

# Multivariate Data & Tables and Graphs



CS 4460 – Intro. to Information Visualization  
Sep. 4, 2014  
John Stasko

## 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 you (in the raw)?

# Example



- Cars
  - make
  - model
  - year
  - miles per gallon
  - cost
  - number of cylinders
  - weights
  - ...

# Example



- 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, ...)

Fall 2014

CS 4460

7

# 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)

Fall 2014

CS 4460

8

# Data Table Format



	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>	
...				

Think of as a function  
 $f(\text{case}_i) = \langle \text{Val}_{1i}, \text{Val}_{2i}, \dots \rangle$

# 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



Baseball statistics

	A	B	C	D	E	F	G	H	I	J	K
1	Name	At Bats	Hits	Home Run	Runs	Rbi	Walks	Years In M	Career At	Career Hit	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	396	101	
8	Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	
9	Al Newman	185	37	1	23	8	21	2	214	42	
10	Argenis Salaza	298	73	0	24	24	7	3	509	108	
11	Andres Thomas	323	81	6	26	32	8	2	341	86	
12	Andre Thornton	401	92	17	49	66	65	13	5206	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	201	42	
19	Buddy Bell	568	158	20	89	75	73	15	8068	2273	
20	Buddy Biancali	190	46	2	24	8	15	5	479	102	
21	Bruce Bochte	407	104	6	57	43	65	12	5233	1478	

# Variable Types



- Three main types of variables
  - 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

Fall 2014

CS 4460

13

# Alternate Characterization



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

From S. Few

Fall 2014

CS 4460

14

# 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 variable1 is "1", then variable3 can only be 3, 7 or 16

# Data Cleaning



- Data may be missing/corrupted
  - Remove?
  - Modify?
- You may want to adjust values
  - Use inverse
  - Map nominal to ordinal/quantitative
  - Normalize values
    - Scale between 0 and 1



# How Many 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



- What are two main ways of presenting multivariate data sets?
  - Directly (textually) → Tables
  - Symbolically (pictures) → Graphs
- When use which?

# Strengths?

S. Few  
*Show Me the Numbers*



- 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

Fall 2014

CS 4460

19

# 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 underused)

Fall 2014

CS 4460

20

# Example



(Image shown in class)

Fall 2014

CS 4460

21

# Example



(Image shown in class)

Fall 2014

CS 4460

22

# Basic Symbolic Displays



- Graphs ←
- Charts
- Maps
- Diagrams

From:  
S. Kosslyn, "Understanding charts and graphs", *Applied Cognitive Psychology*, 1989.

Fall 2014

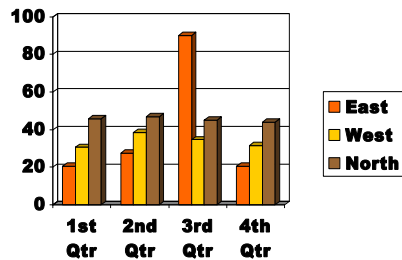
CS 4460

23

## 1. Graph



Showing the relationships between variables' values in a data table



Fall 2014

CS 4460

24

# 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

Fall 2014

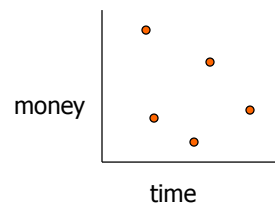
CS 4460

25

# 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?



Fall 2014

CS 4460

26



# Quick Aside



- Other symbolic displays
  - Chart
  - Map
  - Diagram

Fall 2014

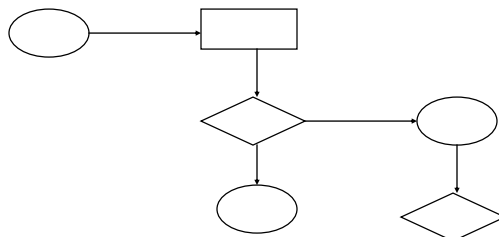
CS 4460

29

## 2. Chart



- Structure is important, relates entities to each other
- Primarily uses lines, enclosure, position to link entities



Examples: flowchart, family tree, org chart, ...

Fall 2014

CS 4460

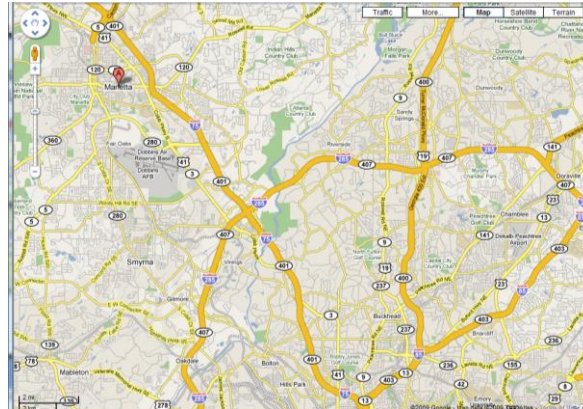
30

# 3. Map



Representation of spatial relations

Locations identified by labels



Fall 2014

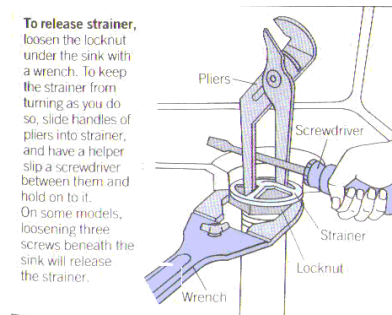
CS 4460

31

# 4. Diagram



- Schematic picture of object or entity
- Parts are symbolic



Examples: figures, steps in a manual, illustrations,...

Fall 2014

CS 4460

32



## Some History



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



Fall 2014

CS 4460

33

## Details



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

Fall 2014

CS 4460

34

# Visual Structures



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

Fall 2014

CS 4460

35

# 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

Fall 2014

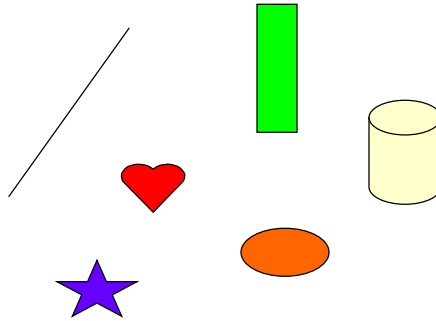
CS 4460

36

# Marks



- Things that occur in space
  - Points
  - Lines
  - Areas
  - Volumes



Fall 2014

CS 4460

37

# Graphical Properties



- Size, shape, color, orientation...

	Spatial properties	Object properties
Expressing extent	Position Size	Grayscale
Differentiating marks	Orientation	Color Shape Texture

Fall 2014

CS 4460

38

# Back to Data



- 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

Fall 2014

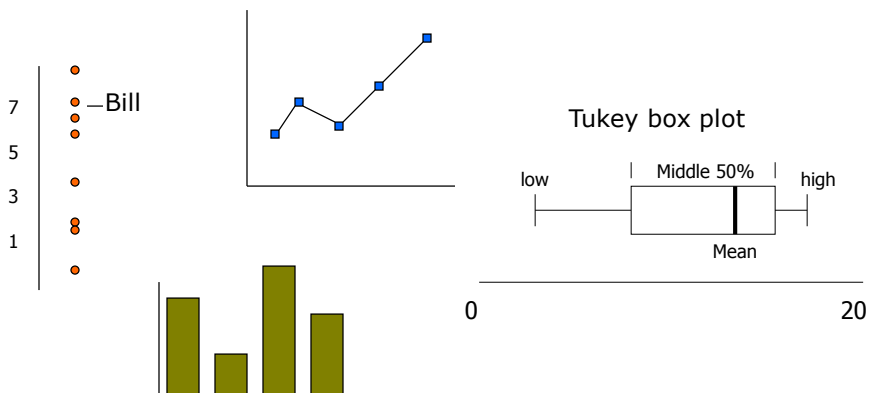
CS 4460

39

# Univariate Data



- Representations



Fall 2014

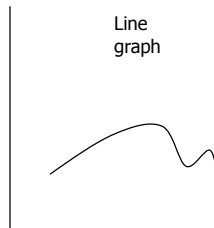
CS 4460

40

# What Goes Where?

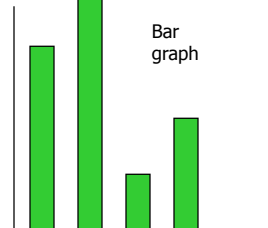


- In univariate representations, we often think of the data case as being shown along one dimension, and the value in another



Y-axis is quantitative variable

See changes over consecutive values



Y-axis is quantitative variable

Compare relative point values

Fall 2014

CS 4460

41

# Alternative View



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

Fall 2014

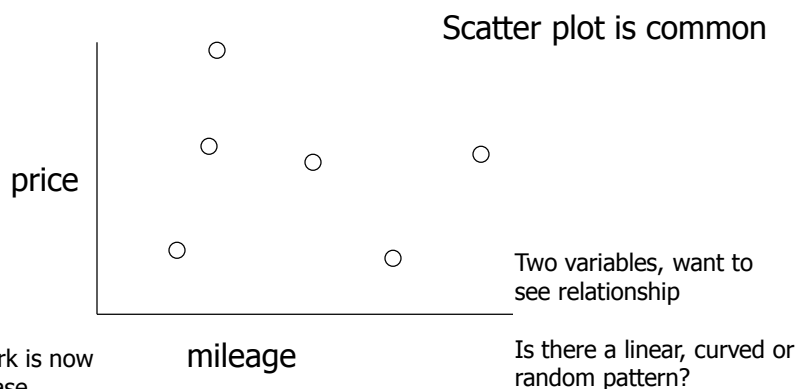
CS 4460

42

# Bivariate Data



- Representations



Each mark is now a data case

Fall 2014

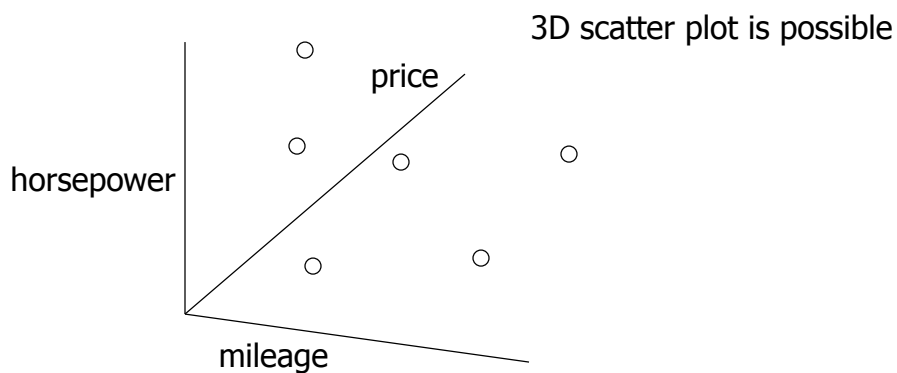
CS 4460

43

# Trivariate Data



- Representations

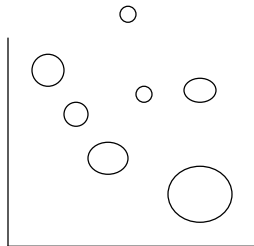


Fall 2014

CS 4460

44

# Alternative Representation



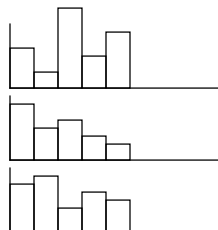
Still use 2D but have  
mark property  
represent third  
variable

Fall 2014

CS 4460

45

# Alternative Representation



Represent each variable  
in its own explicit way

Fall 2014

CS 4460

46

# Hypervariate Data



- Ahhh, the tough one
- 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

Fall 2014

CS 4460

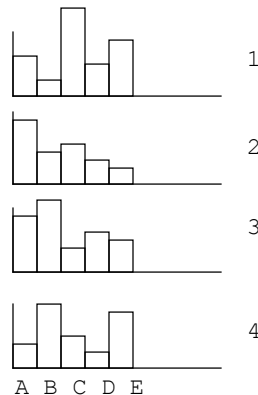
47

# 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



Fall 2014

CS 4460

48

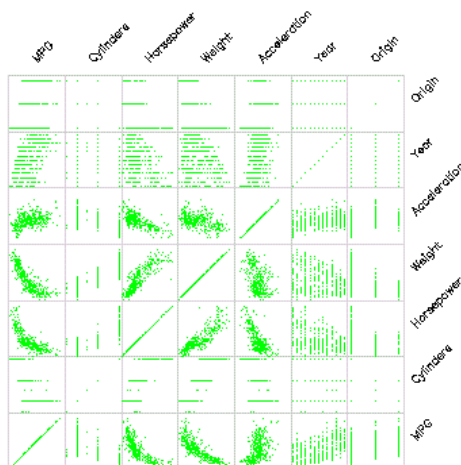


# Scatterplot Matrix



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

Useful for what?  
Misses what?



Fall 2014

CS 4460

49

## More to Come...



- Subsequent day will explore other general techniques for handling hypervariate data

Fall 2014

CS 4460

50

# Back to Graphs



- Design guidance
  - Few provides many helpful principles to design effective graphs

S Few  
"Effectively Communicating Numbers"  
[http://www.perceptualedge.com/articles/Whitepapers/Communicating\\_Numbers.pdf](http://www.perceptualedge.com/articles/Whitepapers/Communicating_Numbers.pdf)

Fall 2014

CS 4460

51

# Few's Selection & Design Process



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

Some  
examples...

Fall 2014

CS 4460

52

# 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

(Image shown in class)

## Multiple Bars



- Can be used to encode another variable

(Image shown in class)

Fall 2014

CS 4460

55

## Multiple Graphs



- Can distribute a variable across graphs too

(Image shown in class)

Sometimes called a  
trellis display

Fall 2014

CS 4460

56



# Examples

Fall 2014

CS 4460

57

## Before



You want to present quantitative sales performance data for the 4 regions of your company for the four quarters of the year

Fall 2014

CS 4460

58

# After?



(Image shown in class)

Fall 2014

CS 4460

59

# Before



(Image shown in class)

Fall 2014

CS 4460

60

# After?



(Image shown in class)

Fall 2014

CS 4460

61

# Before



(Image shown in class)

Fall 2014

CS 4460

62

# After?



(Image shown in class)

Fall 2014

CS 4460

63

# Before



(Image shown in class)

Fall 2014

CS 4460

64



# After?



(Image shown in class)

Fall 2014

CS 4460

65

# Book Recommendation



Loaded with examples of how to redesign ineffective tables and graphs

Fall 2014

CS 4460

66

# Advice



- Take DB & IR courses
  - Learn about query languages, relational data models, datacubes, data warehouses, ...

Fall 2014

CS 4460

67

# Project



- Keep working on topics...
  - Proposal due Sept. 16
  - More topics being added (AJC, ...)
- Things to watch out for
  - Nice HCI project, but not infovis
  - Is the dataset rich enough (enough variables)?

Fall 2014

CS 4460

68

# Upcoming



- S. Few's Design Guidance
  - Reading:  
*Now You See It*, chapters 5-12
- Multivariate Visual Representations 1
  - Reading:  
Munzner chapter 7

Fall 2014

CS 4460

69

# Sources Used



Few book  
CMS book  
Referenced articles  
Marti Hearst SIMS 247 lectures  
Kosslyn '89 article  
A. Marcus, *Graphic Design for Electronic Documents  
and User Interfaces*  
W. Cleveland, *The Elements of Graphing Data*

Fall 2014

CS 4460

70