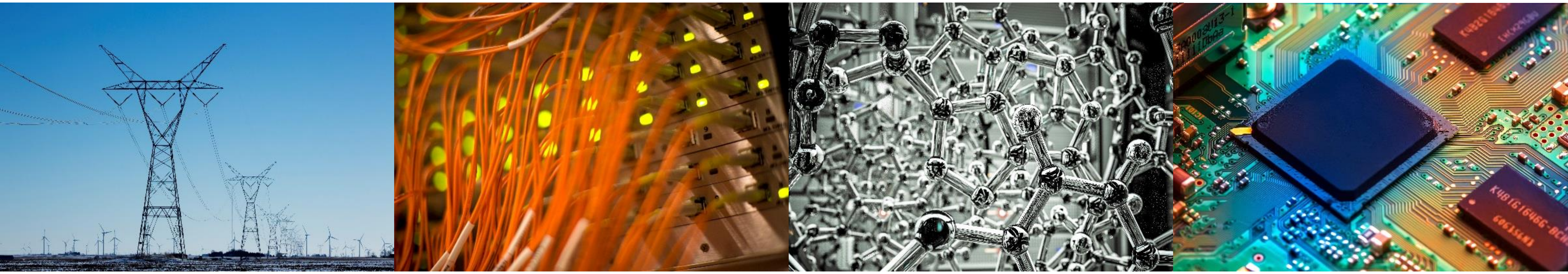
Human-Interpretable Concept Learning via Information Lattices

Lav R. Varshney
University of Illinois at Urbana-Champaign

28 February 2019



AFRL

CHAN
ZUCKERBERG
INITIATIVE

ILLINOIS
Electrical & Computer Engineering
COLLEGE OF ENGINEERING

C³SR
center for
cognitive computing
systems research

Haizi Yu

H. Yu, I. Mineyev, and L. R. Varshney, "A Group-Theoretic Approach to Abstraction: Hierarchical, Interpretable, and Task-Free Clustering," arXiv:1807.11167 [cs.LG].

H. Yu, T. Li, and L. R. Varshney, "Probabilistic Rule Realization and Selection," in *NeurIPS 2017*.

H. Yu and L. R. Varshney, "Towards Deep Interpretability (MUS-ROVER II): Learning Hierarchical Representations of Tonal Music," in *ICLR 2017*.

H. Yu, L. R. Varshney, G. E. Garnett, and R. Kumar, "Learning Interpretable Musical Compositional Rules and Traces," in *ICML WHI 2016*.

Smart sensors: modular, autonomous, reagentless, inline, minimal calibration, low-cost

- On-device intelligence
- Limited training data
- Human interpretability

ENSARAS
Intelligence for Wastewater

The National Academies of SCIENCES ENGINEERING MEDICINE

[N. Kshetry and L. R. Varshney, “Optimal Wastewater Management Using Noisy Sensor Fusion,” presented at *Fifth Arab-American Frontiers of Science, Engineering, and Medicine Symposium*, Rabat, Morocco, Nov. 2017]

“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.]

Five meshing gears are arranged in a horizontal line much like a row of quarters on a table. If you turn the gear on the furthest left clockwise, what will the gear on the furthest right do?

Five meshing gears are arranged in a horizontal line much like a row of quarters on a table. If you turn the gear on the furthest left clockwise, what will the gear on the furthest right do?

- In process of solving many gear problems, collaborative groups often discovered underlying rule: if gears add up to odd number, first and last gear will turn in the same direction (abstract parity rule)
- Once pair discovered rule, stopped motioning with their hands and were able to solve a problem with 131 gears in about 1/10 the time
- Because pairs had to communicate to solve the problems, they developed collaborative representations that neither would have alone, and those representations were more abstract to accommodate the two perspectives that each started out with

[Daniel L. Schwartz, “The Emergence of Abstract Representations in Dyad Problem Solving,” *Journal of the Learning Sciences*, vol. 4, no. 3,, pp. 321-354, 1995.]

Five meshing gears are arranged in a horizontal line much like a row of quarters on a table. If you turn the gear on the furthest left clockwise, what will the gear on the furthest right do?

- Dyads learned a kind of general knowledge useful in later tasks
- Learned rule was very intuitive and human-interpretable
- Rule was based on mod-2 induced symmetry (group-theoretic invariance)
- Rule learning process proceeded through an interaction between two agents

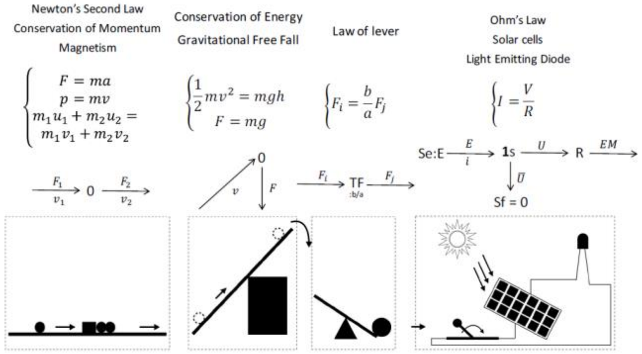
Dimensions of interpretability [Selbst and Barocas, 2018]

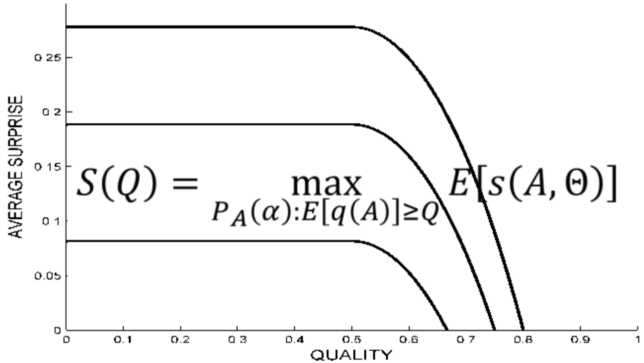
- What sets machine learning models apart from other decision-making mechanisms are their *inscrutability* and *nonintuitiveness*
  - Inscrutability suggests that models available for direct inspection may defy understanding,
  - Nonintuitiveness suggests that even where models are understandable, they 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

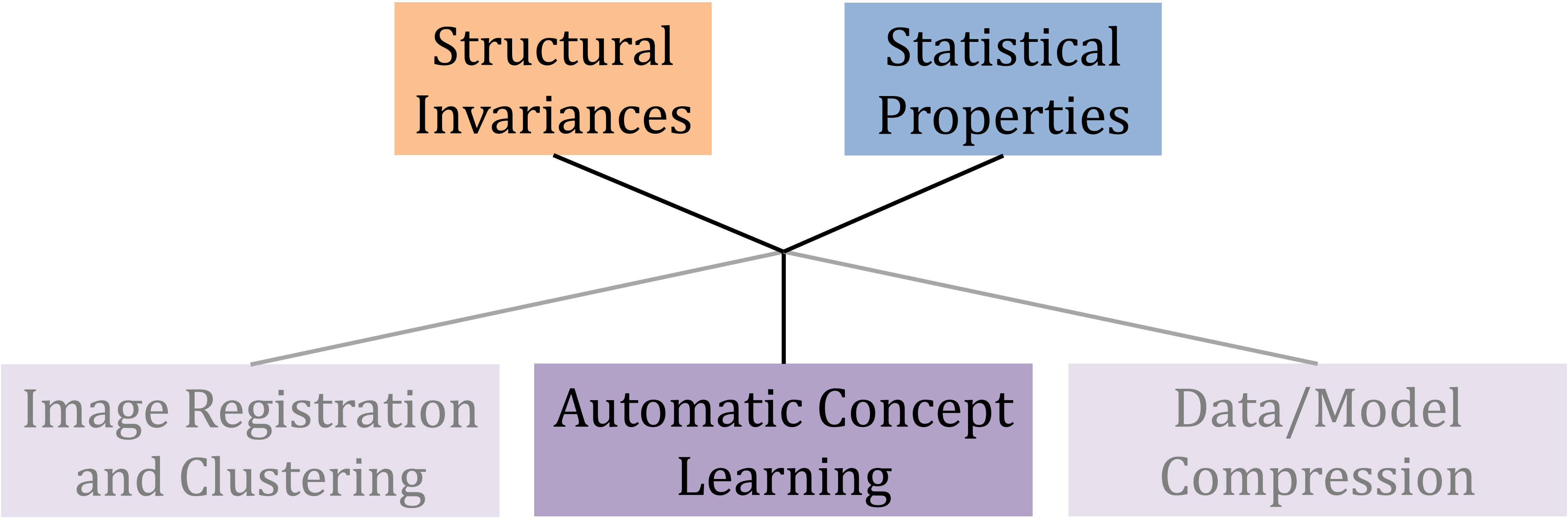
Human-interpretable concept learning

- 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]





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



Structural Invariances

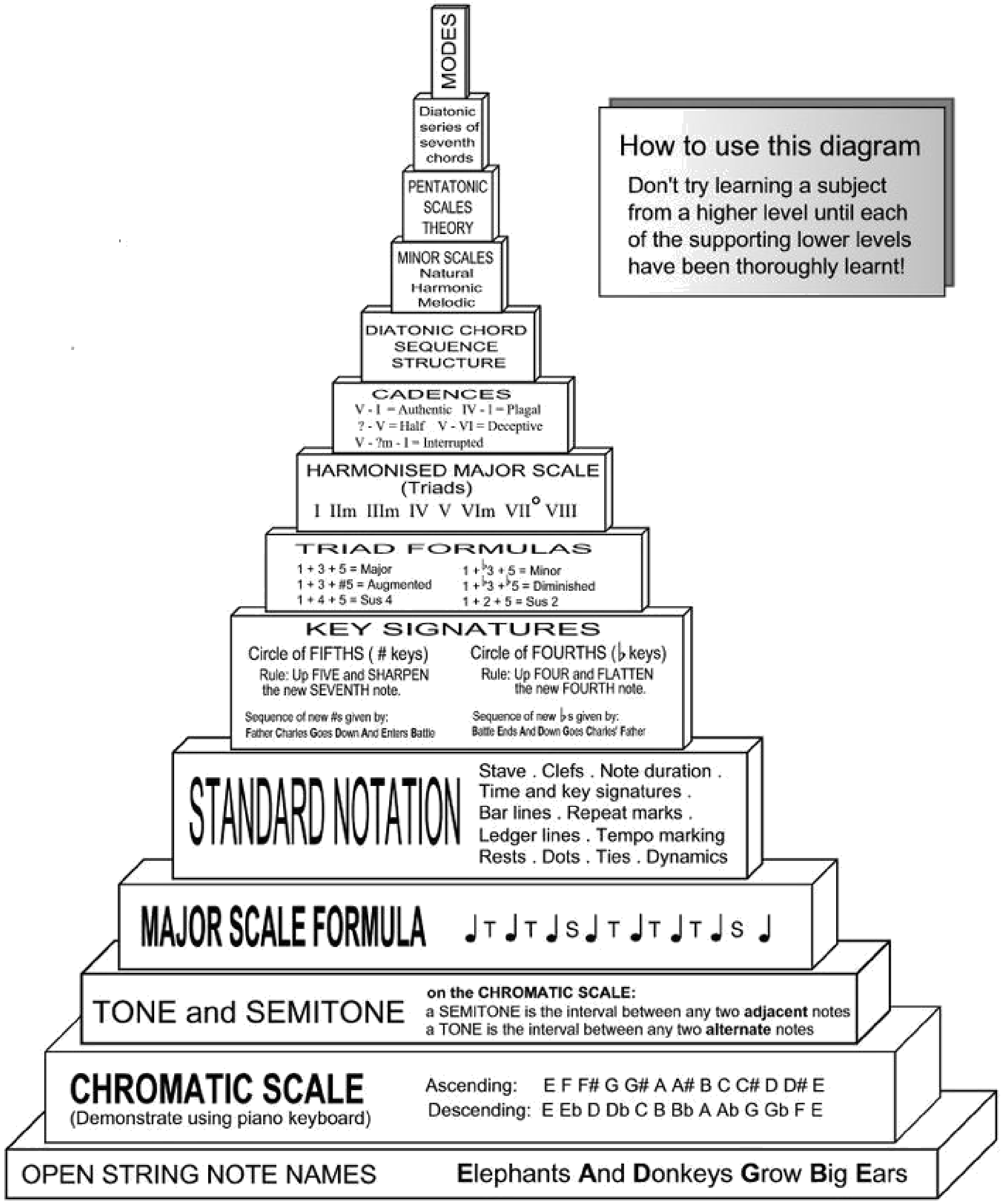
Statistical Properties

Image Registration and Clustering

Automatic Concept Learning

Data/Model Compression

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



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]

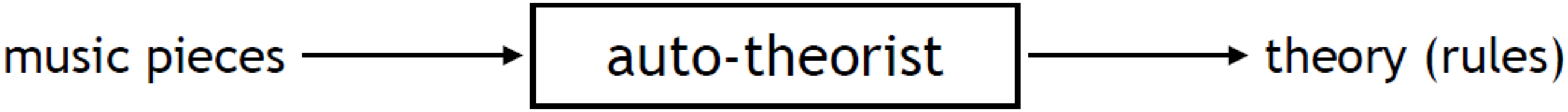
> "Fundamentally, most current deep-learning based language models represent sentences as mere sequences of words, whereas Chomsky has long argued that language has a hierarchical structure, in which larger structures are recursively constructed out of smaller components."
>
> – Gary Marcus [arXiv:1801.00631]

Outline

- Automatic concept learning
  - Group-theoretic framework for abstraction
  - Connections to Shannon’s information lattice
  - Iterative student-teacher algorithm
  - Examples in music theory and elsewhere
  - Applications in safety, ethics, and creativity

Automatic concept learning: An automatic music theorist

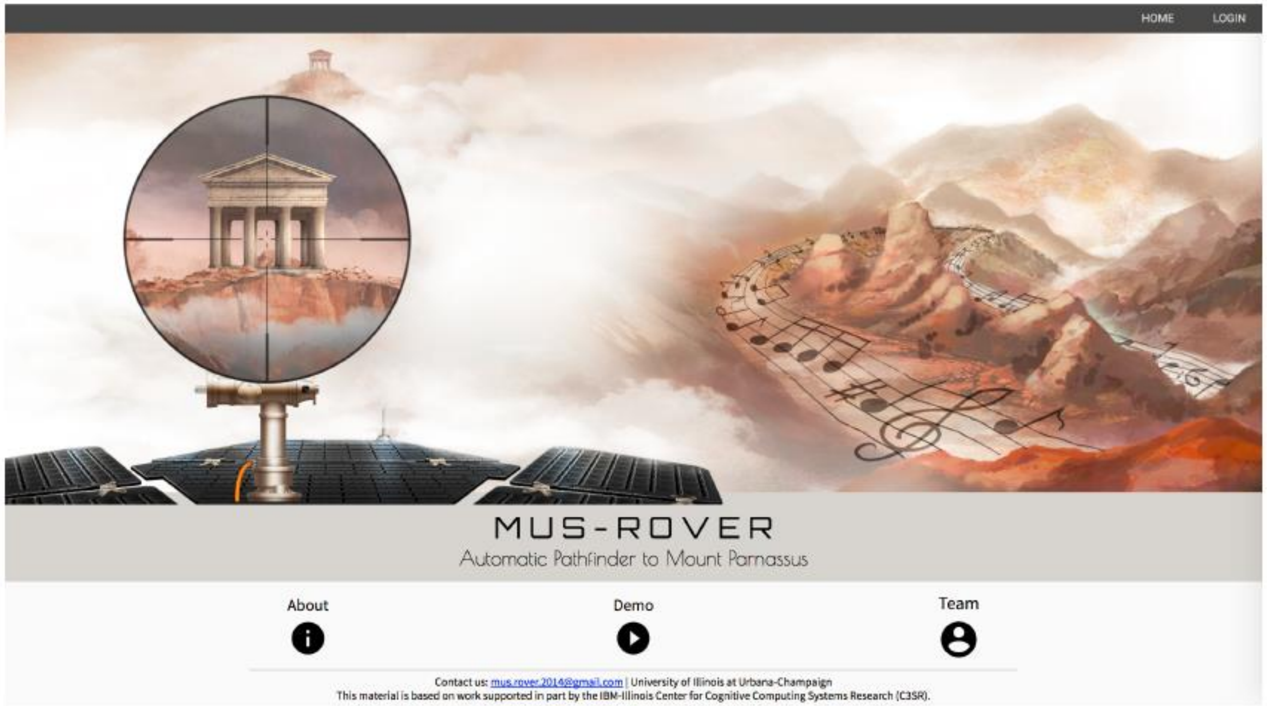
MUS-ROVER, a way to learn the principles of quality (laws of music theory)



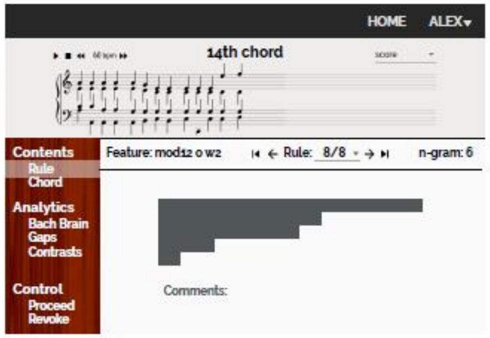
music pieces → auto-theorist → theory (rules)

Computational creativity algorithms for music composition

theory (rules) → auto-composer → music pieces



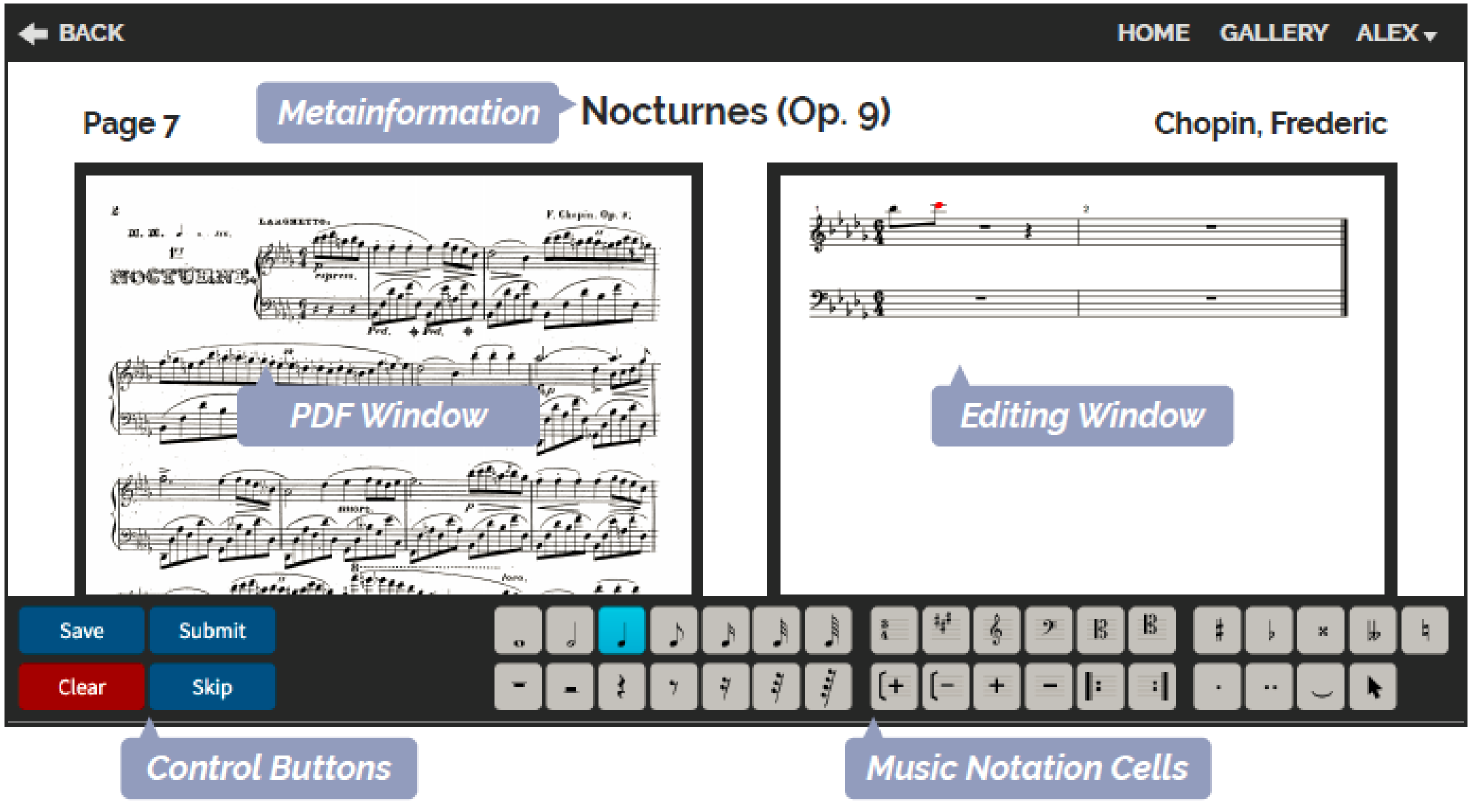
(a) User portfolio



(b) E-classroom

MUS-ROVER

MUS-NET

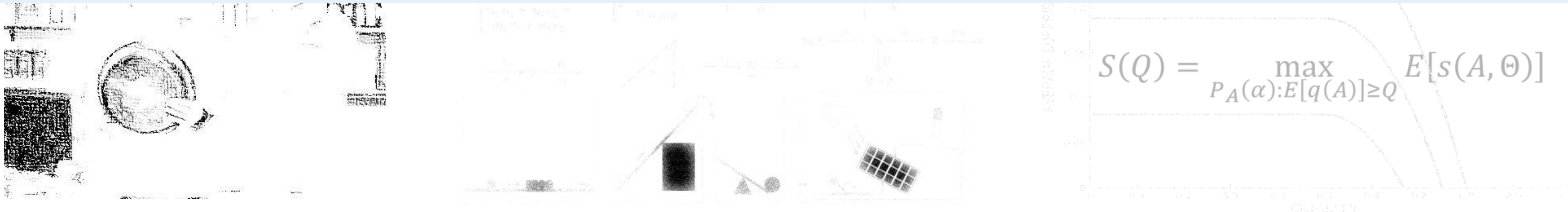


Automatic concept learning: An automatic music theorist

MUS-ROVER, a way to learn the principles of quality (laws of music theory)

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



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

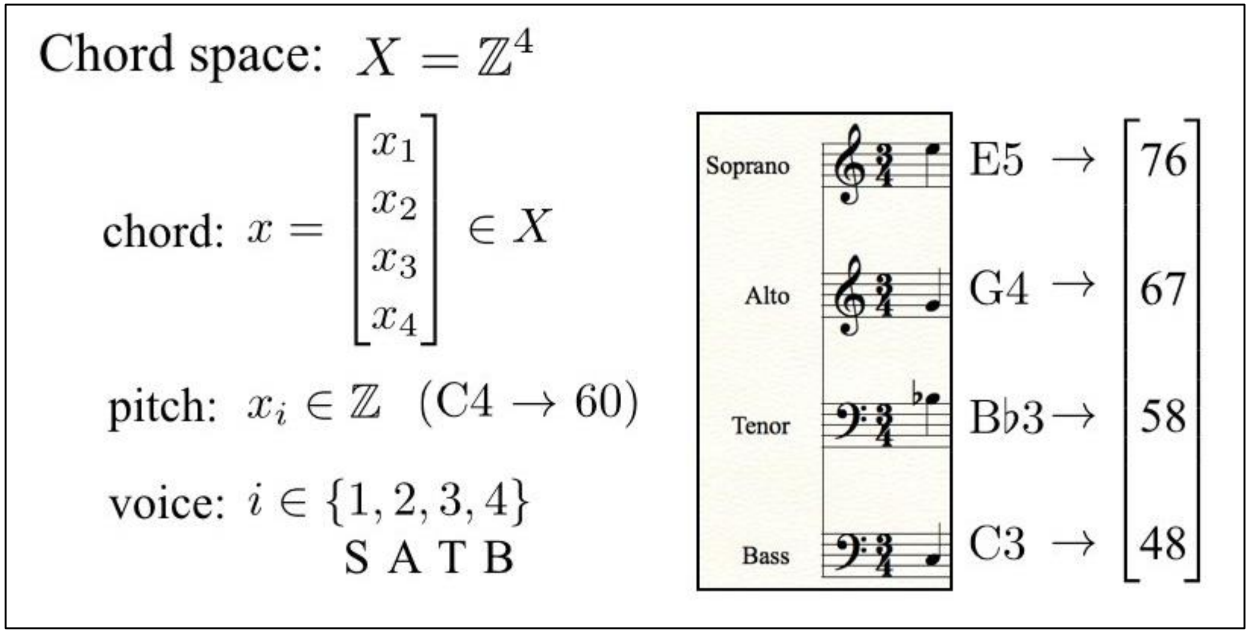
Concept learning as a kind of abstraction process

- Algorithms to automate abstraction in various concept learning tasks almost all require handcrafted priors [Saitta and Zucker, 2013; LeCun et al., 2015; Bredeche et al., 2006; Yu et al., 2016], similar to innate biology [Marcus, 2018; Dietterich, 2018]
  - Whether rules in automatic reasoning, distributions in Bayesian inference, features in classifiers, or architectures in neural networks, typically task-specific and/or domain-specific
- Aim to establish both a theoretical and an algorithmic foundation for abstraction
  - Consider the general question of conceptualizing a domain, a task-free preparation phase before specific problem solving
  - Consider symmetries in nature (or groups in mathematics), a universal prior that encodes no domain knowledge
- Goal is to learn domain concepts/knowledge when our group-theoretic abstraction framework is connected to statistical learning

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: $x_i \in \mathbb{Z}$ $(\mathrm{C4} \to 60)$

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

$\begin{matrix} \mathrm{E5} \to \\ \mathrm{G4} \to \\ \mathrm{B\flat 3} \to \\ \mathrm{C3} \to \end{matrix} \begin{bmatrix} 76 \\ 67 \\ 58 \\ 48 \end{bmatrix}$

Representation: Abstraction

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

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

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

cells (or less formally, clusters)

An **concept** is a partition cell.

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

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

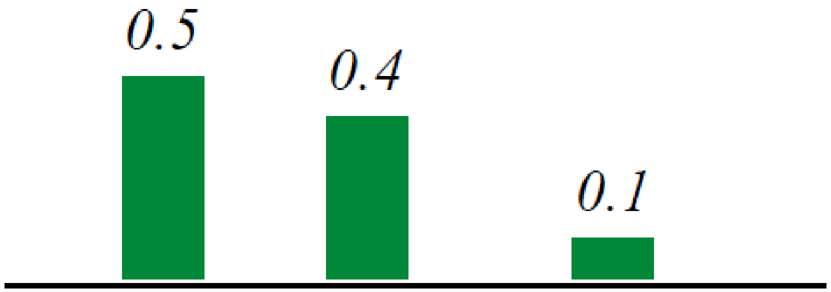
Representation: Probabilistic Rule

A **probabilistic rule** is a pair:

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

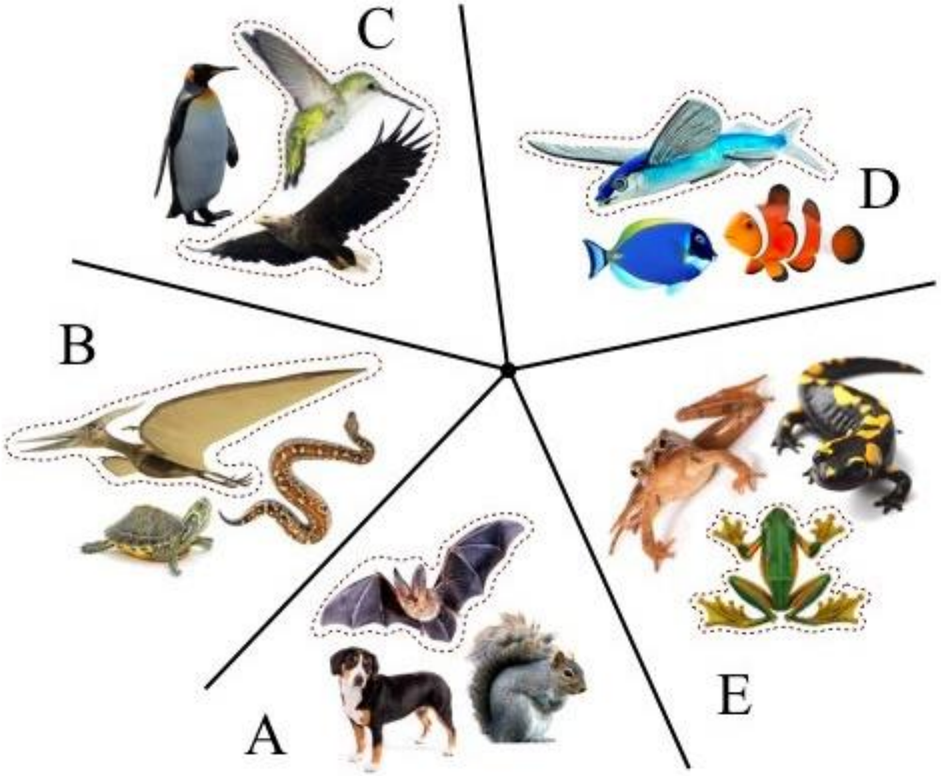
where $\mathcal{A}$ is an abstraction (partition);

$p_{\mathcal{A}}$ 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

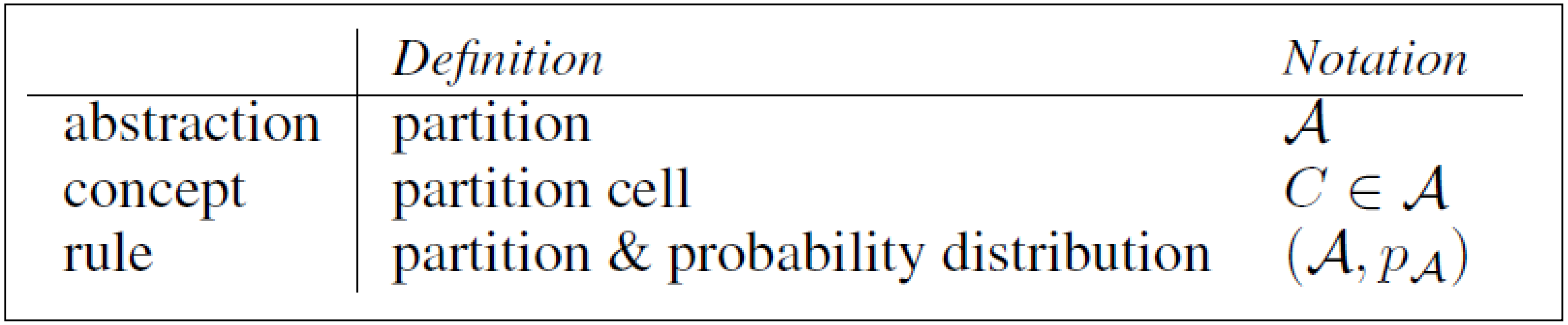
Concepts:

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*

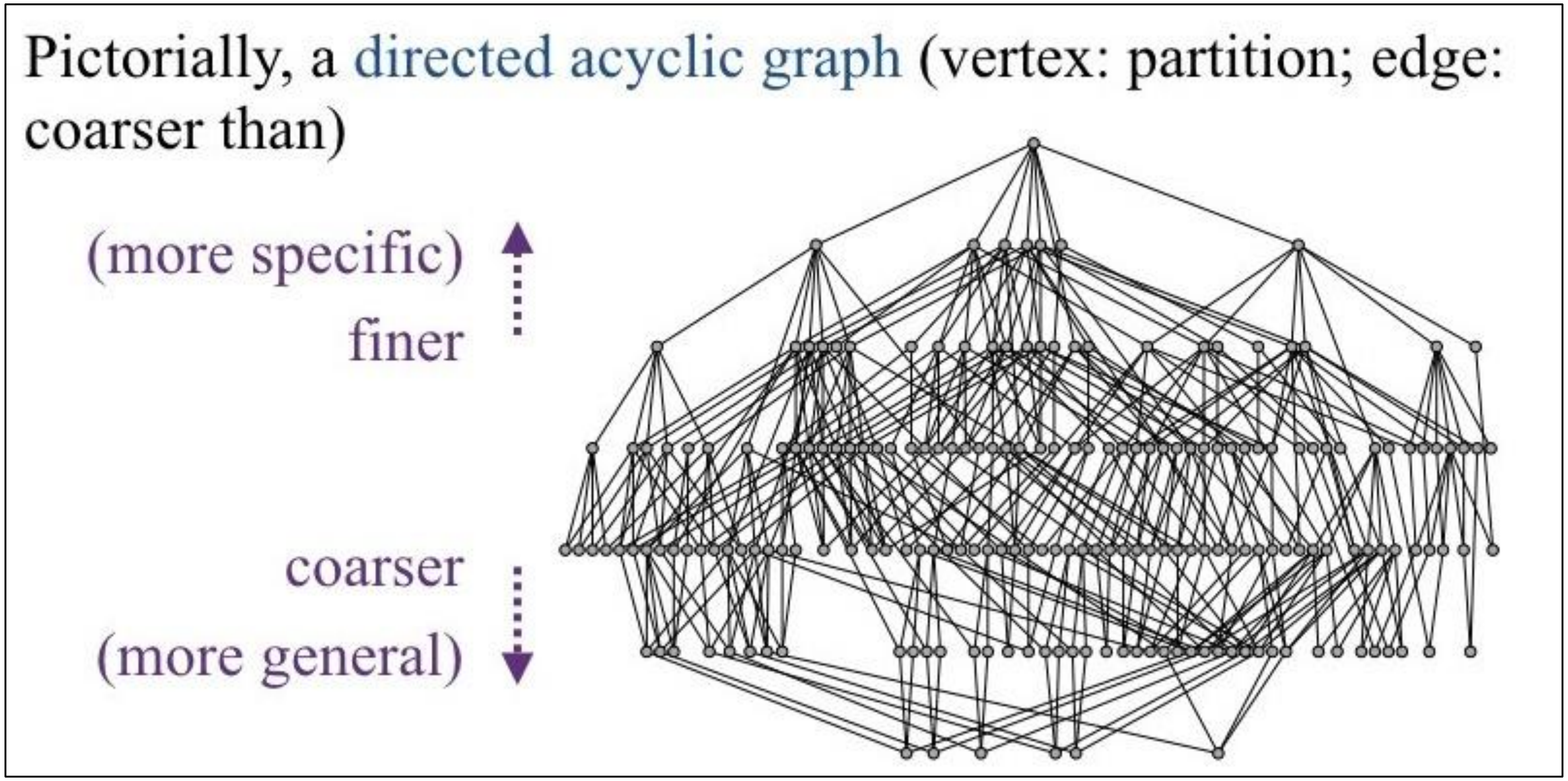


	Definition	*Notation*
abstraction	partition	$\mathcal{A}$
concept	partition cell	$C \in \mathcal{A}$
rule	partition & probability distribution	$(\mathcal{A}, p_{\mathcal{A}})$

- 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

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
  - Induced by equivalent preimages of some feature function $\phi: X \to V$ where V is set of possible feature values
  - 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

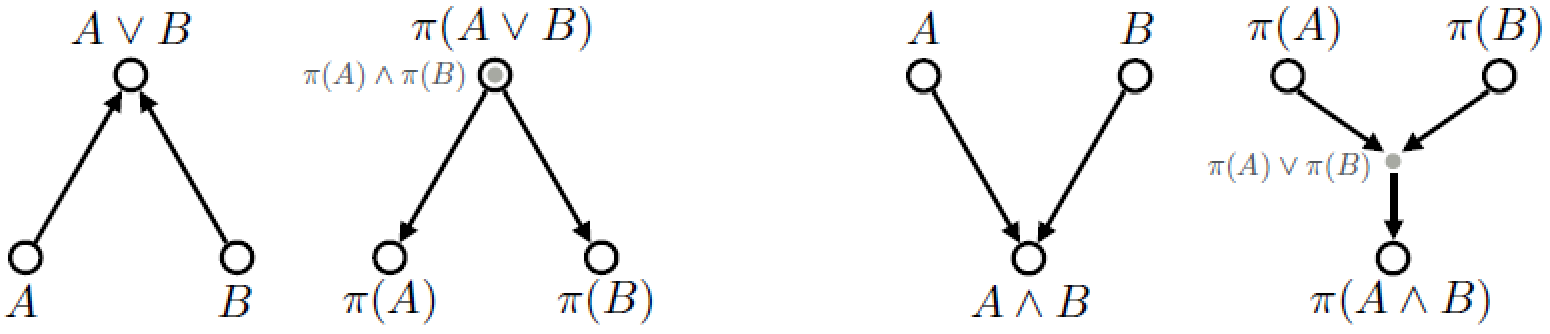
$$\boxed{\text{a subgroup of } \mathrm{F}(X) \xrightarrow{\textit{group action}} \text{orbits} \xrightarrow{\textit{equiv. rel.}} \text{a partition} \xrightarrow{\textit{is}} \text{an abstraction of } X}$$

Duality: From subgroup lattice to abstraction (semi)universe

Definition The *abstraction generating function* is the mapping $\pi: \mathcal{H}^*_{\mathrm{F}(X)} \to \mathfrak{B}^*_X$, 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:

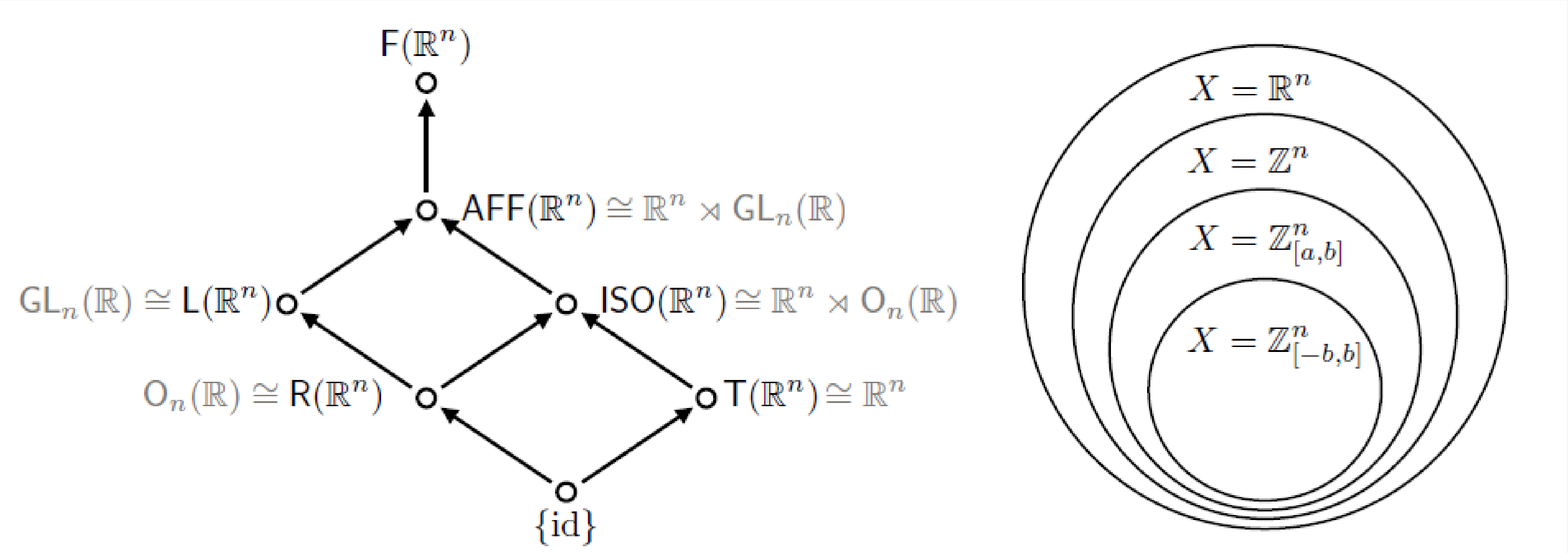
1. partial-order reversal: if $A \leq B$, then $\pi(A) \succcurlyeq \pi(B)$
2. strong duality: $\pi(A \vee B) = \pi(A) \wedge \pi(B)$
3. weak duality: $\pi(A \wedge B) \succcurlyeq \pi(A) \vee \pi(B)$



(a) From join to meet. (b) From meet to join.

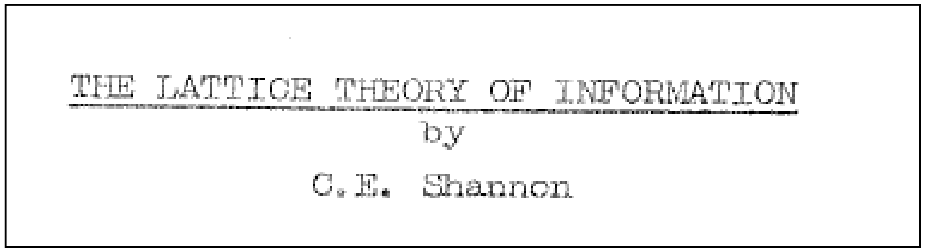
Duality: From subgroup lattice to abstraction (semi)universe

- 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



(a) Subgroups under consideration.

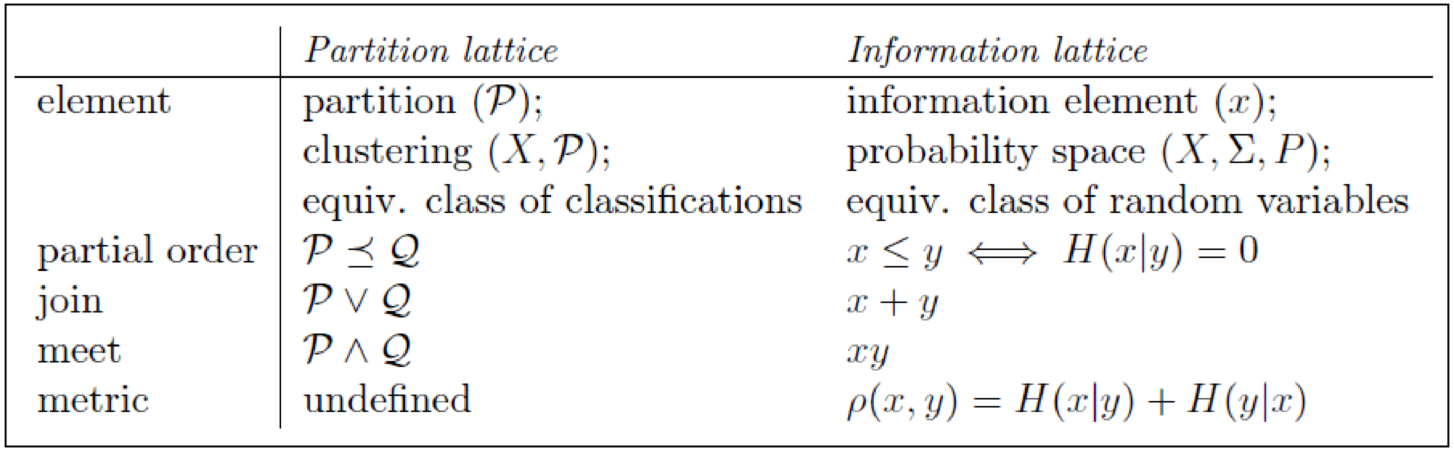
(b) Spaces under consideration.



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 $x \leq y \Leftrightarrow H(x|y) = 0$ 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



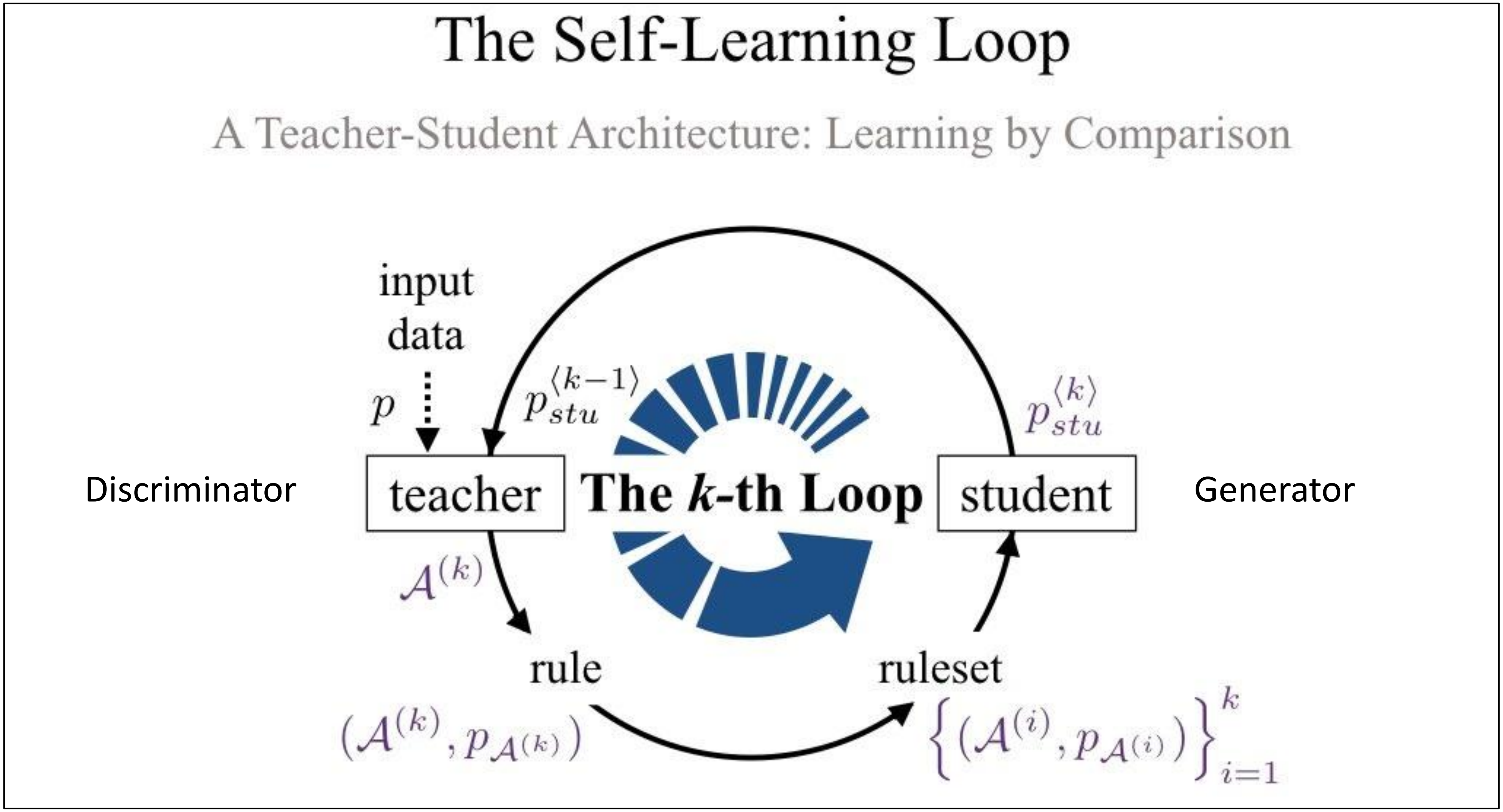
	Partition lattice	*Information lattice*
element	partition ($\mathcal{P}$); clustering ($X, \mathcal{P}$); equiv. class of classifications	information element (x); probability space (X, Σ, P); equiv. class of random variables
partial order	$\mathcal{P} \preceq \mathcal{Q}$	$x \leq y \iff H(x\|y) = 0$
join	$\mathcal{P} \vee \mathcal{Q}$	$x + y$
meet	$\mathcal{P} \wedge \mathcal{Q}$	xy
metric	undefined	$\rho(x, y) = H(x\|y) + H(y\|x)$

Outline

- Automatic concept learning
  - Group-theoretic framework for abstraction
  - Connections to Shannon’s information lattice
  - Iterative student-teacher algorithm
  - Applications in music theory and elsewhere

Information-theory inspired algorithm for rule learning

Learning is achieved by statistical inference on a partition lattice



Discriminator

Generator

Teacher: a Discriminative Model

The teacher solves an optimization problem:

$$\begin{array}{ll} \underset{\mathcal{A} \in \mathfrak{P}_X}{\text{maximize}} & D_{KL}\left(p_{\mathcal{A},stu}^{\langle k-1 \rangle} \,\|\, p_{\mathcal{A}}\right) \\ \text{subject to} & \text{the abstraction } \mathcal{A} \text{ satisfying} \\ & \text{the memorability condition} \\ & \text{the hierarchy condition} \\ & \cdots \end{array}$$

$\mathfrak{P}_X \subseteq \mathfrak{P}_X^*$: abstraction hierarchy
part of the complete partition lattice

Try to adjust the student's information lattice
to match the target information lattice (input).

Student: a Generative Model

The student solves another optimization problem:

$$\begin{aligned} \underset{p_{stu}^{\langle k\rangle}\in\Delta_{|\mathcal{X}|}}{\text{maximize}} \quad & \overbrace{S_q(p_{stu}^{\langle k\rangle})}^{\text{Tsallis entropy: measures randomness}} := (q-1)^{-1}\left(1-\|p_{stu}^{\langle k\rangle}\|_q^q\right) \\ \text{subject to}\quad & \underbrace{\underbrace{A^{(i)}}_{\text{partition matrix: represents abstraction}} p_{stu}^{\langle k\rangle} = p_{\mathcal{A}^{(i)}},}_{\text{linear equality constraint}} \quad i=1,\dots,k \end{aligned}$$

$q = 2$: gini impurity function

Linear Least-Squares Problem!

Information-theory inspired algorithm for rule learning

Learning is achieved by statistical inference on a partition lattice

MUS-ROVER's self-learning loop:

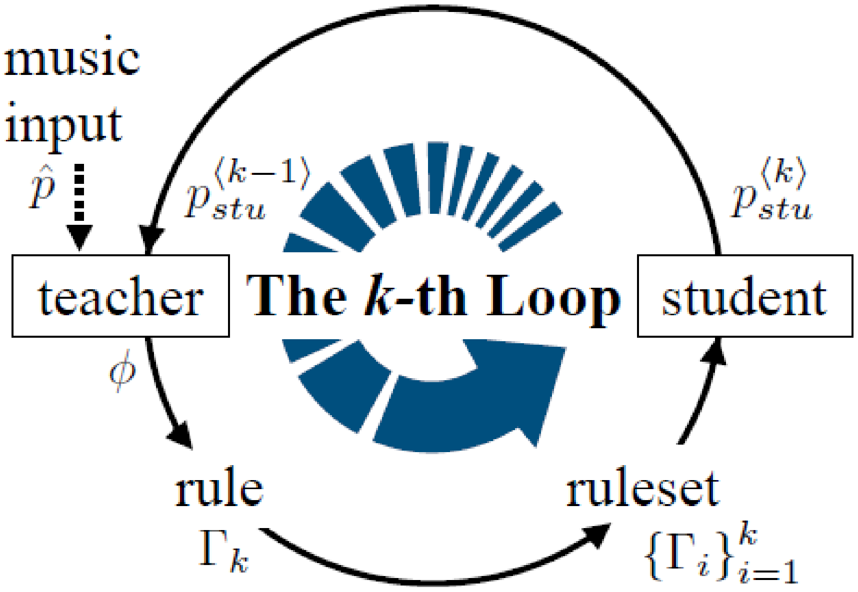
The iterative cooperation between a discriminator (teacher) and a generator (student).

The teacher solves:

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

(discrete optimization)

max. Bayesian surprise



The student solves:

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

(linear least-squares)

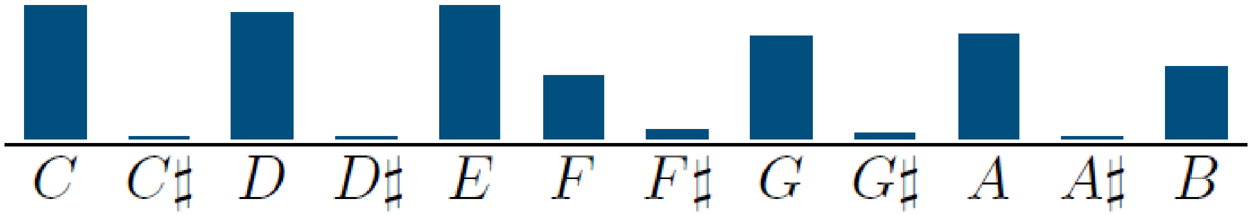
max. creativity

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:

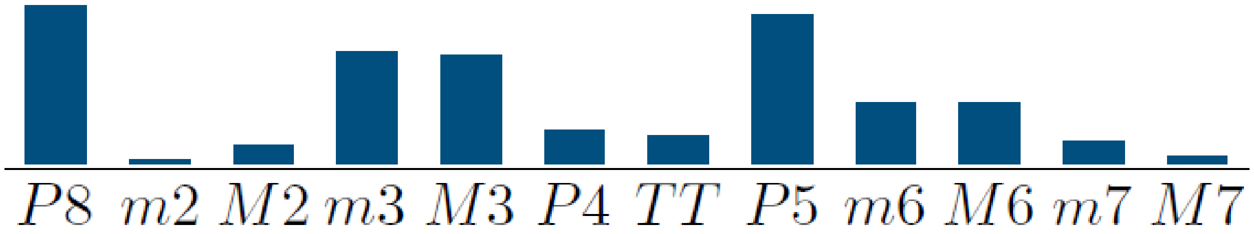
“The soprano voice is built on a diatonic scale.”

Hierarchical concept learning

Compositional Rule Examples:

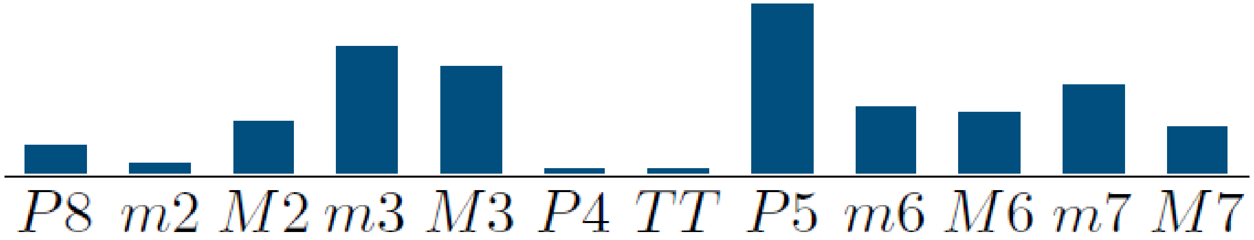
ϕ : interval class between soprano and bass

p_ϕ



1-gram

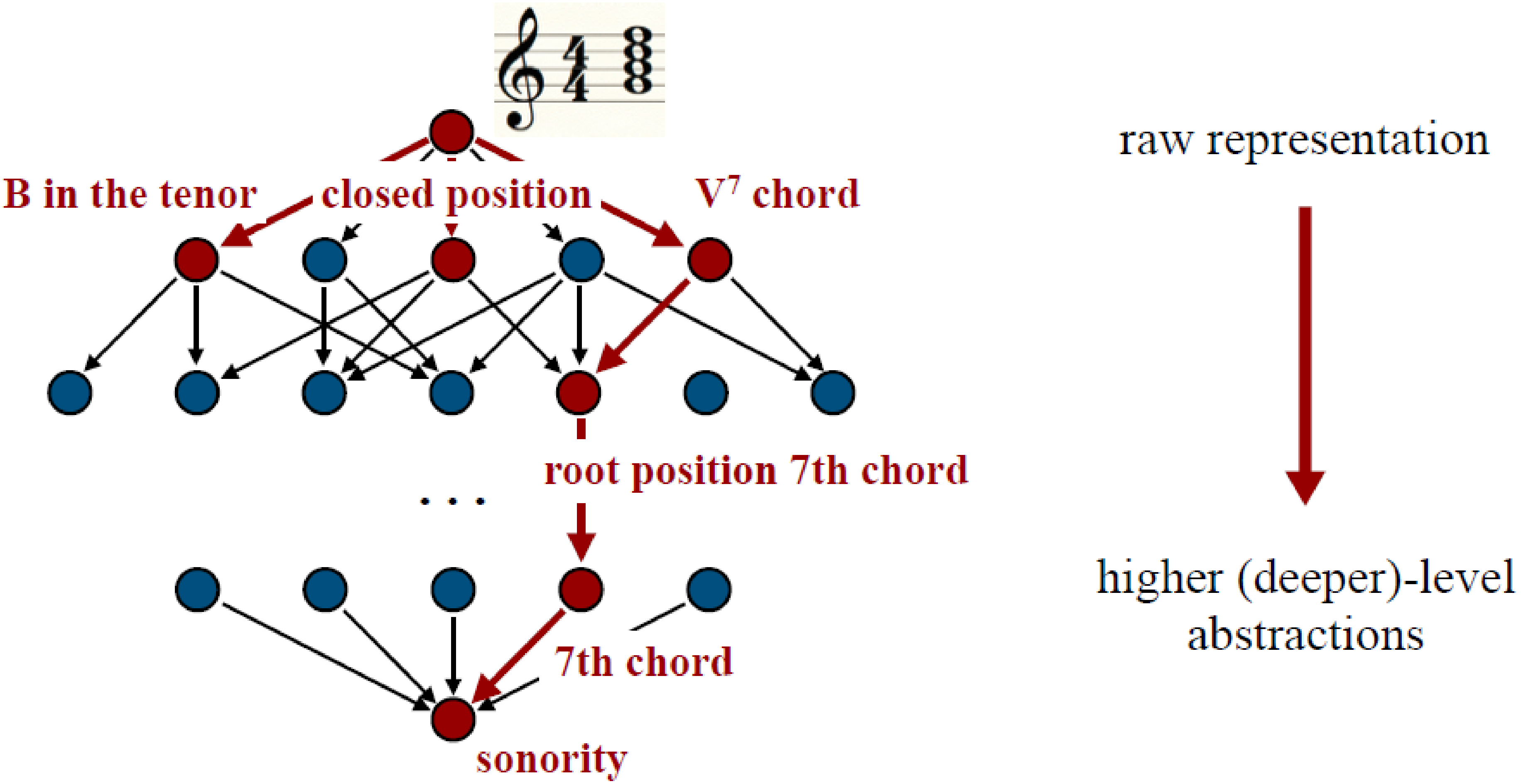
"Individual perfect octaves (P8s) are favored as most consonant."



2-gram
conditioned on P8

"Parallel perfect octaves (P8s) are uncommon."

Hierarchy of music theory concepts



B in the tenor

closed position

V^7 chord

root position 7th chord

. . .

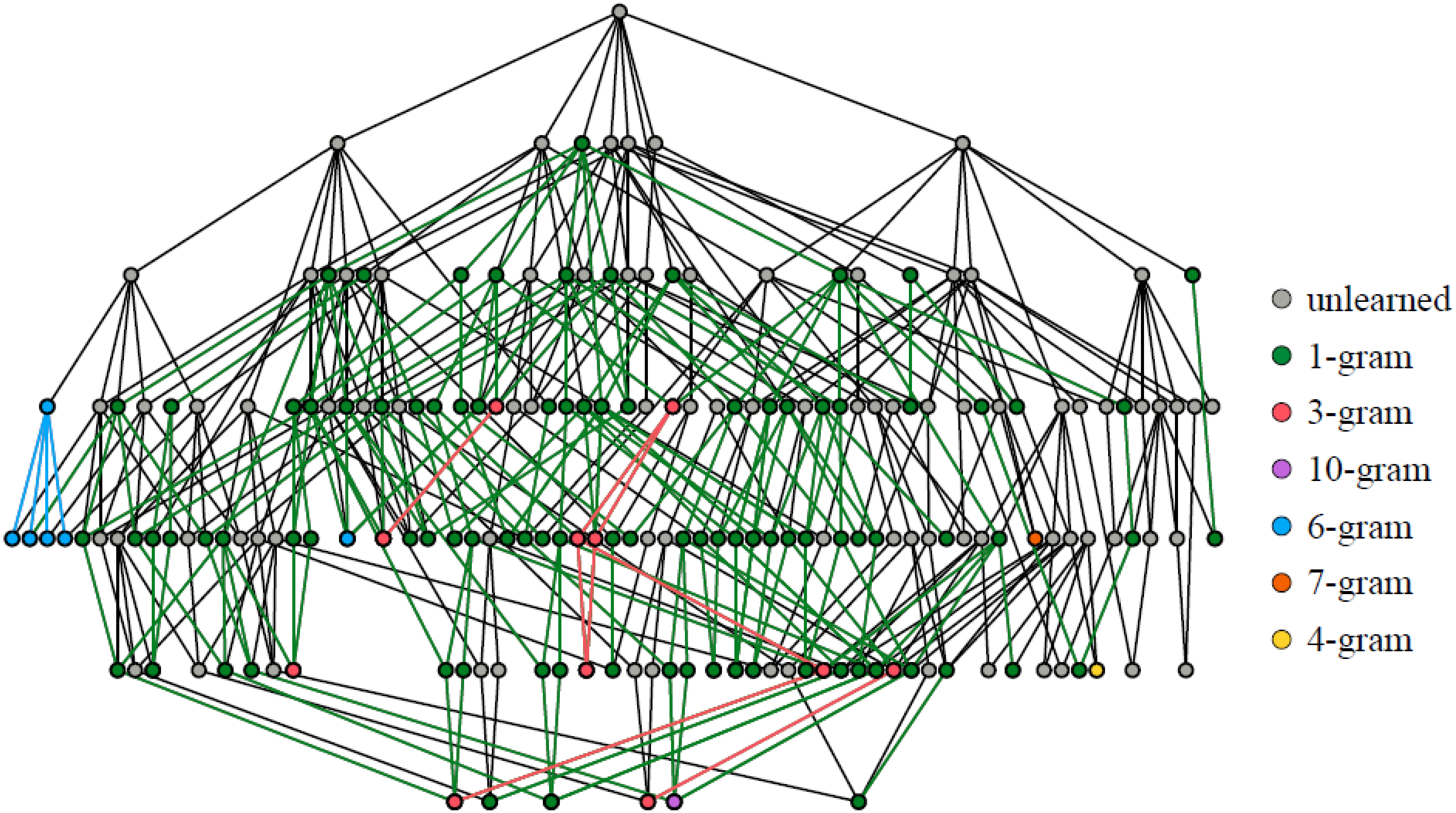
7th chord

sonority

raw representation

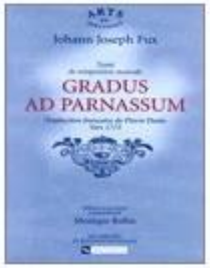
higher (deeper)-level abstractions

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

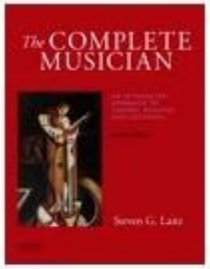


Visualization of Bach's music mind for writing chorales. The underlying directed acyclic graph signifies an upside-down information lattice.

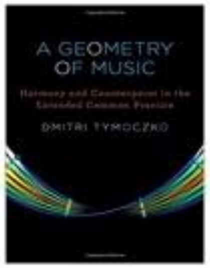
MUS-ROVER recovers nearly all known music theory



- voice leading
- counter point

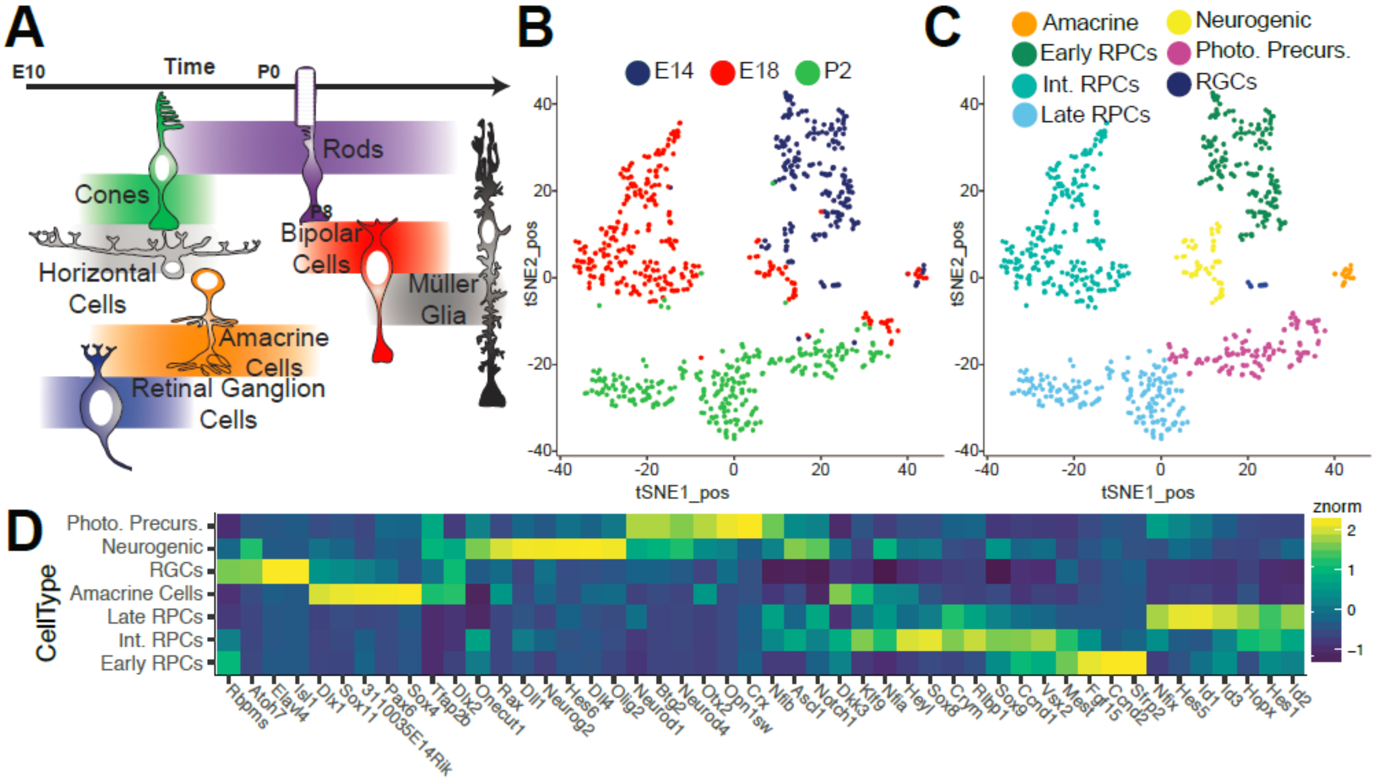


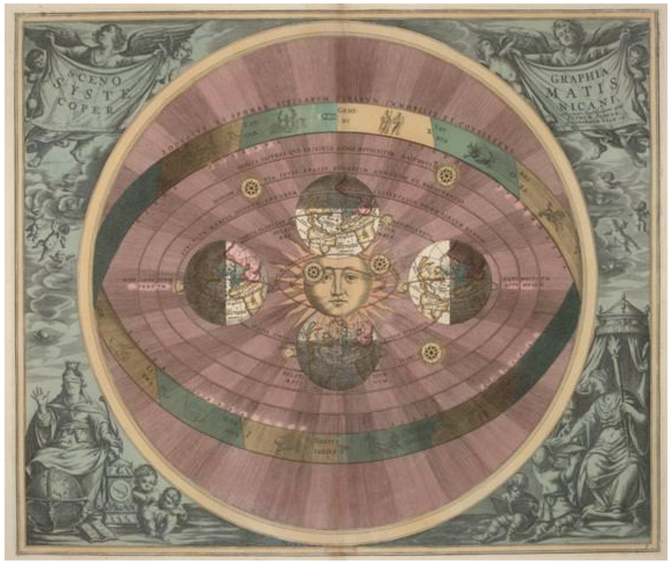
- scale, consonance & dissonance
- voice spacing, crossing, overlap
- chord quality, inversion, progression



- music transformations: OPTIC

Generalizing to other topic domains





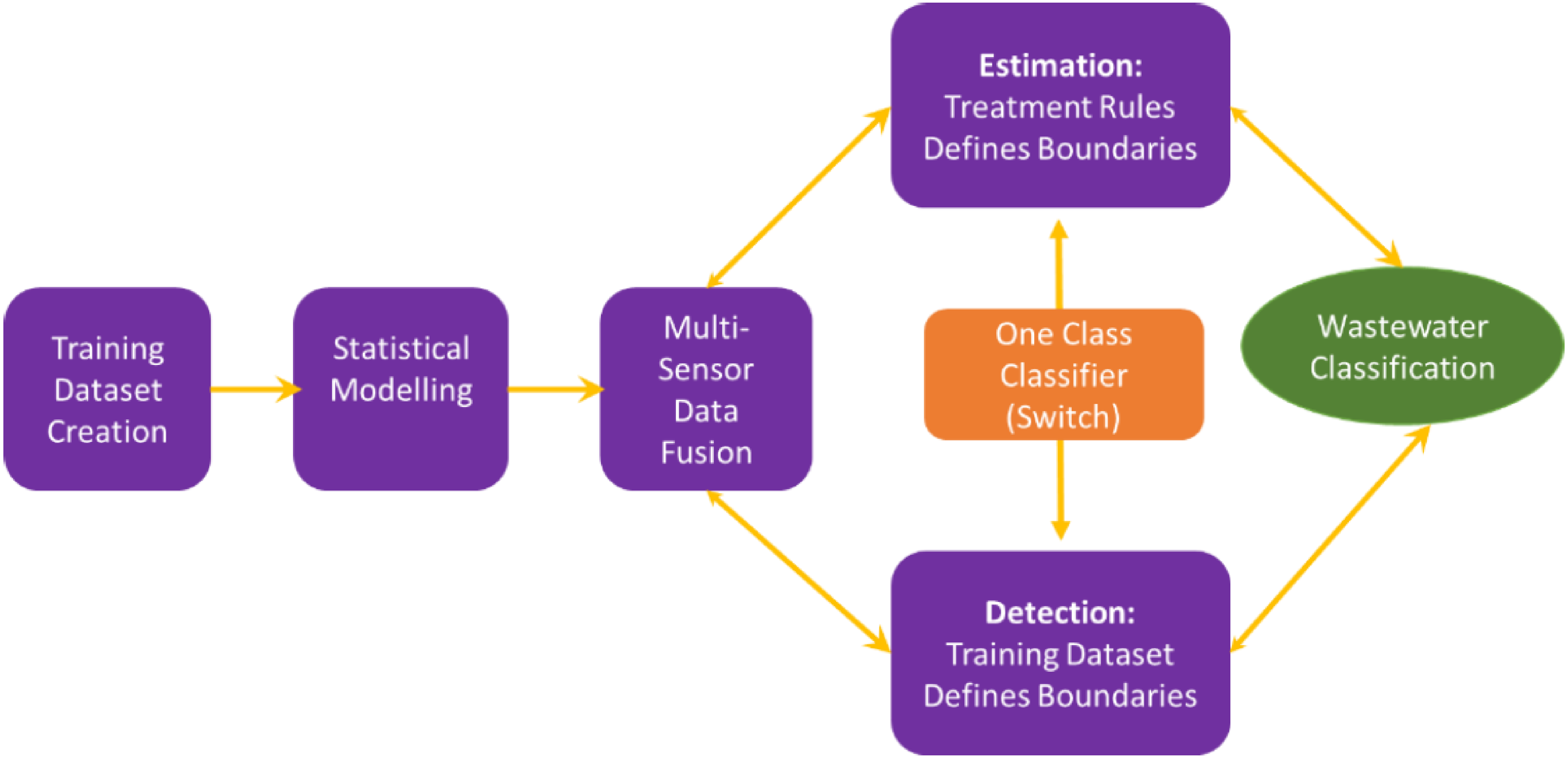
[B. Clark, et al., "Comprehensive analysis of retinal development at single cell resolution identifies NFI factors as essential for mitotic exit and specification of late-born cells," bioRxiv, 2018.]

- Single-cell RNA sequence data analysis for understanding the rules that govern pattern formation in neurodevelopment
- Rediscover physical laws and principles such as heliocentrism using data from Copernicus

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]

Algorithm fusion to deal with epistemic uncertainty



Training Dataset Creation

Statistical Modelling

Multi-Sensor Data Fusion

Estimation: Treatment Rules Defines Boundaries

One Class Classifier (Switch)

Detection: Training Dataset Defines Boundaries

Wastewater Classification

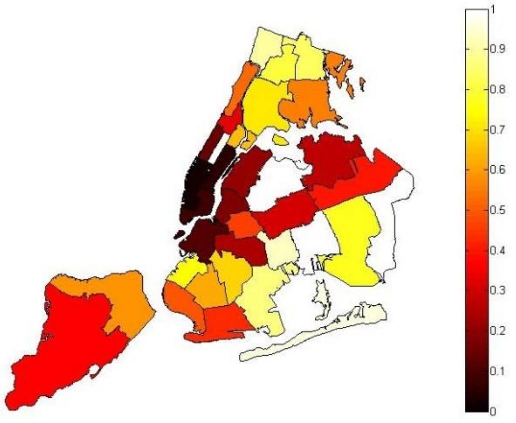
[N. Kshetry and L. R. Varshney, “Safety in the Face of Unknown Unknowns: Algorithm Fusion in Data-Driven Engineering Systems,” to appear in *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, England, 12-17 May 2019.]

AI for social 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**

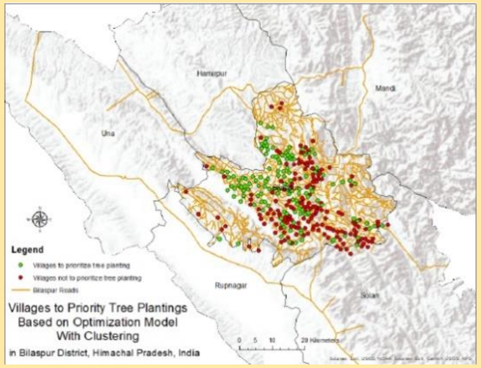
- **bipartite ranking + spatiotemporal modeling**

[*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]



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]

The need to control unintended consequences (FAT)

Human Resources: **Rank candidates according to likelihoods of quality, onboarding, and attrition using historical training data**

- **bipartite ranking**

[*Int. Conf. Extending Database Technology (EDBT)*, 2013]

KRISHNA PATTABHIRAMAN
508-410-0168 (m) | pat_krishna@yahoo.com
19 Butterfield Drive, Westborough, MA 01581

TECHNOLOGY INNOVATION EXECUTIVE
PRODUCT MANAGEMENT & STRATEGY | R & D

Strategic and forward-thinking Product Management and Software/Hardware Development Executive with more than twenty years of success leading innovative product development in networking, RTLS, wireless and software segments. Not only ideates needs-based and profitable products, but creates actionable go-to-market strategies, drives the entire development lifecycle, and forges relationships that accelerate revenue and expand into new markets. Business success complemented by exceptional academic background including doctorate in **Computer Science**.

Research & Development, Go-to-Market Strategy, Product Lifecycle Management, Standards Development
IT Strategy, IT Infrastructure/Design, Software Development, Project Management

PROFESSIONAL EXPERIENCE

Reva Systems – Westford, MA 2004 – Present
Provider of RFID and RTLS products and solutions currently deployed at 600+ sites worldwide.

VICE PRESIDENT
One of the founding team members, acted as the chief architect, subject matter expert, and made key contributions to Reva's product development. Authored worldwide standard and represented [illegible] at standards groups. As the head of solutions and technology group, aligned Reva's technology and product strategy with the needs and requirements of global customers planning to deploy RFID and RTLS solutions. Led solution engineering team and managed solutions development teams in China and US.

- **Re-strategized and re-focused the company** to Operational Patient Logistics in Healthcare - market sizing, go-to-market, competitive analysis, product positioning and validation, and execution plans.
- **Established channel partnerships with Europe and Middle East** to build recurring revenue. Negotiated and closed a diverse set of clients to maximize growth.

Racial Discrimination: **Bounded rationality by human agents, together with segregation, yields information-based model of discrimination**

- **statistical signal processing + quantization theory**

[*Proceedings of the IEEE*, 2017]

Child Abuse: **Prioritize reported cases of child abuse according to likelihoods of indication and severity, under fairness constraints**

- **bipartite ranking + queuing theory**

[*18th National Conference on Child Abuse and Neglect*, 2012]

Impact Sourcing: **Improve efficiency of routing and scheduling tasks to workers in Samasource system to enable more people out of poverty**

- **constrained max-weight scheduling**

[*IEEE/ACM Transactions on Networking*, 2018]

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

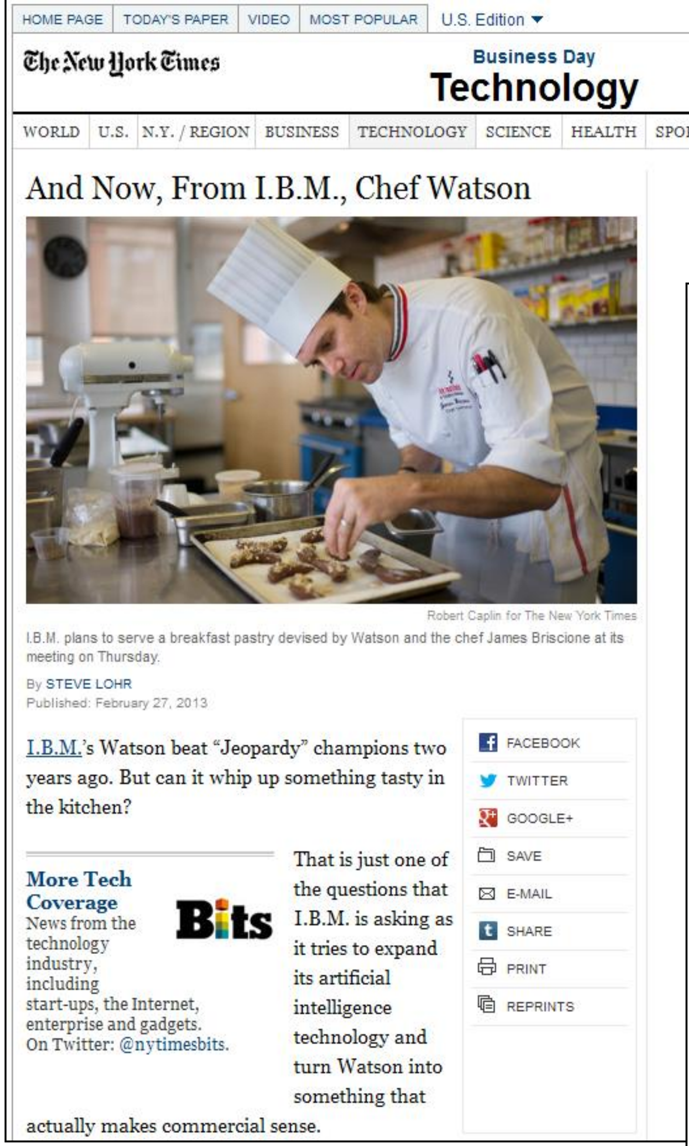
(All of these principles should, *prima facie*, be held and when in conflict should be given equal weight)

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

An ethical framework from biomedicine [Beauchamp and Childress]

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

Human-interpretable machine learning may provide avenues for addressing these ethical challenges, especially autonomy



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.

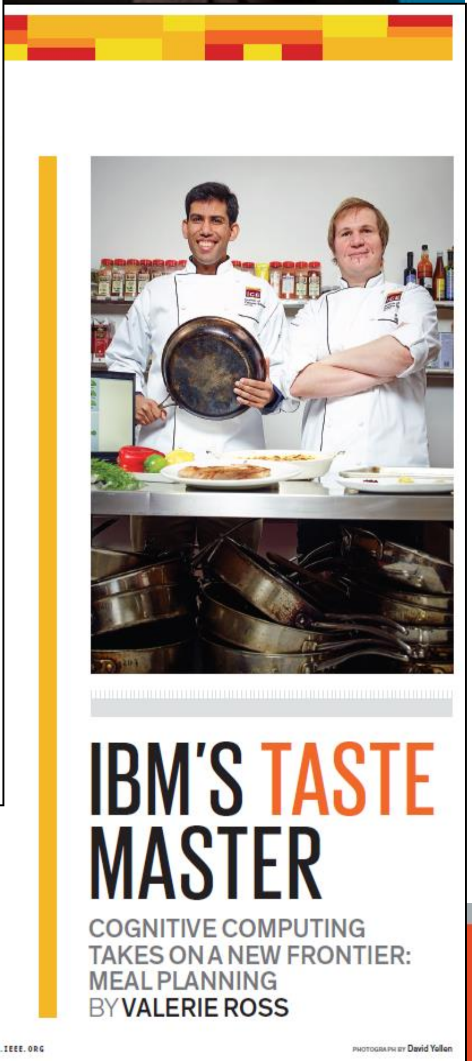
IBM's AI-like computer systems aren't limited to Watson, the *Jeopardy*-winning supercomputer that schooled Ken Jennings on national television. In fact, IBM researchers foresee a not-so-distant future when algorithms will be a replacement for inefficient customer service models, a diagnostic tool for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

Prop styling: Laurie Raab | Justin Fantl



IBM'S TASTE MASTER

COGNITIVE COMPUTING TAKES ON A NEW FRONTIER: MEAL PLANNING

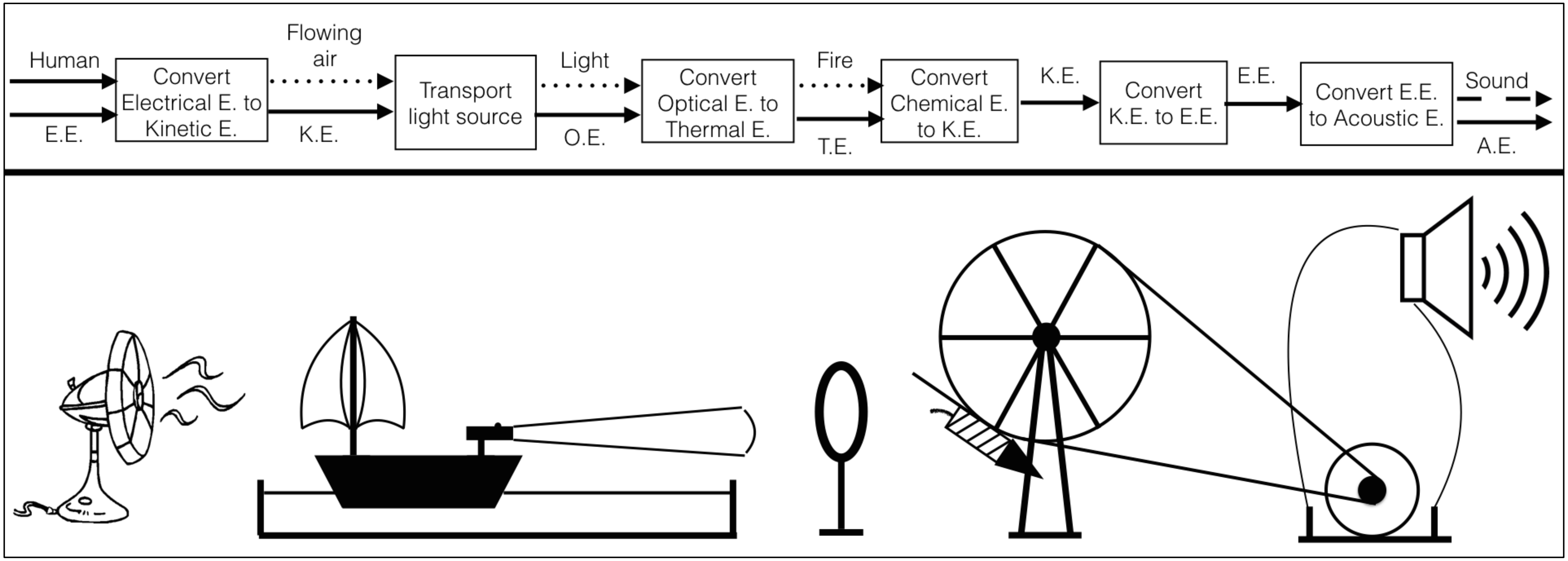
BY VALERIE ROSS

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

PHOTOGRAPH BY David Yellen

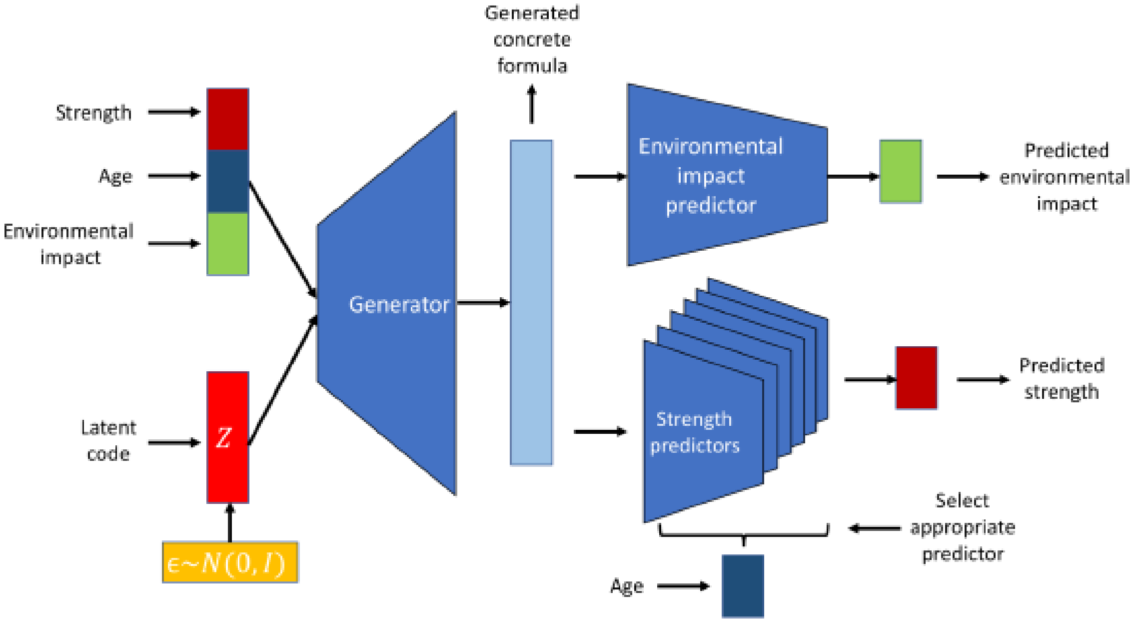
[*The New York Times*, 27 Feb. 2013]
[*San Jose Mercury News*, 28 Feb. 2013]
[*IEEE Spectrum*, 31 May 2013]
[*Wired*, 1 Oct. 2013]

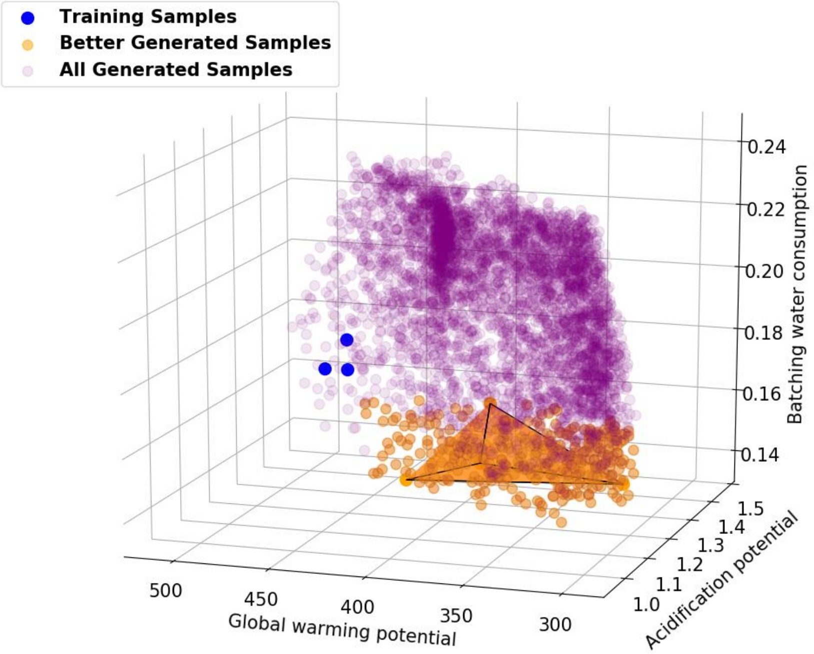
Engineering processes: Rube Goldberg Machines



[X. Ge, J. Xiong, and L. R. Varshney, “Computational Creativity for Valid Rube Goldberg Machines,” in *Proceedings of the Ninth International Conference on Computational Creativity (ICCC)*, Salamanca, Spain, 25-29 June 2018.]

Sustainable building materials

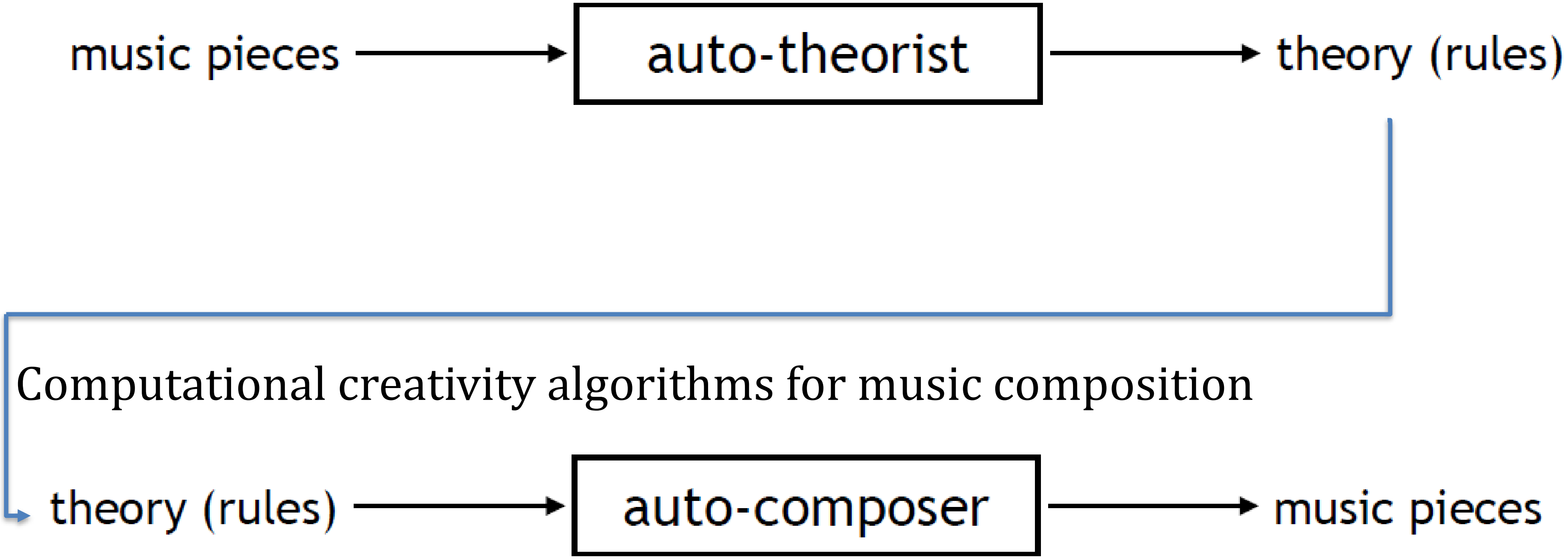




[X. Ge, R. T. Goodwin, J. R. Gregory, R. E. Kirchain, J. Maria, and L. R. Varshney, "Accelerated Discovery of Sustainable Building Materials," to appear in *Proceedings of the AAAI Spring Symposium on Towards AI for Collaborative Open Science*, Palo Alto, California, 25-27 March 2019.]

From automatic music theorist to compose

MUS-ROVER, a way to learn the principles of quality (laws of music theory)



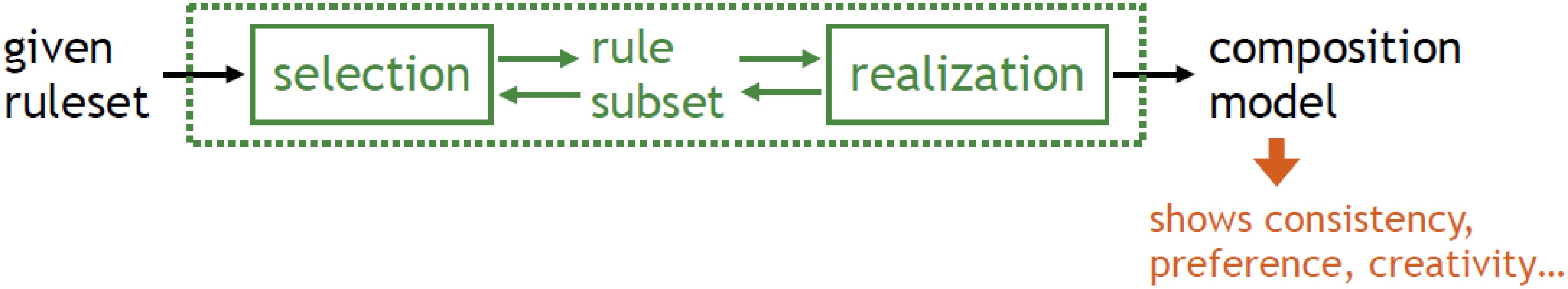
music pieces → **auto-theorist** → theory (rules)

Computational creativity algorithms for music composition

theory (rules) → **auto-composer** → music pieces

In creative composition, want to break rules with a consistent style

Simultaneous Rule Realization and Selection



Solve a simplex constrained bi-convex problem:

$$\begin{aligned} \text{minimize} \quad & \mathcal{E}(p, w; A, b) + \lambda_p P_p(p) + \lambda_w P_w(w) \\ \text{subject to} \quad & p \in \Delta^n, w \in \Delta^m. \end{aligned}$$

- p: composition model
- w: weights of individual rule components
- $P_p(p)$: group penalty
- $P_w(w)$: group elastic-net

Interpretable concept learning to enable augmented intelligence

Problem Solving
Moral Reasoning
Safety
Creativity
Transfer

Lav R. Varshney, University of Illinois at Urbana-Champaign

varshney@Illinois.edu @lrvarshney