Projeto Laboratório de Redes de Conhecimento

Instituto Federal do Sudeste de Minas Gerais, Campus Barbacena

Mineração de Dados Aplicada

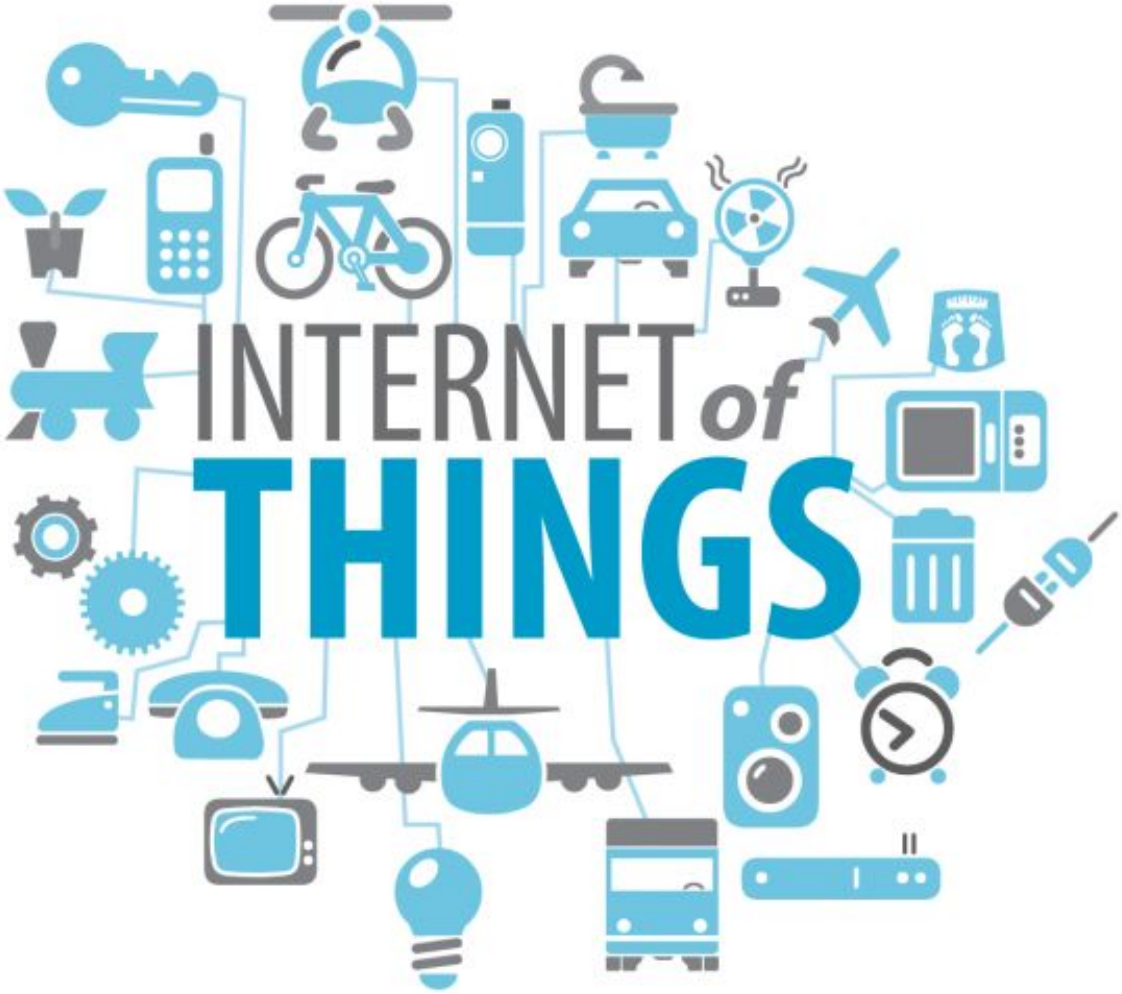
Prof. Rafael José de Alencar Almeida
rafael.alencar@ifsudestemg.edu.br

Aula 1: Introdução à Mineração de Dados

A Internet em 60 segundos (2018)

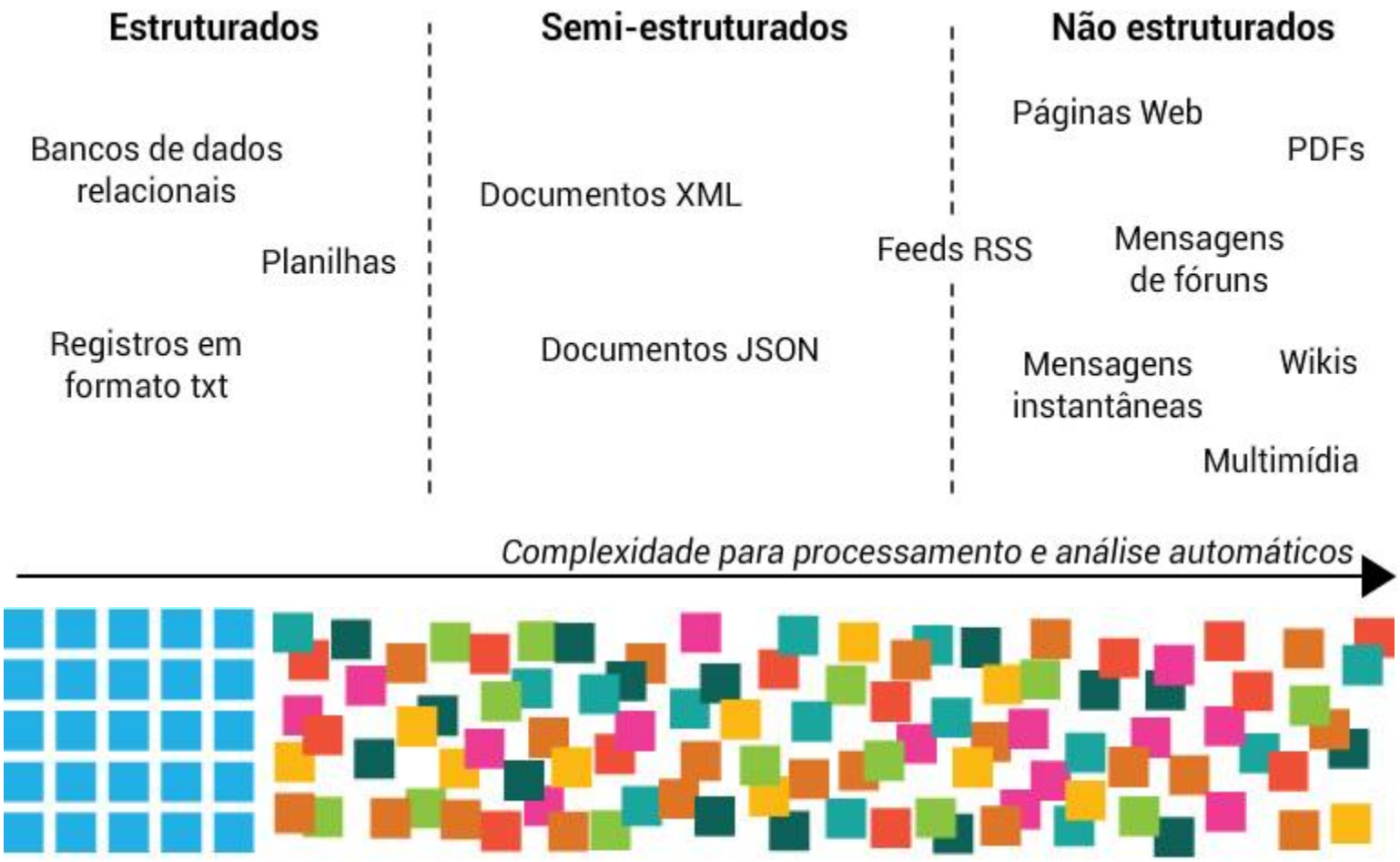
Fonte: http://www.visualcapitalist.com/internet-minute-2018/

+ Internet das Coisas / Internet de Tudo



Fonte: https://mspalliance.com/iot-really-security-things/

Contexto: dados não estruturados



No geral, 80% das informações criadas e utilizadas por uma empresa são dados não estruturados, o que torna a manipulação e interpretação mais complexa [1].

Mineração de Dados

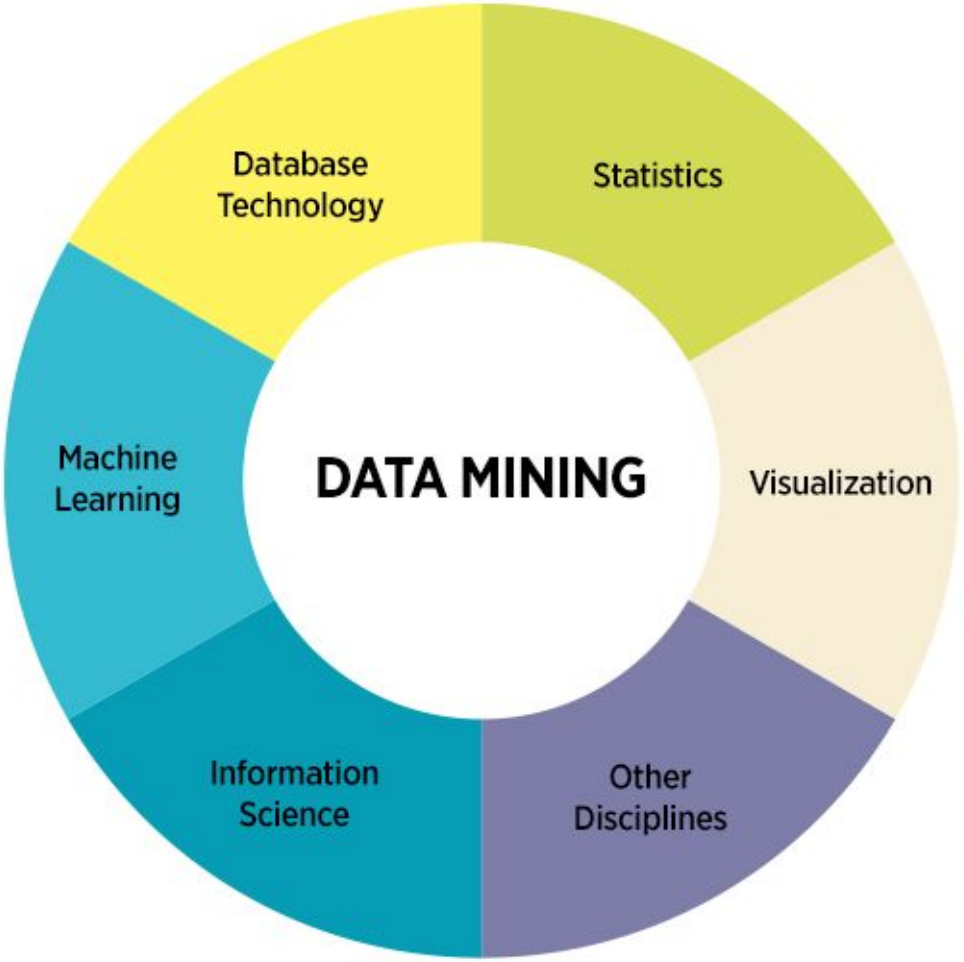
- Processo de descoberta automática de informações úteis em grandes depósitos de dados [2], muitas vezes não estruturados.

- Envolve a busca por **padrões** e **relacionamentos entre os dados**, com aplicabilidade nas mais diversas áreas:
  - Análise e previsão de padrões de compra de consumidores
  - Descoberta dos assuntos principais em discussões *online*
  - Análise das interações entre genes em doenças
  - Detecção de anomalias e fraudes

Nem todas tarefas de recuperação de informação podem ser consideradas mineração de dados, como por exemplo a consulta por registros usando um SGBD.

Disciplinas

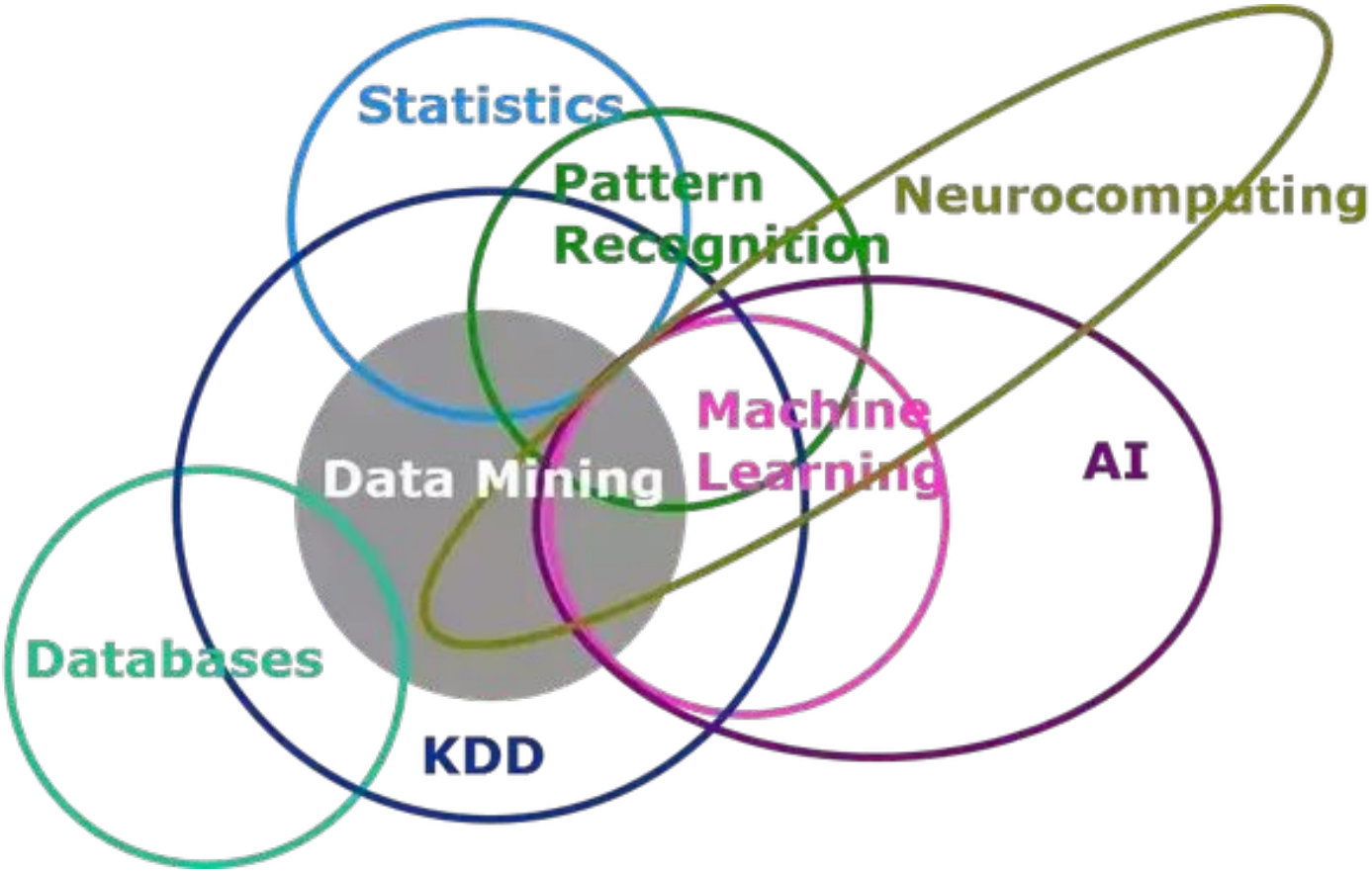
Mineração de dados é uma **confluência de diversas disciplinas**:



Fonte: https://www.simplilearn.com/data-mining-vs-statistics-article

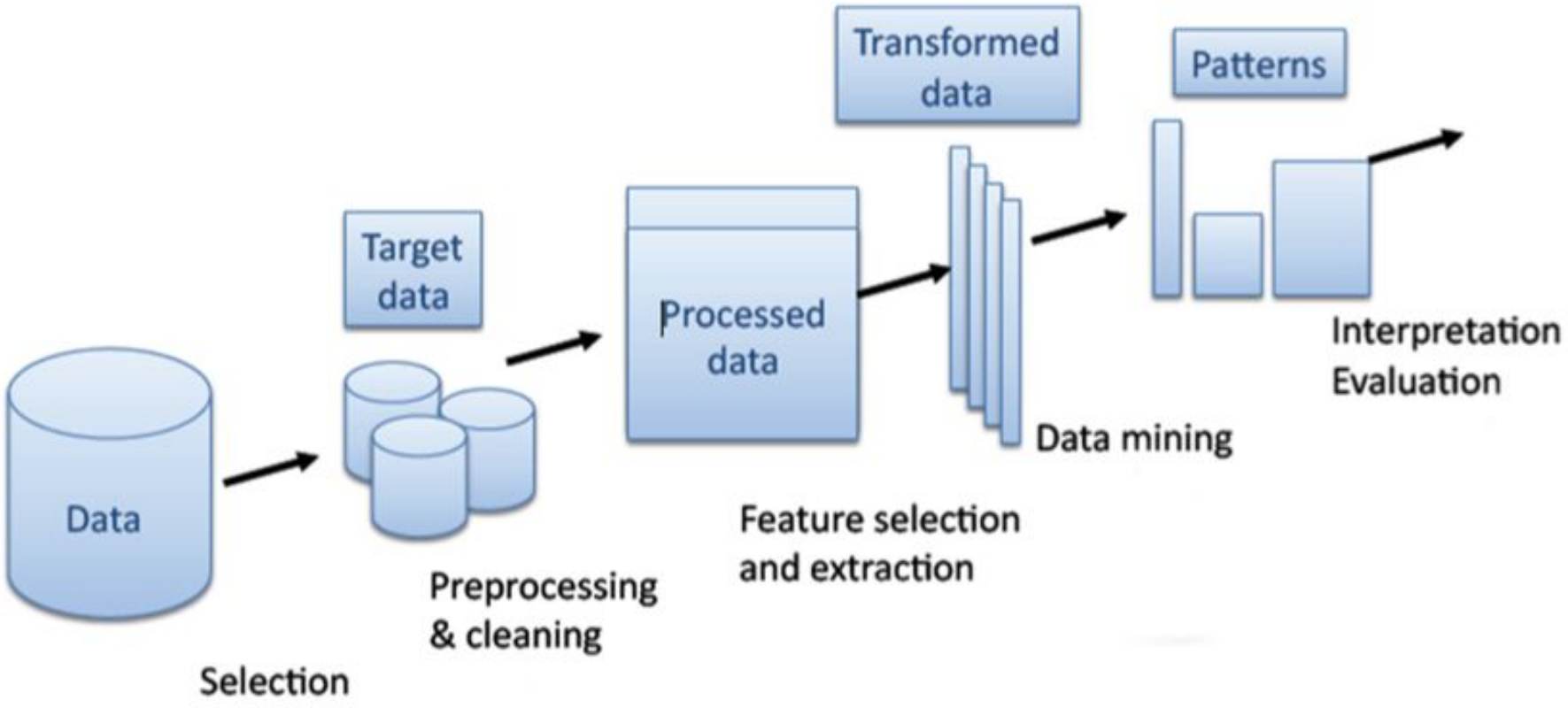
Disciplinas

Mineração de dados é uma **confluência de diversas disciplinas**:



Processo KDD

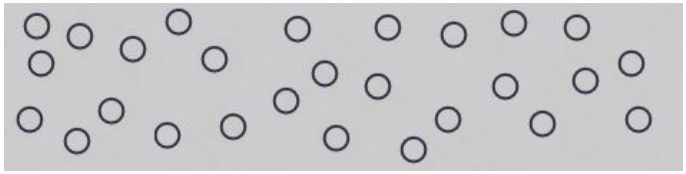
É uma parte integral da área de **KDD** (Knowledge Discovery in Databases - Descoberta de Conhecimento em Banco de Dados), que é o processo de conversão de dados brutos em informações úteis:



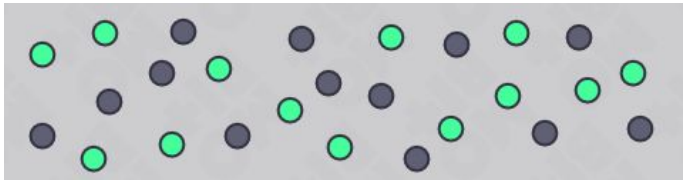
Fonte: http://recommender-systems.readthedocs.io/en/latest/datamining.html

Dados / Informação / Conhecimento

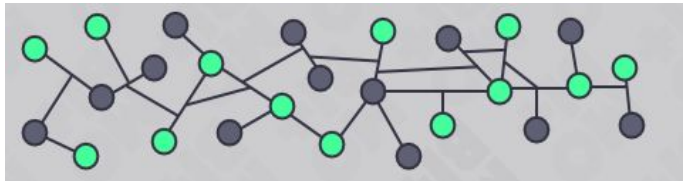
Dados: valores que não estão contextualizados ou organizados. São registros que isoladamente não são significativos ou úteis. Ex.: “fralda”



Informação: dados tratados e contextualizados possuindo significado, e podendo contribuir no processo de tomada decisões. Ex.: “fralda” -> “cerveja”

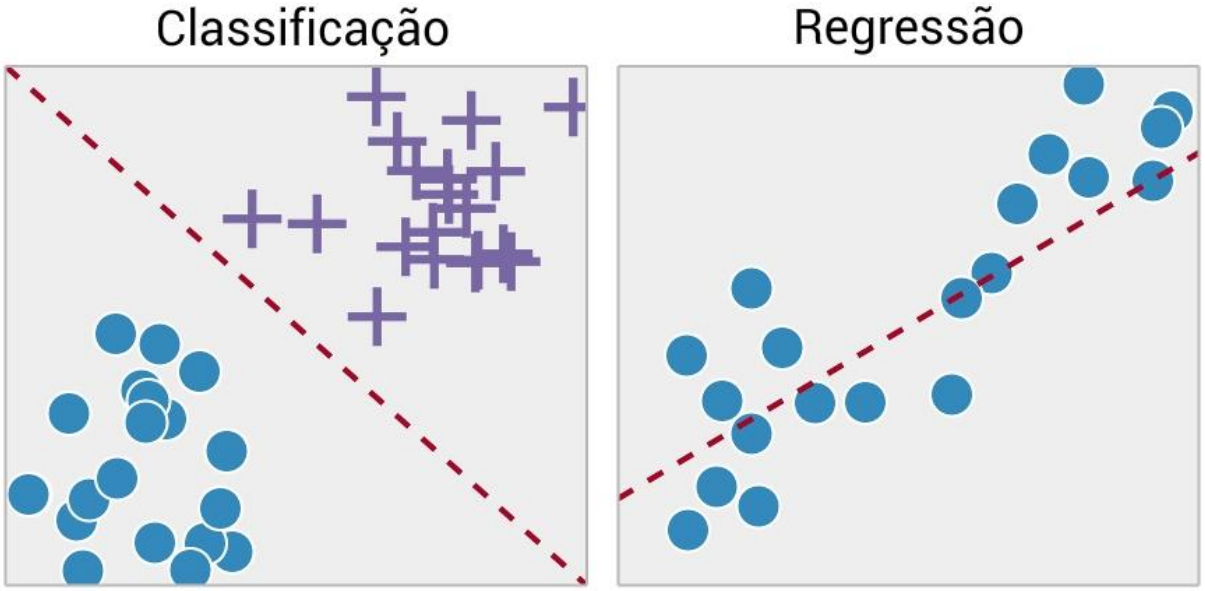


Conhecimento: processamento das informações, proporcionando um entendimento (modelo) sobre pessoas, objetos ou eventos. Ex.: Quando pais vão à noite no mercado comprar fraldas, aproveitam para comprar cerveja.



Tarefas de mineração de dados

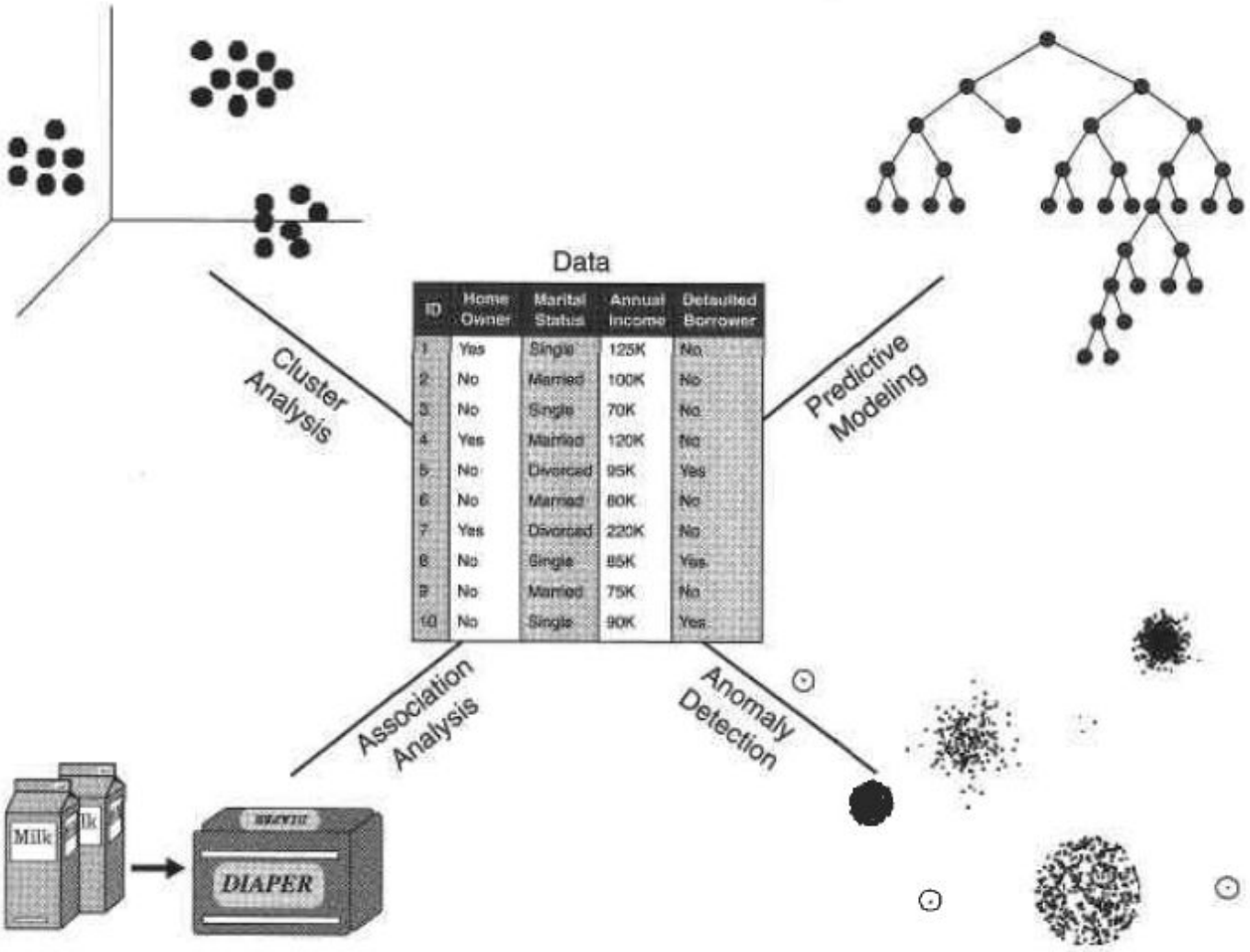
- **Tarefas de previsão:** prever o valor de um determinado atributo baseado nos valores de outros atributos.



Fonte: http://ipython-books.github.io/featured-04/

- **Tarefas descritivas:** derivar padrões (correlações, tendências, grupos e anomalias) que descrevam o relacionamento subjacente dos dados.

Tarefas de mineração de dados



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Fonte: http://recommender-systems.readthedocs.io/en/latest/datamining.html

Ementa do curso

1) Introdução à Mineração de Dados (11/10/2018 - 18:30h)

- Processo KDD
- Jupyter
- NumPy
- SciPy
- Scikit-learn
- Pandas
- Gráficos básicos
- **Atividade sala:** configuração e exploração do ambiente
- **Atividade da semana:** implementar um widget interativo no Jupyter

2) Análise Exploratória de Dados (25/10/2018 - 18:30h)

- Datasets (Kaggle, UCI, etc.)
- Conceitos e técnicas
- Gráficos
- **Atividade sala:** escolher um dataset e iniciar análise
- **Atividade da semana:** concluir análise completa e apresentar para a turma

Ementa do curso

3) Crawling e APIs (01/11/2018 - 18:30h)

- Coleta e persistência de dados remotos
- Consumo de APIs
- **Atividade sala:** construir um dataset a partir da coleta de um site
- **Atividade da semana:** realizar a análise exploratória do dataset coletado e apresentar para a turma

4) Processamento de Linguagem Natural (08/11/2018 - 18:30h)

- Manipulação e pré-processamento de textos
- Vetorização
- Métricas de similaridade
- Análise de sentimentos
- **Atividade sala:** realizar análise de sentimentos dos dados de um site ou API
- **Atividade da semana:** implementar um sistema de recomendação de notícias semelhantes para o site do campus

Ementa do curso

5) Transformação de dados (14/11/2018 - 18:30h)

- Normalização
- Redução de dimensionalidade
- Modelagem de tópicos
- **Atividade sala:** gráfico PCA em 3 dimensões
- **Atividade da semana:** aplicar modelagem de tópicos a um site ou API

6) Agrupamentos (22/11/2018 - 18:30h)

- Clusterização
- Clusterização hierárquica
- Regras de associação
- **Atividade sala:** desenvolver um analisador de co-ocorrência de palavras
- **Atividade da semana:** implementar modelagem de tópicos utilizando clusterização

Ementa do curso

7) Classificação (29/11/2018 - 18:30h)

- Conceitos
- Modelos
- Métricas
- **Atividade sala:** competição Titanic no Kaggle
- **Atividade da semana:** escolher uma competição de classificação no Kaggle e apresentar para a turma o problema, sua análise e modelagem para solução proposta

8) Regressão (06/12/2018 - 18:30h)

- Conceitos
- Modelos
- Métricas
- **Atividade sala:** competição House Prices no Kaggle
- **Atividade da semana:** escolher uma competição de regressão no Kaggle e apresentar para a turma o problema, sua análise e modelagem para solução proposta

Ementa do curso

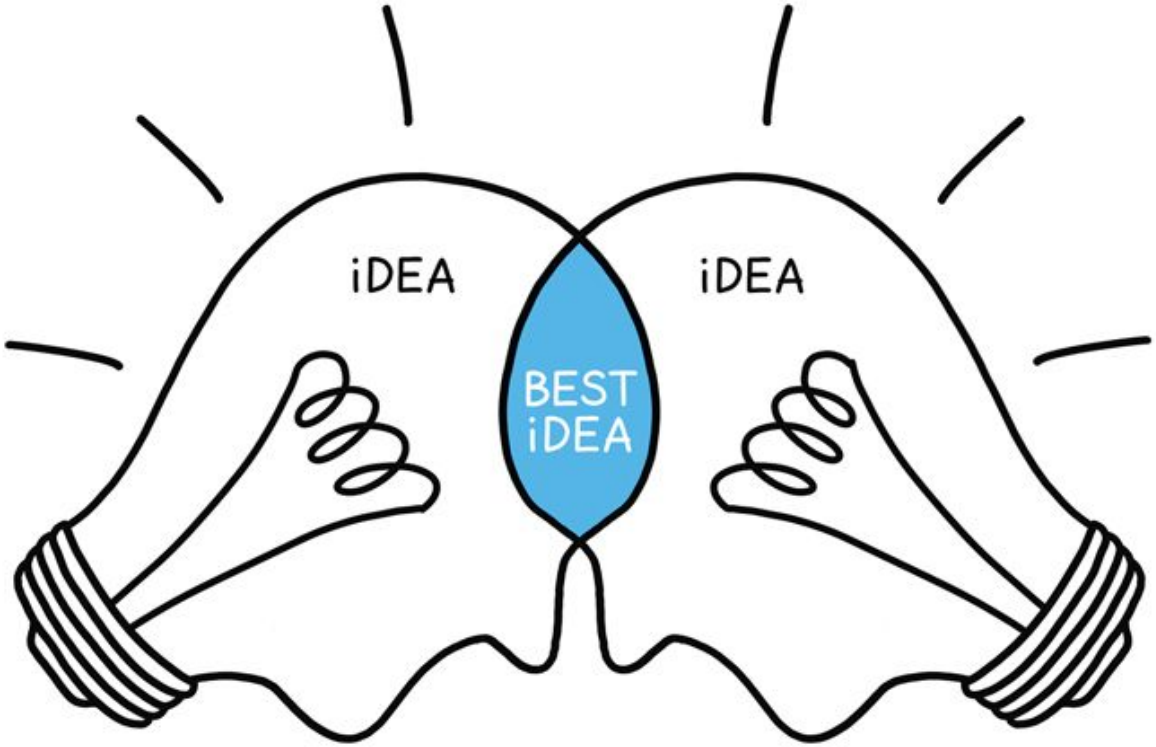
9) Deploy web (13/12/2018 - 18:30h)

- Persistência e deploy de um modelo treinado
- Desenvolvimento de uma API RESTful
- Front-end web
- Chart.js
- **Atividade sala:** desenvolvimento de uma API

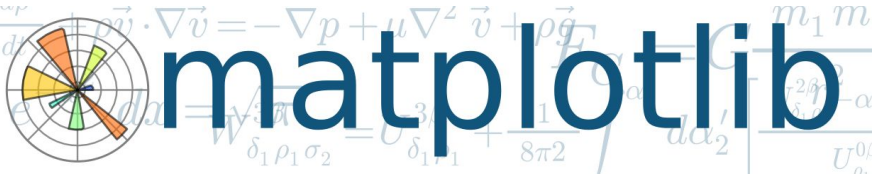
10) Apresentação dos trabalhos finais (20/12/2018 - 18:30h)

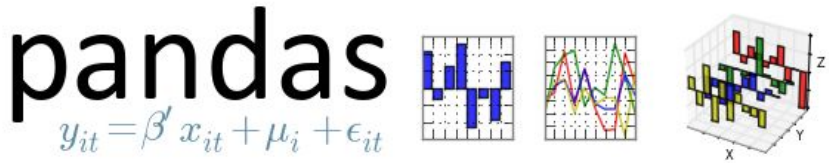
Brainstorm

Discutir com a turma sugestões e possibilidades de trabalhos.



Tecnologias e ferramentas





Tecnologias e ferramentas

Python 3.x

https://www.python.org/

Revisão básica:

- Sintaxe
- Iterações
- Indexação
- Lambda
- List Comprehensions

Tecnologias e ferramentas

Conda

https://conda.io/docs/index.html


```
conda create --name labredes

source activate labredes

source deactivate
```

Tecnologias e ferramentas

Jupyter Notebook

http://jupyter.org/


```
# Com o ambiente ativado
pip install jupyter


jupyter notebook
```

Tecnologias e ferramentas

NumPy

http://www.numpy.org/


```
# Com o ambiente ativado
pip install numpy
```

Exemplos NumPy

```
np.array([1, 2]) # array([1, 2])

np.matrix([1, 2]) # matrix([[1, 2]])

np.zeros([3, 4])
array([[0., 0., 0., 0.],
       [0., 0., 0., 0.],
       [0., 0., 0., 0.]])

m = np.matrix([[1.5, 2, 3], [4, 5, 6]], dtype=float)
matrix([[1.5, 2. , 3. ],
        [4. , 5. , 6. ]])

m.shape # (2, 3)

m.flatten() # array([1.5, 2. , 3. , 4. , 5. , 6. ])

m.reshape(3, 2)
matrix([[1.5, 2. ],
        [3. , 4. ],
        [5. , 6. ]])
```

Exemplos NumPy

```
np.array([1, 2, 3]) * 5  # array([ 5, 10, 15])

m * 10
matrix([[15., 20., 30.],
        [40., 50., 60.]])

np.array([1, 2, 3]) * np.array([1, 0, 2])
array([1, 0, 6])

np.min([1, 2, 3, 4]) # 1

np.max([1, 2, 3, 4]) # 4

np.sum([1, 2, 3, 4]) # 10

np.mean([1, 2, 3, 4]) # 2.5

np.median([1, 2, 3, 4, 5]) # 3.0

np.std([1, 2, 3, 4]) # 1.118033988749895
```

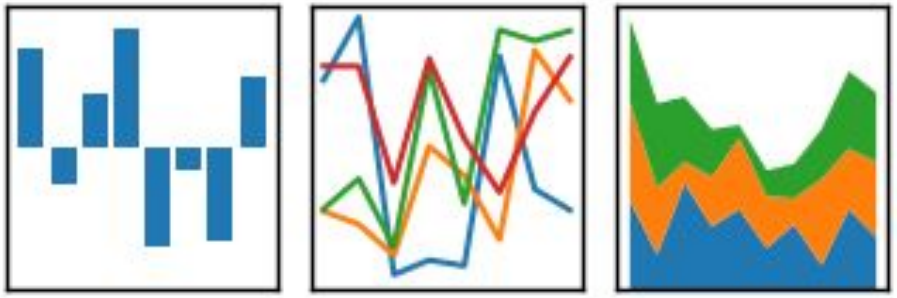
Tecnologias e ferramentas

Pandas

https://pandas.pydata.org/

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



```
# Com o ambiente ativado
pip install pandas
```

Tecnologias e ferramentas

scikit-learn

http://scikit-learn.org/


```
# Com o ambiente ativado
pip install scikit-learn
```

Tecnologias e ferramentas

Matplotlib

https://matplotlib.org


```
# Com o ambiente ativado
pip install matplotlib
```

Exemplo matplotlib

```
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np

t = np.arange(0.0, 2.0, 0.01)
s = 1 + np.sin(2 * np.pi * t)

fig, ax = plt.subplots()
ax.plot(t, s)

ax.set(
    xlabel='Eixo X',
    ylabel='Eixo Y',
    title='Meu primeiro gráfico'
)

ax.grid()
plt.show()
```

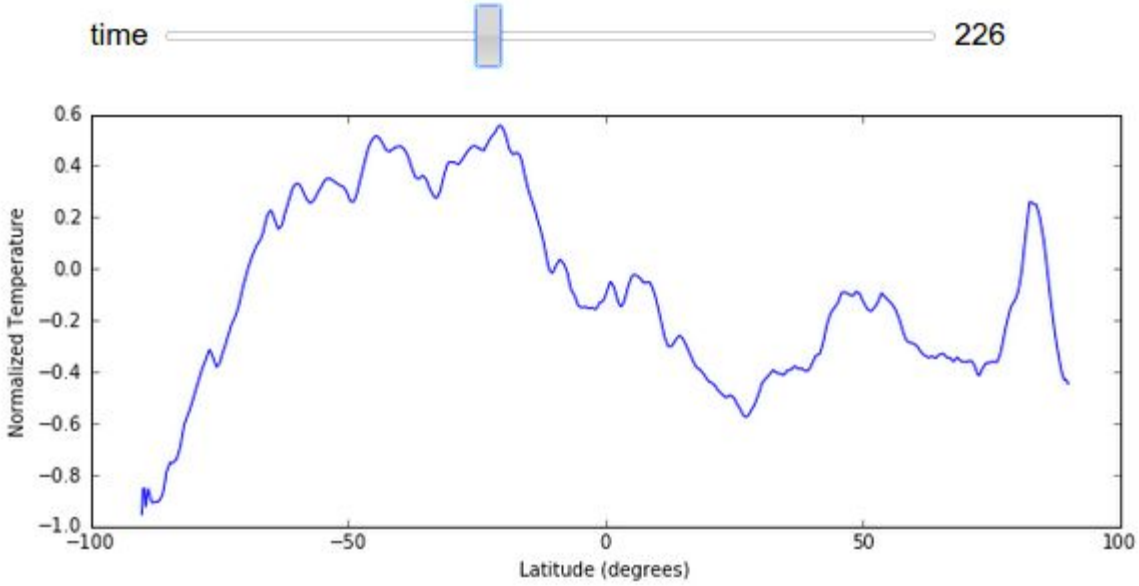
Galeria de exemplos: https://matplotlib.org/gallery/index.html

Atividade da semana

Implementar um **widget** no Jupyter para controlar um gráfico interativamente.

http://jupyter.org/widgets

Exemplo:



Referências

[1] GHAREHCHOPOGH, Farhad Soleimanian; KHALIFELU, Zeinab Abbasi. Analysis and evaluation of unstructured data: text mining versus natural language processing. In: Application of Information and Communication Technologies (AICT), 2011 5th International Conference on. IEEE, 2011. p. 1-4.

[2] PANG-NING, T.; STEINBACH, M.; KUMAR, V. Introdução ao “Data Mining”. Rio de Janeiro: Ciência Moderna, 2009.