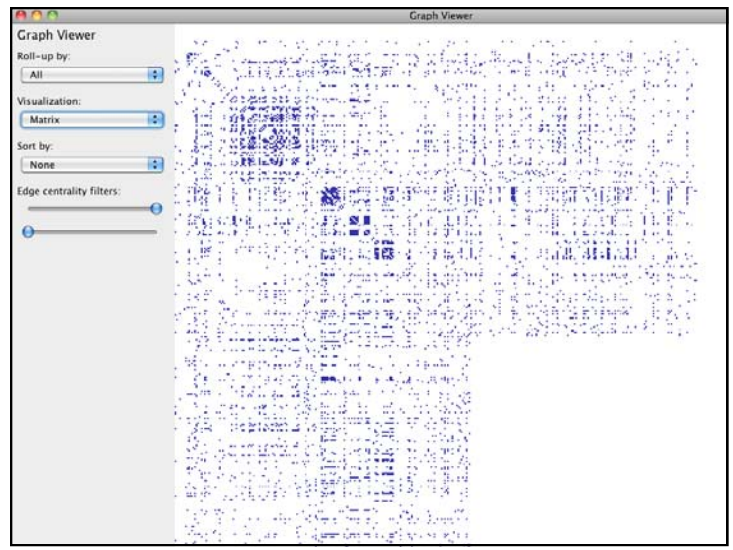
Linear Regression
$Y = 3 + 0.5\ X$
$R^2 = 0.67$

Anscombe 1973


Set A
Set B
Set C
Set D
Y
Y
X
X


Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
Edge centrality filters:


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

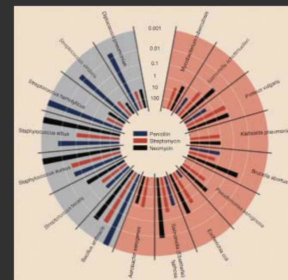

Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
None
Edge centrality filters:

Antibiotic Effectiveness

Table 1: Burtin's data.

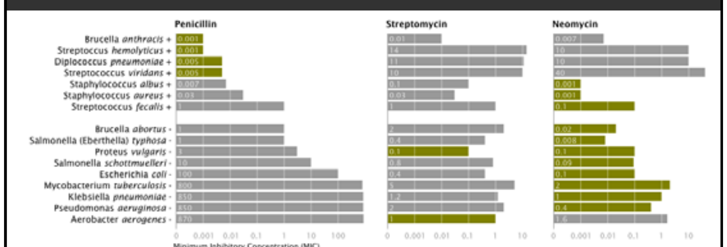
Bacteria	Penicillin	Streptomycin	Neomycin	Gram Staining
	Antibiotic			
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

Will Burtin, 1951

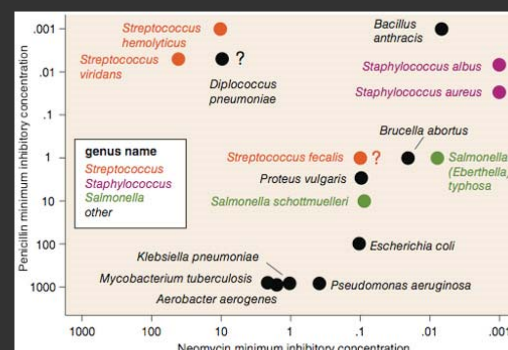


Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

How do the drugs compare?



Mike Bostock, CS448B Winter 2009



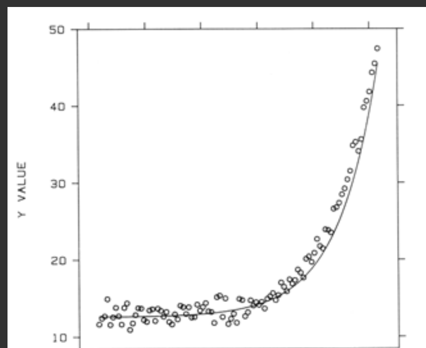
How do the bacteria group w.r.t. resistance?

Do different drugs correlate?

Wainer & Lysen
American Scientist, 2009

Transforming data

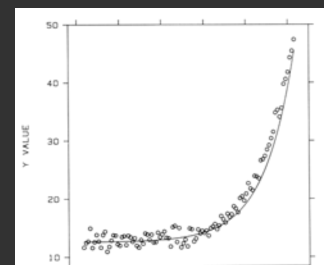
How well does the curve fit data?

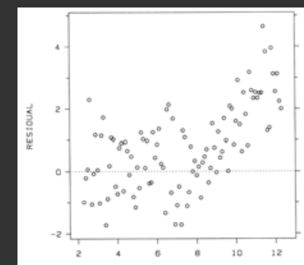


[Cleveland 85]

Plot the Residuals

Plot vertical distance from best fit curve
Residual graph shows accuracy of fit

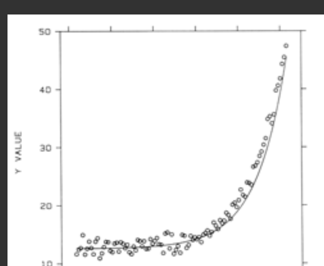




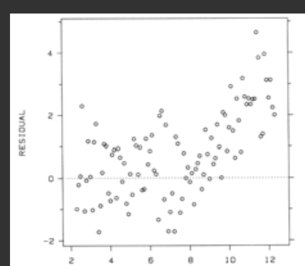
[Cleveland 85]

Multiple Plotting Options

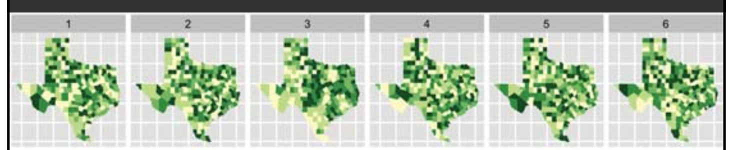
Plot model in data space



Plot data in model space



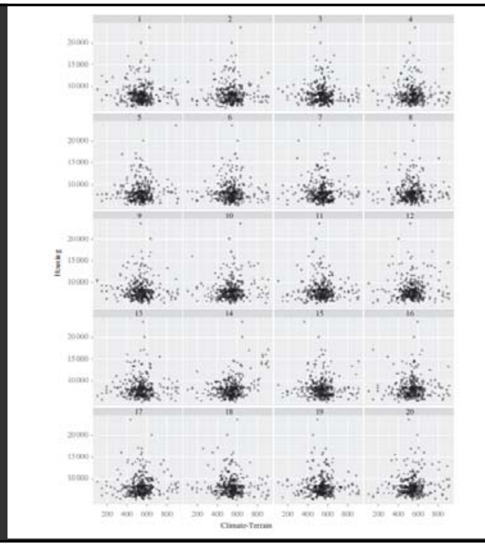
[Cleveland 85]



Choropleth maps of cancer deaths in Texas.

One plot shows a real data sets. The others are simulated under the null hypothesis of spatial independence.

Can you spot the real data? If so, you have some evidence of spatial dependence in the data.

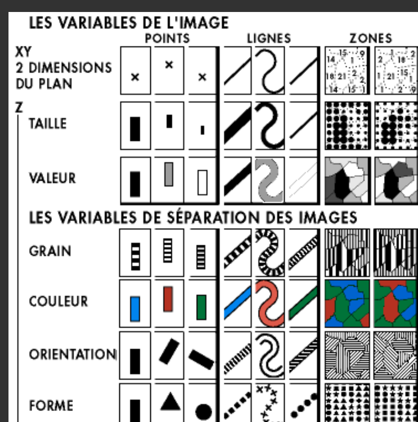


Multidimensional Visualization

Visual Encoding Variables

Position
Length
Area
Volume
Value
Texture
Color
Orientation
Shape

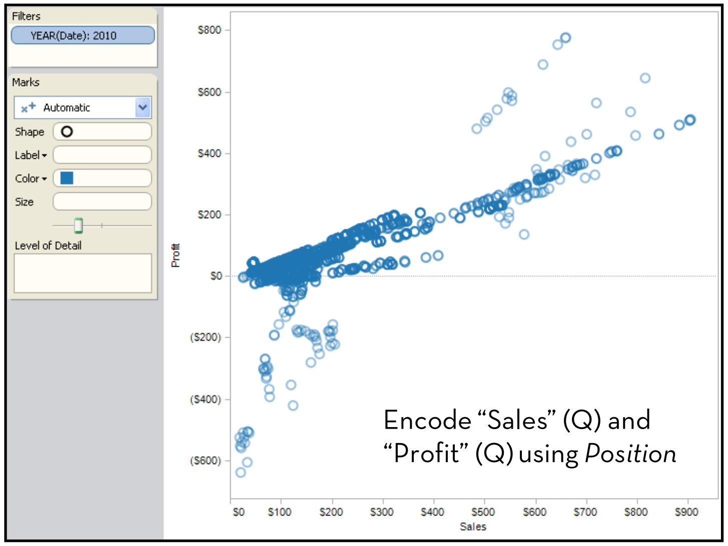
~8 dimensions?

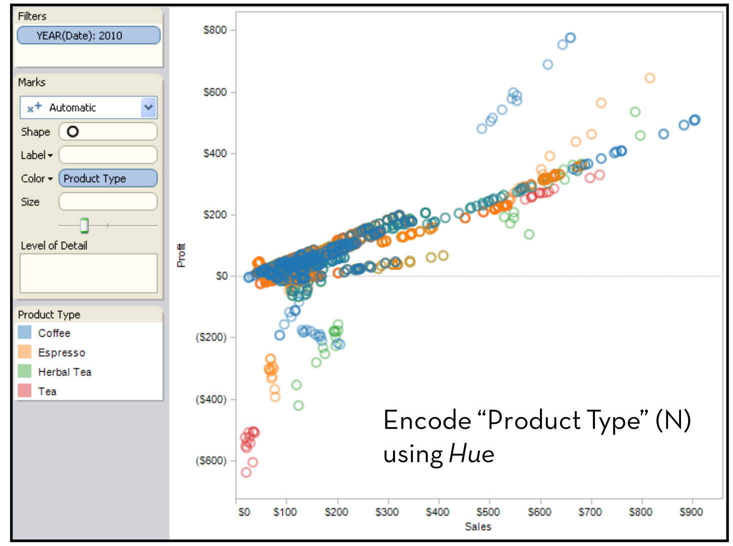


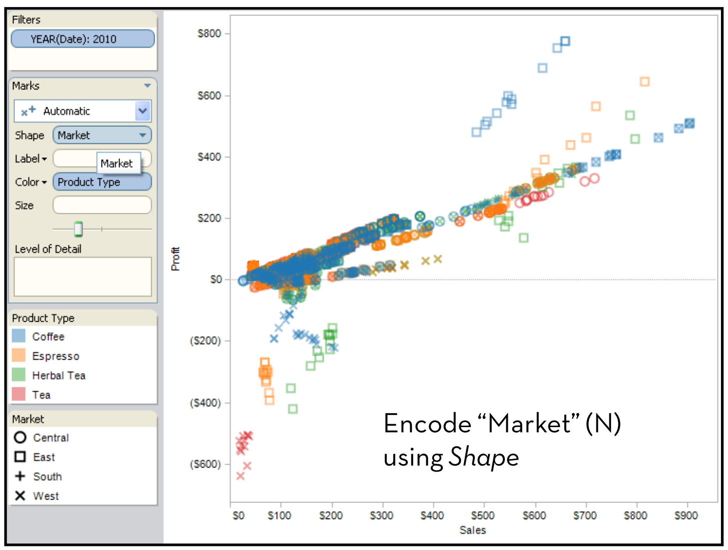
Example: Coffee Sales

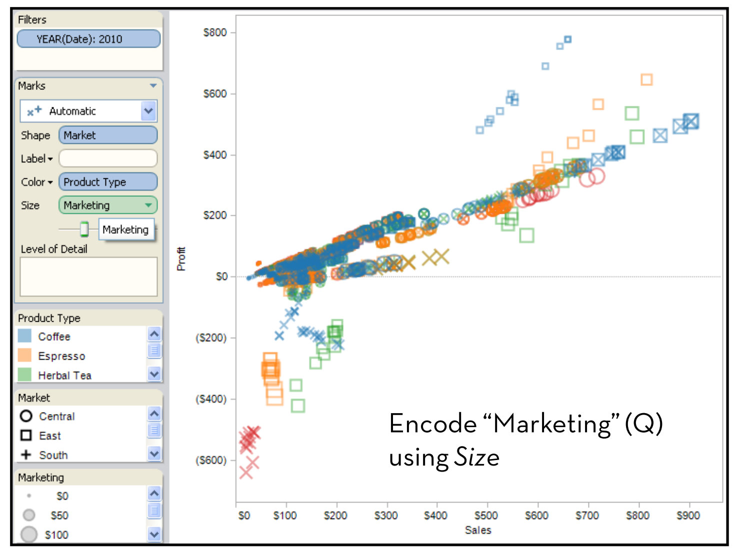
Sales figures for a fictional coffee chain:

Sales	Q-Ratio	
Profit	Q-Ratio	
Marketing	Q-Ratio	
Product Type	N	{Coffee, Espresso, Herbal Tea, Tea}
Market	N	{Central, East, South, West}


Encode “Sales” (Q) and
“Profit” (Q) using *Position*


Encode “Product Type” (N)
using *Hue*


Encode “Market” (N)
using *Shape*

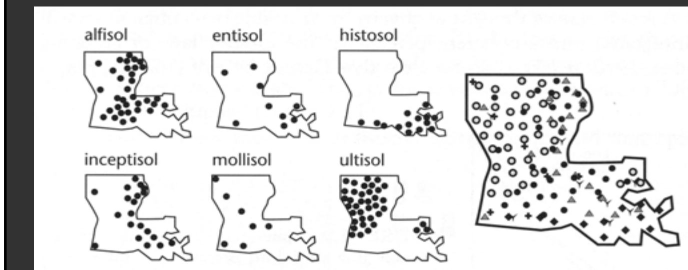

Encode “Marketing” (Q)
using *Size*

Trellis Plots

A *trellis plot* subdivides space to enable comparison across multiple plots.

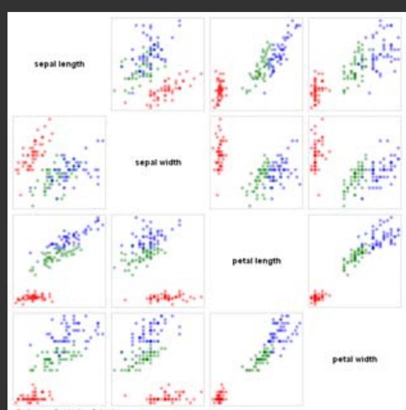
Typically nominal or ordinal variables are used as dimensions for subdivision.

Separation: Small Multiples

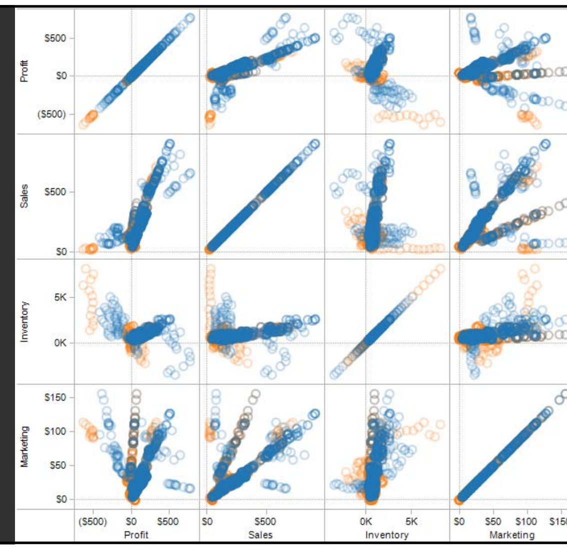


[Figure 2.11, p. 38, MacEachren 95]

Scatterplot Matrix (SPLOM)

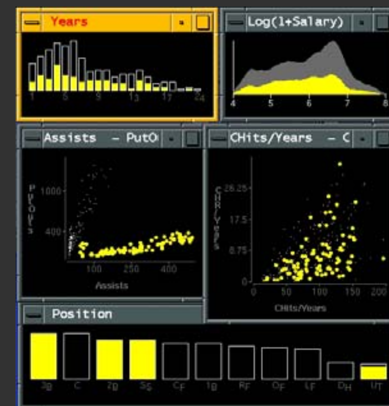


Scatter plots enabling pair-wise comparison of each data dimension.



Multiple Coordinated Views

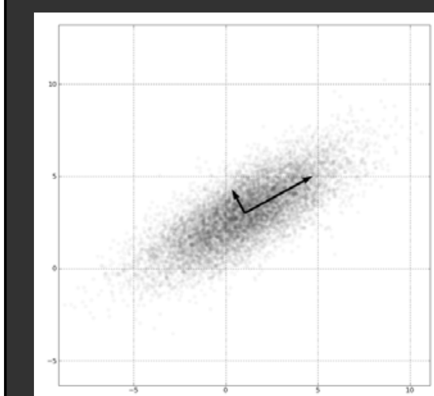
Linking Assists to Positions



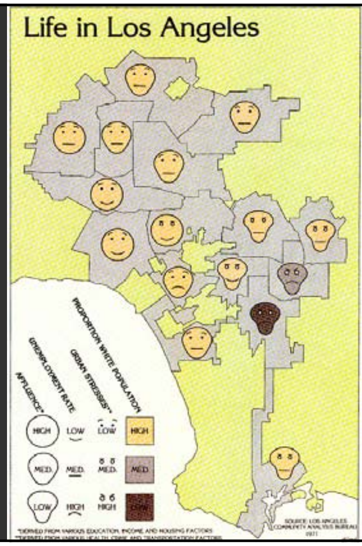
Dimensionality Reduction

http://www.ggobi.org/

Principal Component Analysis

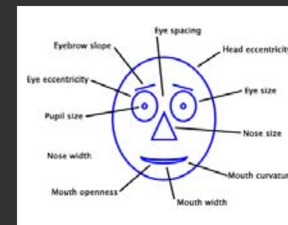


1. Mean-center the data.
2. Find ⊥ basis vectors that maximize the data variance.
3. Plot the data using the top vectors.



Chernoff Faces (1973)

Insight: We have evolved a sophisticated ability to interpret facial expression.



Idea: Map data variables to facial features.

Question: Do we process facial features in an uncorrelated way? (i.e., are they *separable*?)

This is just one example of nD "glyphs"

Visualizing Multiple Dimensions

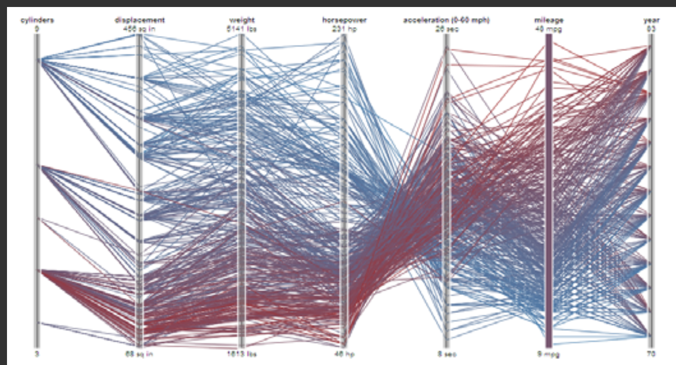
Strategies

- Avoid "over-encoding"
- Use space and small multiples intelligently
- Reduce the problem space
- Use interaction to generate *relevant* views

There is rarely a single visualization that answers all questions. Instead, the ability to generate appropriate visualizations quickly is key.

Parallel Coordinates

Parallel Coordinates [Inselberg]



Parallel Coordinates [Inselberg]

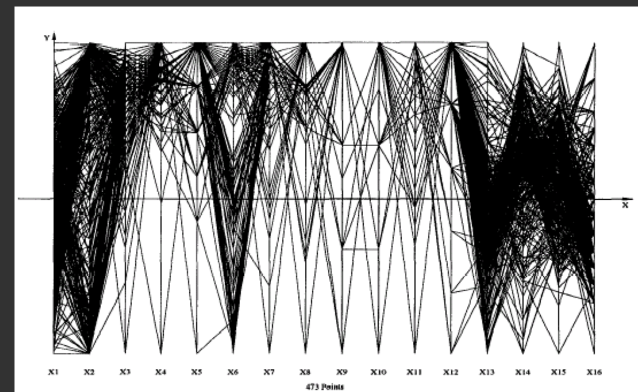


Figure 1: The full dataset consisting of 473 batches

The Multidimensional Detective

The Dataset:

- Production data for 473 batches of a VLSI chip
- 16 process parameters:

X1: The yield: % of produced chips that are useful
X2: The quality of the produced chips (speed)
X3 ... X12: 10 types of defects (zero defects shown at top)
X13 ... X16: 4 physical parameters

The Objective:
Raise the yield (X1) and maintain high quality (X2)

A. Inselberg, Multidimensional Detective, Proceedings of IEEE Symposium on Information Visualization (InfoVis '97), 1997

Parallel Coordinates

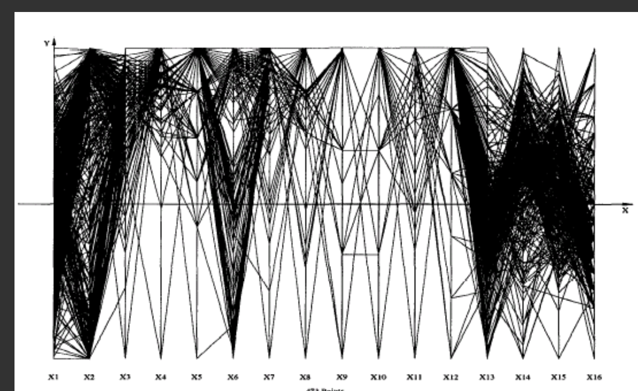


Figure 1: The full dataset consisting of 473 batches

Inselberg's Principles

1. Do not let the picture scare you
2. Understand your objectives
   - Use them to obtain visual cues
3. Carefully scrutinize the picture
4. Test your assumptions, especially the "I am really sure of's"
5. You can't be unlucky all the time!

Each line represents a tuple (e.g., VLSI batch)
Filtered below for high values of X1 and X2

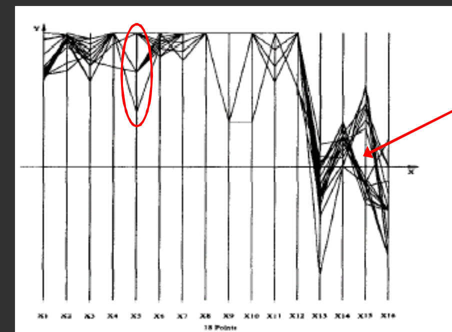
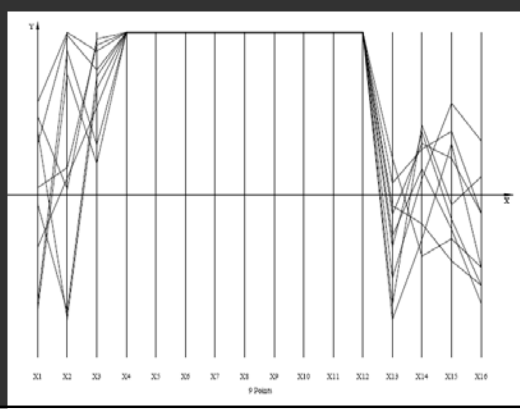


Figure 2: The batches high in Yield, $X1$, and Quality, $X2$.

Look for batches with *nearly* zero defects (9/10)
Most of these have low yields → defects OK.



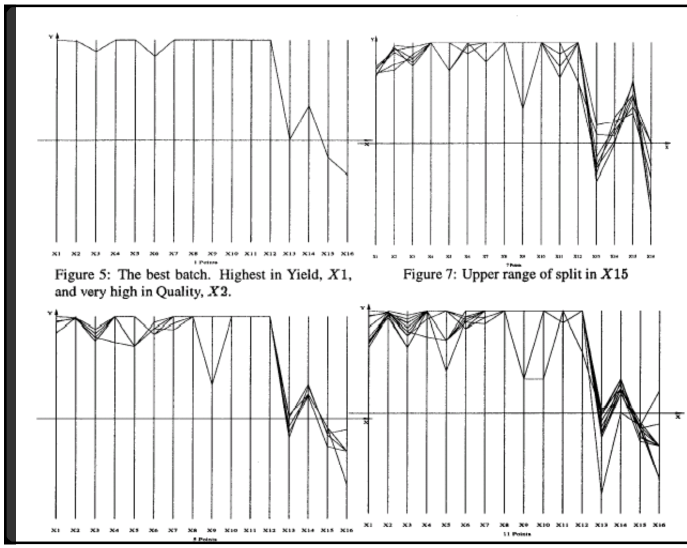


Figure 5: The best batch. Highest in Yield, $X1$, and very high in Quality, $X2$.

Figure 7: Upper range of split in $X15$

Notice that X6 behaves differently.
Allow 2 defects, including X6 → best batches

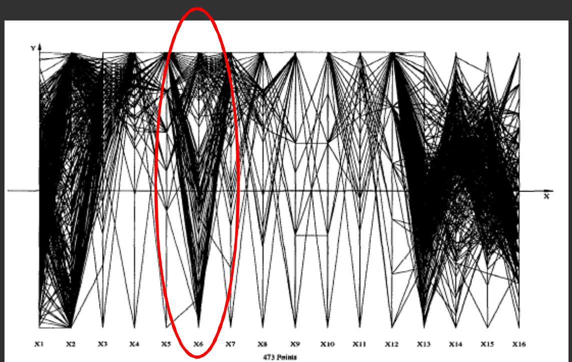
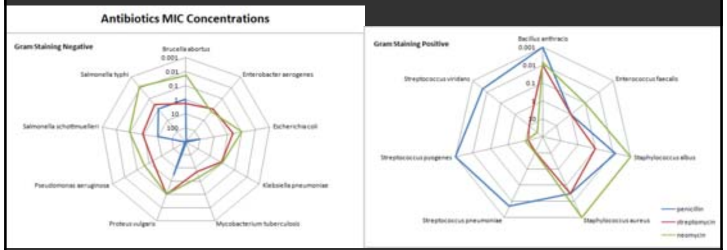


Figure 1: The full dataset consisting of 473 batches

Radar Plot / Star Graph

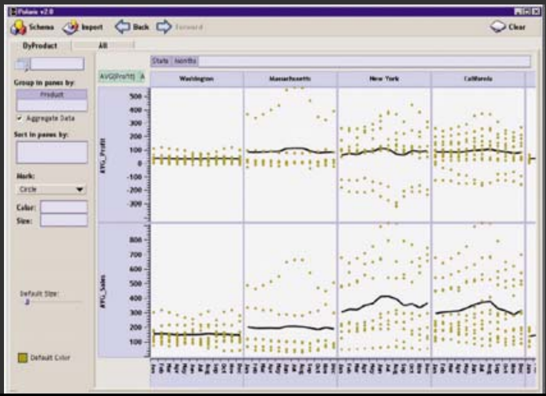


"Parallel" dimensions in polar coordinate space
Best if same units apply to each axis

Tableau / Polaris

Polaris

Research at Stanford by Stolte, Tang, and Hanrahan.



Tableau

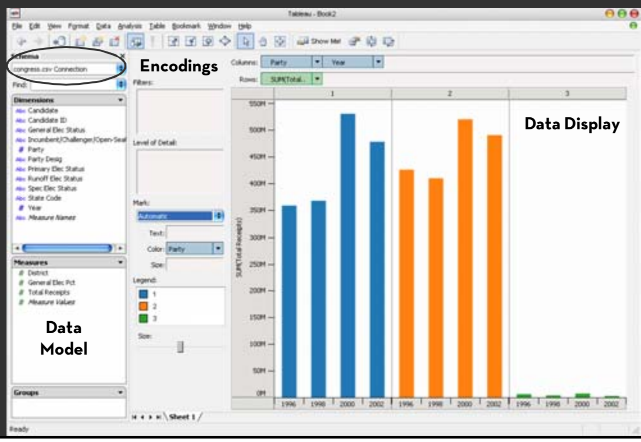


Tableau Demo

The dataset:

Federal Elections Commission Receipts
Every Congressional Candidate from 1996 to 2002
4 Election Cycles
9216 Candidacies

Data Set Schema

Year (Qi)
Candidate Code (N)
Candidate Name (N)
Incumbent / Challenger / Open-Seat (N)
Party Code (N) [1=Dem,2=Rep,3=Other]
Party Name (N)
Total Receipts (Qr)
State (N)
District (N)

This is a subset of the larger data set available from the FEC

Hypotheses?

What might we learn from this data?

- ??

Hypotheses?

What might we learn from this data?
Correlation between receipts and winners?
Do receipts increase over time?
Which states spend the most?
Which party spends the most?
Margin of victory vs. amount spent?
Amount spent between competitors?

Tableau Demo

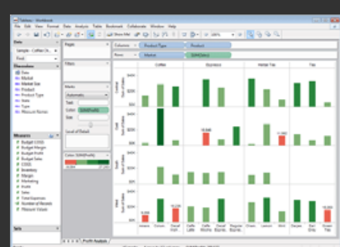
Assignment 2: Exploratory Data Analysis

Use visualization software (Tableau) to form & answer questions

First steps:
- Step 1: Pick domain & data
- Step 2: Pose questions
- Step 3: Profile the data
- Iterate as needed

Create visualizations
- Interact with data
- Refine your questions

Make wiki notebook
- Keep record of your analysis
- Prepare a final graphic and caption



Due by end-of-day
Tuesday, October 18

Polaris/Tableau Approach

Insight: can simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

More recently: automate visualization design

Specifying Table Configurations

Operands are the database fields

- Each operand interpreted as a set {…}
- Quantitative and Ordinal fields treated differently

Three operators:

- **concatenation** (+)
- **cross product** (x)
- **nest** (/)

Table Algebra: Operands

Ordinal fields: interpret domain as a set that partitions table into rows and columns.

Quarter = {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} →

Qtr1	Qtr2	Qtr3	Qtr4
95892	101760	105282	98225

Quantitative fields: treat domain as single element set and encode spatially as axes:

Profit = {(Profit[-410,650])} →

Concatenation (+) Operator

Ordered union of set interpretations

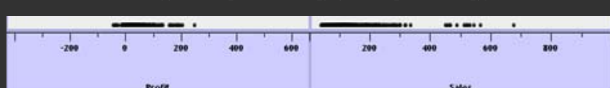
Quarter + Product Type

= {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} + {(Coffee), (Espresso)}

= {(Qtr1),(Qtr2),(Qtr3),(Qtr4),(Coffee),(Espresso)}

Qtr1	Qtr2	Qtr3	Qtr4	Coffee	Espresso
48	59	57	53	151	21

Profit + Sales = {(Profit[-310,620]),(Sales[0,1000])}



Cross (x) Operator

Cross-product of set interpretations

Quarter x Product Type

= {(Qtr1,Coffee), (Qtr1, Tea), (Qtr2, Coffee), (Qtr2, Tea), (Qtr3, Coffee), (Qtr3, Tea), (Qtr4, Coffee), (Qtr4,Tea)}

Qtr1		Qtr2		Qtr3		Qtr4	
Coffee	Espresso	Coffee	Espresso	Coffee	Espresso	Coffee	Espresso
131	19	160	20	178	12	134	33

Product Type x Profit =

Nest (/) Operator

Cross-product filtered by existing records

Quarter x Month

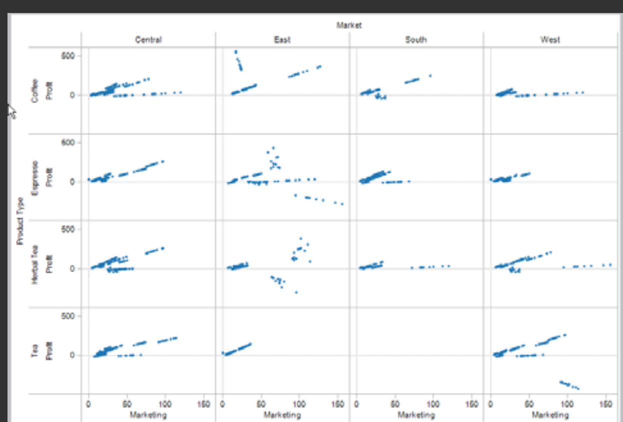
creates twelve entries for each quarter. i.e., (Qtr1, December)

Quarter / Month

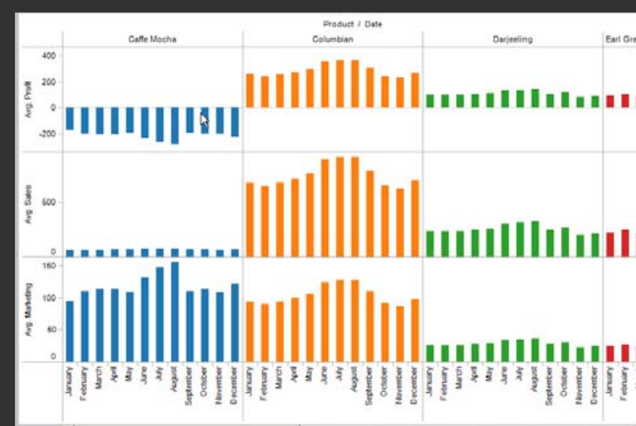
creates three entries per quarter based on tuples in database (not semantics)

Ordinal - Ordinal

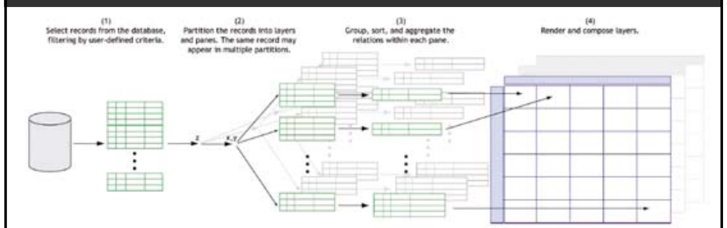
Quantitative - Quantitative



Ordinal - Quantitative



Querying the Database



Visualizing Multiple Dimensions

Strategies

- Start by visualizing individual dimensions
- Avoid “over-encoding”
- Use space and small multiples intelligently
- Use interaction to generate *relevant* views

There is rarely a single visualization that answers all questions. Instead, the ability to generate appropriate visualizations quickly is key.