Big Data & Machine Learning

Daniela Maria Uez
dani.uez@gmail.com

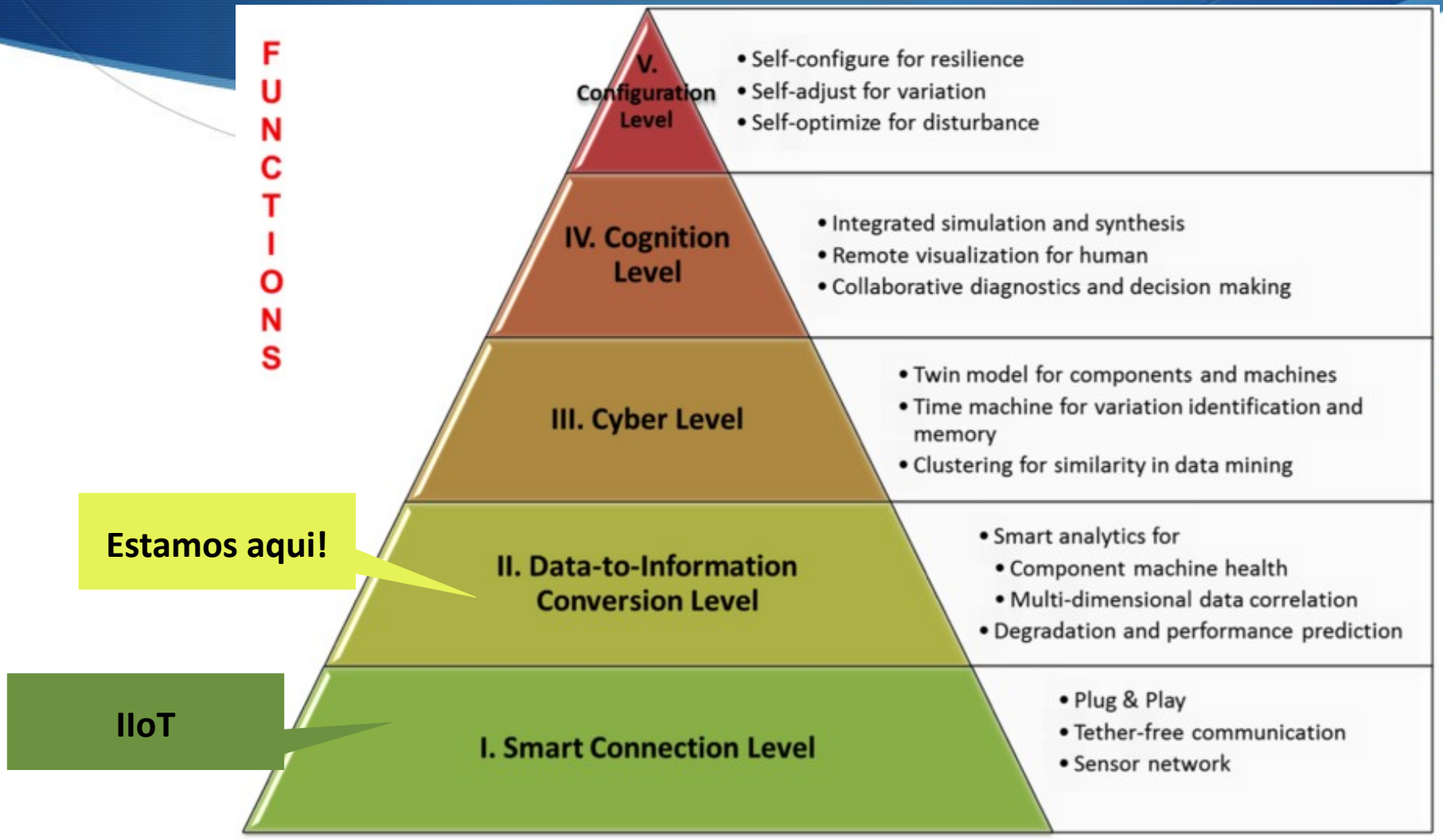
18 JULHO 2019

UCS
UNIVERSIDADE
DE CAXIAS DO SUL

Anteriormente…

Uma rápida recapitulação da aula passada...

Arquitetura 5C



O que é Big data?

“**Grandes** dados”

Dados com maior **variedade** que chegam em **volumes** crescentes e com **velocidade** cada vez maior.

Variedade: diversas fontes (documentos, BDs, JSON, imagem, video) – estruturadas, não-estruturadas, semi-estruturadas

Trabalhando com os dados: Análise exploratória

RESUMIR E ORGANIZAR os dados coletados através de tabelas, gráficos ou medidas numéricas

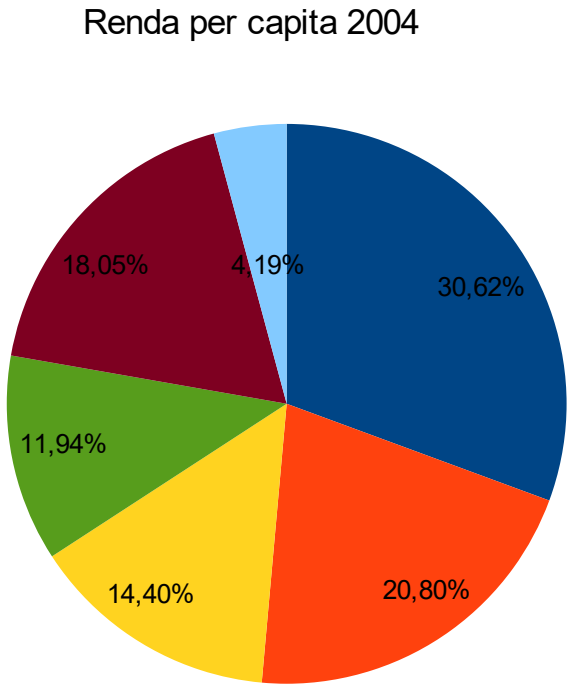
Passo1: normalizar os dados ajustando dados perdidos, discrepantes, com registro errado, etc

Passo2: verificar a frequência com que cada valor aparece na tabela

Passo3: analisar as medidas de posição (média, mediana, moda), de dispersão (intervalo, variância, desvio padrão...) e separatrizes (quartis, centis...)

Análise exploratória

Faixas de renda	Qtd	%
Até 3000,00	1703	30,62%
3000,00 até 5500,00	1157	20,80%
5500,00 até 8000,00	801	14,40%
8000,00 até 11000,00	664	11,94%
Acima de 11000,00	1004	18,05%
Não Informado	233	4,19%
Total	5562	100%



Análise exploratória (II)

Média	7571,95
Moda	1714,97
Mediana	5050,96
1º Quarti	2596,30
3º Quartil	9365,695
Variância	115357767,63
Desvio Padrão	10740,47
Intervalo	315188,03
Mínimo	20,05
Máximo	315208,07
Contagem	5329

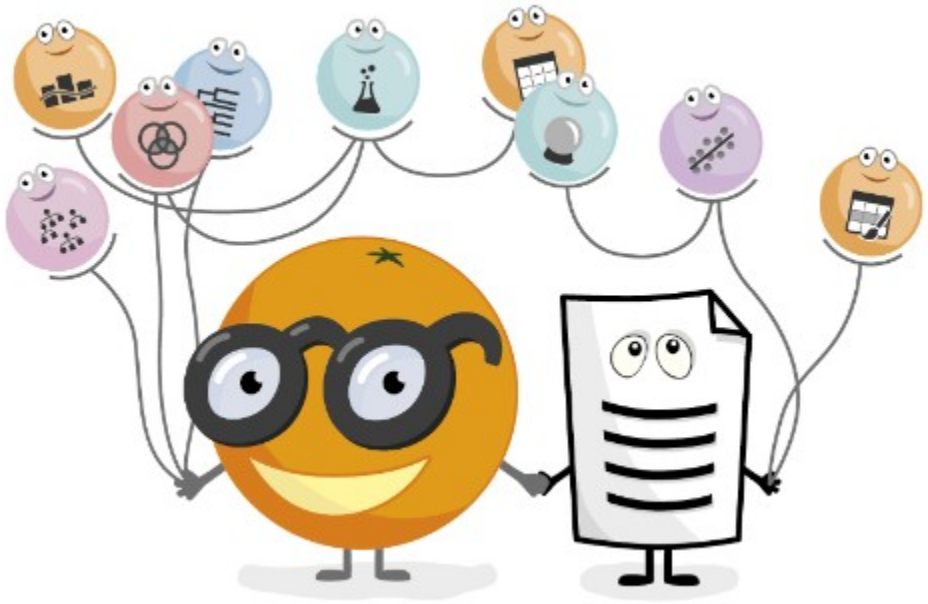
E agora seguimos a programação normal...

[…]

Orange

Kit de ferramentas de visualização de dados
Permite trabalhar com dados de forma visual

https://orange.biolab.si/



Machine Learning
Aprendizagem de Máquina

Área da IA que investiga técnicas que permitem aos computadores adquirir novas habilidades e conhecimentos sem que sejam especificamente programados para isso

É um campo de estudo que dá aos computadores a capacidade de aprender sem estarem programados explicitamente

[Arthur L. Samuel, 1959]

Aprendizagem de Máquina

- **O que é aprender?**
  - Adquirir conhecimento
  - Adquirir habilidade prática
  - Ter melhor compreensão de algo

- **Por que máquinas precisam aprender?**
  - Nem sempre um programa simples resolve os problemas.

    Ex.: filtro de spam
    - Quais tipos de e-mail são spam?

Inferência indutiva

- **Tirar uma conclusão sobre todos os membros de uma classe com base em informações sobre poucos membros**

Inferência indutiva

- **Tirar uma conclusão sobre todos os membros de uma classe com base em informações sobre poucos membros**

Esse e-mail é spam
Tem a palavra **desconto** no título
Todos os e-mails com a palavra desconto no título são **spams**

Inferência indutiva

- **Tirar uma conclusão sobre todos os membros de uma classe com base em informações sobre poucos membros**

- Raciocínio indutivo nem sempre está certo, mas é uma boa maneira de generalizar um comportamento
- Eu posso avisar o algoritmo que o e-mail não é spam para melhorar a classificação

Aprendizagem de Máquina

- Um programa aprende a partir da experiência **E** em relação a uma classe de tarefas **T** com medida de desempenho **P** se o desempenho **P** em **T** melhora com **E**

  Ou seja:

  O programa **aprendeu** se o desempenho na execução de uma classe de tarefas melhorar conforme as tarefas forem sendo executadas

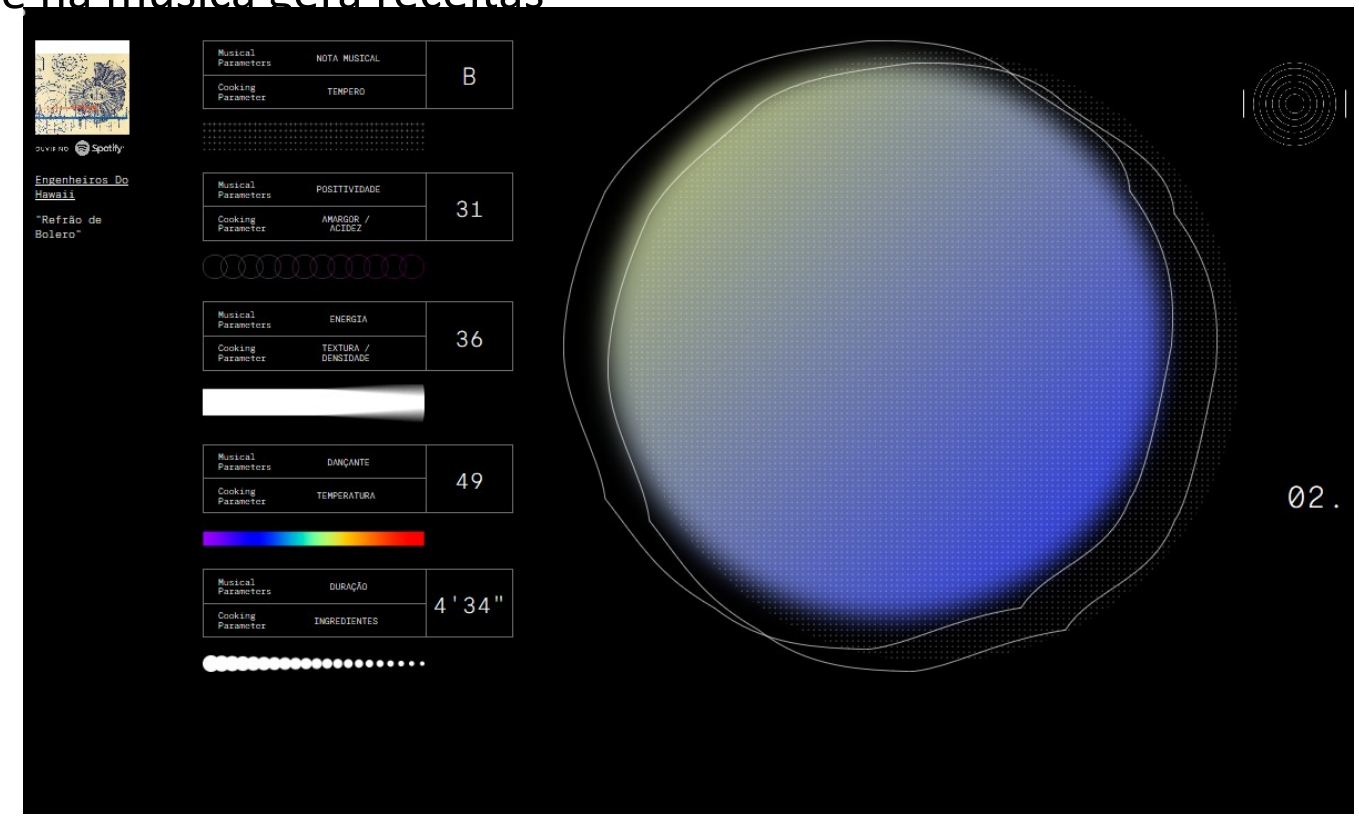
- Ex.: Conforme for classificando e-mails, a quantidade de classificações erradas diminui

Onde usam?

- Reconhecimento facial
  - Busca por pessoas desaparecidas
  - Busca por criminosos
  - Identificação de comportamentos criminosos
- Mecanismos de busca como o Google
- Sistemas de recomendação
- Reconhecimento de linguagem natural:
  - Reconhecimento de fala – conversar com as máquinas
  - Assistentes pessoais
- Reconhecimento de escrita
- Detecção de anomalias
  - Manutenção preventiva

Sabor das músicas

- Tramontina + Spotify: https://sabordasmusicas.withspotify.com
- Com base na música gera receitas



Tipos de aprendizagem

- **Aprendizagem supervisionada**
- **Aprendizagem não supervisionada**
- **Aprendizagem por reforço**

Referências
- Bases de Dados -

- Censo do Legislativo - http://dados.gov.br/dataset/censo-do-legislativo
- IBGE - https://downloads.ibge.gov.br/downloads_estatisticas.htm
- 33 Brilliant And Free Data Sources Anyone Can Use - https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#39b6b626b54d
- Portal da transparência – dados do governo federal

Referências

Orange - https://orange.biolab.si/

Tutorial sobre Orange: https://orange3.readthedocs.io/projects/orange-development/tutorial.html

Dados da aula: www.uez.com.br/ucs