Data Analytics

CS301

Machine Learning:
K-Nearest Neighbours (KNN)

Week 13: 16 November

Fall 2021

Oliver BONHAM-CARTER

On Exam 2,
18th Nov 2021

- Starts 11:10am, finishes at 11:59pm (HARD DEADLINE)
- Work on exam using a GitHub repository (be sure to commit your work)
- **Open book but study your slides and notes**
- Choose your own Exam!
- You will be given a data set. Your grade will be assessed based on your ability to provide relevant questions of the data and then to provide convincing solutions using code for plots, models, or whatever you feel is necssary to respond to the question. Some leading questions will be provided
- For each question: you are to argue that your analysis answers your particular question.

On Exam 2,
18th Nov 2021

- Grading:
  - Inquiry basis and quality
  - Approach to resolving your inquiry
  - Conclusions and explanations
- Review your notes from the class
- Use whatever means *necessary* (according to you) to resolve your question using R code.
  - *Revealing plots*
  - *Basic stats*
  - *Summaries*
  - *Correlations*
  - *P*-values: t-tests, models, hypotheses
  - Other approaches

You’ve Got *This*!

Learning Relationships

THIS IS NOT WAR, IT'S EXTINCTION
TRANSFORMERS
AGE OF EXTINCTION
IN THEATRES, RealD 3D AND IMAX 3D
JUNE 27

Are these films all the same or what?

Machine Learning:
A Subset of Artificial Intelligence

- People learn from experiences
- Computers can also learn from “experiences”
- Computer program(s) with adaptive mechanisms that enable computer / machine to learn from historical data, experience, examples, analogy, rewards.

Types of ML?

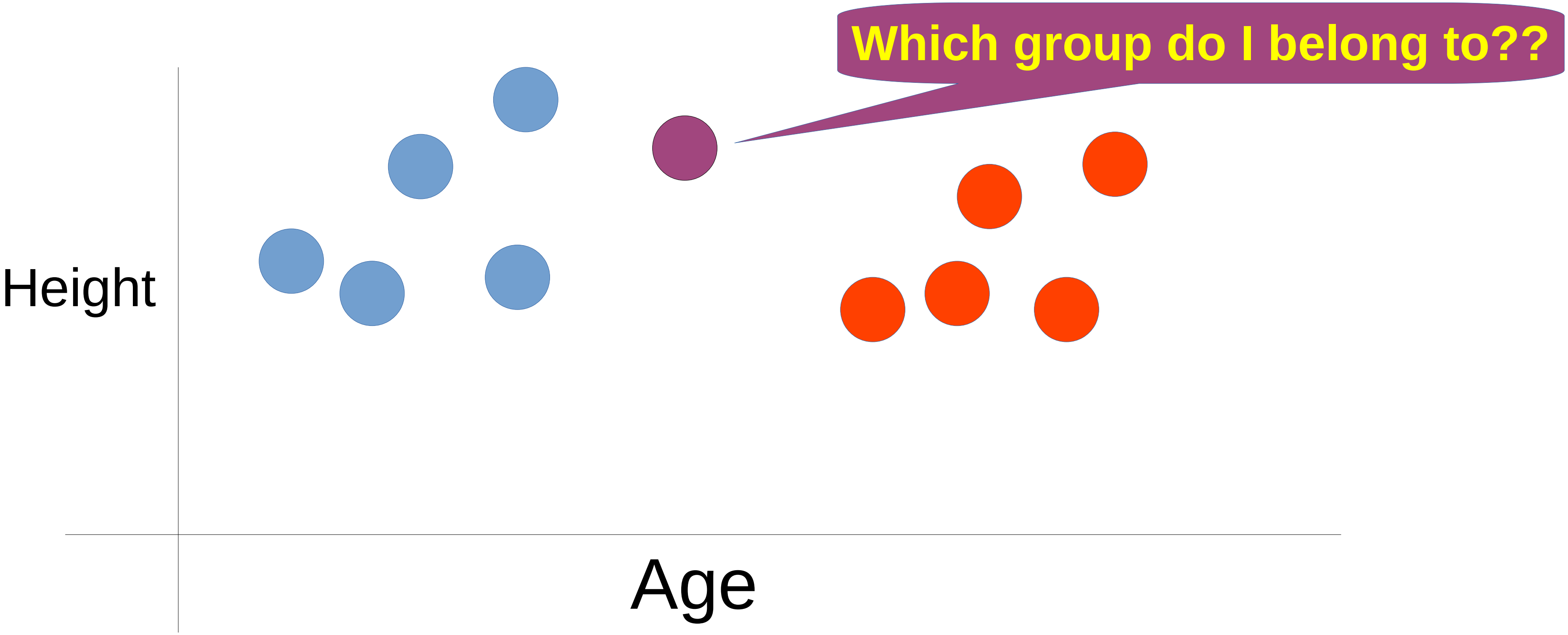
- **Supervised learning**
  - input-output relationships
  - The researcher knows the relationships she wants to find in data

- **Unsupervised learning**
  - relationship among inputs
  - The researcher discovers types of relationships in data

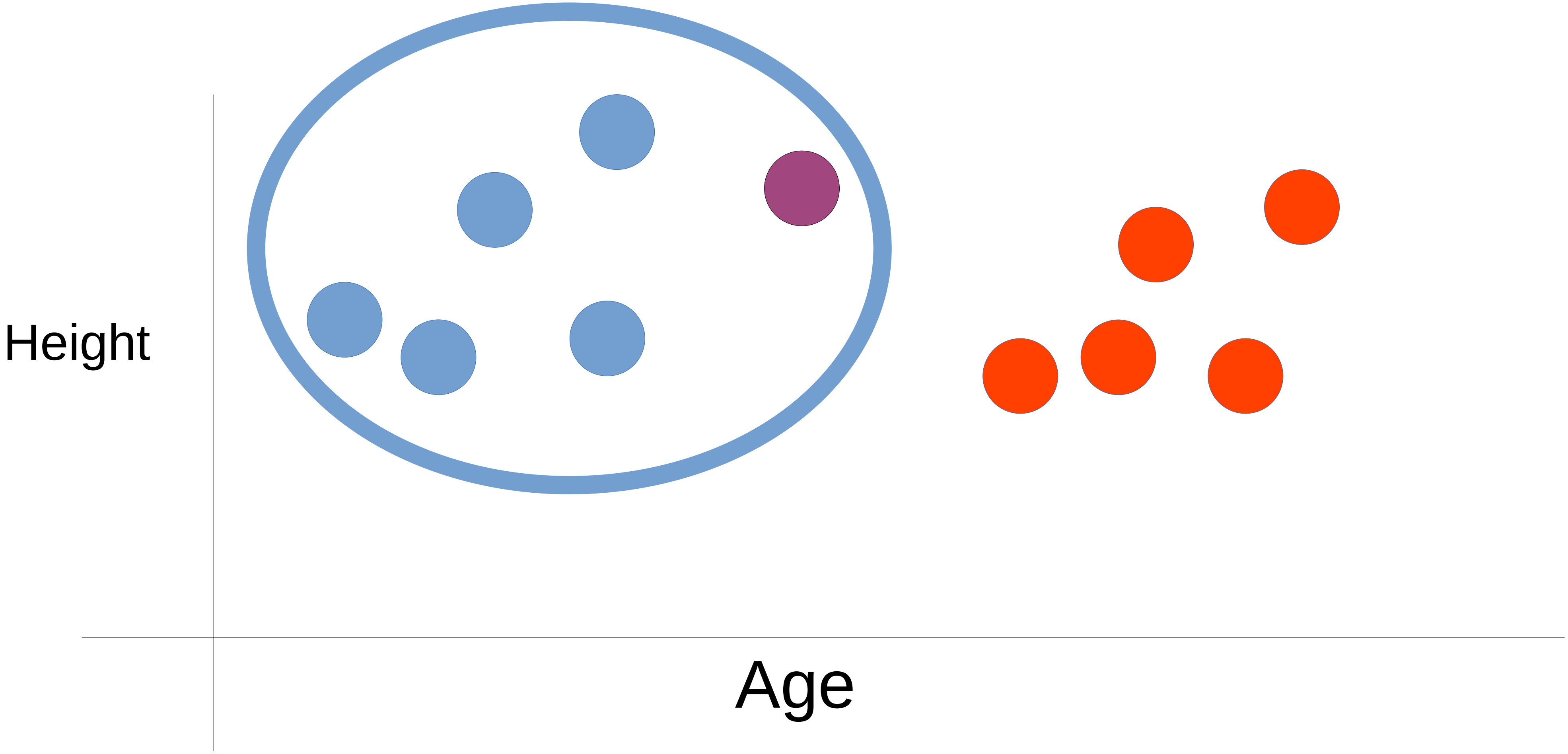
- **Reinforcement learning**
  - input-action relates to rewards / punishment
  - The algorithm earns points for hits and loses them for misses

K-Nearest Neighbours (KNN)

- **Classification problems**
  - I have a point, which group does this point belong to?
- **Regression problems**
  - I have a point, how to I estimate which group this point *would* land inside?

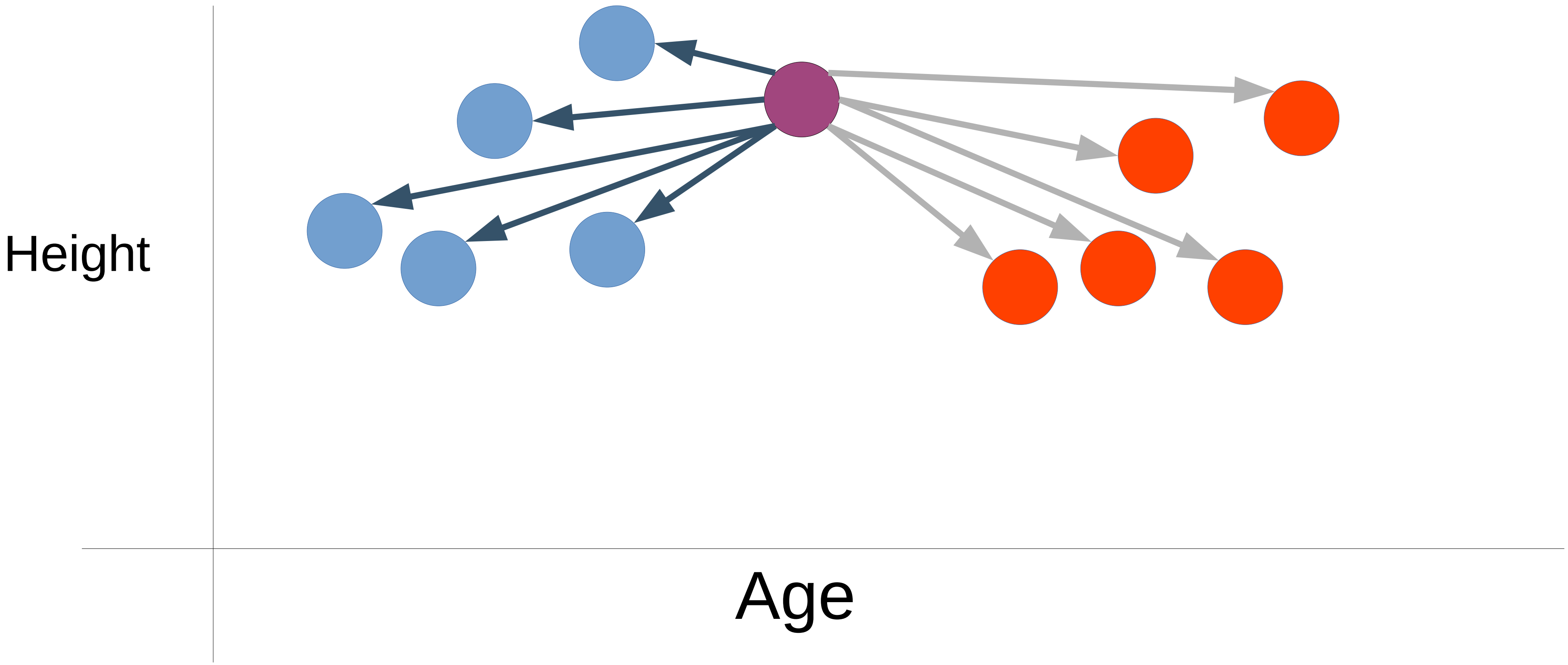


K-Nearest Neighbours (KNN)



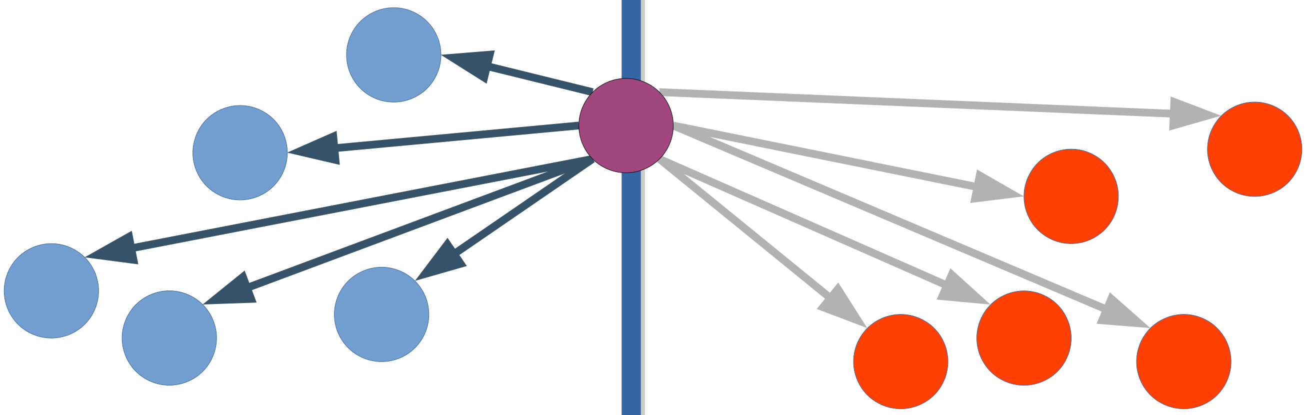
Feature Similarity

- Introduced points assigned weights based on their resemblance to an existing set of points



Feature Similarity

- Introduced points assigned weights based on their resemblance to an existing set of points

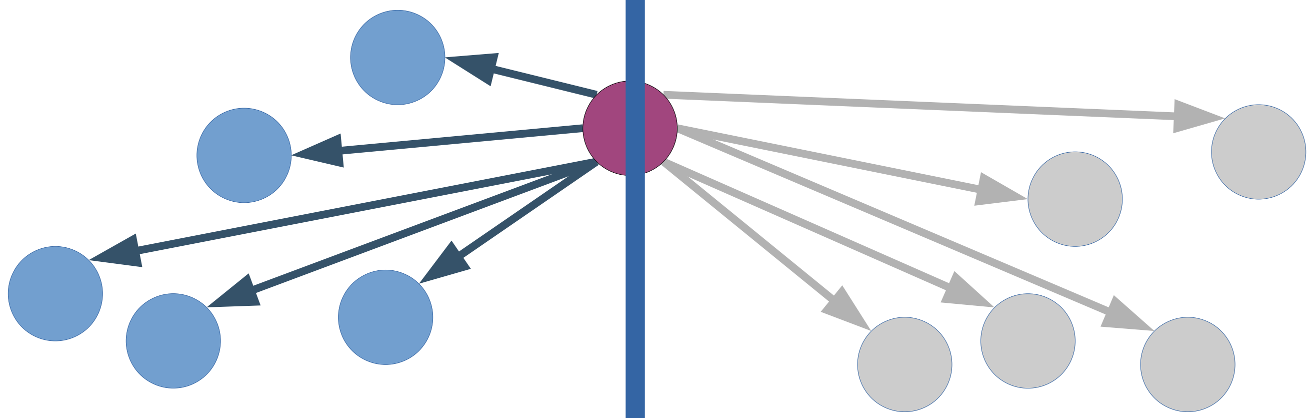


The blue lines are generally **shorter**. This means that the introduced point is likely closer to the blue set.

The grey lines are generally **longer**. This means that the introduced point is likely farther from the red set.

Okay, what’s the *K*?

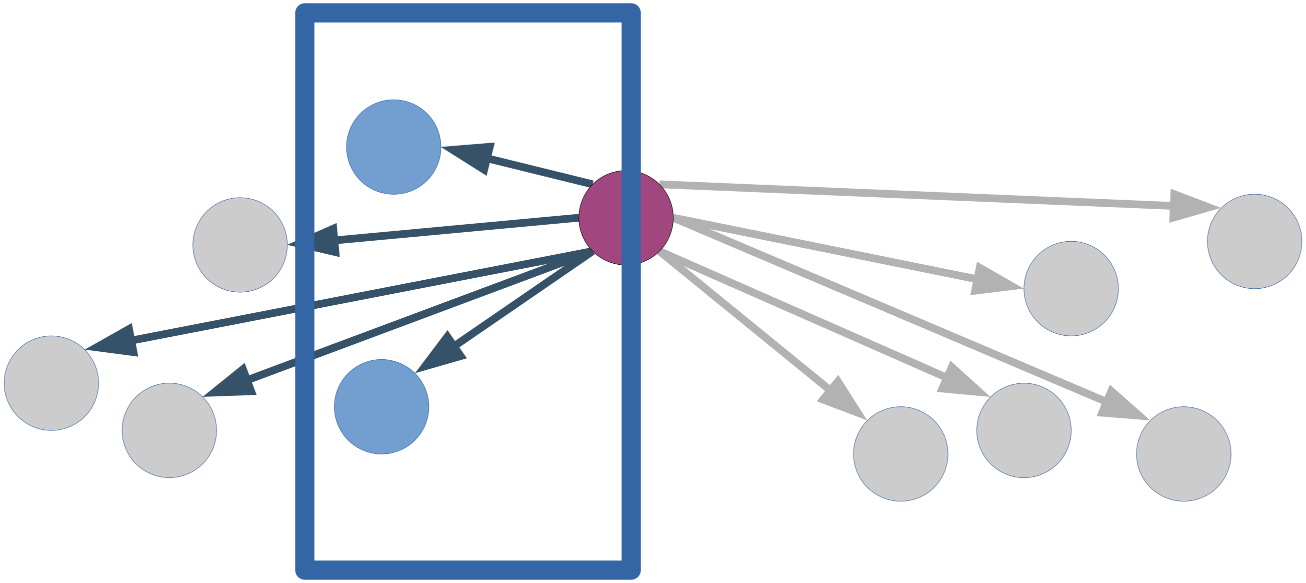
- The specified number of examples (*K*) closest to the query point.



For *K* = 5, the blue set has five points that make up the closest group for the query point

Okay, what's the *K*?

- At *K* = 2, we select only two points near the query point for the group.



For *K* = 2, the blue set has two points that make up the closest group for the query point

Let’s Code!

File: `sandbox/LM_iris.r`