GENERAL ASSEMBLY

Decision Trees

Tui Britton
DSi Instructor, GA Sydney

Learning Objectives

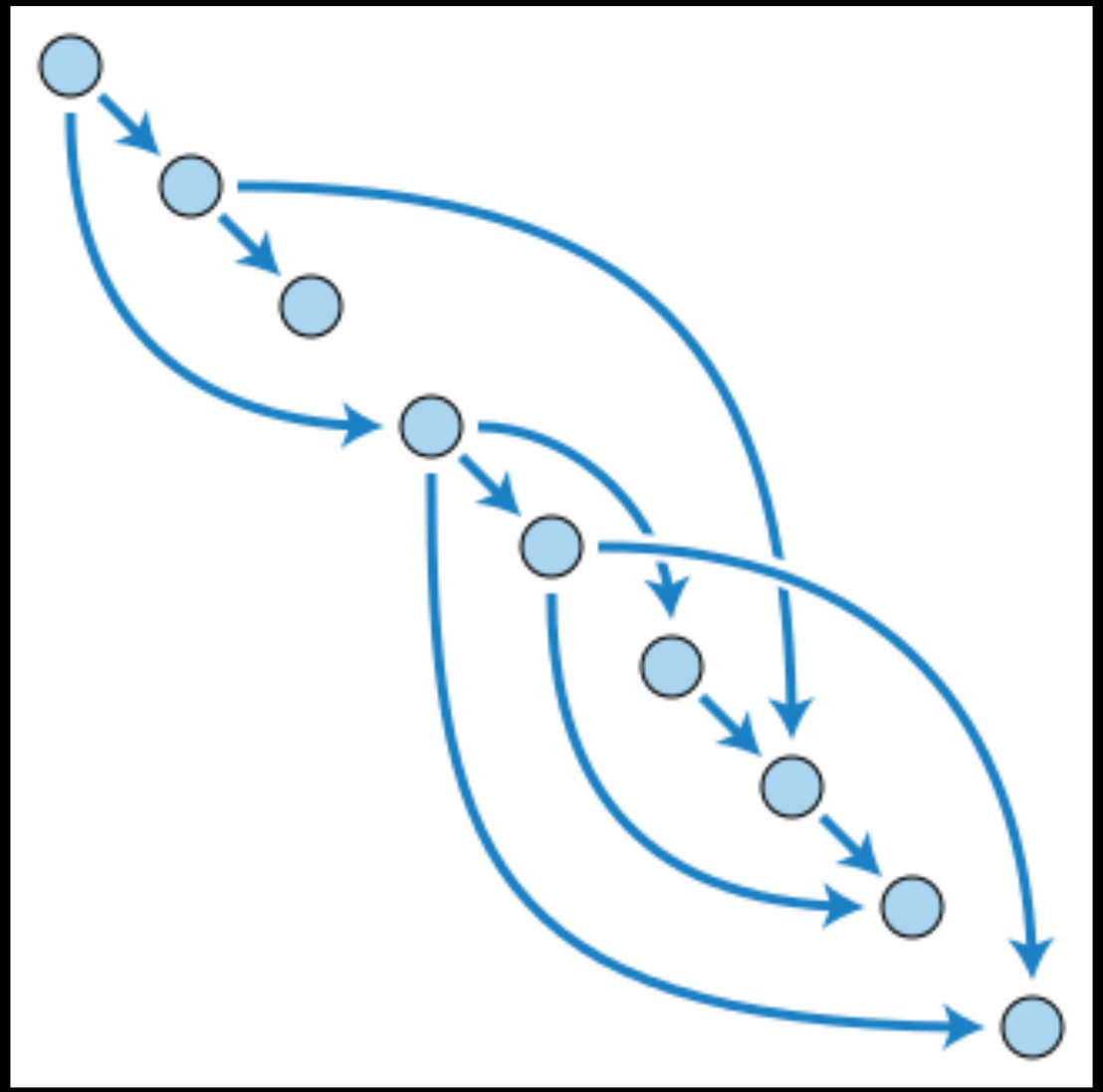
What are they?

Why do we use them?

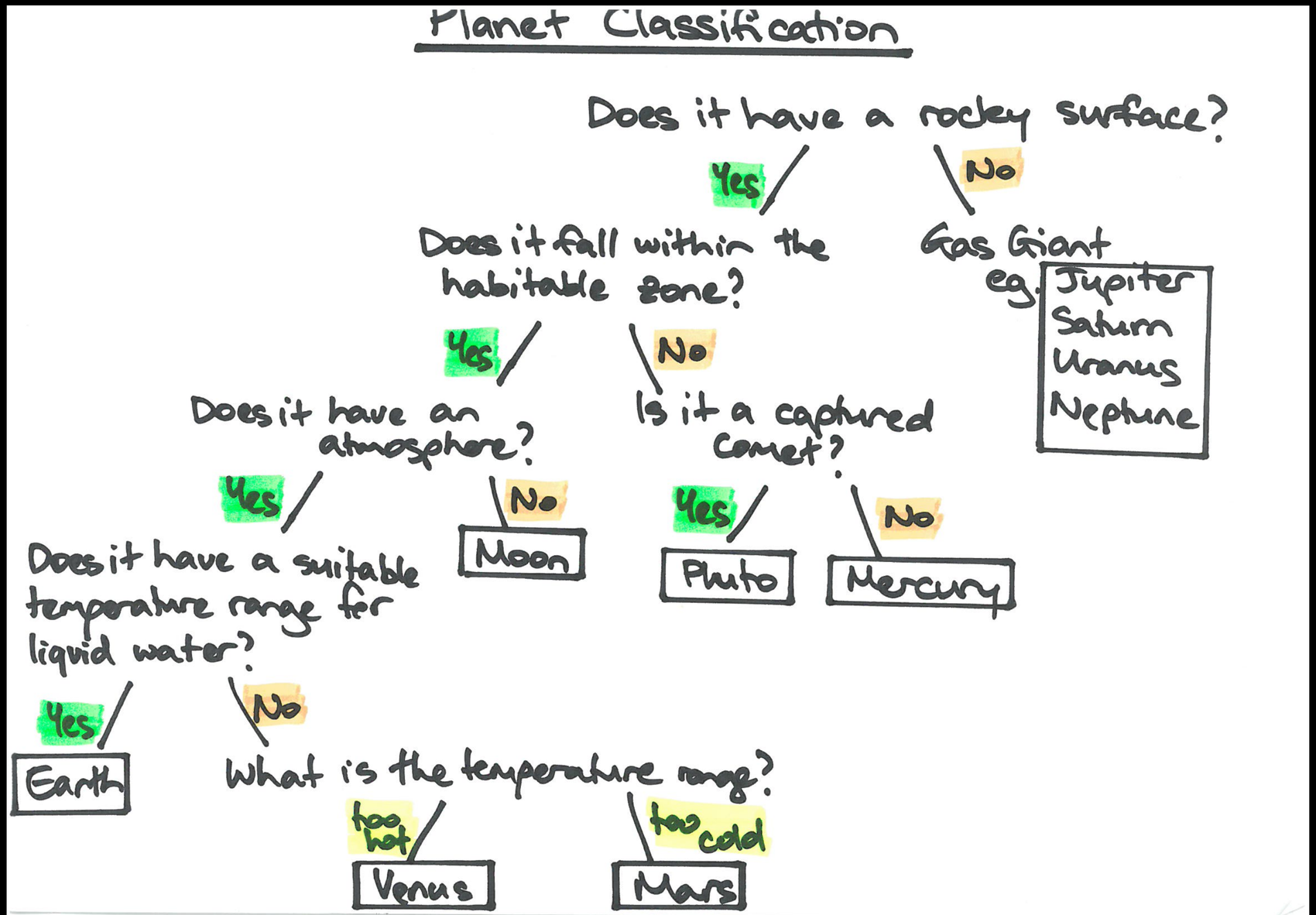
How do we use them?

What is my favourite animal?

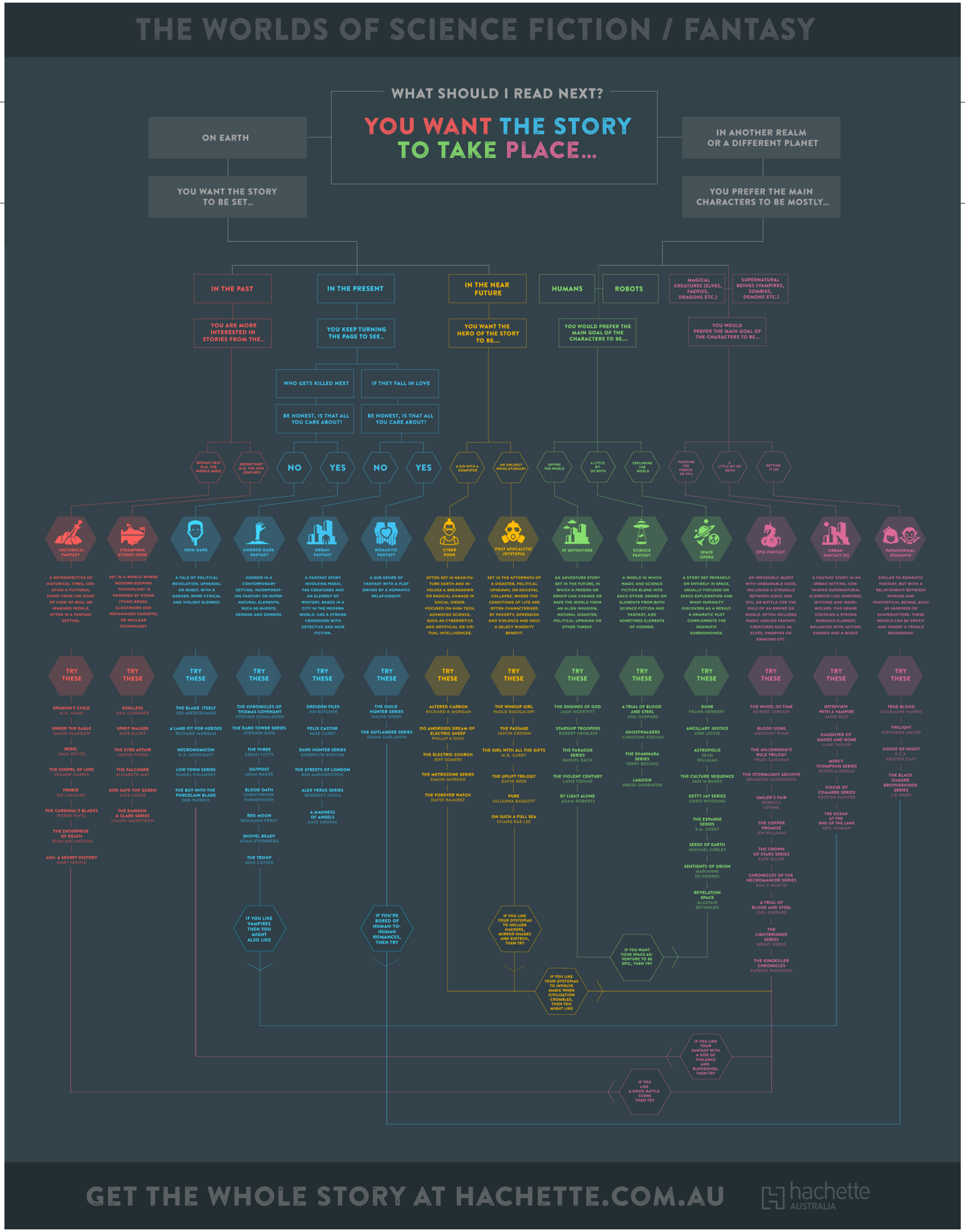
Topological map – not cyclic!



Classification



Outcome



THE WORLDS OF SCIENCE FICTION / FANTASY
WHAT SHOULD I READ NEXT?
YOU WANT THE STORY TO TAKE PLACE...
ON EARTH
YOU WANT THE STORY TO BE SET...
IN ANOTHER REALM OR A DIFFERENT PLANET
YOU PREFER THE MAIN CHARACTERS TO BE MOSTLY...
IN THE PAST
IN THE PRESENT
IN THE NEAR FUTURE
HUMANS
ROBOTS
YOU ARE MORE INTERESTED IN STORIES FROM THE...
YOU KEEP TURNING THE PAGE TO SEE...
YOU WANT THE HERO OF THE STORY TO BE...
YOU WOULD PREFER THE MAIN GOAL OF THE CHARACTERS TO BE...
WHO GETS KILLED NEXT
IF THEY FALL IN LOVE
BE HONEST, IS THAT ALL YOU CARE ABOUT?
BE HONEST, IS THAT ALL YOU CARE ABOUT?
NO
YES
NO
YES
GET THE WHOLE STORY AT HACHETTE.COM.AU

The dog chased the cat

Regression

Different from classification:
Determines “badness” of each node by summing up the overall reduction of optimization criteria like SSE.

linear regression models —> calculate the coefficients of predictors

tree regression models —> calculate the relative importance of predictors

Coding

`sklearn.tree.`DecisionTreeClassifier
The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain.

`sklearn.tree.`DecisionTreeClassifier
#The function to measure the quality of a split. Supported criteria are “mse” for the mean squared error, which is equal to variance reduction as feature selection criterion, and “mae” for the mean absolute error

Metrics

Impurity measurements:
Gini impurity
- is a measure of how often a randomly chosen element from the set would be misclassified

Information gain (entropy)
- used to decide which feature to split on in each step
- simplest is best
- good for exploratory analysis

Mean Squared Error

Mean Absolute Error

Summary

What?
- Visual tool for decision making

Why?
Useful for predicting:
- class to which data belong —> classification
- value of a target variable —> regression

How?
- by learning simple decision rules inferred from the input data features

Cons:
- Prone to design error i.e. asking the wrong questions!
- Can be very complex i.e. too deep
- Can overfit by learning irregular patterns

Topics for future investigation

- Boosted trees
- Bootstrap aggregating
- Random forest
- Rotation forest
- Conditional inference tree