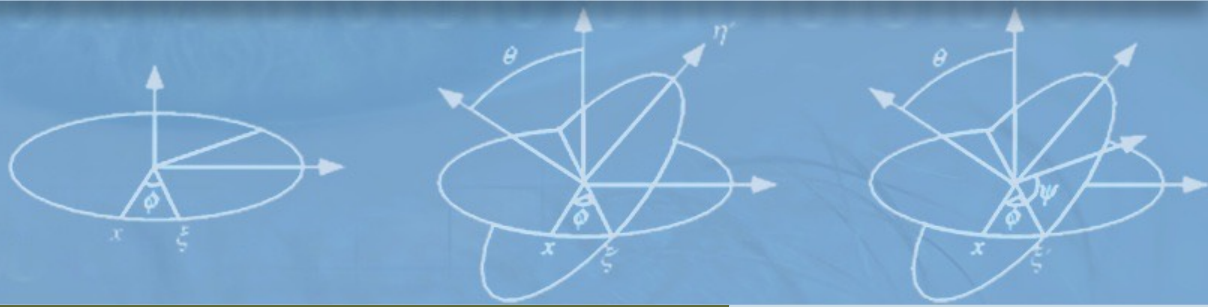
JHU vision lab

Semantic Information Pursuit for Multimodal Data Analysis

Aditya Chattopadhyay, Benjamin Haeffele, Donald Geman, René Vidal

Director of the Mathematical Institute for Data Science and Herschel Seder Professor of Biomedical Engineering at Johns Hopkins University, Amazon Scholar, and Chief Scientist at NORCE



THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

Introduction

- Shannon's "Mathematical Theory of Communication" 1948:
  - Signal represented by bits: message is irrelevant to its transmission.
- Weaver's "Three Levels of Communication" 1953:
  - Technical: how accurately can communication symbols be transmitted.
  - Semantic: how precisely the transmitted symbols convey the meaning.
  - Effectiveness: how effectively the received meaning affects conduct.
- Bar-Hillel and Carnap's "Semantic Information" 1953:
  - Shannon's information measures have nothing to do with what the symbols symbolize, but only with the frequency of their occurrence.
  - Bar-Hillel and Carnap's semantic information theory is based primarily on logic rules that are applicable only to a restricted class of signals.
- Dretske's "Knowledge and the Flow of Information" 1981:
  - Shannon's theory deals with amount of information & ignores content.

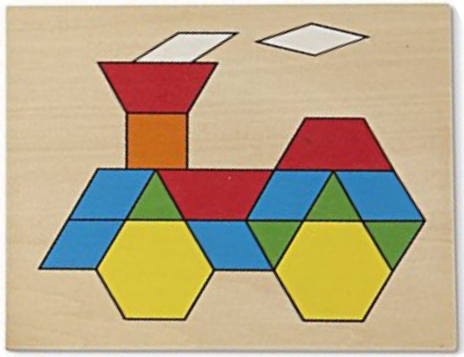
C E Shannon. A mathematical theory of communication. Bell System Technical Journal, 27:379 – 423; 623 –656, 1948.
W. Weaver. Recent contributions to the mathematical theory of communication. A Review of General Semantics, 1953.
Y. Bar-Hillel, R. Carnap. Semantic information. The British Journal for the Philosophy of Science, 4(14):147–157, 1953.
F. Dretske. Knowledge and the Flow of Information. The MIT Press, Cambridge, 1981.

Fundamental Challenges: Role of the Task

- The semantic content of a signal is irrelevant for transmission
  - Any signal can be optimally encoded by a sequence of bits (Huffman coding), transmitted and then reconstructed by the receiver.
- The semantic content of a signal is essential for recognition.
- Measuring semantic content is non trivial due to nuisances:
  - Different people may utter the same words using different speeds and accents, but the semantic content of the message remains the same.
  - An image of an object may be affected by lighting conditions, viewpoint, etc., but the identity of the object remains the same.
- This motivates the quest for new measures of information and new data representations that are relevant for a task, but are ideally invariant with respect to the task nuisances.

Fundamental Challenges: Role of Context

- The semantic content of a signal depends not only on the symbols, but also on how the symbols are arranged:
  - A sentence may have a different meaning depending on context.
  - A scene may only be recognizable once its constituent objects appear in a certain spatial arrangement.



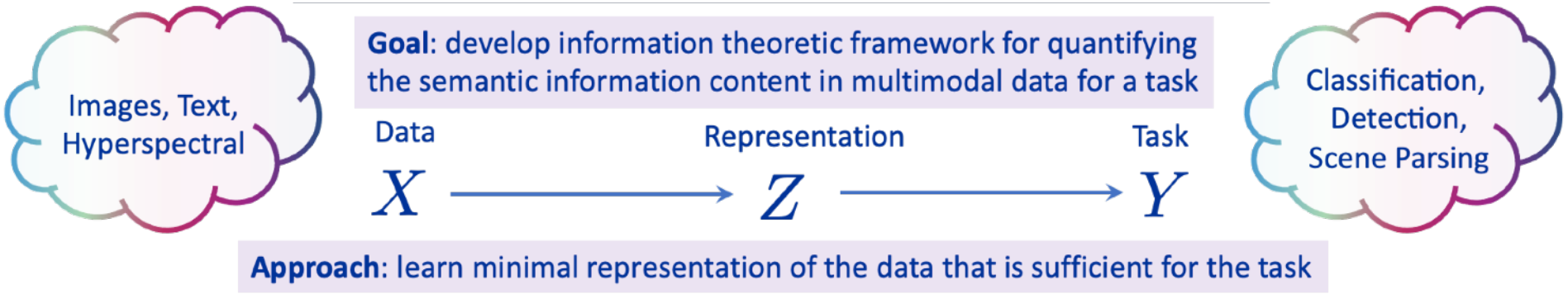
- This motivates the quest for new measures of information and new data representations that capture rich semantic and contextual relationships among scene entities.

Fundamental Challenges: Role of Modalities

- Depending on the task, some pieces of information may be more relevant than others, or the relevant information may be observable only in certain data modalities.
  - How do we know which modality is more relevant for a task?
  - How do we integrate information from multiple modalities, or from multiple spatial or temporal scales?
  - How do we take into account the fact that data from different modalities could be acquired at very different rates, or that the processing times of different modalities could be very different?
- This motivates the quest for new measures of uncertainty and new information fusion methods that depend upon the semantic content of the data.

Overall Goal, Significance and Approach

- **Goal**: develop information-theoretic framework for quantifying semantic information content in complex multimodal data.
- **Rationale**: why do we want to quantify semantic information?
  - Such a measure could help assess which data features or which data modalities are most important/relevant/informative for a task.
  - Such a measure could help assess the complexity of a task, which tasks are "harder" to solve than others, or a "distance" between tasks.
- **Proposed approach**: learn suitable representation for a task.



Talk Outline

- **Defining Semantic Entropy**
  - Semantic bits
  - Semantic coding
  - Semantic entropy
  - Relations to classical information measures and other properties

- **Computing Semantic Entropy via Information Pursuit**
  - Information Pursuit (IP) algorithm
  - Implementation IP using VAEs and Normalizing Flows

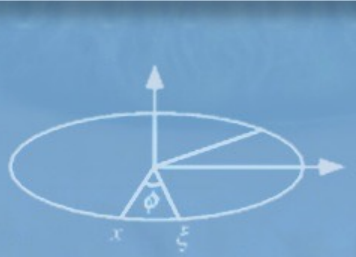
- **Experiments on Binary Image Classification**
  - MNIST < KMNIST < Fashion MNIST < Caltech Silhouettes

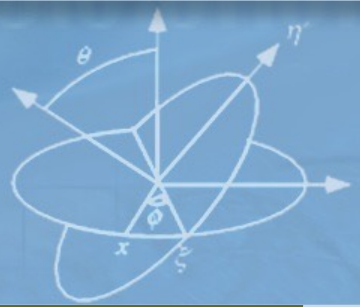
JHU vision lab

Defining Semantic Entropy

Aditya Chattopadhyay, Benjamin Haeffele, Donald Geman, René Vidal

Mathematical Institute for Data Science
Johns Hopkins University





THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

Shannon Entropy, Bits and Huffman Coding

- **Vocabulary** (set of symbols): $V = \{v_1, \ldots, v_N\}$
- **Probability** of each symbol: $p_i = P(X = v_i)$
- **Information content** (in bits): $h(v_i) = \log(1/p_i)$
- **Entropy**: minimum expected #bits needed to recover X. $H(X) = \sum_{i=1}^{N} p_i \log(1/p_i)$
- **Huffman coding**
  - Short code for common symbols, long code for rare symbols.
  - Create a binary tree whose leaves are the symbols.
  - Recursively merge notes with lowest probability.
- **Source coding theorem**

$$H(X) \leq \text{average length of Huffman code} \leq H(X) + 1$$

C E Shannon. A mathematical theory of communication. Bell System Technical Journal, 27:379 – 423; 623 –656, 1948
Huffman. A Method for the Construction of Minimum-Redundancy Code. 1952

More Recent Notions of Information

- **Representation** [Soatto '16]: a function of the data that is
  - Sufficient: as informative as the data
  - Invariant: discount the effect of uninformative data transformations
  - Minimal: "simpler" than the data, ideally minimal
- **Information bottleneck** [Tishby et al. '99]: trade-off sufficiency (H) and minimality (I)

$$\min_{q(z|x)} \mathcal{L} \doteq H_{p,q}(y \mid z) + \beta I(z; x)$$

- **Information on network weights** [Achille-Soatto '18]: minimize upper bound that induces minimality and invariance

$$\min_{q_w(y|x)} L \doteq H_{p,q}(\mathcal{D} \mid w) + \beta I(\mathcal{D}; w)$$

S. Soatto and A. Chiuso. Visual representations: Defining properties and deep approximations. ICLR, 2016.
N. Tishby, F. Pereira, W. Bialek. The information bottleneck method. Allerton, 1999.
A. Achille and S. Soatto, Emergence of Invariance and Disentangling in Deep Representations; JMLR. 2018
A. Achille & S. Soatto, Information Dropout, PAMI 2018.

Proposed Approach: Semantic Task

- **Task**
  - X = random variable denoting multimodal data or derived features.
  - Y = random variable we wish to predict from the data.
  - T = task of estimating p(Y | X) from samples of p(X,Y).
- **Semantic Task**
  - S = (Q, Z) = variables associated with a scene from vocabulary V.
  - Q = relevant variables for task T.
  - Z = nuisance variables for task T.
- **Example**
  - X = images of street scenes
  - T/Y/V = describe objects/relations
  - Q = semantic queries about scene
  - Z = viewpoint, illumination

1. Q: Is there a person in the blue region? A: yes
2. Q: Is there a unique person in the blue region? A: yes
   (Label this person 1)
3. Q: Is person 1 carrying something? A: yes

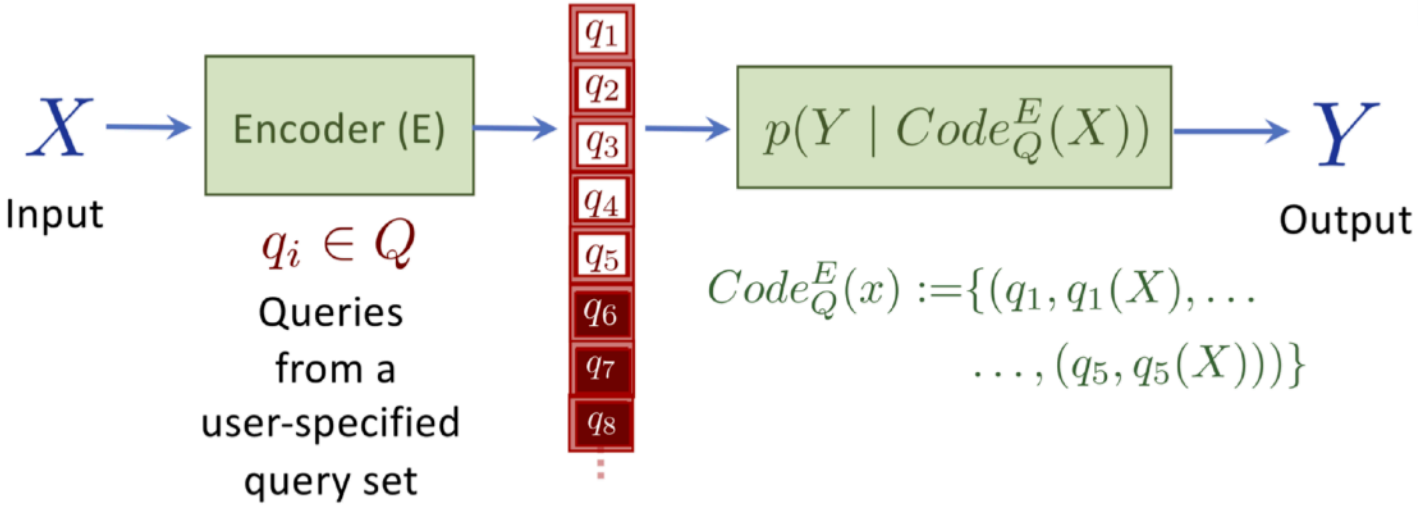
Proposed Approach: Semantic Bits/Code

- **Query set Q**
  - Q = relevant variables for task T = queries about scene content.
  - Q(X) = features obtained from data = answers to queries.
- **Sufficiency of Q**
  - We solve a task T by answering queries about X in order to predict Y. Hence, the answers to the queries need to be sufficient to predict Y.
  - The **query set Q is sufficient** if whenever (x,x') have identically answers, then their posteriors are equal

$$p(y \mid x) = p\big(y \mid \{x' : q(x') = q(x)\ \forall q \in Q\}\big)$$

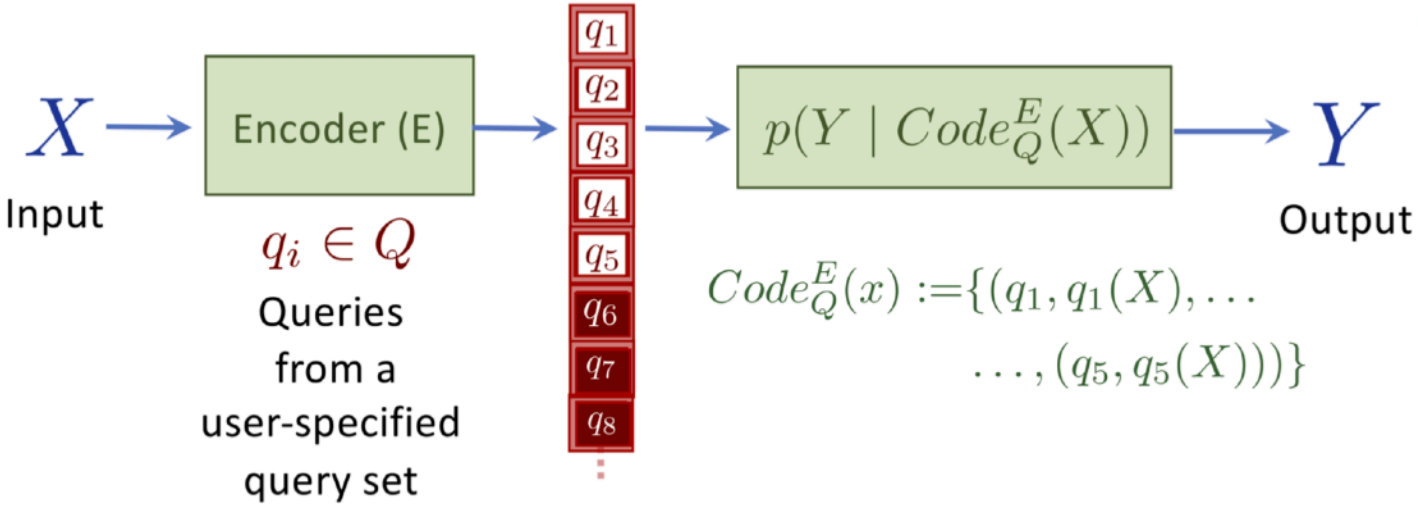
- **Semantic bits and semantic code**
  - Each query q_i is a basic unit of semantic information, or **semantic bit**.
  - The set of query-answer pairs, $Code_Q^E(X)$, defines a **semantic code**.

Proposed Approach: Semantic Encoder



- Given a query set Q, the **encoder** maps an input x to a sequence query-answer pairs to produce code $Code_Q^E(x)$.
  - $q_1 = E(\emptyset)$: the first query is independent of x.
  - $q_{k+1} = E(\{q_i, q_i(x)\}_{i=1}^k)$: subsequent queries depend on previous query-answer pairs.
  - $q_{L+1} = q_{STOP}$: stop when $Code_Q^E(x) = \{q_i, q_i(x)\}_{i=1}^L$ is sufficient for y.

Proposed Approach: Semantic Entropy



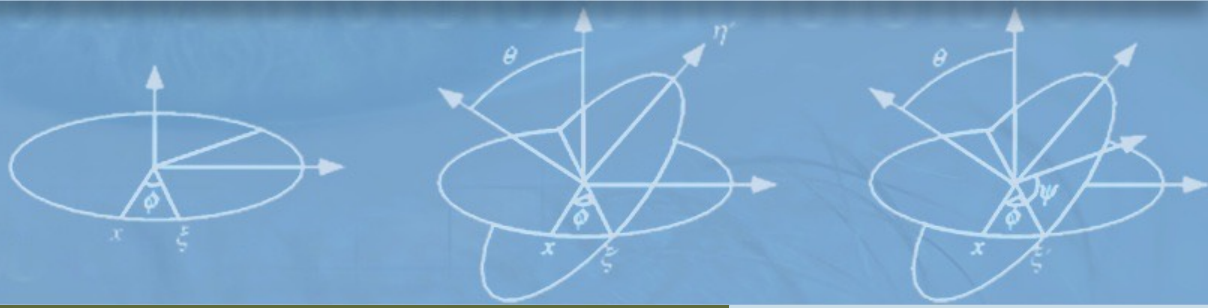
- **Semantic entropy**: minimum expected number of semantic queries about X whose answers are sufficient to predict Y:
  - Minimality: $SE_Q(X;Y) := \min_E \mathbb{E}_X[|(Code_Q^E(X))|]$
  - Sufficiency: $\text{s.t. } \forall x, y, \quad p(y \mid Code_Q^E(x)) = p(y \mid x)$

JHU vision lab

Computing Semantic Entropy Via Semantic Information Pursuit

Aditya Chattopadhyay, Benjamin Haeffele, Donald Geman, René Vidal

Mathematical Institute for Data Science
Johns Hopkins University



Computing Semantic Entropy is Intractable

- Computing the semantic entropy SE(X;Y) is generally intractable due to the huge number of possible semantic bits.

$$SE_Q(X;Y) := \min_E \mathbb{E}_X[|(Code_Q^E(X))|]$$
$$\text{s.t. } \forall x, y, \quad p(y \mid Code_Q^E(x)) = p(y \mid x)$$

- For a given scene and task, a small subset of queries may be enough to provide substantial information.

1. Q: Is there a person in the blue region? A: yes
2. Q: Is there a unique person in the blue region? A: yes
   (Label this person 1)
3. Q: Is person 1 carrying something? A: yes

- A more efficient procedure is to compute SE(X,Y) sequentially, by choosing at each step the most informative query given the history of queries and answers thus far.

Proposed Approach: Information Pursuit

- **Information Pursuit (IP)**: greedy strategy where the encoder chooses queries sequentially in order of information gain.

Definition: IP Encoder

Queries are chosen according to observed x.

- First query: $q_1 = \arg\max_{q \in Q} I(q(X); Y)$
- Next query: $q_{k+1} = \arg\max_{q \in Q} I(q(X); Y \mid q_{1:k}(x))$
- Termination: $q_{L+1} = q_{STOP}$ if $\max_{q \in Q} I(q(X); Y \mid q_{1:L}(x)) = 0$

$q_{1:k}(x)$ is the event that contains all realizations of X that agree on the first k query-answers for x.

- Approximation guarantees for IP are hard to obtain.
- **Theorem**: If Y is a discrete-valued function of X and Q is the set of all binary queries on X, $SE_Q^{IP}(X; Y) \leq SE(X; Y) + 1$.

Geman and Jedynak, An active testing model for tracking roads from satellite images, TPAMI, 1996.
Sznitman, Jedynak. Active Testing for Face Detection and Localization. TPAMI, 2010.
Jahangiri, Yoruk, Vidal, Younes, Geman. Progressive scene annotation by information pursuit. In ArXiv, 2017.
Chattopadhyay, Haeffele, Geman, Vidal. Quantifying Task Complexity Through Generalized Information Measures, 2021.

Computing Mutual Information is Intractable

- The selection of the first query requires computing $I(q(X); Y)$
  - Need a joint distribution of $q(X)$ and Y.

History

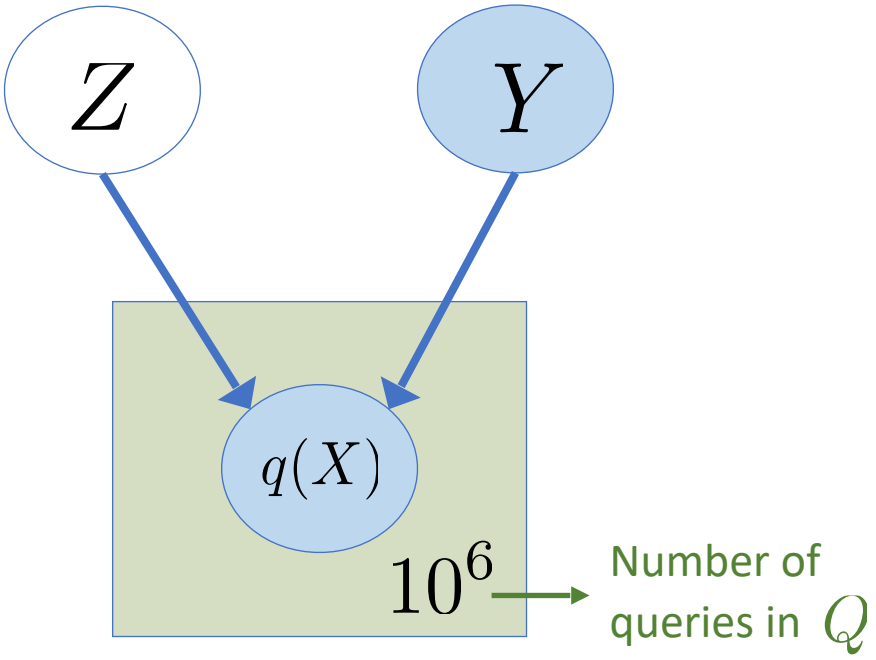
- Later queries require computing $I(q(X); Y \mid q_{1:k}(x))$
  - Need a joint distribution of $(q(X), Y)$ given History.
  - As histories get longer, we run out of samples that match History.
  - Extremely hard to estimate empirically.

- The above two problems need to be solved $\forall q \in Q$.
- What do we assume to make computation tractable?

Chattopadhyay, Haeffele, Geman, Vidal. Quantifying Task Complexity Through Generalized Information Measures, 2021.

Queries are Independent Given Nuisances

- **Insight**: query answers are conditionally independent given "some" latent nuisance variables Z.



- Reasonable assumption unless queries are causally related

- **Examples**:
  - Z = pose and lighting conditions
  - Z = phonemes in speech

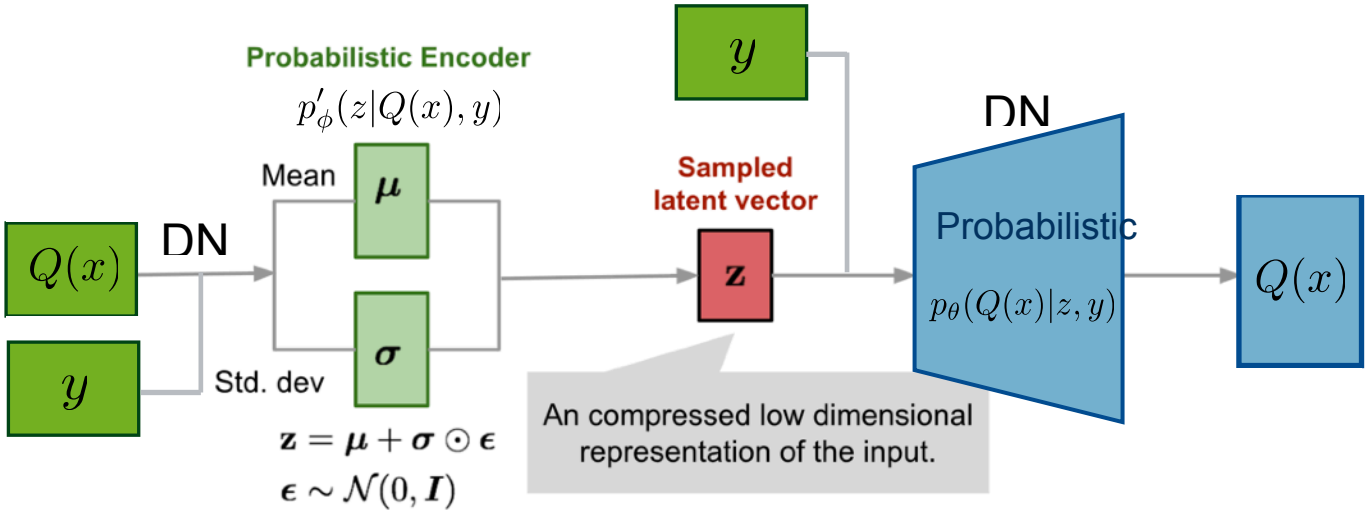
Implementation Using Conditional VAEs

- Learn generative model for $p(Q(X), Y)$
  - Assume queries are conditionally independent given (Z,Y):

$$p(Q(X), Y) = \int_Z \prod_q p(q(X) \mid Z, Y) p(Z) p(Y) dZ$$

  - Maximize Evidence Lower Bound (ELBO)

$$\max_{\omega,\phi} [\mathbb{E}_{Z \sim p'_\phi(Z|Q(X),Y)} [\log \prod_{q \in Q} p_\omega(q(X)|Z,Y)] - KL(p'_\phi(Z|Q(X),Y)) \mid\mid p(Z))]$$



Chattopadhyay, Haeffele, Geman, Vidal. Quantifying Task Complexity Through Generalized Information Measures, 2021.

Summary

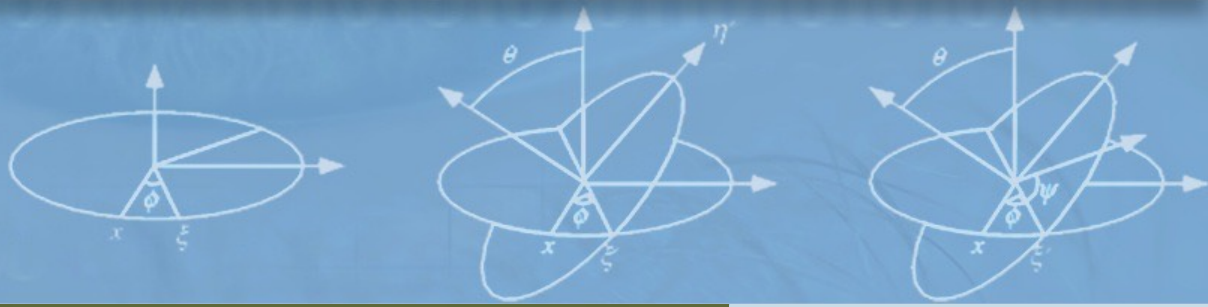
- Computing SE(X;Y) is generally intractable
- Information Pursuit is a greedy algorithm that makes the computation more efficient, but requires computing mutual information
- One can use a combination of latent variable models, VAEs, normalizing flows and MCMC to arrive at an efficient implementation

JHU vision lab

Semantic Information Pursuit for Binary Image Classification

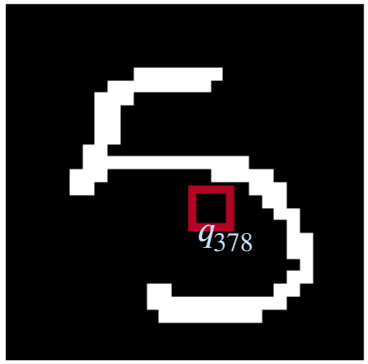
Aditya Chattopadhyay, Benjamin Haeffele, Donald Geman, René Vidal

Mathematical Institute for Data Science
Johns Hopkins University

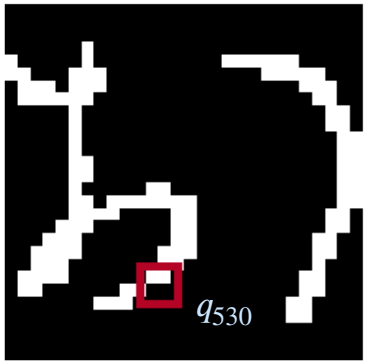


IP for binary image classification

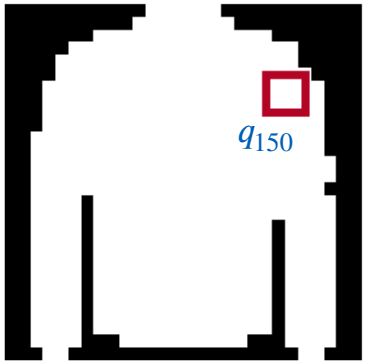
- Task is image classification.
- Queries q_i : "What are the image intensities at the i^{th} 3×3 patch?"



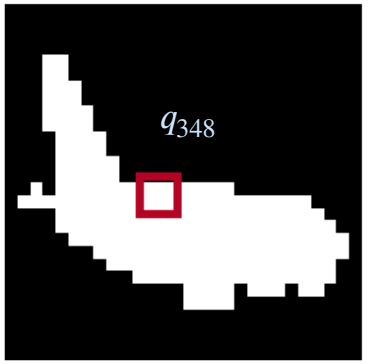
MNIST



KMNIST

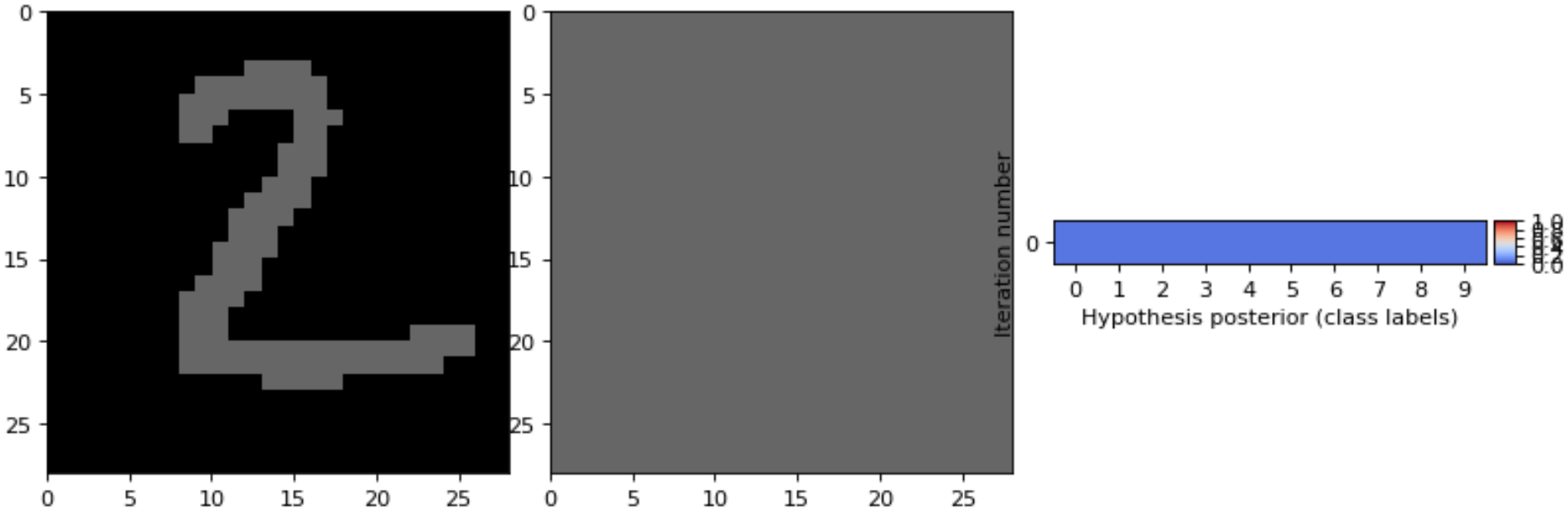


Fashion MNIST

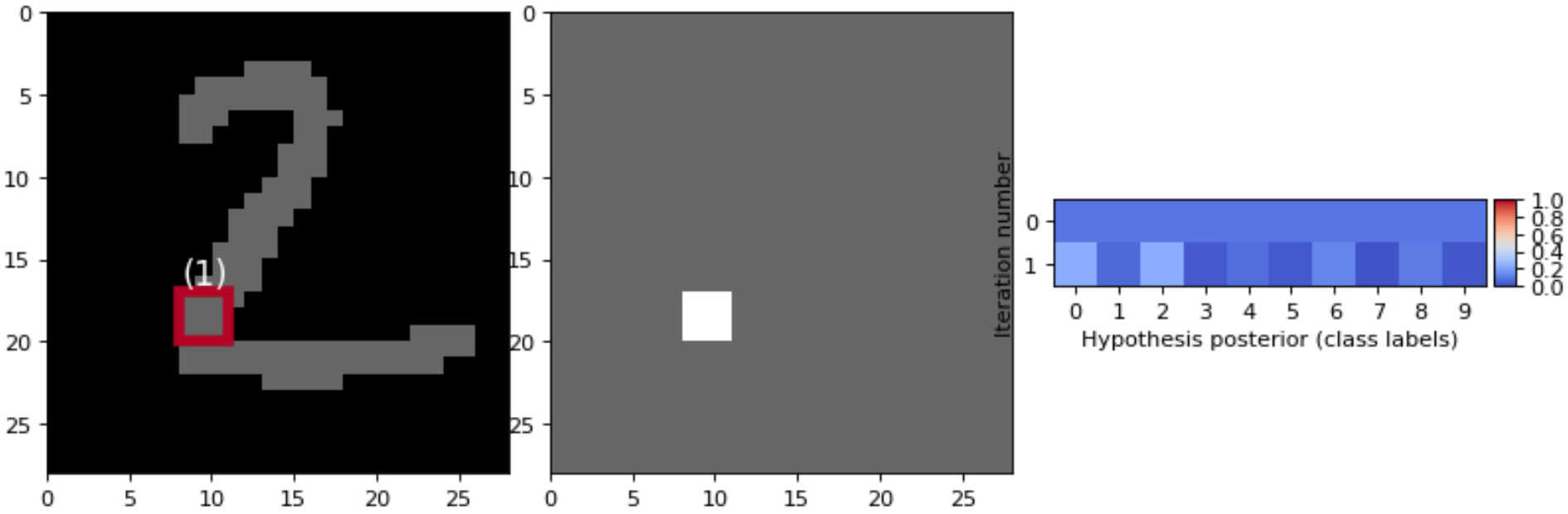


Caltech Silhouettes

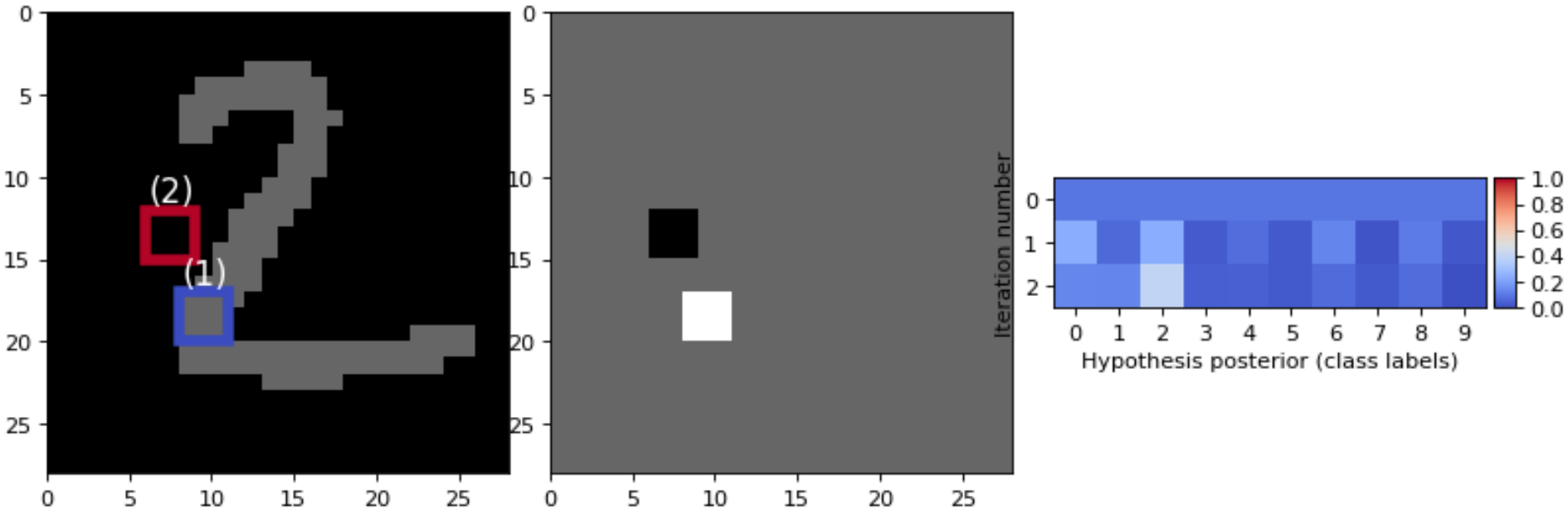
Experiments: IP in action (Iteration 0)



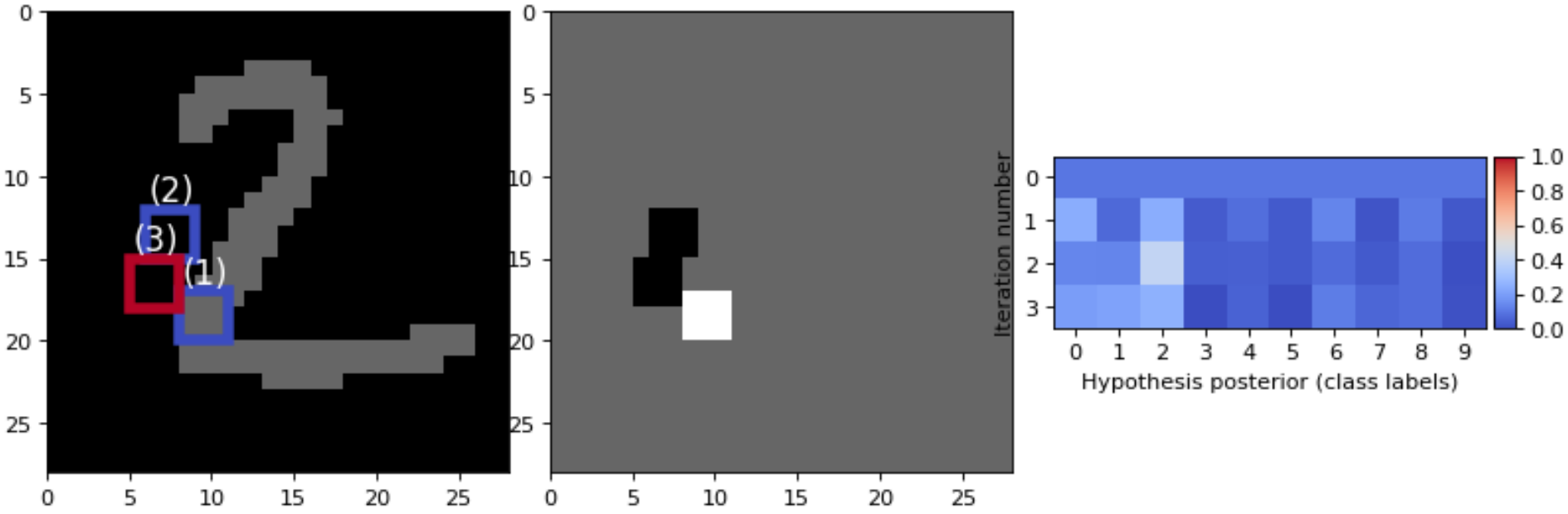
Experiments: IP in action (Iteration 1)



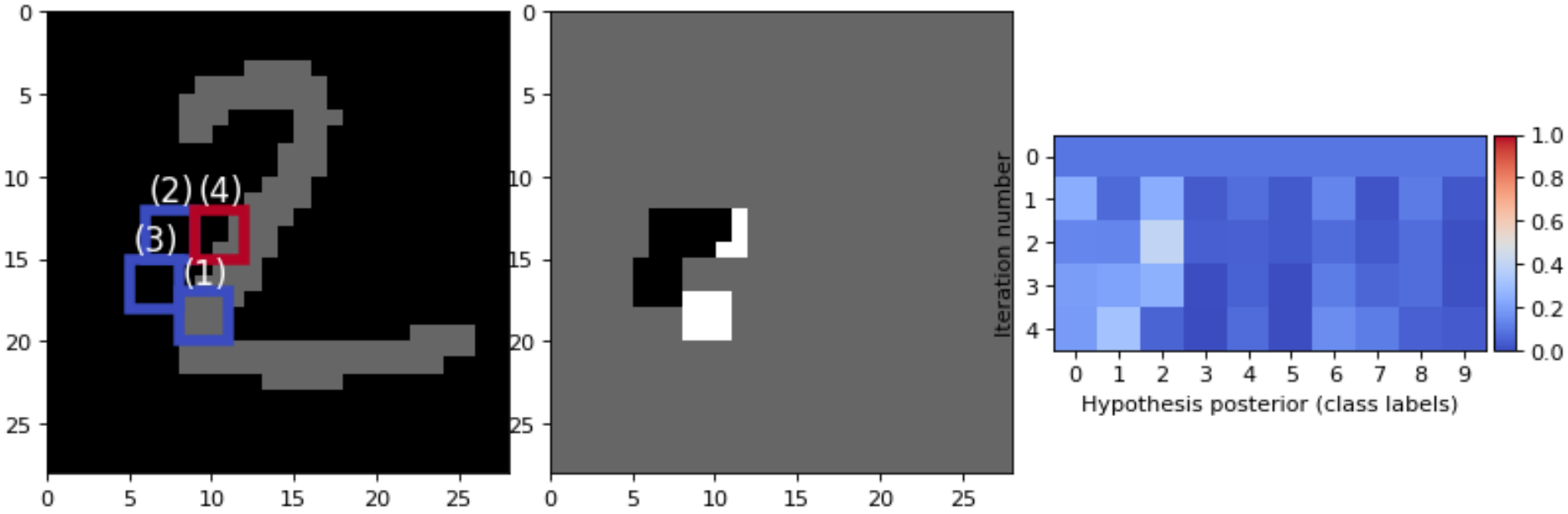
Experiments: IP in action (Iteration 2)



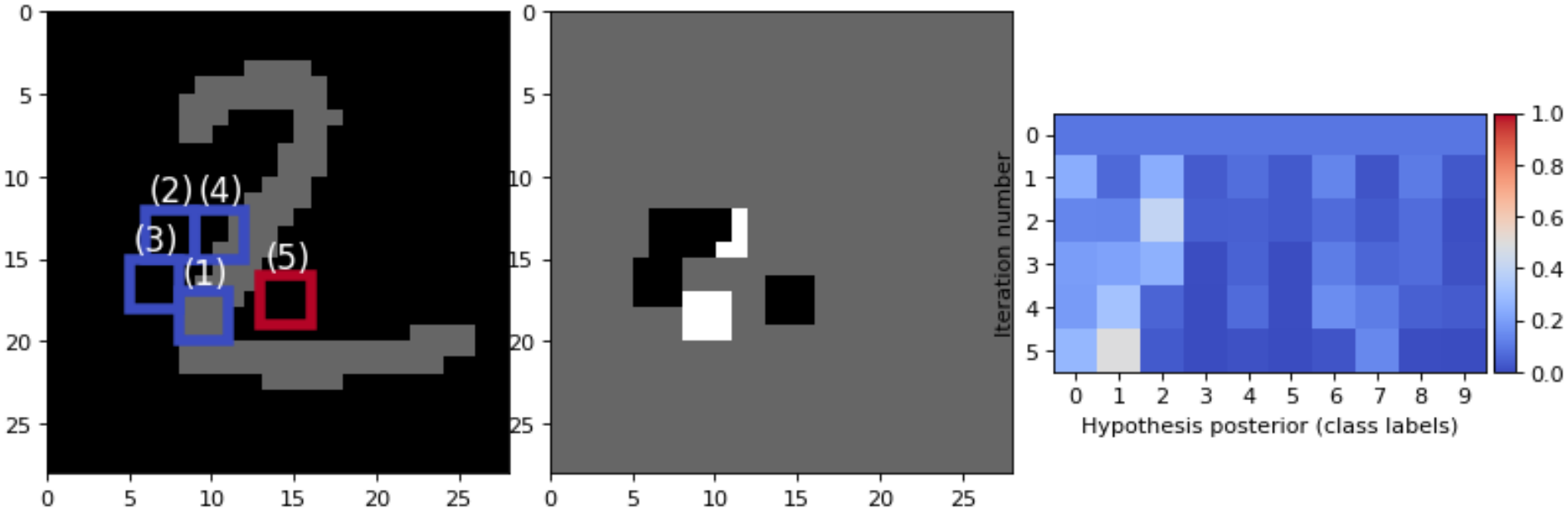
Experiments: IP in action (Iteration 3)



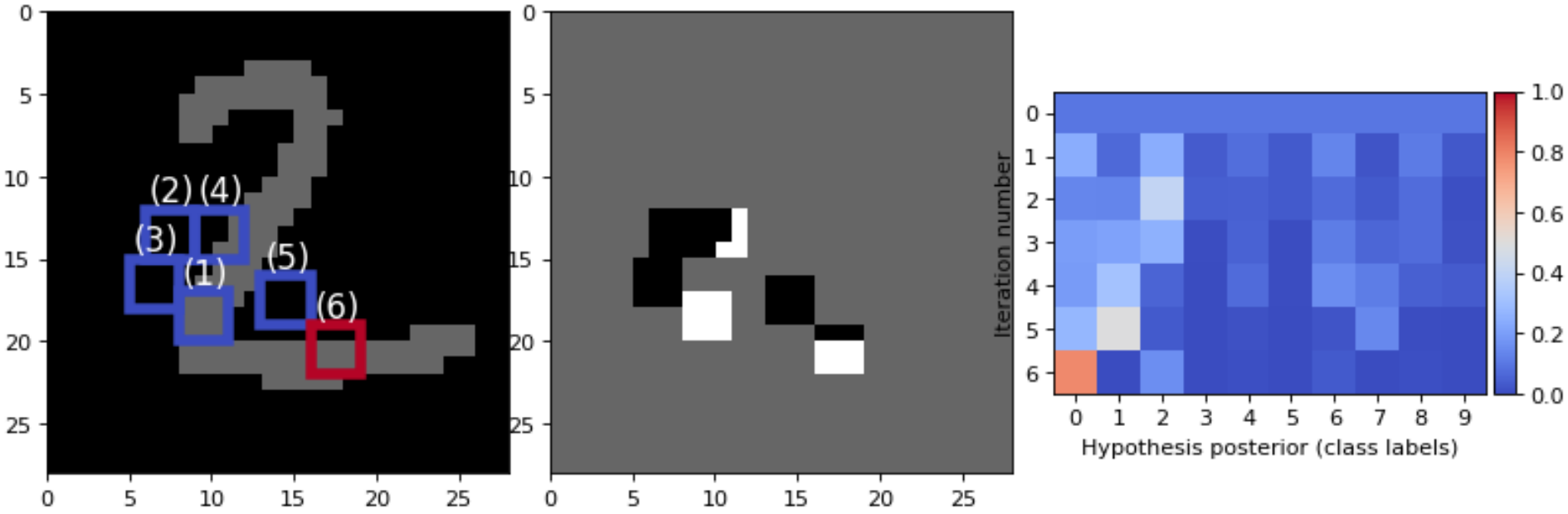
Experiments: IP in action (Iteration 4)



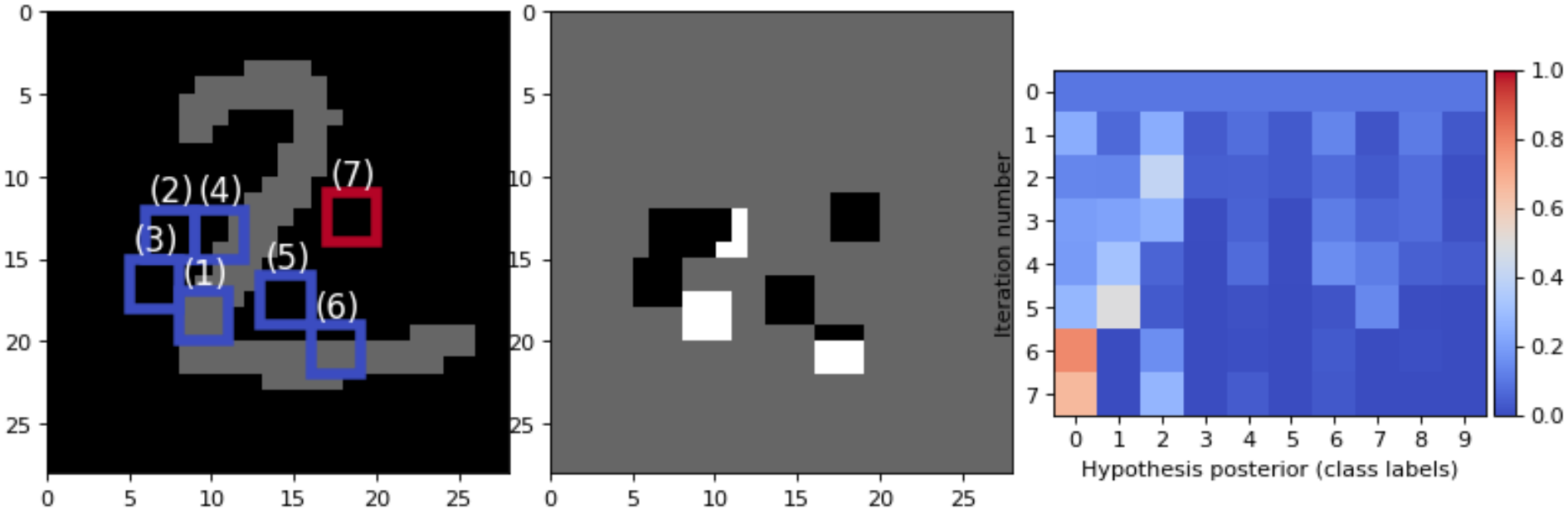
Experiments: IP in action (Iteration 5)



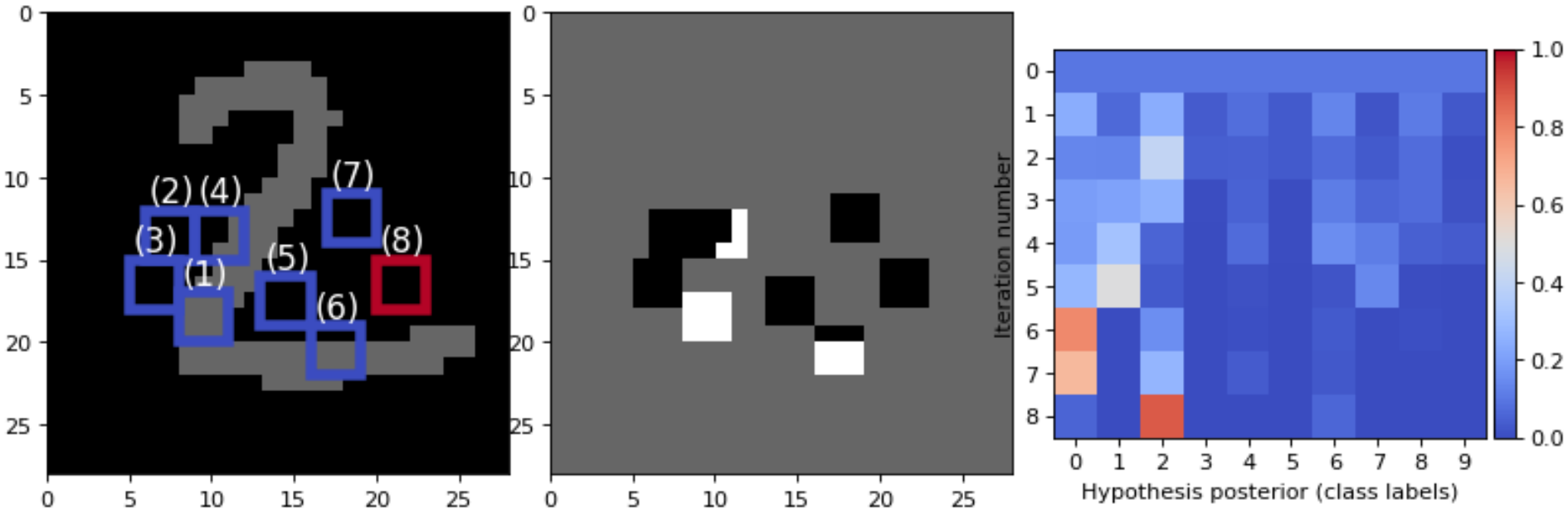
Experiments: IP in action (Iteration 6)



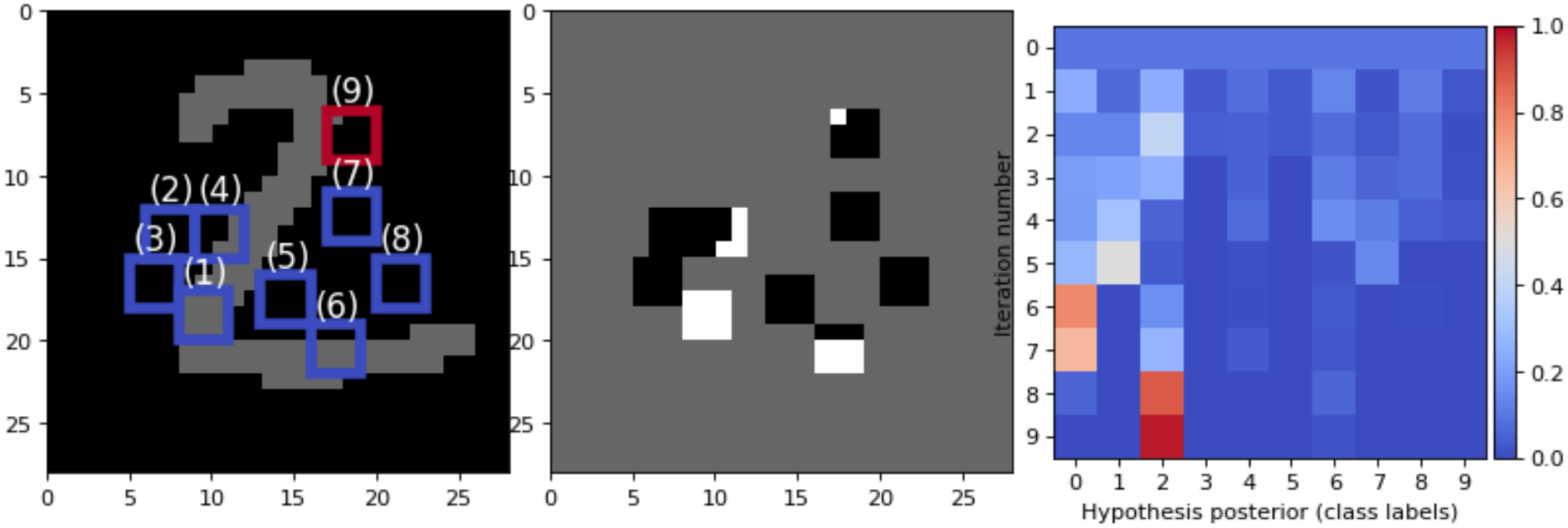
Experiments: IP in action (Iteration 7)



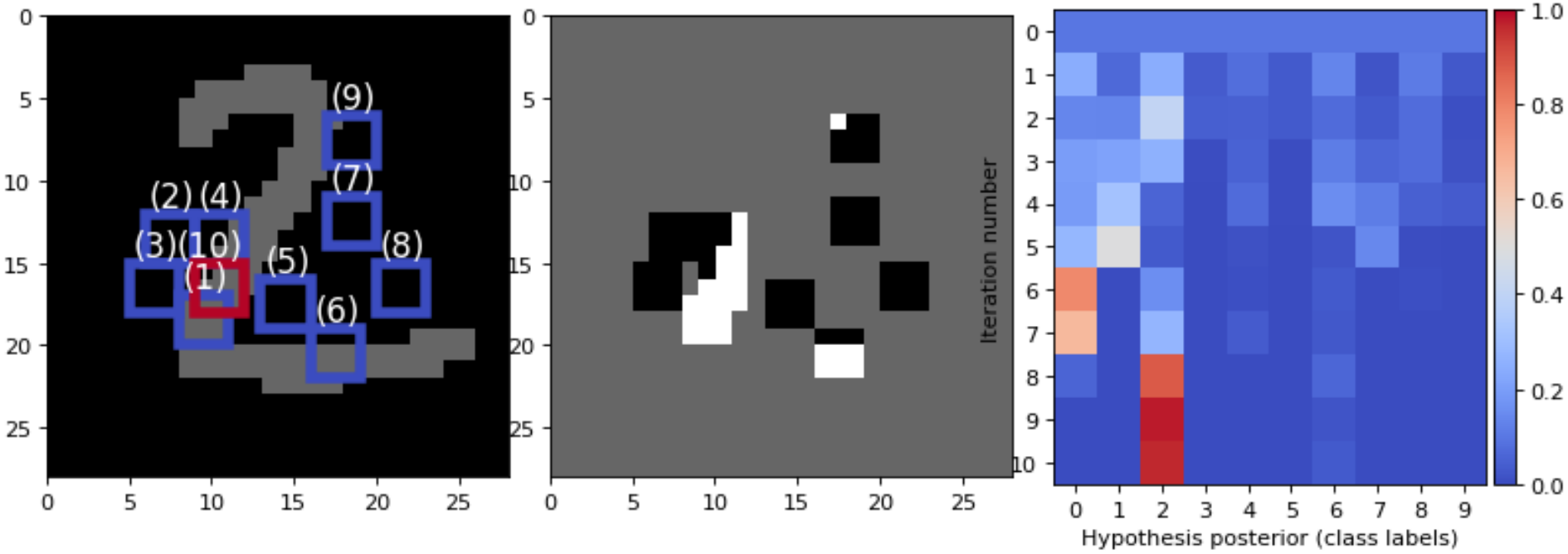
Experiments: IP in action (Iteration 8)



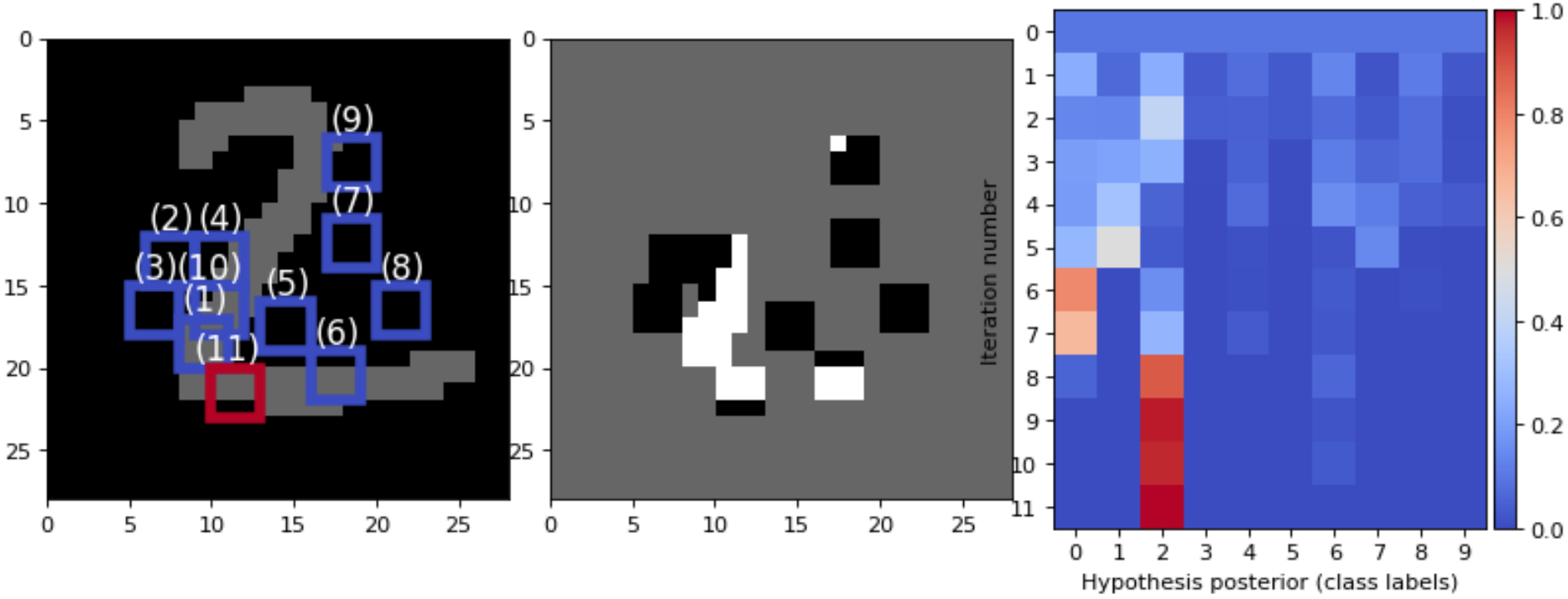
Experiments: IP in action (Iteration 9)



Experiments: IP in action (Iteration 10)



Experiments: IP in action (Iteration 11)



Experiments: Binary Image Classification

- Semantic Entropy correlates with task complexity.

	MNIST	K-MNIST	Fashion-MNIST	Caltech Silhouettes
				
$SE_Q(X;Y)$ (approx.)	11.54	26.89	40.60	61.58
CNN Test Accuracy	99.15	95.1	86.96	65.15

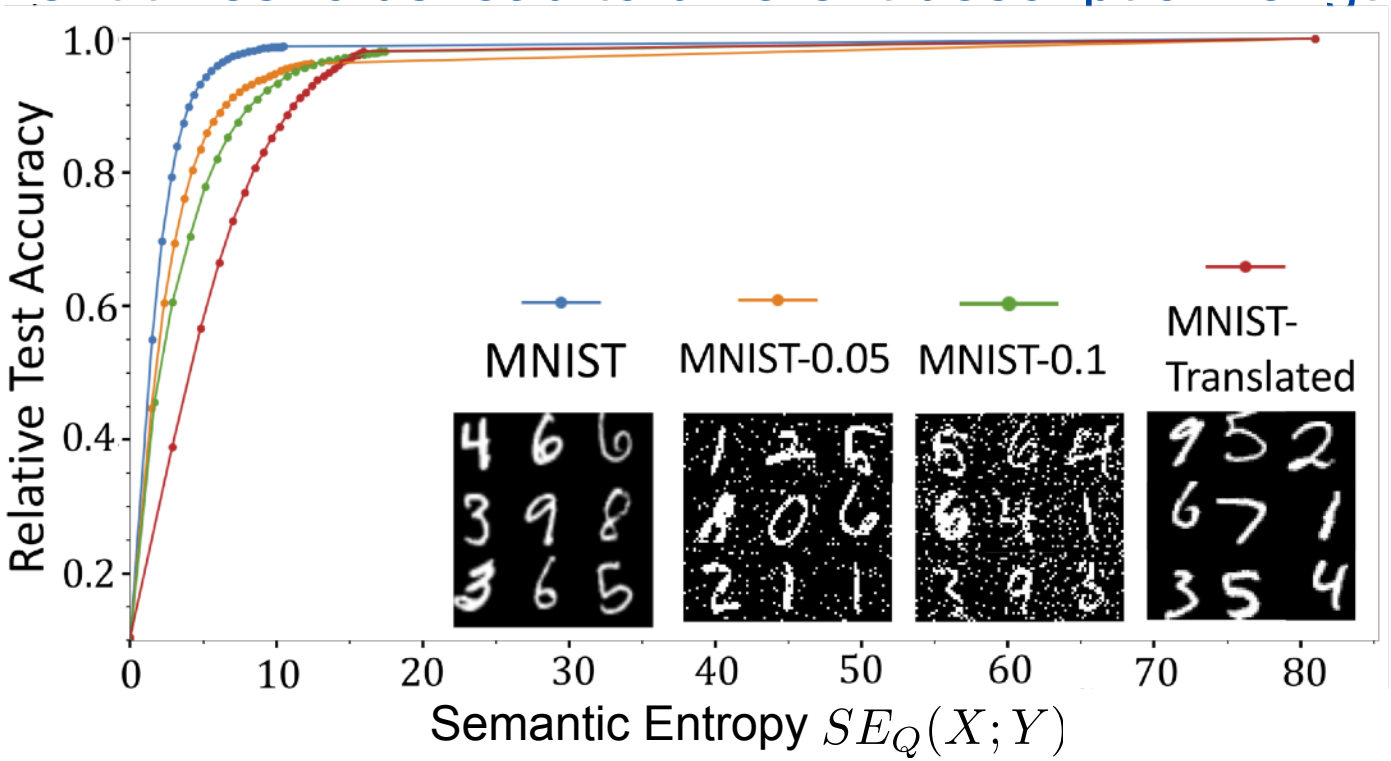
Figure 2. *The results conform with intuition of more complex datasets having higher semantic entropy. For instance, Caltech Silhouettes, a dataset of binarized images of 101 classes from the Caltech dataset is obviously semantically more complex than handwritten digits in the MNIST dataset.*

Adding nuisances increases complexity

- Stop IP when posterior is above user-defined threshold ϵ:

$$\max_Y p(Y \mid q_{1:L}(x)) > \epsilon$$

- Different thresholds lead to different description lengths.



Conclusions

- Defined a new notion of Semantic Entropy
  - Semantics is encoded via queries specified by the user.
- Computing SE(X;Y) is generally intractable
  - Information Pursuit is a greedy algorithm that makes the computation more efficient, but requires computing mutual information
  - One can use a combination of latent variable models, VAEs, normalizing flows and MCMC to arrive at an efficient implementation
- Demonstrated a proof-of-concept for the IP framework on a nontrivial task of image classification using patches queries.
- **Future work**:
  - Extend framework to more complex tasks.
  - Extend framework to multimodal data.

More Information,

Research supported by the Army Research Office (ARO)
Multidisciplinary University Research Initiative (MURI)
W911NF-17-1-0304.

Vision Lab @ JHU
http://www.vision.jhu.edu

Center for Imaging Science @ JHU
http://www.cis.jhu.edu

Mathematical Institute for Data Science @ JHU
http://www.minds.jhu.edu

Thank You!