

An Introduction to Information Visualization Techniques for Exploring Large Database

Jing Yang
Fall 2005

1

Motivation

Class 1

2

Data Explosion

- Between 1 and 2 exabytes of unique info produced per year
 - 100000000000000000 bytes
 - 250 meg for every man, woman and child
 - Printed documents only .003% of total

Peter Lyman and Hal Varian, 2000
Cal-Berkeley, Info Mgmt & Systems
www.sims.berkeley.edu/how-much-info

3

Data Overload

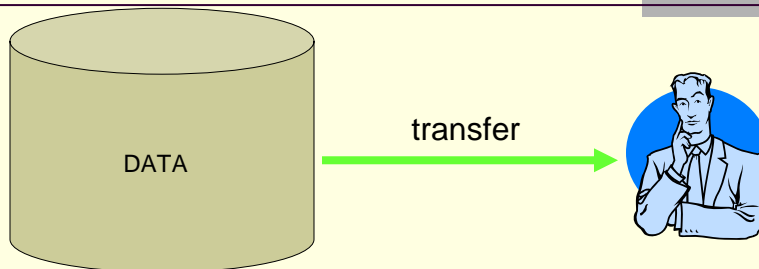
- Confound: How to make use of the data
 - How do we make sense of the data?
 - How do we harness this data in decision-making processes?
 - How do we avoid being overwhelmed?



- Dr. John Stasko, Slides of CS7500 at Gatech

4

Smell it, taste it?



- Vision: 100 MB/s – highest bandwidth sense
- Ears: <100 b/s
- Smell
- Taste

- Dr. John Stasko, Slides of CS7500 at Gatech

5

Visualization Makes Difference

- A simple experiment
Count the number of 3s in the following text:

124356428978301243256721352
453691263813797802183745902

6

Visualization Makes Difference

■ A simple experiment

Count the number of 3s in the following text:

124**3**56428978**3**0124**3**256721**3**52
45**3**69126**3**81**3**79780218**3**745902

7

The Need is There



In five years, 100 million people will be using an information-visualization tool on a near-daily basis. And products that have visualization as one of their top three features will earn \$1 billion per year.

Ramana Rao, founder and chief technology officer, Inxight Software Inc., Sunnyvale, Calif.

<http://www.computerworld.com/databasetopics/data/story/0,10801,80243,00.html>

- Dr. John Stasko, Slides of CS7500 at Gatech

8

Example 1 - Infocanvas

The Infocanvas project

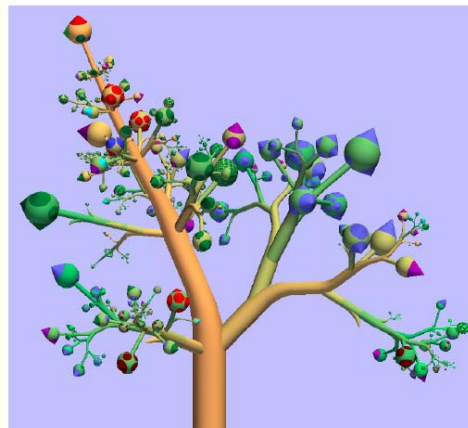
Team Members: [John Stasko](#), [Dave McColgin](#), [Todd Miller](#), [Chris Plaue](#), [Zach Pousman](#)

<http://www.cc.gatech.edu/gvu/ii/infoart/>

9

Example 2 – Botanical Tree

- The Unix home-directory of Dr. Kleiberg?



E. Kleiberg at.el. Infovis 2001

10

Example 3 – Stock Visualization

- Visualization 1:
 - Yahoo stock quotes for one stock
 - <http://finance.yahoo.com/>
- Visualization 2:
 - Smartmoney Map of Market
 - <http://www.smartmoney.com/marketmap/>

11

Example 4 – Home Finder

- Dynamic home finder
 - Human-Computer Interaction Lab / Univ. of Maryland
 - <http://www.cs.umd.edu/hcil/pubs/products.shtml>

12

Example 5 – Dynamic History

- Online American history textbook
 - <http://www.digitalhistory.uh.edu/timeline/timelineO.cfm>

13

Conclusion

- Information visualization techniques are used more and more in our daily life.

14

Terminology

15

Information Visualization

- Data and information
 - Human beings wish to be informed by data
 - Visualization tools facilitate the derivation of information (understanding, insight) from data.
- Information Visualization
 - To form a mental image or vision of ...?
 - To imagine or remember as if actually seeing?
 - To gain insight and understanding from data !
 - The purpose of visualization is insight, not pictures
 - Insight: discovery, decision making, explanation

16

Scientific Visualization vs. Information Visualization

- | | |
|------------------------------|----------------------------------|
| ■ Represents physical things | ■ Represents abstracted concepts |
| ■ Examples: | ■ Examples: |
| ■ Air flow over a wing | ■ Baseball statistics |
| ■ Stresses on a girder | ■ Stock trends |
| ■ Weather over Pennsylvania | ■ Connections between criminals |

17

Tasks in Info Vis

- Search
 - Finding a specific piece of information
 - Find an image with sky in it from an image collection
- Browsing
 - Look over or inspect something in a more casual manner, seek interesting information
 - Learn history of USA and find interesting events

18

Tasks in Info Vis

- Analysis
 - Comparison-Difference
Compare the house prices in different areas
 - Outliers, Extremes
Find the best stocks to investigate
 - Patterns
 - Prediction
 - Hypothesis generation and confirmation
- Monitoring
- Awareness

19

What's next?

- Multi-dimensional visualization

20

Multi-dimensional Visualization

21

Multi-dimensional (Multivariate) Dataset

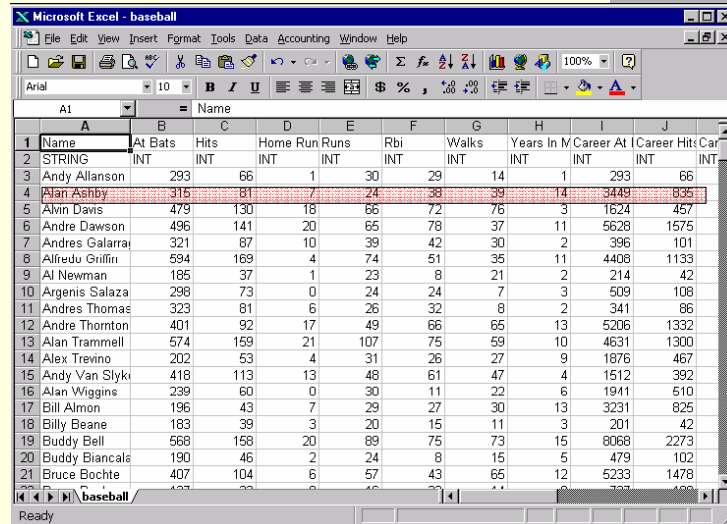
Microsoft Excel - baseball

	A	B	C	D	E	F	G	H	I	J	K
1	Name	At Bats	Hits	Home Run	Runs	Rbi	Walks	Years In	Career At	Career Hit	Career
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	610	
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	

22

- Dr. John Stasko, Slides of CS7500 at Gatech

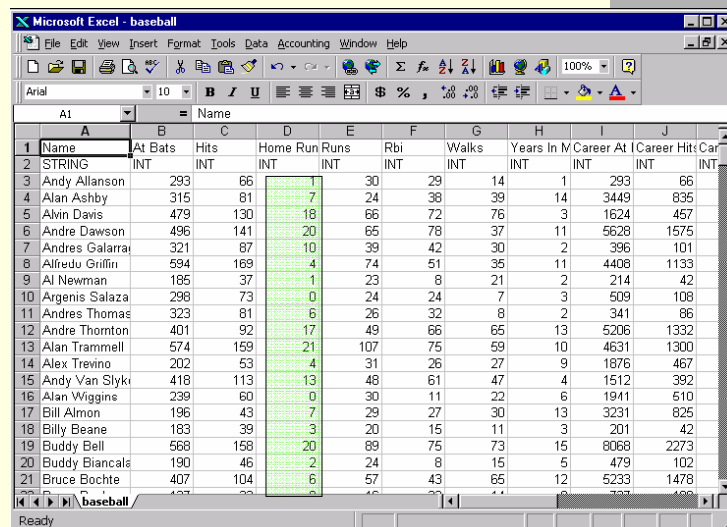
Data Item (Object, Record, Case)



1	Name	At Bats	Hits	Home Run	Runs	Rbi	Walks	Years In	Career At	Career Hit	Career
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 Slyki	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	

23

Dimension (Variable, Attribute)



1	Name	At Bats	Hits	Home Run	Runs	Rbi	Walks	Years In	Career At	Career Hit	Career
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 Slyki	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	

24

1-Dimensional Visualization

■ Discussion:

I have price information of 200 houses. Please find ways to visualization this dataset.

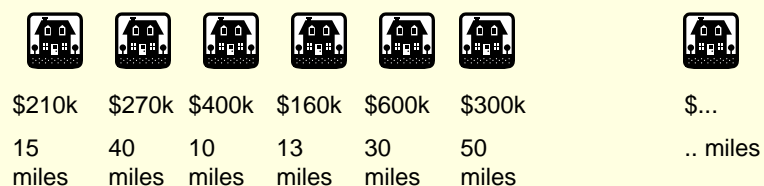


25

2-Dimensional Visualization

■ Discussion:

I also know the distances from the houses to UNCC. Please find ways to visualization both the distances and prices of the houses.



26

3-Dimensional Visualization

■ Discussion:

I also know the builders of the houses. Please find ways to visualization the distances, prices, and builders of the houses.



\$210k

15 miles

Sheahome



\$270k

40 miles

Weirland



\$400k

10 miles

Ryan



\$160k

13 miles

Ryan



\$600k

30 miles

Sheahome



\$...

.. Miles

..

27

Multidimensional Data

Example: Iris Data



- Scientists measured the sepal length, sepal width, petal length, petal width of many kinds of iris...

28

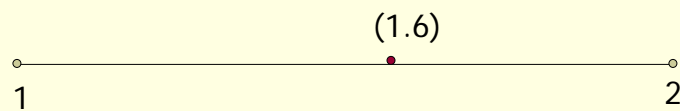
Multidimensional Data

Example: Iris Data

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
...
5.9	3	5.1	1.8

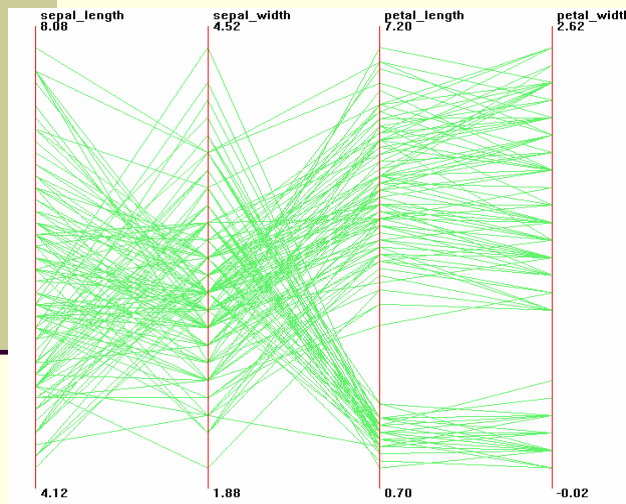
29

Recall 1-Dimensional Visualization



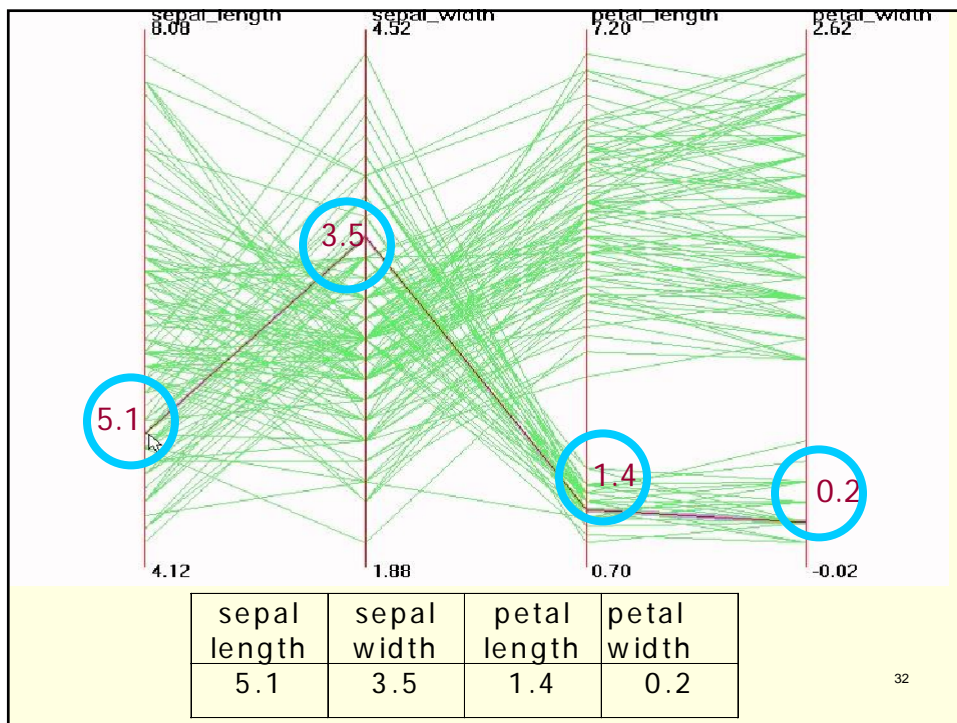
30

Parallel Coordinates



sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
...
5.9	3	5.1	1.8

31

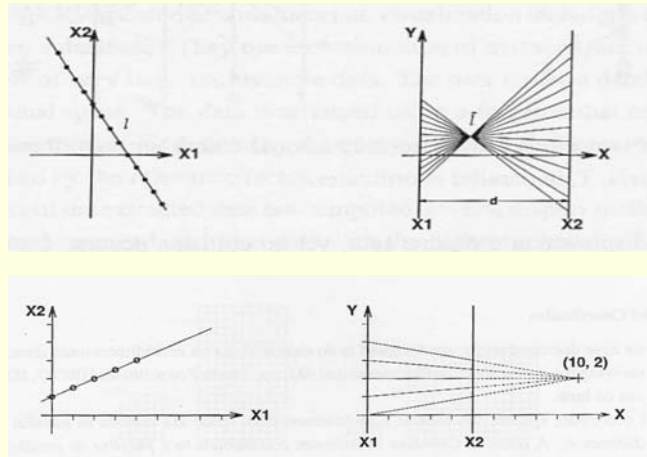


sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

32

Geometry of Data Items

■ The straight lines

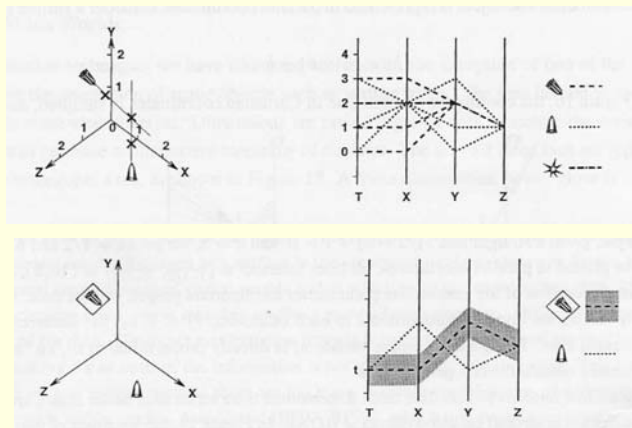


Pak Wong, 1997

33

Geometry of Data Items

■ The aircraft collision example



Pak Wong, 1997

34

Cluster and Outlier

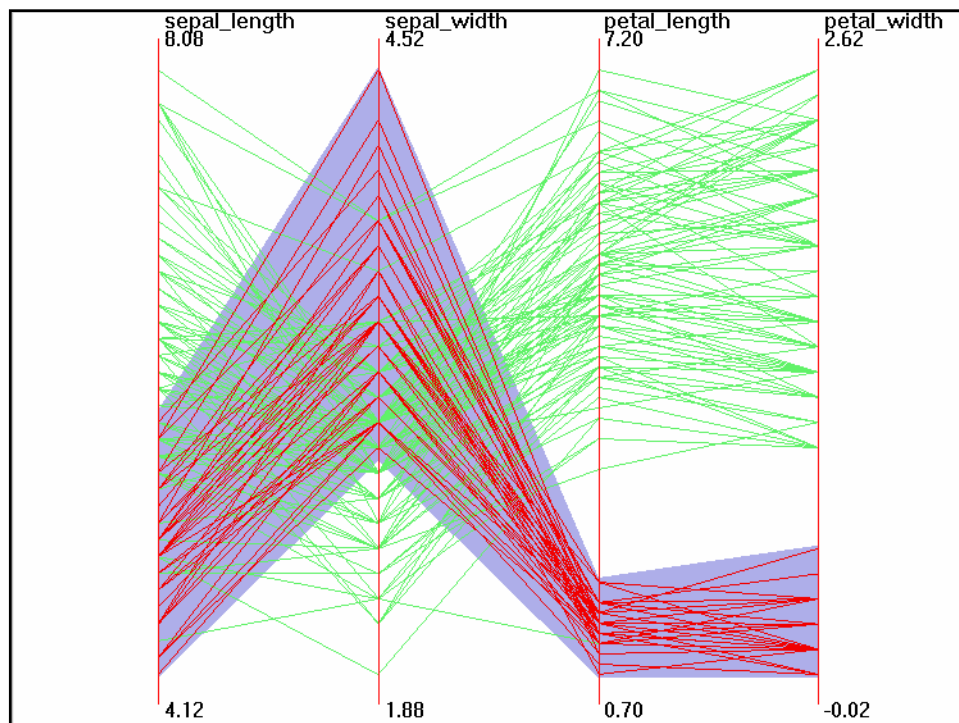
■ Cluster

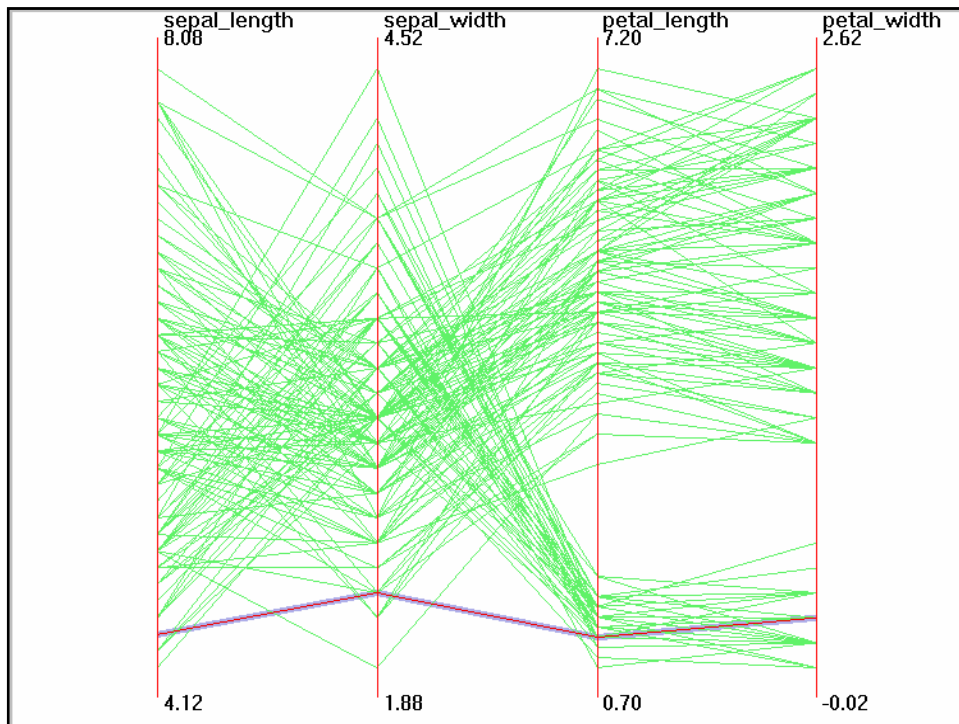
- A group of data items that are similar in all dimensions.

■ Outlier

- A data item that is similar to FEW or No other data items.

35

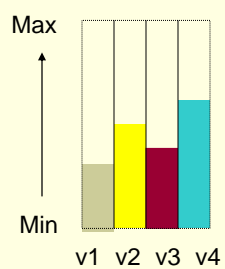




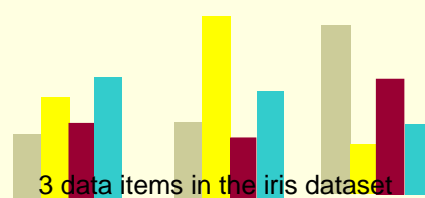
Glyphs

Profile Glyphs

Each bar encodes a variable's value



1 data item with
4 attributes

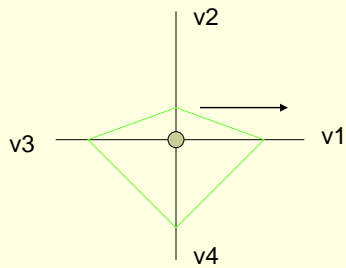


38

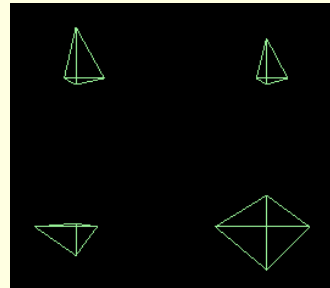
Glyphs

■ Star glyphs

- Space out variables at equal angles around a circle
- Each arm encodes a variable's value



1 data item with
4 attributes

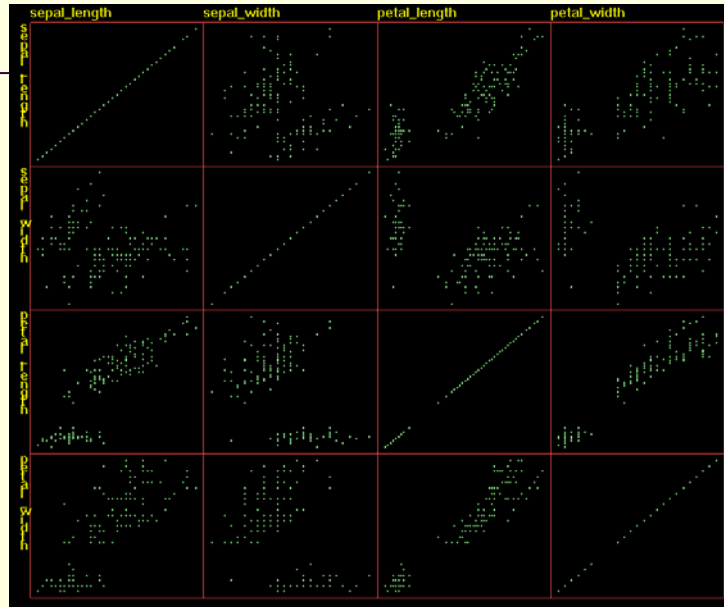


4 data items in the iris dataset ³⁹

Glyphs

- Many more. You find other glyphs -
Assignment

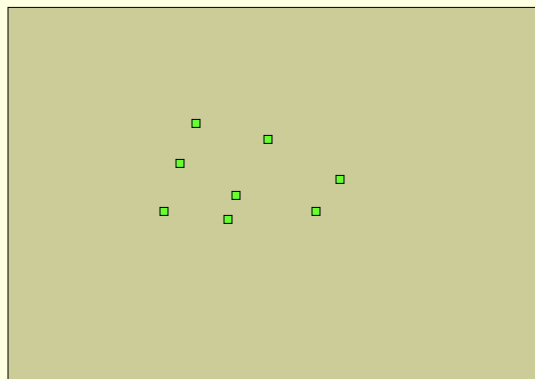
Scatterplot Matrix



41

Dimensional Stacking

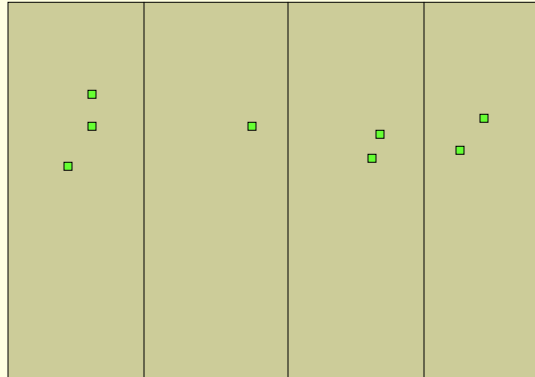
- Imagine each data item (4 attributes) as a small block. We place all blocks on a table.



42

Dimensional Stacking

- Add grids on the table. Place the blocks in the grids according to their values of attribute1.

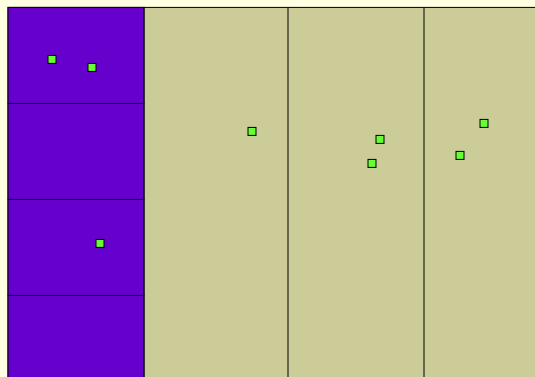


According to
values of
attribute 1

43

Dimensional Stacking

- Add grids in grids. Place the blocks in the grids according to their values of attribute2.



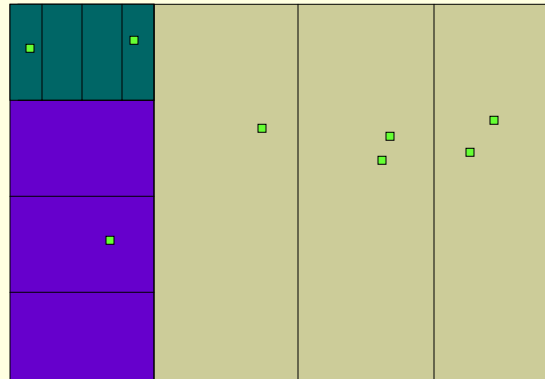
According to
values of
attribute 2

44

Dimensional Stacking

- Add grids in grids. Place the blocks in the grids according to their values of attribute3.

According to
values of
attribute 3

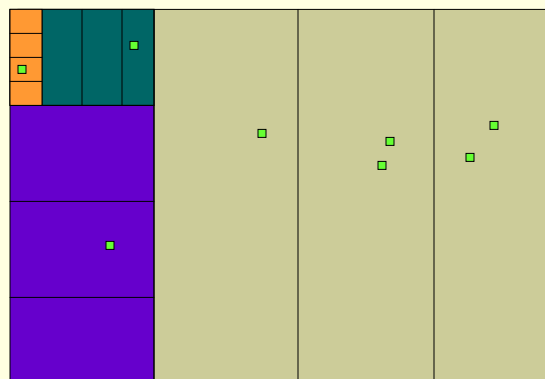


45

Dimensional Stacking

- Add grids in grids. Place the blocks in the grids according to their values of attribute4.

According to
values of
attribute 4

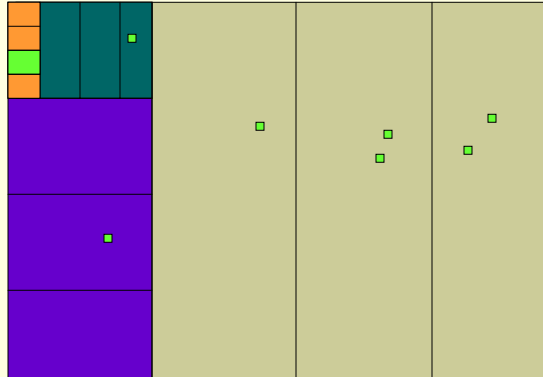


46

Dimensional Stacking

- Fix one block!

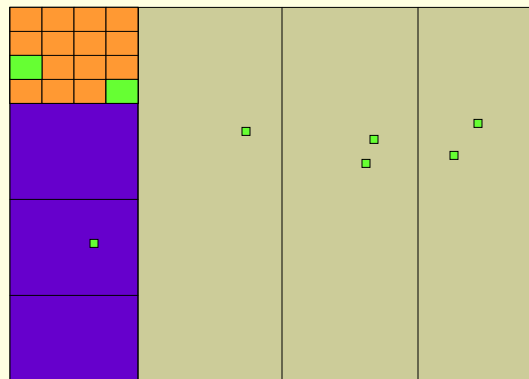
Expand
the block



47

Dimensional Stacking

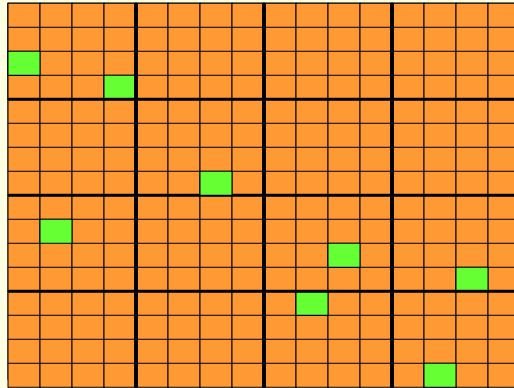
- Fix another block



48

Dimensional Stacking

- Dimensional stacking!



49