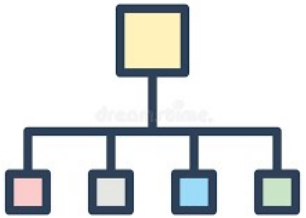
SIN5023

Reconhecimento Sintático e Estrutural de Padrões

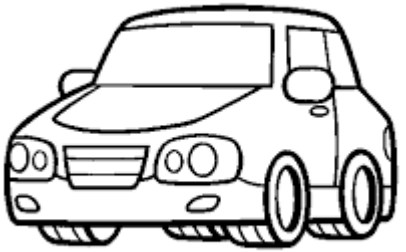
Professora:
Ariane Machado Lima



Objetivo

- Que o aluno compreenda os conceitos teóricos envolvidos no reconhecimento sintático e estrutural de padrões e seja capaz de aplicá-los em problemas práticos.
- Tais conceitos envolverão noções gerais de Grafos e Reconhecimento de Padrões, e grande ênfase em Linguagens Formais, tanto do ponto de vista determinístico quanto estocástico.
- Serão abordados exemplos de problemas práticos em várias áreas como Bioinformática, Processamento de Linguagem Natural e Visão Computacional.

Justificativa



- Padrões sintáticos e/ou estruturais são comuns em várias áreas de aplicação. Exemplos:
  - caracterizações de famílias de genes e de proteínas no campo da **Bioinformática**
  - caracterização de linguagens ou estilos e correção de erros em **Processamento de Linguagem Natural**
  - detecção de componentes e classificação de objetos no campo de **Visão Computacional**.
- Para a caracterização estrutural adequada desses problemas são necessárias técnicas específicas voltadas a esse fim, as quais serão abordadas nesta disciplina.

Conteúdo

- Noções básicas de reconhecimento de padrões: aprendizado supervisionado x não supervisionado, principais técnicas e métricas de estimação de desempenho.
- Noções básicas de linguagens formais, hierarquia de Chomsky, gramáticas, inferência gramatical e analisadores sintáticos. Modelos estocásticos: modelos ocultos de Markov, gramáticas estocásticas, análise e estimação de parâmetros.
- Noções básicas de grafos, grafos and-or.
- Dispositivos adaptativos.

De que conhecimento prévio vamos precisar?

De que conhecimento prévio vamos precisar?

- Quase nada

De que conhecimento prévio vamos precisar?

- Quase nada
- Conceitos básicos de probabilidade ajudarão

Aplicativo de apoio à disciplina

- Usaremos o sistema edisciplinas para
  - Repositório do material das aulas
  - Troca de mensagens / Fórum de discussões
  - Submissão de trabalhos
  - Provas Online

- Basta se cadastrar em https://edisciplinas.usp.br que a disciplina aparece para você (inscrição automática)

Avaliação

- 4 apresentações de artigos (15 min)
  - MA: Média de apresentações (ver próximo slide)
- 2 provas online – provavelmente 30/10 e 11/12
  - Sexta e sábado – 2 dias inteiros (edisciplinas)
  - SEM consulta
  - MP = Média simples das notas das duas provas
- Nota = (MA + MP)/2
- Conceito (notas):
  - R (< 5), C (5,0 a 6,99), B (7,0 a 8,49), A (8,5 a 10)

Avaliação

- 4 apresentações de artigos (15 min)
  - Entregar vídeo de uma apresentação (algumas opções de artigos a escolher)
    - A ser disponilizada para todos
  - Na aula: apresentação de 3 vídeos por sorteio (sem repetir artigo) – aula talvez um pouco mais longa… (uns 30 min a mais – das 8:30 às 10:00h)
    - Cada um dá uma nota para cada apresentação/discussão (NC – nota dos colegas), inclusive eu (NP – nota da professora)
    - MAA = (NP + NCmédia)/2
  - Todos devem escolher durante a semana mais 3 apresentações que não foram apresentadas em aula para assistir e avaliar (cada apresentação terá uma NCmédia)
  - Se o aluno teve uma apresentação sua sorteada para apresentação em aula
    - MA = (2*MAA + Σ NCmédia das 3 outras apresentações)/5
    - Senão MA = (Σ NCmédia das 4 apresentações)/4

RASCUNHO de calendário (SUJEITO A ALTERAÇÕES)

Nr Aula	Data	Tema
1	21/08/20	Noções básicas de reconhecimento de padrões: aprendizado supervisionado x não supervisionado
2	28/08/20	Hierarquia de Chomsky, gramáticas regulares, autômatos e cadeias e modelos ocultos de Markov
3	04/09/20	Estimação de desempenho
	11/09/20	NÃO HAVERÁ AULA – Semana da Pátria
	18/09/20	NÃO HAVERÁ AULA
4	25/09/20	Inferência gramatical (de autômatos) / treinamento de HMMs
5	02/10/20	Artigos com aplicações
6	09/10/20	Apresentações de artigos/Livre de contexto (análise sintática, inferência, Modelos de covariância)
7	16/10/20	Artigos com aplicações
8	23/10/20	Apresentações de artigos/semana para estudo – não há material novo (???)
9	30/10/20	PROVA 1 / Sensibilidade a contexto: complexidade, dispositivos adaptativos
10	06/11/20	Artigos com aplicações
11	13/11/20	Apresentações de artigos/Sensibilidade a contexto: grafos AND-OR
12	20/11/20	Artigos com aplicações
13	27/11/20	Apresentações de artigos/ ???
14	04/12/20	semana para estudo – não há material novo
15	11/12/20	PROVA 2

Perguntas?

Tema 1

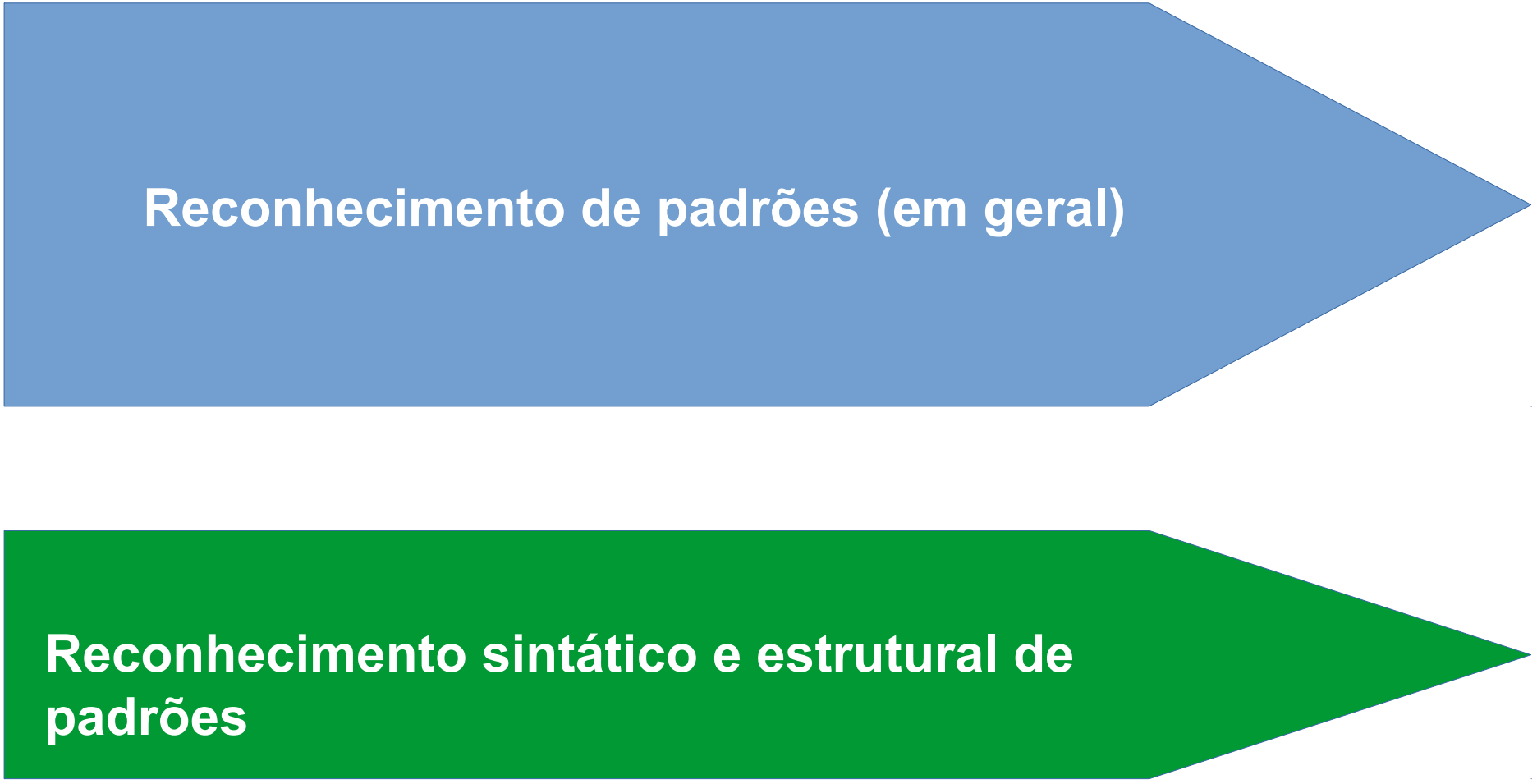
Introdução a Problemas de Reconhecimento de Padrões e Conceitos Básicos

Profa Ariane Machado Lima

Vídeo 1

Problemas de Reconhecimento de Padrões

Profa Ariane Machado Lima


Reconhecimento de padrões (em geral)
Reconhecimento sintático e estrutural de padrões

O que é um padrão?

O que é um padrão?

- “A descrição de um objeto”
- “Modelo oficial de pesos e medidas; qualquer objeto que serve de modelo à feitura de outro; desenho decorativo estampado em tecido ou outra superfície” *Aurélio*

O que é um padrão?

- “A descrição de um objeto”
- “Modelo oficial de pesos e medidas; qualquer objeto que serve de modelo à feitura de outro; desenho decorativo estampado em tecido ou outra superfície” *Aurélio*
- Um padrão representa um único objeto?

O que é um padrão?

- “A descrição de um objeto”
- “Modelo oficial de pesos e medidas; qualquer objeto que serve de modelo à feitura de outro; desenho decorativo estampado em tecido ou outra superfície” *Aurélio*
- Um padrão representa um único objeto?
- Um padrão representa uma população de objetos, um conceito, uma classe

O que é reconhecimento de padrões?

O que é reconhecimento de padrões?

- Problema de discriminar um objeto de entrada não dentre padrões individuais mas dentre populações, por meio da busca de atributos (mais ou menos) invariantes dentre os membros de uma população.

O que é reconhecimento de padrões?

- Problema de discriminar um objeto de entrada não dentre padrões individuais mas dentre populações, por meio da busca de atributos (mais ou menos) invariantes dentre os membros de uma população.

- Todos os atributos de um objeto são relevantes?

O que é reconhecimento de padrões?

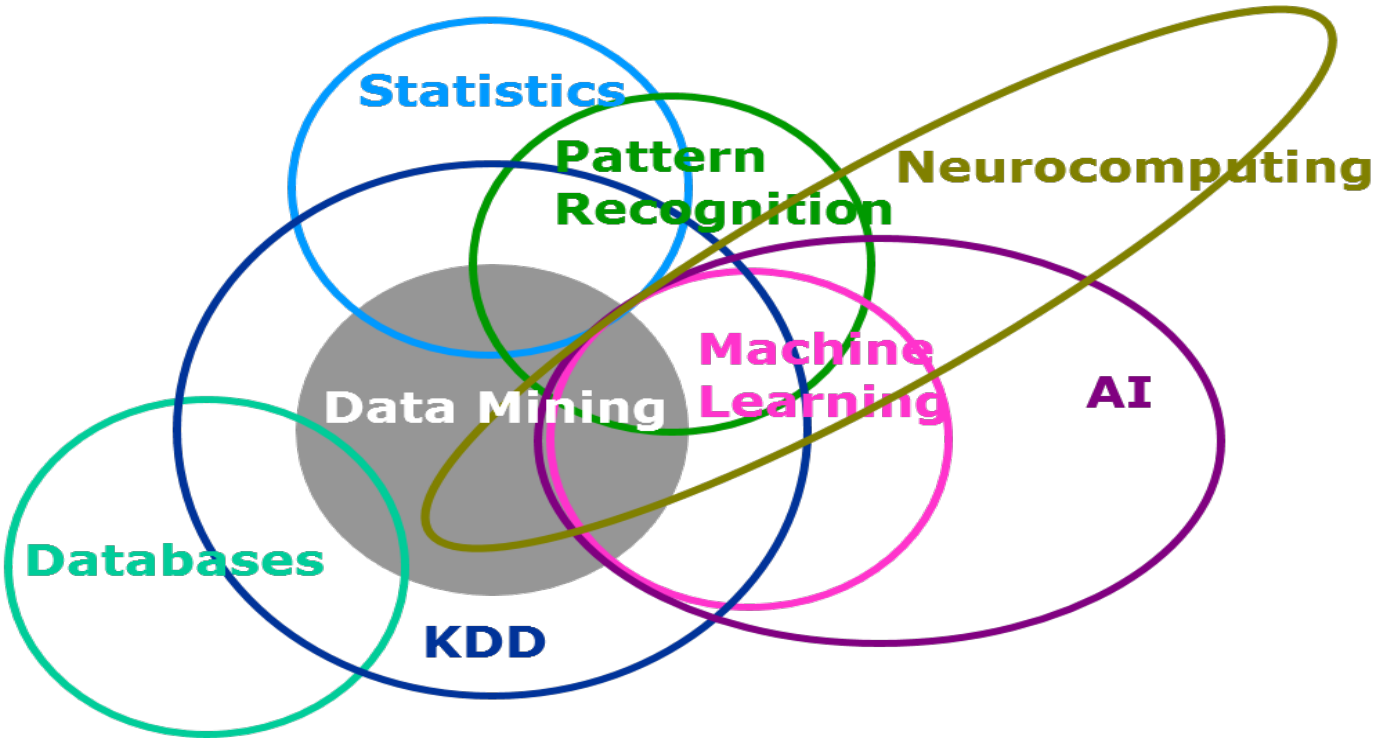
- Problema de discriminar um objeto de entrada não dentre padrões individuais mas dentre populações, por meio da busca de atributos (mais ou menos) invariantes dentre os membros de uma população.

- Todos os atributos de um objeto são relevantes? Normalmente não...

O que é reconhecimento de padrões?

- Categorização dos dados de entrada em classes identificáveis por meio da extração de atributos significantes dos dados, dentre muitos outros atributos irrelevantes.

Termos relacionados - Significados e limites não unânimes (abaixo só um exemplo, que eu nem concordo...)



https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/

Exemplos de problemas de reconhecimento de padrões

Exemplos de problemas de reconhecimento de padrões

- Reconhecimento de caracteres

Exemplos de problemas de reconhecimento de padrões

- Reconhecimento de caracteres
- O que é o padrão neste caso?

Exemplos de problemas de reconhecimento de padrões

- Reconhecimento de caracteres
- O que é o padrão neste caso?
  - Cada caracter?
  - Letras x algarismos?
  - Algarismos arábicos?
  - Alfabeto chinês, russo, árabe, etc...

O que são os padrões

- Reconhecimento de caracteres
- O que é o padrão neste caso?
  - Cada caracter?
  - Letras x algarismos?
  - Algarismos arábicos?
  - Alfabeto chinês, russo, árabe, etc...

- Padrões podem ser hierárquicos
- Os padrões dependem da aplicação
- O problema pode envolver 2 ou mais classes

O que são os padrões

- Reconhecimento de caracteres
- O que é o padrão neste caso?
  - Cada caracter?
  - Letras x algarismos?
  - Algarismos arábicos?
  - Alfabeto chinês, russo, árabe, etc...

- **Padrões podem ser hierárquicos**
- Os padrões dependem da aplicação
- O problema pode envolver 2 ou mais classes

O que mais?

- Teoria da Decisão
  - Tomada de decisões complexas
  - Sistemas de apoio à decisão
  - Tomar uma decisão é semelhante a fazer uma classificação
  - Podemos tomar decisões erradas
  - Há um custo para cada decisão a ser tomada
  - Queremos minimizar o custo total

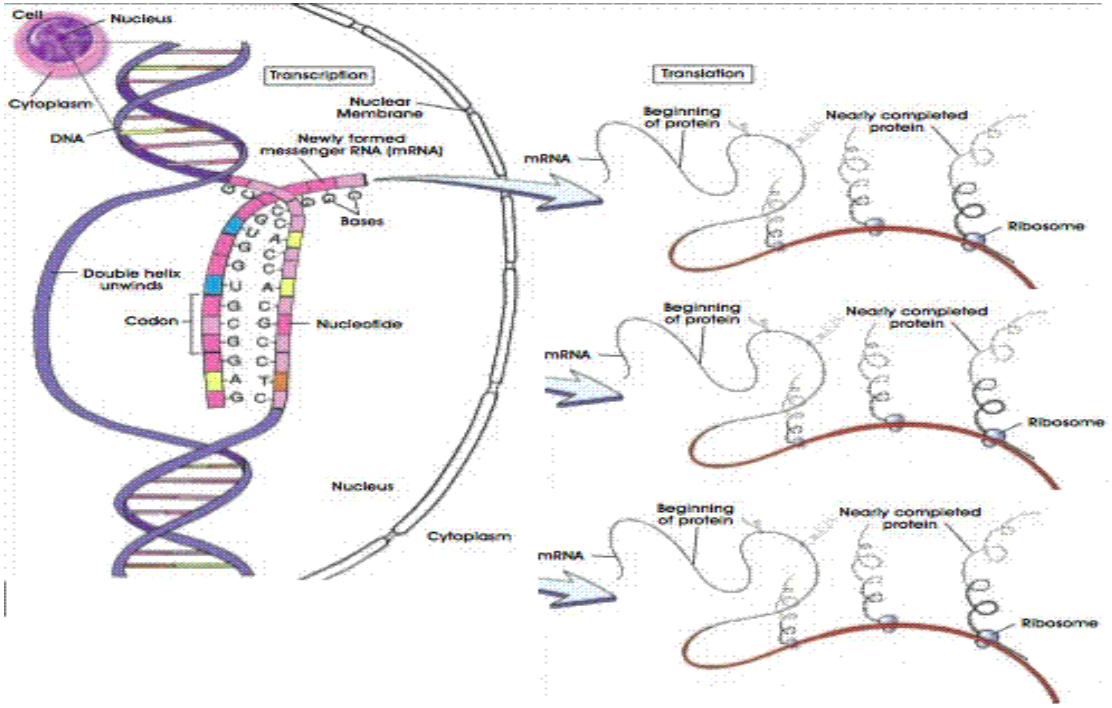
Mais exemplos de problemas de Reconhecimento de Padrões

Mais problemas

- Quais genes estão envolvidos em determinadas doenças?

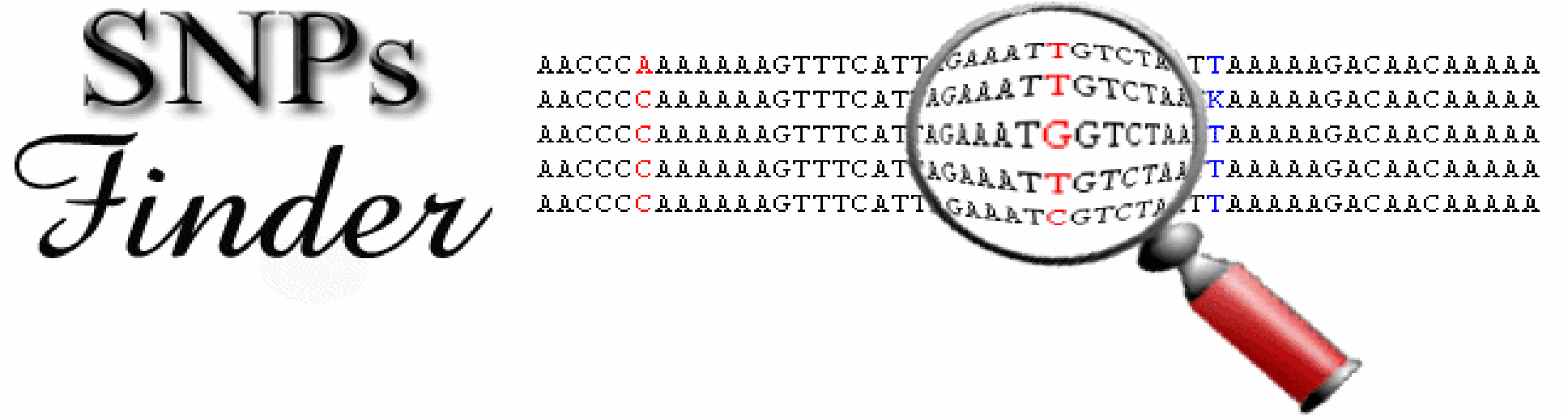
Mais problemas

- Quais genes estão envolvidos em determinadas doenças?
  - Olhando a expressão



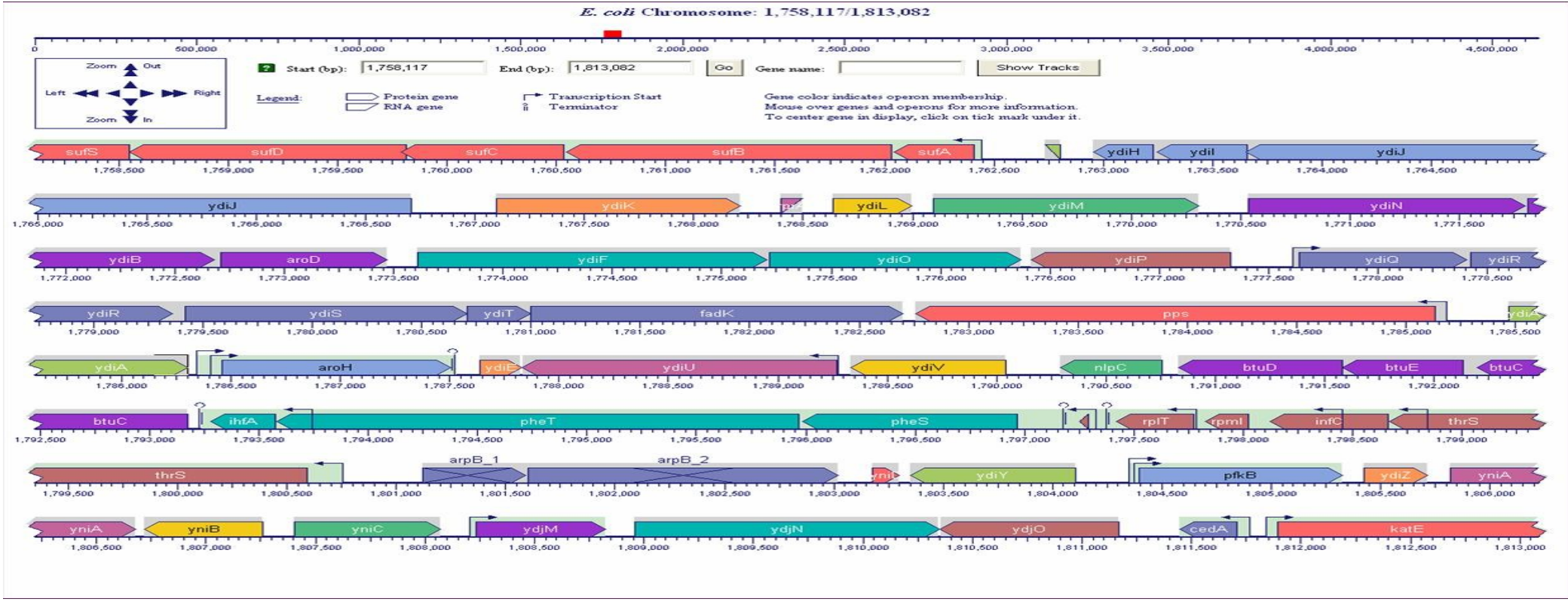
Mais problemas

- Quais genes estão envolvidos em determinadas doenças?
  - Olhando a sequência



Mais problemas

- Dado um genoma recém-sequenciado, como achar onde estão os genes?



Mais problemas

- Como classificar músicas por gênero musical?
- Como distinguir se um instrumento está afinado ou não?

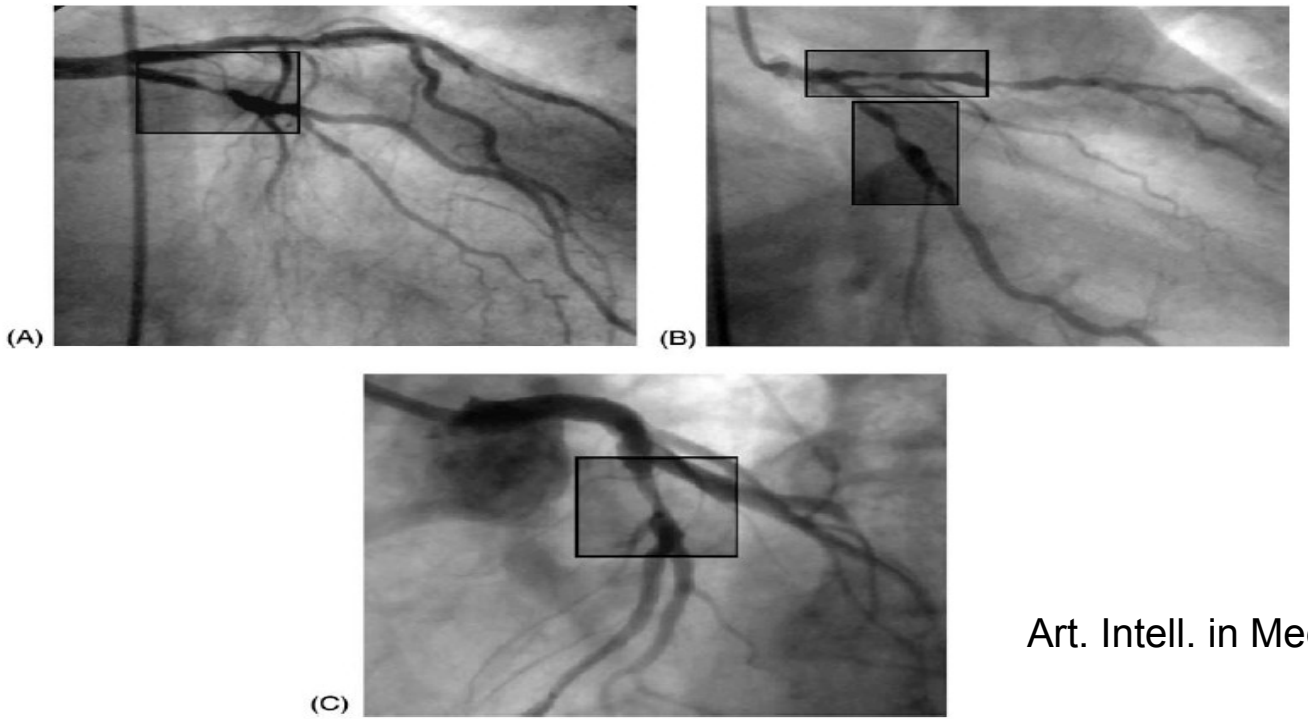
Mais problemas

- Sistemas de reconhecimento de voz

Free Up Your Other Hand With
Voice Recognition TV

Mais problemas

- Sistemas de reconhecimento de imagens:
  - Detecção de “defeitos biológicos” em imagens médicas (traumatismos, aneurisma, tumor, etc)

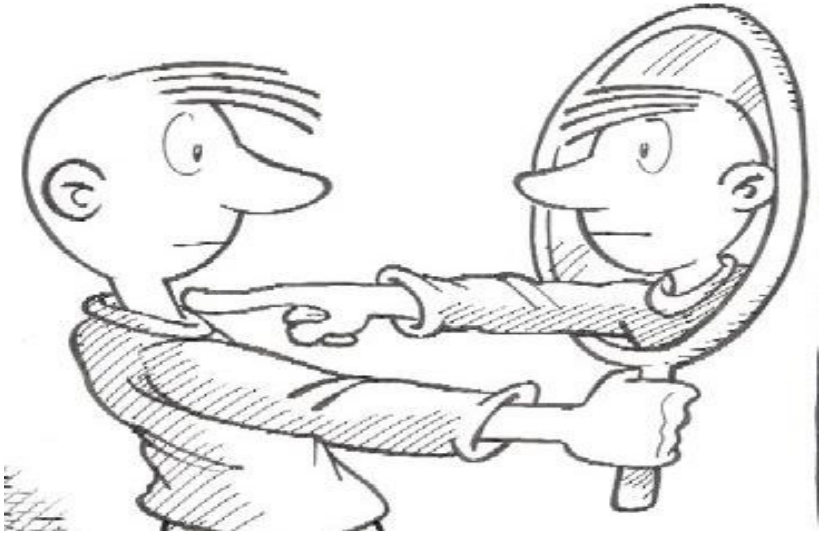


Art. Intell. in Med 26 (2002)

Fig. 1. (A–C) Coronographic images of coronary arteries with stenoses.

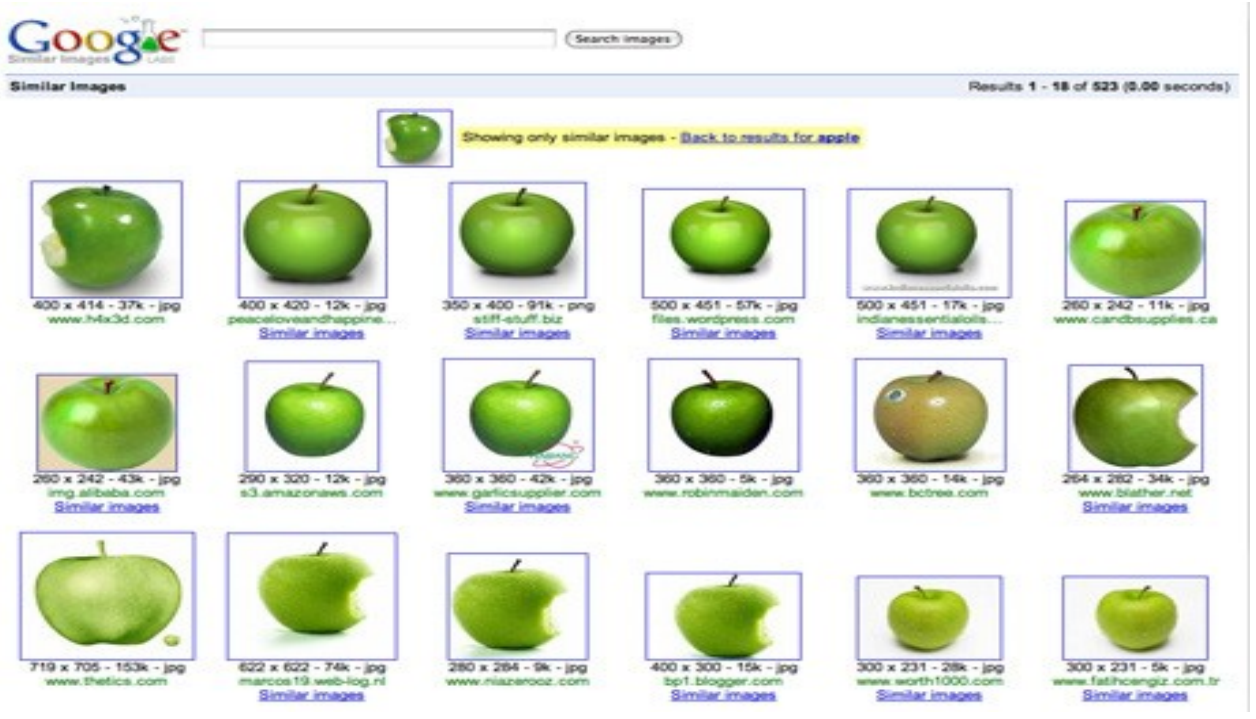
Mais problemas

- Sistemas de reconhecimento de imagens:
  - Reconhecimento de indivíduos em cenas criminais, retrato falado, etc



Mais problemas

- Sistemas de reconhecimento de imagens:
  - Busca por imagens



Mais problemas

- Detecção de fraude em cartões de crédito
- Comportamento de clientes (bons e maus)

Fim do vídeo 1

Problemas de Reconhecimento de Padrões

Profa Ariane Machado Lima

Vídeo 2

Conceitos básicos de reconhecimento de padrões

Profa Ariane Machado Lima

Tipos de classificação

- Binária: 2 classes. Exemplos:
  - RNA: codificante x não codificante
  - Diagnóstico: tem ou não uma doença
  - Tumor maligno/benigno
- Multiclasse: mais de 2 classes. Exemplos:
  - Famílias de RNAs não codificantes
  - Idioma de um texto
- Multirrótulo: há mais de 2 classes envolvidas, e cada instância pode pertencer a mais de uma
  - Diagnóstico com comorbidades

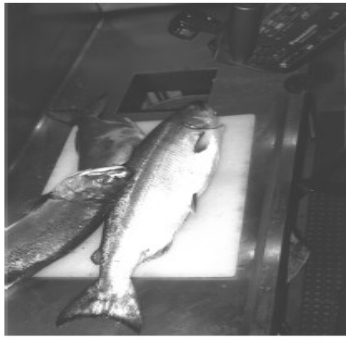
Para aprendermos mais, mais um exemplo...

- Robalos e salmões: precisam ser processados e embalados separadamente

- Fotografias

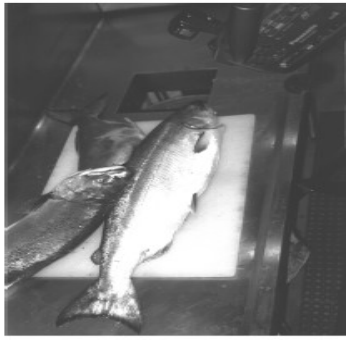
[DUDA, HART & STORK, 2001]

Para aprendermos mais, mais um exemplo...



- Objetivo:
  - Dada uma foto de um peixe, classificá-lo em 1 de 2 categorias possíveis: robalo ou salmão (classificação binária)
- Como podemos fazer isso?

Para aprendermos mais, mais um exemplo...



- Objetivo:
  - Dada uma foto de um peixe, classificá-lo em 1 de 2 categorias possíveis: robalo ou salmão (classificação binária)
- Como podemos fazer isso? (abordagem de reconhecimento de padrões clássico)
  - Escolhemos algum atributo (ou característica) e vemos se ele consegue diferenciar (discriminar) robalos e salmões
  - Queremos descrever um modelo para robalos e salmões, no caso f(x) <= l é um e f(x) > l é outro, x sendo a característica e l um limiar

x, por exemplo, comprimento

Tamanhos dos peixes

- Note que x (cada objeto, representado pelo seu tamanho) é uma variável aleatória

- Gostaríamos de conhecer a distribuição de x para robalos e salmões para podermos fazer a classificação (acharmos um limiar)
  - Distribuições condicionais: P(x | robalo) e P(x | salmão)

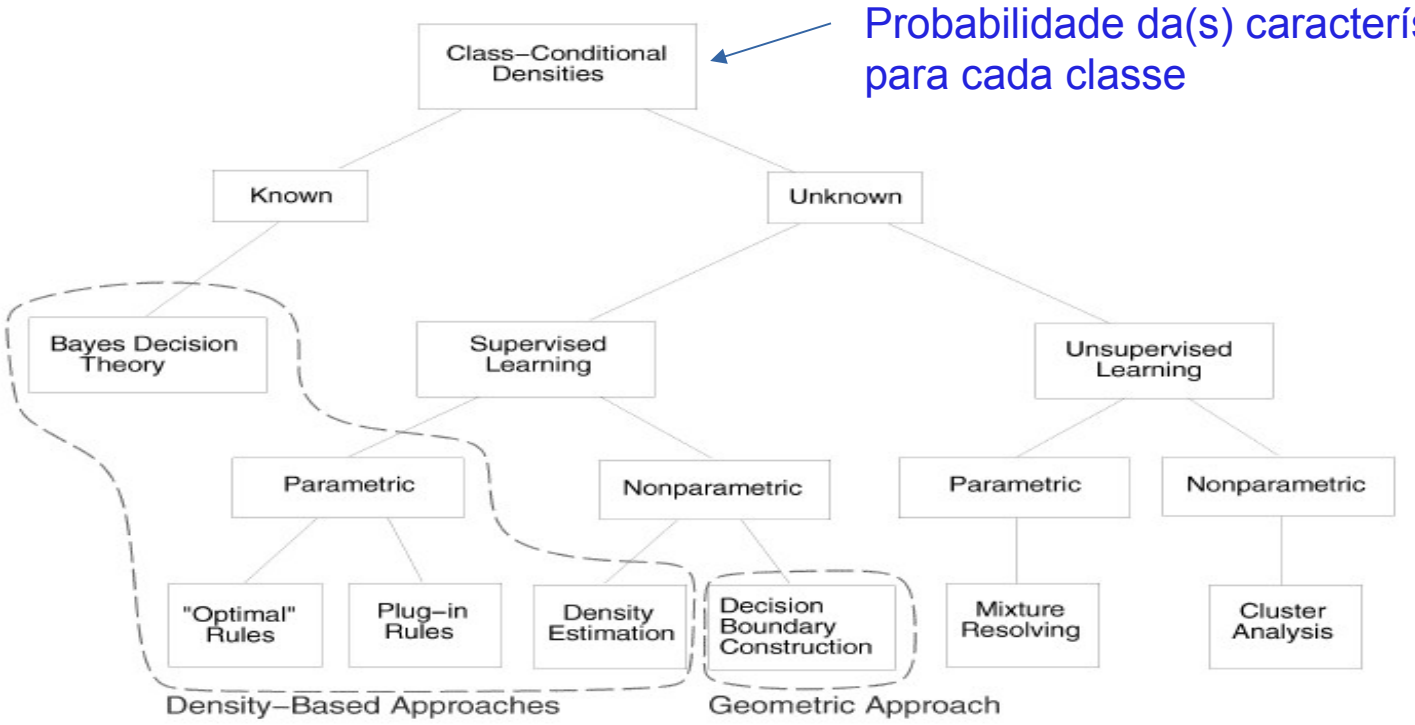
Aprendizado computacional

- O processo de criar um modelo, uma hipótese acerca do conceito real, é conhecido como aprendizado computacional
- Como aprendemos algo?
- Como poderíamos aprender o que é robalo e o que é salmão?

Aprendizado computacional

- O processo de criar um modelo, uma hipótese acerca do conceito real, é conhecido como aprendizado computacional
- Como aprendemos algo?
- Como poderíamos aprender o que é robalo e o que é salmão?
- Uma das formas é através de exemplos
- Aprendizado supervisionado:
  - Amostra de treinamento: exemplos rotulados de cada classe
- Aprendizado não supervisionado: não tenho exemplos, mas apenas os objetos a serem classificados

Métodos de Classificação



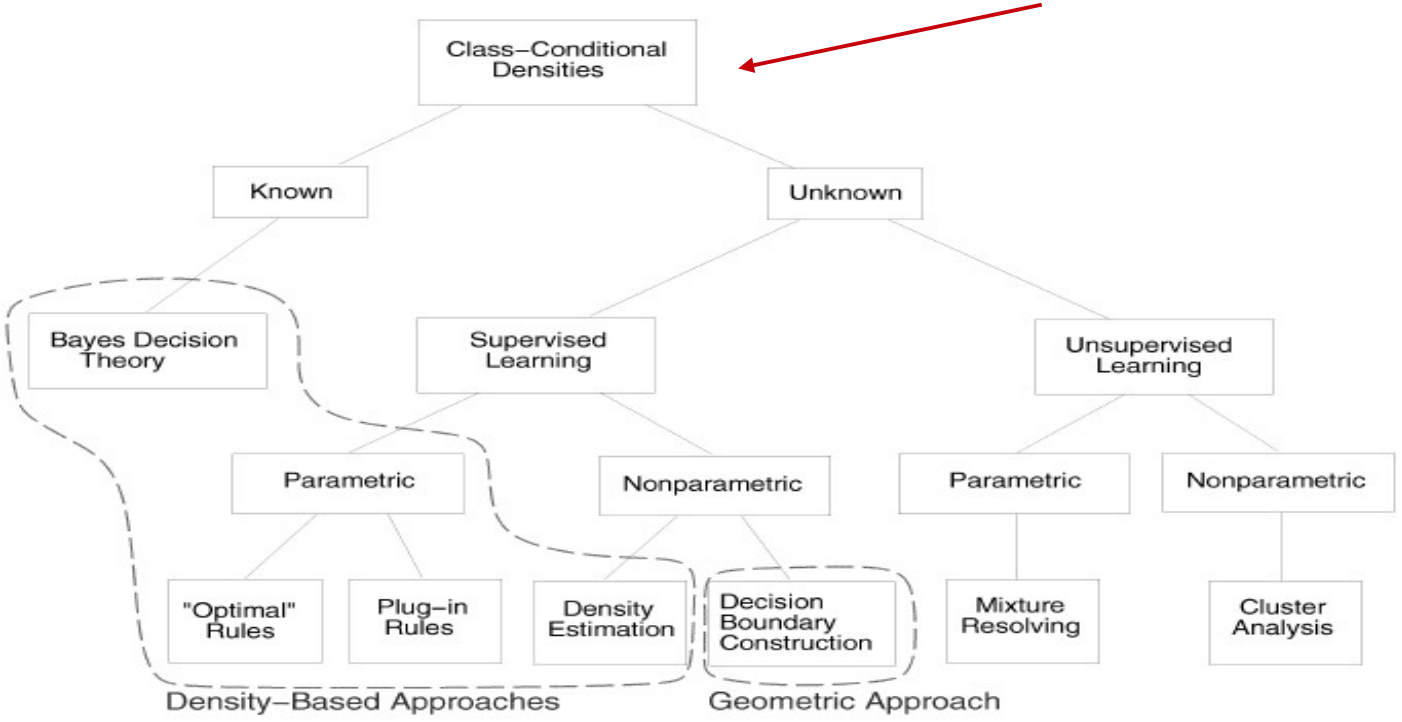
[JAIN et al, 2000]

Tamanhos dos peixes

- Note que x (cada objeto, representado pelo seu tamanho) é uma variável aleatória

- Gostaríamos de conhecer a distribuição de x para robalos e salmões para podermos fazer a classificação (acharmos um limiar)
  - Distribuições condicionais: P(x | robalo) e P(x | salmão)

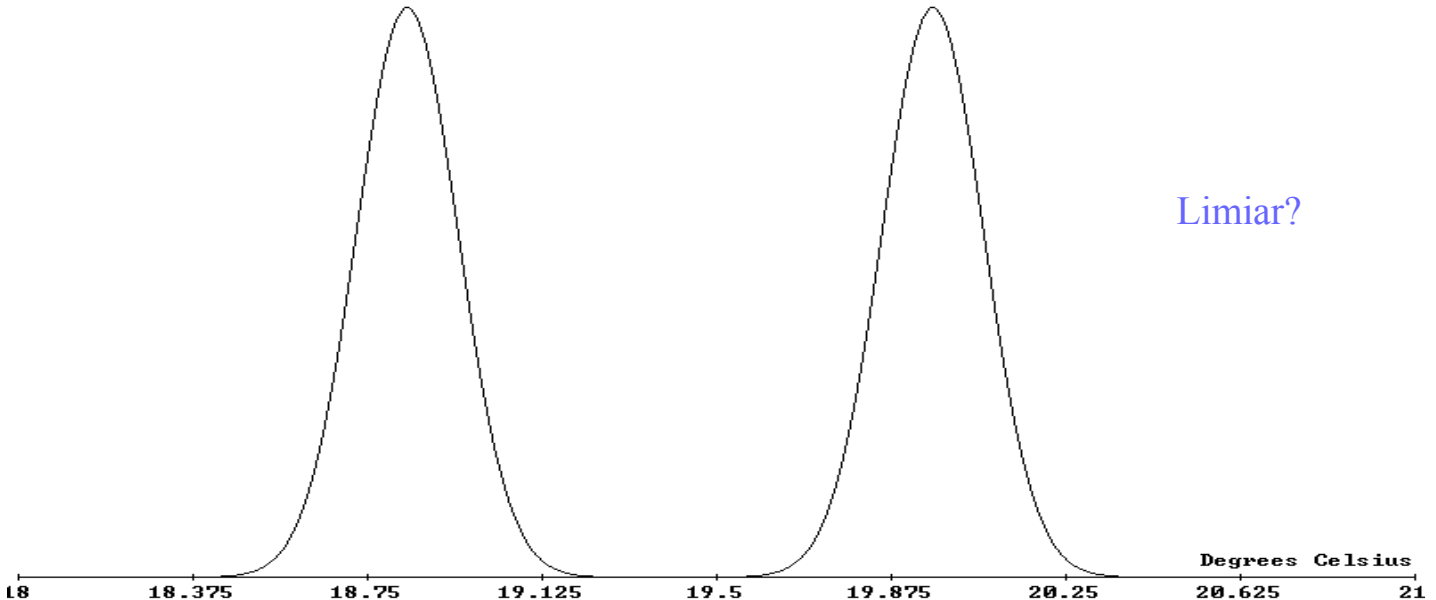
Métodos de Classificação



[JAIN et al, 2000]

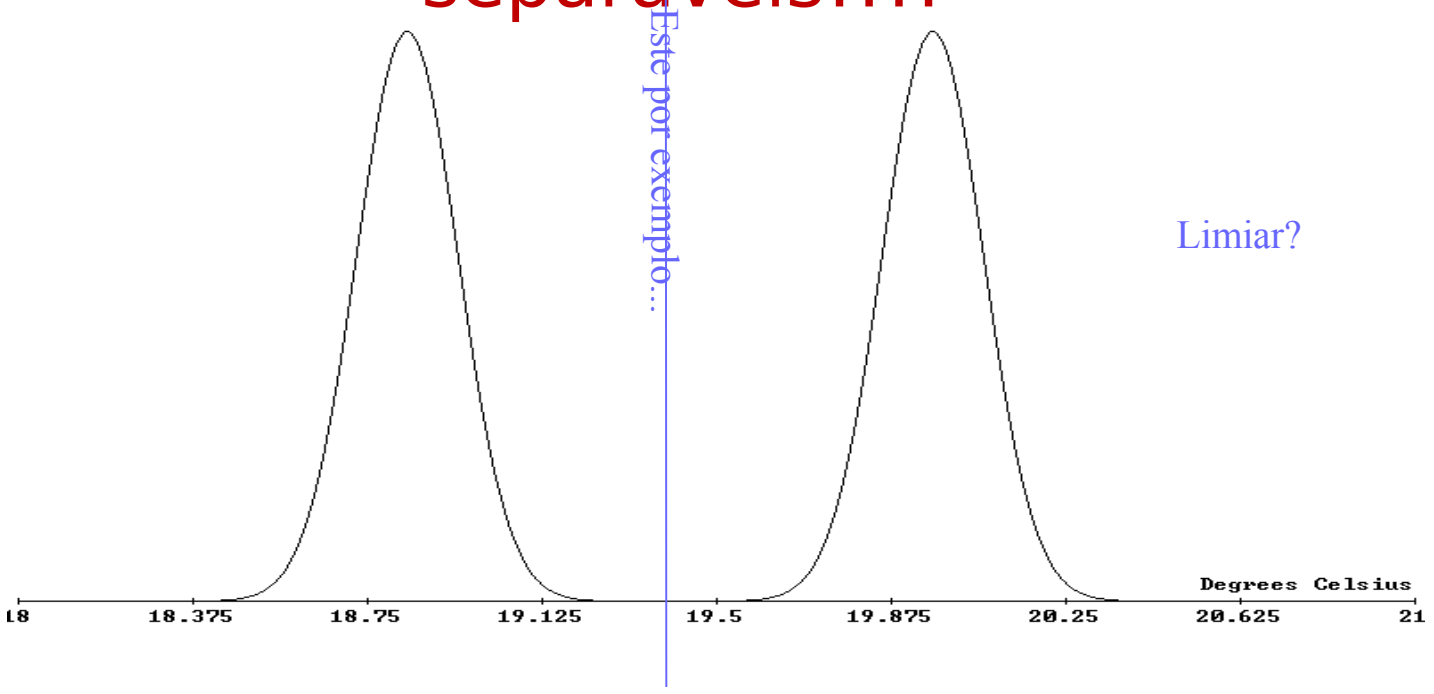
Caso Ideal: temos as condicionais reais (ex: Deus nos deu)

Caso ideal dos sonhos: e as classes são totalmente separáveis!!!!



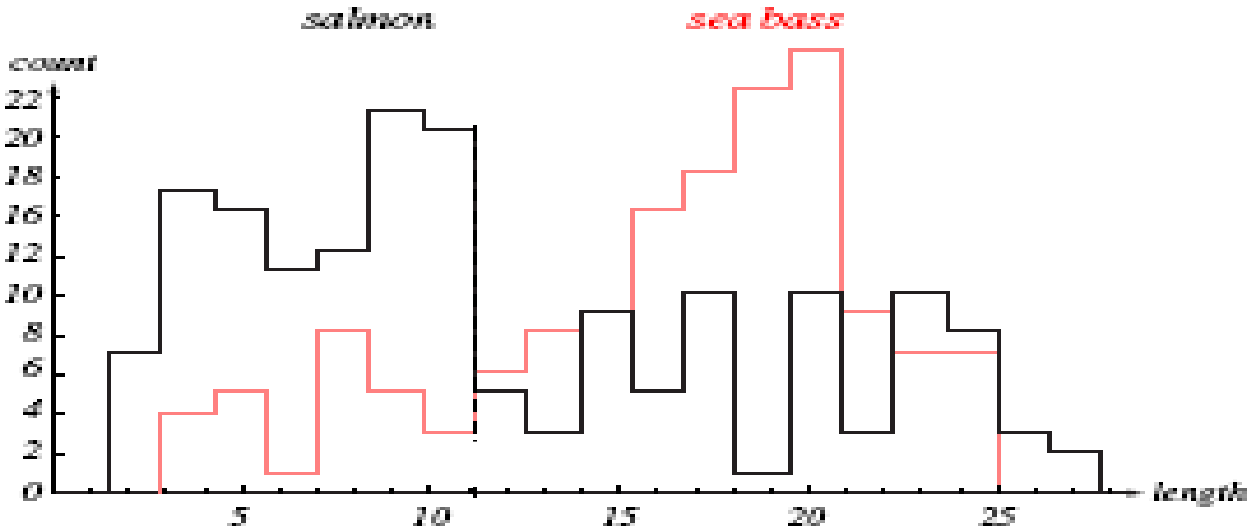
Caso Ideal: temos as condicionais reais (ex: Deus nos deu)

Caso ideal dos sonhos: e as classes são totalmente separáveis!!!!



Se não temos as condicionais (da população), vamos estimá-las (a partir de uma amostra de exemplos)... Veja esse histograma dos tamanhos dos peixes de exemplo...

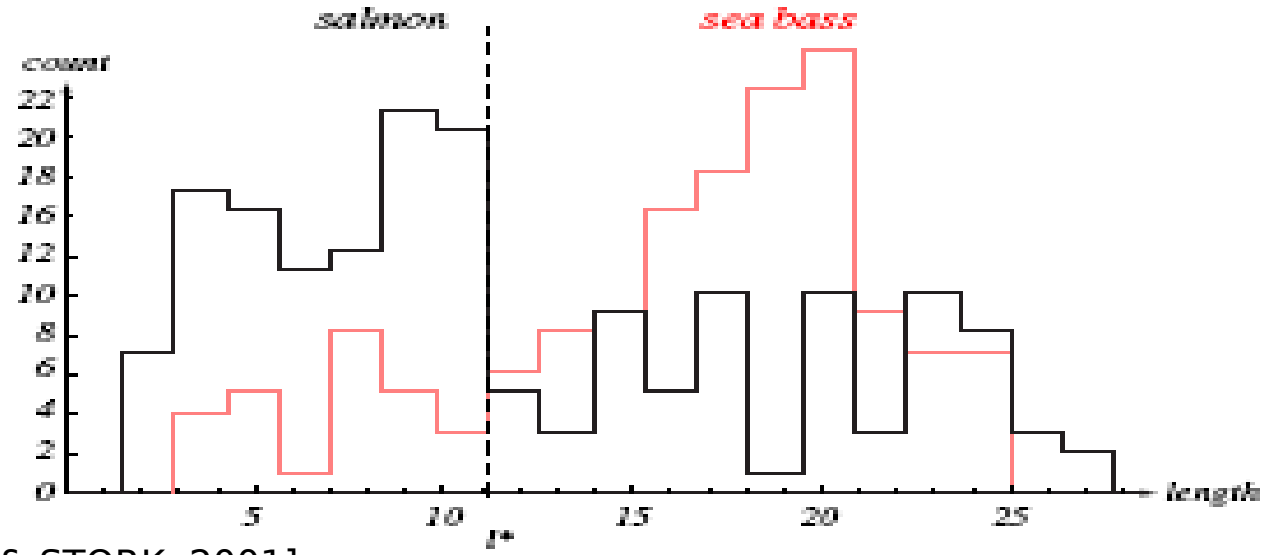
- Qual limiar escolhemos?



[DUDA, HART & STORK, 2001]

Se não temos as condicionais (da população), vamos estimá-las (a partir de uma amostra de exemplos)... Veja esse histograma dos tamanhos dos peixes de exemplo...

- Qual limiar escolhemos?

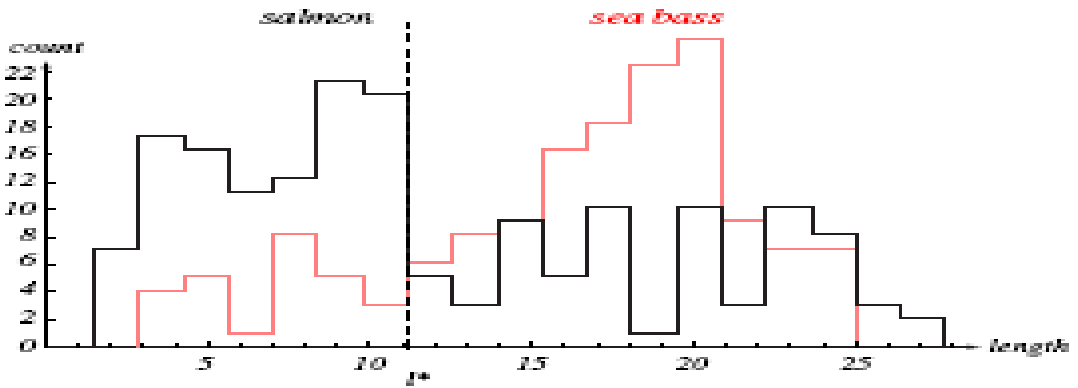


[DUDA, HART & STORK, 2001]

Histograma dos tamanhos dos peixes

- Qual limiar escolhemos?
- A cada limiar há um erro de classificação associado

[DUDA, HART & STORK, 2001]

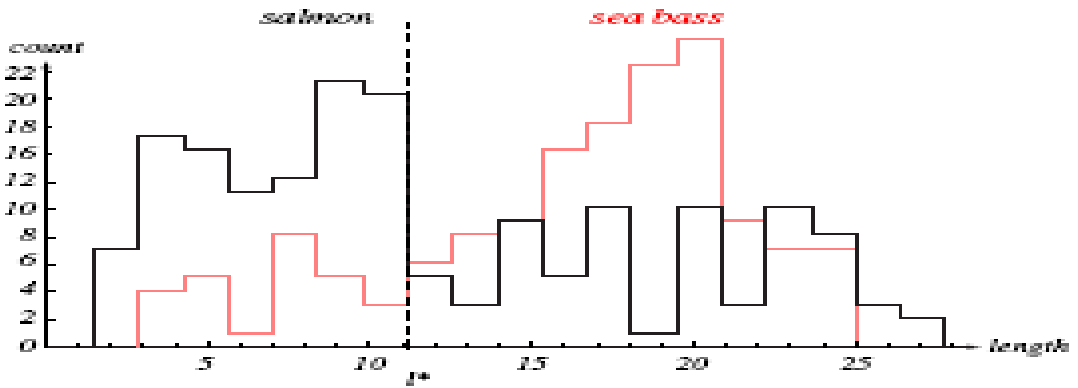


Histograma dos tamanhos dos peixes

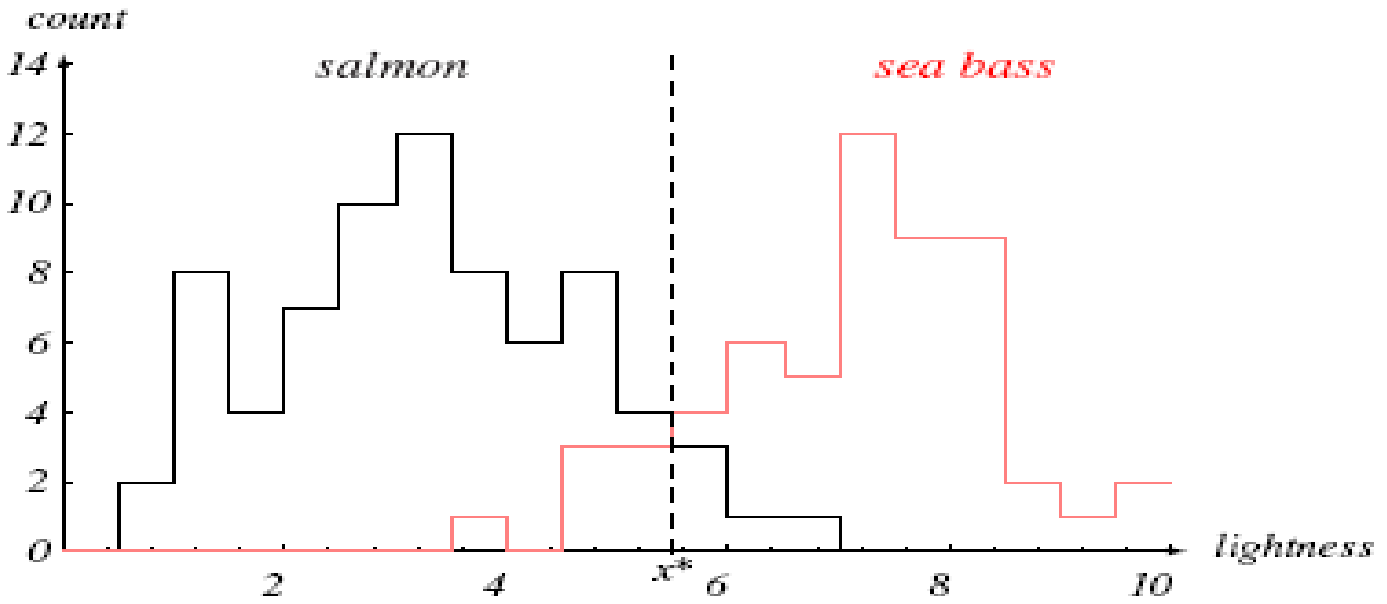
- Qual limiar escolhemos?
- A cada limiar há um erro de classificação associado

Tamanho não parece
ser uma boa característica...
Vamos escolher outra!

[DUDA, HART & STORK, 2001]

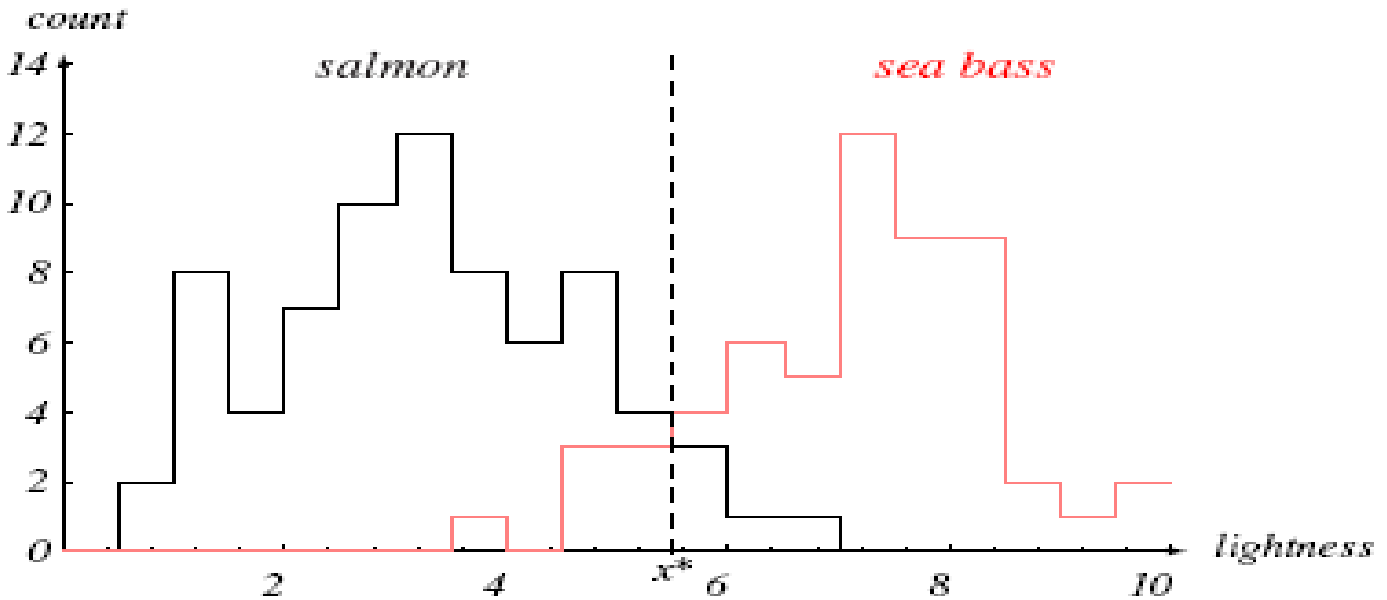


Histograma do Brilho



[DUDA, HART & STORK, 2001]

Histograma do Brilho



MELHOROU!

[DUDA, HART & STORK, 2001]

Podemos melhorar?

- E se combinássemos ambas?

Podemos melhorar?

- E se combinássemos ambas?
- $\mathbf{x} = (x_1, x_2)^T$ no qual:

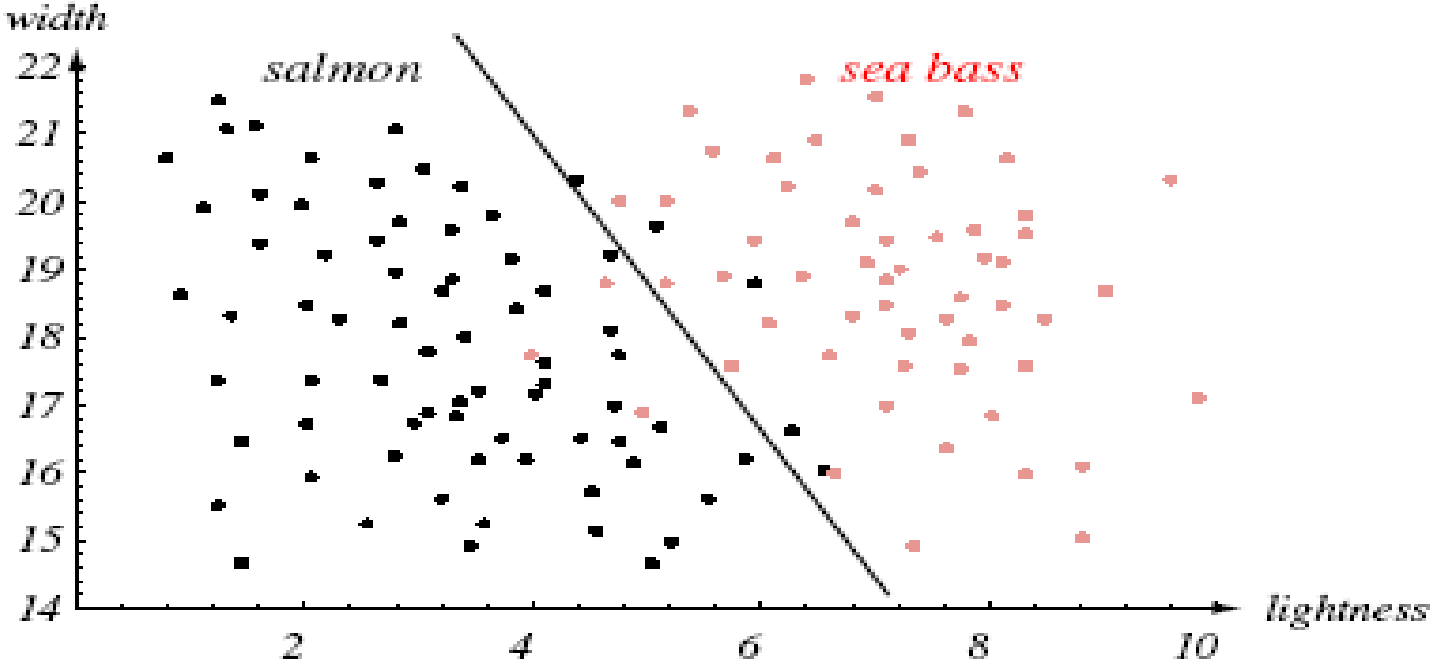
  x_1 é o comprimento (em cm)

  x_2 é a intensidade de brilho
- **x** é um vetor de características de um objeto (em um espaço de características bidimensional) – note o negrito para vetor
- x_1 e x_2 são variáveis aleatórias
- **x** é um vetor aleatório

Tamanho e brilho

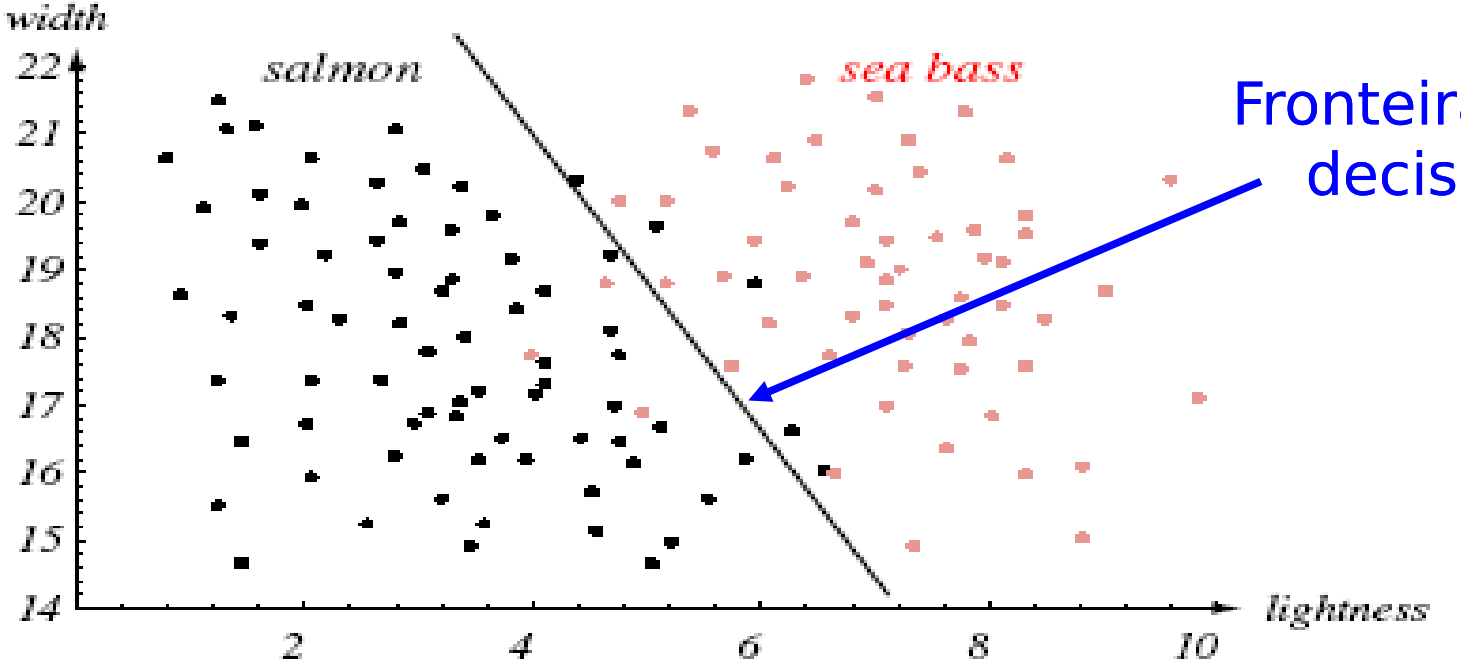
- Como representar?

Tamanho e brilho



[DUDA, HART & STORK, 2001]

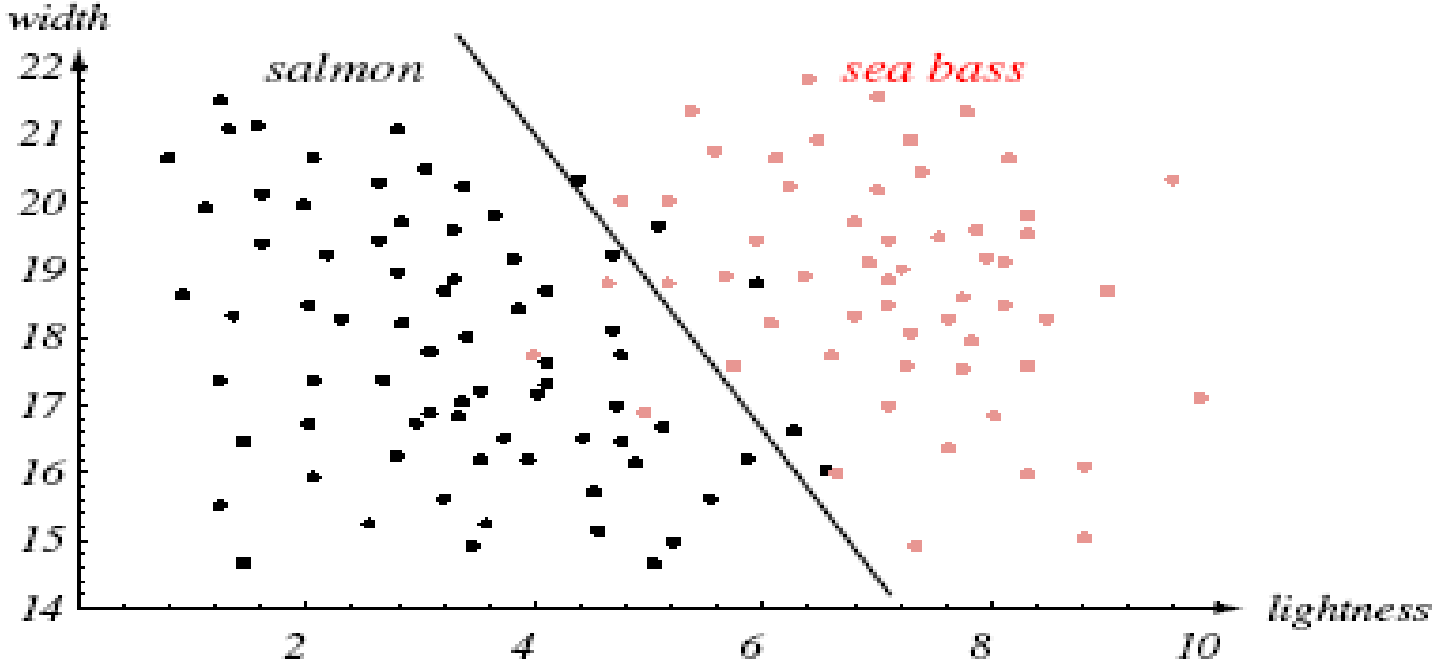
Tamanho e brilho



Classificador linear: a fronteira de decisão é um hiperplano (curvatura nula no espaço n-dimensional)

[DUDA, HART & STORK, 2001]

Tamanho e brilho



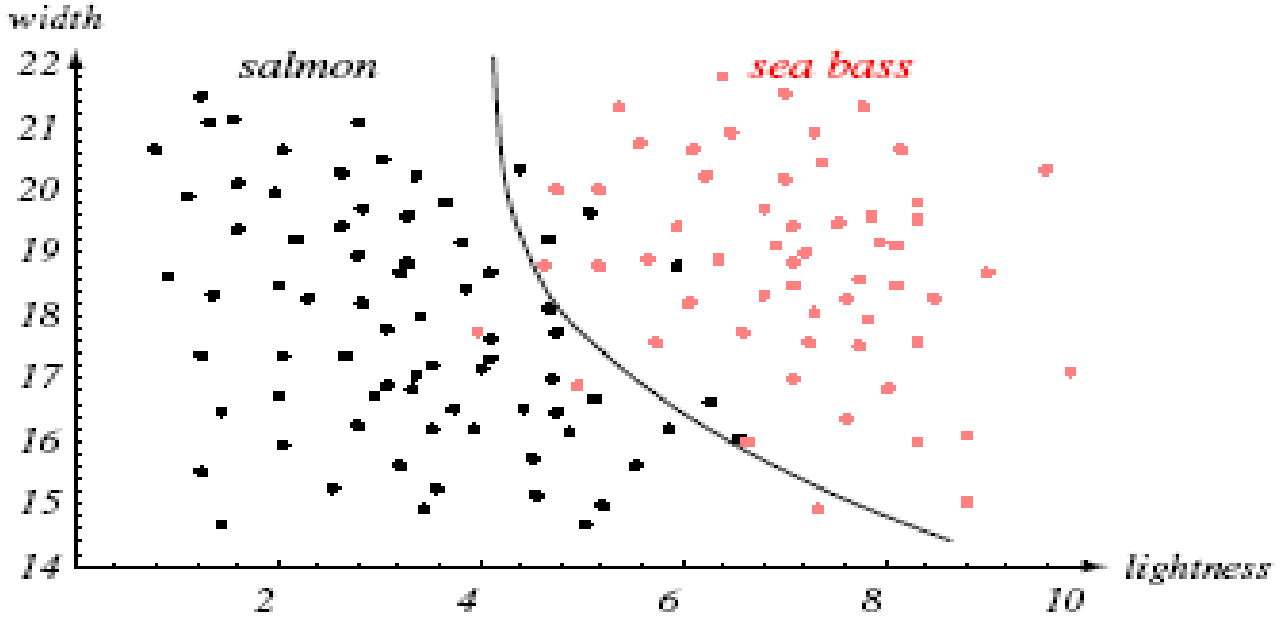
Podemos “desenhar” essa fronteira de uma forma diferente?

[DUDA, HART & STORK, 2001]

Tamanho e brilho

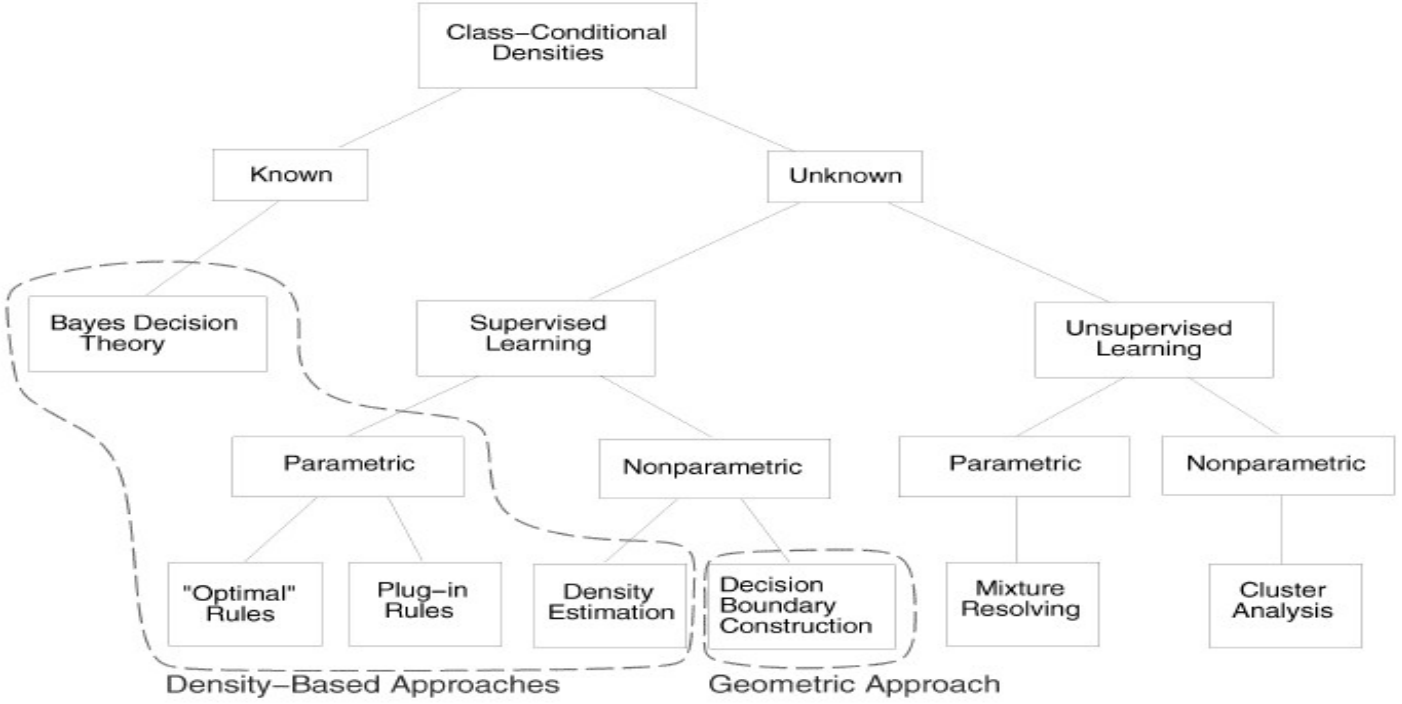
- Que tal assim?

Classificador não é mais linear



[DUDA, HART & STORK, 2001]

Classificação supervisionada x não-supervisionada



[JAIN et al, 2000]

Classificação

- Classificação supervisionada
  - Número de classes definido
  - Amostra de treinamento (exemplos rotulados com a classe a qual pertence)
  - Duas etapas:
    - Aprendizado (ou treinamento)
    - Reconhecimento

Classificação supervisionada

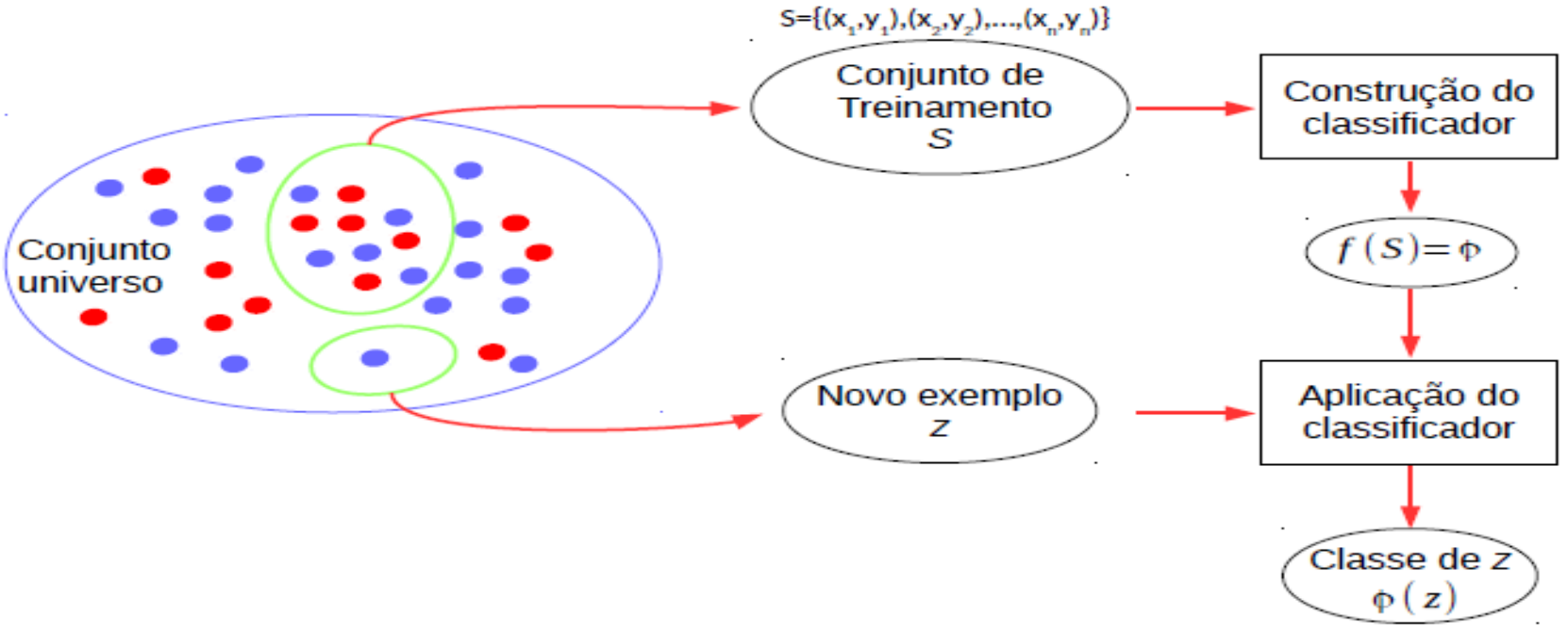


Figura: Representação do processo de aprendizado supervisionado

Classificação - observações

- Um mesmo conjunto de dados ⇒ diferentes classificadores
  - Por quê?

Classificação - observações

- Um mesmo conjunto de dados ⇒ diferentes classificadores
  - Diferentes características
  - Diferentes técnicas de construção de classificadores
  - Diferentes parâmetros

Classificação - observações

- Um mesmo conjunto de dados ⇒ diferentes classificadores
  - Diferentes características
  - Diferentes técnicas de construção de classificadores
  - Diferentes parâmetros

- Qual é o melhor?

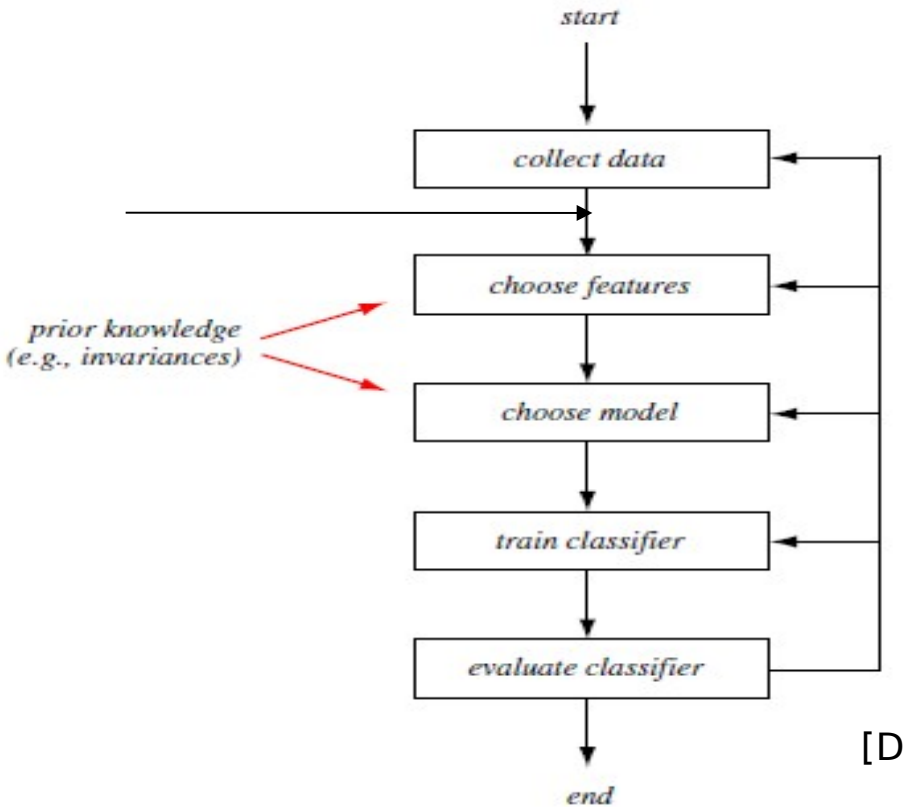
Classificação - observações

- Um mesmo conjunto de dados ⇒ diferentes classificadores
  - Diferentes características
  - Diferentes técnicas de construção de classificadores
  - Diferentes parâmetros

- Qual é o melhor?
  - Necessidade de avaliação dos classificadores
  - Comparação de desempenho entre opções

Ciclo de construção de classificadores baseados em aprendizado supervisionado

Pré-processamento



[DUDA, HART & STORK, 2001]

Fim do vídeo 2

Conceitos básicos de reconhecimento de padrões

Profa Ariane Machado Lima

Vídeo 3

Problemas de Reconhecimento Sintático ou Estrutural de Padrões

Profa Ariane Machado Lima

Os 2 tipos de reconhecimento de padrões

- Não estrutural
  - Normalmente não se considera a composição dos objetos
  - Os objetos são representados por suas características (mais ou menos independentes)
  - cuja ordem não importa

Os 2 tipos de reconhecimento de padrões

- Não estrutural
  - Normalmente não se considera a composição dos objetos
  - Os objetos são representados por suas características (mais ou menos independentes)
  - cuja ordem não importa

Wine Quality Data Set

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Attribute Information:

For more information, read [Cortez et al., 2009].
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

Os 2 tipos de reconhecimento de padrões

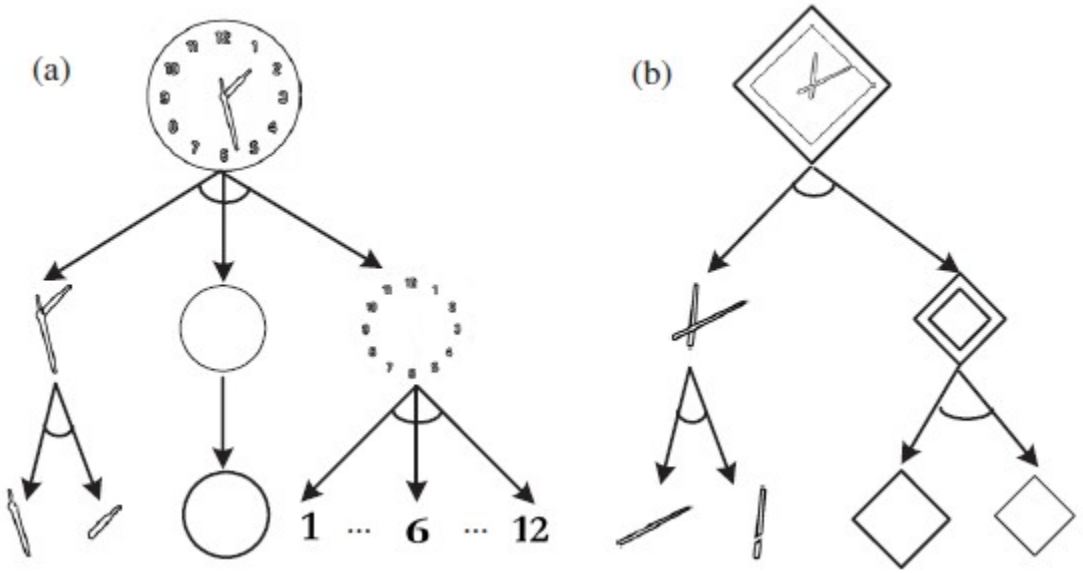
- Não estrutural
  - Normalmente não se considera a composição dos objetos
  - Os objetos são representados por suas características (mais ou menos independentes)
  - cuja ordem não importa
  - Os modelos do problema muitas vezes não possuem interpretabilidade (caixa-preta)
    - O que importa é a classificação

Os 2 tipos de reconhecimento de padrões

- Estrutural / sintático
  - A composição interna dos objetos é FUNDAMENTAL
    - Pode-se definir uma hierarquia de subpadrões, até chegar no nível de primitivas (facilmente identificáveis)
  - Os objetos são representados por constituição, a ORDEM é importantíssima
    - Sequências de DNA, RNA, proteínas
    - Textos
    - Imagens estruturadas (ex: carro)
  - Aprender/modelar a estrutura é muitas vezes um dos objetivos (além de classificar)

Exemplos

Ex: imagens de componentes estruturados



DOI: 10.1561/0600000018

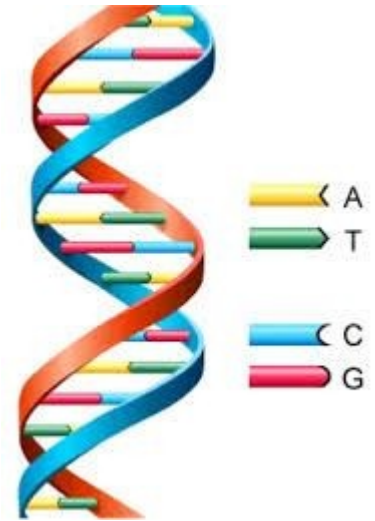
Ex: imagens de componentes estruturados

A identificação dos componentes (moldura, ponteiros) e a relação entre eles (posição relativa) SÃO IMPORTANTES

Não bastam características como cor, número de dígitos, etc…

Exemplos

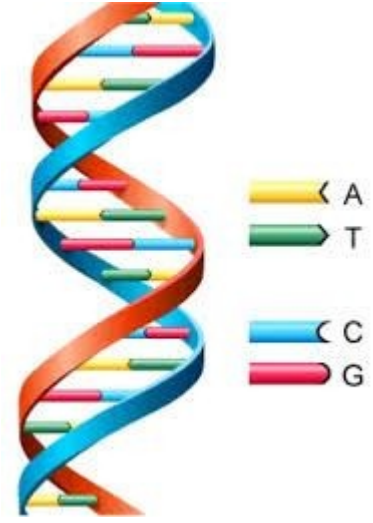
- Predição de genes
- codificantes de proteínas



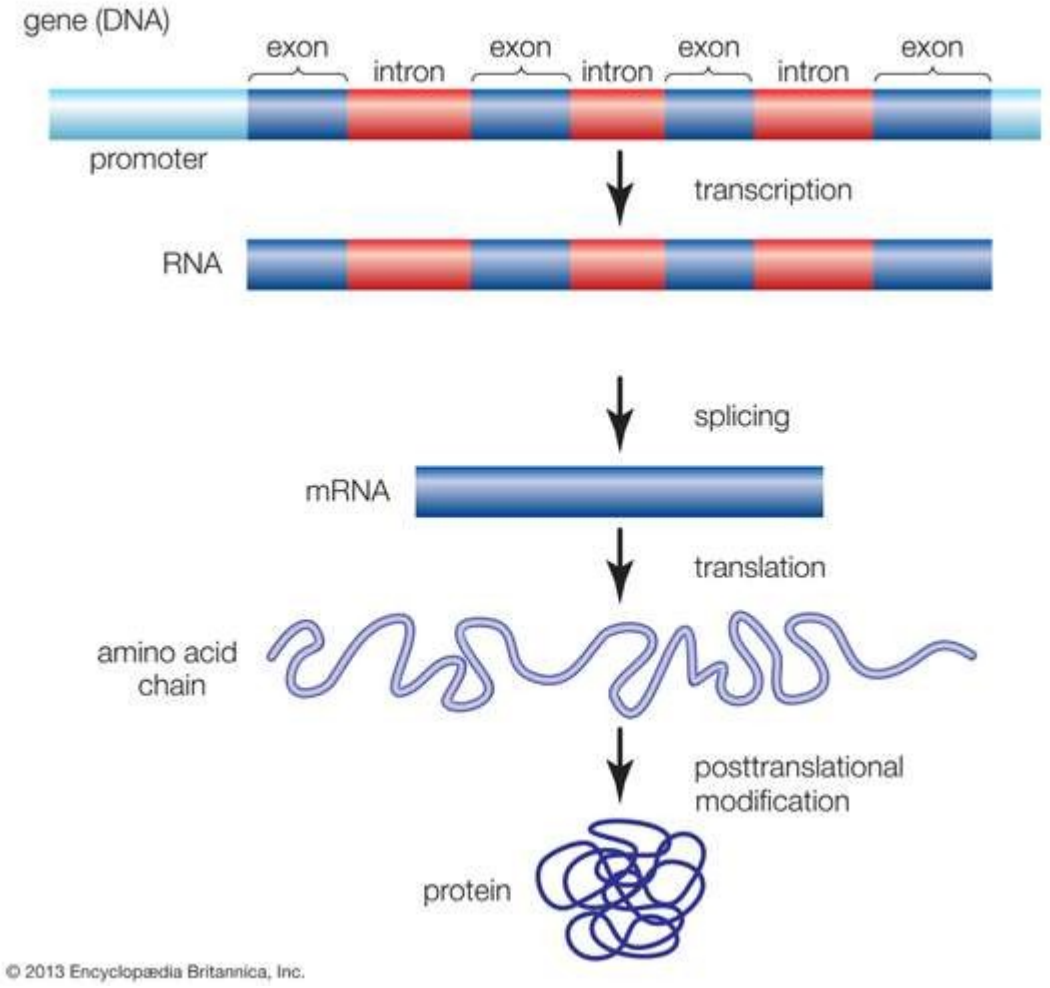
https://www.todamateria.com.br/dna/

Exemplos

- Predição de genes
- codificantes de proteínas

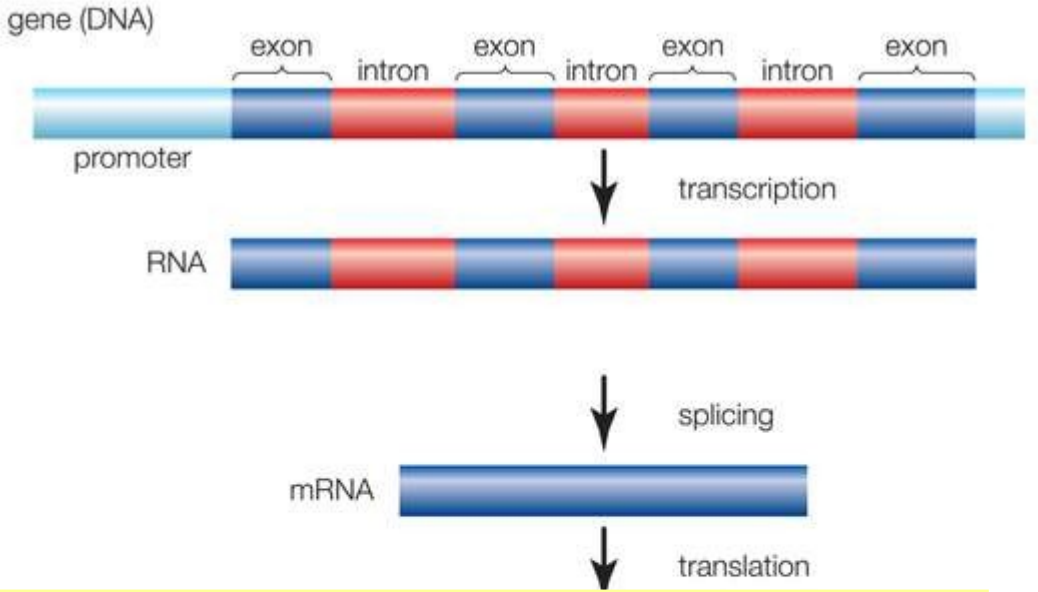


https://www.todamateria.com.br/dna/



https://www.britannica.com/science/gene

Exemplos



- Predição de genes
- codificantes de proteínas

A sequência dos "símbolos" (letras → nucleotídeos) é importante

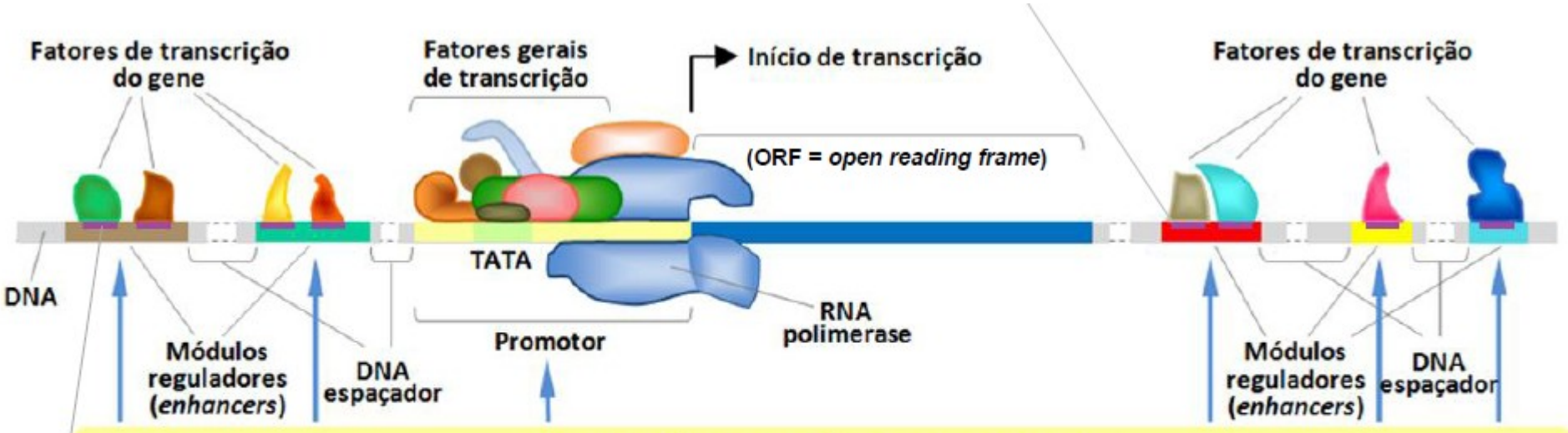
Não bastam características como comprimento, composição de bases (ex: %G+C), etc…

https://www.todamateria.com.br/dna/

https://www.britannica.com/science/gene

Mais um exemplo que nos ajudará a entender mais alguns conceitos da área

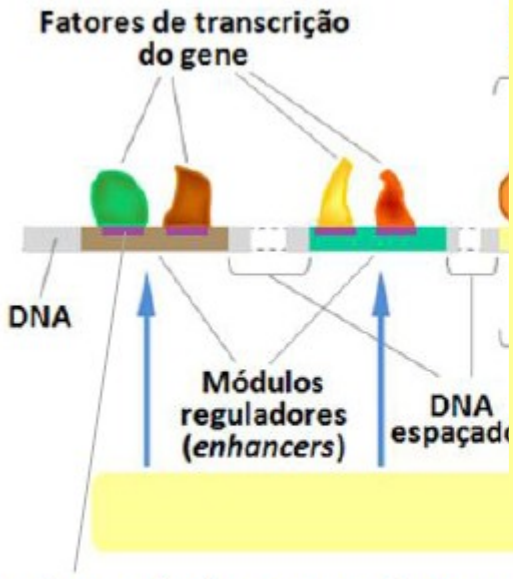
Ex: Sítios de ligação de fatores de transcrição



Sítios de Ligação dos Fatores de Transcrição:
Trechos de 5 a 30pbs no DNA, reconhecidos pelos Fatores de Transcrição ( )

Adaptado de Alberts et al., 2002.

Ex: Sítios de ligação de fatores de transcrição



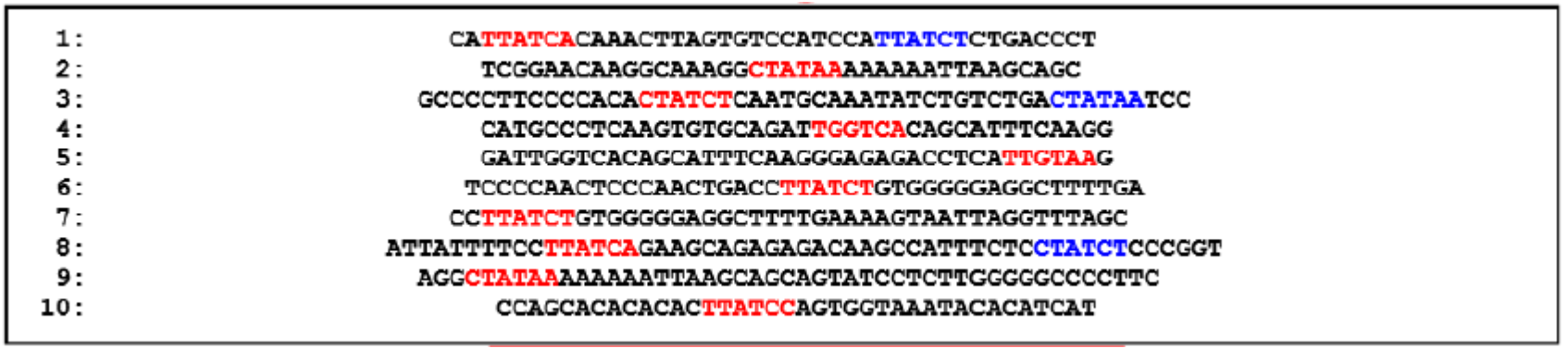
- Existem vários fatores de transcrição (FTs) diferentes
- Cada módulo regulador normalmente possui vários sítios (de FT's diferentes e TAMBÉM para um mesmo FT)
- Cada FT reconhece um padrão de sequência (motivo) cujas
- sequências são:
- 1) parecidas porém não idênticas
- 2) de mesmo tamanho

Região de controle para o Gene X

Sítios de Ligação dos Fatores de Transcrição:
Trechos de 5 a 30pbs no DNA, reconhecidos pelos Fatores de Transcrição (▬)

Adaptado de Alberts et al., 2002.

Sítios de ligação de UM DADO FT



[Dissertação USP – Antônio Ferrão Neto, 2017]

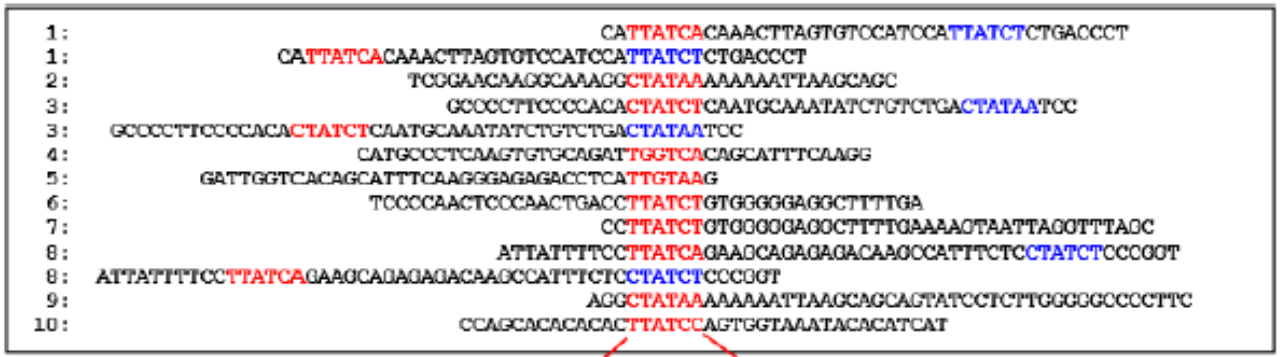
Sítios de ligação de UM DADO FT

```
 1:                          CATTATCACAAACTTAGTGTCCATCCATTATCTCTGACCCT
 2:                            TCGGAACAAGGCAAAGGCTATAAAAAAAATTAAGCAGC
 3:                        GCCCCTTCCCCACACTATCTCAATGCAAATATCTGTCTGACTATAATCC
 4:                            CATGCCCTCAAGTGTGCAGATTGGTCACAGCATTTCAAGG
 5:                            GATTGGTCACAGCATTTCAAGGGAGAGACCTCATTGTAAG
 6:                          TCCCCAACTCCCAACTGACCTTATCTGTGGGGGAGGCTTTTGA
 7:                         CCTTATCTGTGGGGGAGGCTTTTGAAAAGTAATTAGGTTTAGC
 8:                    ATTATTTTCCTTATCAGAAGCAGAGAGACAAGCCATTTCTCCTATCTCCCGGT
 9:                      AGGCTATAAAAAAAATTAAGCAGCAGTATCCTCTTGGGGGCCCCTTC
10:                           CCAGCACACACACTTATCCAGTGGTAAATACACATCAT
```

ALINHAMENTO MÚLTIPLO SEM GAPS

```
 1:                                                  CATTATCACAAACTTAGTGTCCATCCATTATCTCTGACCCT
 1:                        CATTATCACAAACTTAGTGTCCATCCATTATCTCTGACCCT
 2:                                  TCGGAACAAGGCAAAGGCTATAAAAAAAATTAAGCAGC
 3:                                    GCCCCTTCCCCACACTATCTCAATGCAAATATCTGTCTGACTATAATCC
 3:   GCCCCTTCCCCACACTATCTCAATGCAAATATCTGTCTGACTATAATCC
 4:                              CATGCCCTCAAGTGTGCAGATTGGTCACAGCATTTCAAGG
 5:                  GATTGGTCACAGCATTTCAAGGGAGAGACCTCATTGTAAG
 6:                              TCCCCAACTCCCAACTGACCTTATCTGTGGGGGAGGCTTTTGA
 7:                                                  CCTTATCTGTGGGGGAGGCTTTTGAAAAGTAATTAGGTTTAGC
 8:                                             ATTATTTTCCTTATCAGAAGCAGAGAGACAAGCCATTTCTCCTATCTCCCGGT
 8: ATTATTTTCCTTATCAGAAGCAGAGAGACAAGCCATTTCTCCTATCTCCCGGT
 9:                                                  AGGCTATAAAAAAAATTAAGCAGCAGTATCCTCTTGGGGGCCCCTTC
10:                                       CCAGCACACACACTTATCCAGTGGTAAATACACATCAT
```

[Dissertação USP – Antônio Ferrão Neto, 2017]

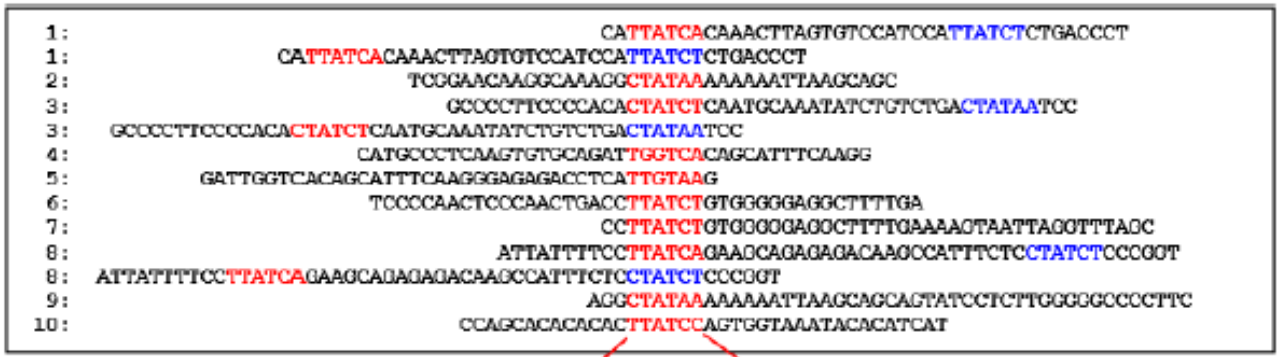


POSIÇÕES

SEQUÊNCIAS

	1	2	3	4	5	6
1	T	T	A	T	C	A
1	T	T	A	T	C	T
2	C	T	A	T	A	A
3	C	T	A	T	C	T
3	C	T	A	T	A	A
4	T	G	G	T	C	A
5	T	T	G	T	A	A
6	T	T	A	T	C	T
7	T	T	A	T	C	T
8	T	T	A	T	C	A
8	C	T	A	T	C	T
9	C	T	A	T	A	A
10	T	T	A	T	C	C

[Dissertação USP – Antônio Ferrão Neto, 2017]



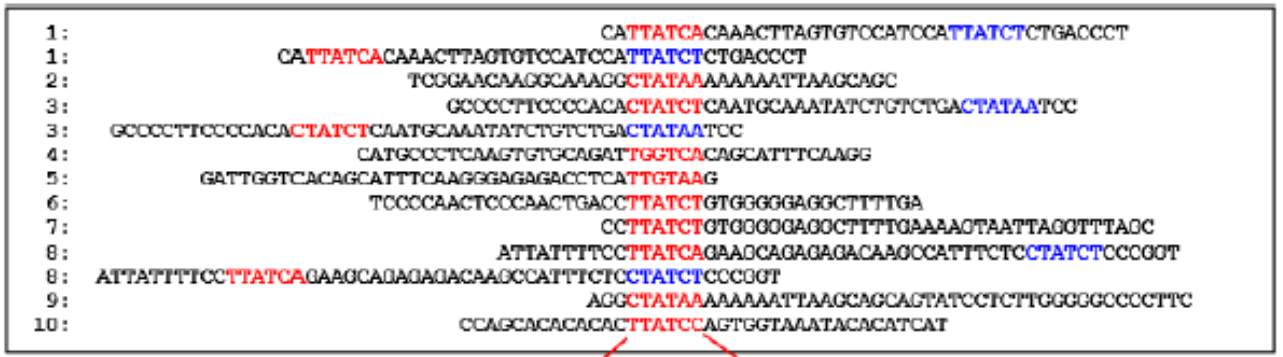
POSIÇÕES

SEQUÊNCIAS

1	2	3	4	5	6
T	T	A	T	C	A
T	T	A	T	C	T
C	T	A	T	A	A
C	T	A	T	C	T
C	T	A	T	A	A
T	G	G	T	C	A
T	T	G	T	A	A
T	T	A	T	C	T
T	T	A	T	C	T
T	T	A	T	C	A
C	T	A	T	C	T
C	T	A	T	A	A
T	T	A	T	C	C

Como podemos criar um modelo para representar esse padrão?

[Dissertação USP – Antônio Ferrão Neto, 2017]



POSIÇÕES

SEQUÊNCIAS

1	2	3	4	5	6
T	T	A	T	C	A
T	T	A	T	C	T
C	T	A	T	A	A
C	T	A	T	C	T
C	T	A	T	A	A
T	G	G	T	C	A
T	T	G	T	A	A
T	T	A	T	C	T
T	T	A	T	C	T
T	T	A	T	C	A
C	T	A	T	C	T
C	T	A	T	A	A
T	T	A	T	C	C

Como podemos criar um modelo para representar esse padrão?

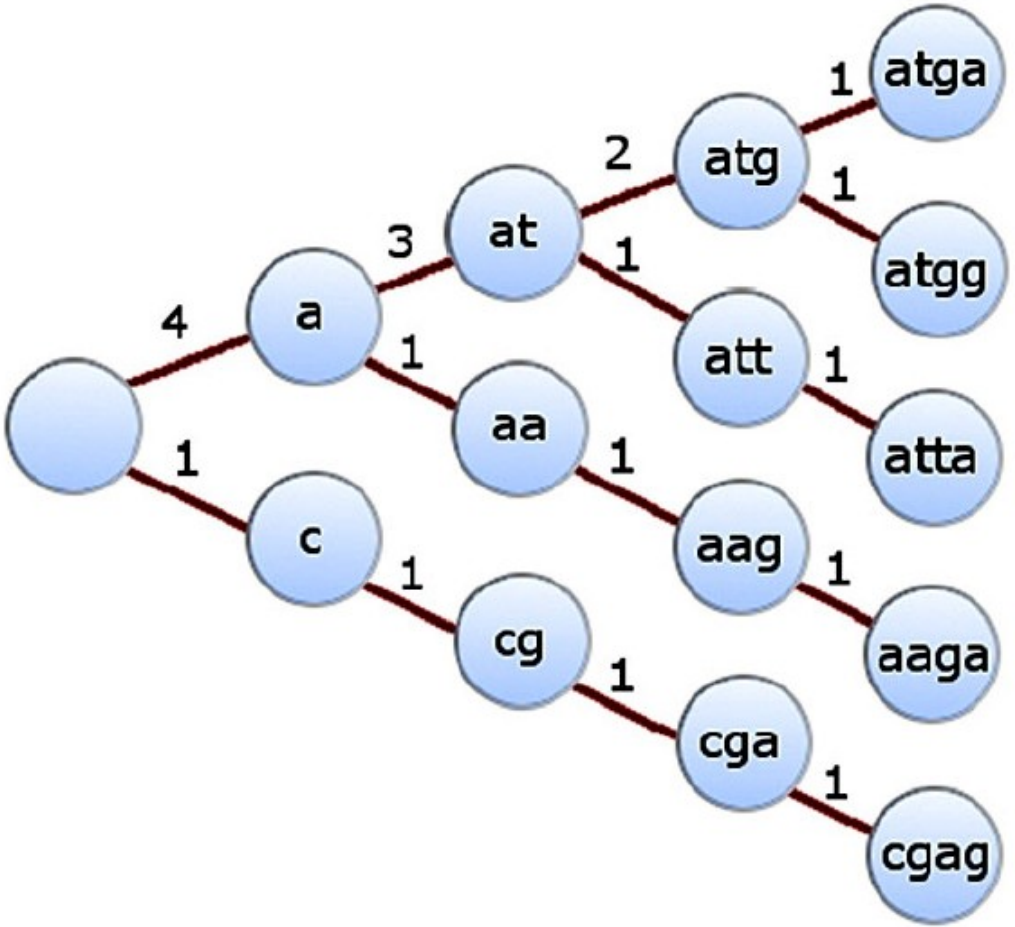
Por ex: árvore de prefixos

[Dissertação USP – Antônio Ferrão Neto, 2017]

Árvore de prefixos

Exemplo para um conjunto menor de de sequências menores:

atga
atgg
atta
aaga
cgag



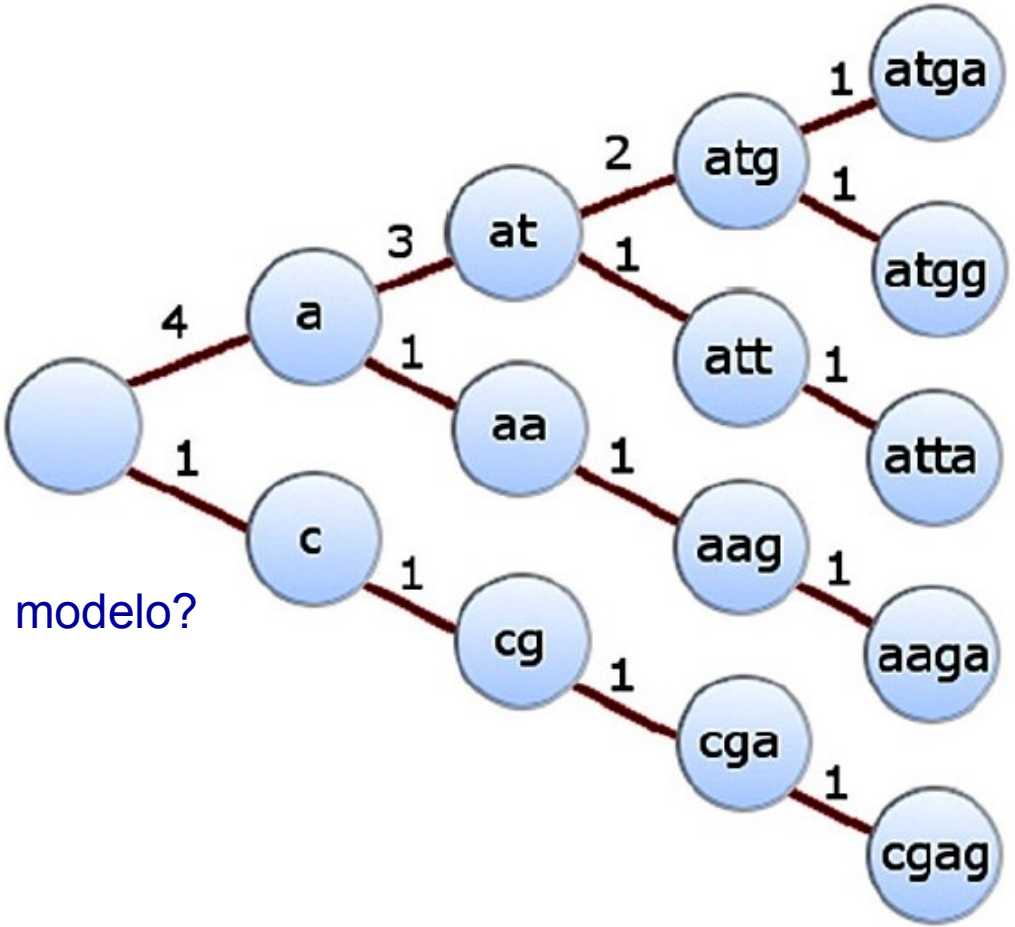
Genetics and Molecular Research 6 (1): 105-115 (2007)

Árvore de prefixos

Exemplo para um conjunto menor de de sequências menores:

atga
atgg
atta
aaga
cgag

Alguém vê um “problema” neste modelo?



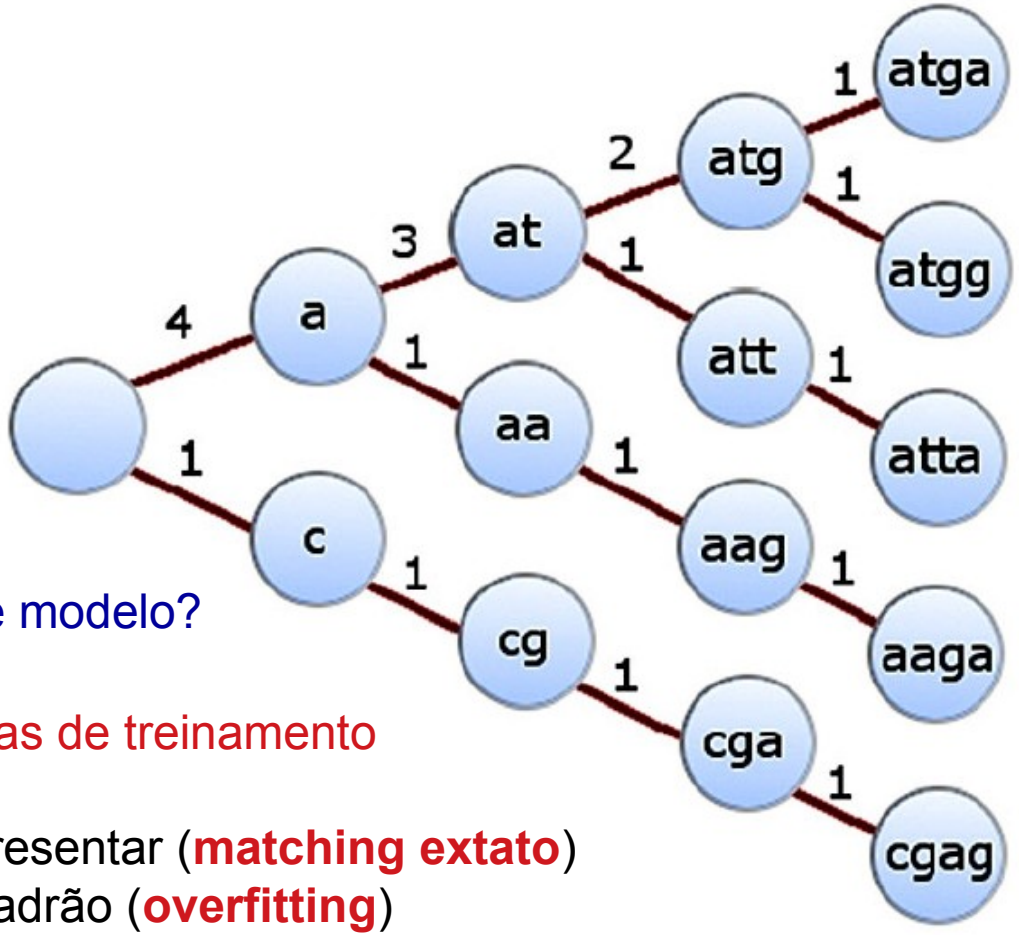
Genetics and Molecular Research 6 (1): 105-115 (2007)

Árvore de prefixos

Exemplo para um conjunto menor de de sequências menores:

atga
atgg
atta
aaga
cgag

Alguém vê um “problema” neste modelo?



Ele representa EXATAMENTE as sequências de treinamento

Isso pode ser:

- maravilhoso se é isso que você quer representar (**matching extato**)
- péssimo se você quer APRENDER um padrão (**overfitting**)

Genetics and Molecular Research 6 (1): 105-115 (2007)

Overfitting x Generalização

- Overfitting: ótimos resultados na amostra de treinamento, mas ruim na amostra de teste.
  - No pior caso, só conhece (todos) os elementos da amostra de treinamento
- Generalização: permite aprender além do que está na amostra de treinamento. Mas se generalizar demais vai errar demais.
  - Ex: ao se querer aprender um conceito, pode erroneamente classificar tudo como sendo daquela classe alvo

Overfitting x Generalização

1. Como será o desempenho (sucesso da classificação) para objetos fora da amostra de treinamento?

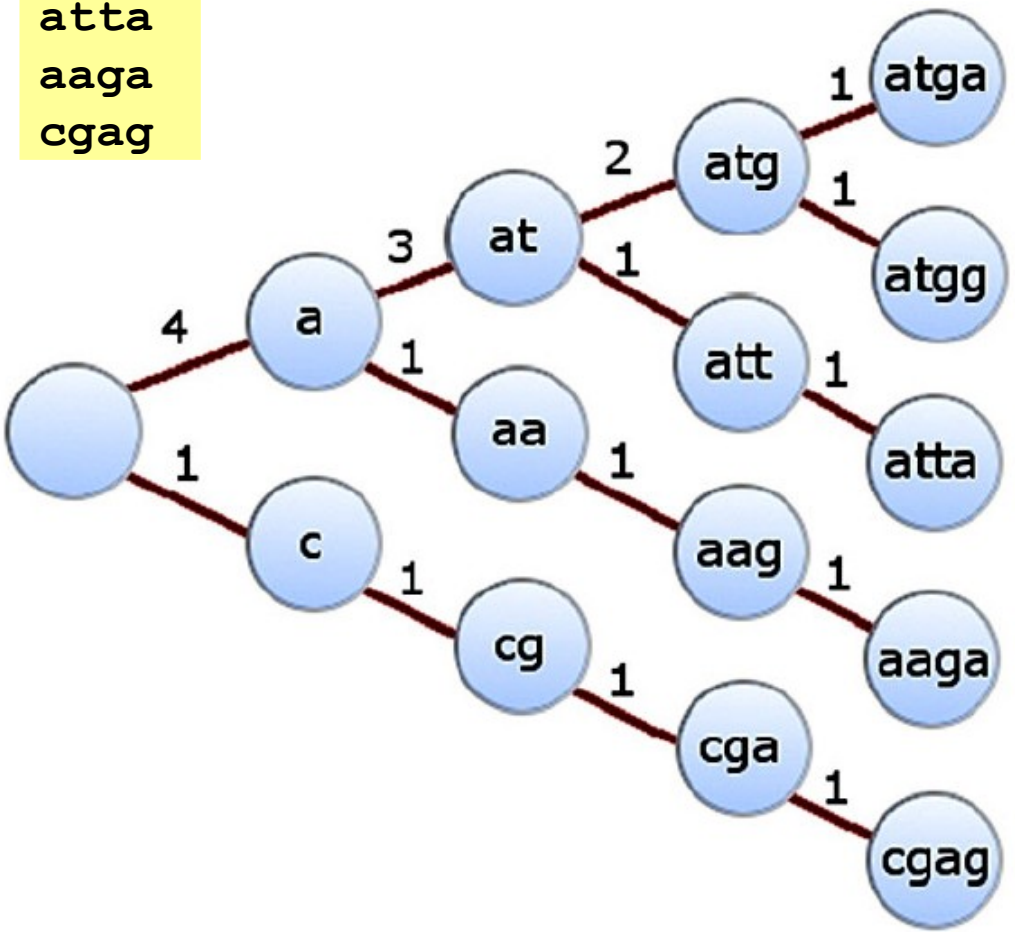
- Não queremos decorar a amostra de treinamento (overfitting), mas aprender um conceito, criar um modelo a partir da generalização dos exemplos, sem generalizar demais
- Maior sucesso na amostra de treinamento não garante maior sucesso nos dados “novos”

Dilema na escolha do modelo

- Há um gradiente de modelos dos mais simples aos mais complexos: qual escolher?
- Tempo de processamento, memória, erros, custos, overfitting, ...
- Como predizer quão boa será a generalização do classificador?
- Discutiremos essas questões durante o curso.

```
atga
atgg
atta
aaga
cgag
```

Overfitting x Generalização



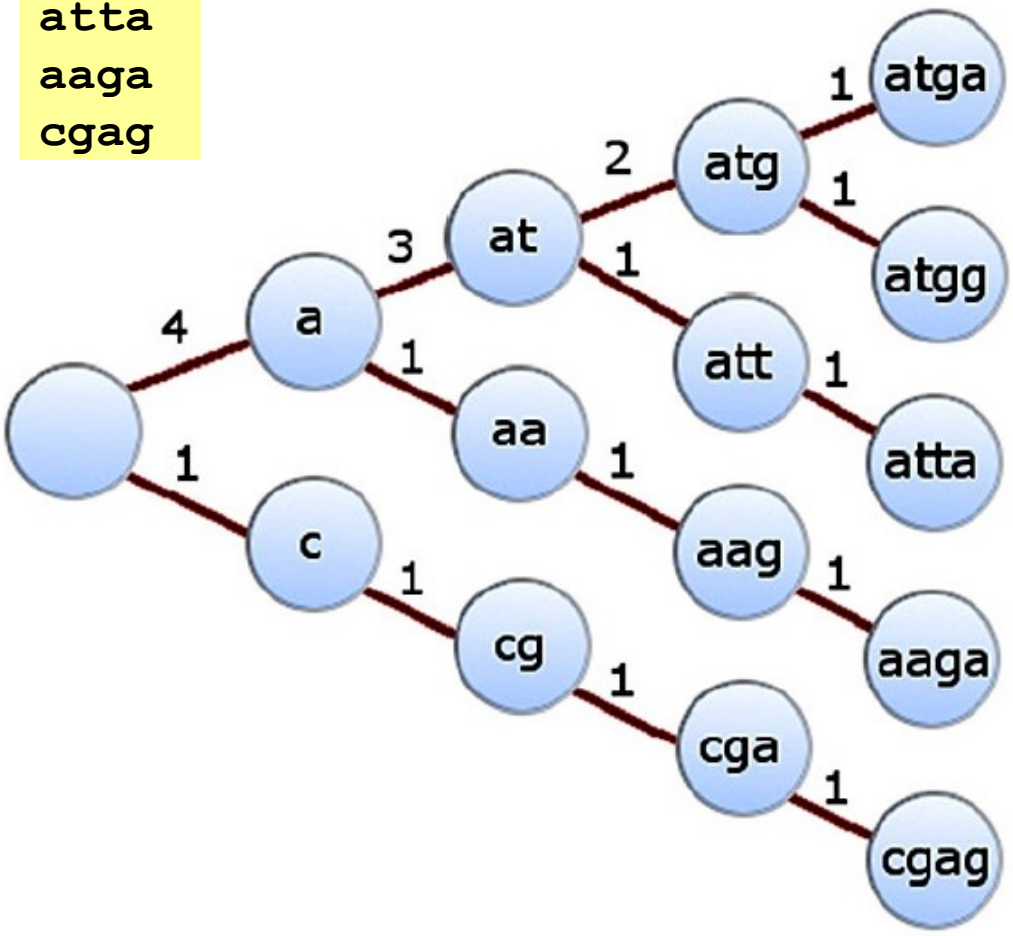
Genetics and Molecular Research 6 (1): 105-115 (2007)

x

Aceitar todas as sequências de tamanho 4

atga
atgg
atta
aaga
cgag

Overfitting x Generalização



x

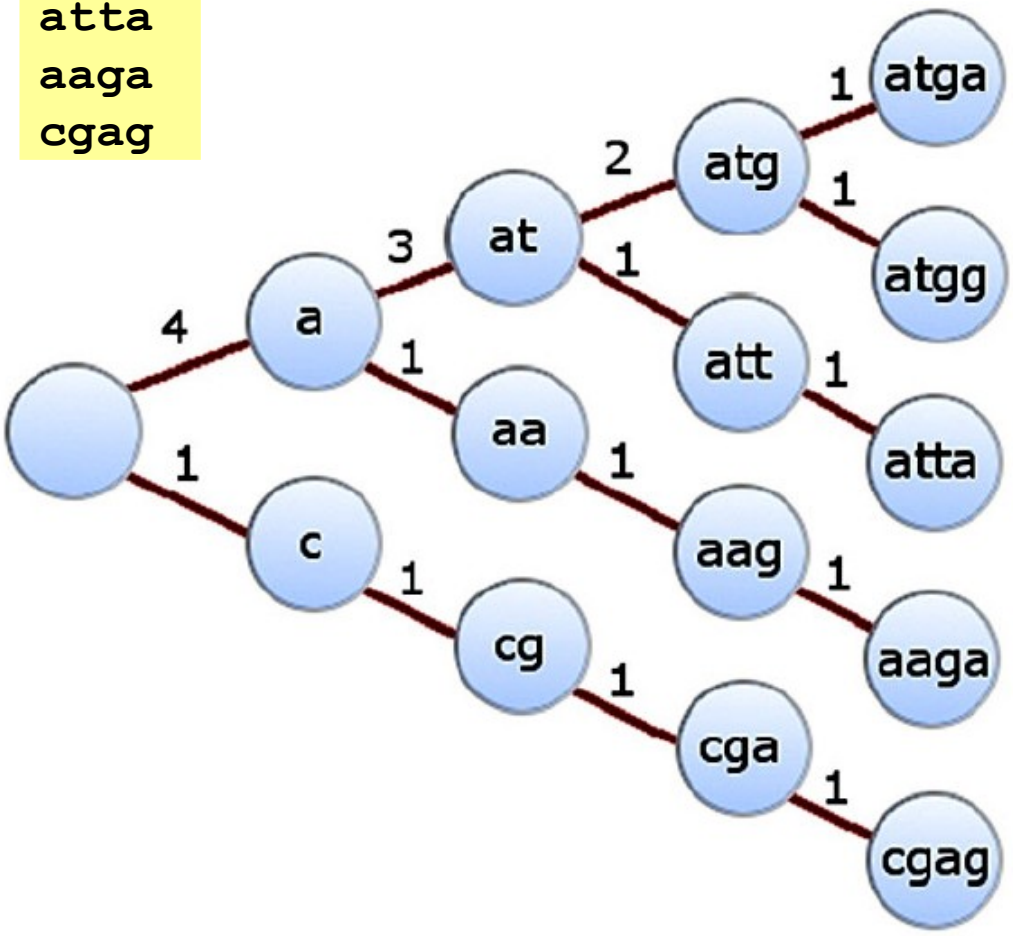
Aceitar todas as sequências de tamanho 4

Outras generalizações (intermediárias), assumindo sempre tamanho 4:

Genetics and Molecular Research 6 (1): 105-115 (2007)

atga
atgg
atta
aaga
cgag

Overfitting x Generalização



Genetics and Molecular Research 6 (1): 105-115 (2007)

X

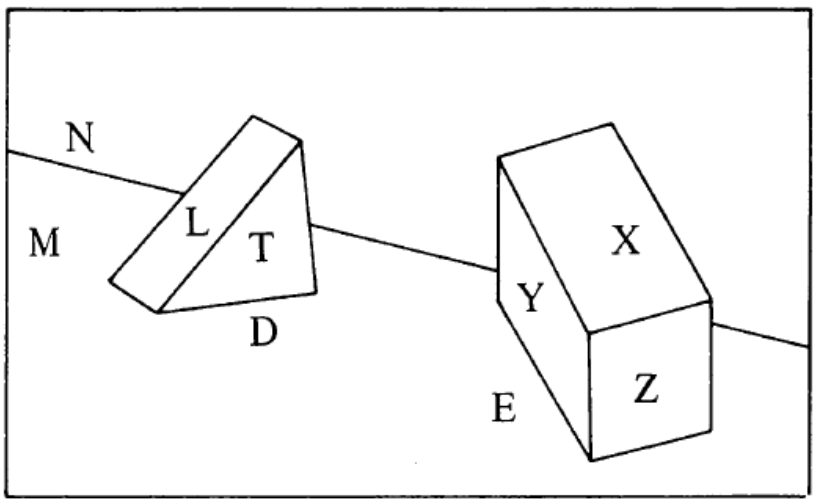
Aceitar todas as sequências de tamanho 4

Outras generalizações (intermediárias), assumindo sempre tamanho 4:

- começa com **a** ou **c**
- começa com **a** ou **c** E termina com **a** ou **g**
- ...

Mais sobre reconhecimento estrutural de padrões

Cena

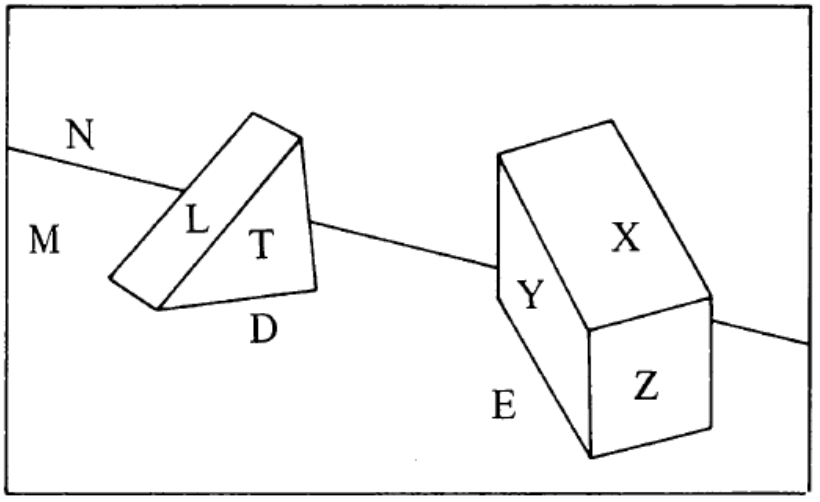


Scene A

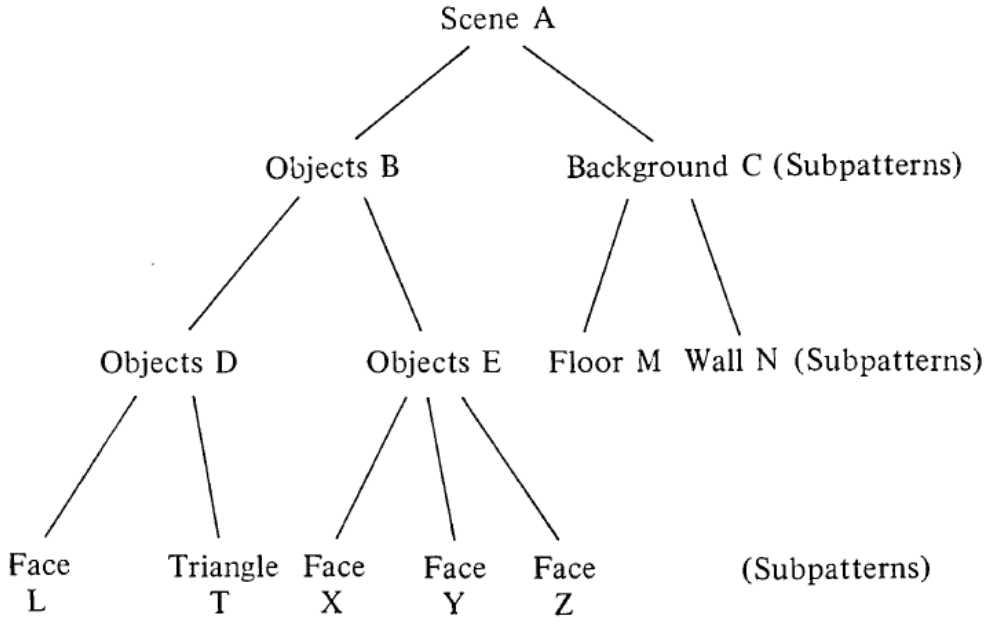
a

Fu, K. S. Syntactic Pattern Recognition and Applications, PrenticeHall, Inc., Englewood Cliffs, 1982

Cena



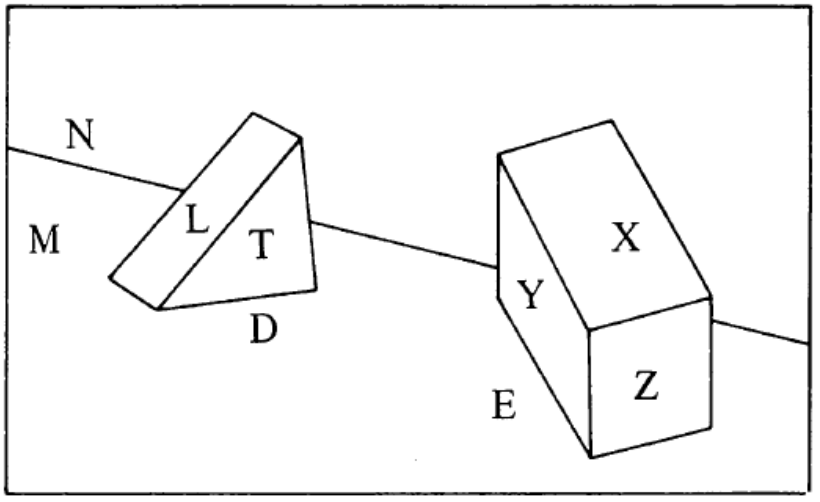
Representação hierárquica (árvore)



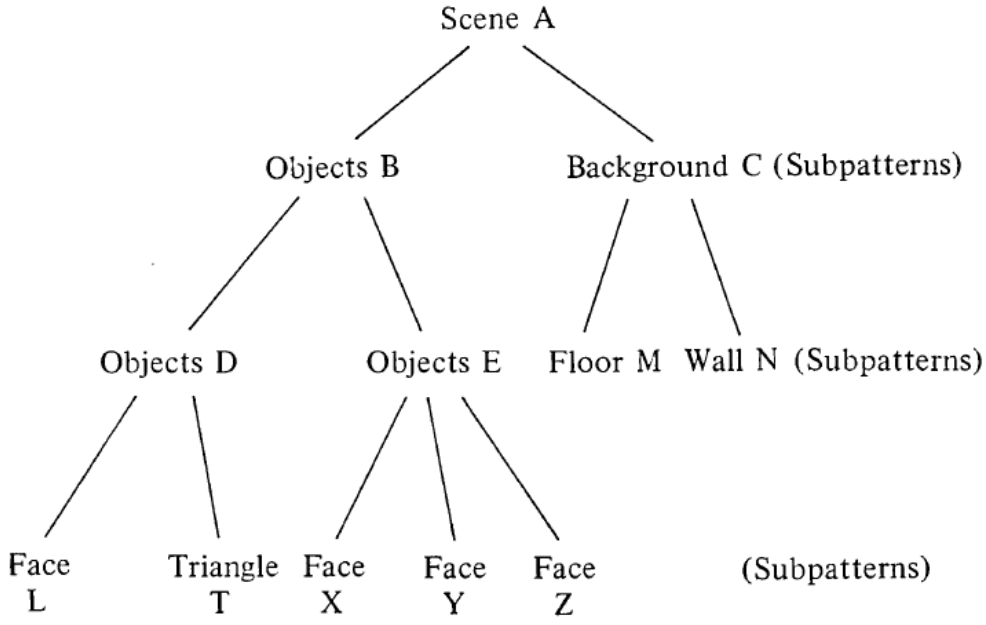
Fu, K. S. Syntactic Pattern Recognition and Applications, PrenticeHall, Inc., Englewood Cliffs, 1982

Padrões definidos a partir de subpadrões, cada vez mais simples, até chegar a primitivas facilmente identificáveis

Cena

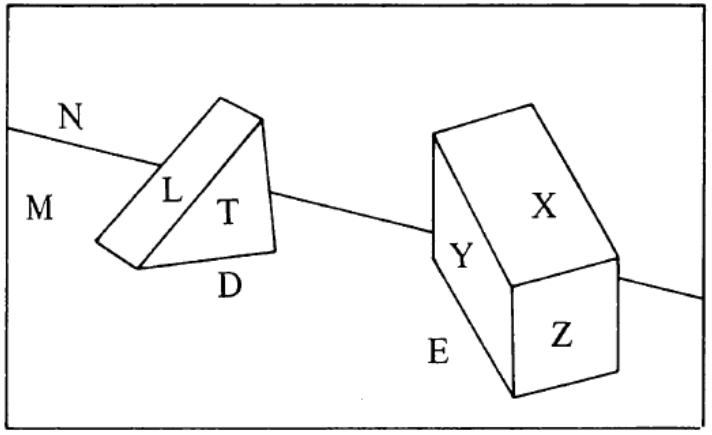


Representação hierárquica (árvore)



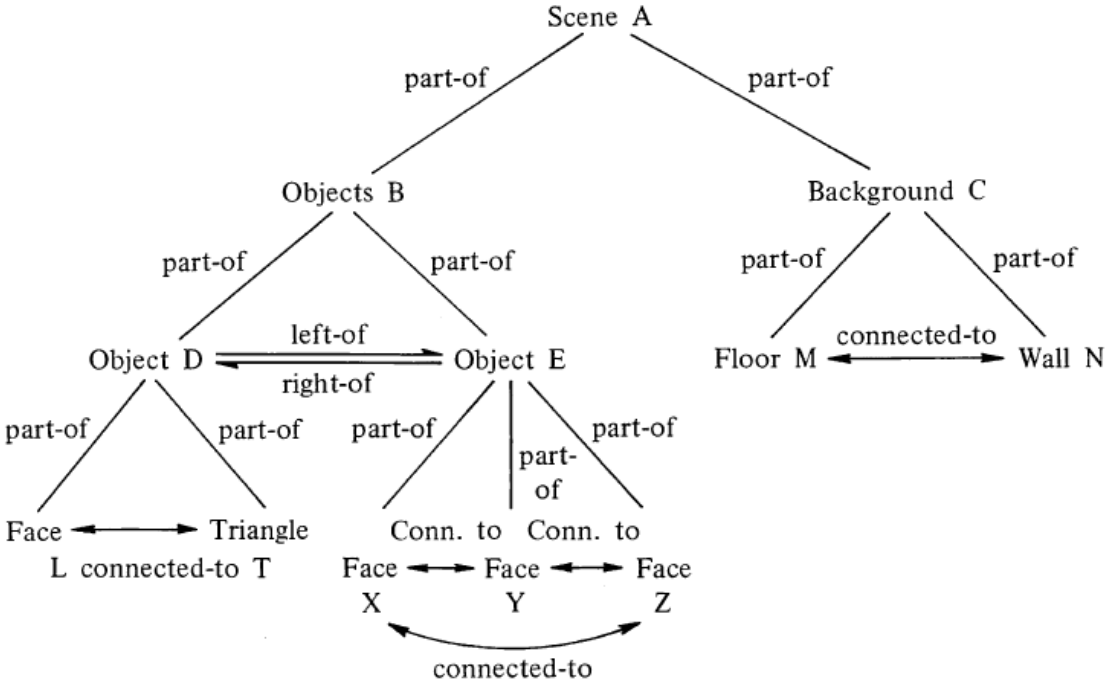
Fu, K. S. Syntactic Pattern Recognition and Applications, PrenticeHall, Inc., Englewood Cliffs, 1982

Cena



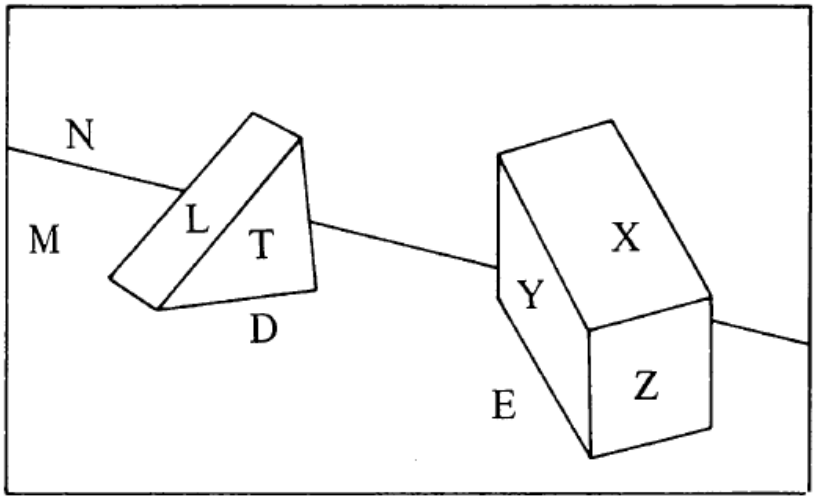
Fu, K. S. Syntactic Pattern Recognition and Applications, PrenticeHall, Inc., Englewo

Representação hierárquica (grafo)

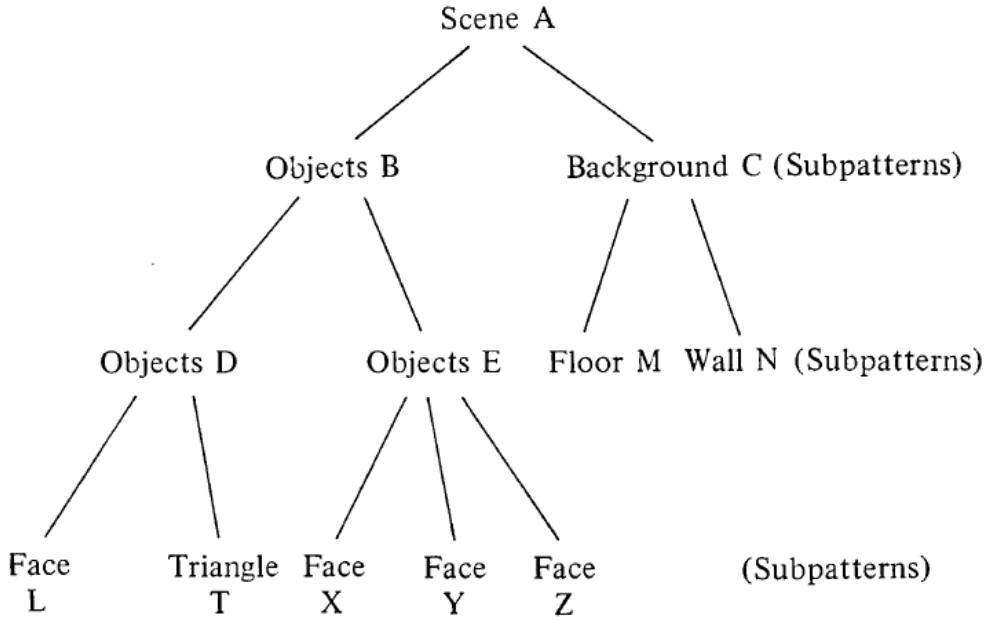


VANTAGEM das árvores: Uso direto de toda a teoria de Linguagens Formais (gramáticas)

Cena



Representação hierárquica (árvore)

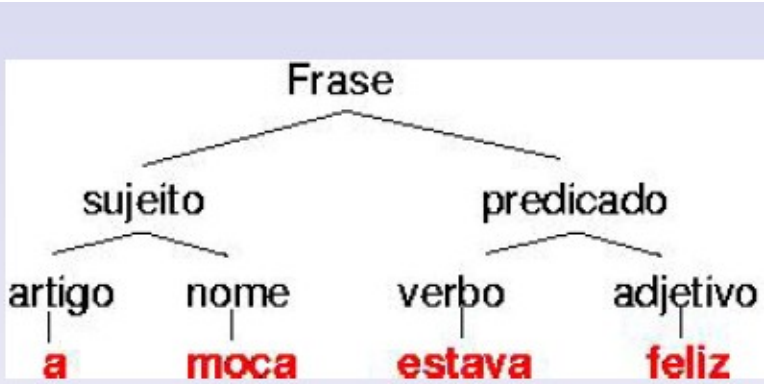


Fu, K. S. Syntactic Pattern Recognition and Applications, PrenticeHall, Inc., Englewood Cliffs, 1982

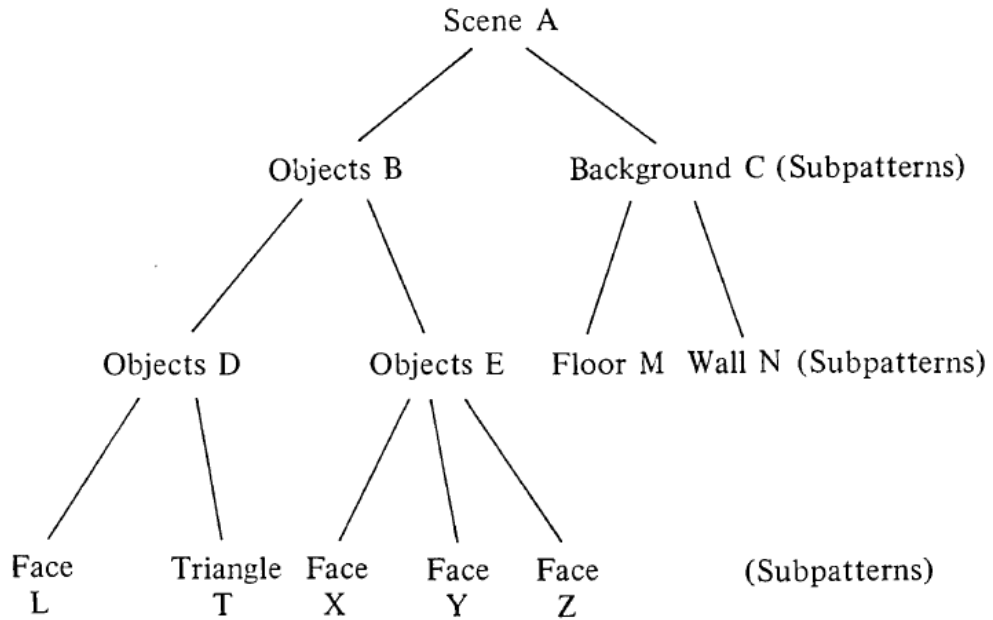
Gramática ?

Gramática ? SIM!

Frase	→	sujeito	predicado
sujeito	→	artigo	nome
artigo	→	a	
artigo	→	o	
nome	→	paletó	
nome	→	moça	
nome	→	dia	
predicado	→	verbo	adjetivo
verbo	→	é	
verbo	→	estava	
adjectivo	→	feliz	
adjectivo	→	azul	



Representação hierárquica (árvore)



Fu, K. S. Syntactic Pattern Recognition and Applications, PrenticeHall, Inc., Englewood Cliffs, 1982

Gramática

<Scene A> → <Objects B> <Background C>
<Objects B> → <Objects D> <Objects E>
<Objects D> → <Face L> <Triangle T>
<Objects E> → <Face X> <Face Y> <Face Z>
<Background C> → <Floor M> <Wall N>

Fim do vídeo 3

Problemas de Reconhecimento Sintático ou Estrutural de Padrões

Profa Ariane Machado Lima

Referências

- COSTA, L. F.; CESAR, R. M. Jr. **Shape Classification and Analysis: Theory and Practice**. CRC Press, 2009
- DUDA, R.; HART, P.; STORK, D**. Pattern Classification**. 2ª ed. John Willey, 2001 (Cap. 1)
- FU, K. S. **Syntatic Pattern Recognition, Applications**. Springer-Verlag, 1977 (Cap. 1).
- MAGALHÃES, M. N.; LIMA, A. C. P. de: **Noções de Probabilidade e Estatística**. Edusp, 2002
- TOU, J. T.; GONZALEZ, R. C. **Pattern Recognition Principles**. Addison Wesley, 1974.

Aula 1

Introdução a Problemas de Reconhecimento de Padrões e Conceitos Básicos

Profa Ariane Machado Lima