DATA SCIENCE IN MANUFACTURING

WEEK 2

ANDREW SHERLOCK, JONATHAN CORNEY, DANAI KORRE

LECTURE: WEEK 2

Data Carpentry

THE UNIVERSITY *of* EDINBURGH
School of Engineering

BY THE END OF THIS LECTURE YOU SHOULD:

- Understand the importance of data quality
- Understand data carpentry

DATA CARPENTRY INTRODUCTION

WHAT IS DATA CARPENTRY?

Data carpentry, you may have heard this referred to as data wrangling, refers to the process of cleaning-up and pre-processing your data in order to be able to analyse them properly and derive actionable insights from them.

WHAT IS DATA CARPENTRY?

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Clock modulation of starch, pigments and nitrog			Timet Project									
2	Sosa M; Pintos A	Study 2019-12-05 to 202(		Period analysed in BioDare									
3	If not indicated differently metabolites reported per g of fresh weight of 6-week-old plant leaf rosettes												
4													
5	**Sample**	**Strain**	**Genotype**	**Media**	**Biomas**	**Starch mg/g FW**	**Sucrose (mg/g)**	**Chloro.**		**Cell**	**Sample**	**Period**	**Phase**
6	A1	D62	*phyB-9*	GM-agar	0.1206g	6	1.2	0.0018 g/g		A1	WT SD	24.2	8.1
7	A2	D64	*phyB-9*	GM-agar	0.1275g	6.5	1.1	0.0016		A2	phyA SD	23.5	7.2
8	A3	D1	*phyA-211*	GM-agar	0.2872 g	5	1	0.0014		A3	phyB SD	24.5	7.7
9	**A4**	**B12**	***elf4-101***	**GM-agar**	**0.1524g**	**3**	**0.6**	**0.002**		A4	elf4 SD	27.1	9
10	**A5**	**B33**	***toc1-2***	**GM-agar**	**0.2035 g**	**0**	**1.1**	**0.0017**		A5	toc1 SD	30.1	11
11	B1	D62	*phyB-9*	GM-agar +SUC	0.2104	6.2	1.3	0.0021		B1	WT LD	24.5	5
12	B2	D64	*phyB-9*	GM-agar +SUC	0.2435	7	1.2	0.0019		B2	phyA LD	24.1	6.1
13	B3	D1	*phyA-211*	GM-agar +SUC	0.3213g	5.8	1.1	error		B3	phyB LD	25	5.7
14	B4	B12	*elf4-101*	GM-agar +SUC	0.2135g	4.9	0.8	0.0022		B4	elf4	-1	-1
15	B5	B33	*toc1-2*	GM-agar +SUC	0.292 g	5.9	0.9	0.0021		B5	toc1 LD	31.1	7
16	C1	D62	*phyB*	short+S	130mg	6	1.2	0.0018					
17	C2	D64	*phyB*	short+S	141.5 mg	6.5	1.1	0.0016					
18	C3	D1	*phyA*	short+S	288 mg	5	1	0.0014					
19	C4	B12	*elf4*	short+S	152mg	3	0.6	0.002					
20	C5	B33	*toc1*	short+S	204mg		1.1	0.0017					
21	D1	D62	*phyB*	LD -S	135mg	6	1.2	0.001					
22	D2	D1	*phyA*	LD -S	695 mg								
23	D3	D64	*phyB*	LD -S	141mg	7	1.1	0.0021					
24	D4	B12	*elf4*	LD -S	1425mg	3.1	0.6	0.003					
25	D5	B33	*toc1*	LD -S	204mg	5	1.1	0.0011					
26		short days 6 h light											
27		long days 18 h light			Updated	21-08-11							

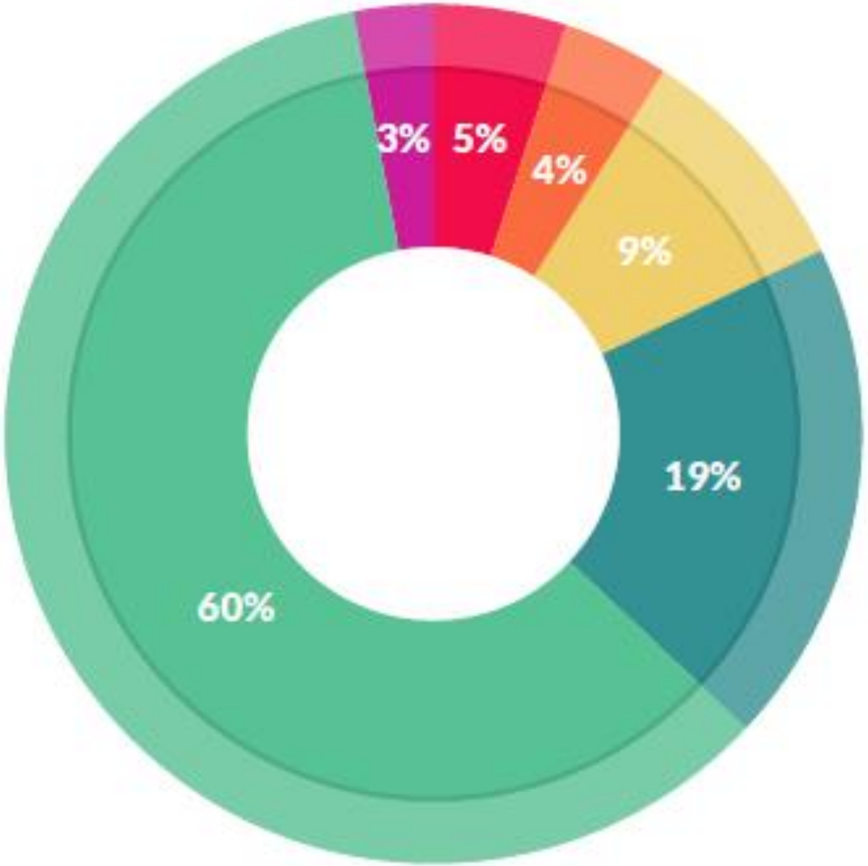
WHY IS IT IMPORTANT?

GIGO (garbage in - garbage out)

Bad data won't produce the right information.

Image source: WellDataLabs

WHY IS IT IMPORTANT?



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: CrowdFlower

KEY STEPS OF WORKING WITH DATA

[8]

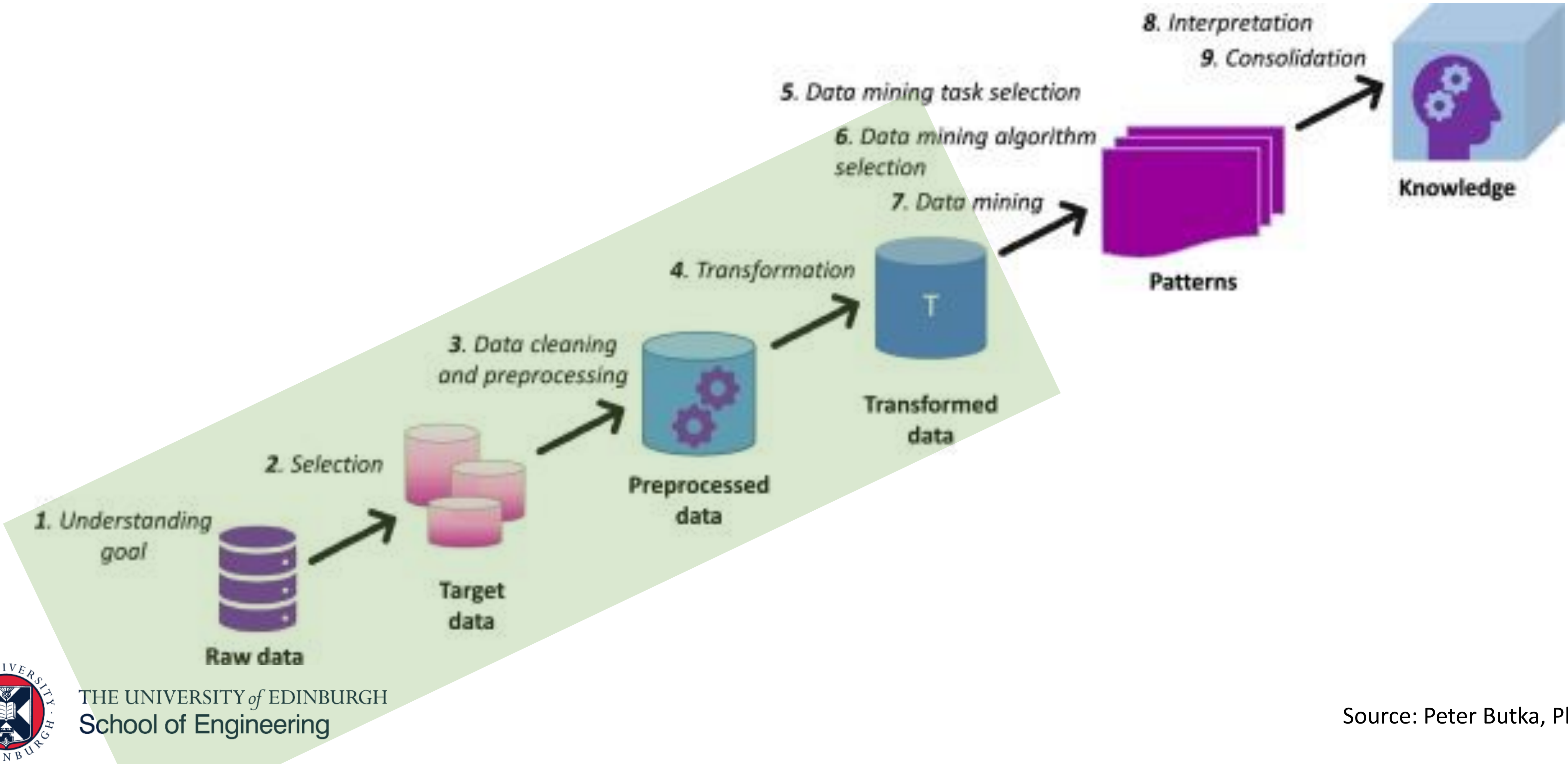
DATA CARPENTRY TAKES THE MOST TIME

DATA MINING AND KNOWLEDGE DISCOVERY PROCESSES

Data carpentry comes before data mining. Data mining is defined as the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, reduce risks and more [2].

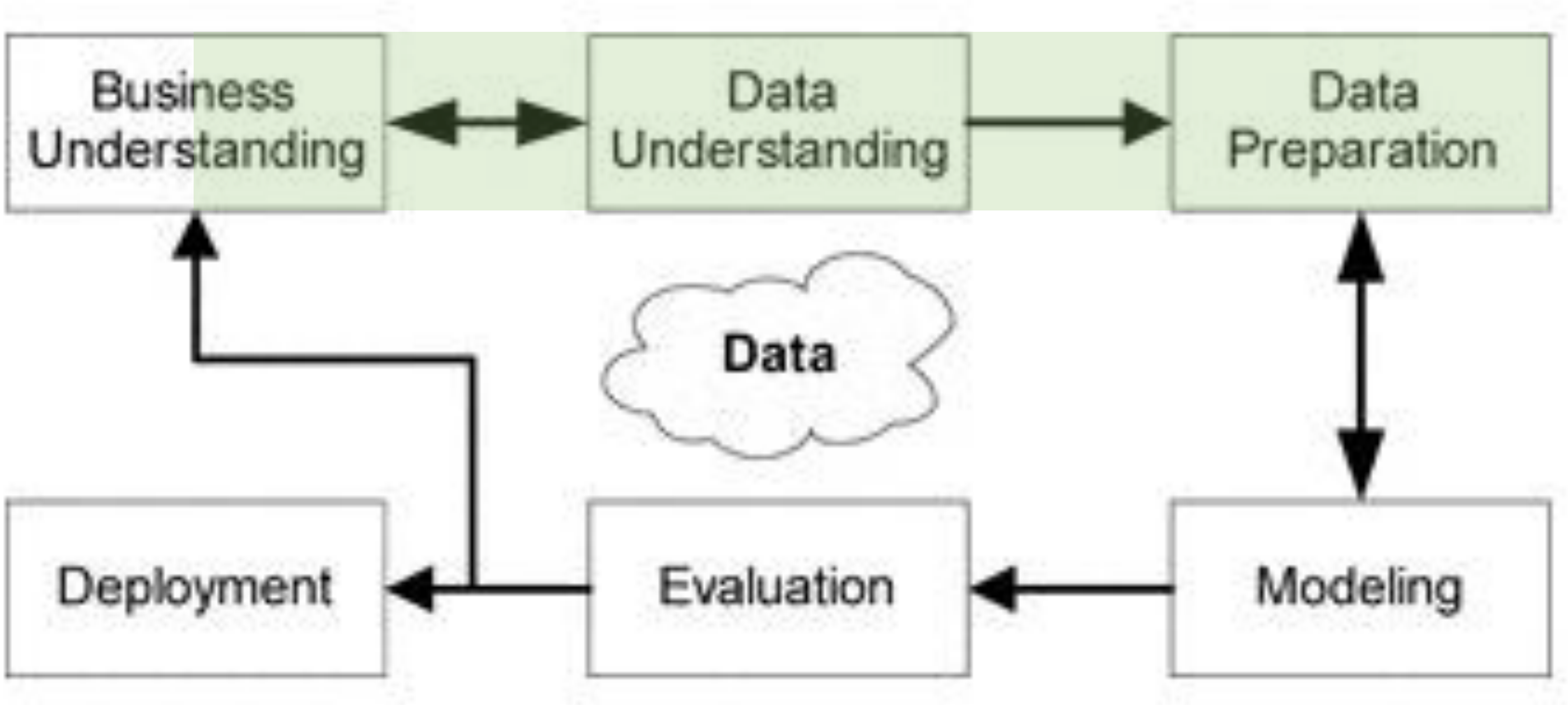
Knowledge Discovery Processes (KDP): The process defines a sequence of steps (with eventual feedback loops) that should be followed to discover knowledge (e.g., patterns) in data. [3]

METHODOLOGIES FOR KNOWLEDGE DISCOVERY PROCESSES: KNOWLEDGE DISCOVERY IN DATABASES (KDD)



Source: Peter Butka, PhD

METHODOLOGIES FOR KNOWLEDGE DISCOVERY PROCESSES: CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)



[4]

CRISP COMPONENTS

CRISP Components	Tasks	Literature and Description
Business understanidng	–Define business objectives – Risk Assessment analysis –Cost and benefit analysis –Technical requirement analysis – Define data analysis objectives and project planning	In Sharma and Osei-Bryson (2009) a framework for implementing various business understanding tasks is presented and highlights dependencies between them. In Sharma and Osei-Bryson (2008); Rao et al. (2012) an organizational-ontology for business understanding is presented. In Nino et al. (2015) various aspects of business understanding and challenges related to big data are discussed
Data understanding and preparation	– Data extraction – Data description – Data quality estimation – Data selection for modeling Data cleaning and feature extraction – Data exploration	In Duch et al. (2004) rule-based data extraction and understanding is discussed. Uddin et al. (2014) discuss various characteristics of big data for its efficient applications. Karkouch et al. (2016); Qin et al. (2016) discuss data properties, life-cycle of data from internet of things (IoT) for maintaining data quality from IoT. Cichy and Rass (2019) reviews various comparisons that provide data quality frameworks from different areas, including industrial production. Hazen et al. (2014); Ardagna et al. (2018) discuss methods for data quality management, monitoring, and assessments. Steed et al. (2017); Zhou et al. (2019); Andrienko et al. (2020)) discuss visualization methods and challenges of manufacturing and big data. Stanula et al. (2018)) discuss guidelines for data selection for understanding business and data in manufacturing
Modeling and evaluation	– Model assumption and selection techniques for modeling, parameter selection – Feature engineering – Model testing, result visualization and analysis – Model evaluation and description – Other data and modeling issues affecting model performance	In Diez-Olivan et al. (2019), Vogl et al. (2019), Bertsimas and Kallus (2020) reviews of various models, models building and evaluation for descriptive, diagnostic, predictive, and prescriptive analysis in industrial production and manufacturing are presented
Deployment	– Model utility assessment – Model monitoring, maintenance and updates – Users response evaluation – Model evaluation for data understanding and business understanding	Issues of model deployment related to human-lefted data science and model safety are discussed in *HUMAN-CENTERED Data Science and Model Safety*

[7]

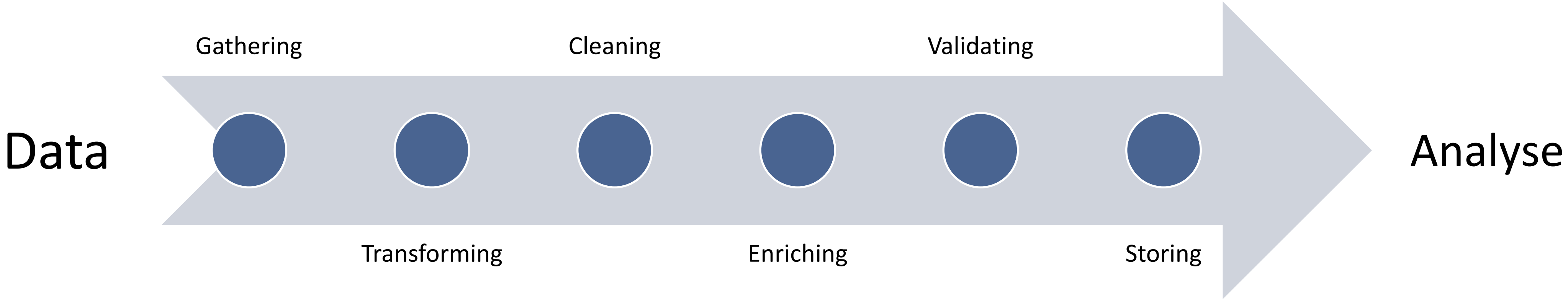
KNOWLEDGE DISCOVERY AND ANALYSIS IN MANUFACTURING (KDAM)

Common applications of KDAM include:

- detection of root causes of deteriorating product quality,
- identification of critical and optimal manufacturing process parameters,
- prediction of effects of manufacturing process changes, and
- identification of root causes and prediction of equipment breakdown. [4]

DATA CARPENTRY PROCESS

DATA CARPENTRY PROCESS



TRANSFORMING

Technique	Definition	Example(s)
Formatting and Encoding	Converting the format of data to the appropriate type	(1) Changing a date that comes to you as a number back to date format; (2) Parsing data that comes from a csv into unique columns; (3) Changing text-based categories to numeric categories so the software can understand it (e.g., Python)
Converting	Transforming data into common units	Transforming global salaries into one common currency so they can be compared
Mathematical Transformation	Using a mathematical process to change data into more useful values	(1) Z-scores (to normalize scale); (2) logarithmic transformations (to make nonlinear data linear)

[5]

TRANSFORMING

Technique	Definition	Example(s)
Append, Merge, Filter, or Join	Bringing data in from different tables into one table.	(1) Adding columns to your data, like adding performance data to employee record data; (2) Adding rows to your data, like combining reps from the South Region to a data table containing reps from the East Region; (3) Using criteria to decide which records or fields are included, like creating a table of only employees who are present in both 2019 and 2020 performance reports
Binning	Turning a continuous variable into a categorical one	Turning an engagement survey score (0–100) into a category like “high,” “medium,” and “low”

DATA TRANSFORMATION

Not an exhaustive list of all the transformation techniques rather an indicative source of information.

- Changing data types (discretisation)
- Changing range of data values (normalisation)

Source: Isabelle Bichindaritz, SUNY Oswego

DATA CLEANING

Video source: Data wrangling with MongoDB

DATA CLEANING

Data Cleaning Vs Data Wrangling

DATA CLEANING	DATA WRANGLING
Process of removing corrupted or inaccurate records from a table, database or record set.	Process of transforming and mapping data from one raw data into another form with the intent of making it more appropriate and valuable for various task.
Also called Data Cleansing	Also called Data Munging

Slide source: Dr. Monica

CLEANING

Technique	**Definition**	**Example(s)**
Imputing	Filling in gaps in your data with educated guesses based on data you do have	Assuming a score on an omitted survey item based on scores provided from similar questions
Check for bad values	Removing corrupted data and bad values. Removing outliers that can potentially skew the results.	In a dataset with values ranging from 1 to 100 there is one that is 300.
Remove duplicates	Removing duplicate entries	Selling data from a specific customer being registered twice
Free-form clean-up	Transforming free-form input data into a standard value.	Changing the values “RU,” “Rutgers,” “Rutgers U,” and “Rutgers New Brunswick” to “Rutgers University” so they can all be tabulated together

CAUSES FOR DATA INACCURACY

- Input errors (human errors)
- Design errors
- Incorrect source data (machine errors)

TOOLS

- Excel
- OpenRefine
- Python

ENRICHING

- The step of enriching your data is optional and refers to the stage of data carpentry where you can augment your data with other data. The combination of your data with data accumulated from other sources can lead to improved accuracy and more meaningful insights [6]. An example would be combining two databases of supplier information where one contains supplier addresses and the other one doesn't.

VALIDATING

According to the Simon (2013), an operational definition of data validation is:

"Data validation could be operationally defined as a process which ensures the correspondence of the final (published) data with a number of quality characteristics."

VALIDATING

The rules of data validation require repetitive programming processes that help to verify the following:

- Quality
- Consistency
- Accuracy
- Security
- Authenticity

[9]

VALIDATING

- Review

Analysis of feedback from stakeholders

Identifying and prioritising problems

- Design

Study of datasets, variables and their relations

Assess quality requirements

Satisfactory set of validation rules

- Execute

Data are checked against the validation rules

Measure results

- Implement

Validation rules are formalized

Metrics for data validation rules

Testing and evaluation of results

Refinement of validation rules

STORING

After completing all the steps of data carpentry correctly you should end up with a high-quality dataset that can then be used for analysis to gain insights.

- Store your data according to your company's policies
- Follow a data management process and be mindful of sensitive data.
- Store your data where they can be easily accessed based on the tool you use for analysis.
- Store your data having backward compatibility in mind in case you’ll need to perform analysis in the future

STORING

After completing all the steps of data carpentry correctly you should end up with a high-quality dataset that can then be used for analysis to gain insights.

- Store your data according to your company's policies
- Follow a data management process and be mindful of sensitive data.
- Store your data where they can be easily accessed based on the tool you use for analysis.

REFERENCES

1. Grieves, Michael. (2005). Product Lifecycle Management: Driving the Next Generation of Lean Thinking.
2. Sas.com. n.d. What is data mining?. [online] Available at: <https://www.sas.com/en_sg/insights/analytics/data-mining.html>.
3. Cios, K., 2010. Data mining. New York: Springer.
4. Mark Polczynski & Andrzej Kochanski (2010) Knowledge Discovery and Analysis in Manufacturing, Quality Engineering, 22:3, 169-181, DOI: [10.1080/08982111003742855](10.1080/08982111003742855)
5. Rosett C.M., Hagerty A. (2021) Data Wrangling. In: Introducing HR Analytics with Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-030-67626-1_13
6. Stefanski, R., Sinha, V. and Poddar, A., 2022. *Data Wrangling in 6 Steps: An Analyst's Guide For Creating Useful Data*. [online] Learn | Hevo. Available at: https://hevodata.com/learn/data-wrangling/#s2
7. Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M. and Emmert-Streib, F., 2021. Ensuring the Robustness and Reliability of Data-Driven Knowledge Discovery Models in Production and Manufacturing. *Frontiers in Artificial Intelligence*, 4.
8. Di Zio, Marco, et al. "Methodology for data validation 1.0." Essnet Validat Foundation, Brussels, Belgium (2016): 1-76.