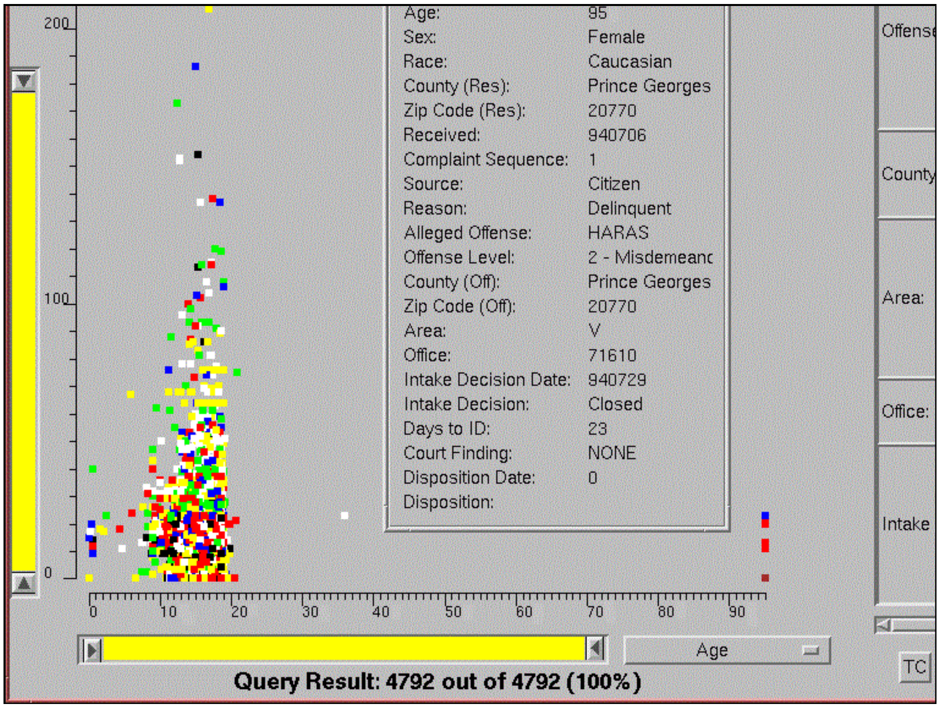
Multidimensional Visualization

Maneesh Agrawala

CS 448B: Visualization
Spring 2016

Last Time:
Exploratory Data Analysis



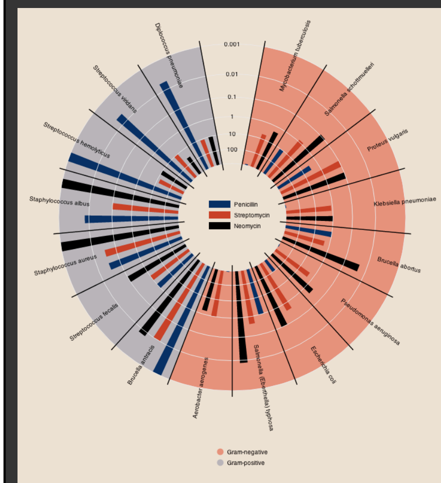
Data Quality & Usability Hurdles

Missing Data	no measurements, redacted, …?
Erroneous Values	misspelling, outliers, …?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

LESSON: **Anticipate problems with your data.**

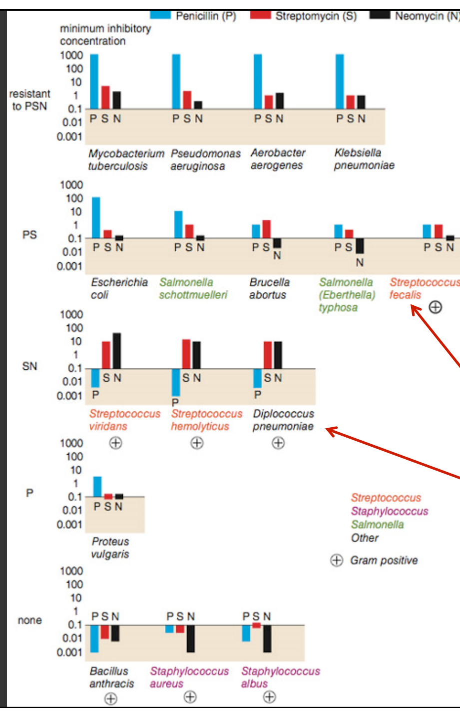
Many research problems around these issues!

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?



How do the bacteria group with respect to antibiotic resistance?

Not a streptococcus! (realized ~30 yrs later)

Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen

American Scientist, 2009

Lessons

Exploratory Process

1. Construct graphics to address questions
2. Inspect "answer" and assess new questions
3. Repeat!

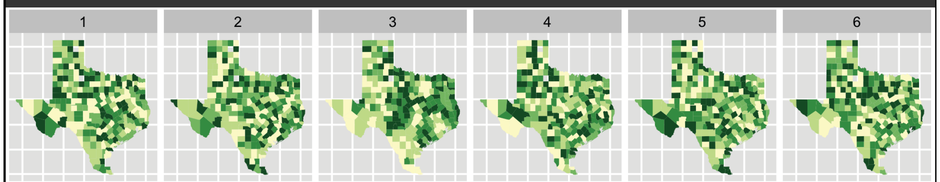
Transform the data appropriately (e.g., invert, log)

"Show data variation, not design variation"

-Tufte

Common Statistical Methods

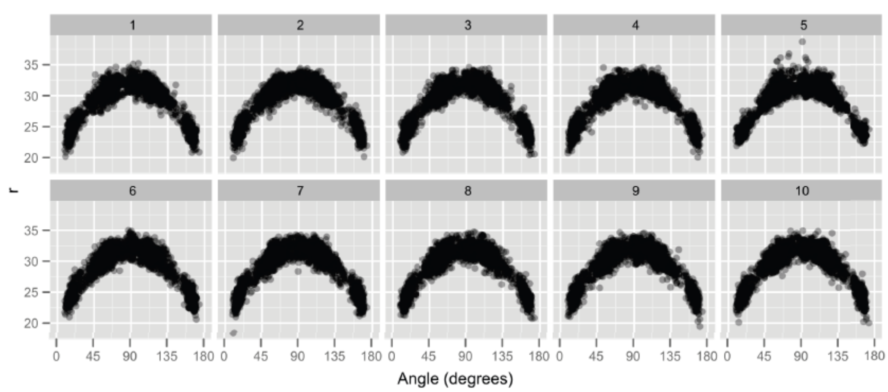
Question	Data Type	Parametric	Non-Parametric
Do data distributions have different "centers"? (aka "location" tests)	2 uni. dists > 2 uni. dists > 2 multi. dists	t-Test ANOVA MANOVA	Mann-Whitney U Kruskal-Wallis Median Test
Are observed counts significantly different?	Counts in categories		χ^2 (chi-squared)
Are two vars related?	2 variables	Pearson coeff.	Rank correl.
Do 1 (or more) variables predict another?	Continuous Binary	Linear regression Logistic regression	



Choropleth maps of cancer deaths in Texas.

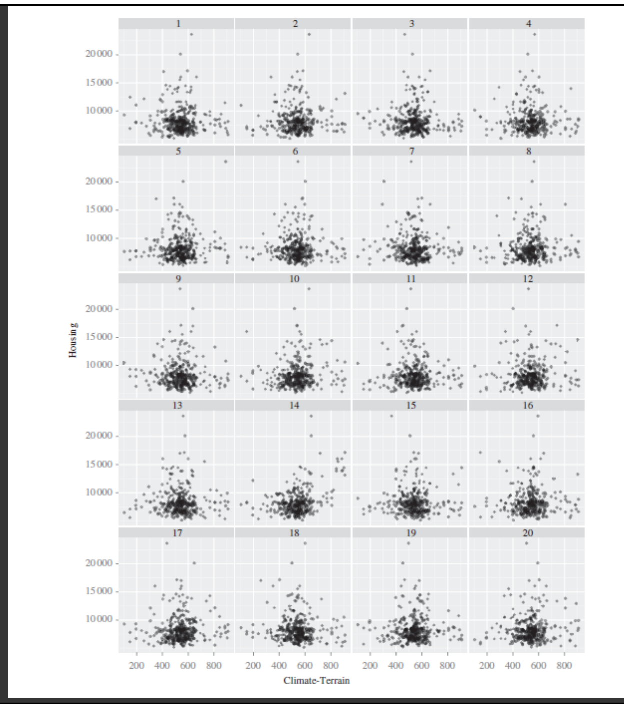
One plot shows a real data sets. The others are simulated under the null hypothesis of spatial independence.

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



Distance vs. angle for 3 point shots by the LA Lakers.

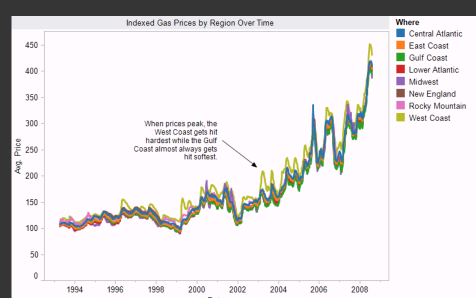
One plot is the real data. The others are generated according to a null hypothesis of quadratic relationship.



Assignment 2: Exploratory Data Analysis

Use Tableau to formulate & answer questions

First steps
- Step 1: Pick a domain
- Step 2: Pose questions
- Step 3: Find data
- Iterate



Create visualizations
- Interact with data
- Question will evolve
- Tableau

Make wiki notebook
- Keep record of all steps you took to answer the questions

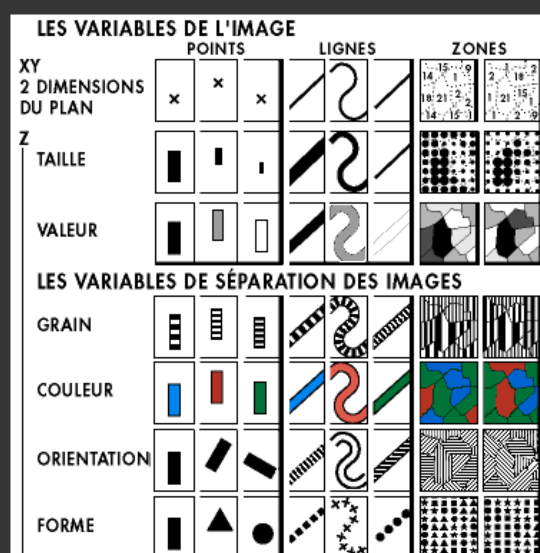
Due before class on Apr 18, 2016

Multidimensional Visualization

Visual Encoding Variables

- Position
- Length
- Area
- Volume
- Value
- Texture
- Color
- Orientation
- Shape

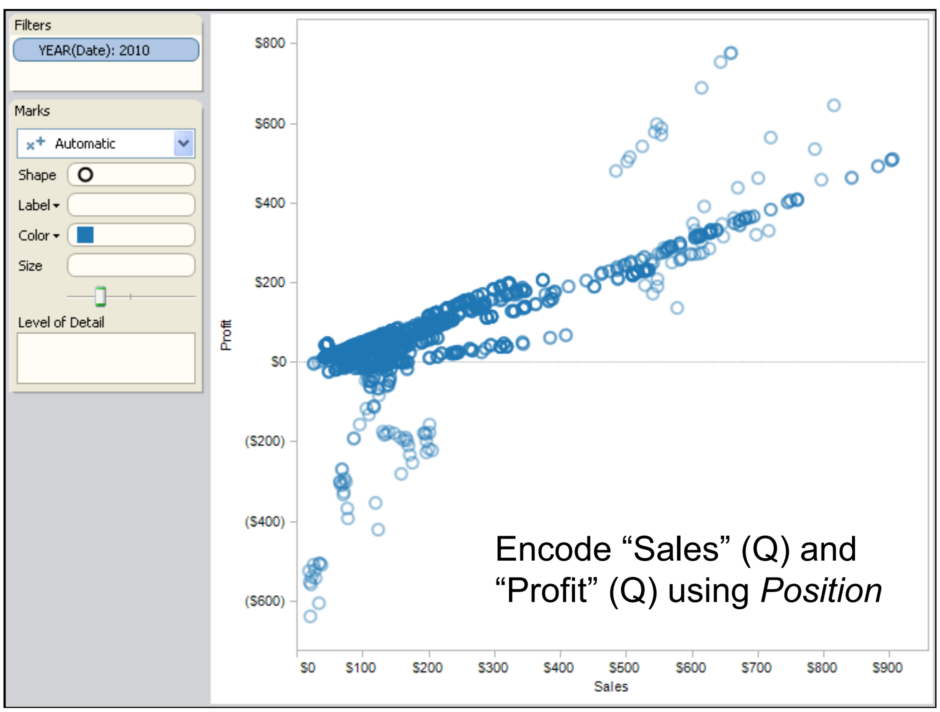
~8 dimensions?

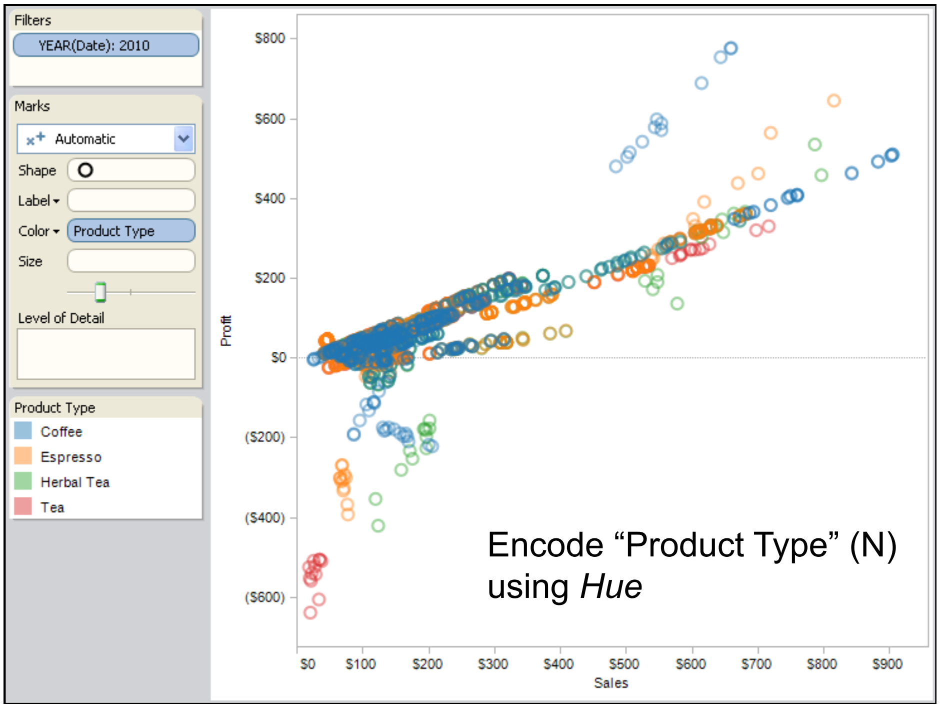


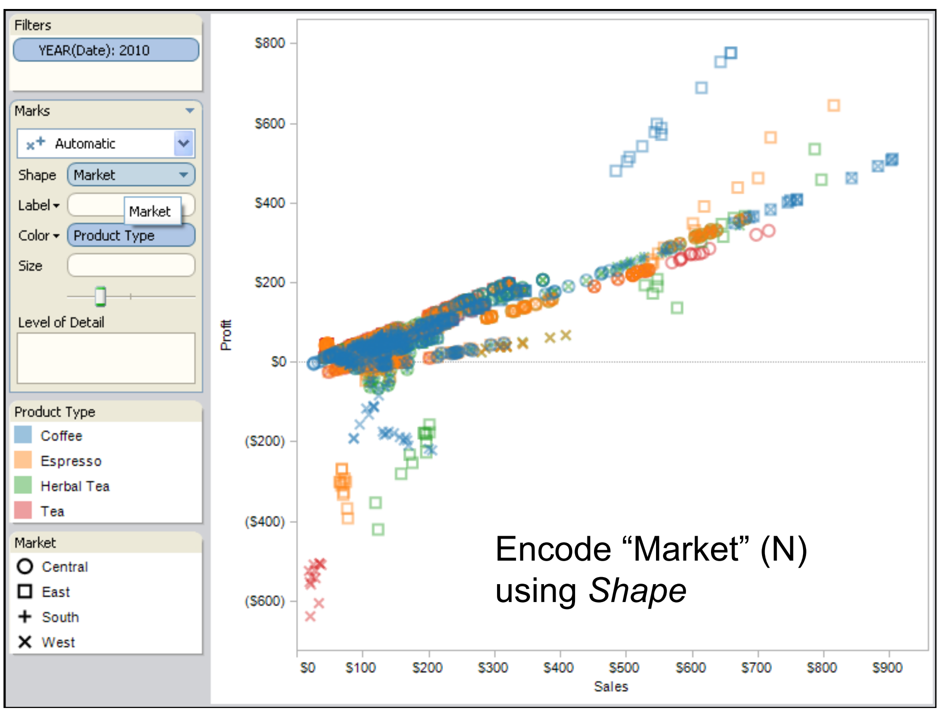
Example: Coffee Sales

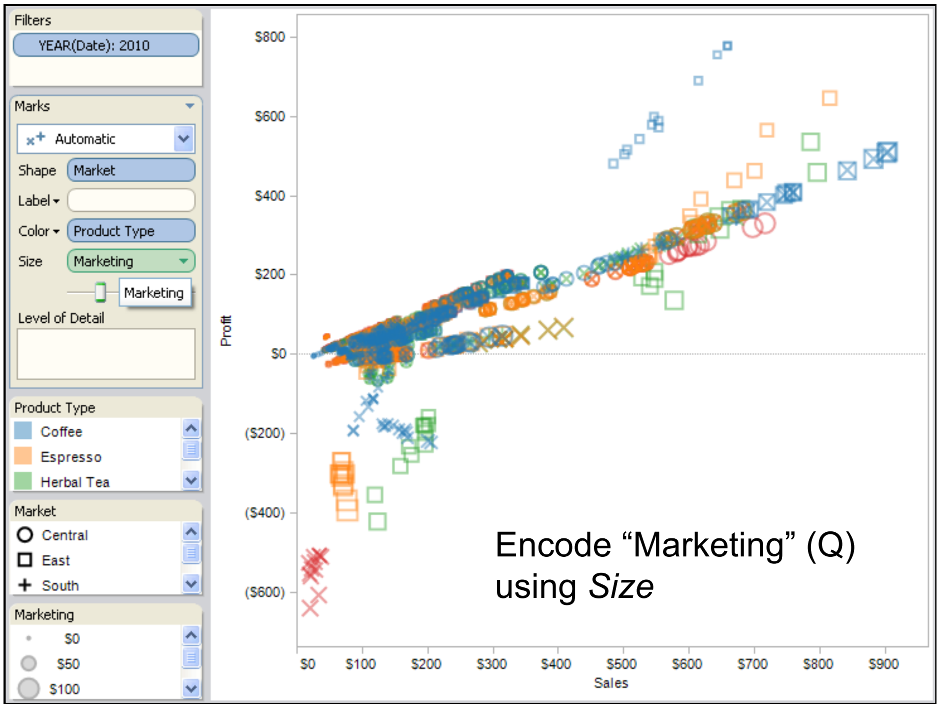
Sales figures for a fictional coffee chain:

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




Filters
YEAR(Date): 2010
Marks
Automatic
Shape
Label
Color
Product Type
Size
Level of Detail
Product Type
Coffee
Espresso
Herbal Tea
Tea
Profit
Sales
Encode "Product Type" (N) using *Hue*


Filters
YEAR(Date): 2010
Marks
Automatic
Shape
Market
Label
Market
Color
Product Type
Size
Level of Detail
Product Type
Coffee
Espresso
Herbal Tea
Tea
Market
Central
East
South
West
Profit
Sales
Encode "Market" (N) using *Shape*

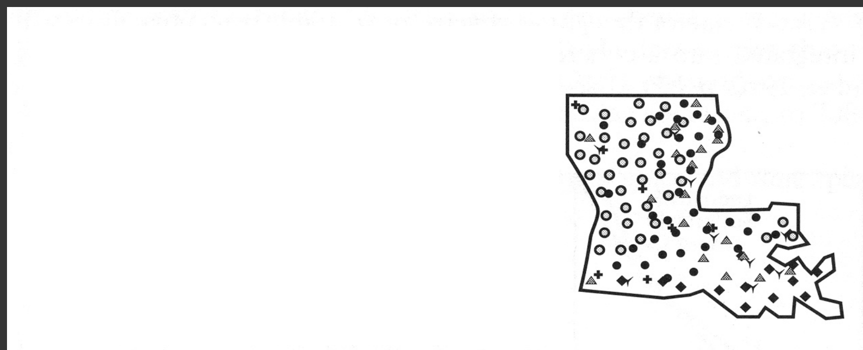


Trellis Plots

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

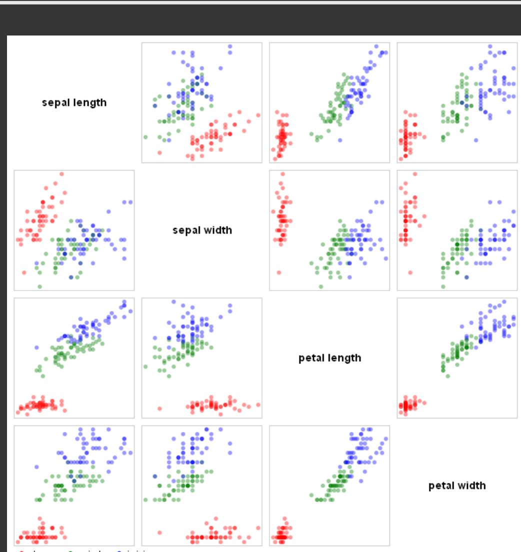
Typically nominal or ordinal variables are used as dimensions for subdivision

Separation: Small Multiples

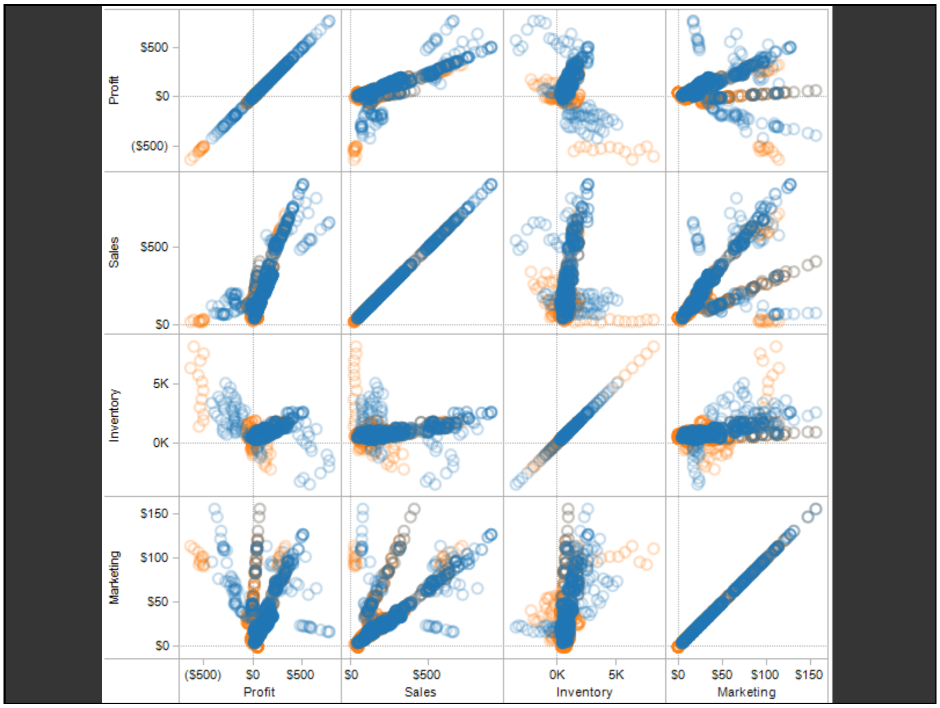


[Figure 2.11, p. 38, MacEachren 95]

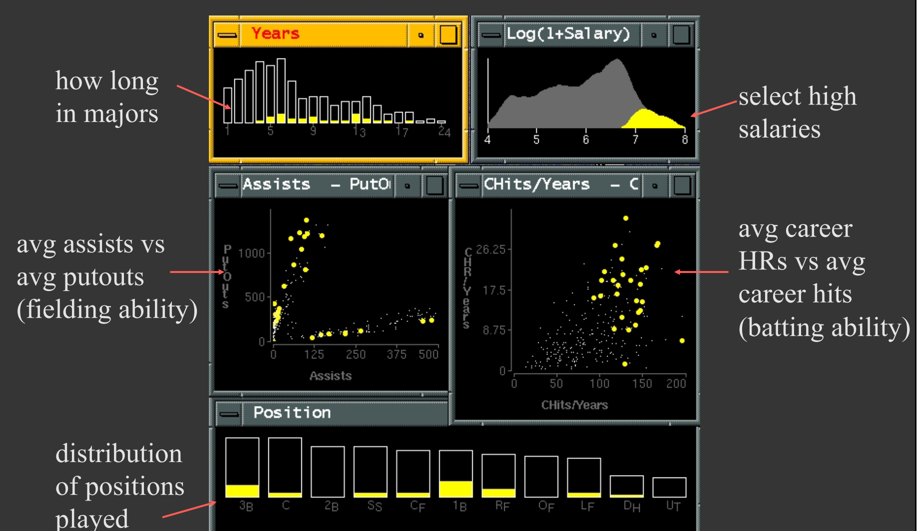
Scatterplot Matrix (SPLOM)



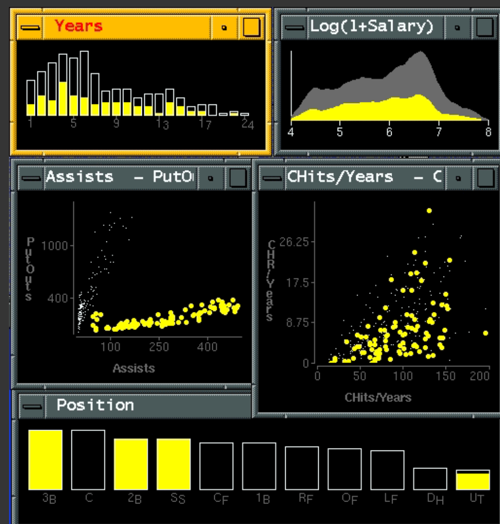
Scatter plots enabling pair-wise comparison of each data dimension



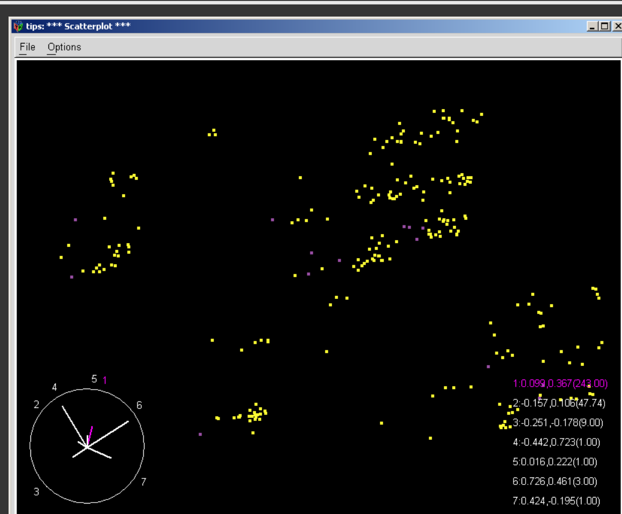
Small Multiples [from Wills 95]



Linking Assists to Positions

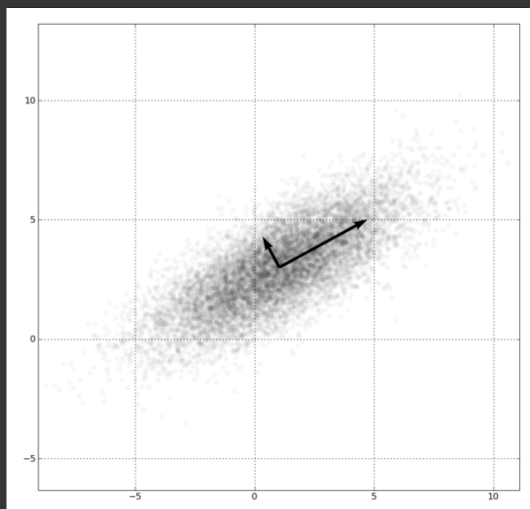


Dimensional Projection

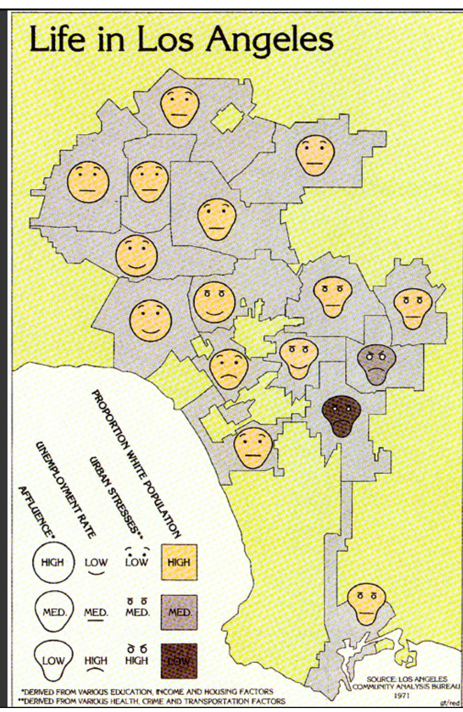


http://www.ggobi.org/

Principal Component Analysis



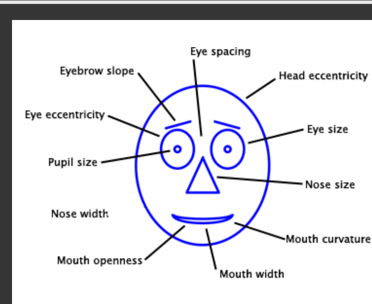
1. Mean-center the data
2. Find ⊥ basis vectors that maximize the data variance
3. Plot the data using the top vectors



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

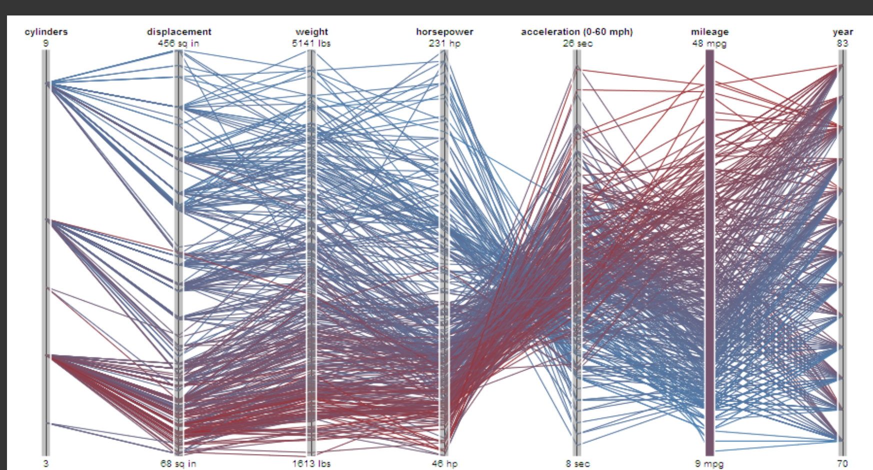
Strategies

- **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]



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

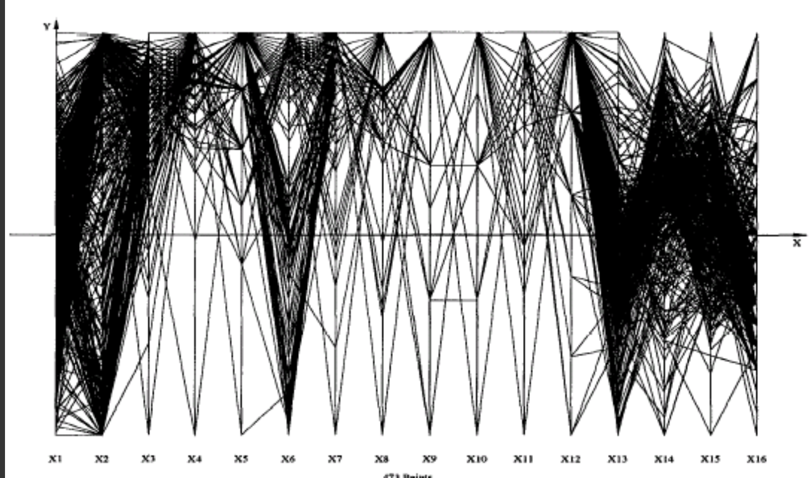


Figure 1: The full dataset consisting of 473 batches

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**

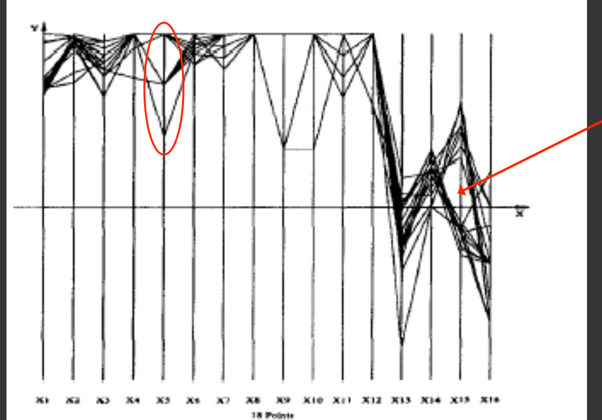
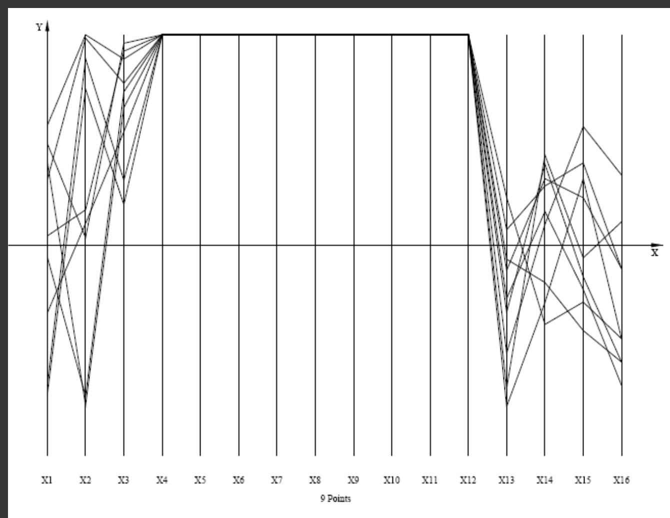


Figure 2: The batches high in Yield, $X1$, and Quality, $X2$.

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



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

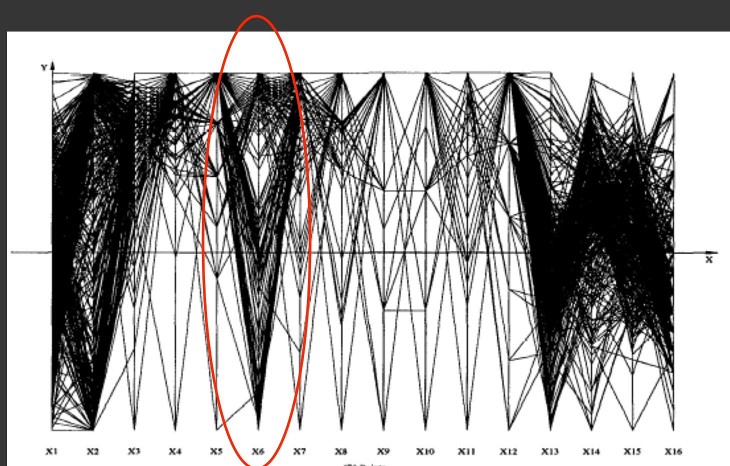
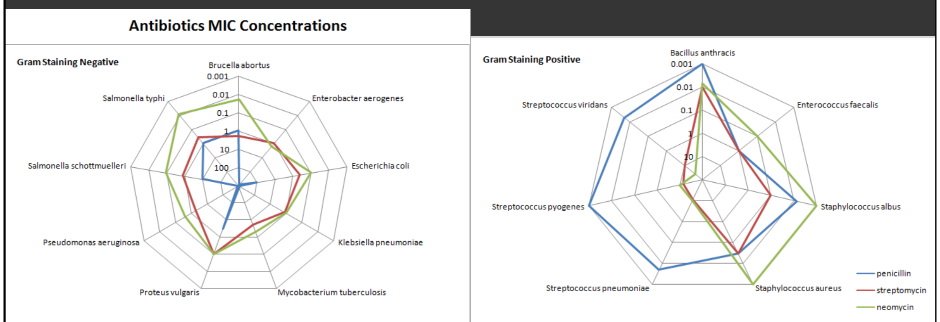


Figure 1: The full dataset consisting of 473 batches

Radar Plot / Star Graph



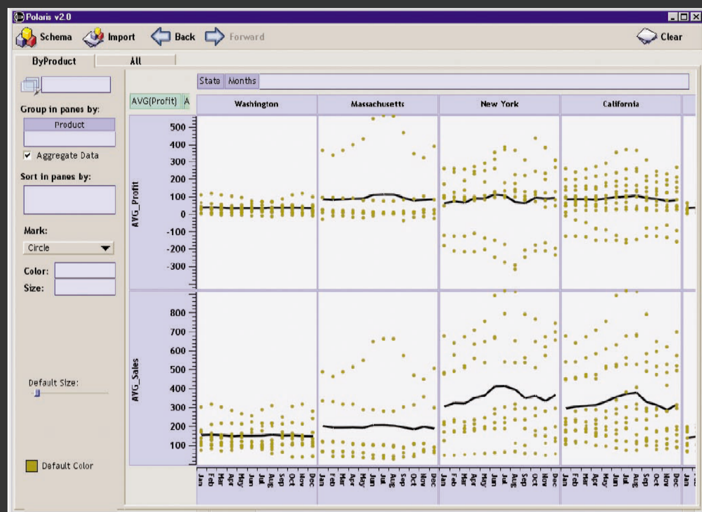
"Parallel" dimensions in polar coordinate space

Best if same units apply to each axis

Tableau / Polaris

Tableau

Research at Stanford: “Polaris” by Stolte, Tang & Hanrahan.



Tableau

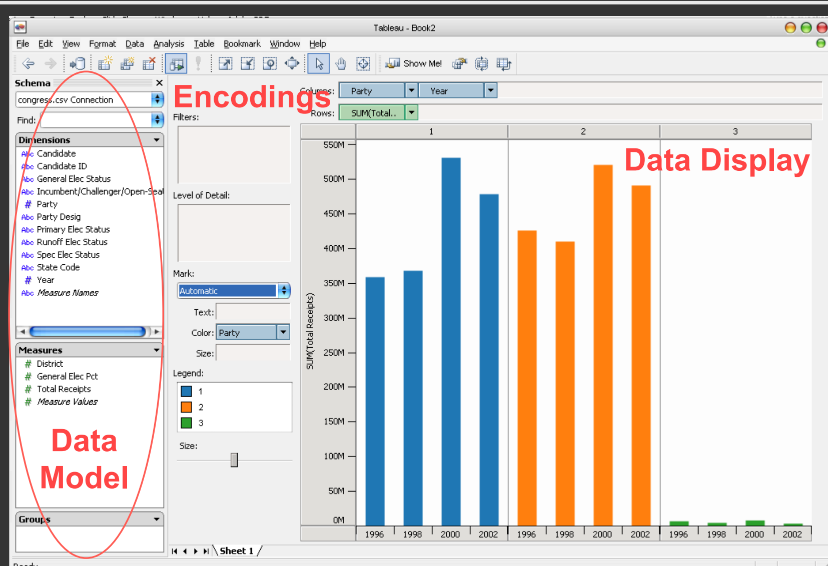


Tableau demo

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, but should be sufficient for the demo

Hypotheses?

What might we learn from this data?

Hypotheses?

What might we learn from this data?

- Has spending increased over time?
- Do democrats or republicans spend more money?
- Candidates from which state spend the most money?

Tableau Demo

Polaris/Tableau Approach

Insight: simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Recently: automate visualization design (ShowMe – Like APT)

Specifying Table Configurations

Operands are names of database fields

Each operand interpreted as a set {…}

Data is either Ordinal or Quantitative

Three operators:

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

Table Algebra: Operands

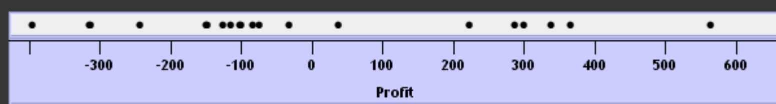
Ordinal fields: interpret domain as a set that partitions table into rows and columns

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

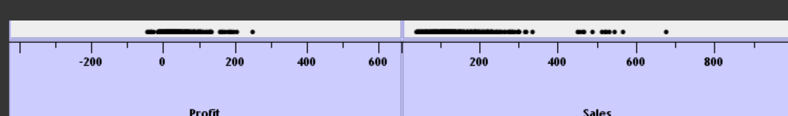
Quarter + Product Type

= {(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

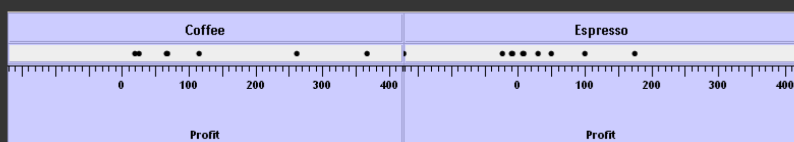
Cross-product of set interpretations

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)

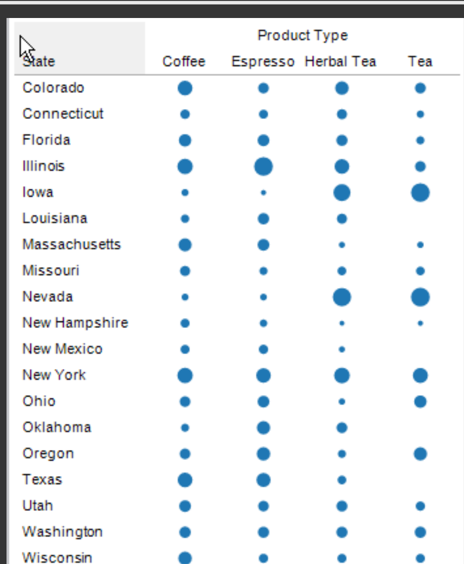
Polaris/Tableau Table Algebra

The operators (+, x, /) and operands (O, Q) provide an *algebra* for tabular visualization.

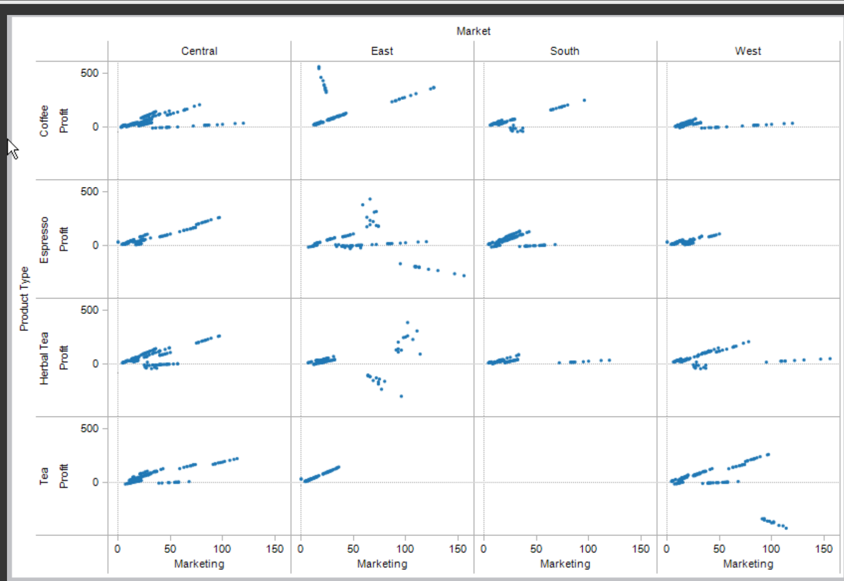
Algebraic statements are then mapped to:
- Visualizations - trellis plot partitions, visual encodings
- Queries - selection, projection, group-by aggregation

In Tableau, users make statements via drag-and-drop
- Note that this specifies operands NOT operators!
- Operators are inferred by data type (O, Q)

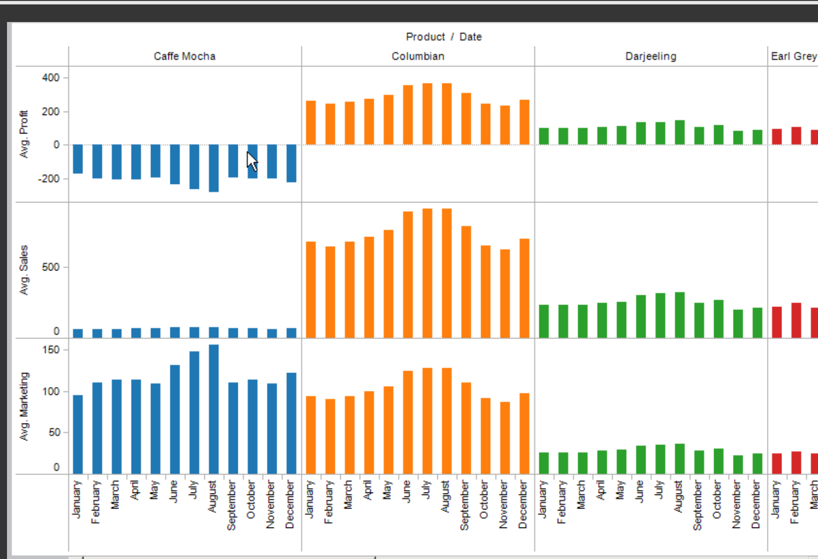
Ordinal - Ordinal



Quantitative - Quantitative



Ordinal - Quantitative



Summary

Visualizing Multiple Dimensions

- 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