Data Analytics with Python
DS102

HACKWAGON

•ACADEMY•

Week 5
Machine Learning

Week 5 Overview

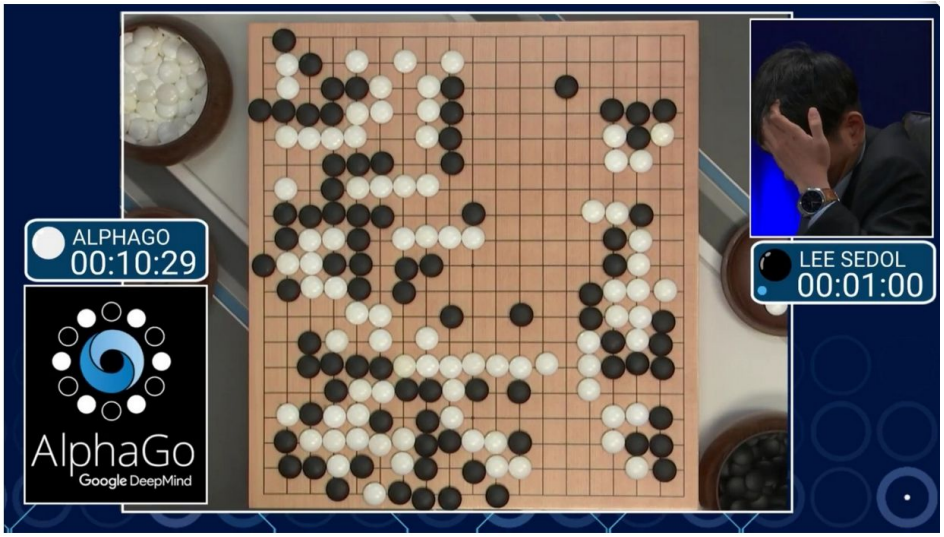
Today, we will learn about:

1. the Definition & Goals of Machine Learning
2. the applications of Machine Learning
3. the key concepts behind Machine Learning

HACKWAGON
•ACADEMY•

What is Machine Learning

Machine Learning is the ability for a computer program to perform a task **without being explicitly programmed**.



Humans vs. Machines
Top: OpenAI, Dota2
Left: AlphaGo, Go

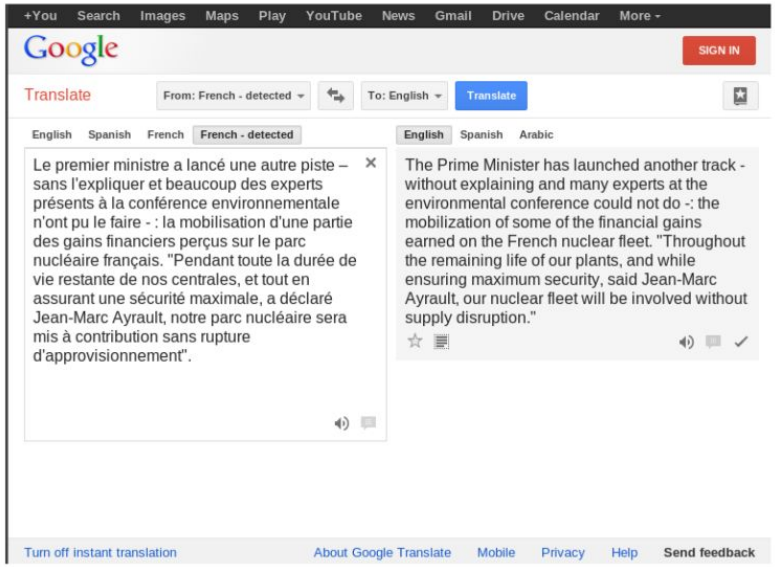
DotA2
https://www.theregister.co.uk/2018/08/24/openai_bots_eliminated_dota_2/
https://qz.com/1348177/why-are-ai-researchers-so-obsessed-with-games/

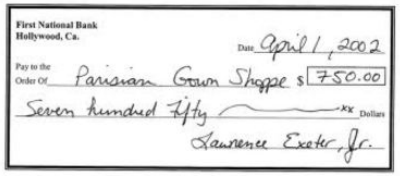
Jeopardy
https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/

AlphaGo
https://mashable.com/2017/10/19/google-alphago-zero/

HACKWAGON
• ACADEMY •

Applications of Machine Learning





- Handwriting Recognition
- Machine Translation
- Autonomous Driving

https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

HACKWAGON
• ACADEMY •

Applications of Machine Learning

- Fraud Detection
- Web Search
- Question & Answers
- Spelling correction
- Image recognition
- Music information retrieval
- Social network analysis
- Product recommendation
- Spam filtering
- Financial trading
- Medical diagnosis...

...and many more!

HACKWAGON
•ACADEMY•

What is Machine Learning

Machine Learning is the ability for a computer program to perform a task **without being explicitly programmed**.

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

HACKWAGON
•ACADEMY•

What is Machine Learning

A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, **improves with experience E**.

HACKWAGON
•ACADEMY•

Application of T, E, P

Problem: You are given the price and floor area of some houses in the North Region. Using linear regression, predict the price of a house that has yet to be valuated.

T (Task): Predict the price of a house
E (Experience): Linear Regression Model takes in existing houses' floor area as inputs and the value as the target variable.
P (Performance): Compare the price given by the model with the actual price of a new house

HACKWAGON
•ACADEMY•

Application of T, E, P - Your Turn

Problem: You are given a bunch of e-mails and they are labelled as spam or not spam (otherwise called ham). Using the Naïve Bayes Classification algorithm, a program attempts to classify if a new e-mail received is spam or ham.

T (Task): ?
E (Experience): ?
P (Performance): ?

HACKWAGON
•ACADEMY•

Application of T, E, P - Your Turn

Problem: You are given a bunch of e-mails and they are labelled as spam or not spam (otherwise called ham). Using the Naïve Bayes Classification algorithm, a program attempts to classify if a new e-mail received is spam or ham.

T (Task): Predict if a new e-mail is spam or ham given the sender and text of the e-mail
E (Experience): A classification model learns using the text and sender as the input variable and the tag (spam/ham) as the target variable
P (Performance): The percentage of new e-mails correctly classified as spam or ham.

HACKWAGON
•ACADEMY•

Classes of Machine Learning Problems

1. **Regression** - Predict a value of a new observation that is unknown to the model e.g. Predict housing prices

2. **Classification** - Predict the class of a new observation that is unknown to the model e.g. predict if an e-mail is spam or not spam

3. **Clustering** - Find out ways to group a bunch of data e.g. T-shirt sizing

4. **Anomaly Detection** - Find out if a datapoint is weird or not (e.g. cancer detection, terrorist profiling)

HACKWAGON
•ACADEMY•

Types of Machine Learning Problems

1. **Supervised Learning** - Where the learner already knows the output class / value of an observation given its inputs. Applied to: Regression, Classification, Anomaly Detection

2. **Unsupervised Learning** - Where the learner does not know the output class / value of the observation and tries to find patterns from the training set. Applied to: Clustering

HACKWAGON
•ACADEMY•

Terms

1. **Target Variable** - The intended class / value to predict given the training examples provided
2. **Input Features** - The set of variables that influence the target variable
3. **Training Set** - A dataset of **Training Examples** used to learn about how the inputs affect the target variable
4. **Hypothesis** - The function learnt by the algorithm to map the input features to the target variable

HACKWAGON
•ACADEMY•

Applied Example - Predict Housing Prices

1. **Target Variable** - The predicted price of a house
2. **Input Features** - The size, location, no. of bedrooms etc. of a house
3. **Training Set** - A historical set of housing price data, with the values of the input features
4. **Hypothesis** - The function that maps the variables to the predicted price of a house

HACKWAGON
•ACADEMY•

Worked Example

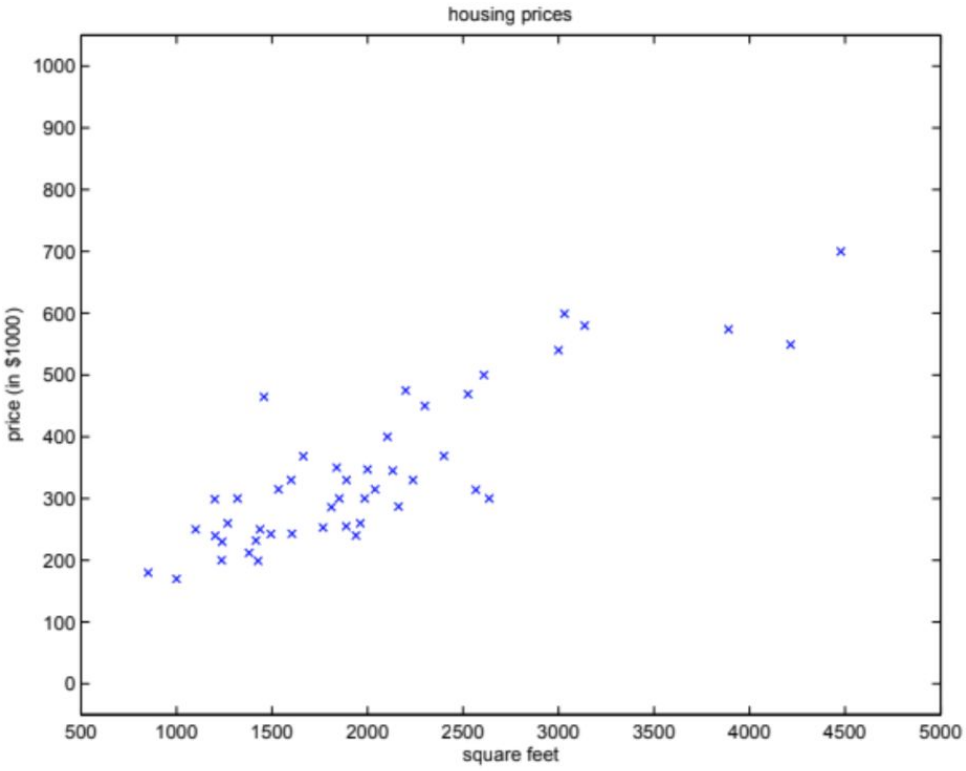
Visit Susan Li, TowardsDataScience

1. Which class of ML does this belong to?
2. Is this a type of supervised or unsupervised learning?
3. What is the target variable and the suggested input features?

HACKWAGON
• ACADEMY •

Linear Regression (In-Class A)

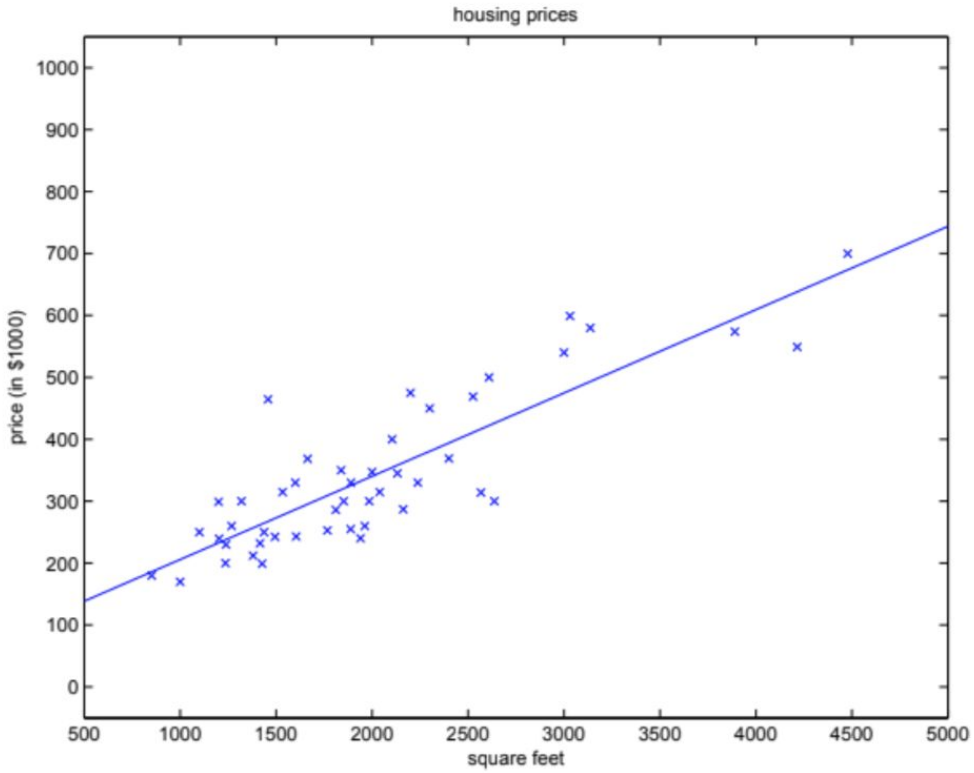
Objective:
Predict the housing prices (Target Variable) using the Input Variables (Housing Floor Area)



HACKWAGON
• ACADEMY •

Linear Regression

Strategy: Use OLS (Ordinary Least Squares) to fit a line to minimise the squared errors



HACKWAGON
• ACADEMY •

K-Means Clustering (In-Class B)

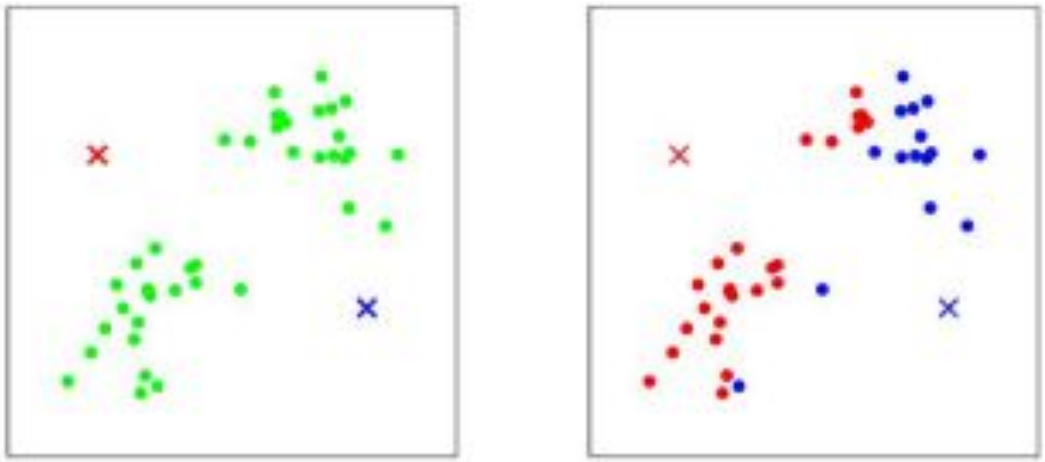
Objective:
Given a set of data points, identify the clusters that they are divided to.

HACKWAGON
• ACADEMY •

K-Means Clustering (In-Class B)

Strategy:

1. Initialise 2 centroids at random points.
2. If a point is nearer to the red centroid, assign it class red. Otherwise if it is nearer to the blue centroid, classify it blue.

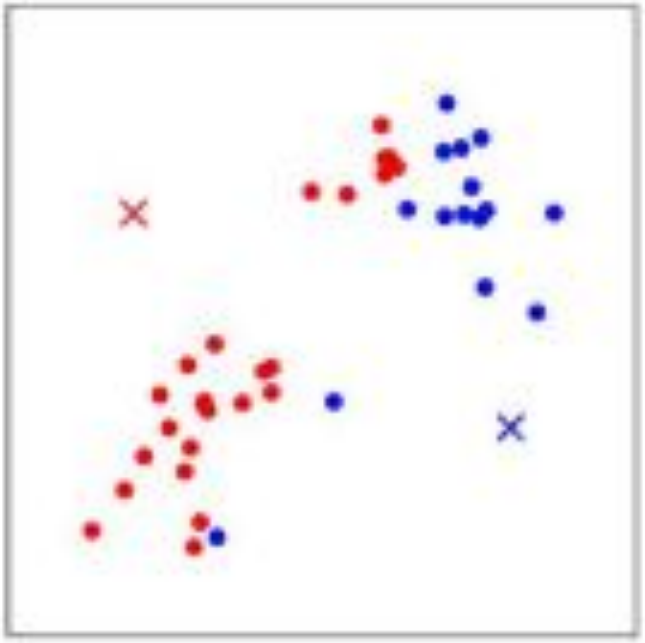


(b) (c)

HACKWAGON
• ACADEMY •

K-Means Clustering (In-Class B)

Observe that after the first round, some points are wrongly clustered.

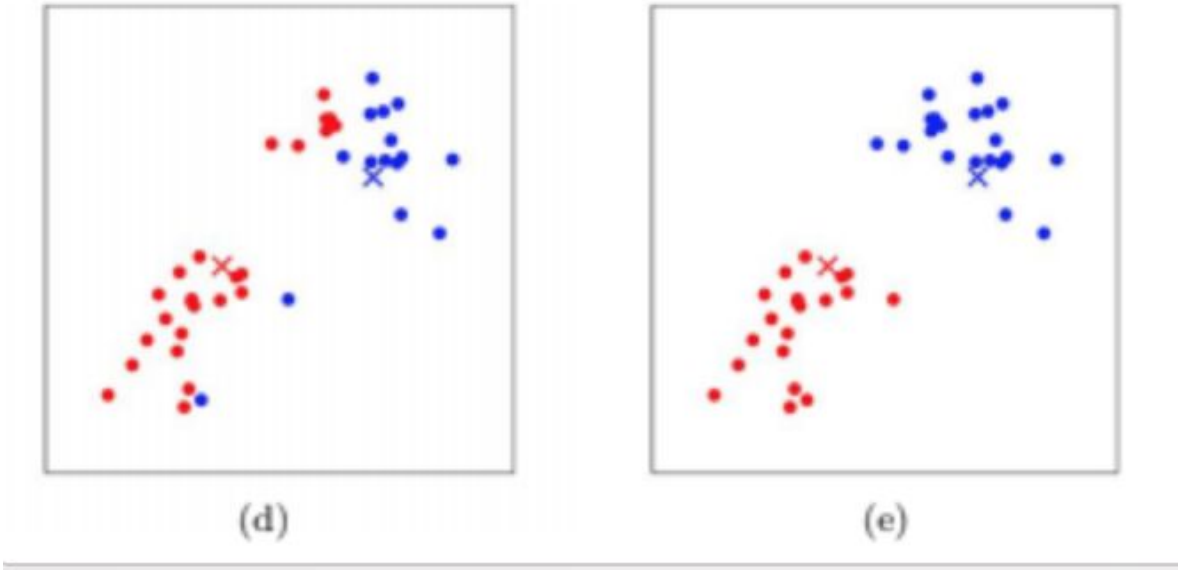


(c)

HACKWAGON
• ACADEMY •

K-Means Clustering (In-Class B)

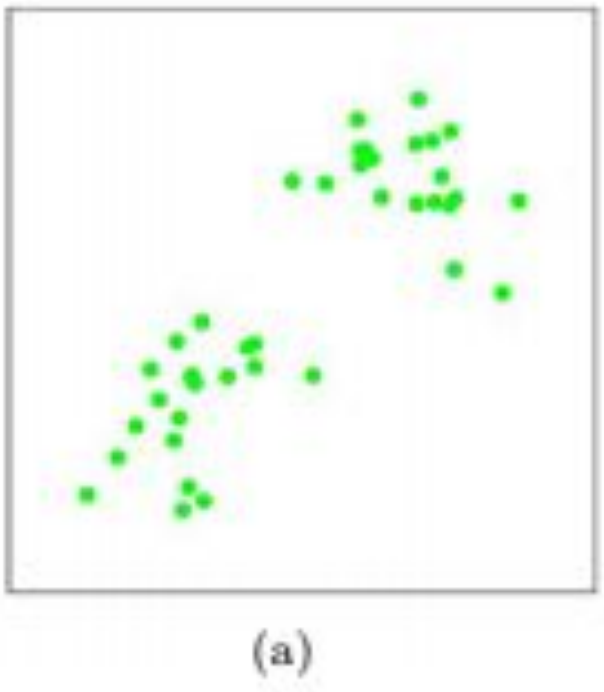
Then, move the centroid to the center of each cluster. Now, perform the assignment to centroids again (If a point is nearer to red , assign red and if a point is nearer to blue, assign blue.)



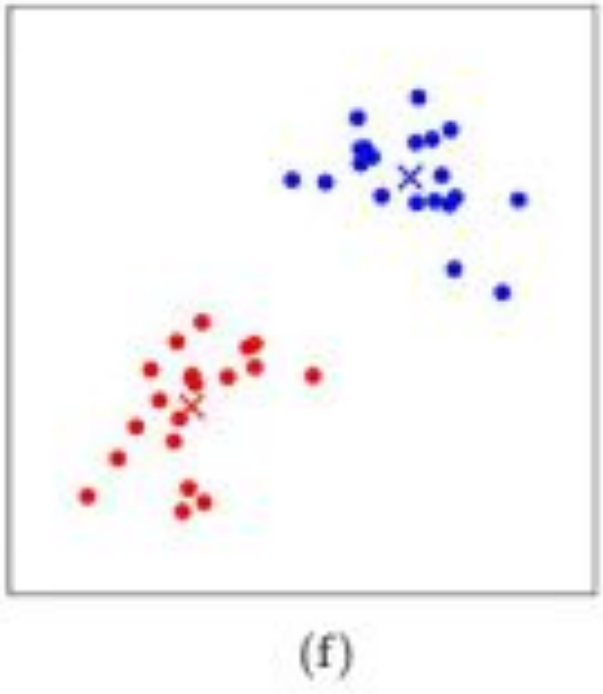
HACKWAGON
• ACADEMY •

K-Means Clustering (In-Class B)

Finally, re-center the centroids again. This is the final result after 2 iterations.



before



after

HACKWAGON
• ACADEMY •

Decision Trees (In-Class D)

Objective:
Given a dataset with features and labels, divide them into “pure” subsets based on each feature-label pair.

HACKWAGON
• ACADEMY •

Decision Trees

3 with diabetes, 2 without

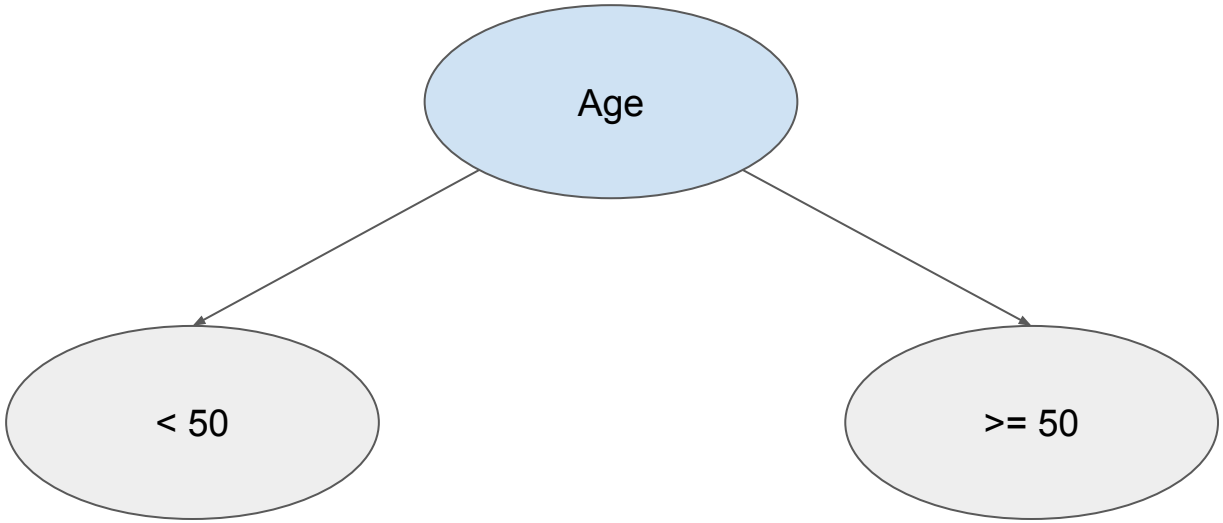
	Age	Insulin	Glucose	Outcome
0	88	141	183	1
1	67	100	175	1
2	21	71	0	0
3	58	86	71	0
4	39	160	175	1

New row: Outcome?

	Age	Insulin	Glucose
5	40	102	165

HACKWAGON
• ACADEMY •

Decision Trees



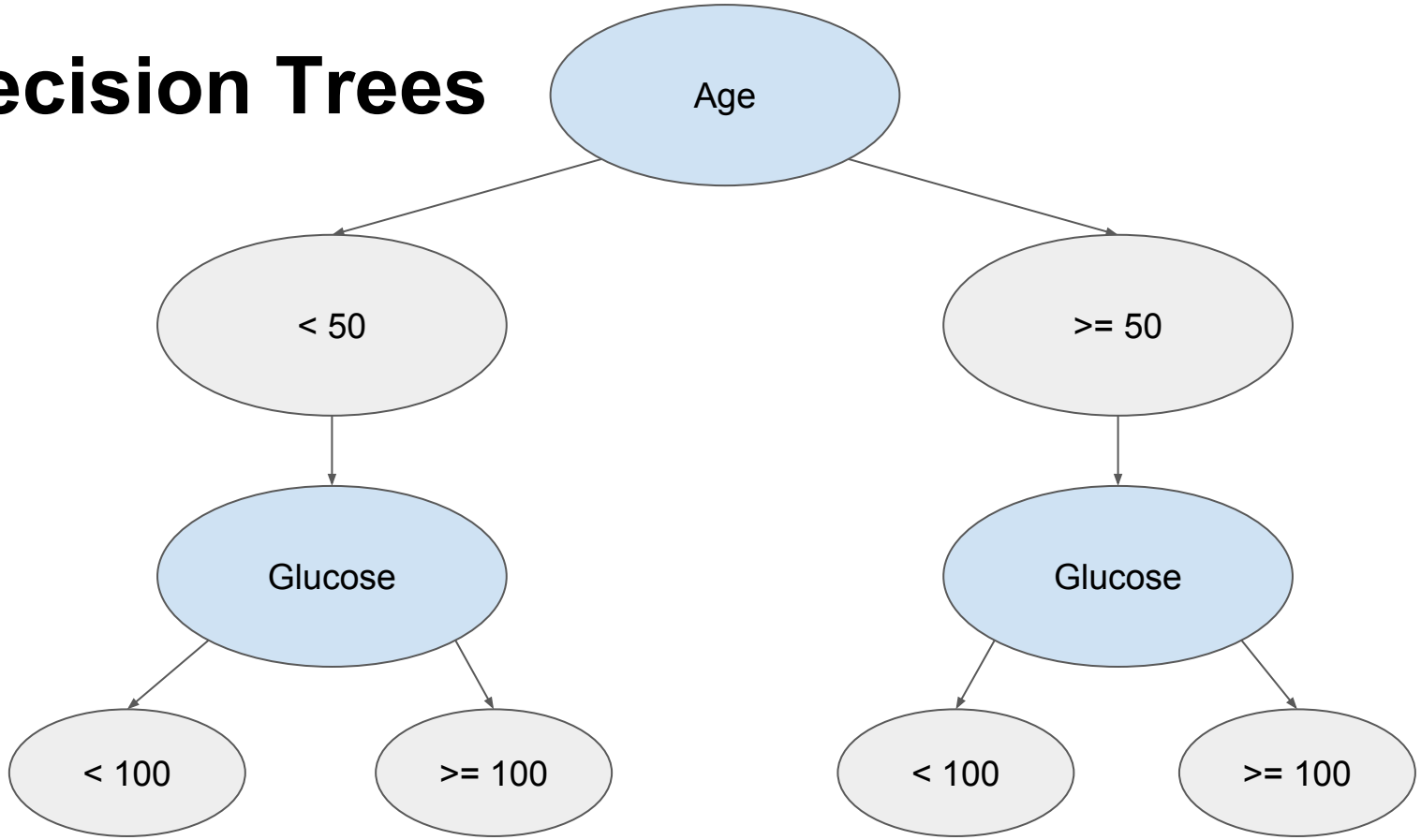
	Age	Insulin	Glucose	Outcome
2	21	71	0	0
4	39	160	175	1

	Age	Insulin	Glucose	Outcome
0	88	141	183	1
1	67	100	175	1
3	58	86	71	0

Not pure. Continue splitting by features!

HACKWAGON
• ACADEMY •

Decision Trees



	Age	Insulin	Glucose	Outcome
2	21	71	0	0

	Age	Insulin	Glucose	Outcome
4	39	160	175	1

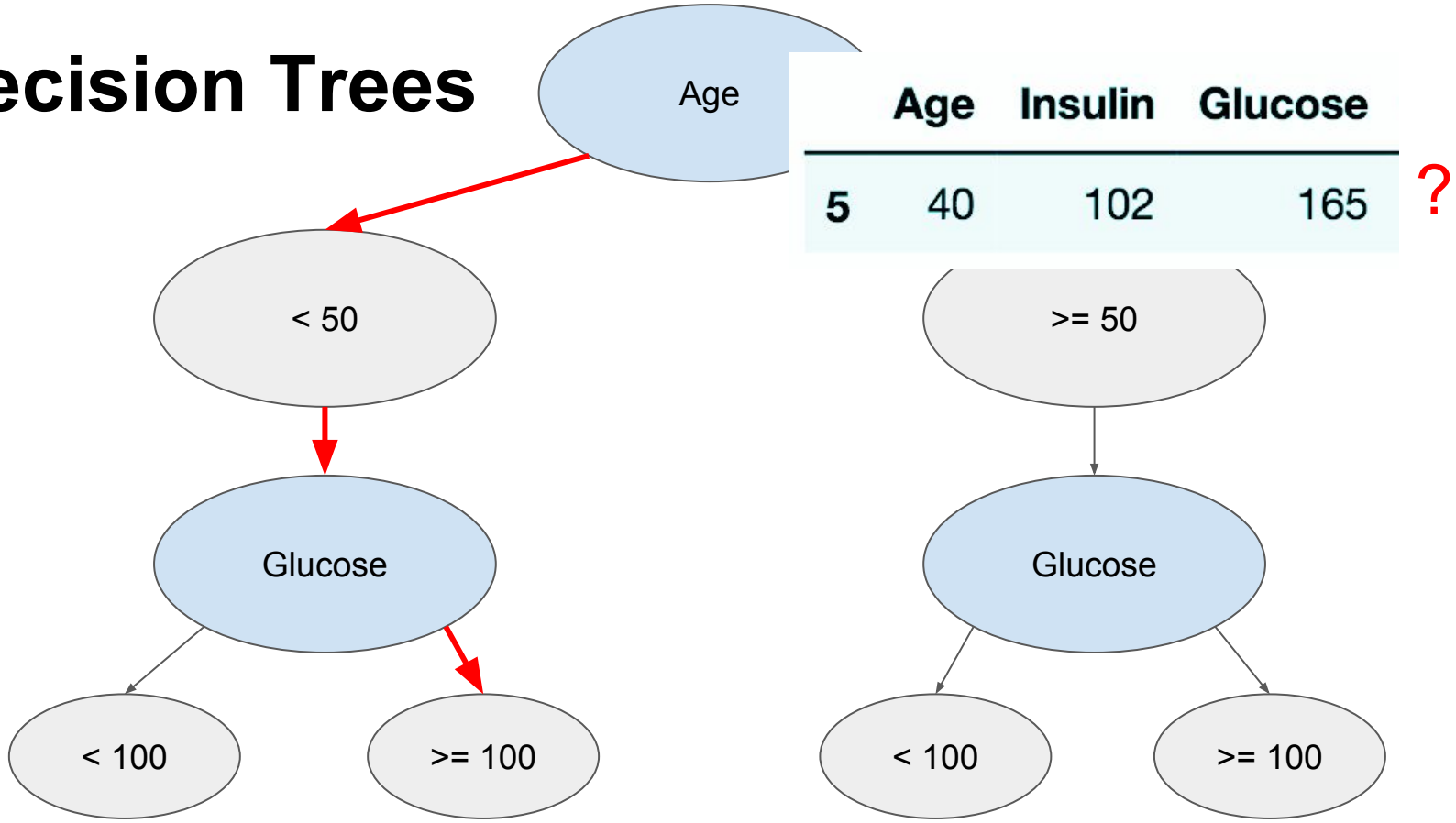
	Age	Insulin	Glucose	Outcome
3	58	86	71	0

	Age	Insulin	Glucose	Outcome
0	88	141	183	1
1	67	100	175	1

Pure subsets. Stop splitting!

HACKWAGON
• ACADEMY •

Decision Trees



	Age	Insulin	Glucose
5	40	102	165

?

	Age	Insulin	Glucose	Outcome
2	21	71	0	0

	Age	Insulin	Glucose	Outcome
4	39	160	175	1

	Age	Insulin	Glucose	Outcome
3	58	86	71	0

	Age	Insulin	Glucose	Outcome
0	88	141	183	1
1	67	100	175	1

Follow the path

HACKWAGON
• ACADEMY •

Credits

CS229: Machine Learning, Stanford University
CS221: Artificial Intelligence: Principles & Techniques, Stanford University
Deep Learning, Goodfellow, Bengio, Courville, 2016

HACKWAGON
• ACADEMY •