

Filter

≡ Introduction (5 min)

≡ Data characteristics (10 min)

≡ Labels (10 min)

≡ Imbalanced datasets (10 min)

≡ Dividing the original dataset (10 min)

≡ Transforming data (5 min)

≡ Generalization (5 min)

≡ Overfitting (10 min)

≡ Model complexity (10 min)

≡ L2 regularization (10 min)

≡ Interpreting loss curves (10 min)

✔ Test your knowledge (10 min)

➡ What's next

Advanced ML models

► Neural networks (75 min)

► Embeddings (45 min)

<

Home > Products > Machine Learning > ML Concepts > Crash Course

Was this helpful?  

Datasets: Transforming data

Send feedback

- On this page
- Sample data when you have too much of it
- Filter examples containing PII

Machine learning models can only train on floating-point values. However, many dataset features are *not* naturally floating-point values. Therefore, one important part of machine learning is transforming non-floating-point features to floating-point representations.

For example, suppose `street names` is a feature. Most street names are strings, such as "Broadway" or "Vilakazi". Your model can't train on "Broadway", so you must transform "Broadway" to a floating-point number. The [Categorical Data module](#) explains how to do this.

Additionally, you should even transform most floating-point features. This transformation process, called **normalization**, converts floating-point numbers to a constrained range that improves model training. The [Numerical Data module](#) explains how to do this.

Sample data when you have too much of it

Some organizations are blessed with an abundance of data. When the dataset contains too many examples, you must select a *subset* of examples for training. When possible, select the subset that is most relevant to your model's predictions.

Filter examples containing PII

Good datasets omit examples containing Personally Identifiable Information (PII). This policy helps safeguard privacy but can influence the model.

See the Safety and Privacy module later in the course for more on these topics.

A Key terms:

- [Normalization](#)

Help Center



Previous

← Dividing the original dataset (10 min)

Next

Generalization (5 min) →

Was this helpful?

Send feedback

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](#), and code samples are licensed under the [Apache 2.0 License](#). For details, see the [Google Developers Site Policies](#). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2024-10-09 UTC.

| | | |
|-------------|--------------------------|-------------------------------|
| Connect | Programs | Developer consoles |
| Blog | Google Developer Groups | Google API Console |
| Instagram | Google Developer Experts | Google Cloud Platform Console |
| LinkedIn | Accelerators | Google Play Console |
| X (Twitter) | Women Techmakers | Firebase Console |
| YouTube | Google Cloud & NVIDIA | Actions on Google Console |
| | | Cast SDK Developer Console |
| | | Chrome Web Store Dashboard |
| | | Google Home Developer Console |