Universidade Federal da Paraíba

Centro de Informática

Departamento de Informática

Aprendizado Profundo

Pré-Processamento

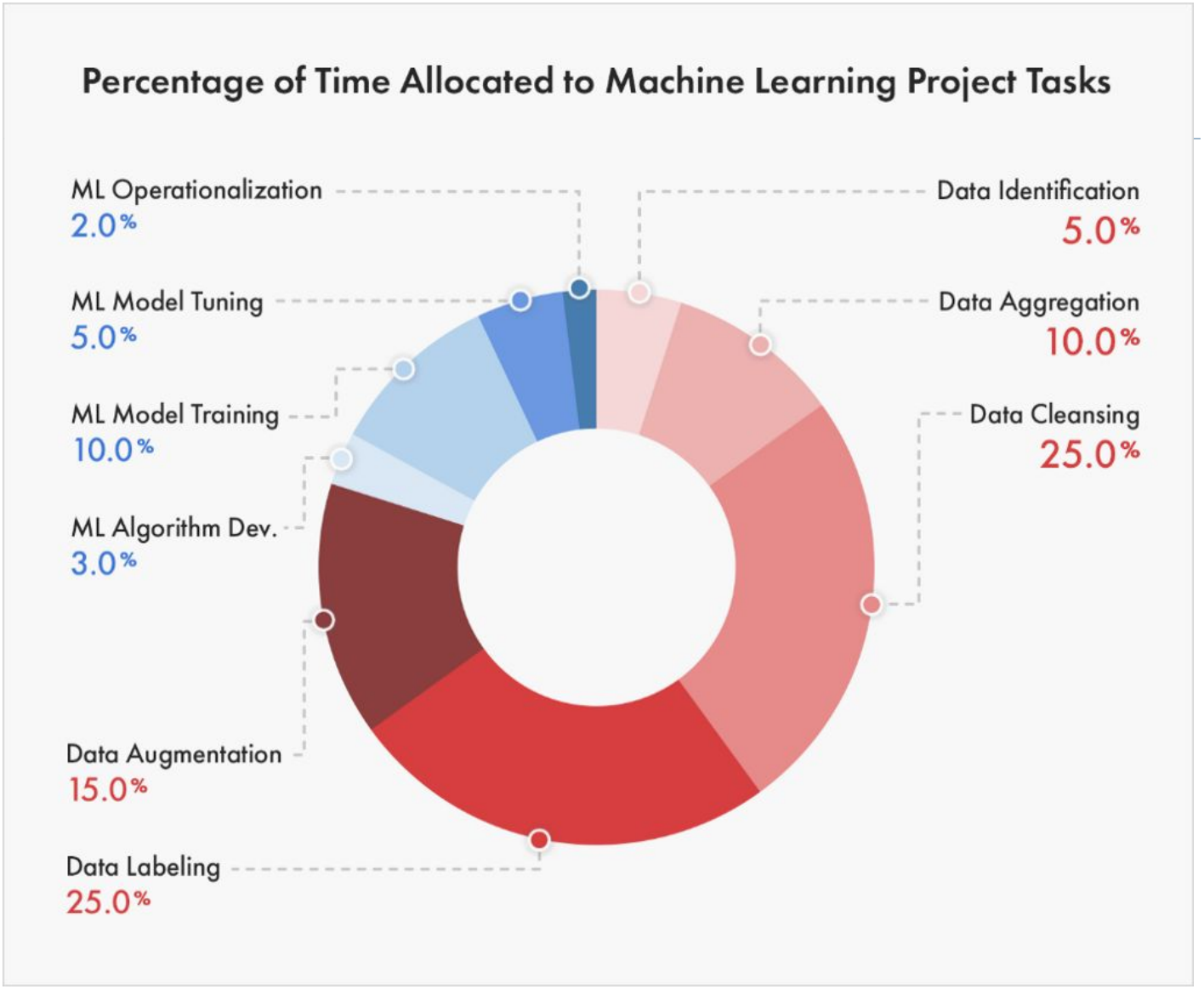
Tiago Maritan

Thaís Gaudêncio

Por que Pré-processar?

- **Dados não padronizados, redundantes ou fora de escala** podem levar a classificações/predições inesperadas;

- **Pré-processamento** corresponde a uma transformação nos dados antes de alimentar os algoritmos de aprendizagem de máquina.



Fonte: https://towardsai.net/p/machine-learning/top-10-reasons-why-87-of-machine-learning-projects-fail

Pré-processamento

- Eliminação manual de atributos
- Balanceamento dos dados
- Limpeza dos dados
- Redução da dimensionalidade
- Transformação dos dados

Eliminação Manual de Atributos

- Exclusão de atributo(s) que não contribuem para a estimativa do atributo alvo;
- Normalmente feito de acordo com a experiência de especialistas no domínio de dados.
- Ex: Para detectar se o paciente está doente ou saudável, o atributo **nome** (do paciente) pode ser irrelevante;

Eliminação Manual de Atributos

- Nesse exemplo, quais atributos vocês eliminariam?

	Identificador	Nome	Idade	Sexo	Peso	Manchas	Temperatura	Internacoes	Estado	Diagnostico
0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
5	2301	Ana	22	F	72	Inexistentes	38.0	3	RJ	Doente
6	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
7	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel

Balanceamento dos Dados

- Num modelo de machine learning, dados desbalanceados podem gerar “alarmes falsos“.

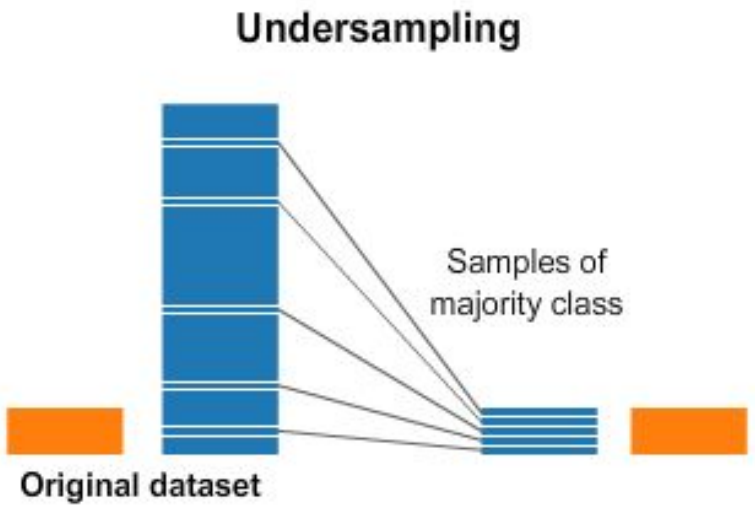
- O sistema responderia bem entradas para as classes majoritárias, mas terá um desempenho inferior para as minoritárias;

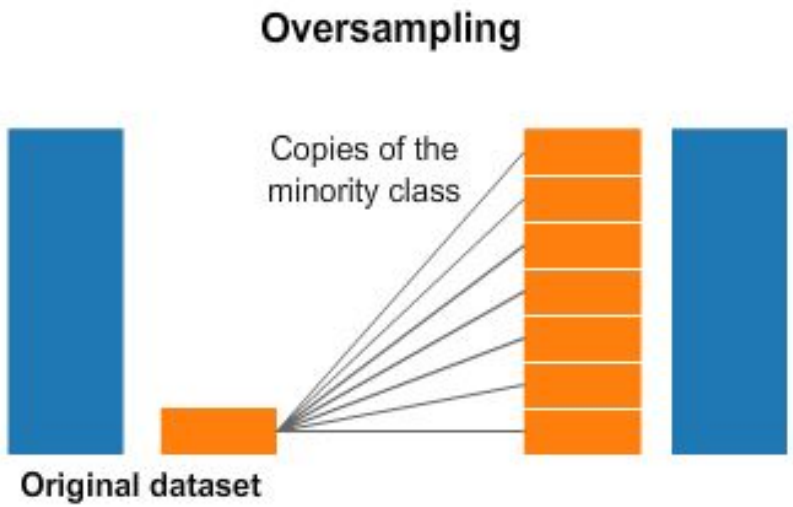
Balanceamento dos Dados

- Ex: Detecção de fraude com cartões de crédito
  - Número de transações financeiras normais é muito maior que o número de transações fraudulentas;
  - Se os dados não forem balanceados, um classificador tenderá a apresentar muitos falsos negativos
  - Situação indesejável para um banco, obviamente.

Técnicas de Balanceamento dos Dados

- Undersampling ou Oversampling





Limpeza de Dados

- Problemas relacionados à **qualidade dos dados**
  - **Dados ruidosos:** possuem erros ou valores diferentes do esperado
  - **Dados inconsistentes:** não combinam ou contradizem valores de outros atributos do mesmo objeto;

Limpeza de Dados

- Problemas relacionados à **qualidade dos dados**
  - **Dados redundantes:** dois ou mais atributos têm os mesmos valores para dois ou mais objetos
  - **Dados incompletos:** ausência de valores em parte dos dados

Limpeza dos Dados

- Exemplo:

Dados incompletos

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	4201	Joao	28	M	NaN	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67.0	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92.0	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43.0	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52.0	Uniformes	NaN	1	PE	Doente
5	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
6	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
7	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
8	1322	Marta	19	F	87.0	Espalhadas	39.0	6	AM	Saudavel
9	3027	Paulo	34	M	67.0	Uniformes	38.4	2	GO	Saudavel

Dados redundantes

Dados Incompletos

- Estratégias para tratar dados incompletos:
  - Eliminar os objetos com valores faltantes;
  - Preencher valores para os atributos com valores ausentes;
    - Ex: usando média, mediana ou moda dos valores conhecidos;
  - Empregar um indutor para estimar o valor do atributo.

Dados Inconsistentes

- Possibilidades!
  - **Problemas na anotação dos dados**
  - **Atributos de entrada não explicam o valor alvo.**

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	4201	Joao	28	M	67	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52	Uniformes	38.5	1	PE	Doente
5	4340	Claudia	21	F	52	Uniformes	38.5	1	PE	Saudavel
6	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
7	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
8	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
9	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
10	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel

Dados Redundantes

- **Redundância de atributos**
  - Ex: **idade** e **data de nascimento**

- **Alta correlação entre os atributos**
  - Os atributos trazem a mesma informação com relação ao alvo - mantenha apenas um!

- Uma forma de identificar dados redundantes é usando **Análise de Componentes Principais (PCA)**

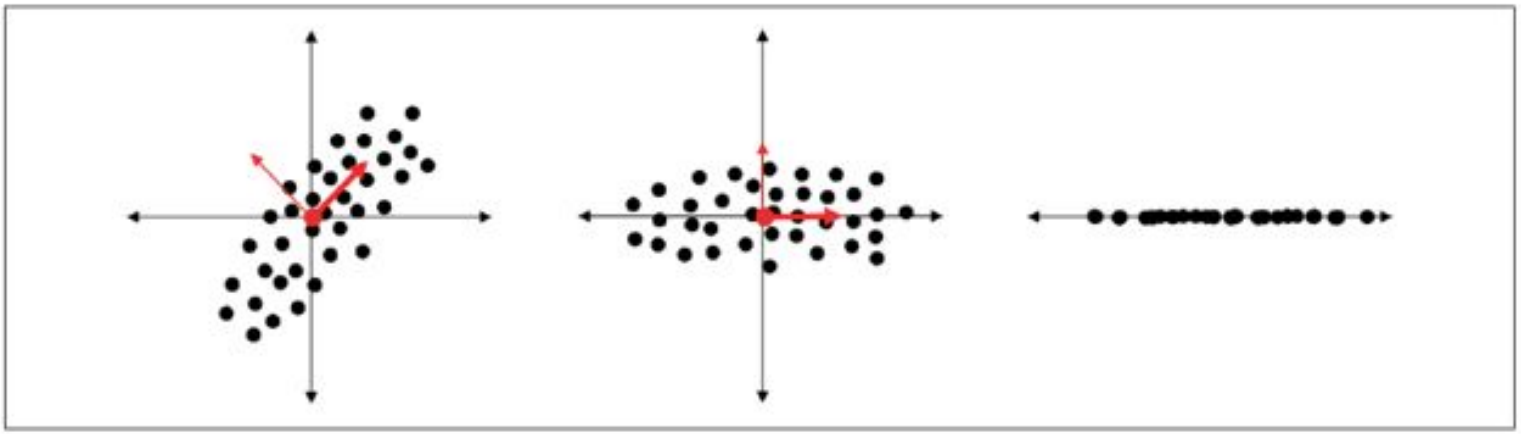
Análise de Componentes Principais (PCA)

- Definição mais formal:
  - Transformação ortogonal que converte um conjunto de **variáveis possivelmente correlacionadas** num conjunto de valores de **variáveis linearmente não correlacionadas**.

- Definição mais informal:
  - Identifica as **variáveis (eixos)** que contém a **informação mais relevante sobre os dados**.
  - Ex: se temos dados com **d dimensões (fatores)**, queremos encontrar um **novo conjunto de m < d dimensões** (fatores) que conserve as informações mais valiosas.

Análise de Componentes Principais (PCA)

- Como funciona?
  - 1º eixo é a direção (fator) que tem a maior variância.
  - A partir dele, pega a direção ortogonal que tem a 2ª maior variância... E assim por diante.
  - Continua o processo até encontrar **d direções (fatores)**.



Transformação de Dados (Numéricos)

- Quando **atributos possuem limites inferior e superior muito diferentes**, isso pode influenciar no desempenho do algoritmo de machine learning.

- Importante **normalizar os dados**

- **Normalização**: faz com que os atributos trabalhem com faixas de valores similares
  - Amplitude
  - Distribuição

Normalização

- **Reescala:** define uma nova escala de valores mínimo e máximo para todos os atributos

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Padronização:** define um **valor central (ex: média)** e um **valor de espalhamento (ex: desvio padrão)** para todos os atributos

$$X_{changed} = \frac{X - \mu}{\sigma}$$

Aumento de Dados
(*Data Augmentation*)

Aumento de Dados (*Data augmentation*)

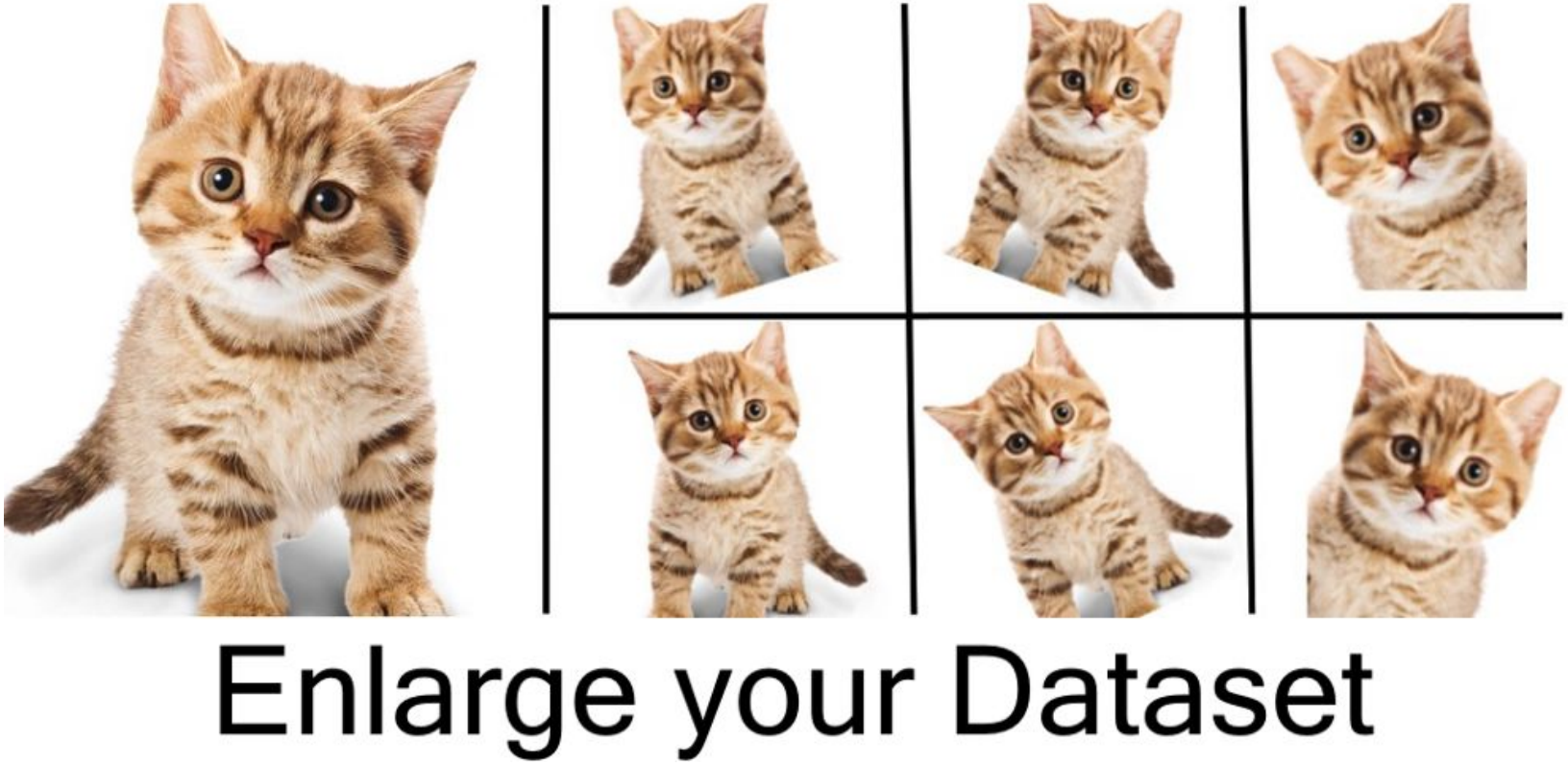
- Inflam artificialmente o conjunto de dados aplicando transformações nesses dados

- Tendem a deixar os modelos mais robustos.

- Exemplos:
  - Rotações e translações de objetos nas imagens;
  - Redimensionamento (*reescale*);
  - Cortes (*crops*)
  - Alterações de brilho e contraste,
  - Desfoque da imagem;

Aumento de Dados (*Data augmentation*)

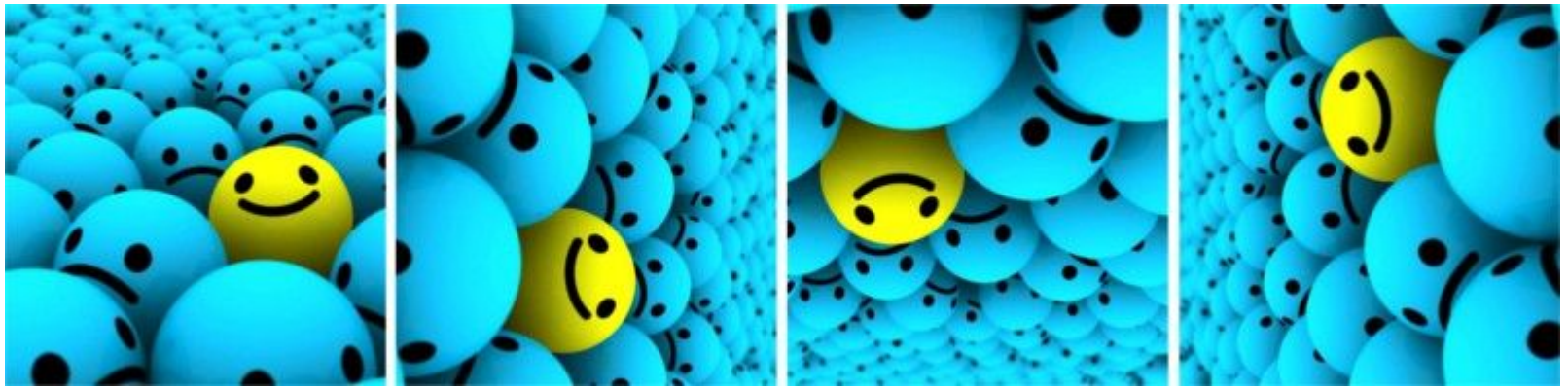
- Exemplo: Rotação



Aumento de Dados (*Data augmentation*)

- Ex: Translação

- Ex: *Flip* (Giro)



Aumento de Dados (*Data augmentation*)

- Ex: Reescala

- Ex: Crop

Aumento de Dados (*Data augmentation*)

- Ex: Ruído Gaussiano

Transferência de Aprendizagem

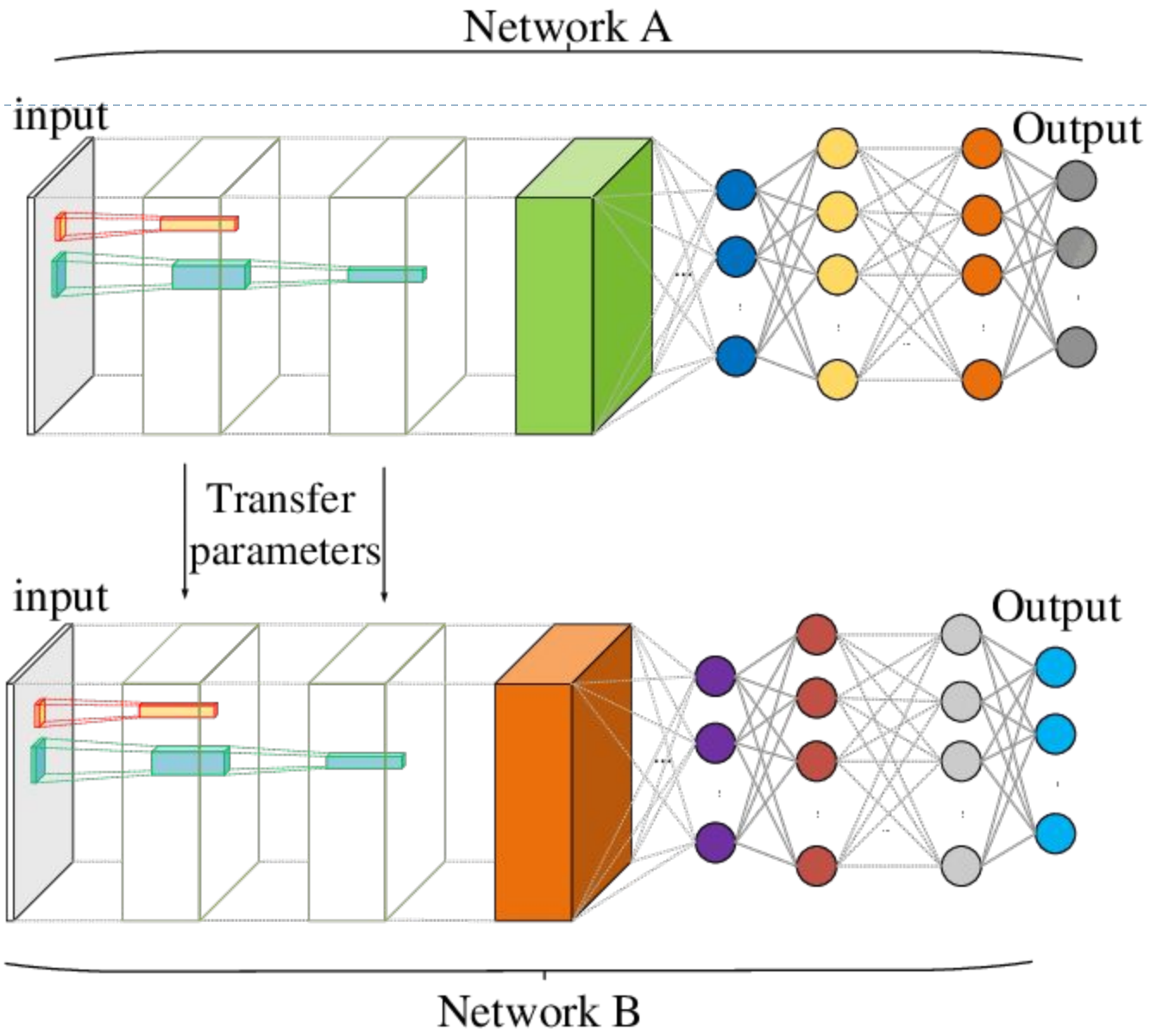
Transferência de Aprendizagem (*Transfer Learning*)

- Aproveita o **conhecimento adquirido** no treinamento de uma **outra tarefa similar**, que possui um conjunto suficientemente grande de dados.

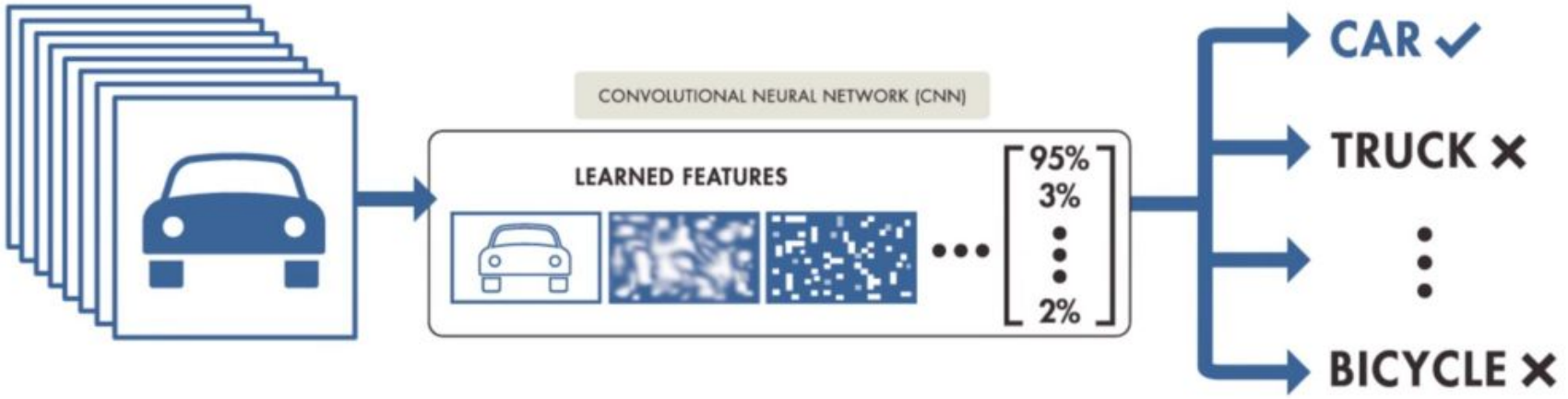
- Ex: **Reconhecimento de sinais em Libras**
  - É possível utilizar modelos pré-treinados em tarefas de reconhecimento de gestos e ações;
  - Ex: Base de dados Kinetics (Kay et al., 2017) contém:
    - 240 mil vídeos para o reconhecimento de ações humanas;
    - 400 classes de ação com 400 vídeos por classe;

Transferência de Aprendizagem (*Transfer Learning*)

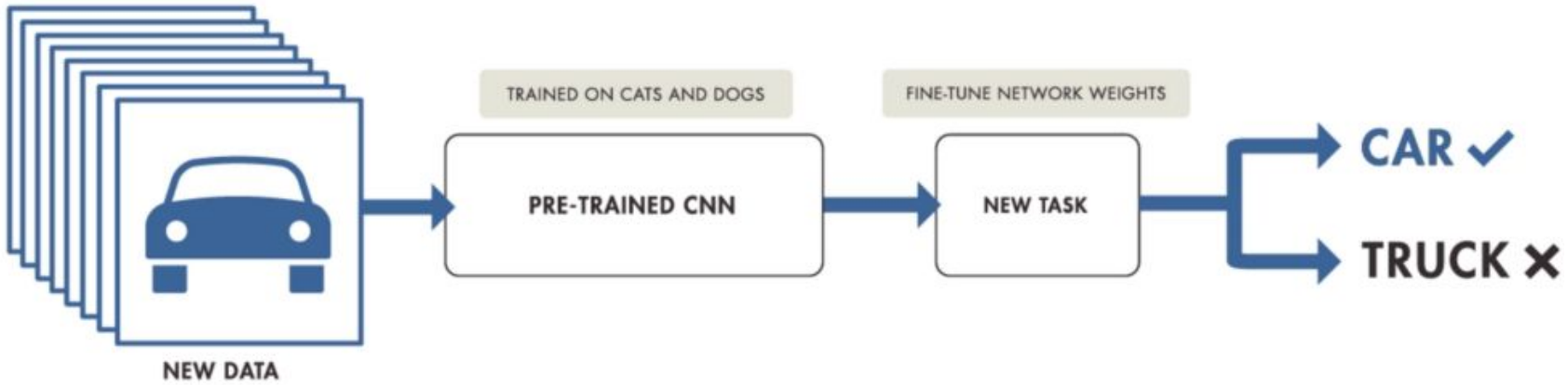
- Como funciona?
  - Usam-se os pesos sinápticos de um modelo pré-treinado;
  - **Congelam-se os pesos das primeiras camadas** (informações hierarquicamente menos abstratas)
  - **Treinam-se os pesos das últimas camadas** (informações mais abstratas)


Network A
input
Output
Transfer
parameters
input
Output
Network B

TRAINING FROM SCRATCH



TRANSFER LEARNING



Universidade Federal da Paraíba

Centro de Informática

Departamento de Informática

Aprendizado Profundo

Pré-Processamento

Tiago Maritan

Thaís Gaudêncio