

L01: Course introduction & motivating example for variable-length encoding

Be prepared to briefly state:

- 1. What area are you working on? Who is your PhD advisor?*
- 2. What do you hope to get out of this course 😊*
- 3. What is your biggest fear for this course 😞*
- 4. What the topic from the course web page that you are most familiar with or excited about?*

Wolfgang Gatterbauer, Javed Aslam

cs7840 Foundations and Applications of Information Theory (fa24)

<https://northeastern-datalab.github.io/cs7840/fa24/>

9/4/2024

A few examples for
what this class is all about

Shannon [1948]: Communicating over a noisy channel

The Bell System Technical Journal

Vol. XXVII

July, 1948

No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

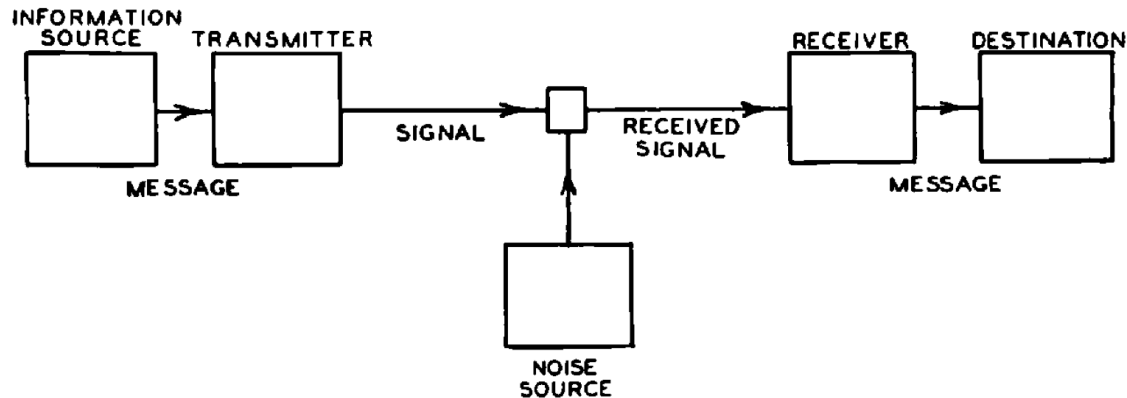


Fig. 1—Schematic diagram of a general communication system.

The entropy in the case of two possibilities with probabilities p and $q = 1 - p$, namely

$$H = -(p \log p + q \log q)$$

is plotted in Fig. 7 as a function of p .

The quantity H has a number of interesting properties which further substantiate it as a reasonable measure of choice or information.

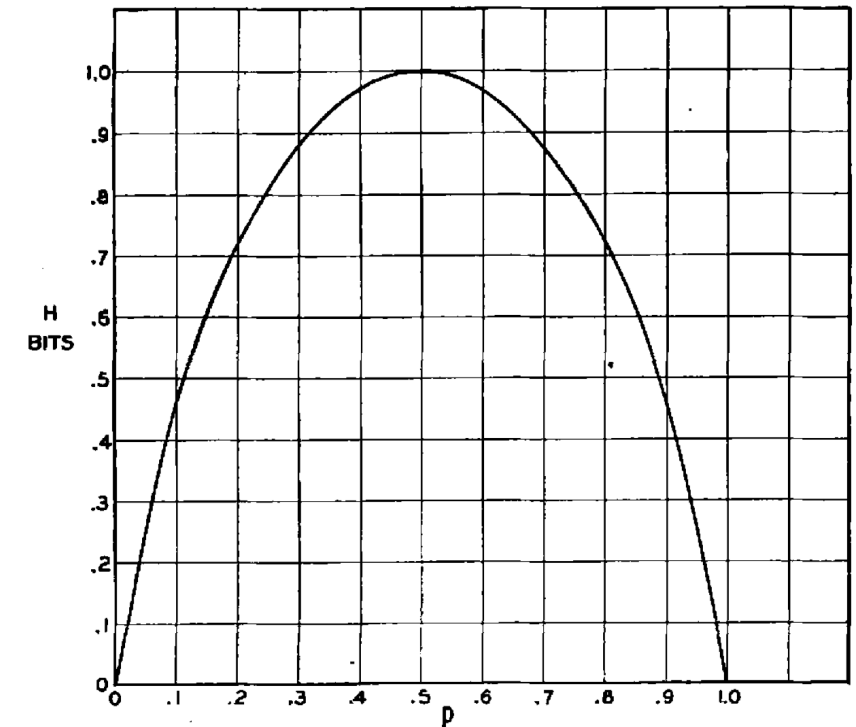


Fig. 7—Entropy in the case of two possibilities with probabilities p and $(1 - p)$.

"Communication": Randomness, Compressibility, Predictability

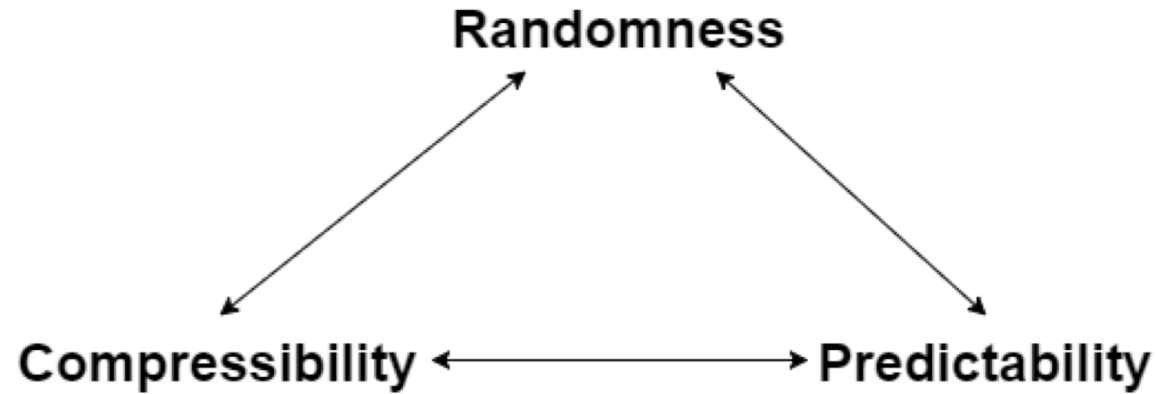


Figure 1.1: Connections between phenomenon properties

- Randomness- How random is a phenomenon
- Predictability- How predictable is a phenomenon
- Compressibility- How compressible is the information associated with a phenomenon

The course is built upon the deep relationships between these three concepts.

Why $\sum_i p_i \log(p_i)$ to measure the "amount" of uncertainty?



Why $\sum_i p_i \log(p_i)$ to measure the "amount" of uncertainty?

Shannon [1948] established that the only meaningful way to measure the amount of uncertainty in evidence expressed by a probability distribution function p_i on a finite set is to use a functional of the form

$$-a \sum_i p_i \log_b(p_i) \quad \text{usually } a = 1, b = 2 \text{ (bits)}$$

How can you establish something like that? 

Why $\sum_i p_i \log(p_i)$ to measure the "amount" of uncertainty?

Shannon [1948] established that the only meaningful way to measure the amount of uncertainty in evidence expressed by a probability distribution function p_i on a finite set is to use a functional of the form

$$-a \sum_i p_i \log_b(p_i) \quad \text{usually } a = 1, b = 2 \text{ (bits)}$$

How can you establish something like that?

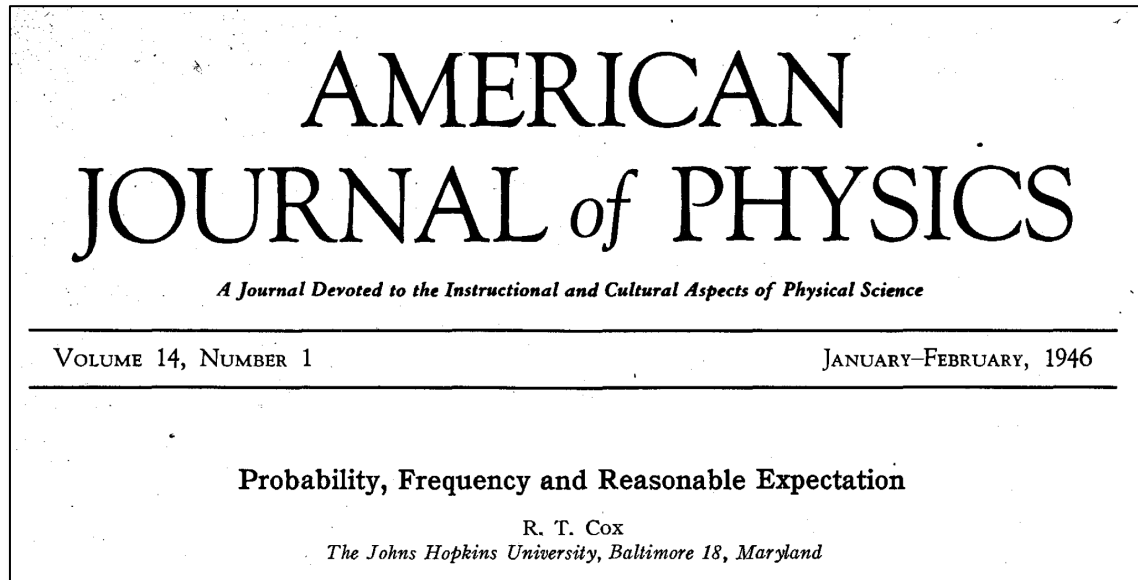
Via an "axiomatic derivation":

To understand the meaning of $-\sum p_i \log(p_i)$, first define an information function I in terms of an event i with probability p_i . The amount of information acquired due to the observation of event i follows from Shannon's solution of the fundamental properties of [information](#):^[12]

1. $I(p)$ is [monotonically decreasing](#) in p : an increase in the probability of an event decreases the information from an observed event, and vice versa.
2. $I(1) = 0$: events that always occur do not communicate information.
3. $I(p_1 \cdot p_2) = I(p_1) + I(p_2)$: the information learned from [independent events](#) is the sum of the information learned from each event.

There are alternative derivations (see e.g. [Jaynes'03 Probability theory: the logic of science])

Cox's axiomatic derivation of probability theory [1946]



"R. T. Cox (1946) published a paper that showed that any set of rules for inference, in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to the Laplace- Jeffreys rules, that is (1)-(3), or inconsistent."
Evans (2002)

"Kolmogorov (1950) is widely quoted as the author of the axiomatic basis of probability calculus, but it was R.T. Cox (1946, 1961) who showed that no other calculus is admissible. The only freedom is to take some monotonic function instead, such as $100 \Pr()$ (percentage) or $\Pr()=(1 \text{ ?? } \Pr())$ (odds), but such changes are merely cosmetic. It follows that other methods are either equivalent to probability calculus (in which case they are unnecessary), or are wrong."
Skilling, 1998

"A third justification for belief as probability (or at least a scaled version of probability) appeared in a paper by R.T. Cox in the American Journal of Physics in 1946 [9]. Cox's proof is not, perhaps, as rigorous as some pedants might prefer and when an attempt is made to fill in all the details some of the attractiveness of the original is lost. Nevertheless his results certainly provide a valuable contribution to our understanding of the nature of belief.

We state here a rigorous version of Cox's main theorem which has aspects which are both stronger and weaker than the original. Slightly stronger versions still can be proved but the increased complications do not seem to justify doing so."

Paris, 1994, page 24

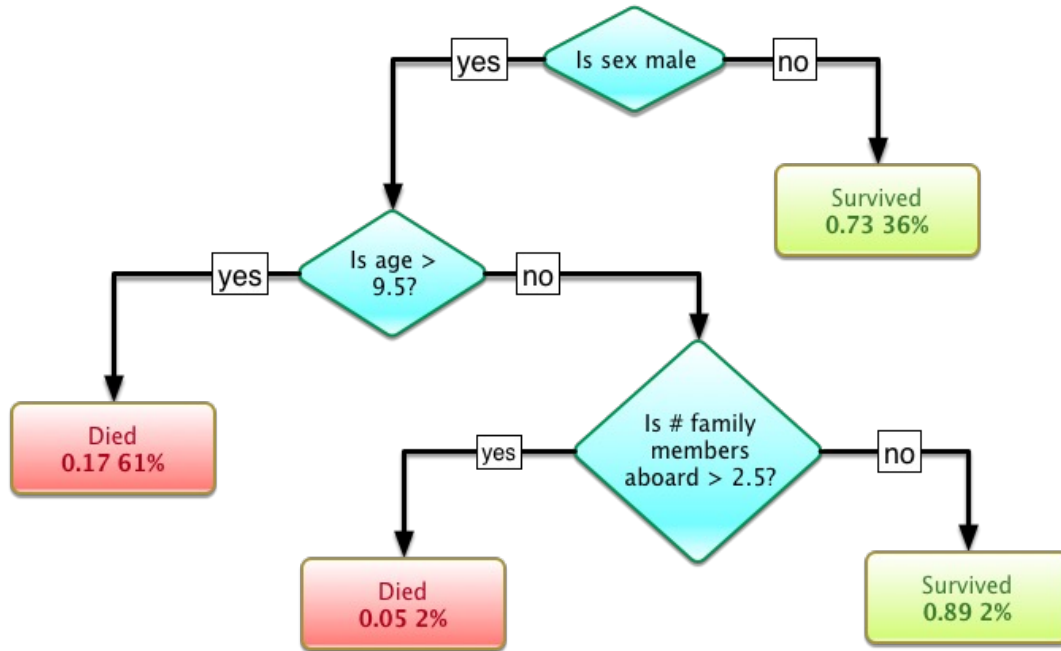
An application in IR: learning decision trees

Try to "best" explain the chances of survival for passengers on the Titanic based on various attributes like gender, age, number of spouses or siblings aboard ("sibsp")?



An application in IR: learning decision trees

Try to "best" explain the chances of survival for passengers on the Titanic based on various attributes like gender, age, number of spouses or siblings aboard ("sibsp")?



How can an algorithm be possible "guided" ?



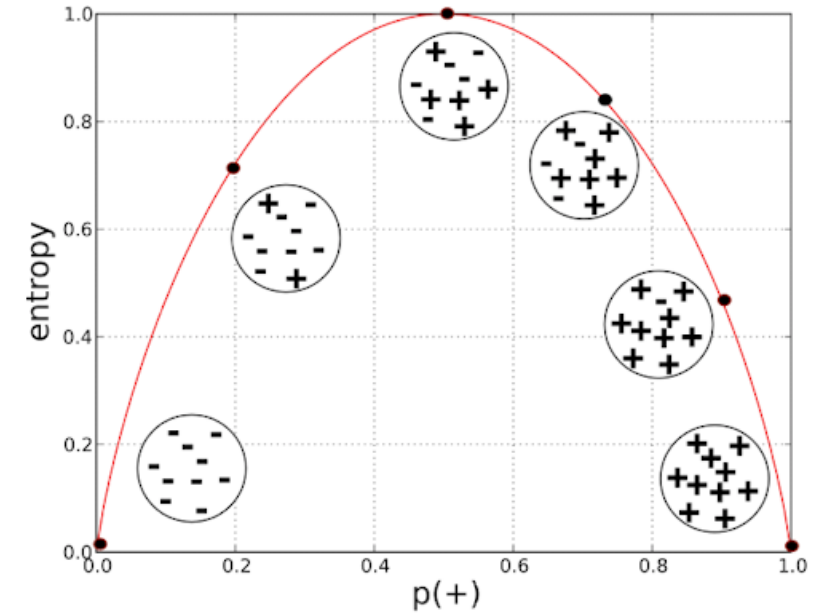
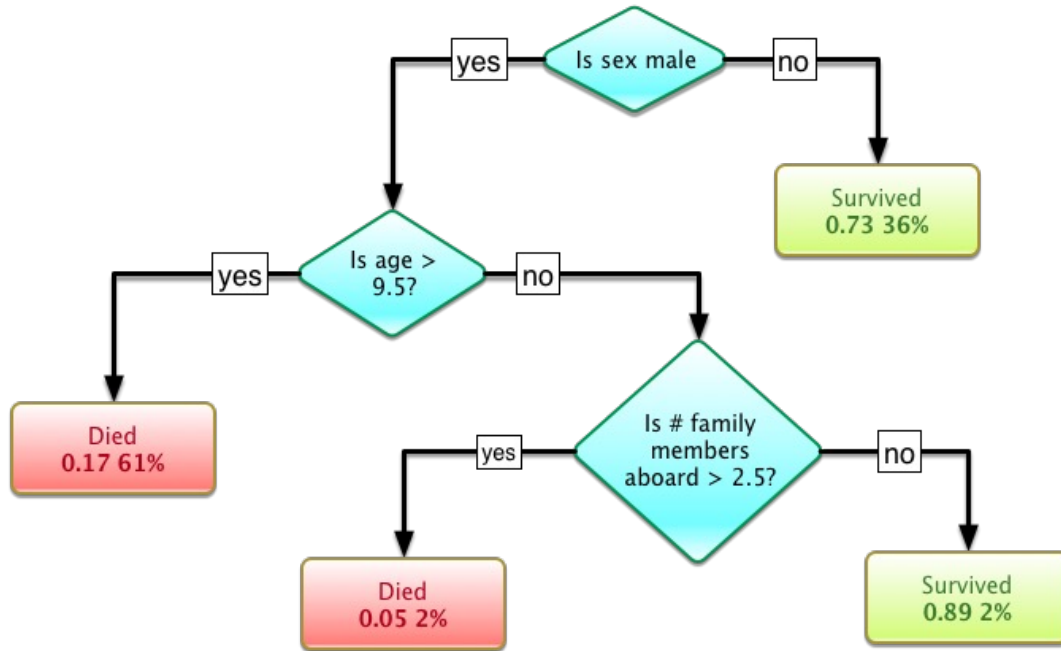
Your chances of survival were good if you were (i) a female or (ii) a male ≤ 9.5 years old with $<$ than 3 siblings.

(The figures under the leaves show the probability of survival and the percentage of observations in the leaf)

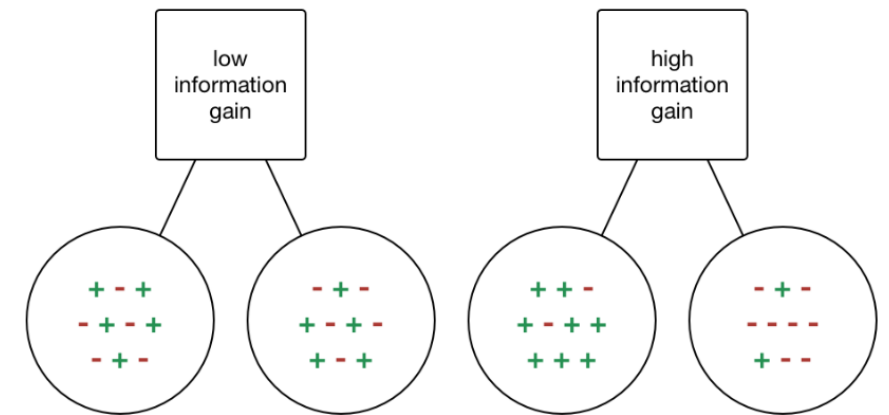
An application in IR: learning decision trees

Try to "best" explain the chances of survival for passengers on the Titanic based on various attributes like gender, age, number of spouses or siblings aboard ("sibsp")?

"learning" =
"compressing"



(simplified for binary choices)



Your chances of survival were good if you were (i) a female or (ii) a male ≤ 9.5 years old with $<$ than 3 siblings.

(The figures under the leaves show the probability of survival and the percentage of observations in the leaf)

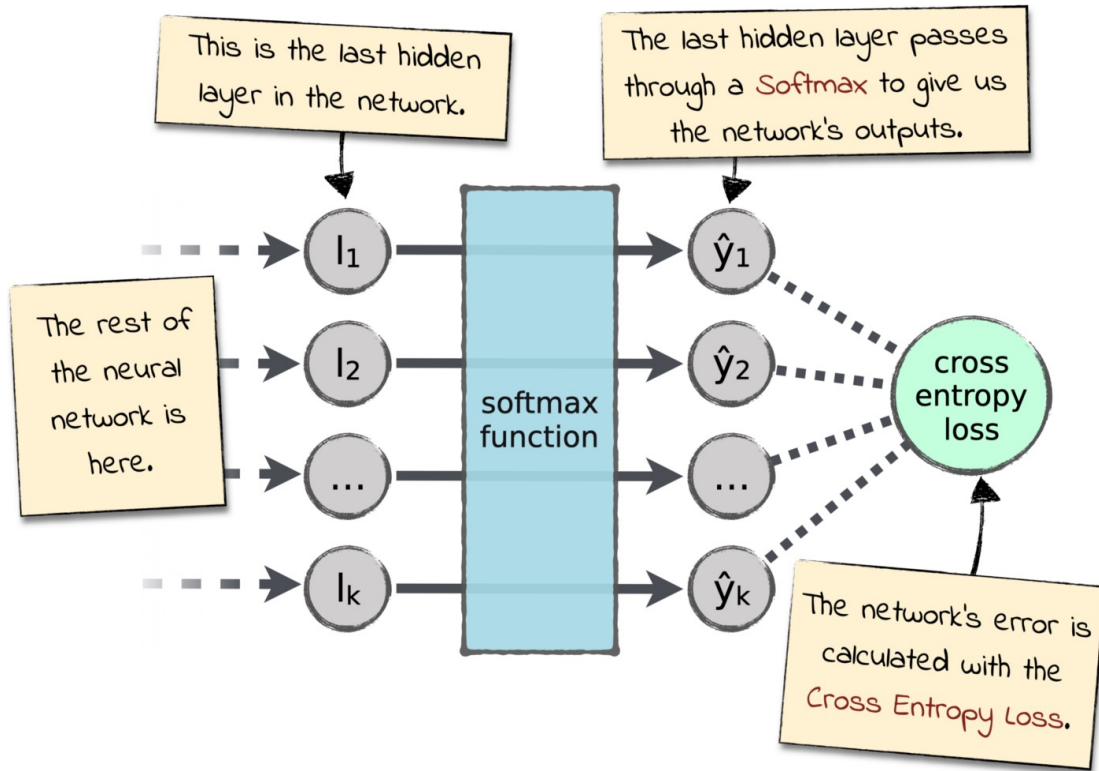
"An application" in ML: DNN

Cross entropy, loss function, softmax, multinomial logistic regression, maximum entropy models, ...



"An application" in ML: DNN

Cross entropy, loss function, softmax, multinomial logistic regression, maximum entropy models, ...



This is the i -th output of the softmax (and also of the entire neural network).

$$L_i = -y_i \log(\hat{y}_i)$$

This is the i -th component of the label.

Ilya Sutskever @ Simons [2023]



An Observation on Generalization

Workshop	<u>Large Language Models and Transformers</u>
Speaker(s)	<u>Ilya Sutskever (OpenAI)</u>
Location	Calvin Lab Auditorium
Date	Monday, Aug. 14, 2023
Time	3 – 4 p.m. PT

Conditional Kolmogorov complexity as the solution

- If C is a computable compressor, then:

For all x ,

$$K(Y|X) < |C(Y|X)| + K(C) + O(1)$$

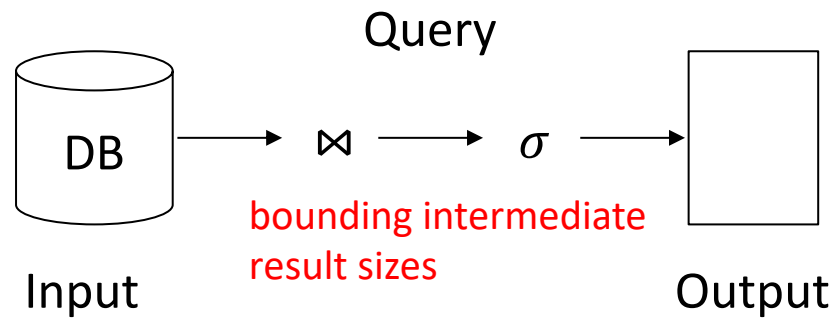
Conditioning on a **dataset**, not an example

Will extract all "value" out of X for predicting Y

So this is the solution
to unsupervised learning--



An application in DB: Query optimization, cardinality estimation



What do Shannon-type Inequalities, Submodular Width, and Disjunctive Datalog have to do with one another?

Mahmoud Abo Khamis
LogicBlox Inc.

Hung Q. Ngo
LogicBlox Inc.

Dan Suciu
LogicBlox Inc. and
University of Washington

PODS'17, May 14 - 19, 2017, Chicago, IL, USA

Overall, our results showed a deep connection between three seemingly unrelated lines of research; and, our results on proof sequences for Shannon flow inequalities might be of independent interest.

Galley: Modern Query Optimization for Sparse Tensor Programs

Kyle Deeds
kdeeds@cs.washington.edu
University of Washington
United States

Magda Balazinska
magda@cs.washington.edu
University of Washington
United States

Willow Ahrens
wahrens@mit.edu
Massachusetts Institute of Technology
United States

Dan Suciu
suciu@cs.washington.edu
University of Washington
United States

If interested, please subscribe to and join our
DATAlab seminar this FRI, Sept 6th @ noon
<https://db.khoury.northeastern.edu/activities/>

Overall motivation for studying the topics of this class

- Information Theory as unified language and mathematical tool to understand and predict phenomena related to data and information
- We cover both:
 - theory:
 - Part 1: basic theory of information theory
 - Part 2: the axiomatic approach (at least for entropy)
 - selected applications to
 - ML,
 - IR,
 - data management (DB)

Very approximate outline (will likely change as we progress)

PART 1: Information Theory (the basics)

Covers the basic mathematical framework behind entropy and its various forms.

- **Lecture 1 (Wed 9/4):** Basics of Probability: Random variables, independence, conditional independence, Markov chains
- **Lecture 2 (Mon 9/9):** Entropy, joint entropy, conditional entropy, mutual information
- **Lecture 3 (Wed 9/11):** KL divergence, conditional mutual information, cross entropy, log loss & basic inequalities
- **Lecture 4 (Mon 9/16):** Data processing theorem
- **Lecture 5 (Wed 9/18):** Fano inequality, max entropy distributions
- **Lecture 6 (Mon 9/23):** More inequalities, “I-measures” can be negative
- **Lecture 7 (Wed 9/25):** Information inequalities, Shannon-type inequalities
- **Lecture 8 (Mon 9/30):** Placeholder

PART 2: The axiomatic approach (deriving formulations from first principles)

Covers the axiomatic approach from multiple angles: a few simple principles (axioms) leading to entropy or the laws of probability up to factors. Starting from a list of postulates leading to particular solution is a powerful approach that has been used across different areas of computer science (e.g. how to define the right scoring function for achieving a desired outcome)

- **Lecture 9 (Wed 10/2):** Cox’ theorem: a derivation of the laws of probability theory from a certain set of postulates
- **Lecture 10 (Mon 10/7):** Contrast with Kolmogorov’s “probability axioms”
- **Lecture 11 (Wed 10/9):** Derivation of entropy function from first principles
- **(Mon 10/14): no class (Indigenous Peoples Day)**
- **Lecture 12 (Wed 10/16):** Maximum entropy solutions via Lagrange formulation

Part 3: Selected Applications to data management, machine learning and information retrieval

Covers example approaches of basic ideas from information theory to practical problems in data management, machine learning, and information retrieval. Topics and discussed papers may vary over years.

- **Lecture 13 (Mon 10/21):** Derivation of ID3 for decision trees
- **Lecture 14 (Wed 10/23):** Connections (multinomial) logistic regression, maximum entropy models, softmax, cross-entropy, loss functions
- **Lecture 15 (Mon 10/28):** Bradley-Terry model, Luce's choice axiom
- **Lecture 16 (Wed 10/30):** Minimum Description Length (MDL)
- **Lecture 17 (Mon 11/4):** Applications of VC dimensions
- **Lecture 18 (Wed 11/6):** An Information-Theoretic Approach to Normal Forms
- **(Mon 11/11): no class (Veterans Day)**
- **Lecture 19 (Wed 11/13):** Information loss in acyclic join schemas
- **Lecture 20 (Mon 11/18):** Information Inequalities for Cardinality estimation
- **Lecture 21 (Wed 11/20):** Measuring approximate Functional Dependencies
- **Lecture 22 (Mon 11/25):** Explanation tables
- **(Wed 4/5): no class (Fall break)**
- **Lecture 23 (Mon 12/2):** Inverse document frequency
- **Lecture 24 (Wed 12/4):** Placeholder

Project presentations

- **Lecture 25 (Mon 12/9):** P4 Project presentations
- **Lecture 26 (Wed 12/11):** P4 Project presentations

Quick background on our two instructors

- Javed Aslam: <https://www.khoury.northeastern.edu/home/jaa/>



- Wolfgang Gatterbauer: <https://gatterbauer.name/>



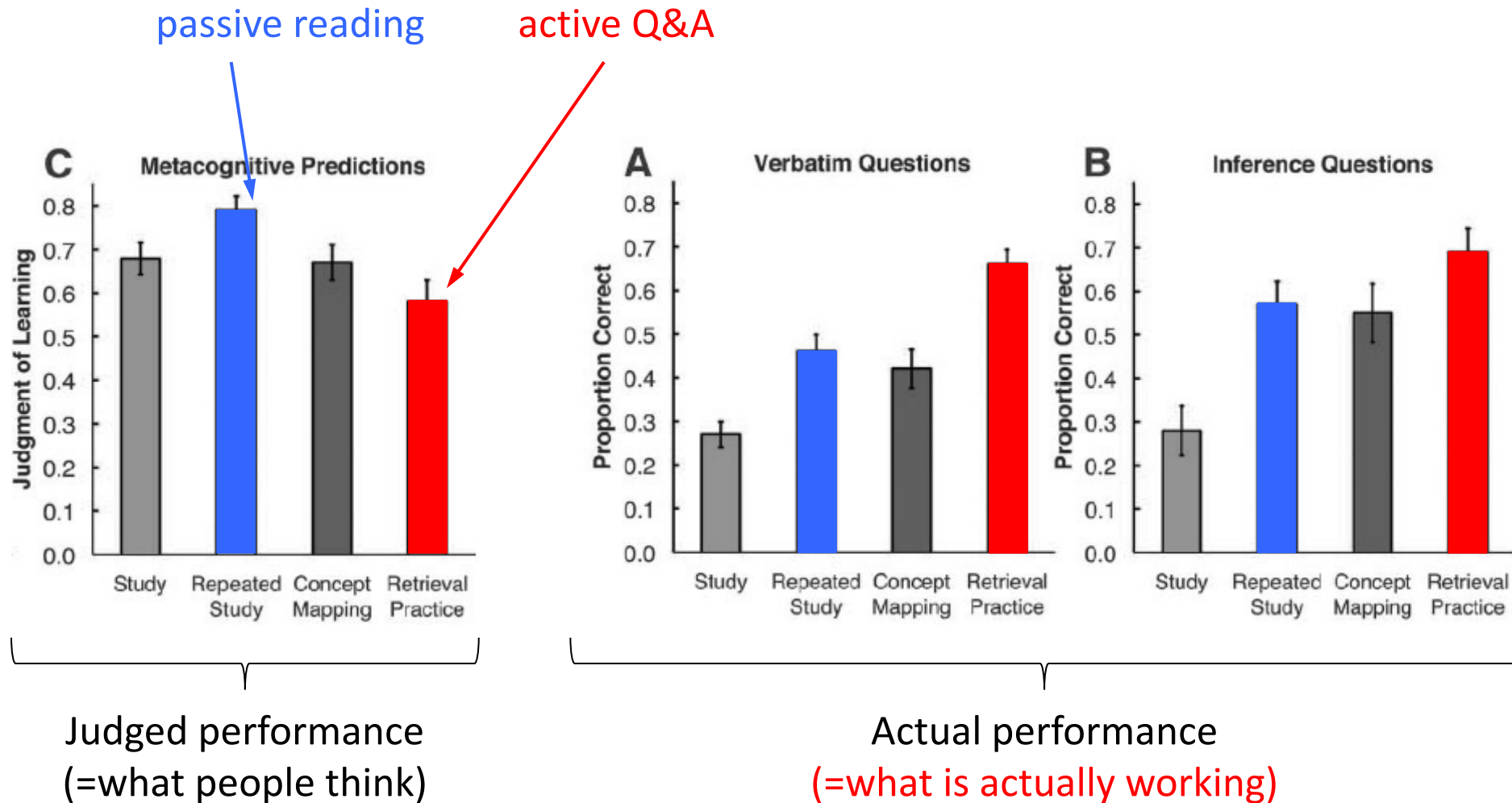
Let's take turns

As you are called, please briefly state:

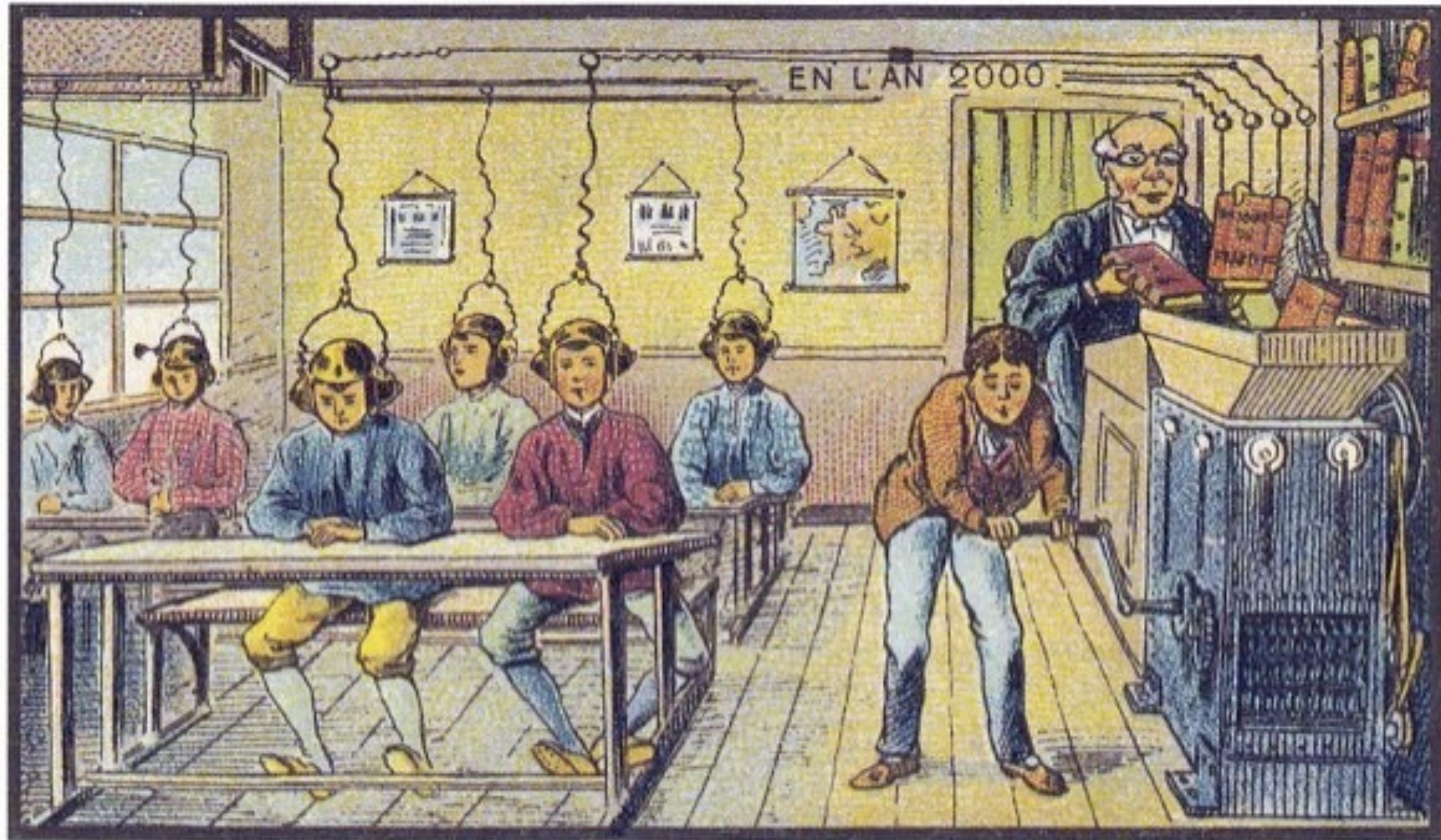
1. What area are you working on? Who is your PhD advisor?
2. What do you hope to get out of this course 😊
3. What is your biggest fear for this course 😞
4. What the topic from the course web page (or more generally information theory) that you are most familiar with, or most excited about?

Pedagogy & Logistics

Studying new material: "Under which study condition do you think you learn better?"



The year 2000 imagined in 1900



At School

Source: <https://publicdomainreview.org/collection/a-19th-century-vision-of-the-year-2000>

Gatterbauer, Aslam. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/>

Late 1950s



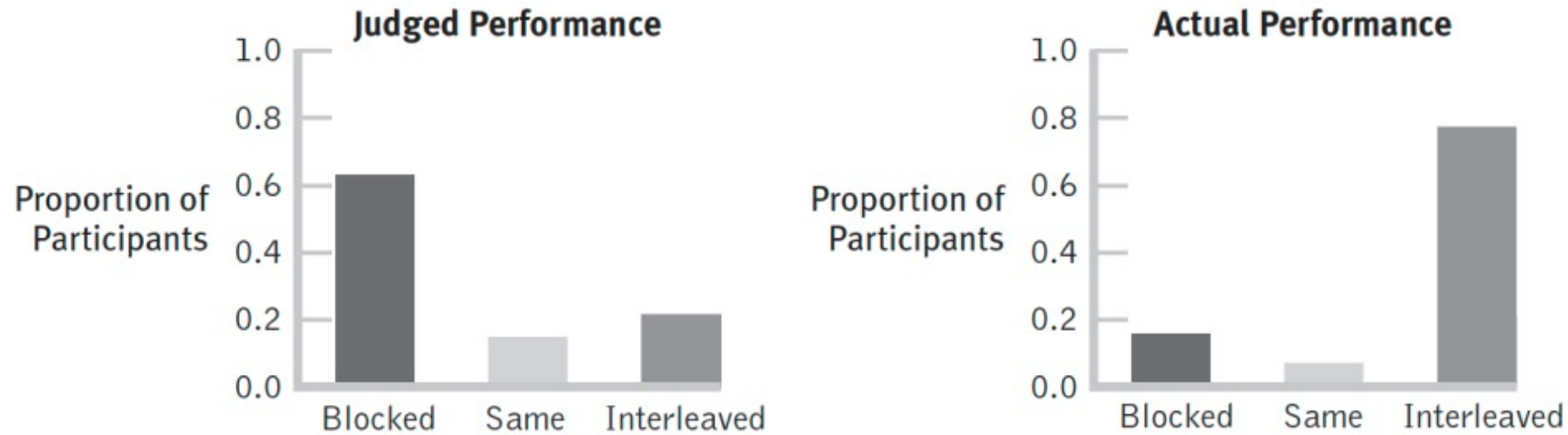
Source: <http://yankeeinexile.wordpress.com/2013/02/17/the-futures-of-the-past/>
Gatterbauer, Aslam. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/>

Predicting the role of IT in Education



◁ Learning by computer in the future will be fun. This computer is displaying a chemistry experiment for the older child and arithmetic problems for the younger one. The computer controls include light pens to draw on the screens. The chemistry student has done something wrong and has caused an explosion!

Sequencing Material: "Under which teaching condition do you think you learn better?"



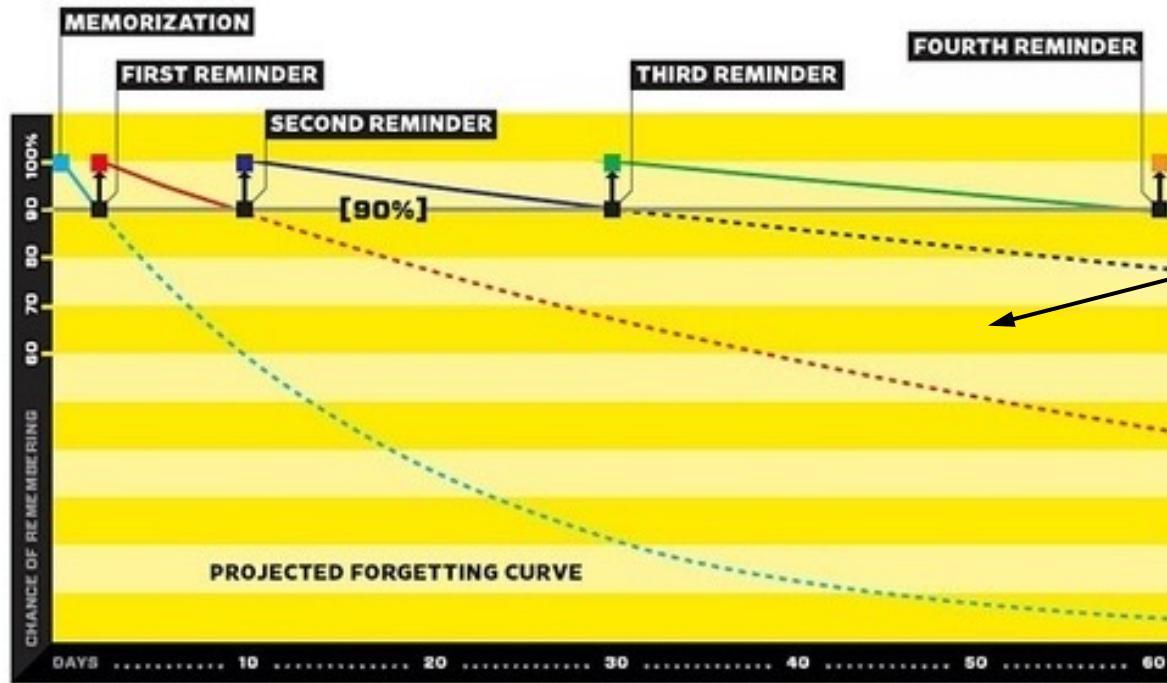
The mix of chapter and cases is also meant to provide a holistic view of how technology and business interrelate. Don't look for an "international" chapter, an "ethics" chapter, a "mobile" chapter, or a "systems development and deployment" chapter. Instead, you'll see these topics woven throughout many of our cases and within chapter examples. This is how professionals encounter these topics "in the wild, so we ought to study them not in isolation but as integrated parts of real-world examples. Examples are consumer-focused and Internet-heavy for approachability, but the topics themselves are applicable far beyond the context presented.

Data from: Bjork & Bjork, "Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning," 2011. <https://psycnet.apa.org/record/2011-19926-008>

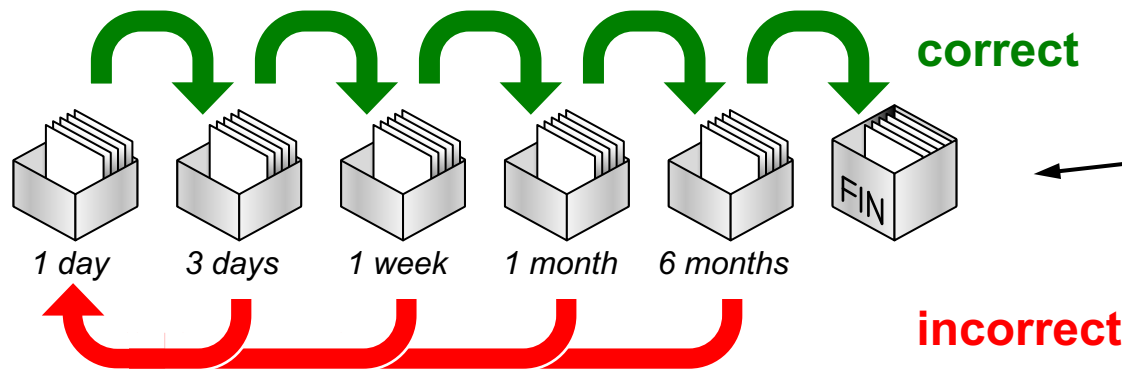
Paragraph from: "Information Systems: A Manager's Guide to Harnessing Technology (book v1.4)," Gallaughier, 2012. <https://gallaughier.com/book/>

Gatterbauer, Aslam. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/>

Spaced Repetition



Ebbinghaus Forgetting Curve



Leitner System
(Pimsleur's graduated interval recall)

Coursework / Evaluation (1/3)

Stop me and ask questions if
I am talking too much!

"your obligation
to dissent"

15%: Class participation: Classes will be interactive and require concentration and participation of the students. We are a big fan of the **Socratic Method (video clip from the 1973 movie "The Paper Chase")**. Participate when we discuss the merits or shortcomings of algorithms, or when we have small group break-out sessions with exercises. Ask questions, during class or on Piazza. Questions that make us ponder or create new illustrating examples are all great examples of class participation. Also, *never* hesitate to point out to us any errors you spot in the slides, even if minor. You can post anonymously to the other students on Piazza (and even anonymously to the instructors, though then we would not be able to associate you with your greatly appreciated participation). Also, don't hesitate to point out to us any interesting links to interesting related material. It can only count towards class participation. Finally notice that while the class provides extensive readings for those interested, these pointers are almost exclusively optional (unless otherwise stated in class).

Lectures are not recorded (1/2)

If gaps in knowledge are the seeds of curiosity, exploration is the sunlight. Hundreds of studies with thousands of students have shown that when science, technology and math courses include active learning, students are less likely to fail and more likely to excel. A key feature of active learning is interaction. But too many online classes have students listening to one-way monologues instead of having two-way dialogues. Too many students are sitting in front of a screen when they could be exploring out in the world.

Lectures are not recorded (2/2)

- We would like to have **an encouraging environment** in which everyone can speak up and discuss ideas freely without concern that discussions will be available outside of classroom.
- The course slides are comprehensive and should allow you to be able to remember the key lessons from class (except for background stories I may tell you). Lecture slides will be posted after each class, usually by end of the day.
- Do not record or otherwise share the classroom video calls yourself. The Commonwealth of Massachusetts's wiretapping law requires "two-party consent". It is a felony to secretly record a conversation, whether the conversation is in person or taking place by telephone or another electronic medium. [See Mass. Gen. Laws ch.272, § 99].

A suggestion on how to best use class time!

- It is ok to make mistakes in class. Making mistakes in class is actually the best thing that can happen to you. You learn and will never make it again 😊
- From Ray Dalio's Principles (2017):
 - "Create a Culture in Which It Is Okay to Make Mistakes and Unacceptable Not to Learn from Them"
 - "Recognize that mistakes are a natural part of the evolutionary process."
 - "Don't feel bad about your mistakes or those of others. Love them!"

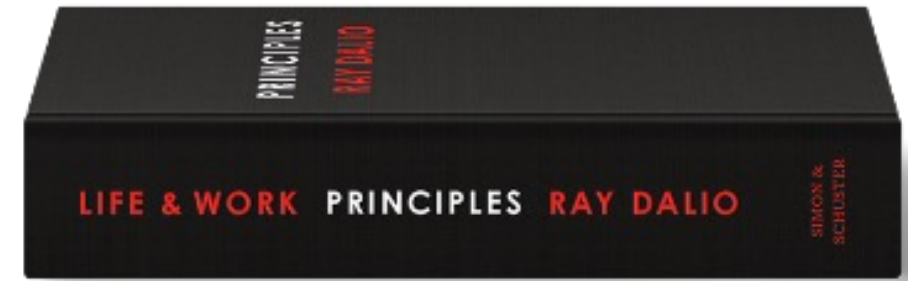
Notice the story around Bridgewater & Ray Dalio is interesting and still being written. See e.g.

<https://www.nytimes.com/2023/11/01/business/how-does-the-worlds-largest-hedge-fund-really-make-its-money.html>

<https://www.vanityfair.com/news/2023/11/james-comey-dalio-bridgewater-the-fund>

<https://nymag.com/intelligencer/article/ray-dalio-rob-copeland-the-fund-book-excerpt.html>

That said, the ideas behind "Principles" are still worth reading



One reason why I don't post slides *before* lecture

From the preamble of one of the best physics books ever: „How to read this book“

The best way to use this book is NOT to simply read it or study it, but to read a question and STOP. Even close the book. Even put it away and THINK about the question. Only after you have formed a reasoned opinion should you read the solution. Why torture yourself thinking? Why jog? Why do push-ups?

If you are given a hammer with which to drive nails at the age of three you may think to yourself, “OK, nice.” But if you are given a hard rock with which to drive nails at the age of three, and at the age of four you are given a hammer, you think to yourself, “What a marvelous invention!” You see, you can't really appreciate the solution until you first appreciate the problem.

...

We will also have in-class whiteboard lectures and exercises!

...

Let this book, then, be your guide to mental push-ups. Think carefully about the questions and their answers *before* you read the answers offered by the author. **You will find many answers don't turn out as you first expect. Does this mean you have no sense for physics? Not at all. Most questions were deliberately chosen to illustrate those aspects of physics which seem contrary to casual surmise. Revising ideas, even in the privacy of your own mind, is not painless work.** But in doing so you will revisit some of the problems that haunted the minds of Archimedes, Galileo, Newton, Maxwell, and Einstein.* The physics you cover here in hours took them centuries to master. Your hours of thinking will be a rewarding experience. Enjoy!

Lewis Epstein

One reason why I don't post slides *before* lecture

From the preamble of one of the best physics books ever: „How to read this book“

The best way to use this book is NOT to simply read it or study it, but to read a question and STOP. Even close the book. Even put it away and THINK about the question. Only after you have formed a reasoned opinion should you read the solution. Why torture yourself thinking? Why jog? Why do push-ups?

If you are given a hammer with which to drive nails at the age of three you may think to yourself, “OK, nice.” But if you are given a hard rock with which to drive nails at the age of three, and at the age of four you are given a hammer, you think to yourself, “What a marvelous invention!” You see, you can't really appreciate the solution until you first appreciate the problem.

...

You must avoid the temptation to look at answers until you have tried to find and ideally write out the solution yourself!

...

We will also have in-class whiteboard lectures and exercises!

Study groups are great for learning material!

- "... The groups of students who were doing best spontaneously formed study groups..."
- Students who were not doing as well tended to do as the instructor suggested-study two hours out of class for every hour in class-but did it by themselves with little social support...
- "... even well-prepared students (high math SATs) are often disadvantaged by high school experiences that lead them to work alone."

The "Surfer Analogy" for time management



Class scribes

"To ask the right question is harder than to answer it."

Georg Cantor

Quote: <https://www.azquotes.com/quote/909685>

Also see: <https://www.inc.com/lisa-calhoun/elon-musk-on-the-1-creative-skill-every-founder-needs-now.html>

Gatterbauer, Aslam. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/>

Coursework / Evaluation (2/3)

35%: Flipped class scribes: Students take turns in "illustrating" 7 lectures (to be consistent with standard notation we refer to it as "scribes"). Graduate theory classes often ask students to scribe the lecture content. However, we change the rule of the game. Rather than scribing (= repeating and summarizing) the content of the class, we ask you to "illustrate" *some interesting aspect in the covered topics with imaginative and ideally tricky illustrating examples*. Those illustrations are great if they in turn can help other students practice and solidify their understanding of the topics discussed. Please start either from our **PPTX template** or use your own template (as long as you include slide numbers), and submit it as PDF to Canvas, naming it "cs7840-fa24-[YOUR NAME]-scribe[NUMBER]-[SOME DESCRIPTIVE TITLE].PDF".

Justification: Georg Cantor is quoted as saying: "To ask the right question is harder than to answer it." In that spirit, our class scribes are closer to research than assignments: What particular aspect in a class is worthy to be "illustrated"? That's already part of the question. Scribes are done in PowerPoint and are due 1 week after class at midnight (Mon for Mon classes, Wed for Wed classes). They can be done individually or in teams of two (if you work in teams of two, you are expected to illustrate 14 classes instead of 7 classes). For some more pedagogic motivation see videos by **Tim Brown on asking questions and reframing problems being key to creativity**, **Dan Meyer on formulating problem being more important than just solving them**, **Derek Muller on increasing learning by including possible misconceptions into stories**, a **blog post on example-based reasoning**, and an older text of mine of the educational **value of temporarily misleading the spectator before giving the correct answer**.

Procedure:

- Optional (but strongly suggested) preliminary submission on Piazza: You can post your first draft of each assignment as PDF to Piazza. If you prefer, you can make your post anonymous to other students (please then simply remove the title page). The instructors will post comments on each submission on Piazza for you and everyone else to see. You may decide to address our feedback in your final submission to Canvas. By posting it visible to other students, both your document and also my feedback may also be helpful to other students. Notice that submitting the preliminary version to Piazza extends the deadline for submitting on Canvas.
- Final submission on Canvas: Submit your final version to Canvas. Please notice that a scribe can be submitted on a given class topic until maximally 7 days after the slides for that class topic are posted (your earlier submission date counts, either for the preliminary on Piazza, or the final on Canvas)

Coursework / Evaluation (3/3)

50%: Course project: The main component of this course will be a **research project** in the latter part of this class. This project can be a new application of one of the techniques presented or theoretically-oriented. The topic will be flexible, allowing students to explore scalable data management and analysis aspects related to their individual PhD research. This will involve an initial project proposal, an intermediate report, a project presentation and a final report. The final report should resemble a workshop paper, and will be evaluated on the basis of soundness, significance, novelty, and clarity. Deliverables and dates are posted on the **project page**.

Project

Project deliverables

- **P1 (Mon 10/7): Project ideas:** Please submit a few tentative ideas you are considering for the project on Piazza.
- **P2 (Wed 10/23): Project proposal:** Prepare a 1-2-page proposal that includes (i) the title of the project, (ii) a short description of the problem you propose to solve, (iii) a brief outline of how you will approach the problem, and how you will evaluate your results. Do not forget to include a list of references!
Use the [1-column ACM latex template on Overleaf](#) for your report. It includes a number of useful packages. Use the latex commands at the end of these instructions to hide unnecessary information from the ACM template. Submit your proposal as PDF on Canvas. We will read your write-up and add comments and clarifying questions to specific line numbers. Optionally, you can additionally share your document on overleaf with our Northeastern email addresses and we will make my comments directly into your reports (please still submit a PDF time-stamped to Canvas). In that case, please rename your document on Overleaf to "cs7840-fa24-[YOUR NAME]-proposal-[PROJECT TITLE]".
- **P3 (Wed 11/13): Intermediate report:** Build upon your proposal and the TEX template and prepare a 2-5 page document that extends your project proposal. The milestones should include (i) a more detailed description of the problem, (ii) related work, (iii) your progress and the unexpected issues you have encountered so far, and (iv) a brief plan for how you plan to continue your project. In our updated write-up, please refer in an easy-to-distinguish way (e.g. extra paragraphs highlighted in color or bold) to our earlier comments on your initial project proposal and how you choose to address or why you prefer to ignore them. This report is not graded, yet the more information you give us, the better we can help you at a time you can still make amendments. Again, submit your intermediate report on Canvas. If you optionally also share your updated latex document with us on Overleaf, rename it first to "cs7840-fa24-[YOUR NAME]-intermediatereport-[PROJECT TITLE]".
- **P4 (Mon 12/9): Project presentation (20%):** The project presentation counts towards 20% of your grade. Design your presentation for approximately 10-15 min, yet add backup slides to be able to answer technical questions. The presentation is interactive, thus be prepared to answer questions during the talk, which may extend the time needed. In case you use PowerPoint, you can optionally share your PPTX presentation slides in Office 365 online with our Northeastern email until 2 days before your presentation and we will have a quick pass and add suggestions to your slides. Please use the same naming conventions for your slides as for the report: "cs7840-fa24-[YOUR NAME]-[PROJECT TITLE].pptx". Please come with your own laptop to present and make sure to test the setup **before** the day you present. Include page numbers on your slides.
- **P5 (Wed 12/11): Final report (30%):** The final report counts towards 30% of your grade. It should be written like a typical research paper that we have read in class. There is no formal length requirement, but a good target would be 8-12 pages in the single column ACM format. Please make sure to address any feedback shared during the presentation and earlier reports as much as possible. Again, please refer in an easy-to-distinguish way to my earlier comments on your project proposals and presentation and summarize how you choose to address those or why you chose to go a different route (which may well be completely legit). And include illustrating examples and visualizations as much as possible. Again, submit your final report on Canvas. If you optionally also share your final document with us on Overleaf, rename it first to "cs7840-fa24-[YOUR NAME]-finalreport-[PROJECT TITLE]".

A story about citations

Richard E. Pattis
Professor of Teaching
[Department of Computer Science](#)
and [Department of Informatics](#)
[Donald Bren School](#) of Information
and Computer Sciences
[University of California, Irvine](#)
Irvine, CA 92697
pattis@ics.uci.edu
Office: 4062 Bren Hall
Phone: (949) 824-2704
Fax: (949) 824-4056



**EBNF: A Notation to
Describe Syntax**

Interesting Snippets

While developing a manuscript for a textbook on the Ada programming language in the late 1980s, I wrote a chapter on [EBNF](#) and began teaching it on the "first" day of my CS-1 class: primarily as a microcosm of programming, but also as a practical tool for later describing the syntax of Ada. These 21 pages (less than 1/4 the size of the original Karel book) discuss the sequence, choice, option, repetition, and recursion control structures (along with "procedural" abstraction via named EBNF rules). They explore various methods of proving that tokens satisfy descriptions, that descriptions are equivalent (and how to simplify them), and the difference between syntax and semantics. I have continued to use this approach until this day in my CS-1 classes. In fact, I have rewritten [this EBNF chapter](#) for an introduction to Python course I am teaching.

EBNF = "Extended Backus-Naur form"

A story about citations

[5] Richard Feynman and Chapter Objectives. "Ebnf: A notation to describe syntax". In: *Cited on* (2016), p. 10.

[40] Richard Feynman. 'EBNF: A Notation to Describe Syntax'. In: - (2016). URL: <http://www.ics.uci.edu/~pattis/misc/ebnf2.pdf>.

Feynman, R., & Objectives, C. (2016). EBNF A Notation to Describe Syntax, 1–19.

Feynman, Richard, and Chapter Objectives. 2016. *EBNF A Notation to Describe Syntax*.

development. The language, presented in Natural Language, and delineated by an EBNF grammar (Feynman & Objectives, 2016), can be used by IoT engineers as a blueprint for the definition of

une grammaire EBNF (Feynman and Objectives 2016).

[9] Richard Feynman. Ebnf: A notation to describe syntax. Cited on page 10.

63. Feynman, R. Ebnf: A Notation to Describe Syntax. 2016. Available online: <http://www.ics.uci.edu/~pattis/misc/ebnf2.pdf> (accessed on 6 May 2022).

EBNF: A Notation to Describe Syntax



Why are these paper citing R. Feynman (and C. Objectives)?

Scholar 2 results (0.03 sec)

[PDF] Ebnf: A notation to describe syntax

[PDF] uci.edu

R Feynman - ics.uci.edu

Chapter Objectives• Learn the four control forms in EBNF• Learn to read and understand EBNF descriptions• Learn to prove a symbol is legal according to an EBNF description• ...

☆ Save 📄 Cite Cited by 14 Related articles ⌕

[PDF] EBNF: A Notation to Describe Syntax

[PDF] uci.edu

R Feynman - ics.uci.edu

Chapter Objectives• Learn the four control forms in EBNF• Learn to read and understand EBNF descriptions• Learn to prove a symbol is legal according to an EBNF description• ...

📄 Cite

A story about citations

[5] Richard Feynman and Chapter Objectives. "Ebnf: A notation to describe syntax". In: *Cited on* (2016), p. 10.

[40] Richard Feynman. 'EBNF: A Notation to Describe Syntax'. In: - (2016). URL: <http://www.ics.uci.edu/~pattis/misc/ebnf2.pdf>.

Feynman, R., & Objectives, C. (2016). EBNF A Notation to Describe Syntax, 1–19.

Feynman, Richard, and Chapter Objectives. 2016. *EBNF A Notation to Describe Syntax*.

development. The language, presented in Natural Language, and delineated by an EBNF grammar (Feynman & Objectives, 2016), can be used by IoT engineers as a blueprint for the definition of

une grammaire EBNF (Feynman and Objectives 2016).

[9] Richard Feynman. Ebnf: A notation to describe syntax. Cited on page 10.

63. Feynman, R. Ebnf: A Notation to Describe Syntax. 2016. Available online: <http://www.ics.uci.edu/~pattis/misc/ebnf2.pdf> (accessed on 6 May 2022).

Chapter 1

EBNF: A Notation to Describe Syntax

Precise language is not the problem. Clear language is the problem.
Richard Feynman

CHAPTER OBJECTIVES

- Learn the four control forms in EBNF
- Learn to read and understand EBNF descriptions
- Learn to prove a symbol is legal according to an EBNF description
- Learn to determine if EBNF descriptions are equivalent
- Learn to write EBNF descriptions from specifications and exemplars
- Learn the difference between syntax and semantics
- Learn the correspondence between EBNF rules and syntax charts
- Learn to understand the meaning of and use recursive EBNF rules

1.1 Introduction

EBNF is a notation for formally describing syntax: how to write the linguistic features in a language. We will study EBNF in this chapter and then use it throughout the rest of this book to describe Python's syntax formally. But there is a more compelling reason to begin our study of programming with We will use EBNF to describe the syntax of Python

☹ Please cite generously and cite what you read, not citations of citations

Tools

- Canvas:
 - Links to website: with preliminary calendar, optional readings, administrative details, some lectures slides (other parts are only done on whiteboard and pen & paper)
 - Links to Piazza: discussions, questions, errors, follow-up instructions beyond web page; but please let me know if there is a strong preference for Canvas announcements!
 - Canvas calendar / assignments: project milestones, submission for scribes
- Other suggestions?

Piazza extends our classroom – please subscribe

We use Piazza as our main online message board. If I have updates to share, I will post them on Piazza. Thus I recommend you to automatically follow every and note.

→ Click on the arrow on the right upper corner from Piazza → Account/Email settings → Edit Email notifications:

Before	After
<p>Edit Email Notification</p> <p>For new Questions or Notes:</p> <p><input type="radio"/> Real Time</p> <p><input type="radio"/> Daily Digest</p> <p><input checked="" type="radio"/> Smart Digest <input type="text" value="2 hours"/></p> <p><input type="radio"/> No Emails</p> <hr/> <p>For updates to Questions or Notes you follow:</p> <p><input checked="" type="radio"/> Real Time</p> <p><input type="radio"/> No Emails</p> <p><input type="checkbox"/> Automatically follow every question and note.</p> <p><input type="button" value="Save Settings"/> <input type="button" value="Cancel"/></p>	<p>Edit Email Notification</p> <p>For new Questions or Notes:</p> <p><input checked="" type="radio"/> Real Time</p> <p><input type="radio"/> Daily Digest</p> <p><input type="radio"/> Smart Digest <input type="text" value="2 hours"/></p> <p><input type="radio"/> No Emails</p> <hr/> <p>For updates to Questions or Notes you follow:</p> <p><input checked="" type="radio"/> Real Time</p> <p><input type="radio"/> No Emails</p> <p><input checked="" type="checkbox"/> Automatically follow every question and note.</p> <p><input type="button" value="Save Settings"/> <input type="button" value="Cancel"/></p>

Feedback throughout the semester

Please use this simple way to let us know what works or not!

<https://forms.gle/6u7Sut8sdpuY7KLM9>

Even if you find minor annoying issues (spelling mistakes, **broken links**, confusing explanations), please spend a moment to let us know. We will notice your participation and contribution, and it will improve our class to everyone.

Piazza is visible to everyone in this class (you can post anonymously). This feedback form is visible only to us instructors.

CS7870: Anonymous feedback

Your comments will help us (Wolfgang, Javeed) tailor the course as we go along. We are the only ones who can read these comments. Notice that you can also post comments anonymously to other students (though not me) to Piazza where everyone can see your comments. Thanks very much for filling this out!

[Sign in to Google](#) to save your progress. [Learn more](#)

Your name

Optional, only if you want me to get back to you

Your answer

1. Content

Do you understand what we are doing?

1 2 3 4 5 6 7 8 9 10

No clue what is going on Super clear

2. Speed

How is the pace of the course?

1 2 3 4 5 6 7 8 9 10

Sooooooooo slow Way too fast

3. Keep (+)

What is working well for you? What is your favorite part of this class and of our teaching?

Your answer

4. Change (-)

What specific suggestions do you have for changes to improve the course or how we teach it? Anything that you have seen in other classes you wished we adopted as well? Any part of the class content you like us to focus more on?

Your answer

5. Help (?)

Which topic from the class preparation do you like us to focus on more? Any particular question you have about the course but prefer to ask anonymously and not visible on Piazza? Any particular topic in class you definitely like to have covered?

Your answer


Submit

Clear form

Other Thoughts

- **Active participation** is important in this class. If you spot any errors or inconsistencies across slides/web page, typos (even if minor) do let me know, in class, office hours, or via Piazza! We appreciate, and it counts towards class participation.
- If we have online classes (unlikely), please keep your webcams on, so we reproduce our in-person setting as much as possible. One camera off encourages all others to switch off (think externalities).
- **Project topics**: do look also through the preliminary class calendar
 - Individual project (except in rare circumstances with good justification)
 - But you should work together on everything else in the class!

Questions

1. Class is 2:50-4:30. Should we have a 5min break? 
2. Comments on sequencing of the topics?
3. Questions on project topics?
4. What would make it easier for you to participate in class?
5. Any other "best practice" from other classes you recommend?

An end-to-end motivation for basic concepts of information theory

Following numeric example is based on Example 5.1.1 from
[Cover,Thomas'06] Elements of Information Theory. <https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X>
Visualizations are based on Christopher Olah's awesome blog post:
<https://colah.github.io/posts/2015-09-Visual-Information/>

Compressing messages

- Assume Alice communicates with Bob about 4 symbols:
(alternatively, consider using the symbols **ACGT** 😊)

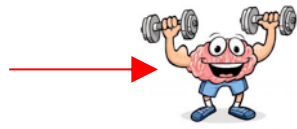
- Alice sends "messages" = sequences of symbols:

- They only communicate in binary. All messages are strings of 0, 1:

- What is a "reasonable" way to encode the symbols?



*Our friend here shows us this
is an active learning exercise*



A B C D

A A C B C D A B B

000010011011000101



Compressing messages

- Assume Alice communicates with Bob about 4 symbols:

A B C D

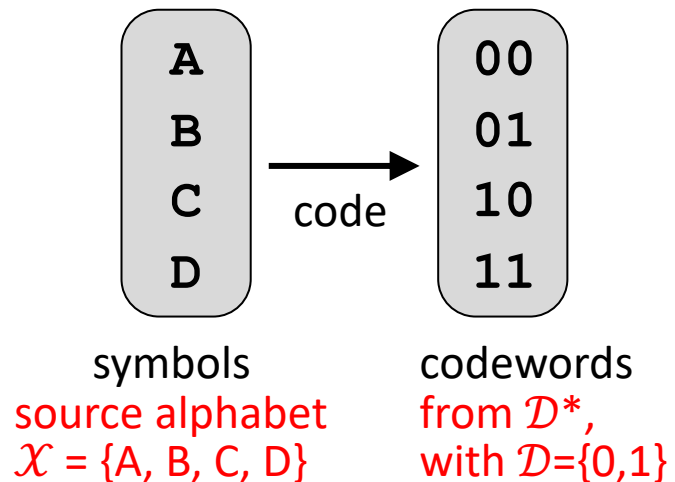
- Alice sends "messages" which are "words" of symbols:

A A C B C D A B B

- They only communicate in binary. All messages are strings of 0, 1:

000010011011000101

- A "reasonable" way to encode the symbols:



Can you decode following message:



0 0 0 1 1 0 1 1

Compressing messages

- Assume Alice communicates with Bob about 4 symbols:

A B C D

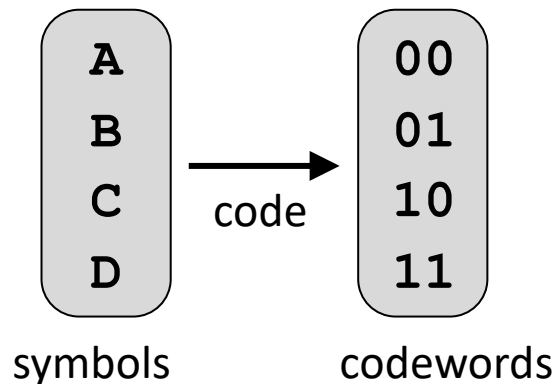
- Alice sends "messages" which are "words" of symbols:

A A C B C D A B B

- They only communicate in binary. All messages are strings of 0, 1:

000010011011000101

- A "reasonable" way to encode the symbols:



Example transmission:

0 0 0 1 1 0 1 1

encoded string

00 01 10 11

codewords

A B C D

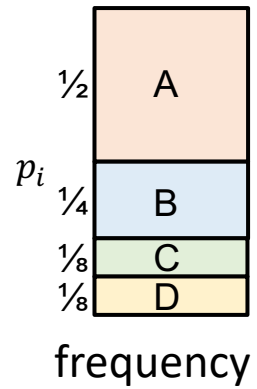
decoded symbols

This is the best you can do for a uniform distribution.



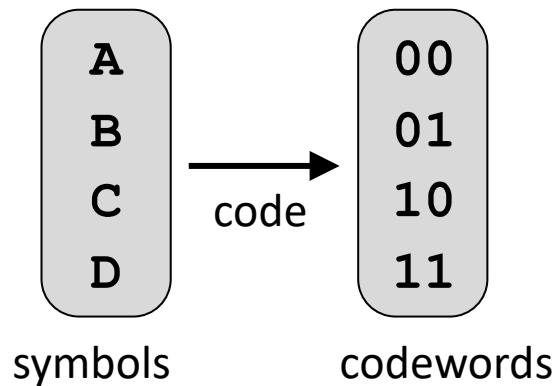
Compressing messages

- Assume we have the following symbol frequency:



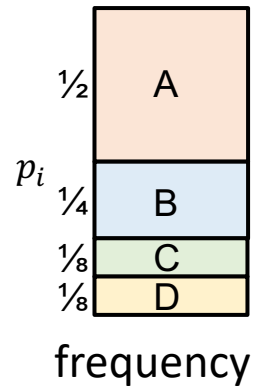
- A "reasonable" way to encode the symbols:

What is our expected message length per symbol?

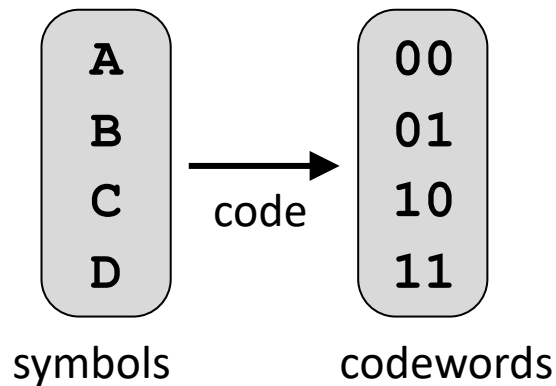


Compressing messages

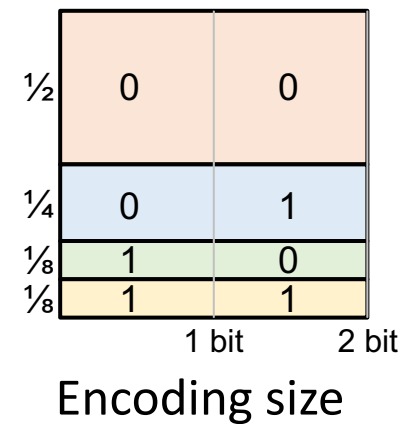
- Assume we have the following symbol frequency:



- A "reasonable" way to encode the symbols:



- Our expected message length per symbol:

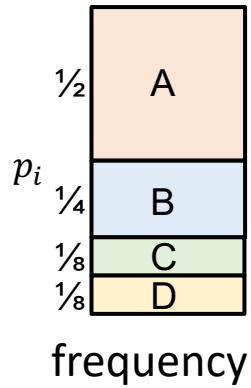


2 bits!



Compressing messages

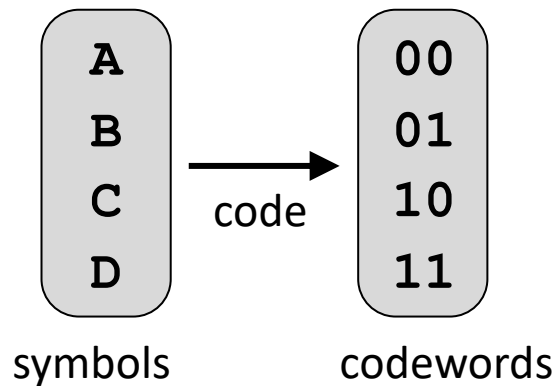
- Assume we have the following symbol frequency:



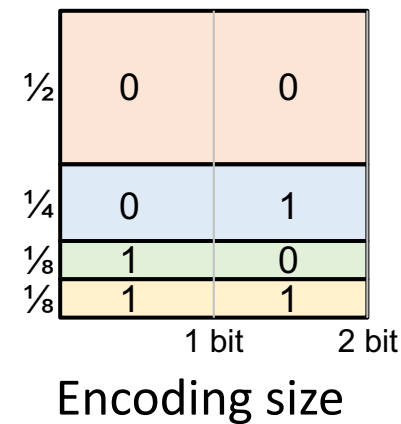
What is our information (expected surprise) we get after seeing each symbol?



- A "reasonable" way to encode the symbols:



- Our expected message length per symbol:

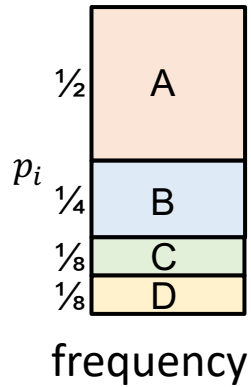


2 bits!



Compressing messages

- Assume we have the following symbol frequency:

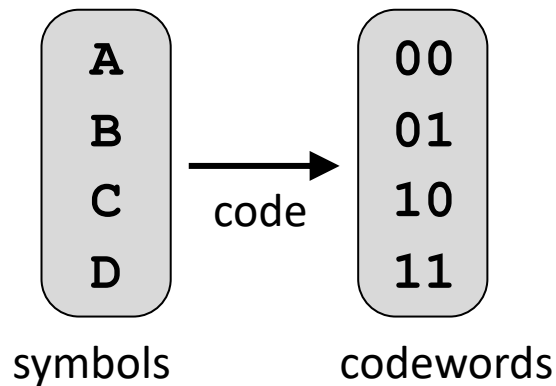


What is our information (expected surprise) we get after seeing each symbol?

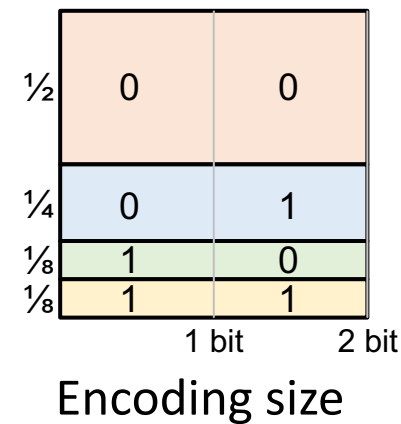
We will write $\log_2(x) = \lg(x)$

Entropy $H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i)$?

- A "reasonable" way to encode the symbols:



Our expected message length per symbol:

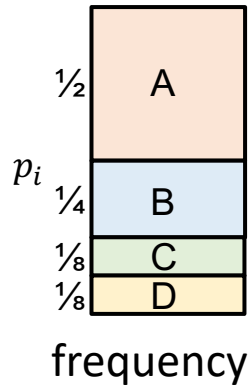


2 bits!



Compressing messages

- Assume we have the following symbol frequency:



What is our information (expected surprise) we get after seeing each symbol?

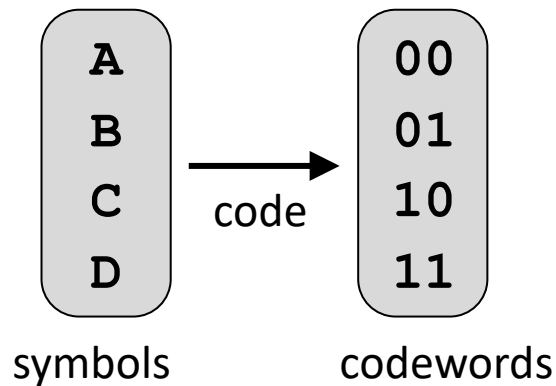
$$\begin{aligned} \lg\left(\frac{1}{2}\right) &= -1 \\ \lg\left(\frac{1}{4}\right) &= -2 \\ \lg\left(\frac{1}{8}\right) &= -3 \end{aligned}$$

$$-\left(\frac{1}{2} \cdot -1 + \frac{1}{4} \cdot -2 + \frac{1}{8} \cdot -3 + \frac{1}{8} \cdot -3\right)$$

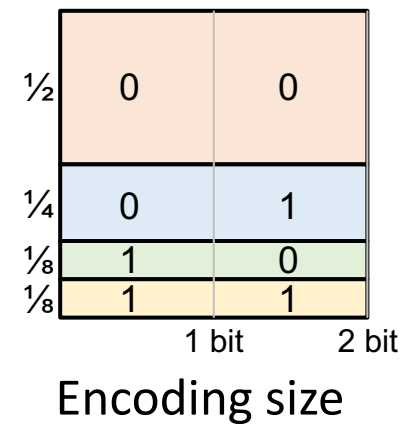
Entropy $H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75$ bits!

Can we match that "bound" w/ some encoding? **?**

- A "reasonable" way to encode the symbols:



Our expected message length per symbol:



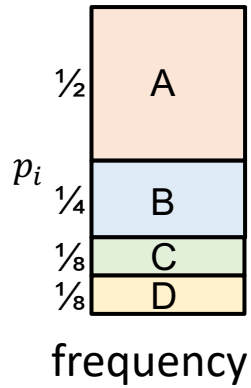
2 bits!

Compressing messages via variable length codes

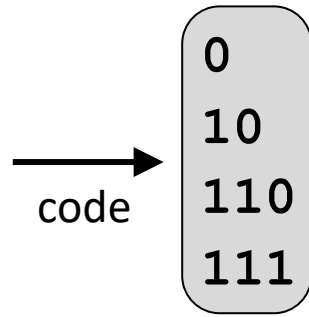


- Assume we have the following symbol frequency:

How can we decode that



symbols



codewords

Intuition: more frequent stuff should use less space!

$$\lg\left(\frac{1}{2}\right) = -1$$

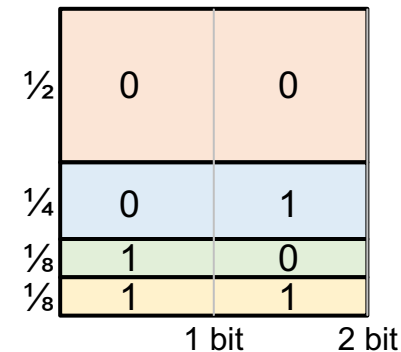
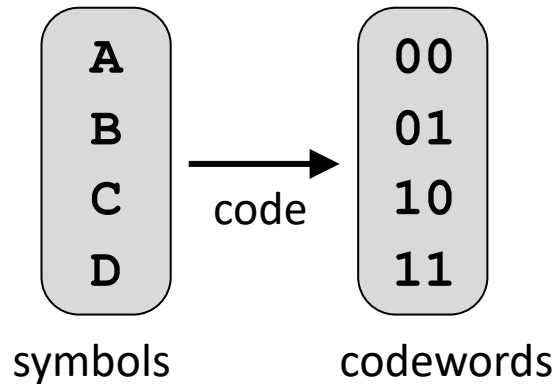
$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

- A "reasonable" way to encode the symbols:

Our expected message length per symbol:



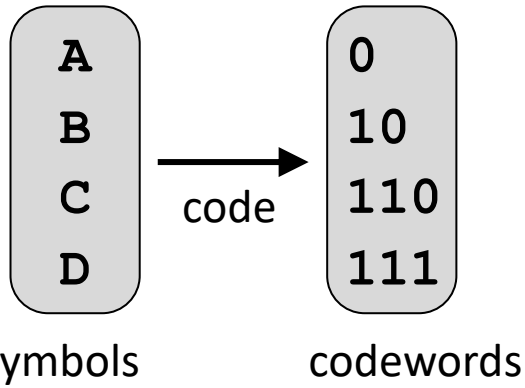
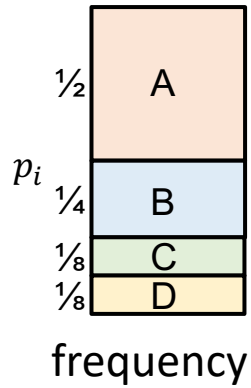
2 bits!

Encoding size

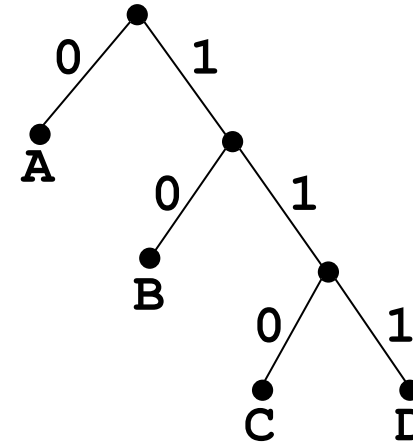
Compressing messages via variable length codes



- Assume we have the following symbol frequency:



Prefix code (shown via binary prefix trees)

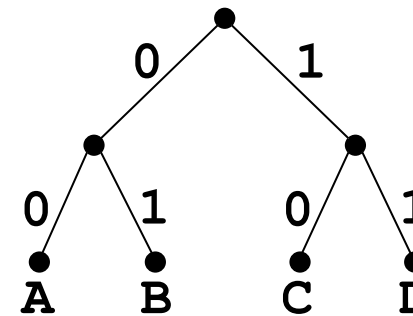
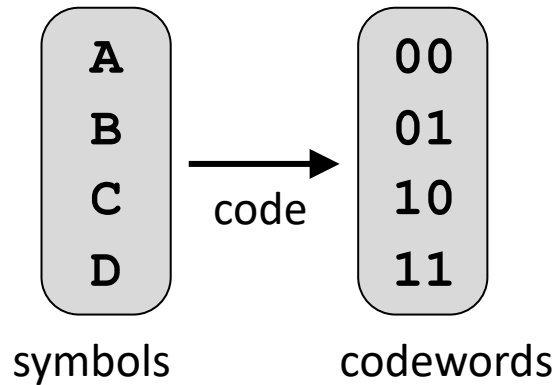


$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

- A "reasonable" way to encode the symbols:

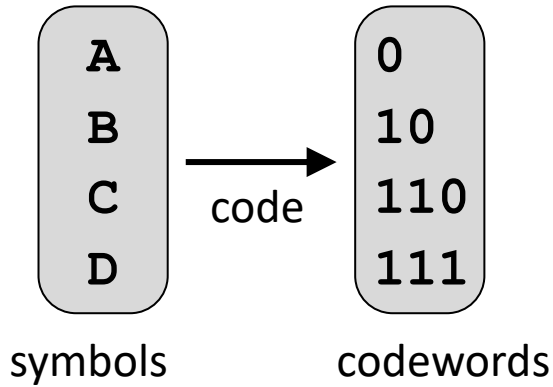
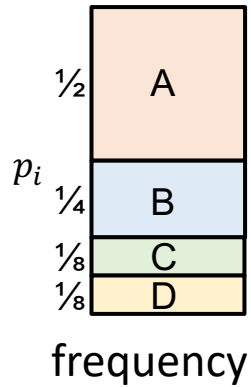


These two are also called a prefix code (or instantaneous or self-punctuating code). Notice that no codeword is a prefix of another codeword and a binary prefix tree can be used to uniquely decode a correctly encoded message

Compressing messages via variable length codes



- Assume we have the following symbol frequency:



What is the new expected length? **?**

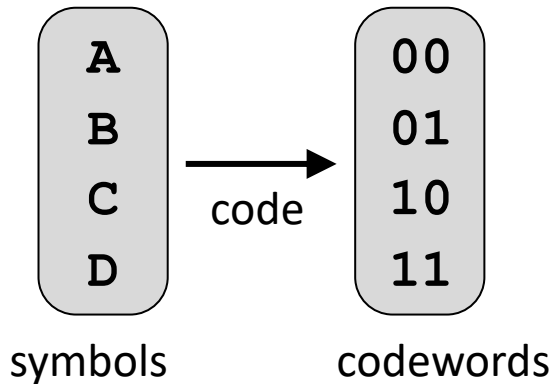
$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

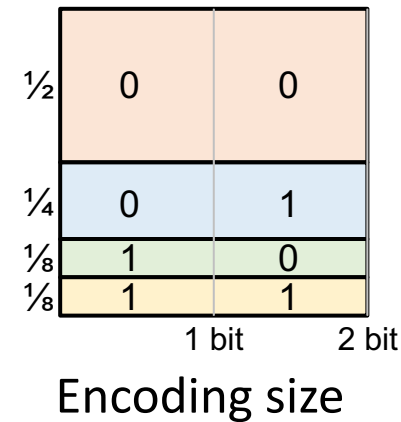
$$\lg\left(\frac{1}{8}\right) = -3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

- A "reasonable" way to encode the symbols:



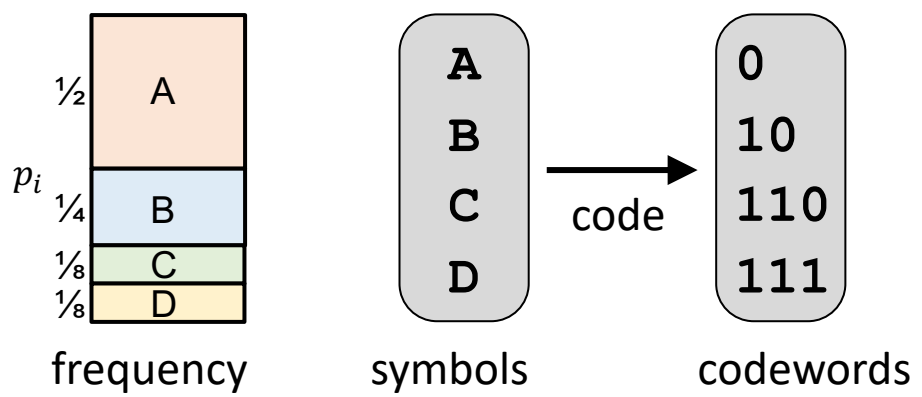
Our expected message length per symbol:



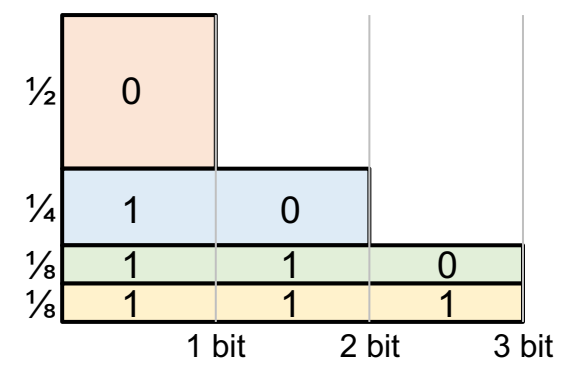
2 bits!

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length:



$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\frac{1}{2} \cdot 1$$

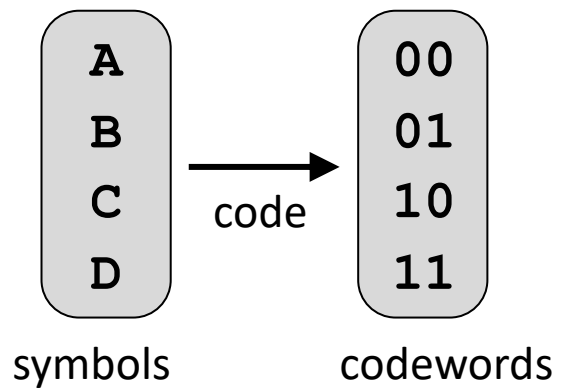
$$\frac{1}{4} \cdot 2 = 1.75 \text{ bits!}$$

$$\frac{1}{8} \cdot 3$$

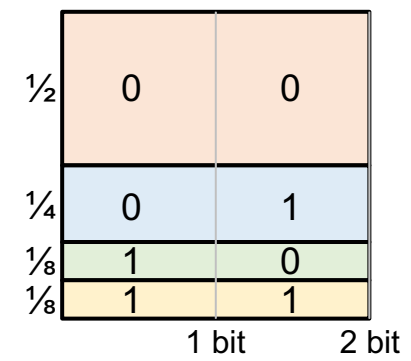
$$\frac{1}{8} \cdot 3$$

Entropy $H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$

- A "reasonable" way to encode the symbols:



Our expected message length per symbol:

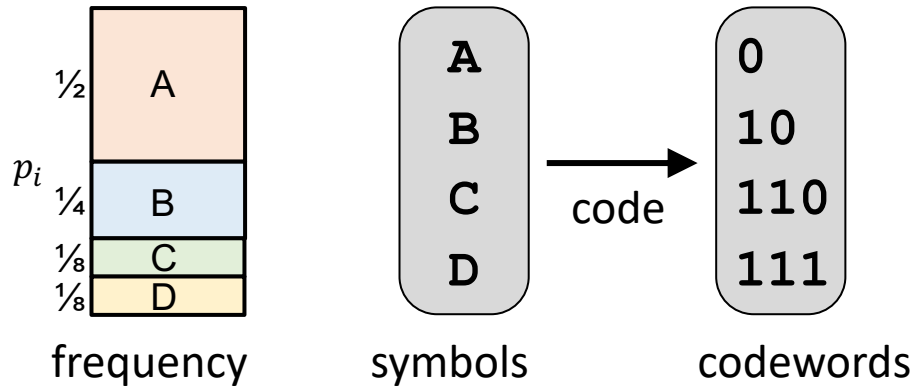


2 bits!

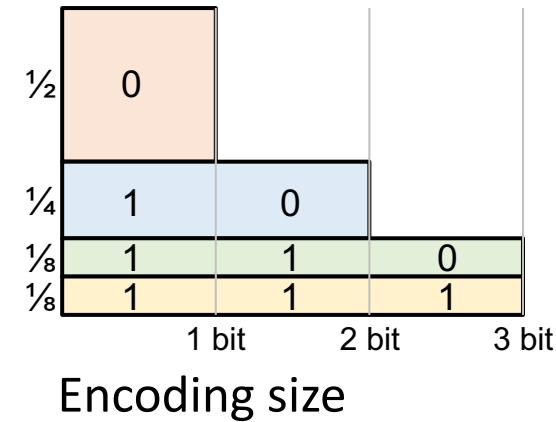
Encoding size

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\frac{1}{2} \cdot 1$$

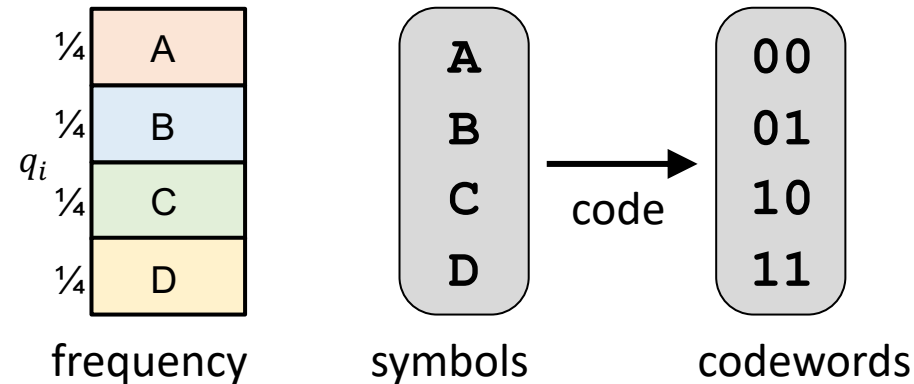
$$\frac{1}{4} \cdot 2 = 1.75 \text{ bits!}$$

$$\frac{1}{8} \cdot 3$$

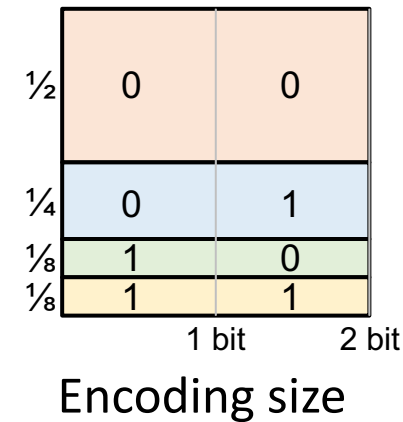
$$\frac{1}{8} \cdot 3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

- Another interpretation: this is our assumed distribution!



Our expected message length per symbol:

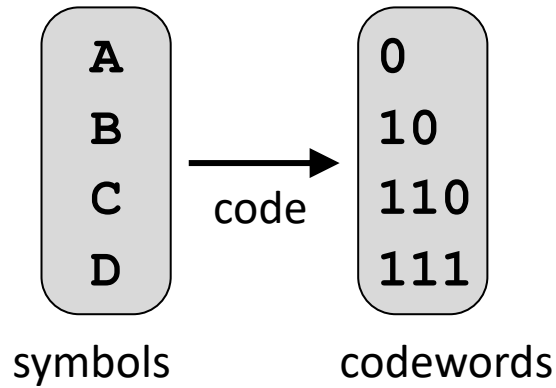
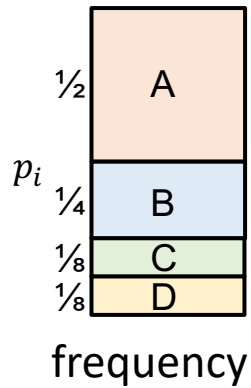


2 bits!

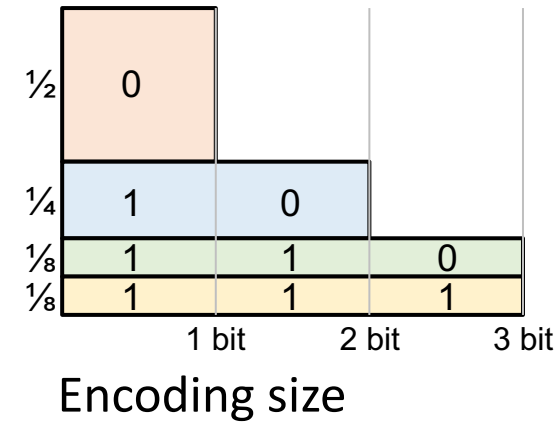


Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\frac{1}{2} \cdot 1$$

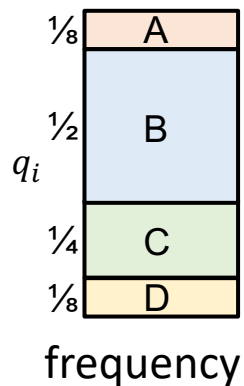
$$\frac{1}{4} \cdot 2 = 1.75 \text{ bits!}$$

$$\frac{1}{8} \cdot 3$$

$$\frac{1}{8} \cdot 3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

- What if we assume following distribution:



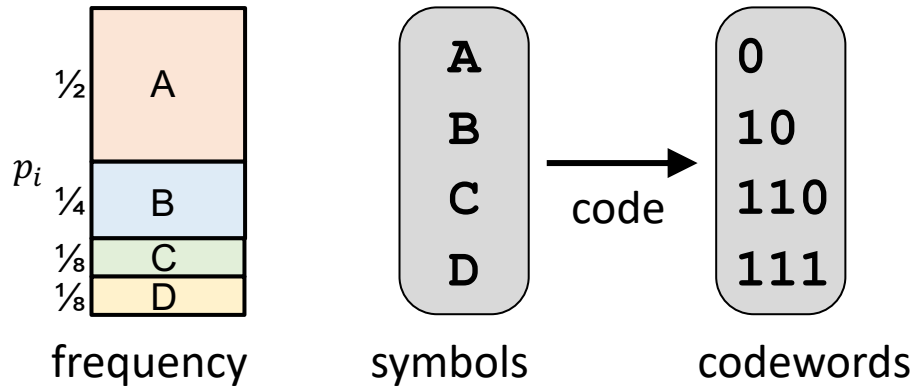
What should be our code if we assumed q as distribution?



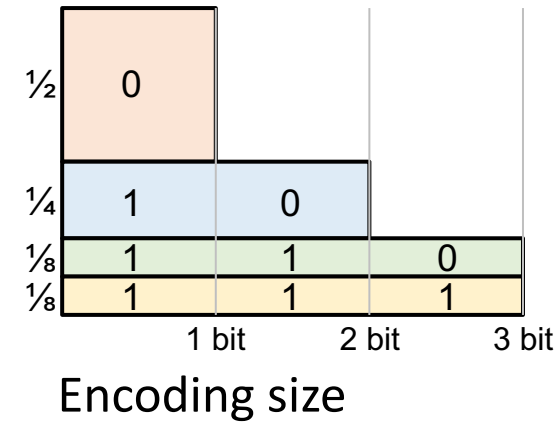


Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\frac{1}{2} \cdot 1$$

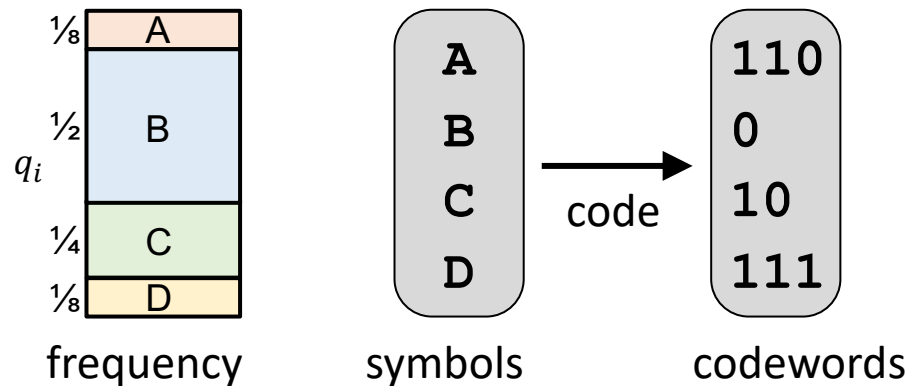
$$\frac{1}{4} \cdot 2 = 1.75 \text{ bits!}$$

$$\frac{1}{8} \cdot 3$$

$$\frac{1}{8} \cdot 3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

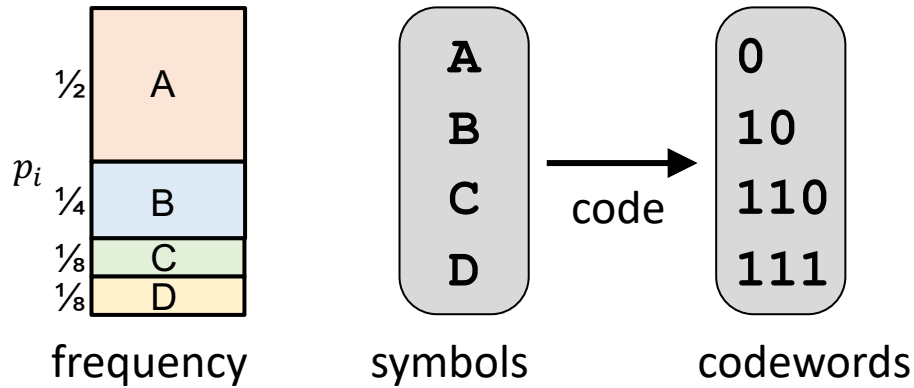
- What if we assume following distribution:



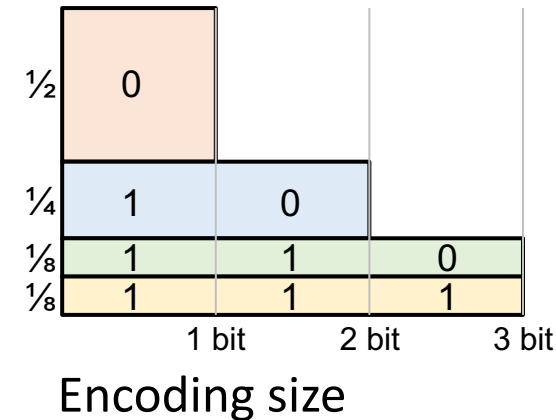
What is our expected message length per symbol if we use that code, but p is the actual distribution ?

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\frac{1}{2} \cdot 1$$

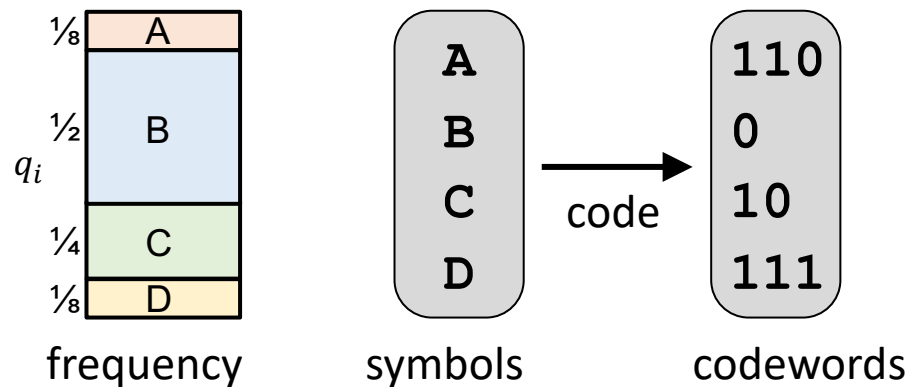
$$\frac{1}{4} \cdot 2 = 1.75 \text{ bits!}$$

$$\frac{1}{8} \cdot 3$$

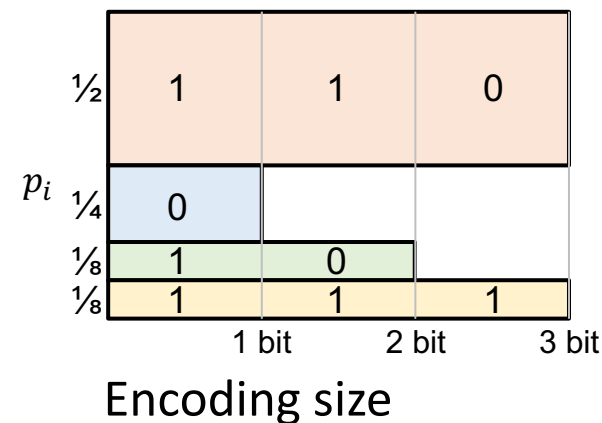
$$\frac{1}{8} \cdot 3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

- What if we assume following distribution:



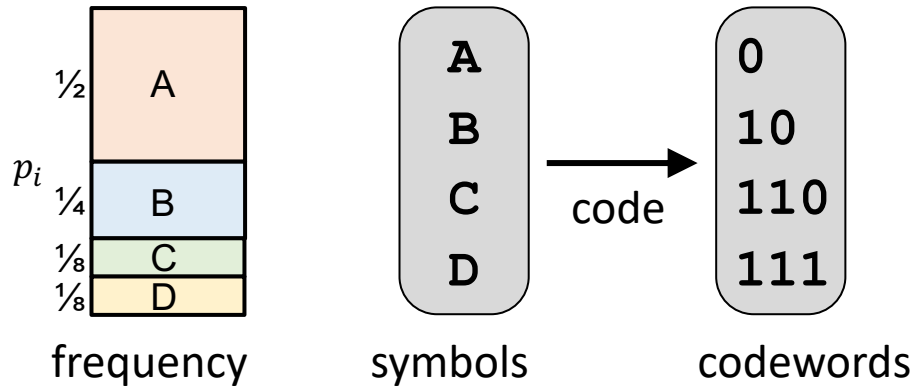
Our new expected message length per symbol:



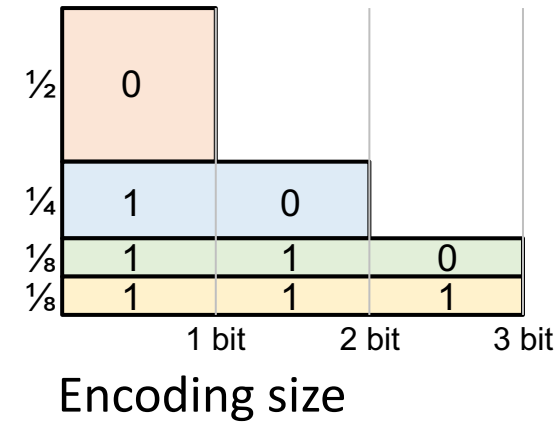


Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\frac{1}{2} \cdot 1$$

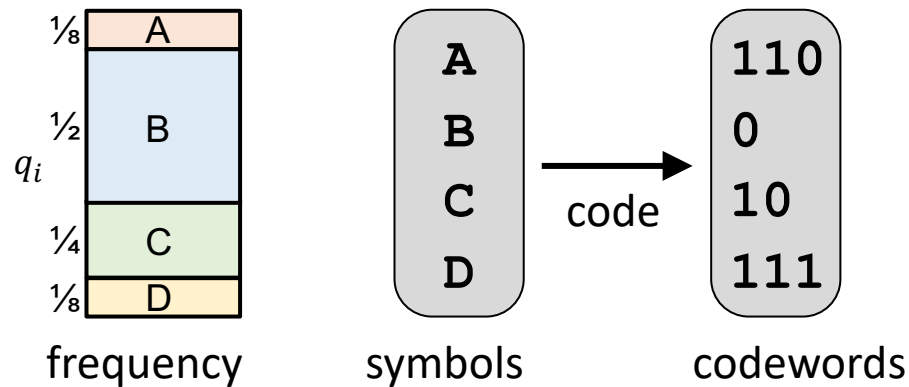
$$\frac{1}{4} \cdot 2 = 1.75 \text{ bits!}$$

$$\frac{1}{8} \cdot 3$$

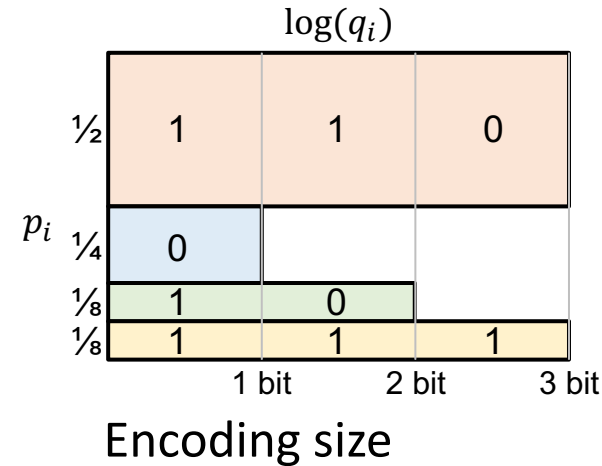
$$\frac{1}{8} \cdot 3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

- What if we assume following distribution:



Our new expected message length per symbol:

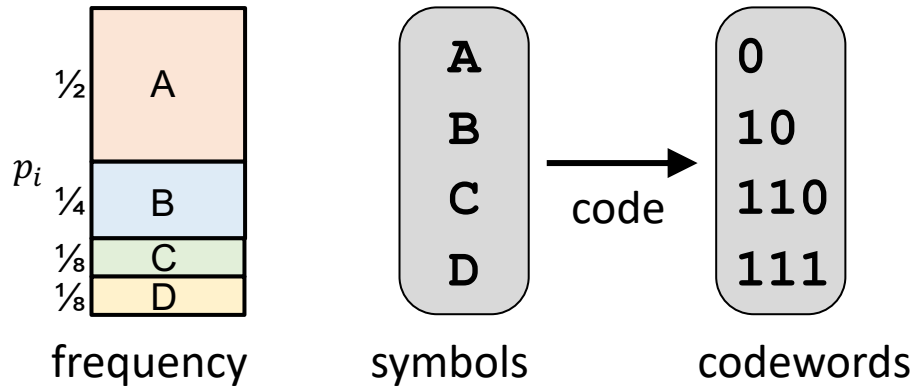


What is the formula we need to evaluate ?

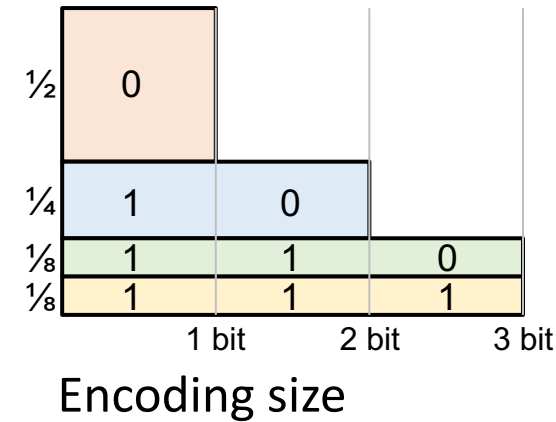
Compressing messages via variable length codes



- Assume we have the following symbol frequency:



New expected length :



$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\frac{1}{2} \cdot 1$$

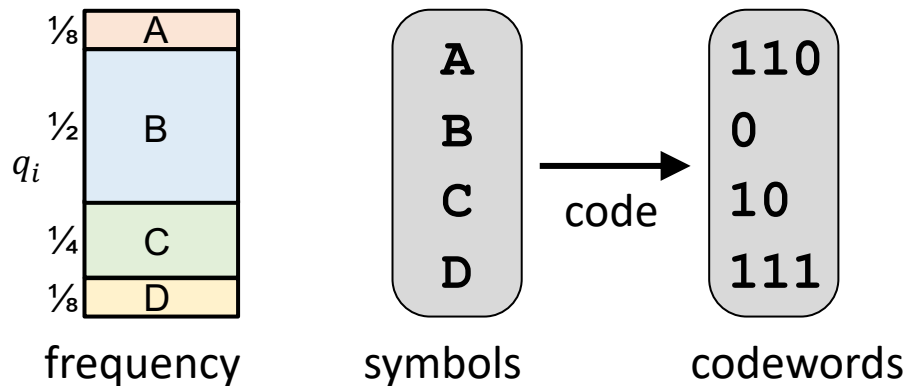
$$\frac{1}{4} \cdot 2 = 1.75 \text{ bits!}$$

$$\frac{1}{8} \cdot 3$$

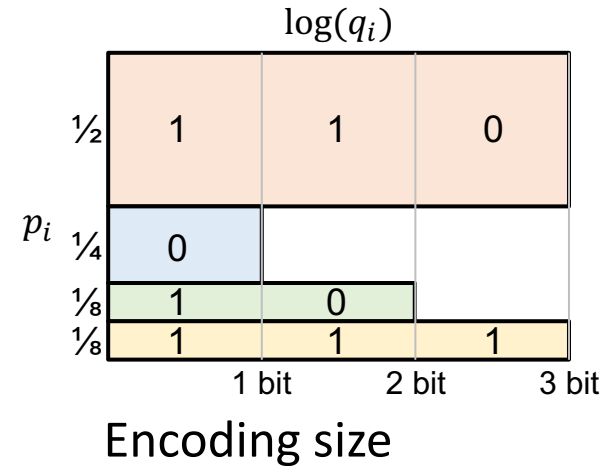
$$\frac{1}{8} \cdot 3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

- What if we assume following distribution:



Our new expected message length per symbol:



$$= 2.375 \text{ bits!}$$

$$-\sum_i p_i \cdot \lg(q_i)$$

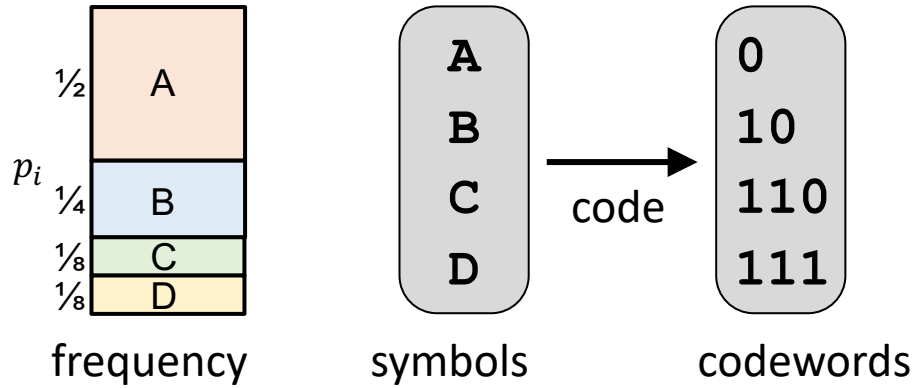
What is this formula called



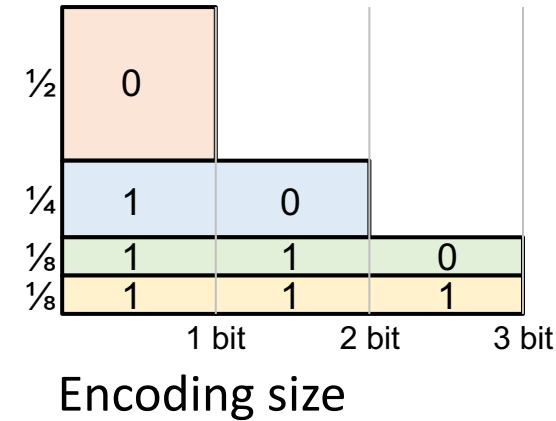
Compressing messages via variable length codes



- Assume we have the following symbol frequency:



New expected length :



$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\frac{1}{2} \cdot 1$$

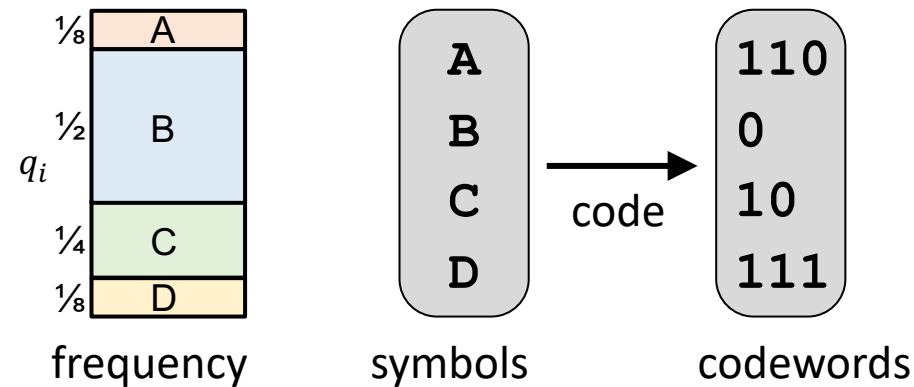
$$\frac{1}{4} \cdot 2 = 1.75 \text{ bits!}$$

$$\frac{1}{8} \cdot 3$$

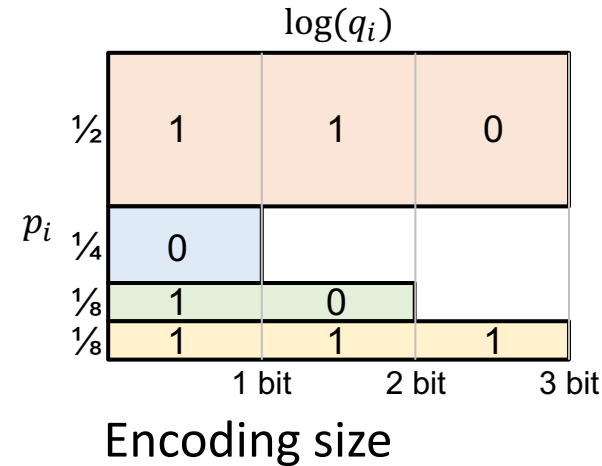
$$\frac{1}{8} \cdot 3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

- What if we assume following distribution:



Our new expected message length per symbol:



$$= 2.375 \text{ bits!}$$

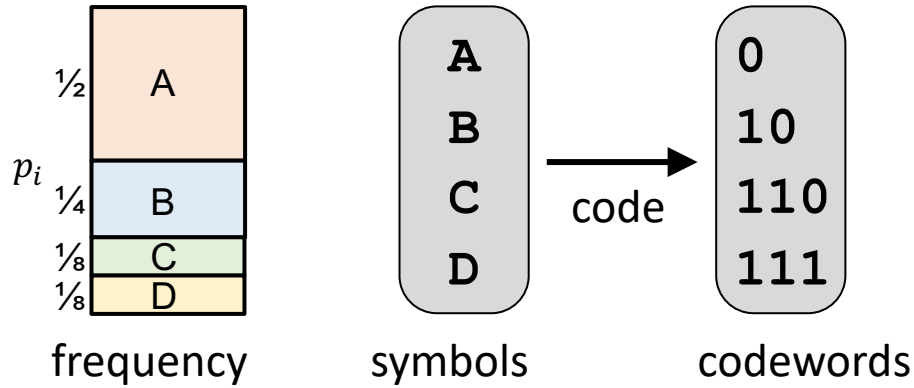
$$-\sum_i p_i \cdot \lg(q_i)$$

Cross entropy $H(p||q)$ ☺

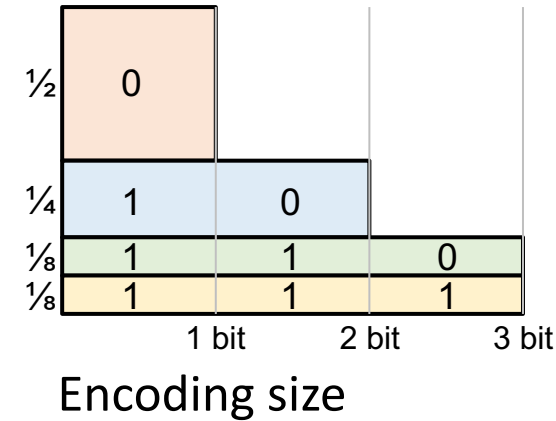
Which distribution q minimizes $H(p||q)$?

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



$$\lg\left(\frac{1}{2}\right) = -1$$

$$\lg\left(\frac{1}{4}\right) = -2$$

$$\lg\left(\frac{1}{8}\right) = -3$$

$$\frac{1}{2} \cdot 1$$

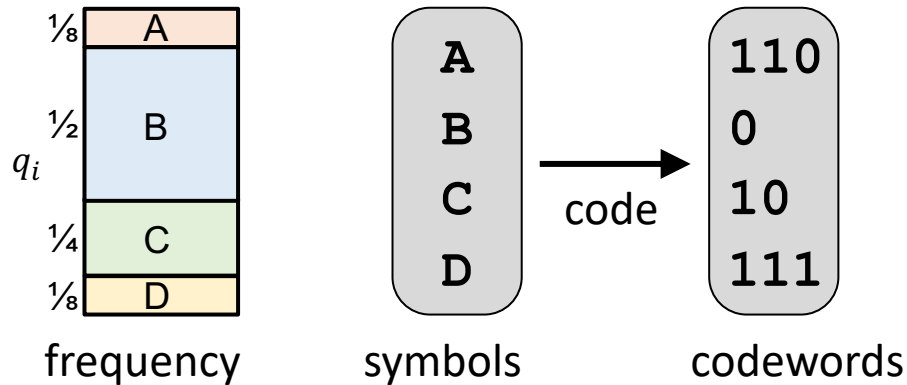
$$\frac{1}{4} \cdot 2 = 1.75 \text{ bits!}$$

$$\frac{1}{8} \cdot 3$$

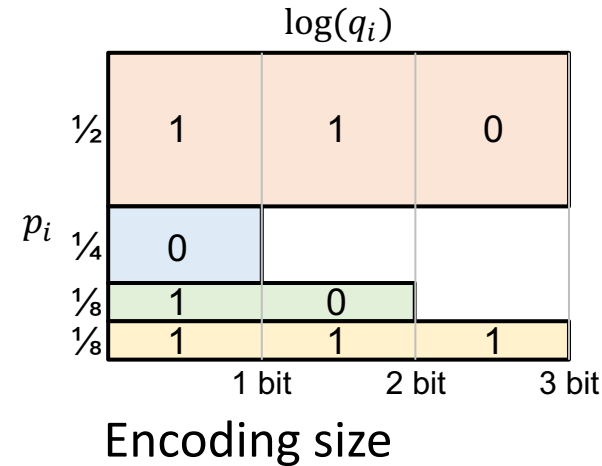
$$\frac{1}{8} \cdot 3$$

$$\text{Entropy } H(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i) = 1.75 \text{ bits!}$$

- What if we assume following distribution:



Our new expected message length per symbol:



$$= 2.375 \text{ bits!}$$

$$-\sum_i p_i \cdot \lg(q_i)$$

Cross entropy $H(p||q)$ ☺

$$q = p \text{ minimizes } H(p||q)$$