ARTIFICIAL INTELLIGENCE 501

Lesson 5
Data Collection and Enhancement

Legal Disclaimers

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting **www.intel.com/design/literature.htm**.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others

Learning Objectives

You will be able to:

- Describe data sources and types.
- Recognize situations where more data samples or features are needed.
- Explain data wrangling, augmentation, and feature engineering.
- Describe different data preprocessing methods.
- Explain ways to label data.
- Identify challenges when working with data.

Data Collection and Preprocessing

Problem Statement	What problem are you trying to solve?
Data Collection	What data do you need to solve it?
Data Exploration & Preprocessing	How should you clean your data so your model can use it?
Modeling	Build a model to solve your problem?
Validation	Did I solve the problem?
Decision Making & Deployment	Communicate to stakeholders or put into production?

Data Collection

There are several things to consider when collecting data.

- Where does the data come from?
- What type of data is there?
- How much data and what attributes do I need?

DATA SOURCES

Data Sources

Data is sourced from many different places, for example:

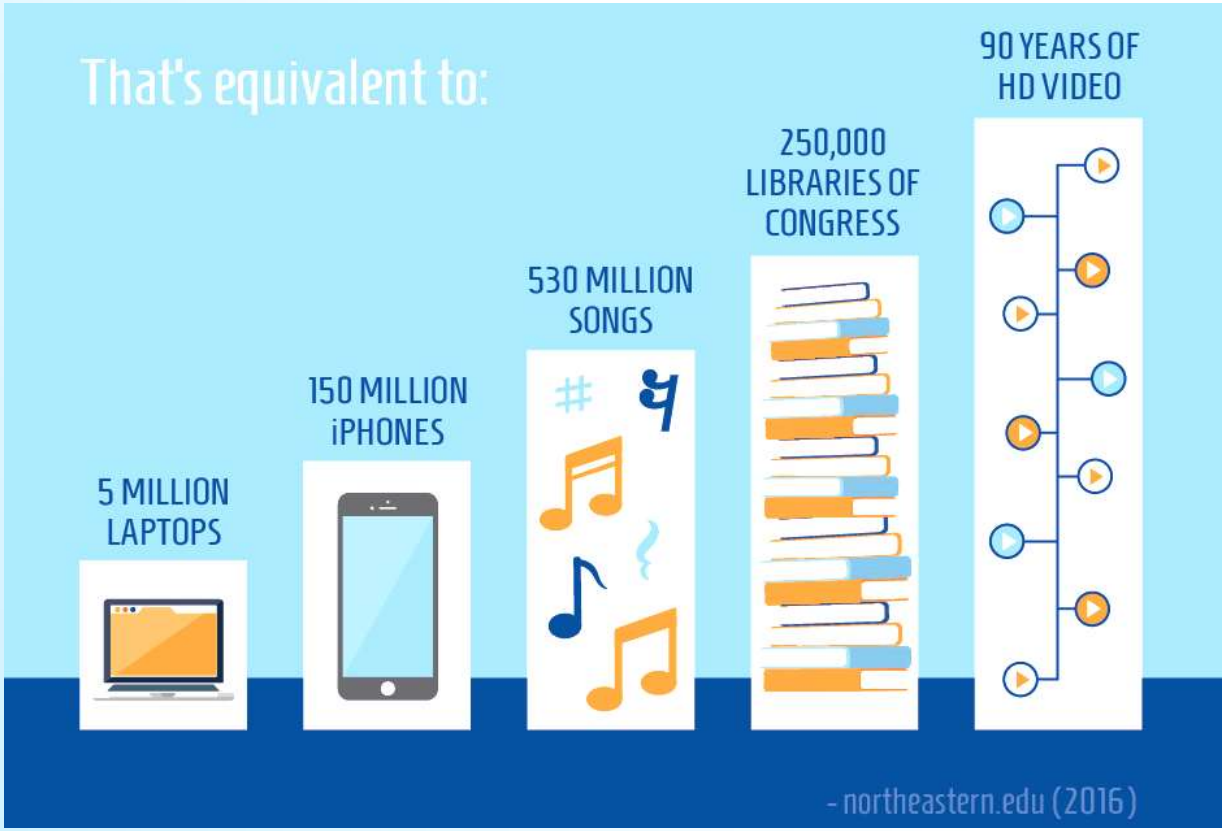
- Human generated
- Internet of Things (IoT) and machine generated
- Public website
- Legacy documents
- Multimedia

Human Generated Data: Social Media

Within the last two years 90% of the world's data was created.

- Experts believe that about 70% of this data is coming from social media.
- Our current amount of data output is 2.5 quintillion bytes per day (2.5 exabytes per day).

2.5 Exabytes Per Day



Human Generated Data: Social Media

There are various APIs to access this data.

- Facebook*: Graph API
- Twitter*: REST API or Streaming API
- Instagram*: Graph API or Platform API
- LinkedIn*: REST API
- Pinterest*: REST API

Human Generated Data: Media & Publications

Social media is not the only source of human generated data; every year humans generate a large amount of media and publications.

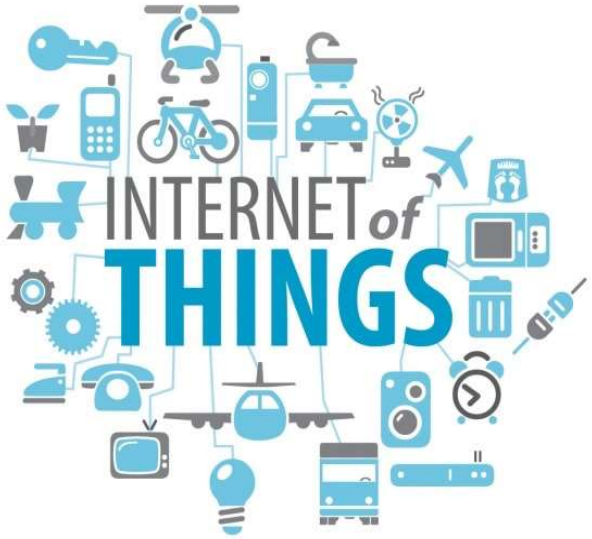
- **2.2 million books** published every *year*
- **2 million blog posts** published every *day*
- **269 billion emails** published every year

- Though not publicly available, this data contains valuable information for companies—e.g. powering services like Google Smart Reply*.

Internet of Things (IoT) Data

IoT data consists of all web-enabled devices that collect, send and respond to data they required from their respective environments.

- By 2020 IoT will include a projected 200 billion smart-and-connected devices.
- The data produced is expected to double every two years to total 40 zettabytes (40 trillion gigabytes).
- Collected via sensors, cameras, and processors.



Internet of Things (IoT) Data: Consumer

Consumer IoT provides new pathways for user experience and interfaces.

- Connected cars, smart home devices, and wearables.

Internet of Things Data: Industrial

Used to monitor and control industrial operations and tools.

- Surgery bots capable of ‘seeing’ during surgery via cameras.
- Numerous sensors and cameras installed on autonomous trucks.
- Jet engine sensors providing real-time feedback

Internet of Things Data: How to Access the data

By design, it is difficult and expensive to access data that organizations maintain and control.

- There are some known open-source datasets.
  (Ex: https://old.datahub.io/dataset/knoesis-linked-sensor-data)
- Design your own with development platforms such as Raspberry* PI* or Netduino* P1.

Public Website

Data that is publically available on the web will allow you access to multiple genres of data as needed.

- Encyclopedia
  (Ex: Wikipedia*)
- Stock data
  (Ex: Quandl*)
- Entertainment
  (Ex: IMDb*)

WIKIPEDIA
The Free Encyclopedia

Public Websites : How to Access the Data

Webscraping:

- Use with caution: the website's stance on crawlers and webscraping is usually within the terms and conditions section of their site.
- Make sure your crawler follows the rules defined in a website's robots.txt file.
  (Ex: don't exceed request rate limit)

APIs:

- Often easier than webscraping

Legacy Documents

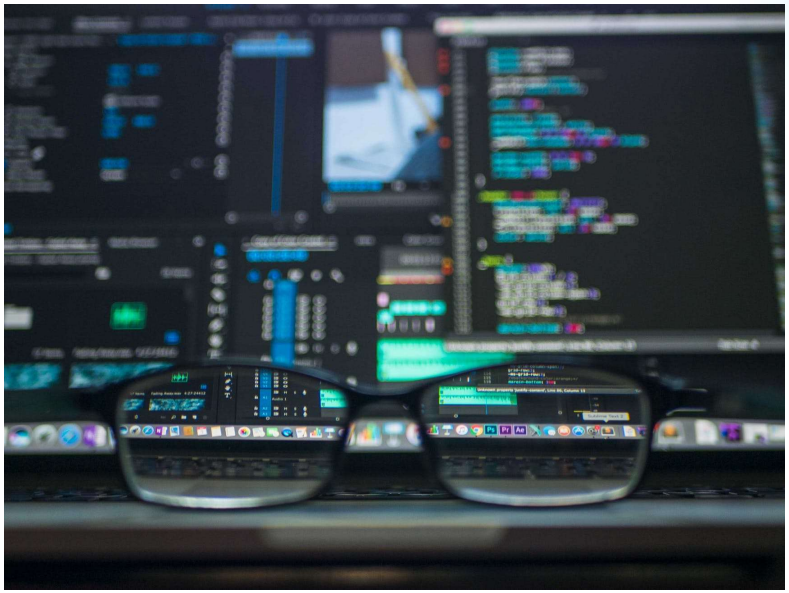
Several industries have traditionally used paper forms to collect data, creating huge potential opportunities to shift to digital.

- Two major industries shifting to more digital approaches are insurance and medicine.
- Medical records have traditionally been paper-and-pencil.
- The push for digital promises to lead to better outcomes for patients.

Multimedia

Data is present in more media types than ever before.

- Companies collect text, images, audio, and video.
- New database technologies have evolved to store this data.
  (MMDBs - multimedia databases)
- New ML techniques (e.g. Deep Learning) have evolved in part due to the necessity of analyzing this data.



DATA TYPES

Data Types

There are different data types.

- Numerical
- Discrete
- Categorical
- Ordinal
- Binary
- Date-time
- Text
- Image
- Audio

Count Data

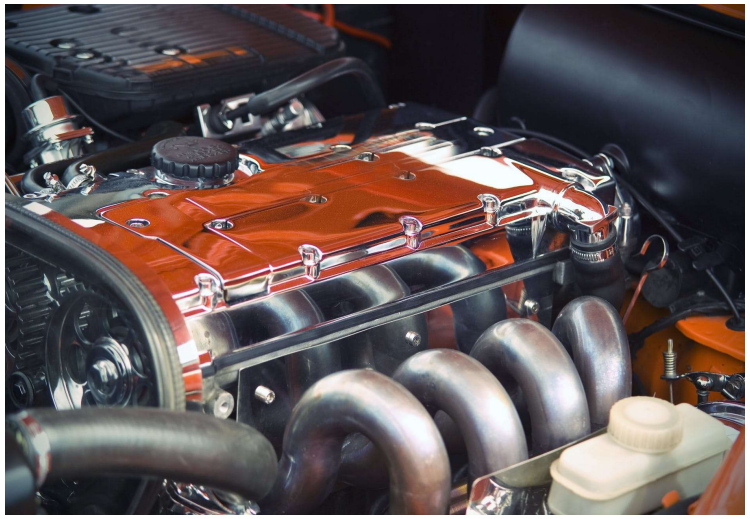
Integer valued data that comes from counting.

- For example, number of cars in a parking lot.
- The programming term for count data is ***integer***.
- In Python* it is an ***int***.

Numerical Data: Continuous

Numerical value that can represent any quantity over a continuous range.

- Can take take on decimal values.
  (Ex: engine stroke = 3.40 in)
- Can be reduced to finer levels of precision.
  (Ex: engine stroke = 3.3775 in)
- The programming term is ***floating-point***.
- In Python* it is a ***float***.



Categorical Data

Data that is restricted to a finite set of known categories.

- Also known as nominal data.
- Raw categorical data can come in the form of different data types.
  (Ex: text data (vehicle color: red)
  or numerical data (number of doors: 4)
- The programming term is ***enumerated type***.

Ordinal Data

Categorical data that is ordered.

- The distance between categories is not known.
  (Ex: car price values of *low*, *medium*, and *high*
  or customer reviews of *poor*, *ok*, and *good*)
- A common mistake is to simply convert ordinal data into integers.
  (Ex: *G*, *PG*, and *R* to *1*, *2*, and *3*)
- The inherent problem is that this assumes that the distance between the categories are known and the same.

Binary Data

Two mutually exclusive categories.

- Binary data is very common, especially in supervised learning problems. (Ex: *true/false*, *heads/tails*).
- The programming term for binary data is ***Boolean***.
- In Python* it is a ***bool***.

Date-time Data

Data that represents date and time information.

- Date, time of day, and fractional seconds based on a 24-hour clock are combined.
- It is possible to convert to primitive types.
- Can be converted to categorical data, or convert to integer via Unix* time stamp.
  (Ex: to only *Date*, *Month*, *Year*, *Time*)
- Most languages have a built-in datatype for date-time data.
- In Python* it is also ***datetime***.

Text Data

Alphanumeric strings, a sequence of characters.

- Typically human-readable.
- Can be encoded into computer-readable formats such as ASCII or unicode.
- Often is unstructured data.
- The programming term for text data is ***string***.
- In Python* it is a ***str*.**

Image Data

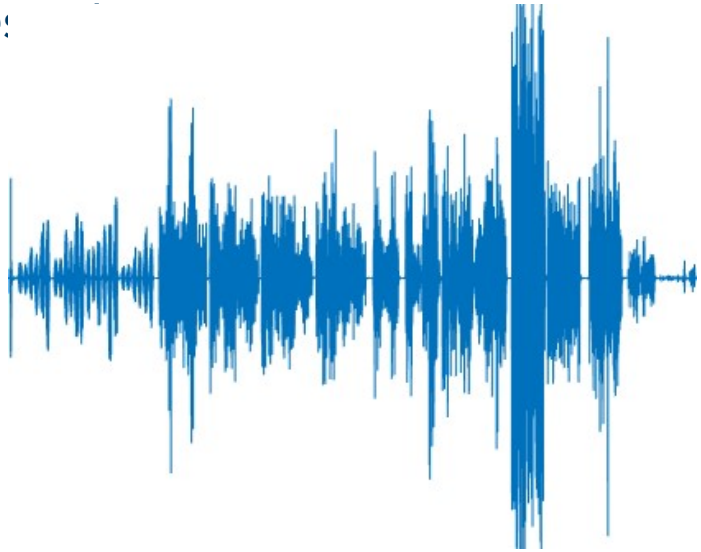
Still and video images are being generated at a rapid rate.

- They are stored in various ways.
  (Ex: PNG and JPEG)
- Images can be very large and computationally intensive.
- They can be multi-dimensional, as in 3D body scans.
- Generated from many types of devices in a diverse set of fields.
  (Ex: satellites, self driving cars, surveillance cameras, mobile devices)
- Medical imaging is an increasingly important area.

Audio

Audio data is used in consumer devices like Alexa*, as well as industrial settings like call centers.

- Stored in various compressed and uncompre
  (Ex: WAV and MP4)



THE SHAPE OF DATA

The Shape of Data

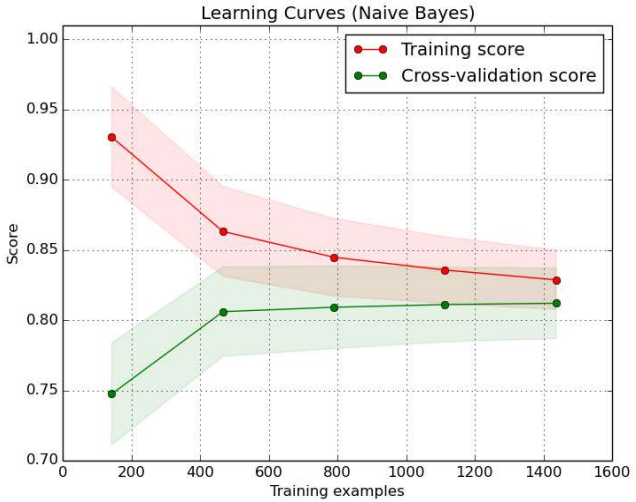
Identify the number of samples and features needed to solve the problem.

- The more features you have, the more samples are generally required.
- If your features don't contain enough information, then adding more samples wont help - additional features are required.
- If you have too few samples, then adding additional features can cause overfitting.
- Difficult problems usually require more features and additional samples.

How Many Samples?

There are several ways to determine the number of sample needed.

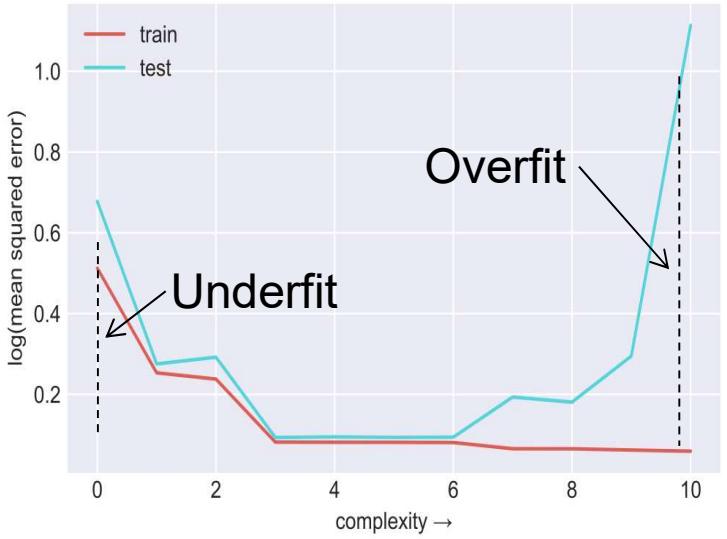
- A **learning curve** plots the performance score by number of samples.
- Use this curve to determine if increasing the number of samples is helping your model.
- **Statistical heuristic**: 10x as many samples as degrees of freedom.
- Peruse similar studies to learn from other successes and failures.



Under-fitting vs Overfitting

A **learning curve** can also help diagnose whether a model is under-fit or overfit.

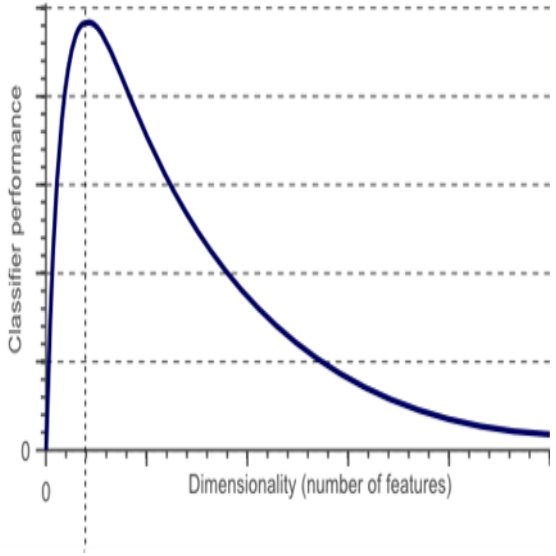
- Plot the performance score by the model complexity.
- More features means more complexity.
- When train and test performance are similar - but low - the model is under-fit.
- When test performance suffers, while train performance improves, the model is overfit.



Under-fitting vs Overfitting

There is a relationship between under-fit and over-fit models and the number of samples and features.

- In general, the number of samples should be greater than your number of features.
- If overfitting:
  - adding more features usually hurts
    adding more samples usually helps
- If under-fitting:
  - adding samples generally doesn't help
  - strong explanatory generally variables helps



How to Increase the Number of Features?

Increasing the number of features can help when the model is under-fitting.

- Revisit the data pipeline and add features that had previously been removed.
- Design new features - this is referred to as **feature engineering**.
- **Polynomial features** are a common method - this builds new features by multiplying existing ones. (Ex: x1*x2 or x2*x2)
- Images/Audio: add transformations. (Ex: **delta features -** a measure of how color changes from pixel-to-pixel - can be added to an existing image)
- Image augmentation for computer vision[1].
- Create new images by shifting, flipping, rotating, existing images.

[1]Refer to ML 501 for more on this.

Reducing Features: Feature Selection

Select a subset of features in order to reduce model complexity and remove redundant or weak features.

- **Filter methods**: filter features via a statistical method (**p-value**).
- Feature is considered on a univariate basis.
- **Regularization methods**: learn which features best contribute to accuracy while the model is being created. [1]
- **Wrapper methods**: selects subsets of features as a search problem, where different combinations are evaluated to capture interactions between features.

[1]Refer to ML 501 for more on this.

Reducing Features: Dimensionality Reduction

Dimensionality reduction is an unsupervised learning technique used to reduce the number of features. [1]

- Reduce dimensions (compress data) without losing too much information. (Ex: using the eigenvectors of the covariance matrix as the reduced dimensional space)

[1] Refer to ML 501 for more on this.

DATASETS

Datasets for AI

There is no standard type of datasets used for AI.

- They can vary in shape and size.
- They can be made up of many different data types.
- They differ depending on the tasks.

Open Source Repositories

There are several repositories to get open source datasets.

- **UCI Machine Learning Repository:** Includes many popular datasets used to analyze algorithms within the ML community.
- **ImageNet:** Large database of images.
- **KDD Cup:** An annual competition organized by the Association for Computing Machinery. Yearly competition datasets are archived.
- **Kaggle*:** A platform for data science competitions. Competition datasets and other datasets are available.
- **Data.gov:** Open data from the US government.
- And many more…

ImageNet

ImageNet is a large image database that is popular in the AI community.

- First presented at the 2009 Conference on Computer Vision and Pattern Recognition.
- Currently has over 14 million images.
- Hosts annual object detection and classification competition.
  - Deep learning models have had breakthrough results that have led to the current focus on deep learning for modern AI.
  - Many of the most popular deep learning models have come from this competition.

Example Image Datasets

Below are some popular image dataset examples.

- **MNIST:** A popular benchmark for many machine learning models.
  - Images of handwritten digits that has served as a
  - 70,000 28 x 28 pixel black and white images.
- **Cifar-10:** A widely used dataset for computer vision research.
  - 60,000 32 x 32 pixel color images.
  - 10 different classes for classification. 6,000 images for each class.
- **ILSVRC:** The annual ImageNet competition has several components each with their own datasets.
  - Object localization: 1.2 million training images of 1,000 object classes
  - Object detection: 456,567 training images of 200 object classes

Example Natural Language Datasets

Below are some popular language datasets.

- **Common Crawl:** Crawls the web four time a year and makes its archives and datasets free for the public.
  - The archive consists of 145 TB of data from 1.81 billion webpages as of 2015.
- **Stanford Question Answering Dataset:** A dataset of over 100,000 question-answer pairs.
- **Project Gutenberg:** Offers over 56,000 free eBooks.

Example Datasets

There are many other datasets available for different tasks.

- **YouTube*-8M Dataset:** Roughly 7 million video URLs containing 450,000 hours of video.
- **MovieLens* Dataset:** A dataset of roughly 20 million ratings on 27,000 movies by 138,000 users.
- **OpenStreetMap**: A crowdsourcing project to create a free map of the world.
  - Over 2 million users collect data using survey, GPS, aerial photographs, and more.
- **1000 Genome Project:** Human genotype and variation data collected 2008-2015
  - 1,000 genomes with 84.4 million variants from 2,504 individuals.

DATA PREPARATION

Data Preprocessing

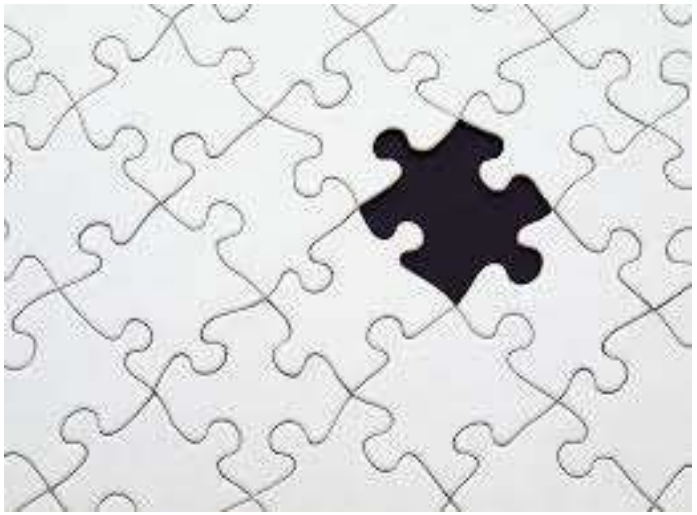
Data preprocessing is needed because the quality and structure of data is often not immediately ready for analysis.

- Data obtained via webscraping or APIs is usually unstructured and must be manipulated into numerical rows and columns for analysis.
- This often involves low-level manipulation of data objects, turning strings into numeric data and vice versa.
- There could be missing data values.
- Different ML models require different data formatting.

Missing Values

Often certain parts of the data are missing.

- Replace missing values with a reasonable default
- Features with symmetric distribution: input the missing values with the mean of the feature.
- If not symmetric or categorical, imputing the median or mode respectively is best.
- If too many features are missing, it is best to drop the observation.



Slide 48

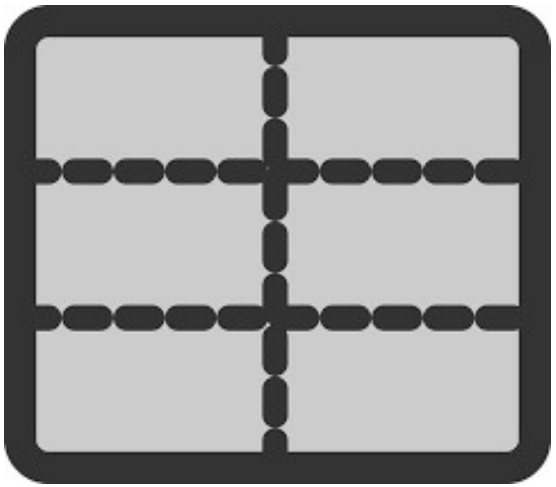
SW1 Build this out

Seth Weidman, 1/9/2018

Data Preprocessing

Some examples of ML models that require data to be formatted differently include:

- Models requiring features to have a similar scale.
  - Option 1: Scale each feature so that the maximum is one and the minimum is zero.
  - Option 2: Standardize each feature to have a mean of zero and a standard deviation of one.
- Others transform the dependent or target so it is not extremely skewed. (Ex: by using logarithms)



LABELING DATA

Labeling Data

We need labeled data for supervised learning tasks, but there is usually more unlabeled than labeled data.

- Data can be hand-labeled by employees, but this can be expensive and time consuming.
- Other options include Amazon Mechanical Turk* and semi-supervised learning techniques.

Amazon Mechanical Turk*

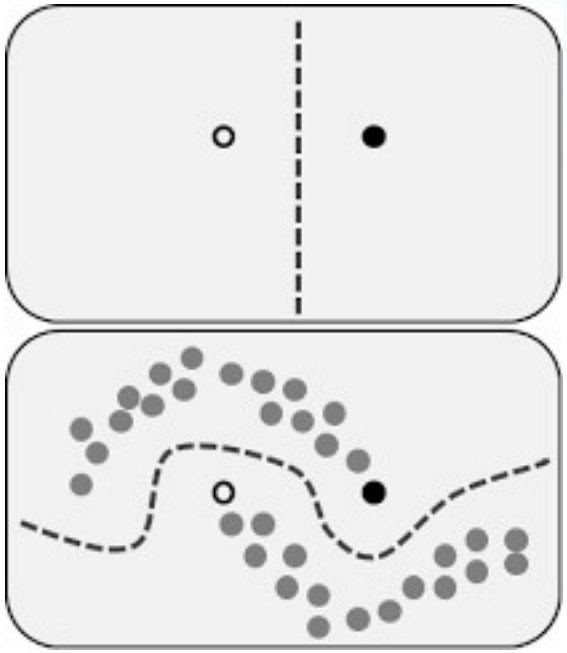
Amazon Mechanical Turk* is an online marketplace for tasks.

- Organizations can post tasks on the website that “Turkers” can then choose to perform.
- Major open source datasets have been complied using Amazon Mechanical Turk.
  - Microsoft Common Objects in Context (COCO) dataset contains images with 5 short descriptions each - from “Turkers”

Semi-supervised

Semi-supervised learning involves using both labeled and unlabeled data to solve tasks

- Build a classifier on the labeled data, use it to label some of the unlabelled data, and then proceed iteratively until all data has been labeled.



Learned Divisions

CHALLENGES

Challenges

There are many challenges with the collection, structure, and usage of data that can lead to poor performance.

- All these can lead to overly optimistic and misleading results or model failure.
- Some examples include: biases and outliers in the data, inappropriate validation and testing, class imbalance in classification problems.

Biases in Data

Algorithms will reflect the data they are trained on.

- When trained on biased data, they will reflect those biases. (Ex: the meaning of the word “man” rather than “woman” may be associated more with “power”, after being trained on a large corpus of human-generated data)
- It’s important for the modeler to know how the data was collected to address any biases that might be present.
  - A lot of the time this cannot be corrected during the analysis stage. There are corrections for some special cases.

Data Leakage

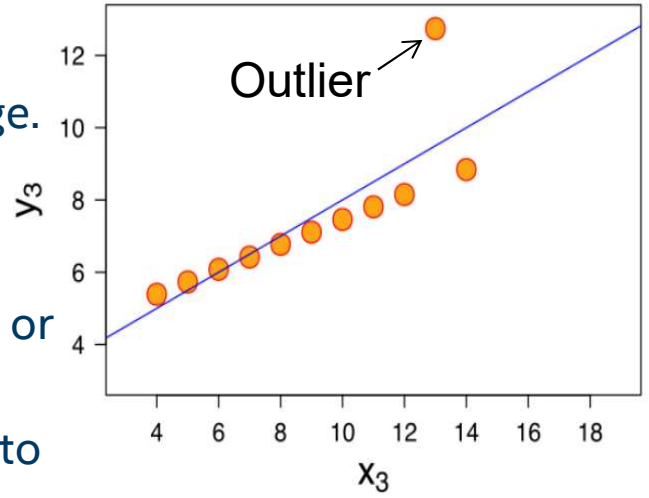
Data leakage is any time the train/test split is violated.

- The model doesn't see test data until evaluation.
- Data leakage can lead to overestimates of how well the model will generalize to unseen examples.
- There is no single solution to data leakage. The modeler should be careful when developing validation strategies and splitting the dataset.
  - Data leakage is common when working with time series data and during preprocessing.
  - For example, models that make predictions about the future based on the past should have train and test sets organized this way or using test data to help preprocess the training data.

Outliers

Outliers are data points far away from other data points.

- They can skew model results by suggesting performance is much lower than it actually is.
- The modeler should detect outliers and try to understand why they're present during the EDA stage.
  - There are many outlier detection methods.
- There are many ways to combat outliers.
  - Often outliers must be removed from the dataset, or adjusted to fit the pattern of remaining dataset.
  - Use models and/or metrics that are less sensitive to outliers.



Imbalanced Datasets

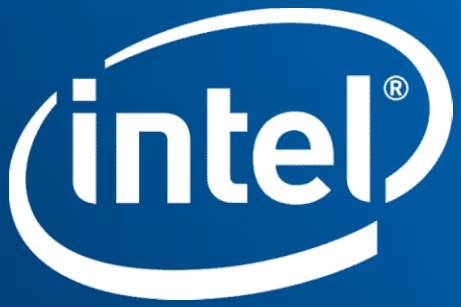
Imbalanced classes can lead to difficulties with training and evaluating models.

- For example, if 99% of labels belong to one class, then a model that always predicts the majority class will be 99% accurate by default.
- Some modeling techniques to address this include **down-sampling** the larger class, or using scoring metrics other than accuracy to assess models.

Learning Objectives Recap

In this lesson, we worked to:

- Describe data sources and types.
- Recognize situations where more data samples or features are needed.
- Explain data wrangling, augmentation, and feature engineering.
- Describe different data preprocessing methods.
- Explain ways to label data.
- Identify challenges when working with data.


intel®

Sources for images used in this presentation

https://www.pexels.com/photo/auto-auto-racing-automobile-automotive-355913/
https://www.pexels.com/photo/automobile-automotive-beautiful-car-210013/
https://www.pexels.com/photo/monochrome-photography-of-round-silver-coin-839351/