AN INFORMATION-THEORETIC OPTIMALITY PRINCIPLE FOR THE FORMATION OF ABSTRACTIONS

Tim Genewein[1,2,3], **Daniel A. Braun**[1,2]

[1]Max Planck Institute for Intelligent Systems
[2]Max Planck Institute for Biological Cybernetics
[3]Graduate Training Center for Neuroscience
Tübingen, Germany

MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

Abstractions and hierarchies

- Why are abstractions important?
  - Separation of structure from noise (instance variation)
  - Fast information processing (only relevant information)
  - Extraction of transferrable knowledge

- What role do hierarchies play?
  - Invariants on multiple different scales (temporal, spatial, …)
  - Different levels of abstraction – leads to hierarchical organization
  - Often hierarchical models are "handcrafted" or formed through heuristics
  - "Self-organization" of hierarchies derived from first principles?

Inference and decision-making

- Inference and decision-making with information processing limits
- Belief/policy is modeled as a probability distribution

- For now: decision-making scenario
- Agent emits an action α conditioned on an observation ω
- Tasks are formalized via a utility function $U(\alpha, \omega)$
- Agent has a default policy $p_0(\alpha)$ that is observation-independent

- Goal:
- Find a posterior $p(\alpha|\omega)$ that maximizes the gain in expected utility while minimizing the transformation cost from $p_0(\alpha)$ to $p(\alpha|\omega)$

Thermodynamic Model for DM

- Find a posterior $p(\alpha|\omega)$ that maximizes the gain in expected utility while minimizing the transformation cost from $p_0(\alpha)$ to $p(\alpha|\omega)$

$$\arg\max_{p(\alpha|\omega)} \mathbf{E}_{p(\alpha|\omega)}[U(\alpha,\omega)] - \frac{1}{\beta} D_{\mathrm{KL}}(p(\alpha|\omega)||p_0(\alpha))$$

- Variational problem has very similar mathematical form as a *free-energy difference* minimization

- Closed-form solution:

$$p(\alpha|\omega) = \frac{1}{Z} p_0(\alpha) e^{\beta U(\alpha,\omega)}$$

Temperature as rationality-parameter

$$\underset{p(\alpha|\omega)}{\arg\max}\ \mathbf{E}_{p(\alpha|\omega)}[U(\alpha,\omega)] - \frac{1}{\beta} D_{\mathrm{KL}}(p(\alpha|\omega)||p_0(\alpha))$$

$$p(\alpha|\omega) = \frac{1}{Z} p_0(\alpha) e^{\beta U(\alpha,\omega)}$$

- Limits:
  - Fully rational actor: $\beta \to \infty$
  - Fully bounded actor: $\beta \to 0$

- Normative framework for changing from **prior** belief/behavior to **posterior** belief/behavior with information processing cost
  - Bayes rule can be recovered as a special case

Rate Distortion for Decision Making

- Extend free energy model by taking the average over observations and optimizing over the prior as well:

$$\underset{p_0(\alpha)}{\arg\max} \sum_{\omega} \mathrm{p}(\omega) \left[\underset{p(\alpha|\omega)}{\arg\max}\, \mathbf{E}_{p(\alpha|\omega)}[U(\alpha,\omega)] - \frac{1}{\beta} D_{\mathrm{KL}}(p(\alpha|\omega)||p_0(\alpha)) \right]$$

- … which can be rewritten:

$$\underset{p(\alpha|\omega)}{\arg\max}\, \mathbf{E}_{p(\alpha,\omega)}[U(\alpha,\omega)] - \frac{1}{\beta} I(\alpha;\omega)$$

- Trade-off: high expected utility and low mutual information
- Rate distortion – a framework for lossy compression
  - Duality between abstraction and lossy compression
  - Channel from observations to actions with limited capacity

Temperature as rationality-parameter

- Well known (self-consistent) solution:

$$p(\alpha|\omega) = \frac{1}{Z} p(\alpha) e^{\beta U(\alpha,\omega)}$$

$$p(\alpha) = \sum_{\omega} p(\omega) p(\alpha|\omega)$$

- $\beta > 0$: favor actions that yield "good" utility for many observations

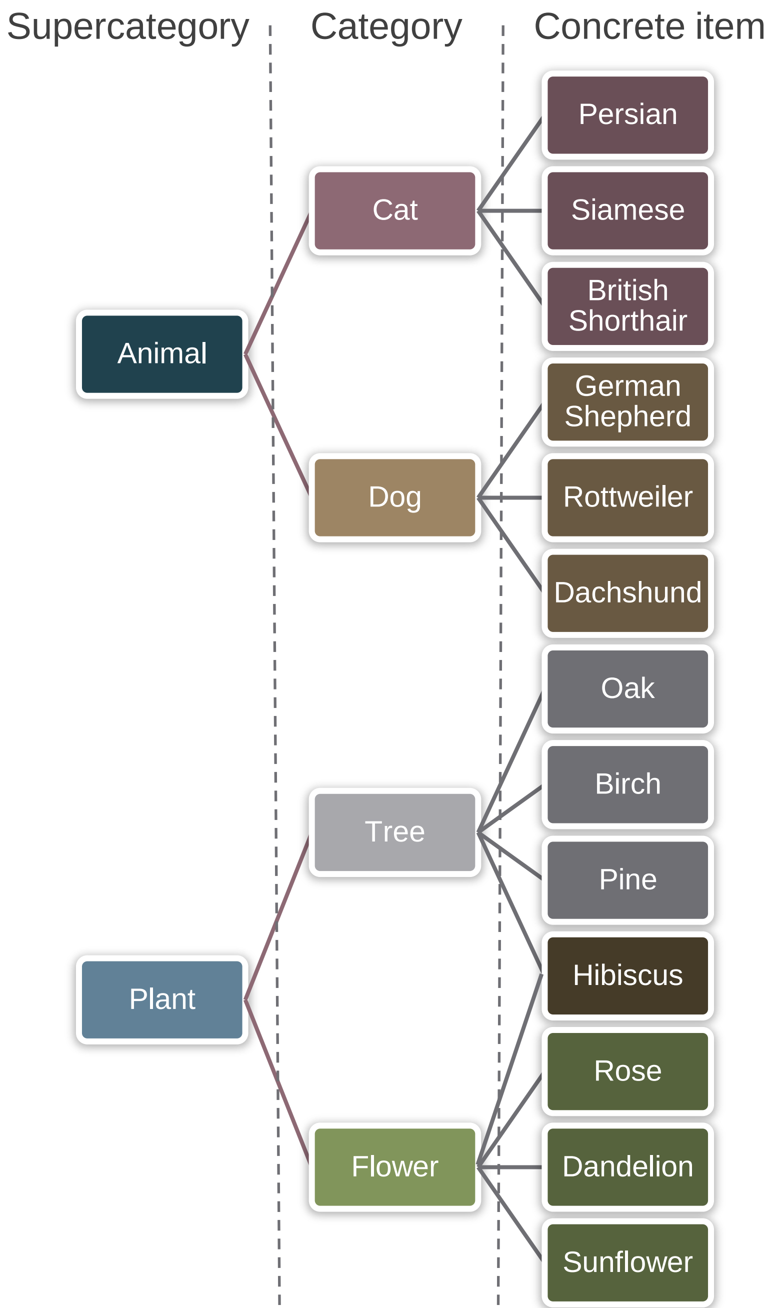
- Temperature governs the granularity of the abstraction
  - --> Example

Toy Example

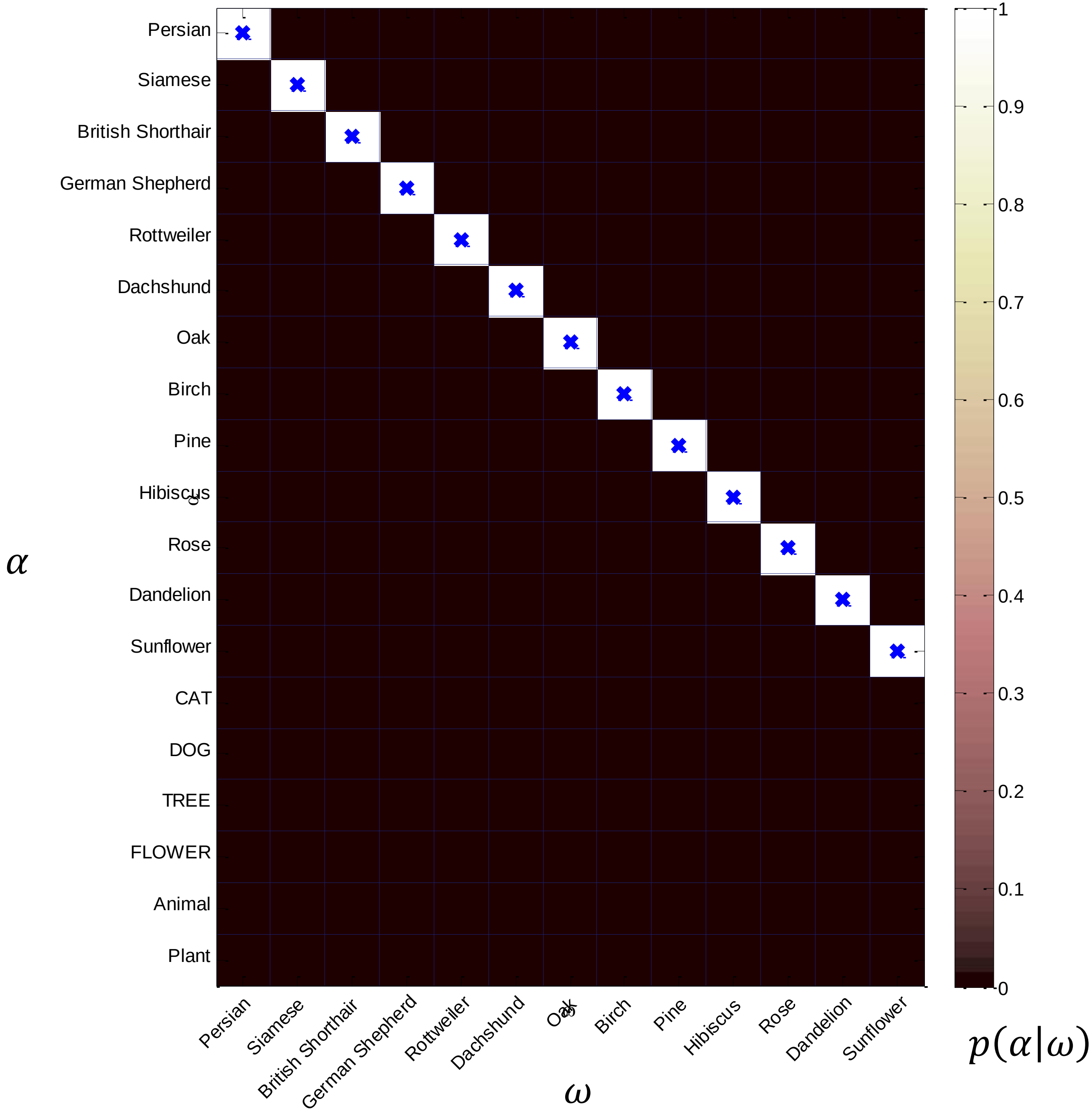
- Simple taxonomy with three layers of abstraction

- Sensory state $\omega \in \{concrete\ items\}$
- Action $\alpha \in \{concrete\ items, categories, supercategories\}$

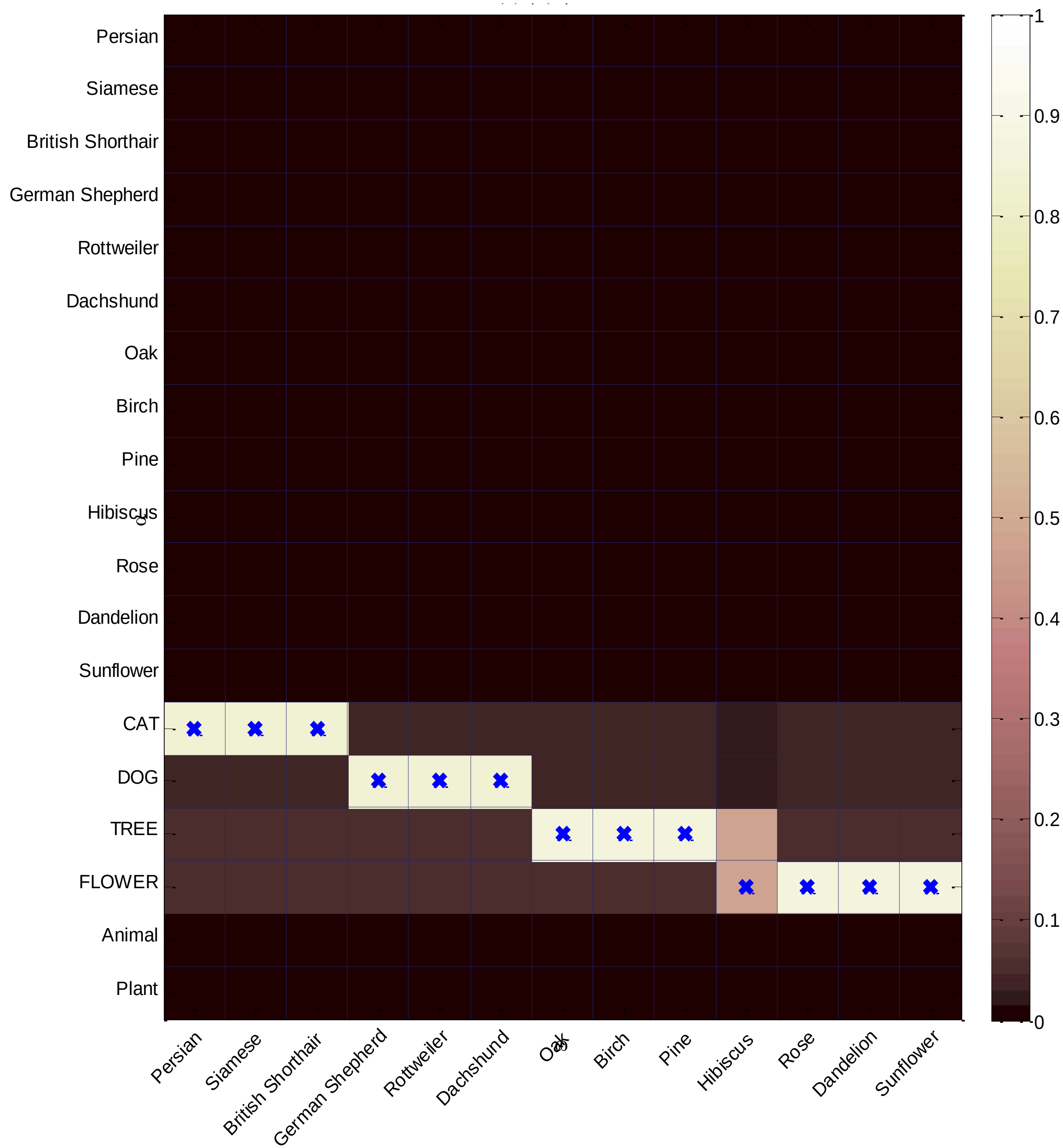
- Rewards/Utilities:
  - 3€ if concrete item correct
  - 2.2€ if category correct
  - 1.6€ if supercategory correct



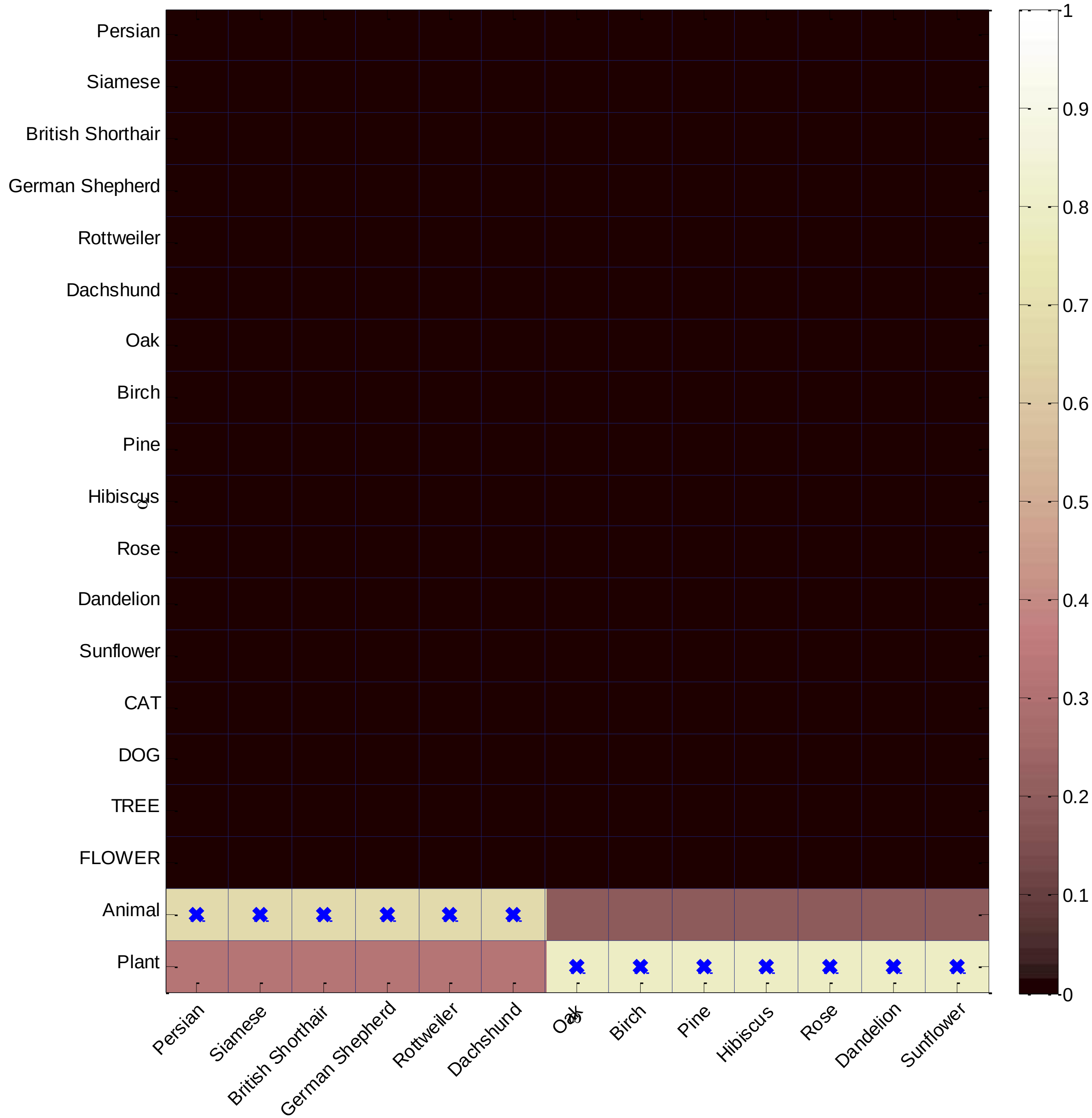
β	10	$[\text{bits}/_{€}]$
I	3.7	[bits]
$\mathbf{E}[U]$	3	[€]



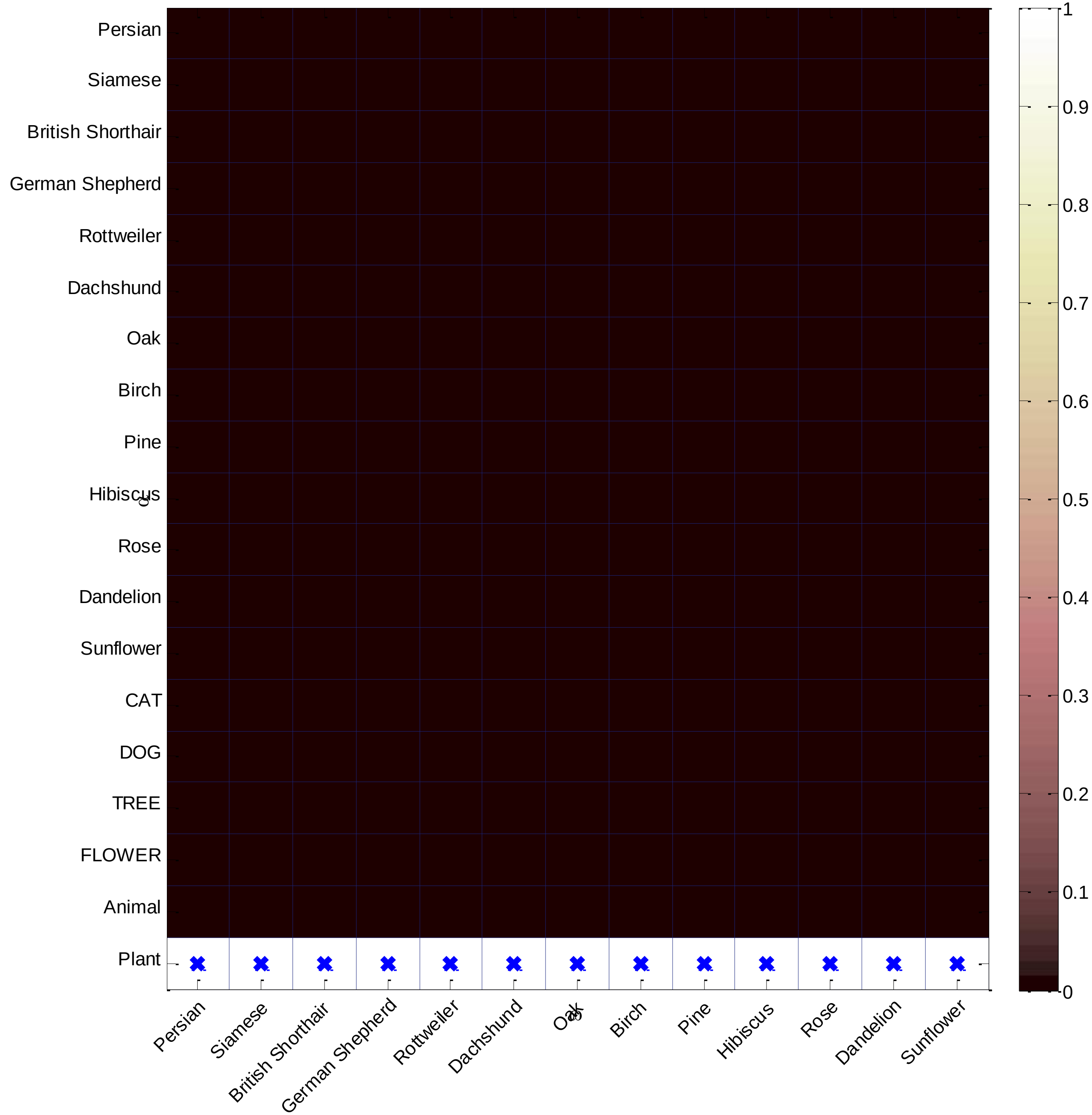
β	1.33	$\left[{}^{\text{bits}}/_{€}\right]$
I	1.7	[bits]
$\mathbf{E}[U]$	2.2	[€]



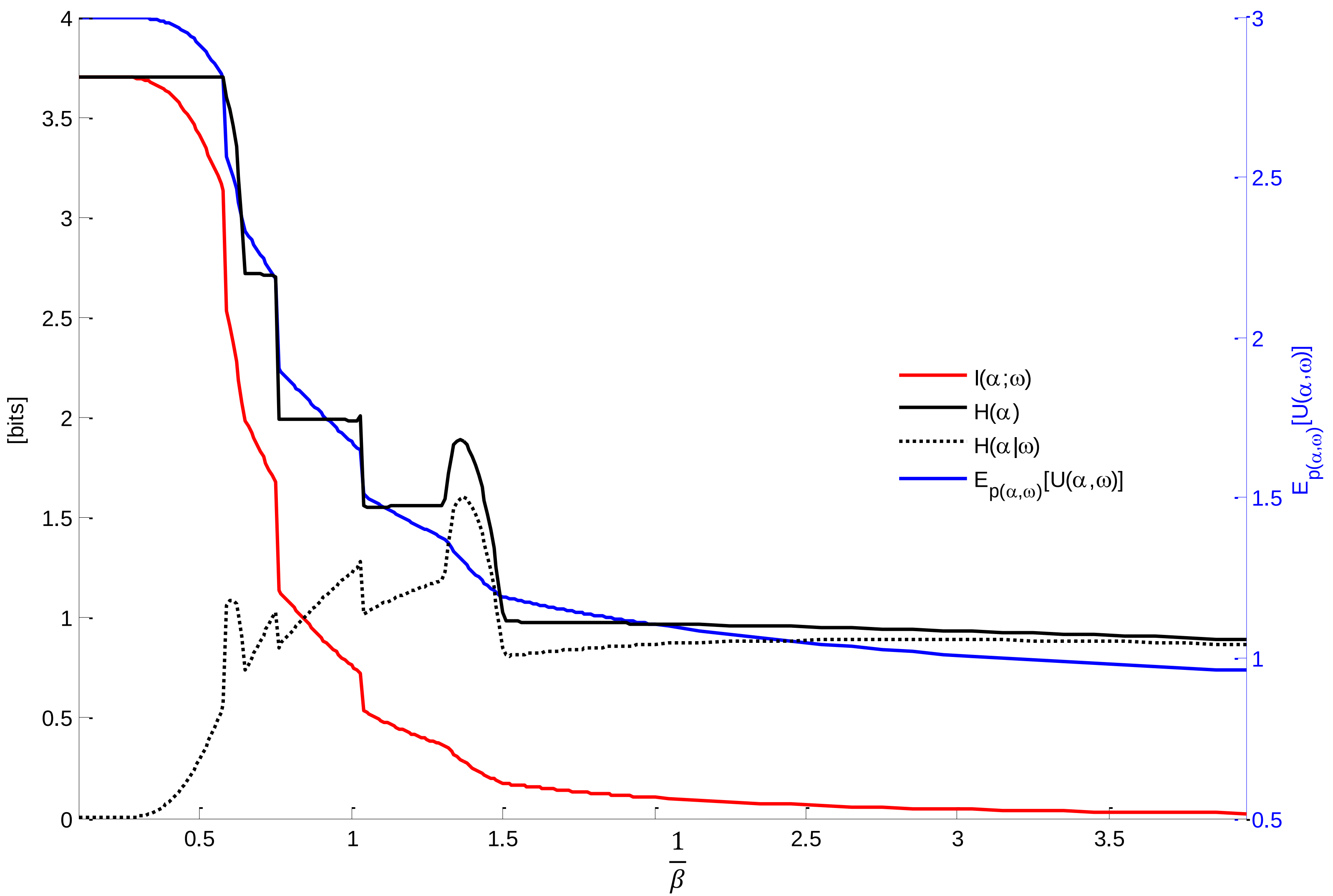
β	0.67	$\left[{}^{\text{bits}}/_{€}\right]$
I	0.2	[bits]
$\mathbf{E}[U]$	1.2	[€]

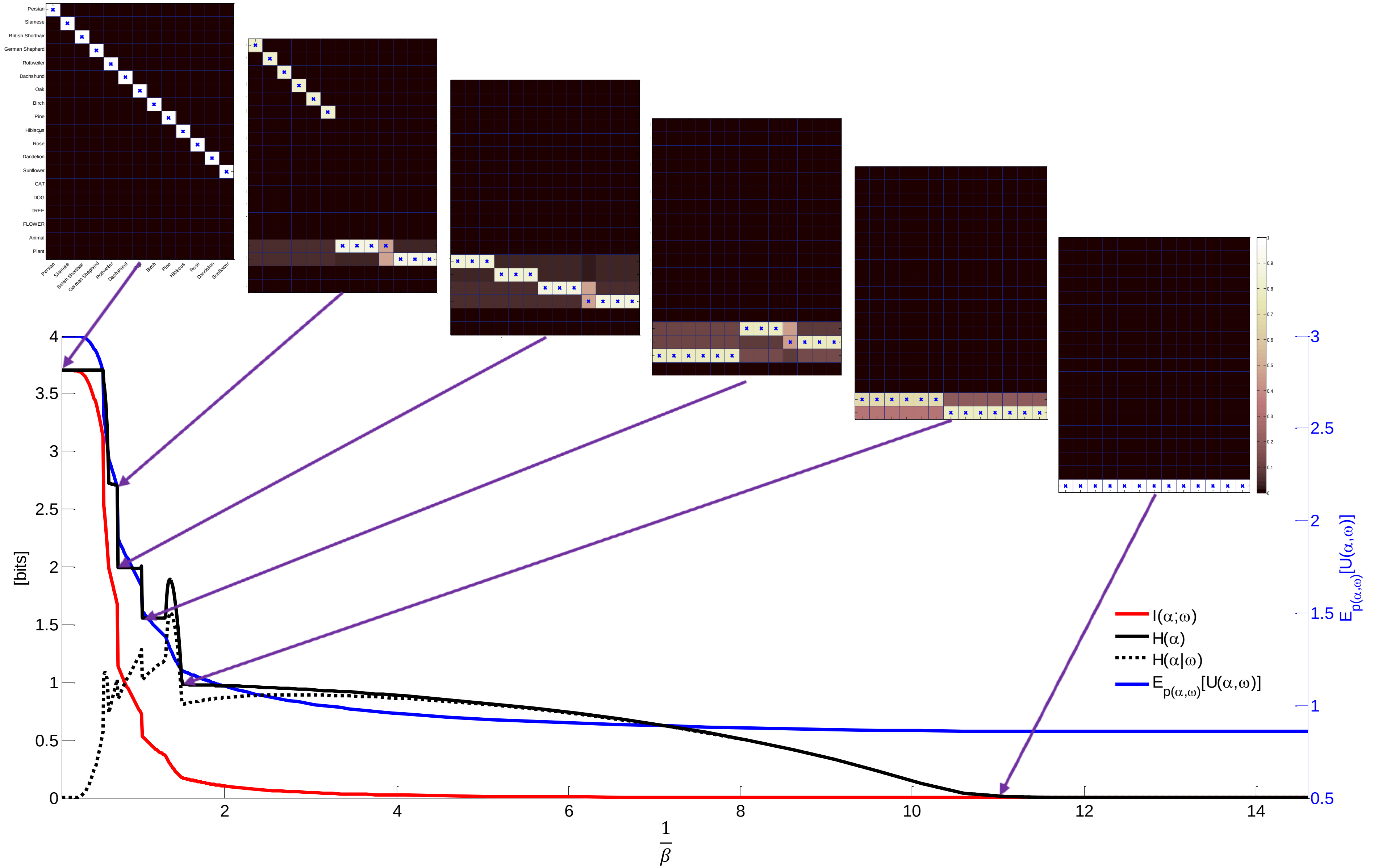


β	0.09	$\left[{}^{\text{bits}}/_{€}\right]$
I	≈ 0	[bits]
$\mathbf{E}[U]$	0.86	[€]



Continuously varying the temperature



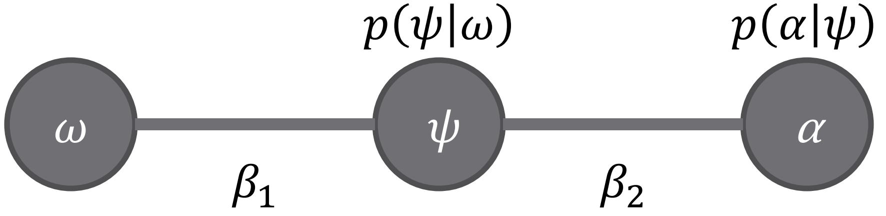


Persian
Siamese
British Shorthair
German Shepherd
Rottweiler
Dachshund
Oak
Birch
Pine
Hibiscus
Rose
Dandelion
Sunflower
CAT
DOG
TREE
FLOWER
Animal
Plant
[bits]
4
3.5
3
2.5
2
1.5
1
0.5
0
2
4
6
8
10
12
14
$\frac{1}{\beta}$
3
2.5
2
1.5
1
0.5
$E_{p(\alpha,\omega)}[U(\alpha,\omega)]$
$I(\alpha;\omega)$
$H(\alpha)$
$H(\alpha|\omega)$
$E_{p(\alpha,\omega)}[U(\alpha,\omega)]$

Extending towards hierarchies

- Temperature changes the granularity of abstraction
- Modelling hierarchies of abstractions?
  - Add variables to the model and apply the principle
  - Multiple ways to do this, here: processing pipeline

- Percept ψ + action α



$$\underset{p(\psi|\omega),\,p(\alpha|\psi)}{\arg\max}\ \mathbf{E}_{p(\alpha,\psi,\omega)}[U(\alpha,\omega)] - \frac{1}{\beta_1} I(\psi;\omega) - \frac{1}{\beta_2} I(\alpha;\psi)$$

Set of self consistent solutions

$$p(\psi|\omega) = \frac{1}{Z_\psi} p(\psi) \exp(\beta_1 \boldsymbol{\Delta F(\alpha|\psi)})$$

Rather than representing the input as good as possible, optimize the utility / computation-cost trade-off downstream $\Delta \mathrm{F} = \mathbf{E}_{p(\alpha|\psi)}[U(\alpha,\omega)] - \frac{1}{\beta_2} D_{KL}(\alpha|\psi\ ||\alpha)$

$$p(\alpha|\psi) = \frac{1}{Z_\alpha} p(\alpha) \exp\left(\beta_2 \sum_{\boldsymbol{\omega}} \boldsymbol{p(\omega|\psi)} U(\alpha,\omega)\right)$$

Take the average utility, using the Bayesian posterior over ω: $p(\omega|\psi)$

$$p(\psi) = \sum_{\omega} p(\omega) p(\psi|\omega)$$

$$p(\alpha) = \sum_{\omega,\psi} p(\omega) p(\psi|\omega) p(\alpha|\psi)$$

Discussion

- Convexity? Convergence?
- Relation to feed forward neural nets, deep architectures?

- Similar work
  - VAN DIJK, S. G. & POLANI, D. (2013). Informational Constraints-Driven Organization in Goal-Directed Behavior. *Advances in Complex Systems.*
  - STILL, S & CRUTCHFIELD, J. P. (2008). Structure or Noise? *arXiv:0708.0654v2 [physics.data-an]*
  - *VER STEEG G. & GALSTYAN A. (2014). Maximally informative Hierarchical Representations of High-Dimensional data*
  - Information Bottleneck Method, Relevant Information
  - Rational Inattention

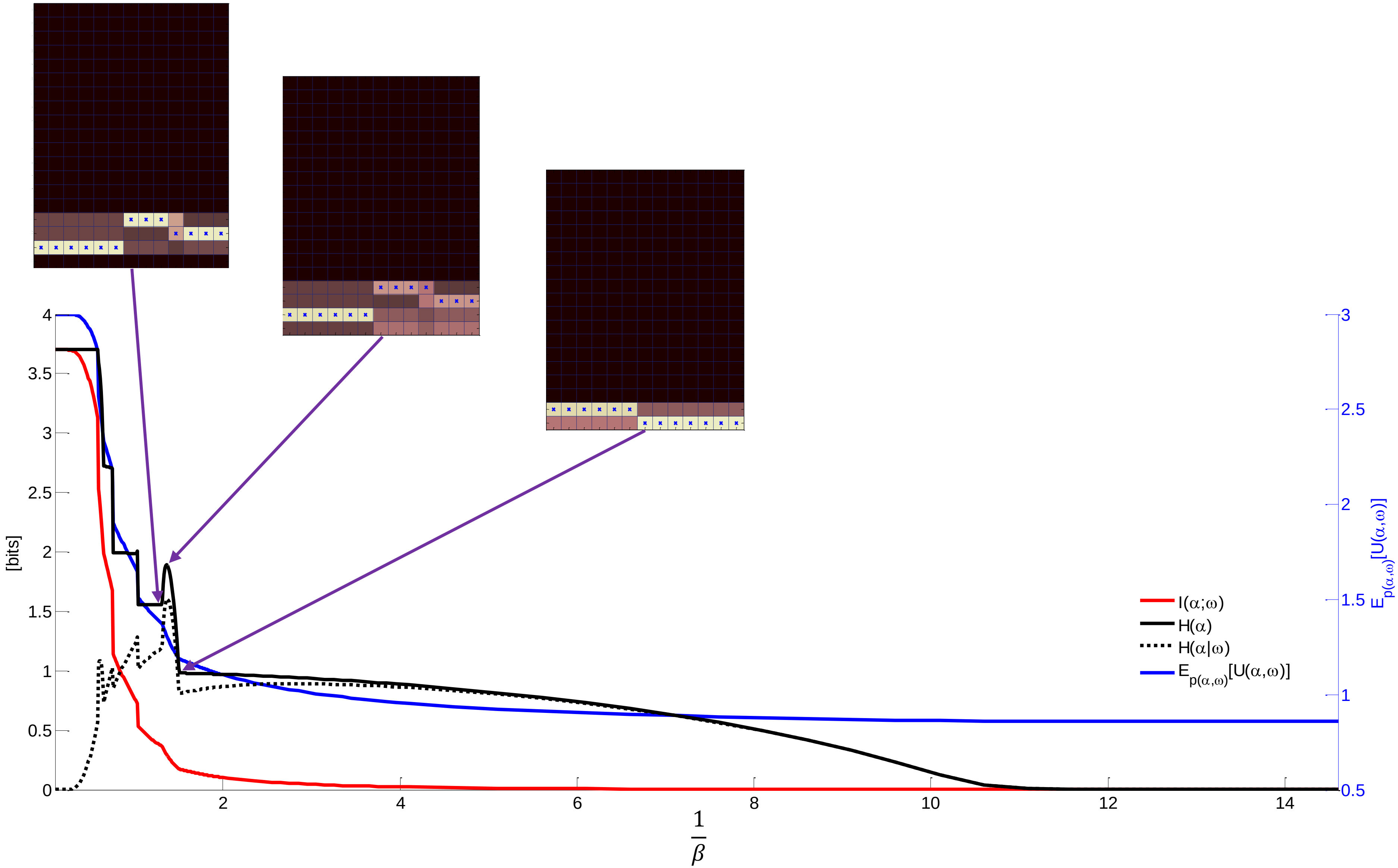
Set of self consistent solutions

$$p(\psi|\omega) = \frac{1}{Z_\psi} p(\psi) \exp\left(\beta_1 \sum_\alpha p(\alpha|\psi) \left(U(\alpha,\omega) - \frac{1}{\beta_2} \log \frac{p(\alpha|\psi)}{p(\alpha)}\right)\right)$$

$$p(\psi) = \sum_\omega p(\omega) p(\psi|\omega)$$

$$p(\alpha|\psi) = \frac{1}{Z_\alpha} p(\alpha) \exp\left(\frac{\beta_2}{p(\psi)} \sum_\omega p(\omega) p(\psi|\omega) U(\alpha,\omega)\right)$$

$$p(\alpha) = \sum_{\omega,\psi} p(\omega) p(\psi|\omega) p(\alpha|\psi)$$



[bits]
4
3.5
3
2.5
2
1.5
1
0.5
0
2
4
6
8
10
12
14
$\frac{1}{\beta}$
3
2.5
2
1.5
1
0.5
$E_{p(\alpha,\omega)}[U(\alpha,\omega)]$
$I(\alpha;\omega)$
$H(\alpha)$
$H(\alpha|\omega)$
$E_{p(\alpha,\omega)}[U(\alpha,\omega)]$