# DataScience

Fernando Anselmo - Versão 1.0

## Data Science


**Cientistas de Dados** são responsáveis por: coletar os dados, limpar e organizar os dados, construir bases de treinamento e garantir que não ocorra *overfitting*, construir algoritmos, gerar *insights* e apresentá-los.

**Habilidades em**: Estatística, entender do negócio e ter experiência no assunto, colaboração, resolução de problemas, ferramentas de visualização, bases de dados SQL e NoSQL, processamento de Big Data, Inteligência Artificial, Aprendizado de Máquina, Mineração de Dados, Linguagens de Programação, comunicação e criatividade.
**Habilidade Principal**: **Curiosidade**.

## Análises

**Descritiva** - responde: "o que aconteceu?" realizada com base em dados complementares e concorrentes. Necessita da Inteligência de Negócio.
**Diagnóstica** - responde: "qual o motivo?" útil para determinar o sucesso/fracasso de qualquer ação com base nos dados.
**Preditiva** - incita: "o que acontecerá?" extrapola a descritiva para prever uma tendência.
**Prescritiva** - incita: "o que deve ser feito?" com base em uma estimativa informada do que acontecerá.

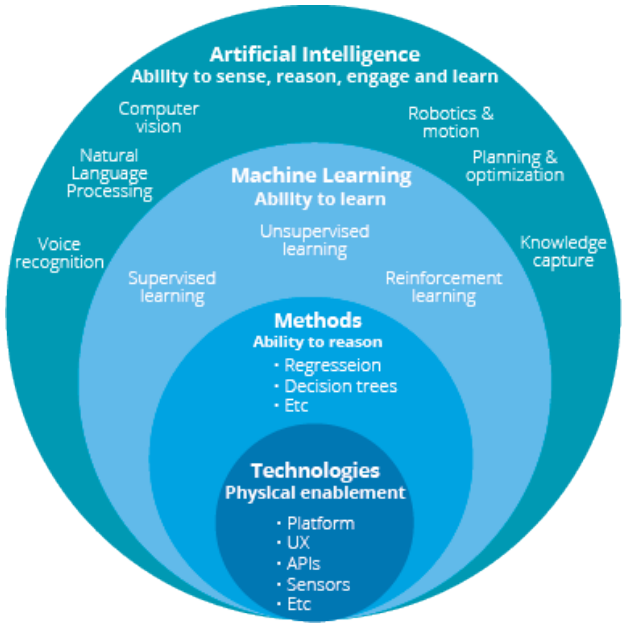
## Linguagens

Encarar a linguagem como uma ferramenta.
Pensar simples e consulte sempre.
Utilizar as bibliotecas, não reinventar a roda.
Documentar todo seu esforço.
Não focar em Orientação a Objetos, manter simples.
Não focar na tecnologia, muda constantemente.

## Data Swamp/Lake

**Data Swamp** são todos os dados que entram, não possuem documentação nem padrão.
**Data Lake** são os dados tratados (normalmente por um ETL como o Pentaho), estão em estado bruto e a disposição para serem explorados. Devem ser altamente acessíveis e passíveis de rápidas atualizações.

## IA & ML

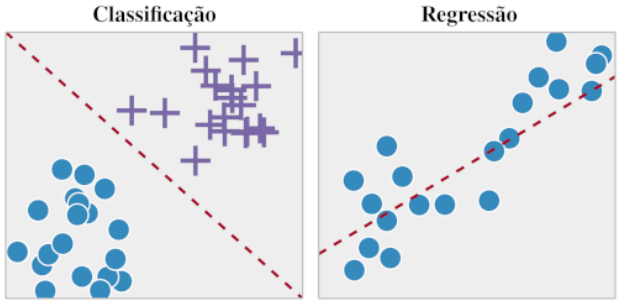



## Big Data

**Volume**: Quantidade. Escala acima de Petabytes.
**Velocidade**: Geração. Em tempo real, *streaming* e *batch*, entre servidores.
**Variedade**: Diferença. Estruturados, semi-estruturados, não estruturados e multi-fator.
**Veracidade**: São reais e podem ser comprovados.
**Valor**: Estatísticas, correlações, predições e hipóteses.
**Complexidade**: Alta dimensionalidade.
**Arquitetura**: Distribuída ou horizontal.
**Globalidade**: Variadas entradas de informação.

## Classificação e Regressão

**Classificação**: prever um rótulo, a variável de resposta é do tipo categórica. Exemplos: Spam/Não, Doentes/Não, Cliente/Não. Algoritmos: KNN, Regressão Logística, SVM, Árvores de Decisão, XGBoost e Redes Neurais.




**Regressão**: prever uma quantidade, a variável de resposta é do tipo contínua. Exemplos: Estimativas de preço, tempo de uso de um serviço. Algoritmos: Regressão Linear, Regressão Polinomial.

## Plataformas de Competição

**AICrowd** - `https://www.aicrowd.com`
**CodaLab** - `https://codalab.org`
**CrowdAnalytix** - `https://www.crowdanalytix.com`
**DataHack** - `https://www.analyticsvidhya.com`
**DrivenData** - `https://www.drivendata.org`
**HackerEarth** - `https://www.hackerearth.com`
**IDAO** - `https://idao.world`
**Iron Viz** - `https://www.tableau.com/iron-viz`
**Kaggle** - `https://www.kaggle.com/`
**MachineHack** - `https://analyticsindiamag.com`
**Tianchi** - `https://tianchi.aliyun.com`
**TopCoder** - `https://www.topcoder.com`
**Zindi** - `https://zindi.africa`

Outros Cartões: `https://github.com/fernandoans/publicacoes/tree/master/Sheet`