What is Data Science?

- **General Definition:** Processes and systems to extract knowledge or insights from data, either structured or unstructured. *(Wikipedia)*

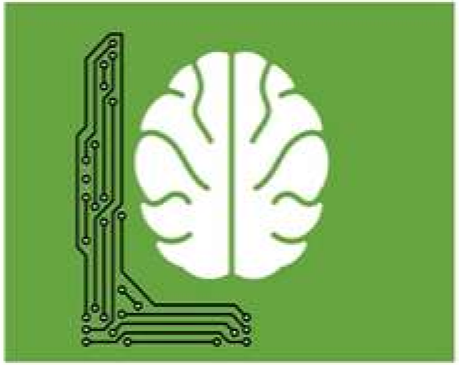
- **For the purposes of this course:** Managing, analyzing, and visualizing data in support of the Machine Learning workflow.

- But **what is Machine Learning?**

What is Machine Learning?

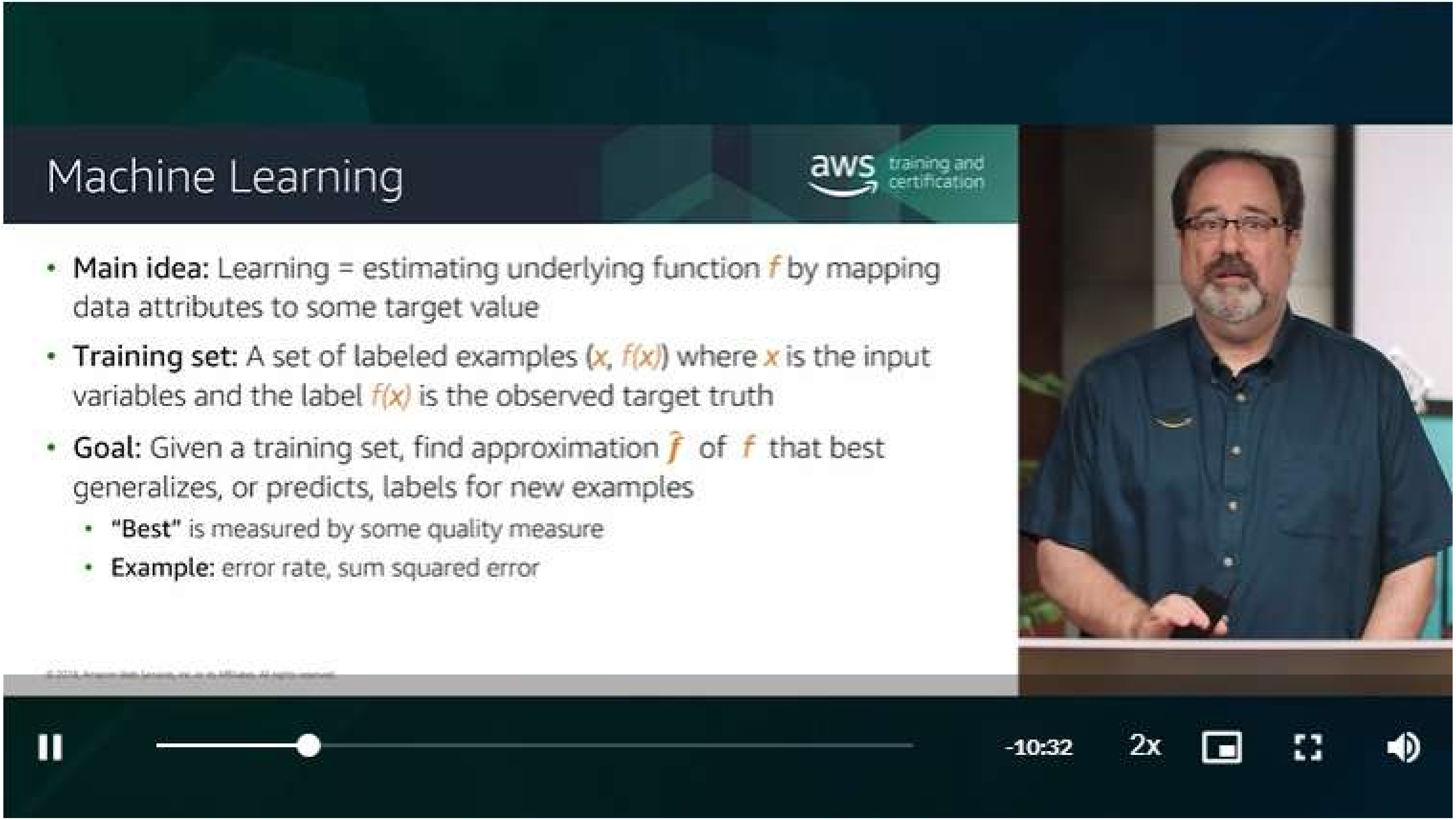
aws training and certification

Artificial Intelligence machines that improve their predictions by learning from large amounts of input data.



Machine Learning

- **Main idea:** Learning = estimating underlying function f by mapping data attributes to some target value
- **Training set:** A set of labeled examples (x, $f(x)$) where x is the input variables and the label $f(x)$ is the observed target truth
- **Goal:** Given a training set, find approximation $\hat{f}$ of f that best generalizes, or predicts, labels for new examples
  - **"Best"** is measured by some quality measure
  - **Example:** error rate, sum squared error

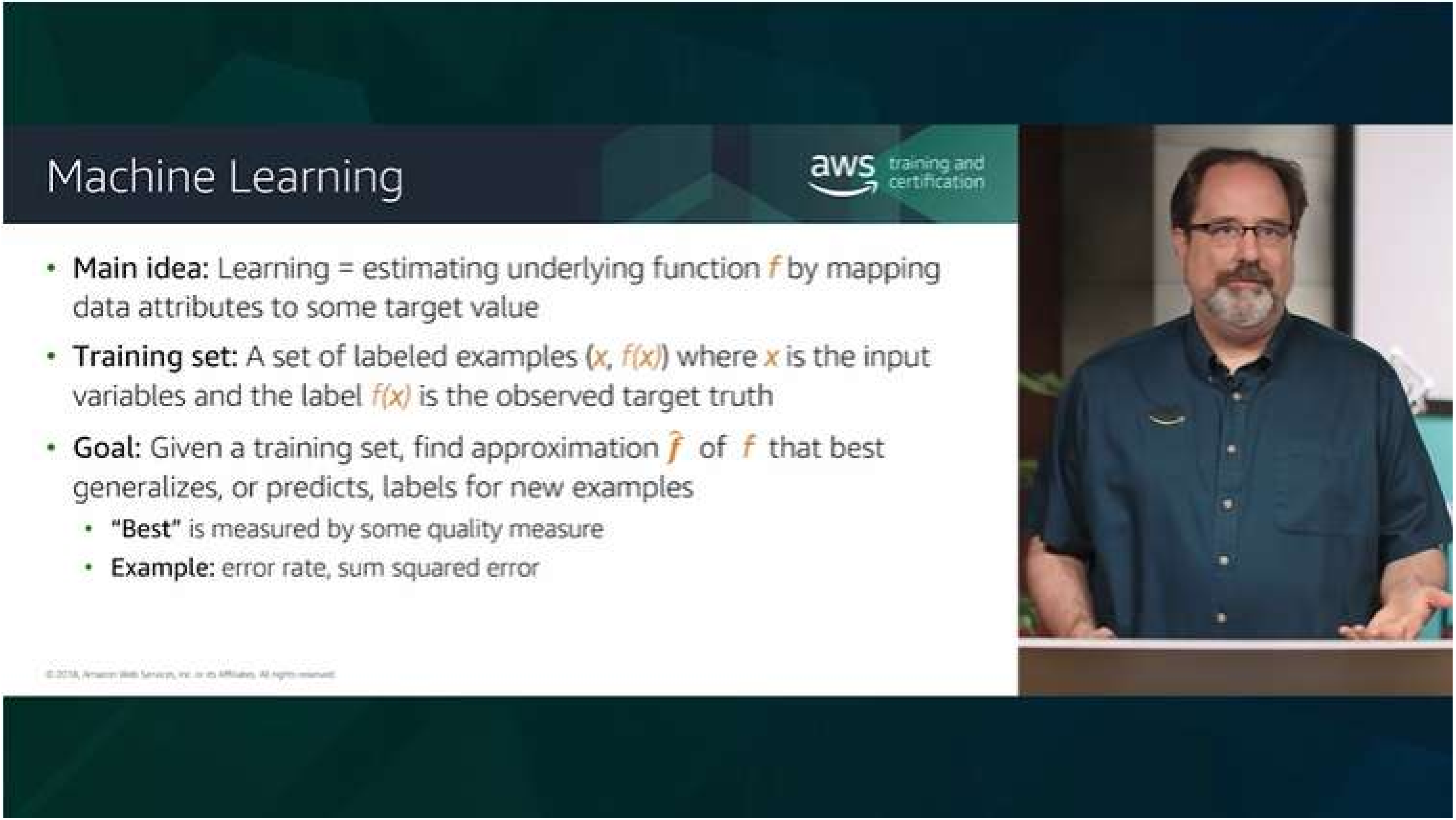


Machine Learning

- **Main idea:** Learning = estimating underlying function *f* by mapping data attributes to some target value
- **Training set:** A set of labeled examples (*x*, *f(x)*) where *x* is the input variables and the label *f(x)* is the observed target truth
- **Goal:** Given a training set, find approximation $\hat{f}$ of *f* that best generalizes, or predicts, labels for new examples
  - **"Best"** is measured by some quality measure
  - **Example:** error rate, sum squared error

Machine Learning

- **Main idea:** Learning = estimating underlying function f by mapping data attributes to some target value
- **Training set:** A set of labeled examples (x, $f(x)$) where x is the input variables and the label $f(x)$ is the observed target truth
- **Goal:** Given a training set, find approximation $\hat{f}$ of f that best generalizes, or predicts, labels for new examples
  - **"Best"** is measured by some quality measure
  - **Example:** error rate, sum squared error

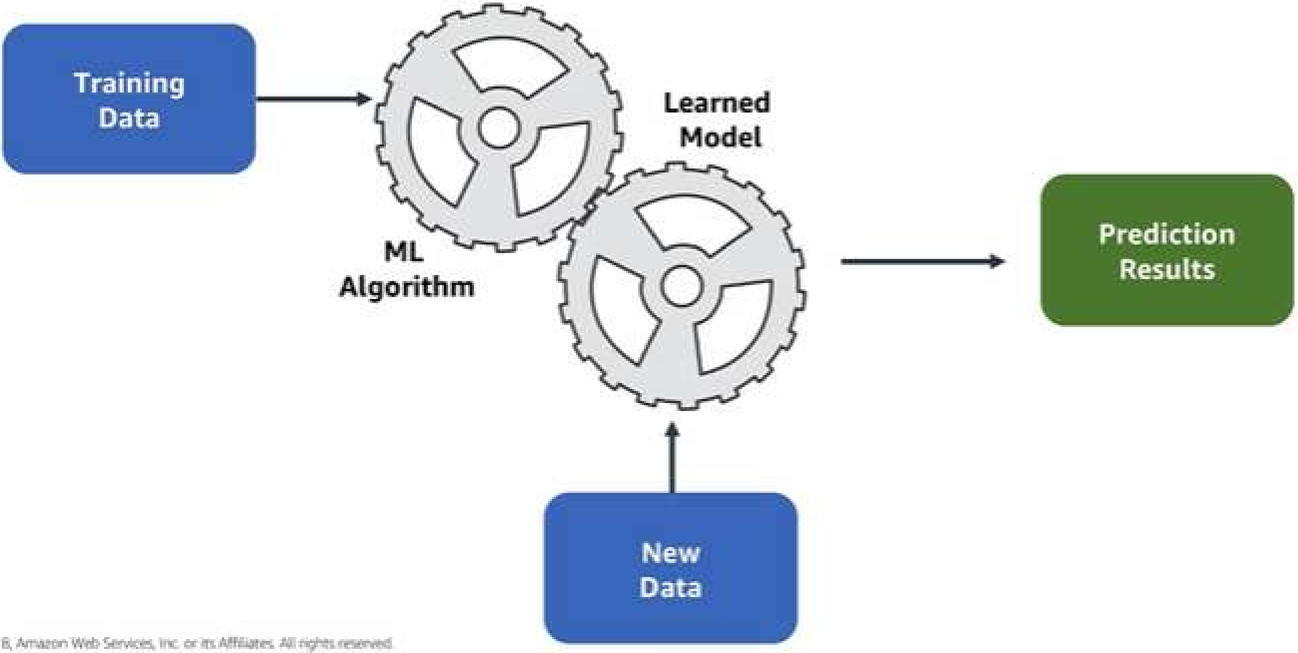


Machine Learning

- **Main idea:** Learning = estimating underlying function *f* by mapping data attributes to some target value
- **Training set:** A set of labeled examples (*x*, *f(x)*) where *x* is the input variables and the label *f(x)* is the observed target truth
- **Goal:** Given a training set, find approximation $\hat{f}$ of *f* that best generalizes, or predicts, labels for new examples
  - **"Best"** is measured by some quality measure
  - **Example:** error rate, sum squared error

Machine Learning

aws training and certification



Why Machine Learning?

Difficulty in writing some programs

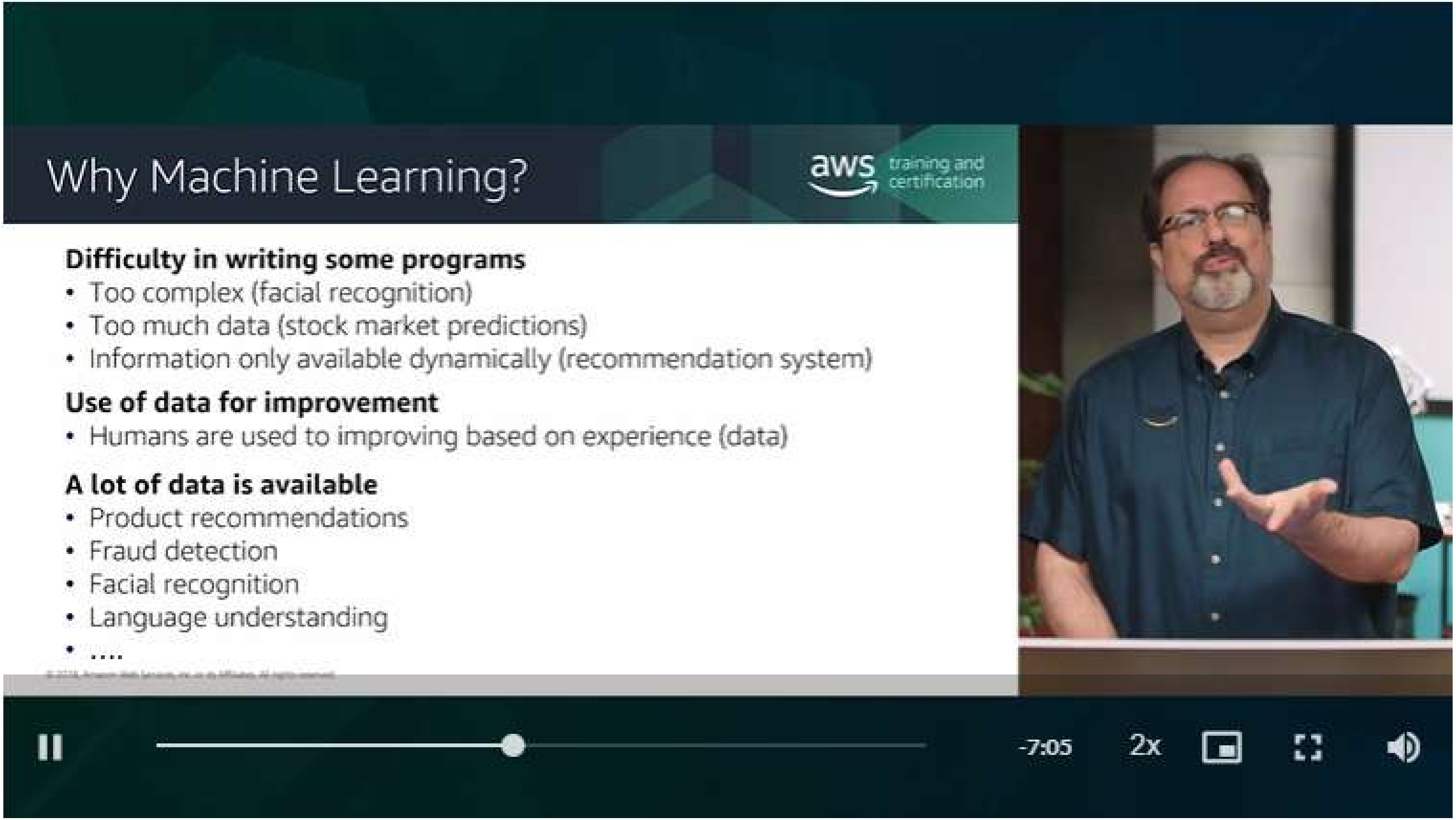
- Too complex (facial recognition)
- Too much data (stock market predictions)
- Information only available dynamically (recommendation system)

Use of data for improvement

- Humans are used to improving based on experience (data)

A lot of data is available

- Product recommendations
- Fraud detection
- Facial recognition
- Language understanding
- ….



Why Machine Learning?

Difficulty in writing some programs

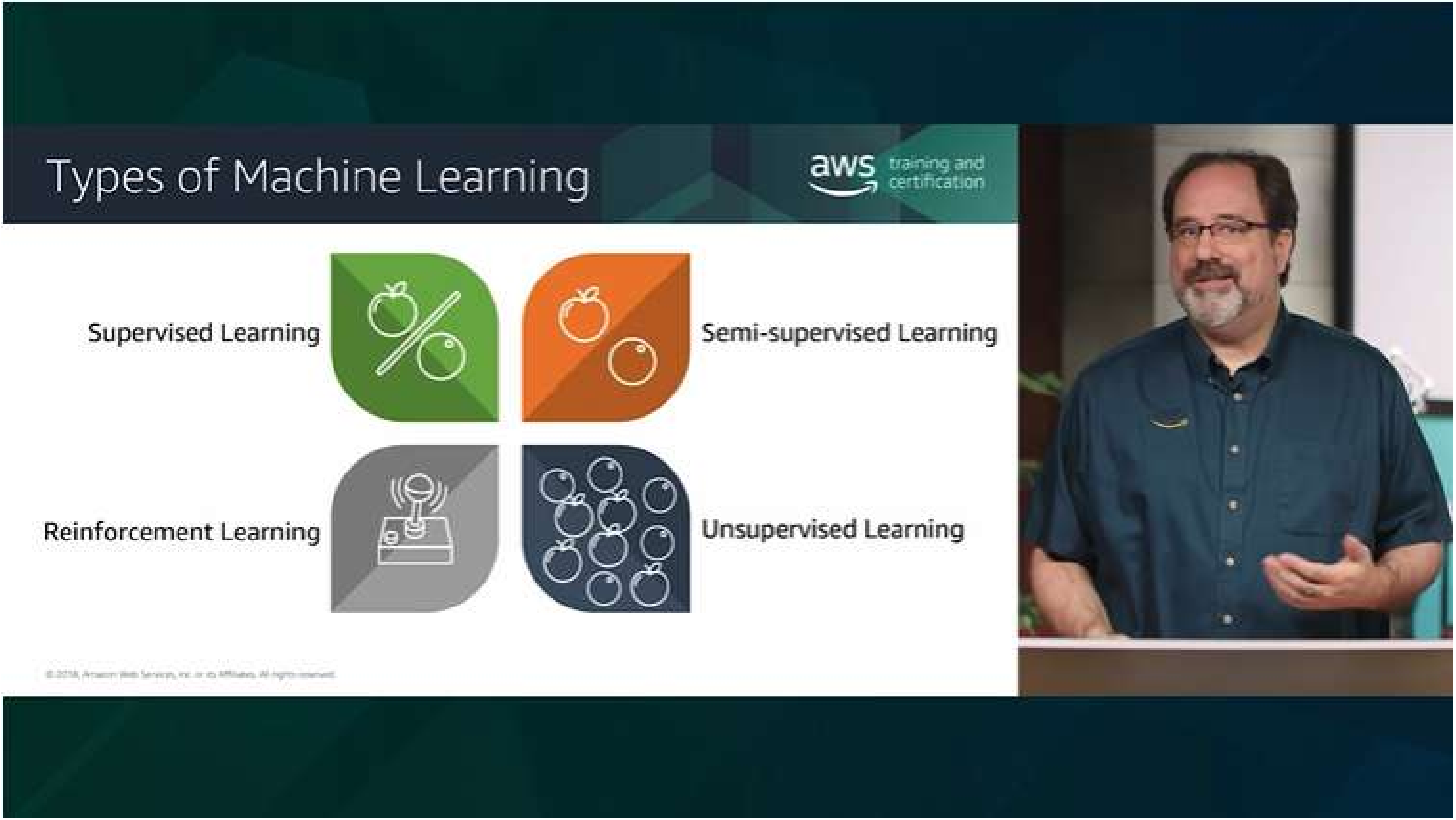
- Too complex (facial recognition)
- Too much data (stock market predictions)
- Information only available dynamically (recommendation system)

Use of data for improvement

- Humans are used to improving based on experience (data)

A lot of data is available

- Product recommendations
- Fraud detection
- Facial recognition
- Language understanding
- ....



Types of Machine Learning

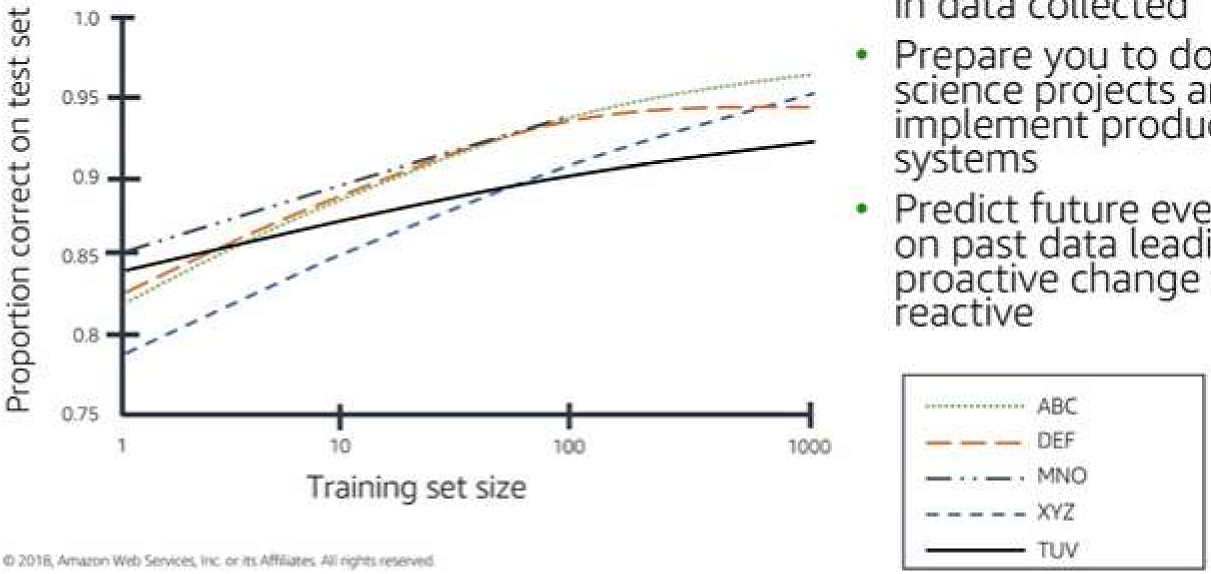
Supervised Learning

Semi-supervised Learning

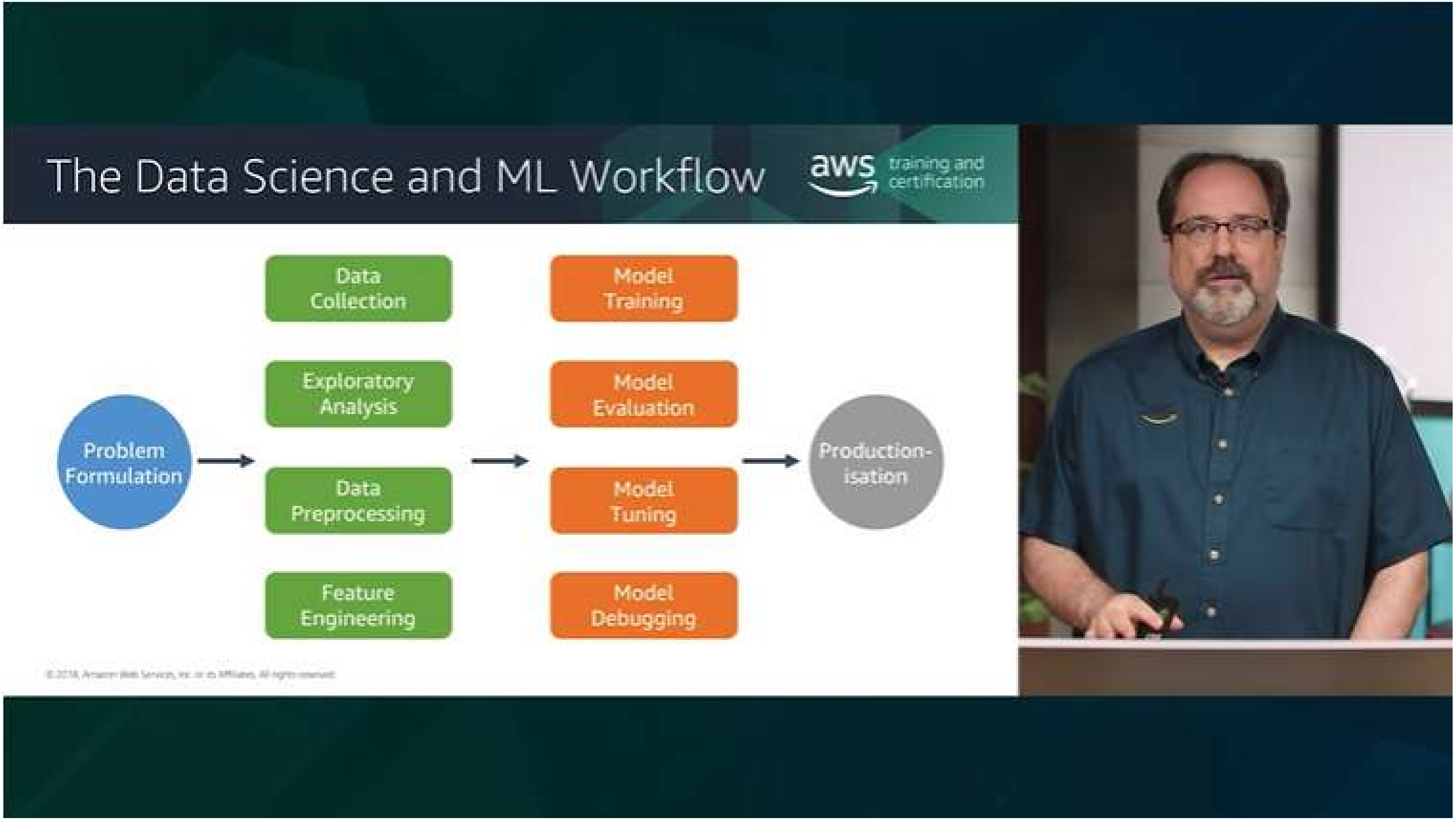
Reinforcement Learning

Unsupervised Learning

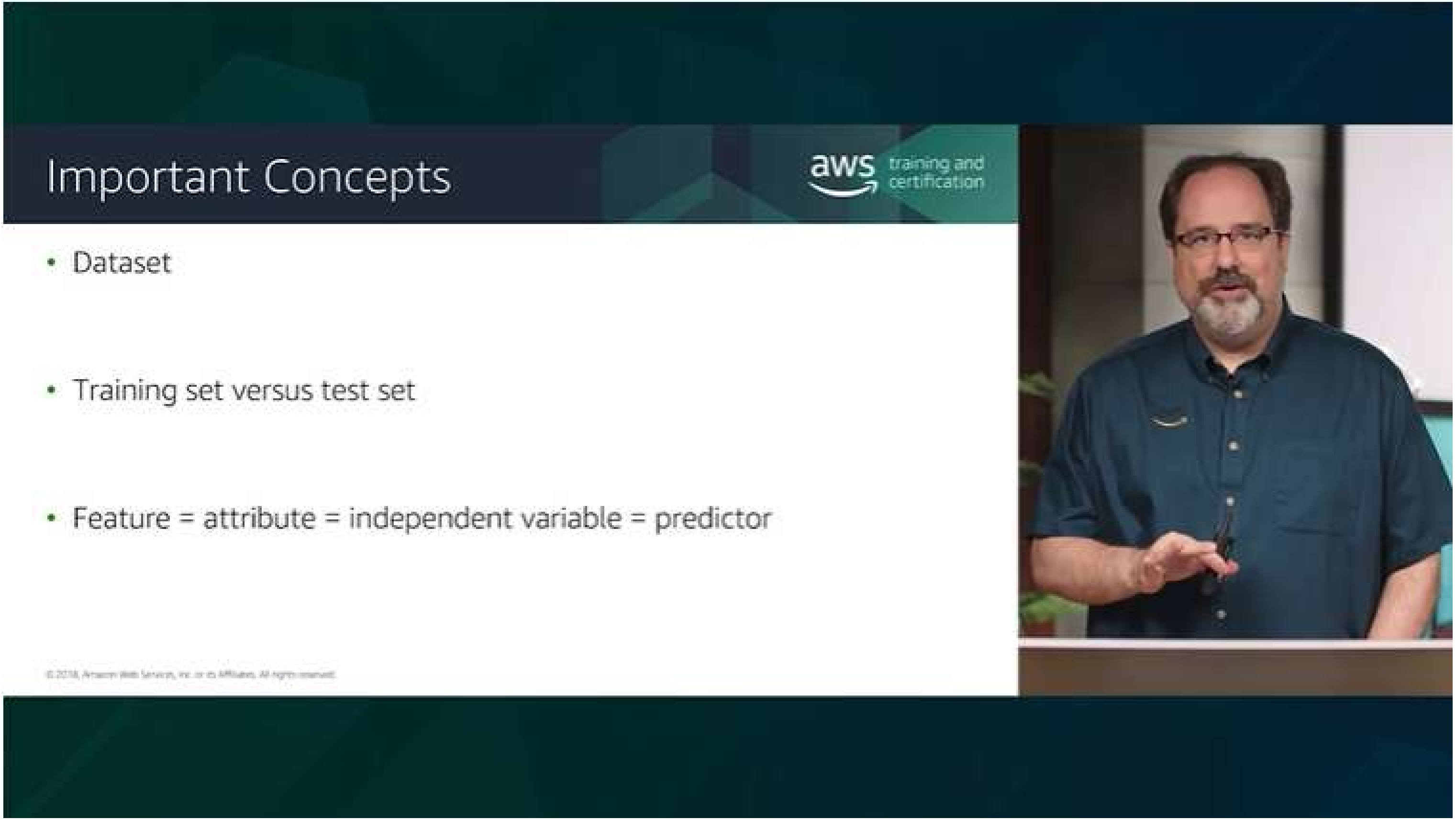
Data Matters



- Unleash the business value in data collected
- Prepare you to do data science projects and to implement production systems
- Predict future events based on past data leading to proactive change than reactive



The Data Science and ML Workflow



Important Concepts

- Dataset
- Training set versus test set
- Feature = attribute = independent variable = predictor

Important Concepts

- Label = target = outcome = class = dependent variable = response

- Dimensionality = number of features

- Model selection

Learning with feedback provided

Supervised learning

A "teacher" provides training examples, each with the correct label.

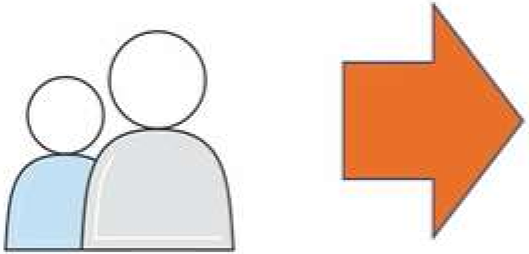
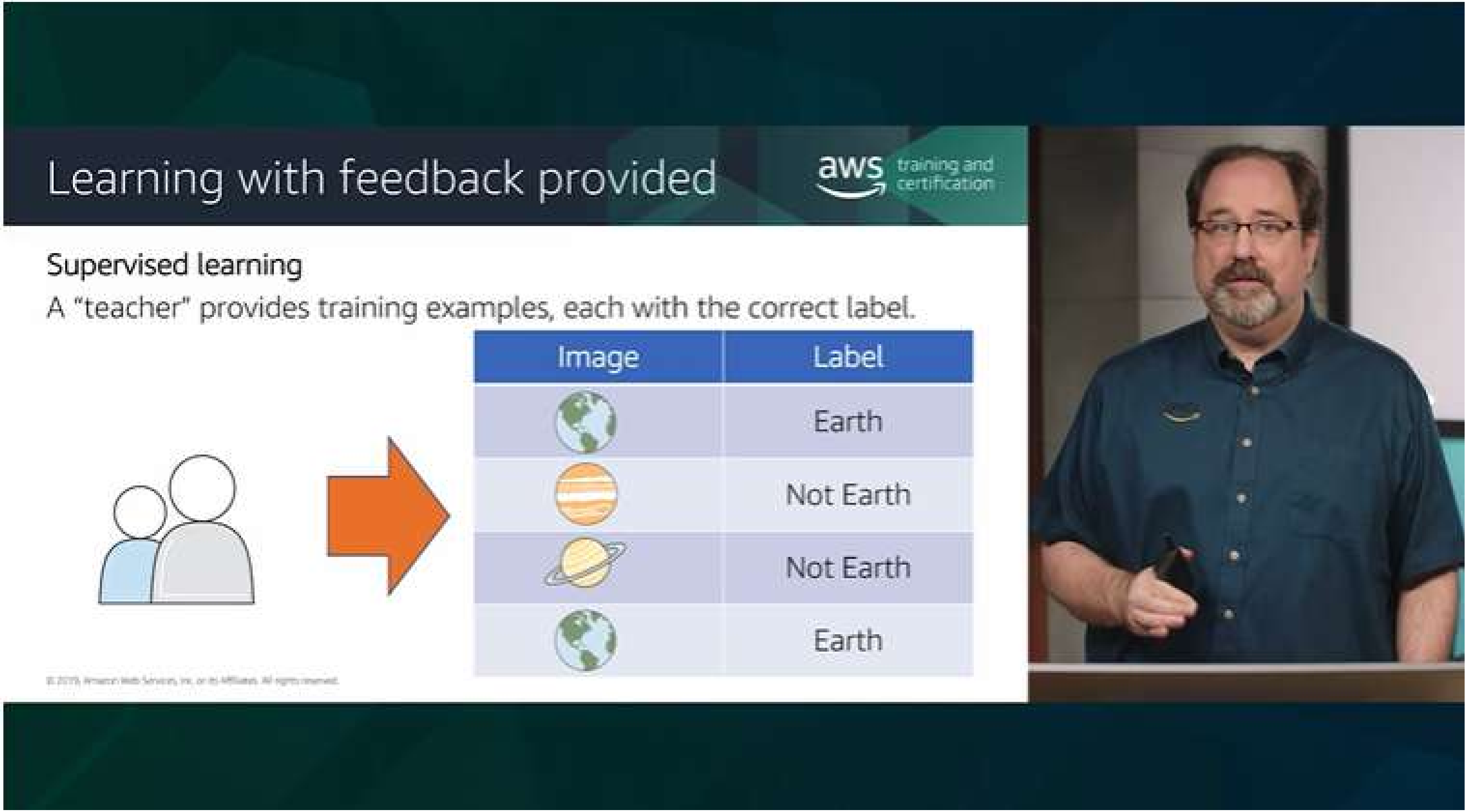


Image	Label
	Earth
	Not Earth
	Not Earth
	Earth

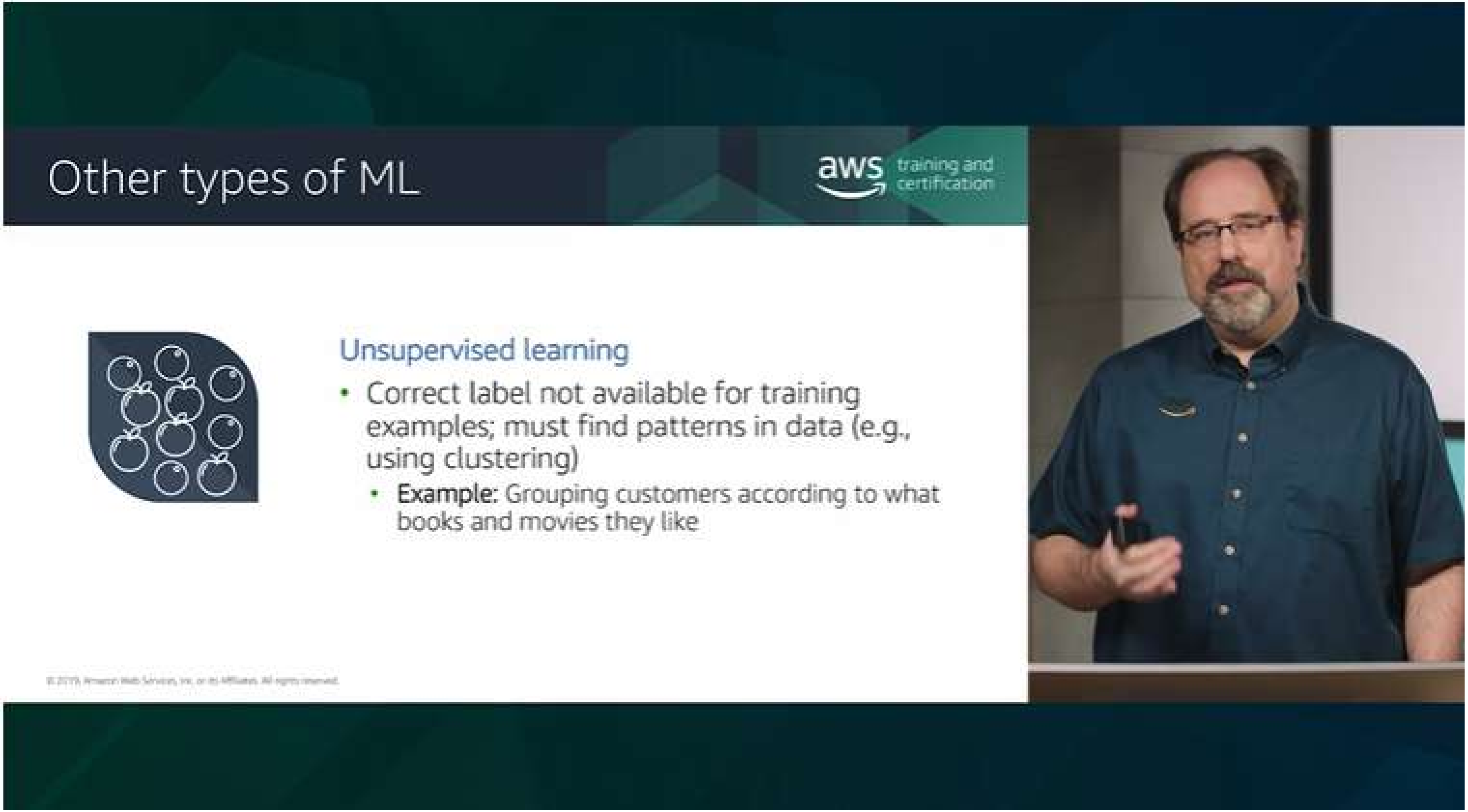


Learning with feedback provided

Supervised learning

A "teacher" provides training examples, each with the correct label.

Image	Label
	Earth
	Not Earth
	Not Earth
	Earth

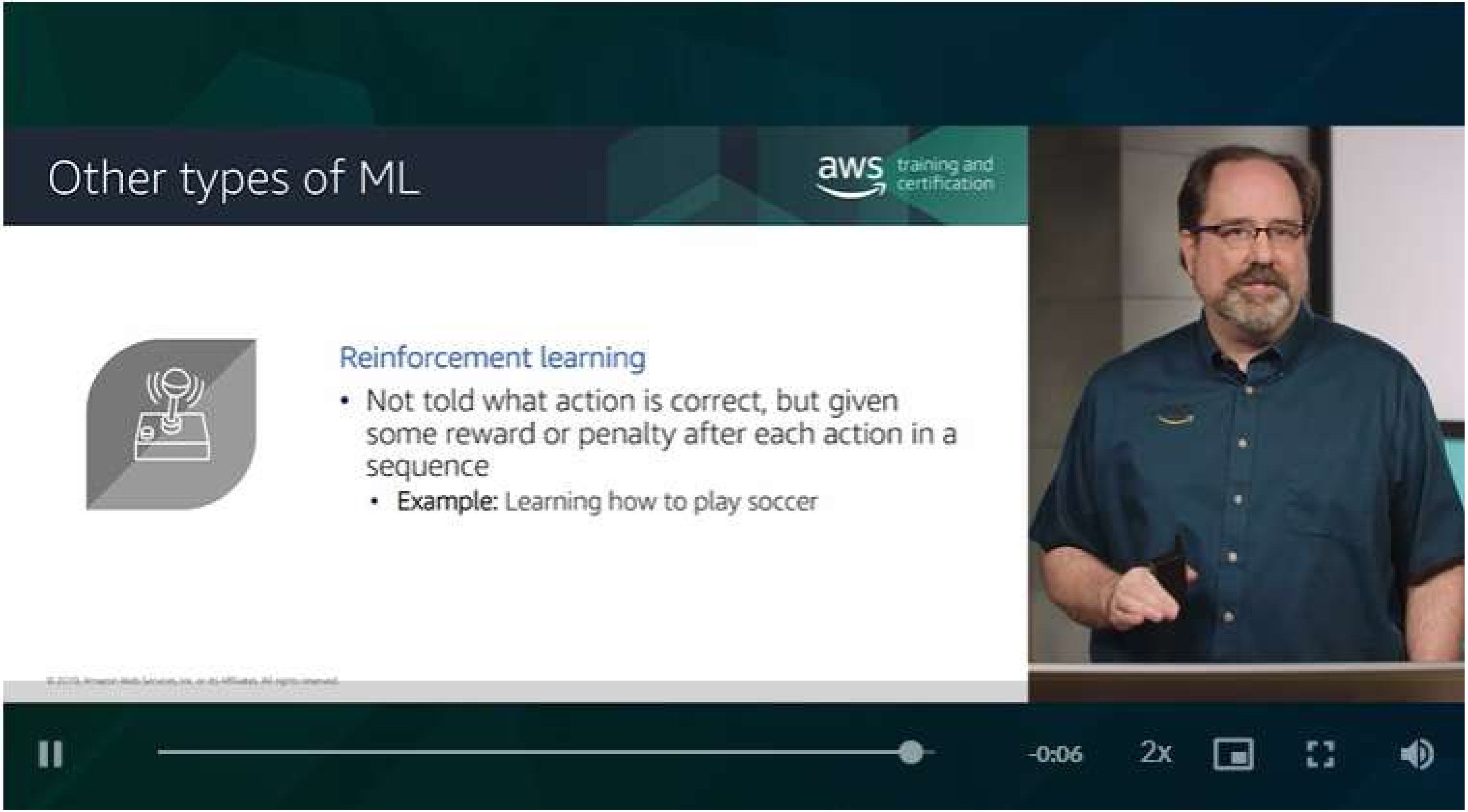


Other types of ML

aws training and certification

Unsupervised learning

- Correct label not available for training examples; must find patterns in data (e.g., using clustering)
  - **Example:** Grouping customers according to what books and movies they like



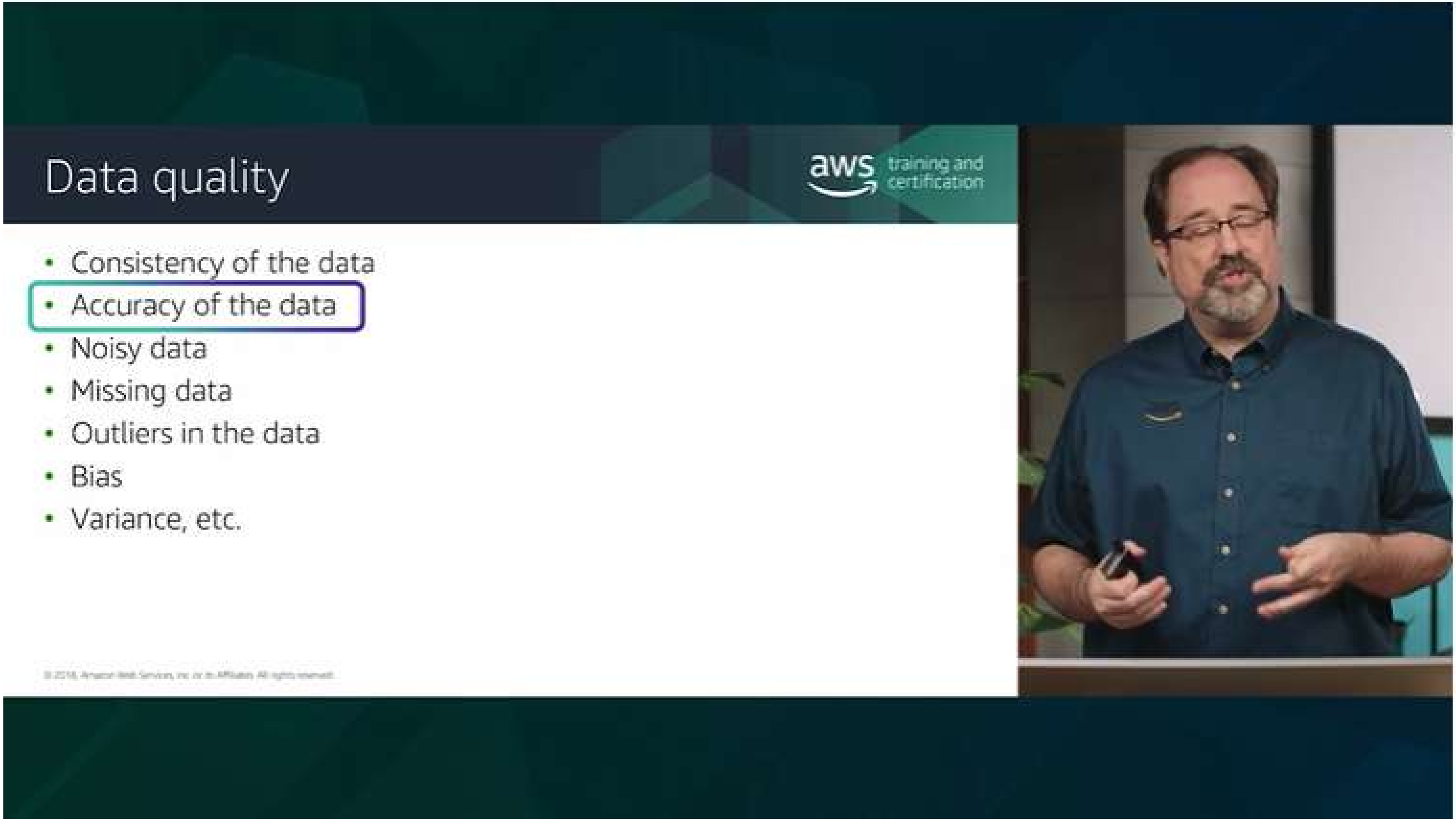
Other types of ML

Reinforcement learning

- Not told what action is correct, but given some reward or penalty after each action in a sequence
  - **Example:** Learning how to play soccer

Data quality

- Consistency of the data
- Accuracy of the data
- Noisy data
- Missing data
- Outliers in the data
- Bias
- Variance, etc.



Data quality

- Consistency of the data
- Accuracy of the data
- Noisy data
- Missing data
- Outliers in the data
- Bias
- Variance, etc.