Updated 12/2/2024

Part 3: Applications
L23: Information Bottleneck Theory

Wolfgang Gatterbauer, Javed Aslam

cs7840 Foundations and Applications of Information Theory (fa24)

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

12/2/2024

Pre-class conversations

- Projects!
- Intended Topics & Feedback

- **Lecture 20 (Mon 11/18):**
  Channel capacity [Cover Thomas'06: Ch 7]
- **Lecture 21 (Wed 11/20):**
  Distortion Theory (1/2) [Cover Thomas'06: Ch 10]
- **Lecture 22 (Mon 11/25):**
  Distortion Theory (2/2) [Cover Thomas'06: Ch 10]
  Python notebooks: 232
- **(Wed 11/27): no class (Fall break)**
- **Lecture 23 (Mon 12/2):** Information Bottleneck Theory
- **Lecture 24 (Wed 12/4):** Information Bottleneck Theory

Project presentations

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

- Rate Distortion & Information bottleneck theory
  - [Cover,Thomas'06] Elements of Information Theory. 2nd ed, 2006: Ch 10 Rate distortion theory
  - [Tishby+'99] Tishby, Pereira, Bialek. The information bottleneck method. The 37th annual Allerton Conference on Communication, Control, and Computing. pp. 368–377.
  - [Harremoes,Tishby'07] The Information Bottleneck Revisited or How to Choose a Good Distortion Measure. International Symposium on Information Theory, 2007.
  - [Zaslavsky+'18] Zaslavsky, Kemp, Regier, Tishby. The Efficient compression in color naming and its evolution. PNAS, 2018.
  - [Webb+'24] Webb, Frankland, Altabaa, Segert, Krishnamurthy, Campbell, Russin, Giallanza, Dulberg, OReilly, Lafferty, Cohen. The Relational Bottleneck as an Inductive Bias for Efficient Abstraction. Trends in Cognitive Science, 2024.
  - [Segert'24] Maximum Entropy, Symmetry, and the Relational Bottleneck: Unraveling the Impact of Inductive Biases on Systematic Reasoning. PhD thesis, Neuroscience @ Princeton, 2024.
  - [Ren,Li,Leskovec'20] Graph Information Bottleneck, NeurIPS, 2020.

- Today:
  - Information Bottleneck Theory (1/2)

Information Bottleneck

Three-step abstractions

$X \longrightarrow Y \longrightarrow Z$ Markov chain

What do we know? ?

Three-step abstractions

$X \longrightarrow Y \longrightarrow Z$ Markov chain $X \perp Z|Y$ $I(X;Y) \geq I(X;Z)$

$p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|\cancel{x},y)$ also: $p(y) \cdot p(x|y) \cdot p(z|\cancel{x},y)$

$X \longrightarrow Y \longrightarrow f(Y)$ Data processing inequality

?

Three-step abstractions

$X \longrightarrow Y \longrightarrow Z$ Markov chain $X \perp Z|Y$ $I(X;Y) \geq I(X;Z)$

$p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|\cancel{x},y)$ also: $p(y) \cdot p(x|y) \cdot p(z|\cancel{x},y)$

$X \longrightarrow Y \longrightarrow f(Y)$ Data processing inequality $I(X;Y) \geq I\big(X;f(Y)\big)$

$\theta \longrightarrow \mathbf{X} \longrightarrow T(\mathbf{X})$ Sufficient statistics

?

Three-step abstractions

$X \longrightarrow Y \longrightarrow Z$ Markov chain $X \perp Z|Y$ $I(X;Y) \geq I(X;Z)$

$p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|\cancel{x},y)$ also: $p(y) \cdot p(x|y) \cdot p(z|\cancel{x},y)$

$X \longrightarrow Y \longrightarrow f(Y)$ Data processing inequality $I(X;Y) \geq I\big(X;f(Y)\big)$

$\theta \longrightarrow \mathbf{X} \longrightarrow T(\mathbf{X})$ Sufficient statistics

A statistic T is sufficient for θ if it preserves all the information in $\mathbf{X}$ about θ:

$\theta \perp \mathbf{X}|T(\mathbf{X}) \Leftrightarrow I\big(\theta;T(\mathbf{X})\big) = I(\theta;\mathbf{X}) \Leftrightarrow \theta \to T(\mathbf{X}) \to \mathbf{X}$ also forms a Markov chain

minimal sufficient: simplest mapping of $\mathbf{X}$ that captures all the information in $\mathbf{X}$ about θ:

We want to determine Y from X. Goal: find a representation $\hat{X}$ of X that captures the relevant features, yet compresses X by removing irrelevant parts that do not contribute to the prediction of Y

?

Three-step abstractions

$X \longrightarrow Y \longrightarrow Z$ Markov chain $X \perp Z|Y$ $I(X;Y) \geq I(X;Z)$

$p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|\cancel{x},y)$ also: $p(y) \cdot p(x|y) \cdot p(z|\cancel{x},y)$

$X \longrightarrow Y \longrightarrow f(Y)$ Data processing inequality $I(X;Y) \geq I\big(X;f(Y)\big)$

$\theta \longrightarrow \mathbf{X} \longrightarrow T(\mathbf{X})$ Sufficient statistics

A statistic T is sufficient for θ if it preserves all the information in $\mathbf{X}$ about θ:

$\theta \perp \mathbf{X}|T(\mathbf{X}) \Leftrightarrow I\big(\theta;T(\mathbf{X})\big) = I(\theta;\mathbf{X}) \Leftrightarrow \theta \to T(\mathbf{X}) \to \mathbf{X}$ also forms a Markov chain

minimal sufficient: simplest mapping of $\mathbf{X}$ that captures all the information in $\mathbf{X}$ about θ:

We want to determine Y from X. Goal: find a representation $\hat{X}$ of X that captures the relevant features "max $I\big(Y;\hat{X}\big)$", yet compresses X by removing irrelevant parts that do not contribute to the prediction of Y: "min $I\big(X;\hat{X}\big)$".

Three-step abstractions

$X \longrightarrow Y \longrightarrow Z$ Markov chain $\quad X \perp Z|Y \quad I(X;Y) \geq I(X;Z)$

$p(x,y,z) = p(x) \cdot p(y|x) \cdot p(z|\cancel{x},y)$ also: $p(y) \cdot p(x|y) \cdot p(z|\cancel{x},y)$

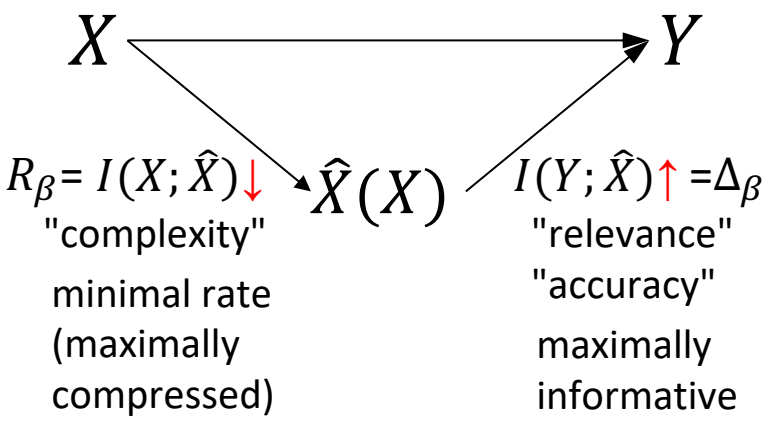
$X \longrightarrow Y \longrightarrow f(Y)$ Data processing inequality $\quad I(X;Y) \geq I\big(X;f(Y)\big)$

$\theta \longrightarrow \mathbf{X} \longrightarrow T(\mathbf{X})$ Sufficient statistics

A statistic T is sufficient for θ if it preserves all the information in $\mathbf{X}$ about θ:

$\theta \perp \mathbf{X}|T(\mathbf{X}) \Leftrightarrow I\big(\theta;T(\mathbf{X})\big) = I(\theta;\mathbf{X}) \Leftrightarrow \theta \to T(\mathbf{X}) \to \mathbf{X}$ also forms a Markov chain

minimal sufficient: simplest mapping of $\mathbf{X}$ that captures all the information in $\mathbf{X}$ about θ:



We want to determine Y from X. Goal: find a representation $\hat{X}$ of X that captures the relevant features "max $I\big(Y;\hat{X}\big)$", yet compresses X by removing irrelevant parts that do not contribute to the prediction of Y: "min $I\big(X;\hat{X}\big)$".

$$\mathcal{L}^* = \min_{p(\hat{X}|X)} [\mathcal{L}(\hat{X})] \qquad \mathcal{L}(\hat{X}) = I(X;\hat{X}) - \beta I(Y;\hat{X})$$

$$\mathcal{L}'^* = \max_{p(\hat{X}|X)} [\mathcal{L}'(\hat{X})] \qquad \mathcal{L}'(\hat{X}) = I(Y;\hat{X}) - \beta' I(X;\hat{X})$$

bigger β (smaller β') allows more complex representations

Information Bottleneck (IB)

Consider an information processing system that receives as input the signal X and tries to predict a target signal Y. We want to process X to get a compressed representation of the input $\hat{X} = f(X)$ (the "bottleneck"), which is then used to predict Y.

$\hat{X}$ is sufficient for predicting Y if it contains all the information that X encodes about Z, i.e. $I(Y;\hat{X}) = I(X;\hat{X})$.

$\hat{X}$ is minimal-sufficient if it is sufficient for Y and does not contain any extraneous information about X which does not help in predicting Y, i.e. $I(X;\hat{X}) \leq I(X;\hat{X}')$ for any other sufficient representation $\hat{X}'$.

The information bottleneck objective tries to strike a balance in achieving max compression (small complexity) while retaining as much relevant information (high accuracy) as possible

bigger β allows more complex representations

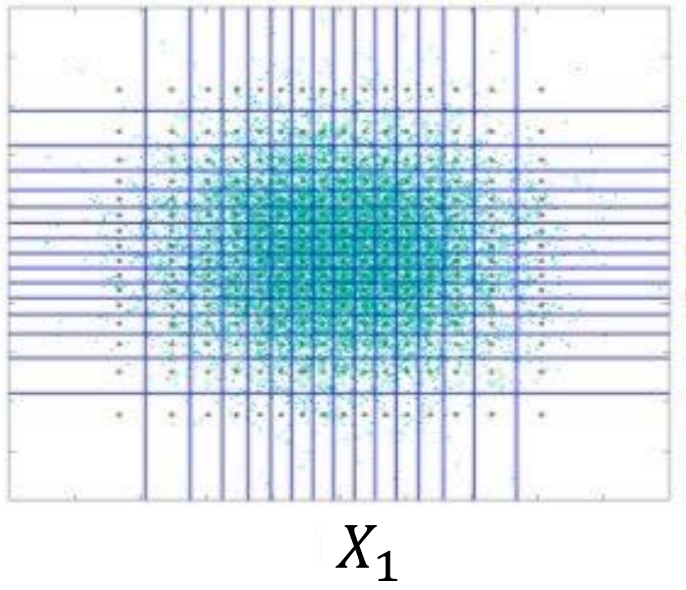
$$\text{minimize } \mathcal{L}(\hat{X}) = I(X;\hat{X}) - \beta I(Y;\hat{X})$$

$$\text{maximize } \mathcal{L}(\hat{X}) = I(Y;\hat{X}) - \beta' I(X;\hat{X})$$

bigger $\beta' = 1/\beta$ penalizes more complex representations

Geometry of longer block lengths:

Independent 4-bit quantization:



Blocklength $n = 2$ and 4-bit per sample

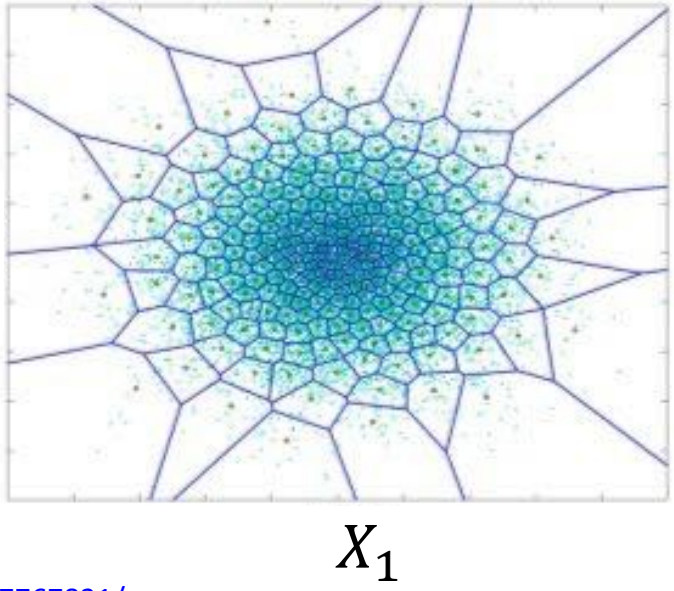
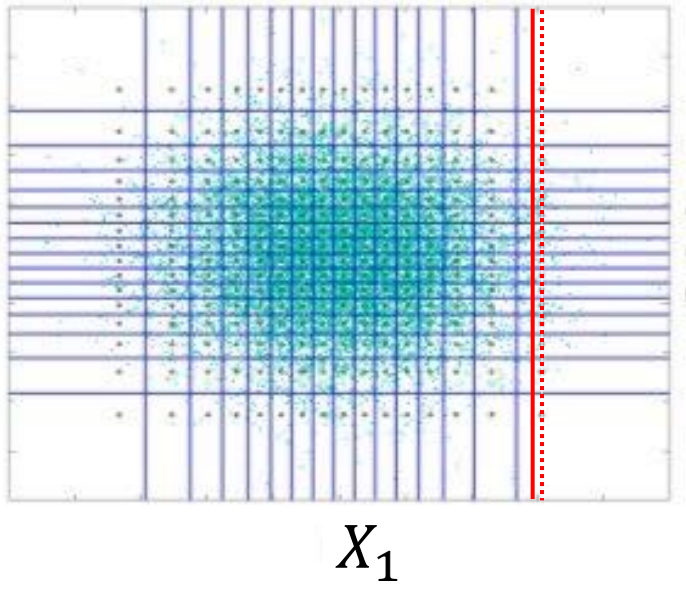


Figure source: https://ieeexplore.ieee.org/document/7767821/

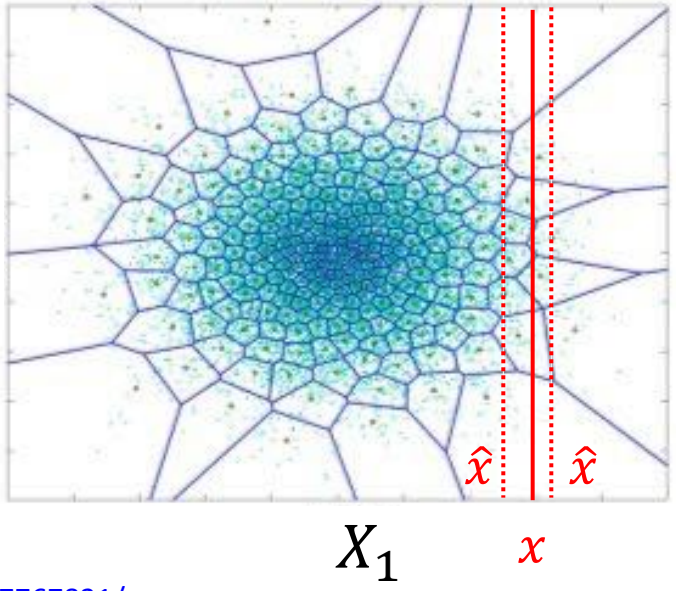
Geometry of longer block lengths:

Independent 4-bit quantization:



$p(\hat{X}|X)$... deterministic

Blocklength $n = 2$ and 4-bit per sample



$p(\hat{X}|X)$... stochastic

"It is simpler to describe an elephant and a chicken with one description than to describe each alone. This is true even for independent random variables."
[Cover, Thomas'06]

Figure source: https://ieeexplore.ieee.org/document/7767821/
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

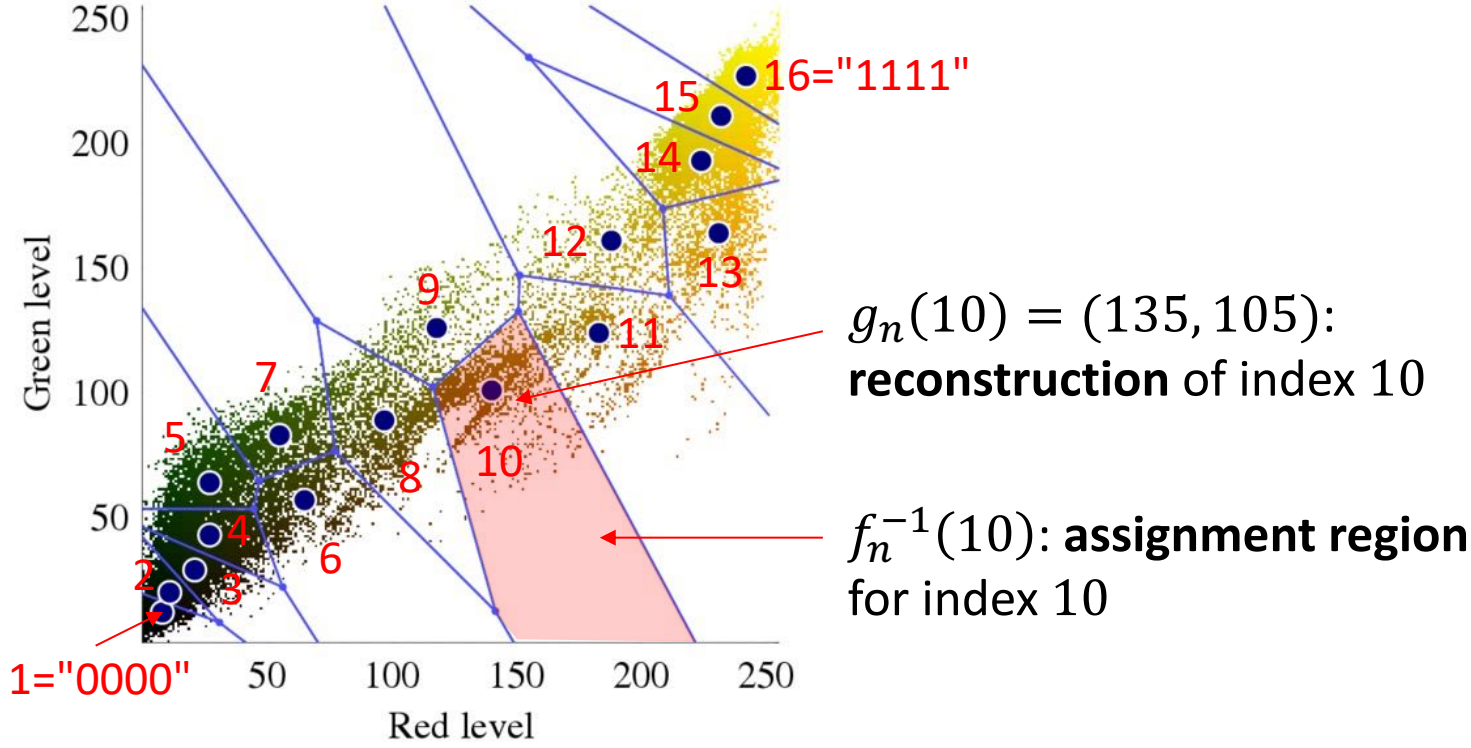
Rate-distortion code vs. k-means

$\mathcal{X} = \hat{\mathcal{X}} = \{0,1, \dots, 255\}$ thus 8 bit resolution

$n = 2$ channels per pixel (will be encoded together), 16 bits per pixel

$nR = 4$ bits per pixel (2 bits per channel level), thus 16 representatives

Example image with only red and green channel (for illustration)



Vector quantization of colors present in the image into Voronoi cells using *k*-means

Source: https://en.wikipedia.org/wiki/K-means_clustering

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

The Information Bottleneck (IB) method was introduced by Tishby et al. [1] as a method for extracting the information that some variable $X \in \mathcal{X}$ provides about another one $Y \in \mathcal{Y}$ that is of interest, as shown in Figure 1.

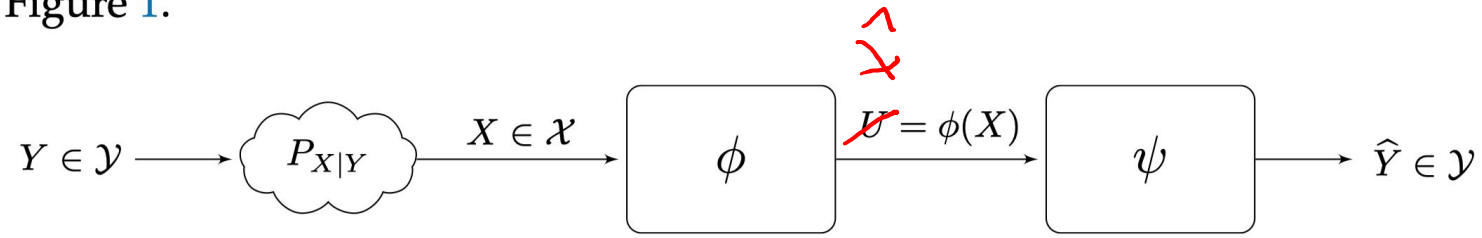


Figure 1. Information bottleneck problem.

$\hat{X} \longrightarrow X \longrightarrow Y$ Markov chain $\hat{X} \perp Y|X$

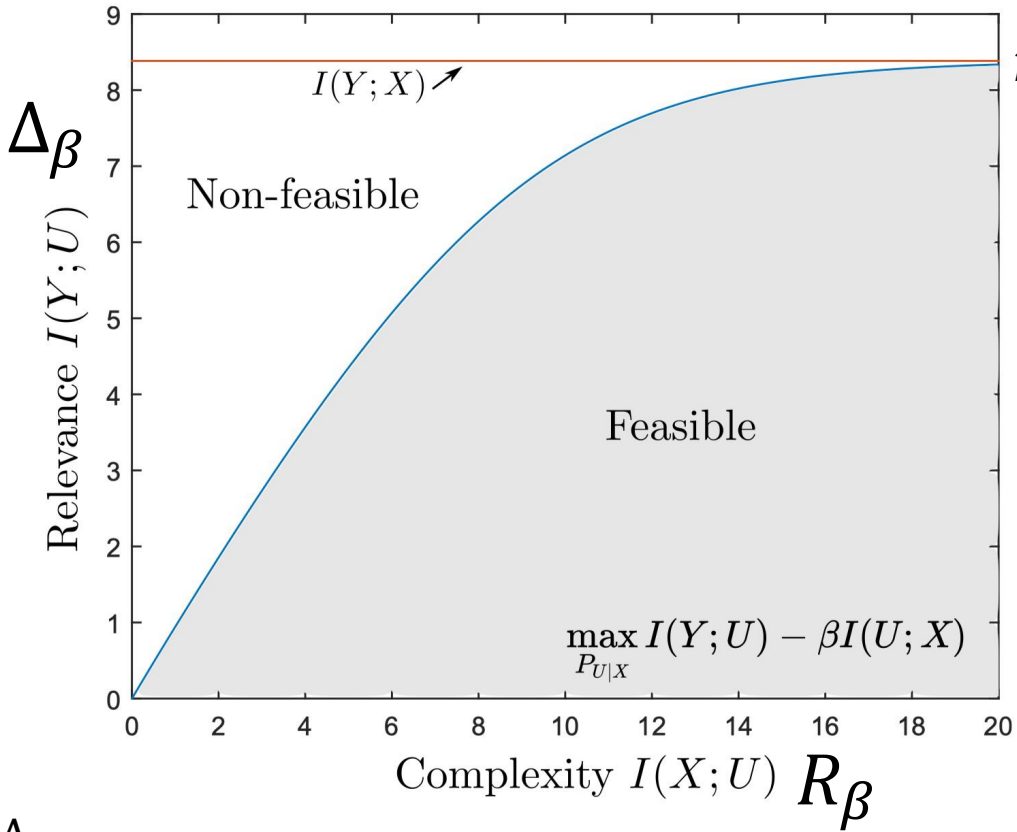
$$p(\hat{x}, x, y) = p(x) \cdot p(\hat{x}|x) \cdot p(y|\hat{x}, x)$$

$$\mathcal{L}'^{*} = \max_{\substack{p(\hat{X}|X) \\ \hat{X} \perp Y|X}} [\mathcal{L}'(\hat{X})] \qquad \mathcal{L}'(\hat{X}) = I(Y;\hat{X}) - \beta' I(X;\hat{X})$$

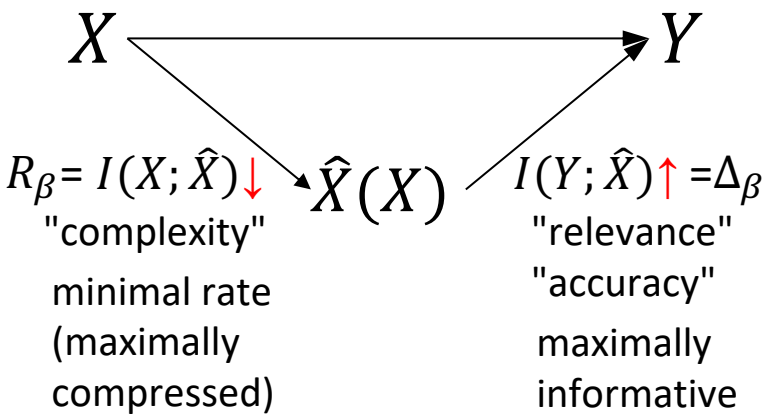
relevance Δ_β rate R_β

Optimization leads to optimal relevance-complexity pairs (Δ_β, R_β)

Source: Zidi, Estella-Aguerri, Shamai. On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views. Entropy, 2020. https://doi.org/10.3390/e22020151
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/



$p(x,y)$ is given



$R_\beta = I(X;\hat{X})\downarrow$
"complexity"
minimal rate
(maximally
compressed)

$I(Y;\hat{X})\uparrow = \Delta_\beta$
"relevance"
"accuracy"
maximally
informative

~~relevance–complexity pairs (Δ, R) that satisfy~~

~~$\Delta \le I(U,Y), \quad R \ge I(X,U)$~~

relevance-complexity pairs (Δ_β, R_β)

with complexity $I(X;\hat{X}) \ge R_\beta$ and relevance $I(Y;\hat{X}) \le \Delta_\beta$

Source: Zidi, Estella-Aguerri, Shamai. On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views. Entropy, 2020. https://doi.org/10.3390/e22020151
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Binary Information Bottleneck

Let X and Y be a doubly symmetric binary sources (DSBS), i.e., $(X, Y) \sim \text{DSBS}(p)$ for some $0 \le p \le 1/2$. (A DSBS is a pair (X, Y) of binary random variables $X \sim \text{Bern}(1/2)$ and $Y \sim \text{Bern}(1/2)$ and $X \oplus Y \sim \text{Bern}(p)$, where $\oplus$ is the sum modulo 2. That is, Y is the output of a binary symmetric channel with crossover probability p corresponding to the input X, and X is the output of the same channel with input Y.)

Then, it can be shown that the optimal U in (4) is such that $(X, U) \sim \text{DSBS}(q)$ for some $0 \le q \le 1$. Such a U can be obtained with the mapping $P_{U|X}$ such that

$$U = X \oplus Q, \quad \text{with } Q \sim \text{DSBS}(q). \tag{6}$$

In this case, straightforward algebra leads to that the complexity level is given by

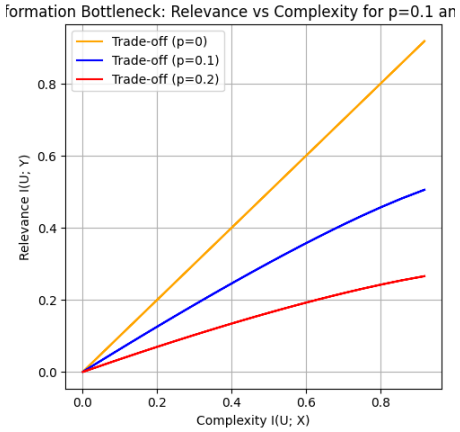
$$I(U;X) = 1 - h_2(q), \tag{7}$$

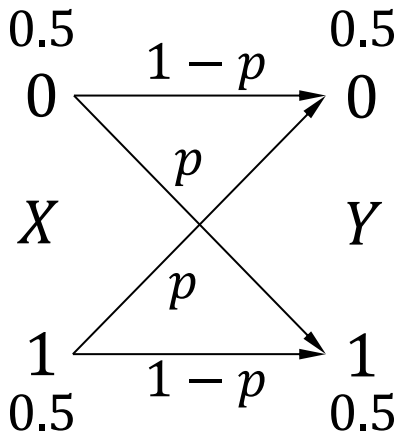
where, for $0 \le x \le 1$, $h_2(x)$ is the entropy of a Bernoulli-(x) source, i.e., $h_2(x) = -x \log_2(x) - (1-x)\log_2(1-x)$

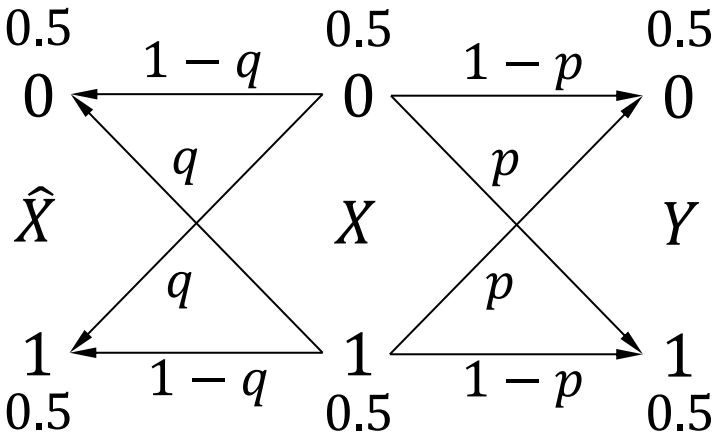
, and the relevance level is given by

$$I(U;Y) = 1 - h_2(p \star q) \tag{8}$$

where $p \star q = p(1-q) + q(1-p)$. The result extends easily to discrete symmetric mappings $Y \longrightarrow X$ with binary X (one bit output quantization) and discrete non-binary Y.

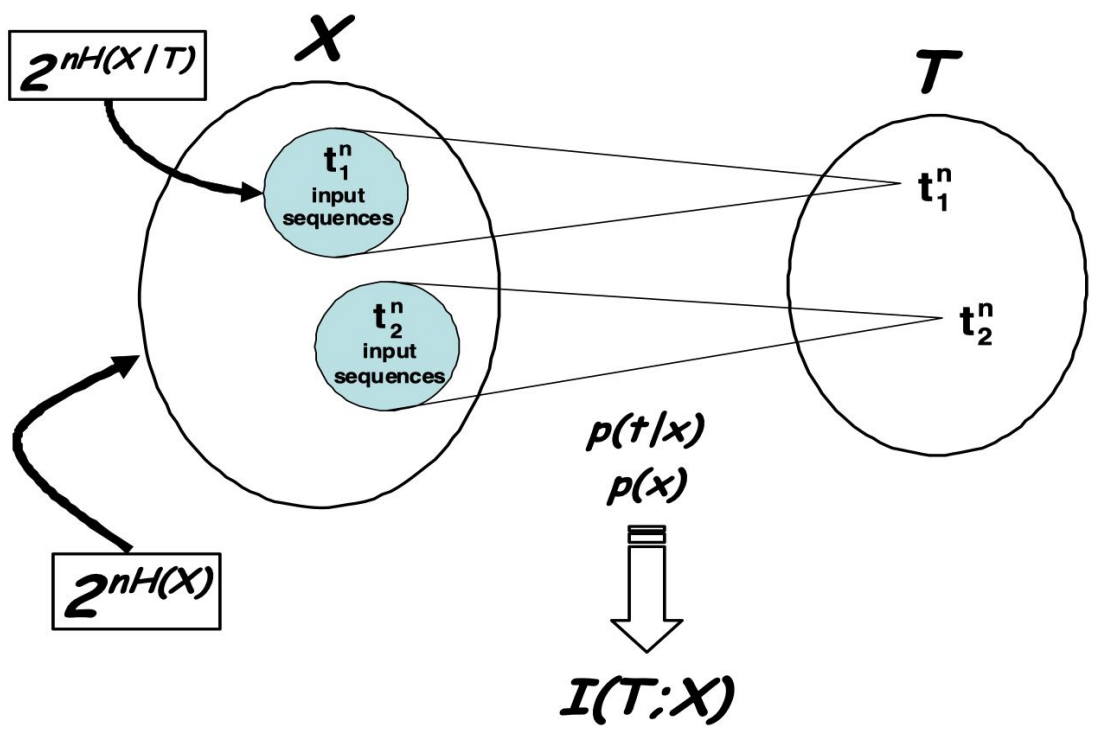






Source: Zidi, Estella-Aguerri, Shamai. On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views. Entropy, 2020. https://doi.org/10.3390/e22020151
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Rate distortion theory



T ... compressed representation (a quantized codebook) of X

representation is defined through a (possibly stochastic) mapping (condition distribution $p(t|x)$) between each value $x \in \mathcal{X}$ to each representative value $t \in \mathcal{T}$.

$I(T;X)$ ... compression information. Also rate of a code.

Is calculated based on the joint distribution $p(t|x) \cdot p(x)$

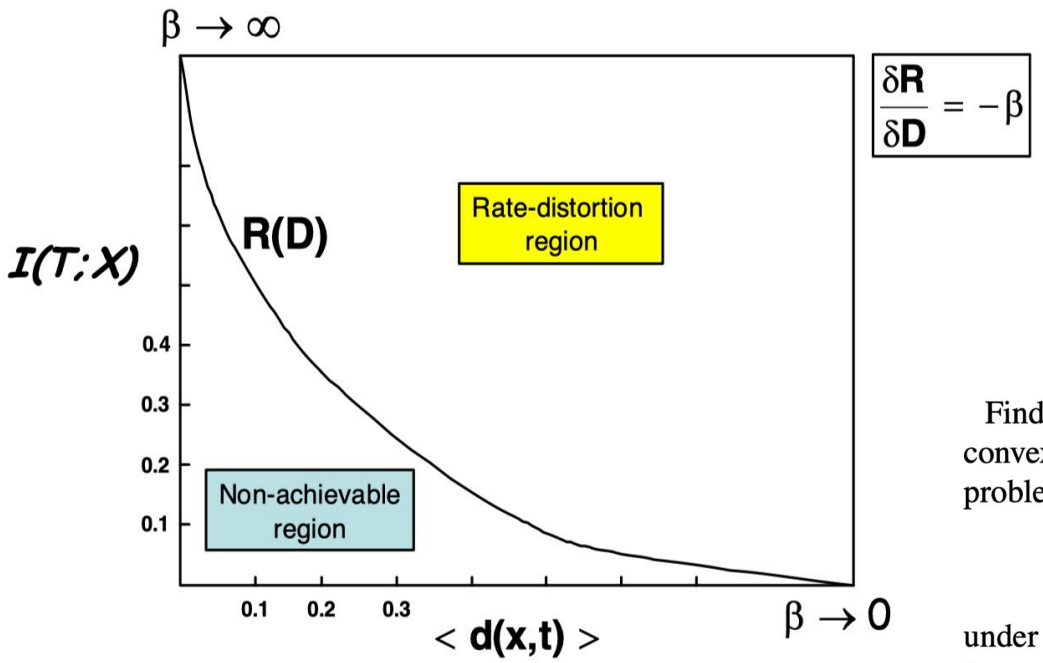
The expected distortion is:

$$D = \mathbb{E}_{X,T}[d(X,T)] = \sum_{x,t} p(x) \cdot p(t|x) \cdot d(x,t)$$

Figure 2.1: An illustration of the relation between the compression-information, $I(T;X)$, and the maximal number of bits that can be reliably transmitted between X and T. For every typical sequence of length n of T symbols there are $\approx 2^{nH(X|T)}$ possible ("input") X sequences. Hence, the total number of $\approx 2^{nH(X)}$ X sequences needs to be divided into disjoint subsets of size $\approx 2^{nH(X|T)}$. The number of such subsets is therefore upper bounded by $2^{n(H(X)-H(X|T))} = 2^{nI(T;X)}$. In other words, we can reliably send at most $\approx 2^{nI(T;X)}$ sequences of length n between X and T.

Source: Slonim. The information bottleneck: theory and applications, PhD thesis 2002. https://citeseerx.ist.psu.edu/document?doi=f9357064ef06a30f4533901cbc956bb25af646ad
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Rate distortion theory



$$R(D) \equiv \min_{\{p(t|x):\ \langle\, d(x,t)\,\rangle \le D\}} I(T;X)$$

Finding the rate-distortion function requires solving a minimization problem of a convex function over the convex set of all the (normalized) conditional distributions $p(t \mid x)$, satisfying the distortion constraint. This problem can be solved by introducing a Lagrange multiplier, β, and then minimizing the functional

$$\mathcal{F}[p(t \mid x)] = I(T;X) + \beta \,\langle\, d(x,t) \,\rangle_{p(x)p(t|x)} \,, \tag{2.3}$$

under the normalization constraints $\sum_t p(t \mid x) = 1,\ \forall x \in \mathcal{X}$. This formulation has the following well known consequences.

Figure 2.2: An illustration of a rate distortion function, $R(D)$. This function defines a monotonic convex curve in the distortion-compression plane with a slope of $-\beta$. When $\beta \to \infty$ we focus solely on minimizing the distortion which corresponds to the extreme case of the curve with $\langle\, d(x,t) \,\rangle_{p(x)p(t|x)} \to 0$. When $\beta \to 0$ we are only interested in compression, which corresponds to the other extreme of the curve with $R \to 0$. This curve characterizes the input (source) statistics, $p(x)$ with respect to a specific distortion measure and a specific choice of representatives, given by T values. The region above the curve is achievable while the region below it is non-achievable.

Source: Slonim. The information bottleneck: theory and applications, PhD thesis 2002. https://citeseerx.ist.psu.edu/document?doi=f9357064ef06a30f4533901cbc956bb25af646ad
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Proposition 2.1.2 : *Let $p(x)p(t \mid x)$ be a given joint distribution. Then the prior distribution $p(t)$ that minimizes $D_{KL}[p(x)p(t \mid x) \| p(x)p(t)]$ is the corresponding marginal distribution, i.e.,*

$$p^*(t) = \sum_x p(x)p(t \mid x) \; . \tag{2.7}$$

Note that at the minimum, $D_{KL}[p(x)p(t \mid x) \| p(x)p(t)]$ is exactly the information, $I(T;X)$ calculated on the basis of the joint distribution $p(x)p(t \mid x)$. Hence, this KL divergence is an upper bound for the compression-information term, and equality holds if and only if $p(t)$ is set to be the marginal distribution of $p(x)p(t \mid x)$. This proposition allows us to rewrite the definition of the rate-distortion function as a double minimization:

$$R(D) = \min_{\{p(t)\}} \min_{\{p(t|x):\, \langle\, d(x,t)\, \rangle \leq D\}} D_{KL}[p(x)p(t \mid x) \| p(x)p(t)] \; . \tag{2.8}$$

If A is the set of all joint distributions $p(t,x)$ with marginal $p(x)$ that satisfy the distortion constraint and if B is the set of the product distributions $p(t)p(x)$ with some normalized $p(t)$, we get

$$R(D) = \min_{b \in B} \min_{a \in A} D_{KL}[a \| b] \; .$$

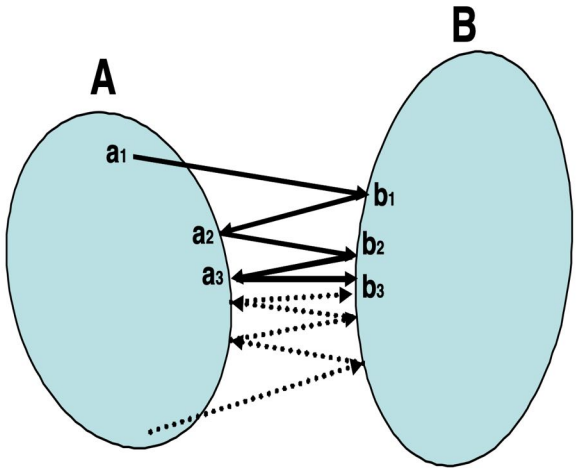


Figure 2.3: An illustration of alternating minimization of the Euclidean distance between two convex sets in $\mathcal{R}^2$. Since the minimized function (i.e., the Euclidean distance between the sets) is convex, the algorithm will always converge to the global minimum distance, independently of the initialization. This is also true for minimizing the KL divergence between two convex sets of probability distributions.

Source: Slonim. The information bottleneck: theory and applications, PhD thesis 2002. https://citeseerx.ist.psu.edu/document?doi=f9357064ef06a30f4533901cbc956bb25af646ad
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Rate distortion theory

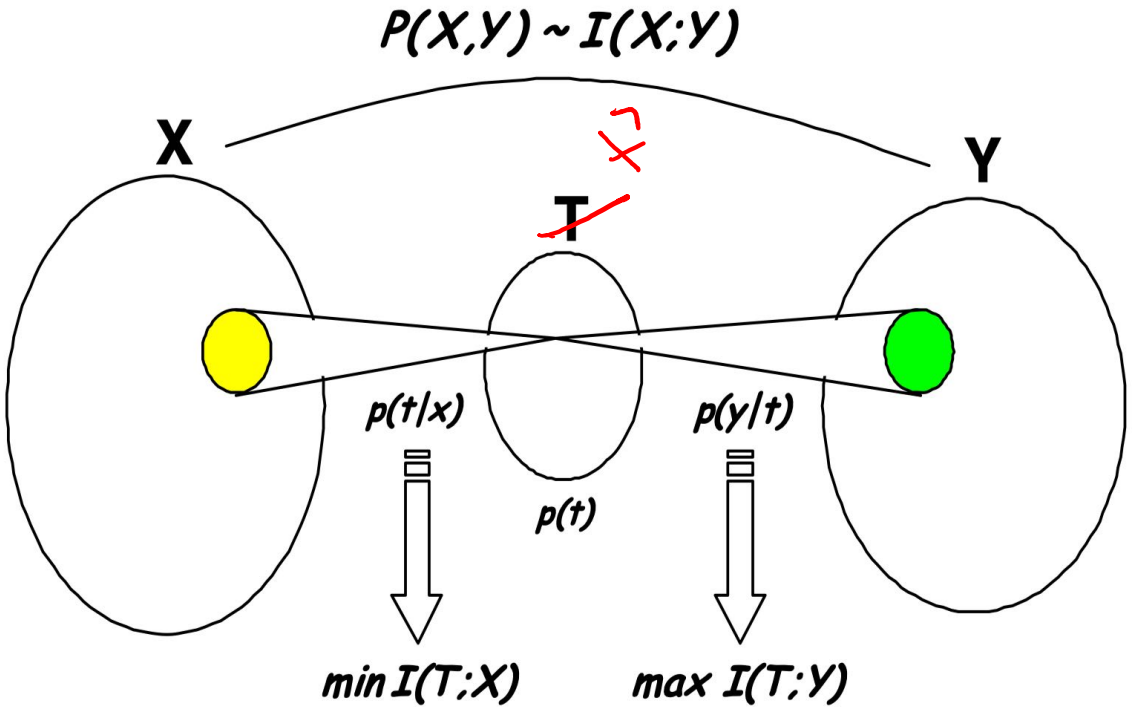


Figure 2.5: The information between X and Y is squeezed through the compact "bottleneck" representation, T. In particular, under some constraint over the minimal level of relevant information, $I(T;Y)$, one is trying to *minimize* the compression-information, $I(T;X)$ (note the similarity of the left part of the figure with Figure 2.1). In this formulation the IB principle extends the rate distortion problem, in the sense that given $p(x,y)$, the setup of the problem is completed and no distortion measure need be defined. An equivalent formulation is to constraint the compression-information to some maximal level, and then try to *maximize* the relevant information term. In this formulation the IB principle is somewhat reminiscent of the channel coding problem. Specifically, in this case one is trying to maximize the information transmitted through a (compact) channel, where the channel properties are governed by the constraint over the compression-information.

Source: Slonim. The information bottleneck: theory and applications, PhD thesis 2002. https://citeseerx.ist.psu.edu/document?doi=f9357064ef06a30f4533901cbc956bb25af646ad
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Updated 12/4/2024

Part 3: Applications
L24: Information Bottleneck Theory

Wolfgang Gatterbauer, Javed Aslam

cs7840 Foundations and Applications of Information Theory (fa24)

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

12/4/2024

Pre-class conversations

- Projects!
- Please start writing extensive feedback (you can post later)
- Today: Information Bottleneck Theory (2/2)

- **Lecture 20 (Mon 11/18):**
  Channel capacity [Cover Thomas'06: Ch 7]
- **Lecture 21 (Wed 11/20):**
  Distortion Theory (1/2) [Cover Thomas'06: Ch 10]
- **Lecture 22 (Mon 11/25):**
  Distortion Theory (2/2) [Cover Thomas'06: Ch 10]
  Python notebooks: 232
- **(Wed 11/27): no class (Fall break)**
- **Lecture 23 (Mon 12/2):** Information Bottleneck Theory
- **Lecture 24 (Wed 12/4):** Information Bottleneck Theory

Project presentations

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

- Rate Distortion & Information bottleneck theory
  - [Cover,Thomas'06] Elements of Information Theory. 2nd ed, 2006: Ch 10 Rate distortion theory
  - [Tishby+'99] Tishby, Pereira, Bialek. The information bottleneck method. The 37th annual Allerton Conference on Communication, Control, and Computing. pp. 368–377.
  - [Harremoes,Tishby'07] The Information Bottleneck Revisited or How to Choose a Good Distortion Measure. International Symposium on Information Theory, 2007.
  - [Zaslavsky+'18] Zaslavsky, Kemp, Regier, Tishby. The Efficient compression in color naming and its evolution. PNAS, 2018.
  - [Webb+'24] Webb, Frankland, Altabaa, Segert, Krishnamurthy, Campbell, Russin, Giallanza, Dulberg, OReilly, Lafferty, Cohen. The Relational Bottleneck as an Inductive Bias for Efficient Abstraction. Trends in Cognitive Science, 2024.
  - [Segert'24] Maximum Entropy, Symmetry, and the Relational Bottleneck: Unraveling the Impact of Inductive Biases on Systematic Reasoning. PhD thesis, Neuroscience @ Princeton, 2024.
  - [Ren,Li,Leskovec'20] Graph Information Bottleneck, NeurIPS, 2020.

Please leave lots of textual feedback on what is most helpful

1. Motivation: foundations & intuitive applications of information theory, n >> 10
2. Topics:
   a. basics → compression/encoding → channel/transmission → distortion → IB
   b. logistic regression, cross entropy, KL divergence as loss function (even k-means)
   c. axioms: intended as separate, ended up mixed into the topics (probability axioms)
   d. AEP, method of types, KL divergence, proofs
   e. Data management applications: information inequalities, cardinality estimation, normal forms, approximate acyclic schemas, explanation tables
3. Regular feedback (in both directions):
   a. Quick and often project feedback. Was not always used. More guidance on projects?
   b. Scribes: quick feedback on Piazza, more time for final versions on Canvas. Any procedural way to improve the scribe process? Scribes vs. homeworks.
   c. Online feedback form for instructors was used very rarely? Why? Can't be that you did not spot any errors in the slides. More interactivity (break-out sessions). Maybe more flipped (posting links upfront)
4. Other?

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 whiteboard lectures and exercises!

Source: "Thinking Physics: Understanding Practical Reality", Lewis Carroll Epstein, 1979-2009. https://www.goodreads.com/book/show/268266.Thinking_Physics
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Efficient compression in color naming and its evolution

Noga Zaslavsky[a,b,1], Charles Kemp[c,2], Terry Regier[b,d], and Naftali Tishby[a,e]

[a]Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem 9190401, Israel; [b]Department of Linguistics, University of California, Berkeley, CA 94720; [c]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213; [d]Cognitive Science Program, University of California, Berkeley, CA 94720; and [e]The Benin School of Computer Science and Engineering, The Hebrew University, Jerusalem 9190401, Israel

Edited by James L. McClelland, Stanford University, Stanford, CA, and approved June 18, 2018 (received for review January 11, 2018)

PNAS

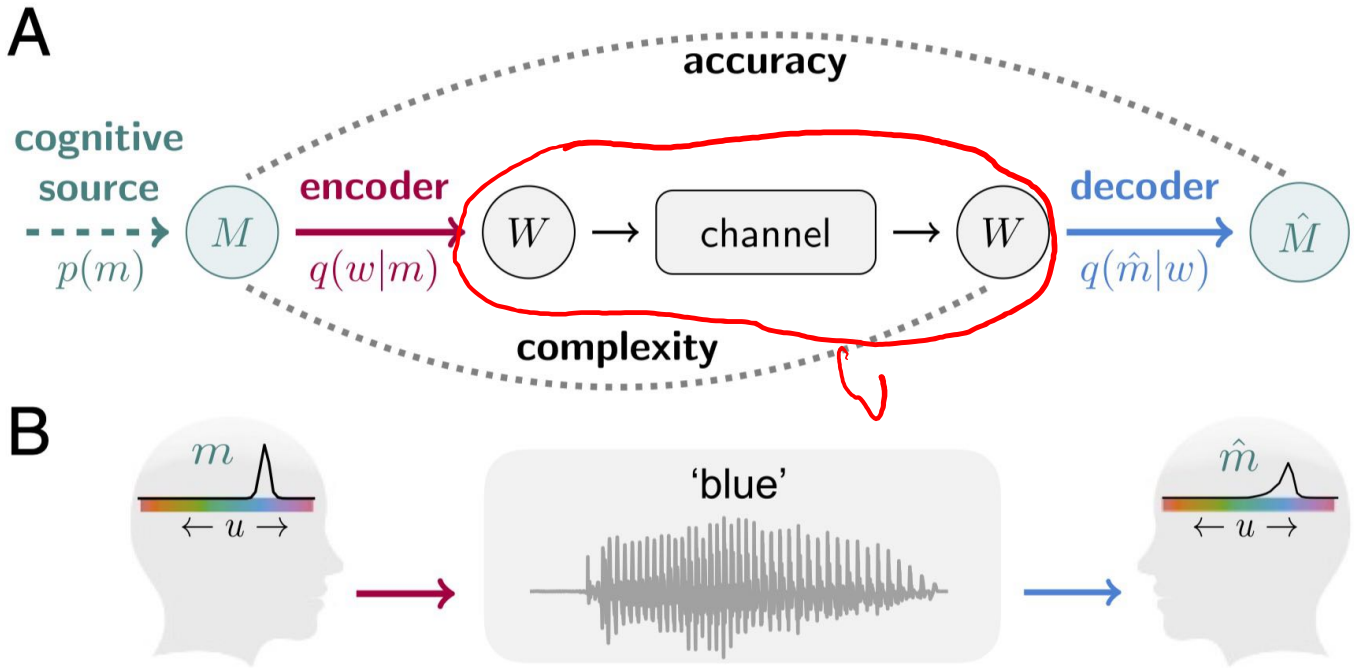
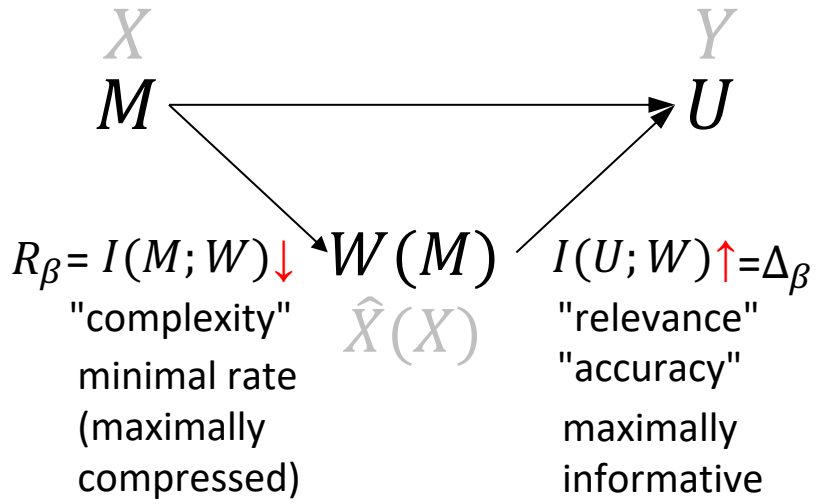


Fig. 1. (*A*) Shannon's (23) communication model. In our instantiation of this model, the source message M and its reconstruction $\hat{M}$ are distributions over objects in the universe $\mathcal{U}$. We refer to these messages as meanings. M is compressed into a code, or word, W. We assume that W is transmitted over an idealized noiseless channel and that the reconstruction $\hat{M}$ of the source message is based on W. The accuracy of communication is determined by comparing M and $\hat{M}$, and the complexity of the lexicon is determined by the mapping from M to W. (*B*) Color communication example, where $\mathcal{U}$ is a set of colors, shown for simplicity along a single dimension. A specific meaning m is drawn from $p(m)$. The speaker communicates m by uttering the word "blue," and the listener interprets blue as meaning $\hat{m}$.



Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

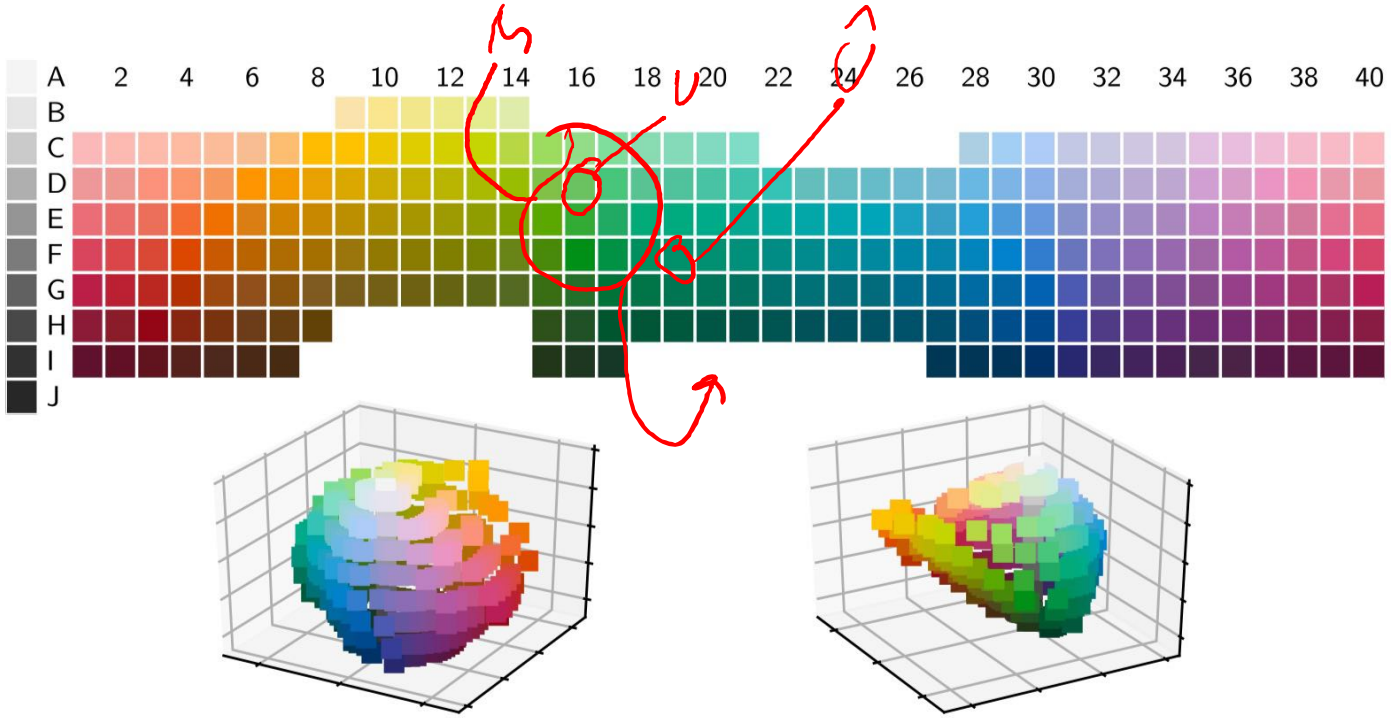


Fig. 2. (*Upper*) The WCS stimulus palette. Columns correspond to equally spaced Munsell hues. Rows correspond to equally spaced lightness values. Each stimulus is at the maximum available saturation for that hue/lightness combination. (*Lower*) These colors are irregularly distributed in 3D CIELAB color space.

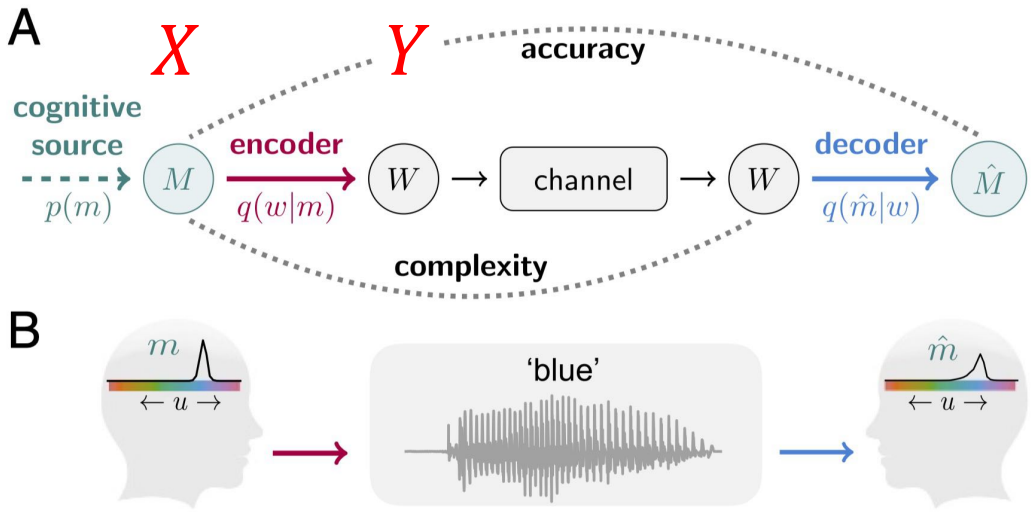
Encoders. Our primary data source for empirically estimating encoders was the World Color Survey (WCS), which contains color-naming data from 110 languages of nonindustrialized societies (24). Native speakers of each language provided names for the 330 color chips shown in Fig. 2, *Upper*. We also analyzed color-naming data from English, collected relative to the same stimulus array (25). We assumed that each color chip c is associated with a unique meaning m_c and therefore estimated an encoder $q_l(w|m_c)$ for each language l from the empirical distribution of word w given chip c (see data rows in Fig. 4 for examples). Each such encoder corresponds to a representative speaker for language l, obtained by averaging naming responses over speakers.

Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

In our formulation the speaker represents her intended meaning M by W, using an encoder $q(w|m)$, and thus the complexity is given by the information rate

$$I_q(M;W) = \sum_{m,w} p(m)q(w|m)\log\frac{q(w|m)}{q(w)}, \qquad [2]$$

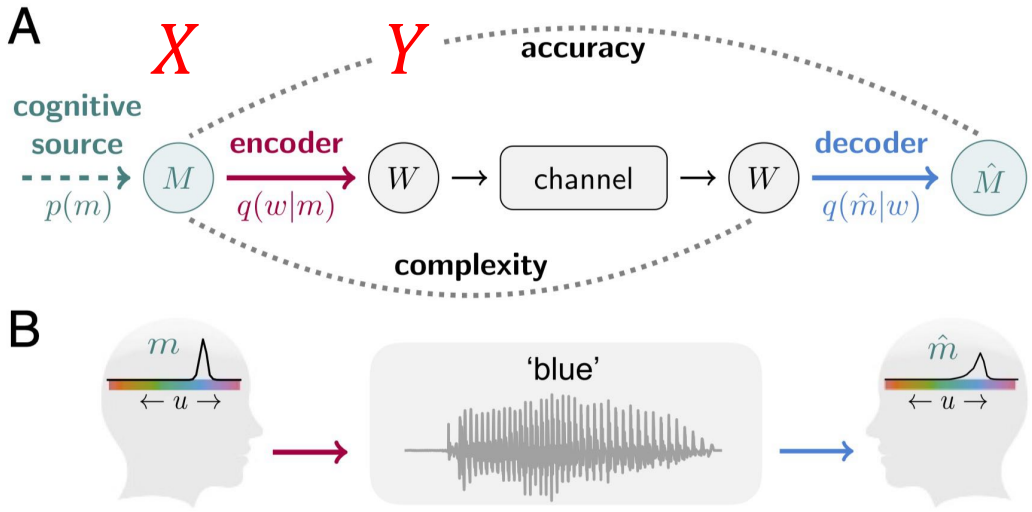
?



Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

In our formulation the speaker represents her intended meaning M by W, using an encoder $q(w|m)$, and thus the complexity is given by the information rate

$$I_q(M;W) = \sum_{m,w} p(m)\,q(w|m) \log \frac{q(w|m)}{q(w)}, \qquad [2]$$



$$I(X;\hat{X}) := H(\hat{X}) - H(\hat{X}|X)$$

$$= \sum_{x,\hat{x}} p(x)\cdot p(\hat{x}|x)\cdot \lg\left(\frac{p(\hat{x}|x)}{p(\hat{x})}\right)$$

$$H(Y|X) = \mathbb{E}_{p(x)}[H(Y|X=x)]$$

$$= \sum_{x} p(x)\cdot H(Y|X=x)$$

$$= \sum_{x} p(x)\cdot \sum_{\hat{x}} p(\hat{x}|x)\cdot \lg\left(\frac{1}{p(\hat{x}|x)}\right)$$

$$H(Y) = \sum_{\hat{x}} p(\hat{x})\cdot \lg\left(\frac{1}{p(\hat{x})}\right) \qquad p(\hat{x}) = \sum_{x} p(x)\cdot p(\hat{x}|x)$$

$$= \sum_{x,\hat{x}} p(x)\cdot p(\hat{x}|x)\cdot \lg\left(\frac{1}{p(\hat{x})}\right)$$

Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

The accuracy of a lexicon is inversely related to the cost of a misinterpreted or distorted meaning. While RDT allows an arbitrary distortion measure, IB considers specifically the Kullback–Leibler (KL) divergence,

$$D\left[m\|\hat{m}\right]=\sum_{u\in\mathcal{U}} m(u)\log\frac{m(u)}{\hat{m}(u)}, \qquad [3]$$

which is a natural distortion measure between distributions. [For a general justification of the KL divergence see ref. 26, and in the context of IB see ref. 18.] Note that this quantity is 0 if and only if the listener's interpretation is accurate; namely, $\hat{m}\equiv m$.

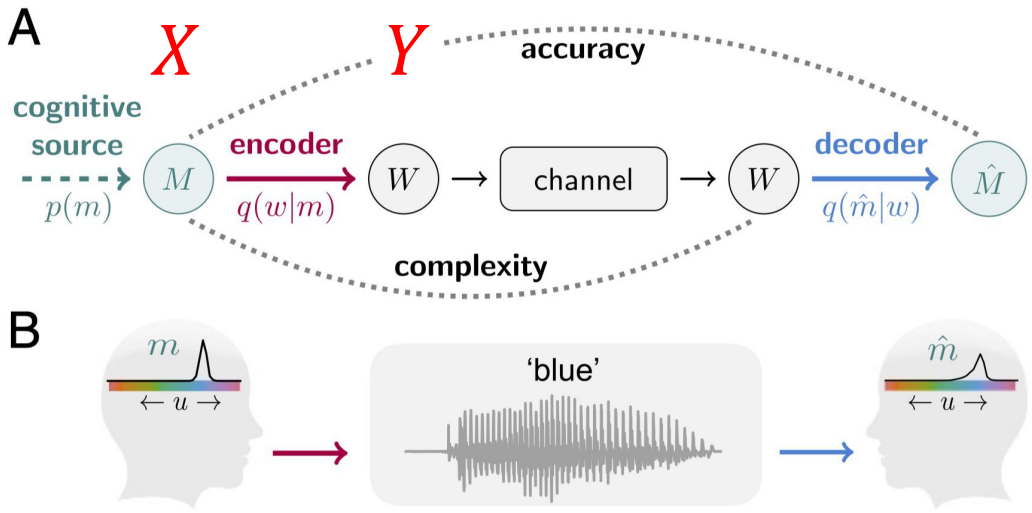
The distortion between the speaker and the ideal listener is the expected KL divergence,

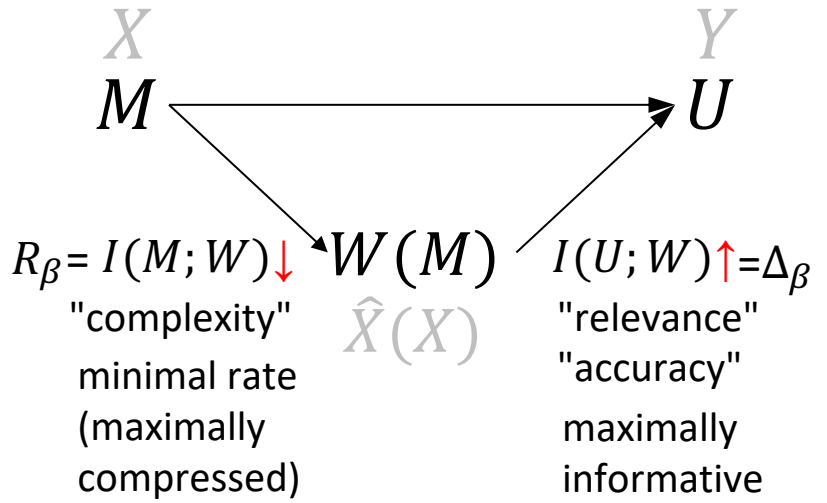
$$\mathbb{E}_q\left[D[M\|\hat{M}]\right]=\sum_{m,w} p(m)q(w|m)D\left[m\|\hat{m}_w\right]. \qquad [4]$$

In this case, the accuracy of the lexicon is directly related to Shannon's mutual information,

$$\mathbb{E}_q\left[D[M\|\hat{M}]\right]=I(M;U)-I_q(W;U). \qquad [5]$$

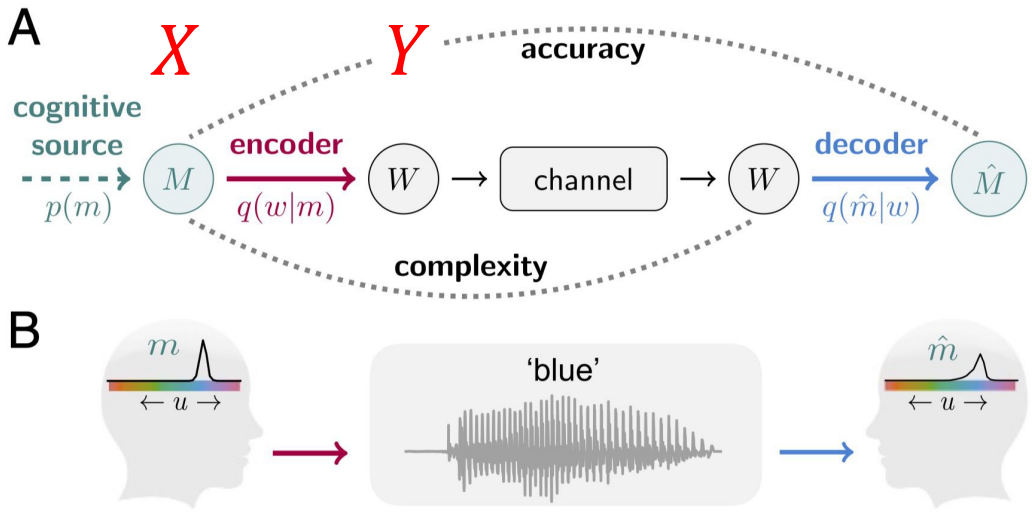
Since $I(M;U)$ is independent of $q(w|m)$, minimizing distortion is equivalent to maximizing the informativeness, or accuracy, of the lexicon, quantified by $I_q(W;U)$. This means that mutual information appears in our setting as a natural measure both for complexity and for semantic informativeness.



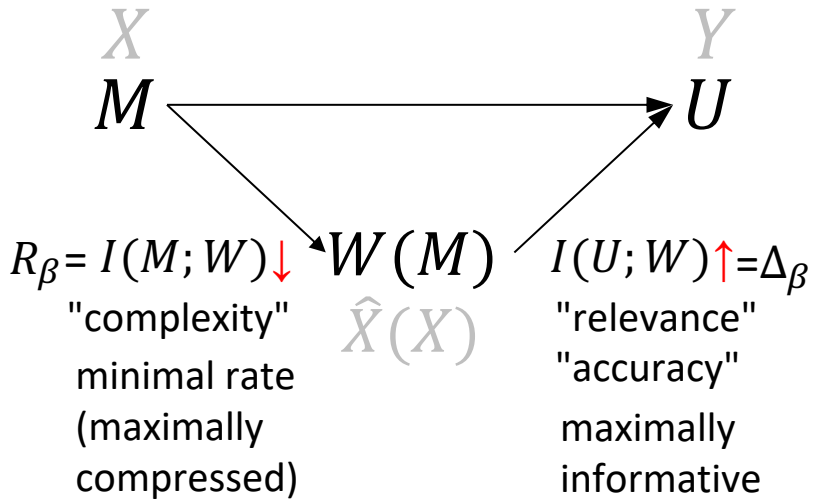


Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

	Component	IB (1999)	IB (current)
Communication model	Target variable / universe	$y \in \mathcal{Y}$	$u \in \mathcal{U}$
	Source variable	$x \in \mathcal{X}$	$m \in \mathcal{M}$
	Speaker's intended meaning	$p(y\|x)$	$m(u)$
	Source distribution / need	$p(x)$	$p(m)$
	Cluster / word	$\hat{x} \in \hat{\mathcal{X}}$	$w \in \mathcal{W}$
	Encoder / naming distribution	$q(\hat{x}\|x)$	$q(w\|m)$
	Decoder	$\hat{x} \mapsto q(y\|\hat{x})$	$q(\hat{m}\|w)$
	Listener's interpreted meaning	$q(y\|\hat{x})$	$\hat{m}_w(u)$
Optimization principle	Complexity	$I_q(X;\hat{X})$	$I_q(M;W)$
	Distortion / communicative cost	$D[p(y\|x)\|\|q(y\|\hat{x})]$	$D[m\|\|\hat{m}]$
	Accuracy	$I_q(\hat{X};Y)$	$I_q(W;U)$
	Tradeoff parameter	β	β



1.1. Summary of notation. We use capital letters to denote random variables (e.g. M and U), calligraphic letters to denote their support (e.g. $\mathcal{M}$ and $\mathcal{U}$), and lower case letters to denote a specific realization (e.g. m and u). In our formulation we consider a finite set of distributions $\mathcal{M}$. Each element in this set (i.e., each $m \in \mathcal{M}$) is a distribution over the set $\mathcal{U}$. In other words, m is a function that takes u as an argument. We use the notation $m(u)$ when we wish to make explicit that m is a function of u, or when we wish to denote the probability of a specific u according to m. It may be helpful to think of $m(u)$ in terms of conditional probabilities, i.e., $m(u) = p(u|m)$. Table S1 summarizes



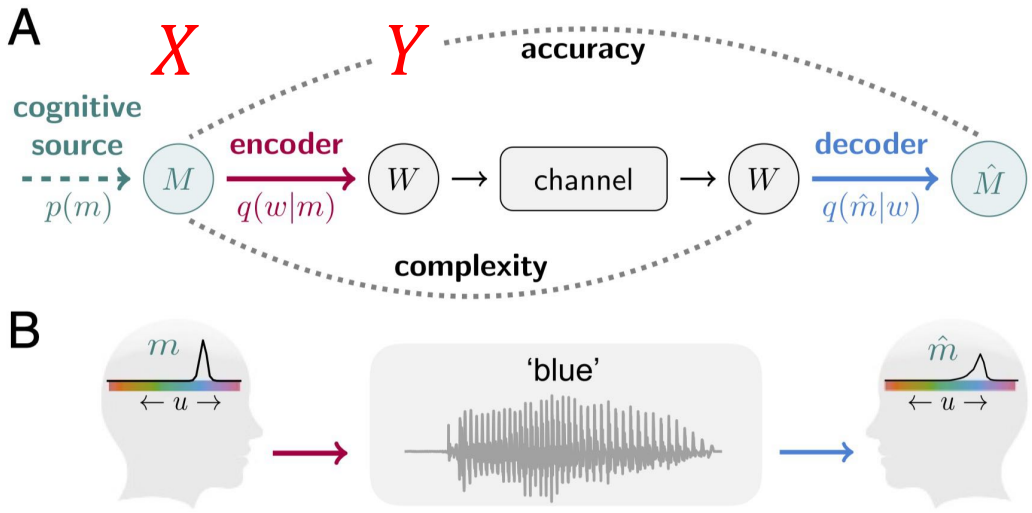
Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

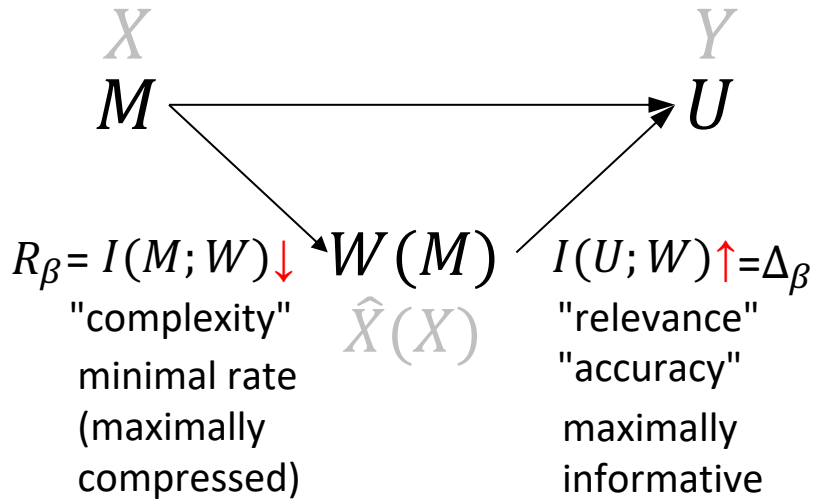
In this case, the listener receives w and interprets it as meaning $\hat{m}$ based on her interpretation policy $q(\hat{m}|w)$, which is a decoder. We focus on the efficiency of the encoder and therefore assume an optimal Bayesian listener with respect to the speaker (see *SI Appendix*, section 1.2 for derivation), who interprets every word w deterministically as meaning

$$\hat{m}_w(u) = \sum_{m \in \mathcal{M}} q(m|w)m(u), \qquad [1]$$

where $q(m|w)$ is obtained by applying Bayes' rule with respect to $q(w|m)$ and $p(m)$.

In this model, different color-naming systems correspond to different encoders, and our goal is to test the hypothesis that encoders corresponding to color-naming systems found in the world's languages are information-theoretically efficient. We next describe the elements of this model in further





Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

$p(u|w)$ $p(u|m)$

1.2. Bayesian listener. We show that the ideal listener with respect to a given speaker is an optimal Bayesian decision maker. In our case, this means that we can assume an ideal listener that always decodes w deterministically by interpreting it as meaning $\hat{m}_w(u) = \sum_{m \in \mathcal{M}} q(m|w)m(u)$, where $q(m|w)$ is obtained by applying Bayes' rule,

$$q(m|w) = \frac{q(w|m)p(m)}{q(w)}, \qquad \text{[S1]}$$

where $q(w) = \sum_{m'} p(m')q(w|m')$. To show that this Bayesian listener is optimal, assume that the speaker's encoder is given by $q(w|m)$. The optimal listener for this speaker is defined by the decoder $q(\hat{m}|w)$ that minimizes

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(W;U) = I_q(M;W) - \beta \left(I(M;U) - \mathbb{E}_q \left[D[M\|\hat{M}] \right] \right), \qquad \text{[S2]}$$

where the second equality follows from Eq. (5). Note that $I(M;U)$ is constant in q and $I_q(M;W)$ depends on the encoder but not on the decoder. Only the last term depends on the decoder, and it holds that

$$\mathbb{E}_q \left[D[M\|\hat{M}] \right] = \sum_{m,w,\hat{m}} p(m)q(w|m)q(\hat{m}|w)D\left[m\|\hat{m}\right] \qquad \text{[S3]}$$

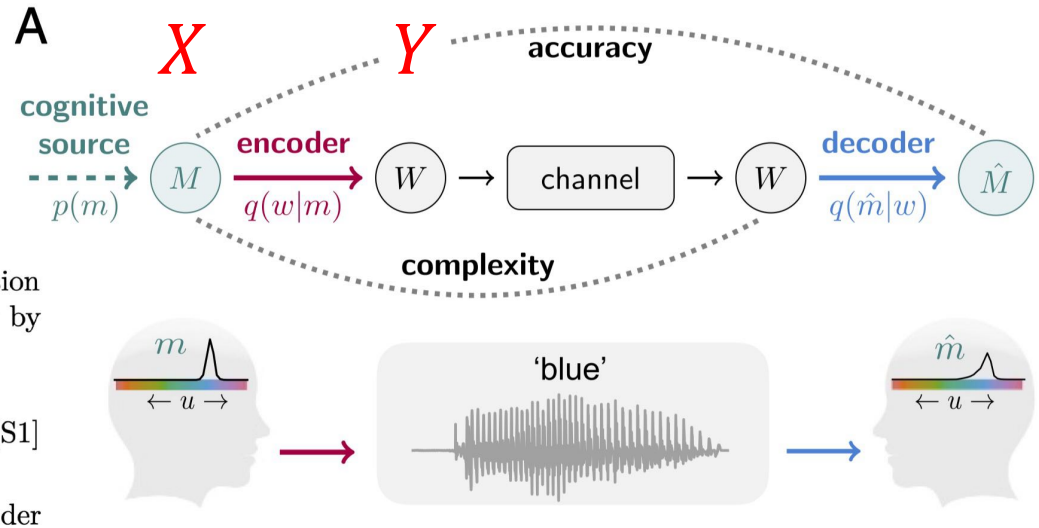
$$= \sum_{m,w,\hat{m}} q(w)q(m|w)q(\hat{m}|w)D\left[m\|\hat{m}\right] \qquad \text{[S4]}$$

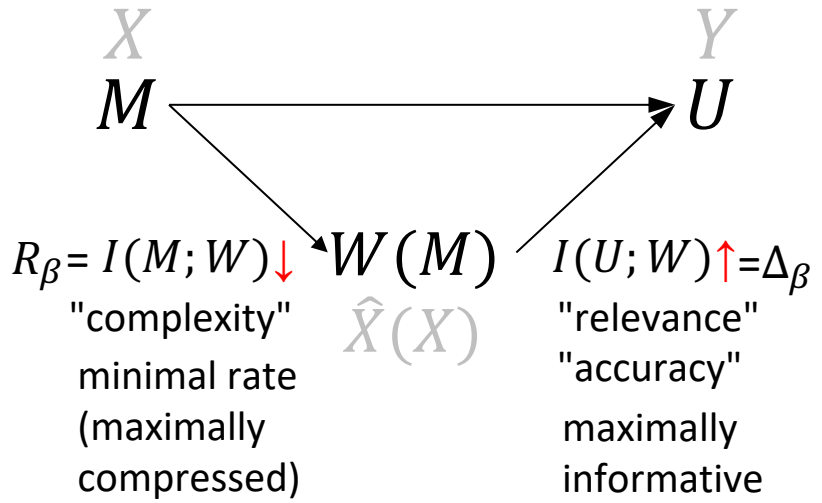
$$\geq \sum_{w} q(w) \operatorname*{argmin}_{\hat{m}'} \sum_{m} q(m|w)D\left[m\|\hat{m}'\right] \qquad \text{[S5]}$$

Therefore, there is a deterministic decoder $q(\hat{m}|w)$ that minimizes Eq. (S2),

$$q(\hat{m}|w) = \begin{cases} 1 & \text{if } \hat{m} = \operatorname*{argmin}_{\hat{m}'} \mathbb{E}_{q(m|w)} \left[D\left[m\|\hat{m}'\right] \right] \\ 0 & \text{otherwise} \end{cases}. \qquad \text{[S6]}$$

Differentiating $\mathbb{E}_{q(m|w)} \left[D\left[m\|\hat{m}'\right] \right]$ with respect to $\hat{m}'$ and equating to 0 gives that the minimum is attained at $\hat{m}_w$. Since $\sum_u \hat{m}_w(u) = 1$ we did not need to impose this normalization constraint on the optimization, and because the KL divergence is convex in both arguments $\hat{m}_w$ is indeed the minimum.





Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

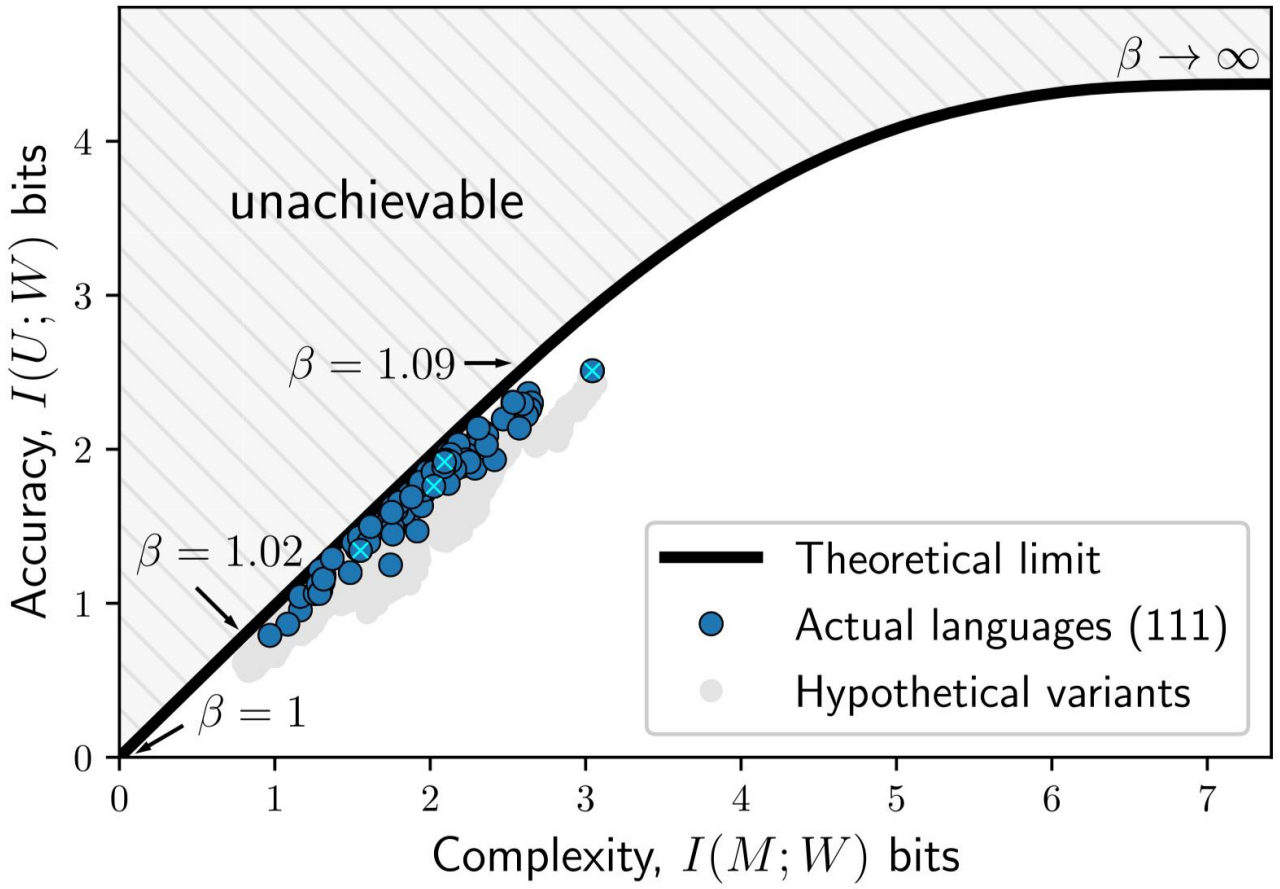
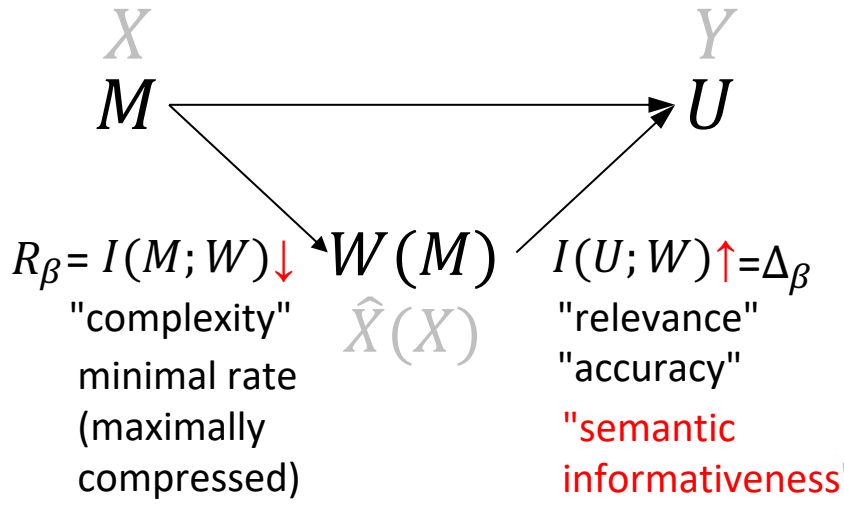


Fig. 3. Color-naming systems across languages (blue circles) achieve near-optimal compression. The theoretical limit is defined by the IB curve (black). A total of 93% of the languages achieve better trade-offs than any of their hypothetical variants (gray circles). Small light-blue Xs mark the languages in Fig. 4, which are ordered by complexity.

If the speaker and the listener are unwilling to tolerate any information loss, the speaker must assign a unique word to each meaning, which requires maximal complexity. However, between the two extremes of minimal complexity and maximal accuracy, an optimal trade-off between these two competing needs can be obtained by minimizing the IB objective function,

$$\mathcal{F}_\beta[q(w|m)] = I_q(M;W) - \beta I_q(W;U), \qquad \textbf{[6]}$$

$\beta \geq 1$



Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

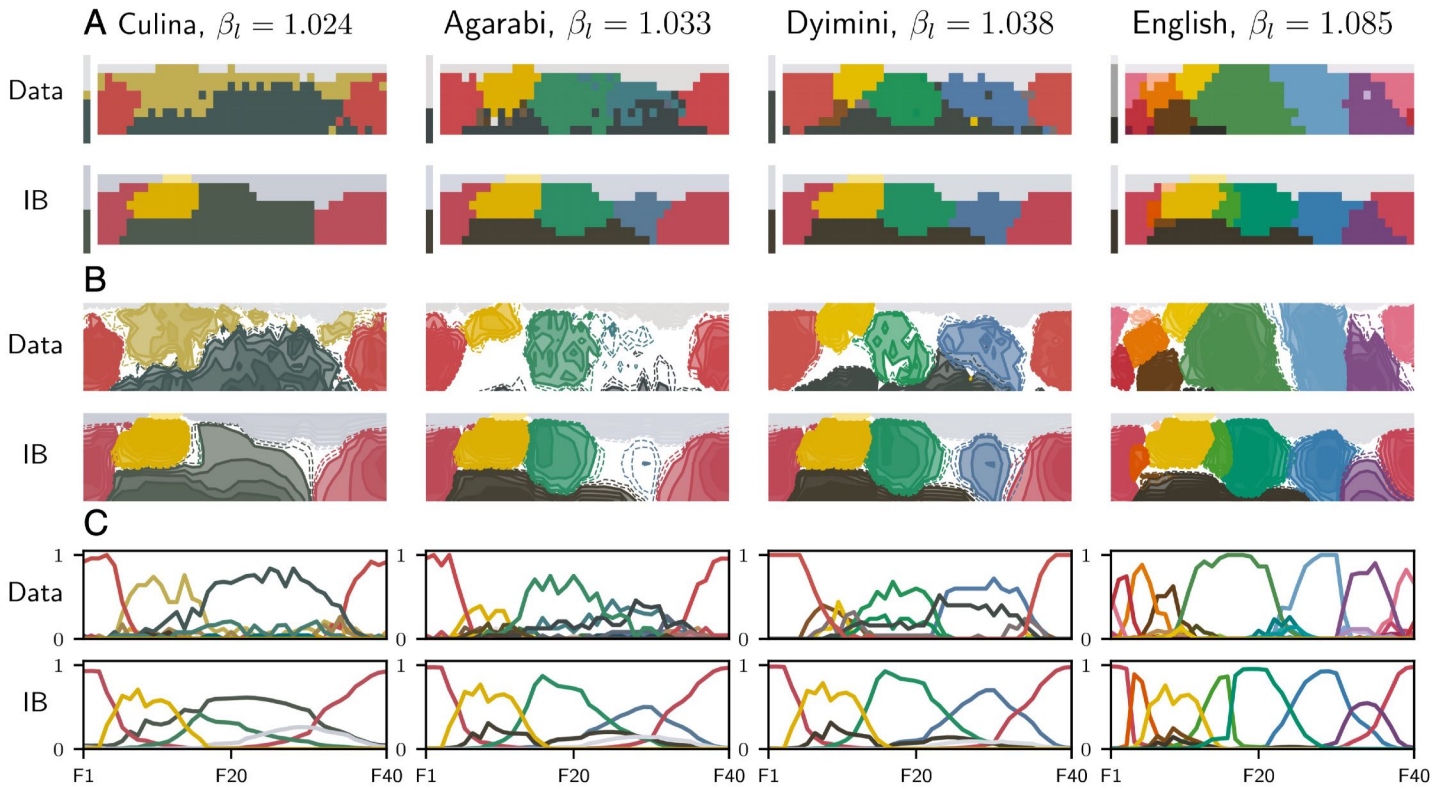


Fig. 4. Similarity between color-naming distributions of languages (data rows) and the corresponding optimal encoders at β_l (IB rows). Each color category is represented by the centroid color of the category. (*A*) Mode maps. Each chip is colored according to its modal category. (*B*) Contours of the naming distribution. Solid lines correspond to level sets between 0.5 and 0.9; dashed lines correspond to level sets of 0.4 and 0.45. (*C*) Naming probabilities along the hue dimension of row F in the WCS palette.

Structure of Semantic Categories. Previous work (e.g., ref. 8) has sometimes summarized color-naming responses across multiple speakers of the same language by recording the modal naming response for each chip, resulting in a hard categorical partition of the stimulus array, called a mode map (e.g., Fig. 4*A*). However, IB predicts that if some information loss is allowed, i.e., $\beta < \infty$, then an efficient encoder would induce soft rather than hard categories. This follows from the structure of the IB optima (14), given by

$$q_\beta(w|m) \propto q_\beta(w) \exp(-\beta D[m \| \hat{m}_w]), \qquad [7]$$

which is satisfied self-consistently with Eq. **1** and with the marginal $q_\beta(w)$. We therefore evaluate how well our model accounts for mode maps, but more importantly we also evaluate how well it accounts for the full color-naming distribution across

Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

2. Least informative source

2.1. Definition for a given language. We begin by defining a least informative prior over color chips, with respect to a given naming distribution $q_l(w|c)$. Because we assumed that each chip c is associated with a unique meaning m_c, any prior $p(c)$ induces a source distribution by setting $p(m_c) = p(c)$. One common approach for obtaining uninformative priors is by invoking the maximum entropy principle. However, in our case the maximum entropy distribution over color chips is simply the uniform distribution. Another natural approach in our setting is to find a distribution that maximizes the entropy of c while minimizing the expected uncertainty over c give a term w in the language. That is,

$$p_l(c) = \underset{p(c)}{\text{argmax}}\ H(C) - H_q(C|W) \qquad \text{[S10]}$$

where $H_q(C|W) = -\sum_{c,w} p(c)q(w|c)\log\frac{q(c|w)}{p(c)}$ is the conditional entropy, and $q(c|w) = \frac{q(w|c)p(c)}{q(w)}$ is the posterior distribution of c given w.

This definition has two interesting interpretations, in addition to being a constrained maximum entropy distribution. First, note that

$$I_q(W;C) = \underset{p(c)}{\text{argmax}}\ H(C) - H_q(C|W)\,, \qquad \text{[S11]}$$

which implies that $p_l(c)$ maximizes the mutual information between colors and words. This type of prior distribution is also called a capacity achieving prior, and can be evaluated using the Blahut-Arimoto algorithm (10, 11). Note that in the IB model, a language l would be maximally complex if the source distribution were defined from $p_l(c)$. This contrasts with the IB principle, which aims to minimize complexity. Second, $p_l(c)$ is considered the least informative prior over c in the sense that it minimizes information about the posterior $q(c|w)$ by maximizing the KL divergence between the prior and posterior. This interpretation follows from the identity

in/of (~~about~~)

$$I_q(W;C) = \sum_w q(w)D[q(c|w)\|p(c)]\,, \qquad \text{[S12]}$$

and it is closely related to the notion of reference priors in Bayesian inference (12). Reference priors are considered objective priors in the sense that they depend solely on the given distribution $q(w|c)$, but not on other assumptions that may reflect subjective prior beliefs.

$$I(X;Y) =$$
$$= D_{\mathrm{KL}}(p(x,y)||p(x)\cdot p(y))$$
$$= D_{\mathrm{KL}}(p_{X,Y}||p_X \cdot p_Y)$$
$$= D_{\mathrm{KL}}(p_{Y|X}||p_Y)$$

Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Color naming

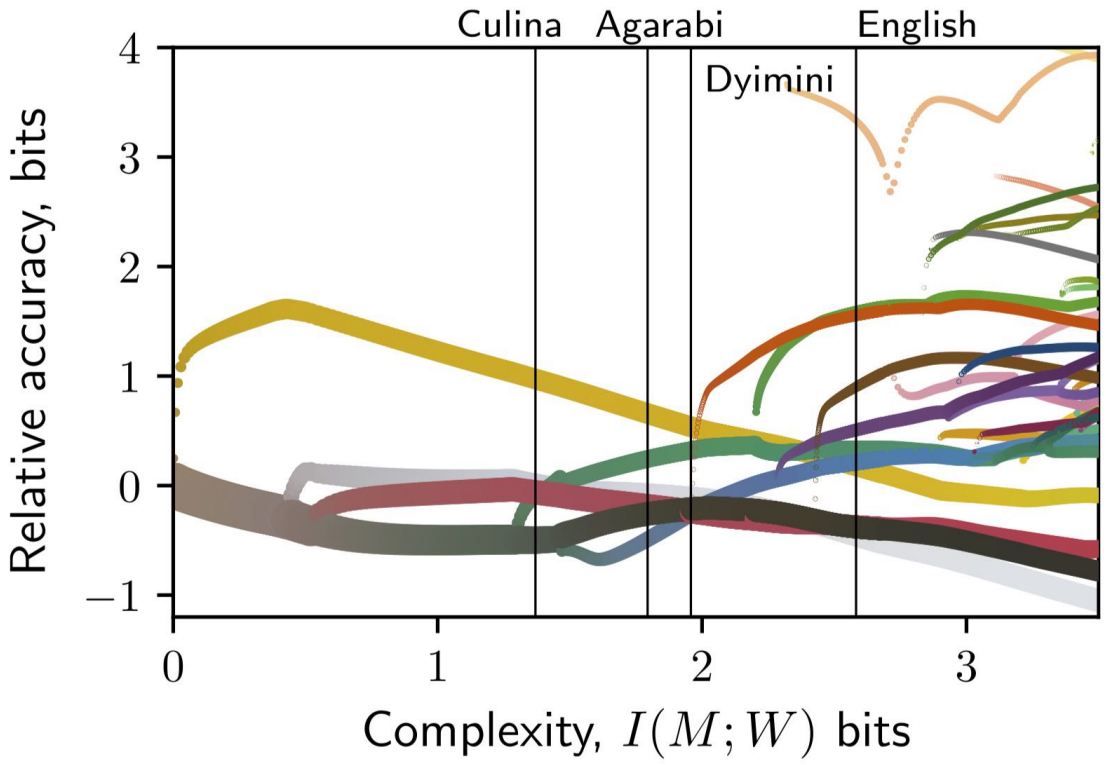


Fig. 5. Bifurcations of the IB color categories (Movie S1). The *y* axis shows the relative accuracy of each category *w* (defined in *Materials and Methods*). Colors correspond to centroids and width is proportional to the weight of each category, i.e., $q_\beta(w)$. Black vertical lines correspond to the IB systems in Fig. 4.

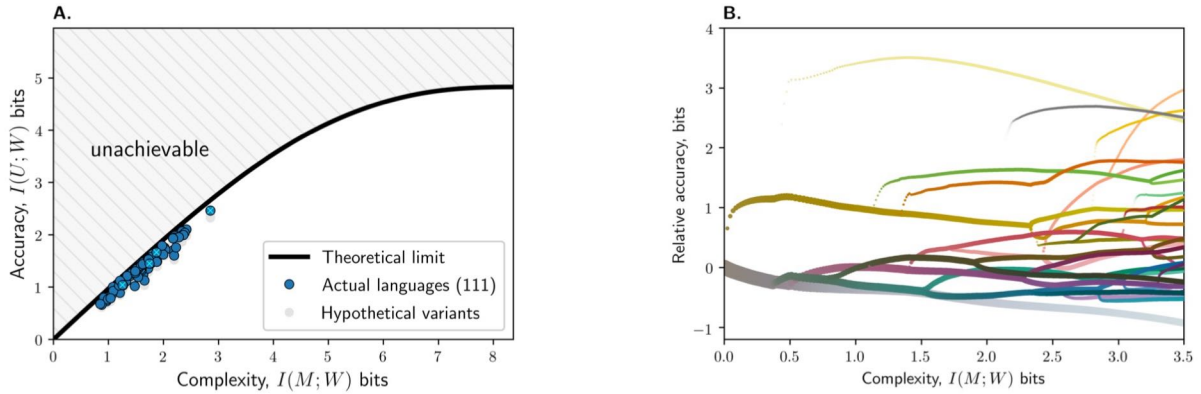


Fig. S7. Uniform source. Information plane (**A**) and bifurcation diagram (**B**) evaluated for the uniform source. For more details see captions of Fig.3 and Fig.5 in main text.

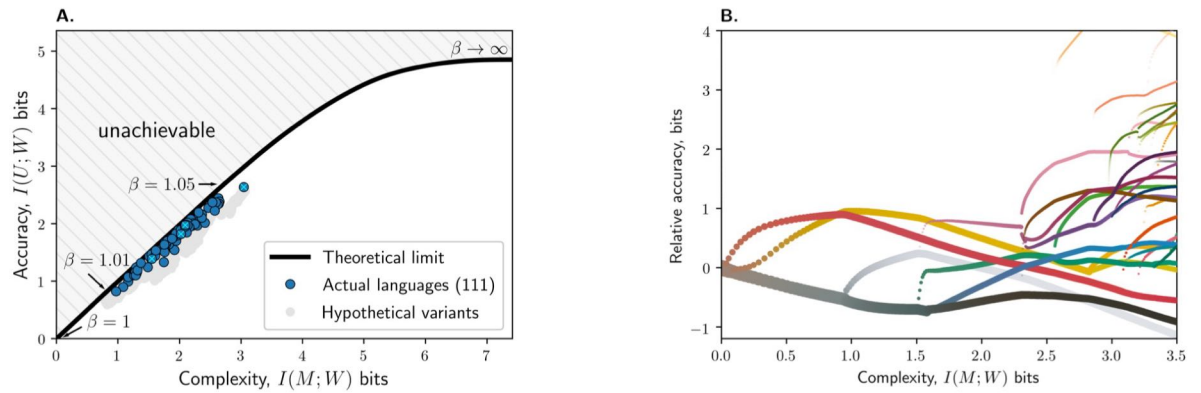


Fig. S5. CIELUV space. Information plane (**A**) and bifurcation diagram (**B**) for the full LI source. These figures are similar to Fig.3 and Fig.5 in main text, but they are based on the results for CIELUV instead of CIELAB.

Source: Zaslavsky, Kemp, Regier, Tishby. Efficient compression in color naming and its evolution, PNAS 2018. https://doi.org/10.1073/pnas.1800521115
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

Trends in
Cognitive Sciences

CellPress

Review

The relational bottleneck as an inductive bias for efficient abstraction

Taylor W. Webb [1,*], Steven M. Frankland [2], Awni Altabaa [3], Simon Segert [4], Kamesh Krishnamurthy [4], Declan Campbell [4], Jacob Russin [5], Tyler Giallanza [4], Randall O'Reilly [6], John Lafferty [3], and Jonathan D. Cohen [4]

Source: Webb, Frankland, Altabaa, Segert, Krishnamurthy, Campbell, Russin, Giallanza, Dulberg, OReilly, Lafferty, Cohen. The Relational Bottleneck as an Inductive Bias for Efficient Abstraction, 2024. Trends in Cognitive Sciences, 2024. https://doi.org/10.1016/j.tics.2024.04.001
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/

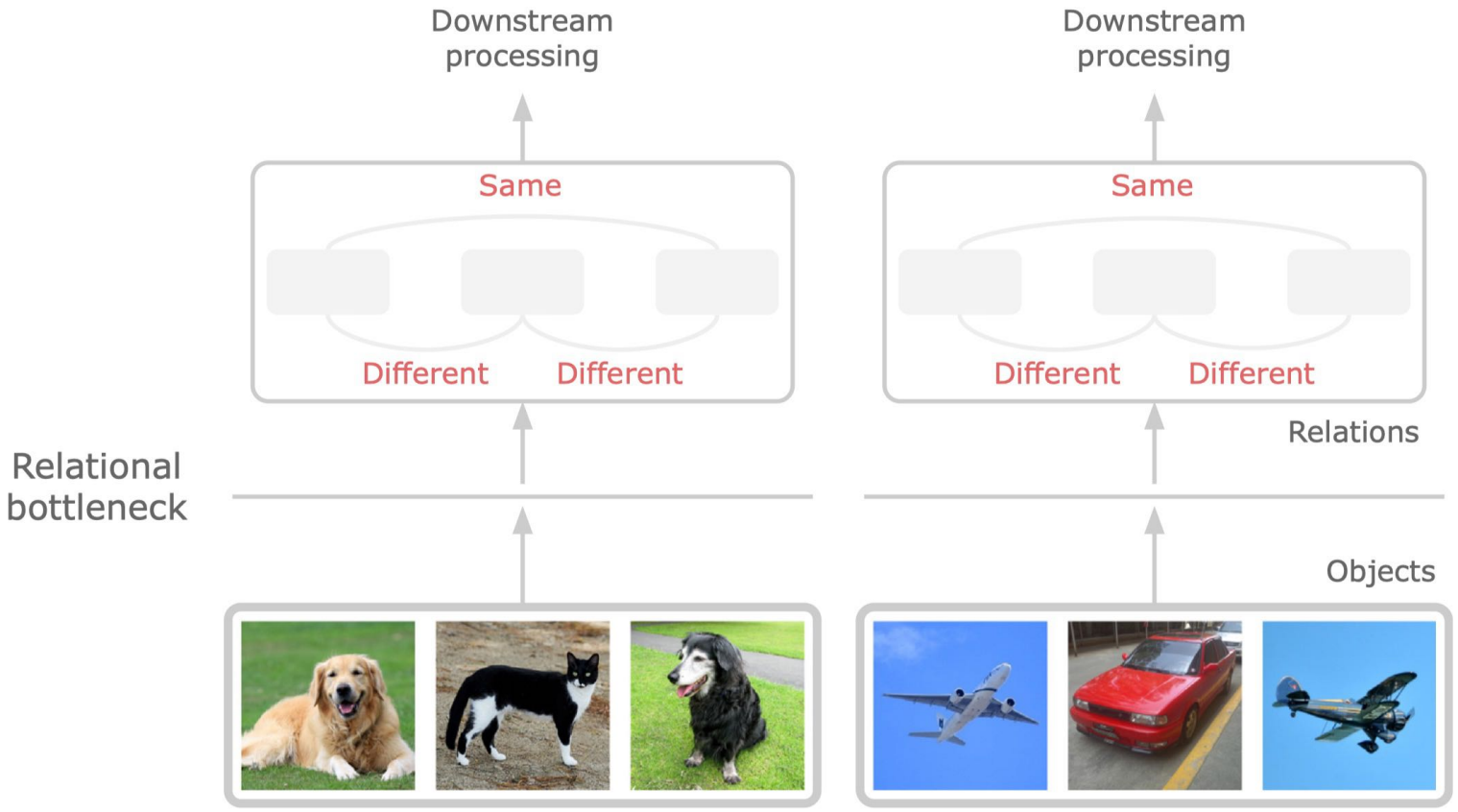


Figure 1. The relational bottleneck. An inductive bias that prioritizes the representation of relations (e.g., 'same' versus 'different') and discourages the representation of the features of individual objects (e.g., the shape or color of the objects in the images). The result is that downstream pr ocessing is driven primarily, or even exclusively, by patterns of relations and can therefore systematically generalize those patterns across distinct instances (e.g., the common ABA pattern displayed on both left and right), even for completely novel objects. The approach is illustrated here with same/different relations, but other relations can also be accommodated. Note that this example is intended only to illustrate the overall goal of the relational bottleneck framework. Figure 2 in the main text depicts neural architectures that implement the approach.

Source: Webb, Frankland, Altabaa, Segert, Krishnamurthy, Campbell, Russin, Giallanza, Dulberg, OReilly, Lafferty, Cohen. The Relational Bottleneck as an Inductive Bias for Efficient Abstraction, 2024. Trends in Cognitive Sciences, 2024. https://doi.org/10.1016/j.tics.2024.04.001

Relational Query Patterns

Definition 9 (Query signature). A *table reference* in a query expression q is any existentially or universally quantified reference to an input table. The *signature* $\mathcal{S}$ of q is the ordered list of its table references.

Definition 10 (Dissociated query). A dissociation of a query expression q with signature $\mathcal{S}$ is a modified query q' with $\mathcal{S}$ being replaced with a table signature $\mathcal{S}'$ of same size (i.e. $|\mathcal{S}'| = |\mathcal{S}|$), where every table in $\mathcal{S}'$ has a different name, and every table $\mathcal{S}'[i]$ has the same schema as table $\mathcal{S}[i]$ for all $i \in [|\mathcal{S}|]$.

Definition 11 (Relational pattern). Given a query expression q with signature $\mathcal{S}$. The *relational pattern* of q is the logical function defined by its dissociated query $q'(\mathcal{S}')$.

Definition 12 (Pattern isomorphism). Given two logically-equivalent queries q_1 and q_2 with signatures $\mathcal{S}_1$ and $\mathcal{S}_2$, and dissociated queries $q'_1(\mathcal{S}'_1)$ and $q'_2(\mathcal{S}'_2)$, respectively. The queries are *pattern-isomorphic* iff $q'_1(\mathcal{S}'_1) = q'_2(\pi(\mathcal{S}'_1))$ for some permutation π. In that case, we call the bijection $\mathcal{S}_1[i] \mapsto \mathcal{S}_2[\pi(i)]$ between the query signatures a *pattern-preserving mapping*.

Definition 15 (Similar Patterns). Given two queries q_1 and q_2. The queries use a *similar pattern* iff there is a schema mapping λ from q_1 to q_2 s.t. $\lambda(q_1)$ and q_2 are pattern-isomorphic.

Source: Gatterbauer, Dunne, On the Reasonable Effectiveness of Relational Diagrams: Explaining Relational Query Patterns and the Pattern Expressiveness of Relational Languages. SIGMOD 2024. https://doi.org/10.1145/3639316 , https://relationaldiagrams.com/

2. Pattern similarity (across different schemas)

R(A,B), S(B)

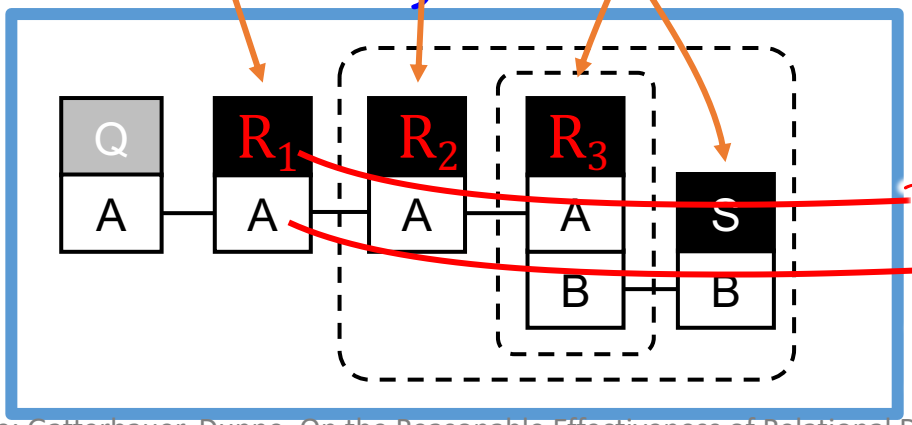
Datalog¬

$$I(x) :- R(x, \_), S(y), \neg R(x, y)$$
$$Q(x) :- R(x, \_), \neg I(x)$$

Relational Algebra

$$\pi_A R - \pi_A((\pi_A R \times S) - R)$$

Relational Diagram:



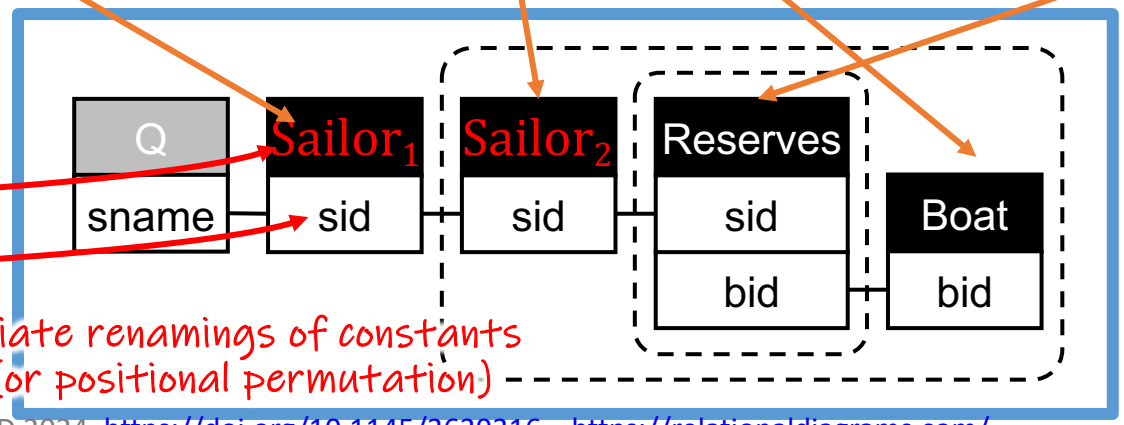
RELATIONAL PATTERN (informal):

1. Only focus on extensional atoms of a query
2. Treat each repeated table as different ("dissociated queries")
3. The logical function defined by that query and table signature is its relational pattern

What about different schemas? "Similar patterns" for queries which are not logically equivalent

Q: "*Find sailors who reserved all boats*"

$$\pi_{sid}\text{Sailor} - \pi_{sid}(\pi_{sid,bid}(\text{Sailor} \times \text{Boat}) - \pi_{sid,bid}\text{Reserves})$$



$R_1 \rightarrow Sailor_1$

$A \rightarrow sid$

via appropriate renamings of constants and names (or positional permutation)

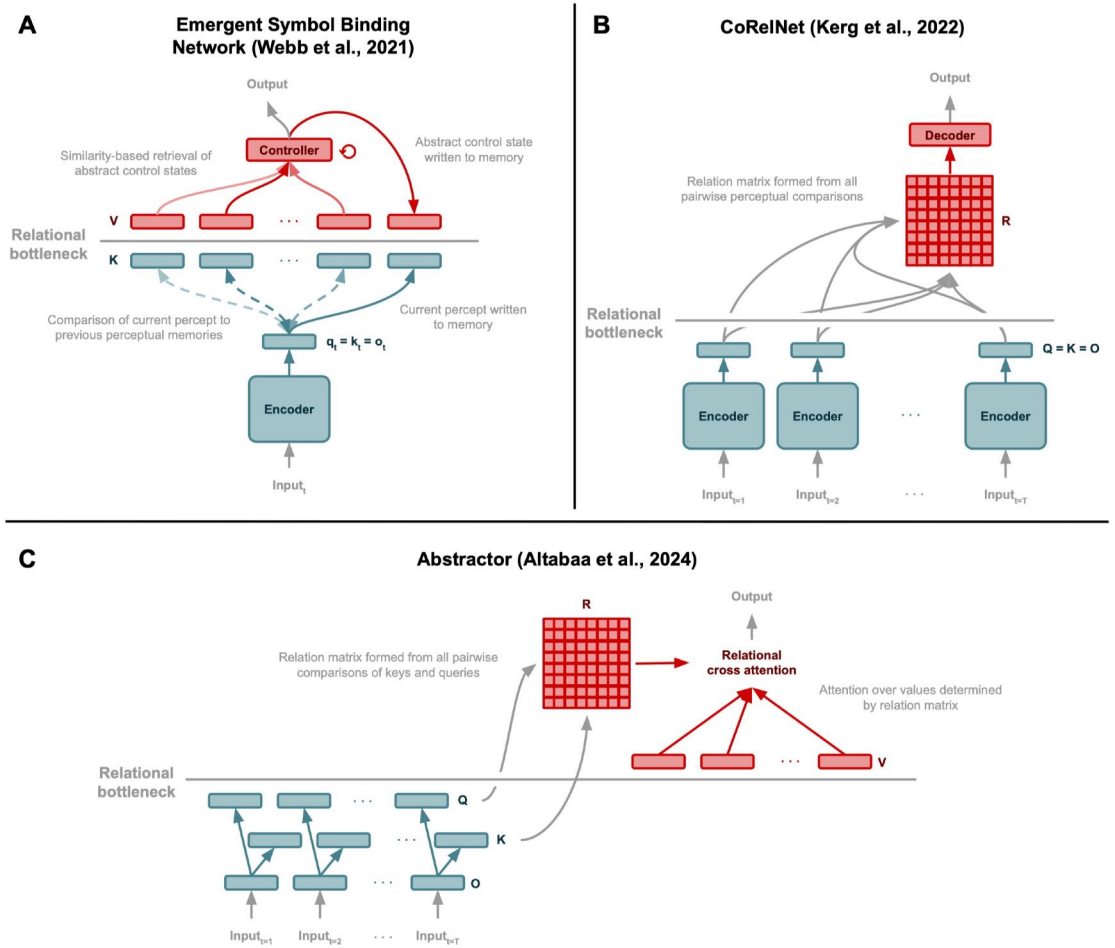


Figure 2: **Implementing the relational bottleneck.** Three neural architectures that implement the relational bottleneck. **(a)** Emergent Symbol Binding Network (ESBN) [52]. **(b)** Compositional Relation Network (CoRelNet) [53]. **(c)** Abstractor [54]. In all cases, high-dimensional inputs (e.g., images) are processed by a neural encoder (e.g., a convolutional network), yielding a set of object embeddings $\mathbf{O}$. These are projected to a set of keys $\mathbf{K}$ and queries $\mathbf{Q}$, which are then compared yielding a relation matrix $\mathbf{R}$, in which each entry is an inner product between a query and key. Abstract values $\mathbf{V}$ are isolated from perceptual inputs (the core feature of the relational bottleneck), and depend only on the relations between them.

Source: Webb, Frankland, Altabaa, Segert, Krishnamurthy, Campbell, Russin, Giallanza, Dulberg, OReilly, Lafferty, Cohen. The Relational Bottleneck as an Inductive Bias for Efficient Abstraction, 2024. Trends in Cognitive Sciences, 2024. https://doi.org/10.1016/j.tics.2024.04.001

In this review, we highlight an emerging approach that suggests a novel reconciliation of these two traditions. The central feature of this approach is an **inductive bias** that we refer to as the *relational bottleneck*: a constraint that biases neural network models to focus on relations between objects rather than the attributes of individual objects. This approach enables the data efficiency associated with symbolic cognitive models, while retaining the scalable training procedures associated with neural network models (see Box 1 for further discussion of neuro-symbolic approaches). In the

- The *relational bottleneck* principle suggests a novel way to bridge the gap. By restricting information processing to focus only on relations, the approach encourages abstract symbol-like mechanisms to emerge in neural networks.

- We present an information theoretic formulation, and review neural network architectures that implement the principle, enabling rapid learning and systematic generalization of relational patterns.

Source: Webb, Frankland, Altabaa, Segert, Krishnamurthy, Campbell, Russin, Giallanza, Dulberg, OReilly, Lafferty, Cohen. The Relational Bottleneck as an Inductive Bias for Efficient Abstraction, 2024. Trends in Cognitive Sciences, 2024. https://doi.org/10.1016/j.tics.2024.04.001
Gatterbauer, Aslam. Foundations and Applications of Information Theory: https://northeastern-datalab.github.io/cs7840/