Data Science e Machine Learning na Prática: Introdução e Aplicações na Indústria de Processos

Apresentação 1 - Conceitos Introdutórios

Afrânio Melo

afraeq@gmail.com
afrjr.weebly.com

Escola Piloto Prof. Giulio Massarani
PEQ-COPPE-UFRJ

2019

1. Data Science
   - Frases interessantes
   - O que é a Ciência de Dados?
   - Big Data

2. Machine Learning
   - Definições
   - Programação tradicional x Aprendizado de máquina
   - Tipos de algoritmos
   - Passo a passo de um projeto de Machine Learning
   - Dificuldades e problemas
   - Exemplos de aplicações

Algumas frases interessantes...

- "**Se os dados tivessem massa, a Terra seria um buraco negro**" - Stephen Marsland, acadêmico;
- "**A grande tendência da tecnologia é tornar os sistemas inteligentes, e para isso a matéria-prima são os dados** " – Amod Malviya, FlipKart CTO;
- "**Aprender a partir dos dados é universalmente útil. Domine essa arte e você será bem-vindo em qualquer lugar**" – John Elder, Elder Research CEO;
- "**Torture os dados, e eles vão confessar tudo**" – Ronald Coase, Prêmio Nobel de Economia;
- "**Guerra é 90% informação**" – Napoleão Bonaparte, militar e monarca francês.

Afinal, o que é a Ciência de Dados?

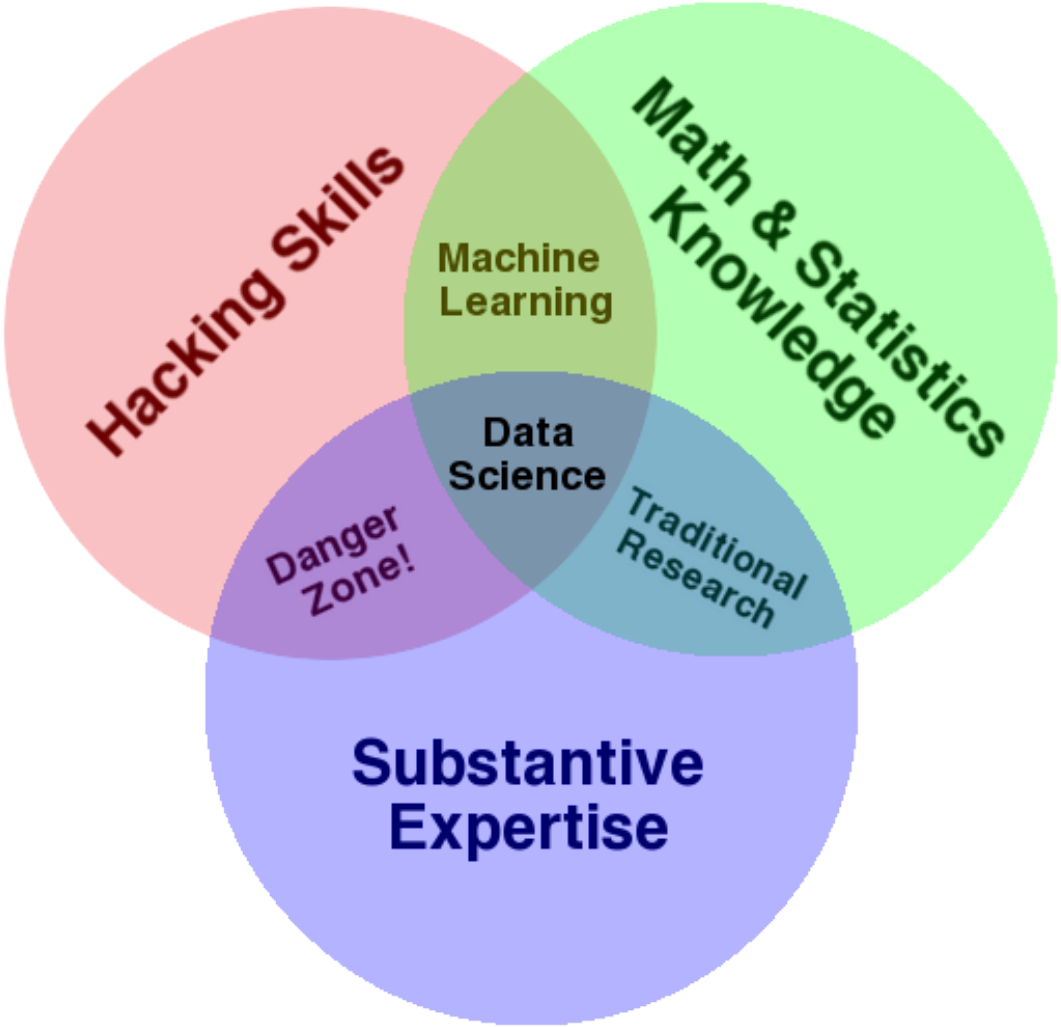
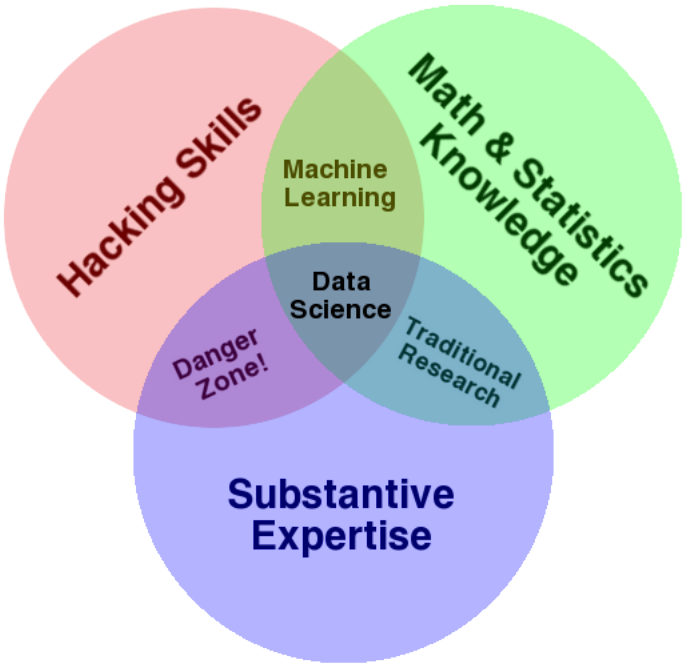


Figura 1: Diagrama de Venn definindo a Ciência de Dados como a interseção de três competências.

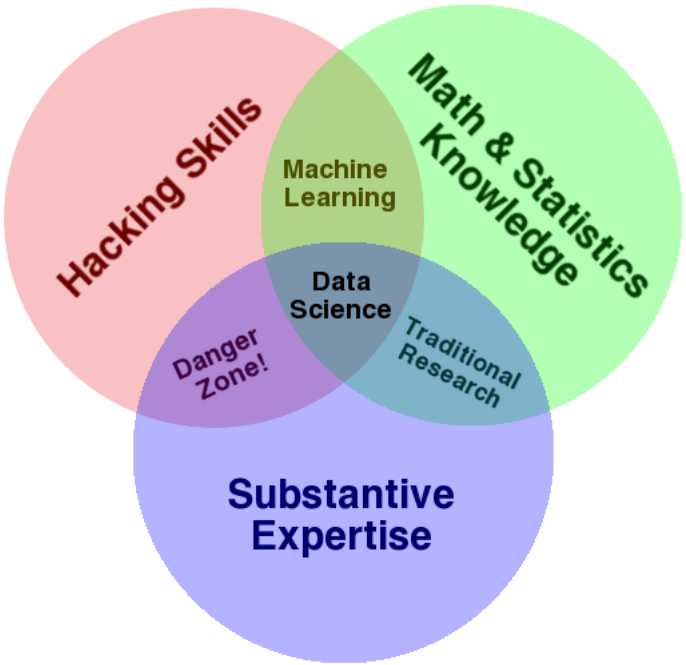
Afinal, o que é a Ciência de Dados?



Competência 1: habilidades computacionais

- Como os dados são bens armazenados e negociados eletronicamente, um cientista de dados precisa ter algumas habilidades computacionais, destacando-se:
  - pensamento algorítmico;
  - capacidade de manipular dados em diferentes formatos;
  - entendimento de operações vetorizadas.

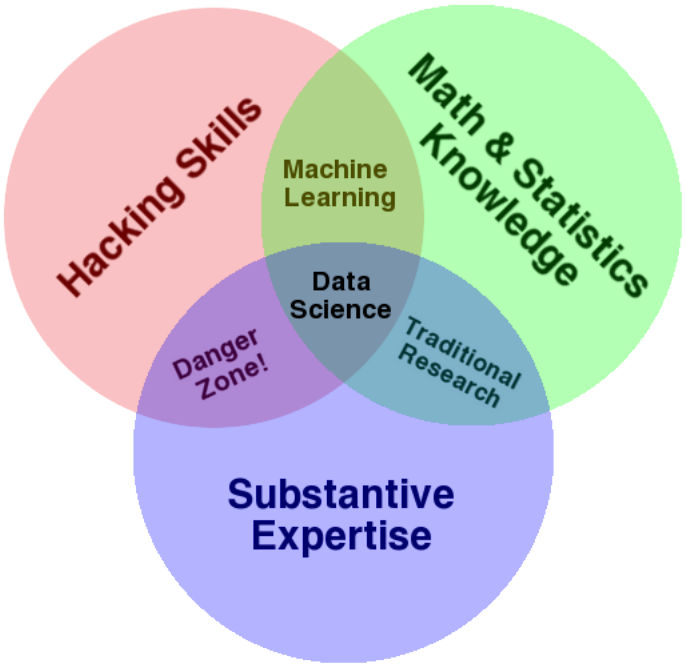
Afinal, o que é a Ciência de Dados?



Competência 2: conhecimento matemático

- Para extrair significado dos dados, é necessário a aplicação de métodos matemáticos e estatísticos.
- Isso requer um conhecimento da base teórica dos métodos e de seus mecanismos de funcionamento.

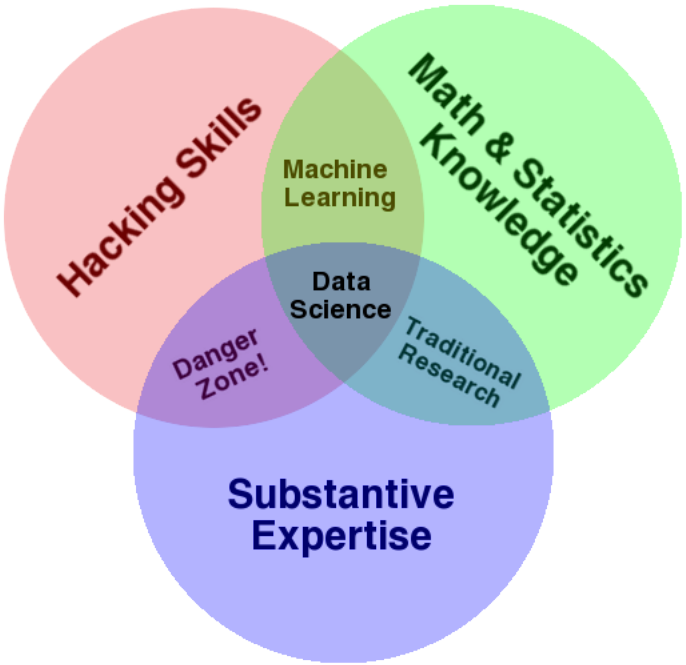
Afinal, o que é a Ciência de Dados?



Competência 3: conhecimento do domínio

- A terceira competência é aquela que justifica o uso do termo "ciência" em "ciência de dados".
- Ter conhecimento do domínio (sobre engenharia química, administração, publicidade, sociologia, etc.) permite usar o significado extraído dos dados para gerar descobertas e construir conhecimento!

Afinal, o que é a Ciência de Dados?



Para pensar

- Reflita sobre as diferentes interseções entre os conjuntos no diagrama ao lado. Onde suas habilidades atuais se encaixam? Onde as habilidades da maioria das pessoas do mundo acadêmico se encaixa? Há algum motivo para isso?

Afinal, o que é a Ciência de Dados?

Figura 2: Outro diagrama que define a Ciência de Dados por meio de seus componentes.

E o tal do *Big Data*?

Definindo *Big Data*

- O termo *Big Data* surgiu para descrever conjuntos de dados com as seguintes características:
  - **Volume**: grande quantidade de dados (da ordem de vários TB);
  - **Velocidade**: taxa na qual novos dados são gerados;
  - **Variedade**: diferentes formas nas quais e fontes das quais os dados são coletados (texto, som, vídeo, diferentes sensores, etc).
- Os três itens acima são comumente conhecidos como os três V's do *Big Data*.
- A ementa desse curso não cobre os métodos de análise de dados especificamente voltados para *Big Data*.
- Os conhecimentos aqui aprendidos, porém, certamente constituem pré-requisito para a compreensão desses métodos.

E o tal do *Big Data*?

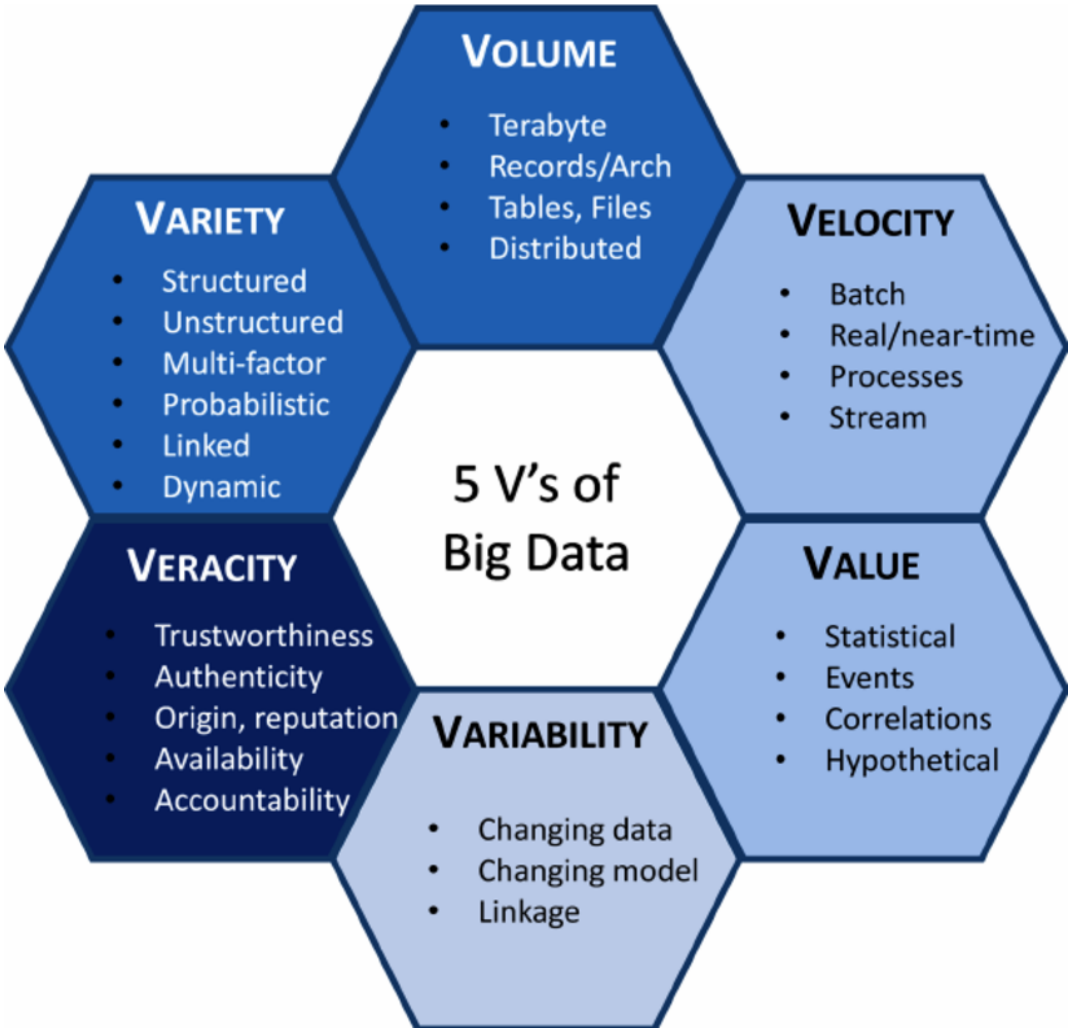


Figura 3: A cada dia surgem mais V's na definição de *Big Data*.

Machine Learning - Definições

- "**O *aprendizado de máquina* é o campo de estudo que fornece aos computadores a habilidade de aprender sem serem explicitamente programados**" – Arthur Samuel;
- "**Diz-se que um programa de computador aprende com a experiência E em relação a alguma tarefa T e alguma medida de desempenho P, se seu desempenho em T, medido por P, melhora com a experiência E.**" – Tom Mitchell.

Machine Learning - Definições

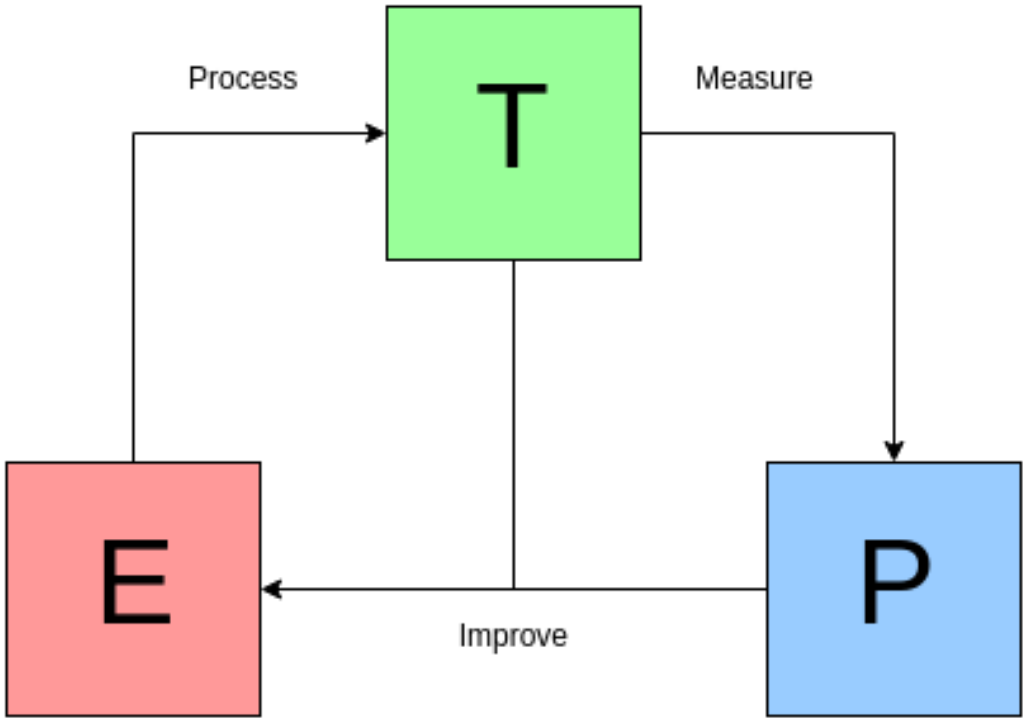


Figura 4: Paradigma de Mitchell.

Machine Learning - Definições

- "**O *aprendizado de máquina* é essencialmente uma forma de estatística aplicada, com ênfase crescente no uso de computadores para estimar estatisticamente funções complicadas e menor ênfase na obtenção de intervalos de confiança em torno dessas funções.**" – Ian Goodfellow;

Machine Learning *vs.* Inteligência Artificial

Machine Learning não é Inteligência Artificial?

- Na verdade, o conceito de Inteligência Artificial (AI) é bem mais amplo e inclui:
  - raciocínio simbólico e prova de teoremas;
  - robótica;
  - visão computacional;
  - sistemas especialistas;
  - aprendizado de máquina;
  - etc...
- Machine Learning pode ser visto como um campo da Inteligência Artificial.

Machine Learning *vs.* Inteligência Artificial

Machine Learning não é Inteligência Artificial?

- Uma definição famosa e abrangente de AI é:

  "**A Inteligência Artificial é o estudo de como fazer computadores realizar tarefas nas quais, no momento, pessoas são melhores**" – Elaine Rich.

- Claramente é uma definição mais ampla do que as de Machine Learning apresentadas anteriormente.

Programação tradicional x Aprendizado de máquina

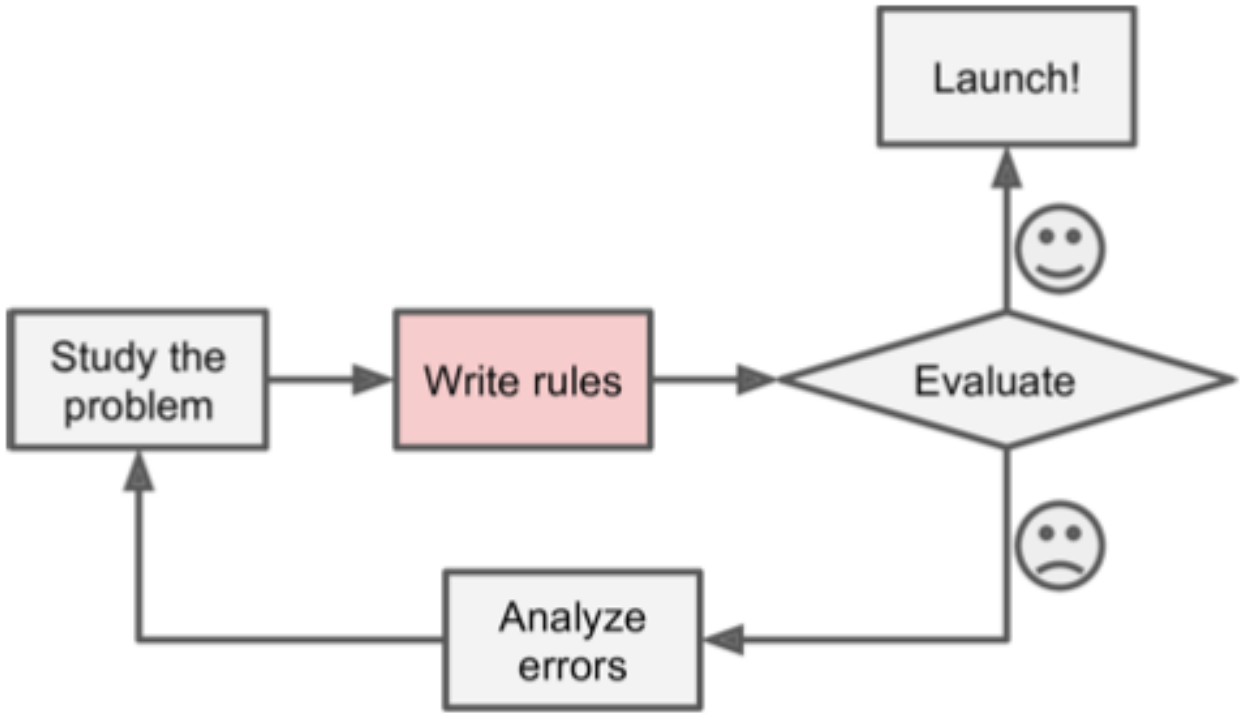


Figura 5: Abordagem tradicional de resolução de problemas. Fonte: GERON, 2017

.

Programação tradicional x Aprendizado de máquina

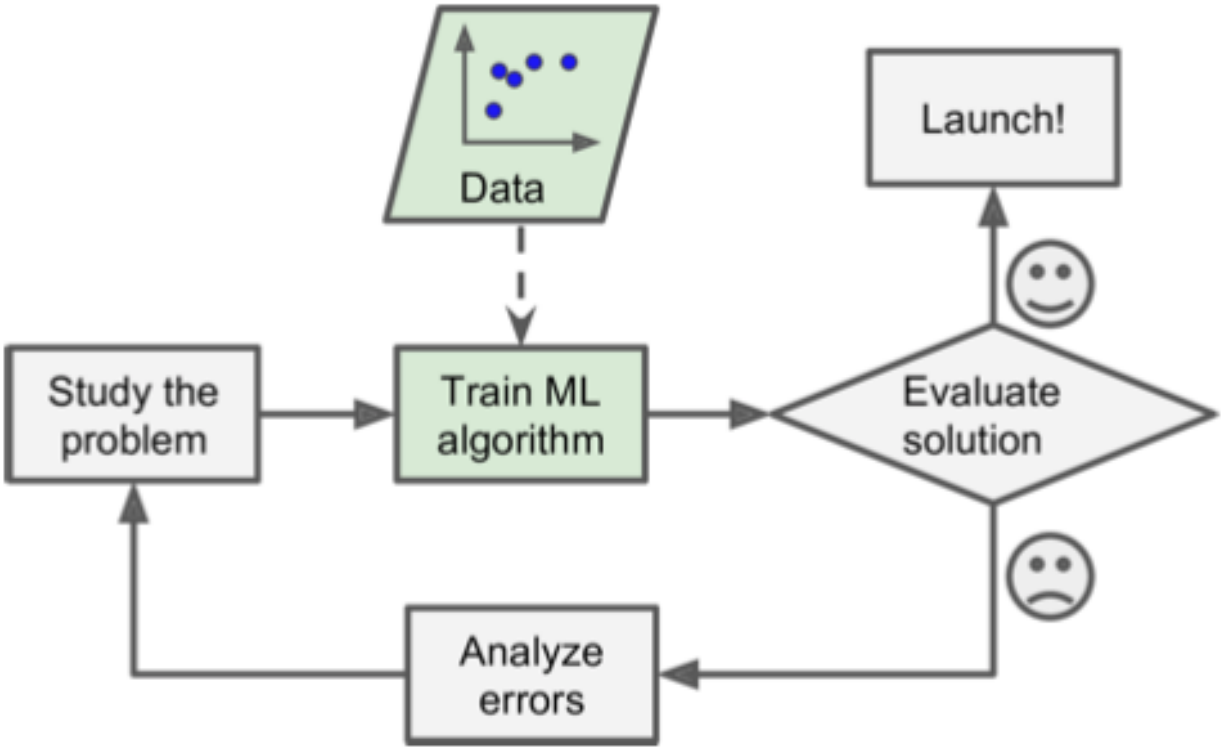


Figura 6: Abordagem de resolução de problemas utilizando aprendizado de máquina. Fonte: GERON, 2017

Programação tradicional x Aprendizado de máquina

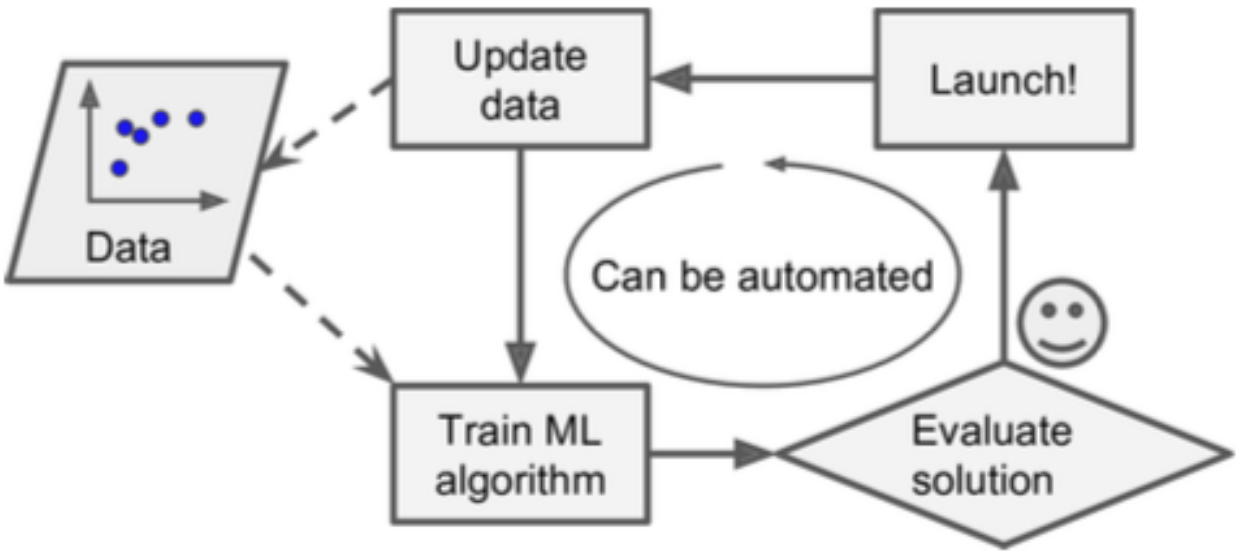


Figura 7: Abordagem de resolução de problemas automatizada utilizando aprendizado de máquina. Fonte: GERON, 2017

Programação tradicional x Aprendizado de máquina

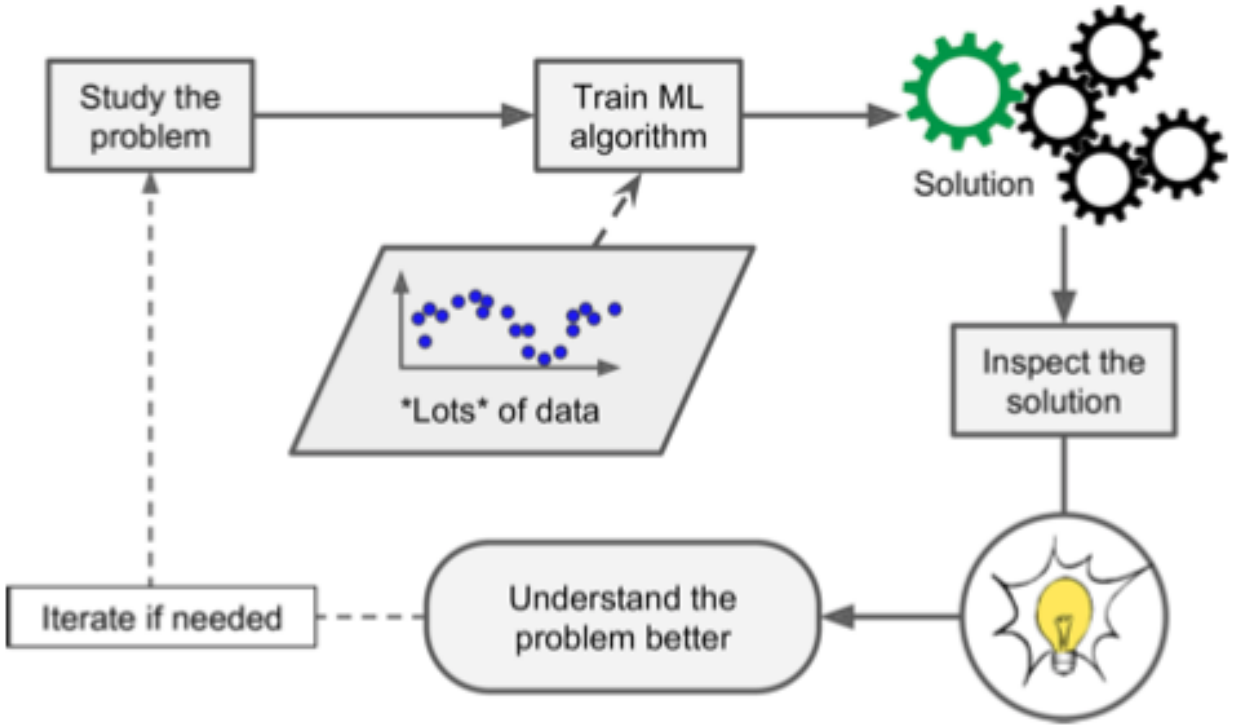


Figura 8: Aprendizado de máquina auxiliando no aprendizado de humanos e vice-versa. Fonte: GERON, 2017

Tipos de algoritmos

Quanto à supervisão humana:

- **Supervisionados**: o conjunto de treinamento possui a solução desejada (tipicamente na forma de rótulos).
- **Não-supervisionados**: o conjunto de treinamento não possui a solução desejada. O algoritmo tenta aprender sem um "professor".
- **Semi-supervisionados**: alguns dados de treinamento possuem a solução desejada e outros não.
- **Por reforço**: o algoritmo observa o ambiente, seleciona e performa ações e ganha recompensas (ou sofre penalidades) em retorno. Com base nessas recompensas ou penalidades, o algoritmo decide o melhor caminho a seguir. Exemplo: `https://www.youtube.com/watch?v=V1eYniJ0Rnk`.

Tipos de algoritmos

Quanto à frequência de treinamento:

- **Em batelada**: o treinamento ocorre de uma só vez, com todos os dados disponíveis. Também chamados de algoritmos de aprendizado *offline*.
- **Contínuos**: os dados de treinamento são fornecidos continuamente e os algoritmos aprendem de forma incremental. Também chamados de algoritmos de aprendizado *online*.

Tipos de algoritmos

Quanto à forma de generalização:

- **Baseados em instâncias**: usa-se medidas de similaridade para comparar os dados a serem avaliados com os dados de treinamento aprendidos.
- **Baseados em modelos**: contrói-se um modelo com os dados de treinamento. Os dados avaliados são então comparados com esse modelo.

Ilustração - algoritmos baseados em instâncias e em modelos

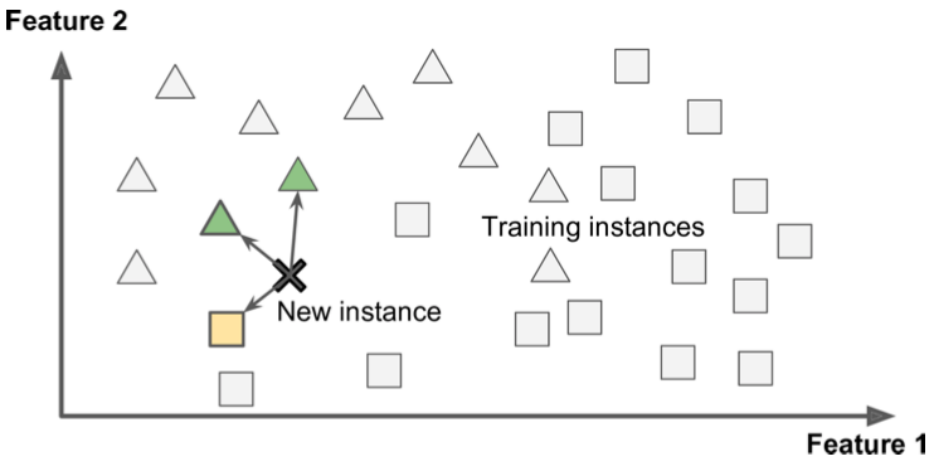


Figura 9: Exemplo de uma tarefa de classificação baseada em instâncias. Fonte: GERON, 2017

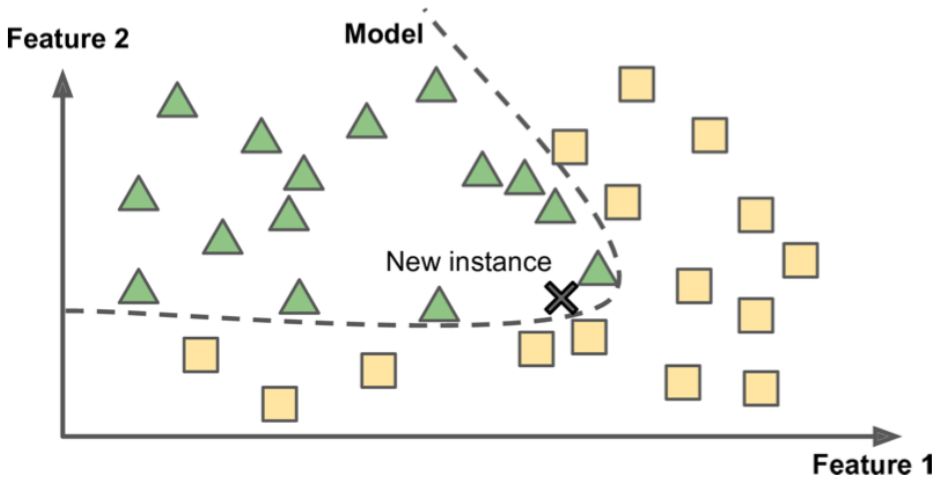


Figura 10: Exemplo de uma tarefa de classificação baseada em modelo. Fonte: GERON, 2017

Passo a passo de um projeto de Machine Learning

- **1** - estabeleça o problema e o analise de forma global;
- **2** - obtenha os dados;
- **3** - explore os dados, de modo a ganhar discernimento e sentimento sobre eles;
- **4** - prepare os dados, de modo a expor seus padrões aos algoritmos de Machine Learning da forma mais eficiente possível;
- **5** - explore vários algoritmos e modelos e escolha alguns com bom desempenho;
- **6** - ajuste os hiperparâmetros dos modelos escolhidos e combine-os para gerar uma única solução;
- **7** - apresente sua solução;
- **8** - implemente, monitore e mantenha seu sistema.

Dificuldades e problemas que podem ser enfrentados em um projeto de ML

- quantidade insuficiente de dados de treinamento;
- dados de treinamento não representativos:
  - se a amostra é muito pequena, pode ocorrer *ruído de amostragem*;
  - mas mesmo se a amostra é grande, pode ocorrer *viés de amostragem*.
- dados de qualidade ruim;
- atributos irrelevantes;
- sobreajuste dos dados;
- subajuste dos dados;

Exemplos de aplicações mais comuns de aprendizado de máquina

- diversos problemas de classificação;
- detecção de anomalias (como e-mails de spam, fraudes em cartões de crédito, falhas em equipamentos industriais, etc).;
- descoberta de relações e associações;
- extrapolação de séries temporais;
- sistemas de recomendação (Google, Netflix, etc);
- robótica;
- processamento de linguagens naturais;
- visão computacional;
- aprendizado de padrões de diferentes tipos.

Bibliografia recomendada – Livros práticos

Peter Flach.

Machine Learning - The Art and Science of Algorithms that Make Sense of Data.

Cambridge, 2012.

Aurélien Géron.

Hands–On Machine Learning with Scikit–Learn and TensorFlow.

O'Reilly, 2017.

Hans Petter Langtangen.

A Primer on Scientific Programming with Python.

Springer, 5 edition, 2016.

Stuart Russell and Peter Norvig.

Artificial Intelligence - A Modern Approach.

Prentice Hall, 3 edition, 2010.

Links para as imagens usadas

- http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
- http://i.imgur.com/4rVgcIx.png
- https://www.optimus360.com/big-data-como-tornar-as-estrategias-de-marketing-mais-assertivas/
- https://www.researchgate.net/figure/The-five-Vs-of-Big-Data-Adapted-from-IBM-big-data-platform-Bringing-big-data-to-the_fig1_281404634
- https://advocaretheblog.wordpress.com/2017/03/13/a-importancia-da-informacao-para-as-organizacoes-empresariais/
- https://www.embalagemmarca.com.br/2017/05/conheca-os-nove-pilares-para-implantacao-da-industria-4-0/
- https://www.embalagemmarca.com.br/2017/05/conheca-os-nove-pilares-para-implantacao-da-industria-4-0/
- https://www.kdnuggets.com/2018/12/essence-machine-learning.html

Obrigado pela atenção!