Como ensinar um computador a classificar galáxias?

Dr. Walter Santos Jr.
walter.augusto@gmail.com

IAG/USP - Astronomia ao meio-dia
30/03/2017

Introdução

Classificação

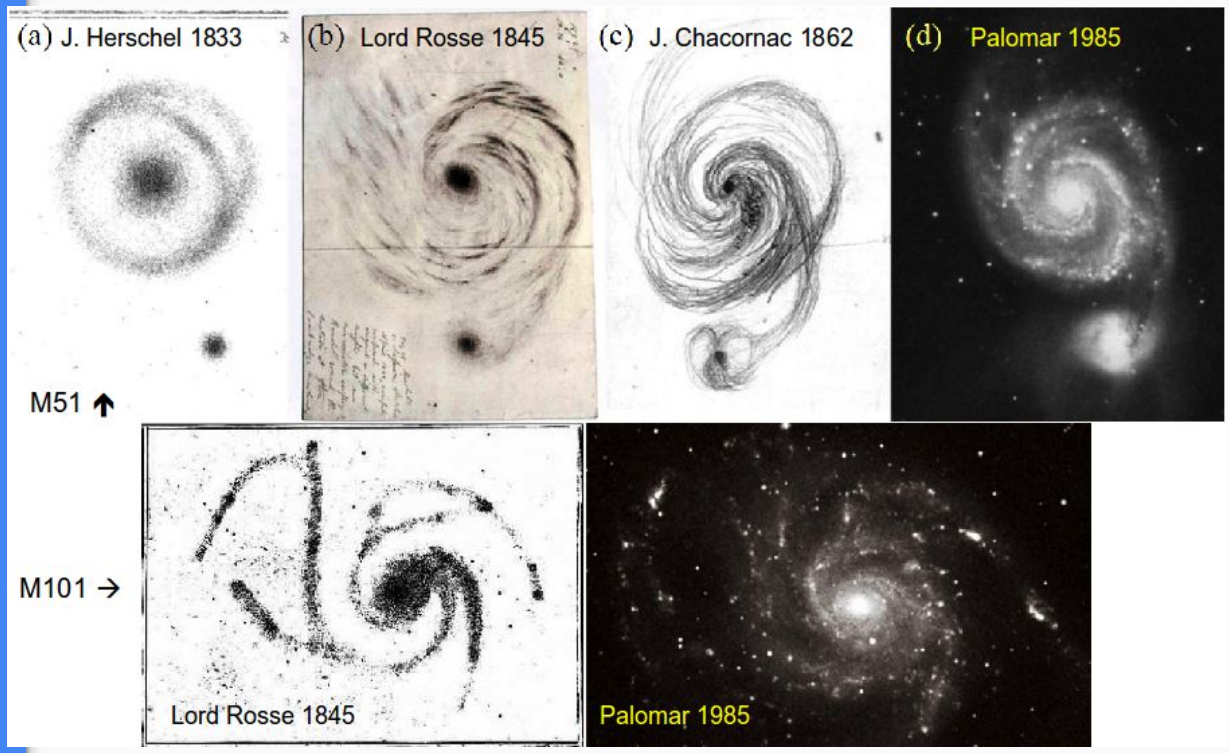
Por que classificamos objetos em ciência (ou Astronomia) em particular?

- Separar objetos com características semelhantes entre si;
- Encontrar padrões, relações entre diferentes classes de objetos, ambiente em que se encontram, etc;
- Também por questões linguísticas, claro;
- Classes de objetos em Astronomia: Estrelas, galáxias, planetas, quasares, ...

Observações de Galáxias

- Desenhos (>~1800s)
- Placas Fotográficas (>~1950s)
- CCD / “Surveys” (>~1990s-2000s)

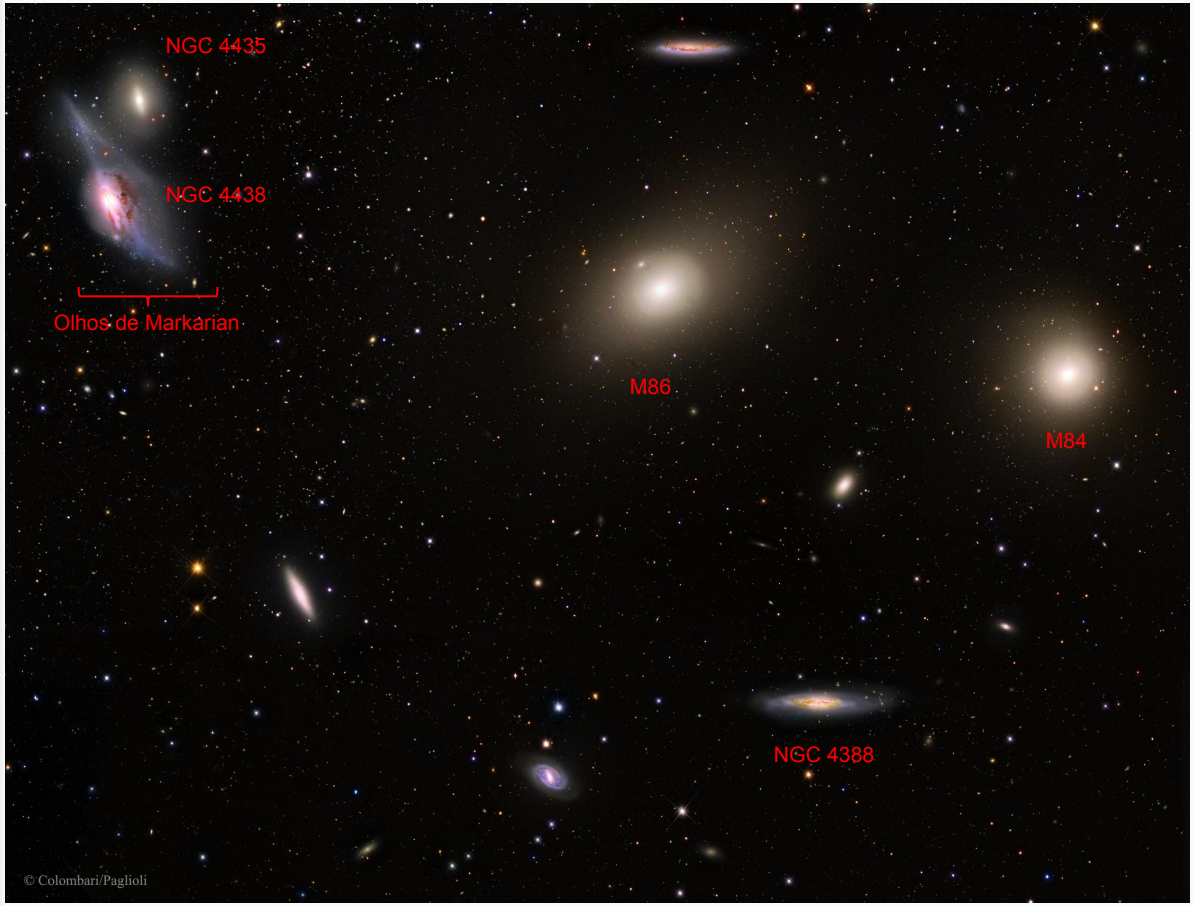
A medida que os telescópios vão se aprimorando, as imagens das galáxias se tornam mais nítidas e bem definidas.



Observações de Galáxias

- Desenhos (>~1800s)
- Placas Fotográficas (>~1950s)
- CCD / “Surveys” (>~1990s-2000s)

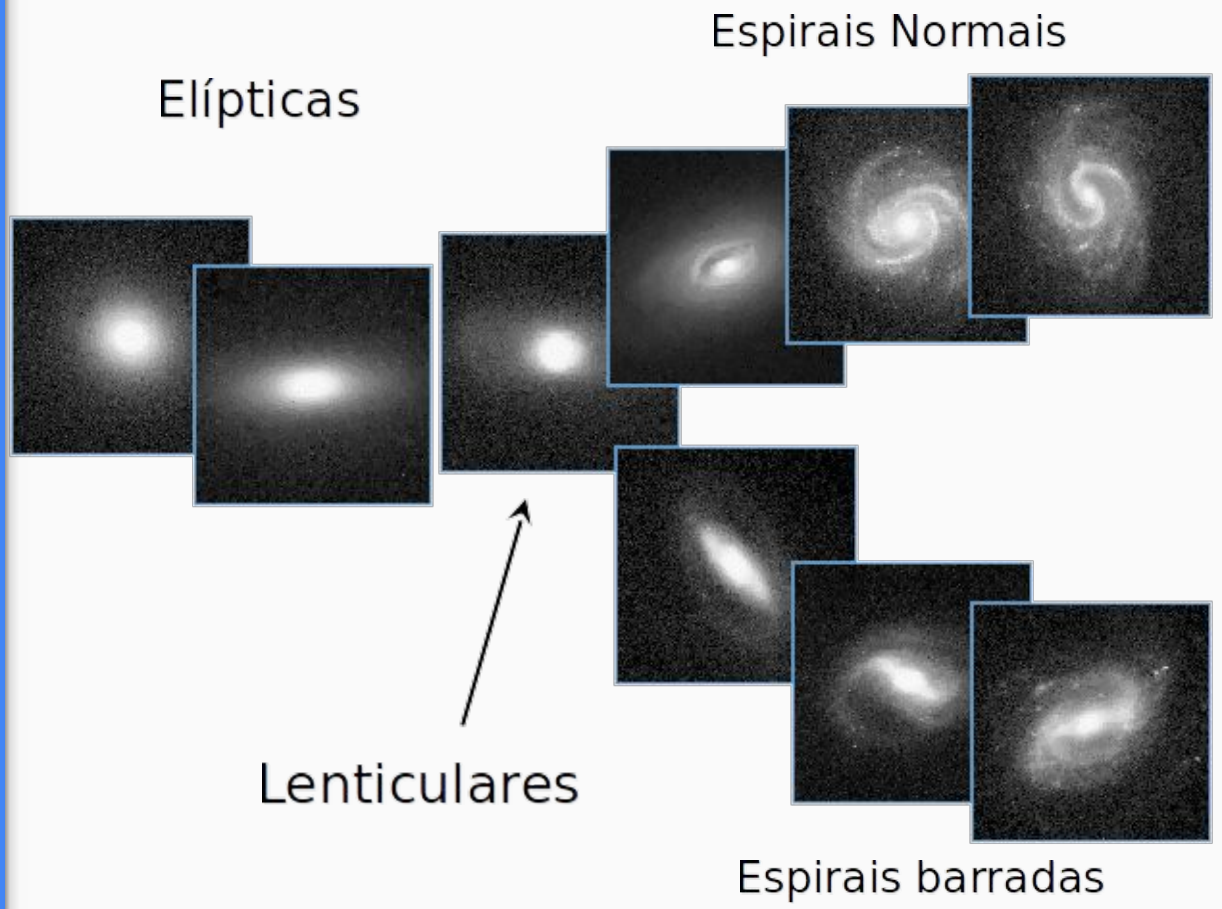
A medida que os telescópios vão se aprimorando, as imagens das galáxias se tornam mais nítidas e bem definidas.



(parte central do Aglomerado de Virgo)

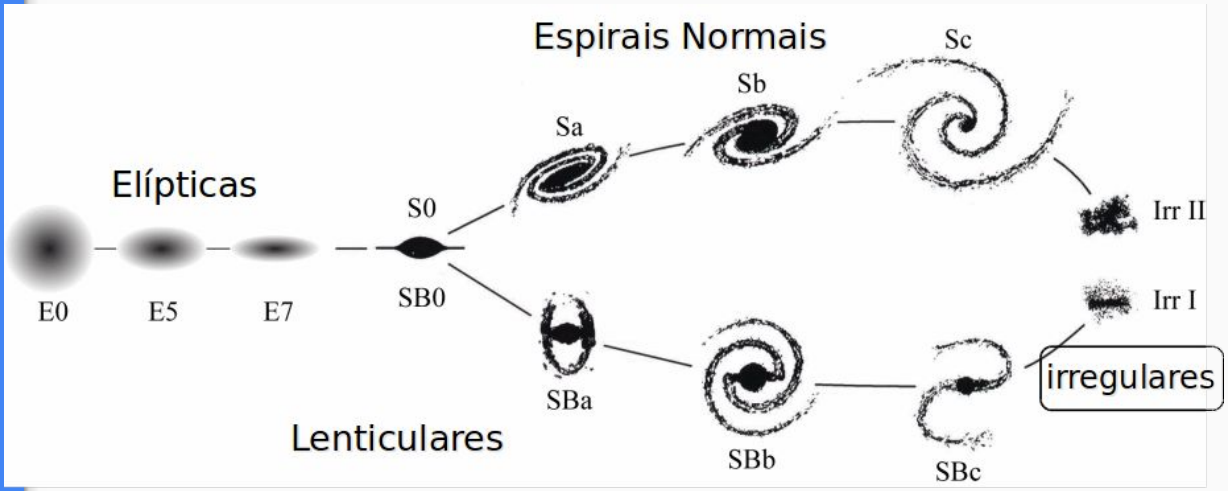
Classificação morfológica de galáxias

- Sequência de Hubble (1926, 1936)
- Sistema de diapasão



Classificação morfológica de galáxias

- Sequência de Hubble revista (1961)
- + irregulares, lenticulares;

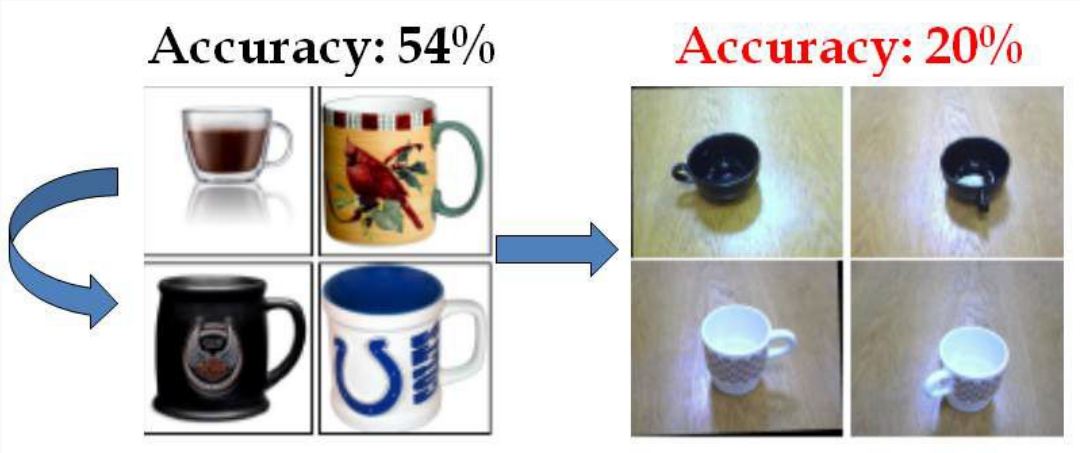


Classificação Humana x Máquina

Humano > Máquina

- O conjunto olho-cérebro-memória humano é incrivelmente capaz de analisar e classificar, de forma rápida e precisa, qualquer tipo de objeto;
- **Em geral, dependendo das imagens or das informações, humanos são melhores classificadores!!**
  - Um processo que começa desde um bebê recém-nascido, com reconhecimento facial, até a generalização (abstração) do conceito de 'objetos' (~2 anos)
- Computador é sensível à limitações de seus algoritmos e como matematizamos (ou modelamos) os objetos

Humano > Máquina



Um dos desafios de classificações de galáxias → diferentes orientações no céu para os mesmos tipos de galáxias

Máquina > Humano

- Humanos cansam; máquinas (pelo menos 99% das vezes), não!
  - Processos cognitivos pioram com o cansaço mental; + necessita dormir, comer, lazer, etc
- Subjetividade é (99% dos métodos) é eliminada:
  - Humanos tem experiências, memórias, etc diferentes entre si → leva a classificações diferentes!
  - Computadores diferentes, com os mesmos métodos e dados, devem levar ao mesmo resultado final!
- **MAS... o grande motivo é simplesmente a imensa quantidade de dados/imagens que estamos (e vamos ainda mais) recebendo!**

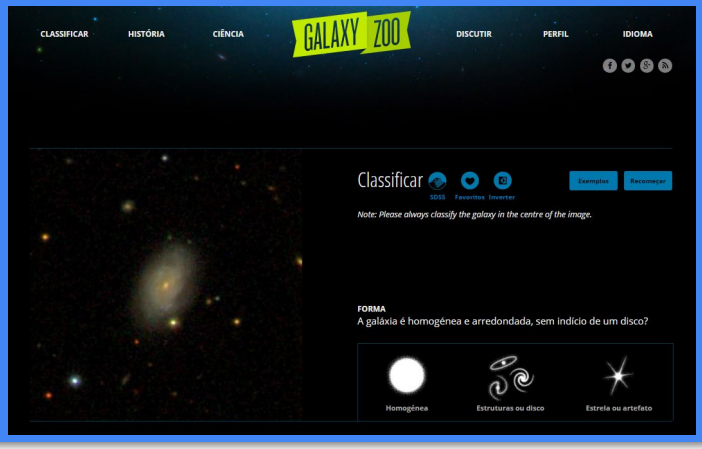
“Há um novo, quarto paradigma de descoberta baseado em ciência de dados intensivos.”

Jim Gray
cientista de computação, 1944-2012 (ano de declaração de sua morte)

Grandes projetos: “surveys” (Big Data)

- Sloan Digital Sky Survey (atual)
  - ~5 terapixels de imagens; ~470 milhões de objetos detectados (~208 milhões galáxias)
- Large Synoptic Survey Telescope (~2020)
  - Em um ano, deve coletar mais dados do que *todos* os outros telescópios combinados (1000 terabytes... **por noite**!); ~37 bilhões de galáxias
- Não há astrônomos (ou mesmo pessoas) pra olhar tudo isso!
  - MAS... computadores ficam mais poderosos exponencialmente + novas arquiteturas (nuvens, *‘clusters’*, ...) + **técnicas de aprendizagem de computadores**
- **MAS... um outro caminho complementar: usar a ajuda da “população leiga” na análise / classificação de imagens! → GalaxyZoo**

GalaxyZoo



- Projeto de astronomia do tipo “Crowdsourced” ou “ciência de cidadãos”: convidar membros da população ‘leiga’ em geral para ajudar em projetos científicos reais!
- 40 milhões de classificações foram feitas em ~ 175 dias, por mais de 100 mil voluntários, uma média de 38 cassificações por galáxia (amostra de 900 mil no total do SDSS)
- O resultado foi impressionante:
  - Incrível participação popular;
  - Mais eficaz que métodos computacionais (pelo menos até recentemente);
  - Astrônomos profissionais não *muito* melhores classificadores que cidadãos comuns;
  - Subjetividade: algo positivo → melhor estimativa de incertezas;
  - Novos objetos “anômalos” foram descobertos!
  - Levou a diversos resultados e artigos científicos, correlacionando morfologia com outras propriedades, ambiente, etc;
  - **Os resultados pode ser ‘incorporados’ para melhores os classificadores de máquinas!**

Aprendizagem de máquina!

Aprendizagem de Máquina (ou como imitar a classificação humana!)

Humanos:

- **Exemplos/generalização:** ver vários tipos de um mesmo objeto (ex: xícara) nos permite dizer se um objeto novo (que nunca vimos antes) é ou não uma xícara!

Máquinas:

- **Conjunto de Treinamento:** fornecer para um computador uma base de dados de diversos objetos, com as devidas classes, já sabidas (por exemplo, xícaras × não-xícaras), com variáveis matemáticas em uma dada modelagem ou medidas

Aprendizagem de Máquina (ou como imitar a classificação humana!)

Humanos:

- **Extrapolação:** mesmo quando algum elemento que consideramos parte essencial para classificar um objeto (ex: xícara faltando ‘alça’), podemos deduzir a classe de um objeto, extrapolando outros elementos conhecidos (ex: talvez simplesmente não podemos ver a ‘alça’)

Máquinas:

- **Ajustes de funções ou extrapolação:** da mesma forma, se estiver faltando dados em um certo intervalo de valores de , por exemplo, uma medida... dependendo do método, o computador pode extrapolar a partir de valores conhecidos através, por exemplo, de um ajuste de uma função

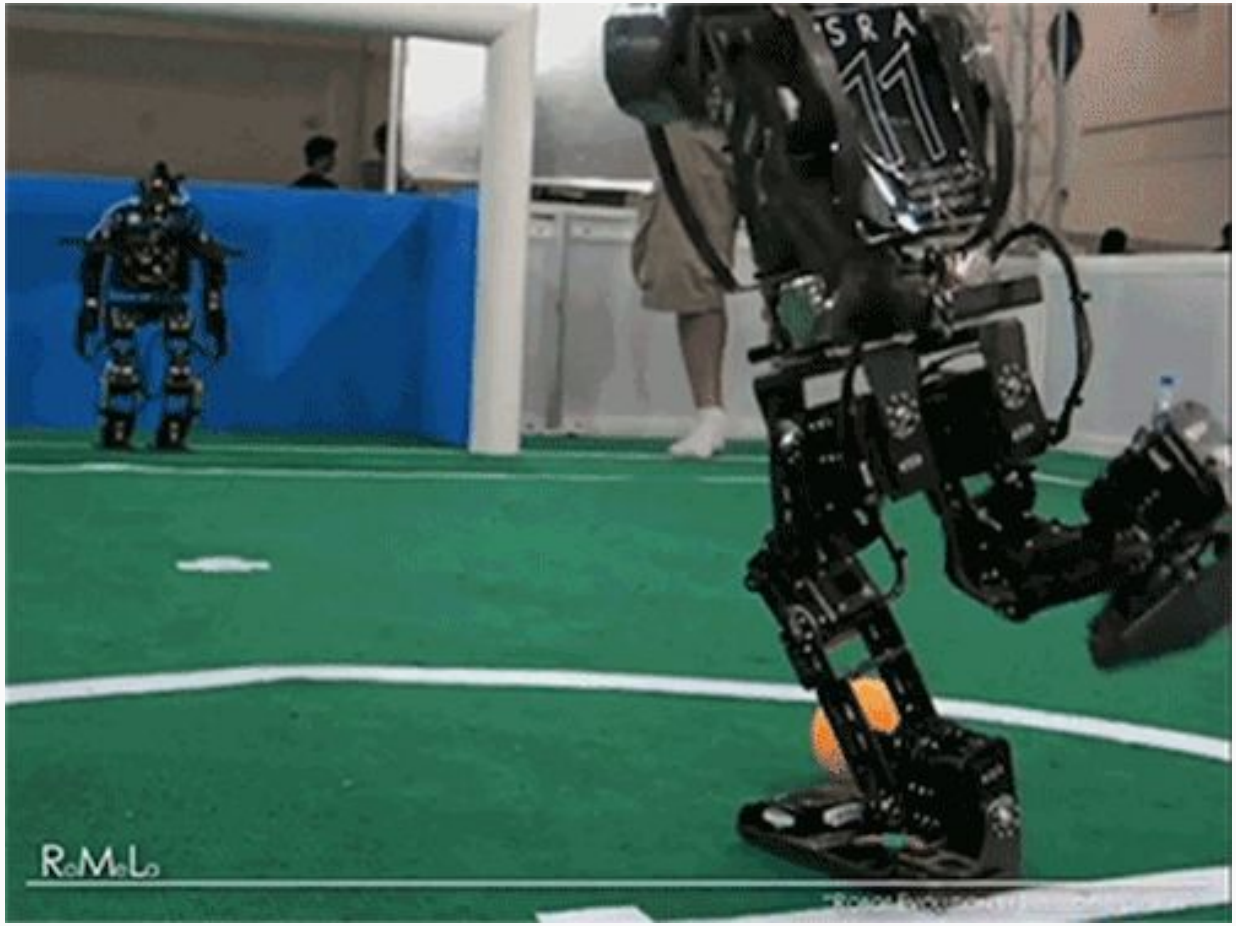
Aprendizagem de Máquina (ou como imitar a classificação humana!) - casos mais difíceis

Humanos:

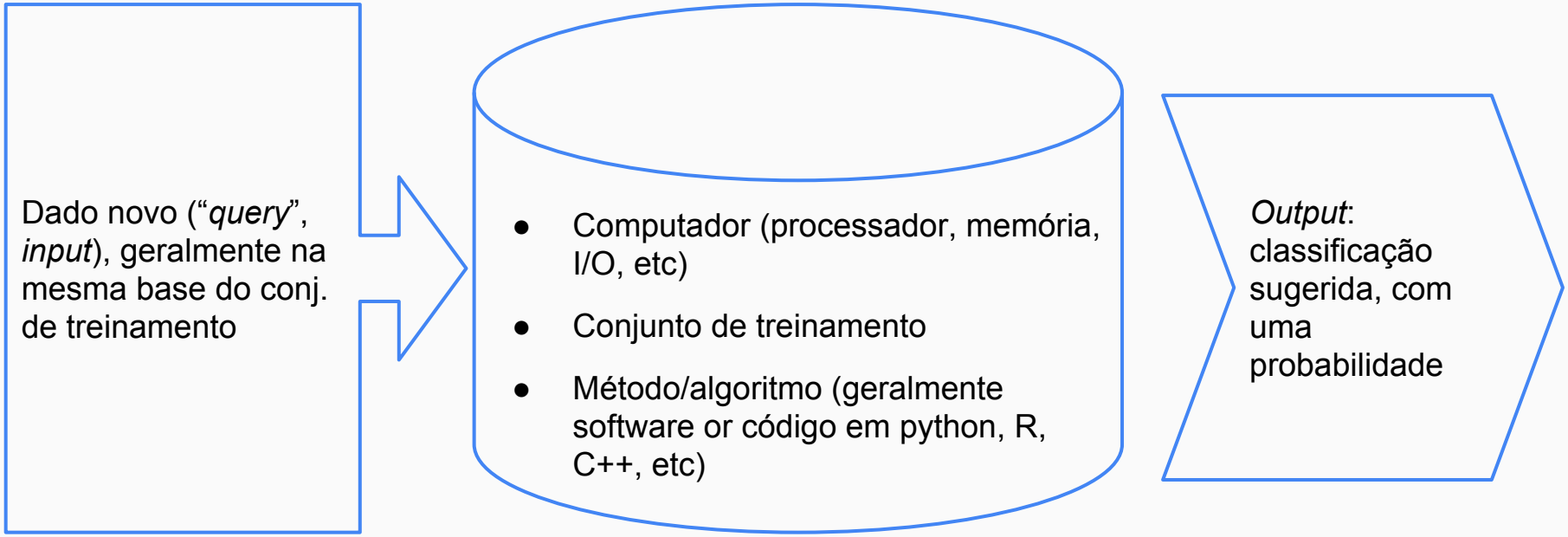
- **Outras relações, outras informações de sentidos, ambiente, etc:** por exemplo, uma xícara tão incomum que ficamos em dúvida! **Forma**: pode conter líquidos; **ambiente**: está numa mesa ou num armário de cozinha; **olfato**: sentimos cheiro de ervas, como um chá; etc

Máquinas:

- **??:** provavelmente, embutimos no **conjunto de treinamento** todas as informações que tivermos sobre o objeto, sem saber, a priori, se são úteis ou não. **Cuidado: isso pode aumentar *muito* o tempo de processamento.** Em astronomia, não é muito crítico, mas em robótica...


SRA
RoMeLo

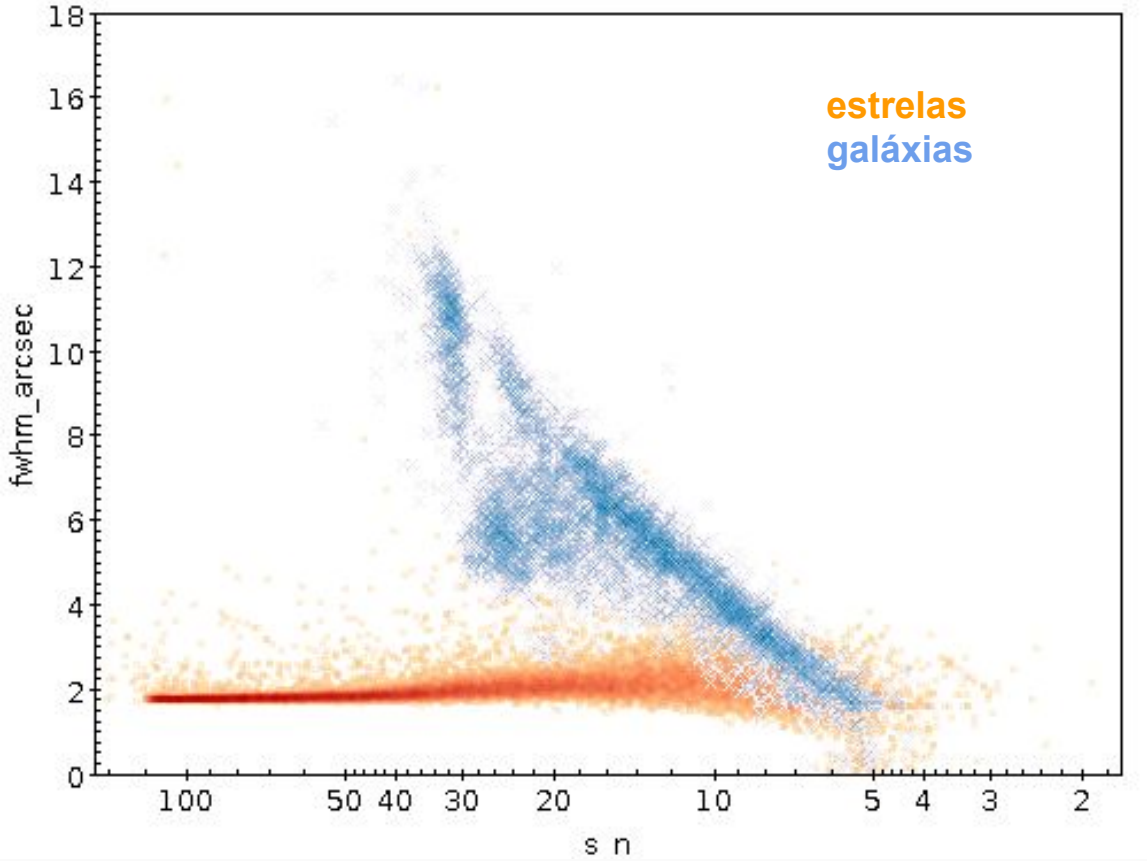
Esquema (receita) de um classificador de máquina



Classificar galáxias (hmm… galáxia ou estrela?)

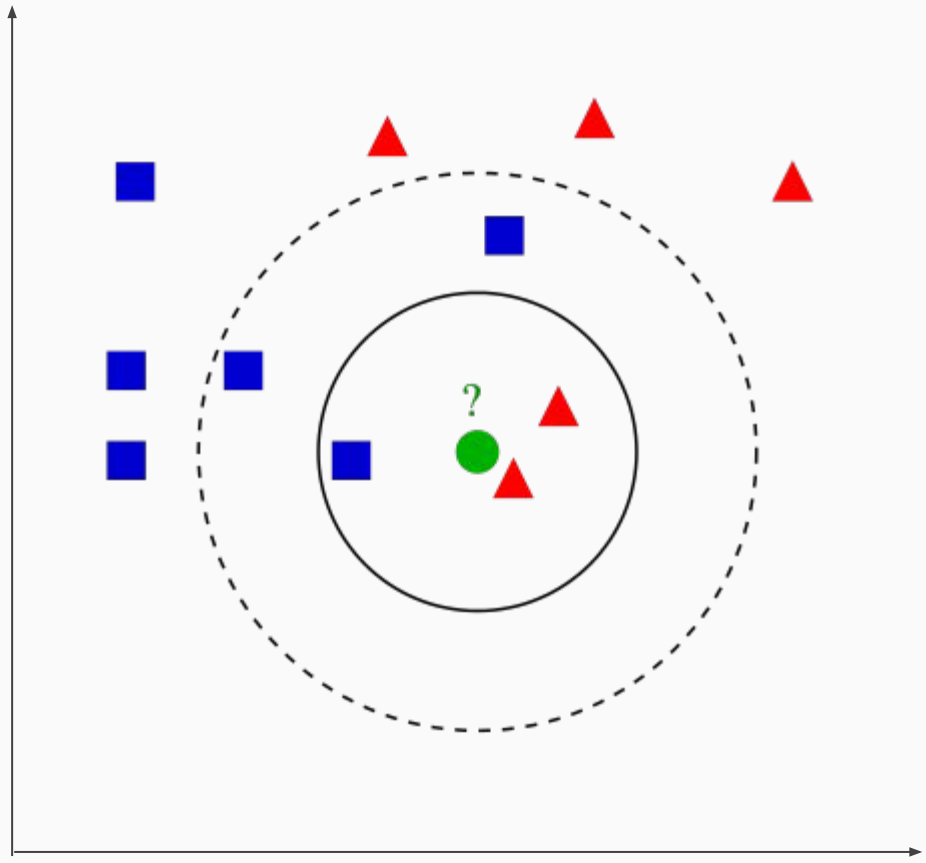
Classificação morfológica entre entrelas e galáxias

- Dados-Modelagem: medidas de **luminosidade**, **tamanho**, cores, etc.
- Quanto maior o objeto (>fwhm), maior a probabilidade de ser galáxia! Quanto menor, maior probabilidade de ser estrela!
- Quanto mais brilhante, maior a probabilidade de ser estrela! Quanto menos, maior probabilidade de ser galáxia!

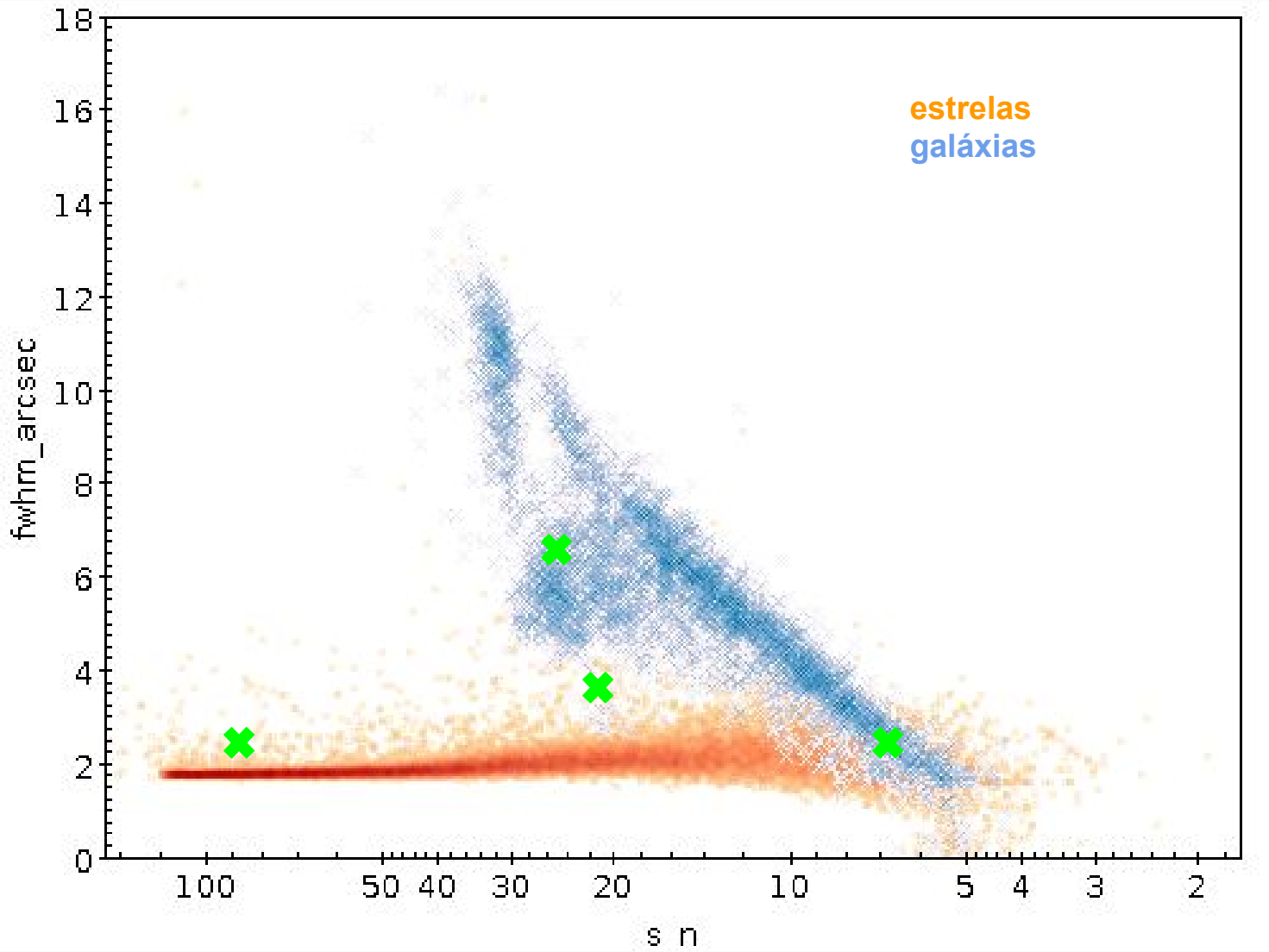


Método dos n-vizinhos mais próximos

- k-Nearest Neighbours (kNN)
- Ponto verde: *query*, queremos saber se é classificado como *triângulo vermelho* ou *quadrado azul*?
- Verificamos os pontos vizinhos, ao redor do ponto de *query*. Para um dado **raio** ou para um dado **número de vizinhos próximos**
- O resultado pode depender disso! Mas, em um exemplo real, os pontos do conj. De treinamento são muito mais concentrados!
- A probabilidade pode ser dada pela fração de vizinhos de classes diferentes.



Classificação morfológica entre entrelas e galáxias

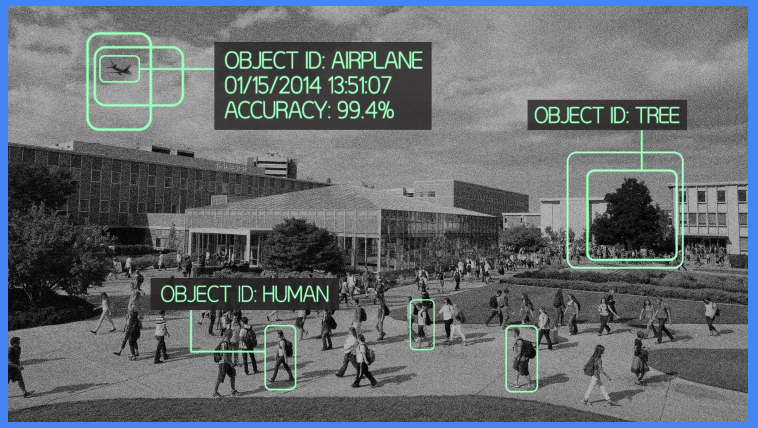


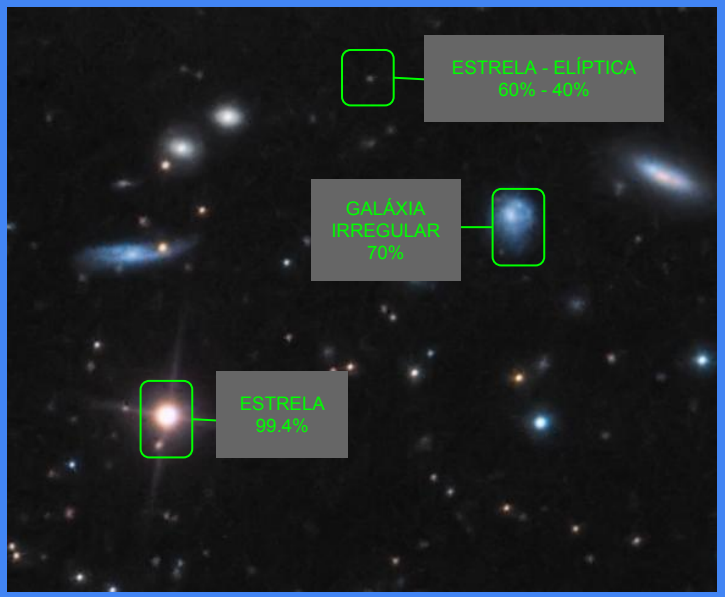
Problemas! (Humano = Máquina)

- Objetos fracos e/ou com pouca resolução; regiões onde não há uma divisão clara entre classes no conjunto de treinamento
- Diferentes orientações no céu;
- Cores nem sempre representativas: poeira, ‘deslocamento para o vermelha’ em galáxias distantes, etc;
- Quasares: ‘quasi-estrelas’ - aparecem como um ponto no céu como estrelas, mas são entendidas como núcleos de galáxias muito brilhantes, muito distantes;
- Categorias de não-classificação (o resto): artefatos, “objetos anômalos”, irregulares;

Astronomia pode e deve utilizar métodos computacionais e estatísticos atuais!

→ Mineração de dados, Aprendizagem de máquina, análise e ciência de dados em geral , métodos estatísticos para ‘big data’, visualização de dados, etc...





Obrigado!

Walter Santos (IAG/USP)
walter.augusto@gmail.com

Referências e agradecimentos

- Prof. Dr. Gastao B Lima Neto / Profa. Dra. Vera Jatenco / Prof. Dr. Laerte Sodré Jr.
- Dra. Emille Ishida / Dr. Rafael Souza
- Imagens: NASA/ESA/ESO/NAOJ/G. Paglioli
- Projeto GalaxyZoo
- Projeto SDSS / website SkyServer
- Todos os astrônomos que no passado trabalharam/desenharam(!) em classificações de galáxias/objetos
- Todos os engenheiros e cientistas de computadores que trabalharam(trabalham) em “aprendizagem de máquinas”, métodos, algoritmos, etc
- A. Szalay, https://www.youtube.com/watch?v=FlcdG4hUn1Q
- Ferramentas: Google, Anaconda python, TOPCAT, Aladin
- Sugestão de leitura:
    - *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*, por Željko Ivezić, Andrew J. Connolly, Jacob T. VanderPlas & Alexander Gray
    - *The Fourth Paradigm: Data-Intensive Scientific Discovery*, por Jim Gray