Information Theory for Data Science

Last modified: 2026-11-21 00:32:35 (PST)

Welcome

Information theory quantifies uncertainty and information: how many bits a message needs, how much one random variable tells us about another, and how fast data can cross a noisy channel. Data science uses its quantities directly: cross-entropy is the quantity that logistic regression and neural networks minimize, Kullback–Leibler divergence links the fit of a model to the data it describes, and mutual information drives the information gain of decision trees.

These notes gather the information theory that data science courses use. They are new. The pages written so far are:

- Entropy and mutual information: surprisal, joint and conditional entropy, mutual information, and the information gain of decision trees
- Source coding: the source coding theorem, and cross-entropy as the cost of coding with a code built for the wrong source
- Huffman codes: prefix codes, the Huffman code and its optimality, and its length within one bit of the entropy
- Stream codes: the arithmetic code and the Lempel–Ziv code
- Channel capacity: discrete memoryless channels, the capacity of the binary symmetric channel and the binary erasure channel, and the noisy-channel coding theorem
- Error-correcting codes: Hamming distance, the Hamming and Reed–Solomon codes, and how locally decodable codes lead to hardness amplification and the PCP theorem

The planned topics still to come are:

- KL divergence and its link to fitting models to data
- proofs of the noisy-channel coding theorem, and codes that approach capacity

Entropy, cross-entropy and KL divergence are defined in *Probability for Data Science*; pages here link to those definitions rather than repeating them. Computational complexity has its own library, *Computational Complexity for Data Science*. Kolmogorov complexity, which relates the shortest description of a single string to Shannon entropy, is covered there.

Where to read about the planned topics

Stanford's information theory courses cover many of the planned topics above, and they publish their course pages:

- EE 276, *Information Theory* (Tsachy Weissman) covers entropy and mutual information, the asymptotic equipartition property, prefix codes and Huffman codes, channel capacity and the channel coding theorem, and lossy compression. Its reading page lists books and the founding papers of the field.
- EE 274, *Data Compression, Theory and Applications* covers entropy coding and how tools such as gzip and bzip2 work, then lossy coding of pictures and video.
- Stats 311/EE 377, *Information Theory and Statistics* (John Duchi) applies these tools to statistics and machine learning. It covers:
  - KL divergence and other divergences
  - the data processing inequality
  - concentration inequalities and PAC-Bayes bounds
  - differential privacy
  - lower bounds on estimation error, by the methods of Le Cam, Fano and Assouad
  - proper loss functions and calibration

  Its textbook, Duchi's *Information Theory, Learning, and Statistics*, is a draft that Cambridge University Press lets him keep free online before publication.

The courses' own notes, slides and textbook draft carry no license, so these notes link to them and do not reuse their text.

Further reading

- Dave's review of MacKay's *Information Theory, Inference, and Learning Algorithms* (2024) describes how the book connects compression, noisy-channel coding and neural networks.

Using these notes in another site

Course sites link to these pages by URL; they do not include this repository as a git submodule. A host site that keeps a copy of this repository at its root, named `ids`, can still include fragments with paths that start with `ids/`. This site includes its own fragments the same way, through a `ids` symlink that points at the repository root.

Quarto resolves `@id` cross-references only within one rendered page, so a host site that links to a result here uses an explicit link to the page and anchor.