dnc

apresenta

Data Science & Machine Learning

data expert.

Data Science & Machine Learning

dnc

Aula 1: Boas Vindas

Consultor: Daniel Soria

O que veremos neste módulo

1. Definindo o Data Preparation;
2. Dataset e Dataset Description;
3. Select Data;
4. Clean Data;
5. Construct Data;

O que veremos neste módulo

6. Integrate Data;
7. Format Data
8. Resumo do Data Preparation;
9. Data Preparation na Prática 1;
10. Data Preparation na Prática 2;

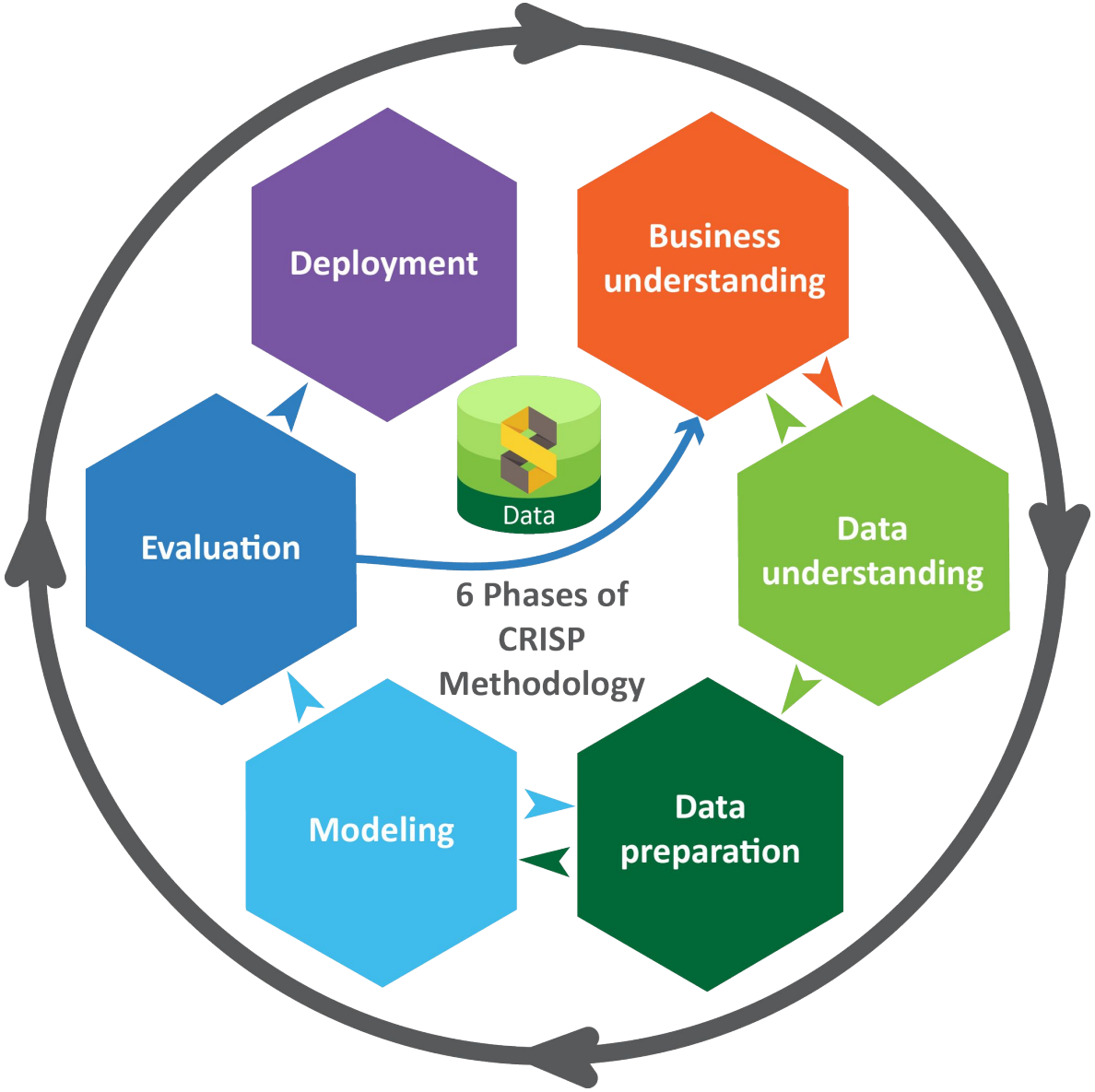
data expert.

Data Science & Machine Learning

dnc

Aula 2: Definindo Data Preparation

Consultor: Daniel Soria

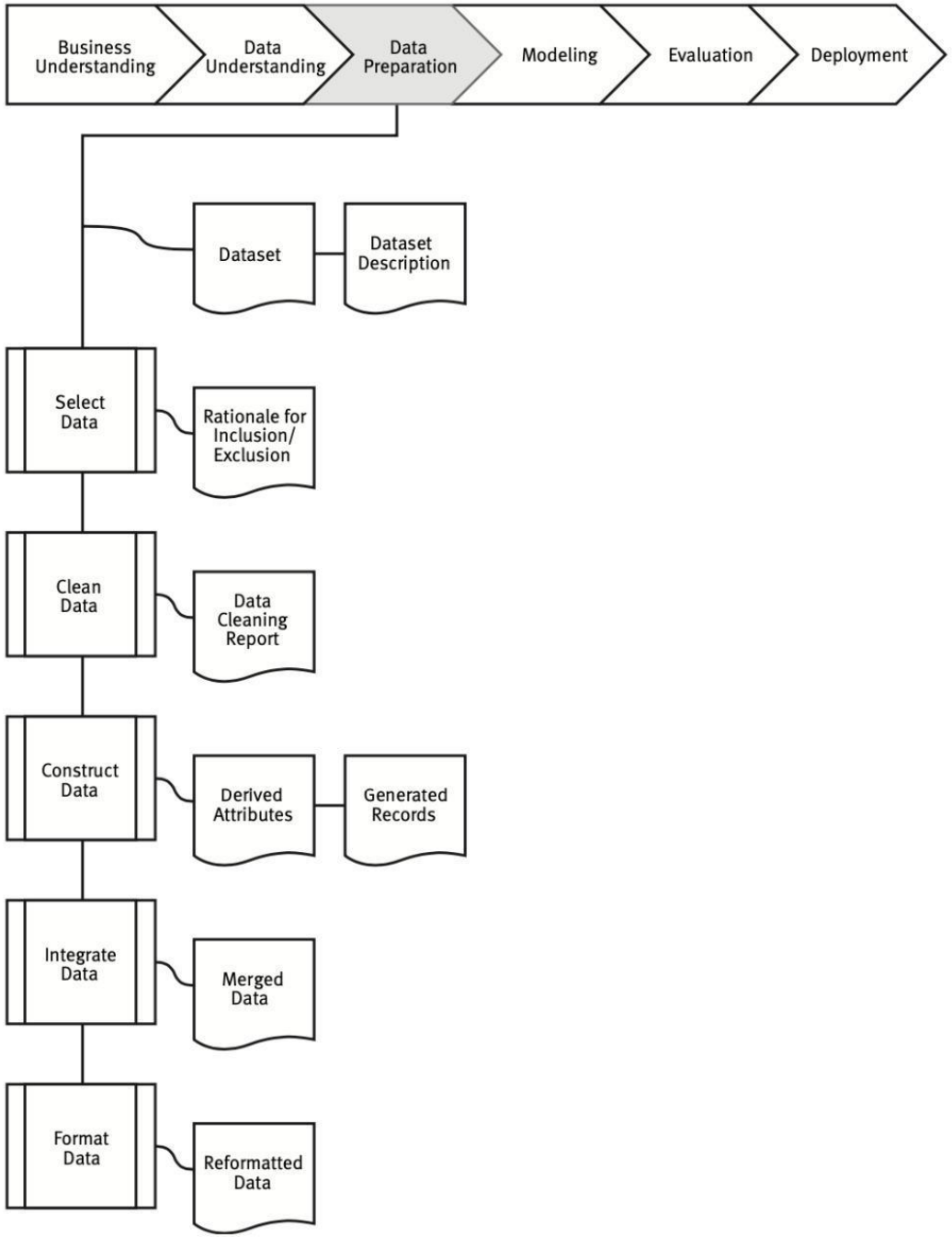

Deployment
Business understanding
Evaluation
Data
Data understanding
6 Phases of CRISP Methodology
Modeling
Data preparation

Data Preparation é a fase mais longa do projeto.

60% - 80% do teu projeto está nessa fase.

Um dos grandes responsável pelo sucesso do projeto

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Dataset
Dataset Description
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data

data expert.

Data Science & Machine Learning

dnc

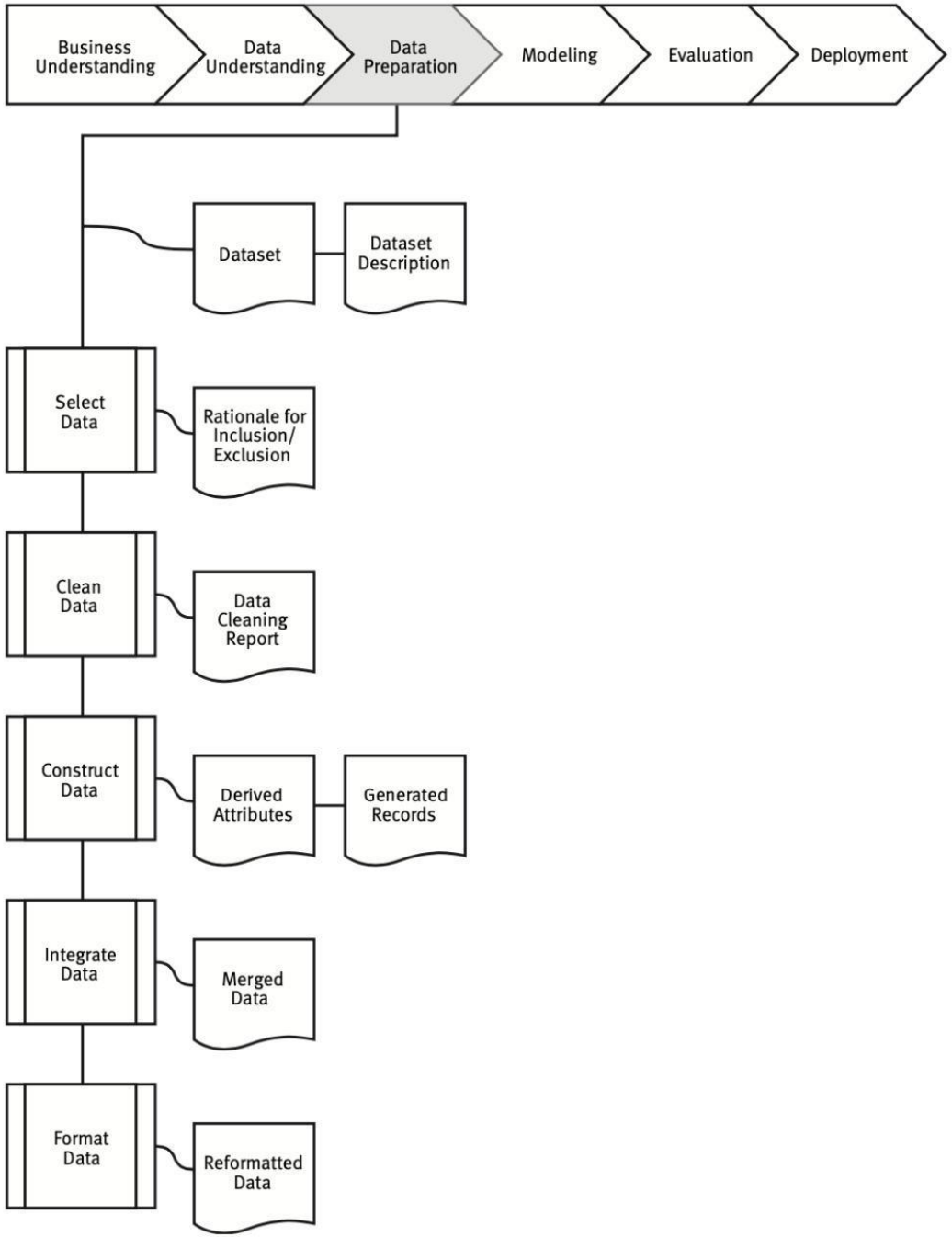
Aula 3: Dataset e Data Description

Consultor: Daniel Soria

Dataset é o conjunto de dados.

Raw Dataset é conjunto de dados bruto.

Dataset é o conjunto de dados que você vai produzir e preparar nessa fase de Data Preparation.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Dataset
Dataset Description
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000.0	2015-08-11 12:12:28	0.00	failed	0	GB	0.00
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000.0	2017-09-02 04:43:57	2421.00	failed	15	US	100.00
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000.0	2013-01-12 00:20:50	220.00	failed	3	US	220.00
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000.0	2012-03-17 03:24:11	1.00	failed	1	US	1.00

dnc

Dataset Description é um outro output

Describe the dataset(s) that will be used for the modeling and the major analysis work of the project.

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000.0	2015-08-11 12:12:28	0.00	failed	0	GB	0.00
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000.0	2017-09-02 04:43:57	2421.00	failed	15	US	100.00
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000.0	2013-01-12 00:20:50	220.00	failed	3	US	220.00
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000.0	2012-03-17 03:24:11	1.00	failed	1	US	1.00

Feature	Type	COUNT	UNIQUE	Description
ID	INT	378661	-	Id único do proj
NAME	OBJECT	378427	375519	Nome do proj
CATEGORY	OBJECT	378390	159	Categoria do proj
MAIN_CATEGORY	OBJECT	378661	15	Cat principal
CURRENCY	OBJECT	378461	14	Moeda do proj
DEADLINE	DATETIME	378661	3164	Tempo final proj
GOAL	FLOAT	378661	378661	Valor objetivo
PLEDGED	FLOAT	378661	378661	Valor levantado

data expert.

Data Science & Machine Learning

dnc

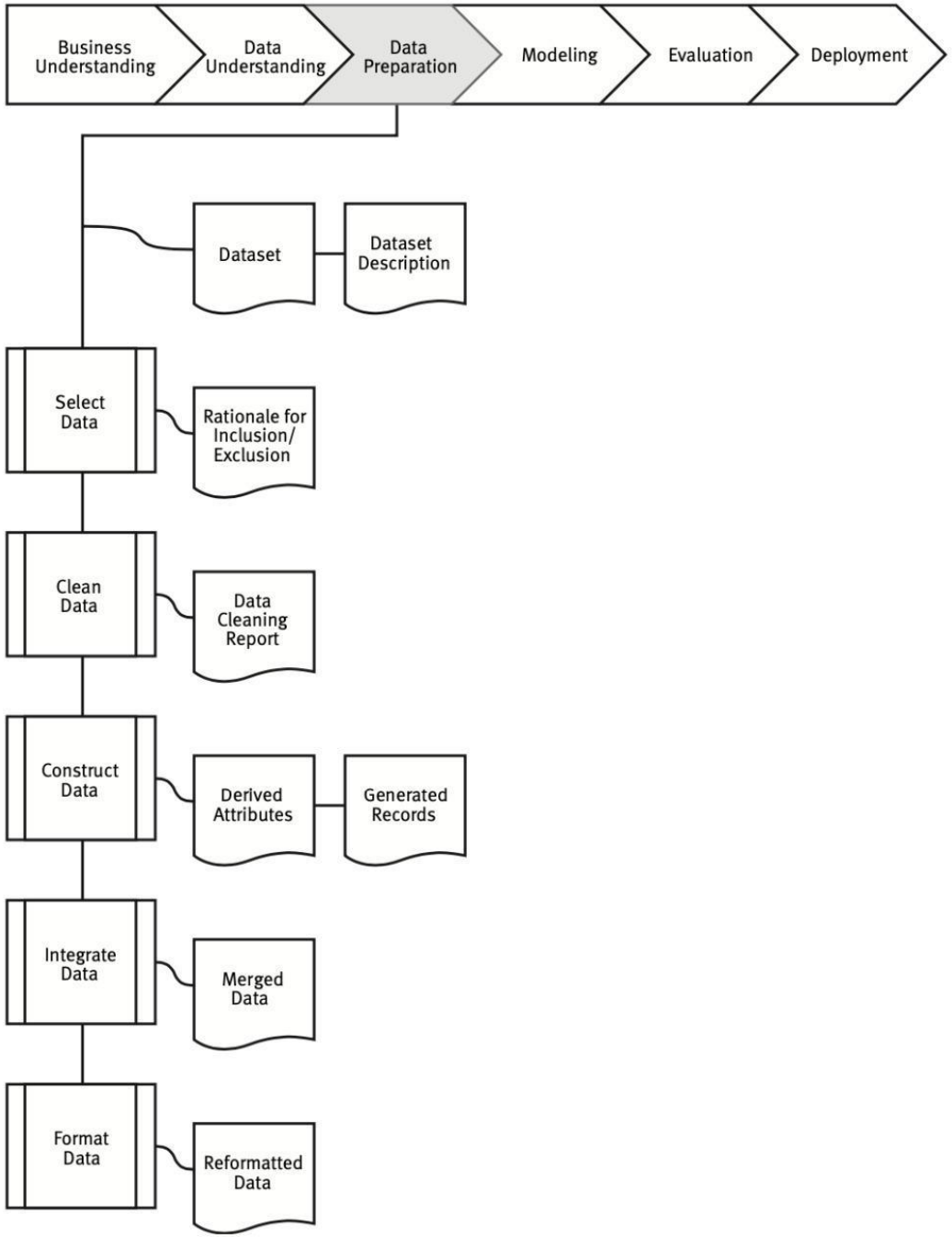
Aula 4: Select Data

Consultor: Daniel Soria

Select Data é a task onde você efetivamente irá selecionar os dados que são relevantes para o teu Dataset

Observe que a seleção de dados cobre a seleção de atributos, features, (colunas), bem como a seleção de registros (linhas) em uma tabela.

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000.0	2015-08-11 12:12:28	0.00	failed	0	GB	0.00
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000.0	2017-09-02 04:43:57	2421.00	failed	15	US	100.00
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000.0	2013-01-12 00:20:50	220.00	failed	3	US	220.00
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000.0	2012-03-17 03:24:11	1.00	failed	1	US	1.00


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Dataset
Dataset Description
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data

Rationale for inclusion/exclusion

Justificativa para inclusão/exclusão

Liste os dados a serem incluídos / excluídos
e as razões para essas decisões.

Exclusão	Motivo
NAME	Não iremos utilizar o nome devido a necessidade de construir uma analise de NLP
CATEGORY	Alta correlação com o a feature Main_Category
GOAL	Vamos utilizar a coluna GOAL_USD devido a ser os mesmos valores da GOAL porém convertidos em dolar

data expert.

Data Science & Machine Learning

dnc

Aula 5: Clean Data

Consultor: Daniel Soria

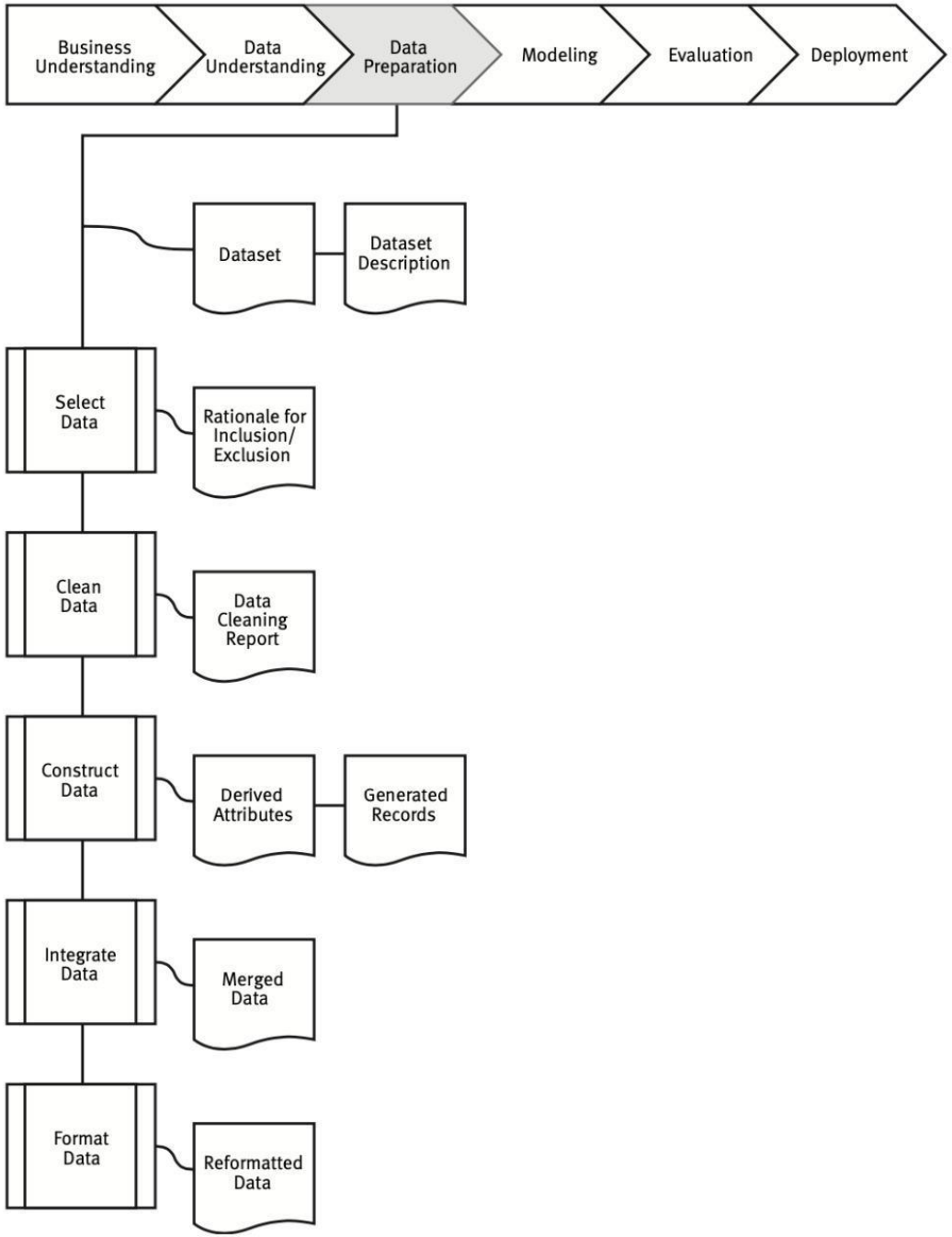
Clean Data é a task onde você vai se aprofundar na qualidade dados.

E iniciar a limpeza dos dados podendo selecionar subconjuntos, selecionar alguns padrões e inclusive fazer estimativas

deadline		goal	launched
2015-10-09	GBP	1000.0	2015-08-11 12:12:28
2017-11-01	USD	30000.0	2017-09-02 04:43:57
2013-02-26	USD	45000.0	2013-01-12 00:20:50
2012-04-16	USD	5000.0	2012-03-17 03:24:11

deadline	goal	launched
2015-10-09	1000.0	2015-08-11 12:12:28
2012-11-01	30000.0	2017-09-02 04:43:57
2013-02-26	45000.0	2013-01-12 00:20:50
2012-04-16	5000.0	2012-03-17 03:24:11

currency	deadline	goal	launched
GBP	2015-10-09	1000.0	2015-08-11 12:12:28
	2017-11-01	30000.0	2017-09-02 04:43:57
USD	2013-02-26	45000.0	2013-01-12 00:20:50
USD	2012-04-16	5000.0	2012-03-17 03:24:11


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Dataset
Dataset Description
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data

Data Cleaning Report

Descrever os métodos e decisões tomadas para resolver os problemas de qualidade encontrados na task de Verify Data Quality na fase de Data Understanding

Feature	Tratamento
GOAL	Excluído o símbolo da moeda e converter os valores para inteiro
DEADLINE	Correção dos valores que são menores que a feature launched utilizando a média de tempo entre launched e deadline.

data expert.

Data Science & Machine Learning

dnc

Aula 6: Construct Data

Consultor: Daniel Soria

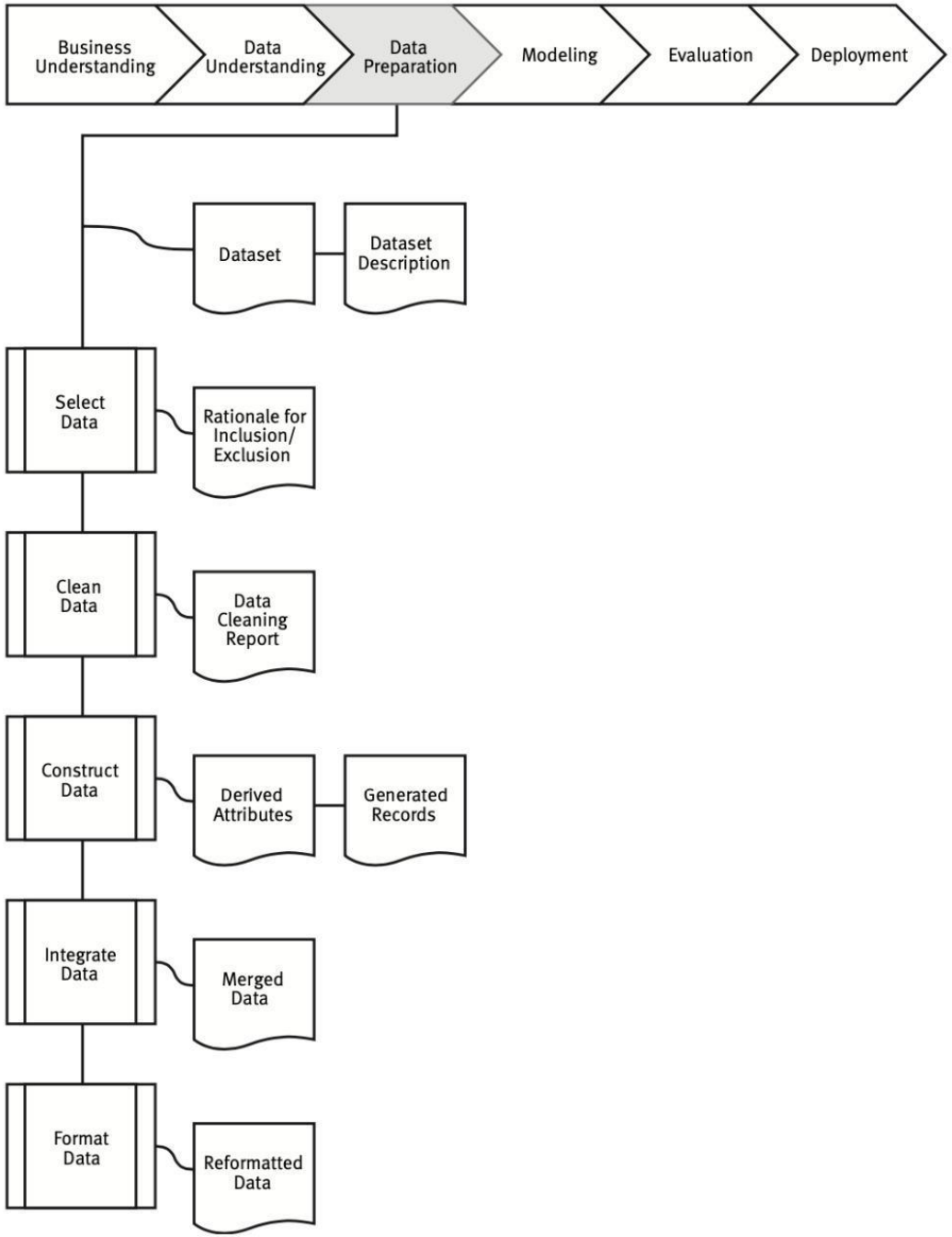
Construct Data é a task onde você vai utilizar de técnicas para construir novas features.

As novas features podem ser geradas da combinação de outras features ou registros inteiros novos ou valores transformados

deadline	launched	time range
2015-10-09	2015-08-11 12:12:28	60
2017-11-01	2017-09-02 04:43:57	59
2013-02-26	2013-01-12 00:20:50	38
2012-04-16	2012-03-17 03:24:11	29

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Dataset
Dataset Description
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data

Derived Atributes

Atributos derivados são novos atributos construídos a partir de um ou mais atributos existentes no mesmo registro. Exemplo: área = comprimento * largura.

Generated Records

Descreva a criação de registros completamente novos. Exemplo: crie registros para clientes que não fizeram nenhuma compra durante o ano anterior. Não havia razão para ter esses registros nos dados brutos, mas para fins de modelagem, pode fazer sentido representar explicitamente o fato de que certos clientes não fizeram compras.

data expert.

Data Science & Machine Learning

dnc

Aula 7: Integrate Data

Consultor: Daniel Soria

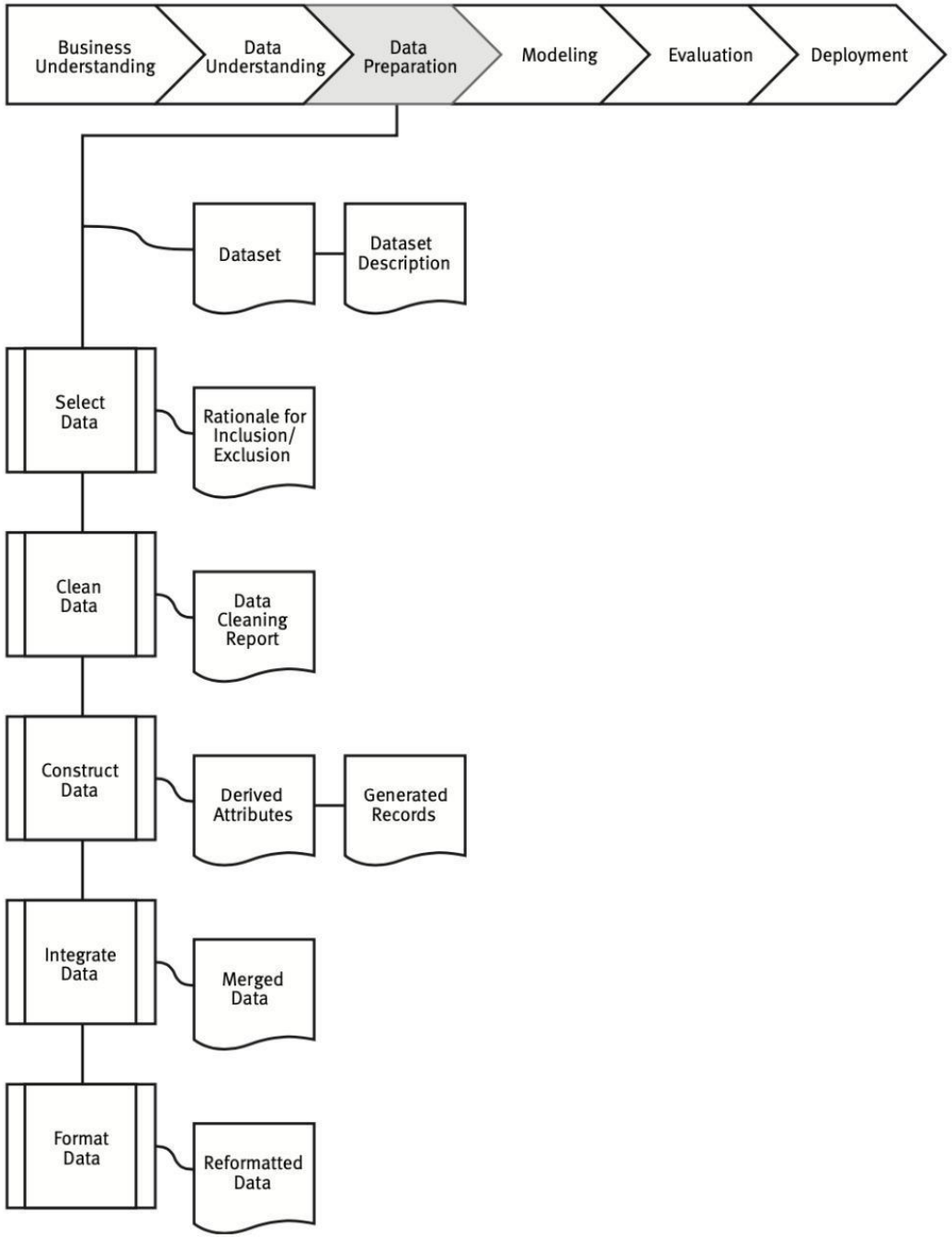
A task de integrate data é sobre integrar dados de diferentes tabelas ou fontes de informação.

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000.0	2015-08-11 12:12:28	0.00	failed	0	GB	0.00
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000.0	2017-09-02 04:43:57	2421.00	failed	15	US	100.00
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000.0	2013-01-12 00:20:50	220.00	failed	3	US	220.00
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000.0	2012-03-17 03:24:11	1.00	failed	1	US	1.00

	ID	Text Description	Vídeo	Image	Infographic	Reviews	FAQ	Risks
0	1000002330	1	1	0	1	1	0	1
1	1000003930	1	0	1	0	0	1	0
2	1000004038	1	1	1	0	1	0	0
3	1000007540	1	1	0	0	0	1	0
4	1000011046	1	0	1	0	1	1	1

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000.0	2015-08-11 12:12:28	0.00	failed	0	GB	0.00
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000.0	2017-09-02 04:43:57	2421.00	failed	15	US	100.00
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000.0	2013-01-12 00:20:50	220.00	failed	3	US	220.00
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000.0	2012-03-17 03:24:11	1.00	failed	1	US	1.00

	ID	Text Description	Vídeo	Image	Infographic	Reviews	FAQ	Risks
0	1000002330	1	1	0	1	1	0	1
1	1000003930	1	0	1	0	0	1	0
2	1000004038	1	1	1	0	1	0	0
3	1000007540	1	1	0	0	0	1	0
4	1000011046	1	0	1	0	1	1	1


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Dataset
Dataset Description
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data

Output são o MERGED DATA

São os próprios dados integrados/mergeados;

Os dados Merged Data também cobrem agregações.

A agregação se refere a operações nas quais novos valores são calculados resumindo informações de vários registros e / ou tabelas.

name	invested	backer location	age
The Songs of Adelaide & Abullah	1000	US	32
The Songs of Adelaide & Abullah	100	GBK	35
Greeting From Earth: ZGAC Arts Capsule For ET	50	BR	20
Where is Hank?	70	US	29
Greeting From Earth: ZGAC Arts Capsule For ET	88	US	30

name	invested	US	GBK	BR	age
The Songs of Adelaide & Abullah	1110	2	1	0	28.333333
Greeting From Earth: ZGAC Arts Capsule For ET	138	1	0	1	25.000000
Where is Hank?	70	1	0	0	29.000000
ToshiCapital Rekordz Needs Help to Complete Album	90	1	0	0	55.000000

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000.0	2015-08-11 12:12:28	0.00	failed	0	GB	0.00
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000.0	2017-09-02 04:43:57	2421.00	failed	15	US	100.00
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000.0	2013-01-12 00:20:50	220.00	failed	3	US	220.00
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000.0	2012-03-17 03:24:11	1.00	failed	1	US	1.00

name	invested	US	GBK	BR	age
The Songs of Adelaide & Abullah	1110	2	1	0	28.333333
Greeting From Earth: ZGAC Arts Capsule For ET	138	1	0	1	25.000000
Where is Hank?	70	1	0	0	29.000000
ToshiCapital Rekordz Needs Help to Complete Album	90	1	0	0	55.000000

data expert.

Data Science & Machine Learning

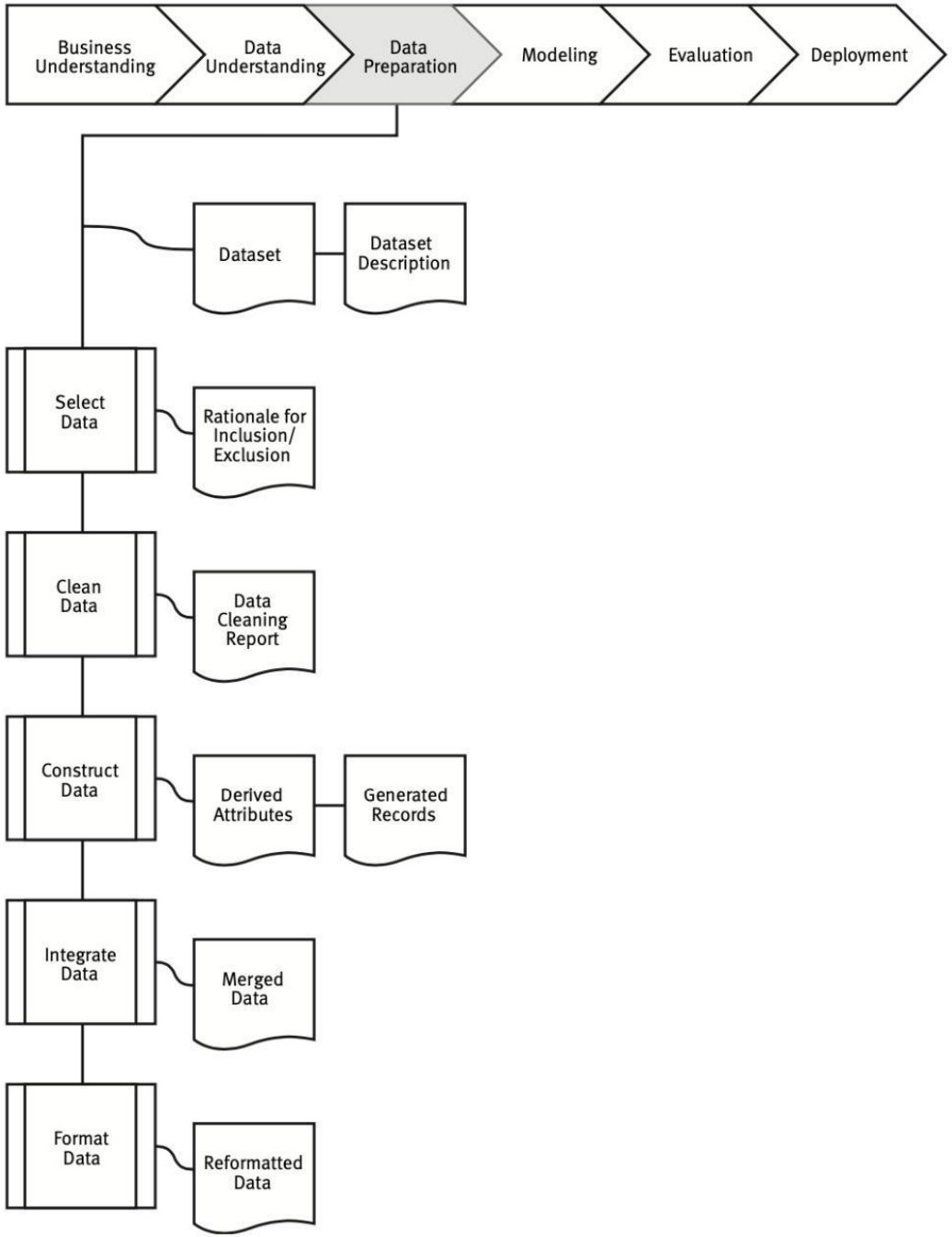
dnc

Aula 8: Format Data

Consultor: Daniel Soria

Format data é a formatação de dados que não mudam seu significado porém facilitam o entendimento e a modelagem.

launched	launched	goal
2015-08-11 12:12:28	2015-08-11	1000.0
2017-09-02 04:43:57	2017-09-02	30000.0
2013-01-12 00:20:50	2013-01-12	45000.0
2012-03-17 03:24:11	2012-03-17	5000.0


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Dataset
Dataset Description
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data

dnc

Dados reformatados

Algumas ferramentas têm requisitos na ordem dos atributos, como o primeiro campo sendo um identificador exclusivo para cada registro ou o último campo sendo o campo de resultado que o modelo deve prever.

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000.0	2015-08-11 12:12:28	0.00	failed	0	GB	0.00
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000.0	2017-09-02 04:43:57	2421.00	failed	15	US	100.00
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000.0	2013-01-12 00:20:50	220.00	failed	3	US	220.00
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000.0	2012-03-17 03:24:11	1.00	failed	1	US	1.00

data expert.

Data Science & Machine Learning

dnc

Aula 9: Resumo Data Preparation

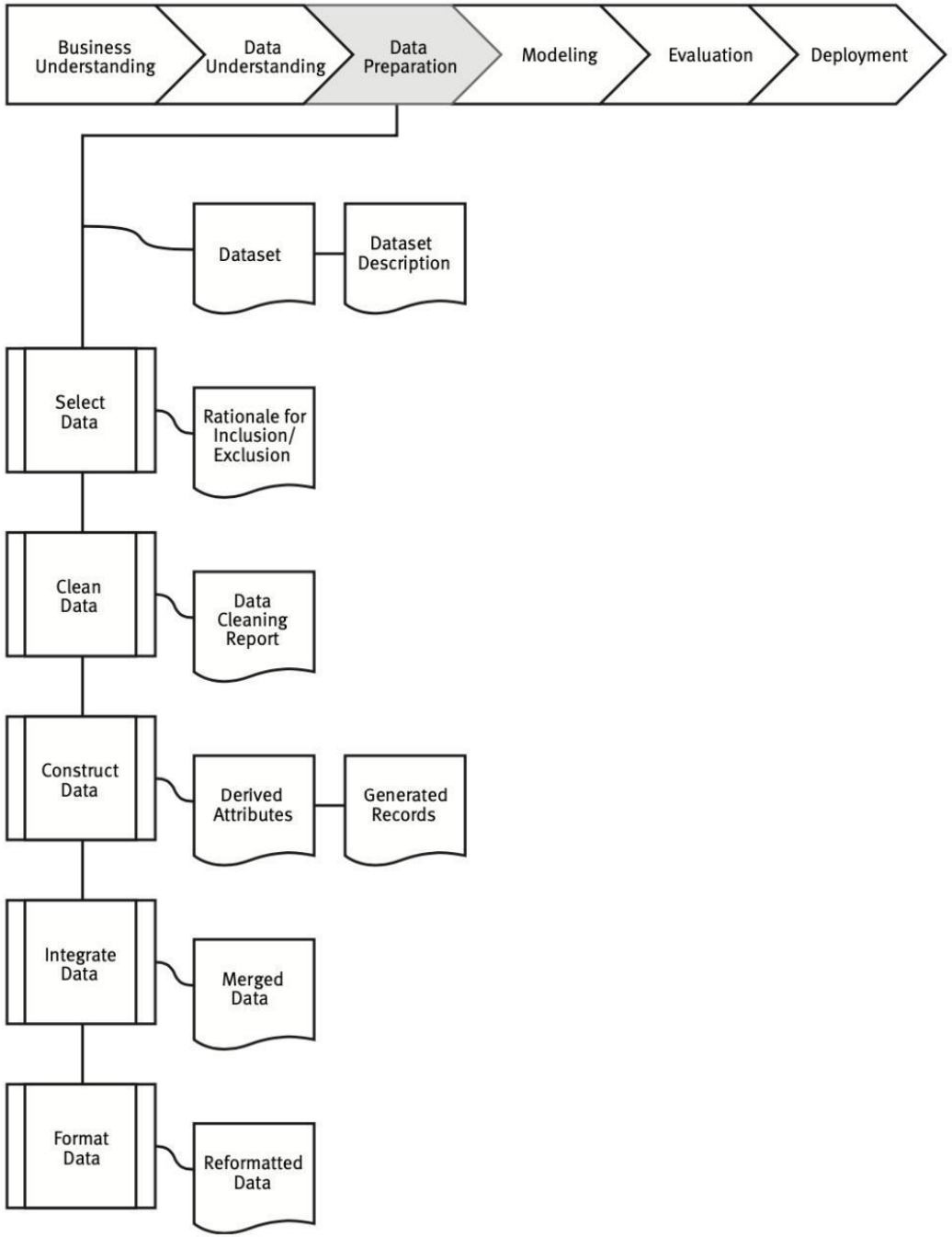
Consultor: Daniel Soria

data expert.
Data Science & Machine Learning

dnc

Aula 10: Data Data Preparation na Prática

Consultor: Daniel Soria


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Dataset
Dataset Description
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data

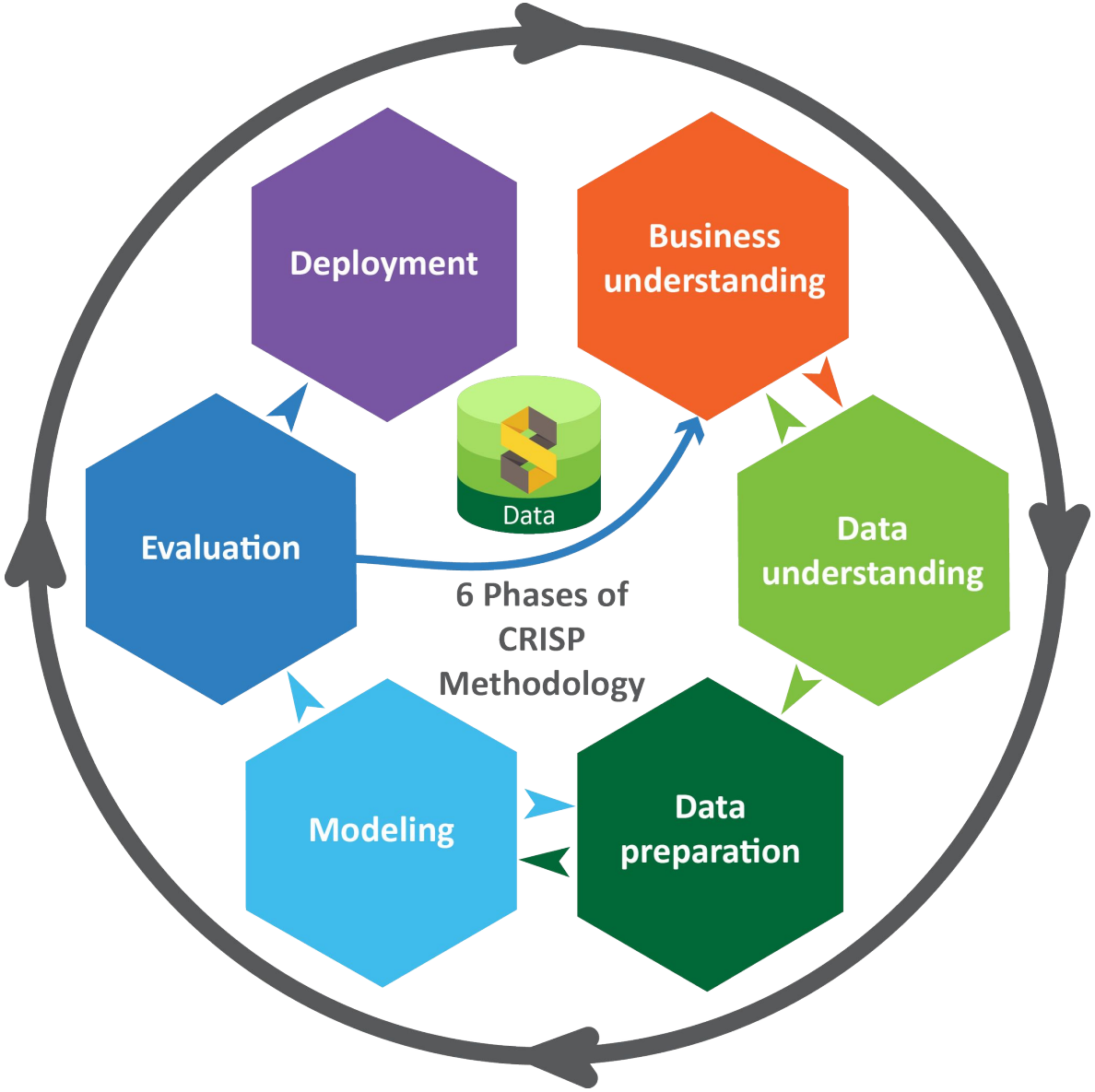
data expert.

Data Science & Machine Learning

dnc

Aula 11: Definindo Modeling

Consultor: Daniel Soria


Deployment
Business understanding
Evaluation
Data
Data understanding
6 Phases of CRISP Methodology
Modeling
Data preparation

Data Preparation é a fase mais longa do projeto.

60% - 80% do teu projeto está nessa fase.

Um dos grandes responsável pelo sucesso do projeto

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

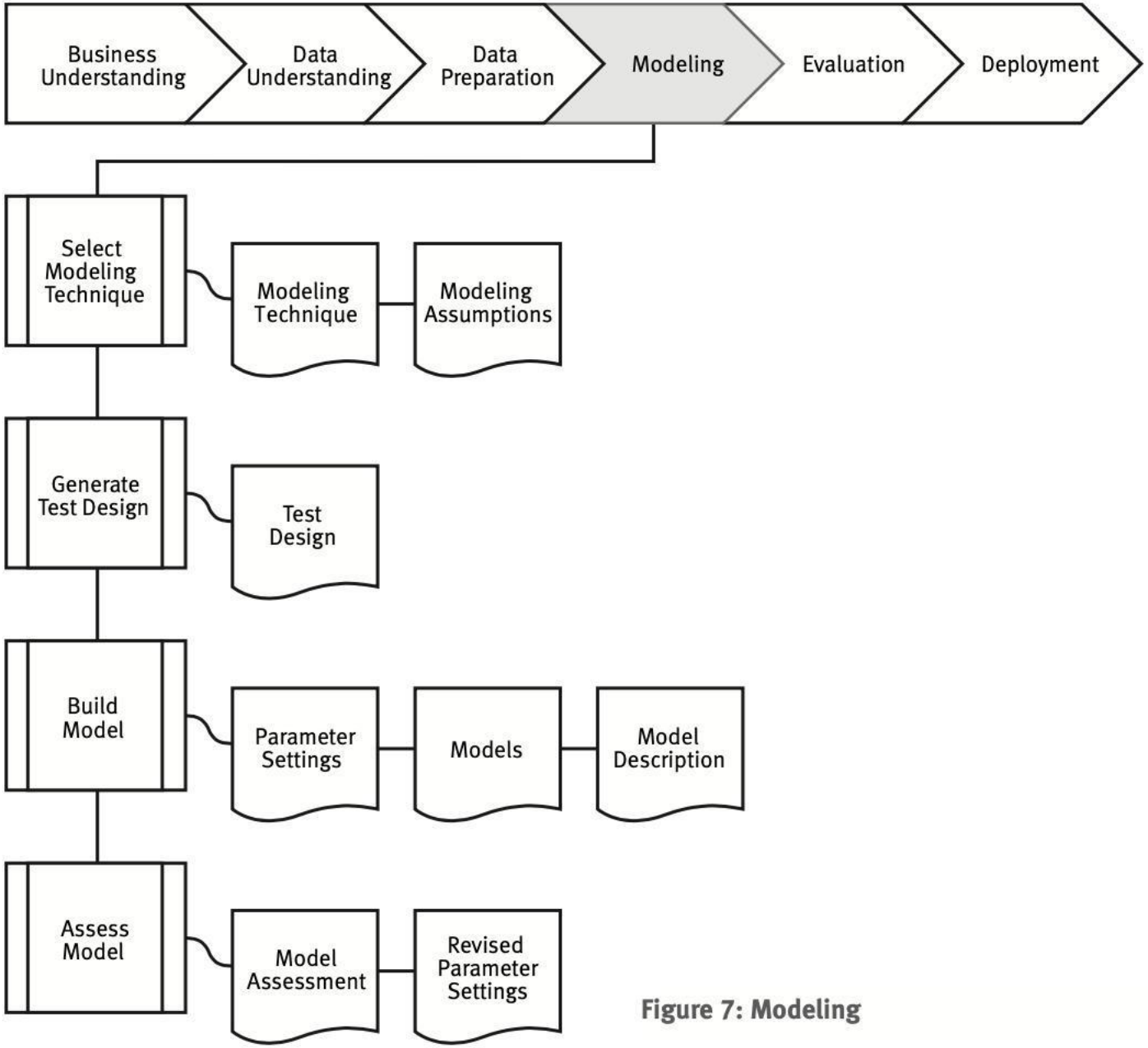


Figure 7: Modeling

O que vimos nesse módulo

1. Definindo Modeling
2. Selecting Modeling
3. Generate Test Design
4. Build Model
5. Assess Model