Machine Learning

| Main Challenges

Prof. Dr. Mostafa Elhosseini

Professor of Smart Systems Engineering

https://youtube.com/drmelhosseini

Main Challenges

- Since your main task is to select a model and train it on some data, the two things that can go wrong are:
  - bad model and
  - bad data

Bad Data

Insufficient Data:

- AI models, especially deep learning ones, require large amounts of data to learn effectively.
- Insufficient data can lead to a model that doesn't perform well.
- Even for very simple problems you typically need thousands of examples, and for complex problems such as image or speech recognition you may need millions of examples (unless you can reuse parts of an existing model).

Bad Data

Insufficient Data:

- Data matters more than algorithms
- it is not always easy or cheap to get extra training data — so don't abandon algorithms just yet

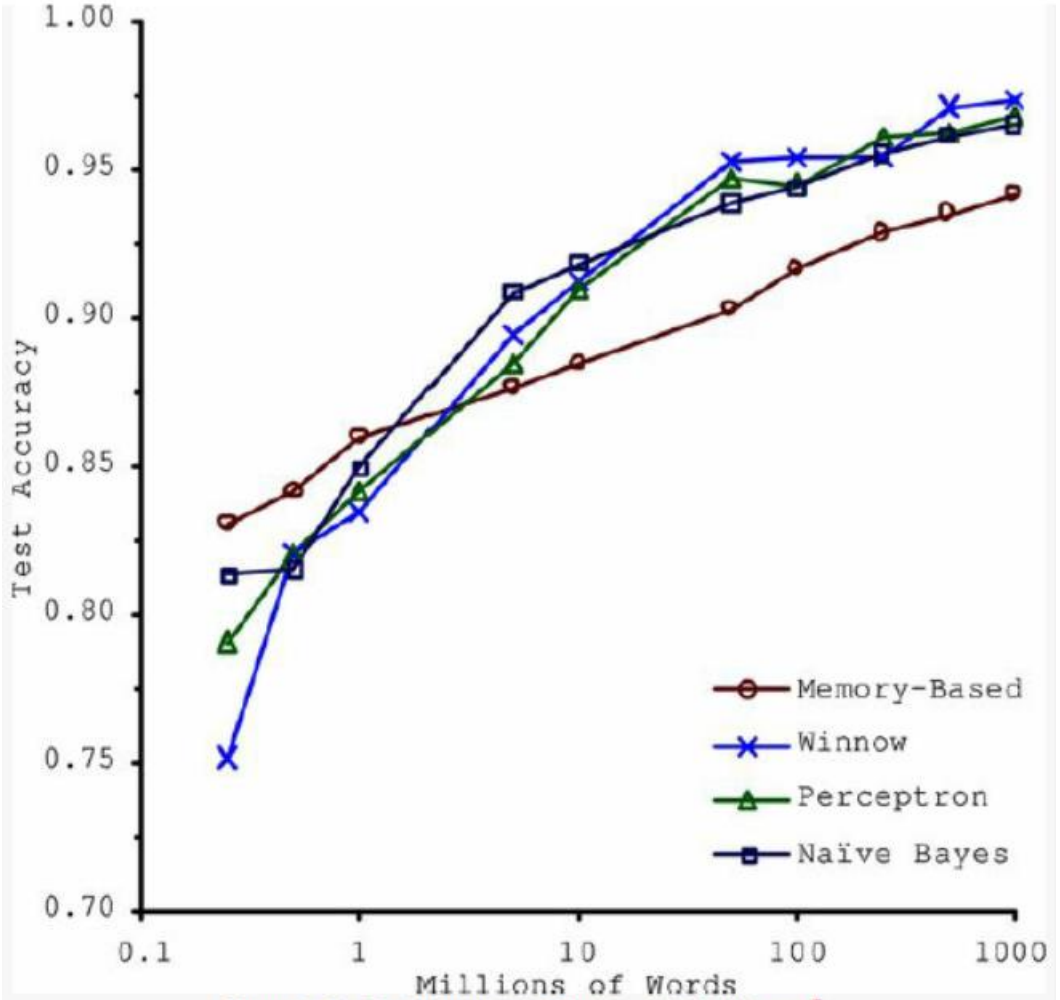


Figure 1-21. The importance of data versus algorithms [8]

Bad Data

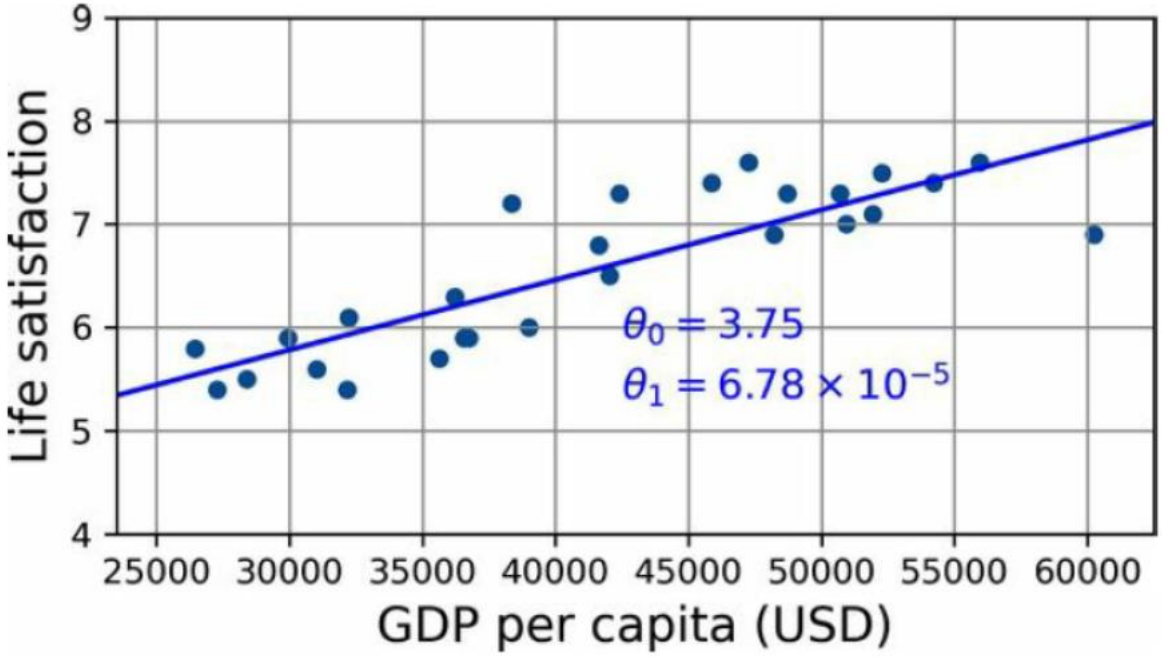
Nonrepresentative Training Data:

- To ensure that a model generalizes well to new, unseen data, The training data must accurately reflect the real-world scenarios and conditions the model will encounter post-deployment.

Bad Data

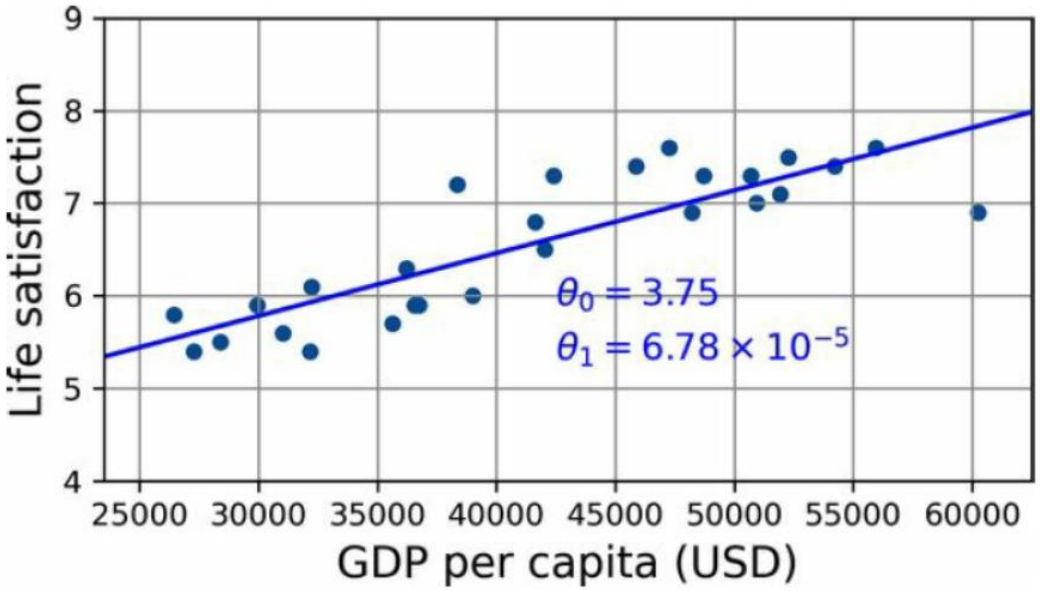
Nonrepresentative Training Data:

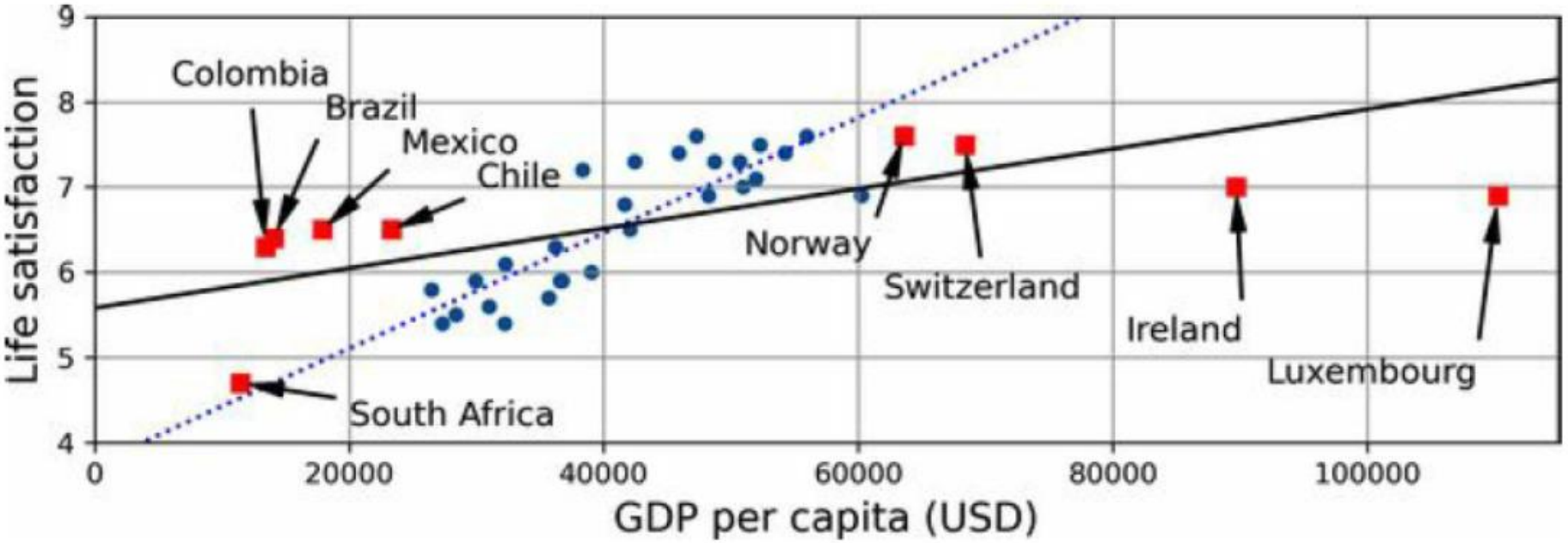
- Recall GDP Example
- Did not contain any country with a GDP per capita lower than $23,500 or higher than $62,500



Bad Data

Nonrepresentative Training Data:

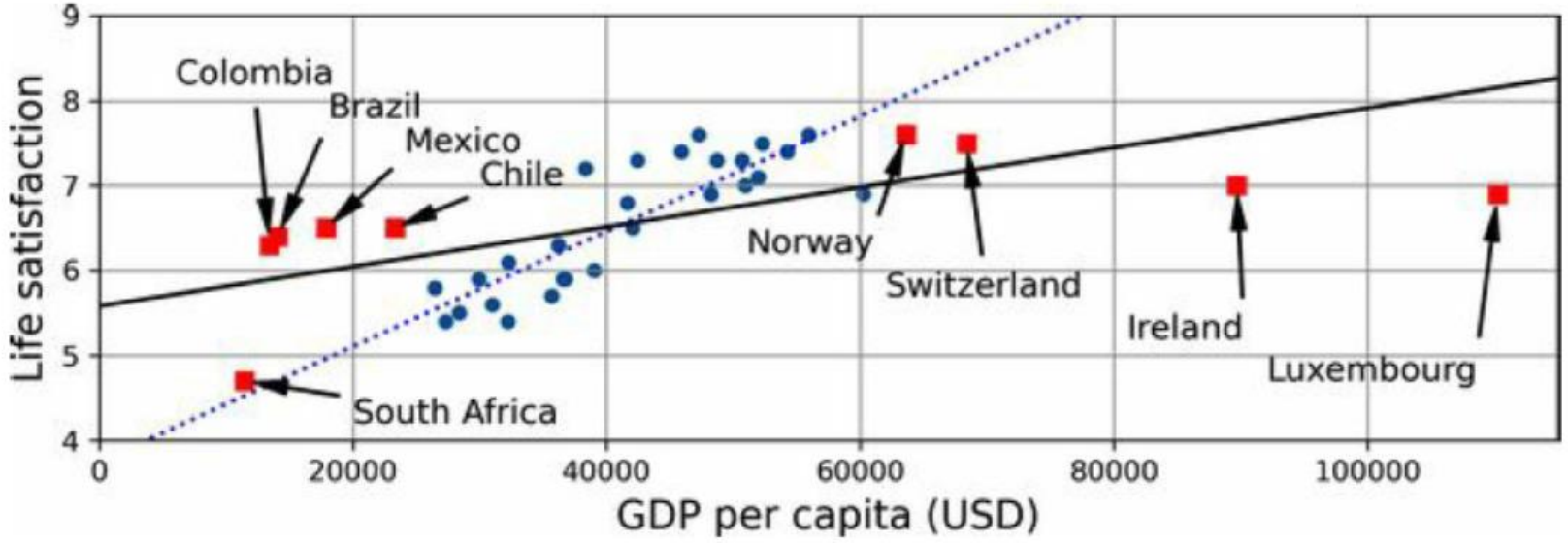




Bad Data

Nonrepresentative Training Data:

- If you train a linear model on this data, you get the solid line, while the old model is represented by the dotted line. As you can see, not only does adding a few missing countries significantly alter the model, but it makes it clear that such a simple linear model is probably never going to work well.



Bad Data

Nonrepresentative Training Data:

- if the sample is too small, you will have **sampling noise** (i.e., nonrepresentative data as a result of chance), but even very large samples can be nonrepresentative if the sampling method is flawed. This is called **sampling bias**.

Bad Data

Nonrepresentative Training Data:

- **Sampling Noise**: This occurs when the training set is too small. A small sample size may not capture the full diversity and range of scenarios in the target population, leading to a model that doesn't generalize well. This issue is more pronounced in datasets with high variability.
- Sampling noise refers to the random variability in a dataset that arises purely by chance because a sample is just a subset of the whole population. This concept is particularly relevant in statistics and data analysis

Bad Data

Nonrepresentative Training Data:

- **Sampling Noise**: It's the random variation that occurs from one sample to another, due to the inherent randomness in which elements are selected for the sample.
- Sampling noise can affect the precision of statistical estimates. For example, in a survey, different samples might give slightly different results purely due to chance variations. This noise is especially pronounced in smaller samples.

Bad Data

Nonrepresentative Training Data:

- **Sampling Noise**: The level of sampling noise is inversely related to the sample size. Larger samples tend to have less sampling noise because they more closely approximate the entire population. In contrast, smaller samples are more likely to show greater variability and therefore greater noise.
- Increasing the sample size is the most direct method to reduce sampling noise. However, this might not always be feasible due to cost, time, or logistical constraints. Another approach is using statistical techniques to estimate and account for the expected level of noise.

Bad Data

Nonrepresentative Training Data:

- **Sampling Noise**: sampling noise can affect the training process of a model. A model trained on a dataset with a high level of sampling noise might not generalize well, as it may learn patterns that are not truly representative of the underlying population.

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**: Sampling bias is a systematic error that occurs in a data collection process, leading to a sample that is not representative of the population from which it was drawn. This bias skews the results of the analysis, as the sample does not accurately reflect the characteristics or patterns present in the larger population.
- Sampling bias can occur due to the way in which the samples are collected. For example, if certain groups or types of data are more likely to be included or excluded in the sample, either intentionally or unintentionally, the resulting dataset will be biased.

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**: If a survey is conducted online, it may exclude those without internet access, thus biasing the results towards tech-savvy or younger demographics.
- In medical research, if a clinical trial includes only a certain age group, the findings might not be applicable to other age groups.
- To avoid sampling bias, it's important to use random and stratified sampling methods that ensure all segments of the population have an equal chance of being represented.

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**: If a survey is conducted online, it may exclude those without internet access, thus biasing the results towards tech-savvy or younger demographics.
- In medical research, if a clinical trial includes only a certain age group, the findings might not be applicable to other age groups.
- To avoid sampling bias, it's important to use random and stratified sampling methods that ensure all segments of the population have an equal chance of being represented.

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**: Example.
- The Literary Digest الملخص الأدبي magazine conducted a poll to predict the result of the 1936 presidential election between Franklin Roosevelt (Democrat and incumbent) and Alf Landon (Republican).
- At the time, the magazine's polls were famous because they had correctly predicted three successive elections.
- In 1936, Literary Digest mailed questionnaires to 10 million people and asked how they planned to vote. The sampling frame was constructed from telephone directories, country club memberships, and automobile registrations.

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**: Example.
- The Literary Digest الملخص الأدبي magazine conducted a poll to predict the result of the 1936 presidential election between Franklin Roosevelt (Democrat and incumbent) and Alf Landon (Republican).
- At the time, the magazine's polls were famous because they had correctly predicted three successive elections.
- In 1936, Literary Digest mailed questionnaires to 10 million people and asked how they planned to vote. The sampling frame was constructed from telephone directories, country club memberships, and automobile registrations.

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**: Example.
- Approximately 2.3 million people returned the questionnaire. The Digest predicted that Landon would win, getting 57% of the vote. Instead, Landon actually got only 36%, and Roosevelt won in a landslide
- This survey had two severe problems
  - Sampling bias due to undercoverage of the sampling frame and a nonrandom sample

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**: Example.
- In 1936, the United States was in the Great Depression. Those who had cars and country club memberships and thus received questionnaires tended to be wealthy. The wealthy tended to be primarily Republican, the political party of Landon. Many potential voters were not on the lists used for the sampling frame
- There was also no guarantee that a subject in the sampling frame was a registered voter
- **Nonresponse bias**: Of the 10 million people who received questionnaires, 7.7 million did not respond — As might be expected, those individuals who were unhappy with the incumbent (Roosevelt) were more likely to respond

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**: Example.
- For this same election, a pollster منظم استطلاع رأي who was getting his new polling agency off the ground surveyed 50,000 people and predicted that Roosevelt would win. Who was this pollster? George Gallup. The Gallup organization is still with us today.
- However, the Literary Digest went out of business soon after the 1936 election

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**:
- Many people think that as long as the sample size is large, it doesn't matter how the sample was selected. This is incorrect as illustrated by the Literary Digest poll. A sample size of 2.3 million did not prevent poor results

Bad Data

Nonrepresentative Training Data:

- **Sampling bias**:
- Consider trying to estimate the average grade point average (GPA) on your campus. If you select 10 students from the library, you can produce an estimate that is biased toward higher GPAs (assuming those students who go to the library are studying more than those who don't). If you, instead, gather a larger sample of 30 students from the library, you still have an estimate that is biased toward higher GPAs.
- The larger sample size did not fix the problem of sampling in the library
- We're almost always better off with a simple random sample of 100 people than with a volunteer sample of thousands of people.