Updated 9/11/2025

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? How did you learn about this class? Why do you consider taking it?
2. What do you hope to get out of this course ☺ What is the topic from the course page (or information theory, in general) that you are most interested in? What could be your project?
3. What is your biggest fear for this course ☹

Wolfgang Gatterbauer

cs7840 Foundations and Applications of Information Theory (fa25)

https://northeastern-datalab.github.io/cs7840/fa25/

9/8/2025

A few examples for what this class is all about

Some intuition

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

- I am thinking of a number from 0 to 255.
  - You can ask me binary questions ("Is it < 30?")
  - How many questions do you have to maximally ask me to get the number?

?

Some intuition

- I am thinking of a number from 0 to 255 (=256 or 2^8 choices).
    - You can ask me binary questions ("Is it < 30?")
    - How many questions do you have to maximally ask me to get the number?
        - You can always halve the range each time. Assume I choose 3:

?

Some intuition

- I am thinking of a number from 0 to 255 (=256 or 2^8 choices).
  - You can ask me binary questions ("Is it < 30?")
  - How many questions do you have to maximally ask me to get the number?
    - You can always halve the range each time. Assume I choose 3:
    - Is it < 128? yes
    - Is it < 64? yes
    - Is it < 32? yes
    - Is it < 16? yes
    - Is it < 8? yes
    - Is it < 4? yes
    - Is it < 2? no
    - Is it < 3? no -> must be 3

Can we represent this sequence of questions in some other (number) format ?

Some intuition

- I am thinking of a number from 0 to 255 (=256 or 2^8 choices).
  - You can ask me binary questions ("Is it < 30?")
  - How many questions do you have to maximally ask me to get the number?
    - You can always halve the range each time. Assume I choose 3:
    - Is it < 128? yes
    - Is it < 64? yes
    - Is it < 32? yes
    - Is it < 16? yes
    - Is it < 8? yes
    - Is it < 4? yes
    - Is it < 2? no
    - Is it < 3? no -> must be 3
  - 3 = 00000011 in binary (the answer to each question corresponds to 1 bit)

Can we represent this sequence of questions in some other (number) format?

Some intuition

- Earlier: I was thinking of a number from 0 to 255 (=2^8 choices).
- Now: I am thinking of a number from 0 to 511 (=2^9 choices).
  - How many questions do you have to ask now me to get the number?

?

Some intuition

- Earlier: I was thinking of a number from 0 to 255 (=2^8 choices).
- Now: I am thinking of a number from 0 to 511 (=2^9 choices).
  - How many questions do you have to ask now me to get the number?
  - just one more than before!
    - 3 =   00000011 in binary with 8 bits
    - 3 = 000000011 in binary with 9 bits

- We have learned: The information content (the level of uncertainty) does not grow proportional to the number of choices!
- It grows with the logarithm!

Some intuition about logarithms

- Earlier: I was thinking of a number from 0 to 255 (=2^8 choices).
- Now: I am thinking of 2 numbers, each from 0 to 15 (=2^4 choices).
  - Do you have to ask me more or fewer questions to get both numbers?

?

Some intuition about logarithms

- Earlier: I was thinking of a number from 0 to 255 (=2^8 choices).
- Now: I am thinking of 2 numbers, each from 0 to 15 (=2^4 choices).
  - Do you have to ask me more or fewer questions to get both numbers?
  - The same! 16 · 16 choices = 256 choices ($2^8 = 2^4 \cdot 2^4$)
    - 3 = 0011 in binary with 4 bits
    - (3, 3) = (0011, 0011) in binary with 4 bits each
    - 51 = 00110011 in binary with 8 bits
- How many questions do you have to ask me as function $f(N)$?
  - where N = number of total choices (256 = 16 · 16)

Some intuition about logarithms

- Earlier: I was thinking of a number from 0 to 255 (=2^8 choices).
- Now: I am thinking of 2 numbers, each from 0 to 15 (=2^4 choices).
  - Do you have to ask me more or fewer questions to get both numbers?
  - The same! 16 · 16 choices = 256 choices ($2^8 = 2^4 \cdot 2^4$)
    - 3 = 0011 in binary with 4 bits
    - (3, 3) = (0011, 0011) in binary with 4 bits each
    - 51 = 00110011 in binary with 8 bits

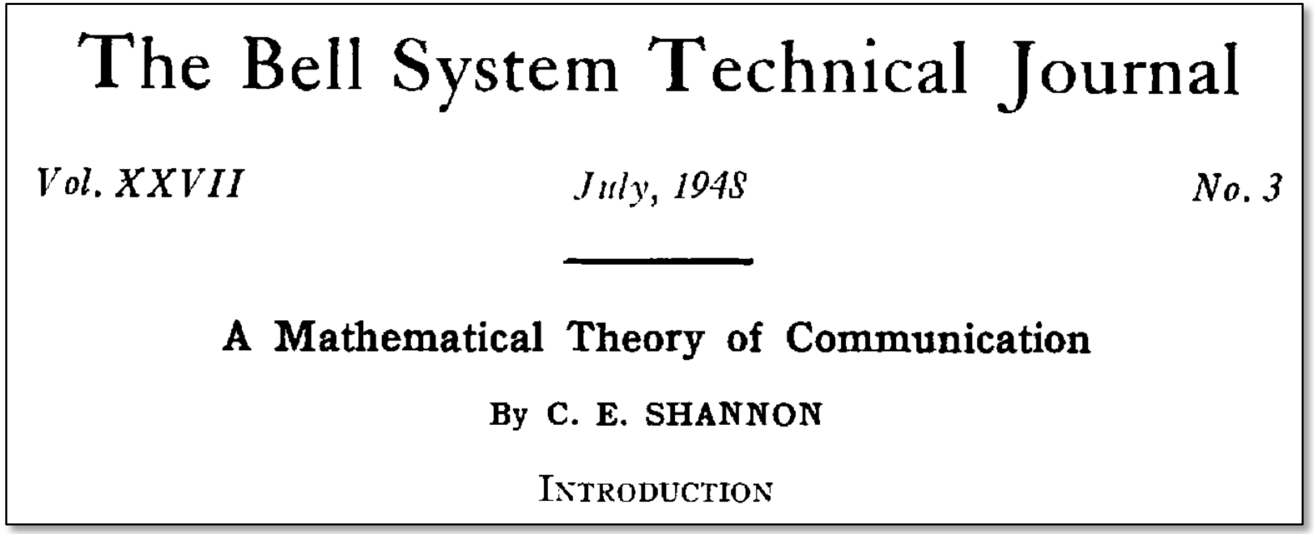
2^8=256 ⇔ $\log_2 256 = 8$

- You have to ask me $\boxed{\log_2 N}$ questions
  - where N = number of total choices (256 = 16 · 16)

$$\log\left(\frac{1}{x}\right) = -\log(x)$$

- Tip for later: $\boxed{\log_2 N = -\log_2(1/N) = -\log_2(P)}$
  - where P = probability of picking any one of the equally likely choices

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

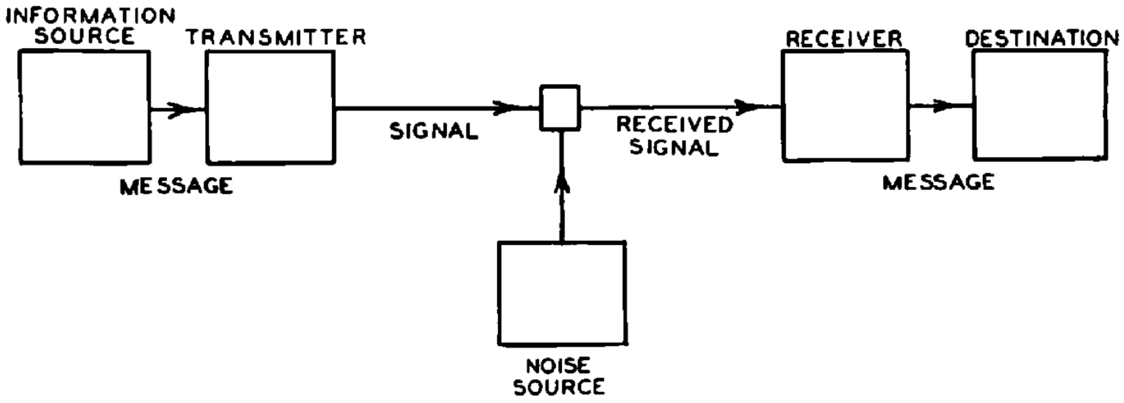
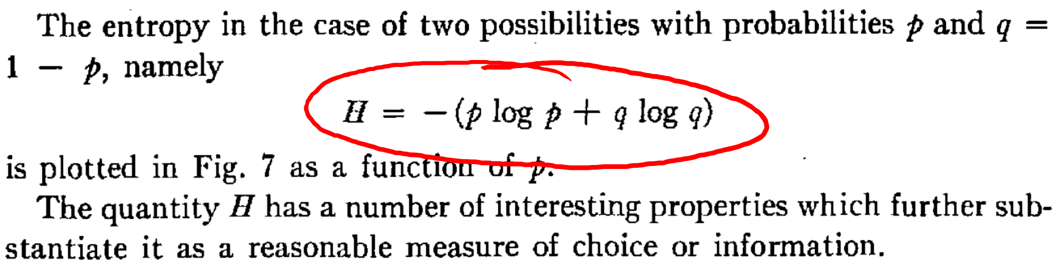


Fig. 1—Schematic diagram of a general communication system.



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

$$\boxed{H = -(p \log p + q \log q)}$$

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.

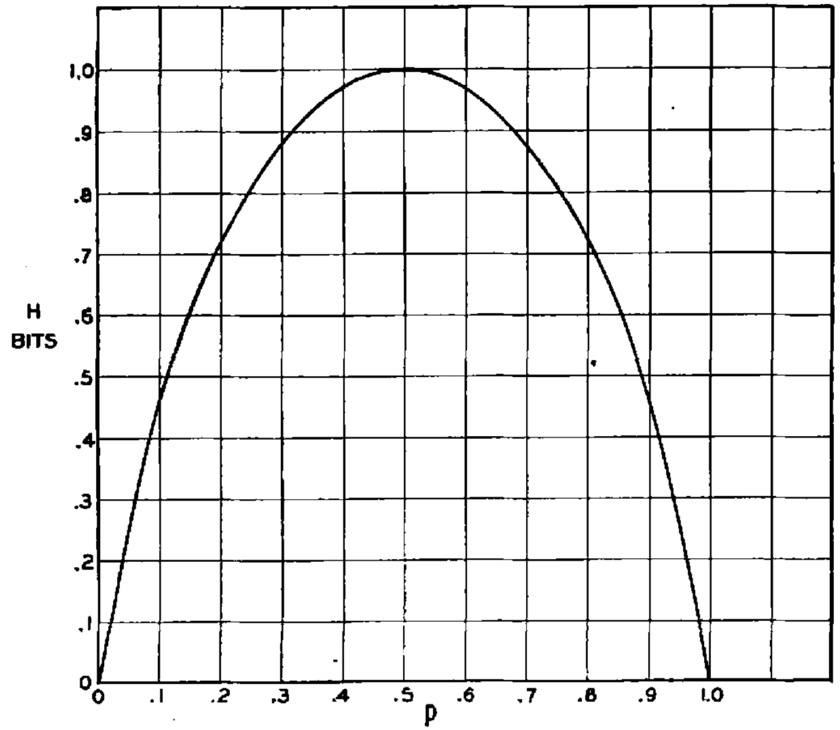


Fig. 7—Entropy in the case of two possibilities with probabilities p and $(1 - p)$.

From: Shannon. A Mathematical Theory of Communication, The Bell System Technical Journal, 1948. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

"Communication": Randomness, Compressibility, Predictability

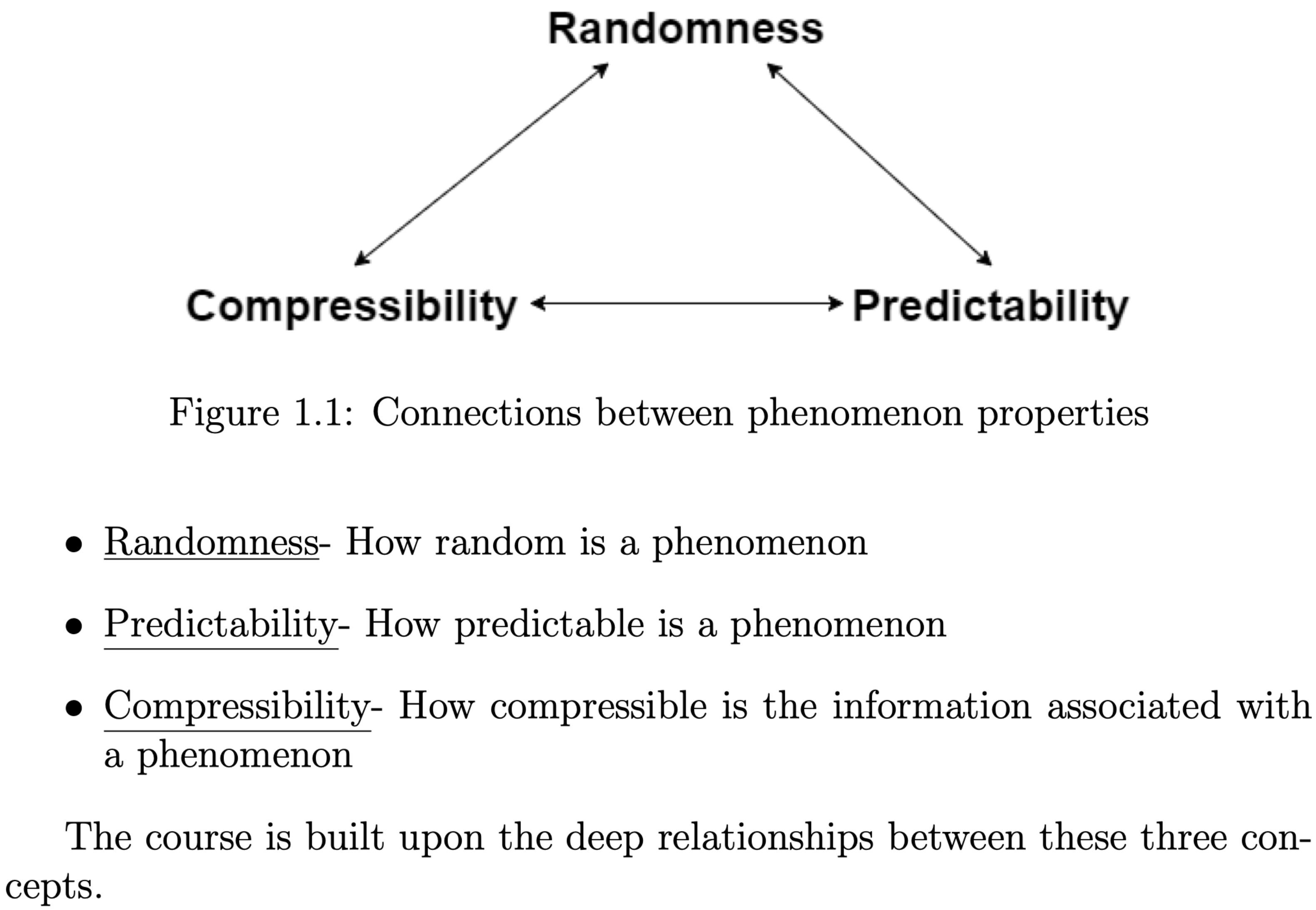

Randomness
Compressibility
Predictability

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.

Source: Scribes from Aslam's cs6800 "Applications of Information Theory to Computer Science" fall 2015: https://www.khoury.northeastern.edu/home/jaa/CS6800.15F/index.html

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

?

Source: https://en.wikipedia.org/wiki/Entropy_(information_theory)
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

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

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

How can you establish something like that? ?

Source: https://en.wikipedia.org/wiki/Entropy_(information_theory)

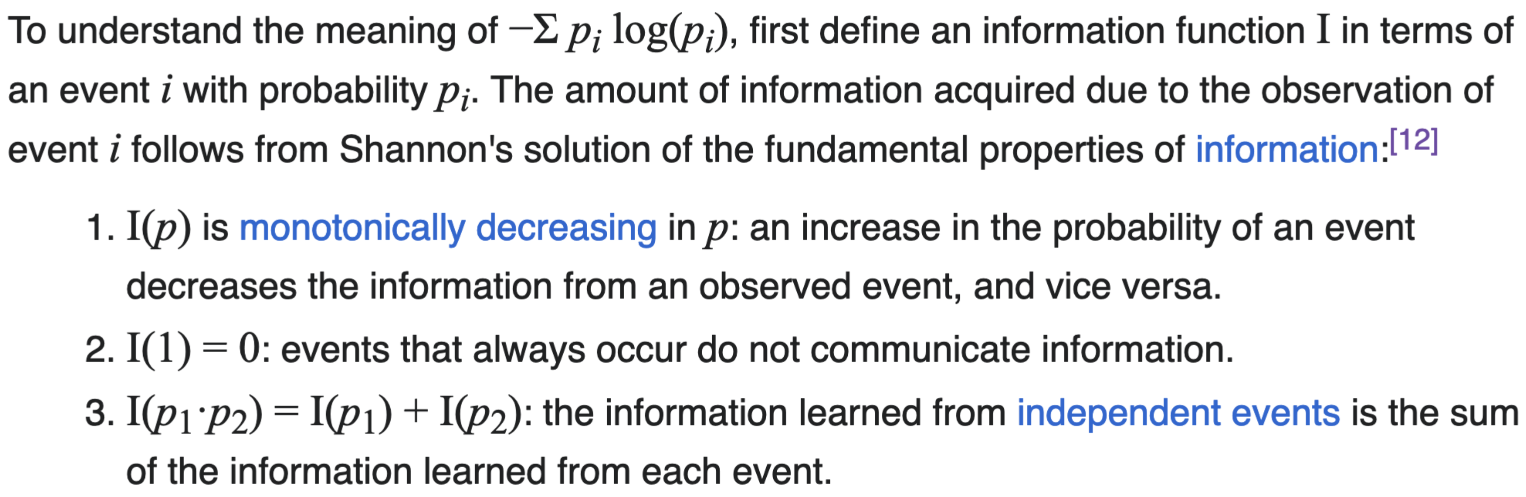
Why $\sum_i p_i \log(p_i)$ to measure the "amount" of uncertainty?

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

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

How can you establish something like that?

Via an "axiomatic derivation":



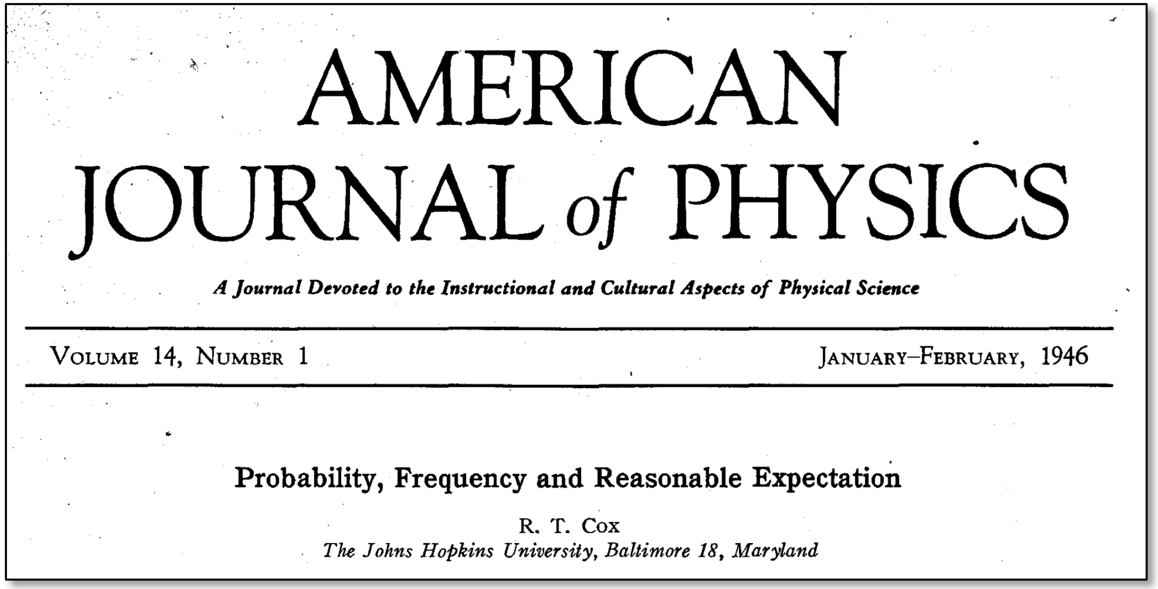
To understand the meaning of $-\Sigma\, p_i \log(p_i)$, first define an information function I in terms of an event i with probability p_i. The amount of information acquired due to the observation of event i follows from Shannon's solution of the fundamental properties of information:[12]

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

Source: https://en.wikipedia.org/wiki/Entropy_(information_theory)

Cox's axiomatic derivation of probability theory [1946]



AMERICAN JOURNAL *of* PHYSICS

A Journal Devoted to the Instructional and Cultural Aspects of Physical Science

VOLUME 14, NUMBER 1 JANUARY–FEBRUARY, 1946

Probability, Frequency and Reasonable Expectation

R. T. COX

The Johns Hopkins University, Baltimore 18, Maryland

"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 - 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

From: Richard T. Cox. Probability, frequency and reasonable expectation, 1946. https://doi.org/10.1119/1.1990764, http://www.stats.org.uk/cox-theorems/

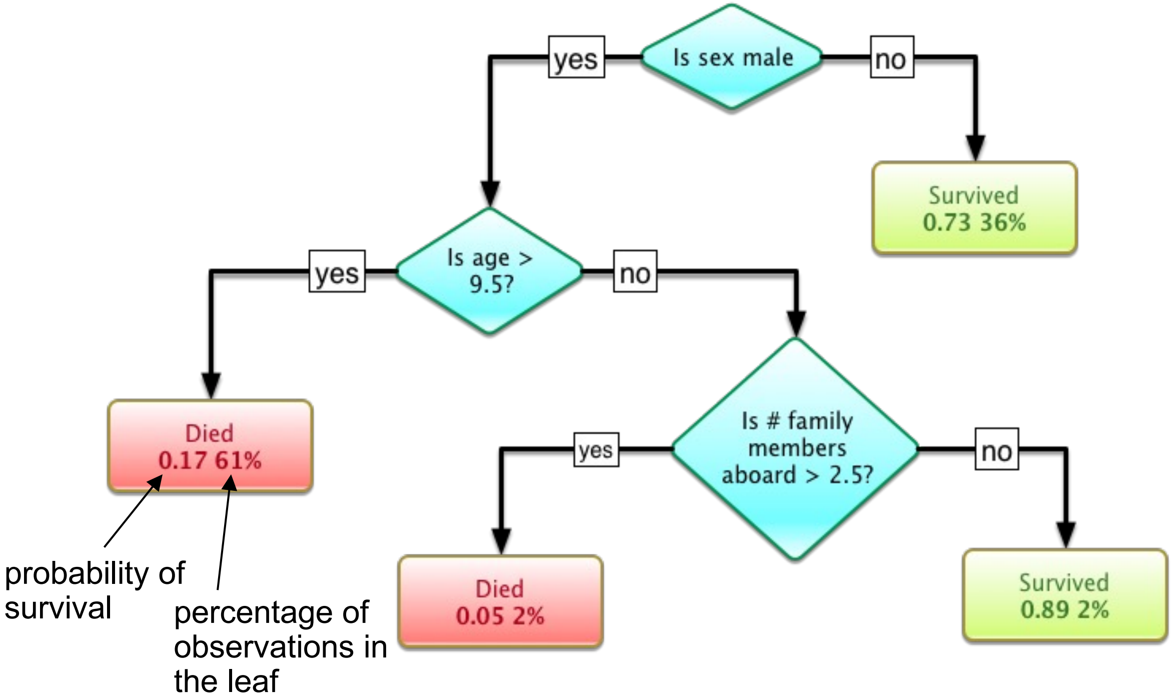
An application in ML

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 ML: 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")?



probability of survival

percentage of observations in the leaf

How can an algorithm be possible "guided" ?

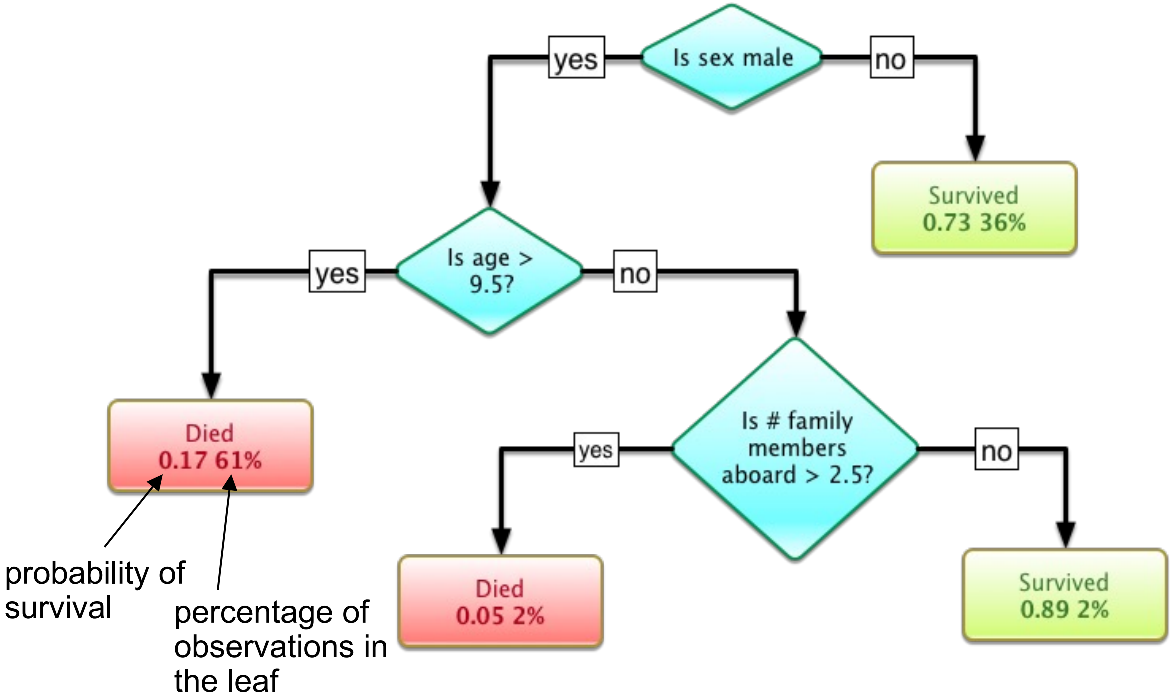
Your chances of survival were good if you were (i) a female or (ii) a male ≤ 9.5 years old with < than 3 siblings.

Sources: https://en.wikipedia.org/wiki/Decision_tree_learning , https://commons.wikimedia.org/wiki/File:Titanic_Survival_Decison_Tree_SVG.png

An application in ML: learning decision trees

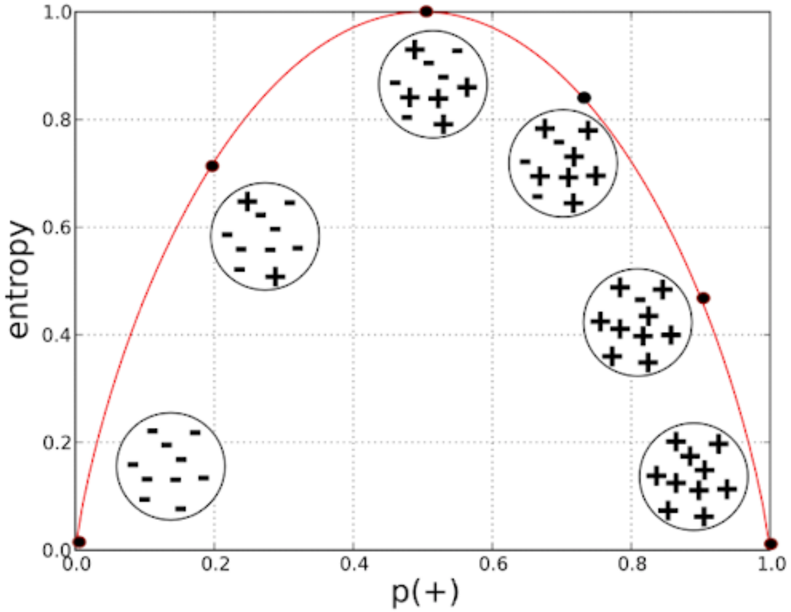
Try to "best" explain the chances of survival for passengers on the Titanic based on various attributes like gender, age, number of spouses or siblings aboard ("sibsp")?

"learning" = "compressing"

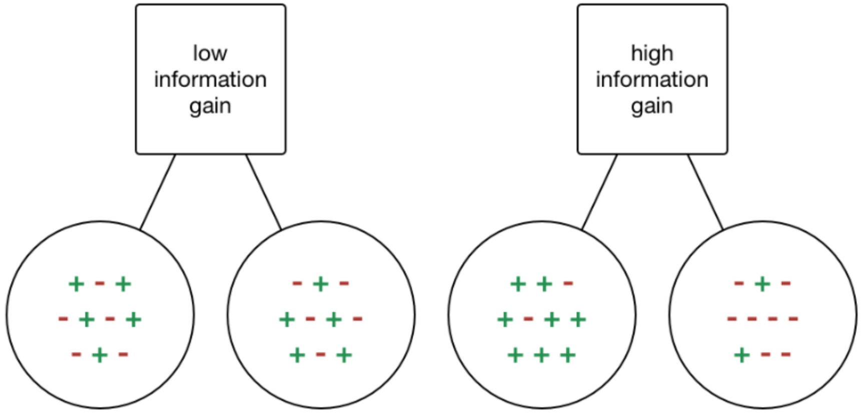


probability of survival

percentage of observations in the leaf



(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.

Sources: https://en.wikipedia.org/wiki/Decision_tree_learning , https://commons.wikimedia.org/wiki/File:Titanic_Survival_Decison_Tree_SVG.png , https://ai.plainenglish.io/simplified-machine-learning-f5ca4e177bac, https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

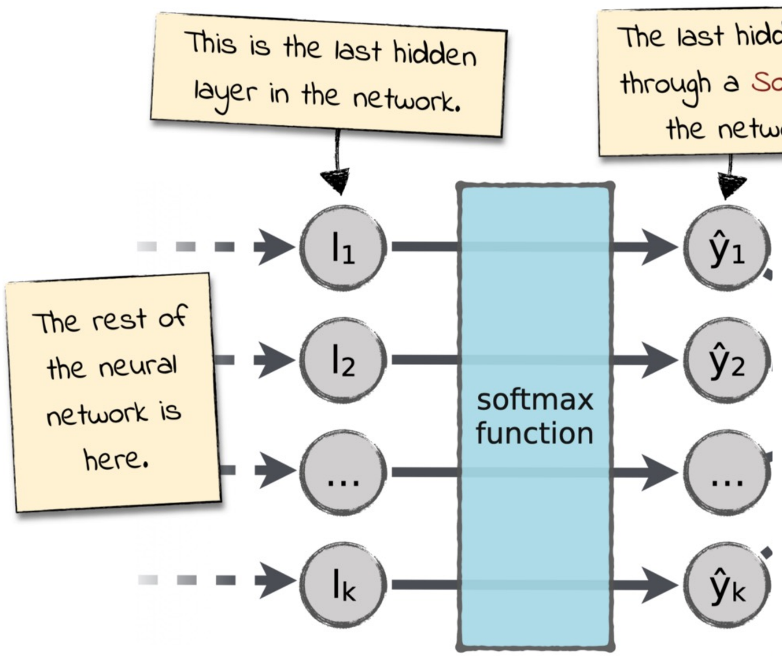
"An application" in ML: DNN

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

https://levelup.gitconnected.com/killer-combo-softmax-and-cross-entropy-5907442f60ba

"An application" in ML: DNN

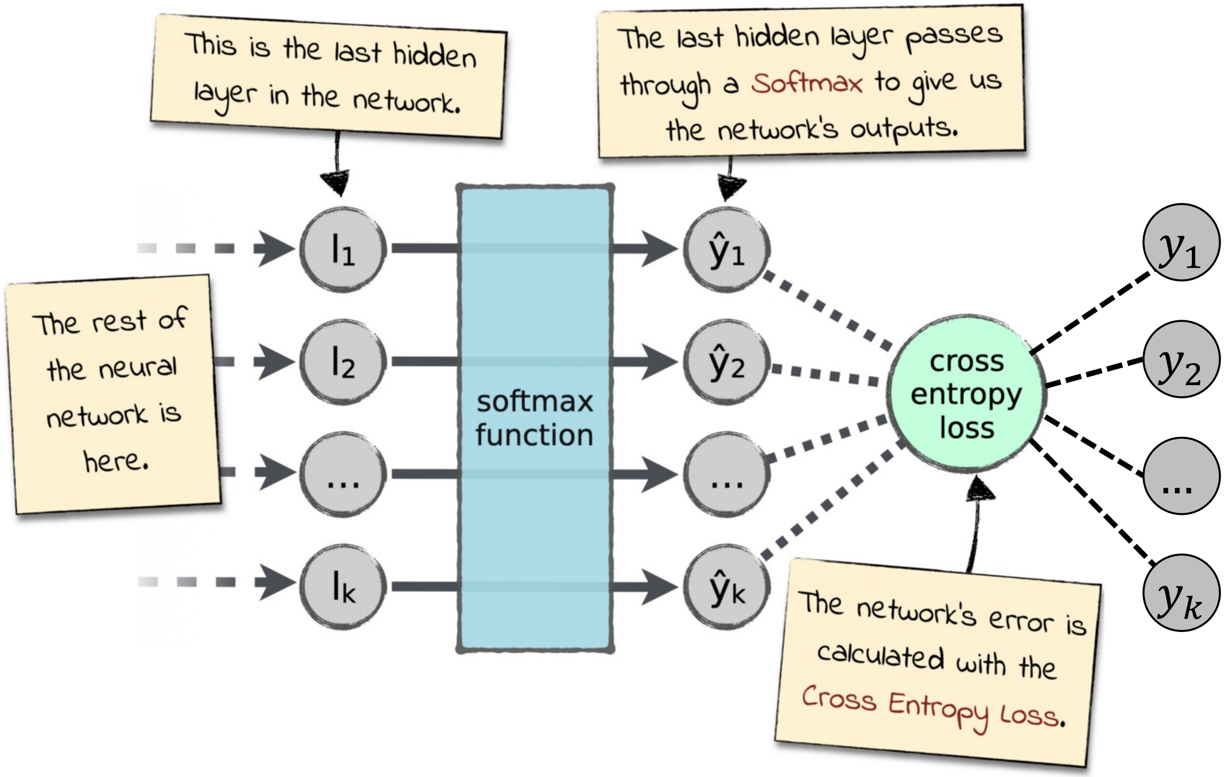
Cross entropy, loss function, softmax, multinomial logistic regression, maximum entropy models, ...


This is the last hidden layer in the network.
The last hidden layer passes through a Softmax to give us the network's outputs.
The rest of the neural network is here.
l_1
l_2
...
l_k
softmax function
$\hat{y}_1$
$\hat{y}_2$
...
$\hat{y}_k$

https://levelup.gitconnected.com/killer-combo-softmax-and-cross-entropy-5907442f60ba

"An application" in ML: DNN

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


This is the last hidden layer in the network.
The last hidden layer passes through a Softmax to give us the network's outputs.
The rest of the neural network is here.
l1
l2
...
lk
softmax function
ŷ1
ŷ2
...
ŷk
cross entropy loss
y_1
y_2
...
y_k
The network's error is calculated with the Cross Entropy Loss.

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.

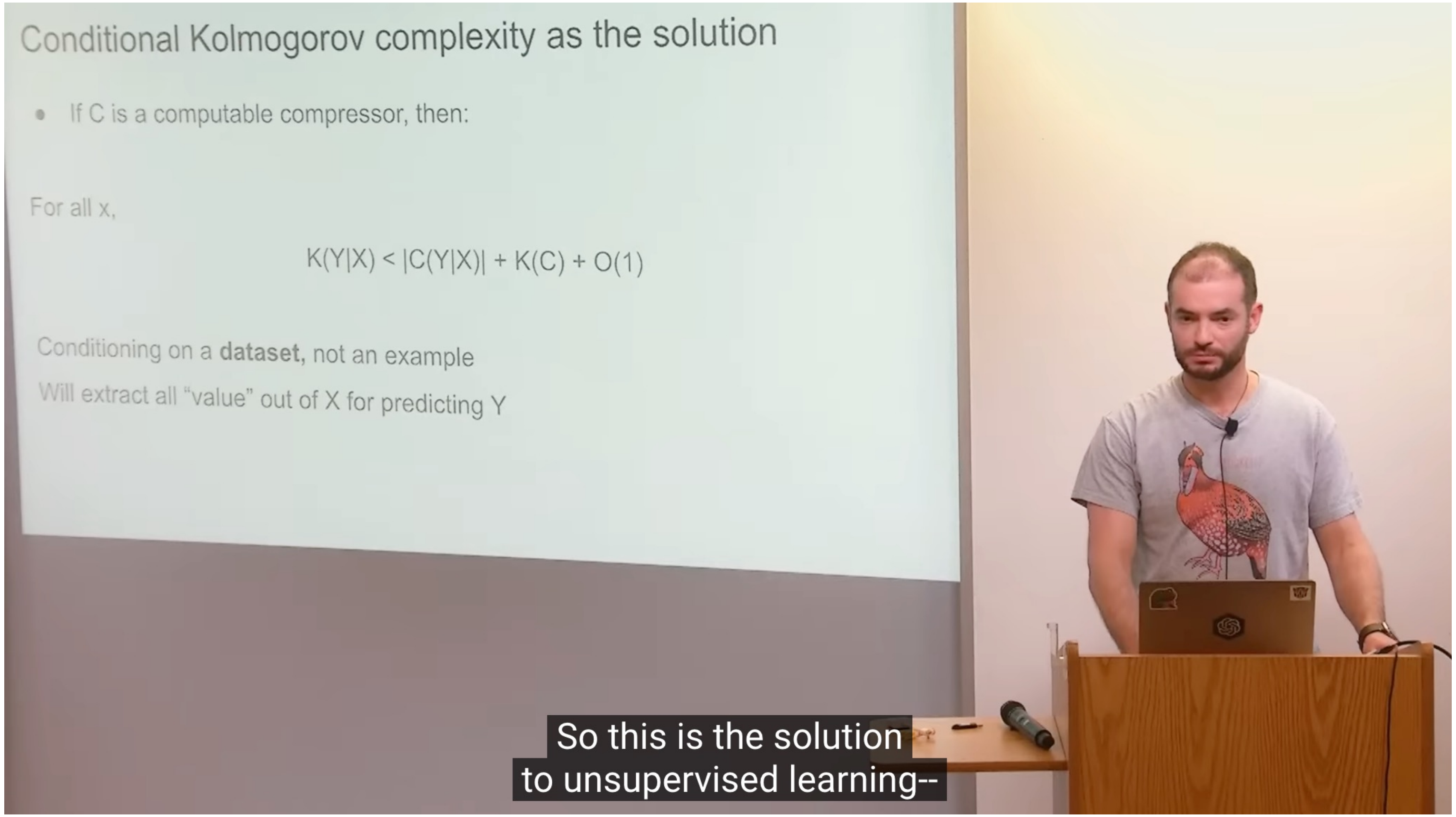
https://levelup.gitconnected.com/killer-combo-softmax-and-cross-entropy-5907442f60ba

Ilya Sutskever @ Simons [2023]

SIMONS INSTITUTE
for the Theory of Computing

An Observation on Generalization

Workshop	Large Language Models and Transformers
Speaker(s)	Ilya Sutskever (OpenAI)
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,

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

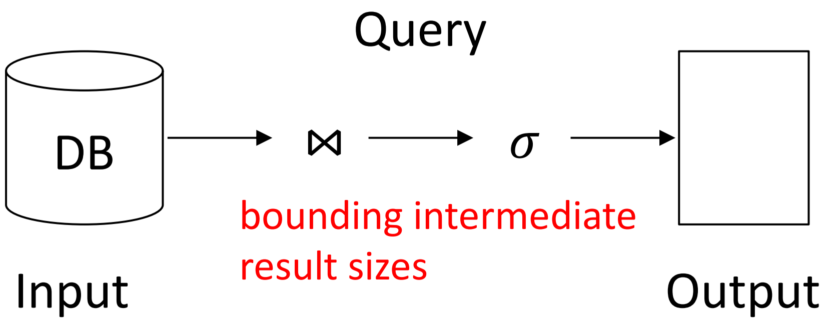
Conditioning on a **dataset**, not an example

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

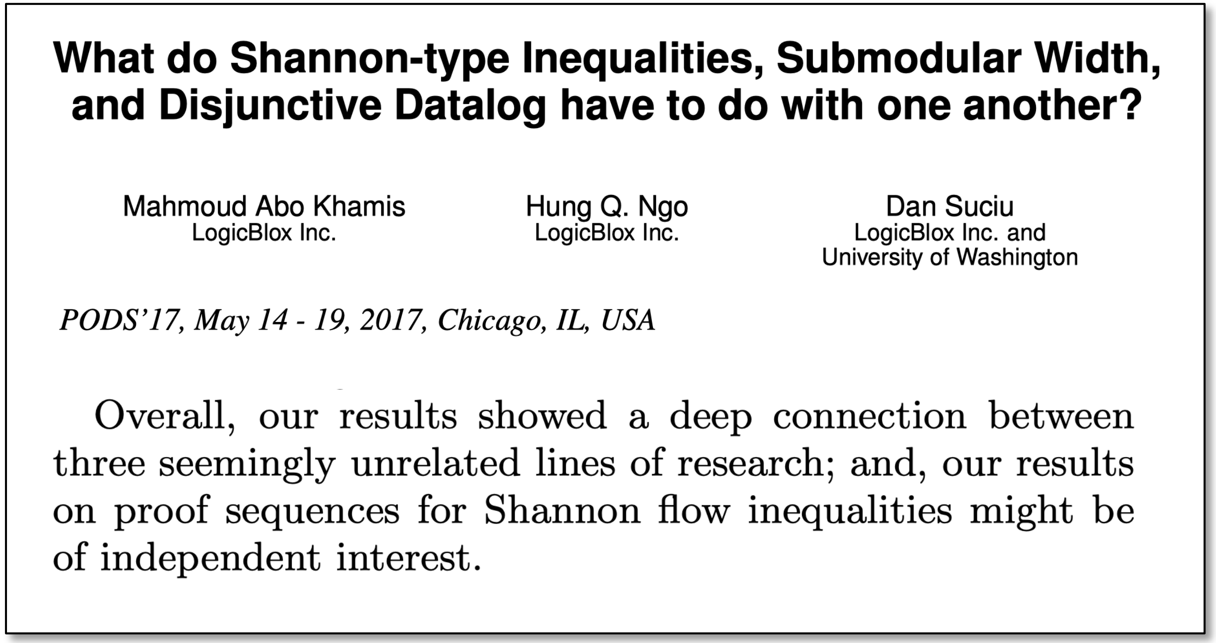
So this is the solution
to unsupervised learning--

Ilya Sutsekever: "An Observation on Generalization". https://simons.berkeley.edu/talks/ilya-sutskever-openai-2023-08-14 , https://www.youtube.com/watch?v=AKMuA_TVz3A&t=1640s
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

An application in DB: Query optimization, cardinality estimation



bounding intermediate result sizes



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

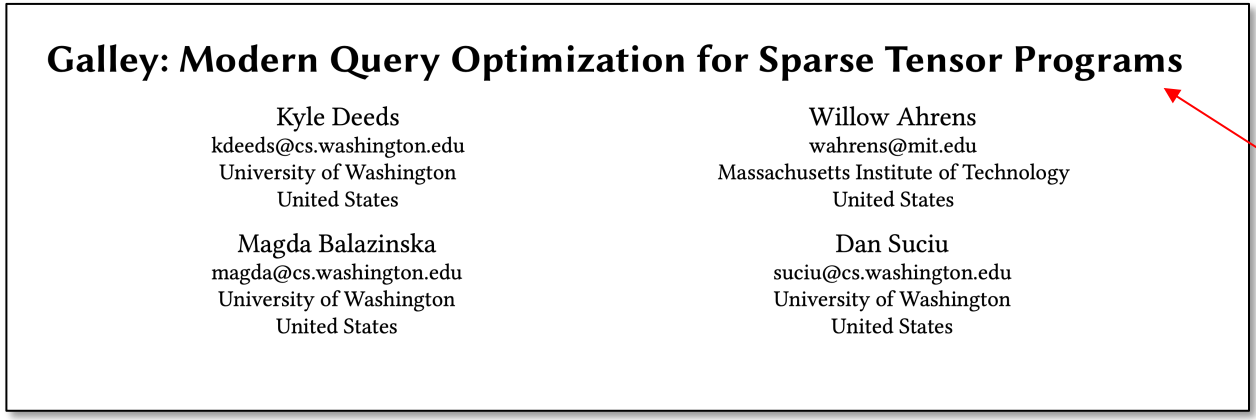
Mahmoud Abo Khamis
LogicBlox Inc.

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

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

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

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

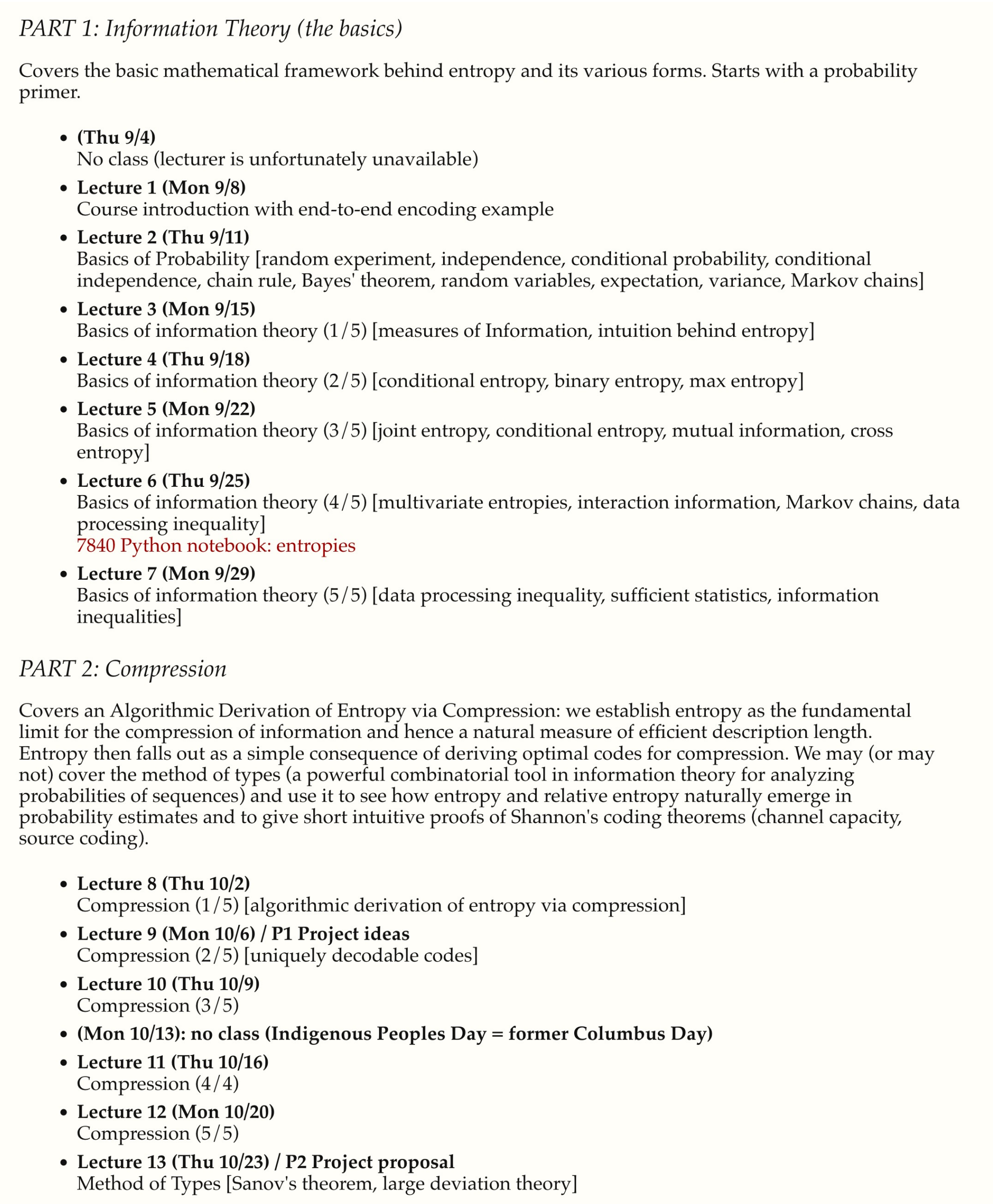
Our DATAlab seminar last year. If interested in similar topic, please subscribe and stop by https://db.khoury.northeastern.edu/activities/

Papers: https://scholar.google.com/scholar?cluster=13416343297965219539 , https://arxiv.org/abs/2408.14706

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: compression
    - Part 3: the axiomatic approach (at least for entropy)
  - practice:
    - Part 4: selected applications to ML, data management (DB), IR

Very approximate outline (will likely change as we progress)



PART 1: Information Theory (the basics)

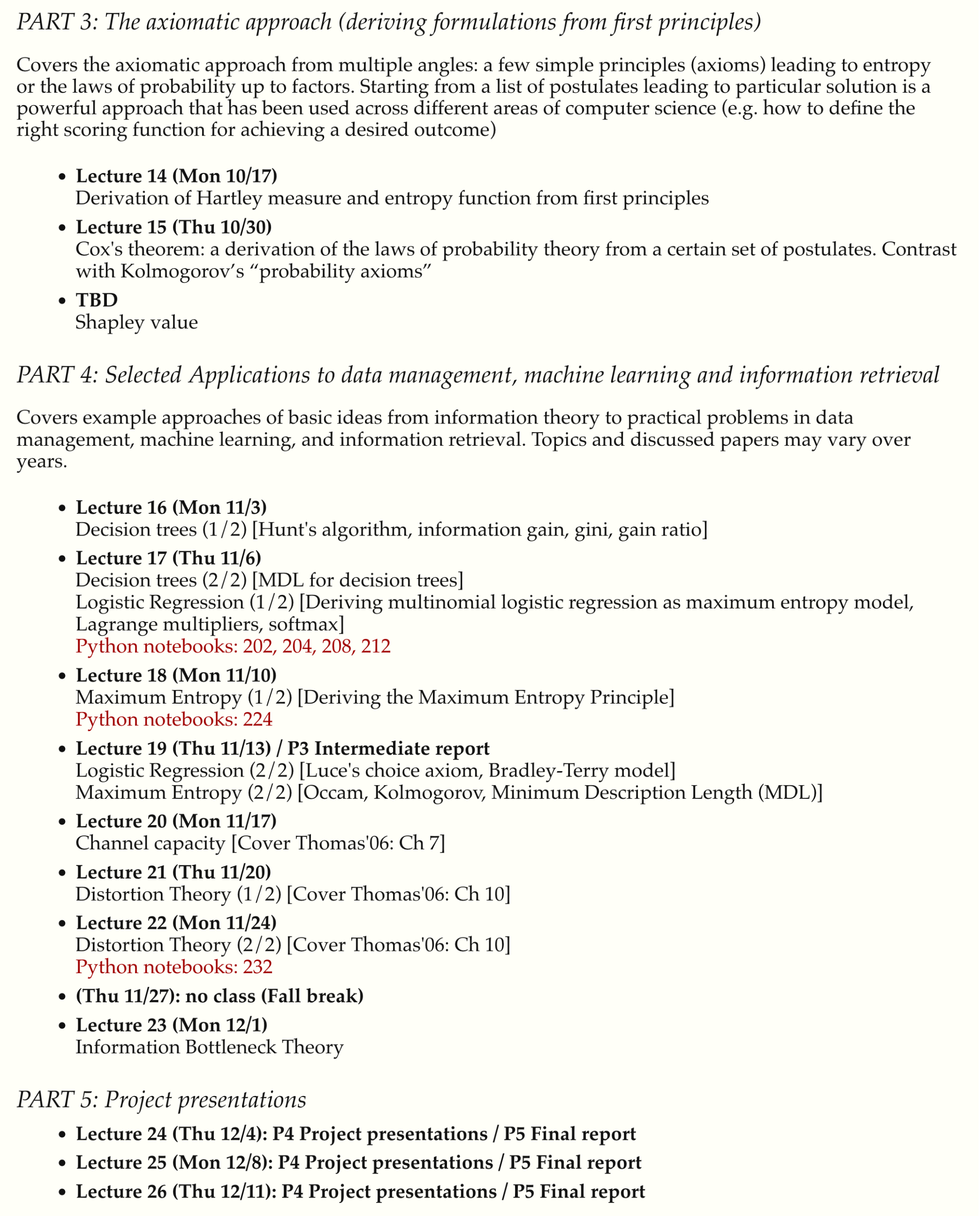
Covers the basic mathematical framework behind entropy and its various forms. Starts with a probability primer.

- **(Thu 9/4)**
  No class (lecturer is unfortunately unavailable)
- **Lecture 1 (Mon 9/8)**
  Course introduction with end-to-end encoding example
- **Lecture 2 (Thu 9/11)**
  Basics of Probability [random experiment, independence, conditional probability, conditional independence, chain rule, Bayes' theorem, random variables, expectation, variance, Markov chains]
- **Lecture 3 (Mon 9/15)**
  Basics of information theory (1/5) [measures of Information, intuition behind entropy]
- **Lecture 4 (Thu 9/18)**
  Basics of information theory (2/5) [conditional entropy, binary entropy, max entropy]
- **Lecture 5 (Mon 9/22)**
  Basics of information theory (3/5) [joint entropy, conditional entropy, mutual information, cross entropy]
- **Lecture 6 (Thu 9/25)**
  Basics of information theory (4/5) [multivariate entropies, interaction information, Markov chains, data processing inequality]
  7840 Python notebook: entropies
- **Lecture 7 (Mon 9/29)**
  Basics of information theory (5/5) [data processing inequality, sufficient statistics, information inequalities]

PART 2: Compression

Covers an Algorithmic Derivation of Entropy via Compression: we establish entropy as the fundamental limit for the compression of information and hence a natural measure of efficient description length. Entropy then falls out as a simple consequence of deriving optimal codes for compression. We may (or may not) cover the method of types (a powerful combinatorial tool in information theory for analyzing probabilities of sequences) and use it to see how entropy and relative entropy naturally emerge in probability estimates and to give short intuitive proofs of Shannon's coding theorems (channel capacity, source coding).

- **Lecture 8 (Thu 10/2)**
  Compression (1/5) [algorithmic derivation of entropy via compression]
- **Lecture 9 (Mon 10/6) / P1 Project ideas**
  Compression (2/5) [uniquely decodable codes]
- **Lecture 10 (Thu 10/9)**
  Compression (3/5)
- **(Mon 10/13): no class (Indigenous Peoples Day = former Columbus Day)**
- **Lecture 11 (Thu 10/16)**
  Compression (4/4)
- **Lecture 12 (Mon 10/20)**
  Compression (5/5)
- **Lecture 13 (Thu 10/23) / P2 Project proposal**
  Method of Types [Sanov's theorem, large deviation theory]



PART 3: 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 14 (Mon 10/17)**
  Derivation of Hartley measure and entropy function from first principles
- **Lecture 15 (Thu 10/30)**
  Cox's theorem: a derivation of the laws of probability theory from a certain set of postulates. Contrast with Kolmogorov's "probability axioms"
- **TBD**
  Shapley value

PART 4: 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 16 (Mon 11/3)**
  Decision trees (1/2) [Hunt's algorithm, information gain, gini, gain ratio]
- **Lecture 17 (Thu 11/6)**
  Decision trees (2/2) [MDL for decision trees]
  Logistic Regression (1/2) [Deriving multinomial logistic regression as maximum entropy model, Lagrange multipliers, softmax]
  Python notebooks: 202, 204, 208, 212
- **Lecture 18 (Mon 11/10)**
  Maximum Entropy (1/2) [Deriving the Maximum Entropy Principle]
  Python notebooks: 224
- **Lecture 19 (Thu 11/13) / P3 Intermediate report**
  Logistic Regression (2/2) [Luce's choice axiom, Bradley-Terry model]
  Maximum Entropy (2/2) [Occam, Kolmogorov, Minimum Description Length (MDL)]
- **Lecture 20 (Mon 11/17)**
  Channel capacity [Cover Thomas'06: Ch 7]
- **Lecture 21 (Thu 11/20)**
  Distortion Theory (1/2) [Cover Thomas'06: Ch 10]
- **Lecture 22 (Mon 11/24)**
  Distortion Theory (2/2) [Cover Thomas'06: Ch 10]
  Python notebooks: 232
- **(Thu 11/27): no class (Fall break)**
- **Lecture 23 (Mon 12/1)**
  Information Bottleneck Theory

PART 5: Project presentations

- **Lecture 24 (Thu 12/4): P4 Project presentations / P5 Final report**
- **Lecture 25 (Mon 12/8): P4 Project presentations / P5 Final report**
- **Lecture 26 (Thu 12/11): P4 Project presentations / P5 Final report**

Quick background on me

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

Let's take turns

As you are called, please briefly state:

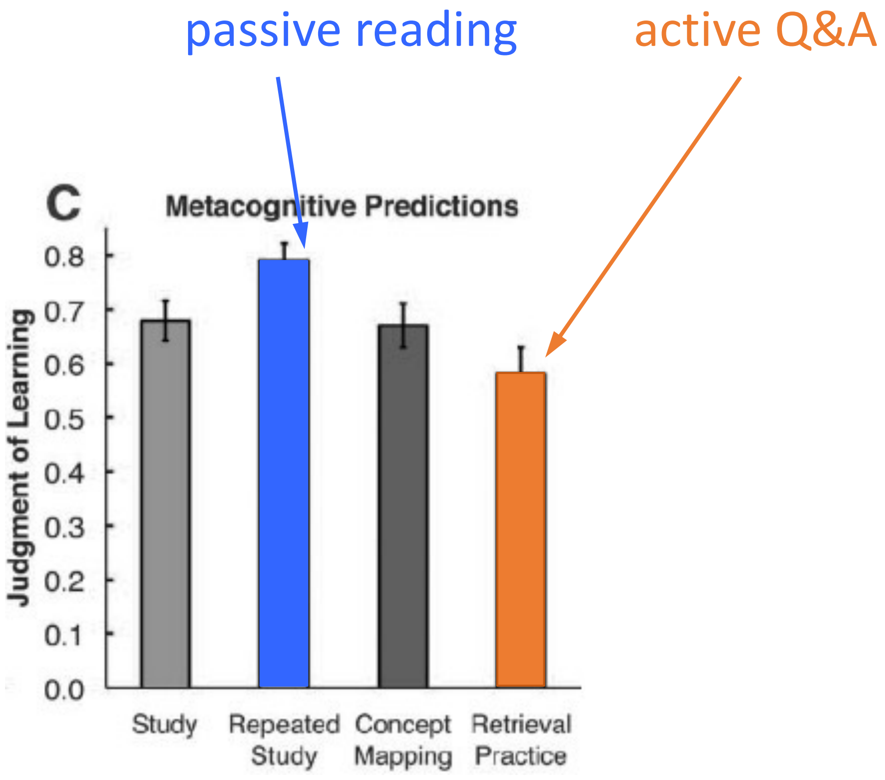
1. What area are you working on? Who is your PhD advisor? How did you learn about this class? Why do you consider taking it?

2. What do you hope to get out of this course ☺? What is the topic from the course page (or information theory, in general) that you are most interested in? What could be your project?

3. What is your biggest fear for this course ☹?

Pedagogy & Logistics

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



Judged performance
(=what people think)

?

Data from: Karpicke & Blunt, "Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping," Science, 2011. https://doi.org/10.1126/science.1199327
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

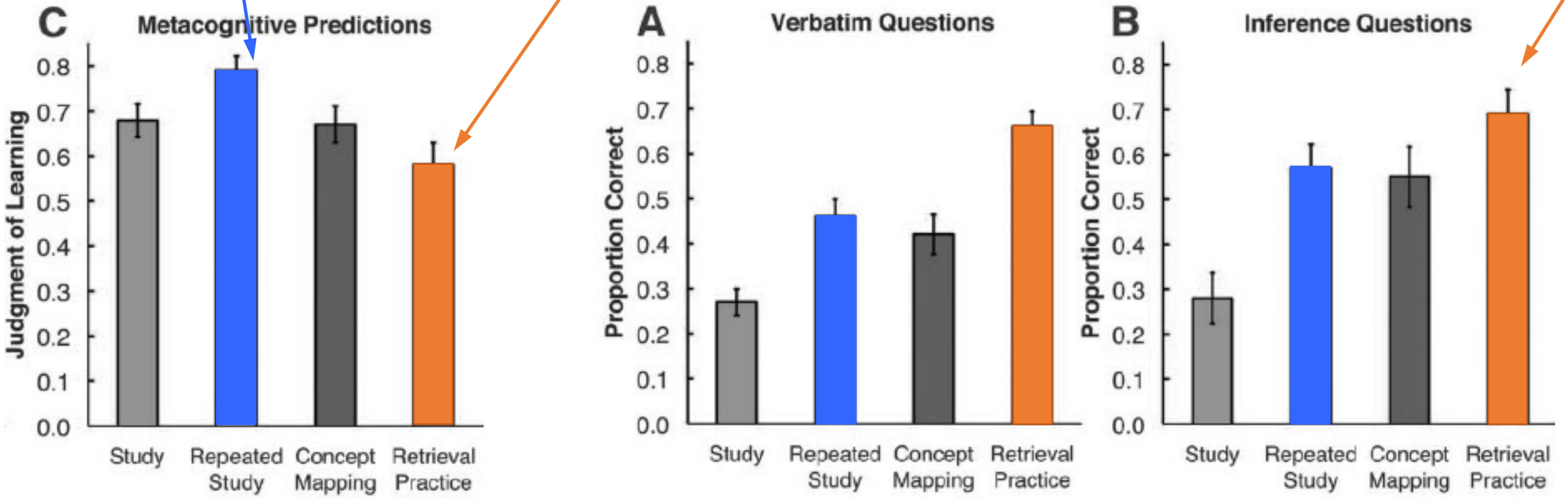
Studying new material: "Under which study condition do you think you learn better?"

Surprise, surprise: active Q&A works better for learning

!

passive reading

active Q&A

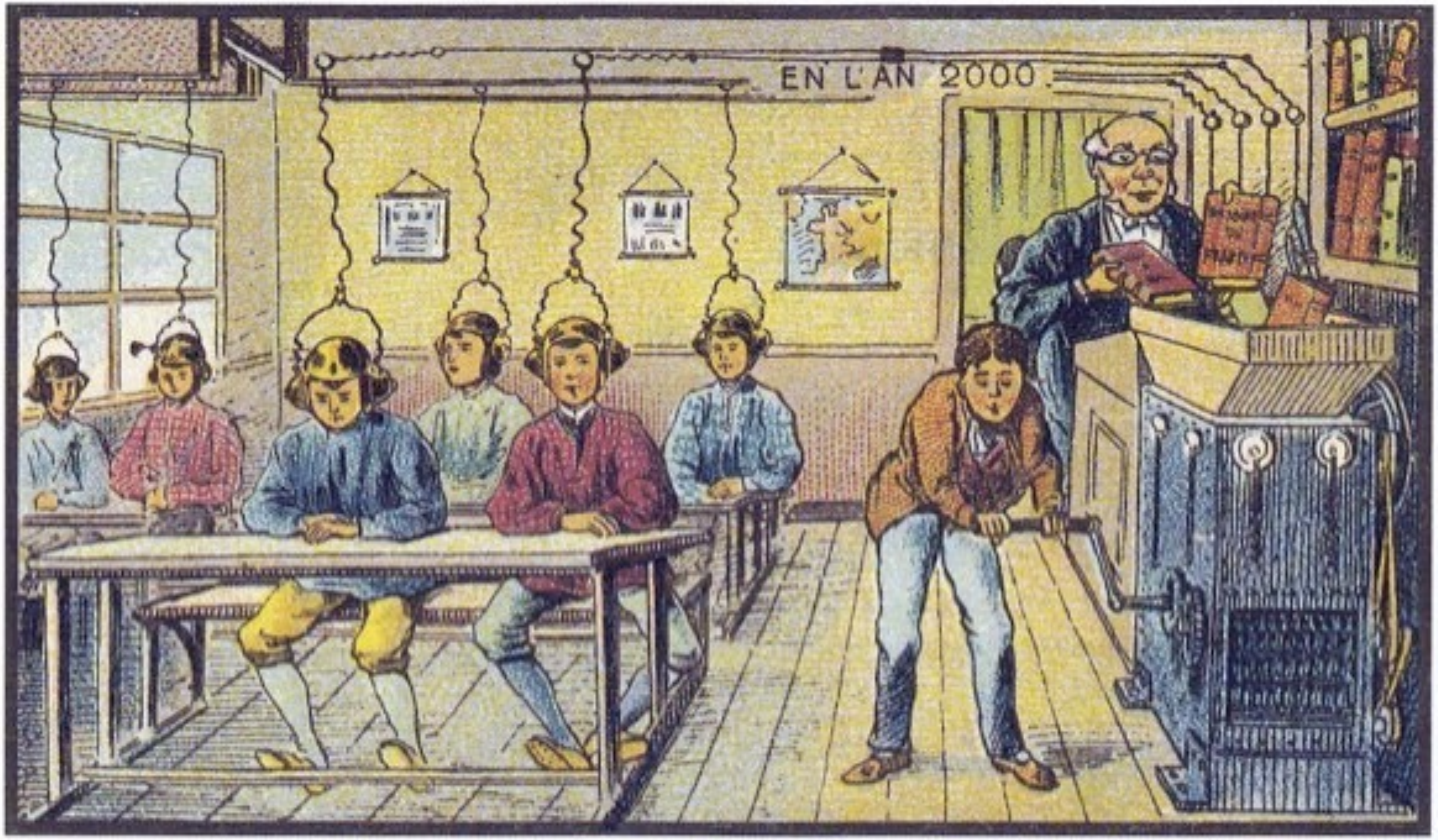


Judged performance
(=what people think)

Actual performance
(=what is actually working)

Data from: Karpicke & Blunt, "Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping," Science, 2011. https://doi.org/10.1126/science.1199327

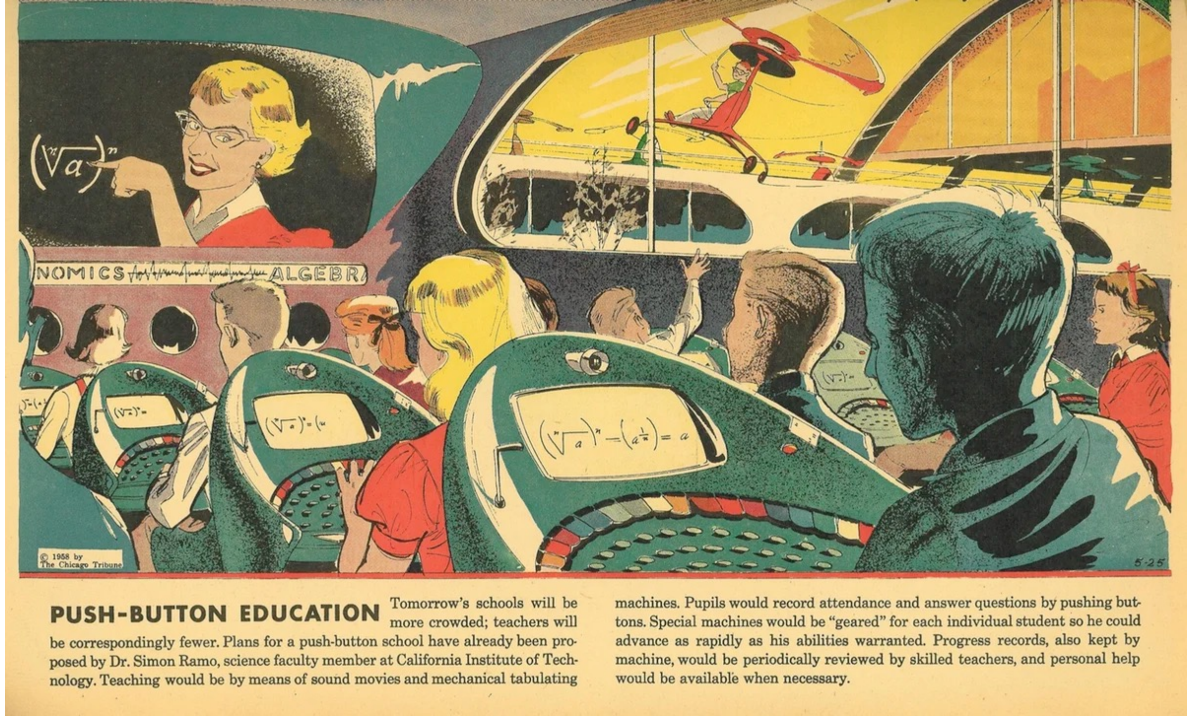
The year 2000 imagined in 1900



At School

Source: Jean-Marc Côté: https://publicdomainreview.org/collection/a-19th-century-vision-of-the-year-2000, https://commons.wikimedia.org/wiki/File%3AFrance_in_XXI_Century._School.jpg

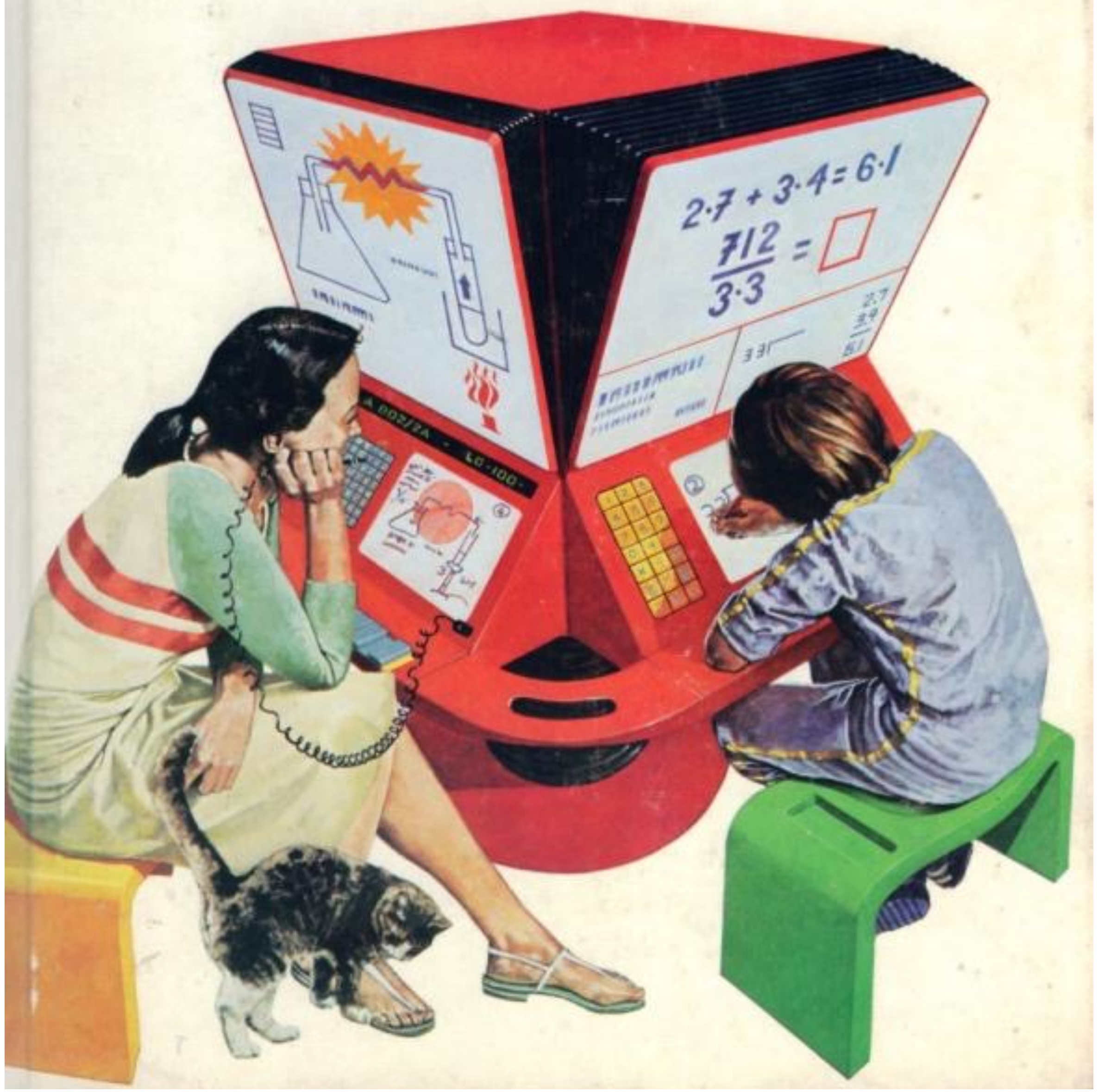
Late 1950s

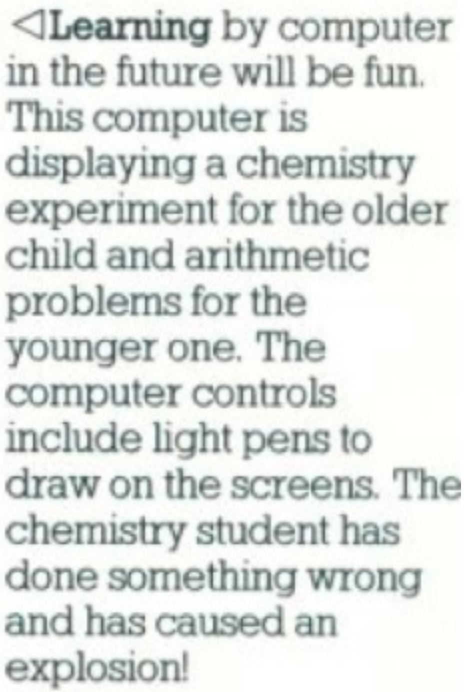


PUSH-BUTTON EDUCATION Tomorrow's schools will be more crowded; teachers will be correspondingly fewer. Plans for a push-button school have already been proposed by Dr. Simon Ramo, science faculty member at California Institute of Technology. Teaching would be by means of sound movies and mechanical tabulating machines. Pupils would record attendance and answer questions by pushing buttons. Special machines would be "geared" for each individual student so he could advance as rapidly as his abilities warranted. Progress records, also kept by machine, would be periodically reviewed by skilled teachers, and personal help would be available when necessary.

Source: https://paleofuture.com/blog/2011/8/24/the-push-button-school-of-tomorrow-1958.html

Predicting the role of IT in Education

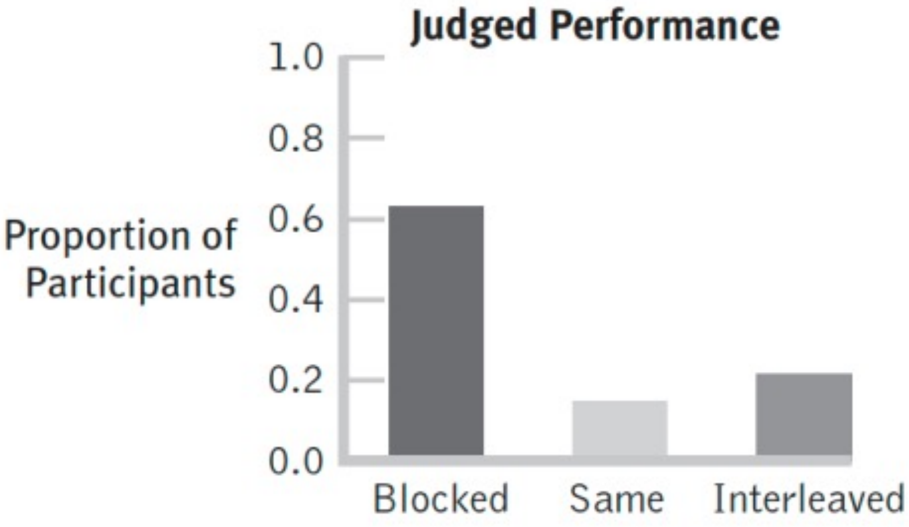




◁**Learning** by computer in the future will be fun. This computer is displaying a chemistry experiment for the older child and arithmetic problems for the younger one. The computer controls include light pens to draw on the screens. The chemistry student has done something wrong and has caused an explosion!

Source: Neil Ardley, World of tomorrow: School, Work and Play, 1981. https://archive.org/details/schoolworkplaywo0000neil/page/40/mode/2up

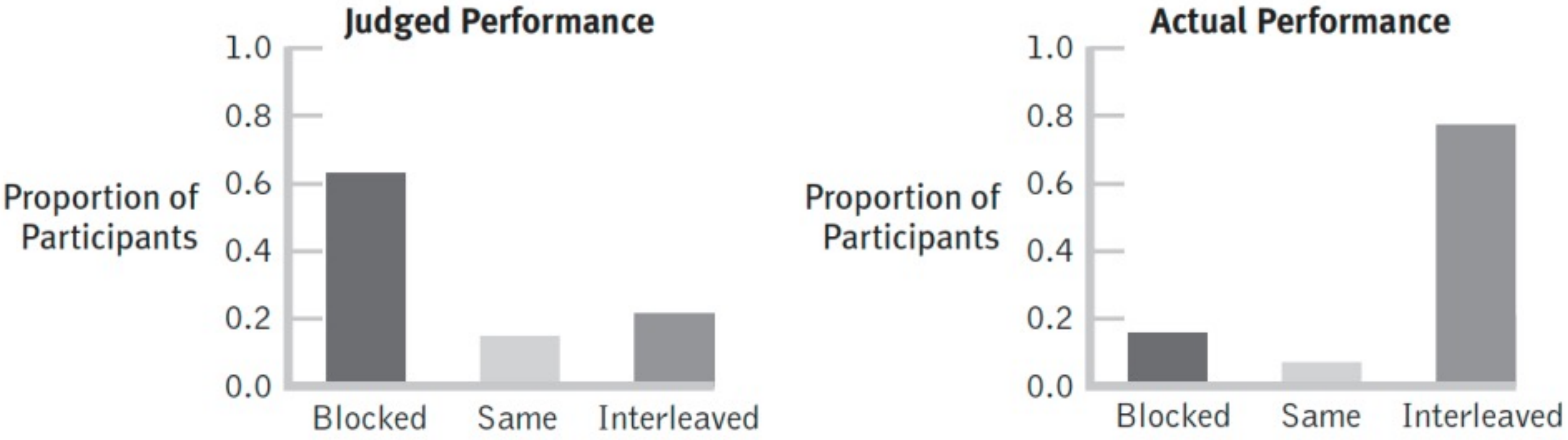
Sequencing Material: "Under which teaching condition do you think you learn better?"

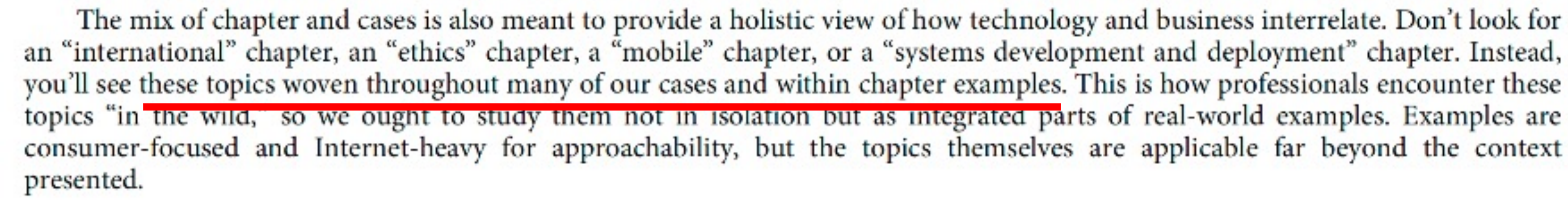


?

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)," Gallaugher, 2012. https://gallaugher.com/book/

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

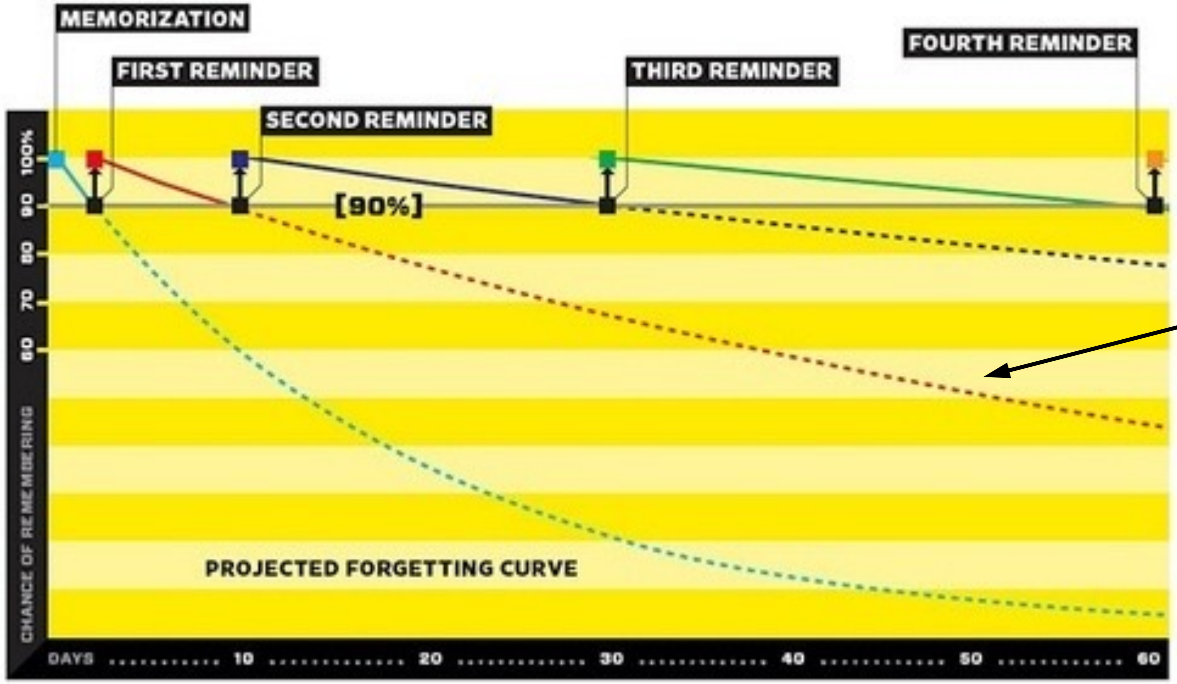




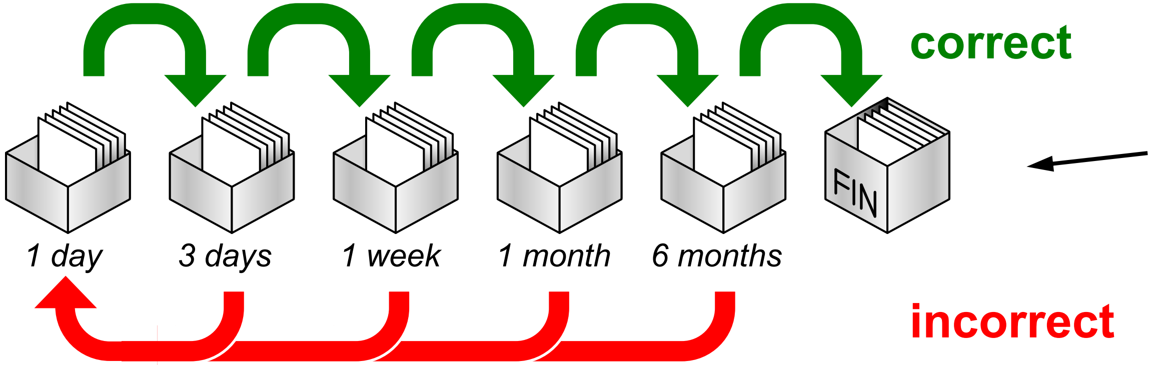
The mix of chapter and cases is also meant to provide a holistic view of how technology and business interrelate. Don't look for an "international" chapter, an "ethics" chapter, a "mobile" chapter, or a "systems development and deployment" chapter. Instead, you'll see these topics woven throughout many 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)," Gallaugher, 2012. https://gallaugher.com/book/

Spaced Repetition



Ebbinghaus Forgetting Curve



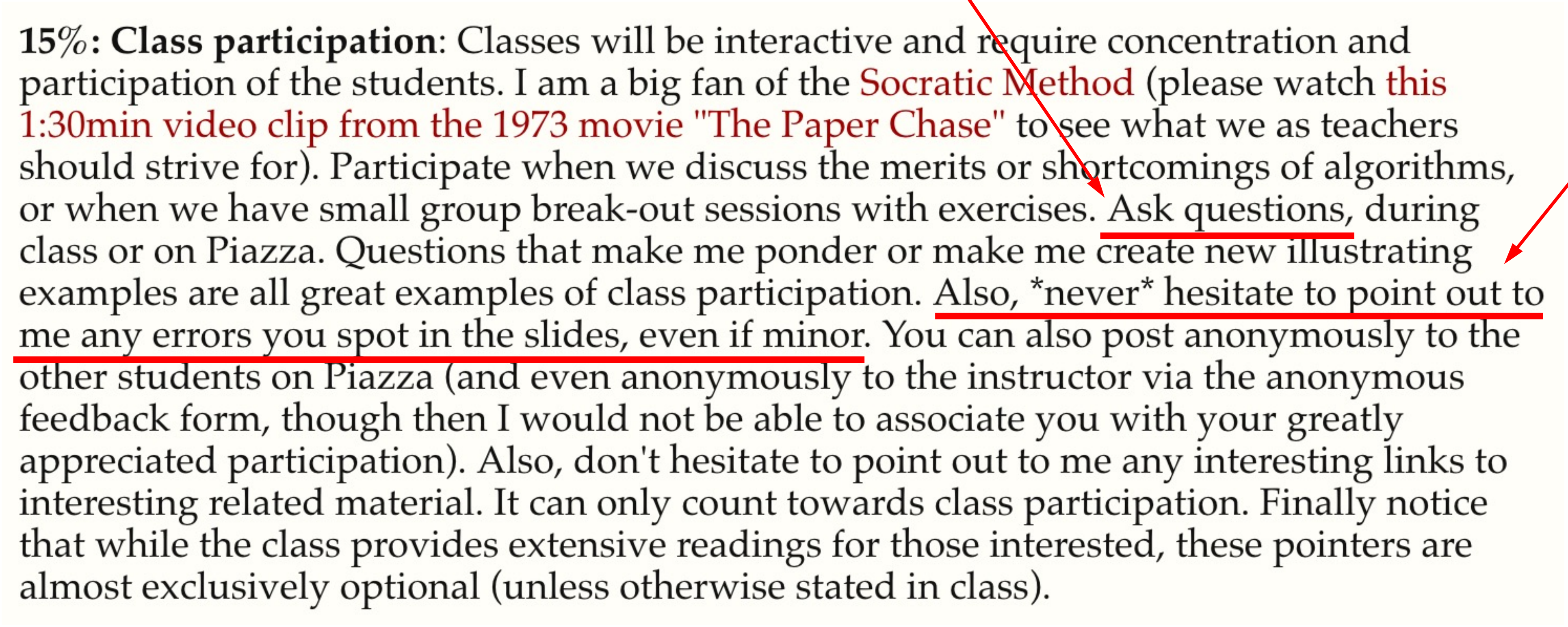
Leitner System (Pimsleur's graduated interval recall)

Sources: http://www.wired.com/2008/04/ff-wozniak/,
Gatterbauer & Suciu, "Managing Structured Collections of Community Data", CIDR 2011. http://cidrdb.org/cidr2011/Papers/CIDR11_Paper28.pdf
Gatterbauer. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Coursework / Evaluation (1/3)

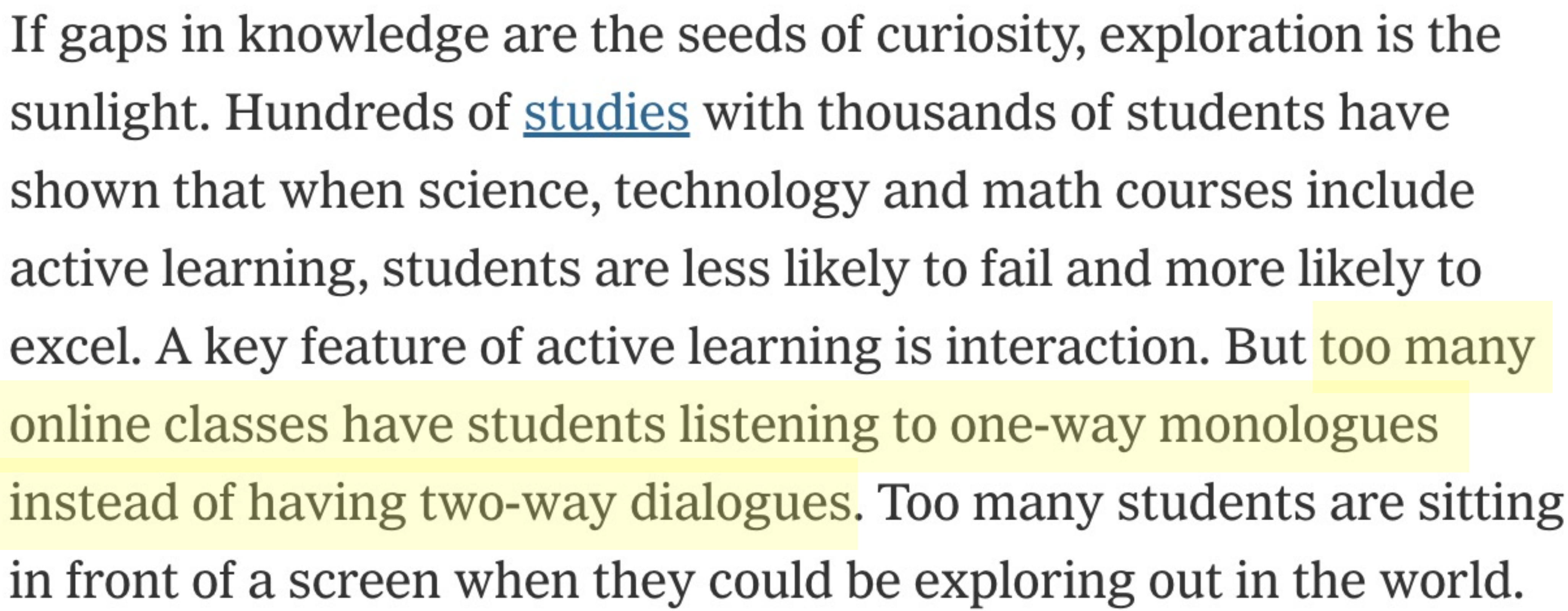
Stop me and ask questions if I am talking too much and not explaining enough!

"your obligation to dissent"



15%: Class participation: Classes will be interactive and require concentration and participation of the students. I am a big fan of the Socratic Method (please watch this 1:30min video clip from the 1973 movie "The Paper Chase" to see what we as teachers should strive for). 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 me ponder or make me create new illustrating examples are all great examples of class participation. Also, *never* hesitate to point out to me any errors you spot in the slides, even if minor. You can also post anonymously to the other students on Piazza (and even anonymously to the instructor via the anonymous feedback form, though then I would not be able to associate you with your greatly appreciated participation). Also, don't hesitate to point out to me 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.

Source: https://www.nytimes.com/2020/09/07/opinion/remote-school.html

Lectures are not recorded (2/2)

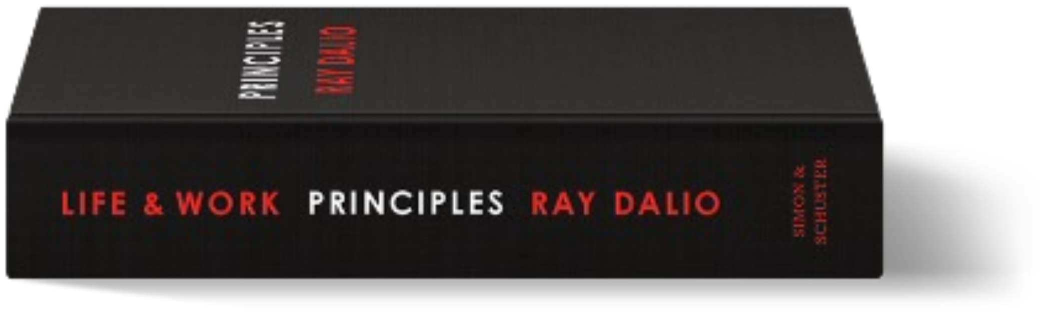
- We would like to have an encouraging environment in which everyone can speak up and discuss ideas freely without concern that discussions will be available outside of classroom.
- The course slides are comprehensive and should allow you to be able to remember the key lessons from class (except for background stories I may tell you). Lecture slides will be posted within 2 days after each class, i.e. WED for MON classes, SAT for THU classes).
- 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].

Mass. Gen. Laws ch. 272, § 99: https://malegislature.gov/laws/generallaws/partiv/titlei/chapter272/section99

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!"

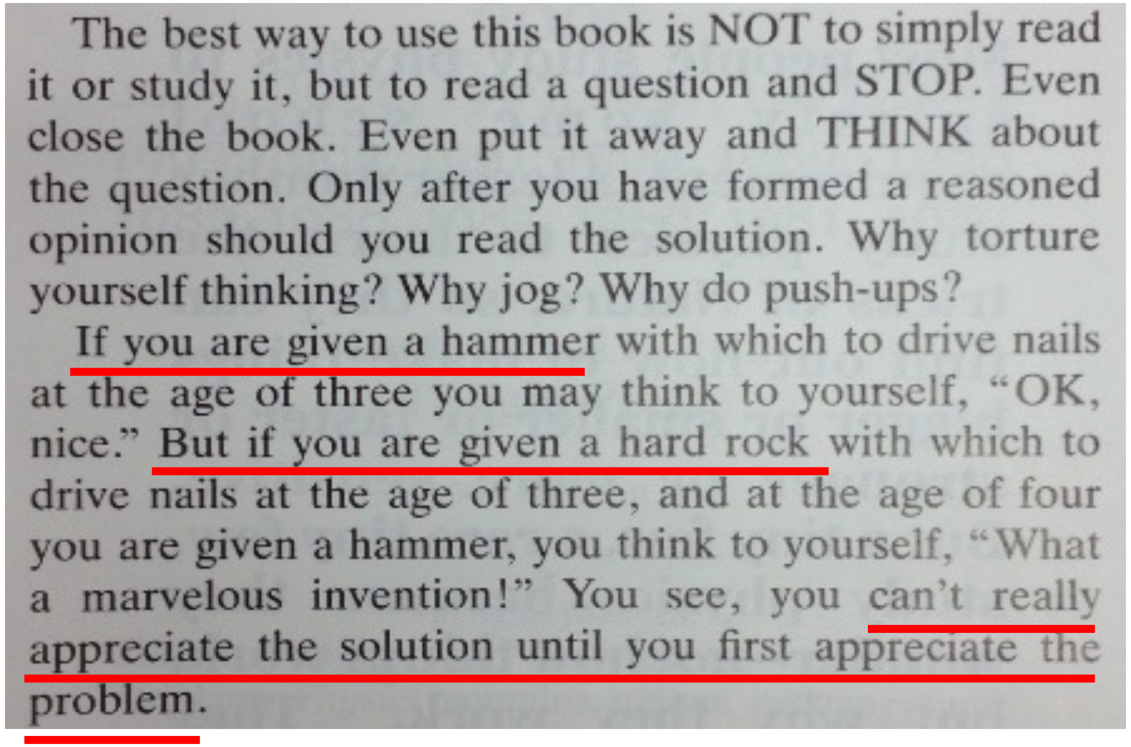
Notice the story around Bridgewater & Ray Dalio is interesting and still being written. See e.g.
https://www.nytimes.com/2023/11/01/business/how-does-the-worlds-largest-hedge-fund-really-make-its-money.html
https://www.vanityfair.com/news/2023/11/james-comey-dalio-bridgewater-the-fund
https://nymag.com/intelligencer/article/ray-dalio-rob-copeland-the-fund-book-excerpt.html
That said, the ideas behind "Principles" are still worth reading



Source: Ray Dalio. https://en.wikipedia.org/wiki/Principles_(book)

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"

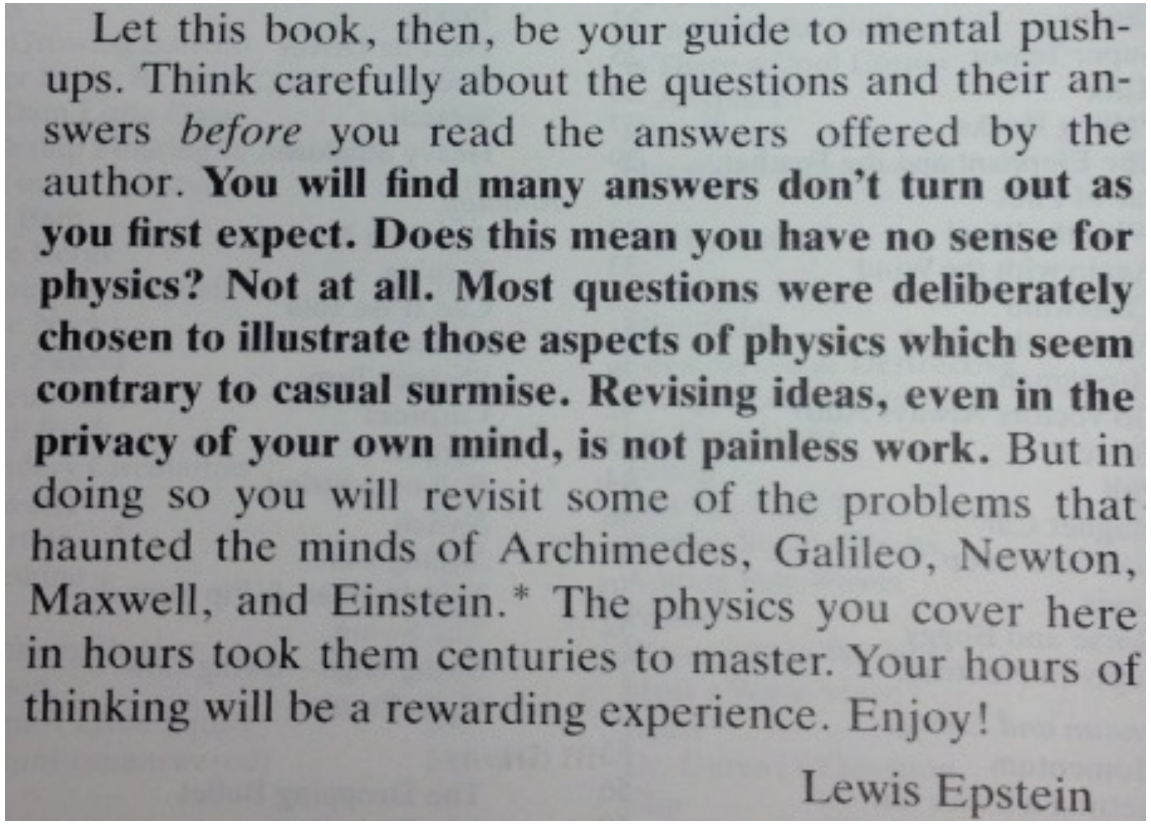


The best way to use this book is NOT to simply read it or study it, but to read a question and STOP. Even close the book. Even put it away and THINK about the question. Only after you have formed a reasoned opinion should you read the solution. Why torture yourself thinking? Why jog? Why do push-ups?

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.

...

...



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!

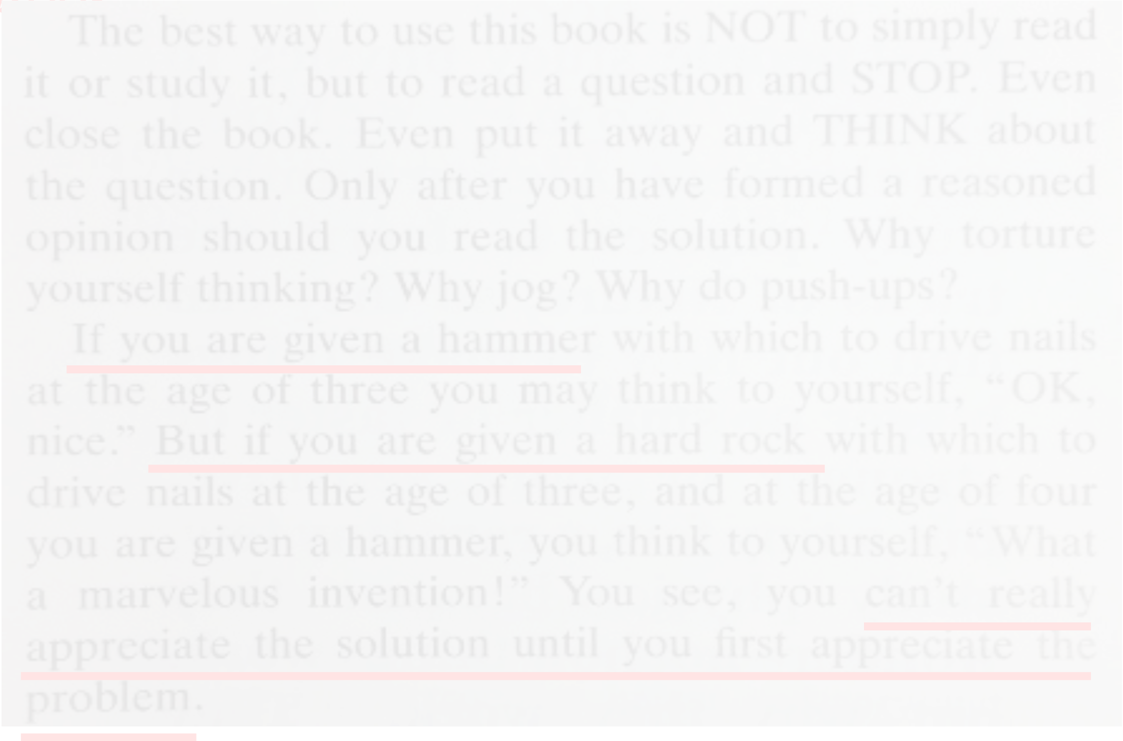
Lewis Epstein

We will also have in-class exercises!

Source: "Thinking Physics: Understanding Practical Reality", Lewis Carroll Epstein, 1979-2009. https://www.goodreads.com/book/show/268266.Thinking_Physics

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 exercises!

Source: "Thinking Physics: Understanding Practical Reality", Lewis Carroll Epstein, 1979-2009. https://www.goodreads.com/book/show/268266.Thinking_Physics

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

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."

Source: Craig Nelson, Dysfunctional Illusions of Rigor: Part 1 - Basic Illusions, http://cgi.stanford.edu/~dept-ctl/tomprof/posting.php?ID=1058

The "Surfer Analogy" for time management

Source: http://stwww.surfermag.com/files/2013/10/Yak_Charlie-970x646.jpg

Class scribes

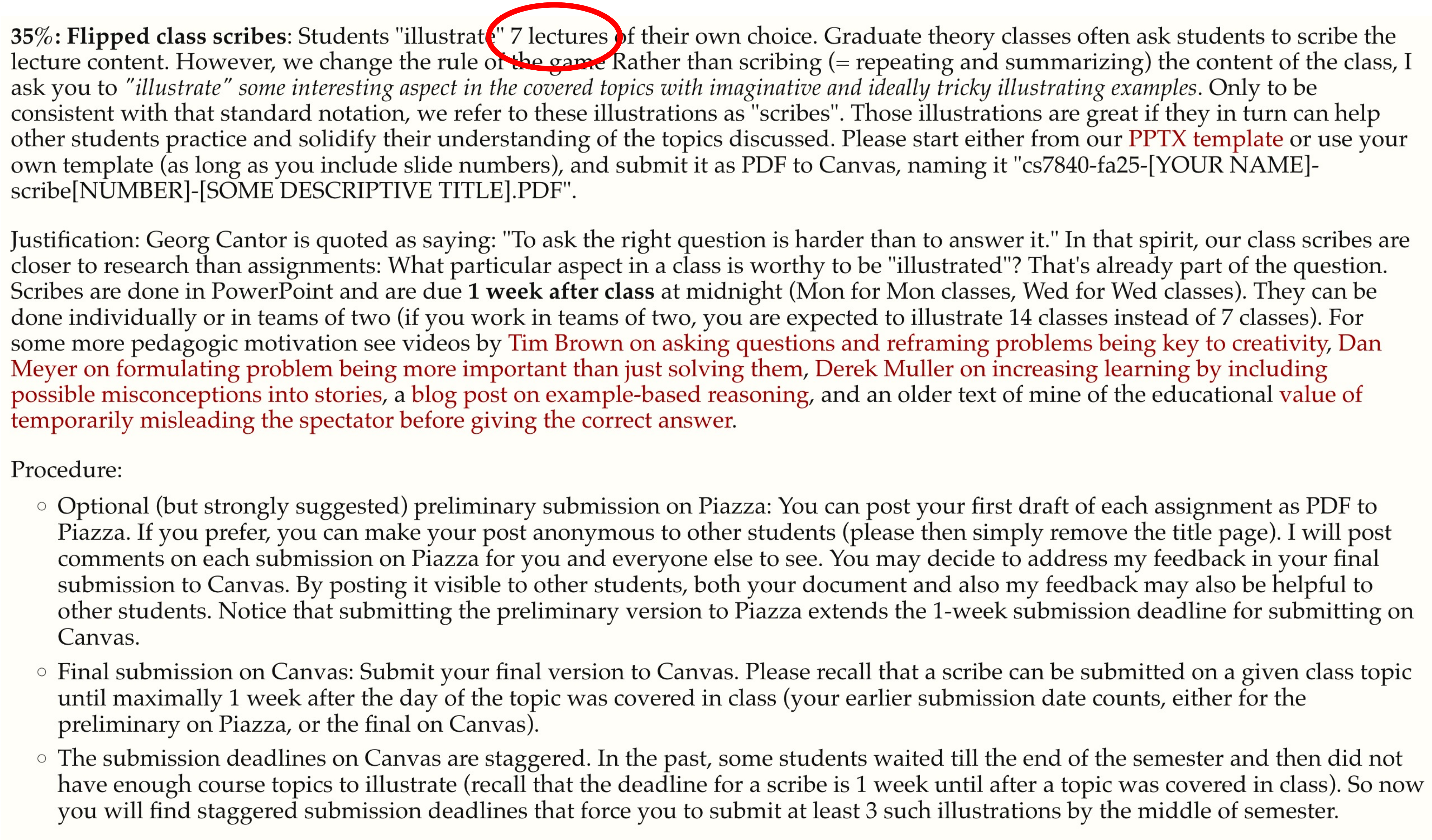
"To ask the right question is harder than to answer it."

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

Coursework / Evaluation (2/3)



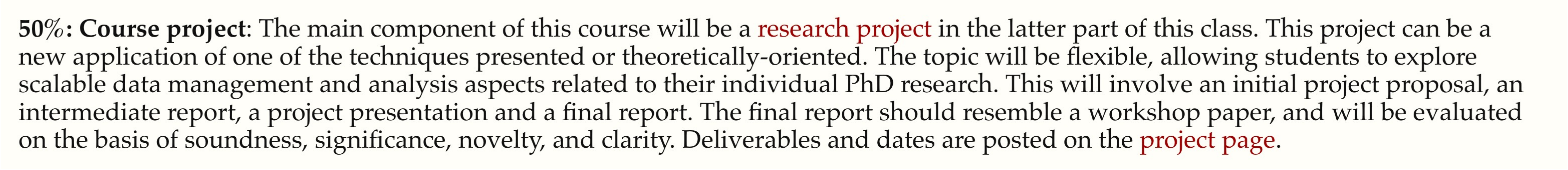
35%: Flipped class scribes: Students "illustrate" 7 lectures of their own choice. Graduate theory classes often ask students to scribe the lecture content. However, we change the rule of the game. Rather than scribing (= repeating and summarizing) the content of the class, I ask you to *"illustrate" some interesting aspect in the covered topics with imaginative and ideally tricky illustrating examples*. Only to be consistent with that standard notation, we refer to these illustrations as "scribes". 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-fa25-[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:

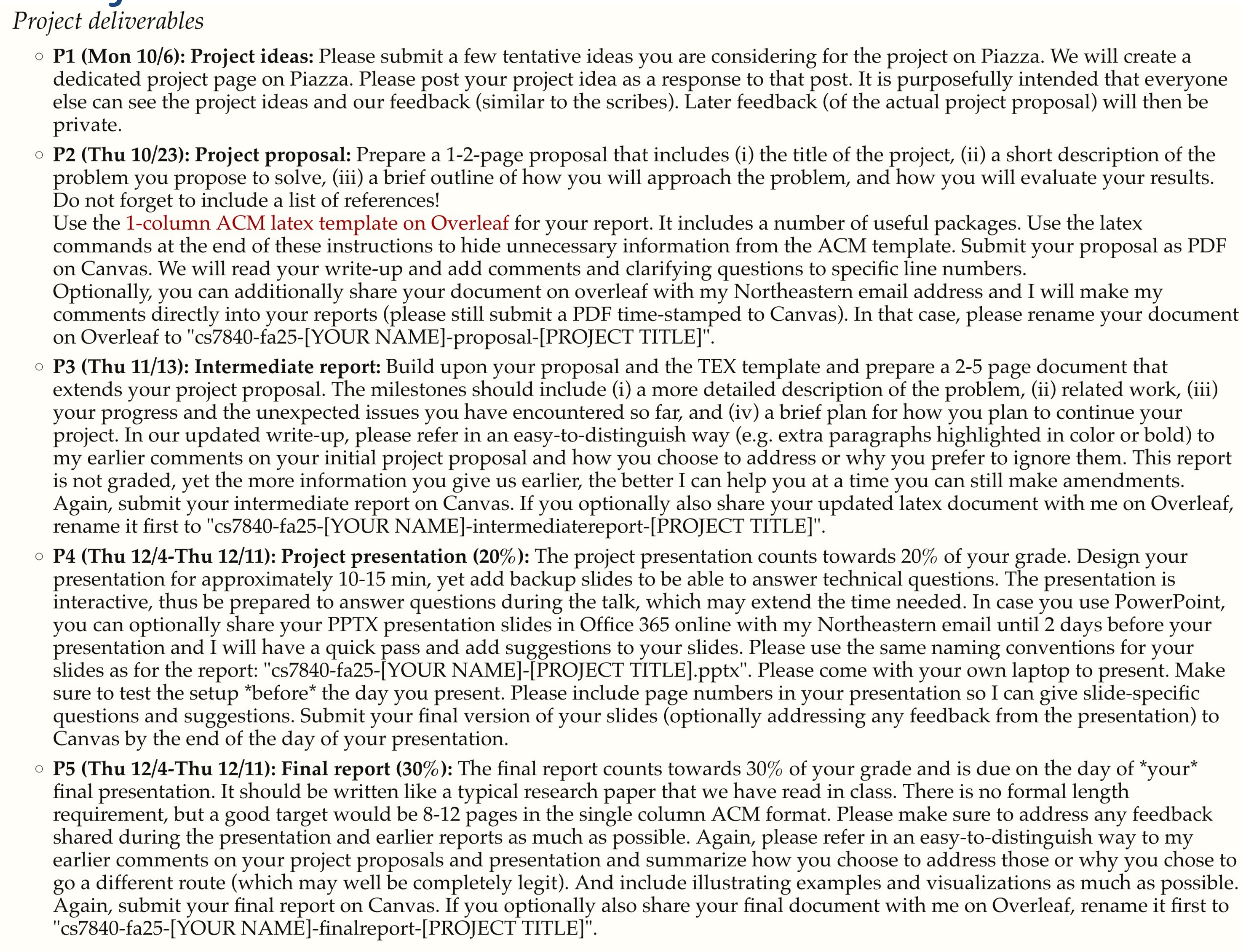
- Optional (but strongly suggested) preliminary submission on Piazza: You can post your first draft of each assignment as PDF to Piazza. If you prefer, you can make your post anonymous to other students (please then simply remove the title page). I will post comments on each submission on Piazza for you and everyone else to see. You may decide to address my 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 1-week submission deadline for submitting on Canvas.
- Final submission on Canvas: Submit your final version to Canvas. Please recall that a scribe can be submitted on a given class topic until maximally 1 week after the day of the topic was covered in class (your earlier submission date counts, either for the preliminary on Piazza, or the final on Canvas).
- The submission deadlines on Canvas are staggered. In the past, some students waited till the end of the semester and then did not have enough course topics to illustrate (recall that the deadline for a scribe is 1 week until after a topic was covered in class). So now you will find staggered submission deadlines that force you to submit at least 3 such illustrations by the middle of semester.

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



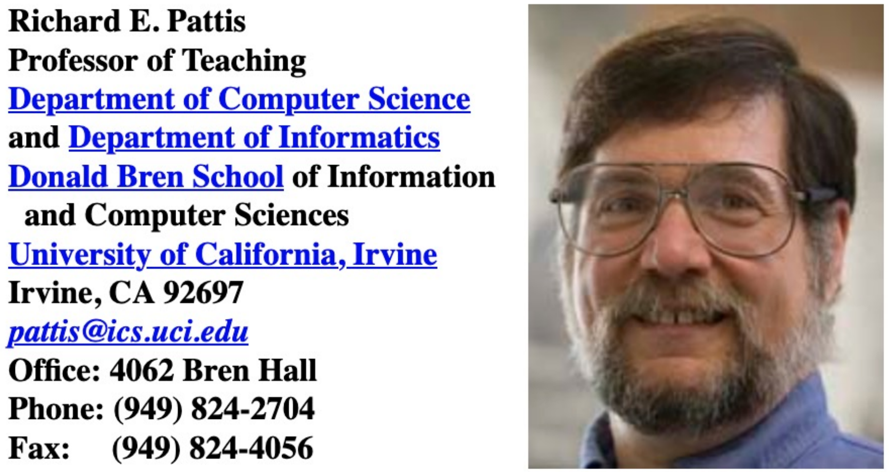
Project deliverables

- **P1 (Mon 10/6): Project ideas:** Please submit a few tentative ideas you are considering for the project on Piazza. We will create a dedicated project page on Piazza. Please post your project idea as a response to that post. It is purposefully intended that everyone else can see the project ideas and our feedback (similar to the scribes). Later feedback (of the actual project proposal) will then be private.
- **P2 (Thu 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 my Northeastern email address and I 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-fa25-[YOUR NAME]-proposal-[PROJECT TITLE]".
- **P3 (Thu 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 my 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 earlier, the better I 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 me on Overleaf, rename it first to "cs7840-fa25-[YOUR NAME]-intermediatereport-[PROJECT TITLE]".
- **P4 (Thu 12/4-Thu 12/11): 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 my Northeastern email until 2 days before your presentation and I 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-fa25-[YOUR NAME]-[PROJECT TITLE].pptx". Please come with your own laptop to present. Make sure to test the setup *before* the day you present. Please include page numbers in your presentation so I can give slide-specific questions and suggestions. Submit your final version of your slides (optionally addressing any feedback from the presentation) to Canvas by the end of the day of your presentation.
- **P5 (Thu 12/4-Thu 12/11): Final report (30%):** The final report counts towards 30% of your grade and is due on the day of *your* final presentation. 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 me on Overleaf, rename it first to "cs7840-fa25-[YOUR NAME]-finalreport-[PROJECT TITLE]".

Project

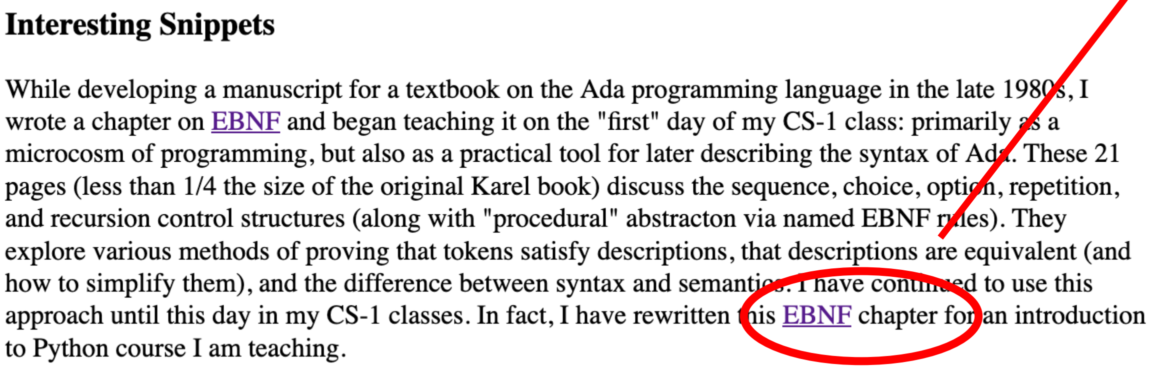
- example past projects
  - chosen from class topics: "Compression via Kraft's Constrained Floor-Ceiling Decision Problem"
  - from current research & class topic: "information bottlenecks in robotics"
  - "applications of information theory in information visualization"
- ?

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



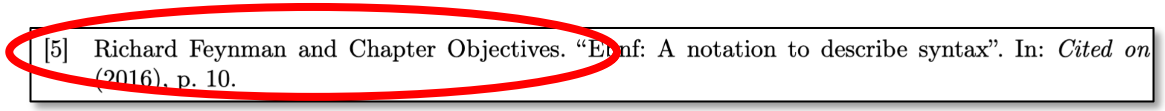
Interesting Snippets

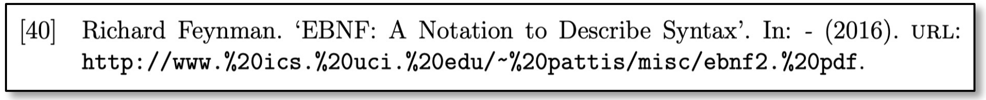
While developing a manuscript for a textbook on the Ada programming language in the late 1980s, I wrote a chapter on EBNF and began teaching it on the "first" day of my CS-1 class: primarily as a microcosm of programming, but also as a practical tool for later describing the syntax of Ada. These 21 pages (less than 1/4 the size of the original Karel book) discuss the sequence, choice, option, repetition, and recursion control structures (along with "procedural" abstracton 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.

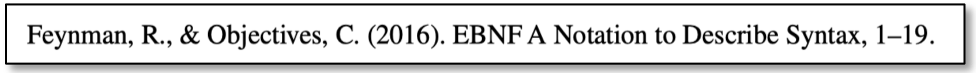
EBNF = "Extended Backus-Naur form"

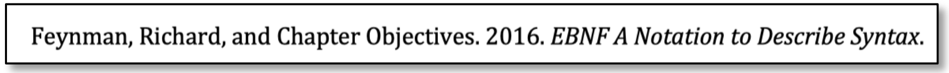
Website of Richard Pattis: https://ics.uci.edu/~pattis/ , EBNF document: https://ics.uci.edu/~pattis/misc/ebnf2.pdf

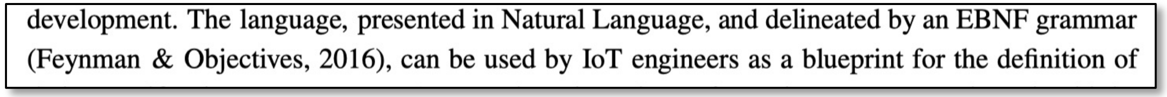
A story about citations

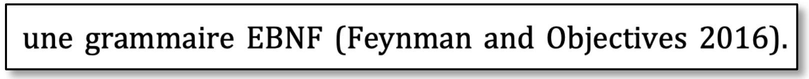

[5] Richard Feynman and Chapter Objectives. “Ebnf: A notation to describe syntax”. In: *Cited on* (2016), p. 10.

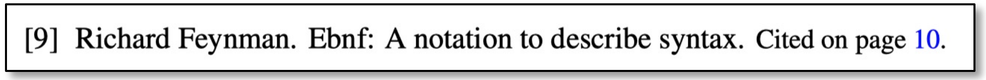

[40] Richard Feynman. ‘EBNF: A Notation to Describe Syntax’. In: - (2016). URL: `http://www.%20ics.%20uci.%20edu/~%20pattis/misc/ebnf2.%20pdf`.

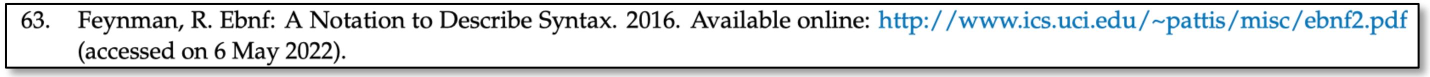

Feynman, R., & Objectives, C. (2016). EBNF A Notation to Describe Syntax, 1–19.

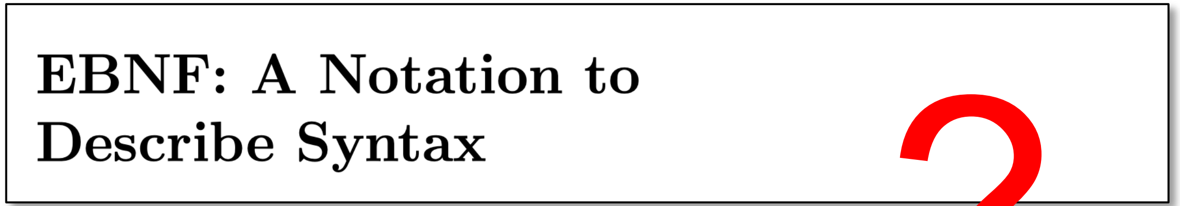

Feynman, Richard, and Chapter Objectives. 2016. *EBNF A Notation to Describe Syntax.*


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

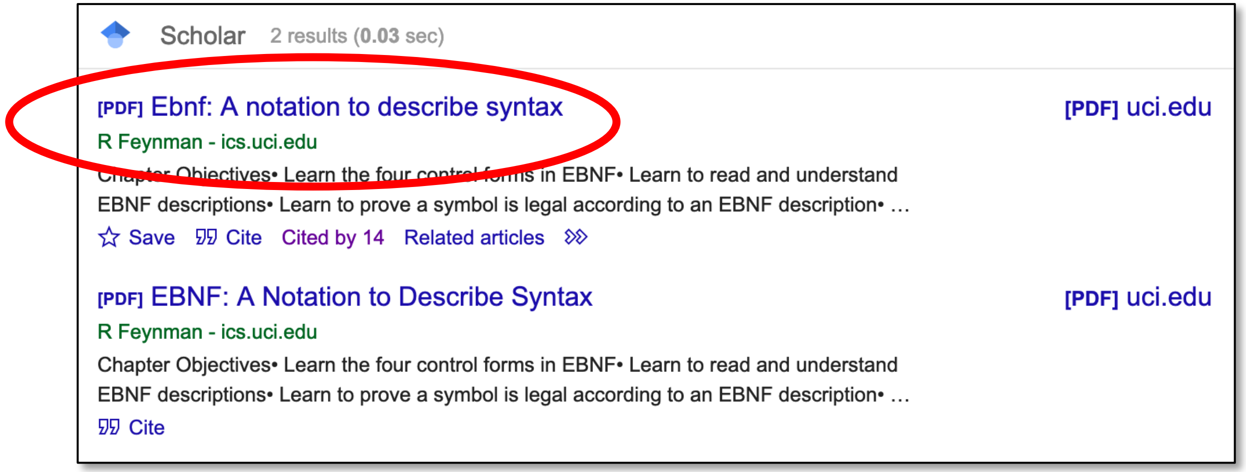

une grammaire EBNF (Feynman and Objectives 2016).


[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• …

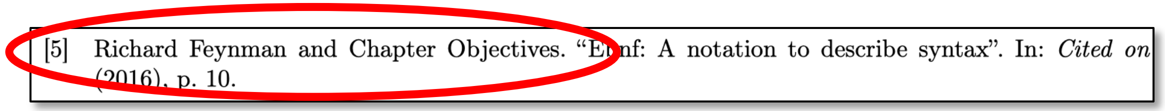
[PDF] EBNF: A Notation to Describe Syntax — [PDF] uci.edu

R Feynman - ics.uci.edu

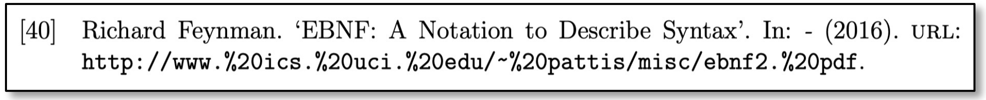
Chapter Objectives• Learn the four control forms in EBNF• Learn to read and understand EBNF descriptions• Learn to prove a symbol is legal according to an EBNF description• …

EBNF document: https://ics.uci.edu/~pattis/misc/ebnf2.pdf , Google citations: https://scholar.google.com/scholar?cluster=13515745032952198904

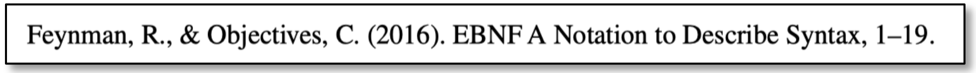
A story about citations



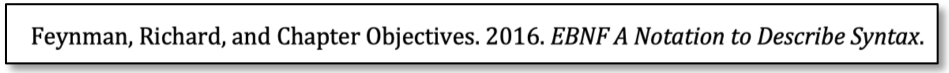
[5] Richard Feynman and Chapter Objectives. “Ebnf: A notation to describe syntax”. In: *Cited on* (2016), p. 10.



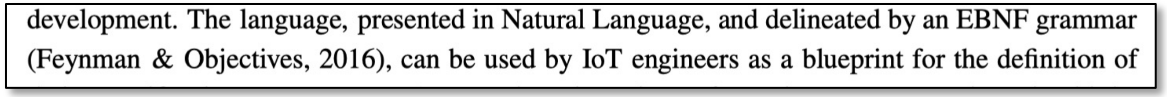
[40] Richard Feynman. ‘EBNF: A Notation to Describe Syntax’. In: - (2016). URL: `http://www.%20ics.%20uci.%20edu/~%20pattis/misc/ebnf2.%20pdf`.



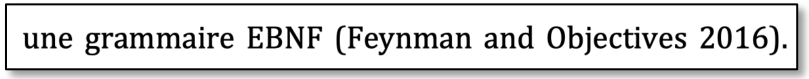
Feynman, R., & Objectives, C. (2016). EBNF A Notation to Describe Syntax, 1–19.



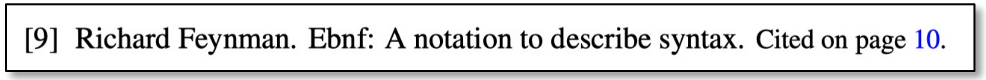
Feynman, Richard, and Chapter Objectives. 2016. *EBNF A Notation to Describe Syntax*.



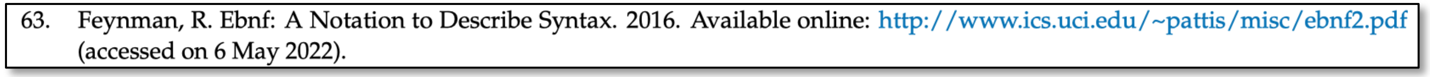
development. The language, presented in Natural Language, and delineated by an EBNF grammar (Feynman & Objectives, 2016), can be used by IoT engineers as a blueprint for the definition of



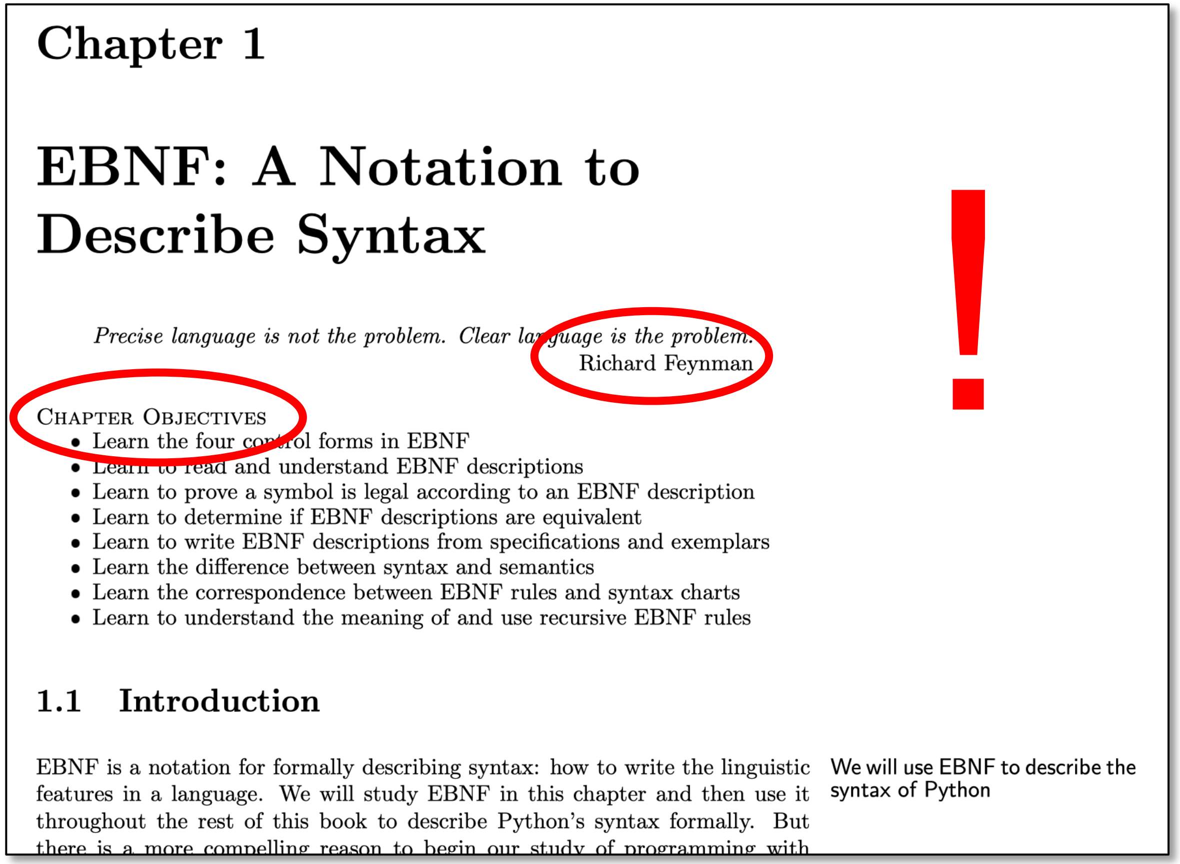
une grammaire EBNF (Feynman and Objectives 2016).



[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

EBNF document: https://ics.uci.edu/~pattis/misc/ebnf2.pdf , Google citations: https://scholar.google.com/scholar?cluster=13515745032952198904

Tools

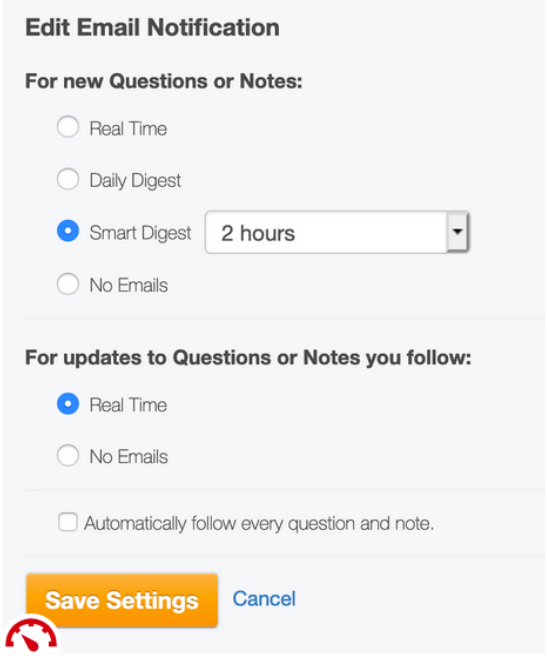
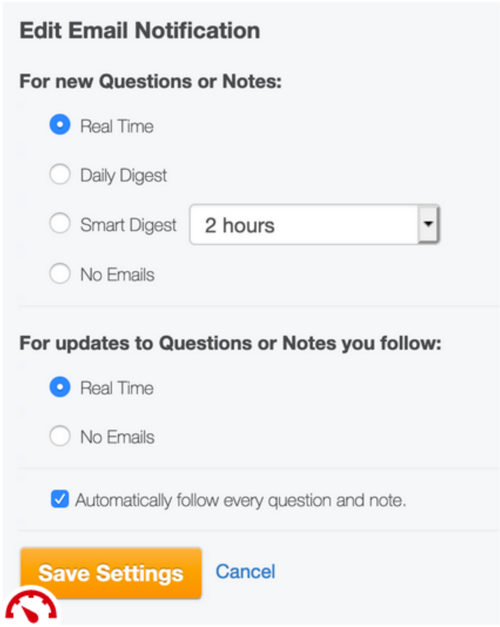
- Canvas:
  - Links to website: with preliminary calendar, optional readings, administrative details, lectures slides (will be posted by end of the 2nd day following a lecture, i.e. WED for MON classes, SAT for THU classes!)
  - Links to Piazza: discussions, questions, errors, follow-up instructions beyond web page! Make sure to subscribe
  - Canvas calendar / assignments: project milestones, submission for scribes (please look at the deadlines)

- 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
 **Edit Email Notification** **For new Questions or Notes:** ○ Real Time ○ Daily Digest ● Smart Digest 2 hours ○ No Emails **For updates to Questions or Notes you follow:** ● Real Time ○ No Emails □ Automatically follow every question and note. 	 **Edit Email Notification** **For new Questions or Notes:** ● Real Time ○ Daily Digest ○ Smart Digest 2 hours ○ No Emails **For updates to Questions or Notes you follow:** ● Real Time ○ No Emails ☑ Automatically follow every question and note.

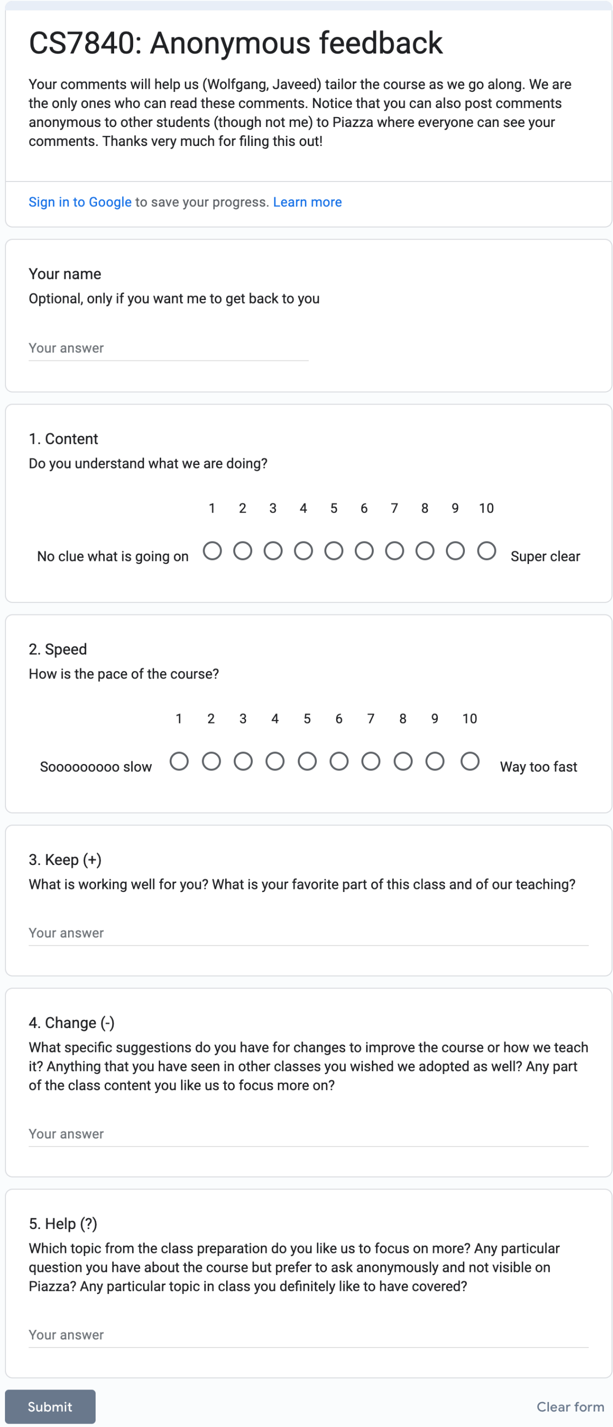
Feedback throughout the semester

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.



CS7840: 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 anonymous to other students (though not me) to Piazza where everyone can see your comments. Thanks very much for filing 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

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! I 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 can work together on everything else in the class!

Questions

1. Class is 11:45-1:25. 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 bascic 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

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

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

  A B C D

- Alice sends "messages" = words (strings/sequences) of symbols:

  A A C B C D A B B

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

  000010011011000101

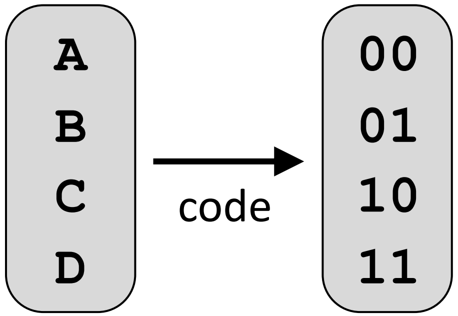
- What is a "reasonable" way to encode the symbols? ?

Compressing messages

- Assume Alice communicates with Bob about 4 symbols: **A B C D**
- Alice sends "messages" = words (strings/sequences) of symbols: **A | A | C | B | C | D | A | B | B**
- They only communicate in binary. All messages are strings of 0, 1: **00|00|10|01|10|11|00|01|01**
- A "reasonable" way to encode the symbols:

Can you decode following message: ?

0 0 0 1 1 0 1 1

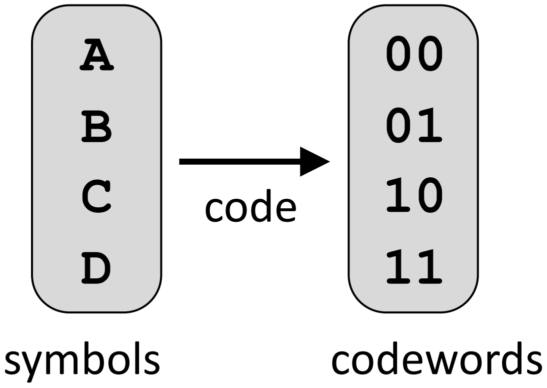


symbols
source alphabet
$\mathcal{X}$ = {A, B, C, D}

codewords
from $\mathcal{D}^*$, with $\mathcal{D}$={0,1}
as coding alphabet

Compressing messages

- Assume Alice communicates with Bob about 4 symbols: **A B C D**
- Alice sends "messages" = words (strings/sequences) 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:



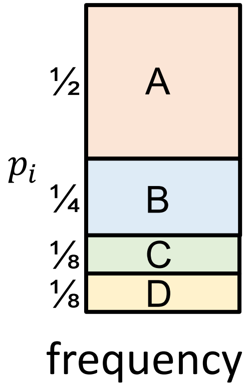
Example transmission:

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.

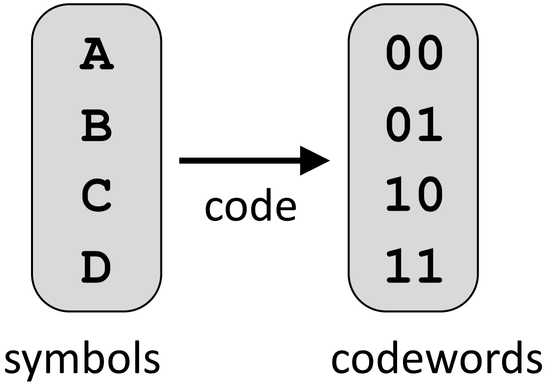
Compressing messages

- Assume we have the following symbol frequency:



- A "reasonable" way to encode the symbols:

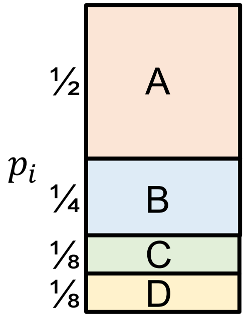
What is our expected message length per symbol? ?



symbols	codewords
A	00
B	01
C	10
D	11

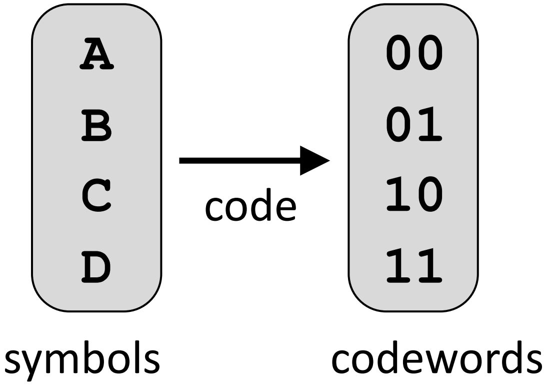
Compressing messages

- Assume we have the following symbol frequency:



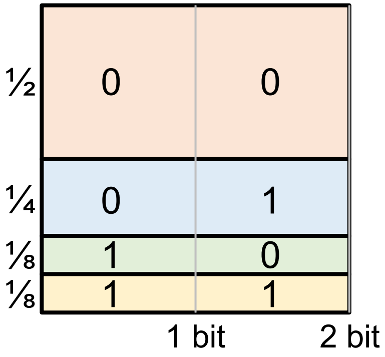
frequency

- A "reasonable" way to encode the symbols:



symbols	codewords
A	00
B	01
C	10
D	11

Our expected message length per symbol:



	1 bit	2 bit
½	0	0
¼	0	1
⅛	1	0
⅛	1	1

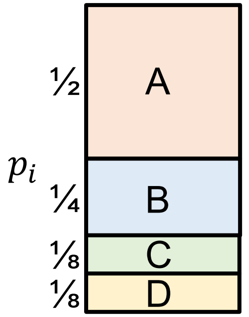
Encoding size

2 bits!

(does not depend on frequency, with current codewords)

Compressing messages

- Assume we have the following symbol frequency:

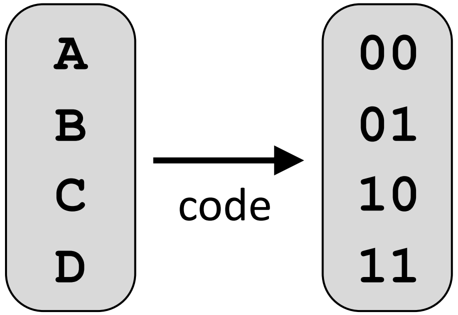


frequency

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

?

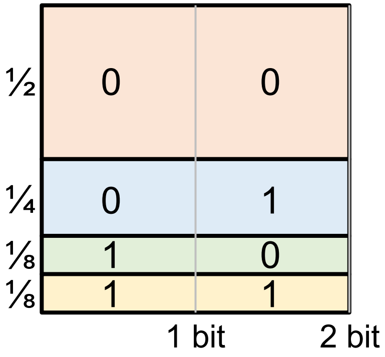
- A "reasonable" way to encode the symbols:



symbols

codewords

Our expected message length per symbol:

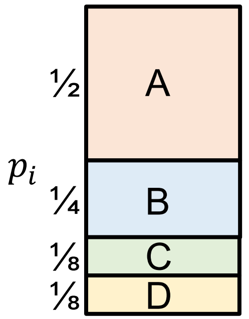


Encoding size

2 bits!

Compressing messages

- Assume we have the following symbol frequency:



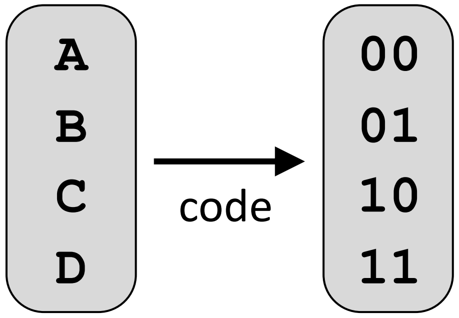
frequency

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

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

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

- A "reasonable" way to encode the symbols:

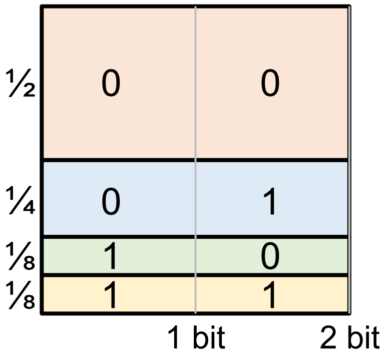


A → 00
B → 01
C → 10
D → 11

code

symbols codewords

Our expected message length per symbol:

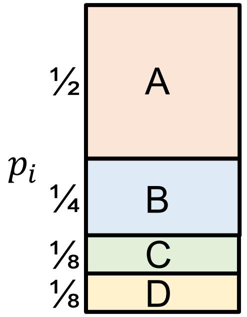


Encoding size

2 bits!

Compressing messages

- Assume we have the following symbol frequency:



frequency

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

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

-(½ · -1 +

¼ · -2 +

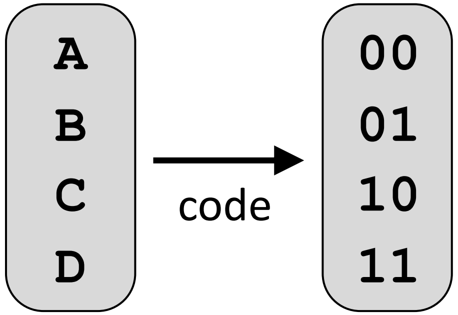
⅛ · -3 +

⅛ · -3)

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

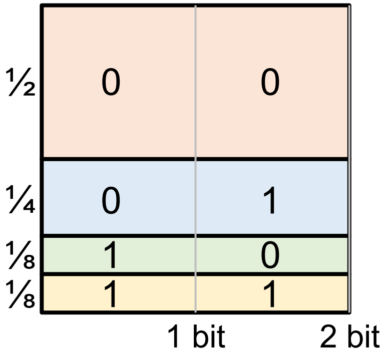
Can we match that "bound" w/ some encoding?

- A "reasonable" way to encode the symbols:



symbols		codewords
A	code	00
B		01
C		10
D		11

Our expected message length per symbol:



½	0	0
¼	0	1
⅛	1	0
⅛	1	1
	1 bit	2 bit

Encoding size

2 bits!

Compressing messages via variable length codes

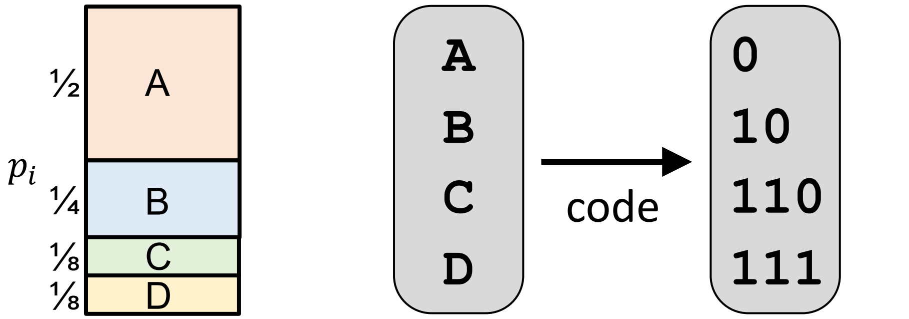
- Assume we have the following symbol frequency:

How can we decode that ?

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

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

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



frequency

symbols

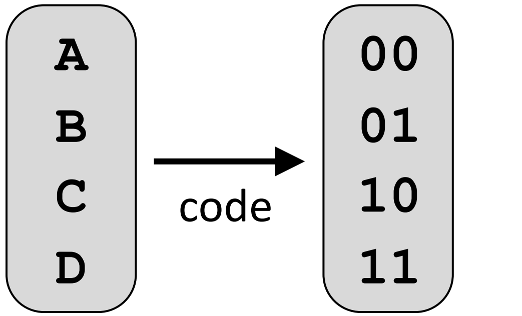
codewords

Intuition: more frequent stuff should use less space!

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

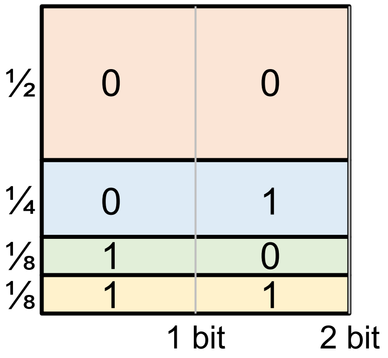
- A "reasonable" way to encode the symbols:

Our expected message length per symbol:



symbols

codewords



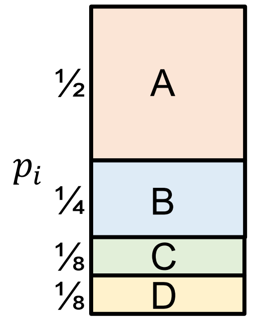
Encoding size

2 bits!

Compressing messages via variable length codes

0 1 1 1 1 1 0 0

- Assume we have the following symbol frequency:



frequency

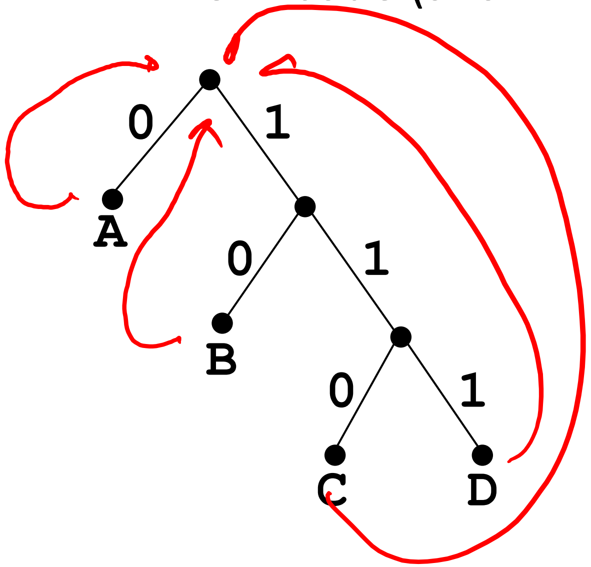
symbols	code	codewords
A	→	0
B		10
C		110
D		111

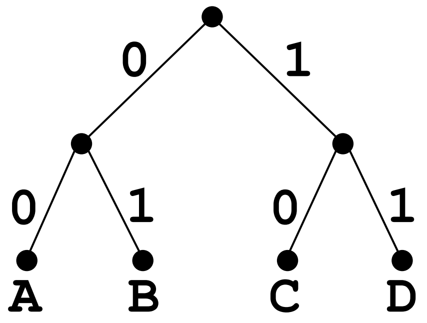
1 1 0 | 1 0 | 0
C B A

- A "reasonable" way to encode the symbols:

symbols	code	codewords
A	→	00
B		01
C		10
D		11

Prefix code (shown via binary prefix trees)





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

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

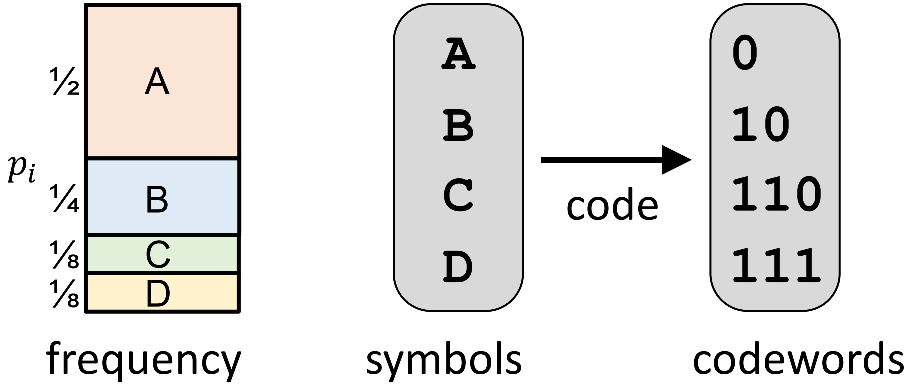
- Assume we have the following symbol frequency:

What is the new expected length? ?

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

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

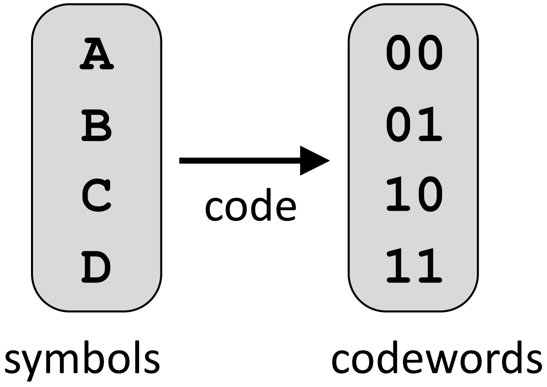
$\lg\left(\frac{1}{8}\right) = -3$

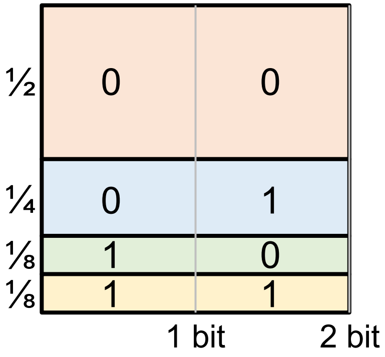


Entropy $\mathrm{H}(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i)$ = 1.75 bits!

- A "reasonable" way to encode the symbols:

Our expected message length per symbol:



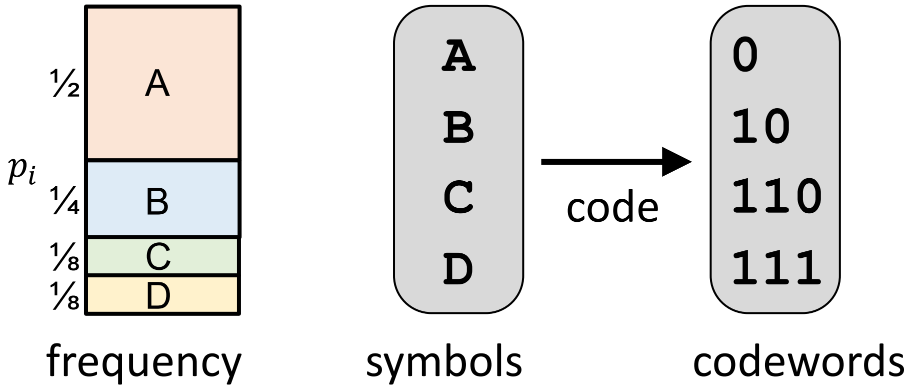


Encoding size

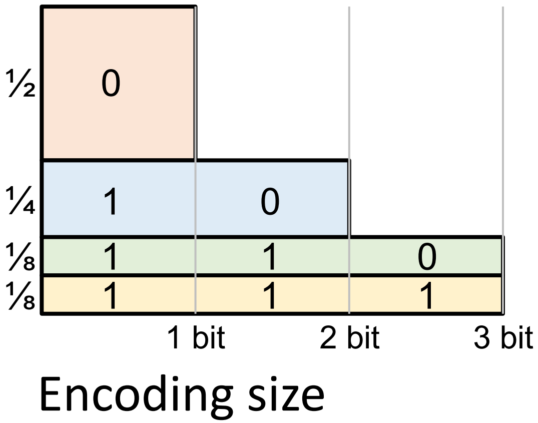
2 bits!

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length:



½ · 1

¼ · 2 = 1.75 bits!

⅛ · 3

⅛ · 3

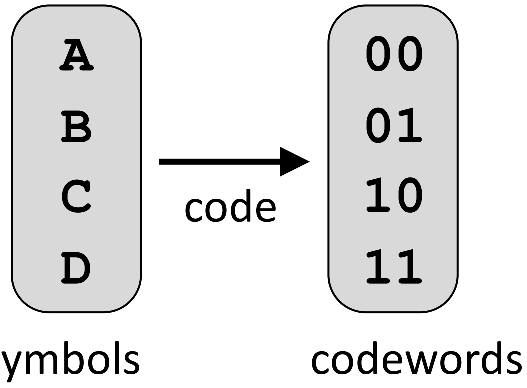
$\lg\left(\frac{1}{2}\right) = -1$

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

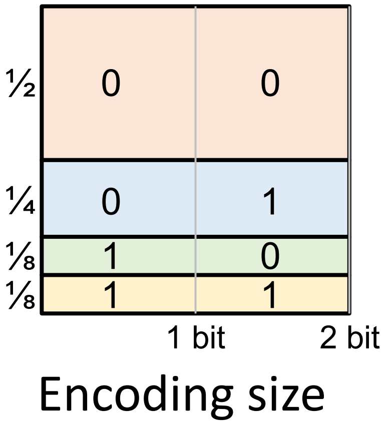
$\lg\left(\frac{1}{8}\right) = -3$

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

- A "reasonable" way to encode the symbols:



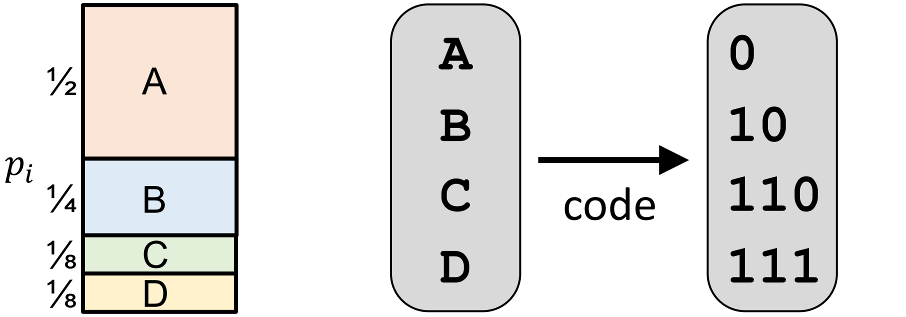
Our expected message length per symbol:



2 bits!

Compressing messages via variable length codes

- Assume we have the following symbol frequency:

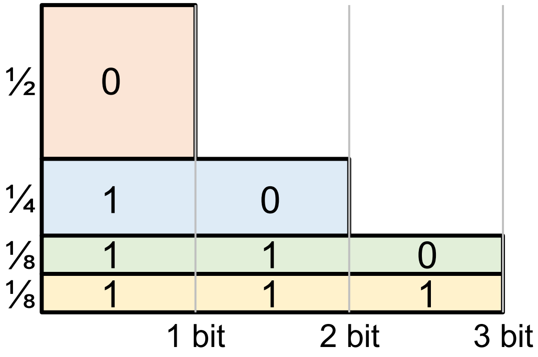


p_i

frequency	symbols		codewords
½	A	code	0
¼	B		10
⅛	C		110
⅛	D		111

frequency symbols codewords

New expected length :



½	0		
¼	1	0	
⅛	1	1	0
⅛	1	1	1
	1 bit	2 bit	3 bit

Encoding size

½ · 1

¼ · 2 = 1.75 bits!

⅛ · 3

⅛ · 3

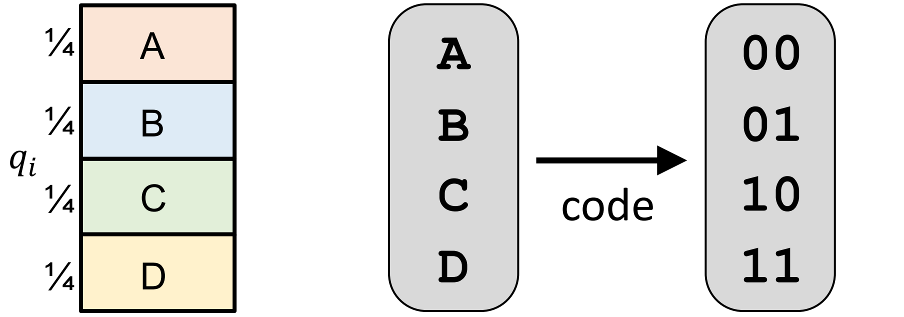
$\lg\left(\frac{1}{2}\right) = -1$

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

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

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

- Another interpretation: this is our assumed distribution!

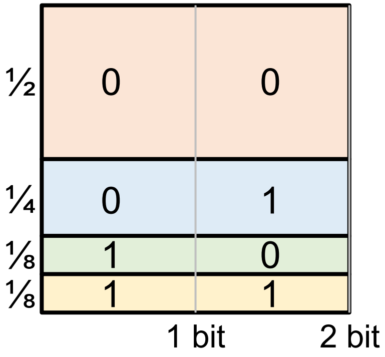


q_i

frequency	symbols		codewords
¼	A	code	00
¼	B		01
¼	C		10
¼	D		11

frequency symbols codewords

Our expected message length per symbol:



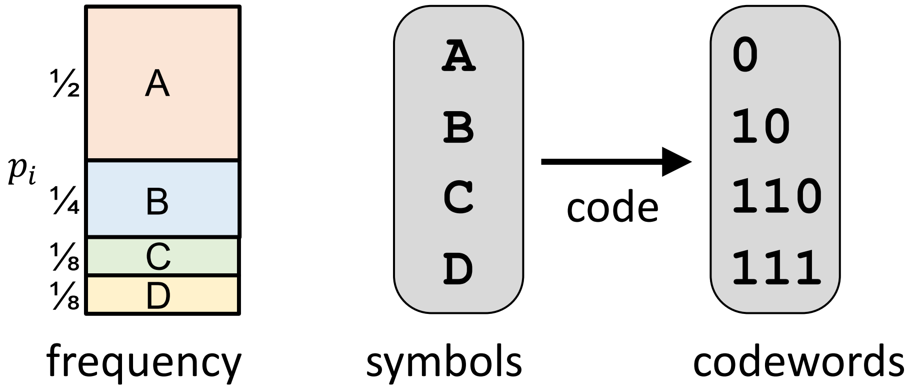
½	0	0
¼	0	1
⅛	1	0
⅛	1	1
	1 bit	2 bit

Encoding size

2 bits!

Compressing messages via variable length codes

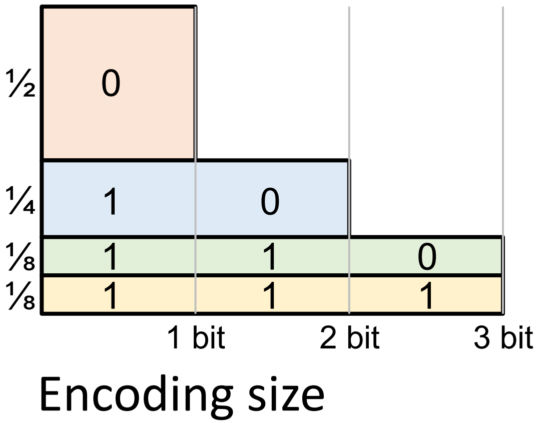
- Assume we have the following symbol frequency:



p_i

frequency		symbols		codewords
½	A	A	code	0
¼	B	B		10
⅛	C	C		110
⅛	D	D		111

New expected length :



Encoding size

½ · 1

¼ · 2 = 1.75 bits!

⅛ · 3

⅛ · 3

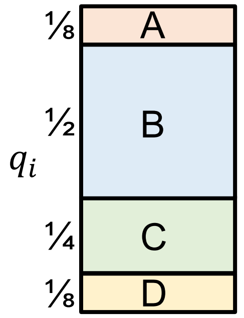
$\lg\left(\frac{1}{2}\right) = -1$

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

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

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

- What if we assume following distribution:



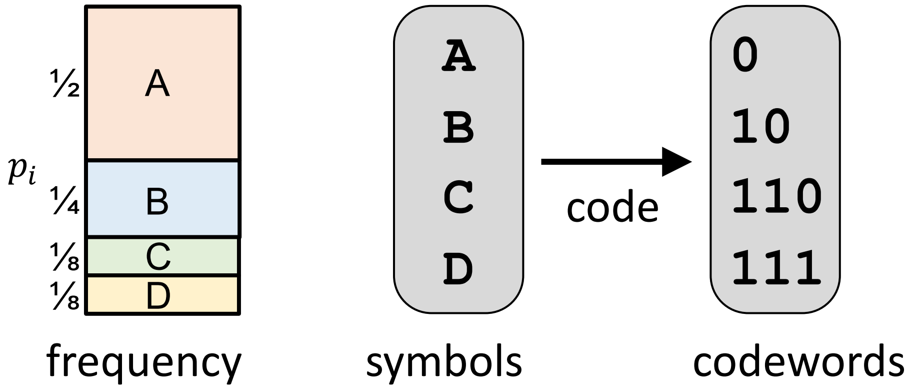
frequency

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

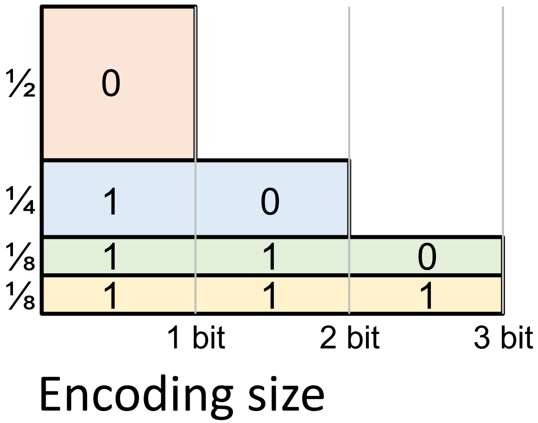
?

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



½ · 1

¼ · 2 = 1.75 bits!

⅛ · 3

⅛ · 3

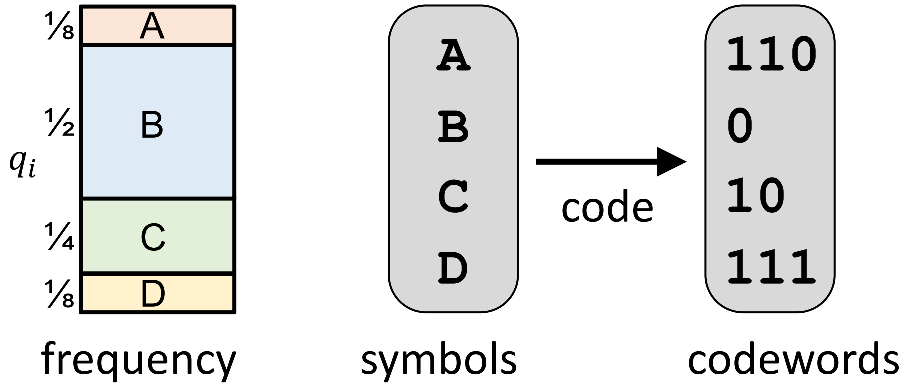
$\lg\left(\frac{1}{2}\right) = -1$

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

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

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

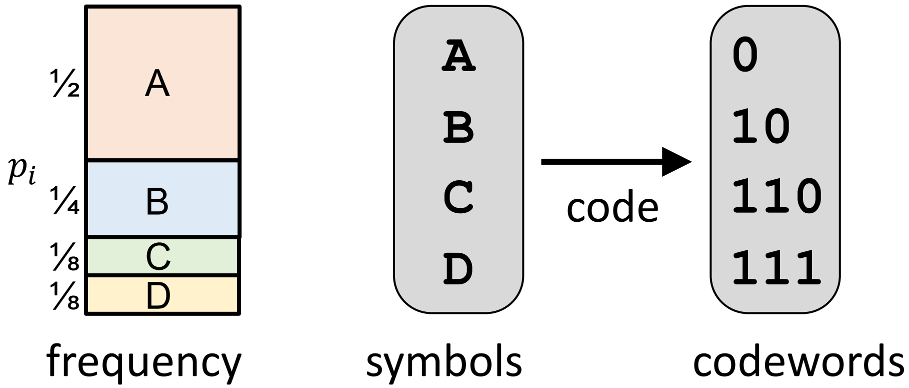
- What if we assume following distribution:



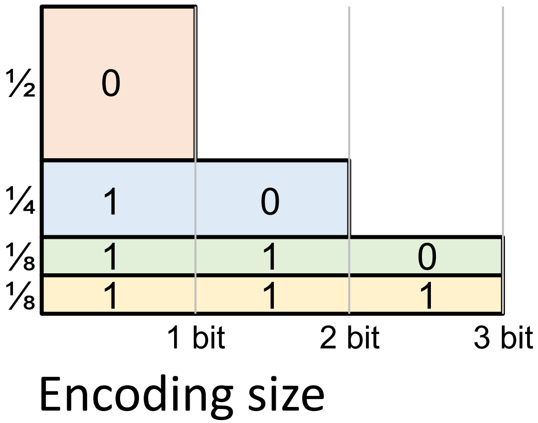
What is our expected message length per symbol if we use that code, but p is the actual distribution ?

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



½ · 1

¼ · 2 = 1.75 bits!

⅛ · 3

⅛ · 3

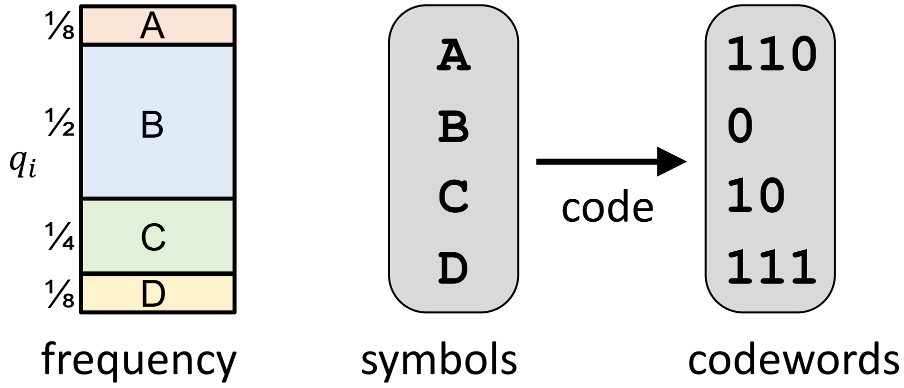
$\lg\left(\frac{1}{2}\right) = -1$

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

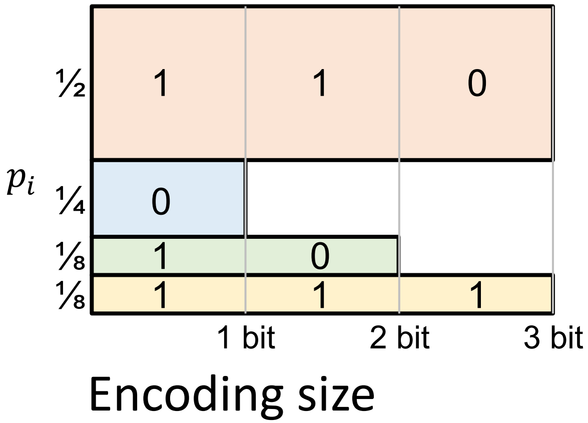
$\lg\left(\frac{1}{8}\right) = -3$

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

- What if we assume following distribution:

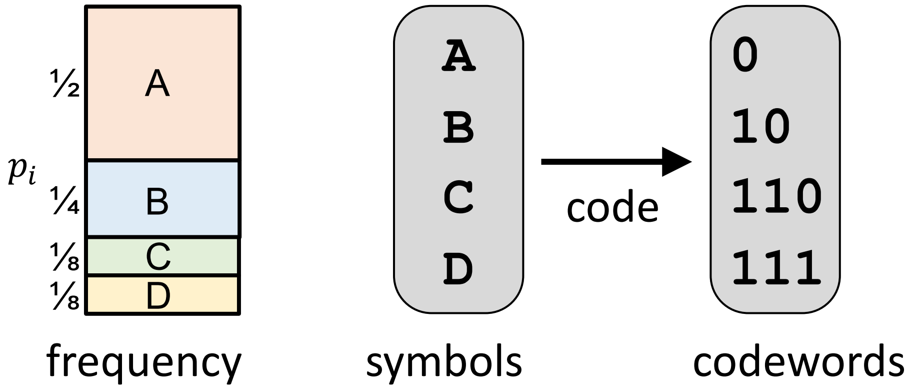


Our new expected message length per symbol:



Compressing messages via variable length codes

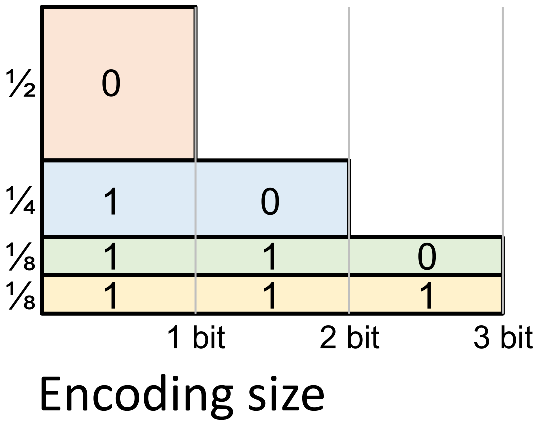
- Assume we have the following symbol frequency:



p_i

frequency	symbols	codewords
½	A	0
¼	B	10
⅛	C	110
⅛	D	111

New expected length :



½ · 1

¼ · 2

⅛ · 3

⅛ · 3

= 1.75 bits!

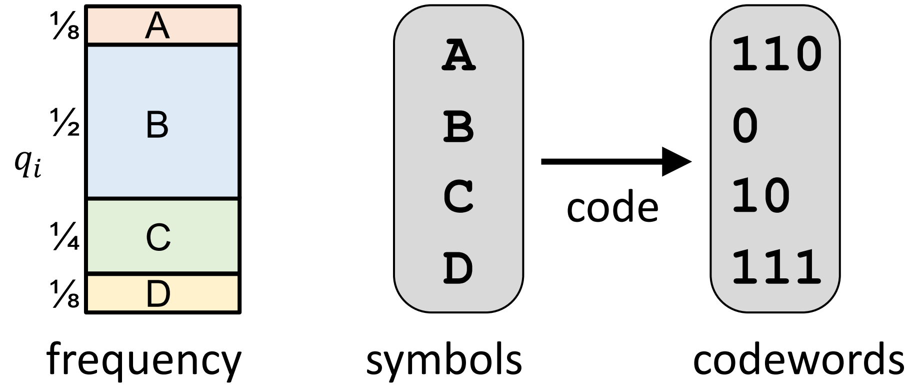
$\lg\left(\frac{1}{2}\right) = -1$

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

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

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

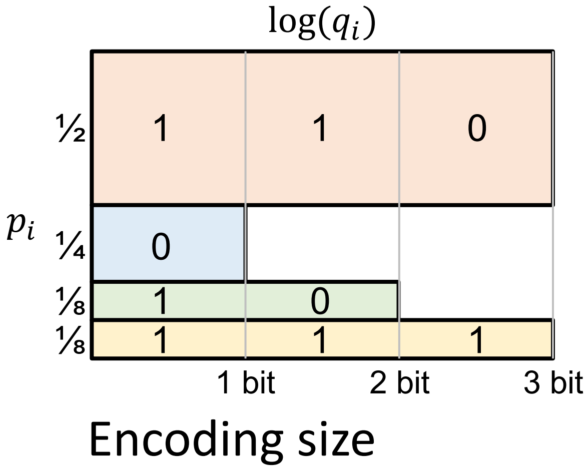
- What if we assume following distribution:



q_i

frequency	symbols	codewords
⅛	A	110
½	B	0
¼	C	10
⅛	D	111

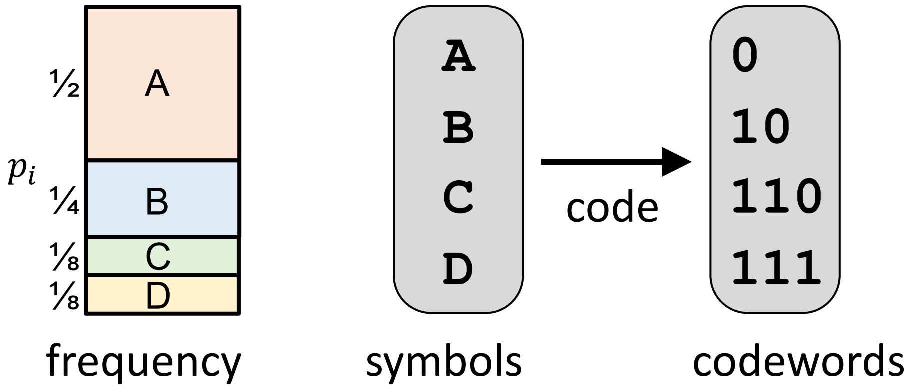
Our new expected message length per symbol:



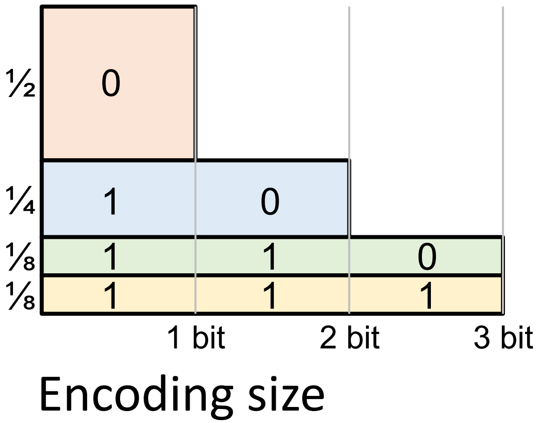
What is the formula we need to evaluate ?

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



½ · 1

¼ · 2

⅛ · 3

⅛ · 3

= 1.75 bits!

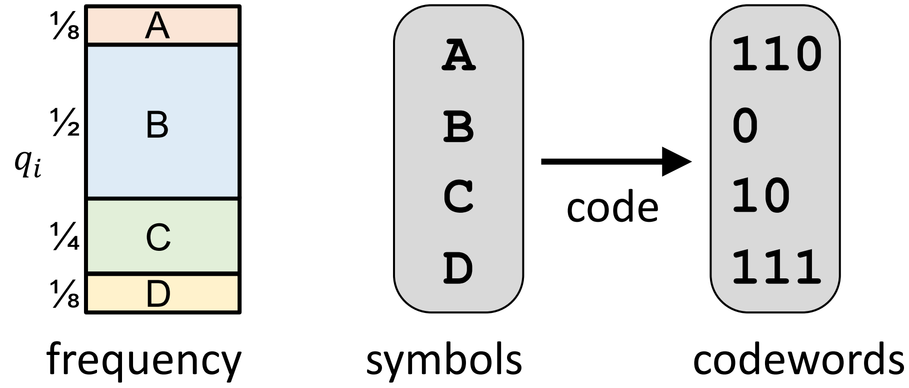
$\lg\left(\frac{1}{2}\right) = -1$

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

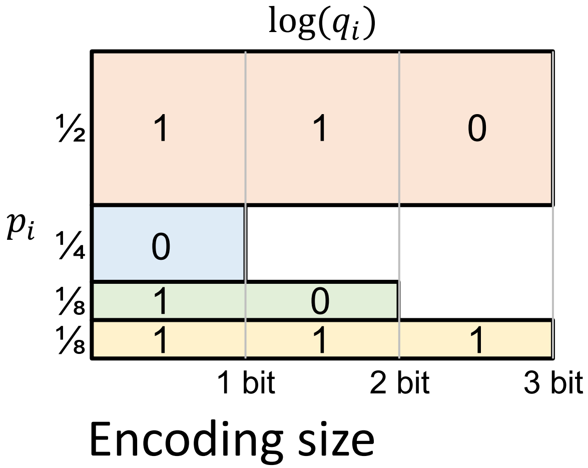
$\lg\left(\frac{1}{8}\right) = -3$

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

- What if we assume following distribution:



Our new expected message length per symbol:



= 2.375 bits!

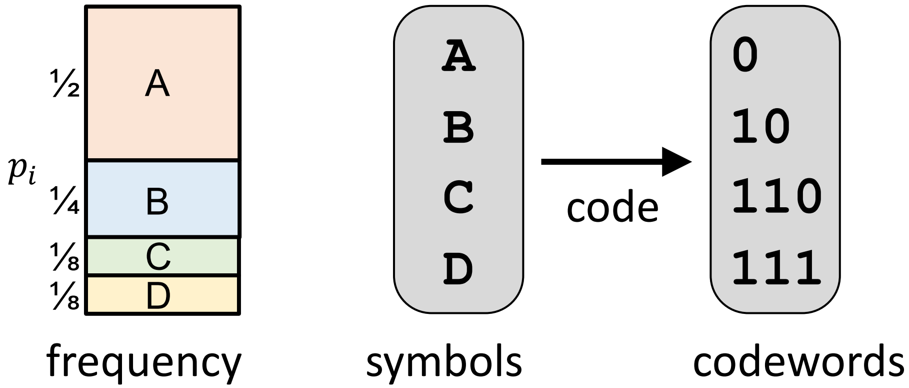
$-\sum_i p_i \cdot \lg(q_i)$

What is this formula called

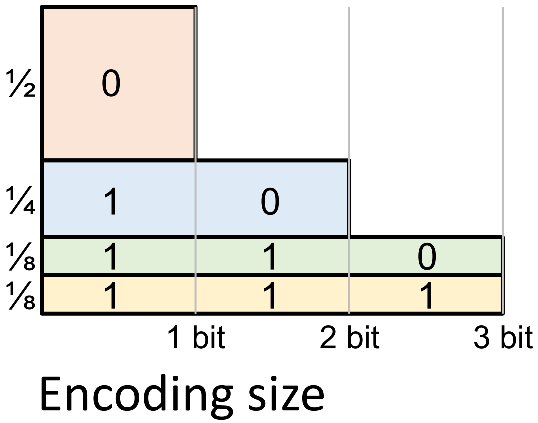
?

Compressing messages via variable length codes

- Assume we have the following symbol frequency:



New expected length :



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

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

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

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

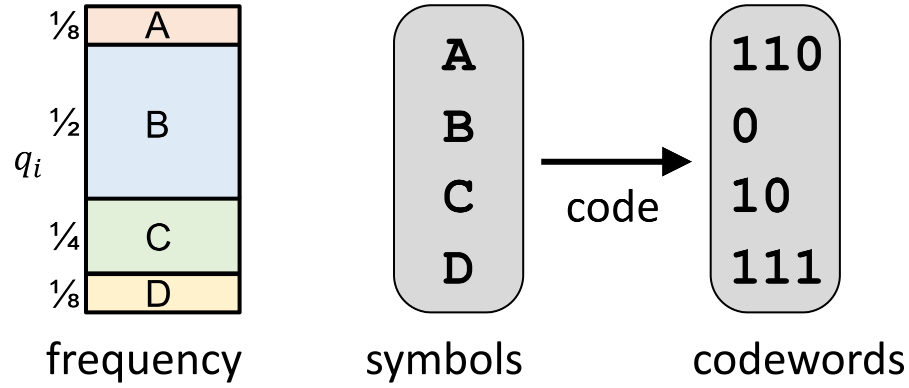
$\lg\left(\frac{1}{2}\right) = -1$

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

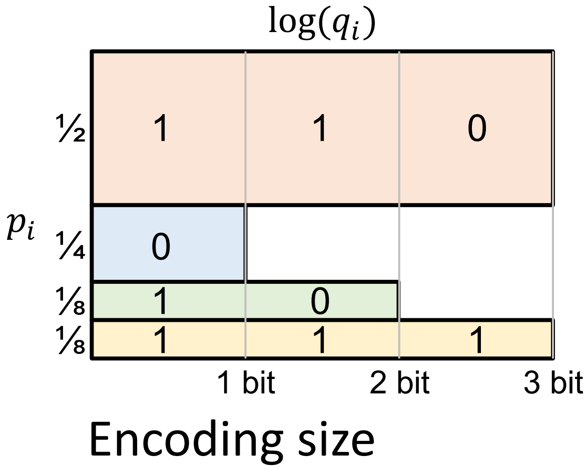
$\lg\left(\frac{1}{8}\right) = -3$

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

- What if we assume following distribution:



Our new expected message length per symbol:



= 2.375 bits!

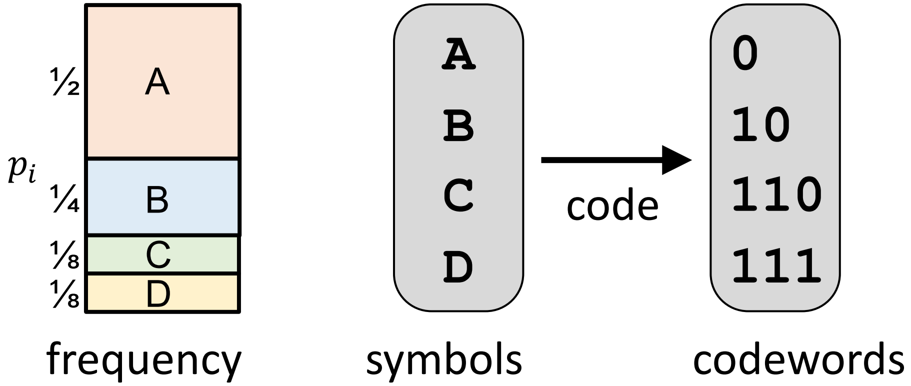
$-\sum_i p_i \cdot \lg(q_i)$

Cross entropy $H(\mathbf{p}||\mathbf{q})$ ☺

Which distribution $\mathbf{q}$ minimizes $H(\mathbf{p}||\mathbf{q})$?

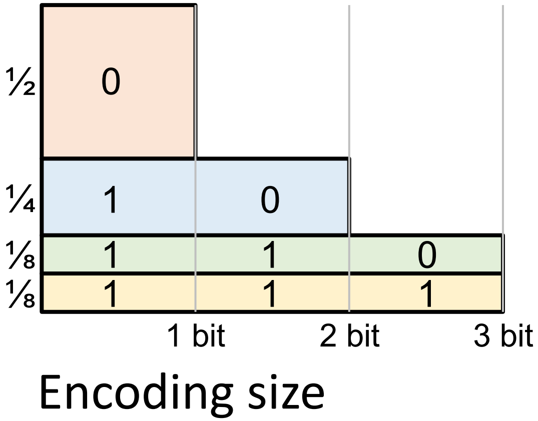
Compressing messages via variable length codes

- Assume we have the following symbol frequency:



symbols		codewords
A	code →	0
B		10
C		110
D		111

New expected length :



½ · 1

¼ · 2 = 1.75 bits!

⅛ · 3

⅛ · 3

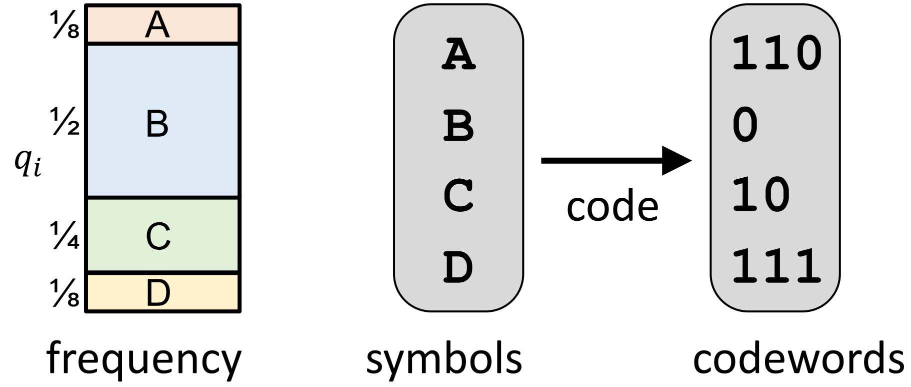
$\lg\left(\frac{1}{2}\right) = -1$

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

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

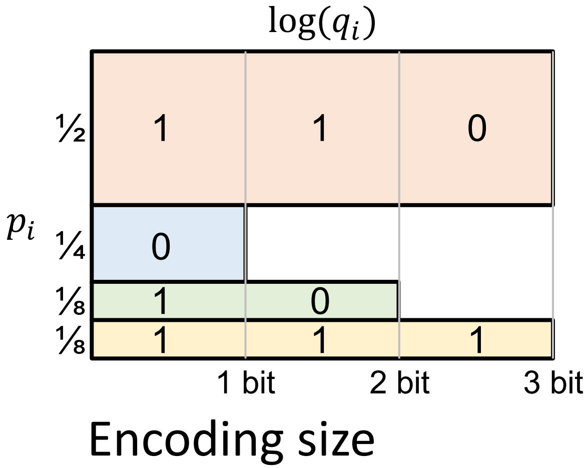
Entropy $\mathrm{H}(\mathbf{p}) := -\sum_i p_i \cdot \lg(p_i)$ = 1.75 bits!

- What if we assume following distribution:



symbols		codewords
A	code →	110
B		0
C		10
D		111

Our new expected message length per symbol:



= 2.375 bits!

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

Cross entropy $H(\mathbf{p}||\mathbf{q})$ ☺

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

= entropy $H(\mathbf{p})$