



# Multivariate Data, Tables & Graphs

BAIS 6140 – Information Visualization

L. Miguel Encarnaç o

# Agenda

- Data and its characteristics
- Tables and graphs
- Design principles

# Data



- Data is taken from and/or representing some phenomena from the world
- Data models something of interest to us



# Data Sets

- Data comes in many different forms
  - Typically, not in the way you want them
  - What is available to me (in the raw)?



# Example (structured)

- Cars
  - Make
  - Model
  - Year
  - Miles per gallon
  - Cost
  - Number of cylinders
  - Weight
  - ...

# Example (unstructured)

- Web pages

# Data Models

- Often characterize data through three components
  - Objects
    - Items of interest  
(students, courses, terms, ...)
  - Attributes
    - Characteristics or properties of data  
(name, age, GPA, number, date, ...)
  - Relations
    - How two or more objects relate  
(student takes course, course during term, ...)

# Data Tables

- We take raw data and transform it into a model/form that is more workable
- Main idea
  - Individual items are called *cases*
  - Cases have *variables* (attributes)
  - Relational: Relations between cases (not our main focus today)



# Data Table Format

Dimension		Case <sub>1</sub>	Case <sub>2</sub>	Case <sub>3</sub>	...
	Variable <sub>1</sub>	Value <sub>11</sub>	Value <sub>21</sub>	Value <sub>31</sub>	
	Variable <sub>2</sub>	Value <sub>12</sub>	Value <sub>22</sub>	Value <sub>32</sub>	
	Variable <sub>3</sub>	Value <sub>13</sub>	Value <sub>23</sub>	Value <sub>33</sub>	
	...	<div>Think of as a function  <math>f(\text{case}_1) = \langle \text{Value}_{11}, \text{Value}_{12}, \dots \rangle</math> </div>			

# Example

	Mary	Jim	Sally	Mitch	...
SSN	145	294	563	823	
Age	23	17	47	29	
Hair	brown	black	blonde	red	
GPA	2.9	3.7	3.4	2.1	
...					

People in class



Or ...

	P1	P2	P3	P4	...
Name	Mary	Jim	Sally	Mitch	
SSN	145	294	563	823	
Age	23	17	47	29	
Hair	brown	black	blonde	red	
GPA	2.9	3.7	3.4	2.1	
...					

People in class

# Example

Microsoft Excel - baseball

File Edit View Insert Format Tools Data Accounting Window Help

Arial 10 B I U

A1 = Name

	A	B	C	D	E	F	G	H	I	J	
1	Name	At Bats	Hits	Home Run	Runs	Rbi	Walks	Years In M	Career At	Career Hits	Car
2	STRING	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT
3	Andy Allanson	293	66	1	30	29	14	1	293	66	
4	Alan Ashby	315	81	7	24	38	39	14	3449	835	
5	Alvin Davis	479	130	18	66	72	76	3	1624	457	
6	Andre Dawson	496	141	20	65	78	37	11	5628	1575	
7	Andres Galarra	321	87	10	39	42	30	2	336	101	
8	Alfredo Griffin	594	169	4	74	51	35	11	4438	1133	
9	Al Newman	185	37	1	23	8	21	2	214	42	
10	Argenis Salaza	298	73	0	24	24	7	3	539	108	
11	Andres Thomas	323	81	6	26	32	8	2	341	86	
12	Andre Thornton	401	92	17	49	66	65	13	5236	1332	
13	Alan Trammell	574	159	21	107	75	59	10	4631	1300	
14	Alex Trevino	202	53	4	31	26	27	9	1876	467	
15	Andy Van Slyke	418	113	13	48	61	47	4	1512	392	
16	Alan Wiggins	239	60	0	30	11	22	6	1941	510	
17	Bill Almon	196	43	7	29	27	30	13	3231	825	
18	Billy Beane	183	39	3	20	15	11	3	231	42	
19	Buddy Bell	568	158	20	89	75	73	15	8058	2273	
20	Ruddy Riancala	190	46	2	24	8	15	5	479	102	
21	Bruce Bochte	407	104	6	57	43	65	12	5233	1478	
22											

baseball

Ready



# Reshaping data (example)

**Long**

Log data

Patient	Date	Item	quantity
a	2011/10/01	Drug-a	2
a	2011/10/01	Drug-b	3
a	2011/10/02	Drug-a	2
b	2011/10/01	Drug-b	3
b	2011/10/02	Drug-c	4



**Wide**

Patient data

Patient	Drug-a			Drug-b		
	Yes/No	quantity	days	Yes/No	quantity	days
a	Yes	4	2	Yes	3	1
b	No			Yes	3	1



# Reshaping data (example)

Wide

	County	LandArea	NatAmenity	College1970	College1980	College1990	College2000	Jobs1970	Jobs1980	Jobs1990	Jobs2000
1	Autauga	599	4	.064	.121	.145	.180	6853	11278	11471	16289
2	Baldwin	1578	4	.065	.121	.168	.231	19749	27861	40809	70247
3	Barbour	891	4	.073	.092	.118	.109	9448	9755	12163	15197
4	Bibb	625	3	.042	.049	.047	.071	3965	4276	5564	6098
5	Blount	639	4	.027	.053	.070	.096	7587	9490	11811	16503



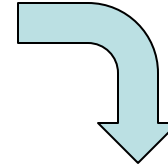
Long

	County	LandArea	NatAmenity	Year	College	Jobs
1	Autauga	599	4	1970	.064	6853
2	Autauga	599	4	1980	.121	11278
3	Autauga	599	4	1990	.145	11471
4	Autauga	599	4	2000	.180	16289
5	Baldwin	1578	4	1970	.065	19749
6	Baldwin	1578	4	1980	.121	27861
7	Baldwin	1578	4	1990	.168	40809
8	Baldwin	1578	4	2000	.231	70247
9	Barbour	891	4	1970	.073	9448
10	Barbour	891	4	1980	.092	9755
11	Barbour	891	4	1990	.118	12163
12	Barbour	891	4	2000	.109	15197
13	Bibb	625	3	1970	.042	3965
14	Bibb	625	3	1980	.049	4276
15	Bibb	625	3	1990	.047	5564
16	Bibb	625	3	2000	.071	6098
17	Blount	639	4	1970	.027	7587
18	Blount	639	4	1980	.053	9490
19	Blount	639	4	1990	.070	11811
20	Blount	639	4	2000	.096	16503

# Reshaping data (example)

**Table A-1. Years of School Completed by People 25 Years and Over, by Age and Sex: Selected Years 1940 to 2017**  
(Numbers in thousands. Noninstitutionalized population except where otherwise specified.)

Age, sex, and years	Total	Years of School Completed						Median	
		Elementary		High school		College			
		0 to 4 years	5 to 8 years	1 to 3 years	4 years	1 to 3 years	4 years or more		
25 YEARS AND OLDER									
Both Sexes									
2017	216,921	2,208	6,600	13,734	62,512	57,765	74,103	(NA)	
2016	215,015	2,414	7,078	13,961	62,002	57,660	71,900	(NA)	
2015	212,132	2,601	7,295	14,686	62,575	56,031	68,945	(NA)	
2014	209,287	2,525	7,388	14,545	62,240	55,709	66,879	(NA)	
2013	206,899	2,344	7,578	14,595	61,704	55,173	65,506	(NA)	
2012	204,579	2,484	7,800	14,993	62,113	53,900	63,291	(NA)	
2011	201,543	2,589	7,688	14,763	61,911	53,249	61,343	(NA)	
2010	199,928	2,615	7,836	15,260	62,456	51,920	59,840	(NA)	
2009	198,285	2,785	8,043	15,587	61,626	51,670	58,574	(NA)	
2008	196,305	2,599	8,226	15,516	61,183	50,994	57,787	(NA)	
2007	194,318	2,830	8,462	16,451	61,490	49,243	55,842	(NA)	
2006	191,884	2,951	8,791	16,154	60,898	49,371	53,720	(NA)	
2005	189,367	2,983	8,935	16,099	60,893	48,076	52,381	(NA)	
2004	186,876	2,858	8,888	15,999	59,811	47,571	51,749	(NA)	
2003	185,183	2,915	9,361	16,323	59,292	46,910	50,383	(NA)	
2002	182,142	2,902	9,668	16,378	58,456	46,042	48,696	(NA)	
2001	180,389	2,810	9,518	16,279	58,272	46,281	47,228	(NA)	



Age	sex	years	Total	median	year level	school level	number of people
25 YEARS AND OLDER	Both Sexes	2017	216921	null	0 to 4 years	Elementary	2208
25 YEARS AND OLDER	Both Sexes	2017	216921	null	1 to 3 years	College	57765
25 YEARS AND OLDER	Both Sexes	2017	216921	null	1 to 3 years	High school	13734
25 YEARS AND OLDER	Both Sexes	2017	216921	null	4 years	High school	62512
25 YEARS AND OLDER	Both Sexes	2017	216921	null	4 years or more	College	74103
25 YEARS AND OLDER	Both Sexes	2017	216921	null	5 to 8 years	Elementary	6600
25 YEARS AND OLDER	Both Sexes	2016	215015	null	0 to 4 years	Elementary	2414
25 YEARS AND OLDER	Both Sexes	2016	215015	null	1 to 3 years	College	57660
25 YEARS AND OLDER	Both Sexes	2016	215015	null	1 to 3 years	High school	13961
25 YEARS AND OLDER	Both Sexes	2016	215015	null	4 years	High school	62002
25 YEARS AND OLDER	Both Sexes	2016	215015	null	4 years or more	College	71900
25 YEARS AND OLDER	Both Sexes	2016	215015	null	5 to 8 years	Elementary	7078
25 YEARS AND OLDER	Both Sexes	2015	212132	null	0 to 4 years	Elementary	2601
25 YEARS AND OLDER	Both Sexes	2015	212132	null	1 to 3 years	College	56031
25 YEARS AND OLDER	Both Sexes	2015	212132	null	1 to 3 years	High school	14686
25 YEARS AND OLDER	Both Sexes	2015	212132	null	4 years	High school	62575
25 YEARS AND OLDER	Both Sexes	2015	212132	null	4 years or more	College	68945
25 YEARS AND OLDER	Both Sexes	2015	212132	null	5 to 8 years	Elementary	7295
25 YEARS AND OLDER	Both Sexes	2014	209287	null	0 to 4 years	Elementary	2525
25 YEARS AND OLDER	Both Sexes	2014	209287	null	1 to 3 years	College	55709
25 YEARS AND OLDER	Both Sexes	2014	209287	null	1 to 3 years	High school	14545
25 YEARS AND OLDER	Both Sexes	2014	209287	null	4 years	High school	62240
25 YEARS AND OLDER	Both Sexes	2014	209287	null	4 years or more	College	66879
25 YEARS AND OLDER	Both Sexes	2014	209287	null	5 to 8 years	Elementary	7388
25 YEARS AND OLDER	Both Sexes	2013	206899	null	0 to 4 years	Elementary	2344
25 YEARS AND OLDER	Both Sexes	2013	206899	null	1 to 3 years	College	55173
25 YEARS AND OLDER	Both Sexes	2013	206899	null	1 to 3 years	High school	14595
25 YEARS AND OLDER	Both Sexes	2013	206899	null	4 years	High school	61704

# Reshaping data: Wide vs. Long

- Wide format

- Longitudinal data analysis

- Same patient, multiple drugs, multiple tests
    - Repeated outcomes are considered different and non-interchangeable variables
      - Each can have its own distribution. Each is distinct.
    - Customized calculations on metrics
      - T-test
      - MANOVA

- Long format

- Exploration of relationships

- Longitudinal dependencies are intentionally deprioritized

- Flexible Viz

- Filters





# Types of Variables



- Three main types
  - N-Nominal (equal or not equal to other values)
    - Example: gender
  - O-Ordinal (obeys  $<$  relation, ordered set)
    - Example: fr,so,jr,sr
  - Q-Quantitative (can do math on them)
    - Example: age

# Types of Variables

- Alternate Characterization
- Two types of data
  - Quantitative
    - Relationships between values:
      - Ranking
      - Ratio
      - Correlation
  - Categorical
    - How attributes relate to each other:
      - Nominal
      - Ordinal
      - Interval
      - Hierarchical



# Metadata

- Descriptive information about the data
  - Might be something as simple as the type of a variable, or could be more complex
  - For times when the table itself just isn't enough
  - Example: if Variable<sub>1</sub> is "I", then Variable<sub>3</sub> can only be 3, 7 or 16

# Number of Variables

- Data sets of dimensions 1, 2, 3 are common
- Number of variables per class
  - 1 -Univariate data
  - 2 -Bivariate data
  - 3 -Trivariate data
  - >3 -Hypervariate data

# Representation

- Two main ways of presenting multivariate data sets
  - Directly (textually)
    - Tables
  - Symbolically (pictures)
    - Graphs
- When to use which?

# Representation

- Use tables when
  - The document will be used to look up individual values
  - The document will be used to compare individual values
  - Precise values are required
  - The quantitative info to be communicated involves more than one unit of measure
- Use graphs when
  - The message is contained in the shape of the values
  - The document will be used to reveal relationships among values

S. Few  
*Show me the Numbers*

# Effective Table Design

- See *Show Me the Numbers*
- Proper and effective use of layout, typography, shading, etc. can go a long way
- (Tables may be underutilized)

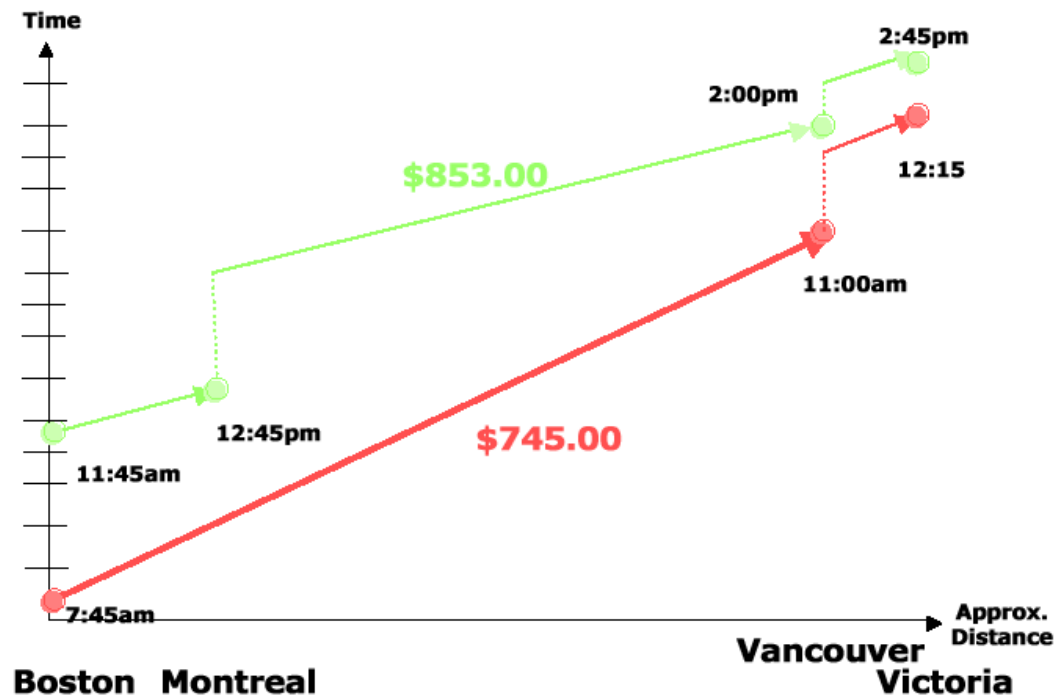


# Basic Symbolic Displays

- Graphs
- Charts
- Maps
- Diagrams

# 1. Graphs

- Showing the relationships between variables' values in a data table



Expedia example from

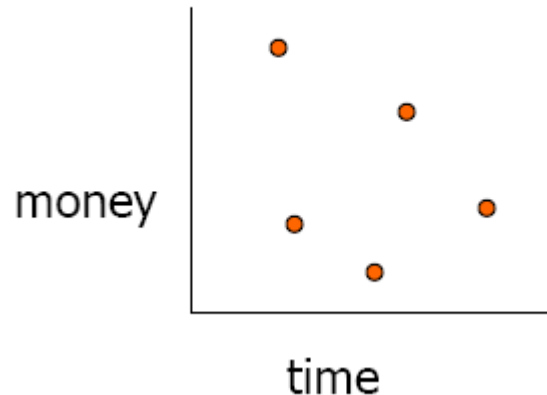
*Storey, M. Information Visualization and Knowledge Management, Dept. of Aero/Astro, MIT, 2003*

# Properties

- Graph
  - Visual display that illustrates one or more relationships among entities
  - Shorthand way to present information
  - Allows a trend, pattern or comparison to be easily comprehended

# Issues

- Critical to remain task-centric
  - Why do you need a graph?
  - What questions are being answered?
  - What data is needed to answer those questions?
  - Who is the audience?




# Components

- Framework
  - Measurement types, scale
- Content
  - Marks, lines, points
- Labels
  - Title, axes, ticks



# Many Examples

 **NationMaster** Categories Countries A-Z Top stats Groups

Break it down: Australian Real Estate, Local Demographic and Amenity Profiles Now On NationMaster.

## [Learn](#) About Neighbourhoods and Amenities For Any Property in Australia

## Compare Countries on Just about Anything!

NationMaster is where stats come alive! We are a massive central data source and a handy way to graphically compare nations.

NationMaster is a vast compilation of data from hundreds of sources. Using the forms below, you can get maps and graphs on all kinds of statistics with ease.

We want to be the web's one-stop resource for country statistics on everything from obesity to murders.

305 countries

[China](#) • [India](#) • [Russia](#) • [United States](#)

43 country groups

[Europe](#) • [Group of 7 countries \(G7\)](#) • [Least Developed Countries](#) • [Emerging markets](#)

5037 categories

[Murder rate](#) • [Crime](#) • [GDP](#) • [Economy](#)


19,814,971 data points

Latest stats


[Import > Northernmost point > Latitude](#) 01.01.2016

[People > Urban and rural > Population living in cities proper](#) 01.01.2016

[Health > Deaths > Percent deaths](#) 01.01.2016



Country	GDP per capita (2015)
United States	54,149
Germany	44,280
France	41,390
United Kingdom	40,350
Canada	39,580
Japan	39,580
China	7,180
India	1,940



Tables, graphs, maps and pie charts

POPULAR STATS

[Geography > Land area > Square miles](#)

[Economy > GDP](#)


[People > Population in 2015](#)


[People > Population](#)

Country facts and stats

Compare any two

Trending now

 [United States & Russian Militaries Compared](#)

 [The Secret of Japan's Mysterious Low Crime Rate](#)

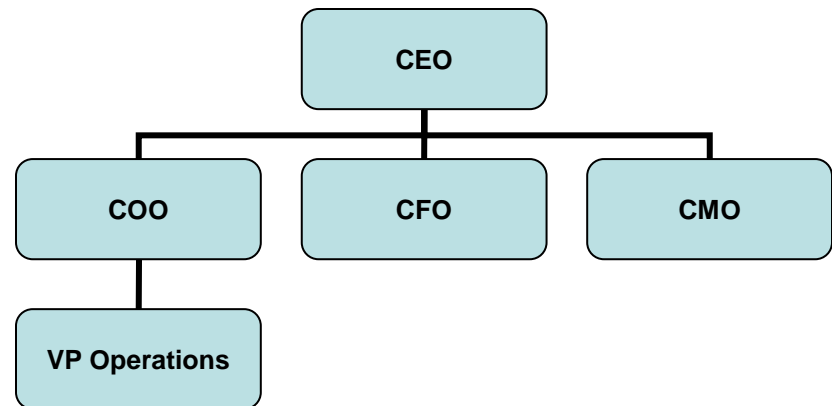
www.nationmaster.com

# Excursion

- Other symbolic displays
  - Chart
  - Map
  - Diagram

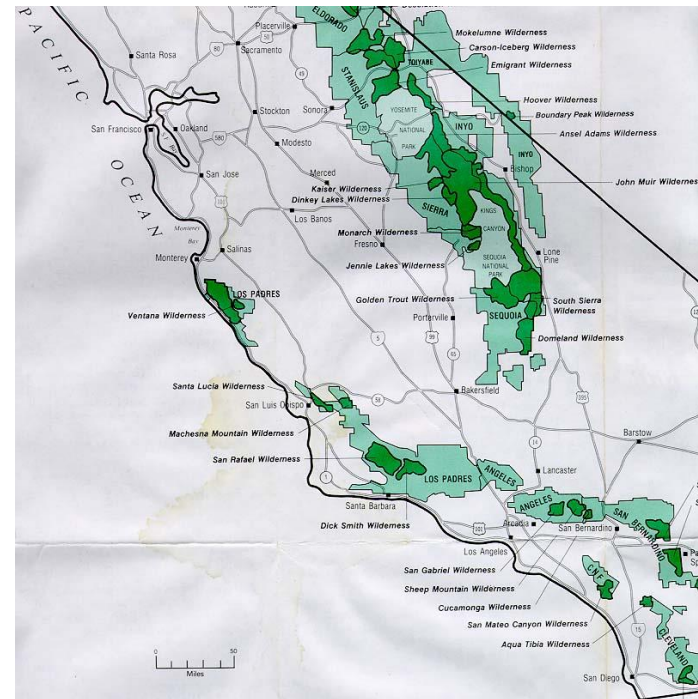
## 2. Charts

- Discrete relations among discrete entities
- Structure relates entities to one another
- Lines and relative position serve as links
- Examples
  - family tree
  - flow chart
  - network diagram



# 3. Maps

- Internal relations determined (in part) by the spatial relations of what is pictured
- Labels paired with locations

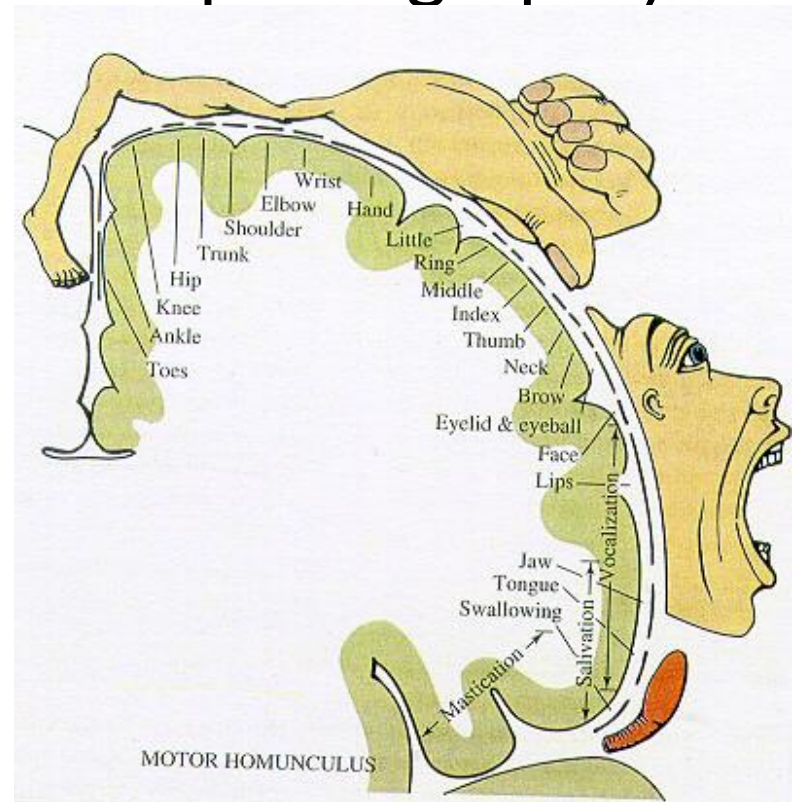


Map of census data topographic maps  
from [www.thehighsierra.com](http://www.thehighsierra.com)



# 4. Diagrams

- Schematic pictures of objects or entities
- Parts are symbolic (unlike photographs)
  - how-to illustrations
  - figures in a manual



From Glietman, Henry. *Psychology*.  
W.W. Norton and Company, Inc. New  
York, 1995

# Some History

- Which is older, map or graph?
- Maps from about 2300 BC
- Graphs from 1600's
  - Rene Descartes
  - William Playfair, late 1700's

# Details

- What are the constituent pieces of these four symbolic displays?
- What are the building blocks?

# Visual Structures

- Composed of
  - Spatial substrate
  - Marks
  - Graphical properties of marks



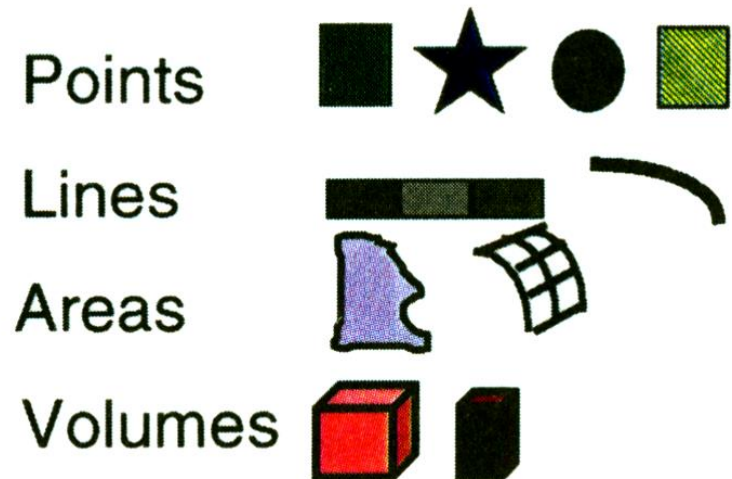
# Space



- Visually dominant
- Often put axes on space to assist
- Use techniques of composition, alignment, folding, recursion, overloading to
  1. increase use of space
  2. do data encodings

# Marks

- Four elementary types of visible things in space
- Point marks and line marks can be combined to signify *graphs* and *trees*



From

Card, Stuart K., *Readings in information visualization: Using vision to think*, Morgan Kaufmann Publishers, 1999.

# Graphical Properties

- Size, shape, color, orientation, ...

	Spatial	Object
Extent	(Position) — — —  Size ● ● ● ●	Gray Scale ■ ■ ■ ■
Dif-feren-tial	Orientation — /   \	Color ■ ■ ■ ■ Texture ■ ■ ■ ■ Shape ■ ★ ● ◆

From

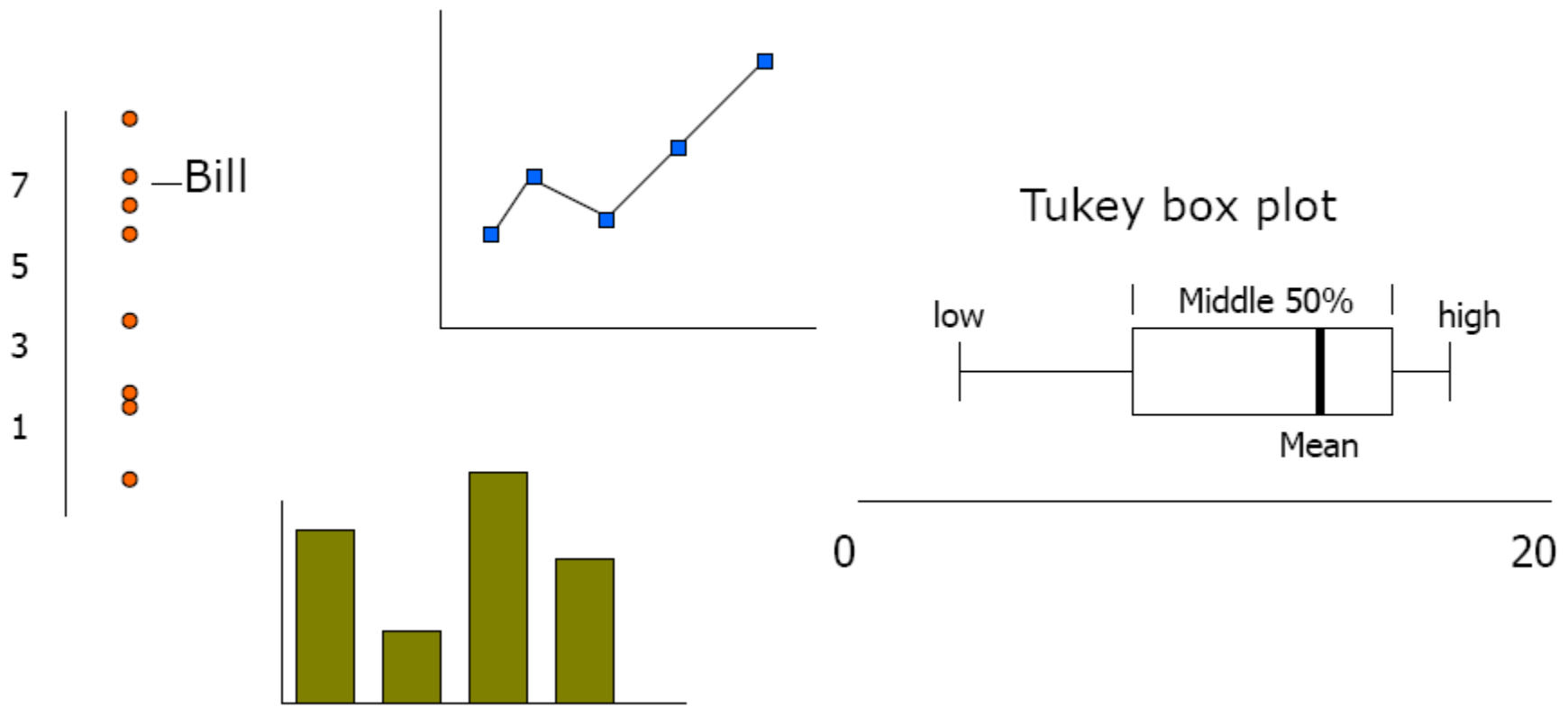
Card, Stuart K., *Readings in information visualization: Using vision to think*, Morgan Kaufmann Publishers, 1999.

# Back to Data Graphs

- What were the different types of data sets?
- Number of variables per class
  - 1 -Univariate data
  - 2 -Bivariate data
  - 3 -Trivariate data
  - >3 -Hypervariate data

# 1 -Univariate Data

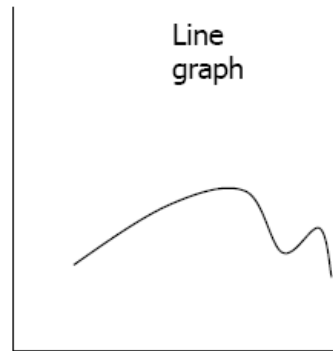
- Representations





# How to ...

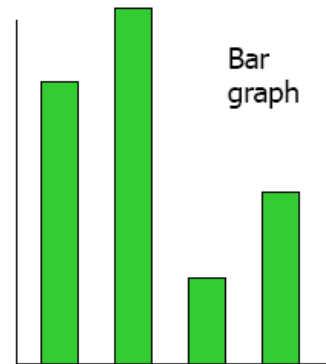
- In univariate representations, we often think of the data case as being shown along one dimension, and the value in another



Line  
graph

Y-axis is quantitative  
variable

See changes over  
consecutive values



Bar  
graph

Y-axis is quantitative  
variable

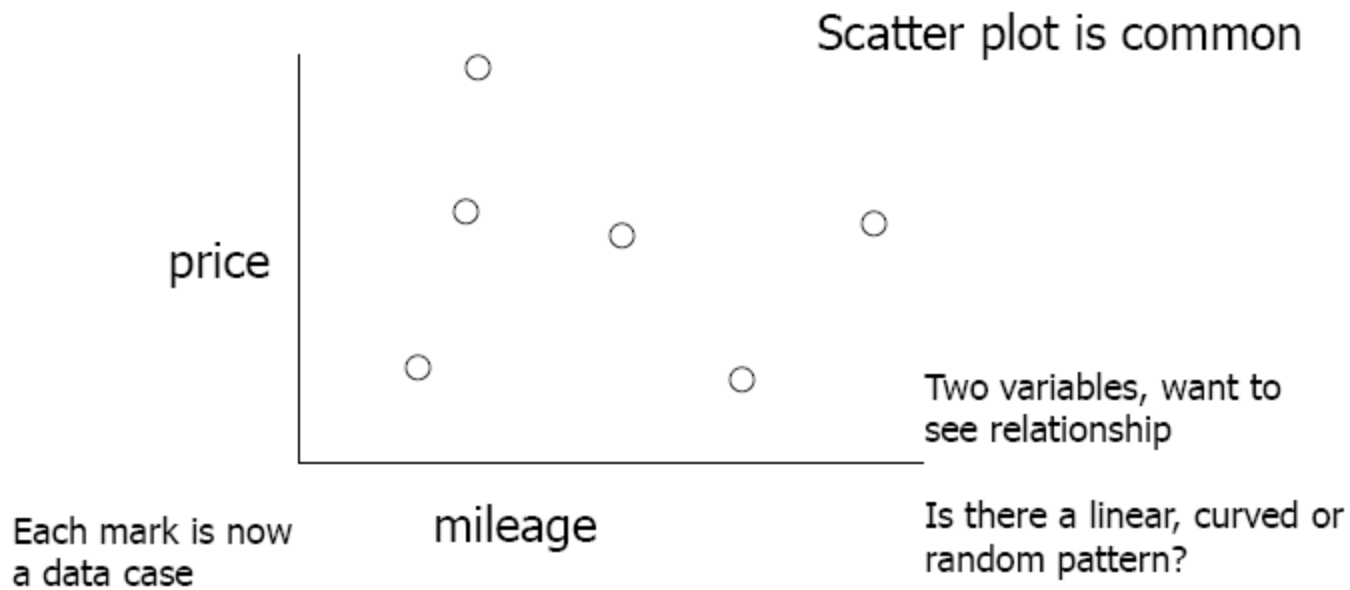
Compare relative point  
values

## ... or (alternate view)

- We may think of graphs as representing independent (data case) variables and dependent (value) variables
- Guideline
  - Independent vs. dependent variables
    - Put independent variables on x-axis
    - See resultant dependent variables along y-axis

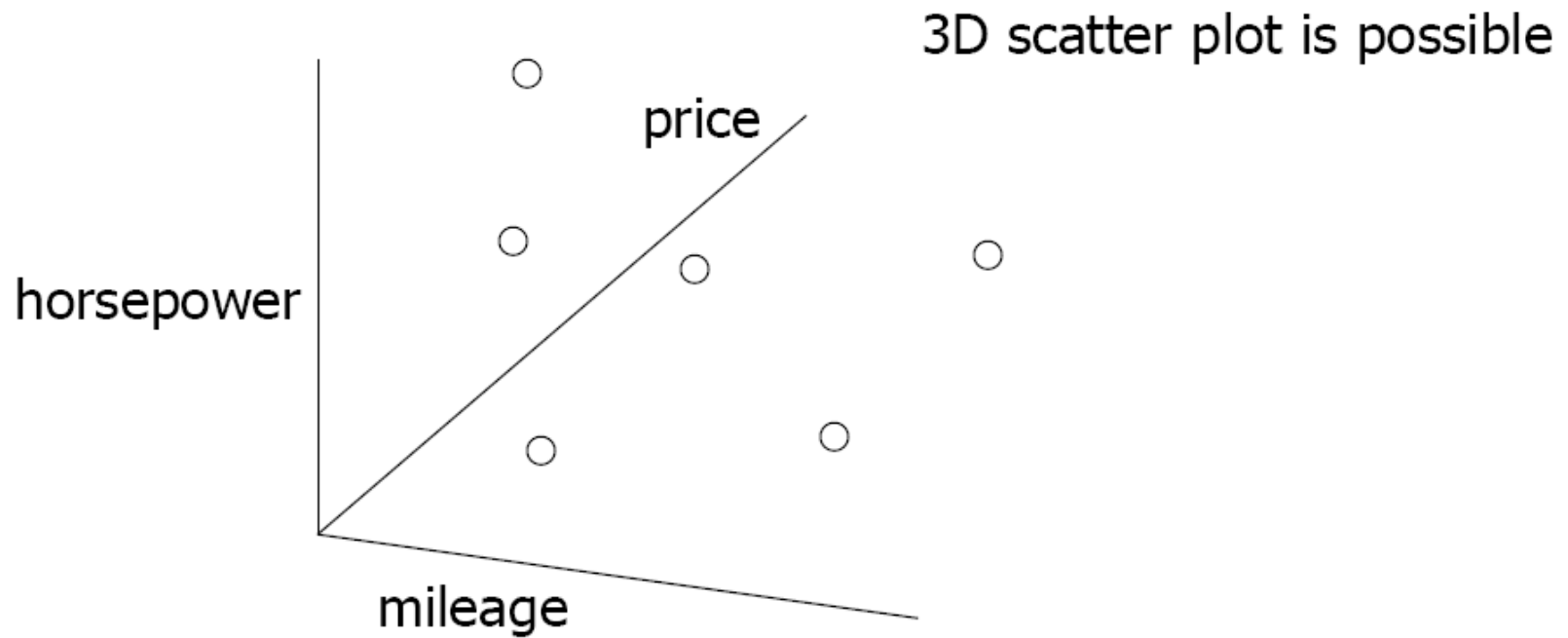
# 2 - Bivariate Data

- Representations

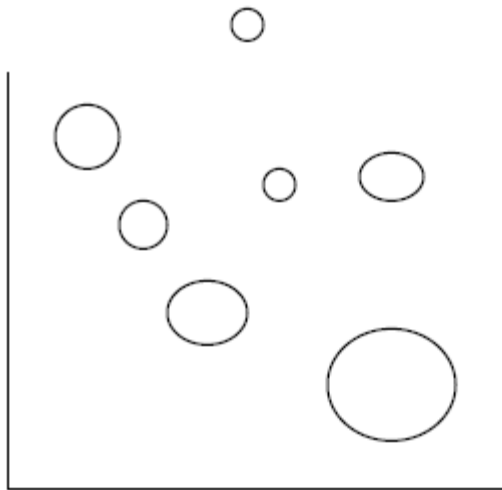


# 3 - Trivariate Data

- Representations



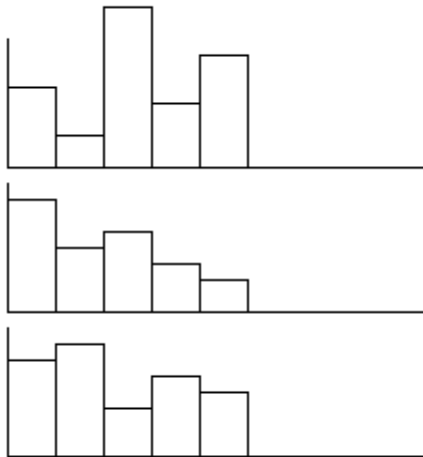
# ... or (alternate representation)



Still use 2D but have  
mark property  
represent third  
variable



# ... or (another alternate rep.)



Represent each variable  
in its own explicit way

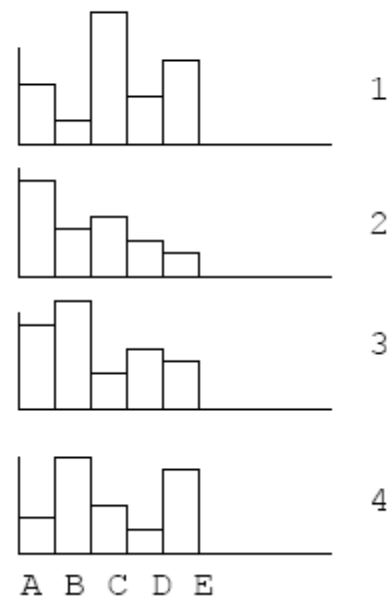
# >3 -Hypervariate Data

- Tough to ‘visualize’
  - Number of well-known visualization techniques exist for data sets of 1-3 dimensions
    - Line graphs, bar graphs, scatter plots
    - We see a 3-D world (4-D with time)
  - What about data sets with more than 3 variables?
    - Often the interesting, challenging ones

# Multiple Views

Give each variable its own display

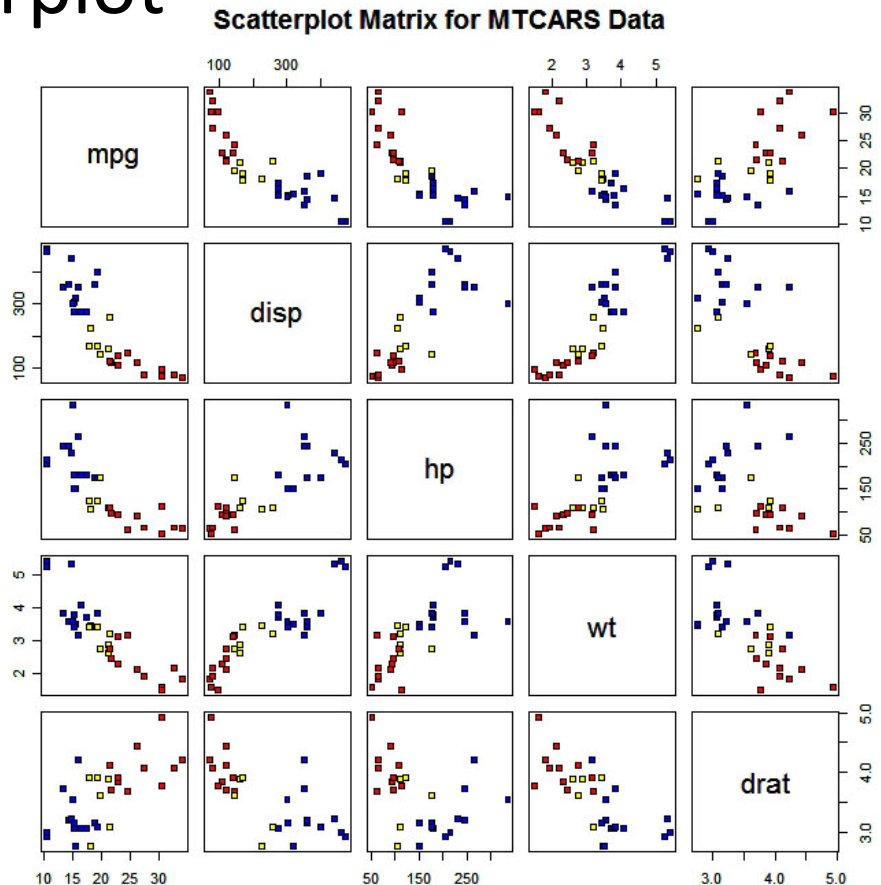
	A	B	C	D	E
1	4	1	8	3	5
2	6	3	4	2	1
3	5	7	2	4	3
4	2	6	3	1	5



# One View – Scatterplot Matrix

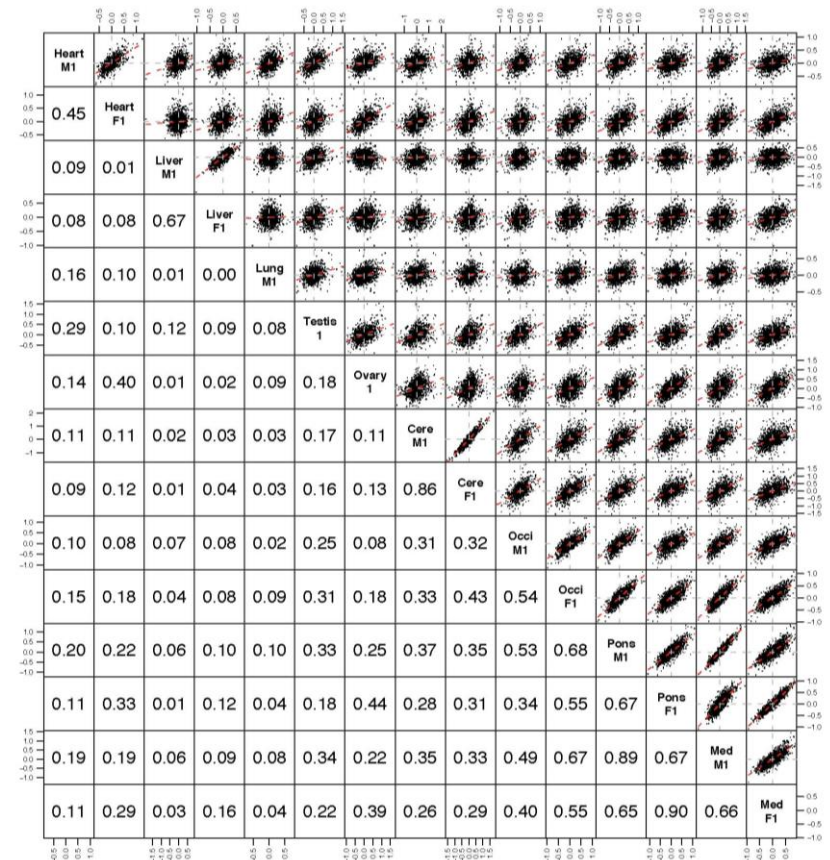
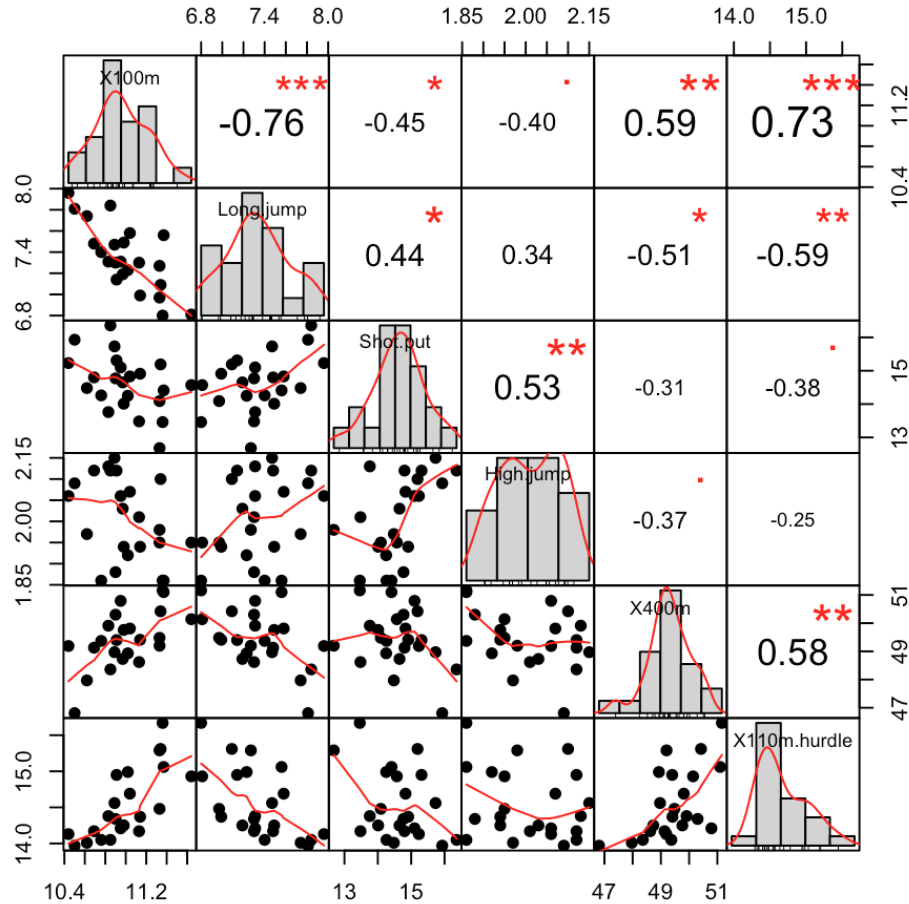
- Represent each possible pair of variables in their own 2-D scatterplot

- Useful for what?
- Misses what?





# One View – Scatterplot Matrix

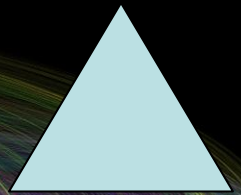




# More to come ...

- Later class will explore other general techniques for handling hypervariate data

# Back to Graphs



- Design guidance
  - S. Few provides many helpful principles to designing effective graphs

# S. Few's Selection & Design Process

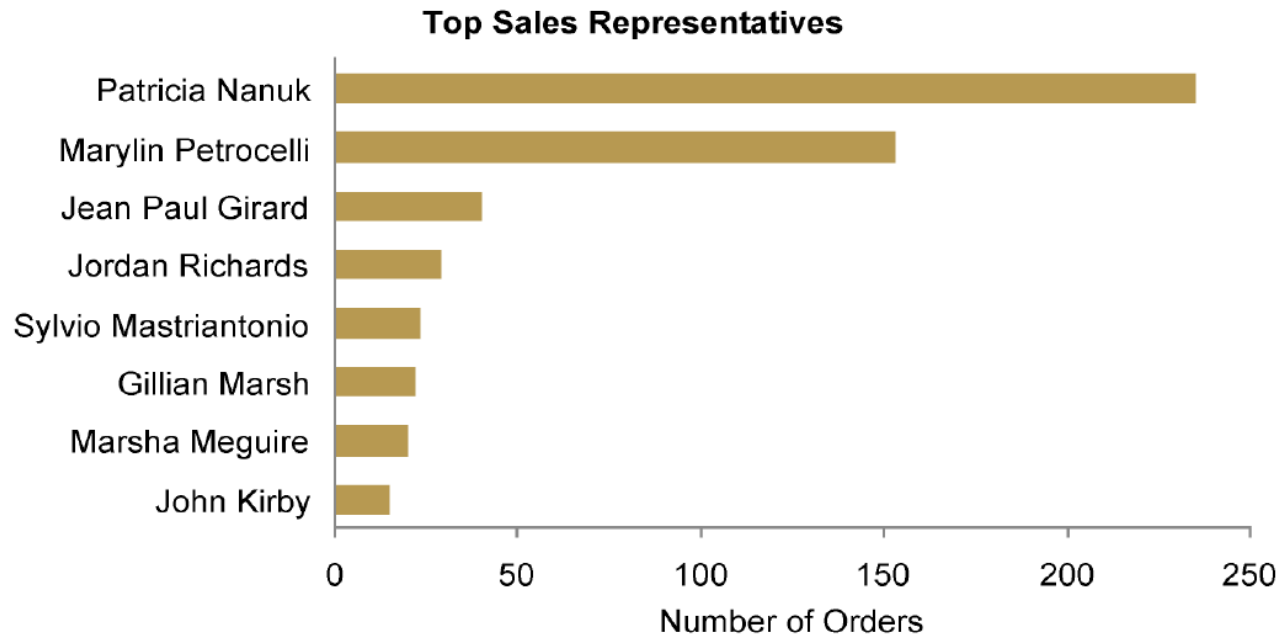
1. Determine your message and identify your data
2. Determine if a table, or graph, or both is needed to communicate your message
3. Determine the best means to encode the values
4. Determine where to display each variable
5. Determine the best design for the remaining objects
  1. Determine the range of the quantitative scale
  2. If a legend is required, determine where to place it
  3. Determine the best location for the quantitative scale
  4. Determine if grid lines are required
  5. Determine what descriptive text is needed
6. Determine if particular data should be featured and how

# Points, Lines, Bars, Boxes

- Points
  - Useful in scatterplots for 2-values
  - Can replace bars when scale doesn't start at 0
- Lines
  - Connect values in a series
  - Show changes, trends, patterns
  - Not for a set of nominal or ordinal values
- Bars
  - Emphasizes individual values
  - Good for comparing individual values
- Boxes
  - Shows a distribution of values

# Vertical vs. Horizontal bars

- Horizontal can be good if long labels or many items

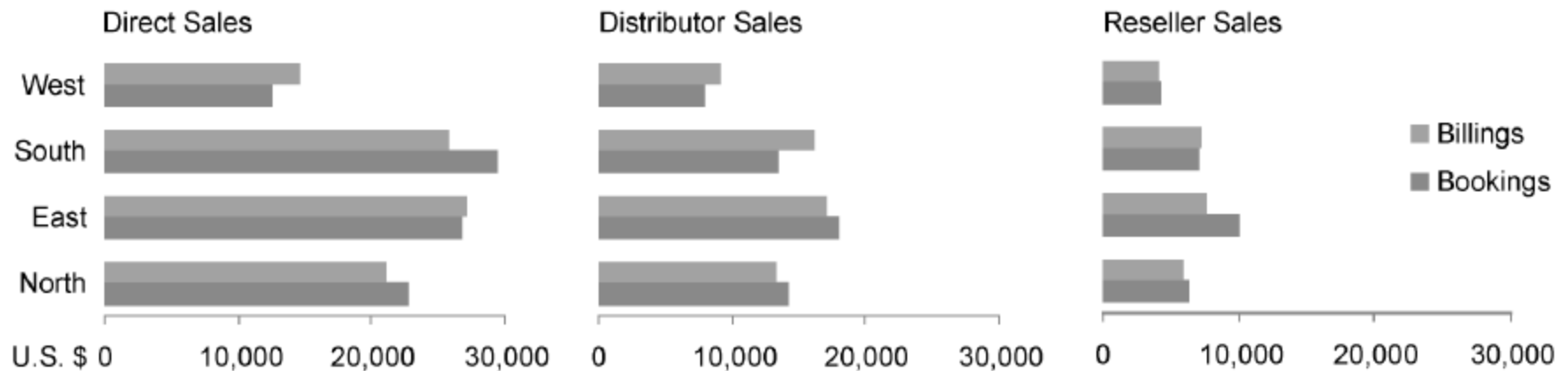


[http://www.perceptualedge.com/articles/Whitepapers/Communicating\\_Numbers.pdf](http://www.perceptualedge.com/articles/Whitepapers/Communicating_Numbers.pdf), Fig. 22



# Multiple Bars

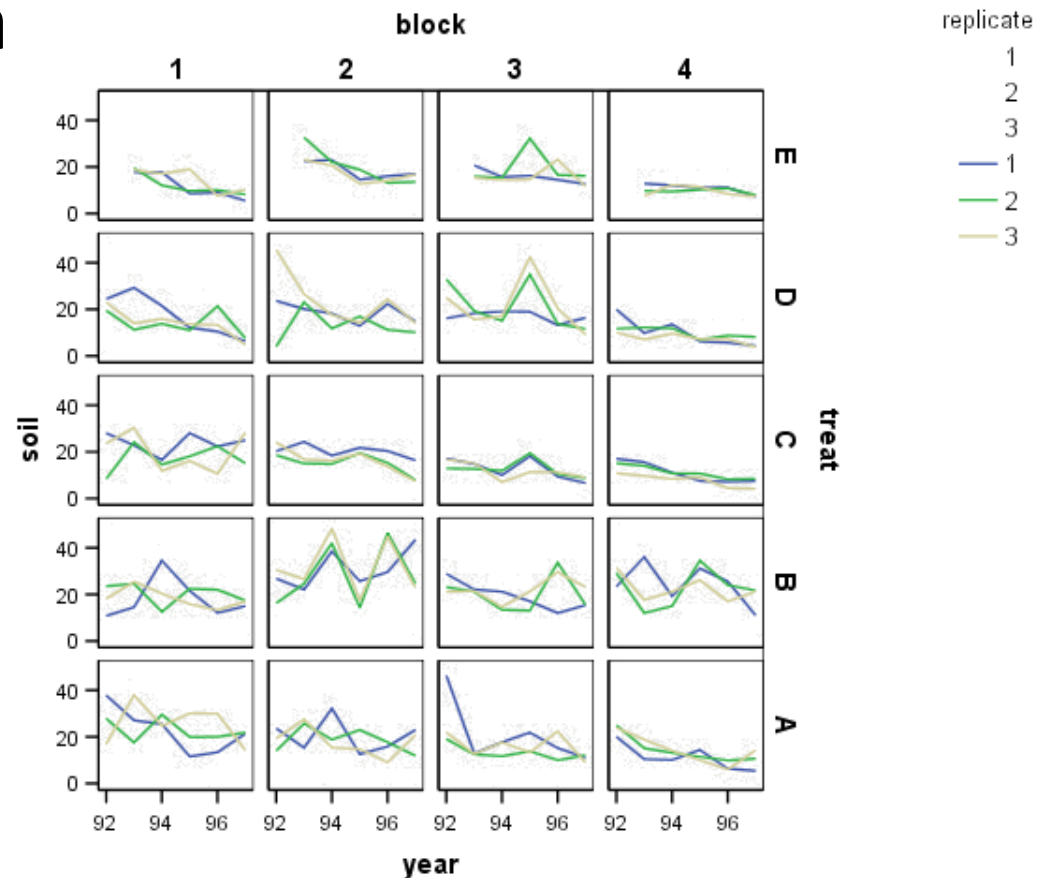
- Can be used to encode another variable



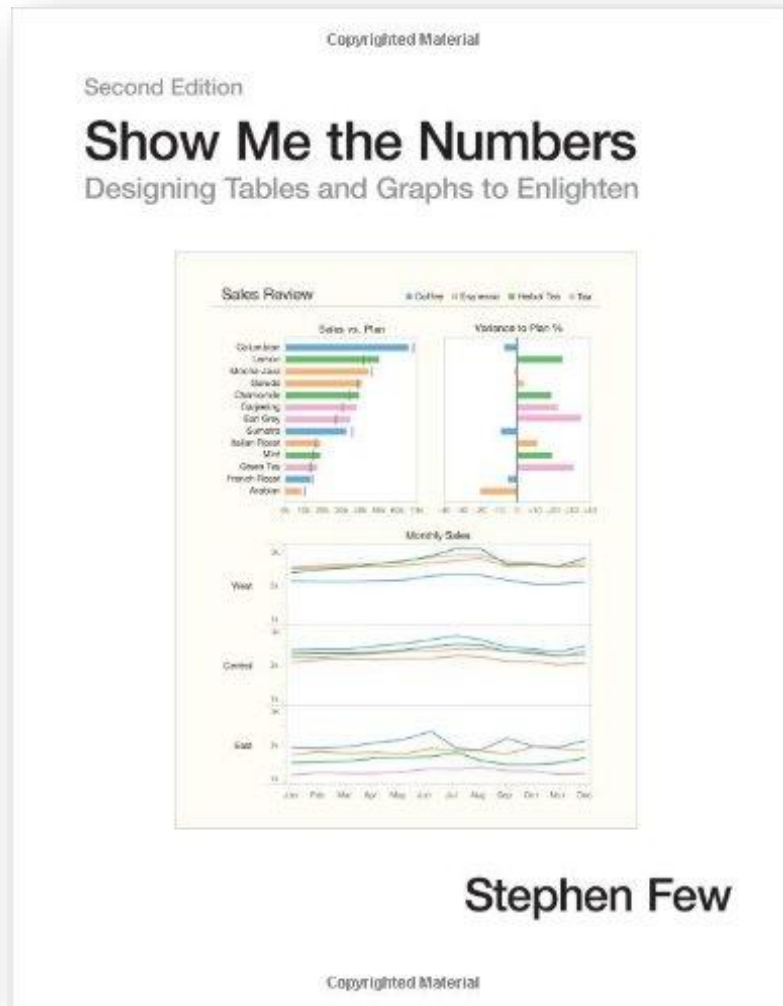
[http://www.perceptualedge.com/articles/Whitepapers/Communicating\\_Numbers.pdf](http://www.perceptualedge.com/articles/Whitepapers/Communicating_Numbers.pdf), Fig. 24

# Multiple Graphs

- Can distribute a variable across graphs too
- Sometimes called a trellis display



# Book Recommendation



- Loaded with examples of how to redesign ineffective tables and graphs