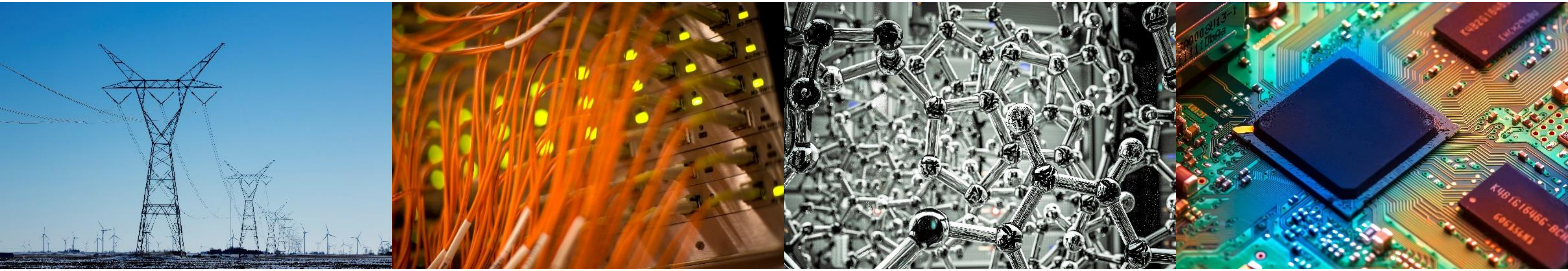
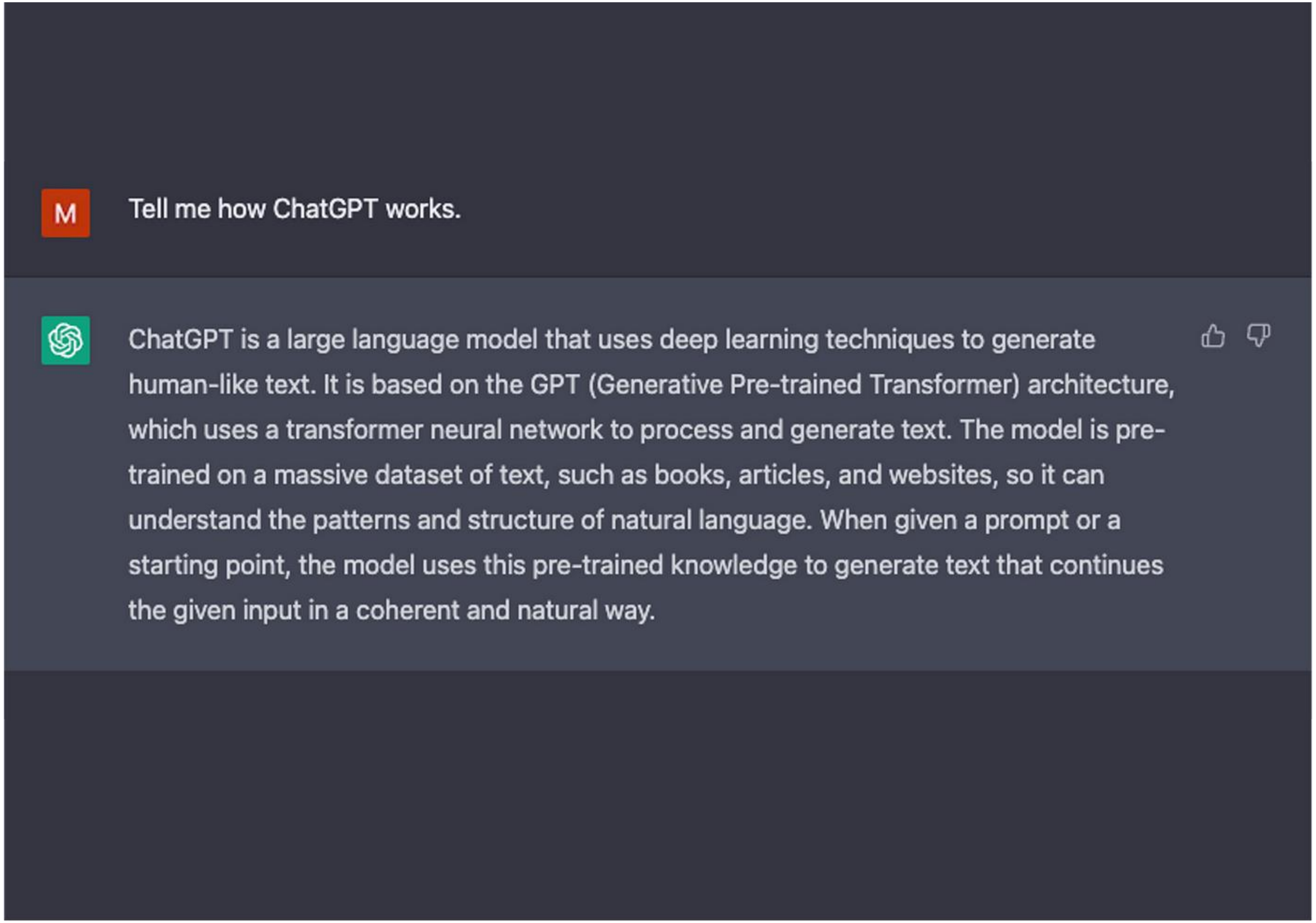
Trustworthy Generative AI

Lav R. Varshney
Founder, Kocree, Inc.
Associate Professor, University of Illinois Urbana-Champaign

6 October 2023

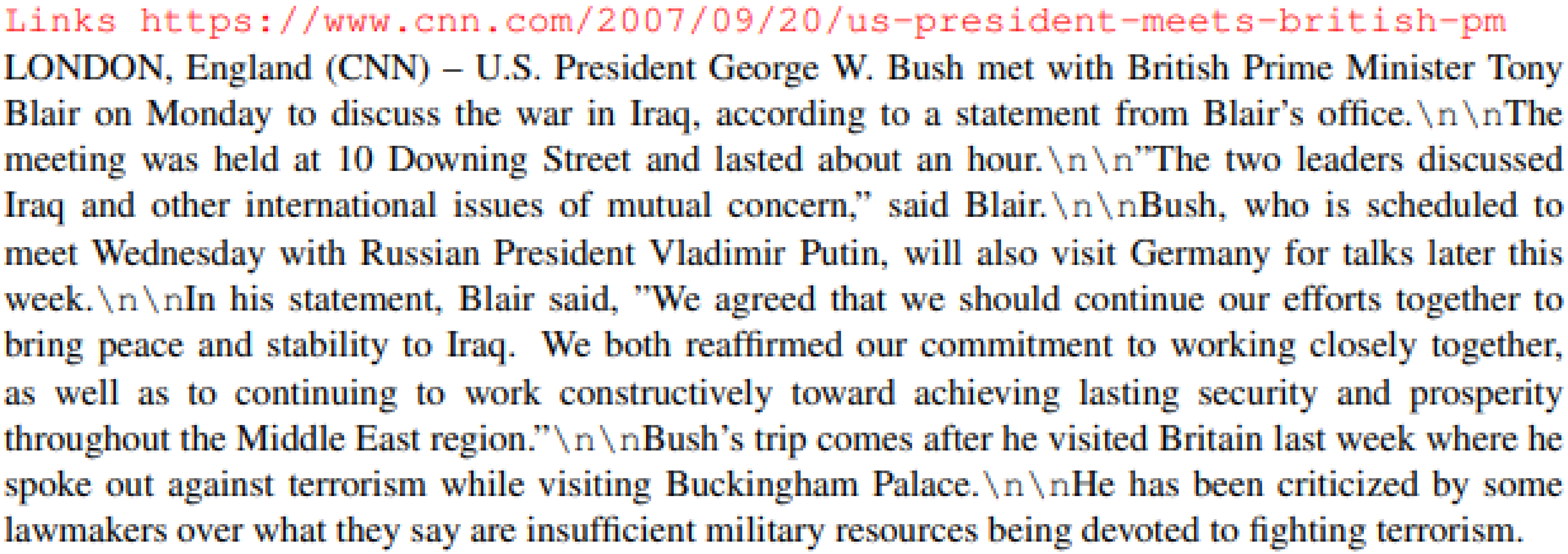


ILLINOIS
Electrical & Computer Engineering
GRAINGER COLLEGE OF ENGINEERING



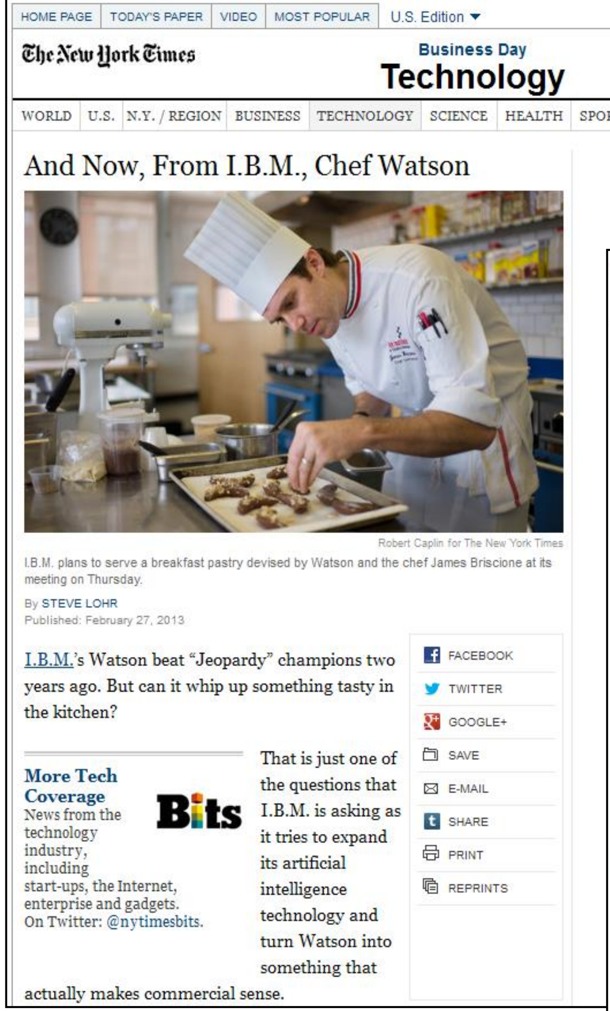
Tell me how ChatGPT works.

ChatGPT is a large language model that uses deep learning techniques to generate human-like text. It is based on the GPT (Generative Pre-trained Transformer) architecture, which uses a transformer neural network to process and generate text. The model is pre-trained on a massive dataset of text, such as books, articles, and websites, so it can understand the patterns and structure of natural language. When given a prompt or a starting point, the model uses this pre-trained knowledge to generate text that continues the given input in a coherent and natural way.



Links https://www.cnn.com/2007/09/20/us-president-meets-british-pm
LONDON, England (CNN) – U.S. President George W. Bush met with British Prime Minister Tony Blair on Monday to discuss the war in Iraq, according to a statement from Blair’s office.\n\nThe meeting was held at 10 Downing Street and lasted about an hour.\n\n”The two leaders discussed Iraq and other international issues of mutual concern,” said Blair.\n\nBush, who is scheduled to meet Wednesday with Russian President Vladimir Putin, will also visit Germany for talks later this week.\n\nIn his statement, Blair said, "We agreed that we should continue our efforts together to bring peace and stability to Iraq. We both reaffirmed our commitment to working closely together, as well as to continuing to work constructively toward achieving lasting security and prosperity throughout the Middle East region."\n\nBush’s trip comes after he visited Britain last week where he spoke out against terrorism while visiting Buckingham Palace.\n\nHe has been criticized by some lawmakers over what they say are insufficient military resources being devoted to fighting terrorism.

[N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, “CTRL: A Conditional Transformer Language Model for Controllable Generation,” Sept. 2019.]



The New York Times

Business Day
Technology

And Now, From I.B.M., Chef Watson

Robert Caplin for The New York Times

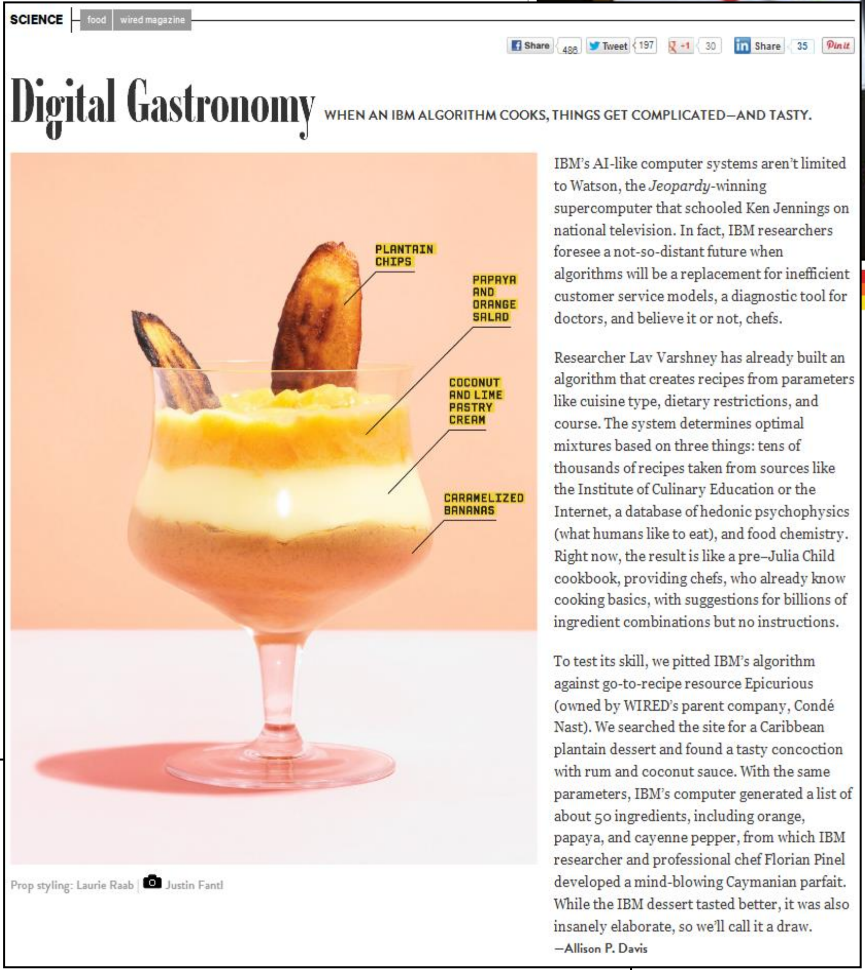
I.B.M. plans to serve a breakfast pastry devised by Watson and the chef James Briscione at its meeting on Thursday.

By STEVE LOHR
Published: February 27, 2013

I.B.M.'s Watson beat "Jeopardy" champions two years ago. But can it whip up something tasty in the kitchen?

More Tech Coverage
News from the technology industry, including start-ups, the Internet, enterprise and gadgets.
On Twitter: @nytimesbits.

That is just one of the questions that I.B.M. is asking as it tries to expand its artificial intelligence technology and turn Watson into something that actually makes commercial sense.



Digital Gastronomy

WHEN AN IBM ALGORITHM COOKS, THINGS GET COMPLICATED—AND TASTY.

Prop styling: Laurie Raab | Justin Fantl

IBM's AI-like computer systems aren't limited to Watson, the *Jeopardy*-winning supercomputer that schooled Ken Jennings on national television. In fact, IBM researchers foresee a not-so-distant future when algorithms will be a replacement for inefficient customer service models, a diagnostic tool for doctors, and believe it or not, chefs.

Researcher Lav Varshney has already built an algorithm that creates recipes from parameters like cuisine type, dietary restrictions, and course. The system determines optimal mixtures based on three things: tens of thousands of recipes taken from sources like the Institute of Culinary Education or the Internet, a database of hedonic psychophysics (what humans like to eat), and food chemistry. Right now, the result is like a pre–Julia Child cookbook, providing chefs, who already know cooking basics, with suggestions for billions of ingredient combinations but no instructions.

To test its skill, we pitted IBM's algorithm against go-to-recipe resource Epicurious (owned by WIRED's parent company, Condé Nast). We searched the site for a Caribbean plantain dessert and found a tasty concoction with rum and coconut sauce. With the same parameters, IBM's computer generated a list of about 50 ingredients, including orange, papaya, and cayenne pepper, from which IBM researcher and professional chef Florian Pinel developed a mind-blowing Caymanian parfait. While the IBM dessert tasted better, it was also insanely elaborate, so we'll call it a draw.

—Allison P. Davis

IBM'S TASTE MASTER

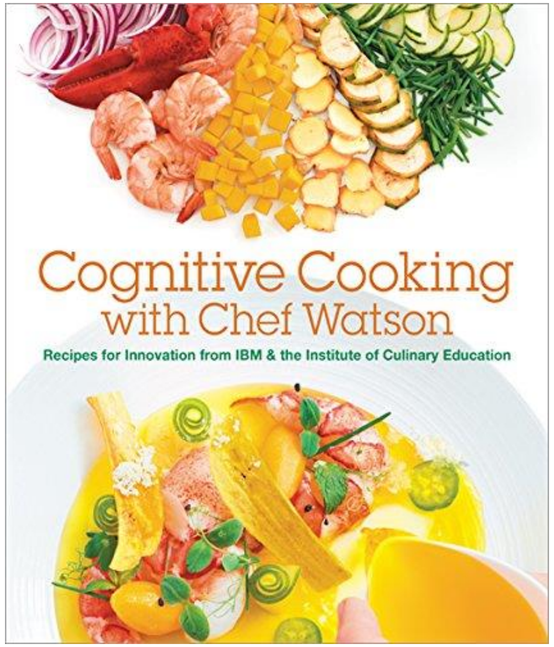
COGNITIVE COMPUTING TAKES ON A NEW FRONTIER: MEAL PLANNING
BY VALERIE ROSS

PHOTOGRAPH BY David Yellen

76 | JUN 2013 | NORTH AMERICAN | SPECTRUM.IEEE.ORG

[*The New York Times*, 27 Feb. 2013]
[*San Jose Mercury News*, 28 Feb. 2013]
[*IEEE Spectrum*, 31 May 2013]
[*Wired*, 1 Oct. 2013]

IBM COGNITIVE COOKING
EAT
TRENDING for Tomorrow
Now Serving
IBM.COM/COGNITIV ING
IBM
Pure and Simple.


Cognitive Cooking with Chef Watson
Recipes for Innovation from IBM & the Institute of Culinary Education

bear naked
Custom Made
GRANOLA
prepared with
IBM Chef Watson™

McCormick
ONE SKILLET
TUSCAN
Chicken & Vegetables
Oregano, Lemon, Sun-Dried Tomato
ALL DONE, IN ONE!

McCormick
ONE SHEET PAN
BOURBON PORK
Tenderloin & Vegetables
NATURALLY FLAVORED
Garlic, Brown Sugar, Black Pepper
ALL DONE, IN ONE!
SEASONING MIX

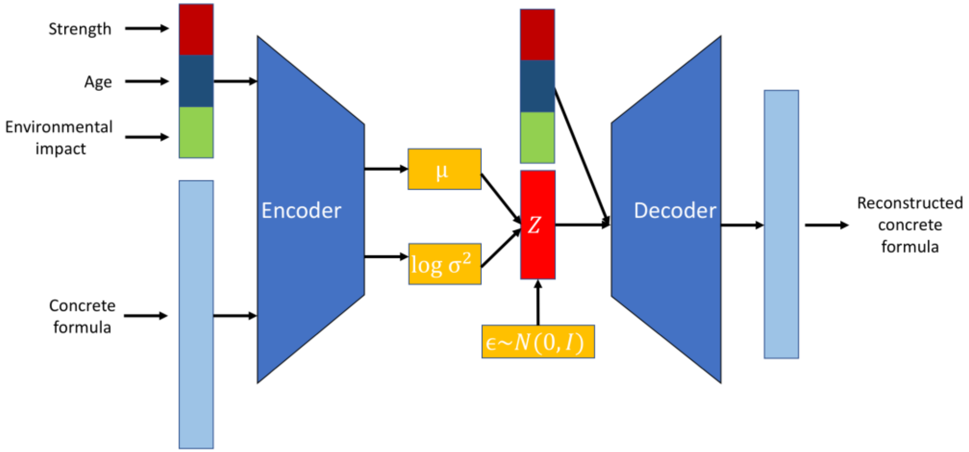
Concrete that has half as much embodied carbon and is much stronger

- 8% of worldwide CO_2 emissions caused by cement production
- Reduce environmental impacts of construction materials while complying with product specifications

- UCI ML repository concrete strength dataset + environmental impact evaluated using the Cement Sustainability Initiative's Environmental Product Declaration tool:
  - 1030 instances
  - 8 input variables (composition)
  - 1 (compressive strength)
  - 12 (environmental impact) output variables

- Train a conditional generative neural network model to be able to create novel formulations of concrete

Conditional Variational Autoencoder (CVAE)

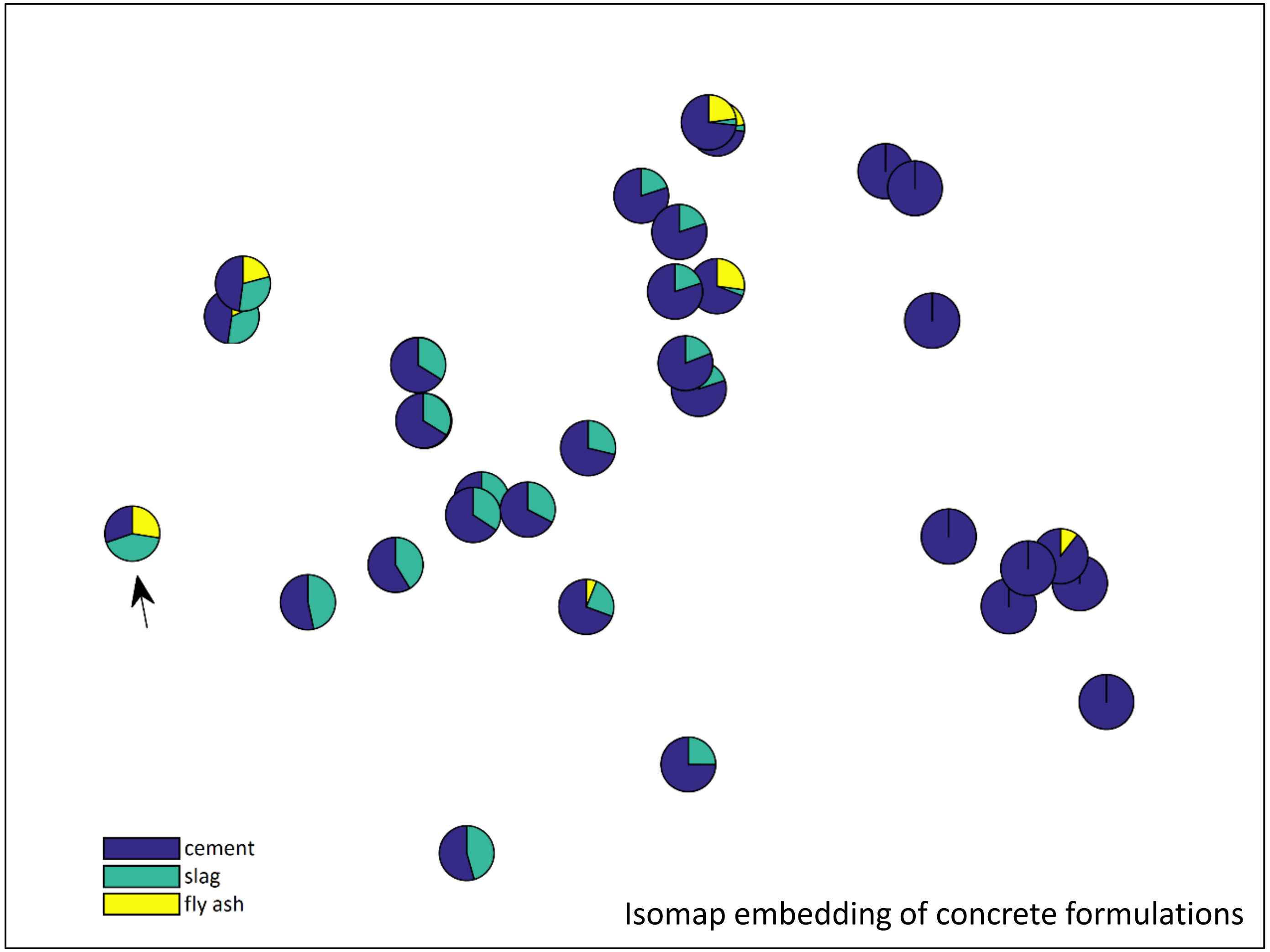


Strength	$[0,1]$
Age	$\{0,1\}^6$
Environmental Impact	$[0,1]^{12}$
Concrete formula	$[0,1]^7$

Concrete that has half as much embodied carbon and is much stronger

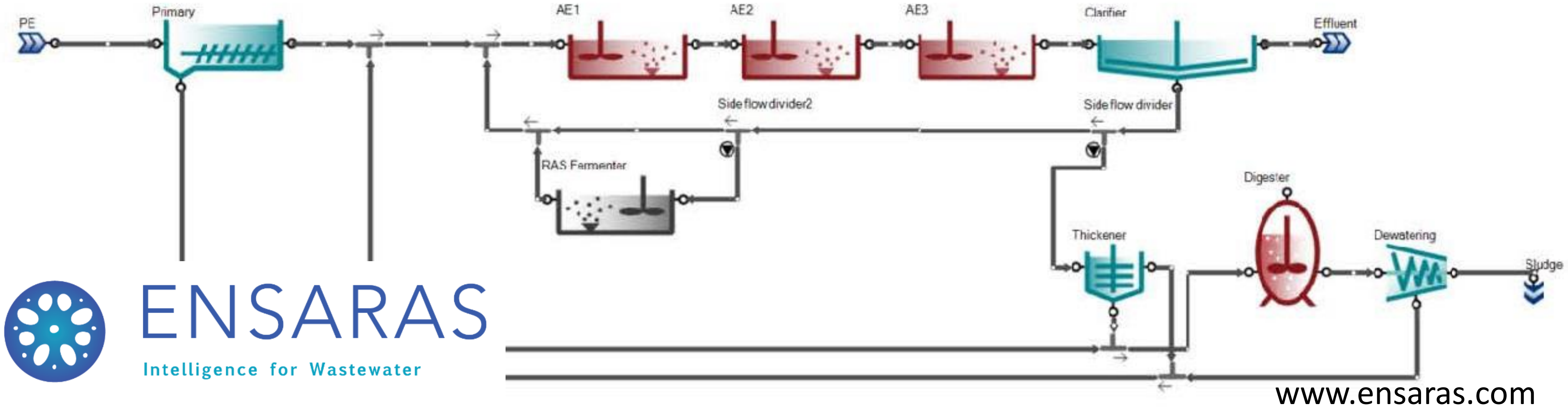
Stronger and more than 50% reduction in carbon emissions

DeKalb data center that has been constructed



Isomap embedding of concrete formulations

[X. Ge, R. T. Goodwin, H. Yu, P. Romero, O. Abdelrahman, A. Sudhalkar, J. Kusuma, R. Cialdella, N. Garg, and L. R. Varshney, “Accelerated Design and Deployment of Low-Carbon Concrete for Data Centers,” in *Proc. 5th ACM SIGCAS Conf. Computing and Sustainable Societies (COMPASS ’22)*, Seattle, July 2022.]

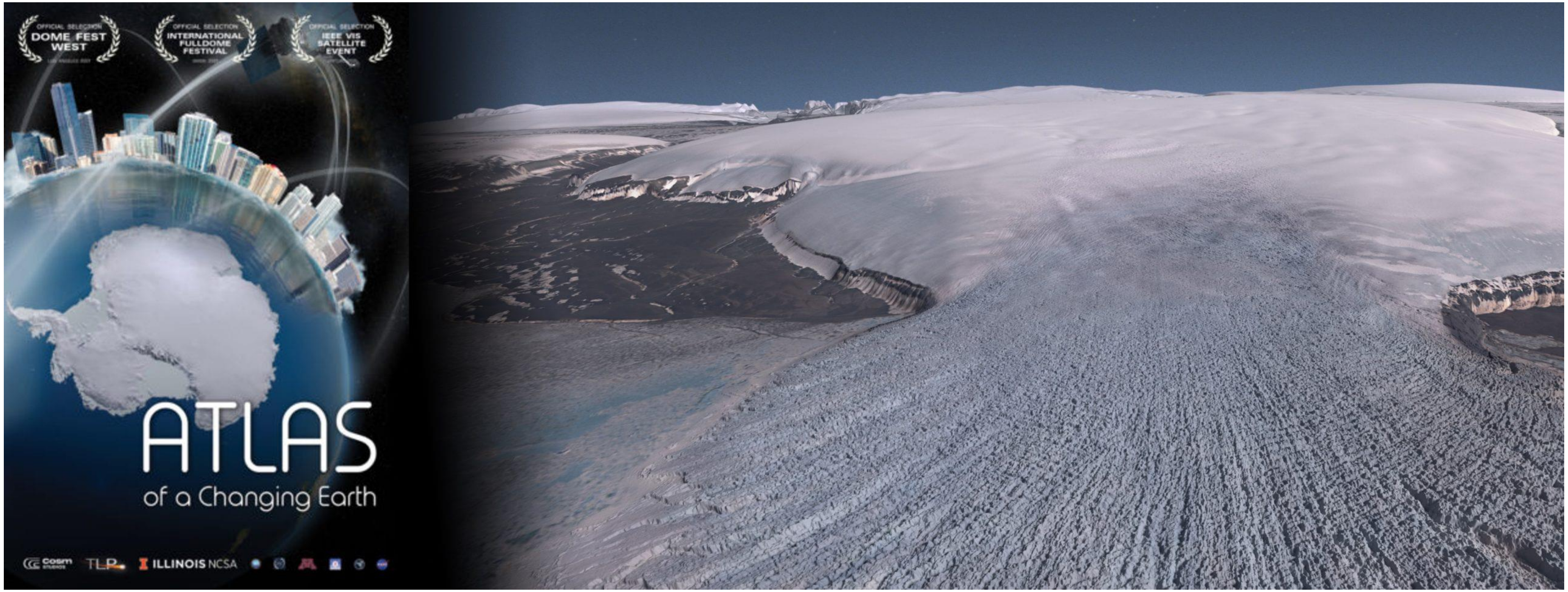


ENSARAS

Intelligence for Wastewater

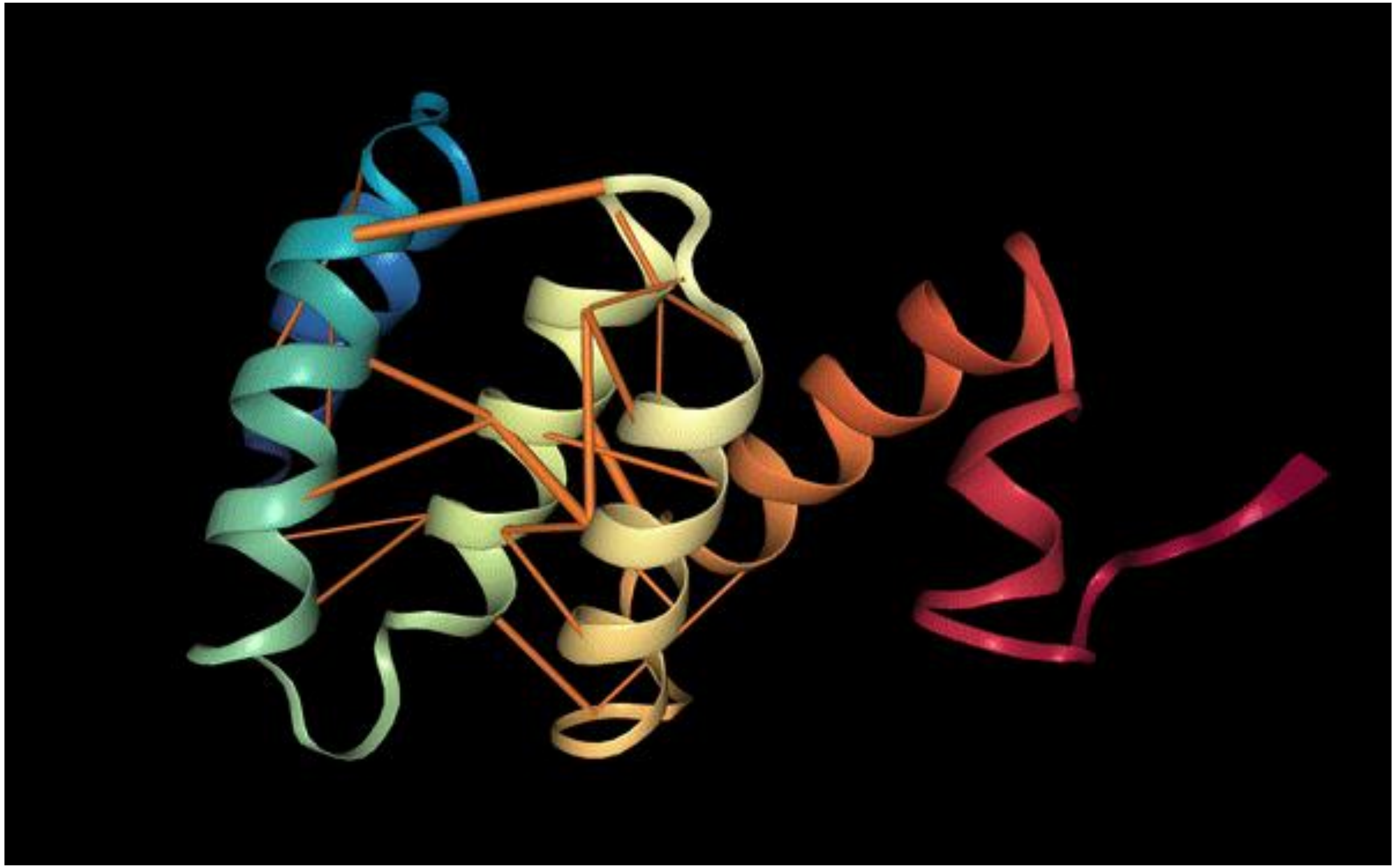
www.ensaras.com

Artificial Weather Generators

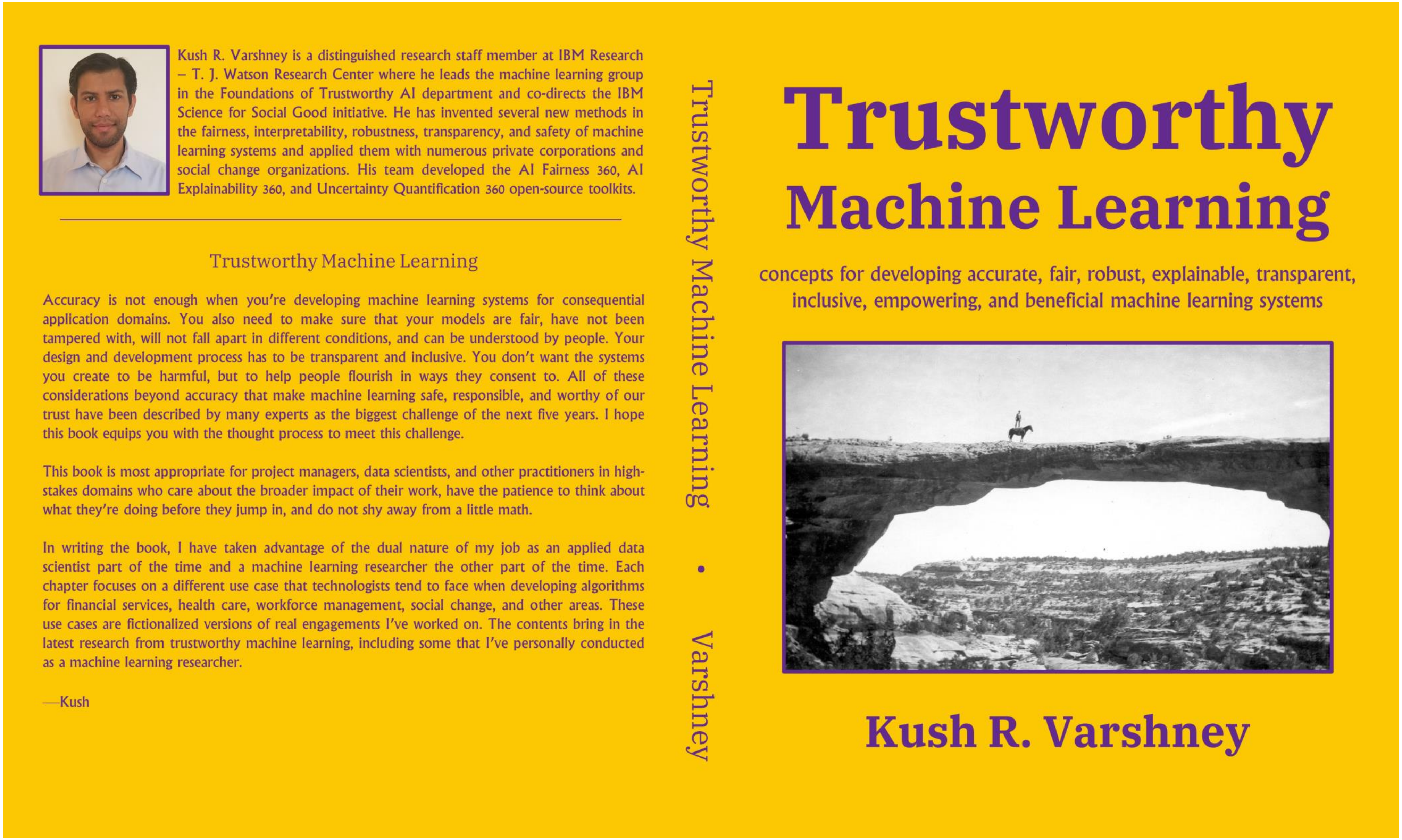

OFFICIAL SELECTION DOME FEST WEST
OFFICIAL SELECTION INTERNATIONAL FULLDOME FESTIVAL
OFFICIAL SELECTION IEEE VIS SATELLITE EVENT
ATLAS
of a Changing Earth
cosm
TLP
ILLINOIS NCSA

[A. Jain, D. Oliveira, A. Sharma, L. R. Varshney, C. Watson, K. Weldemariam, D. Wuebbles, and B. Zadrozny, “Toward an AI-based Framework for Accelerated Discovery of Climate Impacts on Agriculture,” presented at *AAAI Fall Symposium on AI Meets Food Security: Intelligent Approaches for Climate-Aware Agriculture*, Nov. 2021.]

Protein Language Models: Using AI to Generate Proteins



[J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani, “BERTology Meets Biology: Interpreting Attention in Protein Language Models,” in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, May 2021.]



Kush R. Varshney is a distinguished research staff member at IBM Research – T. J. Watson Research Center where he leads the machine learning group in the Foundations of Trustworthy AI department and co-directs the IBM Science for Social Good initiative. He has invented several new methods in the fairness, interpretability, robustness, transparency, and safety of machine learning systems and applied them with numerous private corporations and social change organizations. His team developed the AI Fairness 360, AI Explainability 360, and Uncertainty Quantification 360 open-source toolkits.

Trustworthy Machine Learning

Accuracy is not enough when you're developing machine learning systems for consequential application domains. You also need to make sure that your models are fair, have not been tampered with, will not fall apart in different conditions, and can be understood by people. Your design and development process has to be transparent and inclusive. You don't want the systems you create to be harmful, but to help people flourish in ways they consent to. All of these considerations beyond accuracy that make machine learning safe, responsible, and worthy of our trust have been described by many experts as the biggest challenge of the next five years. I hope this book equips you with the thought process to meet this challenge.

This book is most appropriate for project managers, data scientists, and other practitioners in high-stakes domains who care about the broader impact of their work, have the patience to think about what they're doing before they jump in, and do not shy away from a little math.

In writing the book, I have taken advantage of the dual nature of my job as an applied data scientist part of the time and a machine learning researcher the other part of the time. Each chapter focuses on a different use case that technologists tend to face when developing algorithms for financial services, health care, workforce management, social change, and other areas. These use cases are fictionalized versions of real engagements I've worked on. The contents bring in the latest research from trustworthy machine learning, including some that I've personally conducted as a machine learning researcher.

—Kush

Trustworthy Machine Learning • Varshney

Trustworthy Machine Learning

concepts for developing accurate, fair, robust, explainable, transparent, inclusive, empowering, and beneficial machine learning systems

Kush R. Varshney

An ethical framework from biomedicine

Beauchamp and Childress

Transfer to engineering so as to capture utilitarian and rights-based approaches to ethical thinking in a simple manner

- Justice: The principle of fairness and equality among individuals
- Beneficence: The principle of acting with the best interests of others in mind
- Non-maleficence: The principle that "above all, do no harm," as in the Hippocratic Oath
- Respect for Autonomy: The principle that individuals should have the right to make their own choices

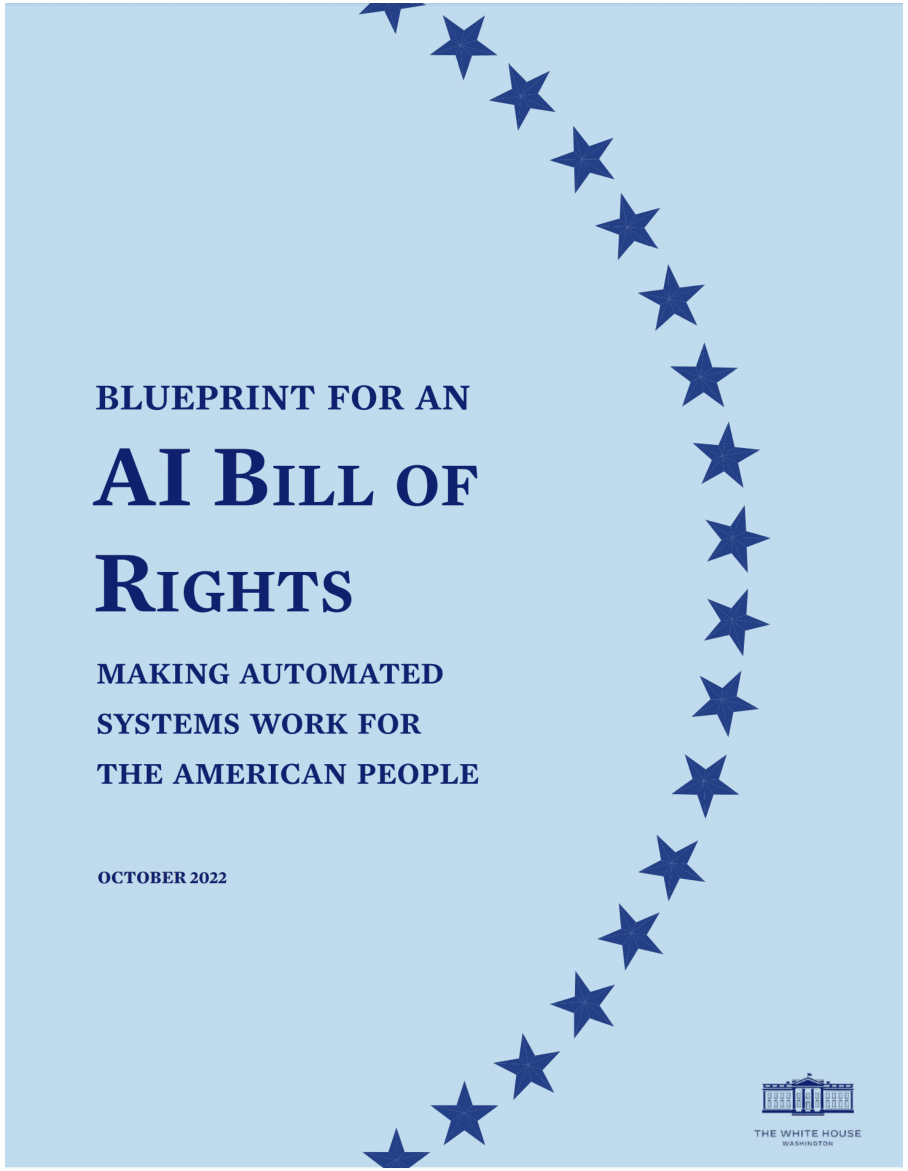
(All of these principles should, prima facie, be held and when in conflict should be given equal weight)

[L. R. Varshney, "Engineering for Problems of Excess," in *Proc. 2014 IEEE Int. Symp. Ethics in Engineering, Science, and Technology*, May 2014.]

Technology Policy

Technology

Policy



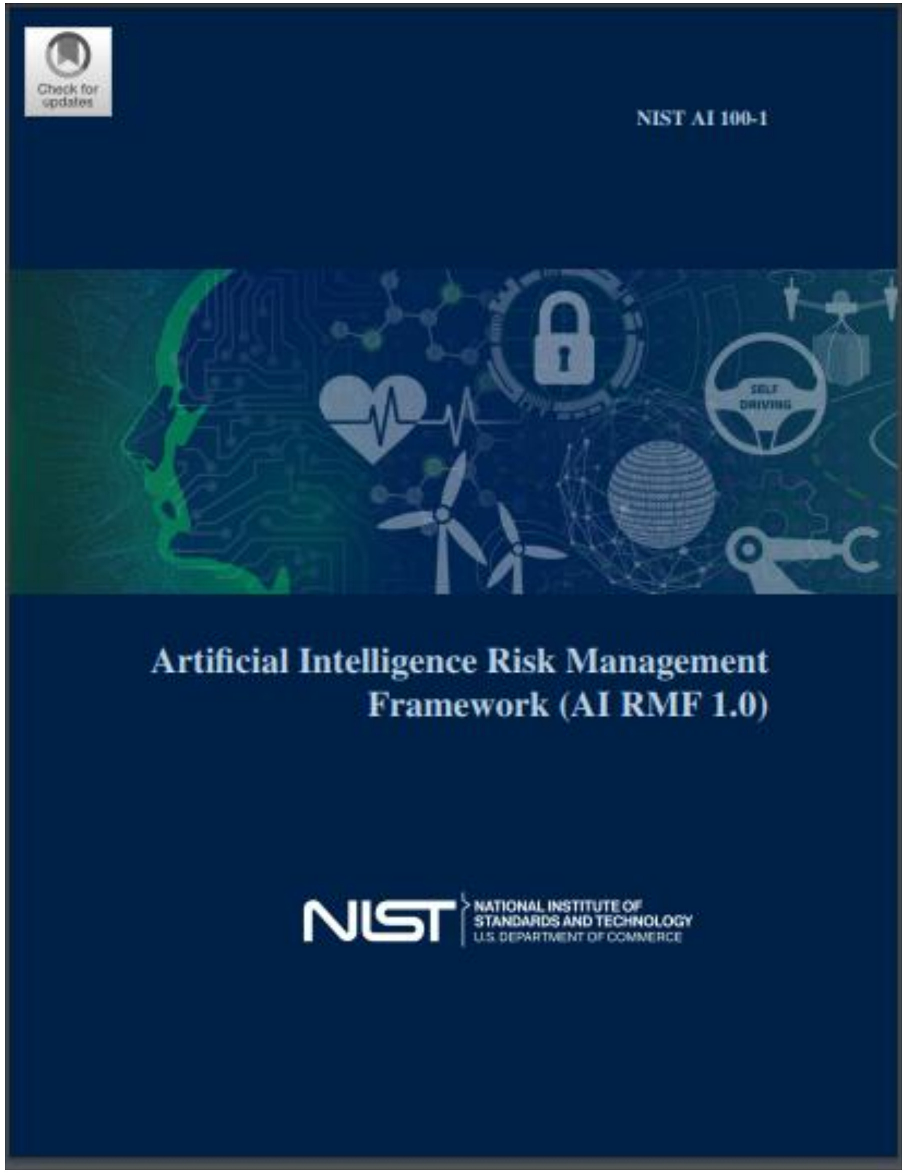
BLUEPRINT FOR AN

AI BILL OF RIGHTS

MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE

OCTOBER 2022

THE WHITE HOUSE
WASHINGTON

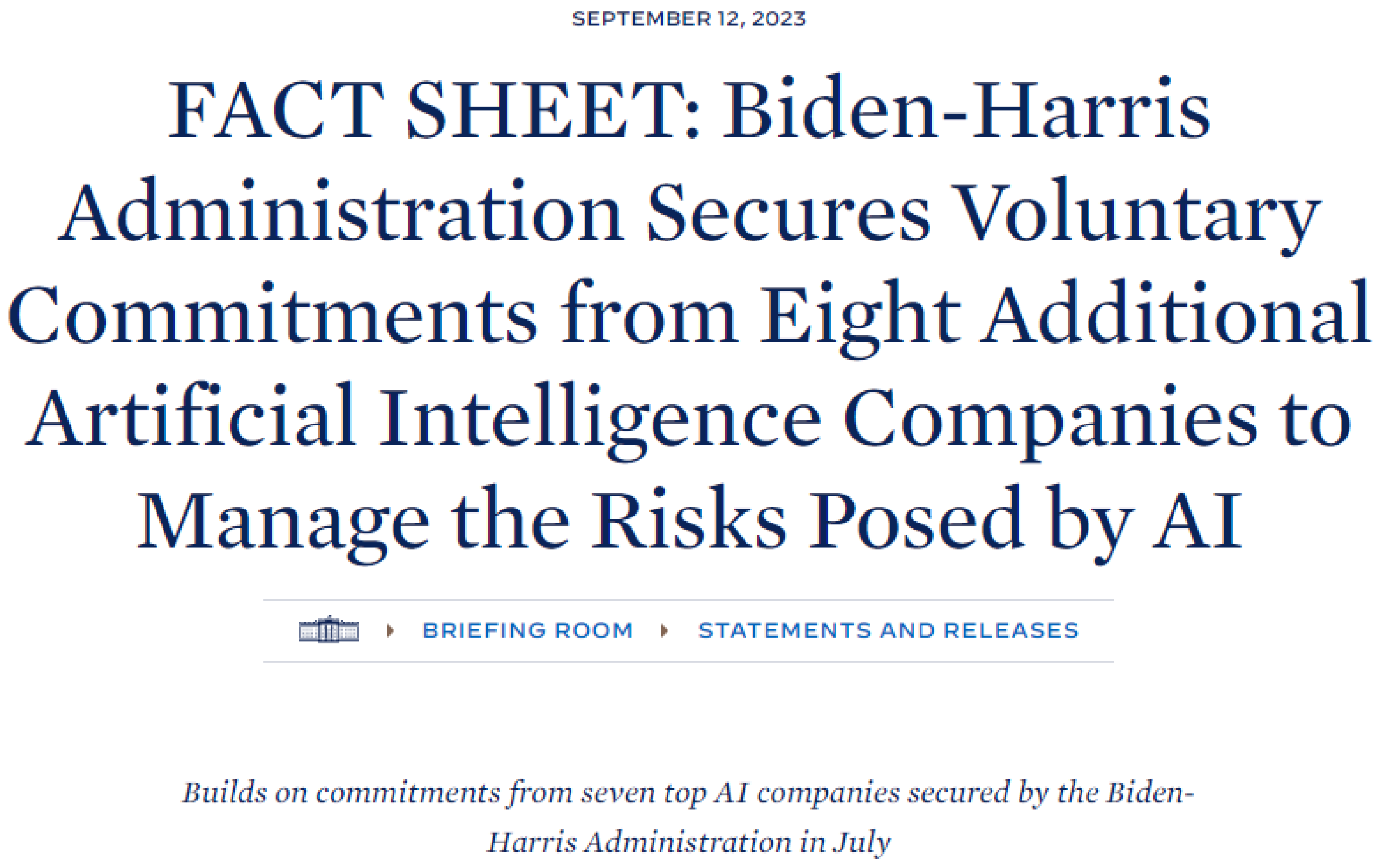


NIST AI 100-1

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Amazon, Anthropic, Google, Inflection, Meta, Microsoft, OpenAI, Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, and Stability



SEPTEMBER 12, 2023

FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI

Builds on commitments from seven top AI companies secured by the Biden-Harris Administration in July

- The companies commit to internal and external security testing of their AI systems before their release.
- The companies commit to sharing information across the industry and with governments, civil society, and academia on managing AI risks.

- The companies commit to investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights.
- The companies commit to facilitating third-party discovery and reporting of vulnerabilities in their AI systems.

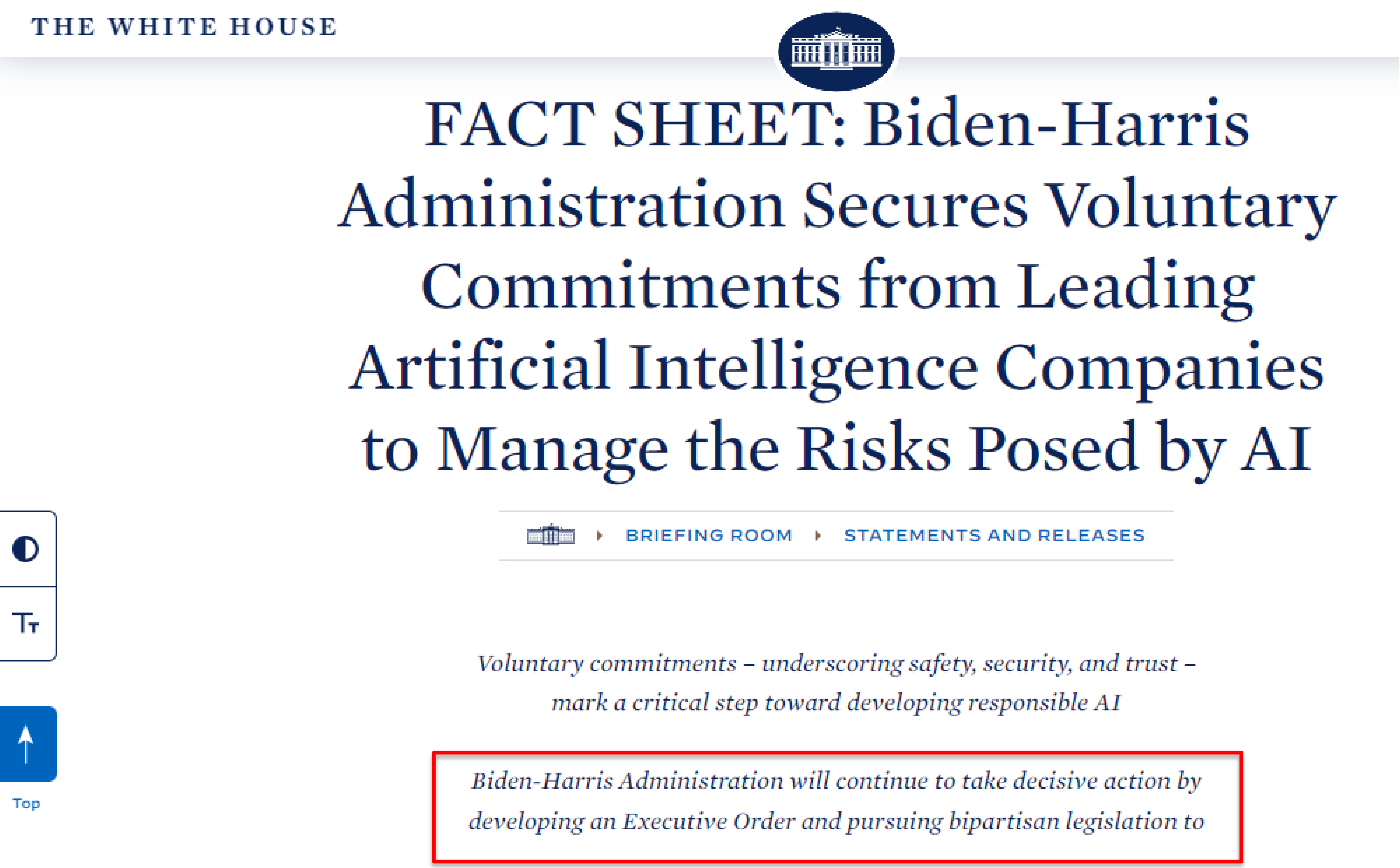
- The companies commit to developing robust technical mechanisms to ensure that users know when content is AI generated, such as a watermarking system.
- The companies commit to publicly reporting their AI systems' capabilities, limitations, and areas of appropriate and inappropriate use.

- The companies commit to prioritizing research on the societal risks that AI systems can pose, including on avoiding harmful bias and discrimination, and protecting privacy.
- The companies commit to develop and deploy advanced AI systems to help address society’s greatest challenges.

FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI

Voluntary commitments – underscoring safety, security, and trust – mark a critical step toward developing responsible AI

Biden-Harris Administration will continue to take decisive action by developing an Executive Order and pursuing bipartisan legislation to



FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI

Voluntary commitments – underscoring safety, security, and trust – mark a critical step toward developing responsible AI

Biden-Harris Administration will continue to take decisive action by developing an Executive Order and pursuing bipartisan legislation to

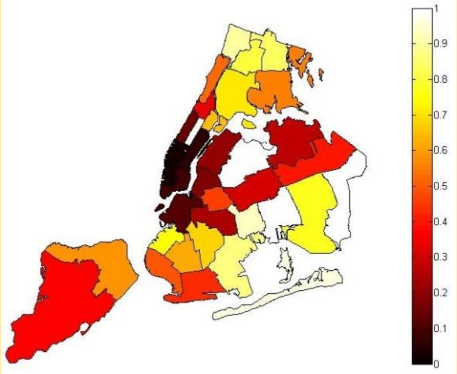
ROOM L-51 13/096
Department of State
Shani Spivak

AI for good

Obesity: Strong association of obesity rates in urban neighborhoods with social capital measures (venues for interaction as per Foursquare)

- **regression models**

[*Data for Good Exchange (D4GX)*, 2015]



Urban Blight: Rank vacant parcels according to likelihoods of occupied status and neighborhood impact

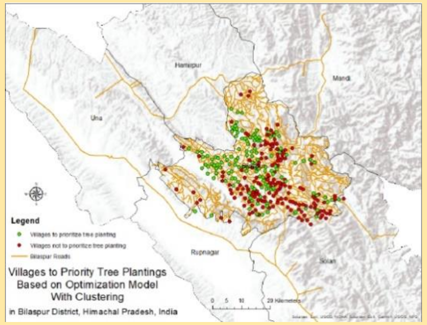
- **bipartite ranking + spatiotemporal modeling**

[*Technological Forecasting and Social Change*, 2014]

Sustainable Farming: Redistribution of permits in Himalayas can significantly improve sustainability (environmental/economic) of timber farming

- **network flow optimization**

[*Data for Good Exchange (D4GX)*, 2018]


Hamirpur
Mandi
Una
Rupnagar
Solan
Legend
Villages to prioritize tree planting
Villages not to prioritize tree planting
Bilaspur Roads
Villages to Priority Tree Plantings
Based on Optimization Model
With Clustering
in Bilaspur District, Himachal Pradesh, India
0
5
10
20 Kilometers

Sustainable/Healthy Food: Computationally create culinary recipes according to perceived flavor and novelty using ingredients such as algae protein

- **computational creativity and hedonic perception**

[*Good Food Conference*, 2018]

Technology

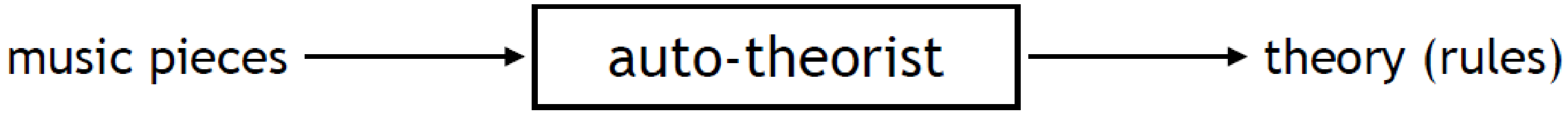
Policy

> “Shannon himself told me that he believes the most promising new developments in information theory will come from work on very complex machines, especially from research into artificial intelligence.”
>
> [J. Campbell, *Grammatical Man,* 1982]

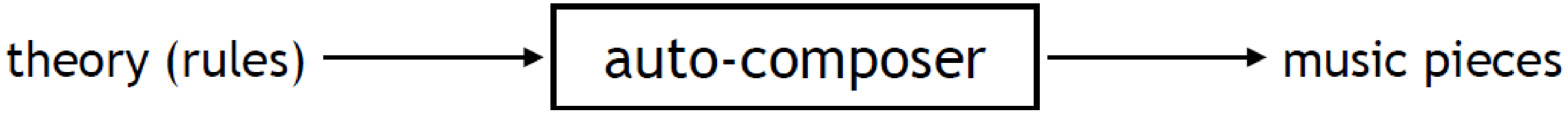
[L. R. Varshney, “Mathematizing the World,” *Issues in Science and Technology*, vol. 35, no. 2, pp. 93–95, Winter 2019.]

Automatic knowledge discovery (An automatic music theorist)

A way to learn the principles of quality (laws of music theory)


music pieces → auto-theorist → theory (rules)

Computational creativity algorithms for music composition


theory (rules) → auto-composer → music pieces

Dimensions of interpretability [Selbst and Barocas, 2018]

- What sets machine learning models apart from other decision-making mechanisms are their *inscrutability* and *nonintuitiveness*
  - Inscrutability suggests that models available for direct inspection may defy understanding
  - Nonintuitiveness suggests that even where models are understandable, they may rest on apparent statistical relationships that defy intuition
  - Most extant work on interpretable ML/AI only addresses inscrutability, but not nonintuitiveness

- Dealing with inscrutability requires providing a sensible description of rules; addressing nonintuitiveness requires providing satisfying explanation for why the rules are what they are

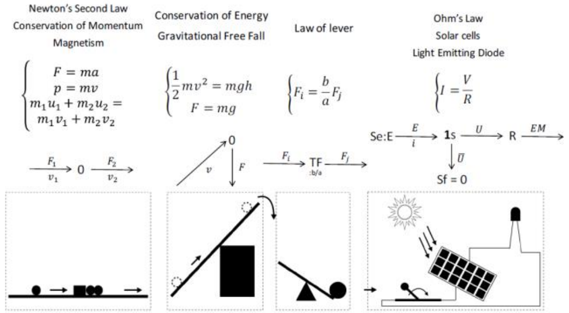
For numerous settings, may need technical solutions to both inscrutability and nonintuitiveness

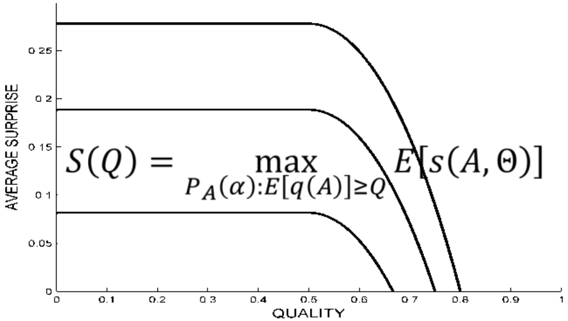
Human-interpretable concept learning

- Learn laws of nature from raw data, e.g. for scientific discovery or for complex systems where epistemic uncertainty (unknown unknowns) can be dangerous [AI safety]

- Learn what black box systems do, whether human or machine, not just in terms of the statistical nature of bias but also the rules that govern behavior [AI ethics]

- Learn principles of human culture, e.g. what are the laws of music theory that make Bach's chorales what they are or psychophysical principles of flavor in world cuisines [AI creativity]





$$S(Q) = \max_{P_A(\alpha):E[q(A)]\geq Q} E[s(A,\Theta)]$$

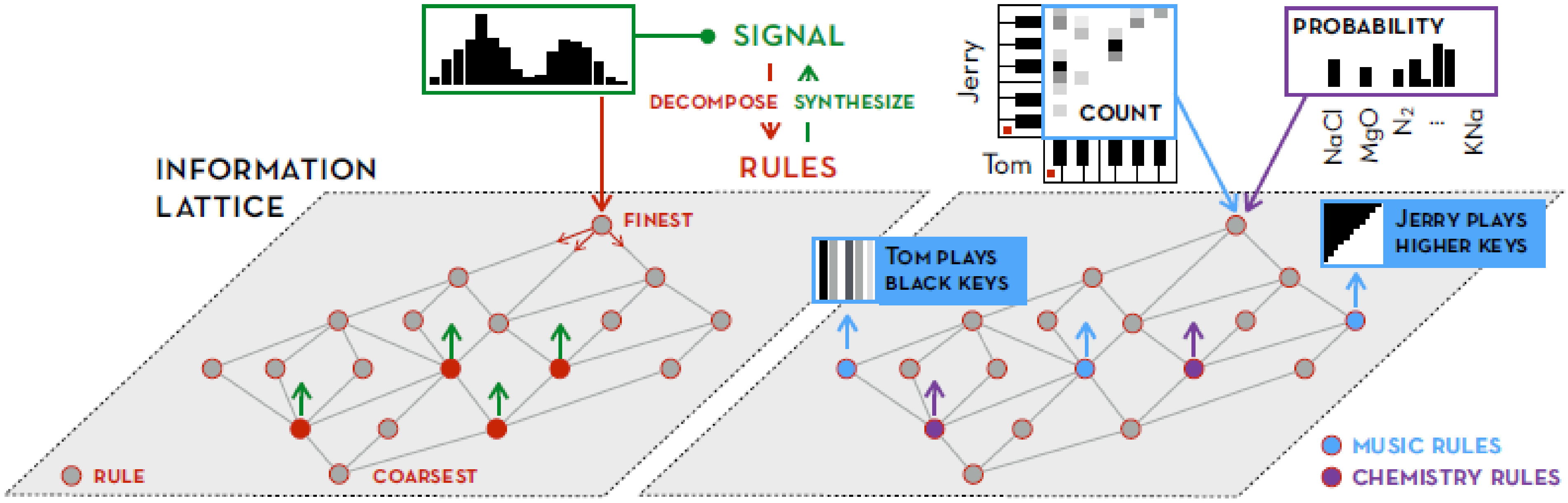
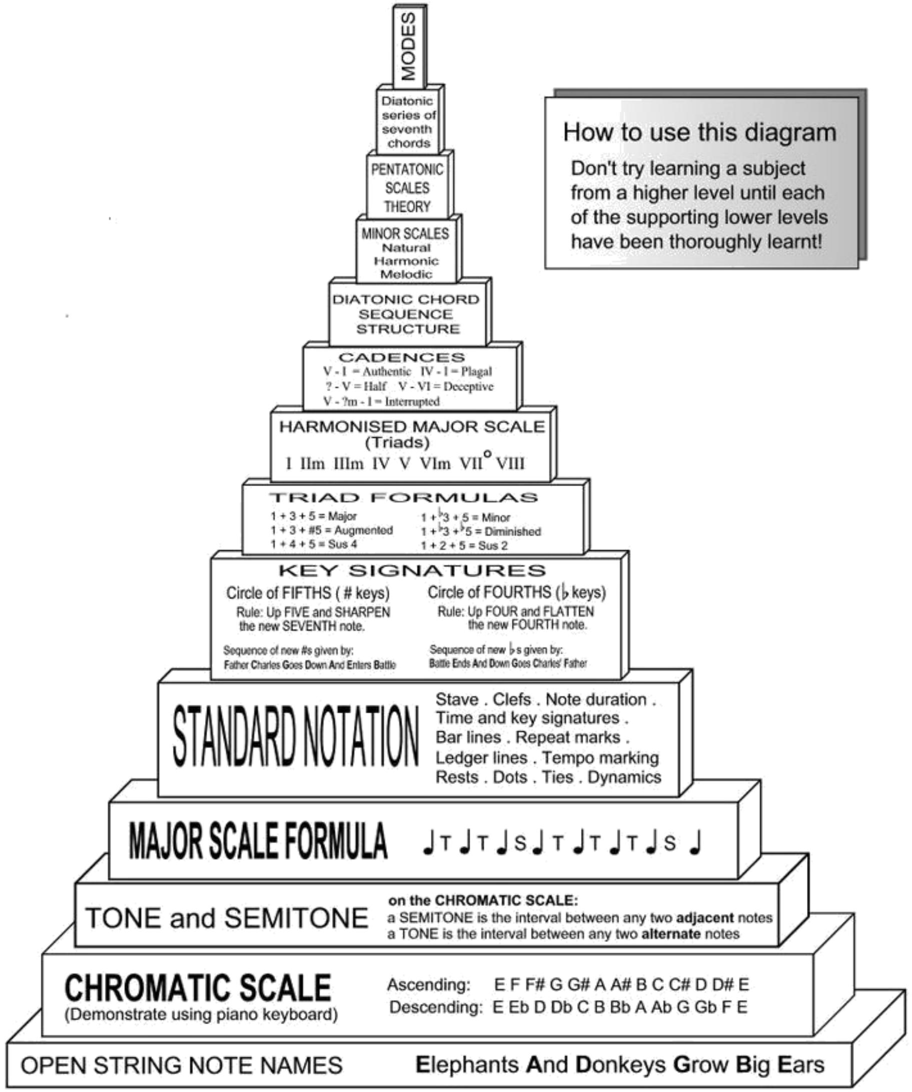


Figure 1: ILL's main idea: decompose the signal into rules that are individually simple but collectively expressive. A lattice is first constructed regardless of the signal (prior-driven), yet the same lattice may be later used to learn rules (data-driven) of signals from different topics, e.g. music and chemistry.

[H. Yu, J. A. Evans, and L. R. Varshney, "Information Lattice Learning," *Journal of Artificial Intelligence Research*, vol. 77, pp. 971–1019, July 2023.]

Learn human-interpretable concept *hierarchies* (not just rules)


MODES
Diatonic series of seventh chords
PENTATONIC SCALES THEORY
MINOR SCALES
Natural
Harmonic
Melodic
DIATONIC CHORD SEQUENCE STRUCTURE
CADENCES
V - I = Authentic IV - I = Plagal
? - V = Half V - VI = Deceptive
V - ?m - I = Interrupted
HARMONISED MAJOR SCALE (Triads)
I IIm IIIm IV V VIm VII° VIII
TRIAD FORMULAS
1 + 3 + 5 = Major
1 + 3 + #5 = Augmented
1 + 4 + 5 = Sus 4
1 + ♭3 + 5 = Minor
1 + ♭3 + ♭5 = Diminished
1 + 2 + 5 = Sus 2
KEY SIGNATURES
Circle of FIFTHS (# keys)
Rule: Up FIVE and SHARPEN the new SEVENTH note.
Sequence of new #s given by:
Father Charles Goes Down And Enters Battle
Circle of FOURTHS (♭ keys)
Rule: Up FOUR and FLATTEN the new FOURTH note.
Sequence of new ♭s given by:
Battle Ends And Down Goes Charles' Father
STANDARD NOTATION
Stave . Clefs . Note duration .
Time and key signatures .
Bar lines . Repeat marks .
Ledger lines . Tempo marking
Rests . Dots . Ties . Dynamics
MAJOR SCALE FORMULA
♩T ♩T ♩S ♩T ♩T ♩T ♩S ♩
TONE and SEMITONE
on the CHROMATIC SCALE:
a SEMITONE is the interval between any two adjacent notes
a TONE is the interval between any two alternate notes
CHROMATIC SCALE
(Demonstrate using piano keyboard)
Ascending: E F F# G G# A A# B C C# D D# E
Descending: E Eb D Db C B Bb A Ab G Gb F E
OPEN STRING NOTE NAMES
Elephants And Donkeys Grow Big Ears
How to use this diagram
Don't try learning a subject from a higher level until each of the supporting lower levels have been thoroughly learnt!

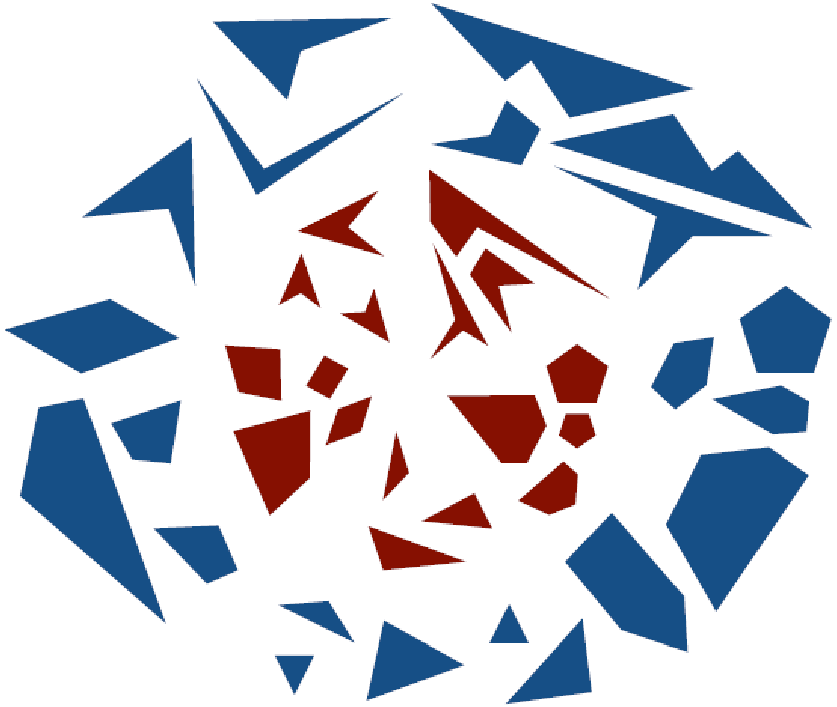
[http://www.teachguitar.com/content/tmpyramid.htm]

> "Fundamentally, most current deep-learning based language models represent sentences as mere sequences of words, whereas Chomsky has long argued that language has a hierarchical structure, in which larger structures are recursively constructed out of smaller components."
>
> – Gary Marcus [*arXiv:1801.00631*]

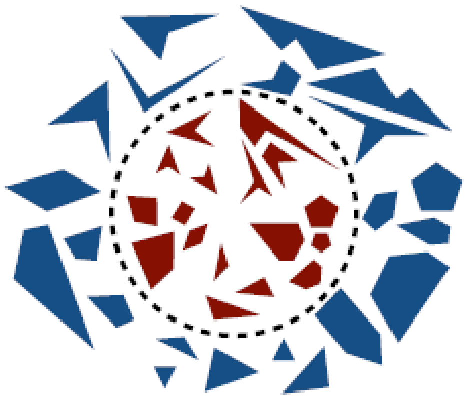
Automatic concept learning

Concept learning is phase before any task solving/performing

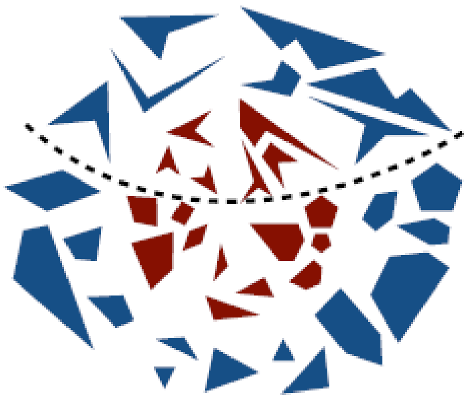
- *Self-exploration*: ultimate goal is learning domain concepts/knowledge from universal priors—priors that encode no domain knowledge
  - Group-theoretic foundations and generalization of Shannon's information lattice
- *Self-explanation*: aim for not only the learned results but also the entire leaning process to be human-interpretable
  - Iterative student-teacher architecture for learning algorithm, which produces interpretable hierarchy of interpretable concepts (with a particular mechanistic cause: symmetry) and its trace



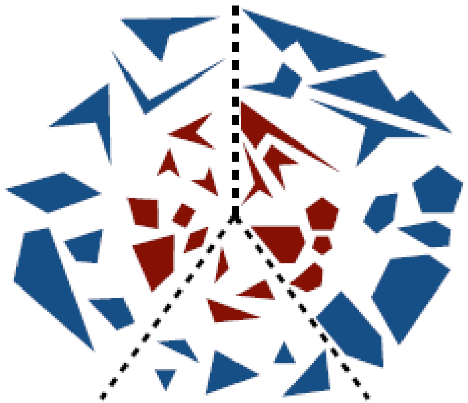
[H. Yu, I. Mineyev, and L. R. Varshney, “A Group-Theoretic Approach to Computational Abstraction: Symmetry-Driven Hierarchical Clustering,” *Journal of Machine Learning Research*, vol. 24, no. 47, pp. 1–61, 2023.]



{red, blue}



{convex, concave}



{trigon, tetragon, pentagon}

Representation: Data space

Data space: (X, p_X) or (X, p) for short

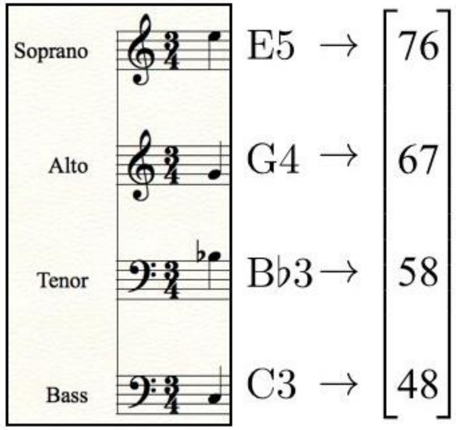
- Assume a data point $x \in X$ is an i.i.d. sample drawn from a probability distribution p
- However, the data distribution p (or an estimate of it) is *known*
- The goal here is not to estimate p but to *explain* it

Chord space: $X = \mathbb{Z}^4$

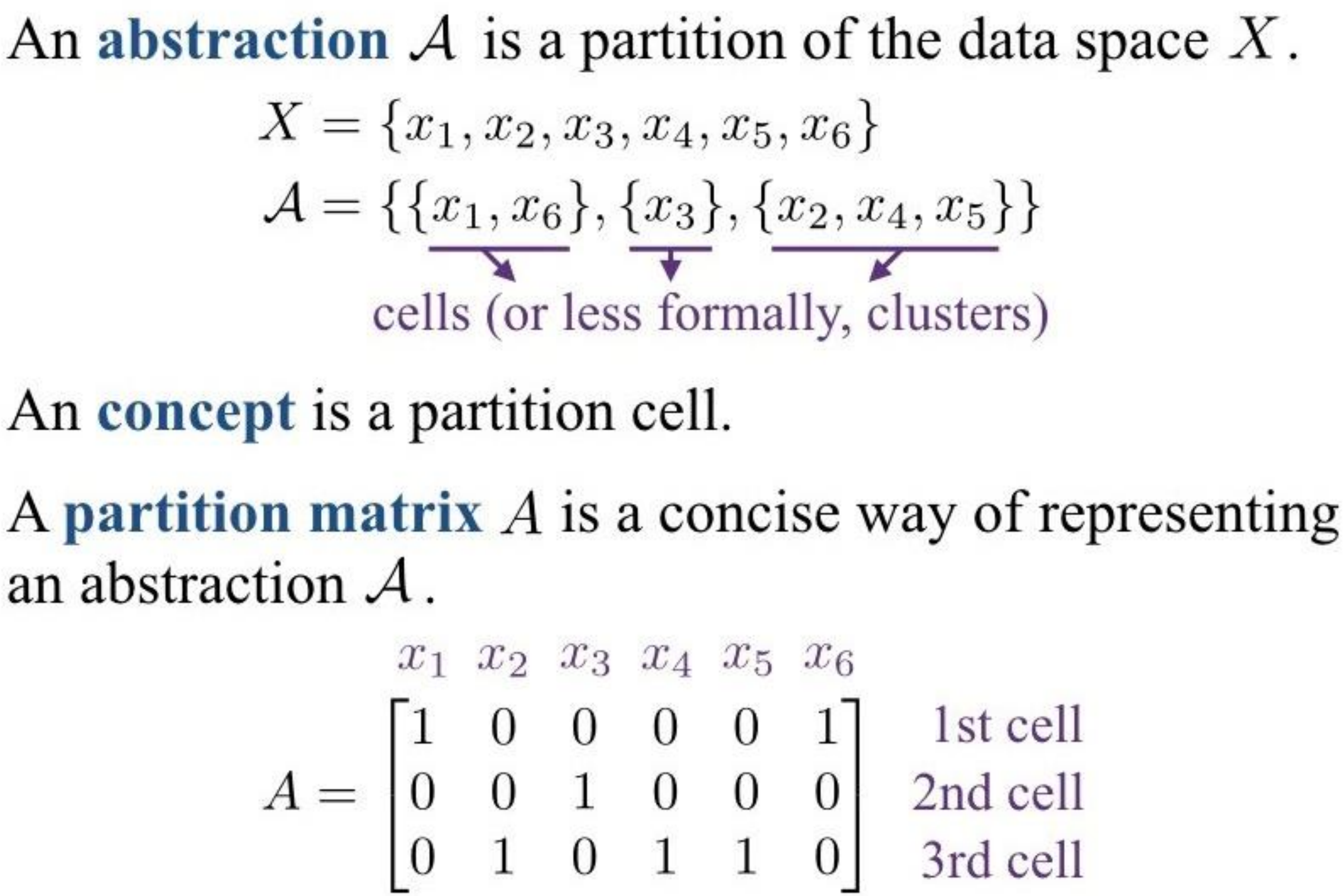
chord: $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \in X$

pitch: $x_i \in \mathbb{Z}$ (C4 → 60)

voice: $i \in \{1, 2, 3, 4\}$
S A T B



Representation: Abstraction



An **abstraction** $\mathcal{A}$ is a partition of the data space X.

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

$$\mathcal{A} = \{\{x_1, x_6\}, \{x_3\}, \{x_2, x_4, x_5\}\}$$

cells (or less formally, clusters)

An **concept** is a partition cell.

A **partition matrix** A is a concise way of representing an abstraction $\mathcal{A}$.

$$A = \begin{array}{c c}
\begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} & \\
\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix} & \begin{array}{l} \text{1st cell} \\ \text{2nd cell} \\ \text{3rd cell} \end{array}
\end{array}$$

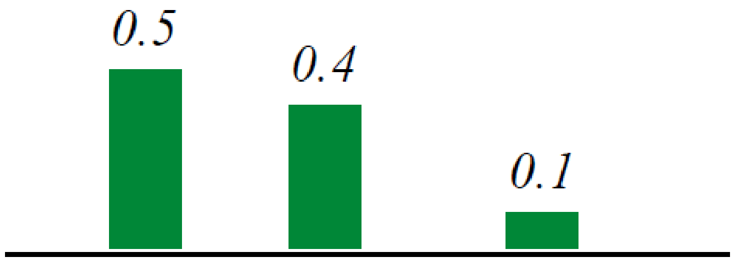
Representation: Probabilistic Rule

A **probabilistic rule** is a pair:

$$(\mathcal{A}, p_{\mathcal{A}})$$

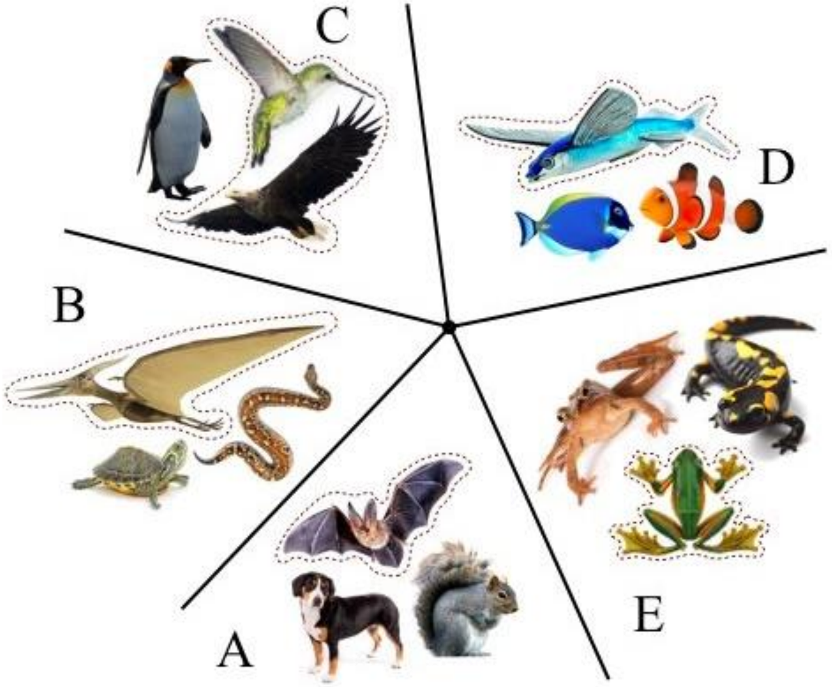
where $\mathcal{A}$ is an abstraction (partition);

$p_{\mathcal{A}}$ is a probability distribution over the abstracted concepts (cells).



$$\mathcal{A} = \{\{x_1, x_6\}, \{x_3\}, \{x_2, x_4, x_5\}\}$$

“Most birds fly; but rare for fish, amphibians, reptiles, mammals.”



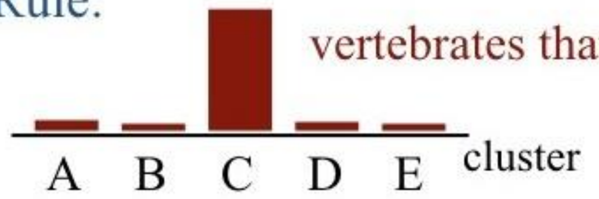
Abstraction (of vertebrates):

Partition vertebrates into five clusters

Concepts:

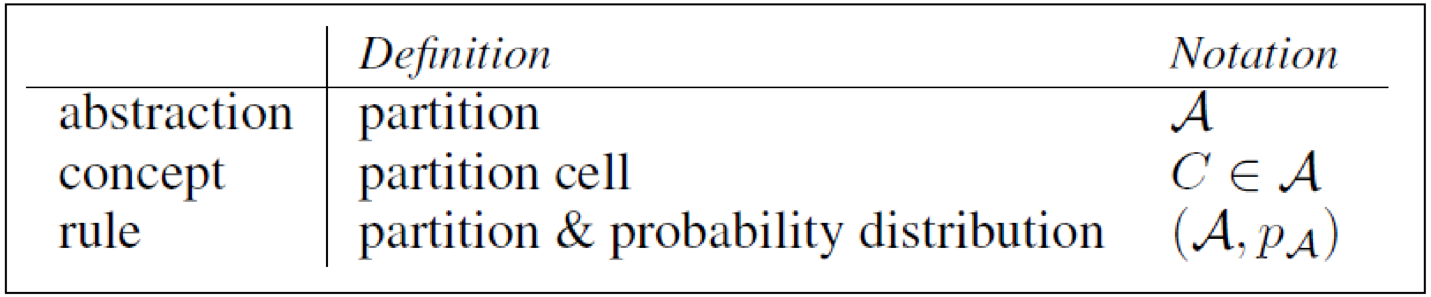
Cluster A: mammals
Cluster B: reptiles
Cluster C: birds
Cluster D: fish
Cluster E: amphibians

Rule:



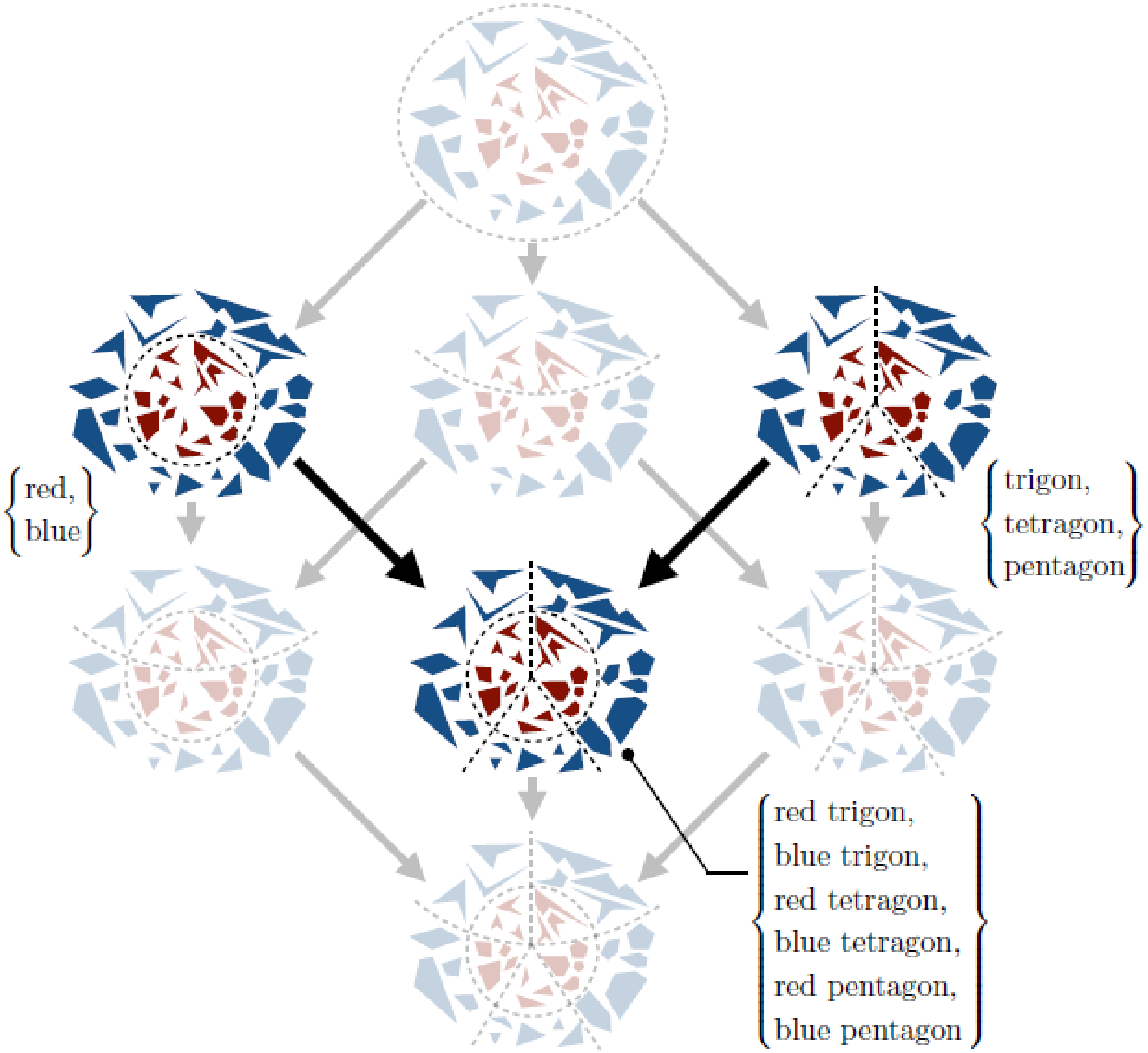
A statistical pattern on abstracted concepts (clusters)

Abstraction as partitioning (clustering) a data space *X*



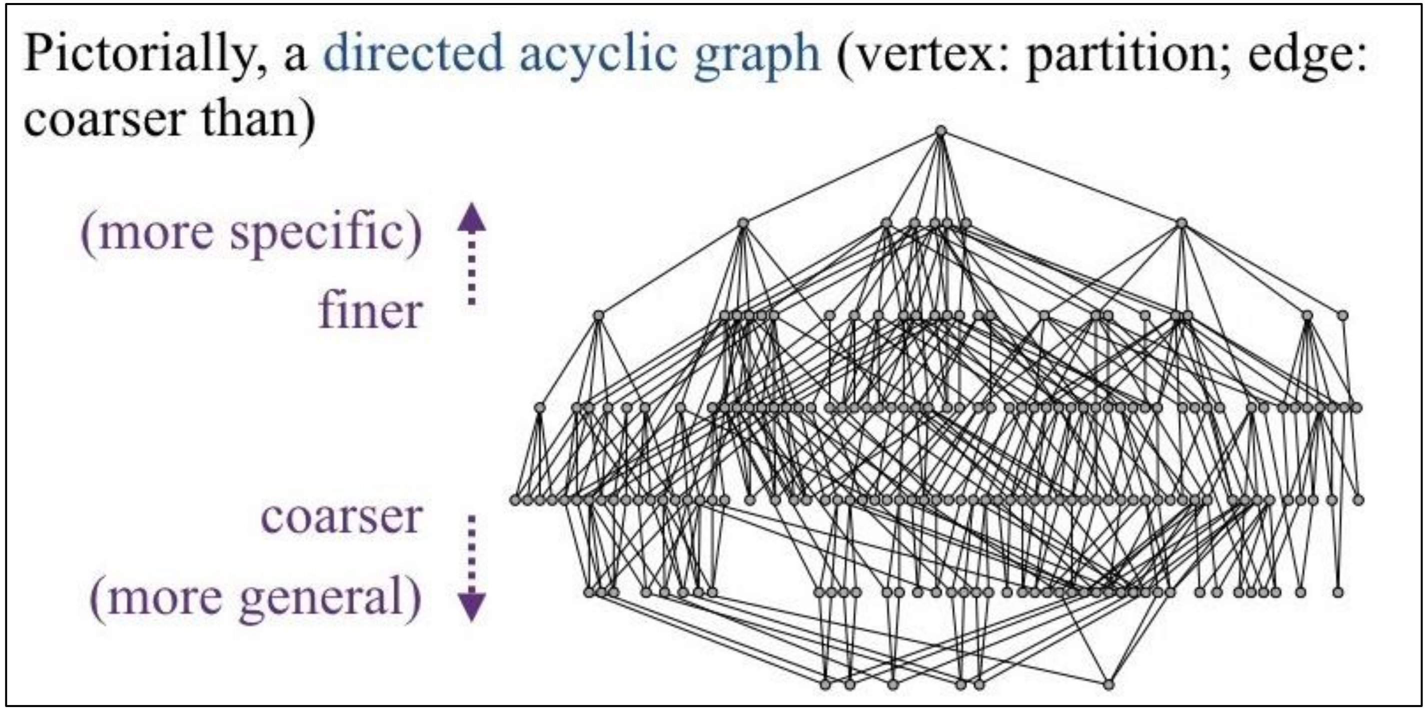
	Definition	*Notation*
abstraction	partition	$\mathcal{A}$
concept	partition cell	$C \in \mathcal{A}$
rule	partition & probability distribution	$(\mathcal{A}, p_{\mathcal{A}})$

- A partition is not an equivalence relation (one is a set, the other is a binary relation), but convey equivalent ideas since they induce each other bijectively
- An equivalence relation explains a partition: elements of a set X are put in the same cell because they are equivalent
- Abstracting the set X involves collapsing equivalent elements in X into a single entity (an equivalence class or partition cell) where collapsing is formalized by taking the quotient


red,
blue
trigon,
tetragon,
pentagon
red trigon,
blue trigon,
red tetragon,
blue tetragon,
red pentagon,
blue pentagon

Abstraction universe as partition lattice

- A set X can have multiple partitions (Bell number $B_{|X|}$)
- Let $\mathfrak{B}_X^*$ denote the family of all partitions of a set X, so $|\mathfrak{B}_X^*| = B_{|X|}$
- Compare partitions of a set by a partial order on $\mathfrak{B}_X^*$
  - Partial order yields a *partition lattice*, a hierarchical representation of a family of partitions



Pictorially, a directed acyclic graph (vertex: partition; edge: coarser than)

Abstraction universe as partition lattice

- Even for a finite set X of relatively small size, the complete abstraction universe $\mathfrak{B}_X^*$ can be quite large and complicated to visualize (Bell number grows very quickly, to say nothing of edges)
- However, not all arbitrary partitions are of interest

What part of $\mathfrak{B}_X^*$ should we focus on?

Abstraction universe as partition lattice

- Even for a finite set X of relatively small size, the complete abstraction universe $\mathfrak{B}_X^*$ can be quite large and complicated to visualize (Bell number grows very quickly, to say nothing of edges)
- However, not all arbitrary partitions are of interest

What part of $\mathfrak{B}_X^*$ should we focus on?

- Feature-induced abstractions
  - Consider a pool of feature functions Φ, spanned by a finite set of basis features that are individually "simple" (e.g. basic arithmetic operators like sort and mod) and easy for people to interpret
  - Key idea is to break a rich pool of domain-specific features into a set of domain-agnostic basis features as building blocks
- Symmetry-induced abstractions

Symmetry-induced abstraction

- Consider the symmetric group $(S_X, \circ)$ defined over a set X, whose group elements are all the bijections from X to X and whose group operation is (function) composition
- A bijection from X to X is also called a *transformation* of X, so the symmetric group S_X comprises all transformations of X, and is also called the transformation group of X, denoted $\mathrm{F}(X)$

- Given a set X and a subgroup $H \leq \mathrm{F}(X)$, we define an H-action on X by $h \cdot x = h(x)$ for any $h \in H, x \in X$ and the orbit of $x \in X$ under H as the set $Hx = \{h(x) | h \in H\}$
- Each orbit is an equivalence class, so the quotient $X/H = X/\sim$ is a partition of X
- We say this abstraction respects H-symmetry or H-invariance

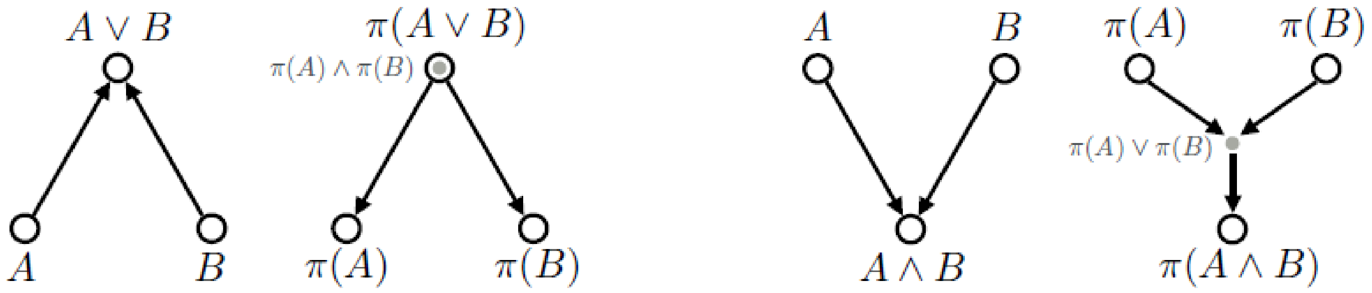
$$\boxed{\text{a subgroup of } \mathrm{F}(X) \xrightarrow{\textit{group action}} \text{orbits} \xrightarrow{\textit{equiv. rel.}} \text{a partition} \xrightarrow{\textit{is}} \text{an abstraction of } X}$$

Duality: From subgroup lattice to abstraction (semi)universe

Definition The *abstraction generating function* is the mapping $\pi\colon \mathcal{H}^*_{\mathrm{F}(X)} \to \mathfrak{B}^*_X$, where $\mathcal{H}^*_{\mathrm{F}(X)}$ is the collection of all subgroups of $\mathrm{F}(X)$, $\mathfrak{B}^*_X$ is the family of all partitions of X, and for any $H \in \mathcal{H}^*_{\mathrm{F}(X)}$, $\pi(H) = X/H$.

Theorem (Duality) Let $\left(\mathcal{H}^*_{\mathrm{F}(X)}, \leq\right)$ be the subgroup lattice for $\mathrm{F}(X)$ and π the abstraction generating function. Then $\left(\pi\left(\mathcal{H}^*_{\mathrm{F}(X)}\right), \preccurlyeq\right)$ is an abstraction meet-semiuniverse for X. That is:

1. partial-order reversal: if $A \leq B$, then $\pi(A) \succcurlyeq \pi(B)$
2. strong duality: $\pi(A \vee B) = \pi(A) \wedge \pi(B)$
3. weak duality: $\pi(A \wedge B) \succcurlyeq \pi(A) \vee \pi(B)$

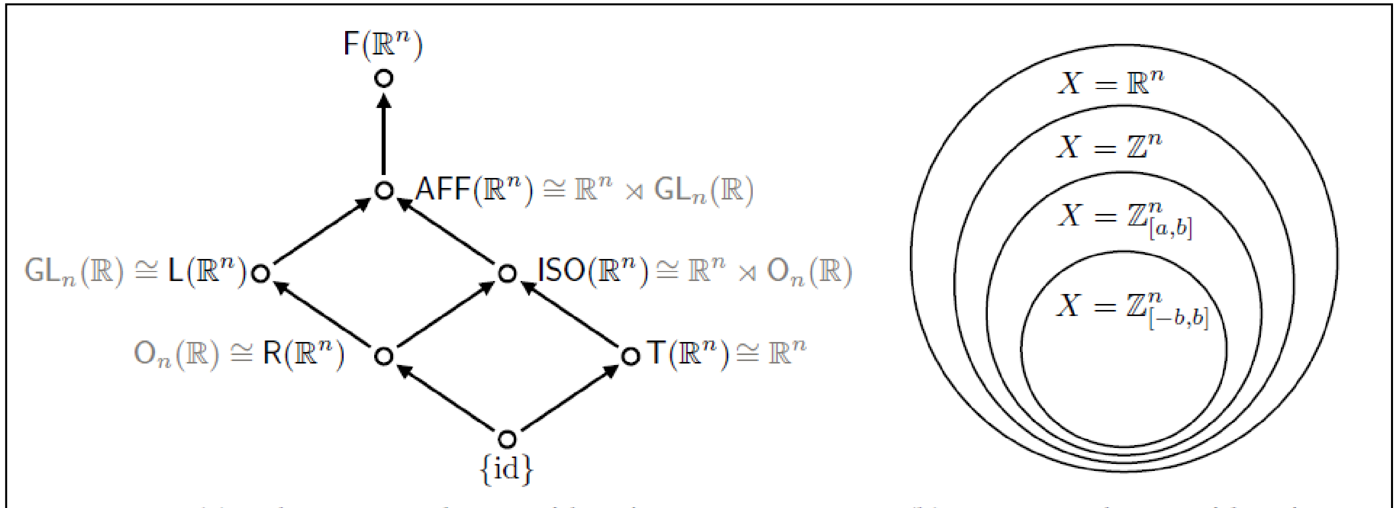


(a) From join to meet.

(b) From meet to join.

Duality: From subgroup lattice to abstraction (semi)universe

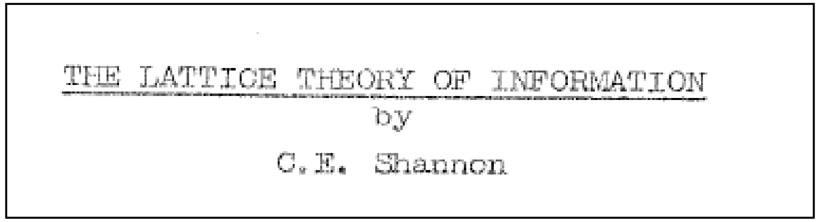
- If one has already computed abstractions $\pi(A)$ and $\pi(B)$, then instead of computing $\pi(A \vee B)$ from $A \vee B$, one can compute the meet $\pi(A) \wedge \pi(B)$, which is generally computationally less expensive than computing $A \vee B$ and identifying all orbits in $\pi(A \vee B)$
- The computer algebra system GAP provides efficient algorithmic methods to construct the subgroup lattice for a given group, and even maintains data libraries for special groups and their subgroup lattices



$\mathsf{F}(\mathbb{R}^n)$
$\mathsf{AFF}(\mathbb{R}^n) \cong \mathbb{R}^n \rtimes \mathsf{GL}_n(\mathbb{R})$
$\mathsf{GL}_n(\mathbb{R}) \cong \mathsf{L}(\mathbb{R}^n)$
$\mathsf{ISO}(\mathbb{R}^n) \cong \mathbb{R}^n \rtimes \mathsf{O}_n(\mathbb{R})$
$\mathsf{O}_n(\mathbb{R}) \cong \mathsf{R}(\mathbb{R}^n)$
$\mathsf{T}(\mathbb{R}^n) \cong \mathbb{R}^n$
$\{\mathrm{id}\}$
$X = \mathbb{R}^n$
$X = \mathbb{Z}^n$
$X = \mathbb{Z}^n_{[a,b]}$
$X = \mathbb{Z}^n_{[-b,b]}$

(a) Subgroups under consideration. (b) Spaces under consideration.

[H. Yu, I. Mineyev, and L. R. Varshney, “Orbit Computation for Atomically Generated Subgroups of Isometries of Z^n,” *SIAM Journal on Applied Algebra and Geometry*, vol. 5, no. 3, pp. 479–505, Sept. 2021.]



THE LATTICE THEORY OF INFORMATION

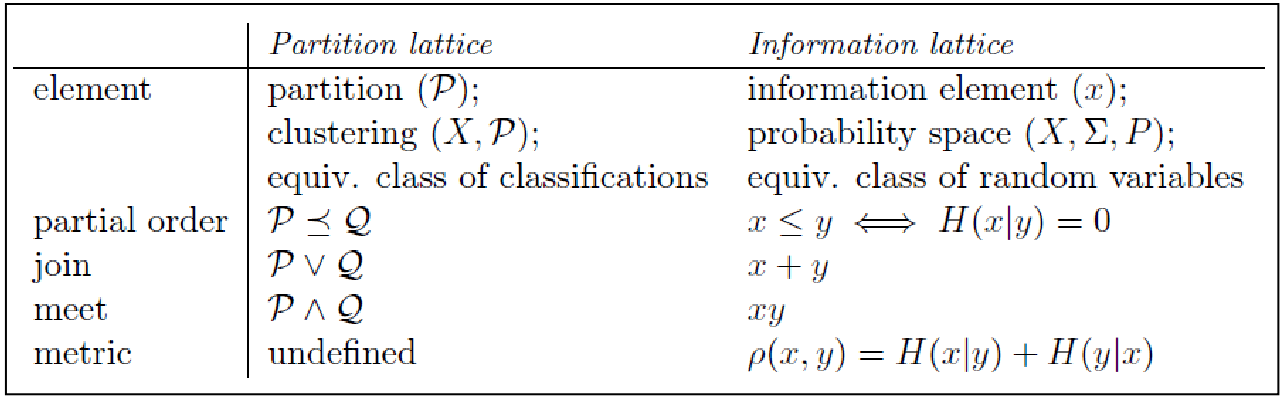
by

C. E. Shannon

- An *information element* is an equivalence class of random variables w.r.t. inducing the same σ-algebra

- An *information lattice* is a lattice of information elements, where partial order defined by $x \le y \Leftrightarrow H(x|y) = 0$ where H is the Shannon entropy. The join of two information elements the *total information*; the meet of two information elements is the *common information*
- Our abstraction-generation framework generalizes Shannon's information lattice, without needing to introduce information-theoretic functionals like entropy
- More importantly gives generating chain to bring learning into picture

Separation of clustering from statistics: partition lattice can be thought as an information lattice without probability measure



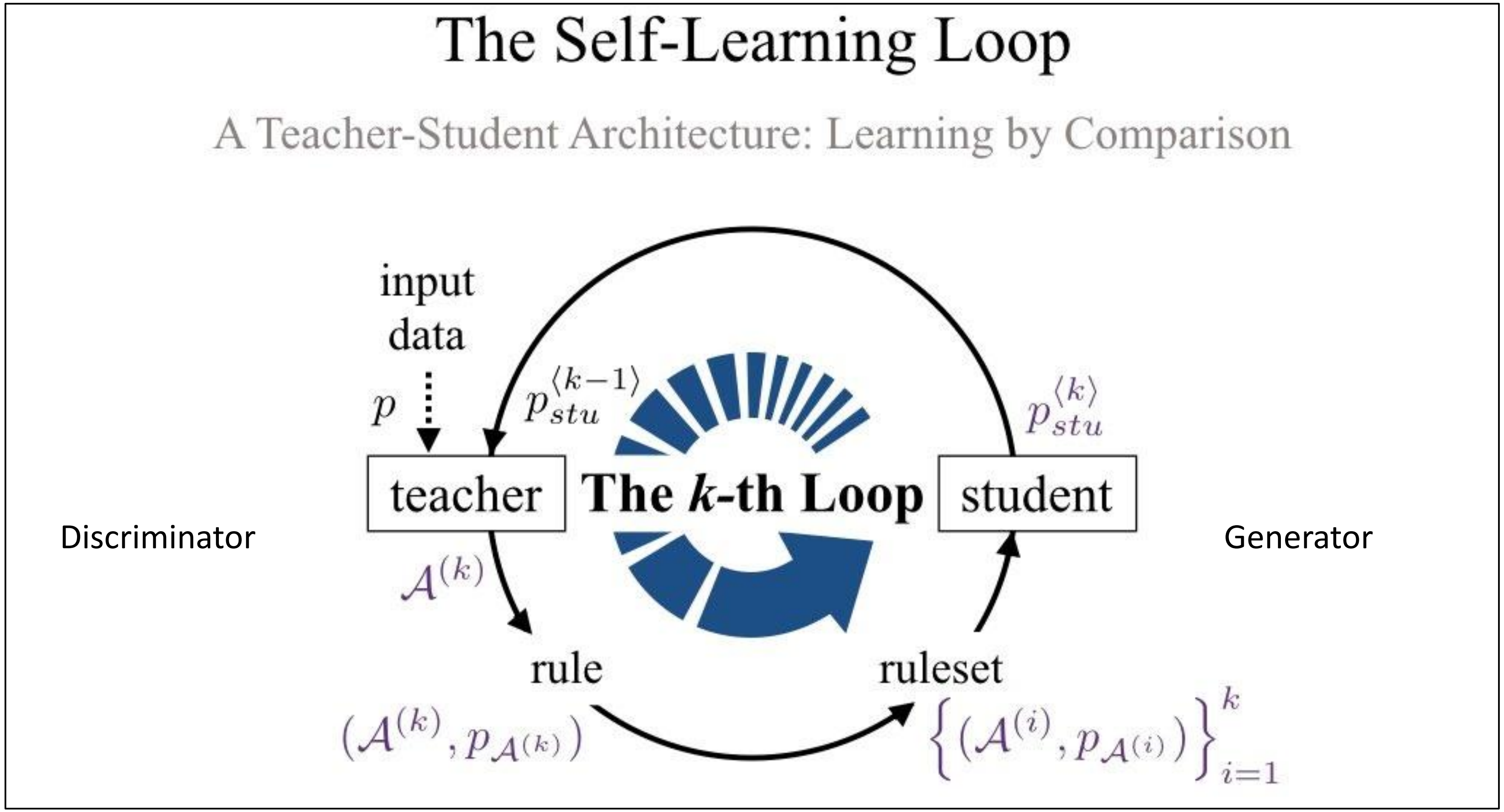
	Partition lattice	*Information lattice*

element	partition ($\mathcal{P}$);
	clustering ($X, \mathcal{P}$);
	equiv. class of classifications
		information element (x);
		probability space (X, Σ, P);
		equiv. class of random variables

partial order	$\mathcal{P} \preceq \mathcal{Q}$	$x \le y \iff H(x\mid y) = 0$
join	$\mathcal{P} \vee \mathcal{Q}$	$x + y$
meet	$\mathcal{P} \wedge \mathcal{Q}$	xy
metric	undefined	$\rho(x, y) = H(x\mid y) + H(y\mid x)$

Information-theory inspired algorithm for rule learning

Learning is achieved by statistical inference on a partition lattice


The Self-Learning Loop
A Teacher-Student Architecture: Learning by Comparison
input data
p
$p_{stu}^{\langle k-1 \rangle}$
$p_{stu}^{\langle k \rangle}$
teacher
The k-th Loop
student
Discriminator
Generator
$\mathcal{A}^{(k)}$
rule
ruleset
$(\mathcal{A}^{(k)}, p_{\mathcal{A}^{(k)}})$
$\left\{(\mathcal{A}^{(i)}, p_{\mathcal{A}^{(i)}})\right\}_{i=1}^{k}$

[H. Yu and L. R. Varshney, “Towards Deep Interpretability (MUS-ROVER II): Learning Hierarchical Representations of Tonal Music,” in *Proc. 5th International Conference on Learning Representations (ICLR)*, April 2017.]

Information-theory inspired algorithm for rule learning

Learning is achieved by statistical inference on a partition lattice

MUS-ROVER's self-learning loop:

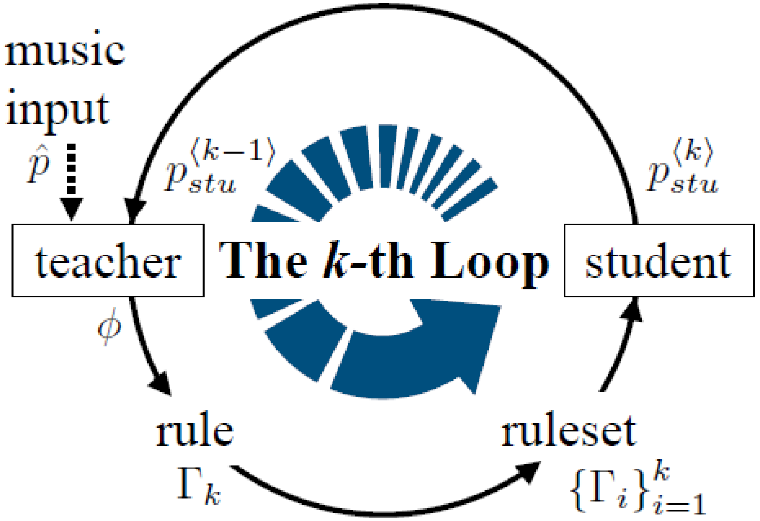
The iterative cooperation between a discriminator (teacher) and a generator (student).

The teacher solves:

$$\begin{array}{ll} \text{maximize} & D\left(p_{\phi,stu}^{\langle k-1\rangle} \,||\, \hat{p}_\phi\right) \\ \text{subject to} & \phi \in \Phi \backslash \Phi^{\langle k-1\rangle} \end{array}$$

(discrete optimization)

max. Bayesian surprise



The student solves:

$$\begin{array}{ll} \text{maximize} & S_q\left(p_{stu}^{\langle k\rangle}\right) \\ \text{subject to} & p_{stu}^{\langle k\rangle} \in \Gamma_1 \\ & \cdots \\ & p_{stu}^{\langle k\rangle} \in \Gamma_k \end{array}$$

(linear least-squares)

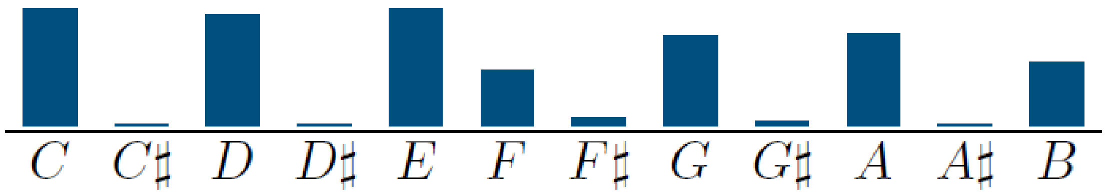
max. creativity

Simple human-interpretable rules

Compositional Rule Examples:

feature → ϕ : pitch class in the soprano voice

feature distribution → p_ϕ :



This rule can be interpreted or translated to:

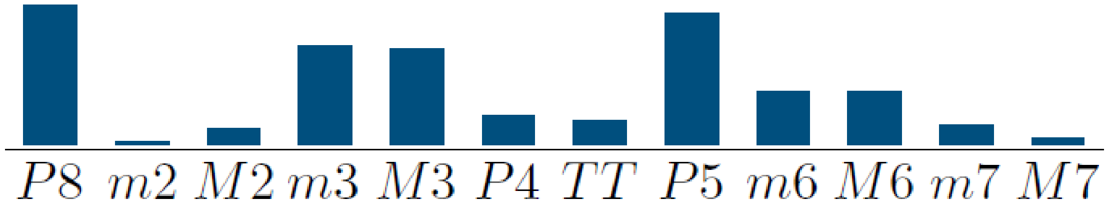
“The soprano voice is built on a diatonic scale.”

Hierarchical concept learning

Compositional Rule Examples:

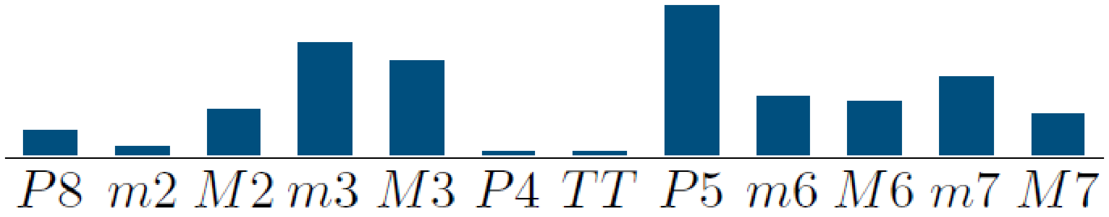
ϕ : interval class between soprano and bass

p_ϕ



1-gram

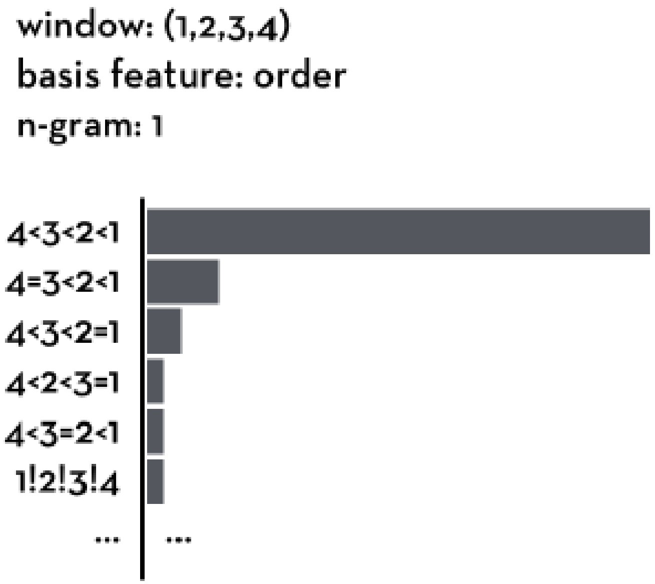
"Individual perfect octaves (P8s) are favored as most consonant."

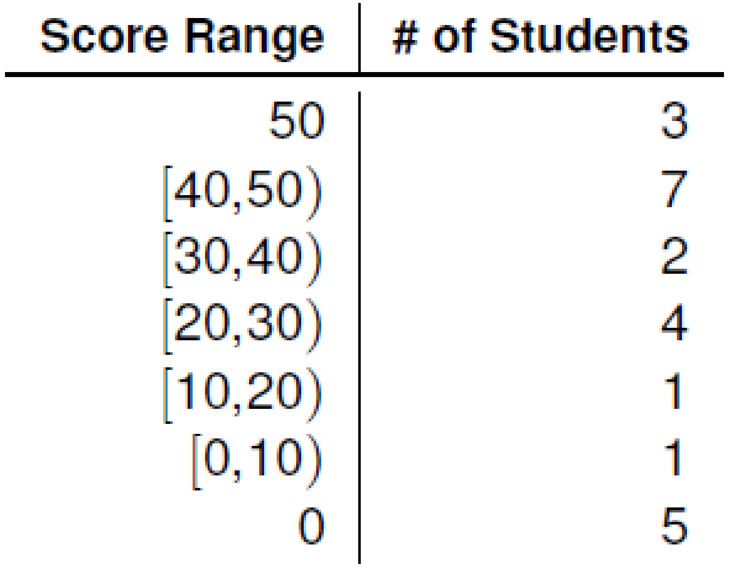


2-gram conditioned on P8

"Parallel perfect octaves (P8s) are uncommon."

This form of compositional rules are in fact human-interpretable



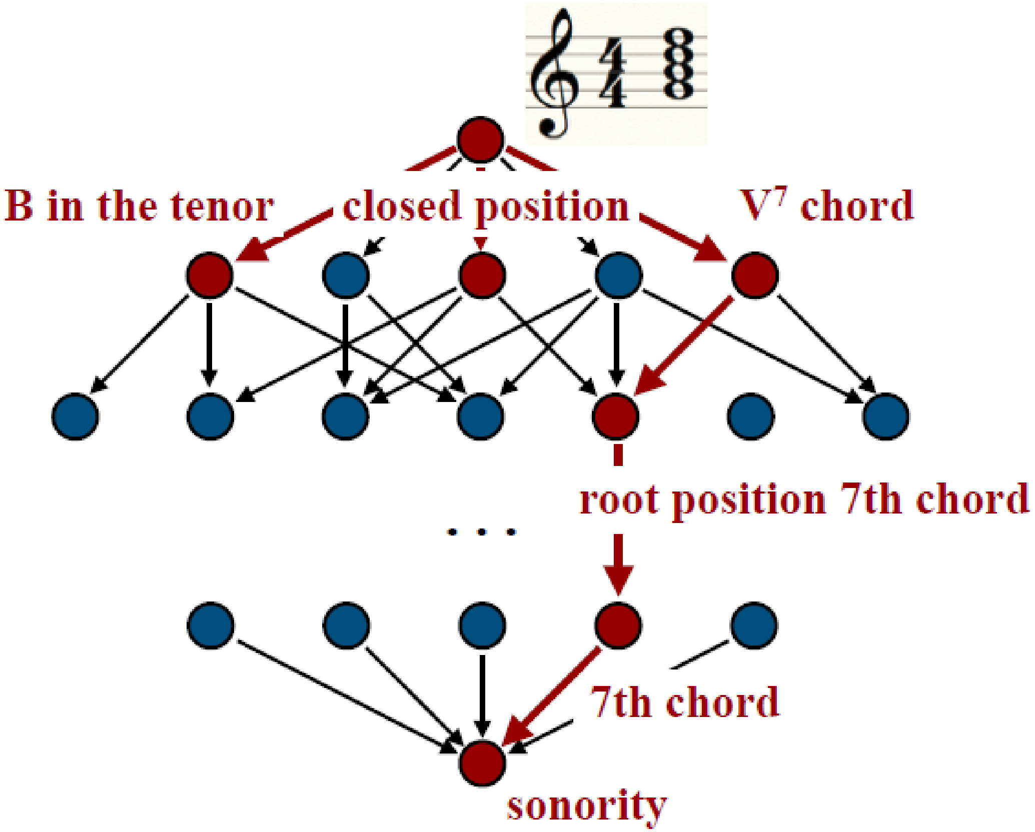


Score Range	# of Students
50	3
[40,50)	7
[30,40)	2
[20,30)	4
[10,20)	1
[0,10)	1
0	5

Table 1: Students' final scores.

[H. Yu, H. Taube, J. A. Evans, and L. R. Varshney, "Human Evaluation of Interpretability: The Case of AI-Generated Music Knowledge," in *ACM CHI 2020 Workshop on Artificial Intelligence for HCI: A Modern Approach*, April 2020.]

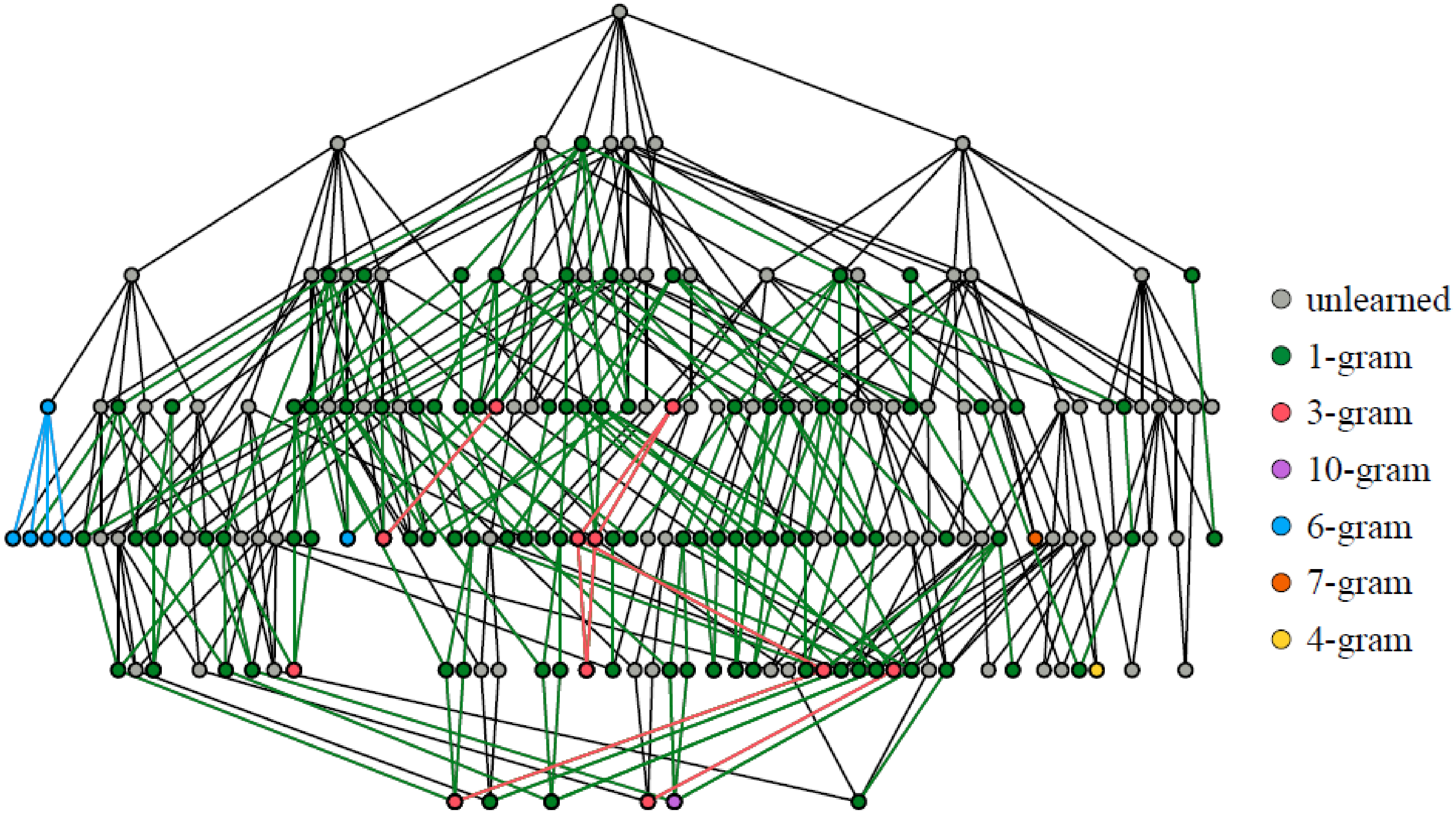
Hierarchy of music theory concepts



raw representation

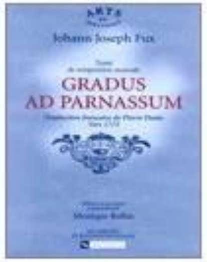
higher (deeper)-level abstractions

Compositional rules are extracted not simply as a linear list, but as hierarchical families and sub-families.



Visualization of Bach's music mind for writing chorales. The underlying directed acyclic graph signifies an upside-down information lattice.

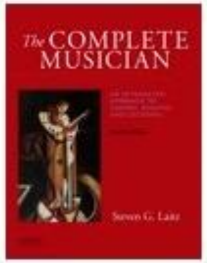
ILL recovers much known music theory



Johann Joseph Fux

GRADUS AD PARNASSUM

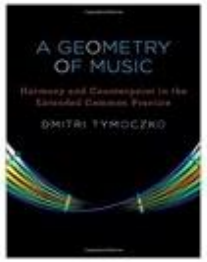
- voice leading
- counter point



The COMPLETE MUSICIAN

Steven G. Laitz

- scale, consonance & dissonance
- voice spacing, crossing, overlap
- chord quality, inversion, progression



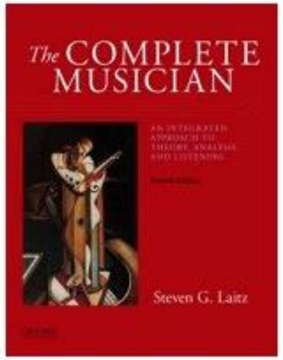
A GEOMETRY OF MUSIC

Harmony and Counterpoint in the Extended Common Practice

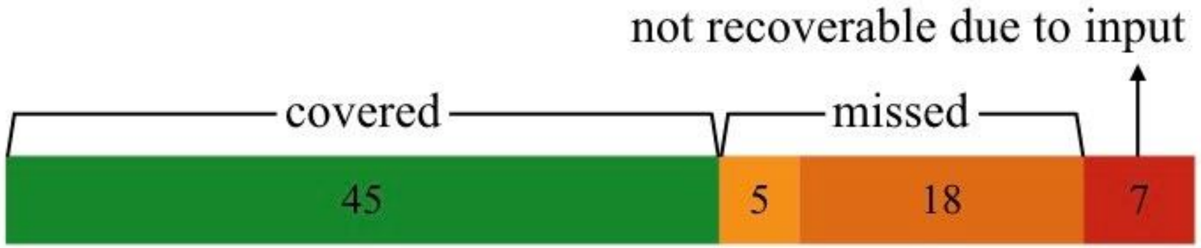
DMITRI TYMOCZKO

- music transformations: OPTIC

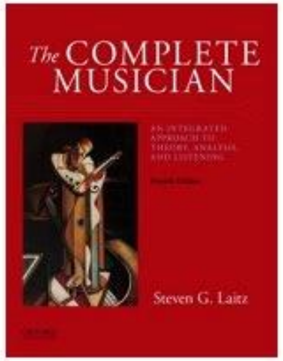
ILL recovers much known music theory



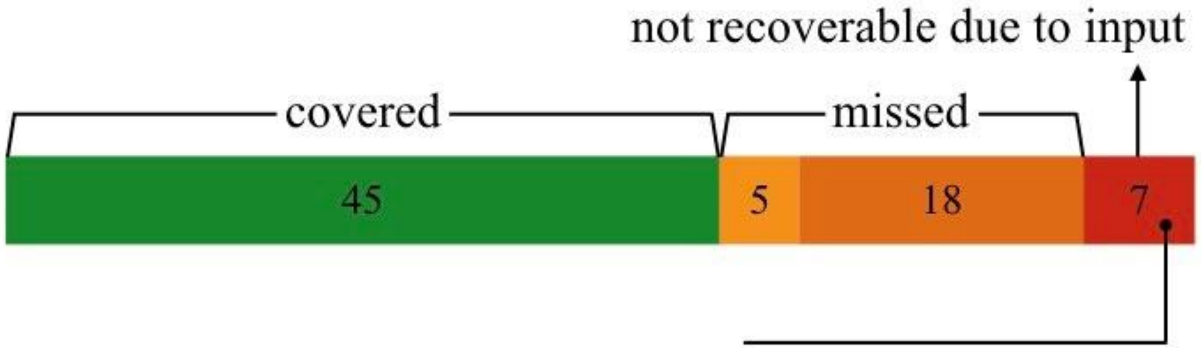
MUS 101, 102, 201 (75 topics in total):



ILL recovers much known music theory



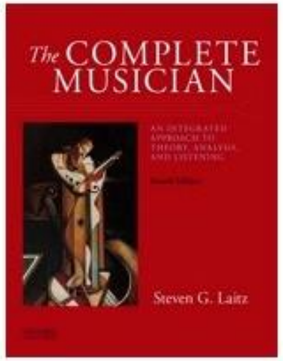
MUS 101, 102, 201 (75 topics in total):



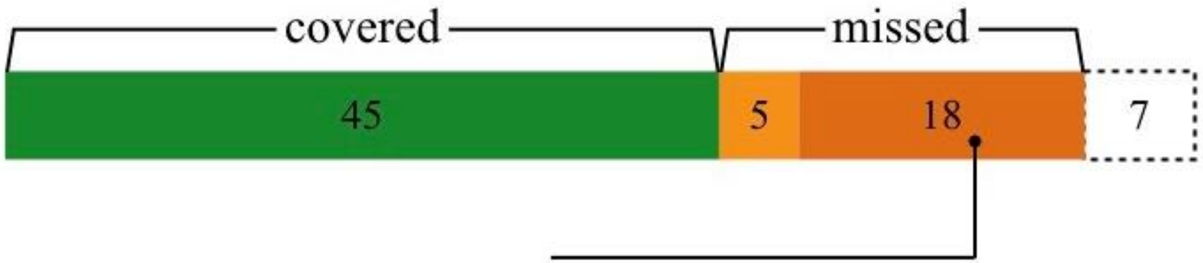
requires info other than MIDI pitches and durations:

- music accents: requires beats, dynamics, etc.
- enharmonic re-spellings: German 6th, fully dim, etc.

ILL recovers much known music theory



MUS 101, 102, 201 (68 recoverable topics in total):



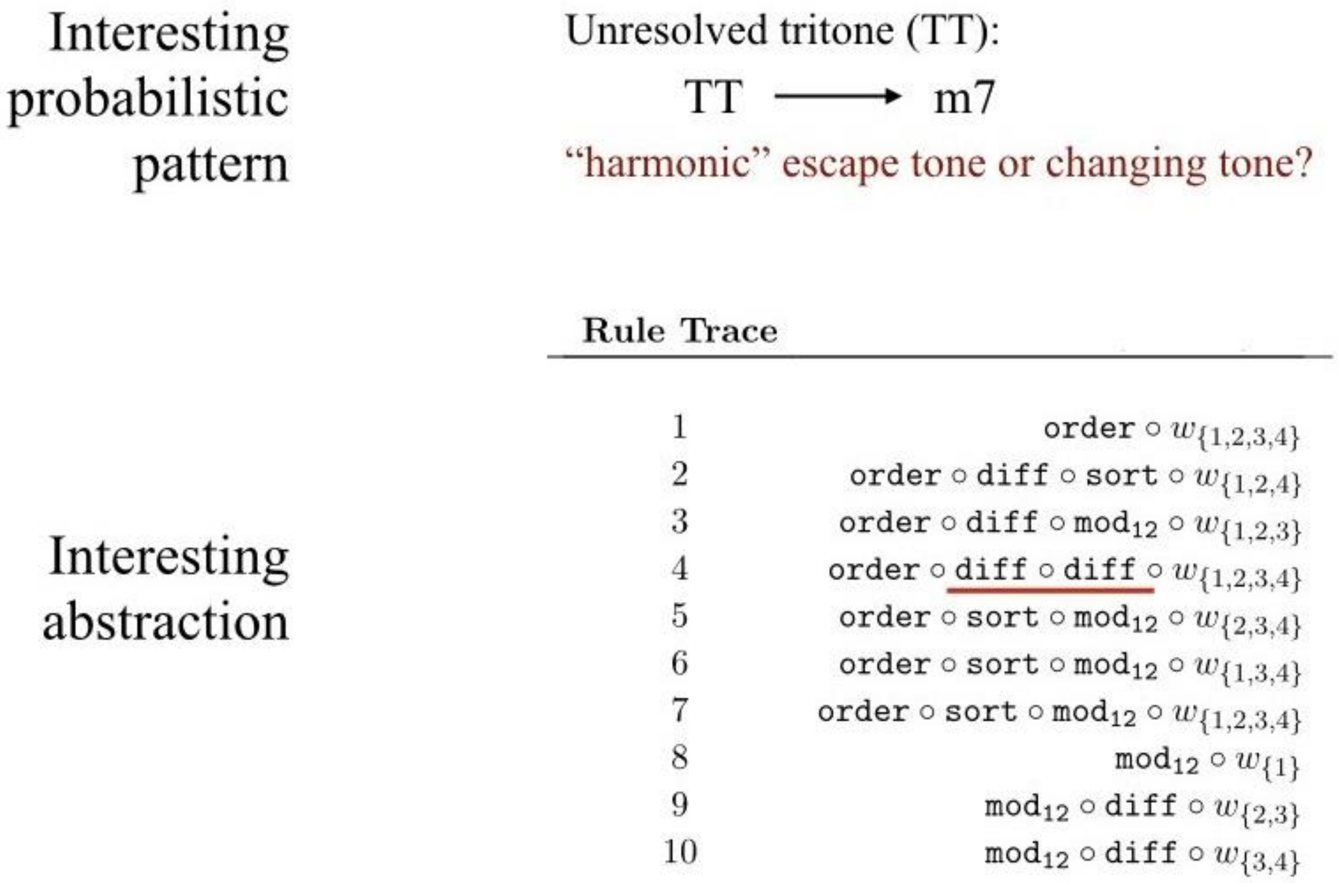
captured but not explicitly presented:

- phrase models, EPMs, sentence structure, etc.
- music forms: binary, ternary, rondo, sonata, etc.

Suggests an extension of the n-gram models to temporal abstractions:

transitions of abstractions → *abstractions of transitions*

ILL discovers new music theory



Interesting probabilistic pattern

Unresolved tritone (TT):

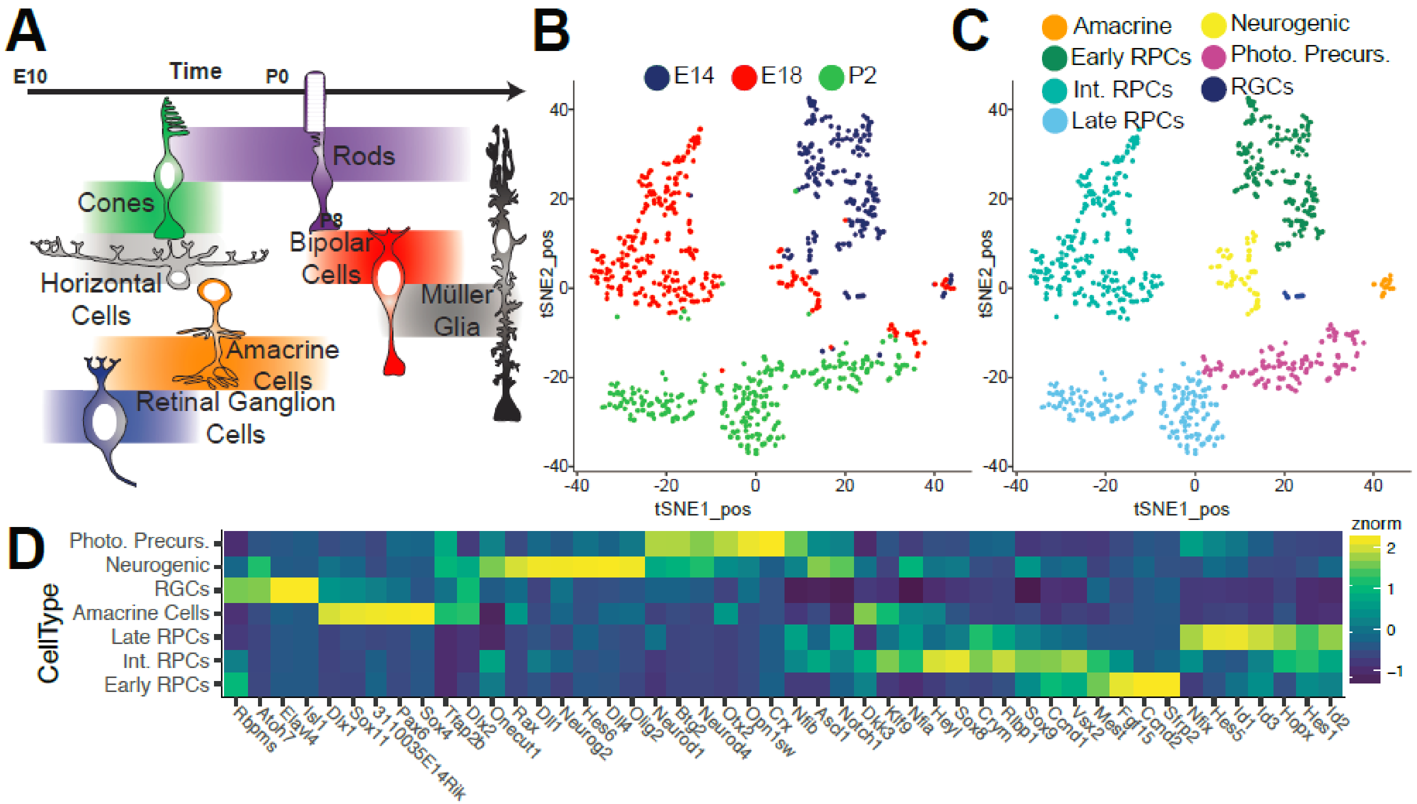
$$\text{TT} \longrightarrow \text{m7}$$

“harmonic” escape tone or changing tone?

Interesting abstraction

Rule	Trace
1	$\texttt{order} \circ w_{\{1,2,3,4\}}$
2	$\texttt{order} \circ \texttt{diff} \circ \texttt{sort} \circ w_{\{1,2,4\}}$
3	$\texttt{order} \circ \texttt{diff} \circ \texttt{mod}_{12} \circ w_{\{1,2,3\}}$
4	$\texttt{order} \circ \underline{\texttt{diff} \circ \texttt{diff}} \circ w_{\{1,2,3,4\}}$
5	$\texttt{order} \circ \texttt{sort} \circ \texttt{mod}_{12} \circ w_{\{2,3,4\}}$
6	$\texttt{order} \circ \texttt{sort} \circ \texttt{mod}_{12} \circ w_{\{1,3,4\}}$
7	$\texttt{order} \circ \texttt{sort} \circ \texttt{mod}_{12} \circ w_{\{1,2,3,4\}}$
8	$\texttt{mod}_{12} \circ w_{\{1\}}$
9	$\texttt{mod}_{12} \circ \texttt{diff} \circ w_{\{2,3\}}$
10	$\texttt{mod}_{12} \circ \texttt{diff} \circ w_{\{3,4\}}$

Learning laws of neurogenesis



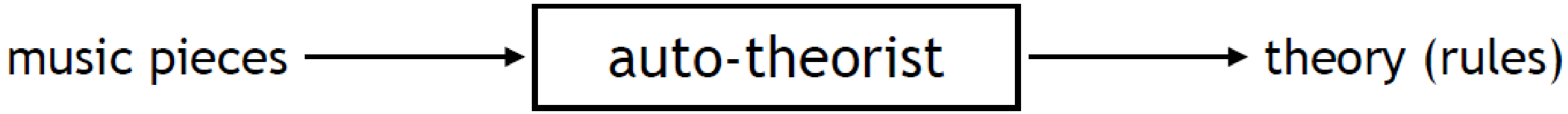
[B. Clark, et al., "Single-Cell RNA-Seq Analysis of Retinal Development Identifies NFI Factors as Regulating Mitotic Exit and Late-Born Cell Specification," *Neuron*, June 2019.]

Single-cell RNA sequence data analysis for understanding the rules that govern pattern formation in neurodevelopment

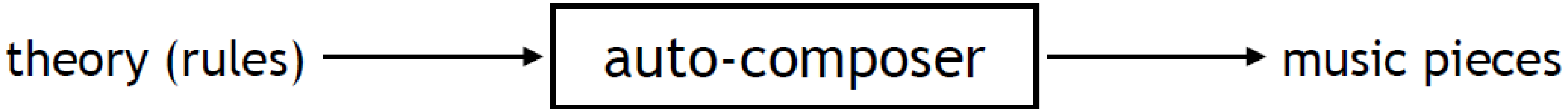
[H. Yu, L. R. Varshney, and G. Stein-O'Brien, "Towards Learning Human-Interpretable Laws of Neurogenesis from Single-Cell RNA-Seq Data via Information Lattices," at *Learning Meaningful Representations of Life Workshop at NeurIPS*, Dec. 2019.]

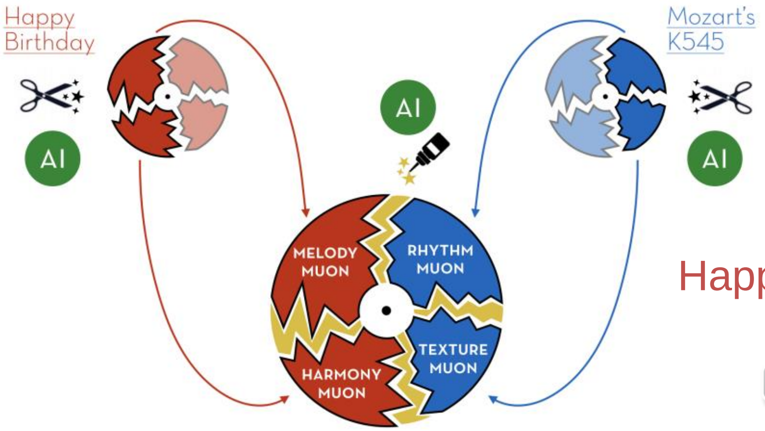
From automatic knowledge discovery to co-creativity

A way to learn the principles of quality (laws of music theory)

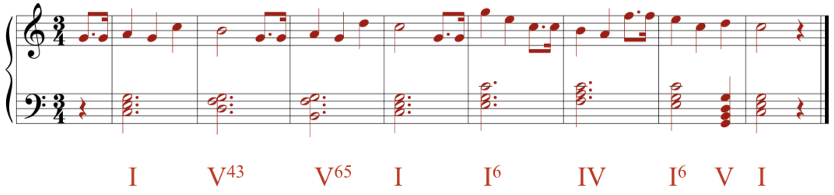


Computational creativity algorithms for music composition

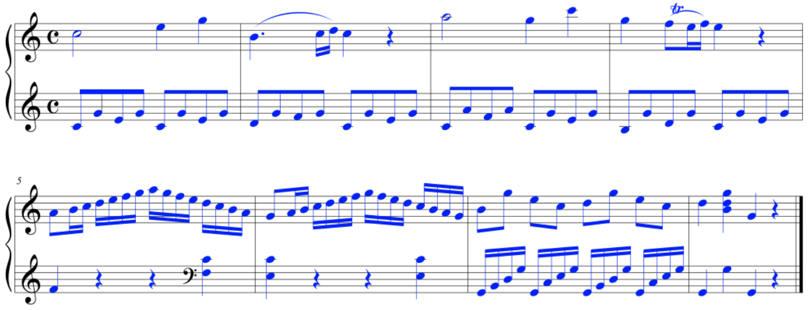




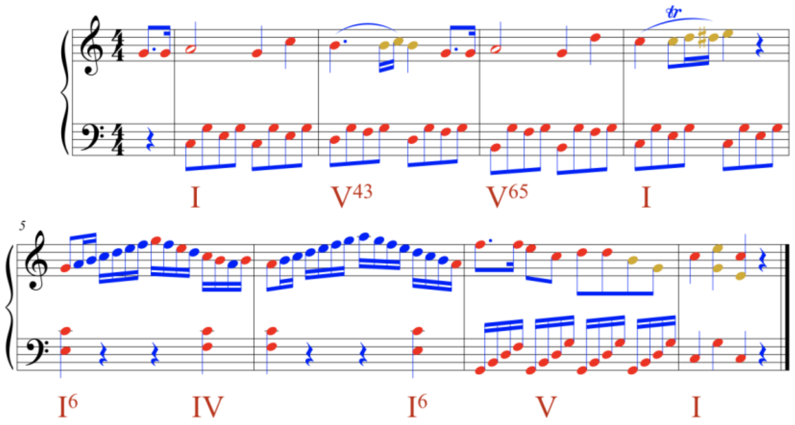
Happy Birthday:

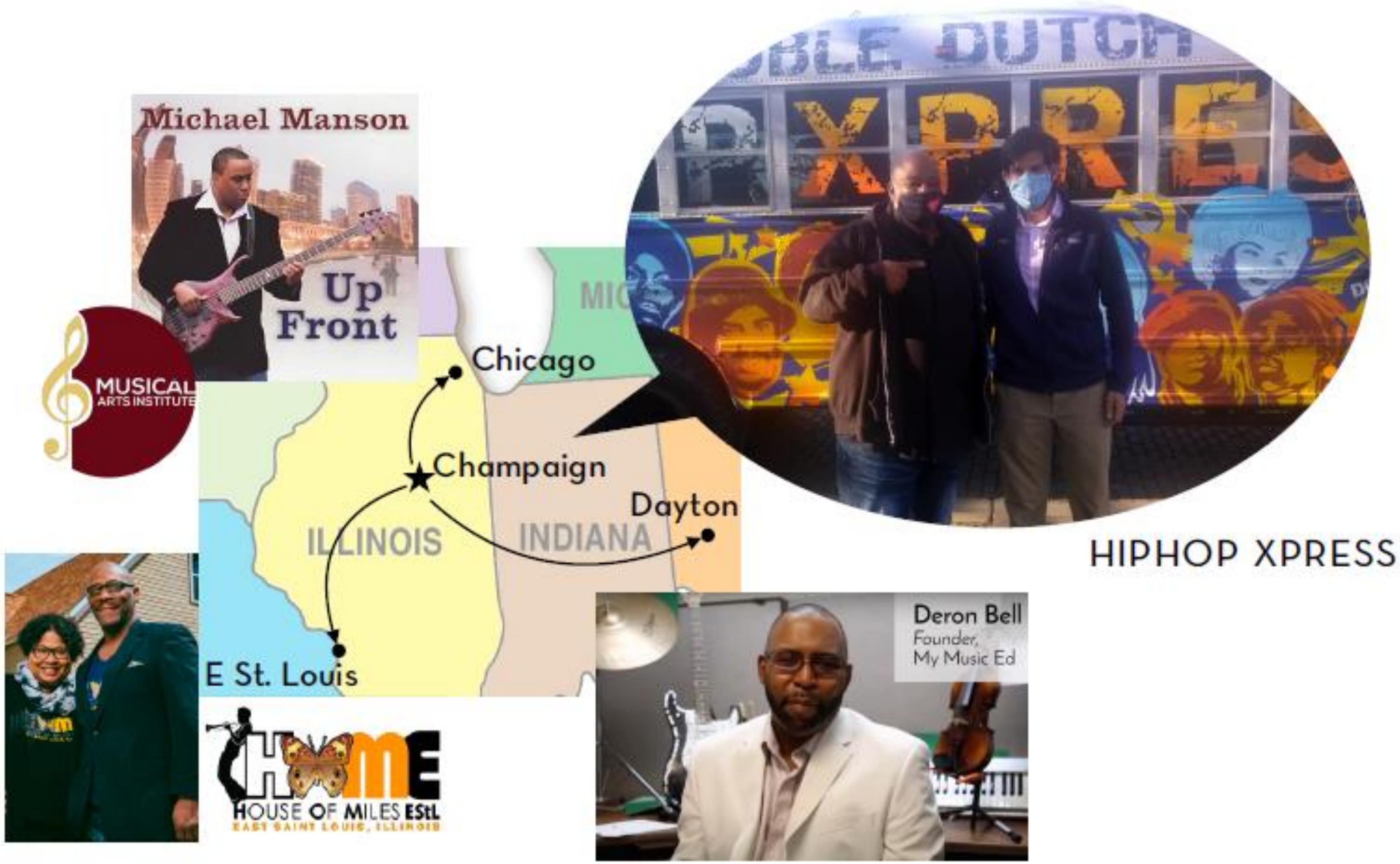


Mozart’s K545:



Music Mosaic:




Michael Manson
Up Front
MUSICAL ARTS INSTITUTE
Chicago
Champaign
Dayton
ILLINOIS
INDIANA
MIC
E St. Louis
HOME
HOUSE OF MILES EStL
EAST SAINT LOUIS, ILLINOIS
DOUBLE DUTCH XPRESS
HIPHOP XPRESS
Deron Bell
Founder,
My Music Ed

Technology

[e.g. Kocree, Inc.]

Policy

[e.g. White House]

Lav R. Varshney

Kocree, Inc.

University of Illinois Urbana-Champaign

lvarshney@kocree.net
varshney@illinois.edu