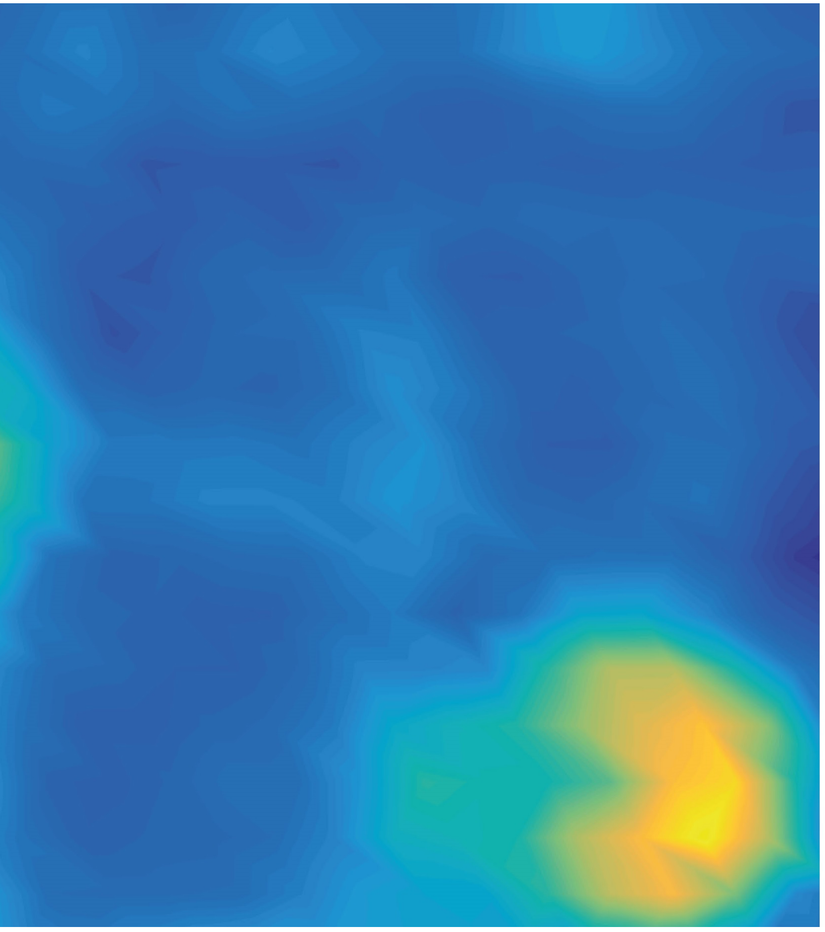
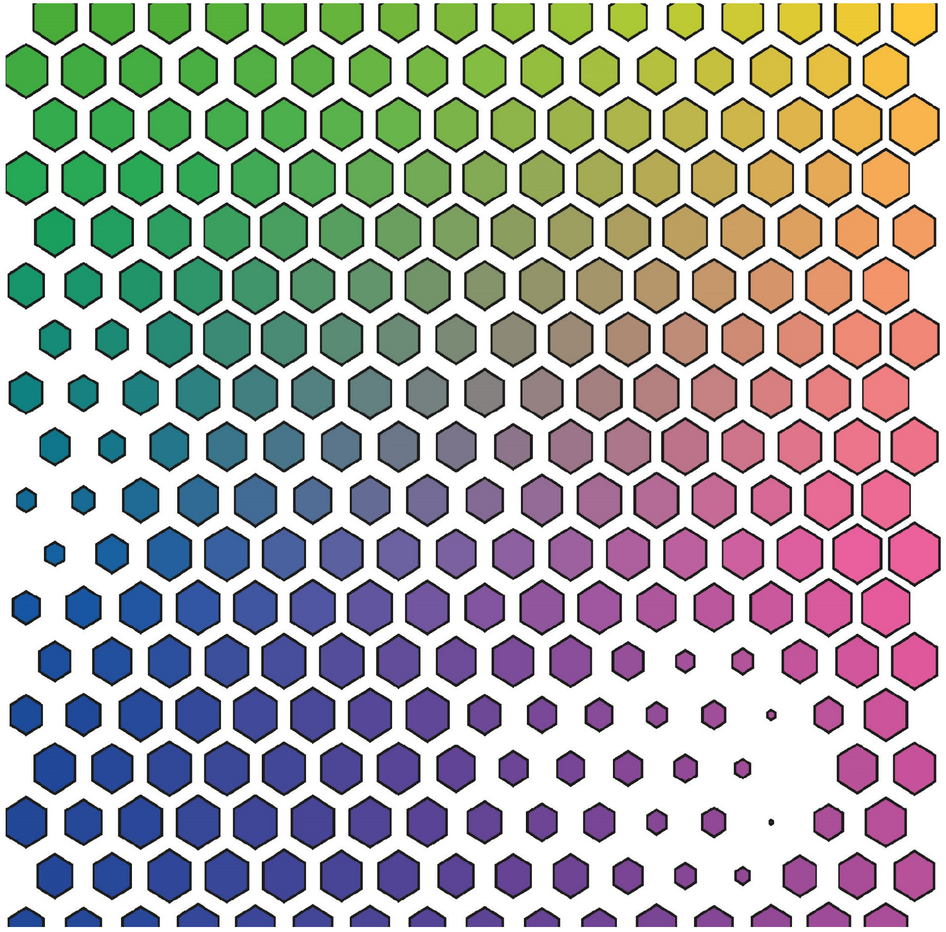
CREATIVE DATA MINING

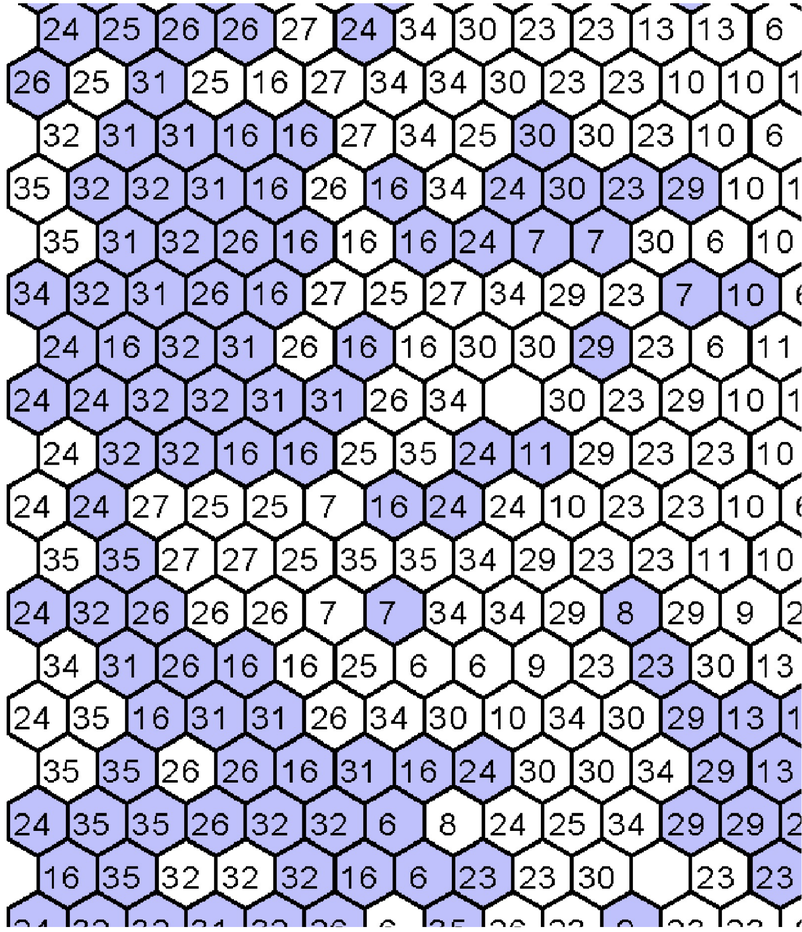
Introduction to Knowledge Discovery Process

Dr. Varun OJHA
Danielle GRIEGO

26.02.2018

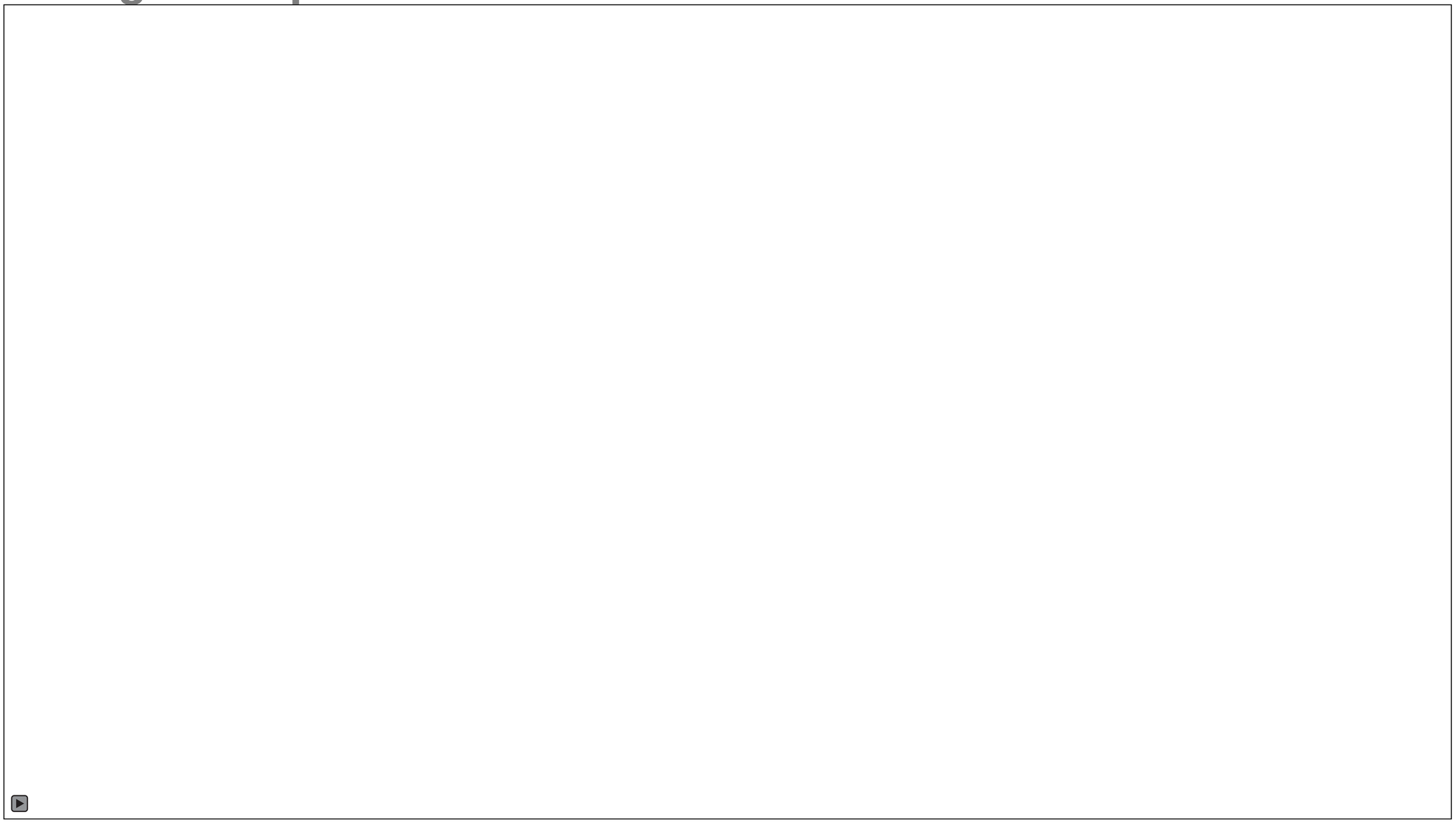






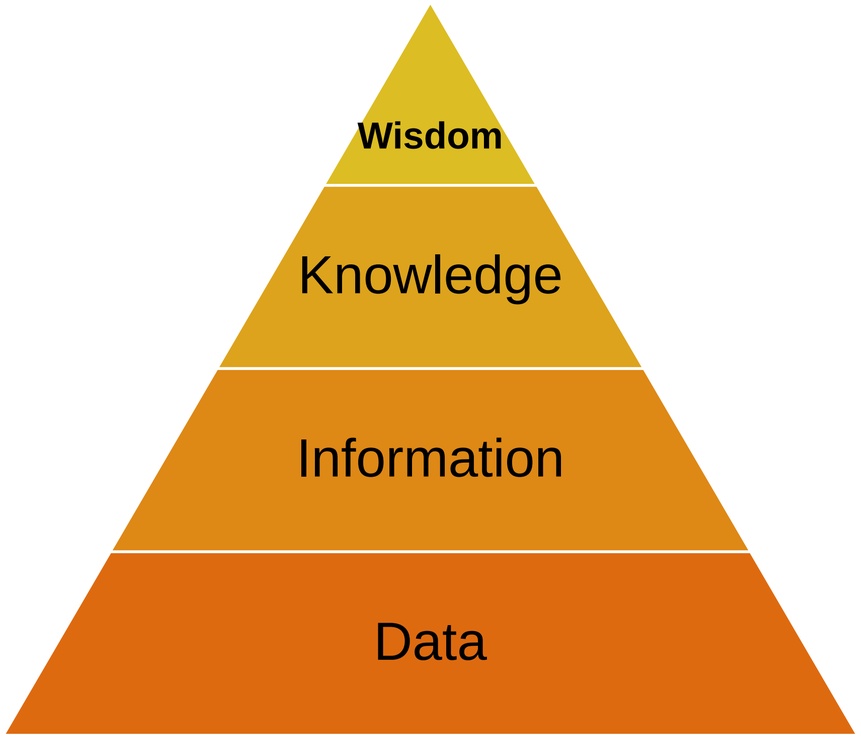
ETH zürich iA Chair of Information Architecture

Data Mining Example from the Chair of Information Architecture



Source:
www.esum.arch.ethz.ch

Why Data Mining?



Applied	• I should stop the car
Context	• Traffic light in the direction of driving has turned red
Meaning	• Traffic light has turned red
Raw	• (47.41, 8.51), Red, 10 kmph, (46.23, 8.11) , 10 m

Source: http://www.allthingy.com/data-information-knowledge-wisdom/

What to do in Data Mining?

Importance of data processing and knowledge discovery

Identifying applications

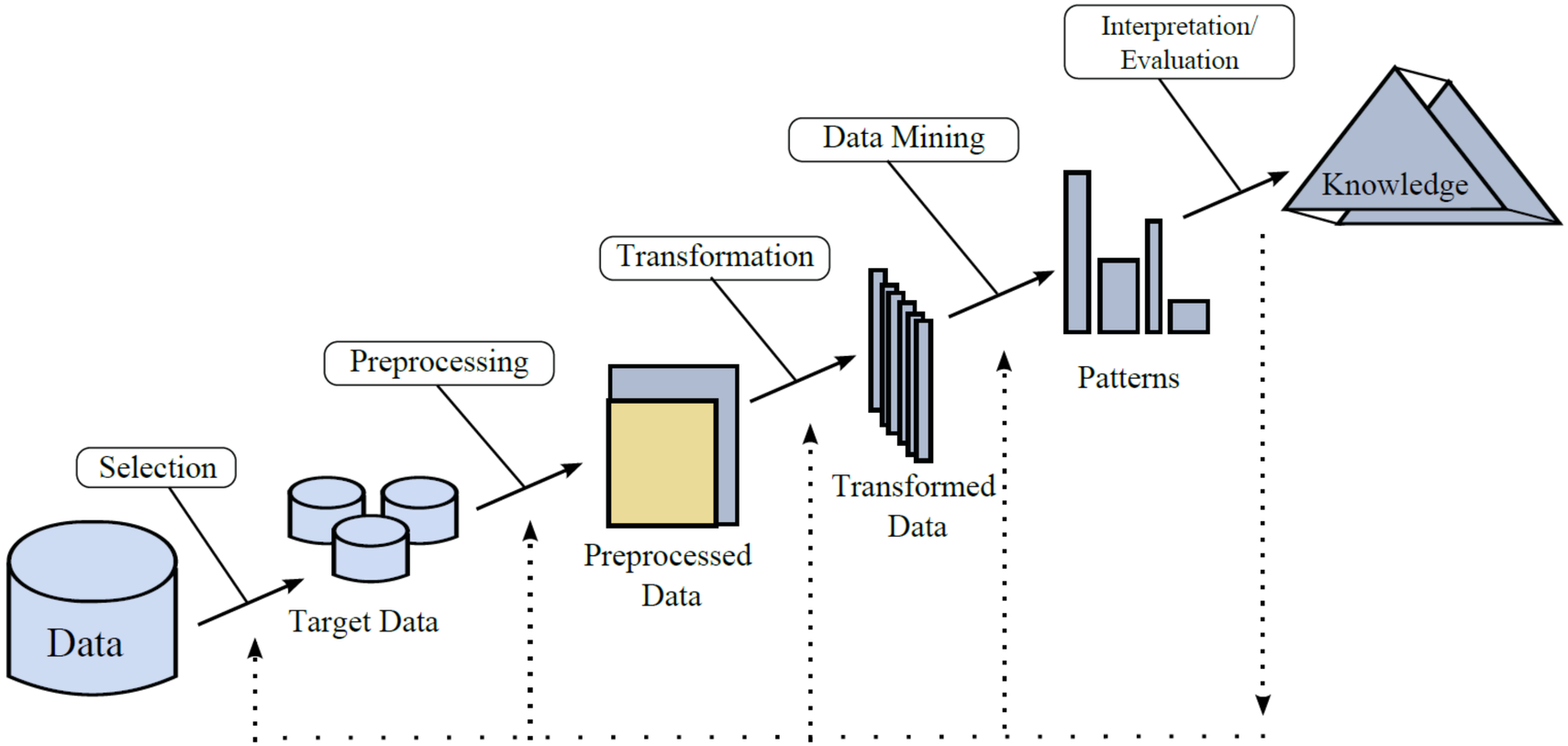
Generating good quality data (data quality assurance)

Cleaning and preprocessing data

Identifying computational methods

Discovering meaning (interpreting discovered knowledge)

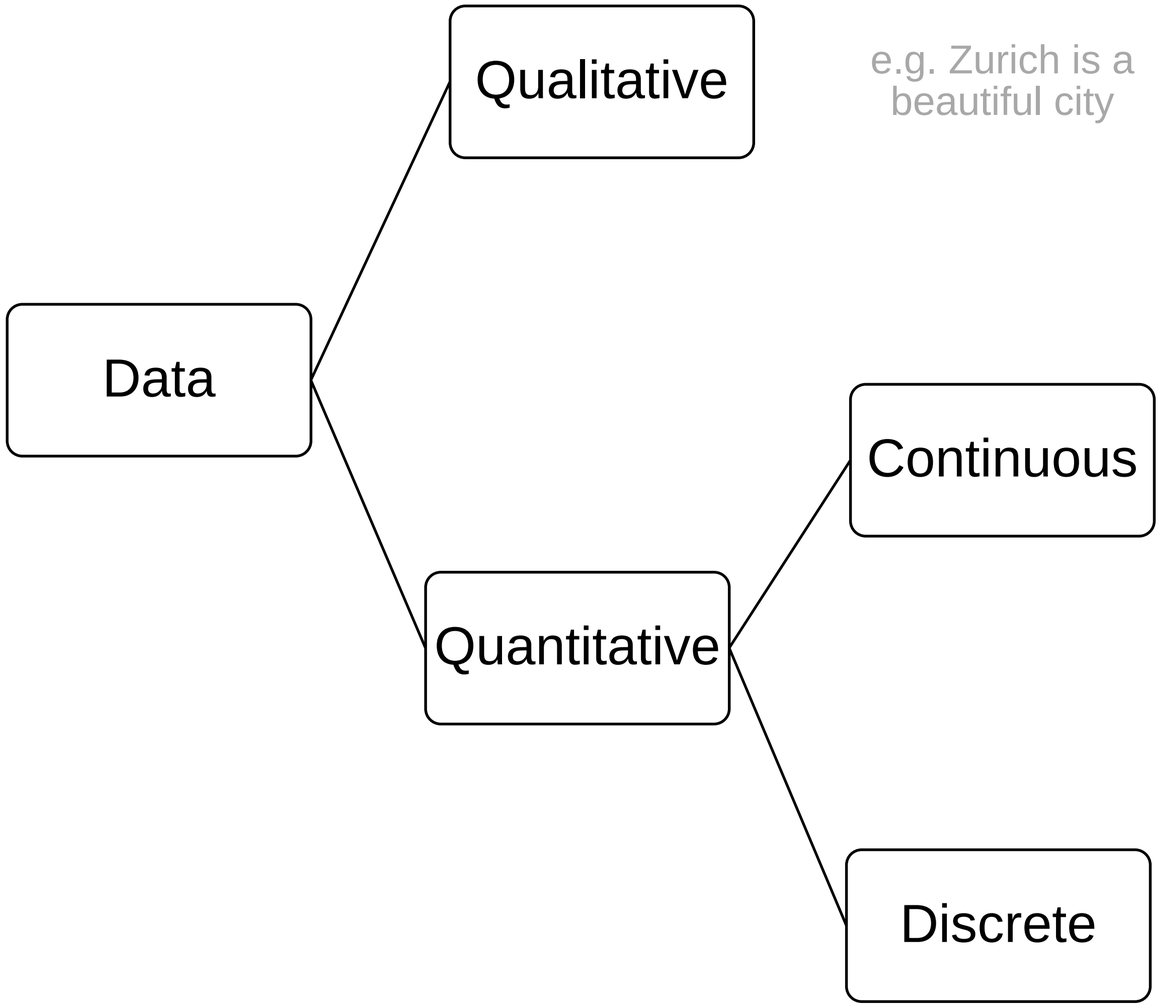
How to Perform Data Mining?



Source: Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, From Data Mining to Knowledge Discovery in Databases, AI Magazine Volume 17 Number 3 (1996)

Data

Defining Variables



e.g. Zurich is a beautiful city

Continuous

- Always numeric
- Can be any number, positive or negative

Something we can measure, e.g., height

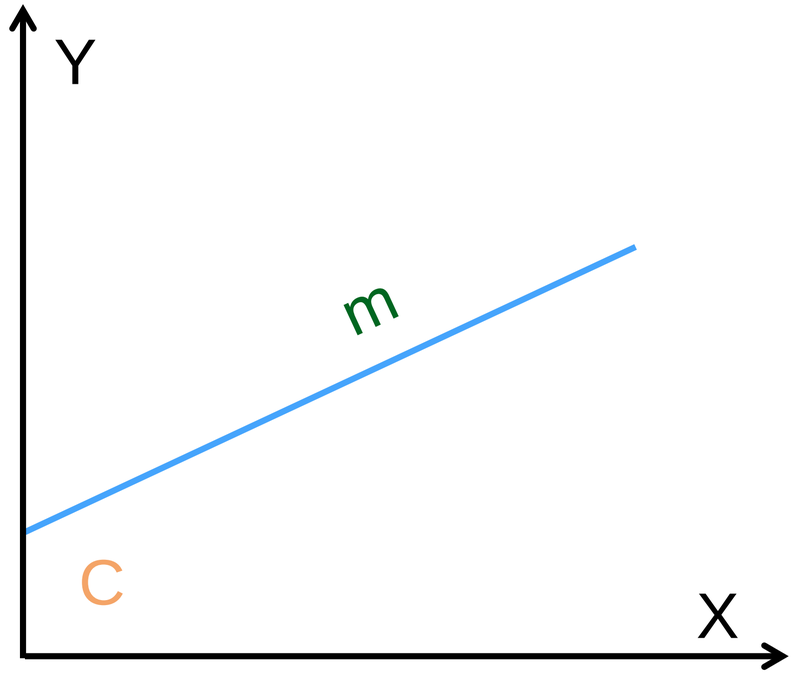
Categorical/Discrete

- Ordinal variable – e.g. rating (excellent, good, fair, poor) or Score (0 – 5)
- Nominal variable – e.g. : gender (male, female)
- Binary variables – Yes or No (0 /1)

Something we can count, e.g., number of students

Data

Defining Variable Types



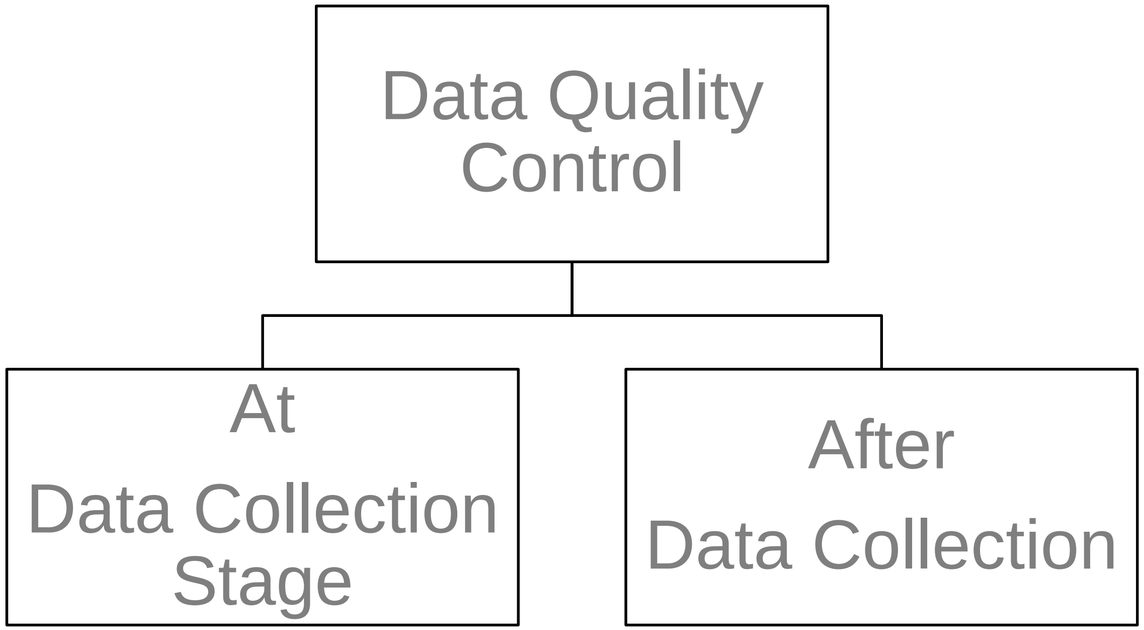
Independent / Dependent

e.g., Y = mX + C

Value of Y depended on m, X, and C

Data Preprocessing

Improving Data Quality



Data Preprocessing

Improving Data Quality



Reference: Mónica Bobrowski, Martina Marré, Daniel Yankelevich, Measuring Data Quality, Report no.: 99-002, Pabellón 1 - Planta Baja - Ciudad Universitaria

Data Preprocessing

Improving Data Quality



Data cleaning

Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

Integration of multiple databases, data cubes, or files

Data transformation

Normalization and aggregation

Data reduction

Obtains reduced representation in volume but produces the same or similar analytical results

Data discretization

Part of data reduction but with particular importance, especially for numerical data

Reference: http://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf

Data Preprocessing

Data Cleaning

	Features (Variables)				
	Input features (Independent Variables)				Output feature (dependent variable)
Samples	Input V1	Input V2	Input V3	Input V4	Output V1
Sample 1	2.3	0.25	Good	1.5	y_1
Sample 2	4.5	43598.21	Good	1.8	y_2
Sample 3	4.7	0.33	Excellent	1.9	y_3
Sample 4	?	0.22	Good	3.9	y_4
Sample 5	?	0.19	Average	1.2	y_5
:	6.7	0.88	Good	1.8	:
Sample N	5.5	0.36	Bad	1.6	y_N

Missing Value

Outlier

Noise (required smoothing)

y_i Output variable can be continuous/discrete

Data Preprocessing

Data Transformation

	Features (Variables)		
	Input features (Independent Variables)		Output feature (dependent variable)
Samples	Input V1	Input V2	Output V1
Sample 1	2.3	0.25	y_1
Sample 2	4.5	0.39	y_2
Sample 3	4.7	0.33	y_3
Sample 4	2.99	0.22	y_4
Sample 5	3.18	0.19	y_5
:	6.7	0.36	:
Sample N	5.5	0.88	y_N

	Features (Variables)		
	Input features (Independent Variables)		Output feature (dependent variable)
Samples	Input V1	Input V2	Output V1
Sample 1	0.00	0.09	y_1
Sample 2	0.50	0.29	y_2
Sample 3	0.55	0.20	y_3
Sample 4	0.16	0.04	y_4
Sample 5	0.20	0.00	y_5
:	1.00	0.25	:
Sample N	0.73	1.00	y_N

Max value

Min value

Transformed Variable

Data Preprocessing

Data Reduction

	Features (Variables)			
	Input features			Output feature
Samples	Input V1	Input V2	Input V4	Output V1
Sample 1	2.3	0.25	1.5	y_1
Sample 2	4.5	0.39	1.8	y_2
Sample 3	4.7	0.33	1.9	y_3
Sample 4	2.99	0.22	1.6	y_4
Sample 5	3.18	0.19	1.2	y_5
:	6.7	0.88	1.8	:
Sample N	5.5	0.36	1.6	y_N

Feature Extraction

Feature Selection

	Features (Variables)			
	Input features			Output feature
Samples	PCA 1	PCA 2		Output V1
Sample 1	-1.97	0.06		y_1
Sample 2	0.25	0.15		y_2
Sample 3	0.45	0.22		y_3
Sample 4	-1.28	0.09		y_4
Sample 5	-1.14	-0.33		y_5
:	2.48	-0.04		:
Sample N	1.21	-0.15		y_N

	Features (Variables)			
	Input features			Output feature
Samples	Input V1		Input V4	Output V1
Sample 1	2.3		1.5	y_1
Sample 2	4.5		1.8	y_2
Sample 3	4.7		1.9	y_3
Sample 4	2.99		1.6	y_4
Sample 5	3.18		1.2	y_5
:	6.7		1.8	:
Sample N	5.5		1.6	y_N

Extracted Features

What’s Next

A: Installing Python



https://anaconda.org/
https://www.youtube.com/watch?v=dgjEUcccRwM

B: Tutorial on Python

Programming practice in Python

C: Machine Learning Algorithms

Supervised and Unsupervised

Thank you!

Varun Ojha ojha@arch.ethz.ch

Danielle Griego griego@arch.ethz.ch