02

Data Foundations

Notice

- **Author**
  - **João Moura Pires (jmp@fct.unl.pt)**

Bibliography

- **Many examples are extracted and adapted from**
  - **Interactive Data Visualization: Foundations, Techniques, and Applications,**

    Matthew O. Ward, Georges Grinstein, Daniel Keim, 2015
  - **Visualization Analysis & Design,**

    Tamara Munzner, 2015

Table of Contents

Some practical Information

Important dates

- Home
- News
- Information
  - Bibliography
  - Sylabus
  - Evaluation Rules
  - Schedule
- Resources
- Summaries
- Training
- Evaluation

Subscribe this calendar:

ICAL

IDV

Today | Friday, March 9 | Print | Week | Month | Agenda

Showing events after 1/1. Look for earlier events

Friday, March 9

4:00pm Teorica de VID

6:00pm Prática de VID

Friday, March 16

4:00pm Teorica de VID

6:00pm Prática de VID

Friday, March 23

4:00pm Teorica de VID

6:00pm Prática de VID

Wednesday, March 28

4:00pm Teorica de VID

6:00pm Prática de VID

Friday, March 30

Team Registration

Friday, April 6

4:00pm Teorica de VID

6:00pm Prática de VID

Friday, April 13

4:00pm Teorica de VID

6:00pm Prática de VID

Monday, April 16

Paper

Friday, April 20

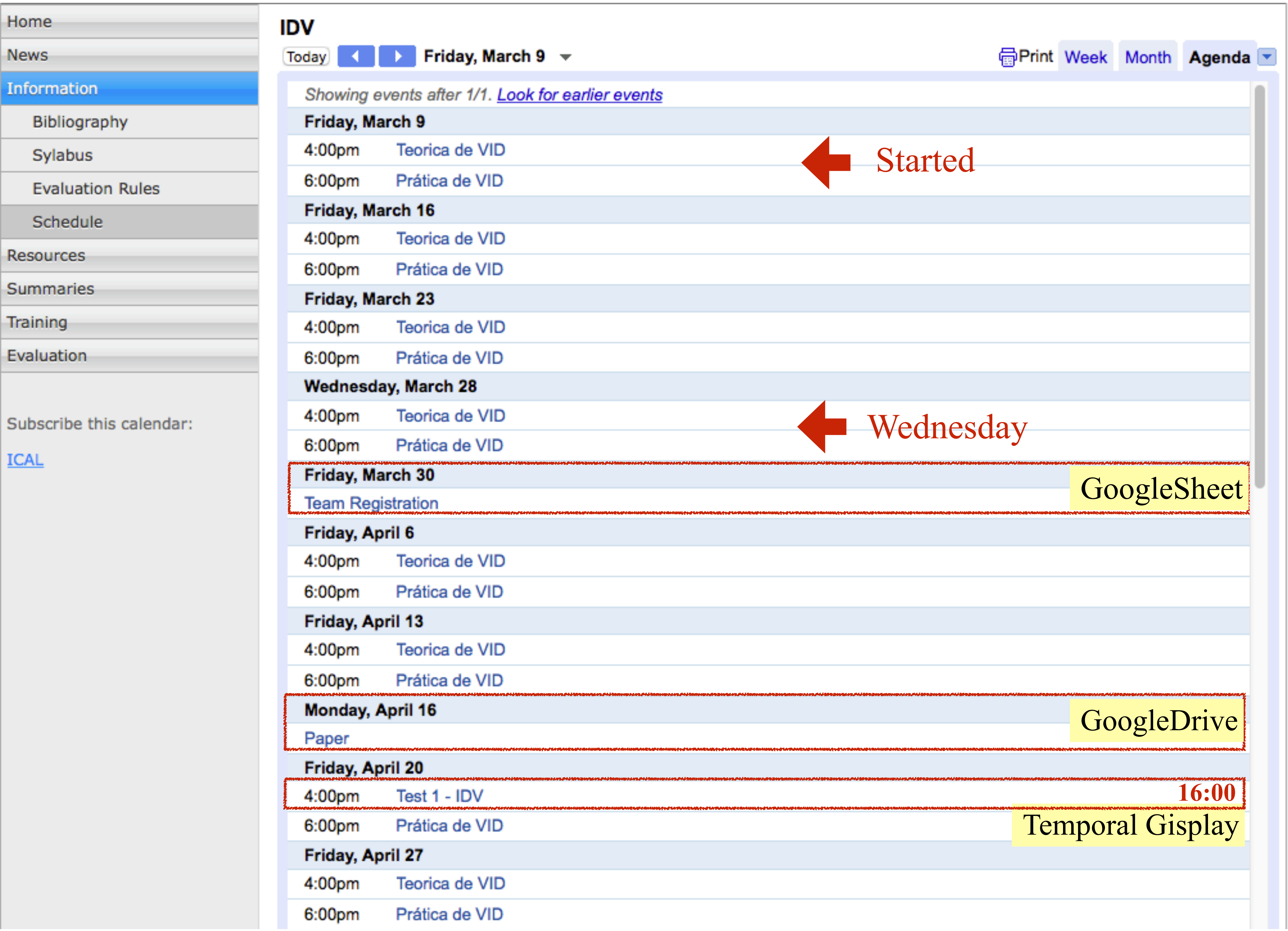
4:00pm Test 1 - IDV

6:00pm Prática de VID

Friday, April 27

4:00pm Teorica de VID

6:00pm Prática de VID


Home
News
Information
Bibliography
Sylabus
Evaluation Rules
Schedule
Resources
Summaries
Training
Evaluation
Subscribe this calendar:
ICAL
IDV
Today
Friday, March 9
Print
Week
Month
Agenda
Showing events after 1/1. Look for earlier events
Friday, March 9
4:00pm Teorica de VID
6:00pm Prática de VID
Started
Friday, March 16
4:00pm Teorica de VID
6:00pm Prática de VID
Friday, March 23
4:00pm Teorica de VID
6:00pm Prática de VID
Wednesday, March 28
4:00pm Teorica de VID
6:00pm Prática de VID
Wednesday
Friday, March 30
Team Registration
GoogleSheet
Friday, April 6
4:00pm Teorica de VID
6:00pm Prática de VID
Friday, April 13
4:00pm Teorica de VID
6:00pm Prática de VID
Monday, April 16
Paper
GoogleDrive
Friday, April 20
4:00pm Test 1 - IDV
16:00
6:00pm Prática de VID
Temporal Gisplay
Friday, April 27
4:00pm Teorica de VID
6:00pm Prática de VID

- Home
- News
- Information
  - Bibliography
  - Sylabus
  - Evaluation Rules
  - Schedule
- Resources
- Summaries
- Training
- Evaluation

Subscribe this calendar:

ICAL

IDV

Today ◀ ▶ Friday, April 27 ▼ Print Week Month Agenda

Friday, April 27

4:00pm Teorica de VID

6:00pm Prática de VID

Friday, May 4

4:00pm Teorica de VID

6:00pm Prática de VID

Friday, May 11

4:00pm Teorica de VID

6:00pm Prática de VID

Friday, May 18

4:00pm Teorica de VID

6:00pm Prática de VID

Friday, May 25

4:00pm Teorica de VID

6:00pm Prática de VID

Friday, June 1

4:00pm Teorica de VID

6:00pm Prática de VID

Team work support

Friday, June 8

4:00pm Test 2 - IDV — 16:00

6:00pm Prática de VID — Team work support

Sunday, June 10

Code and Implementation — GoogleDrive

Monday, June 11

IDV Discussions

Tuesday, June 12

IDV Discussions

With previous appointment

Wednesday, June 13

IDV Discussions

Thursday, June 14

IDV Discussions

Friday, June 15

IDV Discussions

Recap from previous lecture

Further Reading

- **Recommend Readings**
  - Interactive Data Visualization: Foundations, Techniques, and Applications, Matthew O. Ward et all, 2015, pages 1 - 38.
- **Supplemental readings:**
  - Cholera map's John Snow:
    - https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak
  - Napoleon
    - https://en.wikipedia.org/wiki/Charles_Joseph_Minard
  - William Playfair:
    - https://en.wikipedia.org/wiki/William_Playfair
  - Florence Nightingale:
    - https://pt.wikipedia.org/wiki/Florence_Nightingale
  - Periodic table:
    - https://en.wikipedia.org/wiki/Periodic_table

Check - vis25timeline

What you should know

- **What is Data Visualization.**
  - Understanding the data => take decisions
- **Data Visualization can be extremely powerful**
  - Uncover new patterns; confirm hypothesis;
- **Why Visualization is important.**
  - Stats not enough; communication needs; exploratory needs
- **Key aspects of today Visualizations.**
  - Interactions; visual abstractions; multiple (linked) visualizations.
- **The general steps of a Visualization Process**
  - **Raw data -> data -> viz structures -> images -> perception + feedback**
- **The role of Perception.**
  - The role and the importance of the user.

Evaluation rules

- **Two mid-term written individual tests (25% each)**
- **One project (for team of two students), with several phases:**
  - **Specification**
  - **Paper (20%)**
  - **Code/implementation (30%)**
  - **(*) an oral discussion will be required to validate the project components**
- **Course approval requires the following minimal grades:**
  - **(mean (Test1; Test2) >= 10) AND (Test1 >= 8) AND (Test2 >= 8)**
  - **(mean(Paper;Code&Implementation) >= 10) AND**
- **Final exam may replace mean (Test1; Test2) if project is approved.**

Recommended Actions

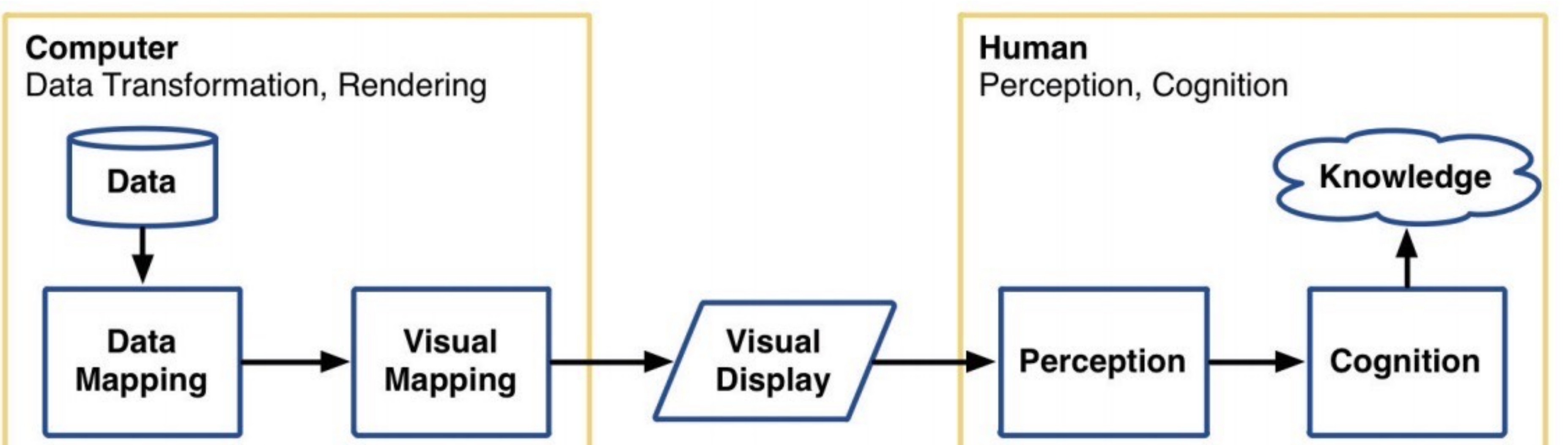
http://vid.ssdi.di.fct.unl.pt/

- **Read the available information on the Web Site**
- **Update your calendar (or subscribe the calendar)**
  - **<u>VID RSS Feed</u>**
- **Find a partner for your team work**
  - **Make the registration until March 30th**
- **Check the <u>Summaries section</u> and follow its recommendations**
- **Install Tableau software**
  - **<u>http://www.tableau.com/academic/students</u>**

Introduction to Data Foundations

Visualization Process: visualization pipeline

- For visualization the stages are:
  - Modeling: the data to be visualized
  - Data Selection: similar to clipping
  - Data to visual mappings: the heart of the visualization is mapping data values to graphical entities or their attributes; may involve scaling, shifting, filtering, interpolating, or subsampling.
  - Scene parameter setting: (ex: color mapping)
  - Rendering or generation of the visualization



Data: Sources

- **Sources**
  - **Sensors;**
  - **Surveys;**
  - **Simulations;**
  - **Computations;**
  - **Log of human and machine activity**

- **Raw versus Processed data**
  - **Raw data (untreated)**
  - **Processed: smoothing, noise removal, scaling, interpolation, aggregation**

Data: typical data set in visualization

- **List of *n* records**
  - $(r_1, r_2, \ldots, r_n)$
  - **a record r_i consists in *m* (one or more) observations or variables**

    $(v_1, v_2, \ldots, v_m)$
  - **one observation may be:**
    - **a single number / symbol / string**
    - **a more complex structure**
  - **A variable may be classified as:**
    - **independent: whose value is not controlled or affected by another variable**
    - **dependent: whose value is affected by the variation in one or more associated independent variables**

Data: typical data set in visualization

- A record *r* consists in *mi* independent variables and *md* dependent variables

$$r = (iv_1, iv_2, \ldots, iv_{mi}, dv_1, dv_2, \ldots, dv_{md})$$

- We **may not know** which variables are dependent and which are independent.
- In general a data set will **not contain an exhaustive list of all possible combinations of values** for the independent variables

- A data set can be seen as a function

Domain of Independent variables → F → **Range of dependent variables**

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Data

(Matthew O. Ward, et all)

Data Types

Types of data. Numeric versus Non-Numeric

- In its **simplest form each variable of a record has a single piece of information** (scalar values)

- **Numeric (ordinal):**
  - binary: assuming only the values 0 and 1;
  - discrete: integer values or from a specific subset (e.g., (2, 4, 6, 8, 10);
  - continuous: representing real values (e.g., [0, 100]).
- **Non Numeric (nominal):**
  - categorial: finite (normally short) list of values (e.g., red, green, blue);
  - ranked: a categorial variable that has an implied order (e.g., small, medium, large);
  - arbitrary: potentially infinite range of values (e.g., names, addresses).

Types of data. Type of scale

- **Properties of scales of measurement:**
  - **Identity. Each value on the measurement scale has a unique meaning.**
  - **Magnitude. Values on the measurement scale have an ordered relationship to one another. That is, some values are larger and some are smaller.**
  - **Equal intervals. Scale units along the scale are equal to one another. This means, for example, that the difference between 1 and 2 would be equal to the difference between 19 and 20. This is also know as distance metric.**
  - **A minimum value of zero. The scale has a true zero point, below which no values exist. When a scale has an absolute zero then it makes sense to apply all the mathematical operations (+, -, *, /).**

Types of data. Type of scale

- Nominal Scale of Measurement:
  - Only satisfies the identity property of measurement
  - Categorial and Arbitrary(*)
- Ordinal Scale of Measurement:
  - Has the property of both identity and magnitude
  - Ranked (and all the numeric)
- Interval Scale of Measurement
  - Has the properties of identity, magnitude, and equal intervals.
  - Discrete. e.g., Fahrenheit (or centigrade) scale to measure temperature
- Ratio Scale of Measurement
  - Satisfies identity, magnitude, equal intervals, and a minimum value of zero.
  - Continuous. e.g., weight, distance, etc. Can apply operations of / and *.

Structure within and between records

Data sets structure

- **The structure of a data set defines:**
  - **Syntactical rules**
  - **The relationships between the components within a record**
  - **The relationship between records**

Scalar, Vector and Tensor

- **Scalar: individual value in a data record.**
  - **e.g.: Age; Color; Weight**
- **Vector: multiple variables in a single record can represent a single item**
  - **e.g.: Position coordinates (2D or 3D); Color using RGB(Red, Green, Blue) components, Phone number (Country code, area code and local number), etc.**
  - **each component (of the vector) can be considered individually but is most appropriate to treat the vector as a whole.**
- **Tensor: a tensor is defined by its *rank* and its *dimensionality.* A scalar is a tensor of rank 0; a vector with *D* components is a tensor of rank 1 and D dimensionality. A tensor of rank 2 and 3 dimensions can be represented as a Matrix 3 x 3.**

More info about tensors -> https://www.youtube.com/watch?v=fu-eMNi_aag

Geometry and Grids

- **Geometry via explicit coordinates for each record in the data set.**
  - **Data set about fires in Portugal. Associated to each fire a coordinate of the starting point;**
  - **Data set about temperature readings from sensors and associated with all the information sensor’s coordinates.**
  - **Data set describing 3D world. The geometry concept is the majority of the data.**
  - **Census data set which associates the data to administrative regions**
- **Geometric structure is implied and it is assumed some form of grid. Successive data records are located at successive positions. It requires to set the starting point, the directions and the step size for each dimension.**
  - **Satellite images.**

Other forms of structure

- **Time**
  - **Present in many data sets**
  - **Uniformly spaced versus non-uniformly spaced**
  - **Relative versus absolute**
  - **Local versus Universal time**
  - **Seen as linear versus as cyclic**
- **Topology**
  - **How the records are connected.**
  - **Geometry and space (spatial neighbors)**
  - **Hierarchy and graphs**
  - **This form of structure can be explicitly included in the data record or as an auxiliary data structure**

http://www.timeviz.net
check to see so many visualization techniques for Time-Oriented Data

Examples

MRI (magnetic resonance imagery). Density (scalar), with three spatial attributes, 3D grid connectivity;

CFD (computational fluid dynamics). Three dimensions for displacement, with one temporal and three spatial attributes, 3D grid connectivity (uniform or nonuniform);

Financial. No geometric structure, n possibly independent components, nominal and ordinal, with a temporal attribute;

CAD (computer-aided design). Three spatial attributes with edge and polygon connections, and surface properties;

Remote sensing. Multiple channels, with two or three spatial attributes, one temporal attribute, and grid connectivity;

Census. Multiple fields of all types, spatial attributes (e.g., addresses), temporal attribute, and connectivity implied by similarities in fields;

Social Network. Nodes consisting of multiple fields of all types, with various connectivity attributes that could be spatial, temporal, or dependent

Interactive Data Visualization: Foundations, Techniques, and Applications, Matthew O. Ward, Georges Grinstein, Daniel Keim, 2015

Data

(Tamara Munzner)

Data Types and Dataset Types

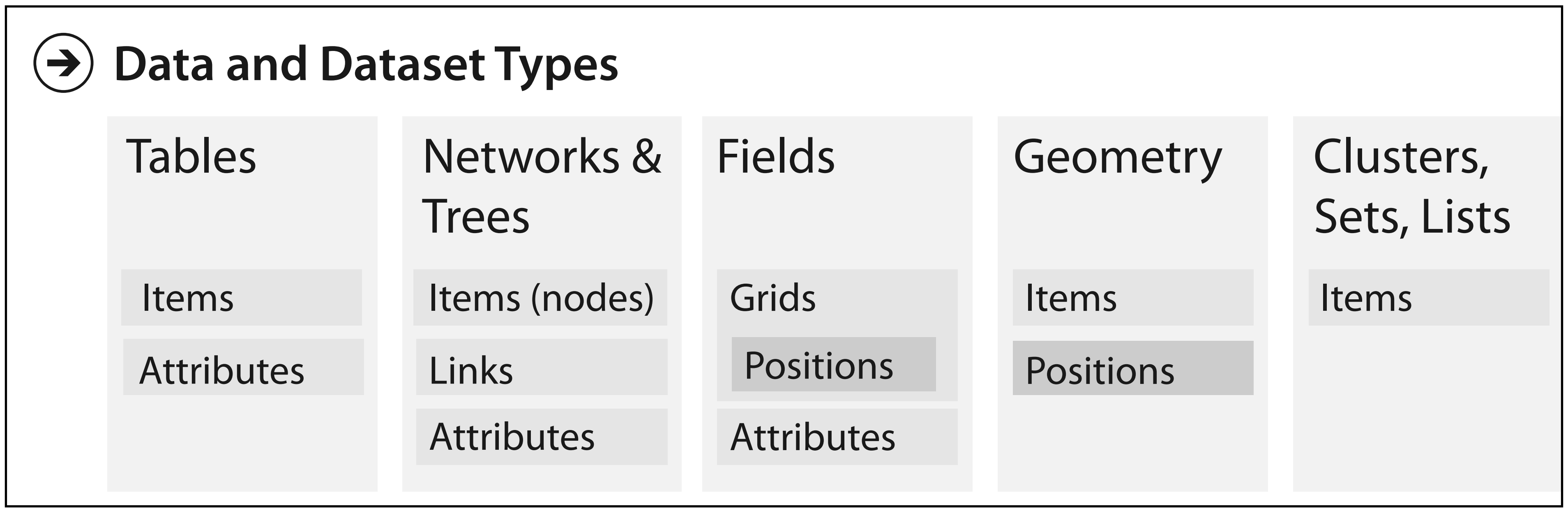
- **Data Types**

Data Types				
➔ Items	➔ Attributes	➔ Links	➔ Positions	➔ Grids

- An **attribute** is some specific property that can be measured, observed, or logged.*
- An **item** is an individual entity that is discrete, such as a row in a simple table or a node in a network
- A **link** is a relationship between items, typ- ically within a network.
- A **grid** specifies the strategy for sampling continuous data in terms of both geometric and topological relationships between its cells
- A **position** is spatial data, providing a location in two-dimensional (2D) or three-dimensional (3D) space.

Data Types and Dataset Types

Dataset Types

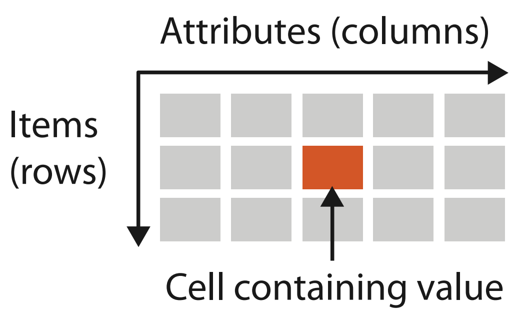


- A **dataset** is any collection of information that is the target of analysis.
- **Other ways** to group items together include **clusters**, **sets**, and **lists**.
- In real-world situations, complex combinations of these basic types are common.

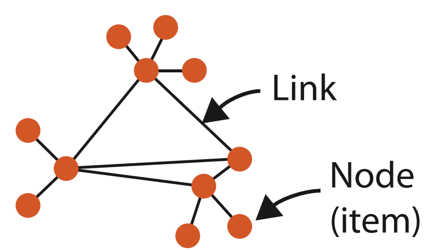
Data Types and Dataset Types

Dataset Types

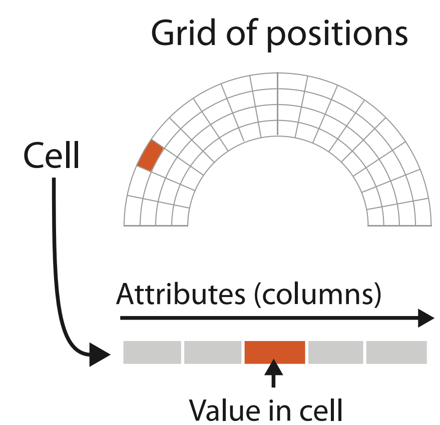
- Tables



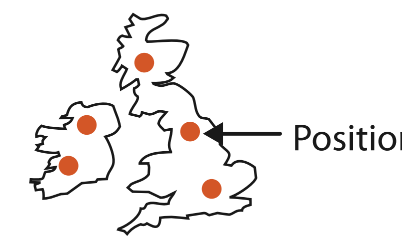
- Networks



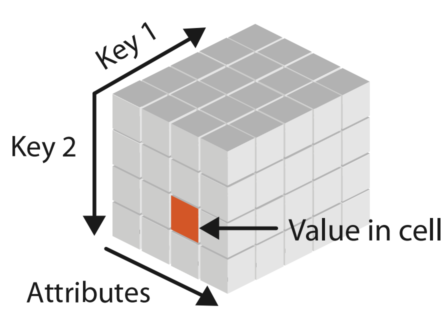
- Fields (Continuous)



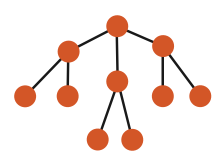
- Geometry (Spatial)



- *Multidimensional Table*

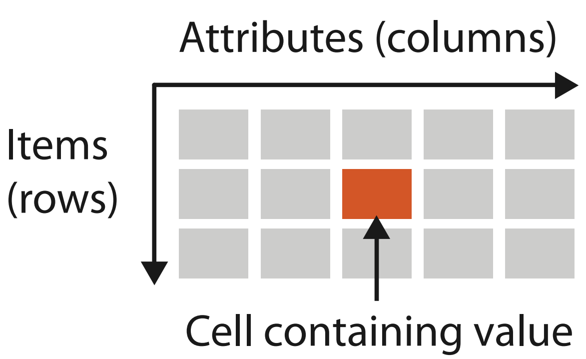


- *Trees*

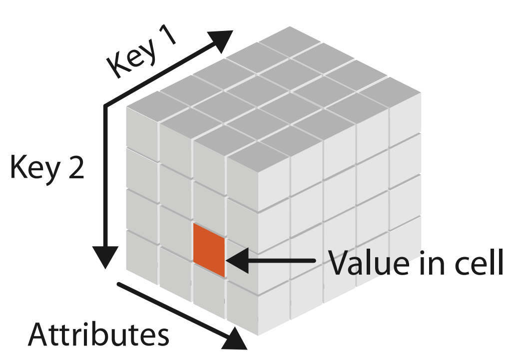


Dataset Types: **Table**

➔ Tables



➔ *Multidimensional Table*



A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	[illegible]	7/17/07
32	7/16/07	2-High	Medium Box	[illegible]	7/18/07
32	7/16/07	2-High	Medium Box	[illegible]	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	[illegible]	4-Not Specified	Small Pack	0.44	6/6/05
69	[illegible]	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

attribute

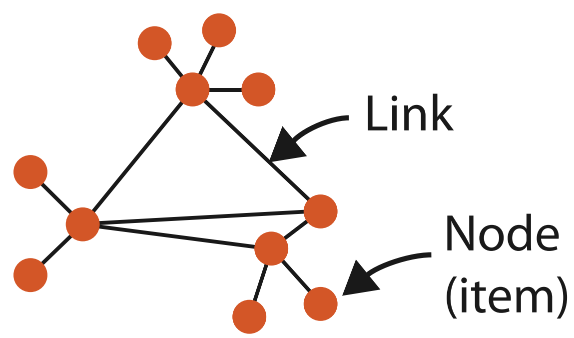
item

cell

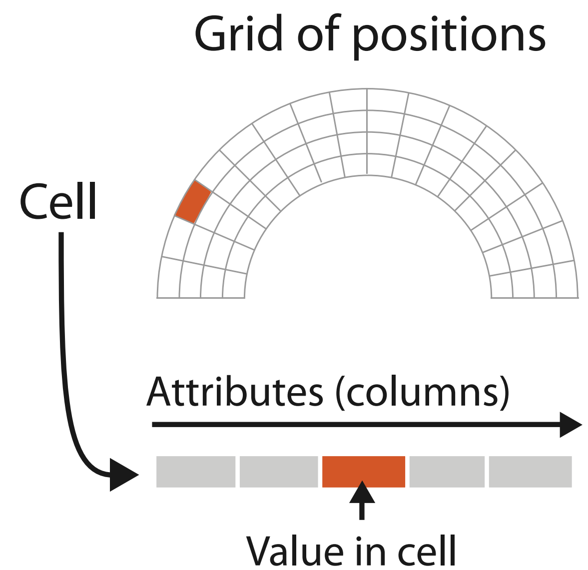
A **multidimensional table** has a more complex structure for indexing into a cell, with multiple keys.

Data Types and Dataset Types

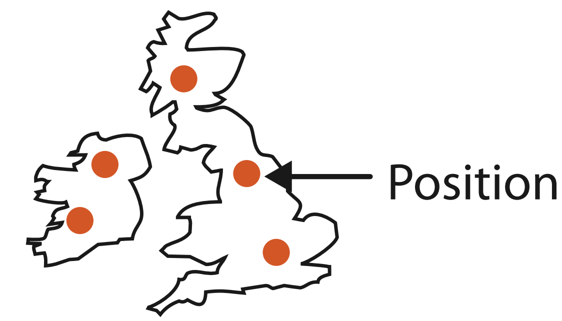
→ Networks



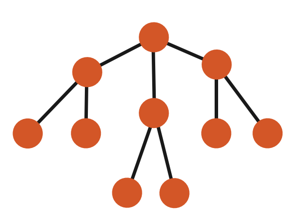
→ Fields (Continuous)



→ Geometry (Spatial)



→ *Trees*



The **field** dataset type also contains attribute values associated with cells.

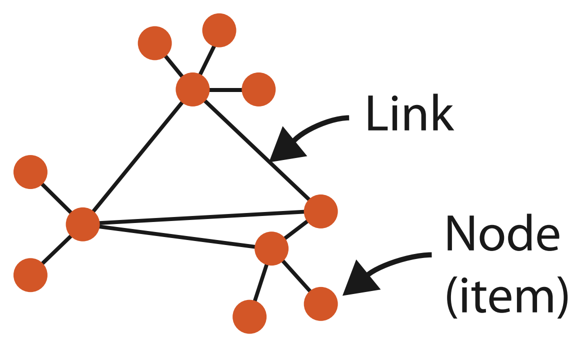
Each **cell** in a field contains measurements or calculations from a **continuous** domain

Continuous data requires careful treatment that takes into account the mathematical questions of **sampling** data **interpolation**

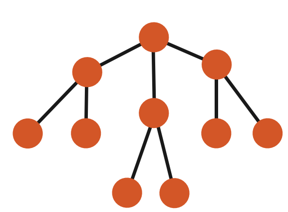
scientific visualization

Data Types and Dataset Types

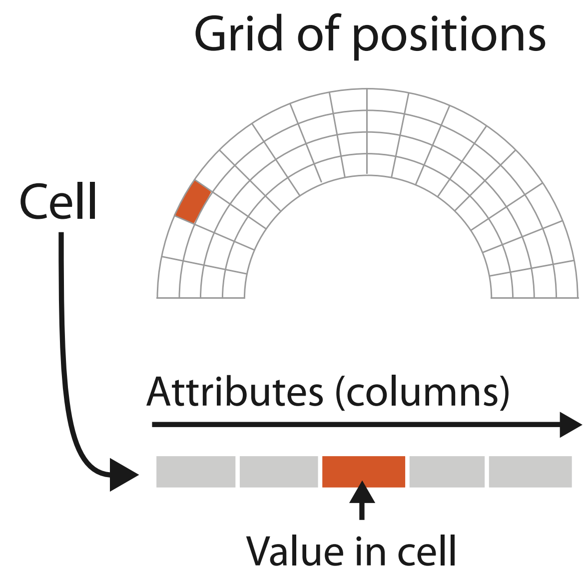
- ➔ Networks



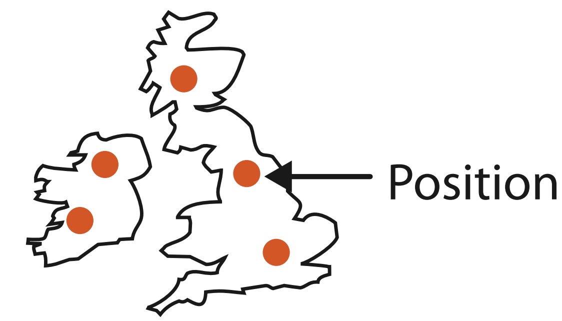
- ➔ *Trees*



- ➔ Fields (Continuous)



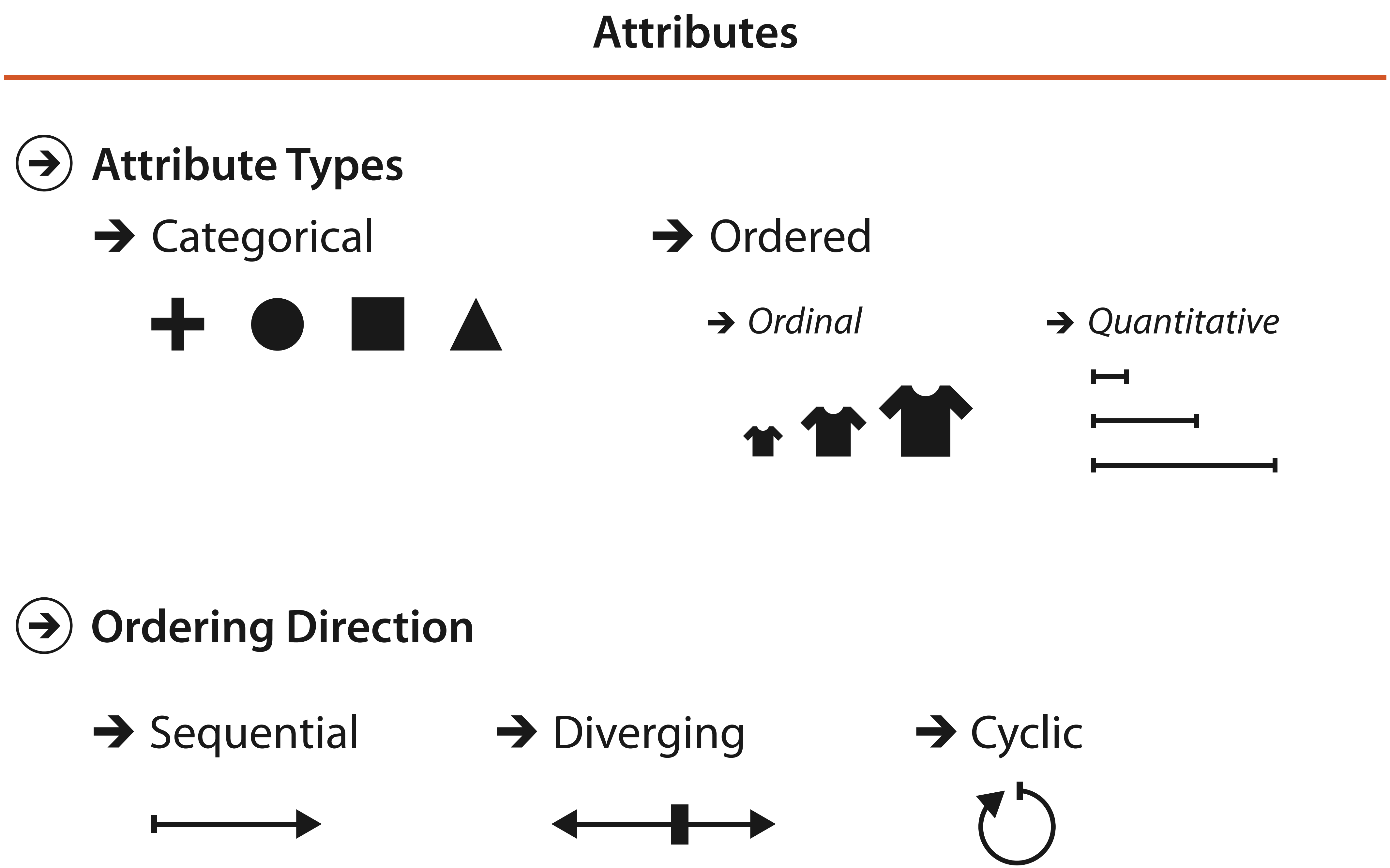
- ➔ Geometry (Spatial)



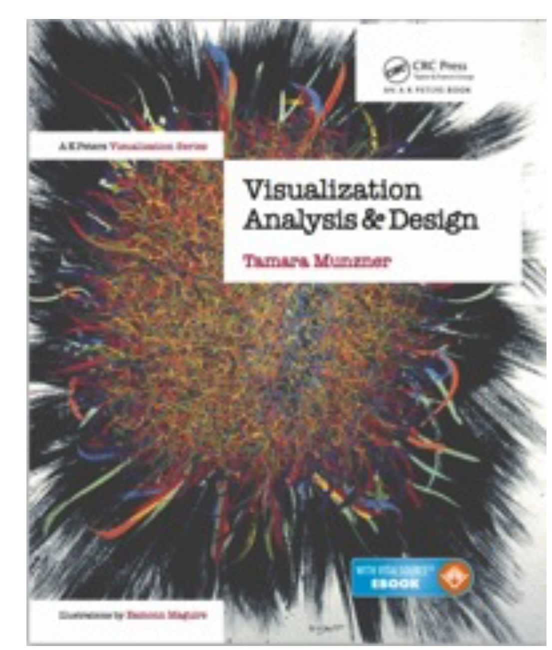
The problem of how to **create images from a geometric description** of a scene falls into another domain: **computer graphics**.

Simply showing a geometric dataset is not an interesting problem from the point of view of a vis designer.

Attribute Types



Tamara Munzner



Data Preprocessing

Data Preprocessing

- **Metadata**
- **Basic statistics about the (scalar) data**
- **Missing Values and Data Cleansing**
- **Normalization**
- **Dimension reduction**
- **Mapping Nominal Dimensions to Numbers**
- **Other data processing topics**

Metadata

- **Sample from the cars data set**

Acura 3.5 RL 4dr	1	0	0	0	0	0	0	0	43755	39014	3,5	6	225	18	24	3880	115	197	72
Acura 3.5 RL w/Navigation 4dr	1	0	0	0	0	0	0	0	46100	41100	3,5	6	225	18	24	3893	115	197	72
Acura MDX	0	0	1	0	0	0	1	0	36945	33337	3,5	6	265	17	23	4451	106	189	77
Acura NSX coupe 2dr manual S	0	1	0	0	0	0	0	1	89765	79978	3,2	6	290	17	24	3153	100	174	71
Acura RSX Type S 2dr	1	0	0	0	0	0	0	0	23820	21761	2	4	200	24	31	2778	101	172	68
Acura TL 4dr	1	0	0	0	0	0	0	0	33195	30299	3,2	6	270	20	28	3575	108	186	72
Acura TSX 4dr	1	0	0	0	0	0	0	0	26990	24647	2,4	4	200	22	29	3230	105	183	69
Audi A4 1.8T 4dr	1	0	0	0	0	0	0	0	25940	23508	1,8	4	170	22	31	3252	104	179	70
Audi A4 3.0 4dr	1	0	0	0	0	0	0	0	31840	28846	3	6	220	20	28	3462	104	179	70
Audi A4 3.0 convertible 2dr	1	0	0	0	0	0	0	0	42490	38325	3	6	220	20	27	3814	105	180	70
Audi A4 3.0 Quattro 4dr auto	1	0	0	0	0	0	1	0	34480	31388	3	6	220	18	25	3627	104	179	70

- **With the exception of first column (Vehicle name) we need more information!**

Vehicle Name	Small/Sporty/ Compact/Large Sedan	Sports Car	SUV	Wagon	Minivan	Pickup	AWD	RWD	Retail Price	Dealer Cost	Engine Size (l)	Cyl	HP	City MPG	Hwy MPG	Weight	Wheel Base	Len	Width
Acura 3.5 RL 4dr	1	0	0	0	0	0	0	0	43755	39014	3,5	6	225	18	24	3880	115	197	72
Acura 3.5 RL w/Navigation 4dr	1	0	0	0	0	0	0	0	46100	41100	3,5	6	225	18	24	3893	115	197	72
Acura MDX	0	0	1	0	0	0	1	0	36945	33337	3,5	6	265	17	23	4451	106	189	77
Acura NSX coupe 2dr manual S	0	1	0	0	0	0	0	1	89765	79978	3,2	6	290	17	24	3153	100	174	71
Acura RSX Type S 2dr	1	0	0	0	0	0	0	0	23820	21761	2	4	200	24	31	2778	101	172	68
Acura TL 4dr	1	0	0	0	0	0	0	0	33195	30299	3,2	6	270	20	28	3575	108	186	72
Acura TSX 4dr	1	0	0	0	0	0	0	0	26990	24647	2,4	4	200	22	29	3230	105	183	69
Audi A4 1.8T 4dr	1	0	0	0	0	0	0	0	25940	23508	1,8	4	170	22	31	3252	104	179	70
Audi A4 3.0 4dr	1	0	0	0	0	0	0	0	31840	28846	3	6	220	20	28	3462	104	179	70
Audi A4 3.0 convertible 2dr	1	0	0	0	0	0	0	0	42490	38325	3	6	220	20	27	3814	105	180	70
Audi A4 3.0 Quattro 4dr auto	1	0	0	0	0	0	1	0	34480	31388	3	6	220	18	25	3627	104	179	70
Audi A4 3.0 Quattro 4dr manual	1	0	0	0	0	0	1	0	33430	30366	3	6	220	17	26	3583	104	179	70
Audi A4 3.0 Quattro convertible 2dr	1	0	0	0	0	0	1	0	44240	40075	3	6	220	18	25	4013	105	180	70

- **With the column names it is much better but it is not enough !**

Metadata

Associated Metadata

```
NAME:   2004 New Car and Truck Data
TYPE:   Sample
SIZE:   428 observations, 19 variables

DESCRIPTIVE ABSTRACT:
Specifications are given for 428 new vehicles for the 2004 year. The variables recorded include price, measurements
relating to the size of the vehicle, and fuel efficiency.

SOURCE:
_Kiplinger's Personal Finance_, December 2003, vol. 57, no. 12, pp. 104-123, http:/www.kiplinger.com (permission to post on
the JSE Web site kindly granted by PARS International Corporation, 102 West 38th Street, New York, NY 10018)

VARIABLE DESCRIPTIONS:

Columns Variables
  1- 45 Vehicle Name
 47             Sports Car? (1=yes, 0=no)
 49             Sport Utility Vehicle? (1=yes, 0=no)
 51             Wagon? (1=yes, 0=no)
 53             Minivan? (1=yes, 0=no)
 55             Pickup? (1=yes, 0=no)
 57             All-Wheel Drive? (1=yes, 0=no)
 59             Rear-Wheel Drive? (1=yes, 0=no)
 61- 66 Suggested Retail Price, what the manufacturer thinks the
                vehicle is worth, including adequate profit for the
                automaker and the dealer (U.S. Dollars)
 68- 73 Dealer Cost (or "invoice price"), what the dealership pays
                the manufacturer (U.S. Dollars)
 75- 77 Engine Size (liters)
 79- 80 Number of Cylinders (=-1 if rotary engine)
 82- 84 Horsepower
 86- 87 City Miles Per Gallon
 89- 90 Highway Miles Per Gallon
 92- 95 Weight (Pounds)
 97- 99 Wheel Base (inches)
101-103 Length (inches)
105-106 Width (inches)

Values are aligned and delimited with blanks.
Missing values are denoted with *.
```

+ Extended variable names and their meaning

+ Used units

+ Special values

+ How to denote missing values

Metadata

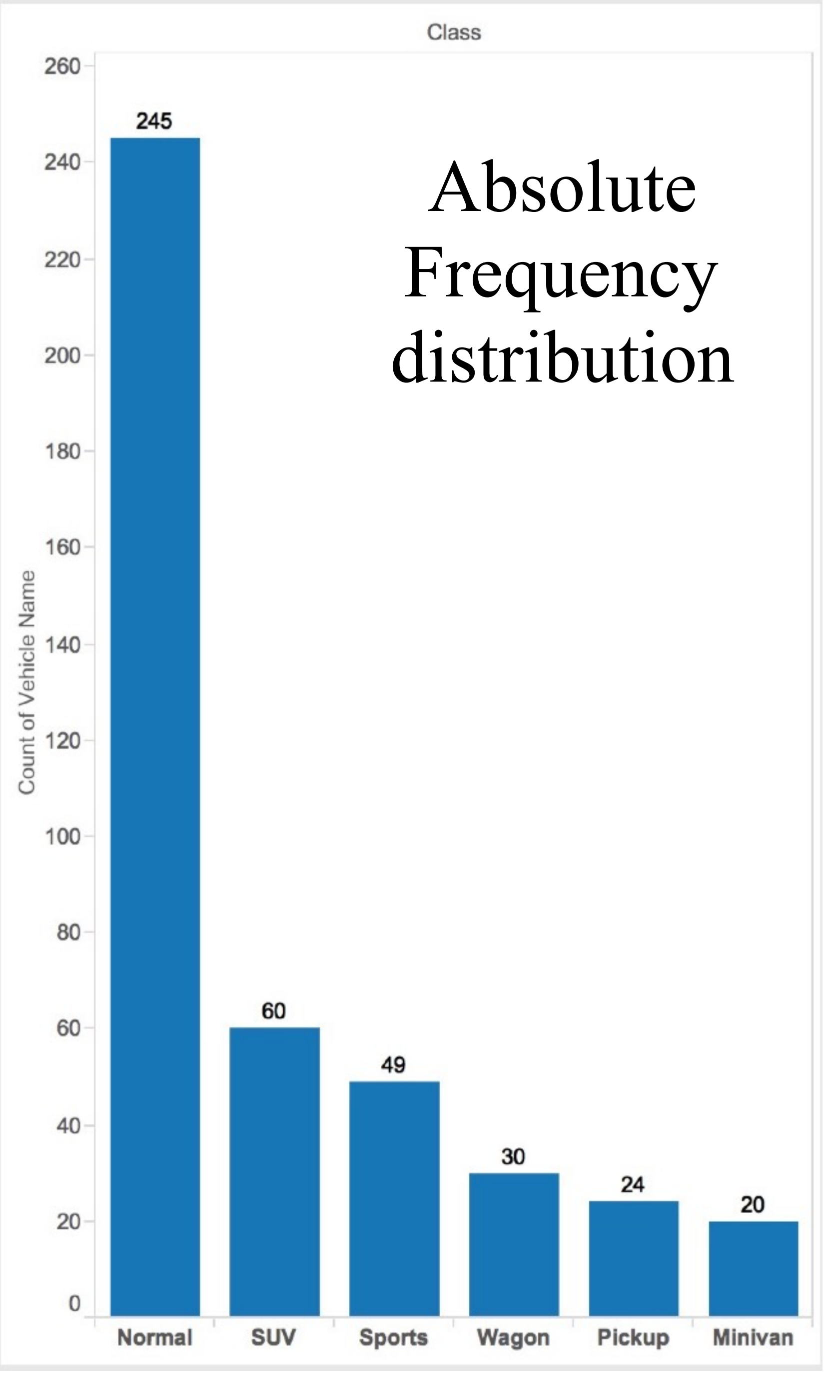
- **Metadata provides:**
  - **Source of data**
  - **Information that facilitates the interpretation of the data set**
  - **Units**
  - **Symbol to indicate a missing value**
  - **Reference point for some measurements**
  - **Resolution at which the measurements were acquired**

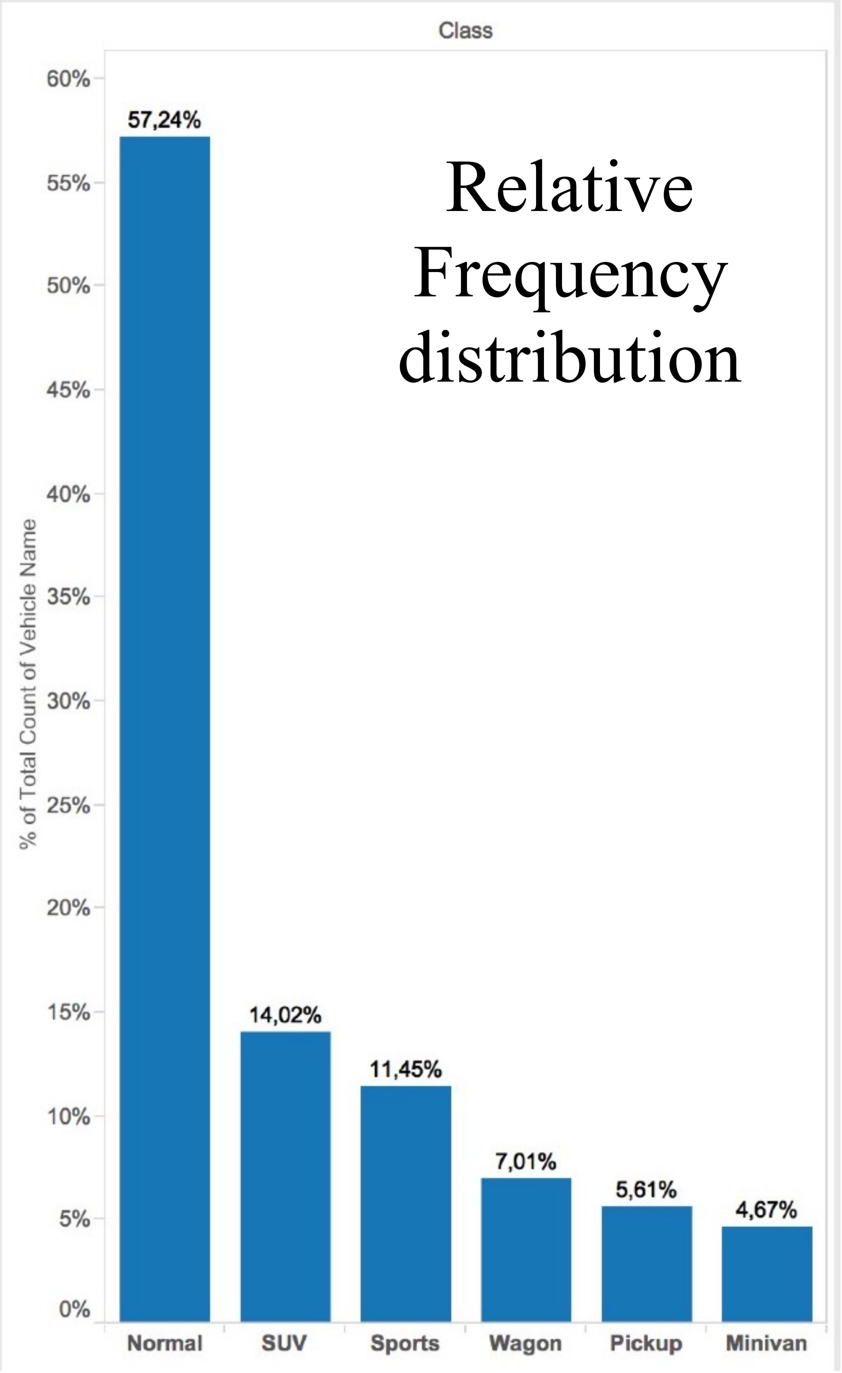
Basic statistics about the (scalar) data

- **For simple data types (scalars)**
- **All data types**
  - **Number of missing values**
- **Excluding the non-numeric arbitrary (names, address, etc)**
  - **Number of values out of range (if the range of variable is provided)**
- **For non-continuous values**
  - **Frequency distribution**
  - **Mode**
- **For numeric variables**
  - **Mean, Variance, etc.**

Basic statistics about the (scalar) data

- **Categorial variable (from Cars data set): Class**

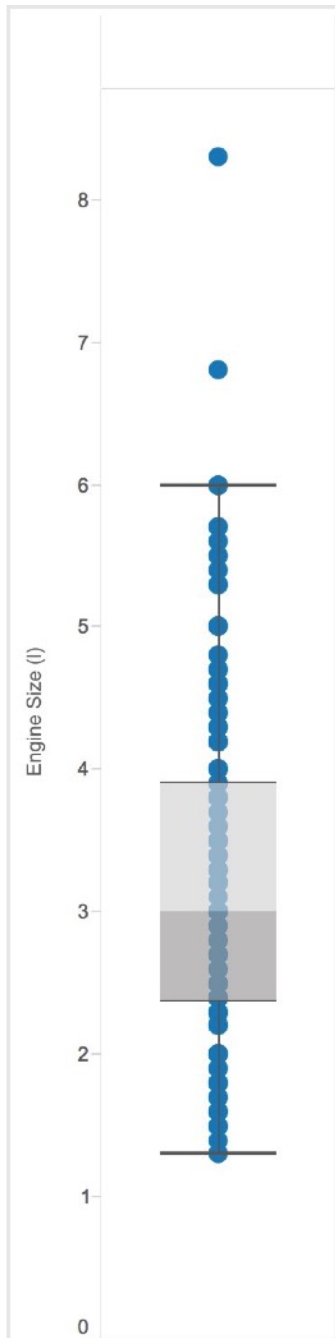




Stats:
- mode
- domain cardinality

Basic statistics about the (scalar) data

- **Numeric (continuous) variable (from Cars data set): Engine Size**



Summary	
Count:	428
SUM(Engine Size (l))	
Average:	3.197
Minimum:	1.300
Maximum:	8.300
Median:	3.000
Standard Deviation:	1.109
First Quartile:	2.375
Third Quartile:	3.900
Skewness:	0.71
Excess Kurtosis:	0.52

Statistics techniques for getting additional insights

- **Outlier detection**
  - **"In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.!"**

https://en.wikipedia.org/wiki/Outlier

https://www.siam.org/meetings/sdm10/tutorial3.pdf

- **Cluster Analysis**
  - **Can help segment the data into groups with strong similarities**

https://en.wikipedia.org/wiki/Cluster_analysis

- **Correlation Analysis**
  - **can help users to eliminate variables (because are redundant or highlight)**

Statistics techniques for getting additional insights

Correlation Analysis

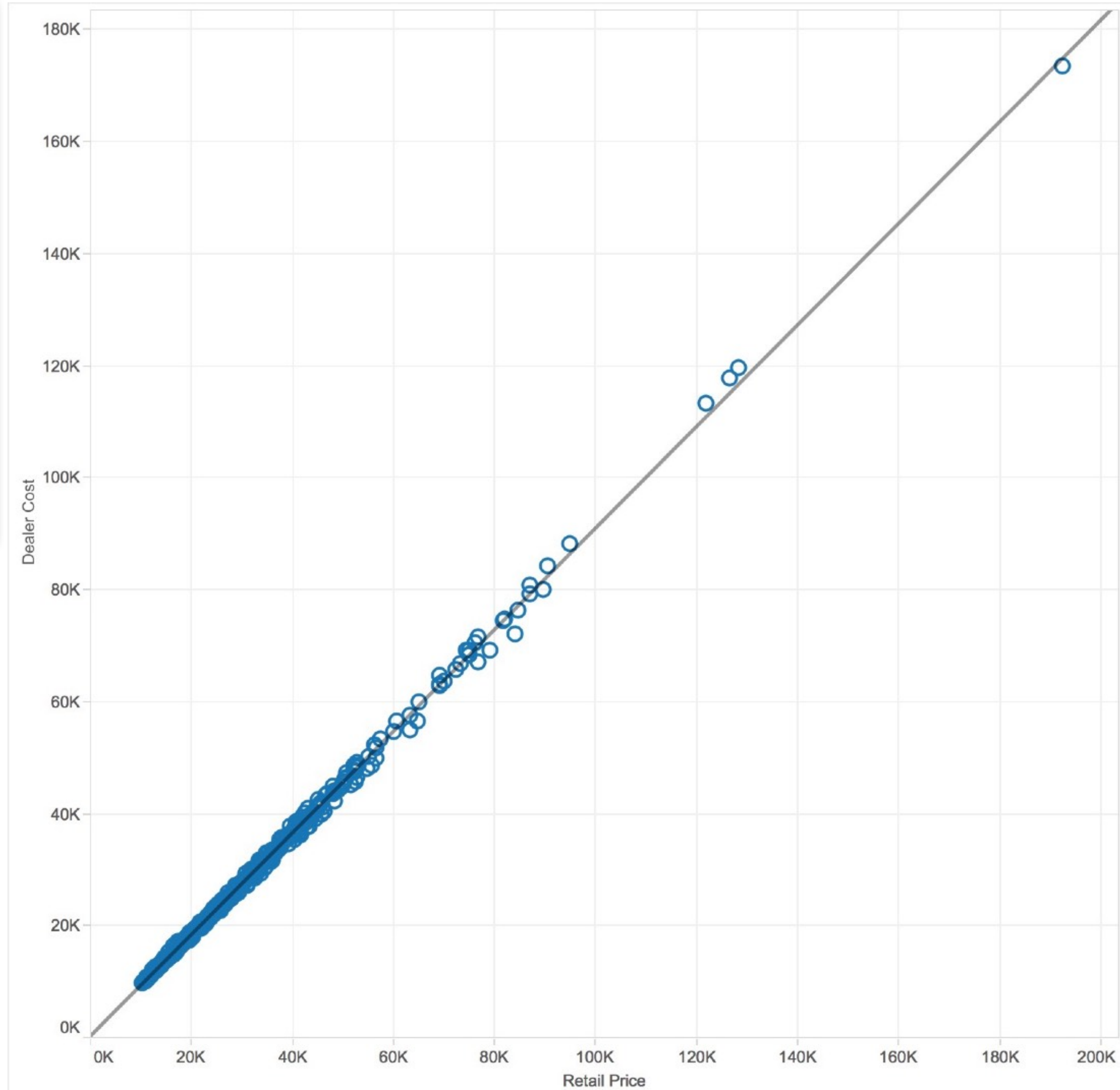
Trend Lines Model

A linear trend model is computed for Dealer Cost given Retail Price. The model may be significant at p <= 0,05.

Model formula:	(Retail Price + intercept)
Number of modeled observations:	428
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	426
SSE (sum squared error):	2,30717e+08
MSE (mean squared error):	541590
R-Squared:	0,998264
Standard error:	735,928
p-value (significance):	< 0,0001

Individual trend lines:

Panes		Line		Coefficients				
Row	Column	p-value	DF	Term	Value	StdErr	t-value	p-value
Dealer Cost	Retail Price	< 0,0001	426	Retail Price	0,907115	0,0018328	494,939	< 0,0001
				intercept	284,145	69,8118	4,07015	< 0,0001



Missing Values and Data Cleansing

- **Missing data:**
  - **malfunctioning sensor; blank entry on a survey; omission on a person entering the data; etc..**
  - **It is necessary to define a strategy to deal with missing data. It should depend on the application domain, the number of missing values, the quality of the other variables.**
- **Erroneous data**
  - **human error; malfunctioning sensor, etc..**
  - **May be very hard to detect unless they are out of range values or obvious outlier.**

Missing Values and Data Cleansing

- **Discard the bad record**
  - **Is the most commonly applied; It implies a loss of information that should be evaluated. Sometimes the records with missing values are the most interesting to be analyzed.**
- **Assign a sentinel value**
  - **Assign a sentinel value for each variable when the real value is in question (missing or erroneous). This value should be carefully considered in the processing.**
- **Assign the average value**
  - **Average value for that variable; Minimally affects the statistics of that variable; The average may not be a good guess; It may mask outliers.**

Missing Values and Data Cleansing

- **Assign value based on nearest neighbor**
  - **Try to find the (missing) value for one variable *i* for one particular record based on the value(s) for that variable based on the records that are the most similar to this particular record (based on the other variables). We are assuming that the variable *i* depends on all other variables and may not be the case.**
  - **When we have connectivity information (spatial or geo-spatial data, graphs) the nearest neighbor may be considered based on the available connections.**
- **Compute a substitute value**
  - **All the previous methods are had hoc ! Some new statistical approaches propose methods and algorithms to make multiple imputations for the missing values**
  - **More info: "Multiple imputation for multivariate missing-data problems: a data analyst's perspective", by Joseph L. Schafer and Maren K. Olsen**

Normalization

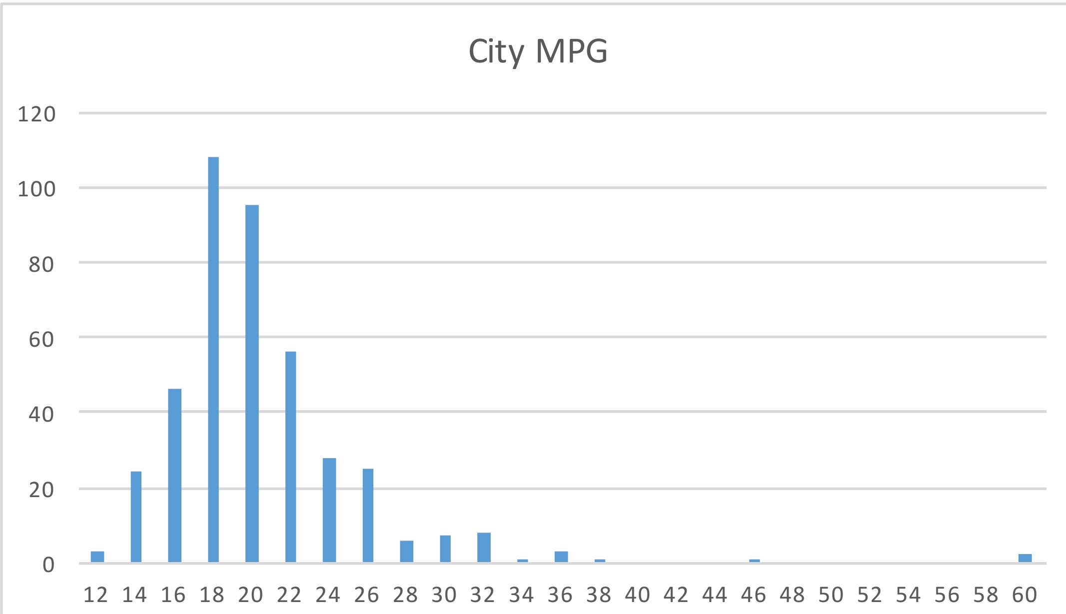
- **Most normalization methods require a distance metric.**
- **One purpose is to scale different variables to comparable range of values.**
- **Another objective is to redistribute the values if they are concentrated on a small part of the available scale**
- **Examples of normalization functions:**

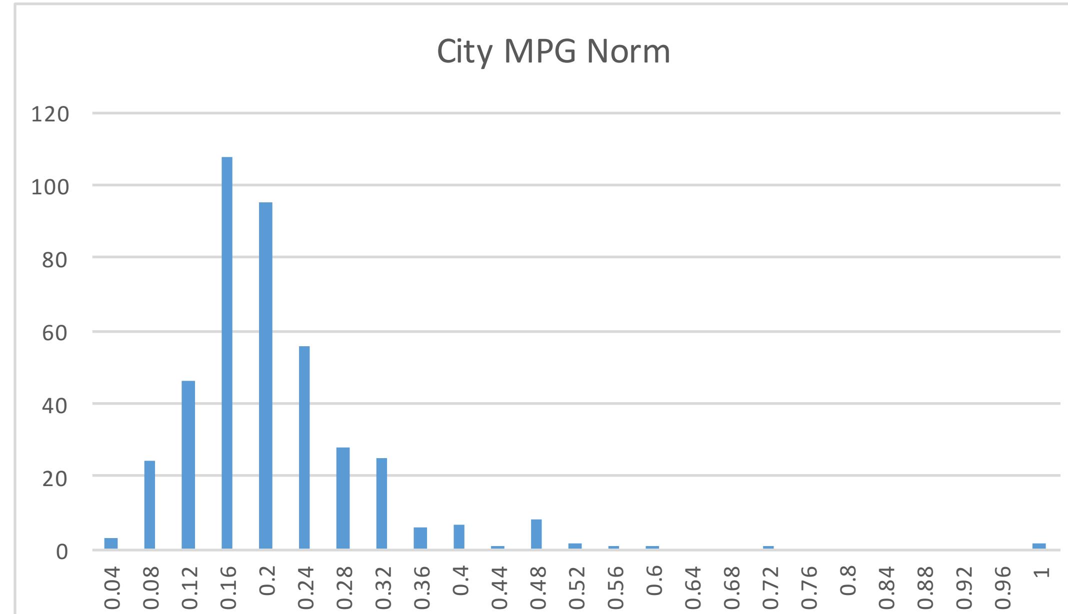
  - $d_{normalized} = \frac{(d_{orignial} - d_{min})}{(d_{max} - d_{min})}$
  - $d_{sqrt-normalized} = \frac{(\sqrt{d_{orignal}} - \sqrt{d_{min}})}{(\sqrt{d_{max}} - \sqrt{d_{min}})}$
  - $d_{log-normalized} = \frac{(\log d_{original} - \log d_{min})}{(\log d_{max} - \log d_{min})}$
  - $d_{z-Score} = \frac{(d_{orignal} - \mu)}{\sigma}$

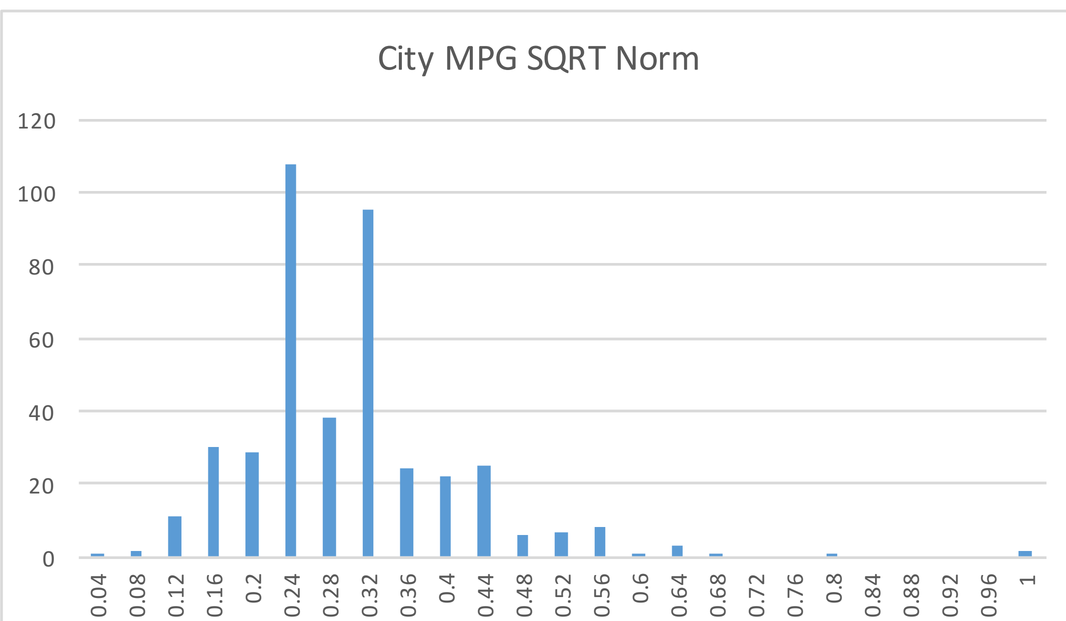
- **Replacing *Min* and *Max* by ∂-*Quantile* and (1-∂)-*Quantile***

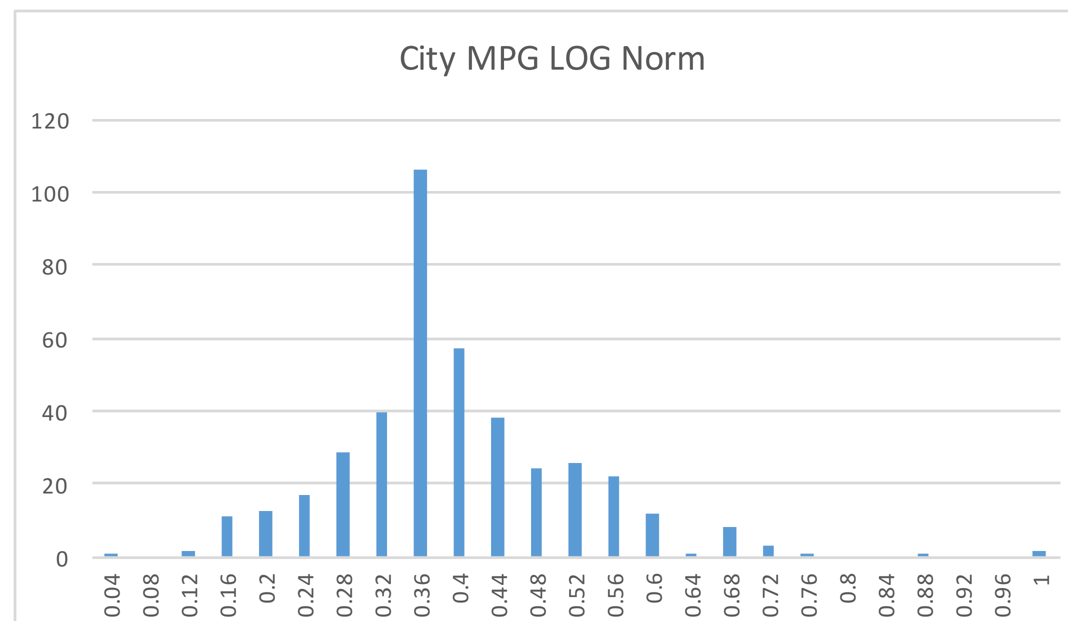
Normalization

- **Data from 414 cars (from 2004); Variable: City Miles Per Gallon (City MPG)**



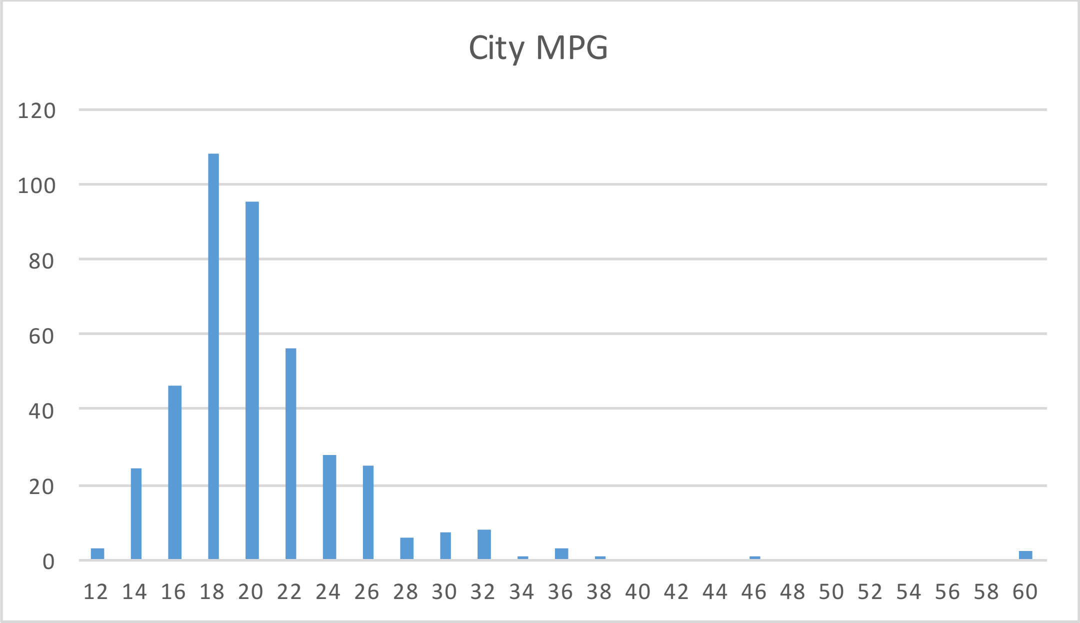


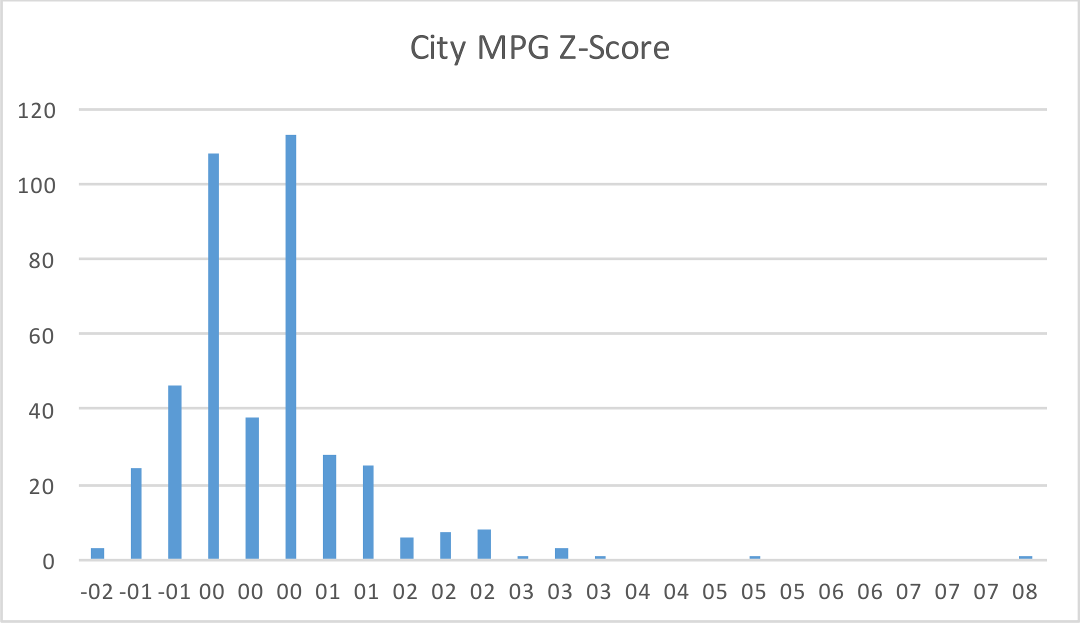


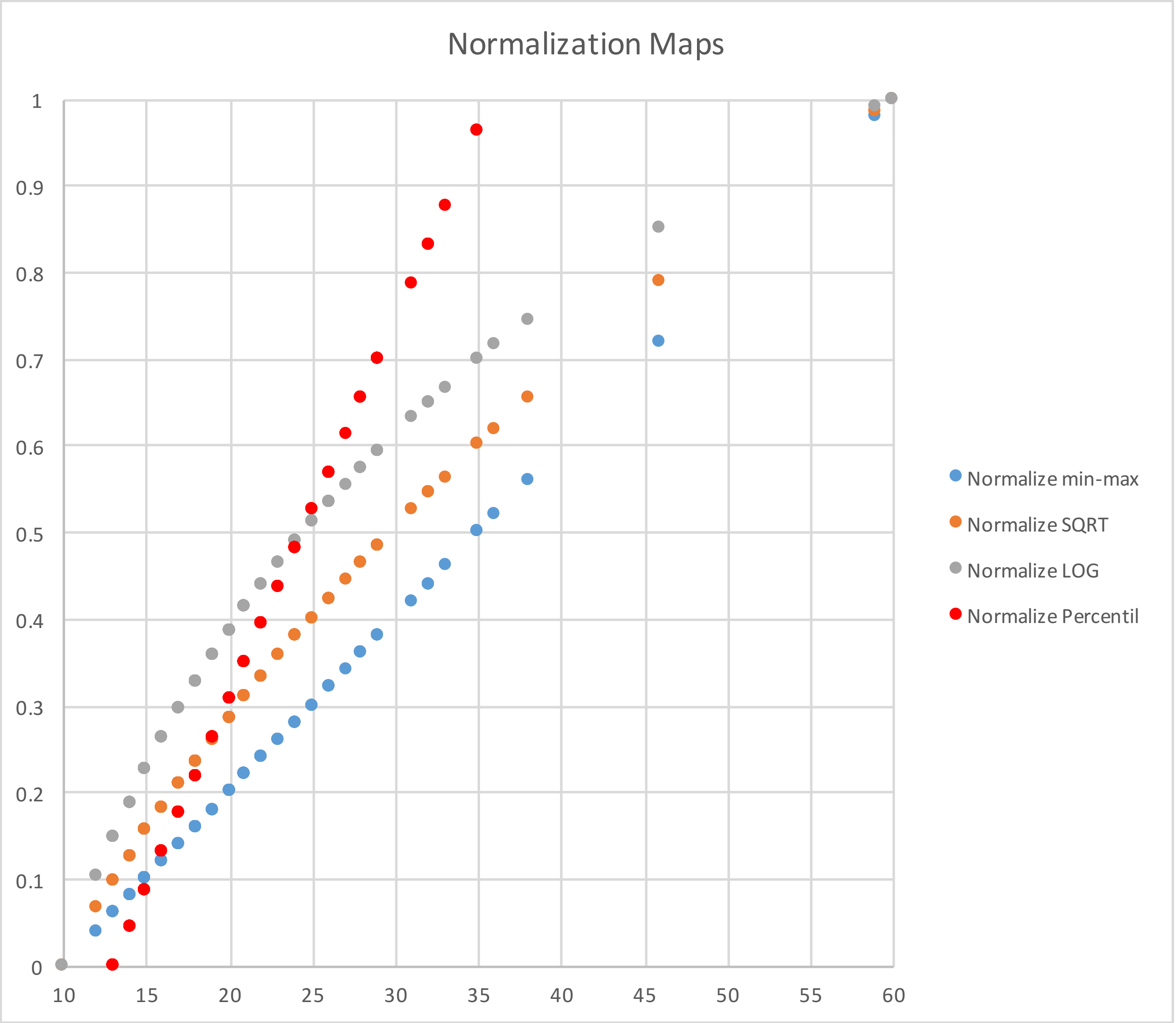


Normalization

- **Data from 414 cars (from 2004); Variable: City Miles Per Gallon (City MPG)**



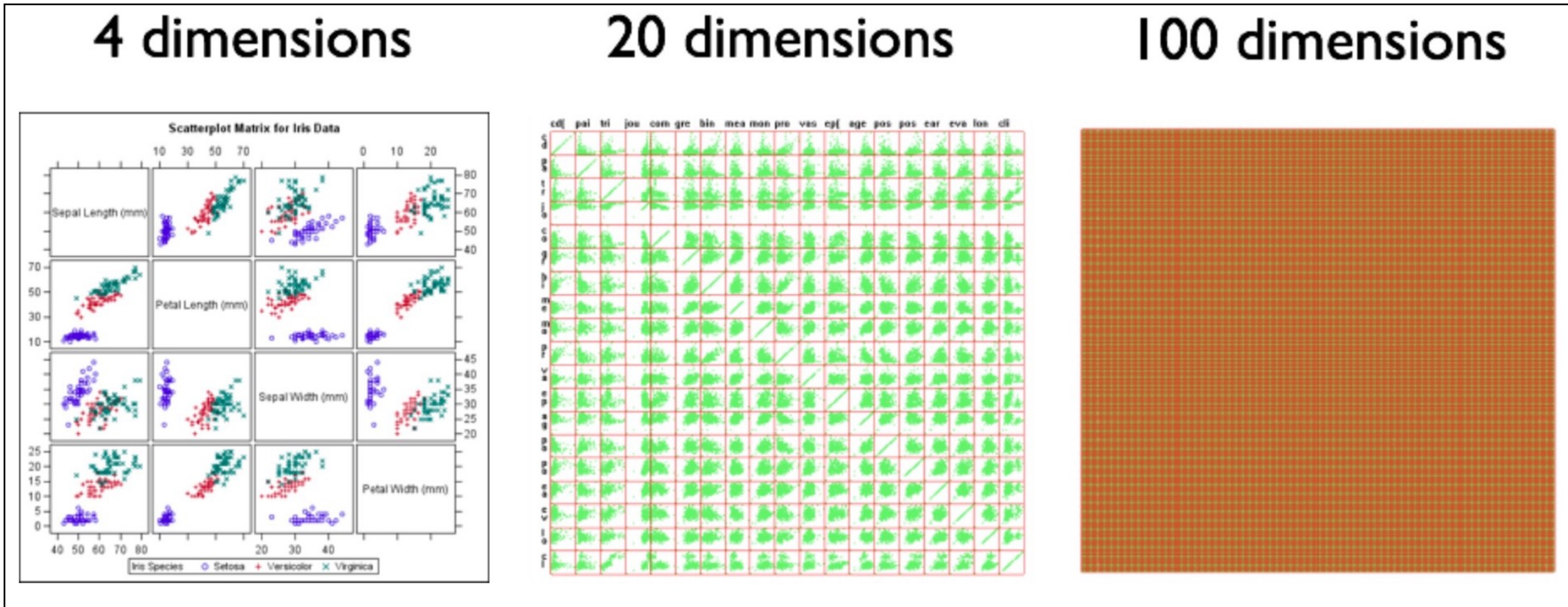




Dimension reduction

- **In situations where the dimensionality of the data exceeds the capabilities of the visualization technique.**

Example of Scatter Plot



* Taken from: Jing Yang's Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets.

Bertini DataScience showcase (2014)

Dimension reduction

- In situations where the dimensionality of the data exceeds the capabilities of the visualization technique. It is necessary to investigate ways to **reduce the data dimensionality, while at the same time preserving, as much as possible, the information contained within**.
- Principal Component Analysis (PCA) - read more
- Multidimensional Scaling (MDS) - read more and more
- Non-linear dimension reduction techniques:
  - Self-organizing Maps (SOMs) - read more
  - Local Linear Embeddings (LLE) - read more

Dimension reduction - Principal Component Analysis (PCA)

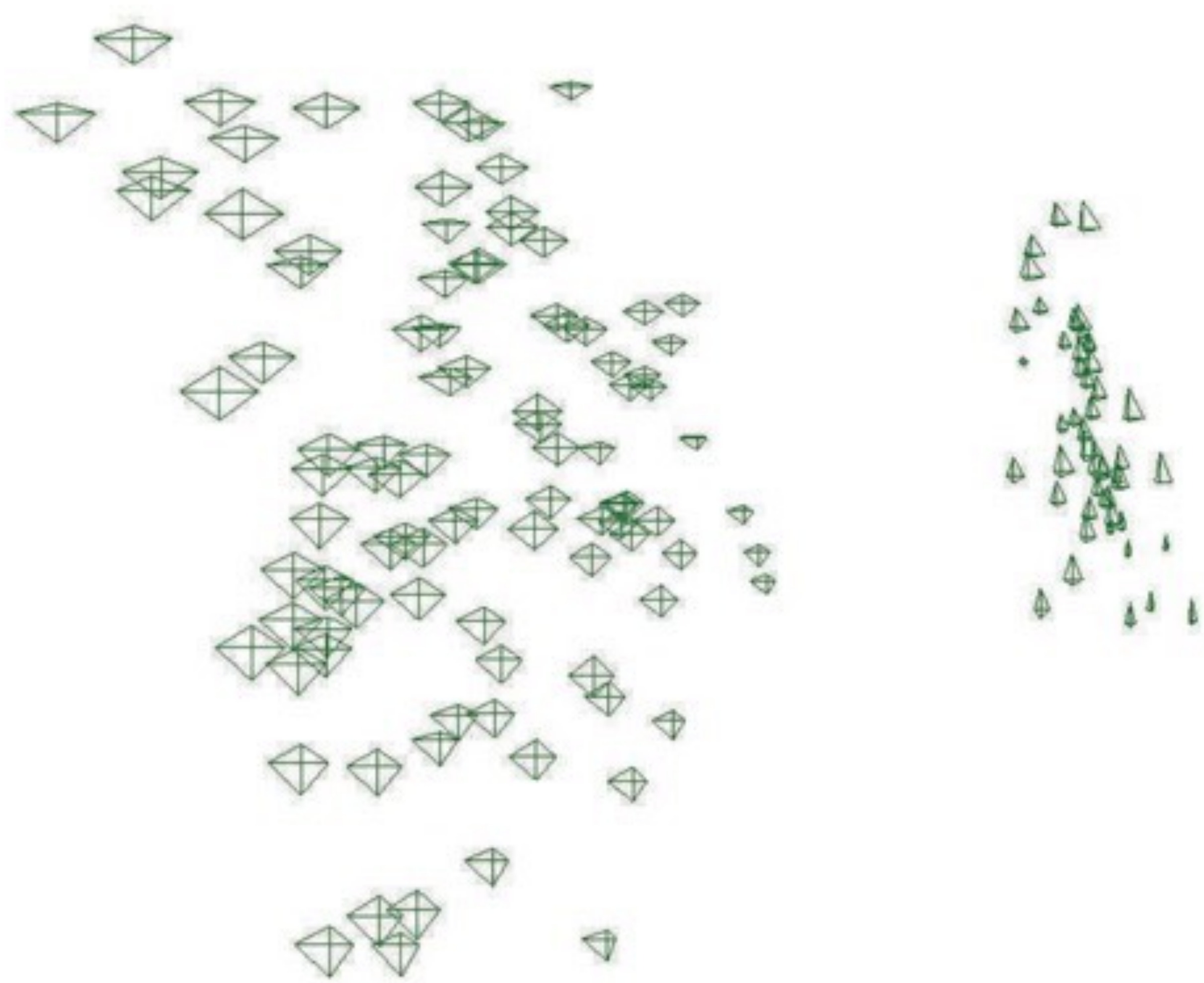
- **PCA computes new dimensions/attributes which are linear combinations of the original data attributes.**

- **The advantage of the new dimensions is that they can be sorted according to their contribution in explaining the variance of the data.**

- **By selecting the most relevant new dimensions, a subspace of variables is obtained that minimizes the average error of lost information**

Dimension reduction - Principal Component Analysis (PCA)

- Figure 2.4 from Interactive Data Visualization: Foundations, Techniques, and Applications, Matthew O. Ward, Georges Grinstein, Daniel Keim, 2010



The Iris data set in star glyphs, with the position of each point based on the first two principal components. The star glyph represents four variables as the lengths of the each of the lines emanating from the center of a four-pointed star. Reasonable clustering can be seen.

Mapping Nominal Dimensions to Numbers

- **How to visualize Nominal dimensions?**
  - **how many nominal dimension exist?**
  - **how many distinct values each variable can take on?**
  - **an ordering or distance relation is available or can be derived?**
- **Warning:**

> Find a mapping of the data to a graphical entity or attribute that
> **doesn’t introduce artificial relationships that don’t exist in the data**

- **Ranked nominal values can be mapped to numbers and so can be easily mapped to many graphical attributes**
- **Non ranked nominal values have to be managed carefully**

Mapping Nominal Dimensions to Numbers

- Non-ranked nominal values have to be managed carefully
  - Variables with only a modest number of different values:
    - map to graphical attributes like color or shape
  - A single nominal variable:
    - Use this variable as the label for the graphical elements being displayed when the number of records to be displayed is modest.
    - Showing random subsets of labels and changing the points with labels being shown on a regular basis, and showing only the labels on objects near the cursor.

Mapping Nominal Dimensions to Numbers

- **Mapping to numbers by looking at similarities between the numeric variables associated with a pair of nominal values**
  - **If the statistical properties of the records associated with one nominal value are sufficiently similar to the properties of a different value, then that implies that these two values should likely be mapped to similar numeric values.**
  - **Conversely, if there are sufficient differences in properties, then likely they should be mapped to quite distinct values.**
- **Given all the pairwise similarities, we could use correspondence analysis to map the different nominal values to positions in one dimension. Applying to all nominal dimensions of the data set - multiple correspondence analysis.**

See more: https://en.wikipedia.org/wiki/Correspondence_analysis

http://www.mathematica-journal.com/2010/09/an-introduction-to-correspondence-analysis/

Other data processing topics

Segmentation

- In many situations, the data can be separated into contiguous regions, where each region corresponds to a particular classification of the data.
- Simple segmentation can be performed by just mapping disjoint ranges of the data values to specific categories.
- it is important to look at the classification of neighboring points to improve the confidence of classification, or even to do a probabilistic segmentation, where each data point is assigned a probability for belonging to each of the available classifications.
- Common in image data or geo-spatial data (satellite images)

Sampling and subsetting

- To transform a data set with one spatial resolution into another data set with a different spatial resolution. For example, we might have an image we would like to shrink or expand, or we might have only a small sampling of data points and wish to fill in values for locations between our samples (assuming that the data is a discrete sampling of a continuous phenomenon).
- The process of interpolation is a commonly used resampling method in many fields, including visualization:
  - Linear interpolation
  - bi-linear interpolation
  - Nonlinear interpolation

Sampling and subsetting

- **Data subsetting is also a frequently used operation both prior to and during visualization.**

- **This is especially helpful for very large data sets, as the visualization of the entire data set may lead to substantial visual clutter.**

- **Query before visualization versus subsetting during visualization**

Aggregation and Summarization

- **it is often useful to group data points based on their similarity in value and/or position and represent the group by some smaller amount of data:**
- **Data Clustering methods**
  - **See More:**
    - https://en.wikipedia.org/wiki/Cluster_analysis
    - http://www.ise.bgu.ac.il/faculty/liorr/hbchap15.pdf
- **Displaying the clusters (or their representation)**
  - **Provide sufficient information for the user to decide whether he or she wishes to perform a drill-down on the data**

Aggregation and Summarization

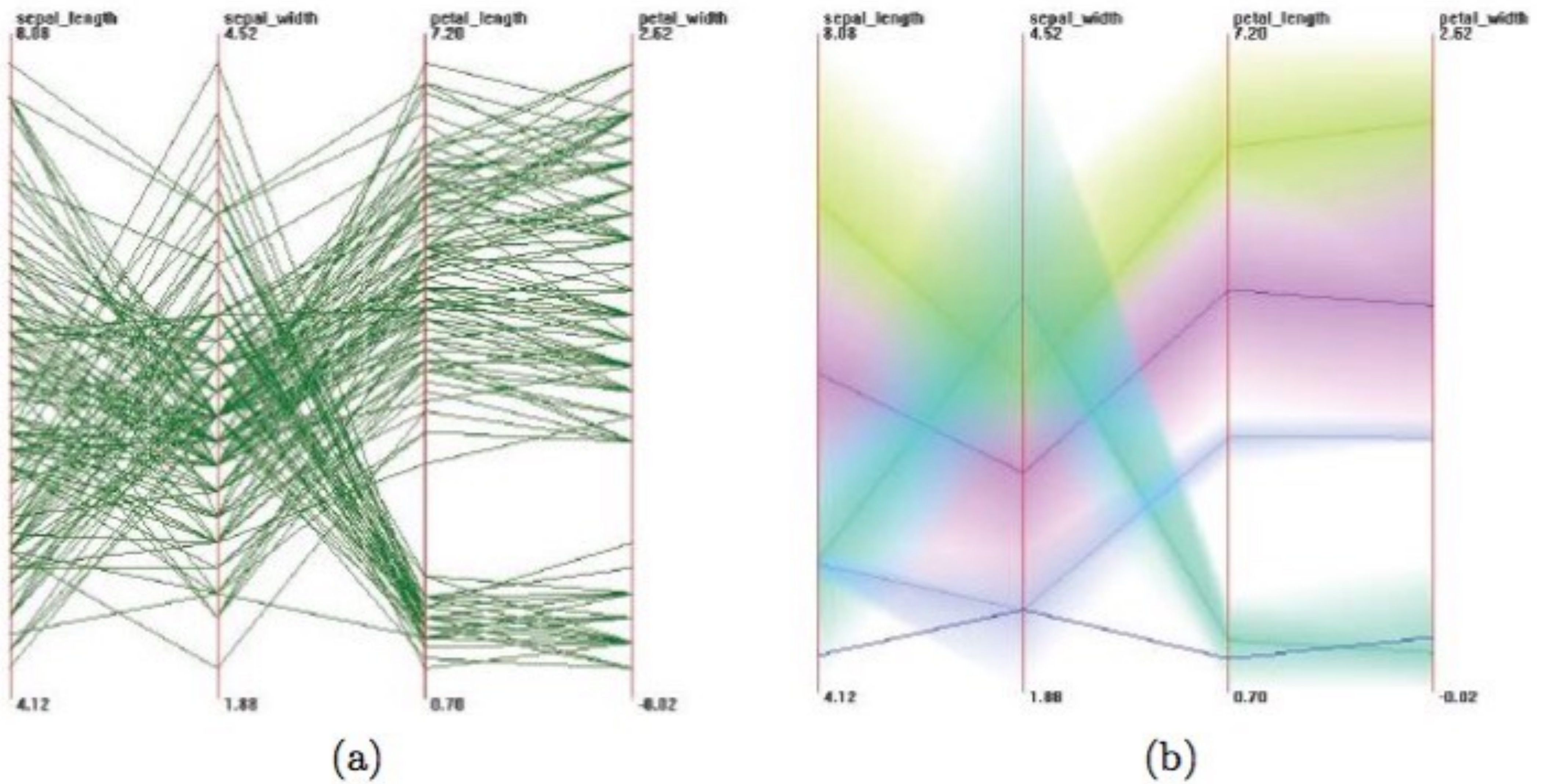


Figure 2.5. The Iris data set in parallel coordinates: (a) the original data; (b) the centers and extents of clusters after aggregation. Each axis in parallel coordinates represents a dimension, with each record being drawn as a polyline through each of the coordinate values on the axes.

Smoothing and Filtering

- **In statistics and image processing, to smooth a data set is to create an approximating function that attempts to capture important patterns in the data, while leaving out noise or other fine-scale structures/rapid phenomena.**
- **In smoothing, the data points of a signal are modified so individual points (presumably because of noise) are reduced, and points that are lower than the adjacent points are increased leading to a smoother signal**
- **See more:**
  - https://en.wikipedia.org/wiki/Smoothing

Raster to vector conversion

- **In Computer Graphics:**
  - **Vector data (vertices, edges, and triangular or quadrilateral patches) => Image (pixel-based)**
- **It can be important to make the reverse:**
  - **Compressing the contents for transmission.**
  - **Comparing the contents of two or more images**
  - **Transforming the data**
  - **Segmenting the data**
- **Read more: IDV: Foundations, Techniques, and Applications, Pag 72 - 74**

Further Reading and Summary

Further Reading

- **Recommend Readings**
  - Pag 51 - 76 from Interactive Data Visualization: Foundations, Techniques, and Applications
  - Pag 30 - 40 from Visualization Analysis & Design, Tamara Munzner
- **Supplemental readings:**
  - https://en.wikipedia.org/wiki/Outlier
  - https://en.wikipedia.org/wiki/Cluster_analysis
  - https://en.wikipedia.org/wiki/Correspondence_analysis
  - https://en.wikipedia.org/wiki/Cluster_analysis

What you should know

- **The concept of variable or dimension and the diference between independent and dependent variables.**
  - grocking the data => take decisions
- **The various data types taxonomies and the impact of a data type in visualization.**
  - numeric vs non numeric; oder vs non-order; Types of scale;
- **The structural aspects of a data set.**
  - Tables, links, position, grid, etc.
- **Data pre-processing techniques: the goal of each one and the most important ones**
  - Outlier detection and process; normalization; dimensionality reduction, Sampling and subsetting; Aggregation and Summarization

Recommended Actions

- **Install Tableau software (desktop version). Activate with a students license.**
  - http://www.tableau.com/academic/students
- **To get an overview of Tableau see the video:**
  - http://www.tableau.com/learn/tutorials/on-demand/getting-started
- **Get familiar with the dataset 2004 Cars and Trucks Data Set**
  - http://www.idvbook.com/teaching-aid/teaching-aid/data-sets/2004-cars-and-trucks-data/