✓ **Congratulations! You passed!**
TO PASS 80% or higher

Keep Learning

GRADE
100%

Understanding Data for ML

LATEST SUBMISSION GRADE

100%

1. Which of the following are sources of data that can be used for machine learning? (click all that apply) `1 / 1 point`

   ☑ Readings from sensors such as temperature, pressure, pH monitors, etc.

   > ✓ **Correct**
   > Correct! Well done!

   ☑ Government data such as census results.

   > ✓ **Correct**
   > Correct! Well done!

   ☐ Personal data collected without permission

   ☑ Data collected by a business about their own operations

   > ✓ **Correct**
   > Correct! Well done!

   ☑ Data collected by a business about their customers

   > ✓ **Correct**
   > Correct! Well done!

   ☑ Text data from the Internet, such as Amazon reviews or Wikipedia

   > ✓ **Correct**
   > Correct! Well done!

   ☐ Data handwritten in a notebook

   ☑ Government archives

   > ✓ **Correct**
   > Correct! Well done!

   ☑ Data purchased from third party data "brokers"

   > ✓ **Correct**
   > Correct! Well done!

2. Which of the following are issues of ethics and responsibility in machine learning? (click all that apply) `1 / 1 point`

   ☑ The fair treatment of the people collecting and processing the data

   > ✓ **Correct**
   > Correct, well done!

   ☑ The security of the data, so that it isn't easily lost or stolen

   > ✓ **Correct**
   > Correct, well done!

   ☑ The proper consent of the original owners of the data

   > ✓ **Correct**
   > Correct, well done!

☑ The representativeness of the data

> ✓ **Correct**
> Correct, well done!

☑ The anonymization of the data, as much as is possible

> ✓ **Correct**
> Correct, well done!

3. How can data be biased? (click all that apply)   **1 / 1 point**

☐ It can't; data is data and it reflects the real world

☑ It might include data collected under different conditions, and so not reflect operational data

> ✓ **Correct**
> Correct, well done!

☑ It might not include enough training data on a range of gender and ethnic groups, and so not reflect operational data

> ✓ **Correct**
> Correct, well done!

☑ It might not include data from underrepresented socioeconomic groups, and so not reflect real-world data

> ✓ **Correct**
> Correct, well done!

4. What is the batch effect?   **1 / 1 point**

○ When hospitals don't have the same scan results

○ When data from different times have included measurements of different things

○ When you train your QuAM several times in different batches

◉ When data from different sources have variations that aren't meaningful, but the algorithm takes as meaningful

> ✓ **Correct**
> Correct, well done!

5. Which of the following statements are true about data and data pipelines?   **1 / 1 point**

☐ Long term data storage is never a concern

☑ Learning data and operational data need to be in the same format

> ✓ **Correct**
> Correct! Well done!

☑ Machine learning is an ongoing process, so new, incoming data is important

> ✓ **Correct**
> Correct! Well done!

☑ Features that were used in the learning data must be present in operational data

> ✓ **Correct**
> Correct! Well done!

☐ Automating data retrieval is a straight-forward process

☑ Transformed data will need to be accessible to your QuAM

> ✓ **Correct**
> Correct! Well done!

☑ Integrating data from multiple sources can cause formatting issues

Correct
Correct! Well done!