AULA 2: TIPOS DE DADOS

INTRODUÇÃO A CIÊNCIA DE DADOS NA ENGENHARIA DE PETRÓLEO

Maria Clara Duque (clara_duque@poli.ufrj.br)

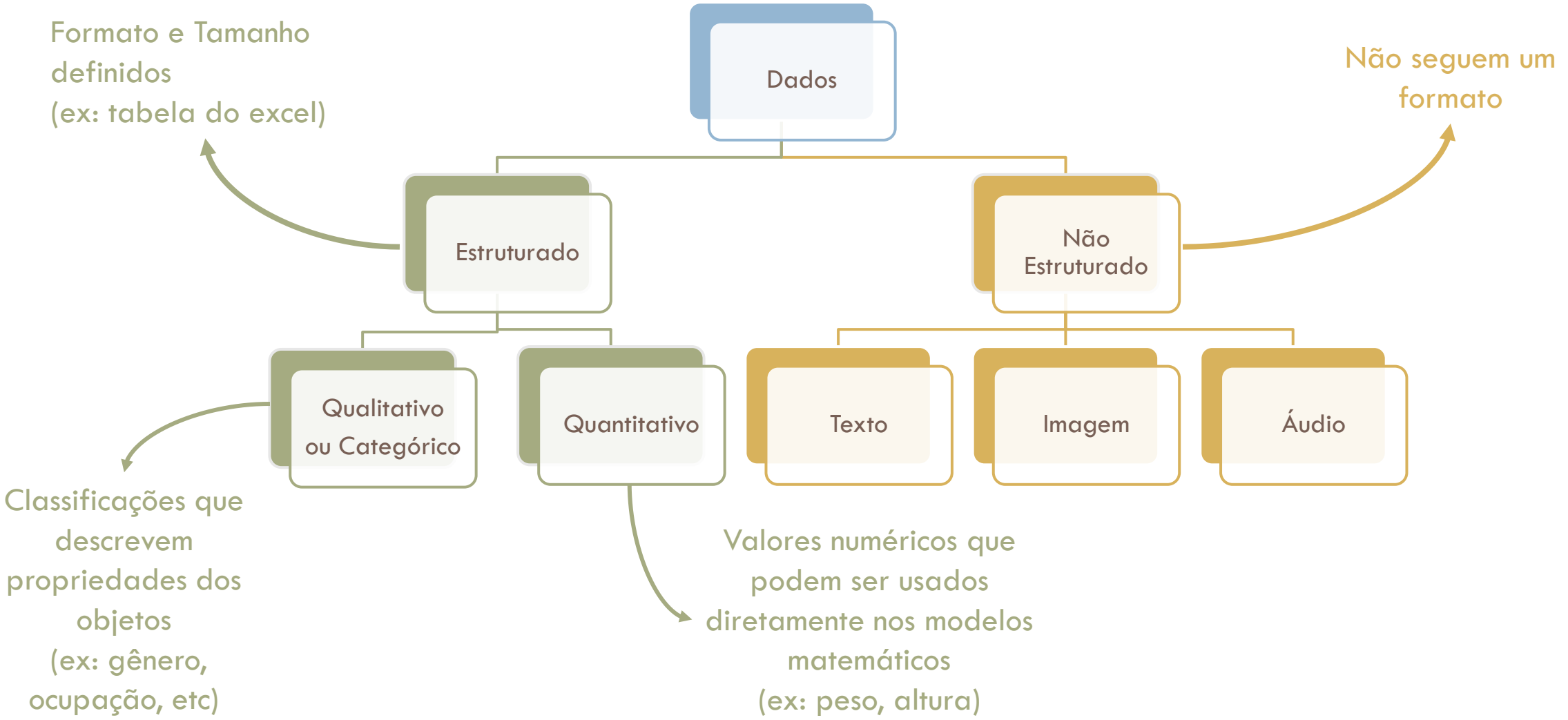
Calendário

DATA	ATIVIDADE
26/08	Introdução/Tipos de dados
02/09	Pré-processamento/Estatística
09/09	Aula Prática 1
16/09	Aula Prática 2
23/09	Introdução ML
30/09	ML Regressão
07/10	Aula Prática 3
14/10	ML Classificação
21/10	ML Agrupamento
28/10	Feriado
04/11	Aula Prática 4
11/11	Entrega dos Trabalhos

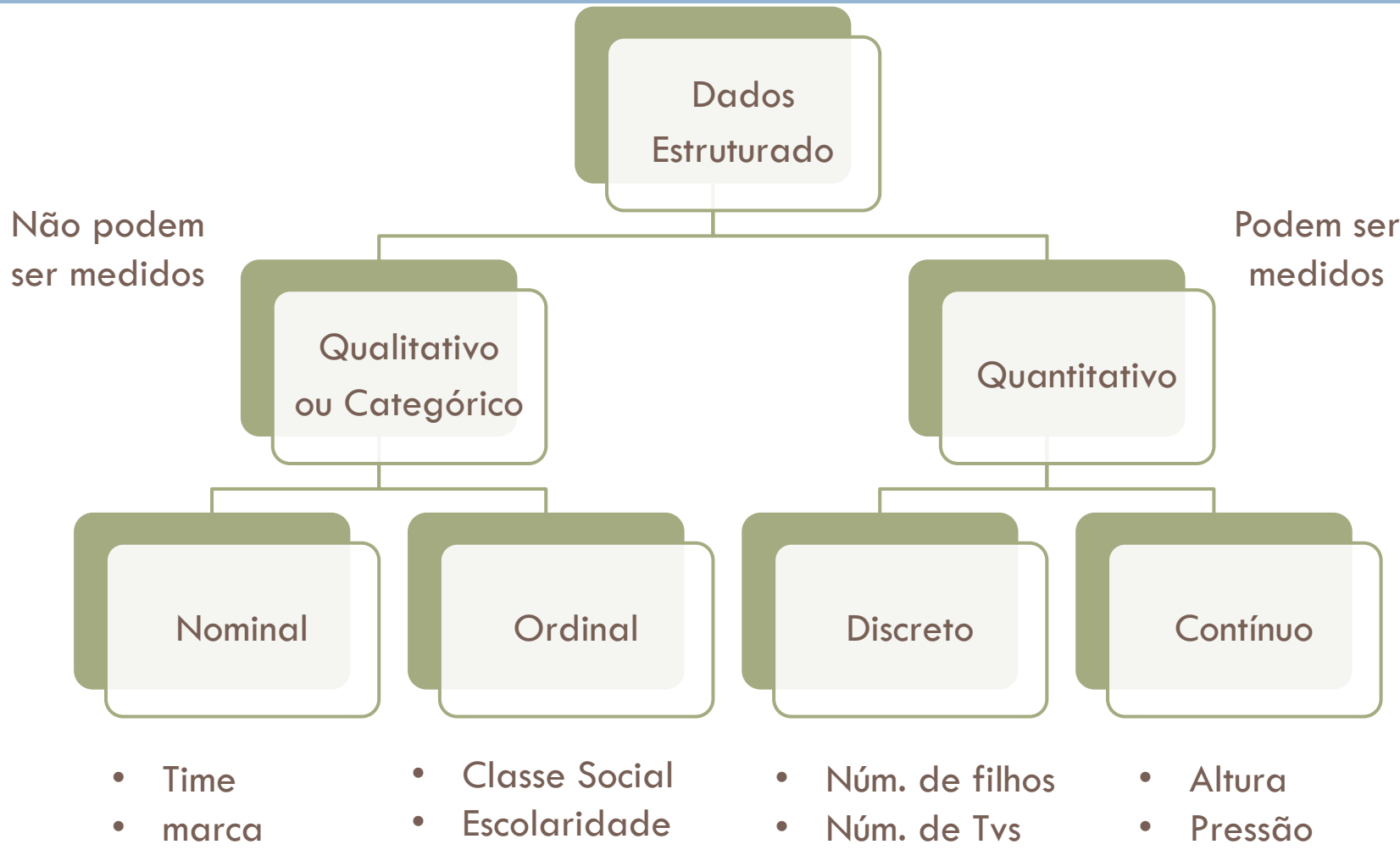
Introdução

- Taxonomia da propriedade dos Dados:
  - Estruturados x Não-estruturados.
  - Quantitativos x Categóricos.
  - Little Data x Big Data.

Tipos dos Dados



1 - Dados Estruturados



1 - Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

Representados por matrizes

Colunas: Variáveis

Linhas: registros

Mês	Estado	Bacia	Campo	Instalação	Produção Óleo (m^3)	Produção de Água (m^3)
Jan	ES	Campos	JUBARTE	FPSO CA	18081,92	16518,41
Jan	ES	Campos	JUBARTE	FPSO CAPIXABA	77391,12	14382,92
Jan	RJ	Campos	MARIMBÁ	P-08	2132,27	12322,65
Jan	RJ	Campos	MARIMBÁ	P-08	3890,63	4031,35
Jan	ES	Campos	JUBARTE	FPSO CA	-	-
Fev	ES	Campos	JUBARTE	FPSO CAPIXABA	-	-
Fev	RJ	Campos	MARIMBÁ	P-08	275,81	724,21

Linha: registro

Coluna: variável ou atributo

Fonte: http://www.anp.gov.br/conteudo-do-menu-superior/31-dados-abertos/5538-producao-de-petroleo-e-gas-natural-nacional

1 - Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

Mês	Estado	Bacia	Campo	Instalação	Produção Óleo (m^3)	Produção de Água (m^3)
Jan	ES	Campos	JUBARTE	FPSO CA	18081,92	16518,41
Jan	ES	Campos	JUBARTE	FPSO CAPIXABA	77391,12	14382,92
Jan	RJ	Campos	MARIMBÁ	P-08	2132,27	12322,65
Jan	RJ	Campos	MARIMBÁ	P-08	3890,63	4031,35
Jan	ES	Campos	JUBARTE	FPSO CA	-	-
Fev	ES	Campos	JUBARTE	FPSO CAPIXABA	-	-
Fev	RJ	Campos	MARIMBÁ	P-08	275,81	724,21

Fonte: http://www.anp.gov.br/conteudo-do-menu-superior/31-dados-abertos/5538-producao-de-petroleo-e-gas-natural-nacional

1 – Quais os tipos de variáveis do *dataset*?

- **Mês**: categórica ordinal
- **Estado**: categórica nominal (binária)
- **Bacia**: categórica nominal
- **Campo**: categórica nominal (binária)
- **Instalação**: categórica nominal
- **Produção de óleo e água**: quantitativa contínua

1 - Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

Mês	Estado	Bacia	Campo	Instalação	Produção Óleo (m^3)	Produção de Água (m^3)
Jan	ES	Campos	JUBARTE	FPSO CA	18081,92	16518,41
Jan	ES	Campos	JUBARTE	FPSO CAPIXABA	77391,12	14382,92
Jan	RJ	Campos	MARIMBÁ	P-08	2132,27	12322,65
Jan	RJ	Campos	MARIMBÁ	P-08	3890,63	4031,35
Jan	ES	Campos	JUBARTE	FPSO CA	-	-
Fev	ES	Campos	JUBARTE	FPSO CAPIXABA	-	-
Fev	RJ	Campos	MARIMBÁ	P-08	275,81	724,21

2 – As variáveis categóricas são: mês, Estado, Bacia, Campo e Instalação.
Precisamos **transformar as categorias em números** para que possamos manipula-las.
Como podemos fazer isso?

Fonte: http://www.anp.gov.br/conteudo-do-menu-superior/31-dados-abertos/5538-producao-de-petroleo-e-gas-natural-nacional

1 - Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

Categórica ordinal

Mês	Mês	Estado	Bacia	Campo	Instalação	Produção Óleo (m^3)	Produção de Água (m^3)
Jan	1	ES	Campos	JUBARTE	FPSO CA	18081,92	16518,41
Jan	1	ES	Campos	JUBARTE	FPSO CAPIXABA	77391,12	14382,92
Jan	1	RJ	Campos	MARIMBÁ	P-08	2132,27	12322,65
Jan	1	RJ	Campos	MARIMBÁ	P-08	3890,63	4031,35
Jan	1	ES	Campos	JUBARTE	FPSO CA	-	-
Fev	2	ES	Campos	JUBARTE	FPSO CAPIXABA	-	-
Fev	2	RJ	Campos	MARIMBÁ	P-08	275,81	724,21

Fonte: http://www.anp.gov.br/conteudo-do-menu-superior/31-dados-abertos/5538-producao-de-petroleo-e-gas-natural-nacional

1 - Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

Mês	Mês	Estado	ES	RJ	Bacia	Campo	Instalação	Produção Óleo (m^3)	Produção de Água (m^3)
Jan	1	ES	1	0	Campos	JUBARTE	FPSO CA	18081,92	16518,41
Jan	1	ES	1	0	Campos	JUBARTE	FPSO CAPIXABA	77391,12	14382,92
Jan	1	RJ	0	1	Campos	MARIMBÁ	P-08	2132,27	12322,65
Jan	1	RJ	0	1	Campos	MARIMBÁ	P-08	3890,63	4031,35
Jan	1	ES	1	0	Campos	JUBARTE	FPSO CA	-	-
Fev	2	ES	1	0	Campos	JUBARTE	FPSO CAPIXABA	-	-
Fev	2	RJ	0	1	Campos	MARIMBÁ	P-08	275,81	724,21

Categórica nominal binária

Fonte: http://www.anp.gov.br/conteudo-do-menu-superior/31-dados-abertos/5538-producao-de-petroleo-e-gas-natural-nacional

1 - Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

Faz-se o mesmo para as variáveis Bacia e Campo.

Mês	Mês	Estado	ES	RJ	Bacia	Campos	Campo	Jub	Mari	Instalação	Produção Óleo (m^3)	Produção de Água (m^3)
Jan	1	ES	1	0	Campos	1	JUBARTE	1	0	FPSO CA	18081,92	16518,41
Jan	1	ES	1	0	Campos	1	JUBARTE	1	0	FPSO CAPIXABA	77391,12	14382,92
Jan	1	RJ	0	1	Campos	1	MARIMBÁ	0	1	P-08	2132,27	12322,65
Jan	1	RJ	0	1	Campos	1	MARIMBÁ	0	1	P-08	3890,63	4031,35
Jan	1	ES	1	0	Campos	1	JUBARTE	1	0	FPSO CA	-	-
Fev	2	ES	1	0	Campos	1	JUBARTE	1	0	FPSO CAPIXABA	-	-
Fev	2	RJ	0	1	Campos	1	MARIMBÁ	0	1	P-08	275,81	724,21

Fonte: http://www.anp.gov.br/conteudo-do-menu-superior/31-dados-abertos/5538-producao-de-petroleo-e-gas-natural-nacional

1 - Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

Podemos fazer isso em instalação?
Não!!

Mês	Mês	Estado	ES	RJ	Bacia	Campos	Campo	Jub	Mari	Instalação	Inst	Produção Óleo (m^3)	Produção de Água (m^3)
Jan	1	ES	1	0	Campos	1	JUBARTE	1	0	FPSO CA	1	18081,92	16518,41
Jan	1	ES	1	0	Campos	1	JUBARTE	1	0	FPSO CAPIXABA	2	77391,12	14382,92
Jan	1	RJ	0	1	Campos	1	MARIMBÁ	0	1	P-08	3	2132,27	12322,65
Jan	1	RJ	0	1	Campos	1	MARIMBÁ	0	1	P-08	3	3890,63	4031,35
Jan	1	ES	1	0	Campos	1	JUBARTE	1	0	FPSO CA	1	-	-
Fev	2	ES	1	0	Campos	1	JUBARTE	1	0	FPSO CAPIXABA	2	-	-
Fev	2	RJ	0	1	Campos	1	MARIMBÁ	0	1	P-08	3	275,81	724,21

Números inteiros dão ideia de ordem.
Ou seja: P-08 > FPSO Capixaba > FPSO CA.
Na prática, isto não faz sentido. Só podemos fazer isso com variáveis categóricas ordinárias (como mês).

1- Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

O que se faz?

Mês	Mês	Estado	ES	RJ	Bacia	Campos	Campo	Jub	Mari	Instalação	F CA	FC CPX	P08	Produção Óleo (m^3)	Produção de Água (m^3)
Jan	1	**ES**	1	0	Campos	1	JUBARTE	1	0	FPSO CA	**1**	**0**	**0**	18081,92	16518,41
Jan	1	**ES**	1	0	Campos	1	JUBARTE	1	0	FPSO CAPIXABA	**0**	**1**	**0**	77391,12	14382,92
Jan	1	**RJ**	0	1	Campos	1	MARIMBÁ	0	1	P-08	**0**	**0**	**1**	2132,27	12322,65
Jan	1	**RJ**	0	1	Campos	1	MARIMBÁ	0	1	P-08	**0**	**0**	**1**	3890,63	4031,35
Jan	1	**ES**	1	0	Campos	1	JUBARTE	1	0	FPSO CA	**1**	**0**	**0**	-	-
Fev	2	**ES**	1	0	Campos	1	JUBARTE	1	0	FPSO CAPIXABA	**0**	**1**	**0**	-	-
Fev	2	**RJ**	0	1	Campos	1	MARIMBÁ	0	1	P-08	**0**	**0**	**1**	275,81	724,21

1 - Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

	Estado			Bacia		Instalação			*Dataset* final	
Mês	ES	RJ	Campos	Jub	Mari	F CA	FC CPX	P08	Produção Óleo (m^3)	Produção de Água (m^3)
1	1	0	1	1	0	1	0	0	18081,92	16518,41
1	1	0	1	1	0	0	1	0	77391,12	14382,92
1	0	1	1	0	1	0	0	1	2132,27	12322,65
1	0	1	1	0	1	0	0	1	3890,63	4031,35
1	1	0	1	1	0	1	0	0	-	-
2	1	0	1	1	0	0	1	0	-	-
2	0	1	1	0	1	0	0	1	275,81	724,21

3 – Existe alguma variável que você removeria da sua análise? Por que?

- **"Campo".** Somente tem um valor. **Não** irá acrescentar informação ao problema. Não tem variabilidade.

1 - Dados Estruturados - Exemplo

- Produção de petróleo no mar no ano de 2020:

Dataset final

	Estado		Bacia		Instalação				
Mês	ES	RJ	Jub	Mari	F CA	FC CPX	P08	Produção Óleo (m^3)	Produção de Água (m^3)
1	1	0	1	0	1	0	0	18081,92	16518,41
1	1	0	1	0	0	1	0	77391,12	14382,92
1	0	1	0	1	0	0	1	2132,27	12322,65
1	0	1	0	1	0	0	1	3890,63	4031,35
1	1	0	1	0	1	0	0		
2	1	0	1	0	0	1	0		
2	0	1	0	1	0	0	1	275,81	724,21

4 – Você consegue identificar algum problema que pode ocorrer ao tentar rodar o *dataset* em algum programa, por exemplo, para rodar um modelo de regressão?

- Note que a produção de óleo e água tem 2 registros com um traço (-). O programa entende como caracteres. O que pode gerar erros de leitura. Devemos retira-los para que as colunas sejam lidas como números.

2 - Dados Não Estruturados - Textos

Informações de atividades coletadas de relatórios de perfuração:

- Update 1: PU, MU & RIH 9 5/8" Csg cutter Assy from surface to 1922m, try to cut casing without successful (No indication of casing cut), trouble with casing cutter, POOH to surface & Re-tested Csg cutter @surface, Observed the bypass of drilling mud from the tool above the cutter which did not exerted enough force on the cylinder of blades to push to cutter knives outwards. LD tools. PU and MU 8 1/2" Taper Mill with 8 1/4" Section Mill Assy and RIH to 1909m. - Performed cut 9 5/8" Csg @1907.6m. FC- Ok. POOH 8 1/4" Section Mill Assy to 1650m.
- Update 2: Continued RIH with 8 1/2"Milling BHA and Ream down 3445-3522m depth, start milling Junk 3522-3523.64m, sweep hole with Hi-vis pill, observed metal swarfs on Ditch magnet Weighted 90 grams, attempt to mill below 3523.64m depth - no success. Circulate hole clean and POOH with 8 1/2" Milling BHA to surface, found several metal Components in the Junk sub, and signs of milling on the Bladed Junk mill. P/U, M/U and RIH with 8 1/2" Directional BHA to 1347m depth.

Kowalchuk, P. (2019)

- Contêm informações importantes.
- Como podemos trabalhar com textos?

- Quais os tipos de eventos dessas atualizações?
- Quão críticas e complexas?

2 - Dados Não Estruturados - Textos

- *Bag of Words*: modelo simplificado para converter sentenças de texto em vetores numéricos.

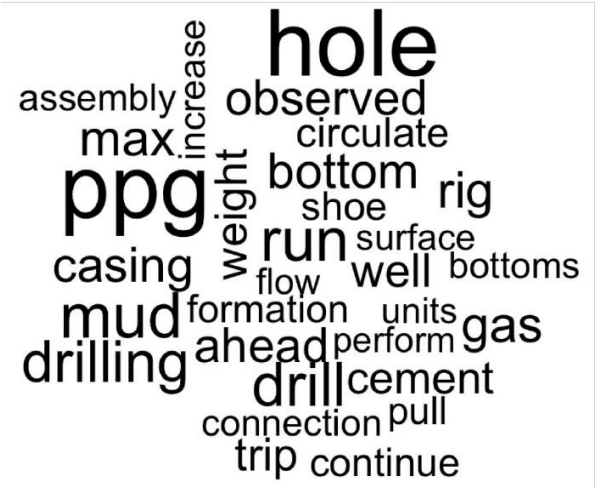
Palavras: variáveis

	hole	ppg	formation	shoe	pipe	perform	pull	...
Document 1	1			1	1			
Document 2		1		2			1	
Document 3	3	1		1				
Document 4		2		2		1	1	
...								

Figure 2—Document term matrix example (partial).

Linhas: documentos

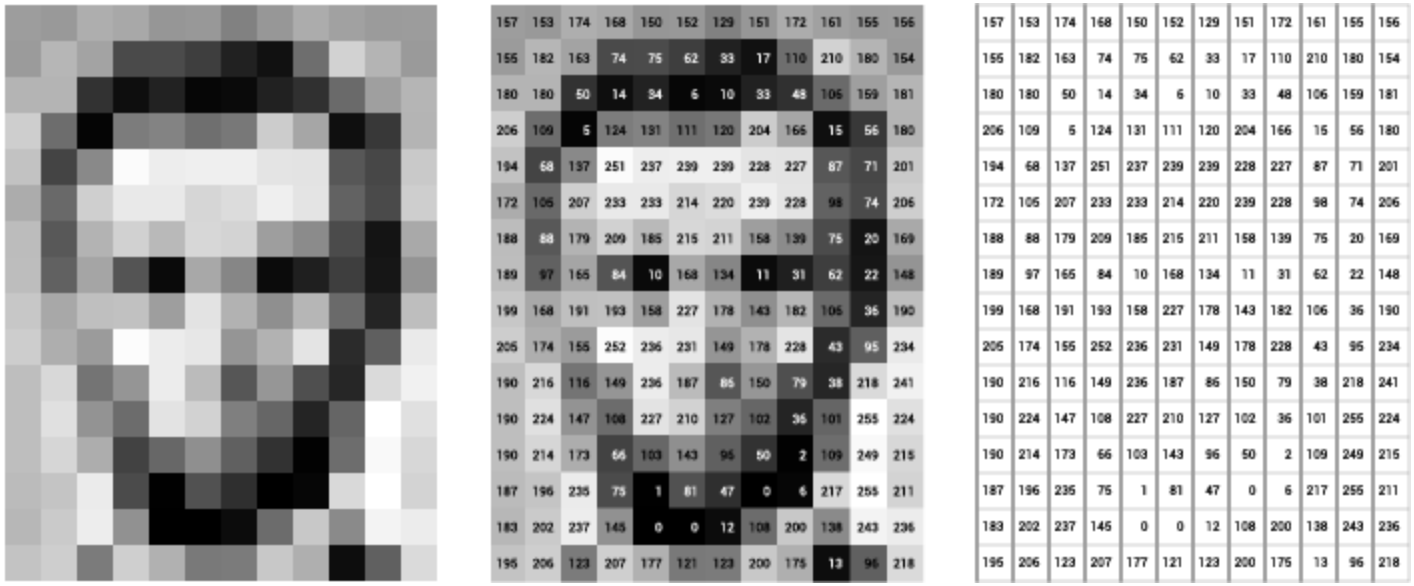
Matriz: Quantas vezes uma palavra aparece em um documento.



Kowalchuk, P. (2019)

2 - Dados Não Estruturados - Imagens

- Imagem: Matriz de pixels com várias dimensões.
- Computadores trabalham com números.
  - Cada pixel é um valor correspondente a intensidade de cor.



https://openframeworks.cc/ofBook/chapters/image_processing_computer_vision.html

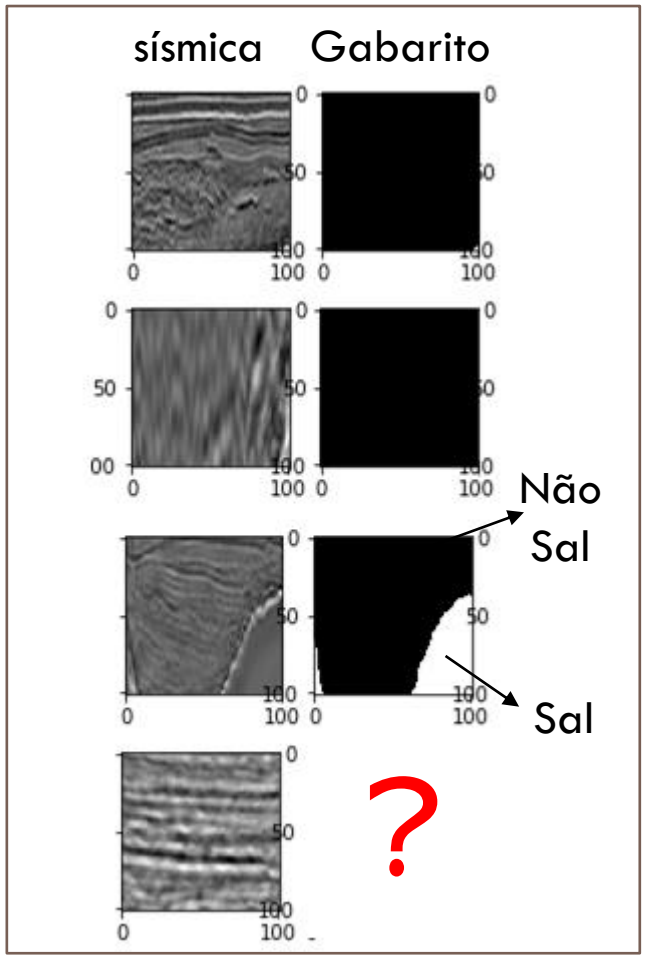
2 - Dados Não Estruturados - Imagens

- Exemplo: Competição *TGS Salt Identification Challenge - Kaggle*
  - https://www.kaggle.com/c/tgs-salt-identification-challenge

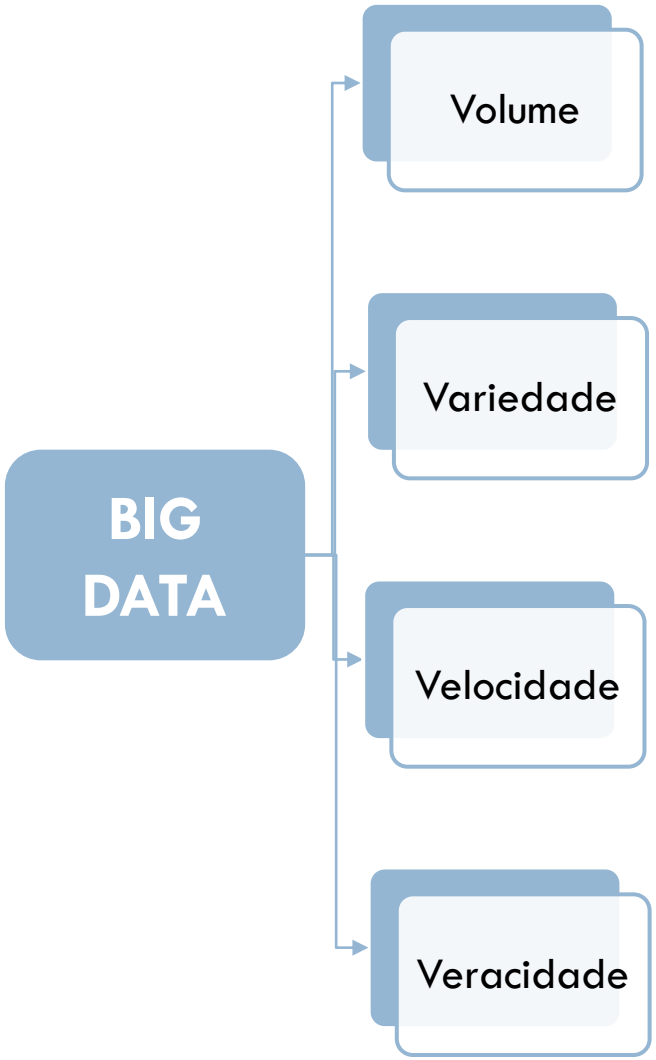
Descrição:
Várias áreas do mundo contém grandes quantidades de sal no subsolo. Um dos desafios da imagem sísmica é identificar a parte da subsuperfície que é o sal.

Dados:
Conjunto de imagens escolhidas em diferentes locais de subsuperfície. As imagens têm 101 x 101 pixels e cada pixel (região) pode ser classificado como sal ou não. Além das imagens sísmicas, a profundidade de cada uma é fornecida.

Objetivo: *Segmentar regiões que contenham sal.*



3 - Big Data



- Grande quantidade de dados gerada (tera, zetabytes,...).
- Onde armazenar e como processar esses dados?

- Diferentes tipos e fontes de dados.
- Como integrar esses diferentes tipos de dados? Quais usar?

- Dados são gerados em tempo real

- Qualidade e origem dos dados
- Podemos confiar nas medidas que temos acesso?

- Sensores, relatórios, textos, imagens, etc.
- Maior processamento computacional com aumento do volume.
- Complexos de visualizar.

Referências Bibliográficas

- Steven S. Skiena. 2017. **The Data Science Design Manual** (1st. ed.). Springer Publishing Company, Incorporated.
- Hodaway, K R. **Harness Oil and Gas Big Data with Analytics**. 2014.
- Kowalchuk, P. (2019). Implementing a Drilling Reporting Data Mining Tool Using Natural Language Processing Sentiment Analysis Techniques. Society of Petroleum Engineers. doi:10.2118/194961-MS
- Sites:
  - http://www.anp.gov.br/
  - https://www.kaggle.com/c/tgs-salt-identification-challenge
  - https://openframeworks.cc/ofBook/chapters/image_processing_computer_vision.html