**Data Science**: segue três áreas

* Business:
* Dev:
* Math: Estatisticas, físicas

**Diferença entre Machine e Deep Learning**

* Machine Learning: uma camada de dados
* Deep Learning: várias camadas de dados

**Data Mining**: é formada por um conjunto de ferramentas e técnicas que através do uso de algoritmos de aprendizagem ou classificação baseados em redes neurais e estatística. Estes são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões nestes dados e auxiliando na descoberta de conhecimento. O conhecimento em Data Mining pode ser apresentado por essas ferramentas de diversas formas: agrupamentos, hipóteses, regras, árvores de decisão, grafos ou dendrogramas.

**Big data:**

é uma entidade. Possui dados (de qualquer “coisa”)

Exemplo de fontes públicas:

* Midias sociais (social analistics)
* - Jornais (noticias)
* - Kaggle

**Data Crawler:**

script montado para fazer capturas de dados, script de automação robótica utilizando conceito de RPA. Caso não tenha uma API é utilizado Data Crawler. É criado um script que vai até a página HTML e devolve as informações mais importantes.

Depois de capturar os dados joga em um **Data Lake** (não precisa organizar os dados)

**Data Lake**: geralmente trabalha com NOSLQ (Not only SQL) com uma estrutura Json.

**ETL (Extract, transform, Load)** – é um conceito, pega as informações de outras fontes (banco de dados, arquivos CSV, XML) por uma chave, gera um outro arquivo tratado e passa para um Dataset.

**Data Warehouse**: É utilizado ETL (extrair, transforma e dar carga) e depois passa para o Data Warehouse, armazenando DB que possui DataSet

**Dados estruturados**: dados que estão em tabelas. É utilizado NLP para estruturar os dados

**Dados não estruturados**: são as informações que não estão estruturados em uma tabela, só possui os dados.

**Crisp-DM (Cross Industry Standard Process for Data Mining)**

* **Business Understanding**: Entender o negócio, entender o problema.
* **Data Understanding**: Verificar os dados, entender se os dados atendem para resolver o problema.
* **Data Preparation**: Preparar/tratar os dados, tirar os blank’s/null’s.
  + Cleaning;
  + Transformation (transformação)/discretization;
  + Discretization;
  + Scaling;
  + Reduction: Técnica escolhe e tira a coluna, pra ver se dar certo e com **baixa** (aleatoriedade) e **alta correlação** (muita relação), retira a coluna/limpa.
* **Modeling**;
* **Evaluation**;
* **Deployment.**

**Passo a passo**

**Data Cleaning**

* Remoção/Preenchimento campos nulos/brancos, dados duplicados, outliers (fora do padrão);
* Verificar campos negativos e/ou < 0 (dependendo da coluna);
* Excluir colunas que não serão usadas.

**Data Transformation**

* Transformação. Exemplo: T-F -> 0-1
* Separação de classes e previsores
* Escalonamento (Escalas distintas);
* Table row.

**Data Discretization**

* Passar os dados continuo para dados numéricos;

Machine Learning – é favibilistico (passível a falhas)

Data Anotation – dar uma classificação, anotar os dados.

**Tipos de dados**

* Estruturados;
* Semi-Estruturados;
* Não estruturados.

**4 tipos de problemas em Data Science**

* **Classificação** (Class);
* **Clustering** (Agrupamento);
  + Agregas
    - Apriori (utilizado no caso da cerveja perto da fralda). Pode associar a um outro tipo de problema: **Association Rules.**
* **Regressão/Forescasting**.

**Regras Sazonais**: funcionam em determinado tempo do ano.

**Altíssima correlação**: Dados iguais. (Poderá ser apagado do data set)

**Baixíssima correlação**: Tende a dados aleatórios (Poderá ser apagado do data set)

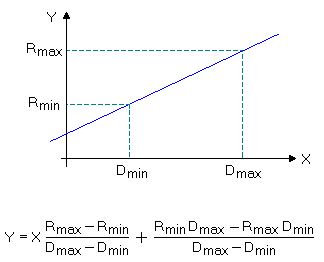
**Data Scaling**

Uma etapa de pré processamento de dados. Deixar os atributos na mesma escala, para quando usar algum algoritmo de **Distância Euclidiana** (é a distância entre dois pontos. Exem.: KNN), não seja dado maior importância para o dado que possui um valor maior.

Necessária quando variáveis ​​abrangem diferentes faixas.

Exemplos de Data Scaling:

**Range Scaling:** transforma os valores em outra faixa que geralmente inclui uma mudança e uma mudança de escala (ampliação ou redução). Equação:



**Mean Centering:** Média dos dados

**Standardization (Padronização):** A padronização (às vezes também chamada de dimensionamento automático ou transformação z ) é o procedimento de dimensionamento que resulta em uma média zero e variação de unidade de qualquer variável do descritor. Para cada valor de dados, a média μ deve ser subtraída e o resultado deve ser dividido pelo desvio padrão σ (observe que a ordem dessas duas operações não deve ser revertida):

Y = (X - μ) / σ

X = X – média(X) / desvio padrão (X)

**Normalization (Normalização)**:

X = X – mínimo(X) / máximo (X) – mínimo (X)

**Variáveis**

Qualitativa: características (Exe.: sexo, cor dos olhos, cabelos)

Ordinal: pertence a algo, atribui ordem de grandeza, pertence a uma ordem ou grupo

Quantitativa: falamos de números, quantidade de coisas.

**Amostra População**

Amostra (amostragem): tira uma parcela da população (Exe. Exame de sangue, tira um pouco o sangue)

**População finita**: dados que consegue contar (Exe. Alunos de uma escola)

**População infinita**: dados que não consegue contar, não tem um limite (Exe. Contar folhas da árvore. Durante o processo pode cair folhas e nascerem novas)

**Técnicas**

**Amostragem sistemática:** elabora um sistema para fazer uma amostra, faz uma regra. (Exe. Pegar todos os alunos que o número da chamada termina com 0). Não pode ser tendencioso.

**Amostragem casual simples:** Tendem ao aleatório, escolher as coisas sem uma “regra” (Exe. Fechar o olho e escolher uma linha).

**Amostragem estratificada**: uma determinada característica/variável é um extrato (porção). Uma amostra composta por elementos provenientes por todos os extratos da população.

**Amostra de conveniência**: são os dados que “eu tenho”, trabalha com o que já possui. Sabe que tem outros lugares, mas só pode usar determinada quantidade, o que já possui.