

Lecture 2

How to Conduct a Data Mining Project?

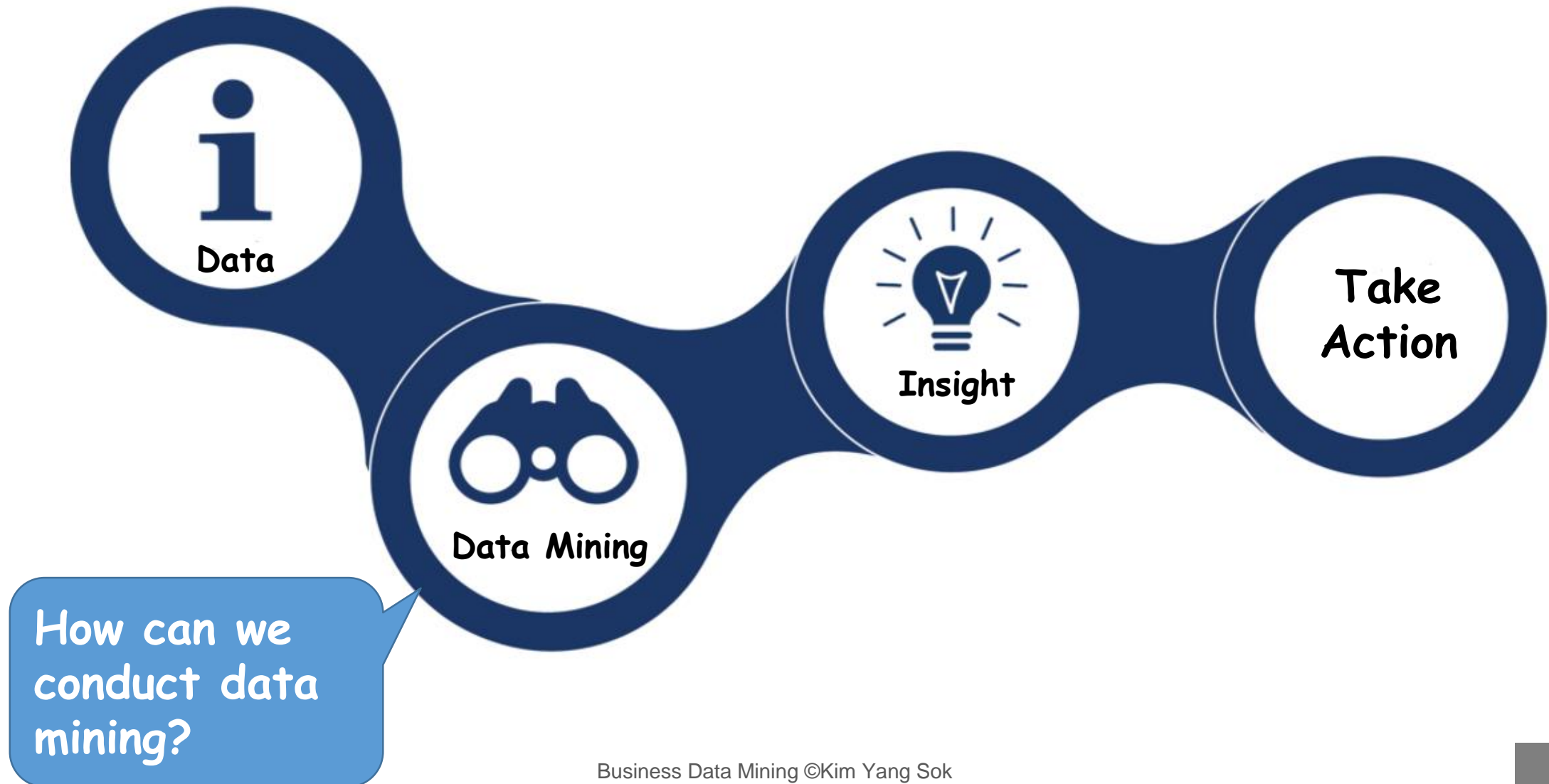
Kim, Yang Sok
Dept. of MIS, Keimyung University

- Introduction
- CRISP-DM Background
- CRISP-DM Reference Model
 - Business Understanding
 - Data Understanding
 - Data Preparation
 - Modeling
 - Evaluation
 - Deployment
- Conclusion

Introduction

Introduction

What Use Data Mining?



CRISP-DM Background

CRISP-DM Background

There are many ideas...



People have different ideas about "how to conduct data mining." So this causes many problems. What kind of problems do you expect?

- **Must be reliable and repeatable by people with little data mining background.**
- **Framework for recording experience**
- **Allows projects to be replicated**
- **Aid to project planning and management**
- **“Comfort factor” for new adopters**
- **Reduces dependency on “stars”**

CRISP-DM Background

Meeting Together



- **Initiative launched in late 1996 by three “veterans” of data mining market.**
 - Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) , NCR
- **Developed and refined through series of workshops (from 1997-1999)**
 - Over 300 organization contributed to the process model
- **Published CRISP-DM 1.0 (1999)**
 - Over 200 members of the CRISP-DM(Cross Industry Standard Process for Data Mining) SIG worldwide
 - Download: <https://www.the-modeling-agency.com/crisp-dm.pdf>

- **CRISP-DM does not belong to a company or an organization.**
- **CRISP-DM is application, industry, and tool neutral**
- **CRISP-DM focus on business issues**
- **CRISP-DM is a framework for guidance based on data experience**

Business Understanding

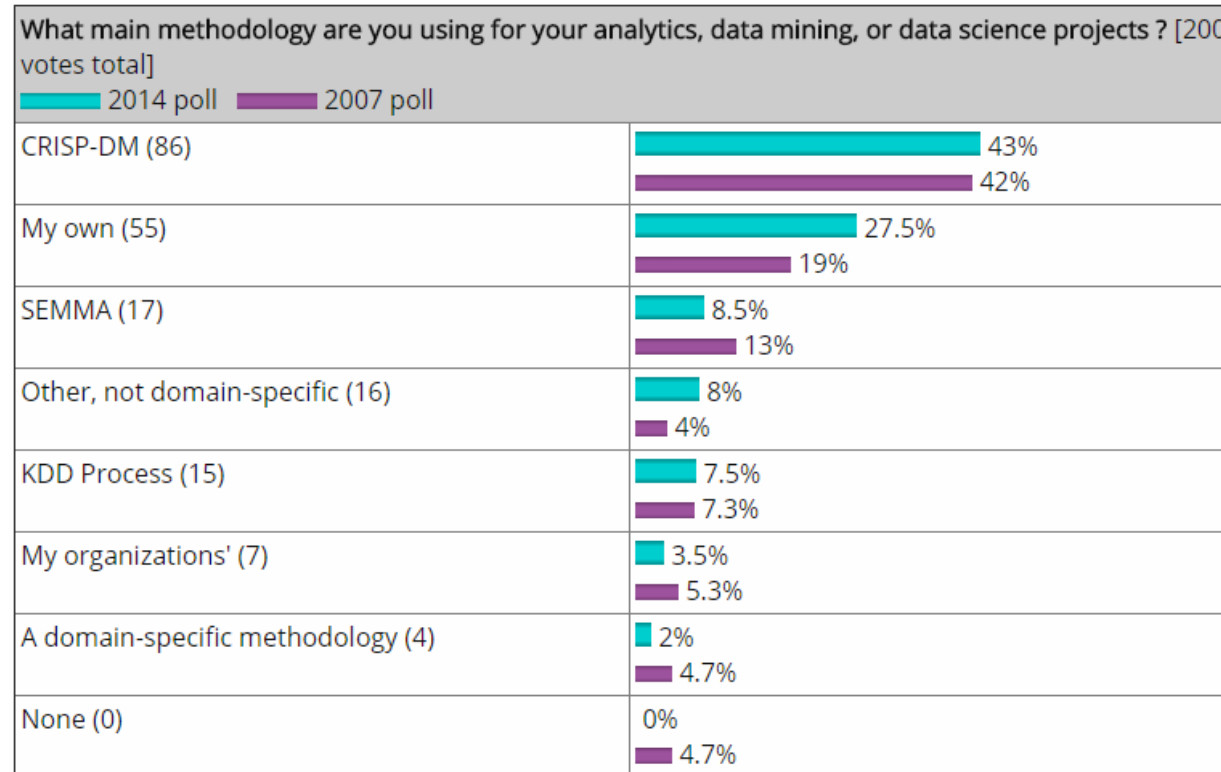
America's highest earners work at least 60 hours a week—more than anyone else in the world



By [Richard Koch](#) • October 13, 2013



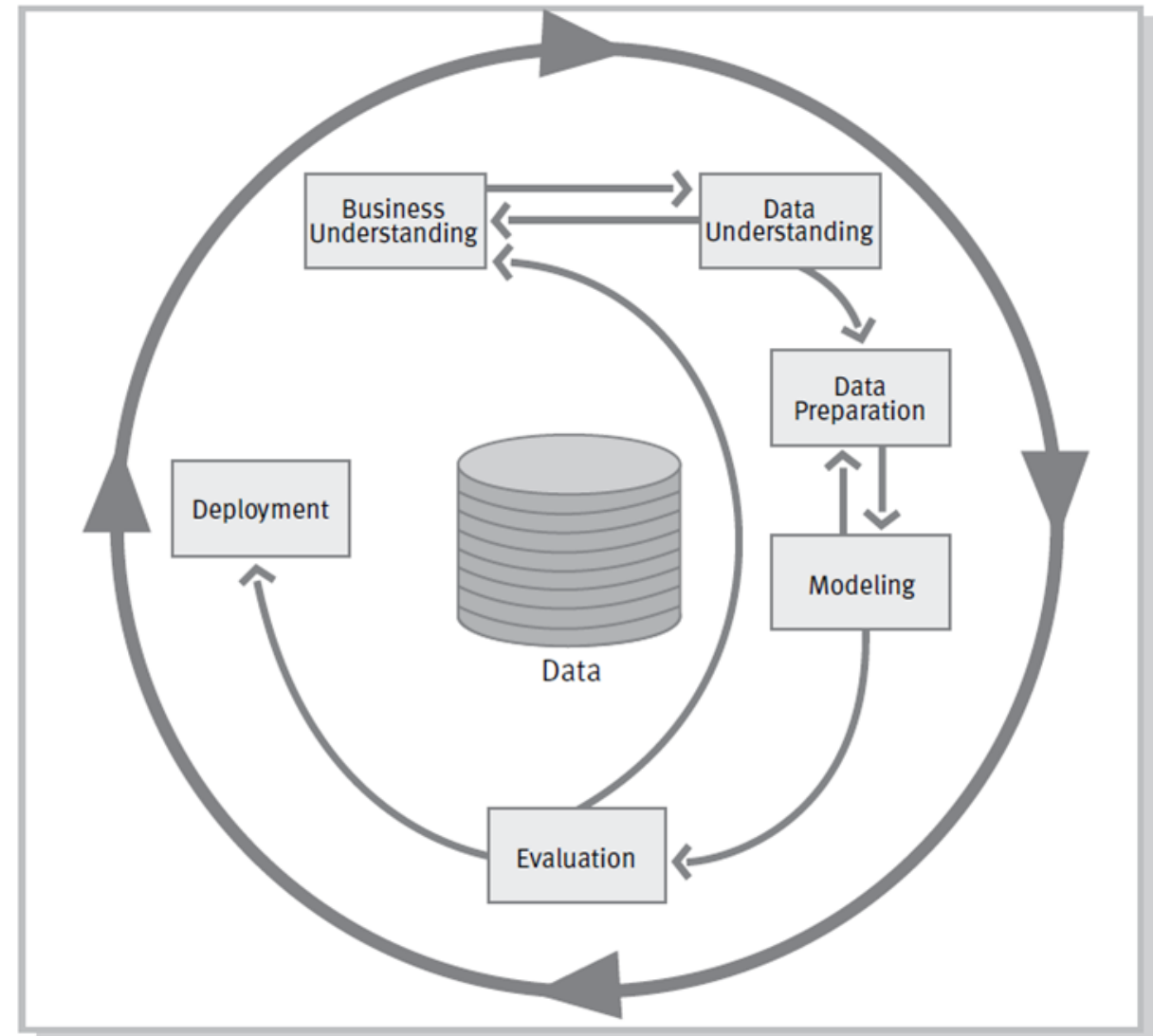
- **CRISP-DM is still the most preferred data mining / data science methodology.**



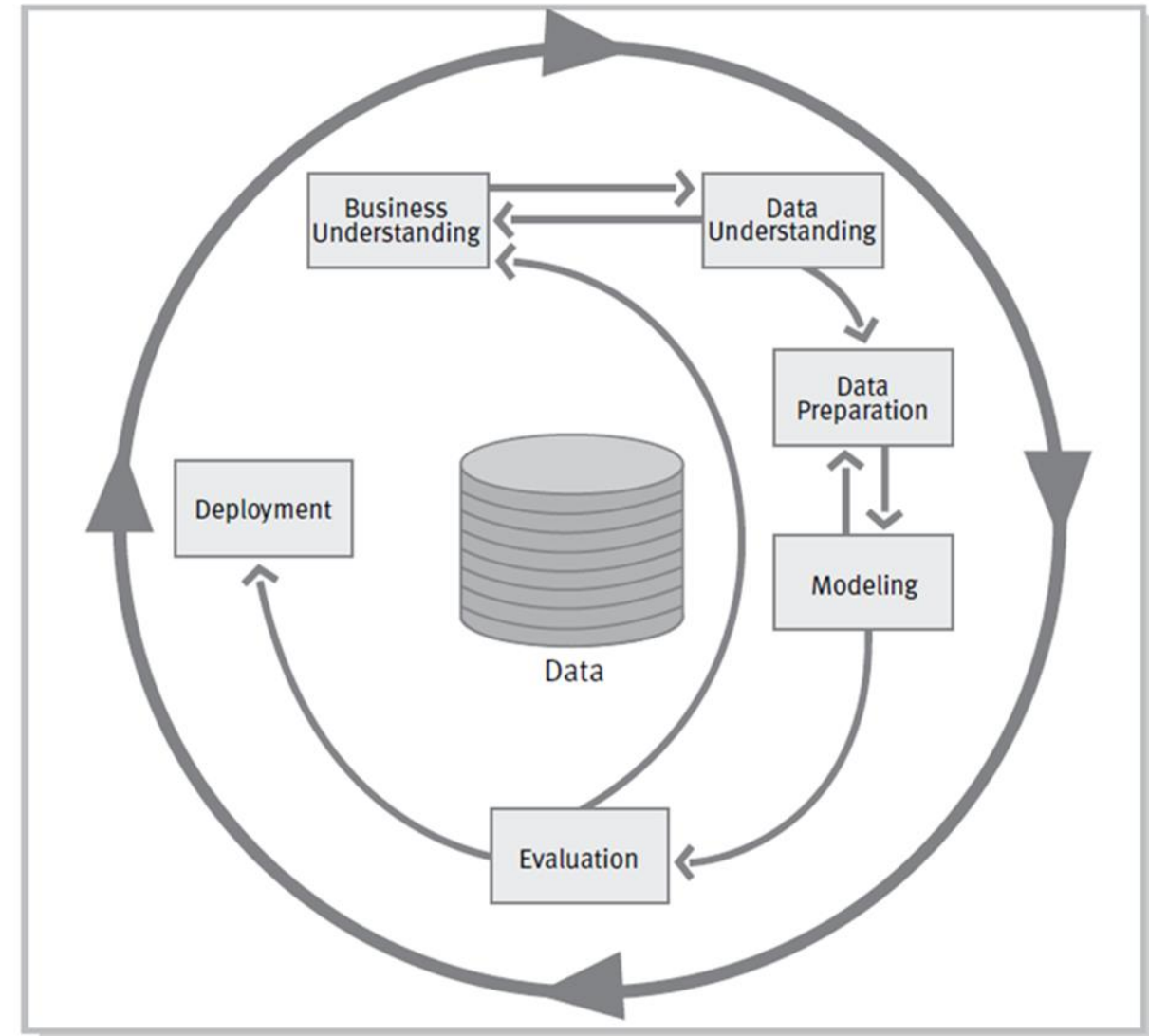
<http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>

CRISP-DM Reference Model

- Provides an overview of the life cycle of a data mining project.
- Contains the phases of a project, their respective tasks and relationships between these tasks.
- The life cycle of a data mining project consists of **six phases**.



- The sequence of the phases is not rigid.
- The arrows indicate the **most important and frequent dependencies** between phases.
- The outer circle symbolizes the **cyclical nature of data mining itself**.
 - Data mining is not over once a solution is deployed.



Business Understanding

- **Determine business objectives**
- **Assess situation**
 - Resources (data!), risks, costs & benefits
- **Determine data mining goals**
 - Ideally with quantitative success criteria
- **Develop project plan**
 - Estimate time line, budget, but also tools and techniques

- **Difficult!**
- **Often, you have to enter a new field**
- **You have to explain data mining limitations to non-experts**
 - No, performance will not be 100%
 - Need much more data to train an accurate model
- **Have a lot of patience for vaguely defined problems**
- **Learn to concretize or even reduce the scope of the initial idea**
 - Data sample
 - Real-life use cases
 - Quantitative success metrics
- **Do not waste your time on ill-defined, unrealistic projects**

Business Understanding

Case: Wine Quality Prediction Project

Do you like wines?



Business Understanding

Case: Wine Quality Prediction Project

Which wine is the best?



Business Understanding

Case: Wine Quality Prediction Project



Three wine specialists give scores for each wine from 1 to 10 and average their scores.



Business Understanding

Case: Wine Quality Prediction Project

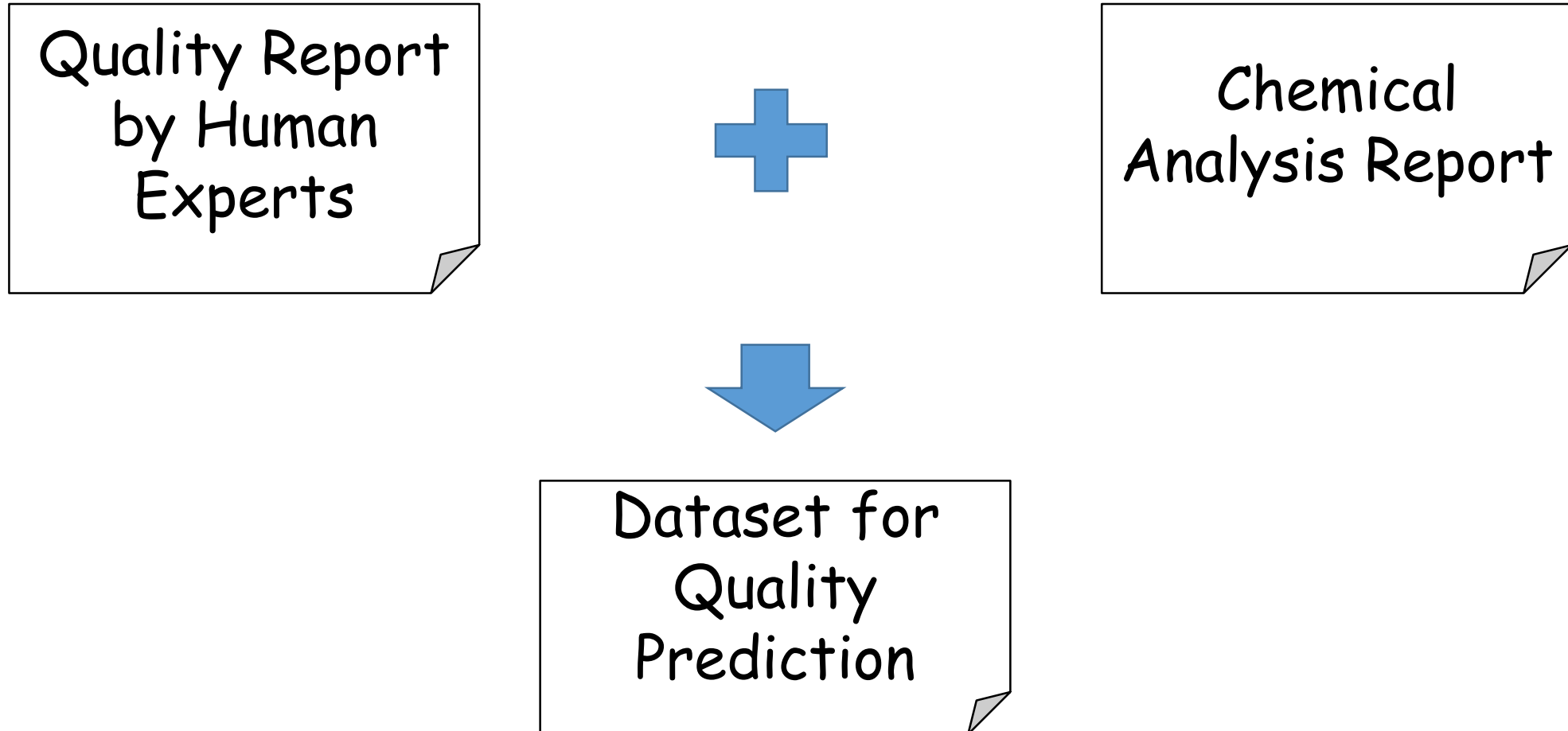
I think this process too much cost and takes too long time. Show me new solution!!



Business Objectives

We want to develop a model that predict wine quality without human expert.

There are two data!



Data Mining Goal

We want to develop a model that predict wine quality with small error
(less than ± 0.5).

Project Plan – Wine Quality Prediction

Business Goal

The project aims to develop a model or multiple models that predict the wine quality.

Business Situation

Measuring wine quality is important because it directly impact on our pricing and marketing strategy. Current quality measuring process depends on human. It causes several problems including cost and delays.

We found two interesting data – one is records of human expert's quality evaluation results and the other is records of chemical evaluation results. We view these data are important because there are studies that show chemical ingredients in wine impact on the quality of wines.

Data mining Goal

Our model should predict the quality with the following error limits : **RMSE <0.5**

Data Understanding

- **Collect initial data**
- **Describe data**
 - Statistical summary
- **Explore data**
 - Use charts
- **Verify data quality**
 - Carefully document problems and issues found!

- **Do not economize on this phase**
 - The earlier you discover issues with your data the better (yes, your data will have issues!)
 - Data understanding leads to domain understanding, it will pay off in the modelling phase
- **Do not trust data quality estimates provided by your customer**
- **Verify as far as you can, if your data is correct, complete, coherent, de-duplicated, representative, independent, up-to-date, stationary**
- **What sort of processing was applied to the raw data**
- **Understand anomalies and outliers**

Data Understanding

Case: Wine Quality Prediction Project

The project team collected data...

Quality Data

	A	B
1	quality	id
2	5	1
3	5	2
4	5	3
5	6	4
6	5	5
7	5	6
8	5	7
9	7	8
10	7	9
11	5	10
12	5	11
13	5	12
14	5	13
15	5	14

Chemical Data

	A	B	C	D	E	F	G	H	I	J	K	L
1	fixed acid	volatile ac	citric acid	residual su	chlorides	free sulfur	total sulfu	density	pH	sulphates	alcohol	id
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	1
3	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	2
4	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	3
5	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	4
6	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	6
8	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	7
9	7.3	0.65	0	1.2	0.0			0.9946	3.39	0.47	10	8
10	7.8	0.58	0.02	2	0.0			0.9968	3.36	0.57	9.5	9
11	7.5	0.5	0.36	6.1	0.0			0.9978	3.35	0.8	10.5	10
12	6.7	0.58	0.08	1.8	0.0			0.9959	3.28	0.54	9.2	11
13	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	12
14	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	13
15	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	14

Generated by
chemical
analysis tool

Use the same key

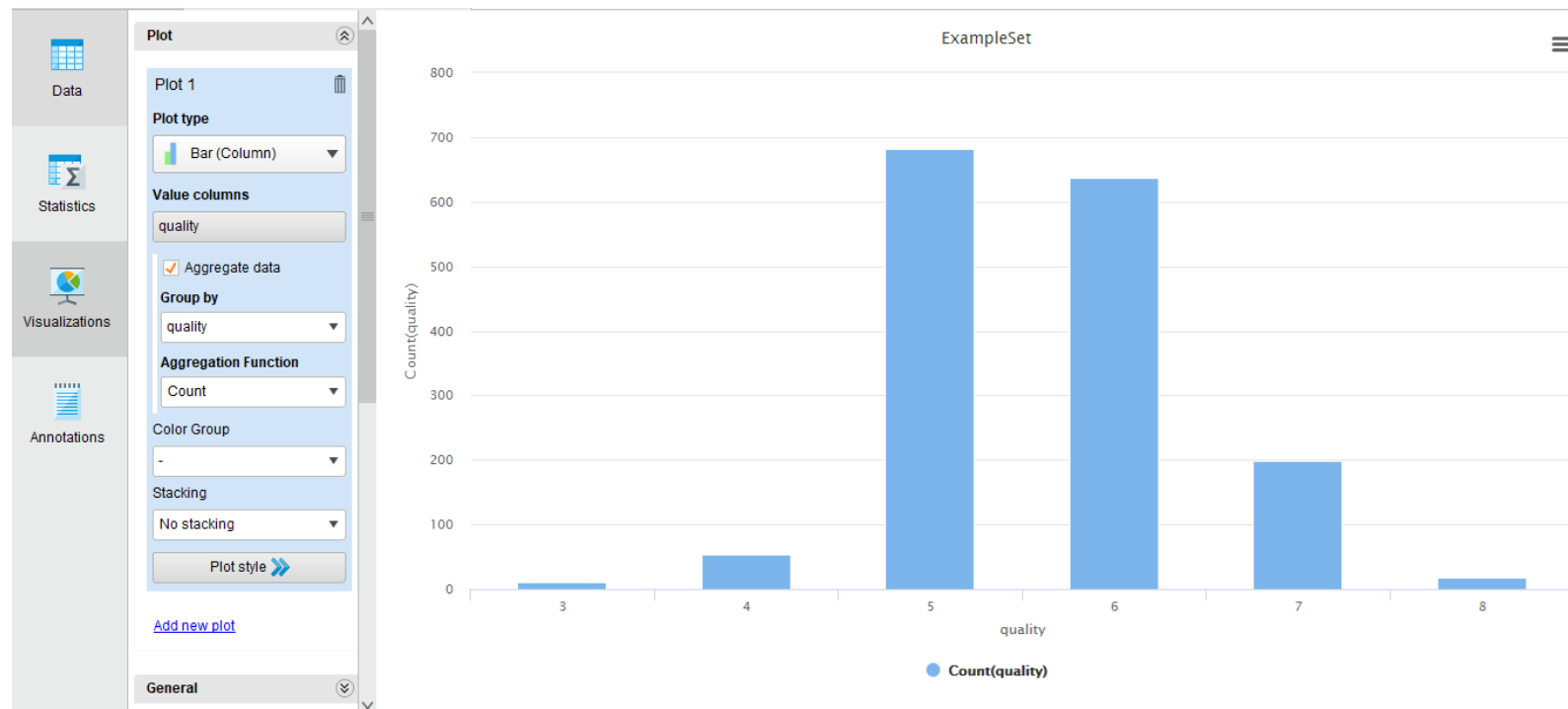
Generated by
Human Experts

Data Understanding

Case: Wine Quality Prediction Project

Get statistics & draw chart.

Name	Type	Missing	Statistics			Filter (2 / 2 attributes): <input type="text" value="Search for Attributes"/>
quality	Integer	0	Min 3	Max 8	Average 5.636	
id	Integer	0	Min 1	Max 1599	Average 800	

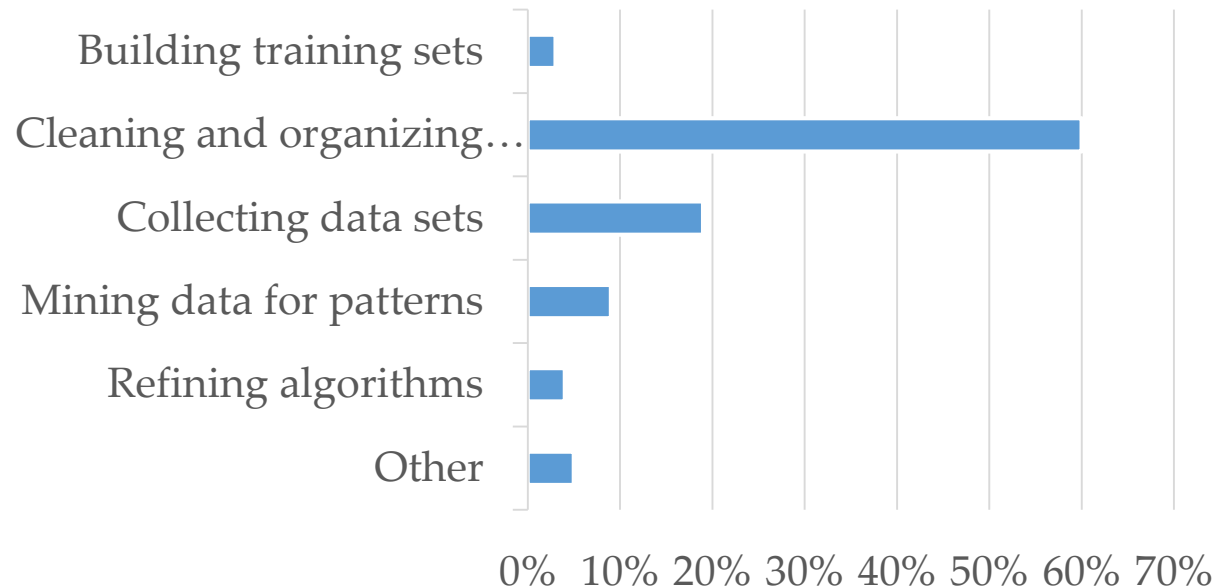


Data Preparation

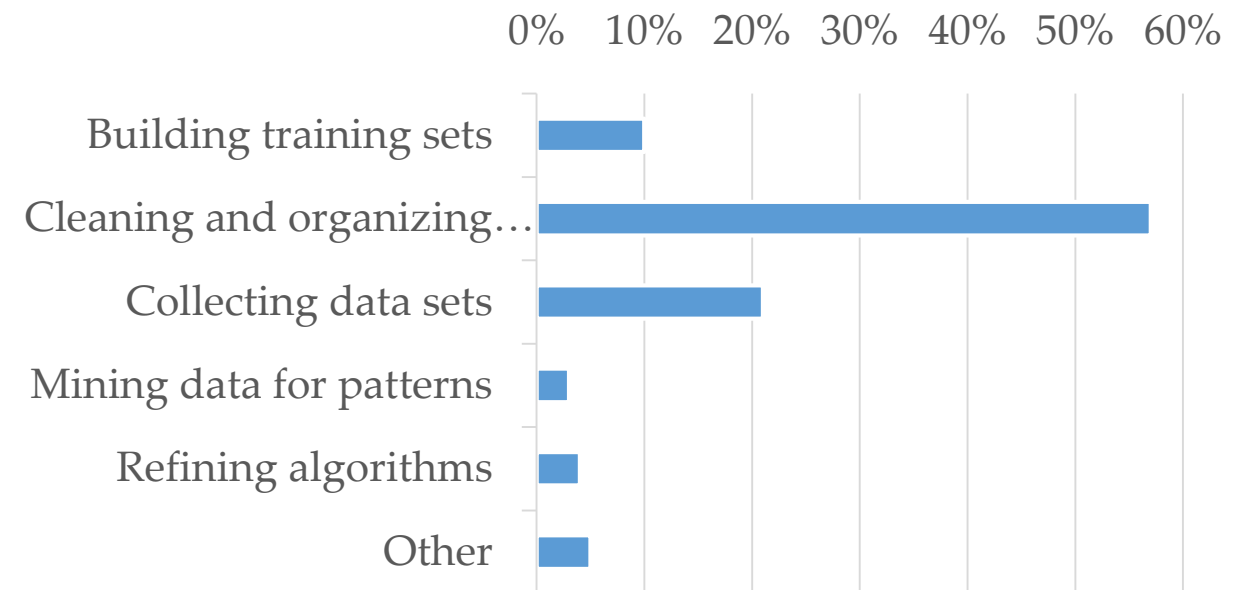
- **Select data**
 - Select attributes (columns) and examples (data, rows)
- **Clean data**
 - Handling missing values, normalization, outliers...
- **Construct data**
 - Generate derived attributes
- **Integrate data**
 - Merge information from different sources
- **Format data**
 - Convert to format convenient for modelling

- **Tedious!**
- **Use workflow tools to document, automate & parallelize data prep.**
- **Data understanding and preparation will usually consume half or more of your project time!**

What data scientists spend the most time doing



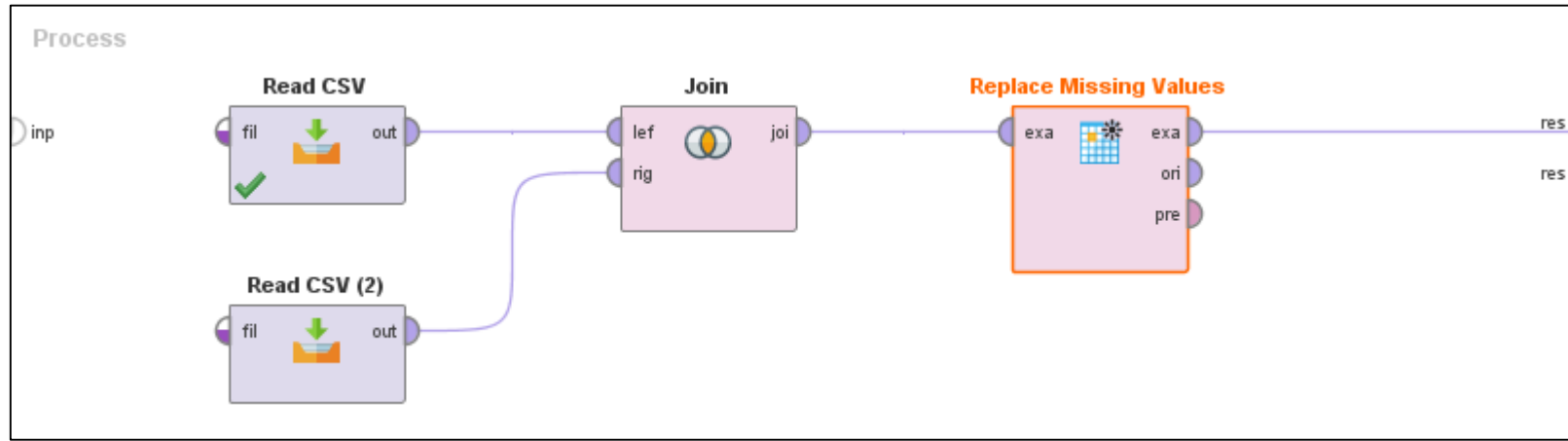
What's the least enjoyable part of data science?



- Automate this phase as far as possible
- When merging multiple sources, track provenance of your data
- Use workflow tools to help you with the above
- Prepare your customer that data understanding and preparation take considerable amount of time

Data Understanding

Case: Wine Quality Prediction Project



Row No.	id	quality	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol
1	1	5	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400
2	2	5	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	9.800
3	3	5	7.800	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650	9.800
4	4	6	11.200	0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580	9.800
5	5	5	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400
6	6	5	7.400	0.660	0	1.800	0.075	13	40	0.998	3.510	0.560	9.400
7	7	5	7.900	0.600	0.060	1.600	0.069	15	59	0.996	3.300	0.460	9.400
8	8	7	7.300	0.650	0	1.200	0.065	15	21	0.995	3.390	0.470	10

Modeling

- **Select modelling technique**
 - Assumptions, measure of accuracy
- **Generate test design**
- **Build model**
 - Conduct feature engineering and optimize model parameters
- **Assess model**
 - Iterate the above

- **Be creative with your features (feature engineering)**
 - Especially from textual data or time-series you can generate a lot of standard features
 - Make conscious decision about missing data and outliers (regression!)
- Allocate time for **hyperparameter optimization**
- **Whenever possible, look inside your model and consult it with domain expert**
 - Assess feature importance
 - Run your model on simulated data

- **Classification**



Who will respond our marketing program?

Modeling

What Data Mining Techniques?

• Prediction (Regression)

What
quality level?

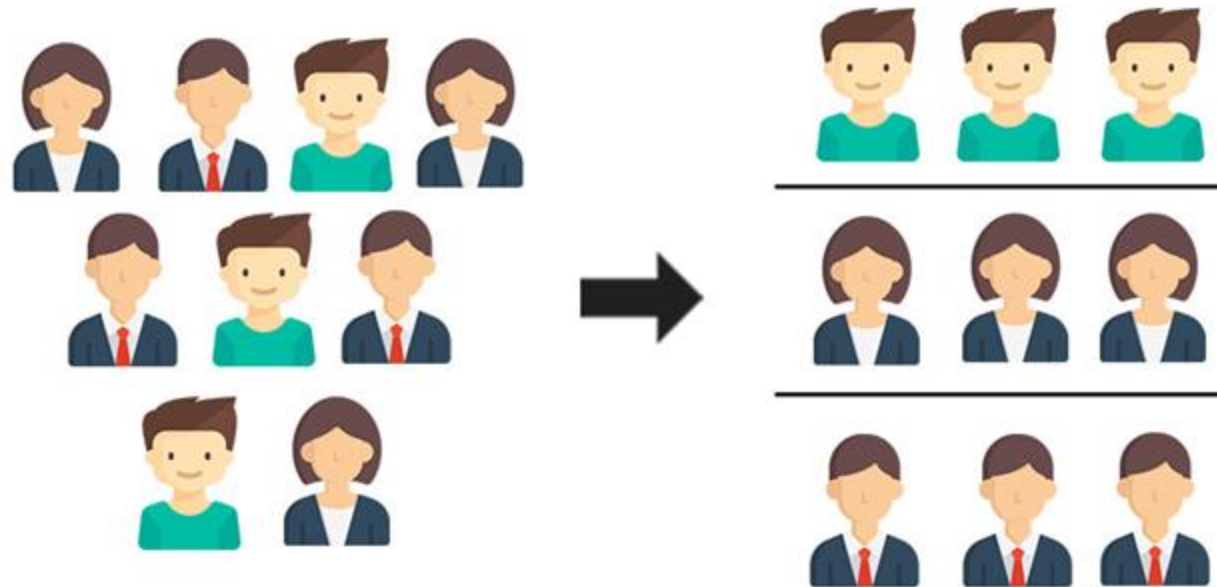
Real
quality

Predicted
quality

$$\text{Error} = 5 - 4.7 = 0.3$$



- **Clustering / Segmentation**

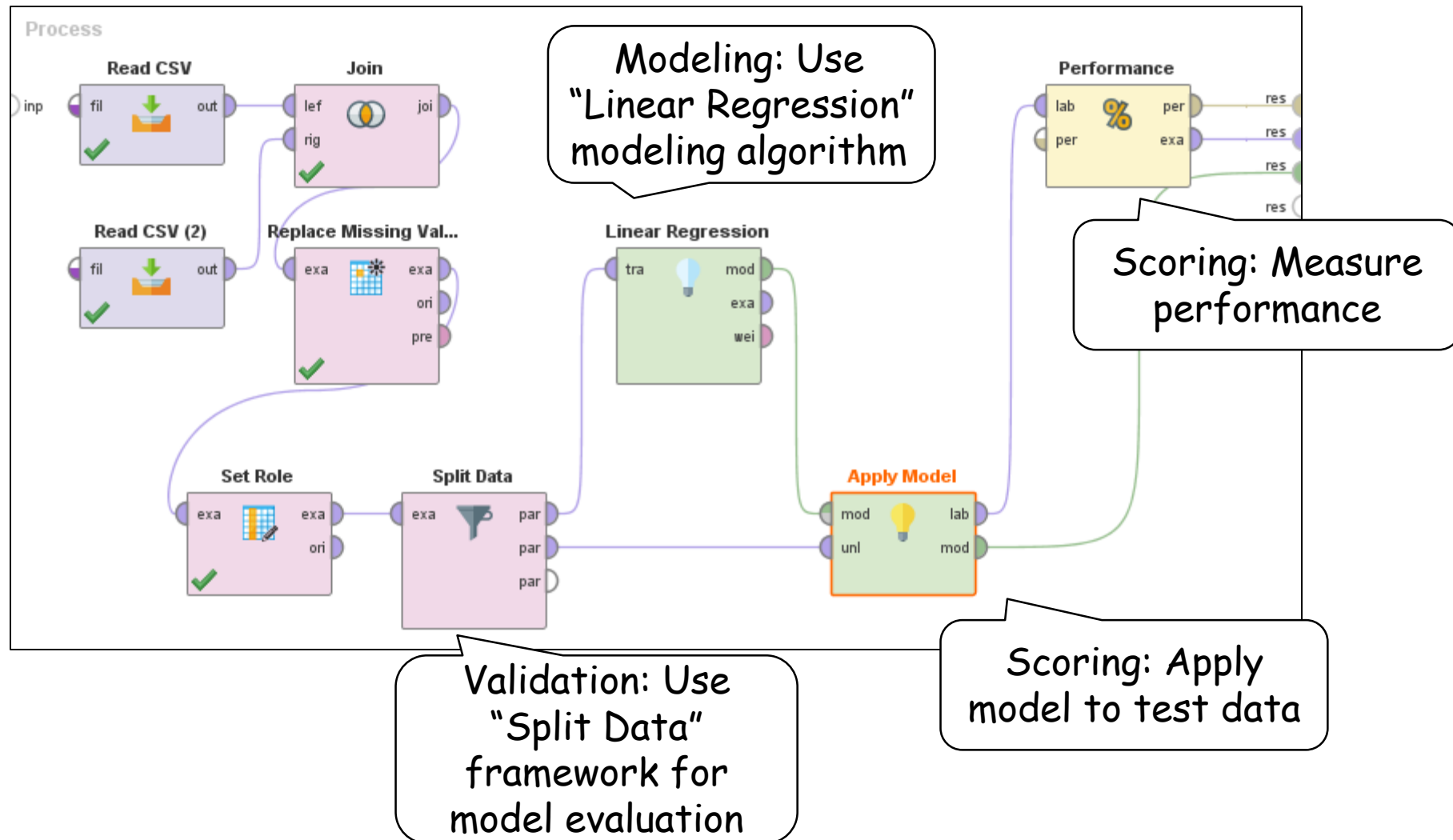


- **Association Rules (Dependency analysis; Basket Analysis)**



Modeling

Case: Wine Quality Prediction Project



Modeling

Case: Wine Quality Prediction Project

LinearRegression (Linear Regression) × ExampleSet (Apply Model) × PerformanceVector (Performance) ×



Criterion
root mean squared error

root_mean_squared_error

root_mean_squared_error: 0.640 +/- 0.000

Performance

LinearRegression (Linear Regression) × ExampleSet (Apply Model) × PerformanceVector (Performance) ×

Open in  Turbo Prep  Auto Model

Filter (480 / 480 examples): all

Row No.	id	quality	prediction(q...	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH
1	1	5	5.012	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510
2	2	5	5.093	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200

Model
Prediction

LinearRegression (Linear Regression) × ExampleSet (Apply Model) × PerformanceVector (Performance) ×

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
fixed acidity	0.019	0.020	0.039	0.956	0.944	0.345	
volatile acidity	-1.256	0.144	-0.270	0.827	-8.740	0	****
citric acid	-0.249	0.178	-0.059	0.816	-1.402	0.161	
residual sugar	0.020	0.015	0.034	1.000	1.372	0.170	
chlorides	-2.012	0.486	-0.122	0.994	-4.141	0.000	****
total sulfur dioxide	-0.002	0.001	-0.072	0.962	-2.719	0.007	***
pH	-0.513	0.186	-0.098	0.997	-2.764	0.006	***
sulphates	0.903	0.133	0.188	0.972	6.810	0.000	****
alcohol	0.301	0.021	0.395	0.883	14.458	0	****
(Intercept)	4.395	0.727	?	?	6.044	0.000	****

Model

Evaluation

- **Evaluate results**
 - Business success criteria fulfilled?
- **Review process**
- **Determine next steps**
 - To deploy or not to deploy?

- **Work with the performance criteria dictated by your customer's business model**
- **Assess not only performance, but also practical aspects, related to deployment, for example:**
 - Training and prediction speed
 - Robustness and maintainability (tooling, dependence on other subsystems, library vs. homegrown code)
- **Watch out for data leakage, for example:**
 - Time series - mixing past and future
 - Meaningful identifiers
 - Other nasty ways of artificially introducing extra information, not available in production

Deployment

- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project
 - Collect lessons learned!

- Deploying analytic models can be a long, slow moving process with many obstacles along the way.
- Many models are abandoned before they ever make it into production because of inefficiencies that slow down or halt the entire process.
- To overcome the challenges of model deployment, we need to identify the problems and learn what causes them.

- Some of the top technical challenges organizations face when trying to deploy a model into production are:
- The model is not compatible with the production environment.
- The model is not portable.
- The organization has a monolithic architecture.
- The model does not scale.

<https://www.opendatagroup.com/blog/technical-challenges-of-model-deployment>

Conclusion

- **Why CRISP-DM?**
 - The data mining process must be reliable and repeatable by people with little data mining skills
- **CRISP-DM provides a uniform framework for**
 - guidelines
 - experience documentation
- **CRISP-DM is flexible to account for differences**
 - Different business/agency problems
 - Different data



QUESTIONS?