CDS 6324
DATA VISUALIZATION

Lecture 7: Interactive Visualization

Exploratory Data Analysis:

A Quick Review

Exploratory Data Analysis (EDA)

- Before designing your visualization, get to know your data!
  - distributions (symmetric, normal, skewed)
  - data quality problems
  - outliers
  - correlations and inter-relationships
  - subsets of interest
  - suggest functional relationships

Exploratory Data Analysis (EDA)

- EDA - coined by Tukey, 1977

- Allows you to **generate hypotheses** as well as get a feel for you data
- Get an idea of how the experiment went without losing any richness in the data

Exploratory Data Analysis (EDA)

- EDA - coined by Tukey, 1977

- **Illuminate underlying pattern** in noisy data
- May lead to different analysis than originally planned

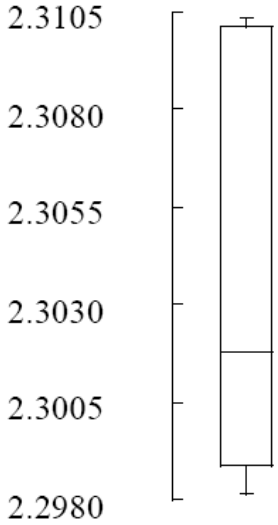
Revealing patterns

- Raw data is hard to understand
- EDA provides ways of presenting data that make the data easier to understand

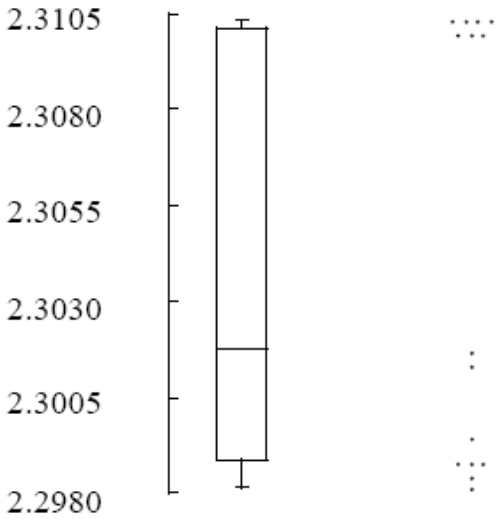
- Example of Lord Rayleigh's research on the weight of nitrogen
  - used a chemical compound to isolate a fixed amount of nitrogen
  - repeated this experiment 15 times

Date	*Source compound*	*Extraction method*	*Weight observed*
29.11.93	NO	hot iron	2.30143
5.12.93	NO	hot iron	2.29816
6.12.93	NO	hot iron	2.30182
8.12.93	NO	hot iron	2.29890
12.12.93	Air	hot iron	2.31017
14.12.93	Air	hot iron	2.30986
19.12.93	Air	hot iron	2.31010
22.12.93	Air	hot iron	2.31001
26.12.93	N_2O	hot iron	2.29889
28.12.93	N_2O	hot iron	2.29940
9.1.94	NH_4NO_2	hot iron	2.29849
13.1.94	NH_4NO_2	hot iron	2.29889
27.1.94	Air	ferrous hydrate	2.31024
30.1.94	Air	ferrous hydrate	2.31030
1.2.94	Air	ferrous hydrate	2.31028

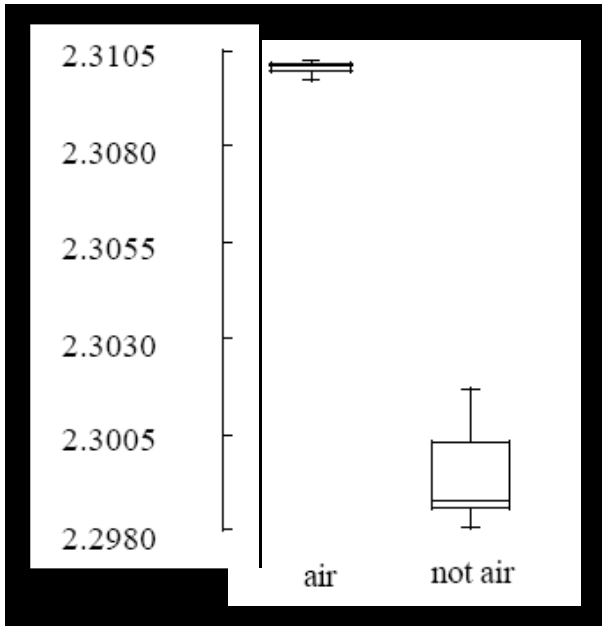
Box & whisker plot



Dot plot

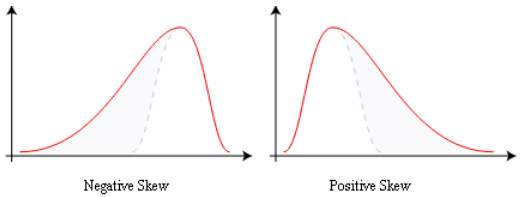


Two separate box & whisker plots



Statistics Analysis

- Some useful basic statistics measures for initial EDA:
  - **mean**: $\mu = \sum_i X_i / n$
  - **mode**: most common value in X
  - **median**: **X**=sort(X), median = $\mathbf{X}_{n/2}$ (half below, half above)
  - **quartiles of sorted X**: Q1 value = $\mathbf{X}_{0.25n}$, Q3 value = $\mathbf{X}_{0.75\ n}$
    - inter quartile range: value(Q3) - value(Q1)
    - range: max(X) - min(X) = $\mathbf{X}_n - \mathbf{X}_1$
  - **variance**: $\sigma^2 = \sum_i (X_i - \mu)^2 / n$
  - **skewness**: $\sum_i (X_i - \mu)^3 \ / \ [(\sum_i (X_i - \mu)^2)^{3/2}]$
    - zero if symmetric; right-skewed more common (what kind of data is right skewed?)



EDA in a Nutshell

- **Goal**: get a general sense of the data
  - means, medians, quantiles, histograms, boxplots
  - You should always look at every variable - you will learn something!

- **Think interactive and visual** :
  - Humans are the best pattern recognizers
  - You can use more than 2 dimensions!
    - x,y,z, space, color, time….

- **EDA is useful for**:
  - detecting outliers (e.g. assess data quality)
  - testing assumptions (e.g. normal distributions or skewed?)
  - identifying useful raw data & transforms (e.g. log(x))

- **Bottom line**: it is always well worth looking at your data!

Precise descriptions are better

- "Most of the key questions in our world sooner or later demand answers to "**by how much**?" rather than merely to "in which direction?“

  (Tukey, 1977)

- The **EDA** approach is precisely that -- **an approach -- not a set of techniques**, but an attitude/philosophy about how a data analysis should be carried out.

- Detailed reference on EDA:

  http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

INTERACTION

Interaction between people and

machines requires *mutual intelligibility*

or *shared understanding.*

Taxonomy of Interaction

Data and View Specification

- Visualize, Filter, Sort, Derive

View Manipulation

- Select, Navigate, Coordinate, Organize

Process and Provenance

- Record, Annotate, Share, Guide

Taxonomy of Interaction

Data and View Specification

- Visualize, Filter, Sort, Derive

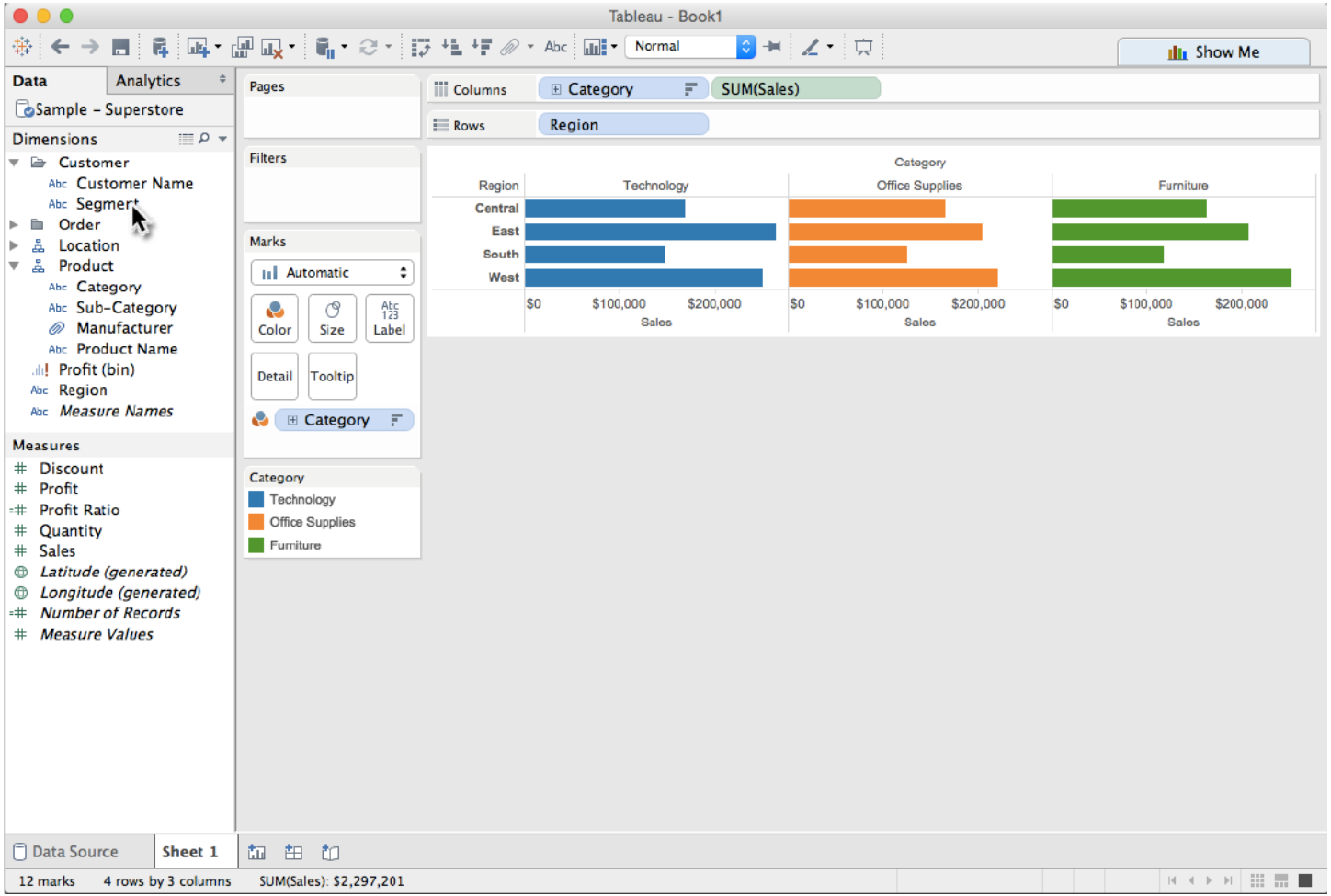


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Category
Category
Technology
Office Supplies
Furniture
Columns
Category
SUM(Sales)
Rows
Region
Category
Region
Technology
Office Supplies
Furniture
Central
East
South
West
$0
$100,000
$200,000
Sales
Data Source
Sheet 1
12 marks
4 rows by 3 columns
SUM(Sales): $2,297,201

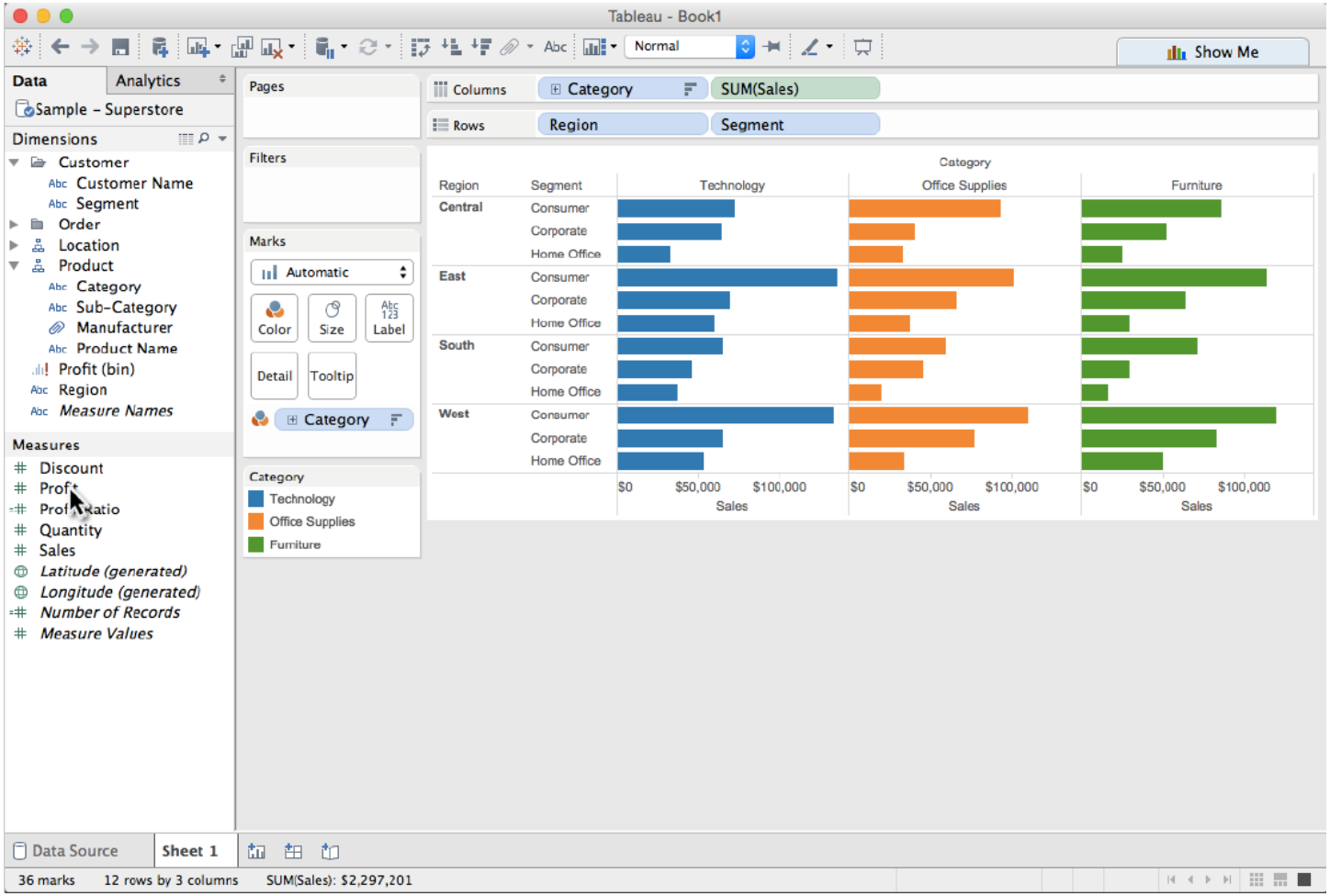


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Category
Category
Technology
Office Supplies
Furniture
Columns
Category
SUM(Sales)
Rows
Region
Segment
Category
Region
Segment
Technology
Office Supplies
Furniture
Central
East
South
West
Consumer
Corporate
Home Office
$0
$50,000
$100,000
Sales
Data Source
Sheet 1
36 marks
12 rows by 3 columns
SUM(Sales): $2,297,201

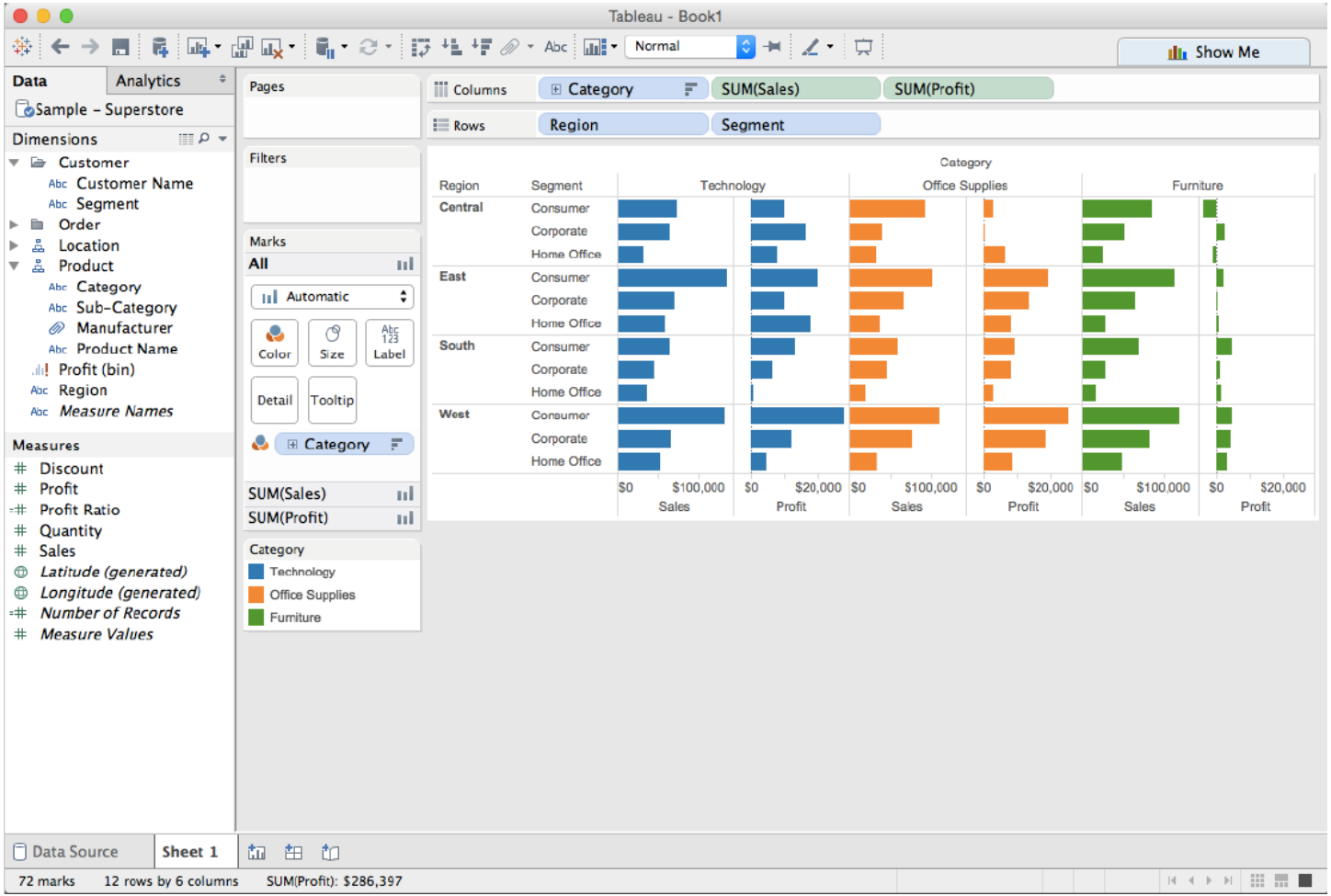
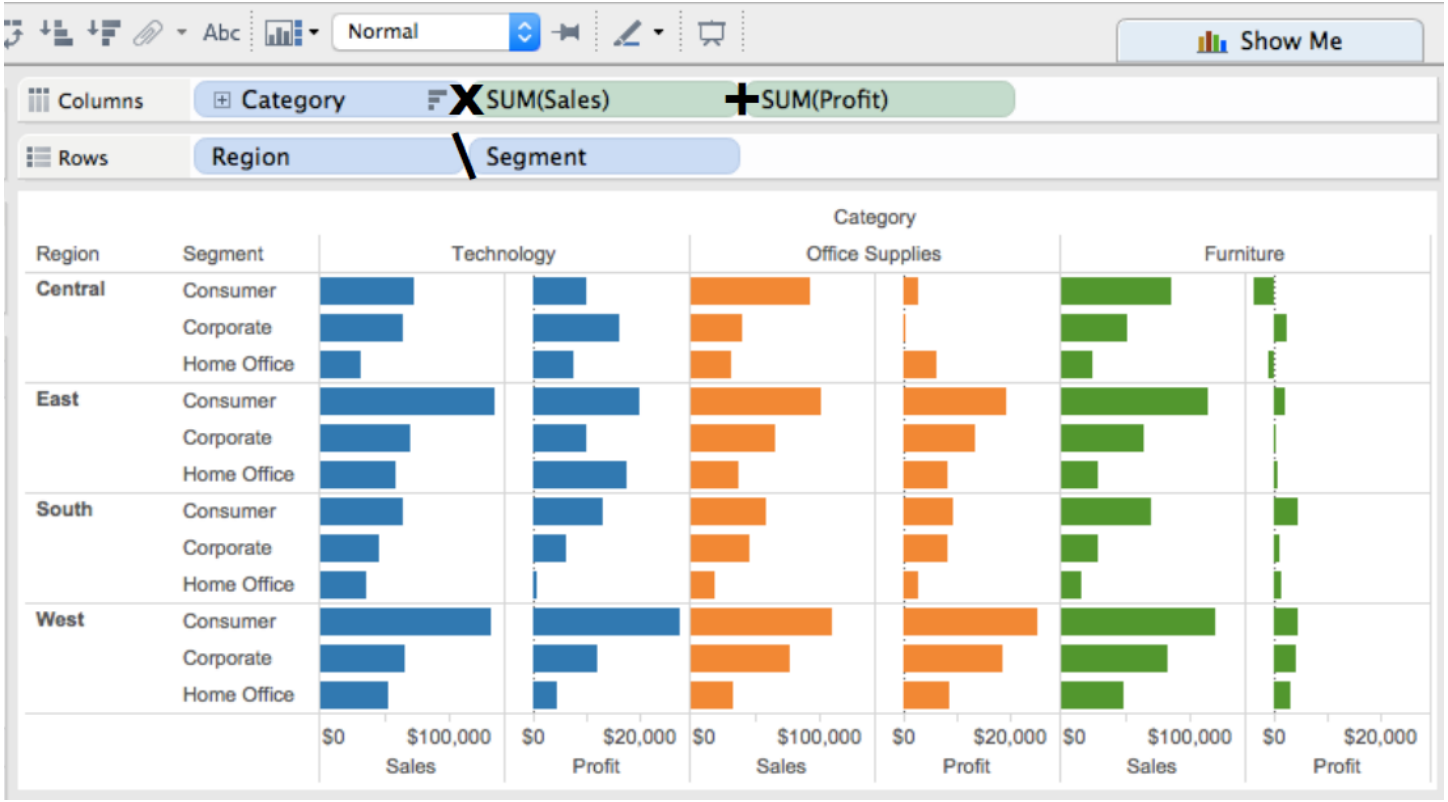


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Marks
All
Automatic
Color
Size
Label
Detail
Tooltip
Category
SUM(Sales)
SUM(Profit)
Category
Technology
Office Supplies
Furniture
Columns
Category
SUM(Sales)
SUM(Profit)
Rows
Region
Segment
Category
Region
Segment
Technology
Office Supplies
Furniture
Central
East
South
West
Consumer
Corporate
Home Office
$0
$100,000
$20,000
Sales
Profit
Data Source
Sheet 1
72 marks
12 rows by 6 columns
SUM(Profit): $286,397


Normal
Show Me
Columns
Category
SUM(Sales)
SUM(Profit)
Rows
Region
Segment
Category
Region
Segment
Technology
Office Supplies
Furniture
Central
East
South
West
Consumer
Corporate
Home Office
$0
$100,000
$20,000
Sales
Profit

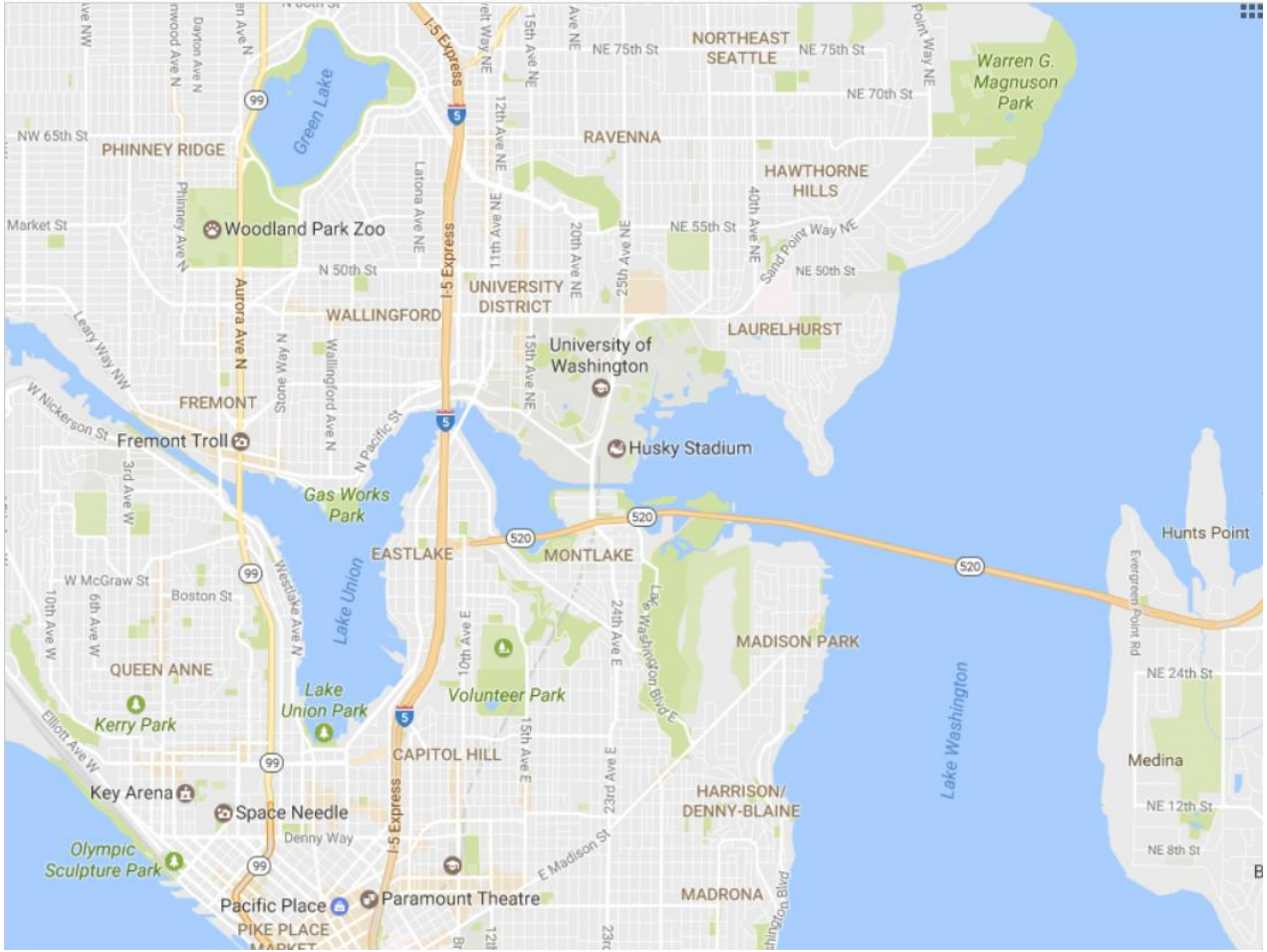
Taxonomy of Interaction

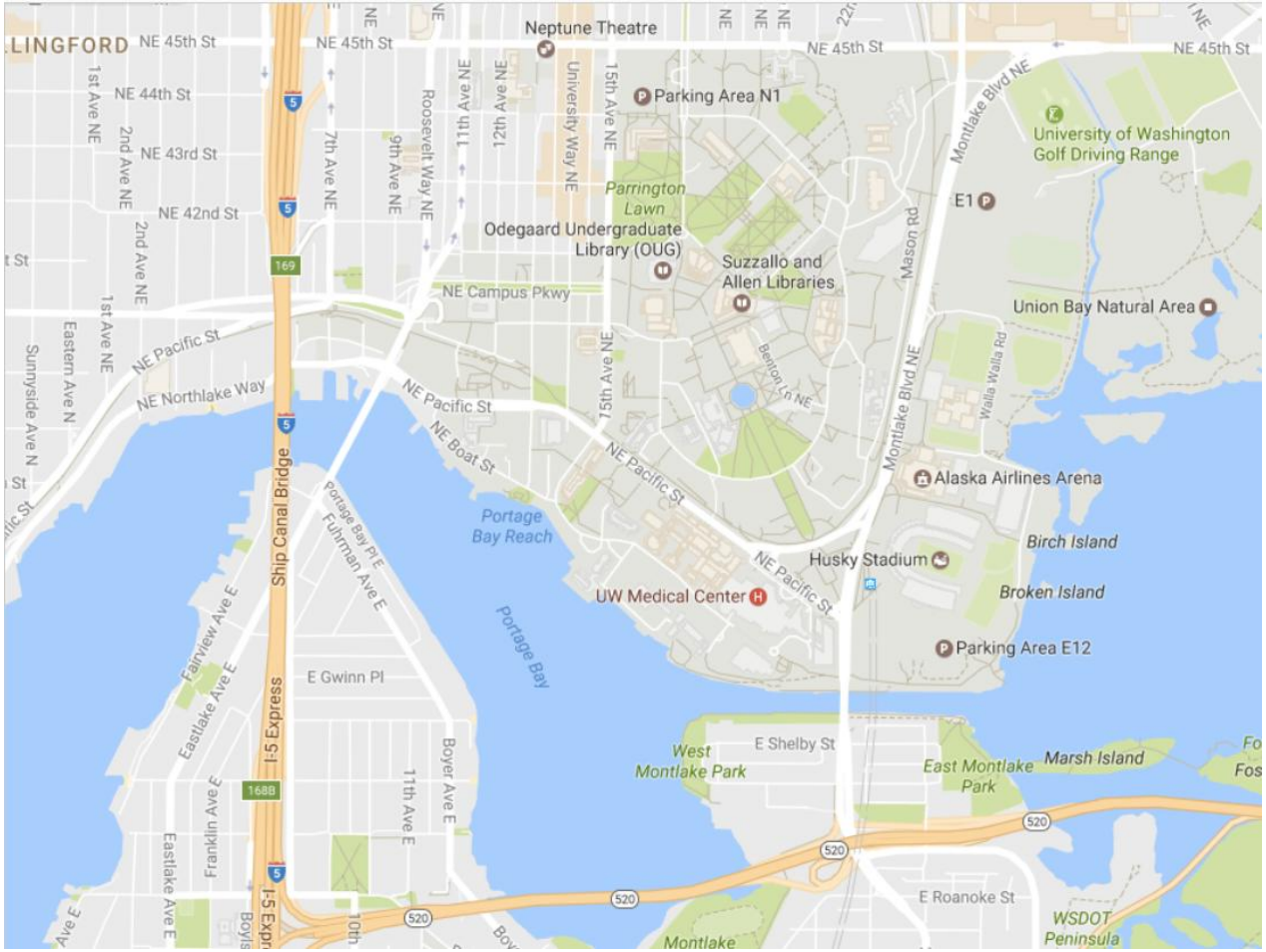
Data and View Specification

- Visualize, Filter, Sort, Derive

View Manipulation

- Select, Navigate, Coordinate, Organize


NORTHEAST SEATTLE
NE 75th St
NE 70th St
Warren G. Magnuson Park
RAVENNA
HAWTHORNE HILLS
PHINNEY RIDGE
NW 65th St
Green Lake
Woodland Park Zoo
Market St
NE 55th St
N 50th St
NE 50th St
UNIVERSITY DISTRICT
WALLINGFORD
LAURELHURST
University of Washington
FREMONT
Fremont Troll
Husky Stadium
Gas Works Park
EASTLAKE
MONTLAKE
Hunts Point
Lake Union
Lake Union Park
Volunteer Park
MADISON PARK
QUEEN ANNE
Kerry Park
NE 24th St
CAPITOL HILL
Medina
Key Arena
Space Needle
HARRISON/ DENNY-BLAINE
Lake Washington
NE 12th St
NE 8th St
Olympic Sculpture Park
Pacific Place
Paramount Theatre
MADRONA
PIKE PLACE
I-5 Express
Aurora Ave N
Lake Washington Blvd E
Sand Point Way NE
Leary Way NW
W Nickerson St
W McGraw St
Boston St
Denny Way
E Madison St
Evergreen Point Rd


Neptune Theatre
NE 45th St
NE 44th St
NE 43rd St
NE 42nd St
Parking Area N1
University of Washington Golf Driving Range
Parrington Lawn
Odegaard Undergraduate Library (OUG)
Suzzallo and Allen Libraries
E1
Union Bay Natural Area
NE Campus Pkwy
NE Pacific St
NE Northlake Way
NE Boat St
Ship Canal Bridge
I-5 Express
Portage Bay Reach
Portage Bay
Alaska Airlines Arena
Husky Stadium
Birch Island
Broken Island
UW Medical Center
Parking Area E12
E Gwinn Pl
E Shelby St
West Montlake Park
East Montlake Park
Marsh Island
E Roanoke St
WSDOT Peninsula
Montlake
Montlake Blvd NE
Mason Rd
Roosevelt Way NE
University Way NE
15th Ave NE
7th Ave NE
9th Ave NE
11th Ave NE
12th Ave NE
Fairview Ave E
Eastlake Ave E
Fuhrman Ave E
Boyer Ave E
11th Ave E
Franklin Ave E
Portage Bay Pl E
Benton Ln NE
Walla Walla Rd
Sunnyside Ave N
Eastern Ave N
1st Ave NE
2nd Ave NE
520
169
168B

Taxonomy of Interaction

Data and View Specification

- Visualize, Filter, Sort, Derive

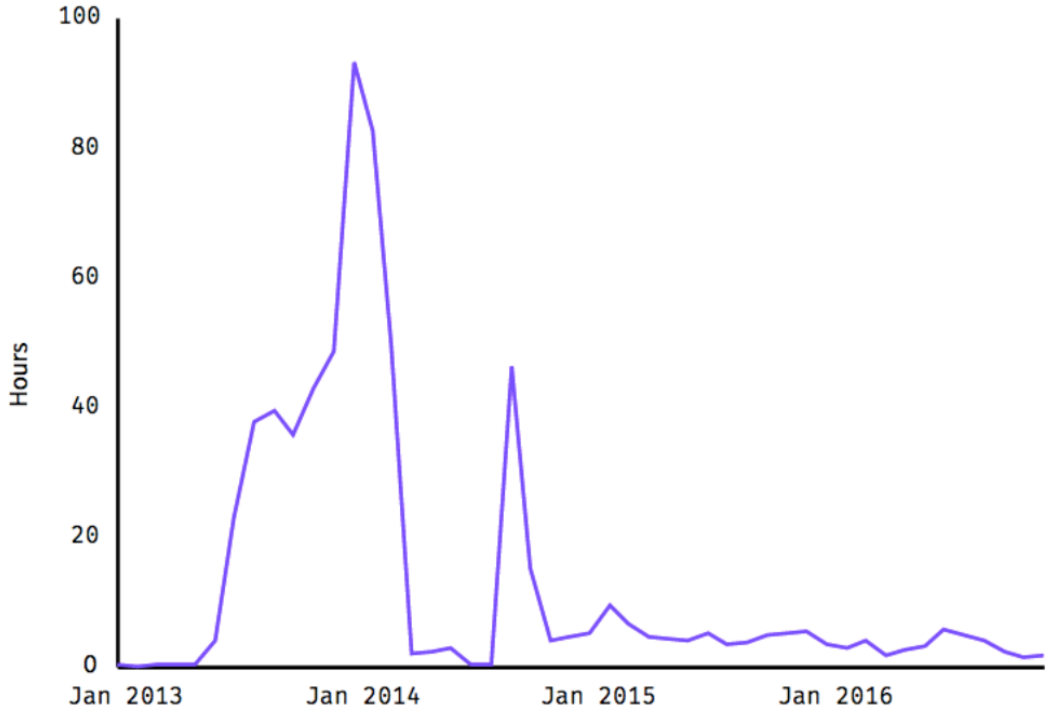
View Manipulation

- Select, Navigate, Coordinate, Organize

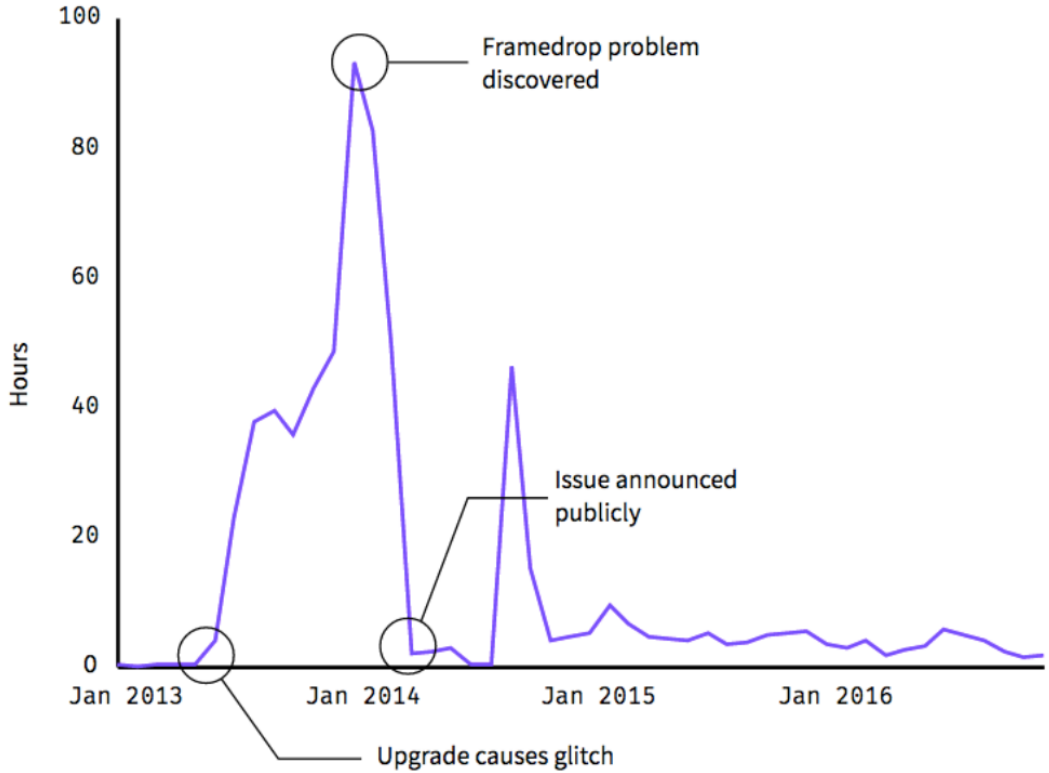
Process and Provenance

- Record, Annotate, Share, Guide

Hours of footage lost each month due to dropped frames



Hours of footage lost each month due to dropped frames



Taxonomy of Interaction

Data and View Specification

- Visualize, Filter, Sort, Derive

View Manipulation

- Select, Navigate, Coordinate, Organize

Process and Provenance

- Record, Annotate, Share, Guide

A Classic Example: Reorderable Matrix

- Objects and characteristics are reorderable (≠≠)
- Permutable in x and y
- Overall relationship is discovered by permutations

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P		
							1			1						1	HIGH SCHOOL
	1	1	1			1					1			1		2	AGRICULTRAL COOP.
							1			1						3	RAILWAY STATION
1	1			1	1			1	1			1	1		1	4	ONE ROOM SCHOOL
	1	1	1			1					1			1		5	VETERINARY
1	1			1	1			1	1			1	1		1	6	NO DOCTOR
									1				1			7	NO WATER SUPPLY
		1					1			1						8	POLICE STATION
	1	1	1			1					1			1		9	LAND REALLOCATION

Special Cases for (≠≠)

- Weighted matrix
  - Areas become meaningful
  - Applicable to a data table in which row and column totals are meaningful
  - Limited in dimension

- Matrix-file
  - When one of the dimensions is too large
  - Constructed similar to image files
  - Use sorting to discover correlations

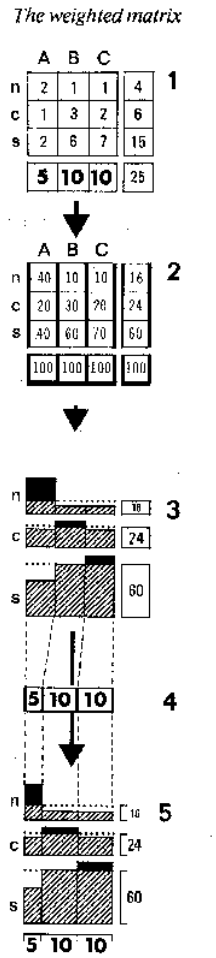
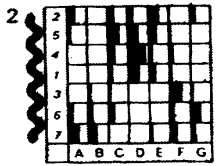
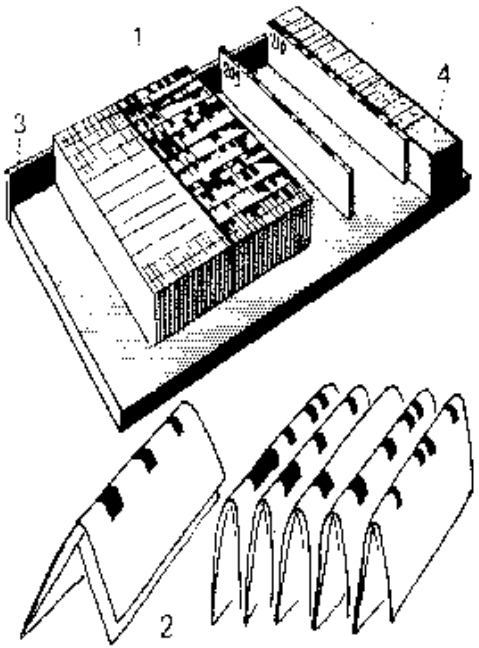


Image File

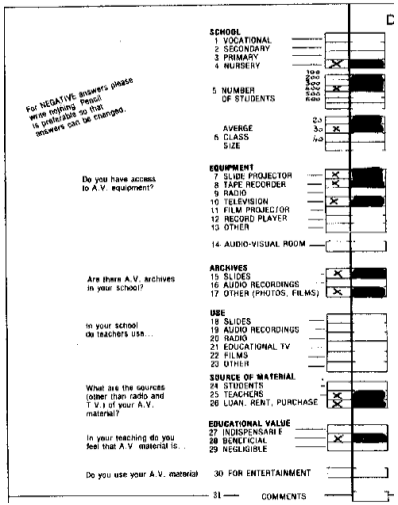
- Used for ordered objects and reorderable characteristics
- One card for each characteristic
- Values greater than the mean of that row are darkened

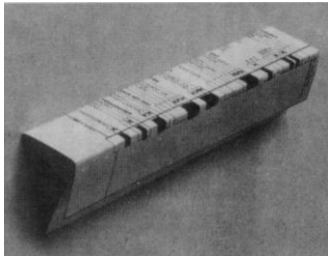




Matrix-File

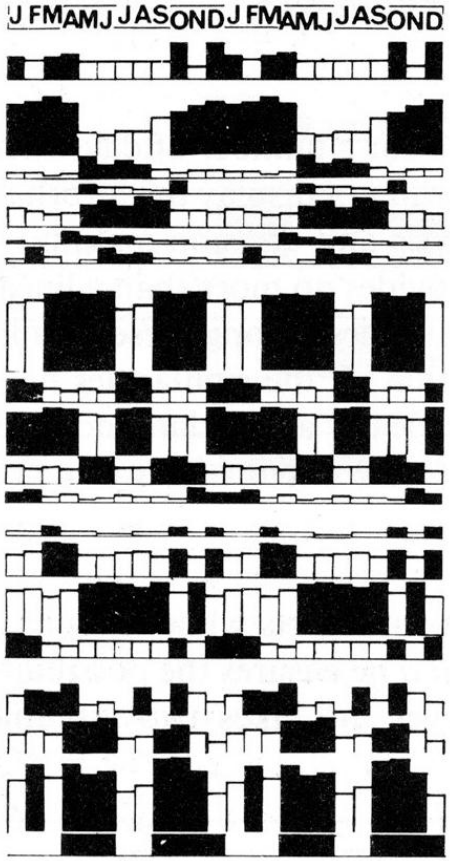
- Special case for permutable matrix; one of the dimensions is too big.
- Large number of objects across a small number of characteristics.
- Constructed similar to image files
- Use sorting to discover correlations

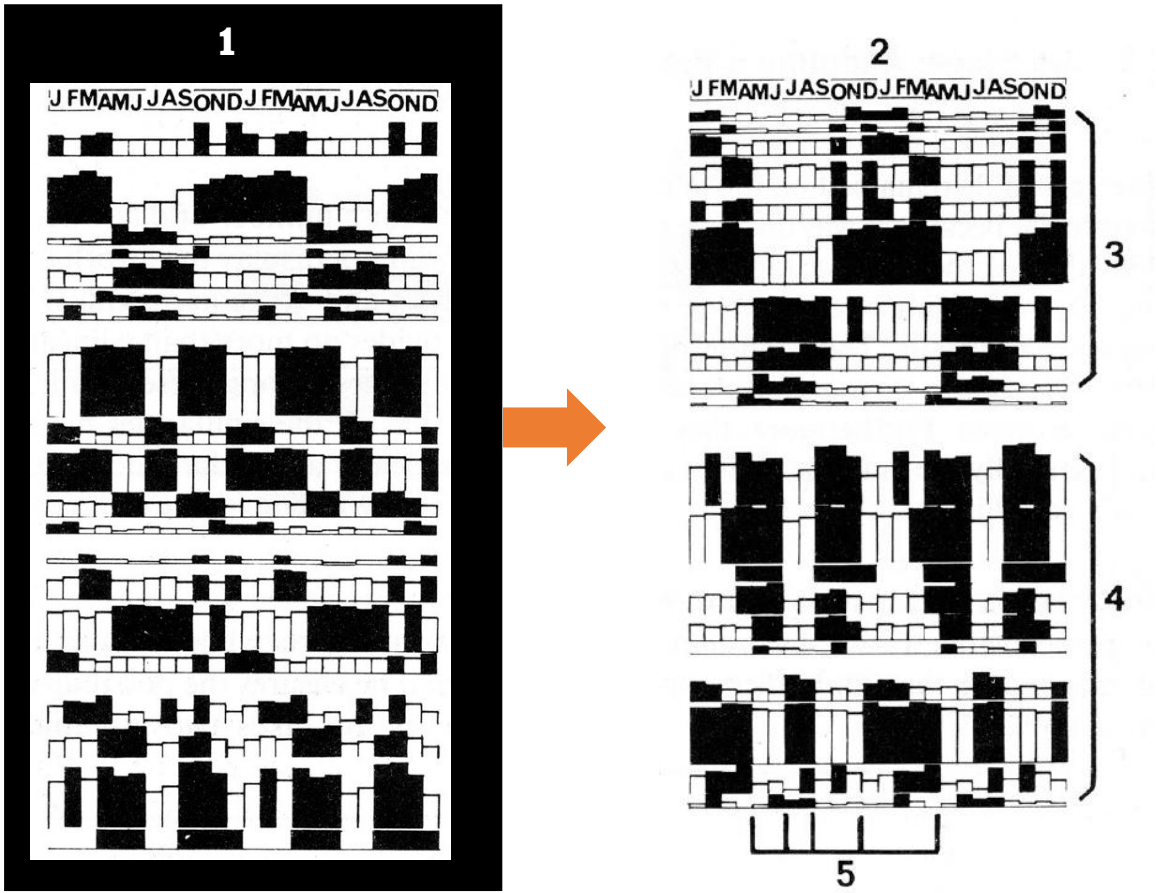


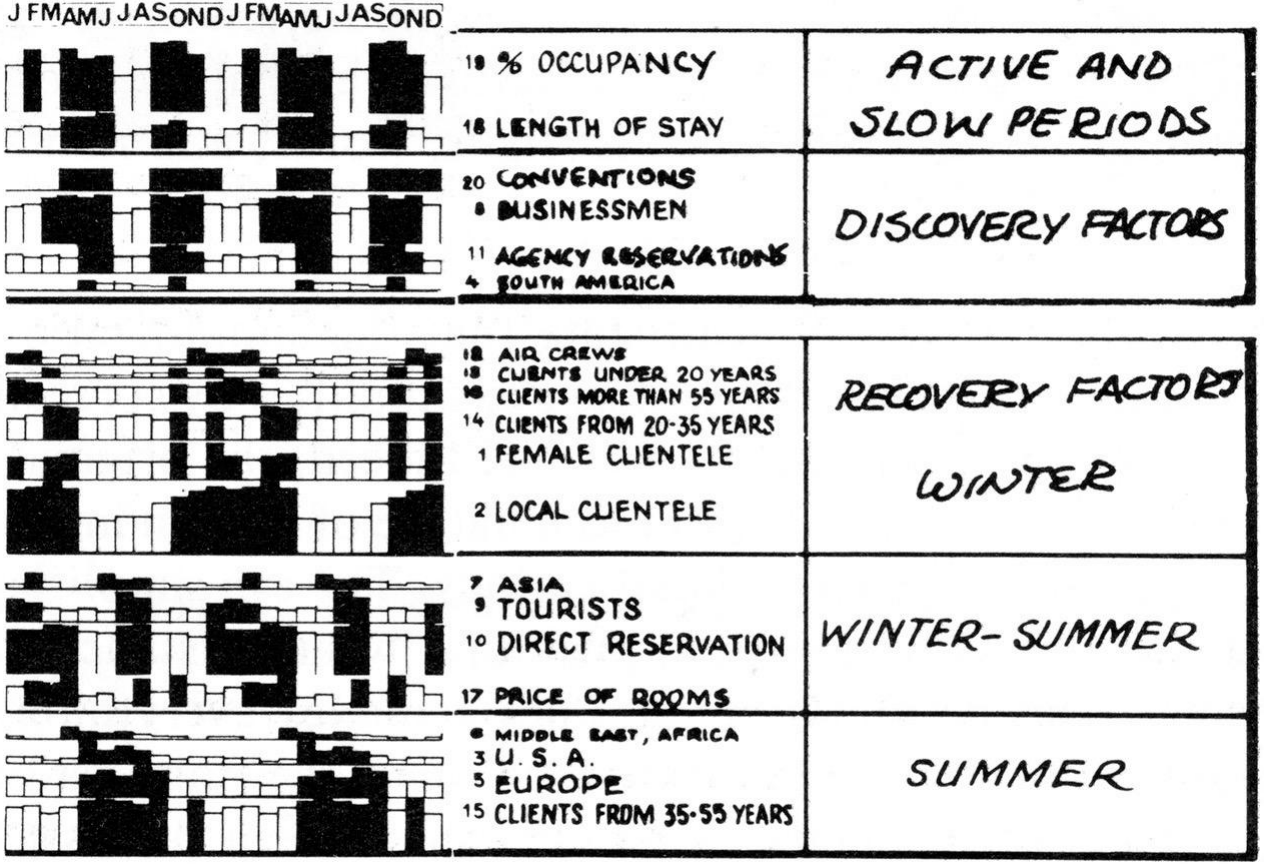


Example: Bertin's Hotel Data

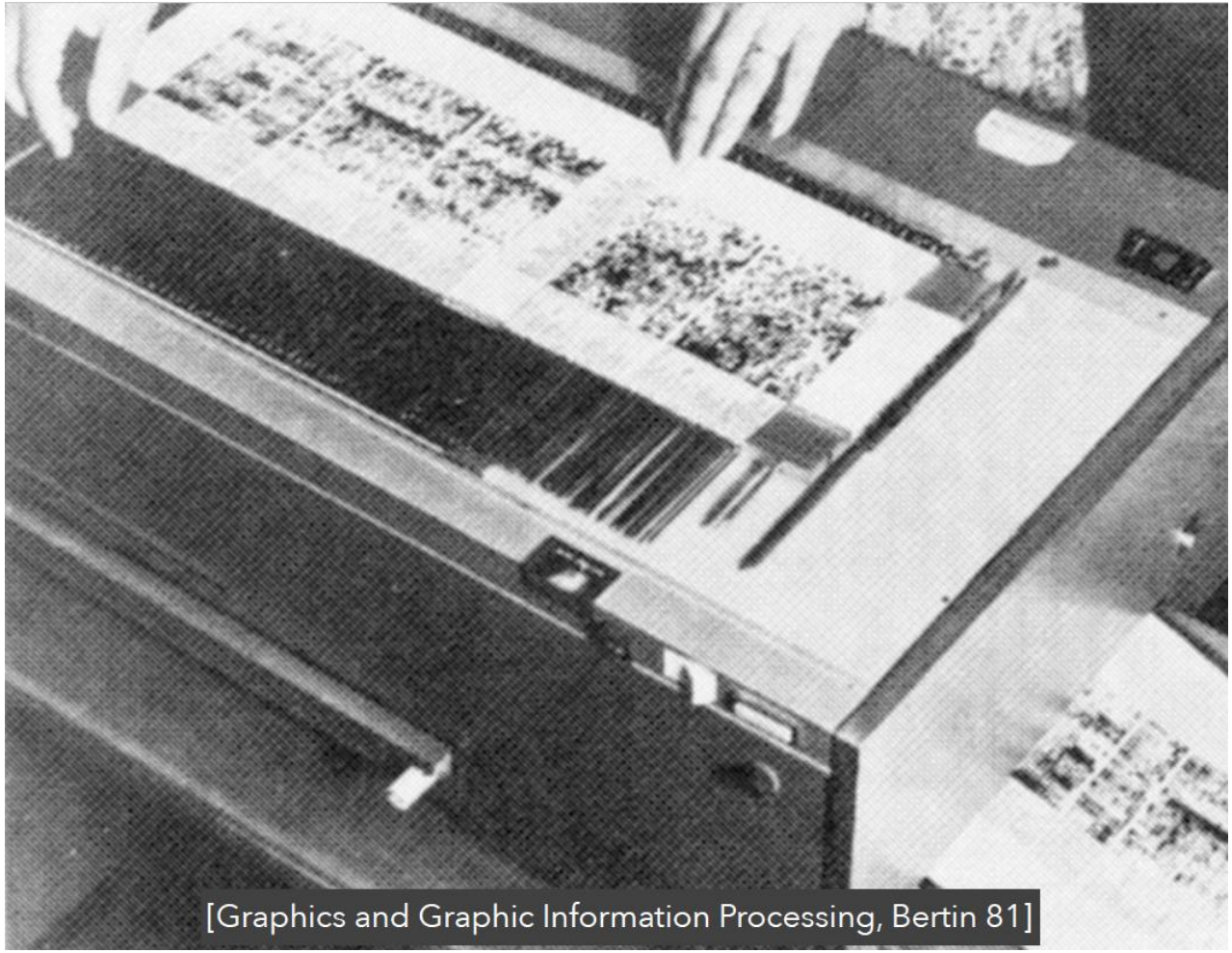
J	F	M	A	M	J	J	A	S	O	N	D		
26	21	26	28	20	20	20	20	20	40	15	40	1	% CLIENTELE FEMALE
69	70	77	71	37	36	39	39	55	60	68	72	2	% —〃— LOCAL
7	6	3	6	23	14	19	14	9	6	8	8	3	% —〃— U.S.A.
0	0	0	0	8	6	6	4	2	12	0	0	4	% —〃— SOUTH AMERICA
20	15	14	15	23	27	22	30	27	19	19	17	5	% —〃— EUROPE
1	0	0	8	6	4	6	4	2	1	0	1	6	% —〃— M.EAST, AFRICA
3	10	6	0	3	13	8	9	5	2	5	2	7	% —〃— ASIA
78	80	85	86	85	87	70	76	87	85	87	80	8	% BUSINESSMEN
22	20	15	14	15	13	30	24	13	15	13	20	9	% TOURISTS
70	70	75	74	69	68	74	75	68	68	64	75	10	% DIRECT RESERVATIONS
20	18	19	17	27	27	19	19	26	27	21	15	11	% AGENCY ———〃———
10	12	6	9	4	5	7	6	6	5	15	10	12	% AIR CREWS
2	2	4	2	2	1	1	2	2	4	2	5	13	% CLIENTS UNDER 20 YEARS
25	27	37	35	25	25	27	28	24	30	24	30	14	% —〃— 20-35 —〃—
48	49	42	48	54	55	53	51	55	46	55	43	15	% —〃— 35-55 —〃—
25	22	17	15	19	19	19	19	19	20	19	22	16	% —〃— MORE THAN 55 —〃—
163	167	166	174	152	155	145	170	157	174	165	156	17	PRICE OF ROOMS
1.65	1.71	1.65	1.91	1.90	2.	1.54	1.60	1.73	1.82	1.66	1.44	18	LENGTH OF STAY
67	82	70	83	74	77	56	62	90	92	78	55	19	% OCCUPANCY
			X	X	X			X	X	X	X	20	CONVENTIONS


J FMAMJ JASOND J FMAMJ JASOND


1
J FMAMJ JASONDJ FMAMJ JASOND
2
J FMAMJ JASONDJ FMAMJ JASOND
3
4
5


J FMAMJ JASONDJ FMAMJ JASOND
18 % OCCUPANCY
18 LENGTH OF STAY
ACTIVE AND SLOW PERIODS
20 CONVENTIONS
8 BUSINESSMEN
11 AGENCY RESERVATIONS
4 SOUTH AMERICA
DISCOVERY FACTORS
18 AIR CREWS
13 CLIENTS UNDER 20 YEARS
16 CLIENTS MORE THAN 55 YEARS
14 CLIENTS FROM 20-35 YEARS
1 FEMALE CLIENTELE
2 LOCAL CLIENTELE
RECOVERY FACTORS WINTER
7 ASIA
9 TOURISTS
10 DIRECT RESERVATION
17 PRICE OF ROOMS
WINTER-SUMMER
6 MIDDLE EAST, AFRICA
3 U.S.A.
5 EUROPE
15 CLIENTS FROM 35-55 YEARS
SUMMER

[Graphics and Graphic Information Processing, Bertin 81]



[Graphics and Graphic Information Processing, Bertin 81]

[Graphics and Graphic Information Processing, Bertin 81]

Example: Bertin's Hotel Data

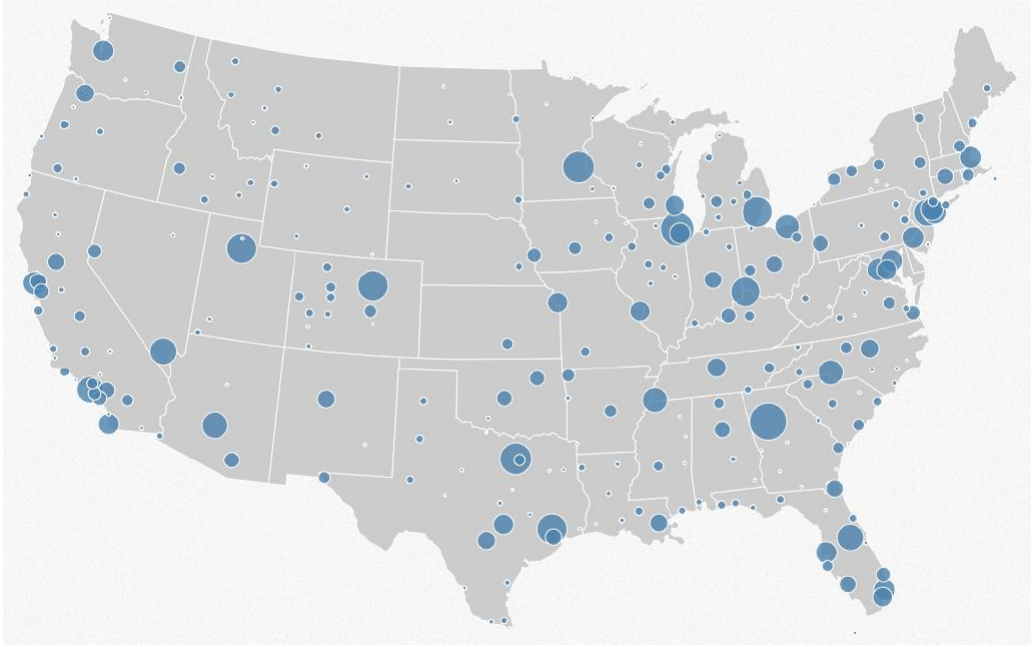
- **Most visualizations are interactive**
  Even passive media elicit interactions
- **Good visualizations are task dependent**
  Pick the right interaction technique
  Consider the semantics of the data domain
- **Fundamental interaction techniques**
  Selection / Annotation, Sorting, Navigation,
  Brushing & Linking, Dynamic Queries

Basic Selection Methods

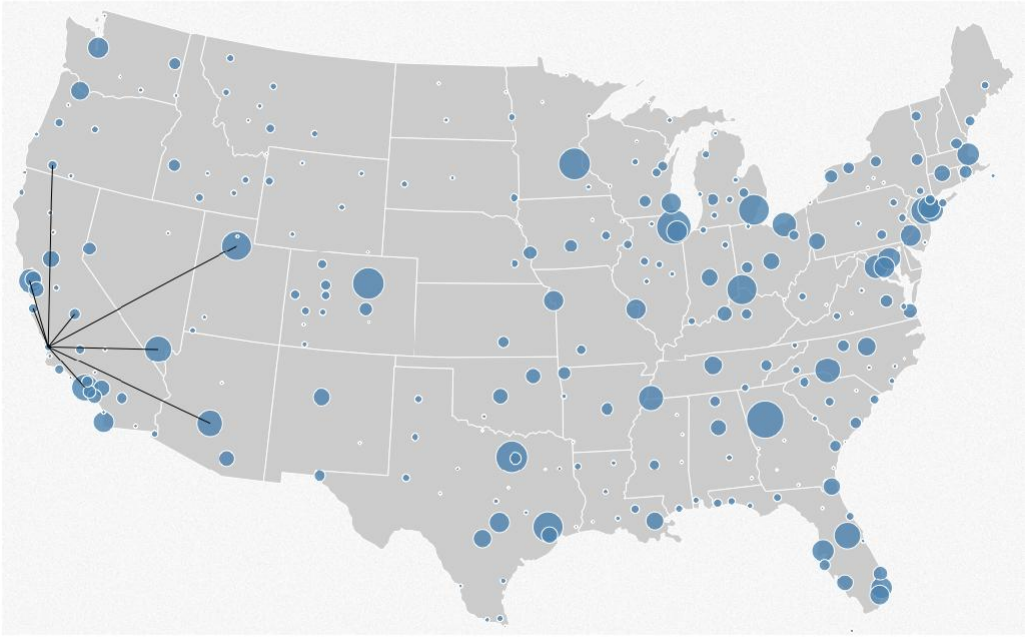
Point Selection

- Mouse Hover / Click
- Touch / Tap
- Select Nearby Element (e.g. Bubble Cursor)

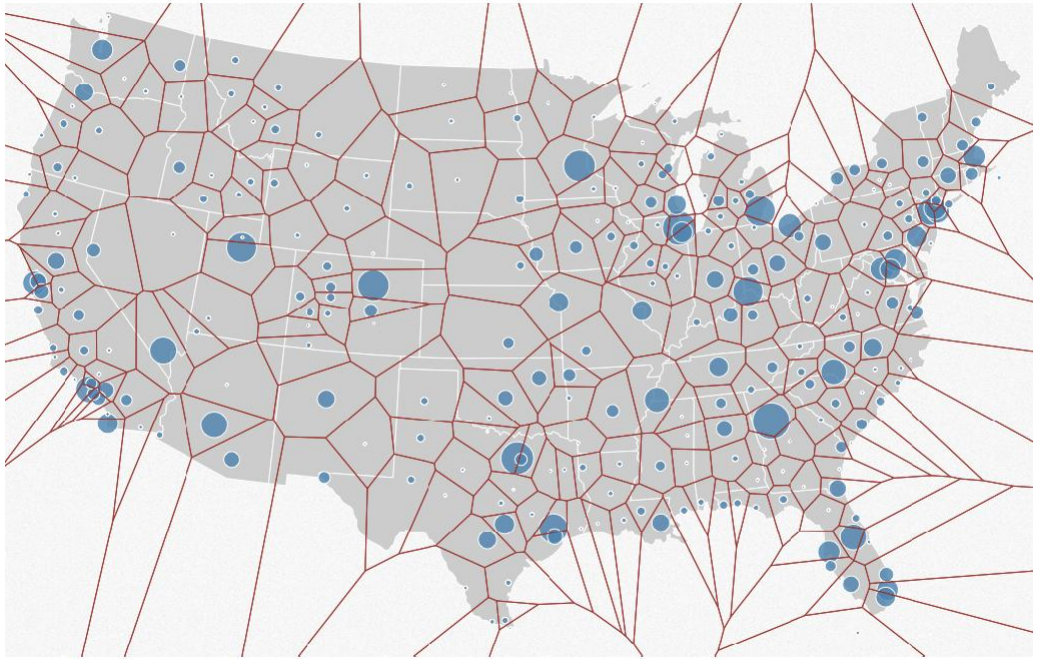
Point Selection



Point Selection



Point Selection



Basic Selection Methods

Point Selection

- Mouse Hover / Click
- Touch / Tap
- Select Nearby Element (e.g. Bubble Cursor)

Region Selection

- Rubber-band or Lasso
- Area Cursors (“Brushes”)

Brushing and Linking

Brushing and Linking

Select (“brush”) a subset of data

See selected data in other views

The components must be linked

- by tuple (matching data points), or
- by query (matching range or values)

Brushing and Linking

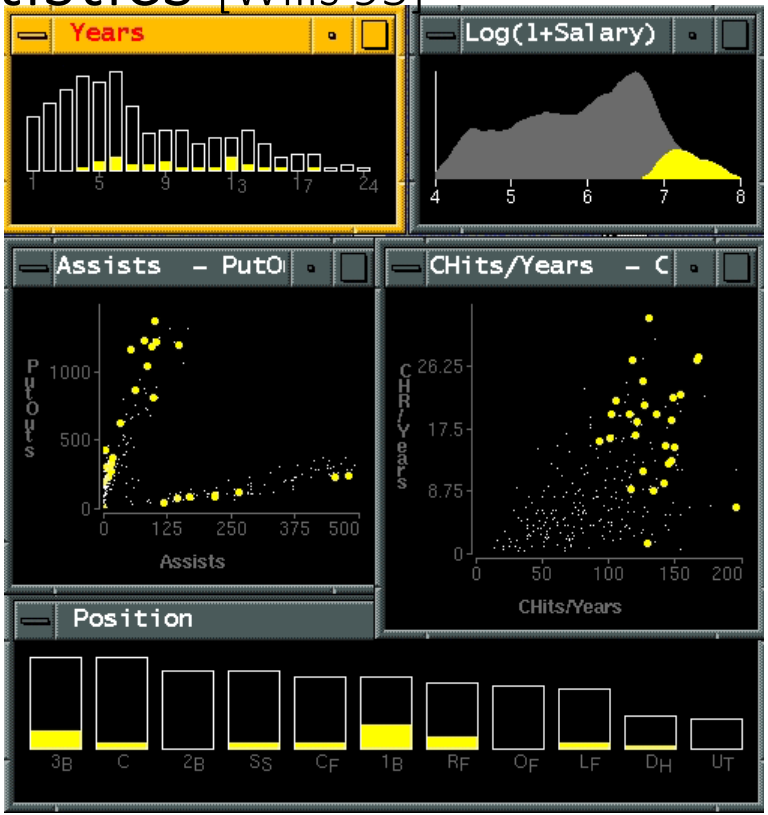
Linked Highlighting

- See how regions contiguous in one view are distributed within another – powerful & pervasive

The components must be linked

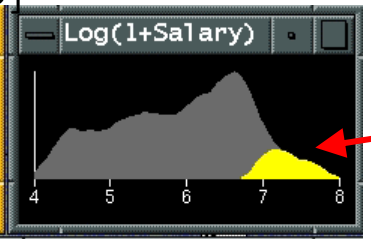
- by tuple (matching data points), or
- by query (matching range or values)

Baseball Statistics [Wills 95]



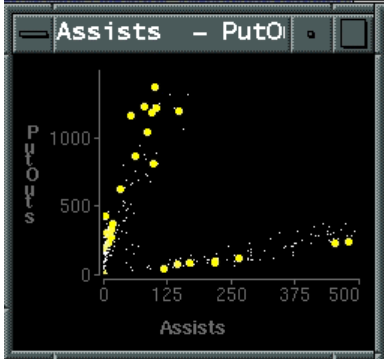
Baseball Statistics [Wills 95]

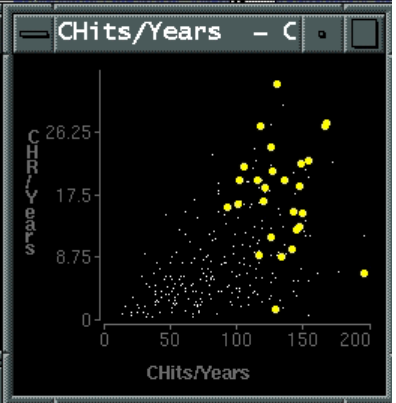
how long in majors



select high salaries

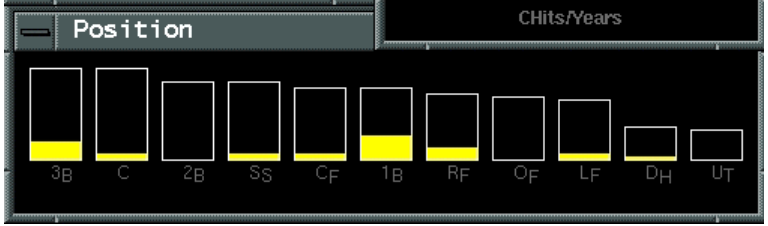
avg assists vs avg putouts (fielding ability)





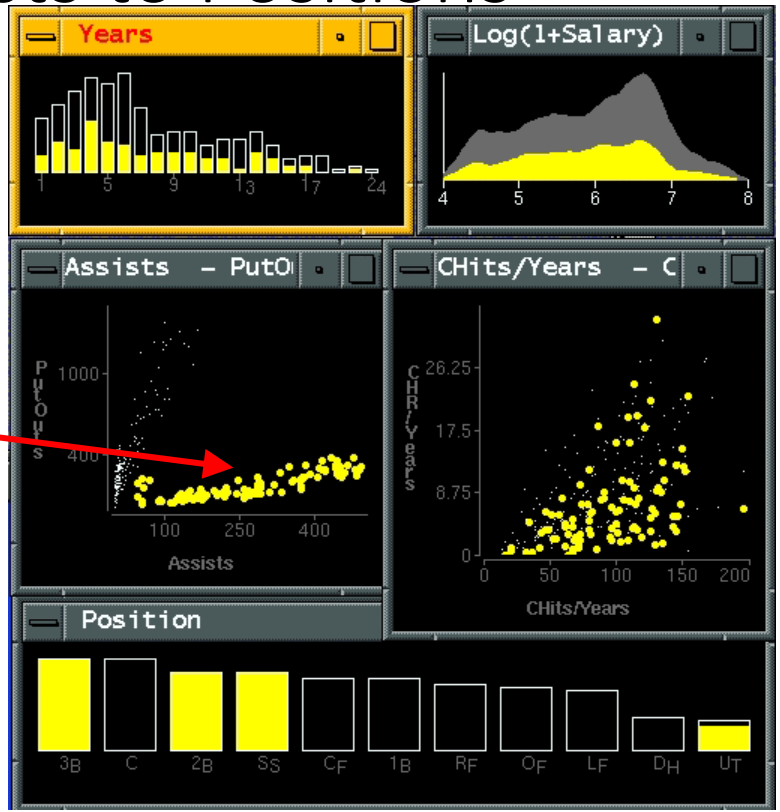
avg career HRs vs avg career hits (batting ability)

distribution of positions played

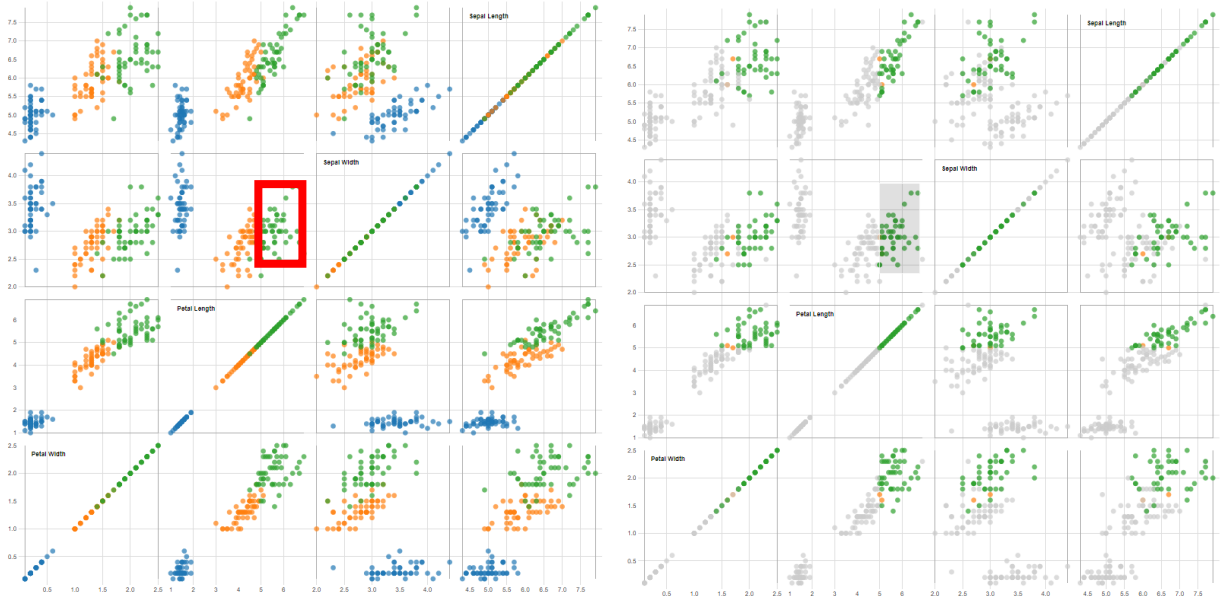


Linking Assists to Positions

select
assists



Brushing Scatterplots



Dynamic Queries

Interaction between people and

machines requires *mutual intelligibility*

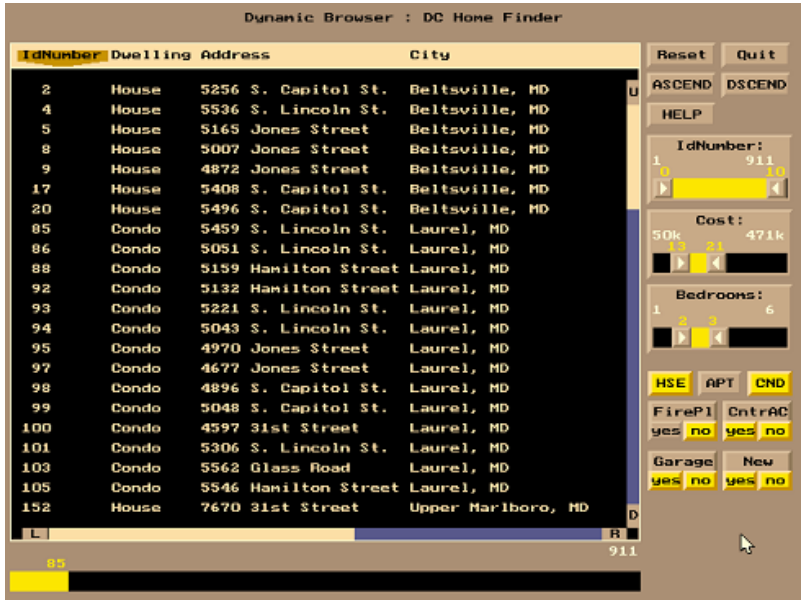
or *shared understanding.*

Query & Results

SELECT house FROM seattle_homes

WHERE price < 1,000,000 AND bedrooms > 2

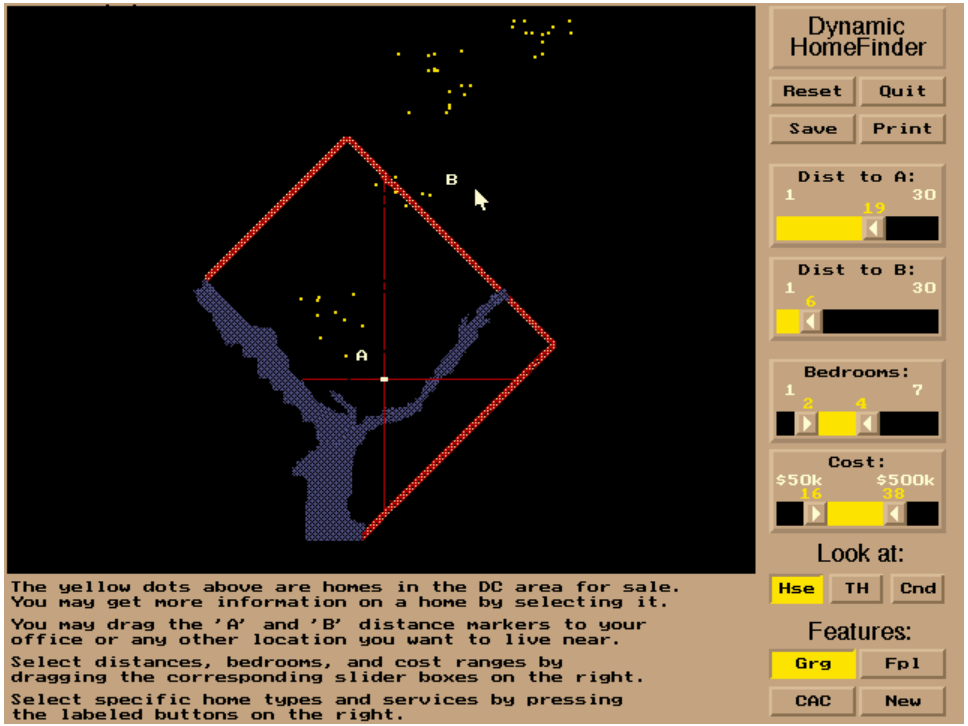
ORDER BY price



Issues with Textual Queries

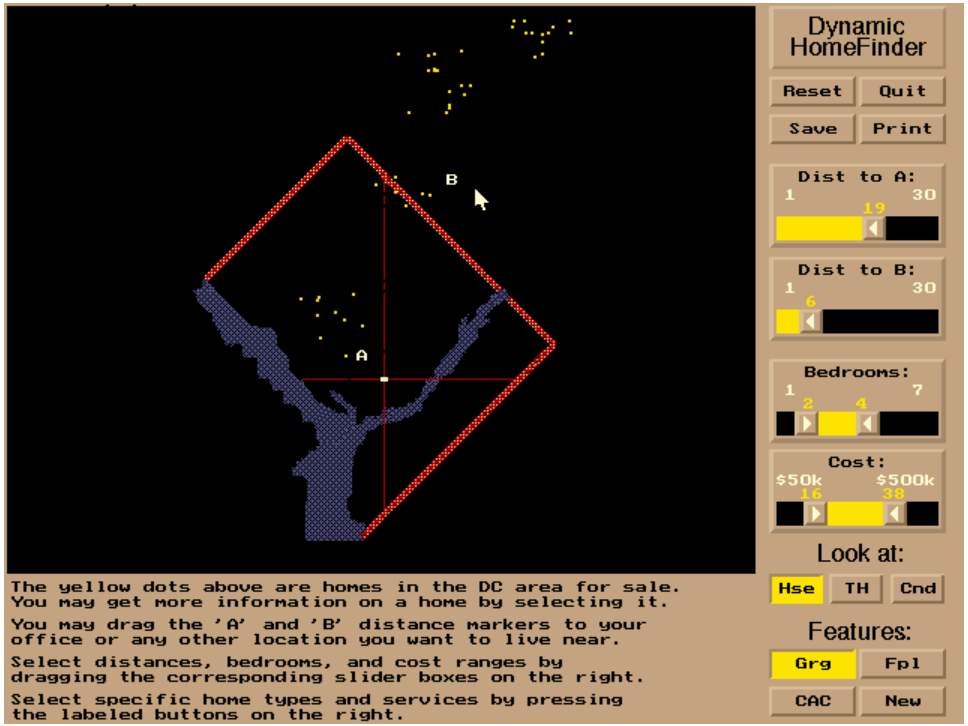
1. For programmers
2. Rigid syntax
3. Only shows exact matches
4. Too few or too many hits
5. No hint on how to reformulate the query
6. Slow question-answer loop
7. Results returned as table

HomeFinder



[Williamson and Shneiderman 92]

HomeFinder



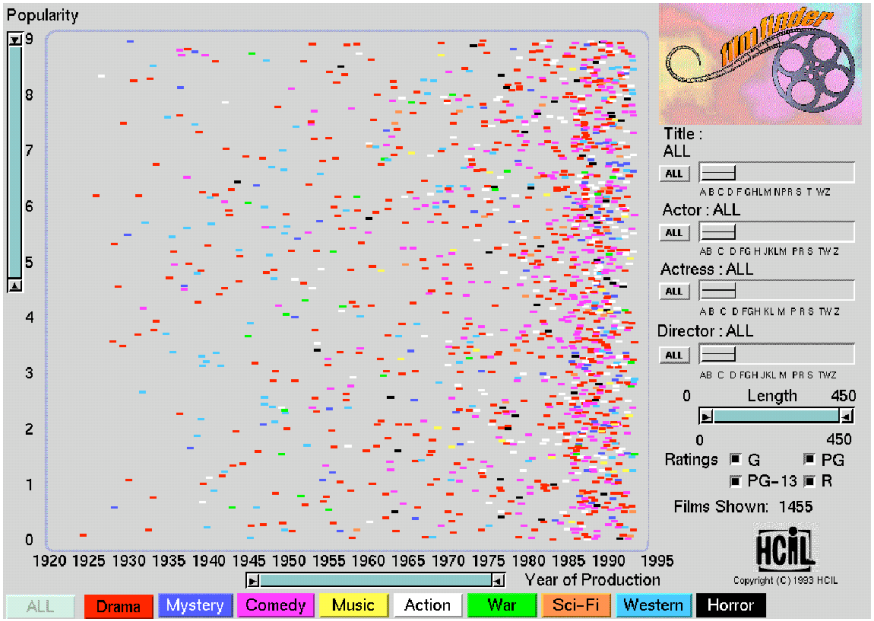
[Williamson and Shneiderman 92]

Direct Manipulation

1. Visual representation of objects and actions
2. Rapid, incremental and reversible actions
3. Selection by pointing (not typing)
4. Immediate and continuous display of results

Product Popularity

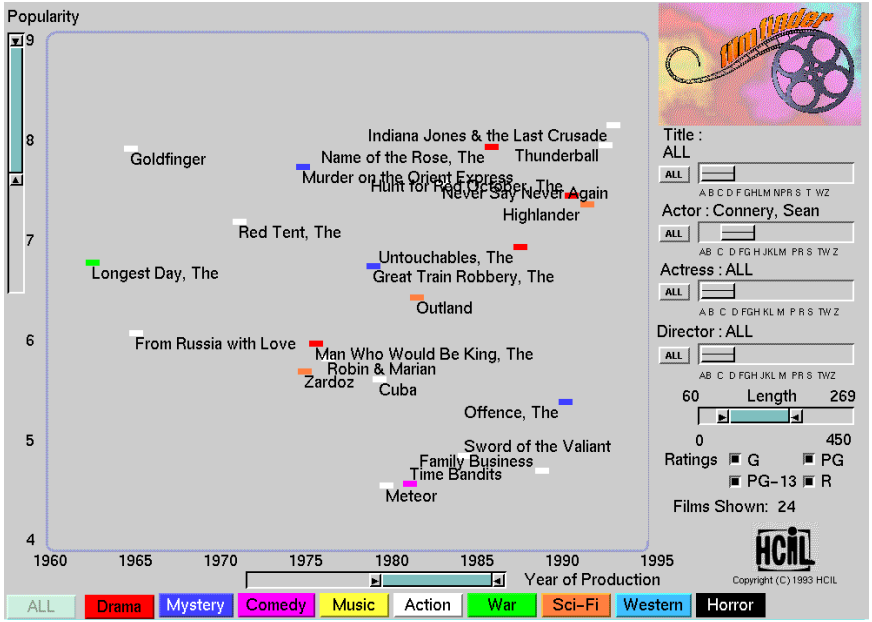
- Item Filtering
- Browse through tightly coupled interaction



[Ahlberg and Shneiderman 94]

Product Popularity

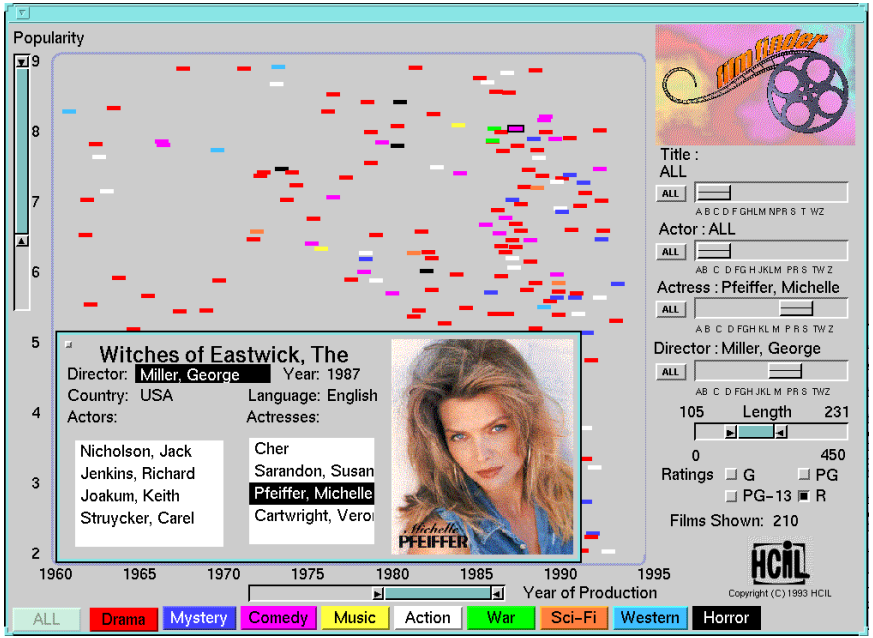
- Item Filtering
- Browse through tightly coupled interaction



[Ahlberg and Shneiderman 94]

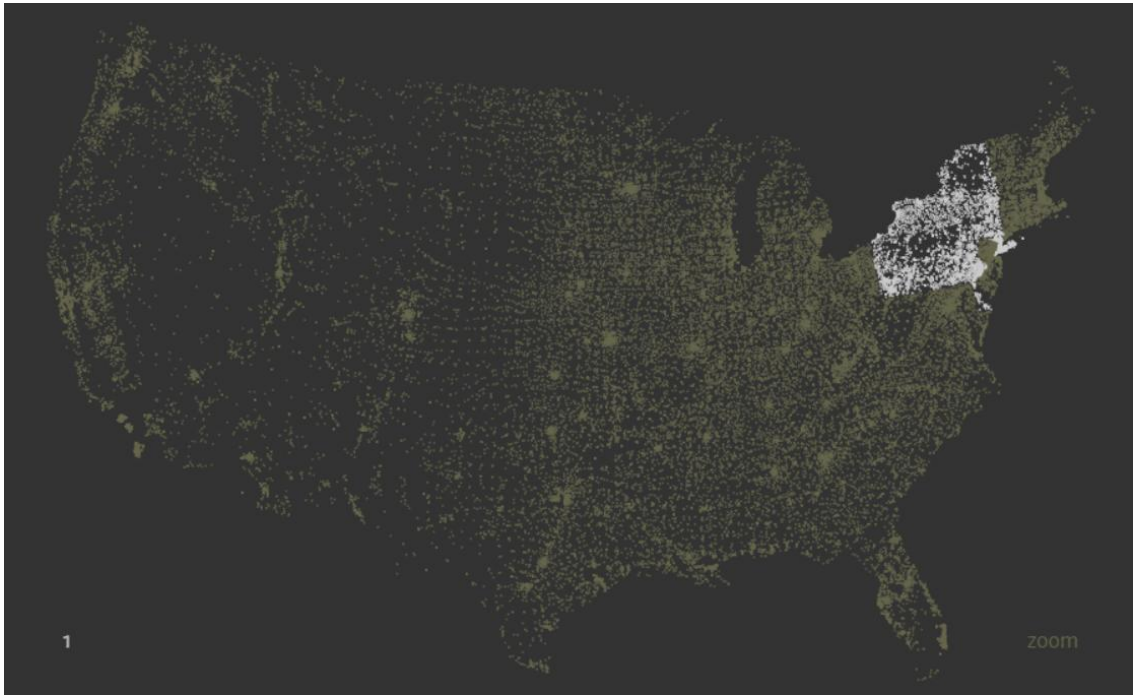
Product Popularity

- Item Filtering
- Browse through tightly coupled interaction



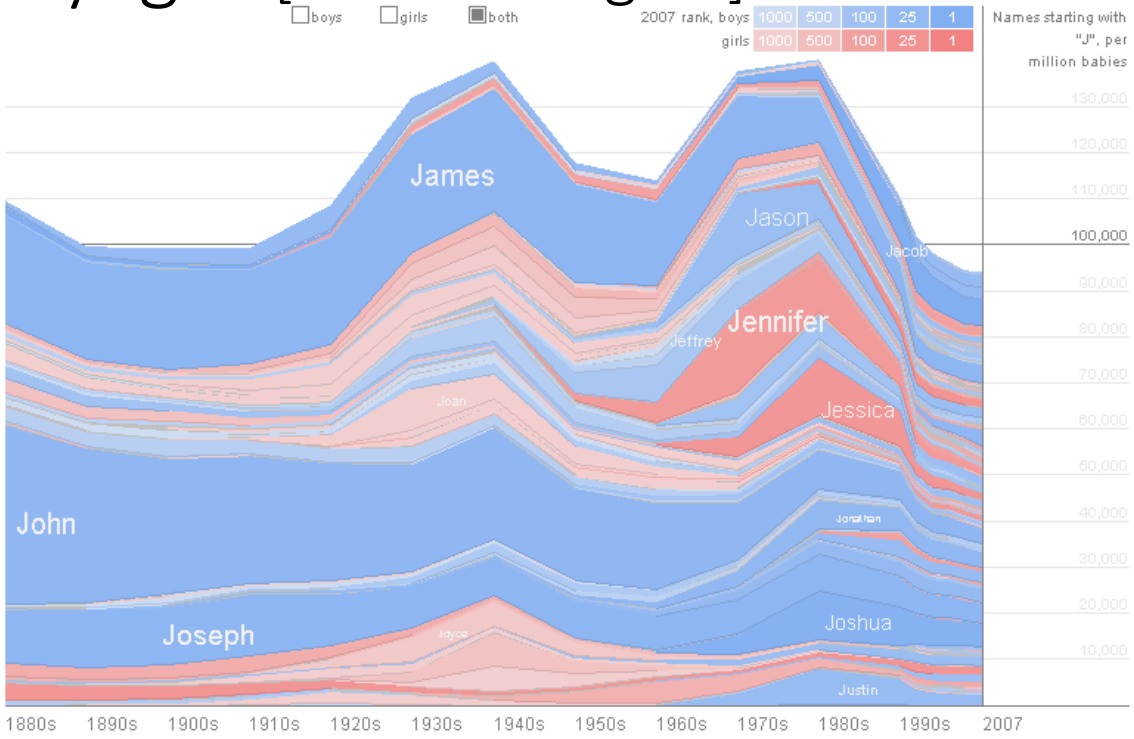
[Ahlberg and Shneiderman 94]

Zipdecode [Fry 04]

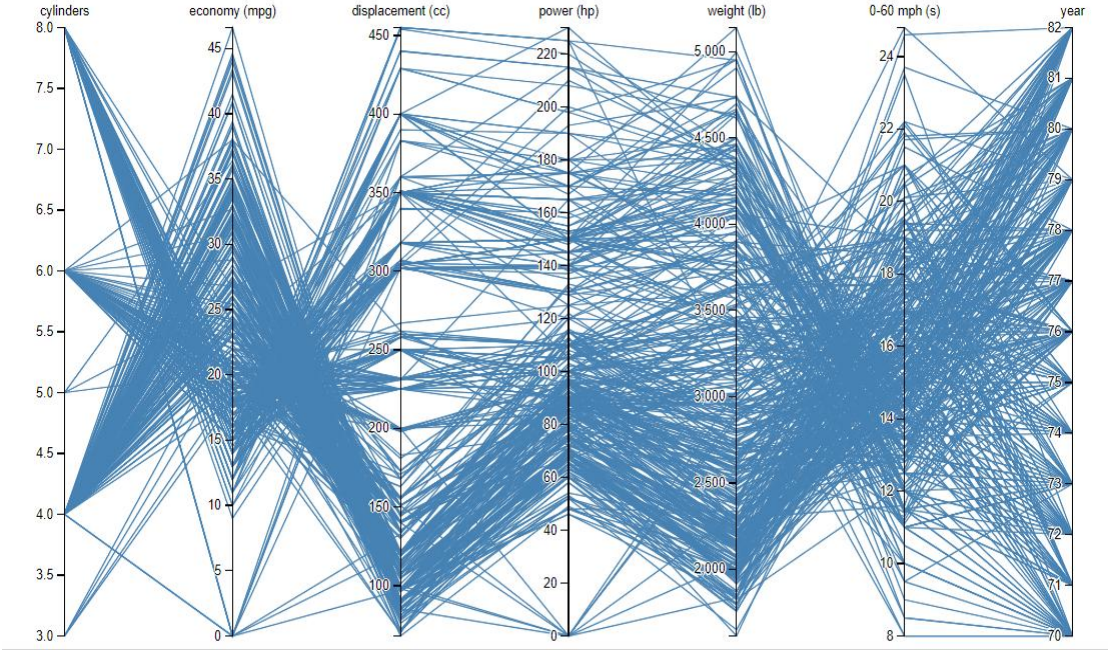


http://benfry.com/zipdecode/

NameVoyager [Wattenberg 06]



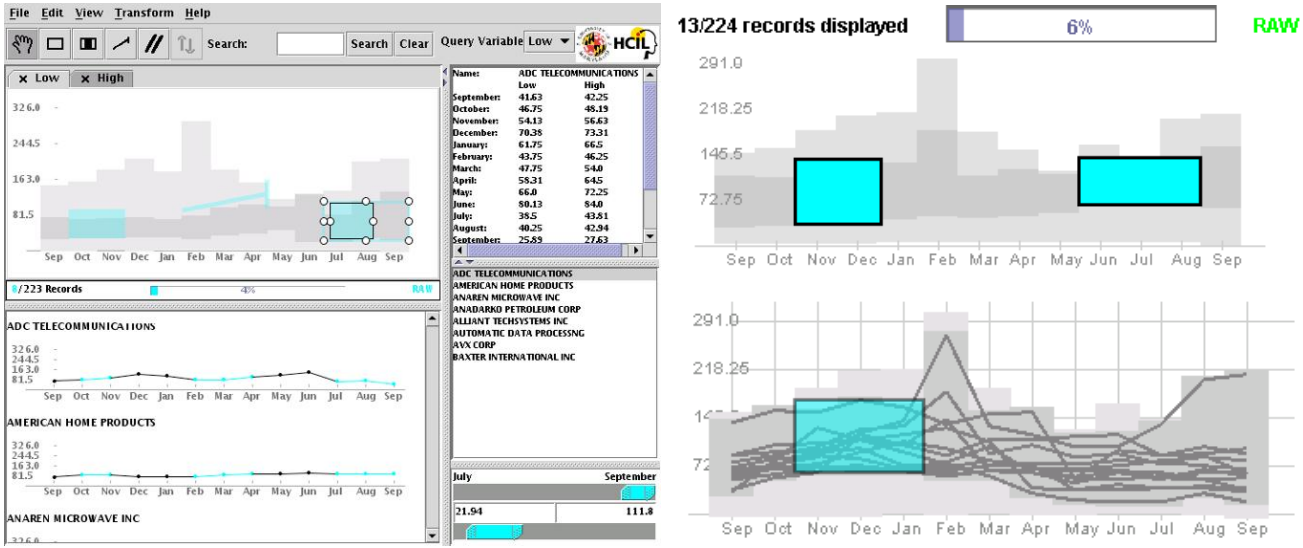
Parallel Coordinates [Mike Bostock]



Modified by John Davies to include re-orderable axes.

https://bl.ocks.org/jasondavies/1341281

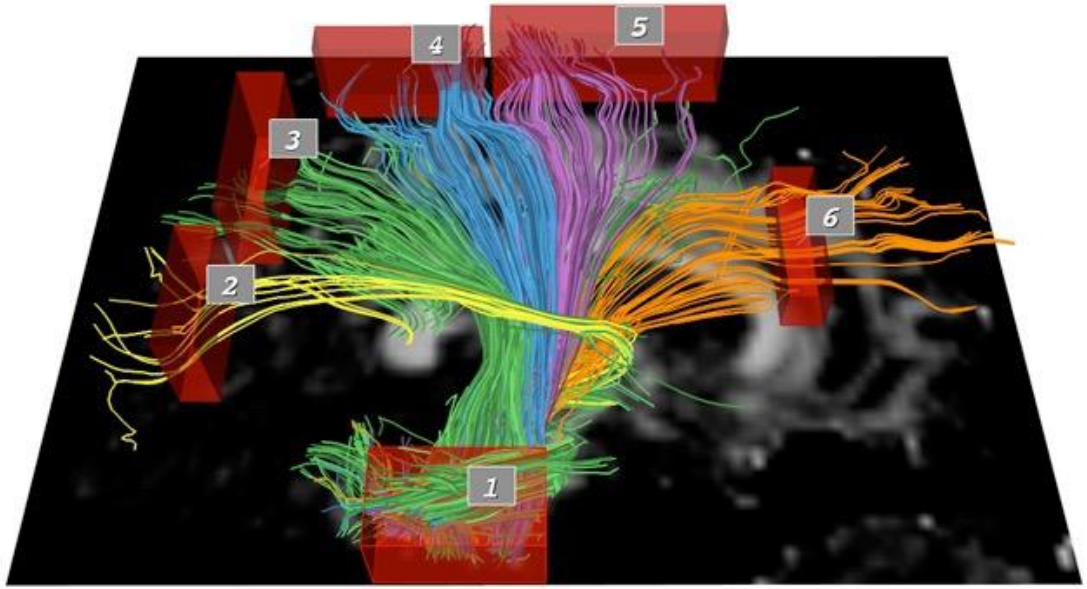
TimeSearcher [Hocheiser 02]



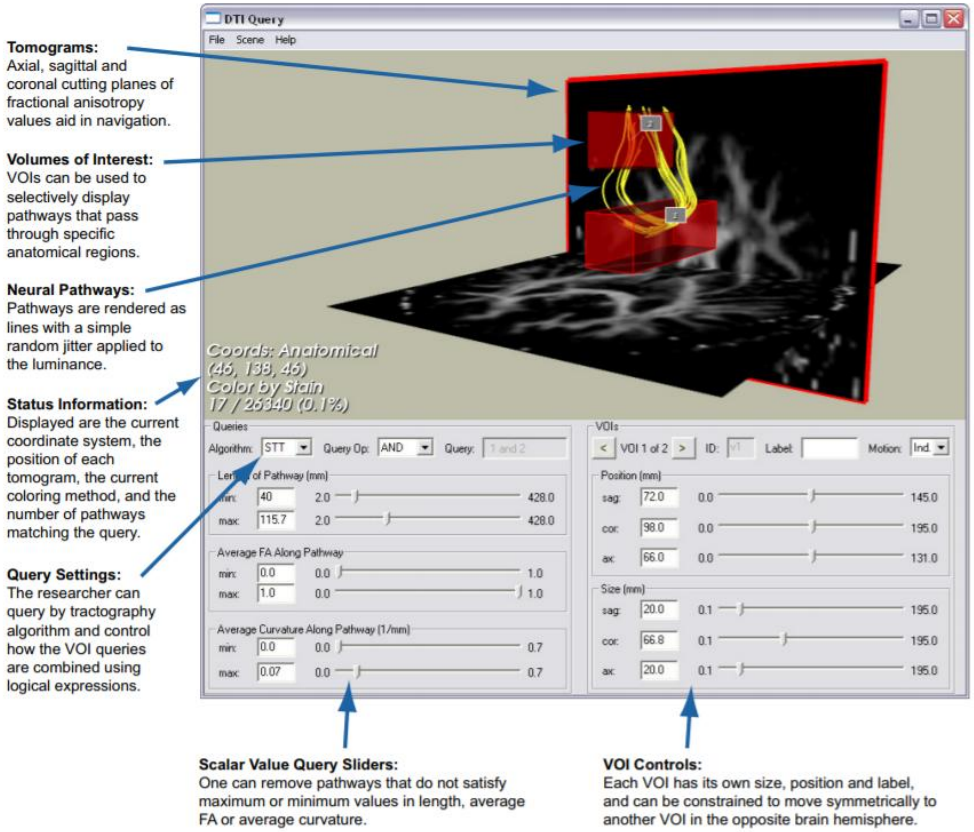
Builds on Wattenberg's [2001] idea for sketch-based queries of time-series data.

3D Dynamic Queries [Akers 04]

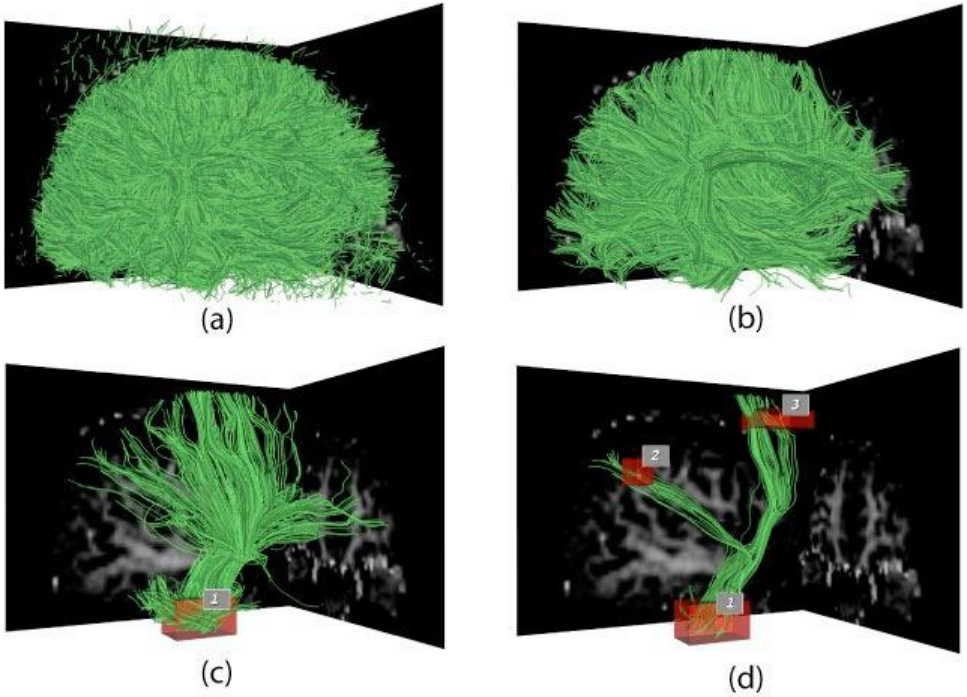
- Uses dynamic queries to find structure in neural pathways suggested by MR tractography.



3D Dynamic Queries [Akers 04]



3D Dynamic Queries [Akers 04]



Pros & Cons

Pros

- Controls useful for both novices and experts
- Quick way to explore data

Cons

- Simple queries
- Lots of controls
- Amount of data shown limited by screen space

Who would use these kinds of tools?

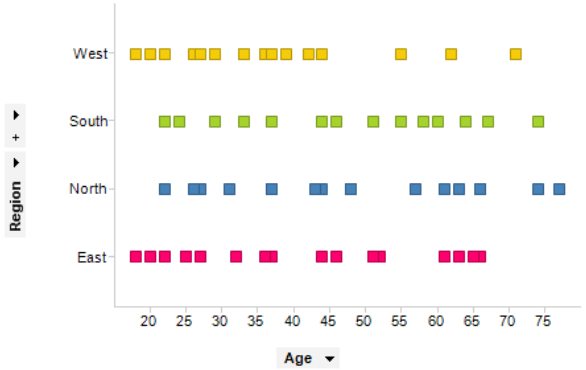
Sorting

Trellis Display [Becker, Cleveland, and Shyu 96]

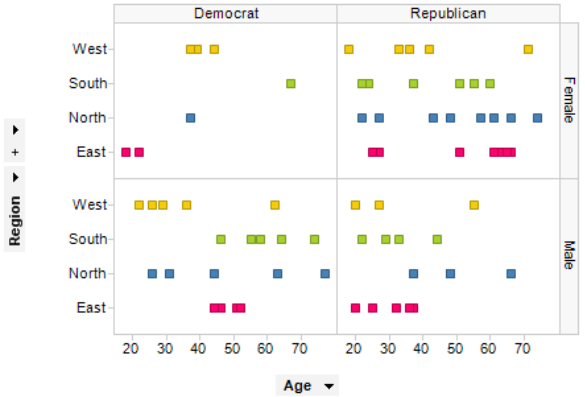
- Trellis display is a framework for the visualization of data.
- One important application is **uncovering the relationships of variables** in multivariable data sets
- Trellised visualizations enable you to quickly recognize similarities or differences between different categories in the data. Each individual panel in a trellis visualization displays a subset of the original data table, where the subsets are defined by the categories available in a column or hierarchy.

Trellis Display [Becker, Cleveland, and Shyu 96]

Standard non-trellised scatter plot:

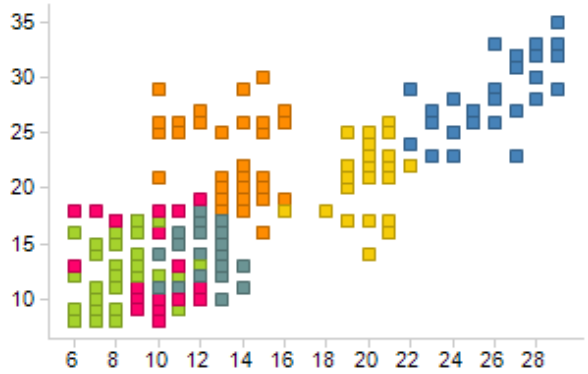


By trellising this visualization based on Gender and Political affiliation, it will look like this:

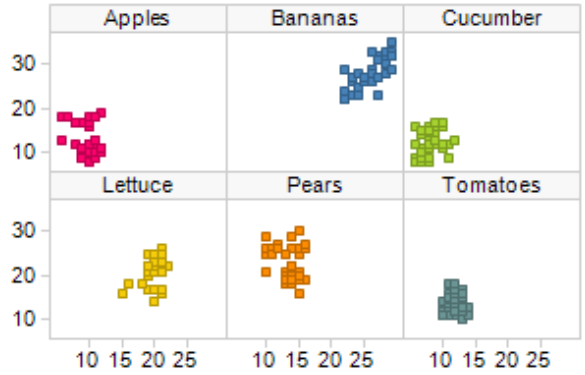


Trellis Display [Becker, Cleveland, and Shyu 96]

In the scatter plot below, sales (Y) is plotted against cost (X) for a number of different products (colored by product), to display a low positive correlation:



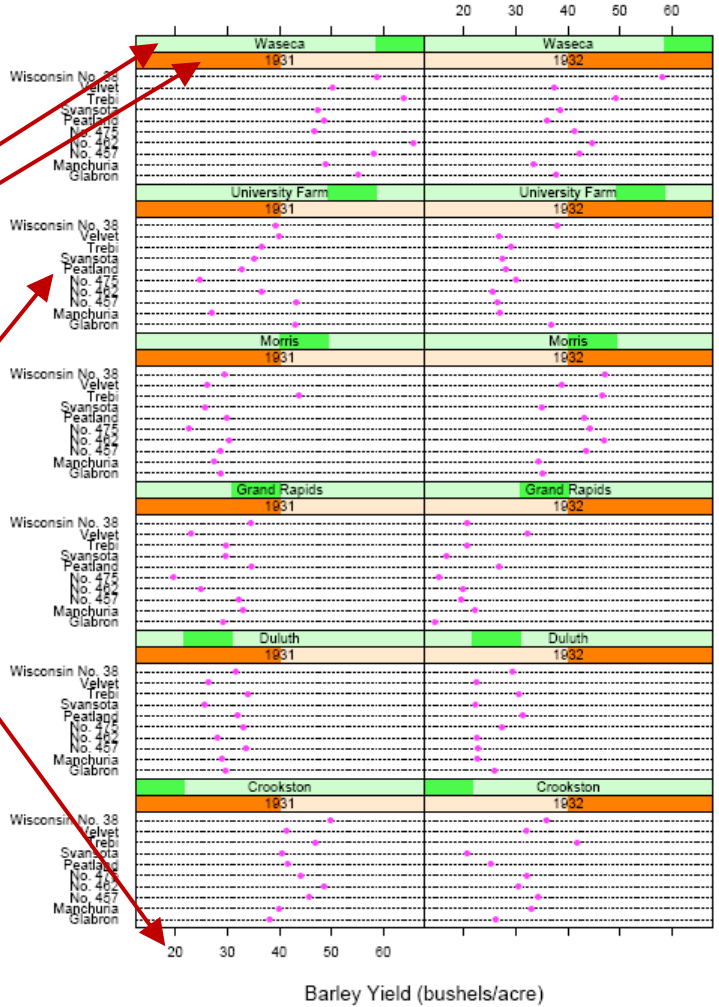
Each product can be shown separately using trellising:



https://docs.tibco.com/pub/spotfire/6.5.3/doc/html/vis/vis_trellis_visualizations.htm

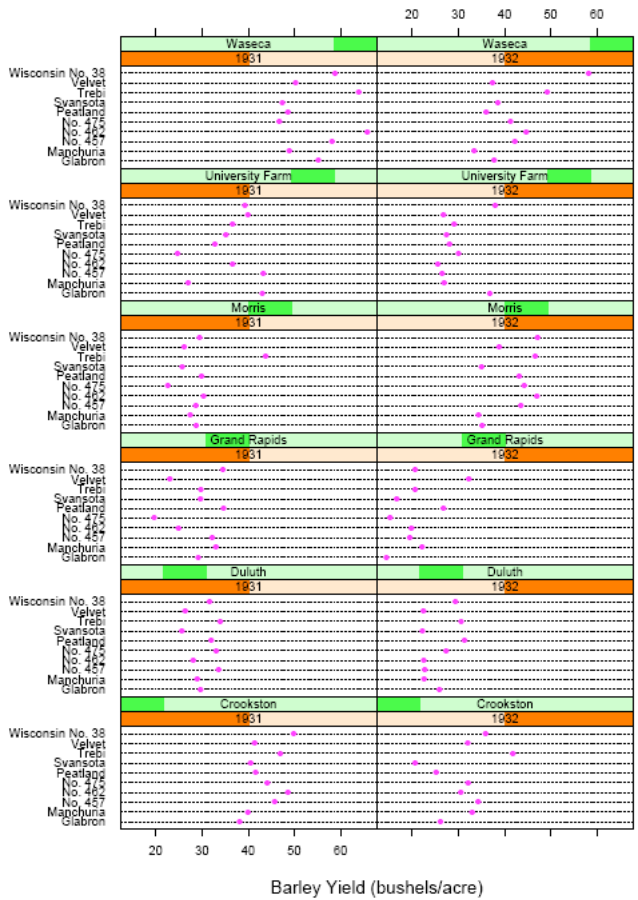
Condition variables
location, year

Panel variables
type, yield

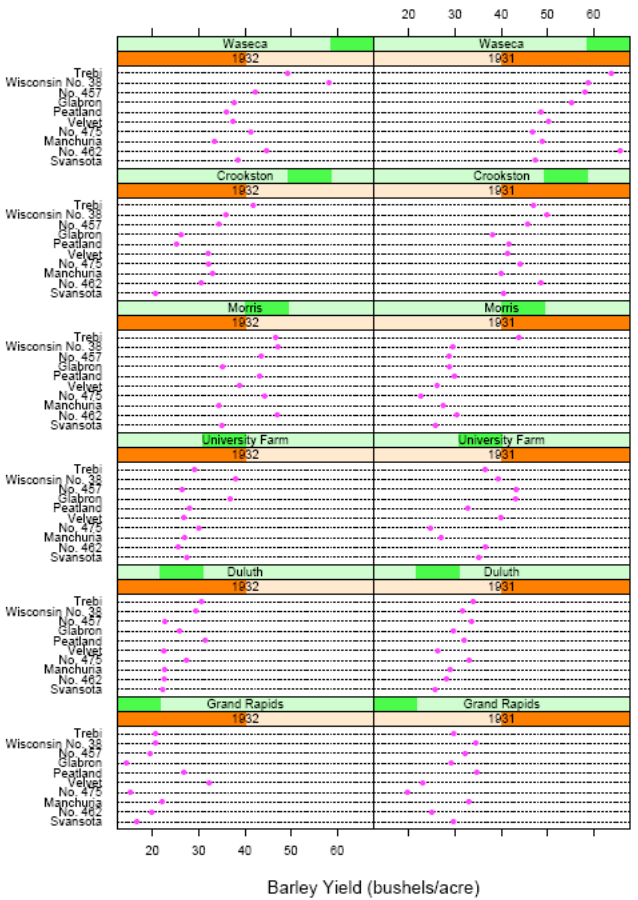


Trellis Display

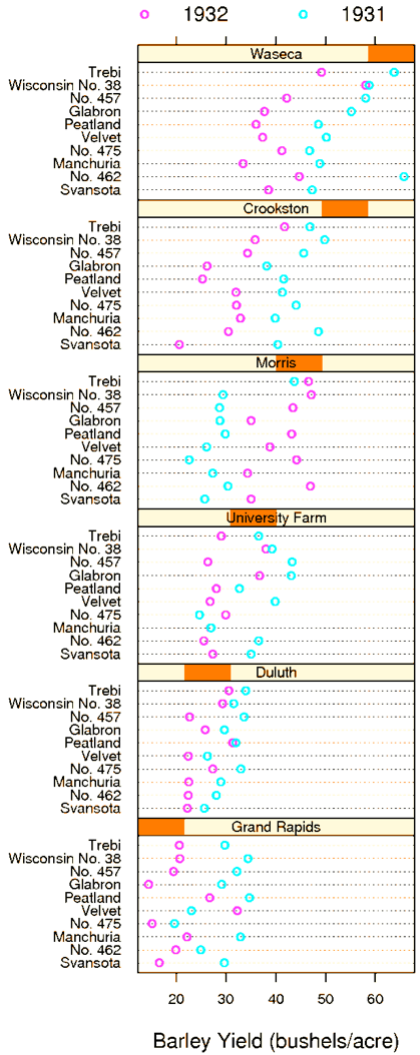
[Becker, Cleveland, and Shyu 96]



Alphabetical ordering



Main-effects ordering


1932
1931
Waseca
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
Crookston
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
Morris
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
University Farm
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
Duluth
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
Grand Rapids
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
20
30
40
50
60
Barley Yield (bushels/acre)

Recap

- **Most visualizations are interactive**

  Even passive media elicit interactions

- **Good visualizations are task dependent**

  Pick the right interaction technique

  Consider the semantics of the data domain

- **Fundamental interaction techniques**

  Selection / Annotation, Sorting, Navigation,

  Brushing & Linking, Dynamic Queries

imMens

[Liu, Jiang & Heer ’13]

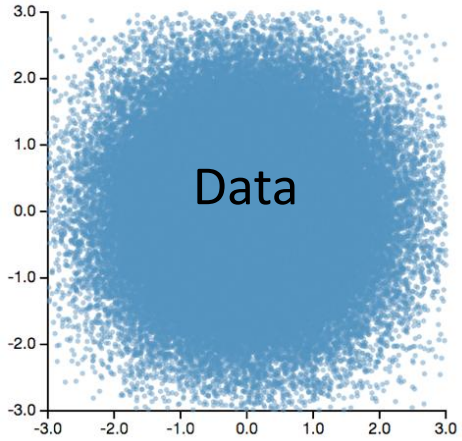
http://vis.stanford.edu/projects/immens/

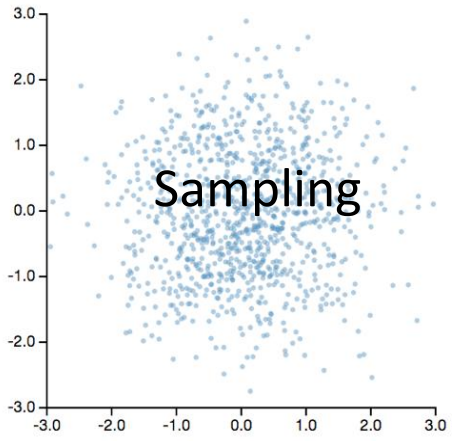
How can we visualize and interact with **billion+ record** databases in real-time?

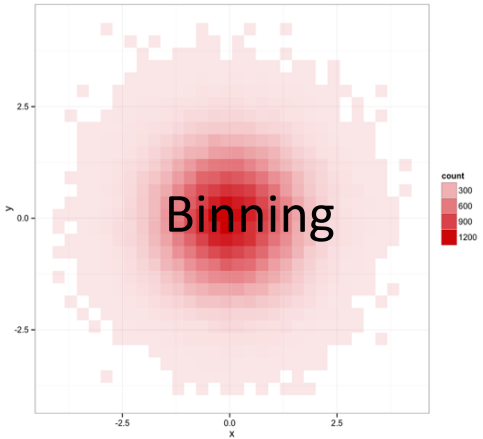
Two Challenges:

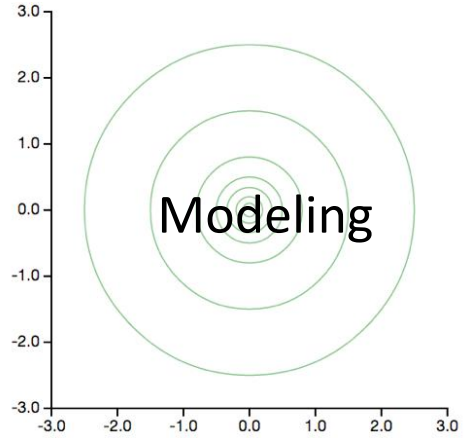
1. Effective **visual encoding**
2. Real-time **interaction**

Perceptual and interactive scalability should be limited by the **chosen resolution of** the visualized data, not the number of records.

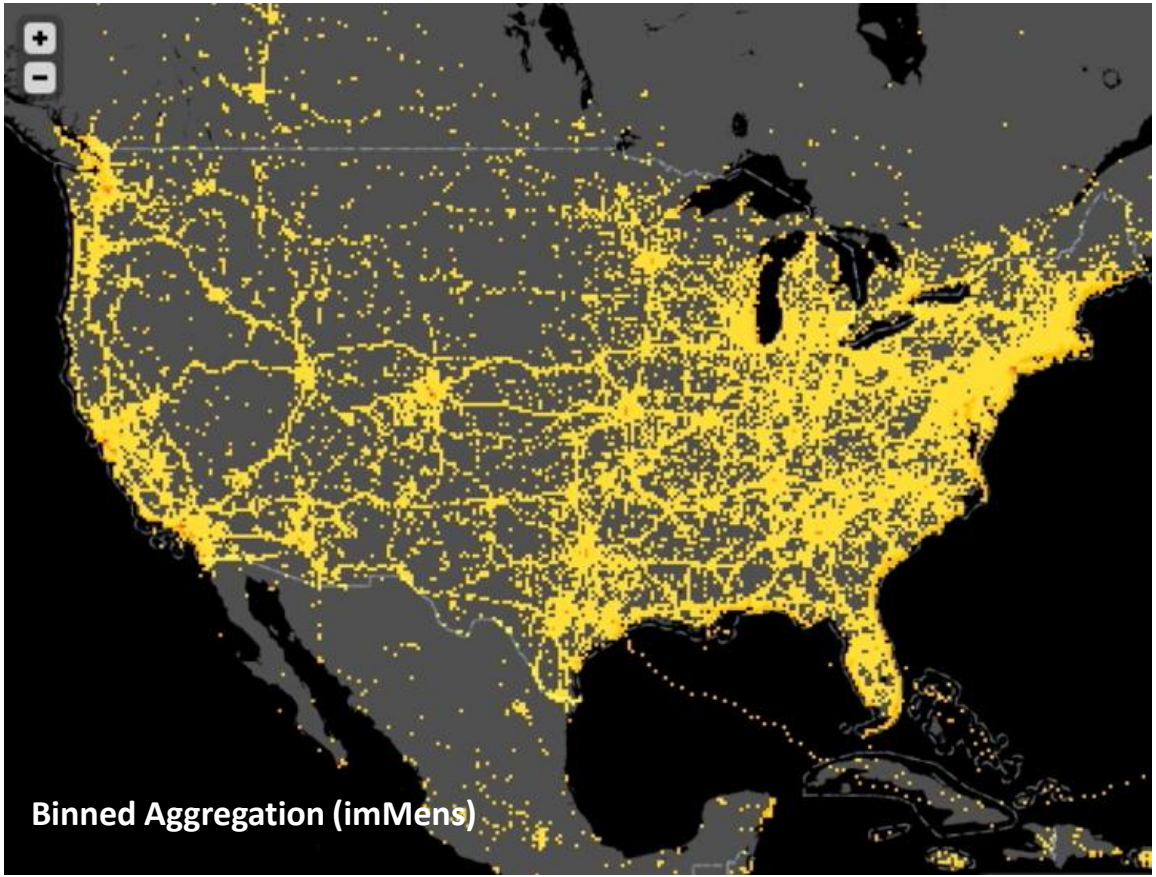

Data
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-3.0 -2.0 -1.0 0.0 1.0 2.0 3.0

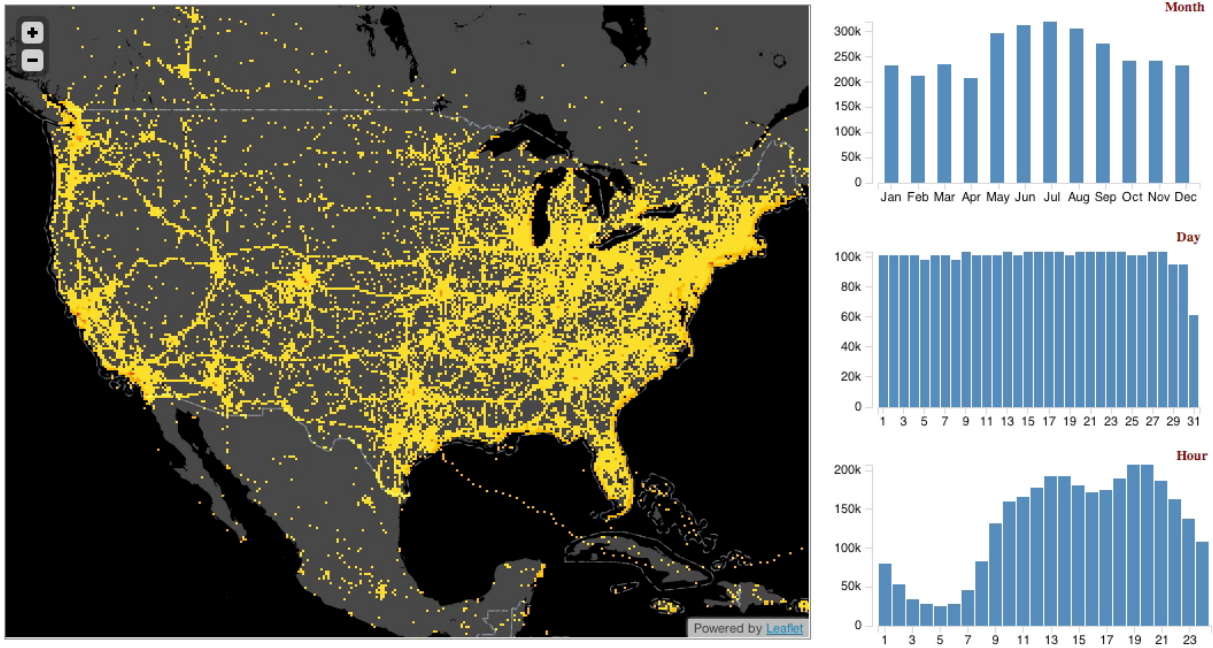

Sampling
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-3.0 -2.0 -1.0 0.0 1.0 2.0 3.0


Binning
count
300
600
900
1200
2.5
0.0
-2.5
y
-2.5 0.0 2.5
x


Modeling
3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
-3.0 -2.0 -1.0 0.0 1.0 2.0 3.0

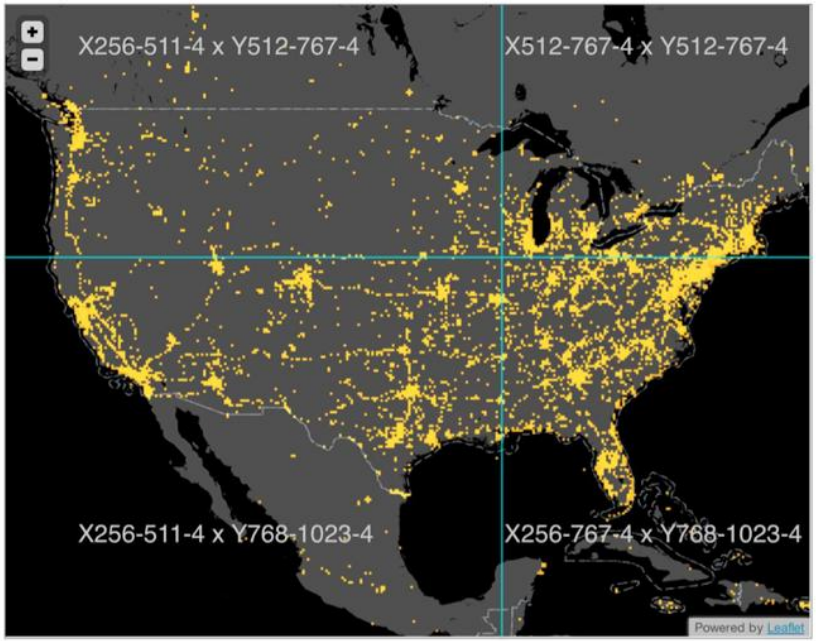
Stratified Sampling (Google Fusion Tables)

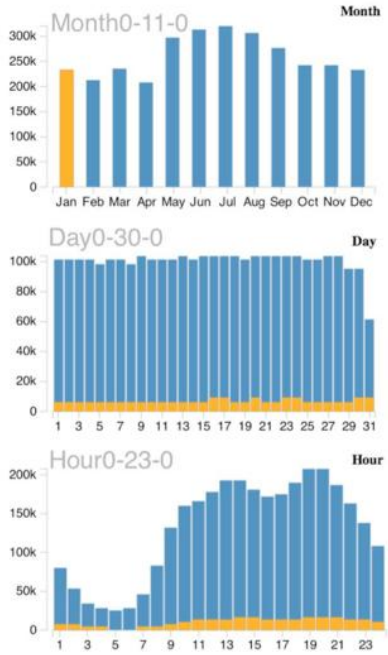

Binned Aggregation (imMens)

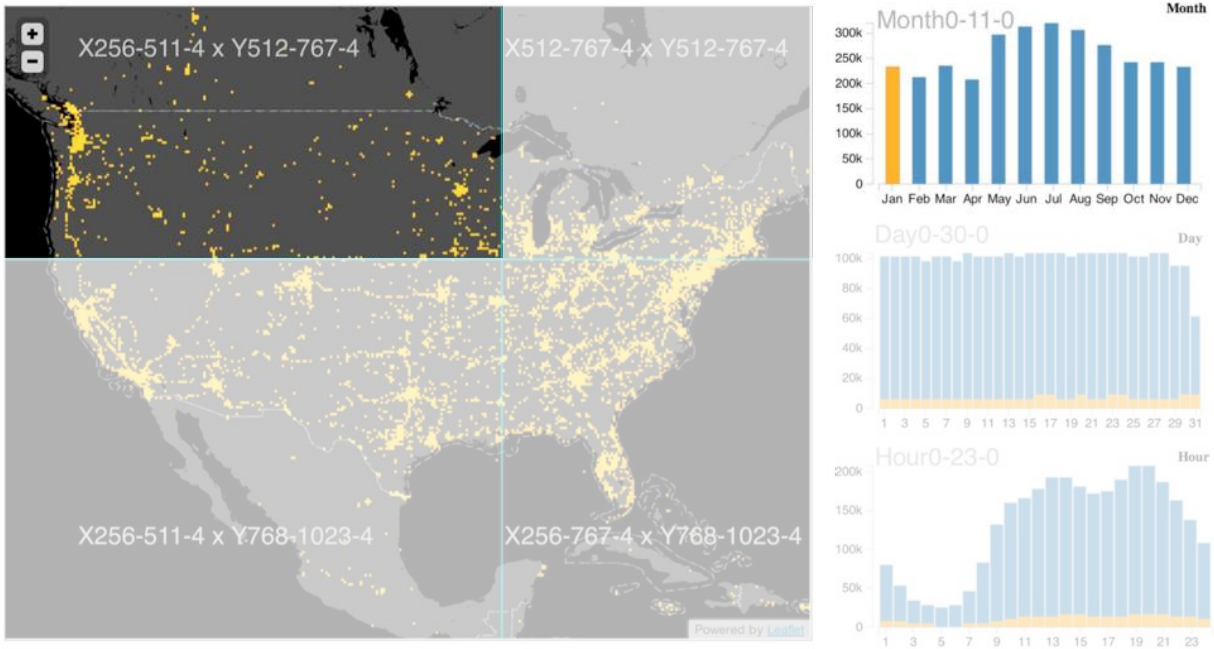


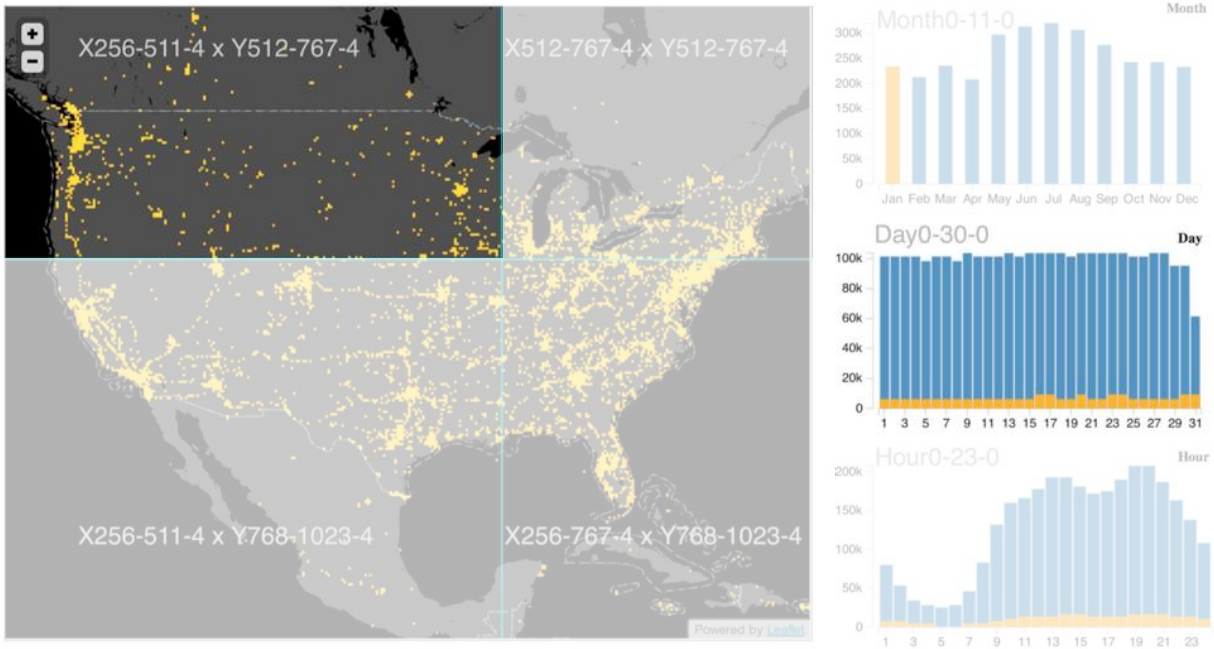
imMens: Real-Time Visual Querying of Big Data

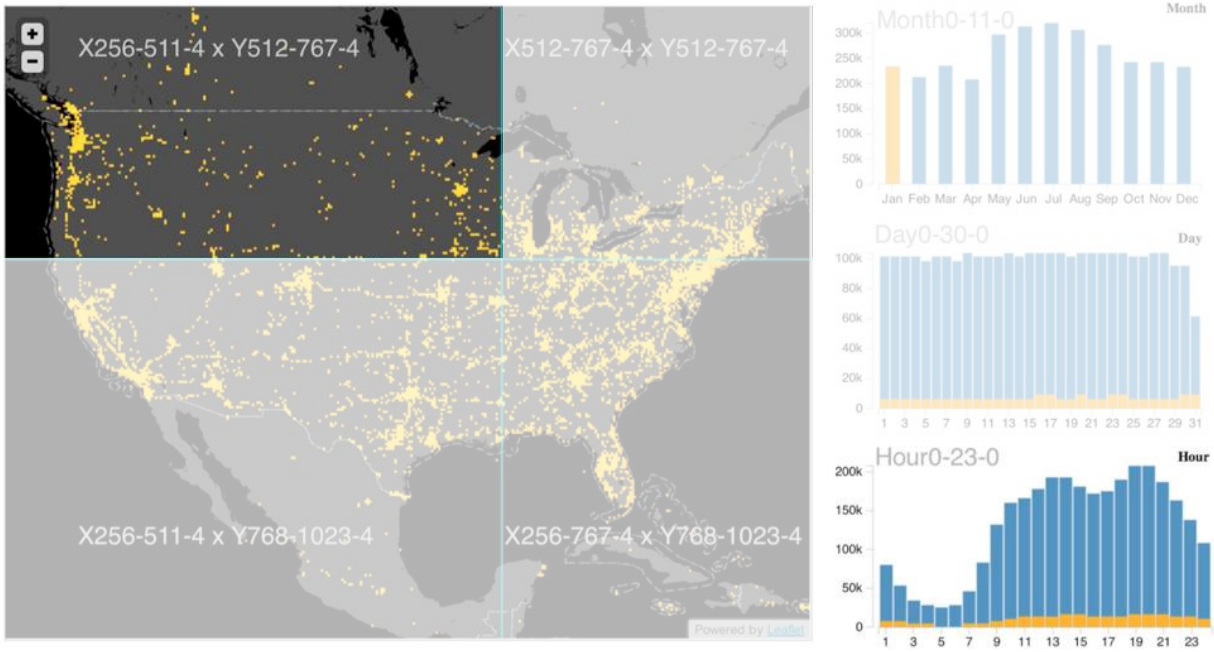
with Zhicheng (Leo) Liu & Biye Jiang


X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet


Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23


X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month
Month0-11-0
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day
Day0-30-0
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour
Hour0-23-0
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

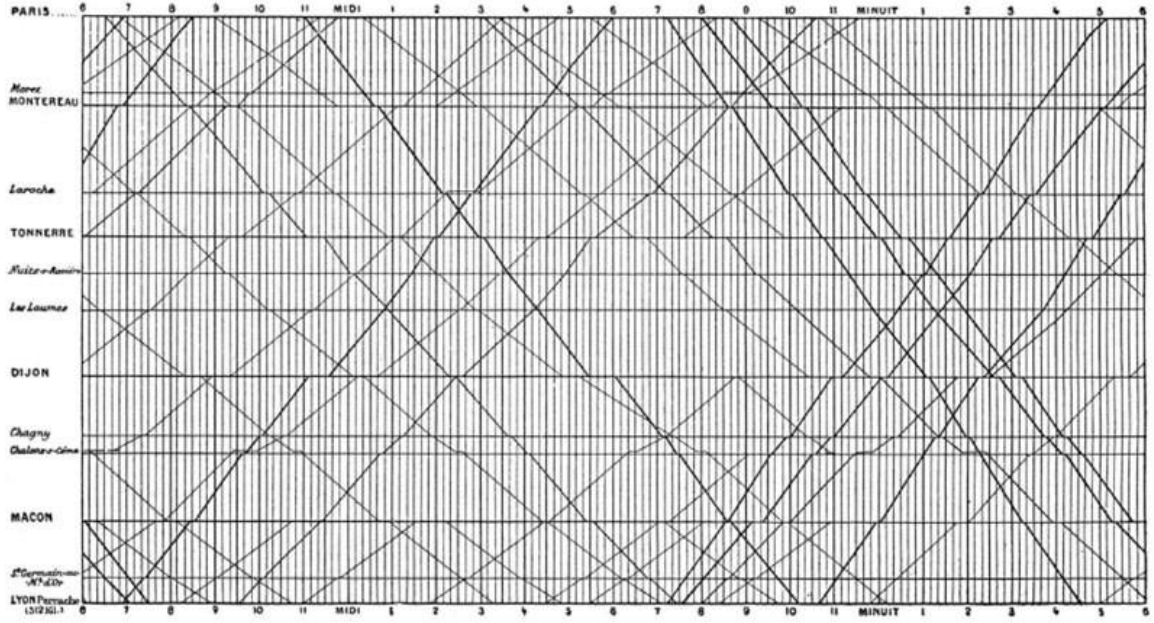

X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23


X256-511-4 x Y512-767-4
X512-767-4 x Y512-767-4
X256-511-4 x Y768-1023-4
X256-767-4 x Y768-1023-4
Powered by Leaflet
Month0-11-0
Month
300k
250k
200k
150k
100k
50k
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Day0-30-0
Day
100k
80k
60k
40k
20k
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Hour0-23-0
Hour
200k
150k
100k
50k
0
1 3 5 7 9 11 13 15 17 19 21 23

More Interactive Examples

Interaction: Example

- Paris-Lyon 1885 Train Schedule

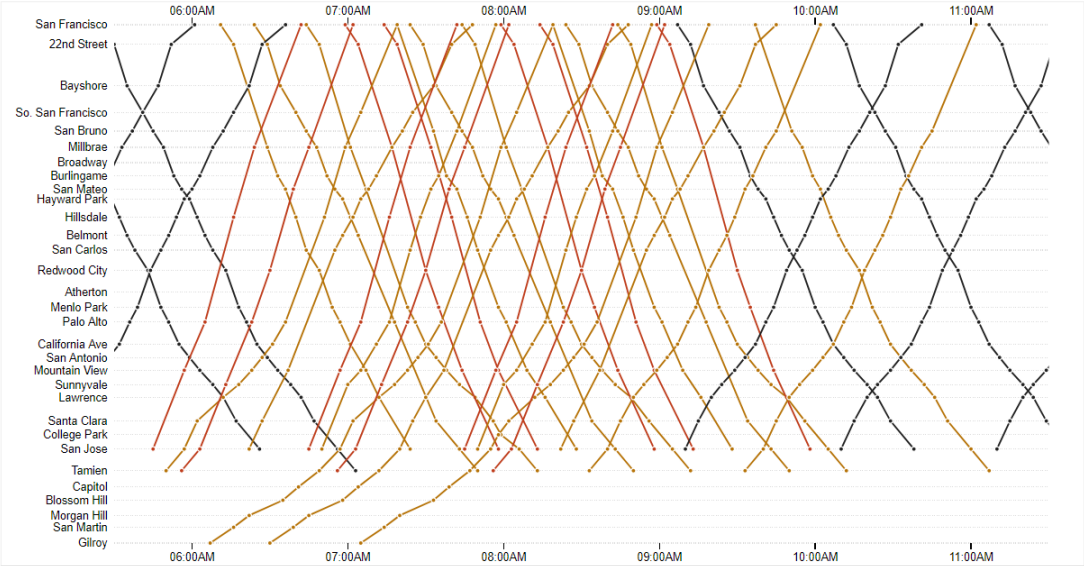


Etienne-Jules Marey , 1885

Interaction: Example

- Paris-Lyon Train Schedule Redesigned with Interaction

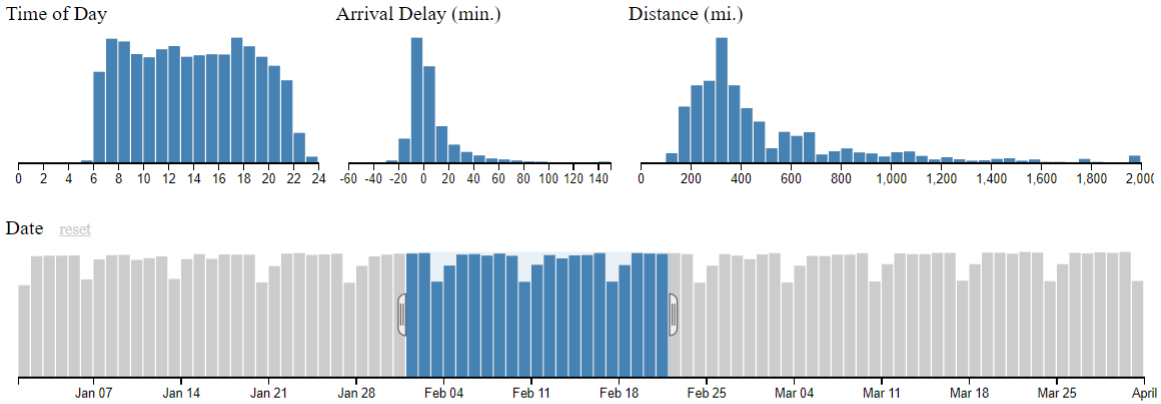
Marey's Trains



https://observablehq.com/@d3/mareys-trains

Interaction: Example

- Airline on-time performance



February 21, 2001

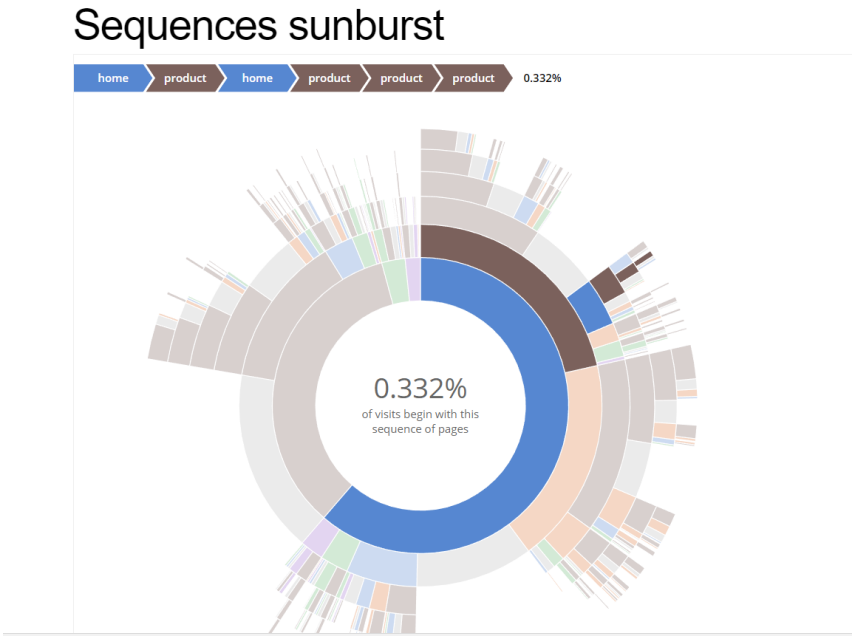
53,937 of 231,083 flights selected.

11:50 PM	PHX	ONT	325 mi.	+33 min.
11:50 PM	MCI	MDW	405 mi.	-9 min.
11:49 PM	BNA	HOU	670 mi.	+118 min.
11:30 PM	LAX	LAS	236 mi.	+21 min.
11:20 PM	LAS	PHX	256 mi.	+9 min.

http://square.github.io/crossfilter/

Interaction: Example

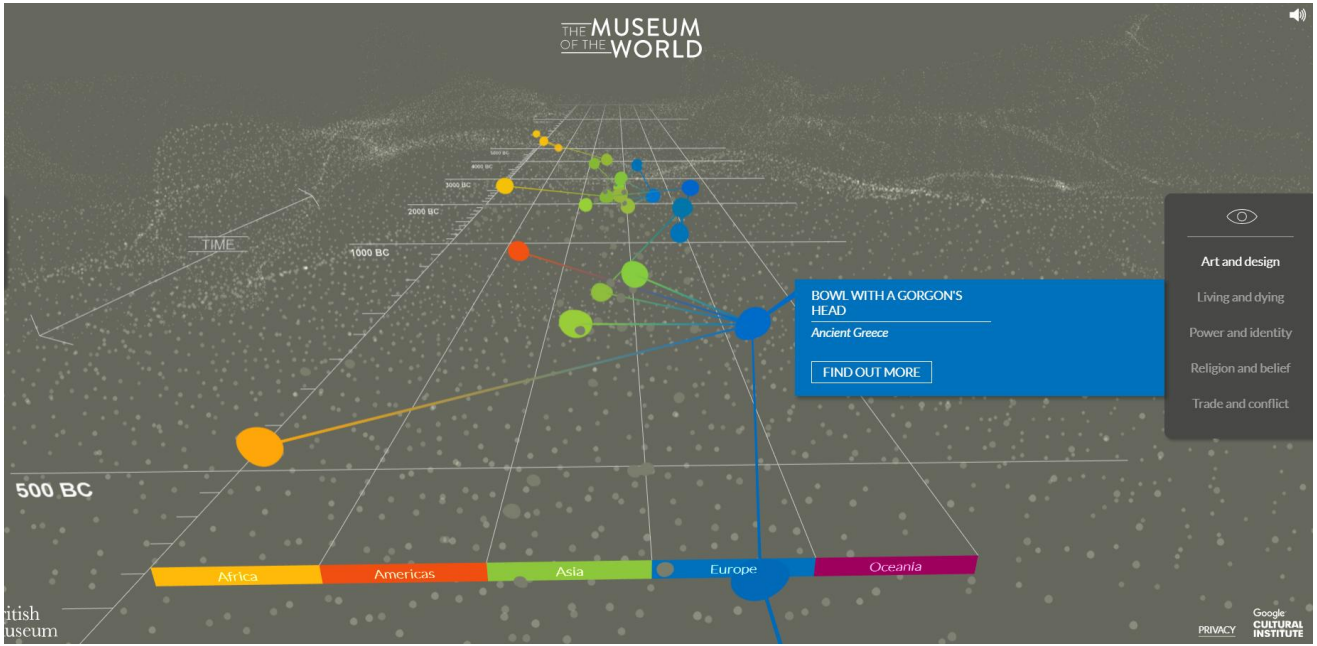
- D3 Sunburst Visualization



https://observablehq.com/@kerryrodden/sequences-sunburst

Interaction: Example

- British Museum



https://britishmuseum.withgoogle.com/

End of Lecture