ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

inf | Informatik
Computer Science

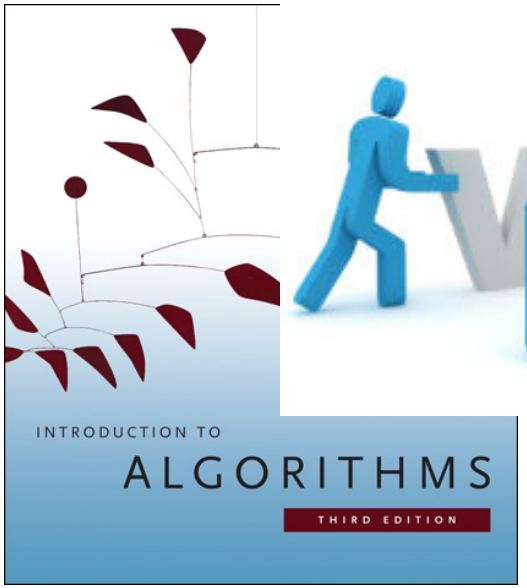
Resilient Algorithmics –

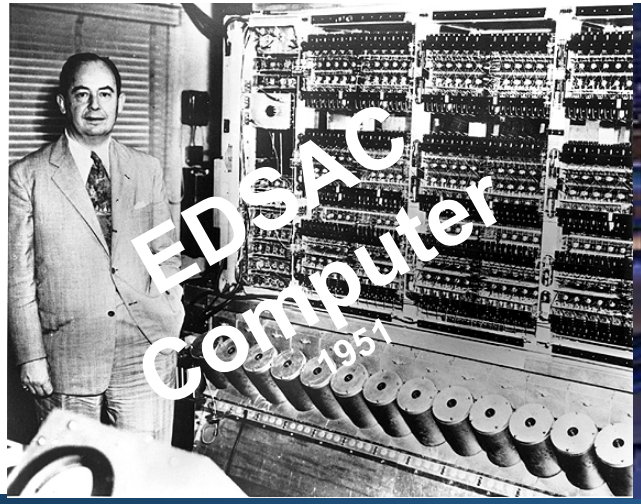
How Much Information is Extracted by an Algorithm from my Data?

Joachim M. Buhmann

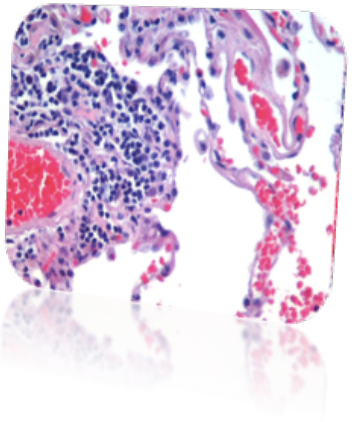
Computer Science Department, ETH Zurich

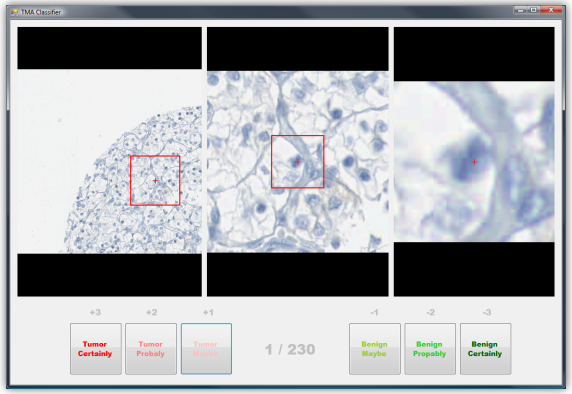
The world, in which we live!

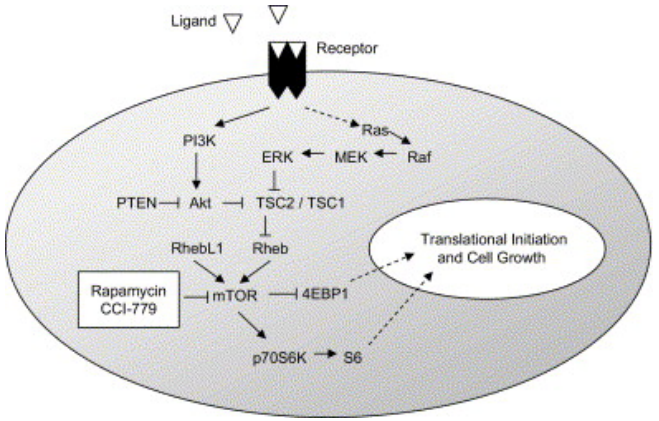




IT value for personalized medicine

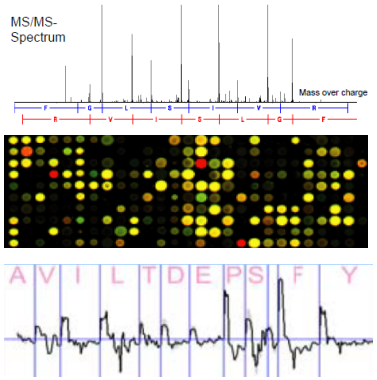


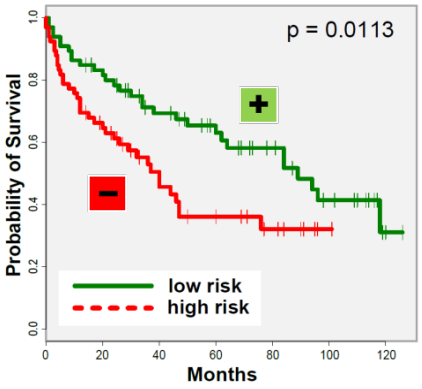




Activation of the mTOR Signaling Pathway in Renal Clear Cell Carcinoma. Robb et al., J Urology 177:346 (2007)

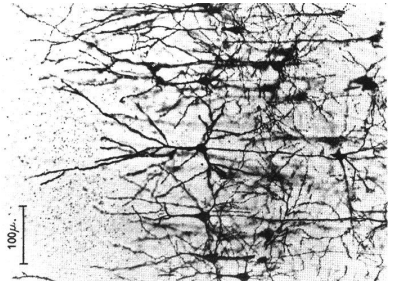
my **Data** → *my* **Information** → *our* **Knowledge**



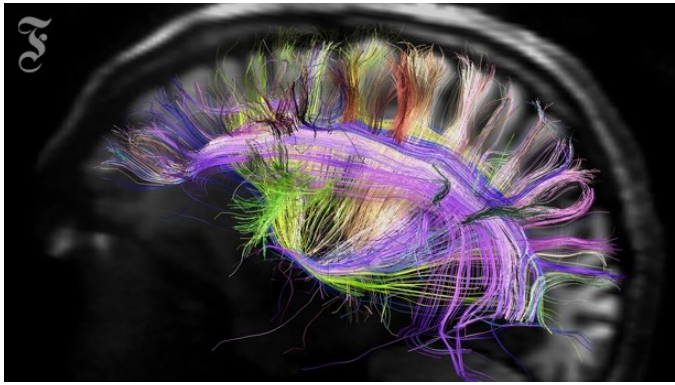


my **Value**

Learning machines master algorithmic induction

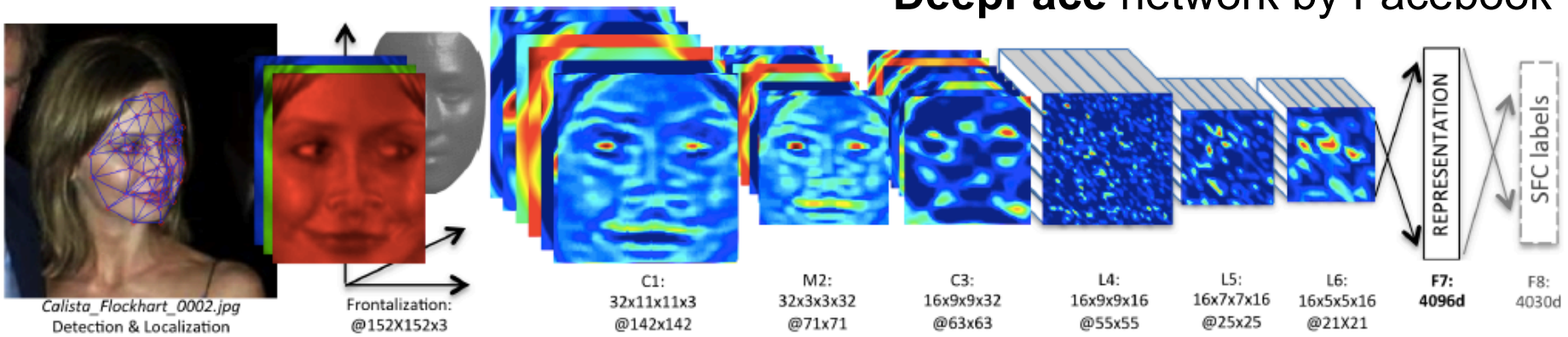


- **Biological neural networks** are adaptive and capable of learning.
- **Artificial neural networks** model these learning capabilities.



Nervenzell-Netze mit Hirnscans sichtbar gemacht. © VAN WEDEEN

DeepFace network by Facebook



Roadmap

- **Algorithms for Data Science: A quest for resilience?!**
- **Algorithm/Model validation** by information theory
  Learning optimal algorithms as open challenge!
- **Examples**
  - Comparing approximate spanning tree algorithms
  - Cortex parcellation
  - Sparse Minimum Bisection of random graphs

The Algorithm: Idiom of Modern Science (Bernard Chazelle)

- Informally, an **algorithm** is any well-defined computational procedure, that takes some value as **input** and produces some value as **output**. (CLRS)

Muḥammad ibn Mūsā al-Khwārizmī[
(c. 780 – c. 850)

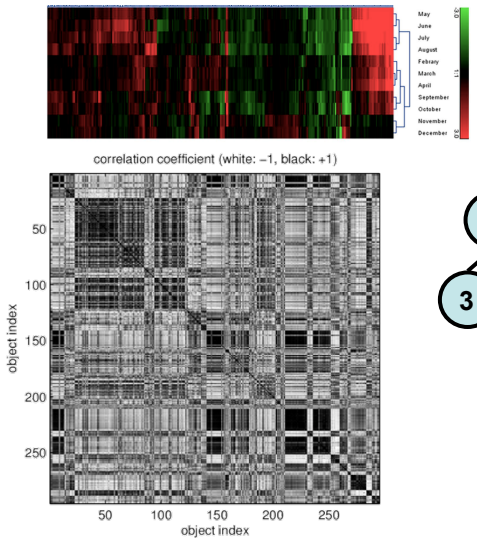
- **Analysis of algorithms**
  - ✔ **Runtime, memory consumption**
  - ✗ **Robustness, generalization!**

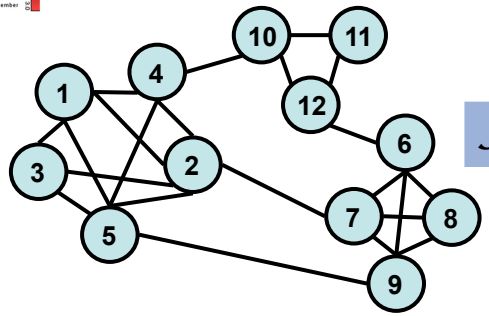
➢ **Learning algorithms „explore“ reality** and its complexity!

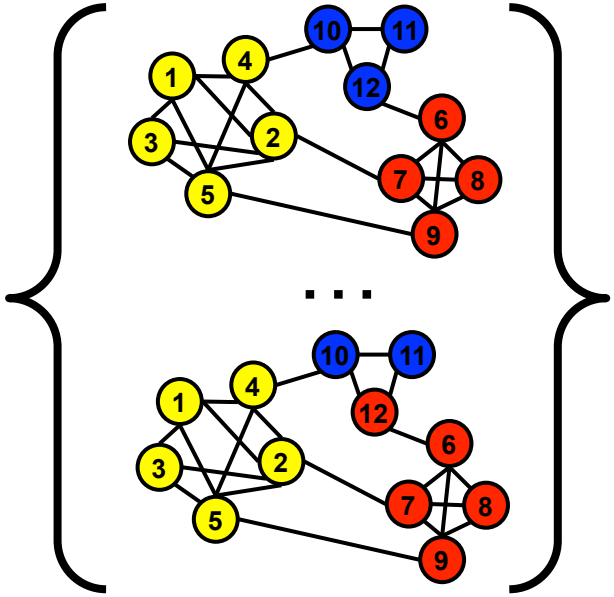
Data-driven algorithmics - what is the problem?

- **Problem**: random input implies random output

$$\underbrace{\text{input } \mathbf{X} \sim \mathbb{P}(\mathbf{X})}_{\text{given}} \Longrightarrow \underbrace{\mathcal{A}}_{\text{algorithm}} \Longrightarrow \underbrace{\text{output } c \sim \mathbb{P}(c|\mathbf{X})}_{\text{design}}$$







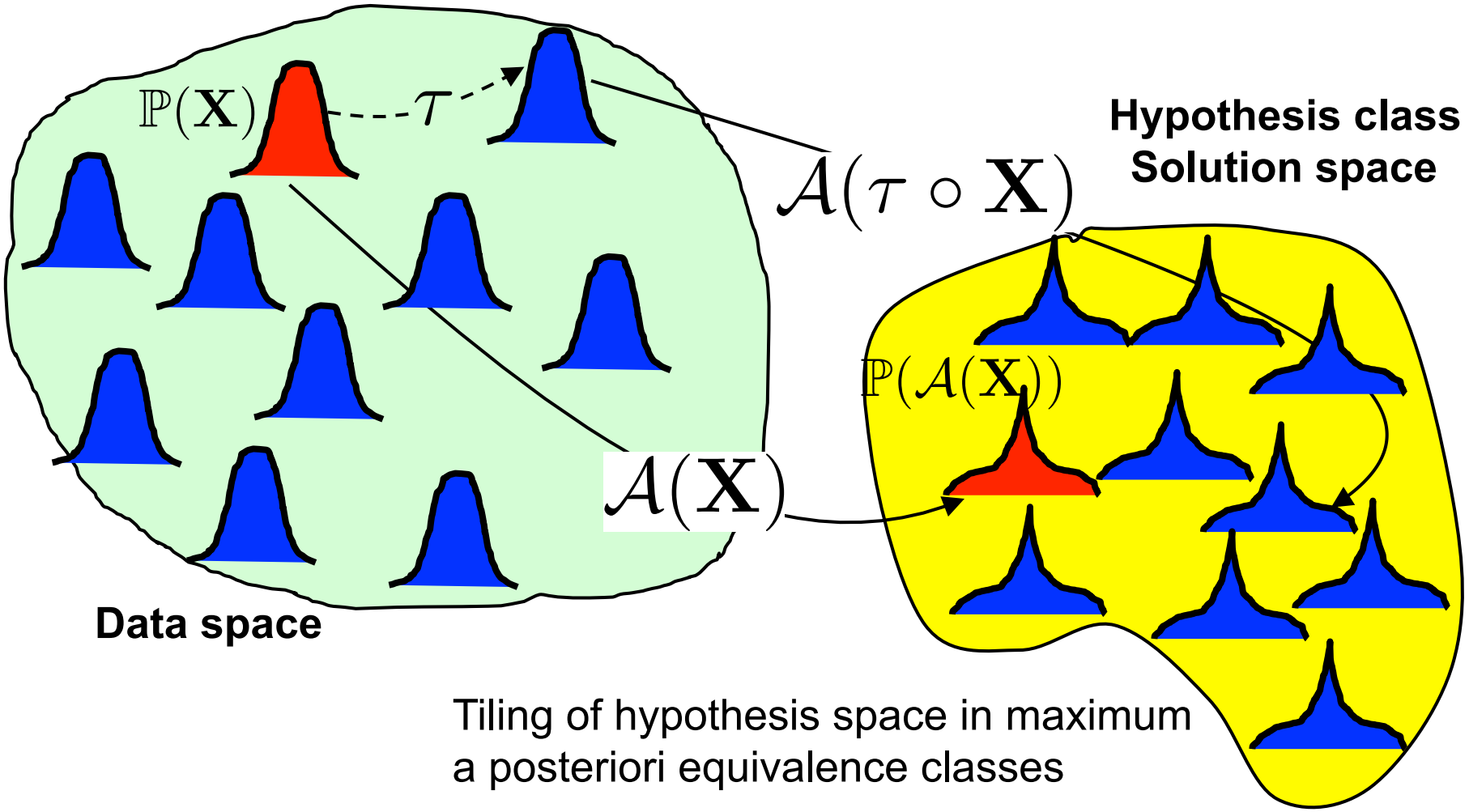
Thoughts on data-driven algorithmics

- **Randomness in data reduces resolution of solution space!** Otherwise, the solution space is too simple.
- **Resilient algorithms localize solutions** / hypotheses given noisy data with a (strong) **signal**.

➢ **Goal**: **rank algorithms** w.r.t. their ***sensitivity to signal*** and their ***robustness to noise***.

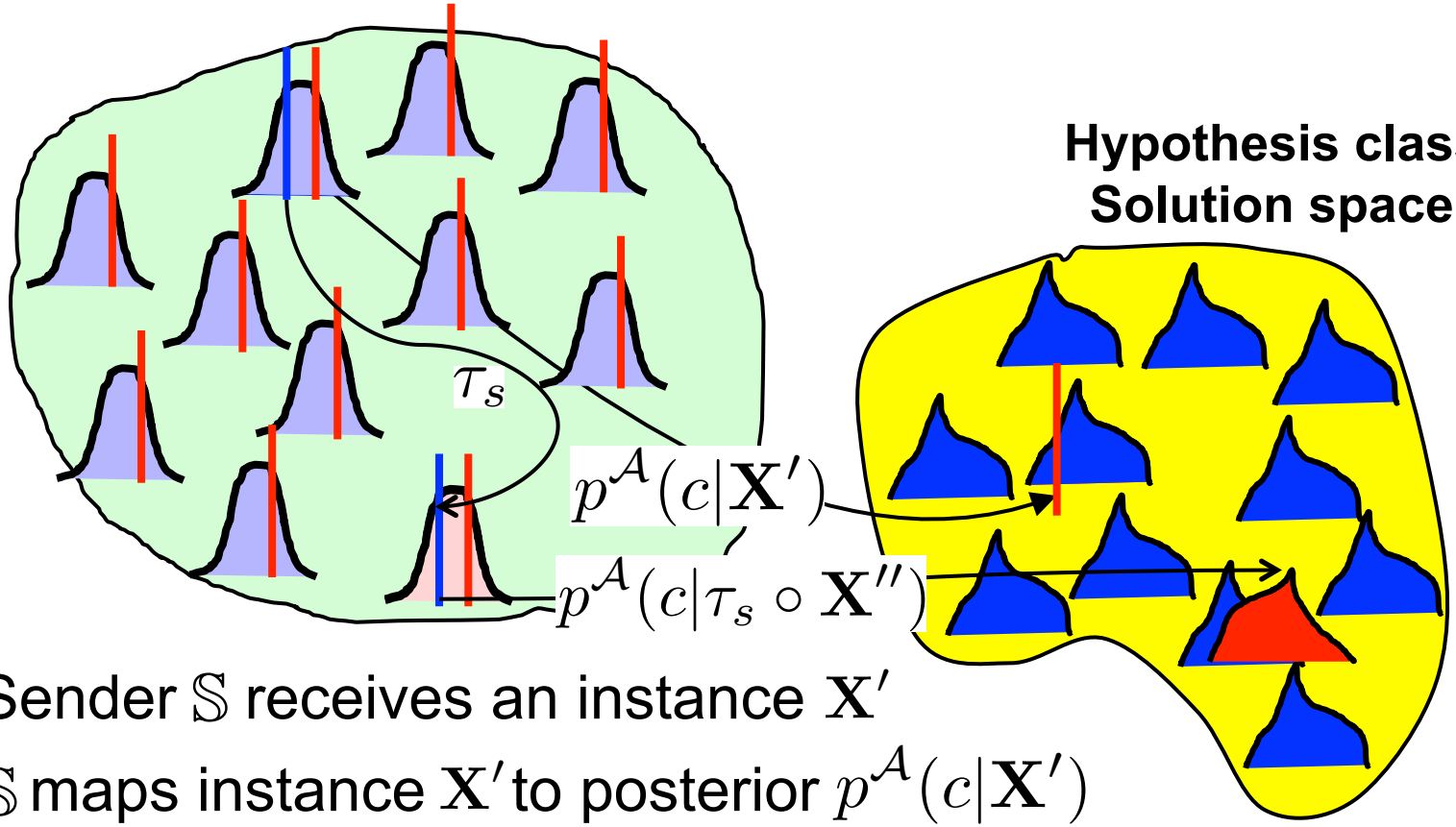
=> study $\left(\mathbb{P}(\mathbf{X}), \mathcal{A}(\mathbf{X})\right)$ => **Information Theory**

Noisy input quantizes a hypothesis class



Tiling of hypothesis space in maximum a posteriori equivalence classes

Information theory for algorithm validation



- Sender $\mathbb{S}$ receives an instance $\mathbf{X}'$
- $\mathbb{S}$ maps instance $\mathbf{X}'$ to posterior $p^{\mathcal{A}}(c|\mathbf{X}')$
- Oracle samples instance $\tilde{\mathbf{X}} = \tau_s \circ \mathbf{X}''$
- Receiver estimates transformation $\hat{\tau}$ by posterior agreement

Algorithms as distributions of solutions

- Let $\mathbf{X}$ denote data and $\mathbb{P}_t(c|\mathbf{X})$ the posterior at iteration t of algorithm $\mathcal{A}$

$$\text{algorithm} \quad \mathcal{A}(\mathbf{X}) \;=\; \langle \mathbb{P}_0(c|\mathbf{X}), \ldots, \mathbb{P}_T(c|\mathbf{X}) \rangle,$$

$$\textbf{init}\ \mathbb{P}_0(c|\mathbf{X}) \;=\; |\mathcal{C}|^{-1},$$

$$\mathbb{P}_t(c|\mathbf{X}) \;\text{ is contracting for } 0 < t \leq T,$$

$$\textbf{return}\ \mathbb{P}_T(c|\mathbf{X}) \;=\; \Delta_{c,c^{\perp}(\mathbf{X})}.$$

- Greedy or monotonically contractive algorithms

$$\mathcal{A}(\mathbf{X}) \;=\; \langle \mathbb{P}_0(c|\mathbf{X}) \gg \ldots \gg \mathbb{P}_{t^*}(c|\mathbf{X}) \gg \ldots \gg \mathbb{P}_T(c|\mathbf{X}) \rangle$$

Localizing solutions

- "**Posteriors**" for probable data $\mathbf{X}'$, $\mathbf{X}''$ should agree!

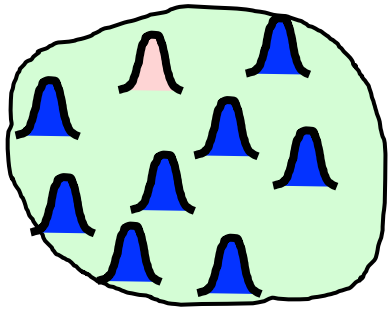
$$k^{\mathcal{A}}(\mathbf{X}', \mathbf{X}'') \;=\; \sum_{c \in \mathcal{C}} p^{\mathcal{A}}(c|\mathbf{X}')\; p^{\mathcal{A}}(c|\mathbf{X}'') \in [0,1]$$

A too broad or too narrow posterior $p^{\mathcal{A}}(.|\mathbf{X})$ yields a small kernel value $k^{\mathcal{A}}(\mathbf{X}', \mathbf{X}'')$! Optimize width of $p^{\mathcal{A}}(.|\mathbf{X})$

- Information theory provides the validation criterion

$$\mathbb{P}^* \in \arg\max_{\{\mathcal{A}\}} \max_{t} \mathbb{E}_{\mathbf{X}',\mathbf{X}''} \log\left(|\mathcal{C}|\, k_t^{\mathcal{A}}(\mathbf{X}', \mathbf{X}'')\right)$$

Generalization capacity as mutual information



$$\mathcal{I} = \mathbb{E}_{\mathbf{X}',\mathbf{X}''} \log \frac{p(\mathbf{X}',\mathbf{X}'')}{p(\mathbf{X}')p(\mathbf{X}'')}$$

$$\frac{p(\mathbf{X}',\mathbf{X}'')}{p(\mathbf{X}')p(\mathbf{X}'')} = \sum_{c\in\mathcal{C}}\sum_{\mathcal{T}} \frac{p(\mathbf{X}',\mathbf{X}''|\mathcal{T},c)\; p(\mathcal{T},c)}{p(\mathbf{X}')p(\mathbf{X}'')}$$

$$= \sum_{c\in\mathcal{C}}\sum_{\mathcal{T}} \frac{p(\mathbf{X}'|\mathcal{T},c)\; p(\mathbf{X}''|\mathcal{T},c)\; p(\mathcal{T},c)}{p(\mathbf{X}')p(\mathbf{X}'')} \qquad \mathbf{X}',\mathbf{X}''|\tau \text{ i.i.d}$$

$$= \sum_{c\in\mathcal{C}} \frac{1}{p(c|\tau_s)}\; p(c|\tau_s\circ\mathbf{X}')\; p(c|\tau_s\circ\mathbf{X}'')$$

$$\leq |\mathcal{C}| \sum_{c\in\mathcal{C}} p(c|\tau_s\circ\mathbf{X}')\; p(c|\tau_s\circ\mathbf{X}'') =: |\mathcal{C}|\, k(\mathbf{X}',\mathbf{X}'')$$

Learning an algorithm: open challenge!

- **Given a set of algorithm**

$$\{\mathcal{A}^{(\alpha)}(\mathbf{X}) \;=\; \langle \mathbb{P}_0^{(\alpha)}(c|\mathbf{X}), \ldots, \mathbb{P}_{t^*}^{(\alpha)}(c|\mathbf{X})\rangle\}$$

- **Select posterior** $\mathcal{A}^{(\alpha)}(\mathbf{X})$ s.t. generalization capacity is maximized

$$\mathbb{P}^* \in \arg\max_{\{\mathcal{A}\}} \max_t \mathbb{E}_{\mathbf{X}',\mathbf{X}''} \log\left(|\mathcal{C}|\, k_t^{\mathcal{A}}(\mathbf{X}',\mathbf{X}'')\right)$$

- **Problem**: We cannot evaluate $\mathbb{E}_{\mathbf{X}',\mathbf{X}''} \log \ldots$ since $\mathbb{P}(\mathbf{X}',\mathbf{X}'')$ is unknown!

Consistent learning of an algorithm

- **Learning** requires that an empirical estimate of the capacity should be close to its expectation

$$\mathbb{E}_{\mathbf{X}',\mathbf{X}''} \log\left(|\mathcal{C}|\, \hat{k}^{\mathcal{A}}(\mathbf{X}',\mathbf{X}'')\right) \geq$$

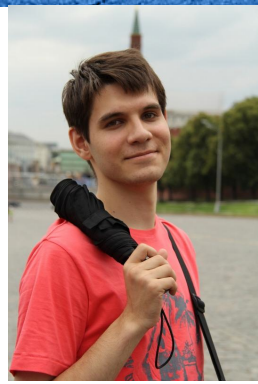
$$\frac{1}{L}\sum_{l\leq L} \log\left(|\mathcal{C}|\, \hat{k}^{\mathcal{A}}(\mathbf{X}'_l,\mathbf{X}''_l)\right) - \text{penalty}$$

$$\text{with optimal empirical posterior } \hat{\mathbb{P}}^*$$

Example: Robust Spanning Trees

- **Given** are graphs with stochastic edge weights.
- **Question**: How robust are MST algorithms in this stochastic setting?
- Measure **robustness of algorithms** like Prim's, Kruskal's algorithm or the Reverse-Delete algorithm.
- Determine a **stable set of approximate spanning trees** by early stopping of an MST algorithm.

Learning to span a graph


Alexey Gronskiy

Consider **Minimum Spanning Tree** algorithms

- **Prim**'s "Growing tree" strategy: add minimal edge to tree.
- **Kruskal**'s "Joining trees" strategy: add minimal edge connecting two trees in a forest.
- **Reverse-Delete**: "Reducing graph" strategy: delete maximal edge without destroying connectivity.

MST Algorithm as a sequence of approximate spanning tree sets

- Let $\mathbf{X}$ denote data and $\mathbb{P}_t(c|\mathbf{X}) = \frac{w_t(c,\mathbf{X})}{\sum_c w_t(c,\mathbf{X})}$ the posterior of algorithm $\mathcal{A}$ at iteration t

$$w_t(c,\mathbf{X}) = \begin{cases} 1 & \text{if partial solution at } t \text{ admits } c, \\ 0 & \text{otherwise.} \end{cases}$$

- algorithm $\mathcal{A}(\mathbf{X}) = \langle \mathbb{P}_0(c|\mathbf{X}), \ldots, \mathbb{P}_T(c|\mathbf{X}) \rangle,$

$$\begin{aligned} \textbf{init } & \mathbb{P}_0(c|\mathbf{X}) = |\mathcal{C}|^{-1}, \\ \textbf{for } & 1 \le t \le T \text{ calculate } \mathbb{P}_t(c|\mathbf{X}), \\ \textbf{return } & \mathbb{P}_{t^*}(c|\mathbf{X}) \in \arg\max_t \mathbb{E}_{\mathbf{X}',\mathbf{X}''} \log\left(|\mathcal{C}|\, k_t^{\mathcal{A}}(\mathbf{X}',\mathbf{X}'')\right) \end{aligned}$$

Cardinality of AST sets

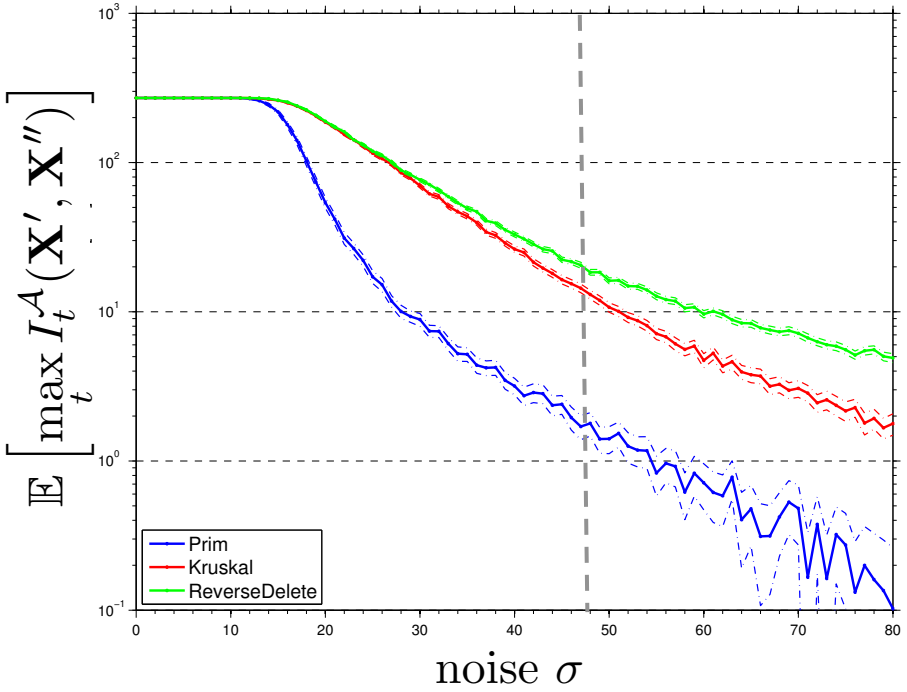
- **Matrix tree theorem**: the number of spanning trees is equal to (any) cofactor of the matrix

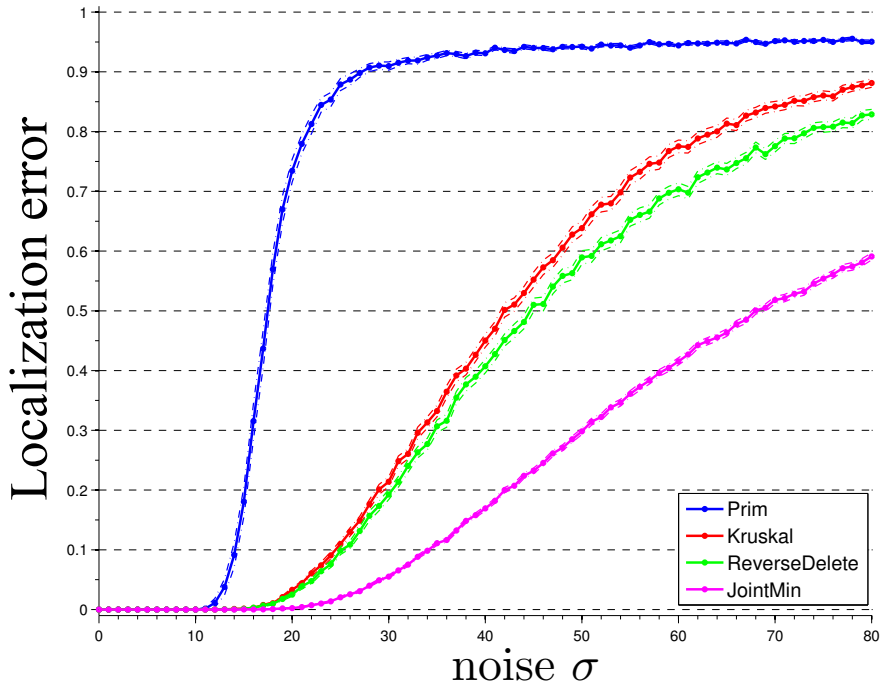
$$L = M^X_{\mathrm{deg}} - M^X_{\mathrm{adj}}$$

- Calculate the cofactor of L after t steps for the effective $X(t)$ where selected edges are contracted (Prim, Kruskal) or removed (Reverse-Delete).

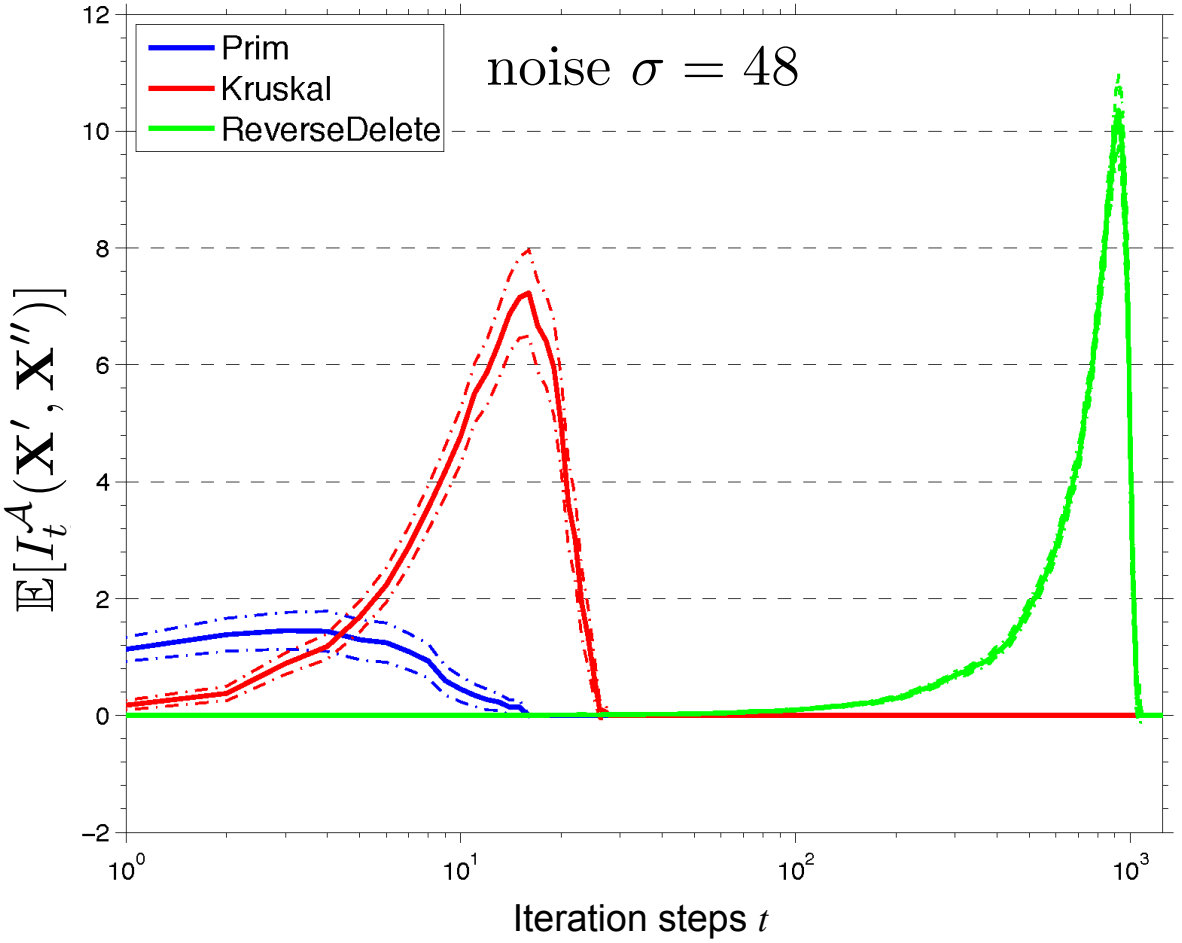
Algorithmic informativeness

- Hierarchical graph generation:
  - ground truth graph: 50 vertices, i.i.d. normal weights $\mathcal{N}(100, 100)$
  - Additive Gaussian noise $\mathcal{N}(0, \sigma^2) \Rightarrow \mathbf{X}', \mathbf{X}''$



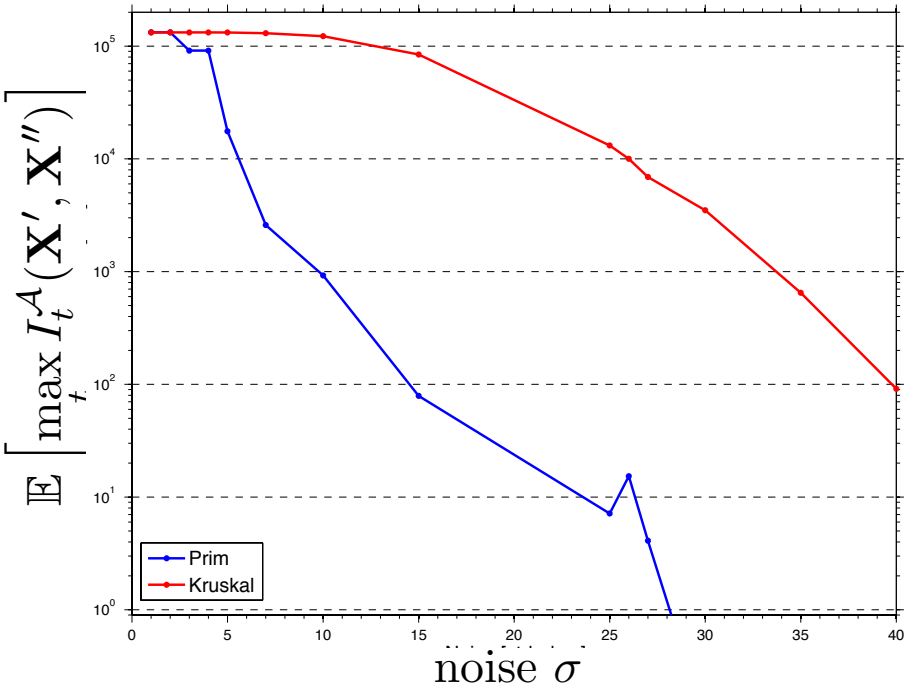


Dynamics of algorithmic informativeness

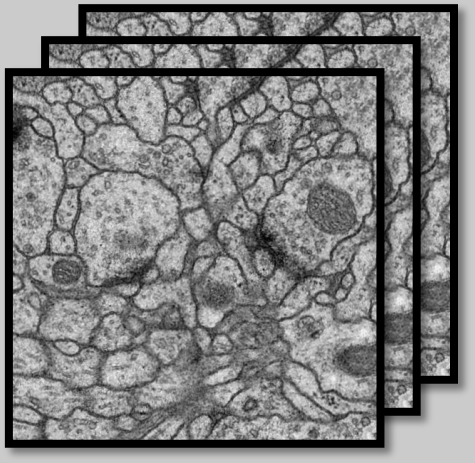


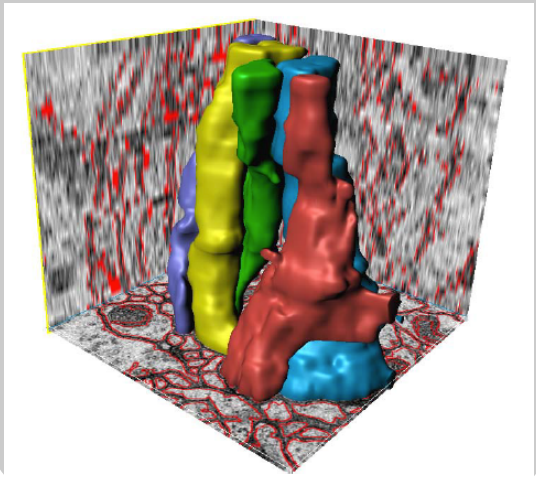
Informativeness of ASTs with 10^4 vertices

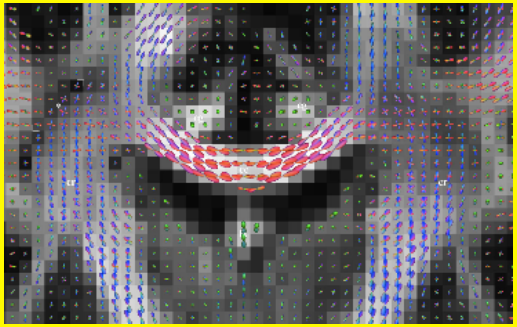
- Hierarchical graph generation:
  - ground truth graph: 10^4 vertices, i.i.d. normal weights $\mathcal{N}(100, 100)$
  - Additive Gaussian noise $\mathcal{N}(0, \sigma^2)$

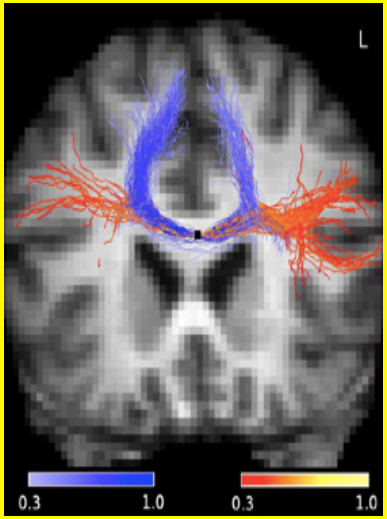


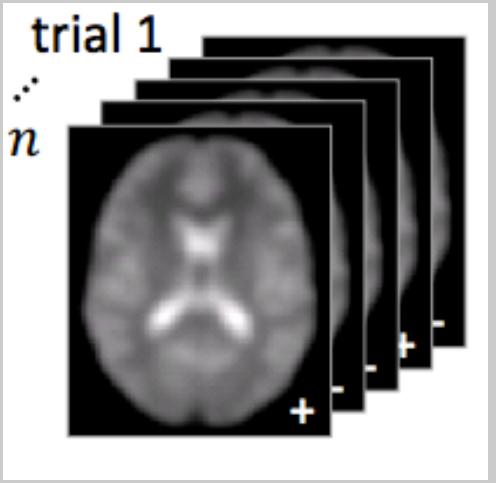
Big Data in Neuroscience

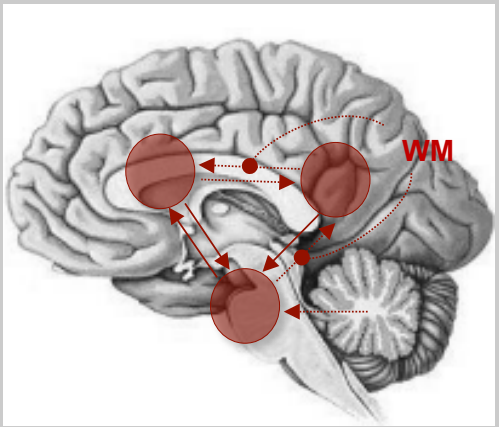












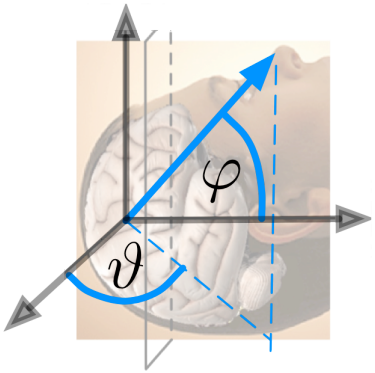
Diffusion weighted tensor imaging: pipeline

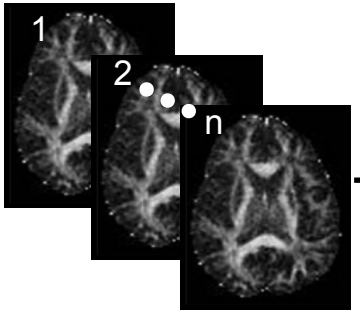

Nico Gorbach

scanner parameters

noise modeling

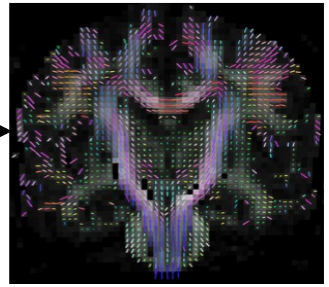
local fiber orientation density function



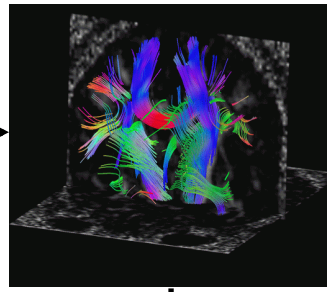


Diffusion weighted images (1-10GB)

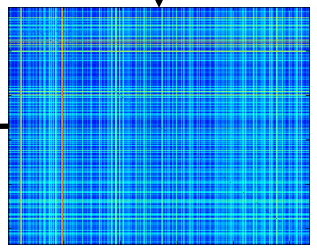
diffusion tensor construction



fiber tracking

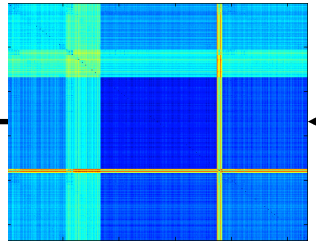


# of samples

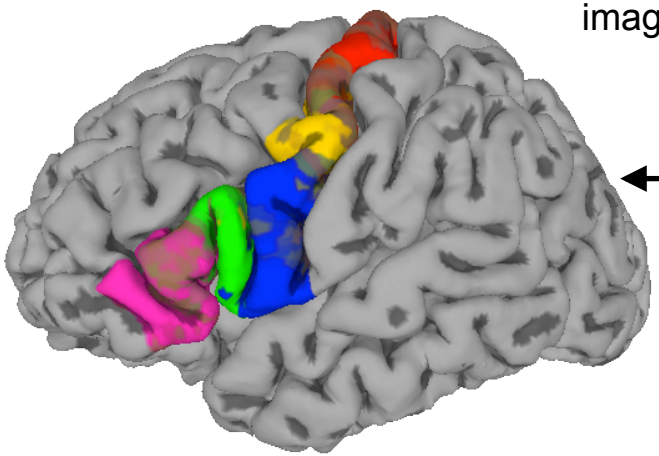


connectivity matrix (3000x1M)

dissimilarity measure + **clustering method**

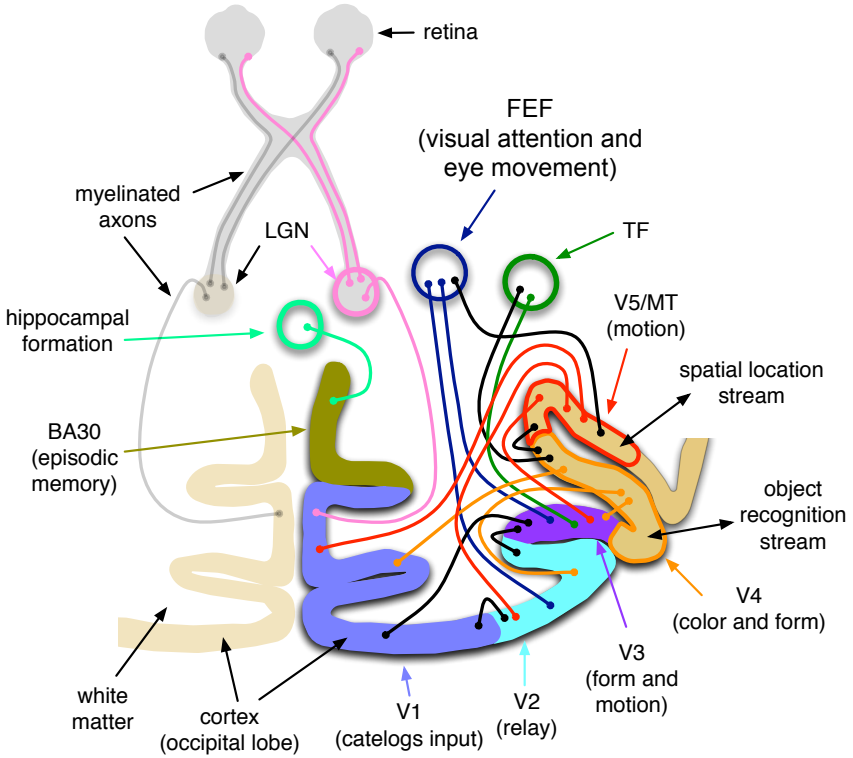


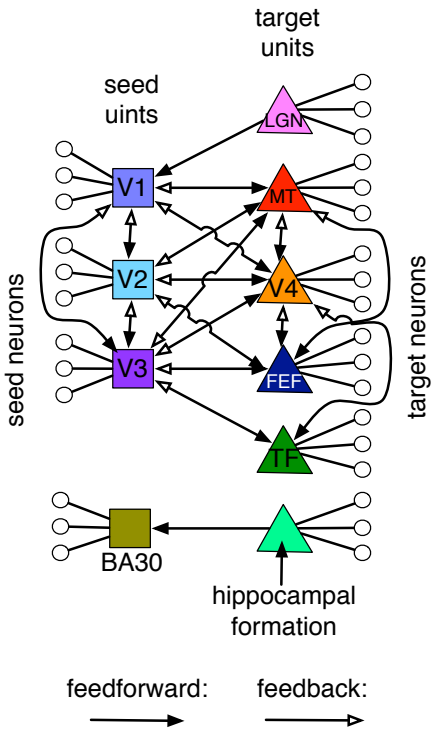
minimalistic structure projection onto dissimilarity matrix



parcellation

Systems Neuroscience



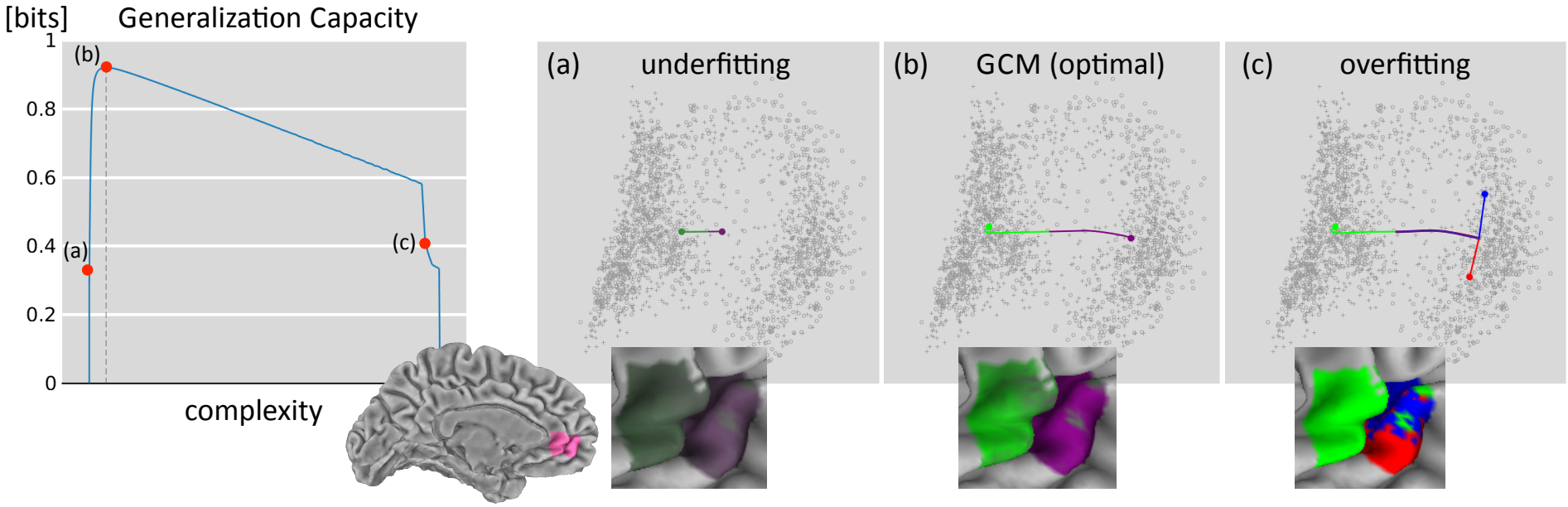


Subset of the visual system in the macaque monkey.
Target connections are limited for illustration purposes.

- The brain is considered as an ensemble of functionally specialized units coupled together in a modulatory fashion (Friston, 2002).

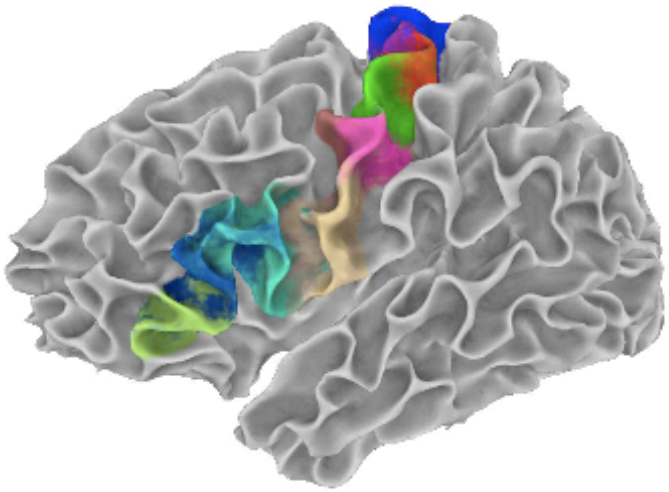
Under- and overfitting in parcellation

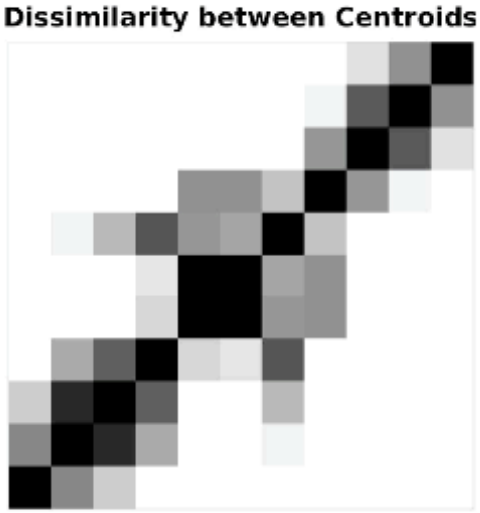
- Connectivity of two brain regions is analyzed
- Generalization capacity maximizer (GCM) outperforms empirical risk minimizer

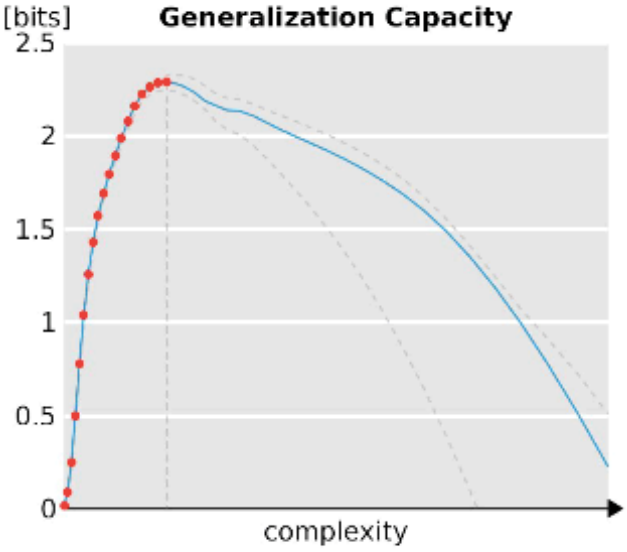


Dynamics of cortex parcellation

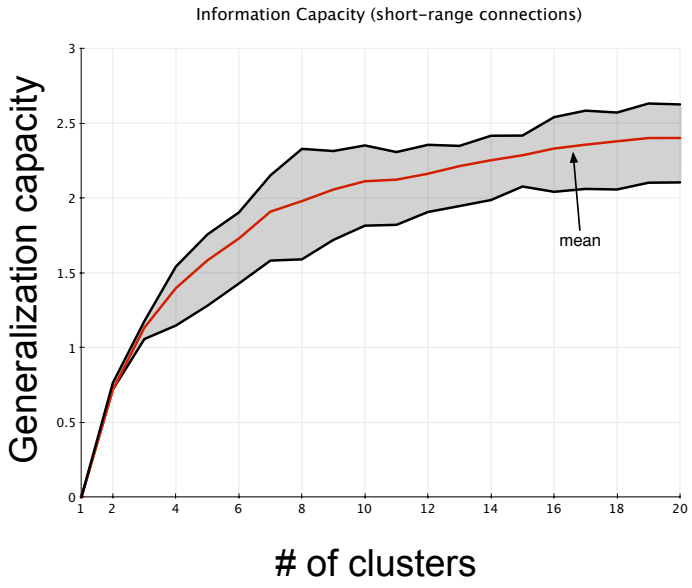
- Start at low resolution
- Estimate parcellations with higher resolution
- Stop at maximal generalization capacity

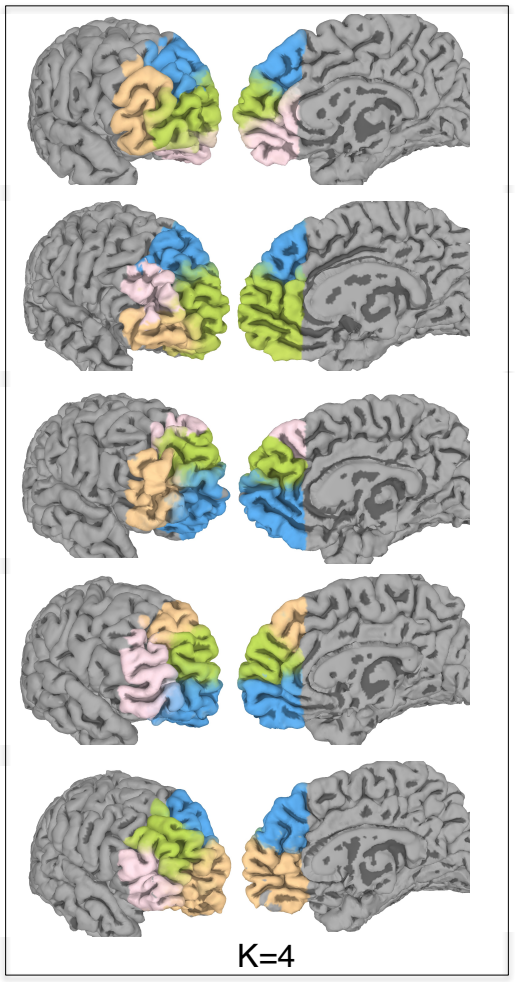


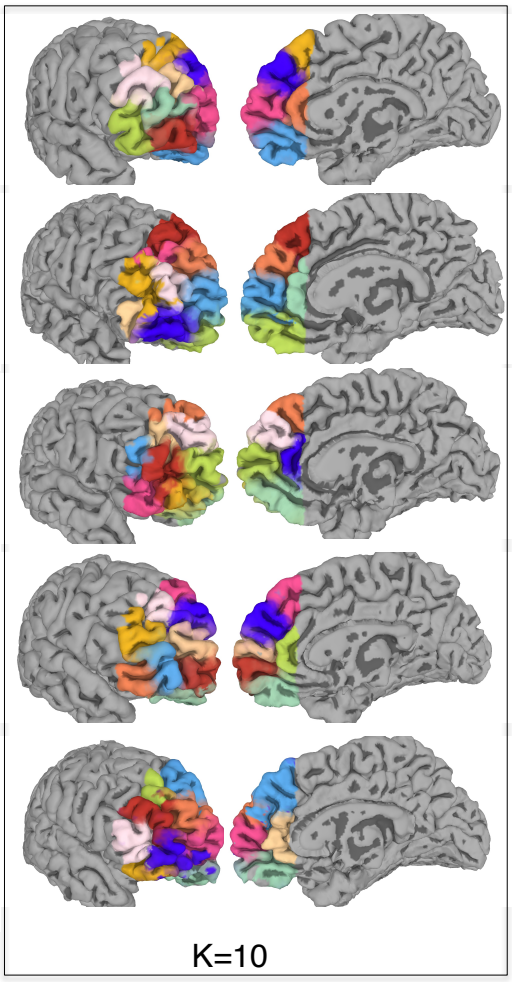




Comparative Cortex Parcellation of Different Patients





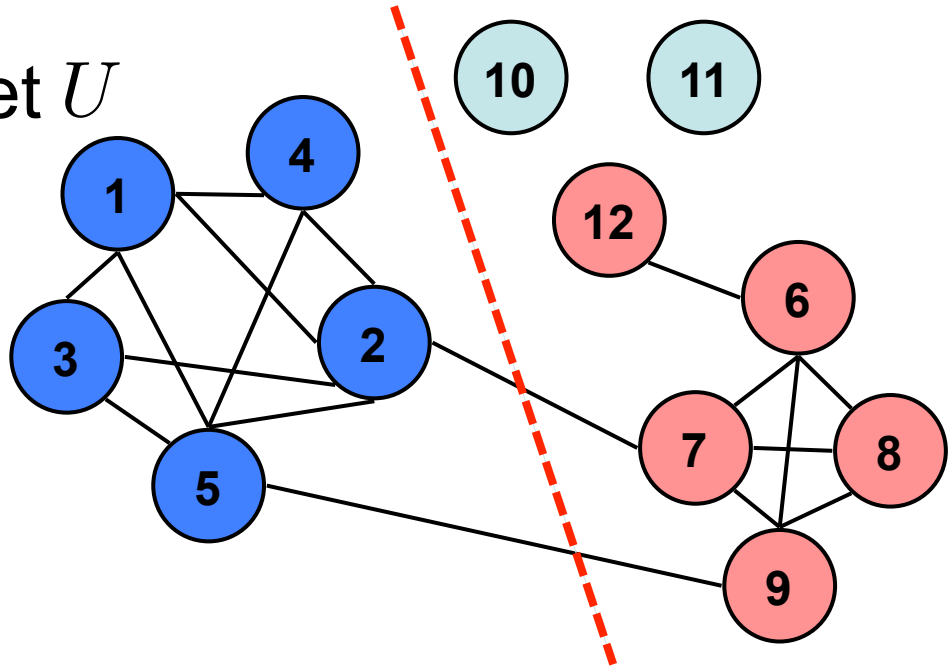


Sparse Minimum Bisection Problem

Given is a graph $G = (V, E, W)$

1. Select a subset of nodes U
2. Find the minimum bisection for this subset U

If $|U| = \Theta(n^{2/7})$ then the problem is conjectured to be NP-hard.



Random energy model (Derrida 1981)

- Given are $M = 2^n$ states or solutions c
- Cost of solution c is normally distributed

$$R(c, \mathbf{X}) \sim \mathcal{N}(0, \sqrt{n})$$

- Remarks
  - Model is defined by M random variables
  - REM is maximally disordered and without any structure for searching.
  - Derrida introduced REM to study approximation schemes for partition functions of disordered systems.

Relation of sMBP to REM

- Assume sparsity $n^s = \Theta(n^{2/7})$

Th.: sMBP is upper bounded by „REM“ for rescaled temperature (JMB, Gronskiy, Szpankowski, AofA ‘12)

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, \mathbf{X})] + \hat{\beta}\mu\sqrt{N \log m}}{\log m} \leq \begin{cases} 1 + \frac{\hat{\beta}^2\sigma^2}{2}, & \hat{\beta}\sigma < \sqrt{2}, \\ \hat{\beta}\sigma\sqrt{2}, & \hat{\beta}\sigma \geq \sqrt{2}. \end{cases}$$

Th.: sMBP is lower bounded by „REM“ (... ANALCO’17)

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, \mathbf{X})] + \hat{\beta}\mu\sqrt{N \log m}}{\log m} \geq \begin{cases} 1 + \frac{\hat{\beta}^2\sigma^2}{2}, & \hat{\beta}\sigma < \sqrt{2}, \\ \hat{\beta}\sigma\sqrt{2}, & \hat{\beta}\sigma \geq \sqrt{2}. \end{cases}$$

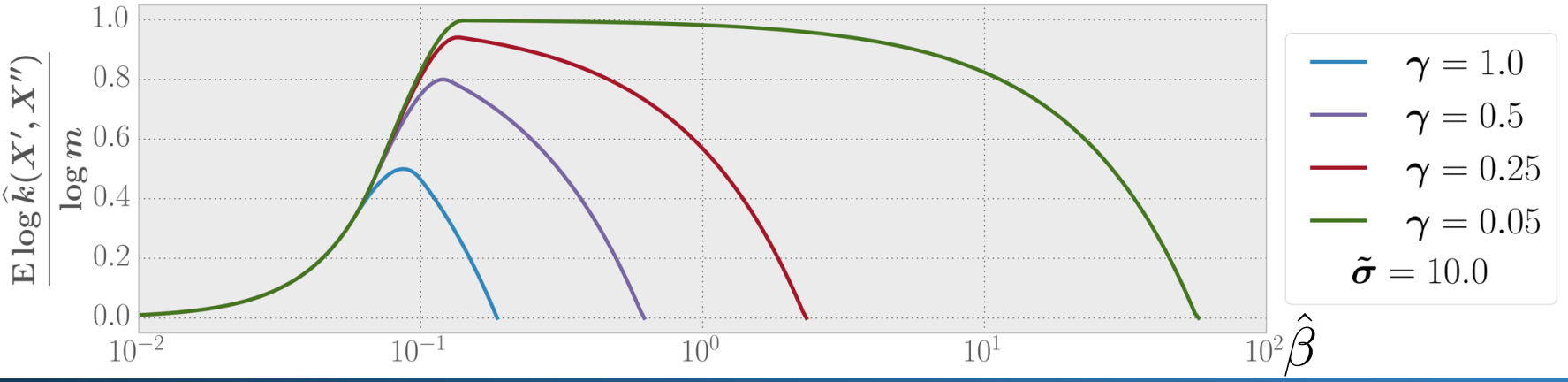
Generalization capacity of sMBP for noise perturbed random graphs

- Choose two random graphs with normal weights

$$\mathbf{X}' = \mathbf{X} + \delta\mathbf{X}', \;\; \mathbf{X}'' = \mathbf{X} + \delta\mathbf{X}'' \qquad \gamma = \sigma/\tilde{\sigma}$$

- Generalization capacity has 2 phase transitions

$$\lim_{n\to\infty} \frac{\mathbb{E}_{X,\delta X',\delta X''} \log\left(|\mathcal{C}|\ \hat{k}_\beta(X', X'')\right)}{\log m} = \eta(\hat{\beta}),$$



Discussion of sMBP – REM relation

- REM has no information for searching, i.e., we have to evaluate $M = 2^n$ cost values.
- Random sMBP is statistically equivalent to REM.
- Question: How can we search for a minimal solution of a random sMBP?
- The equivalence sMBP – REM relates statistical complexity with computational complexity!

ML Challenges for HPC

- We need **efficient sampling schemes** for posteriors.
- Algorithms have to efficiently estimate their informativeness. **Adaptive regularization**!
- **Statistical resilience** and **computational efficiency** have to match to maximize prediction performance.

✓ Avoid computationally expensive overfitting.

Conclusion

- **Algorithms are models of posteriors!**
- **Uncertainty** in a given input space reduces resolution of output space (hypotheses)
  => quantization of mathematical structures

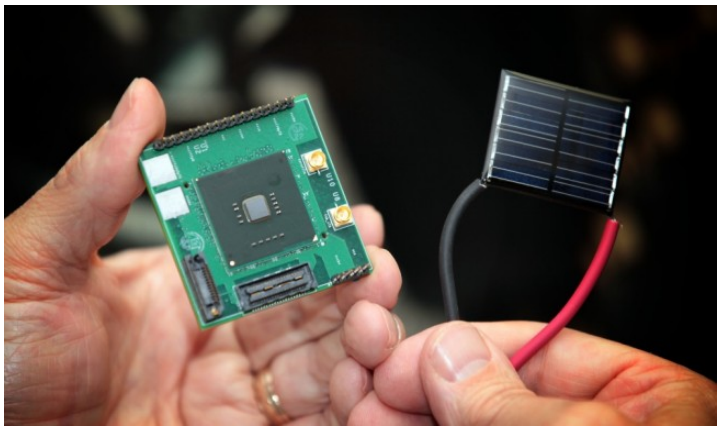
$\Rightarrow$ **Information theory yields** a **generalization capacity** for **algorithms given an input distribution!**

$\Rightarrow$ **structure specific information** in data.

Relate **statistical** to **computational complexity**!

Low-Energy Architecture Trends

- Novel low-power architectures operate **near transistor threshold voltage** (NTV)
  - e.g., Intel Claremont
  - 1.5 mW @10 MHz (x86)
- NTV promises 10x more energy efficiency at 10x more parallelism!



source: Intel

  - 10^5 times more soft errors (bits flip stochastically)
  - Hard to correct in hardware → expose to programmer?

@ Thorsten Höfler