

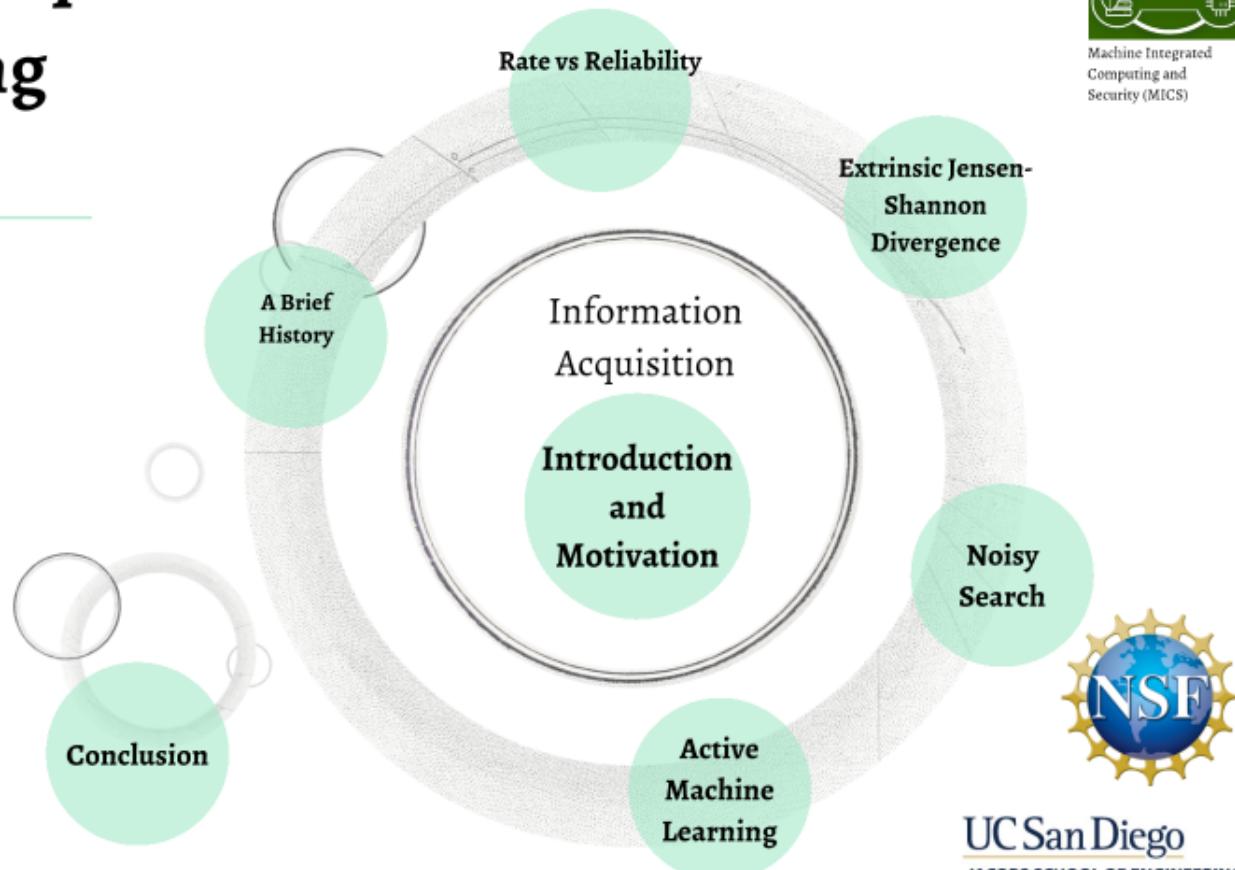
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz



Machine Integrated
Computing and
Security (MICS)



UC San Diego
JACOBS SCHOOL OF ENGINEERING
Center for Wireless Communications

(Labeled) Data Collection Story

First generation data analytics ignored the control over (big) data collection

- Learning models from given (passively collected/labeled) data
- Inference on arbitrary instances

(Labeled) Data Collection Story

First generation data analytics ignored the control over (big) data collection

- Learning models from given (passively collected/labeled) data
- Inference on arbitrary instances

In many practical/engineering settings some control over data collection

- Sensor management (run time)
- Data collection/labeling (training)

(Labeled) Data Collection Story

First generation data analytics ignored the control over (big) data collection

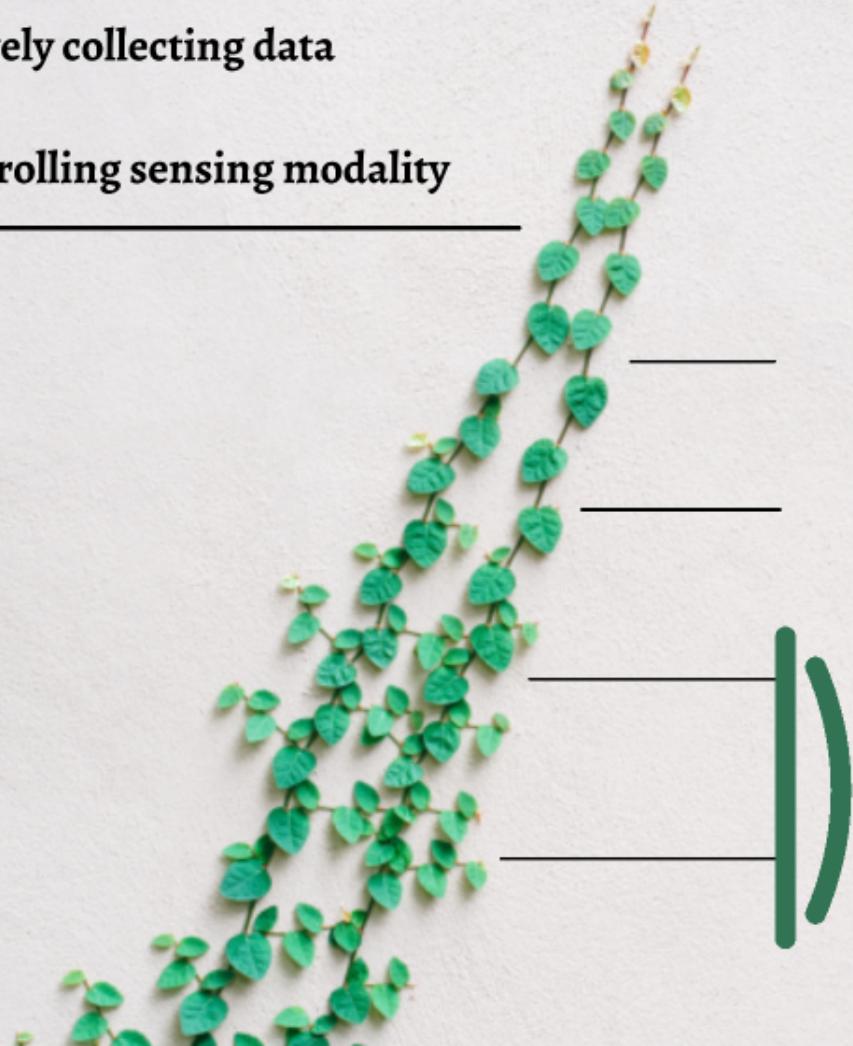
- Learning models from given (passively collected/labeled) data
- Inference on arbitrary instances

In many practical/engineering settings some control over data collection

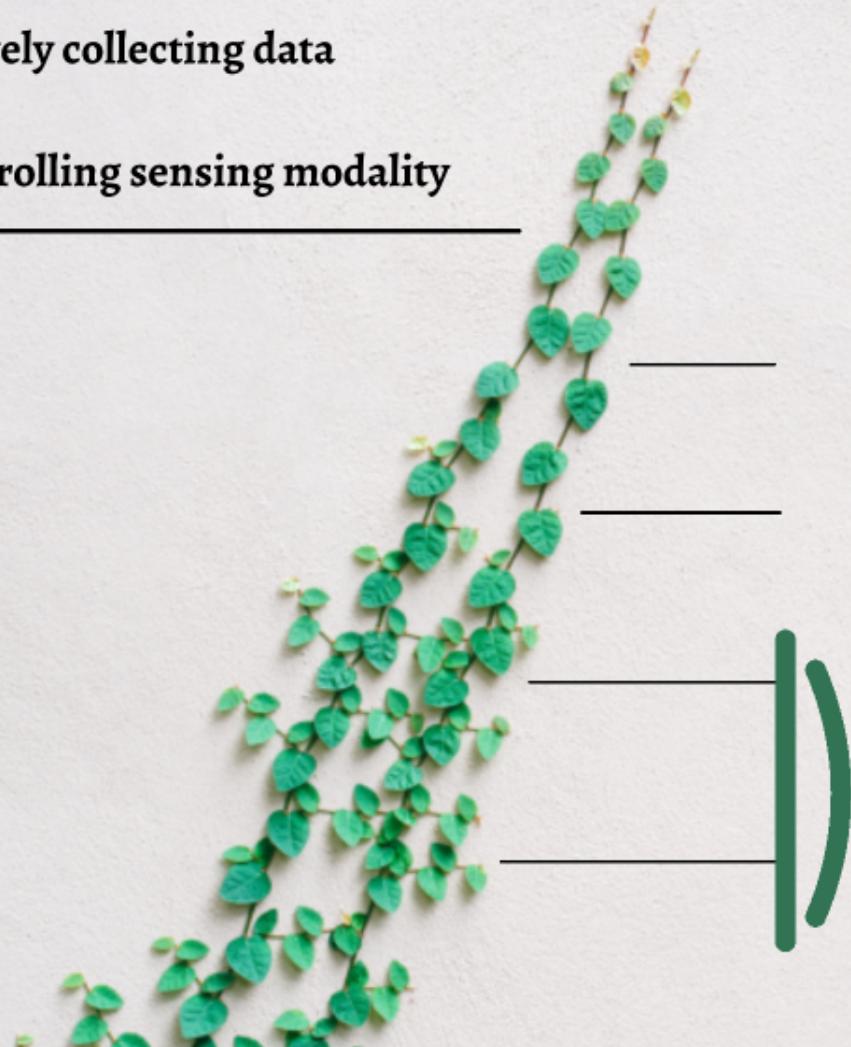
- Sensor management (run time)
- Data collection/labeling (training)

**Opportunities
and
Challenges**

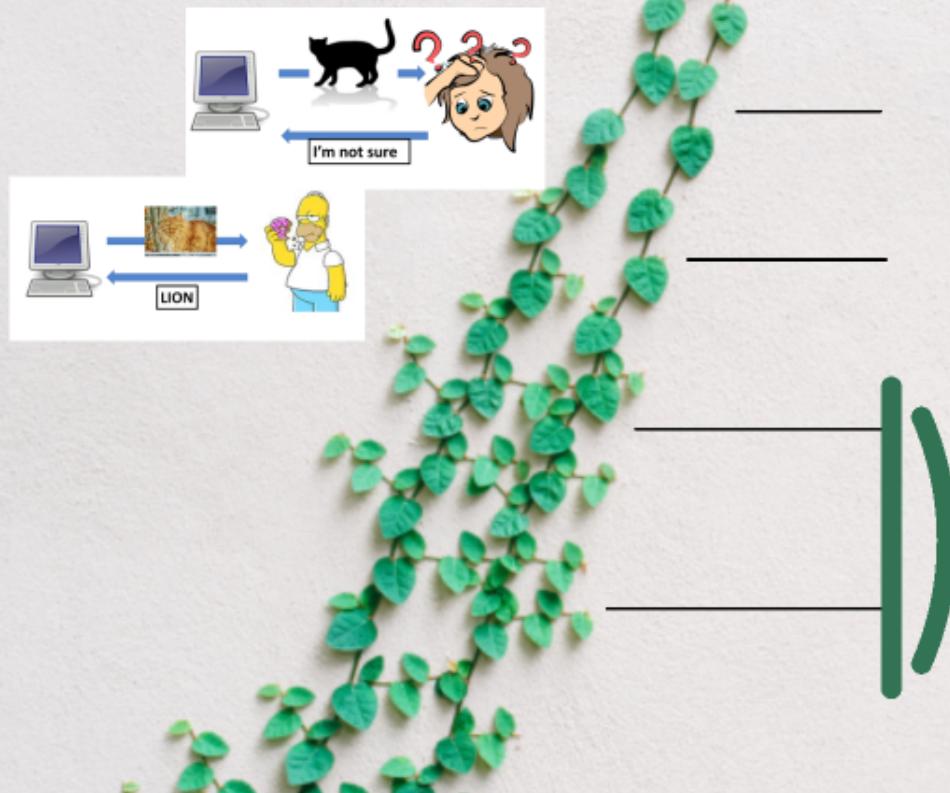
**Actively collecting data
and
Controlling sensing modality**

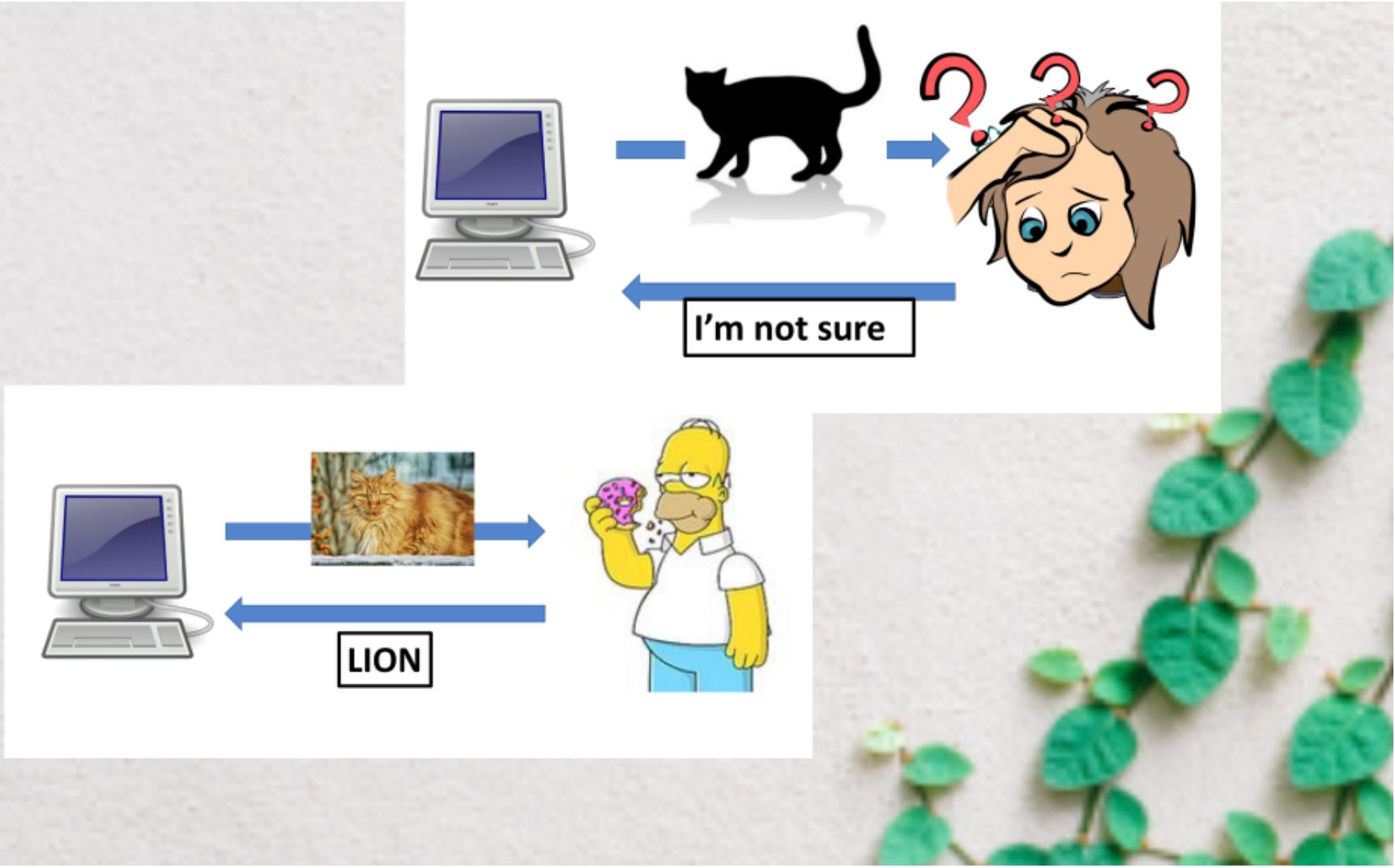


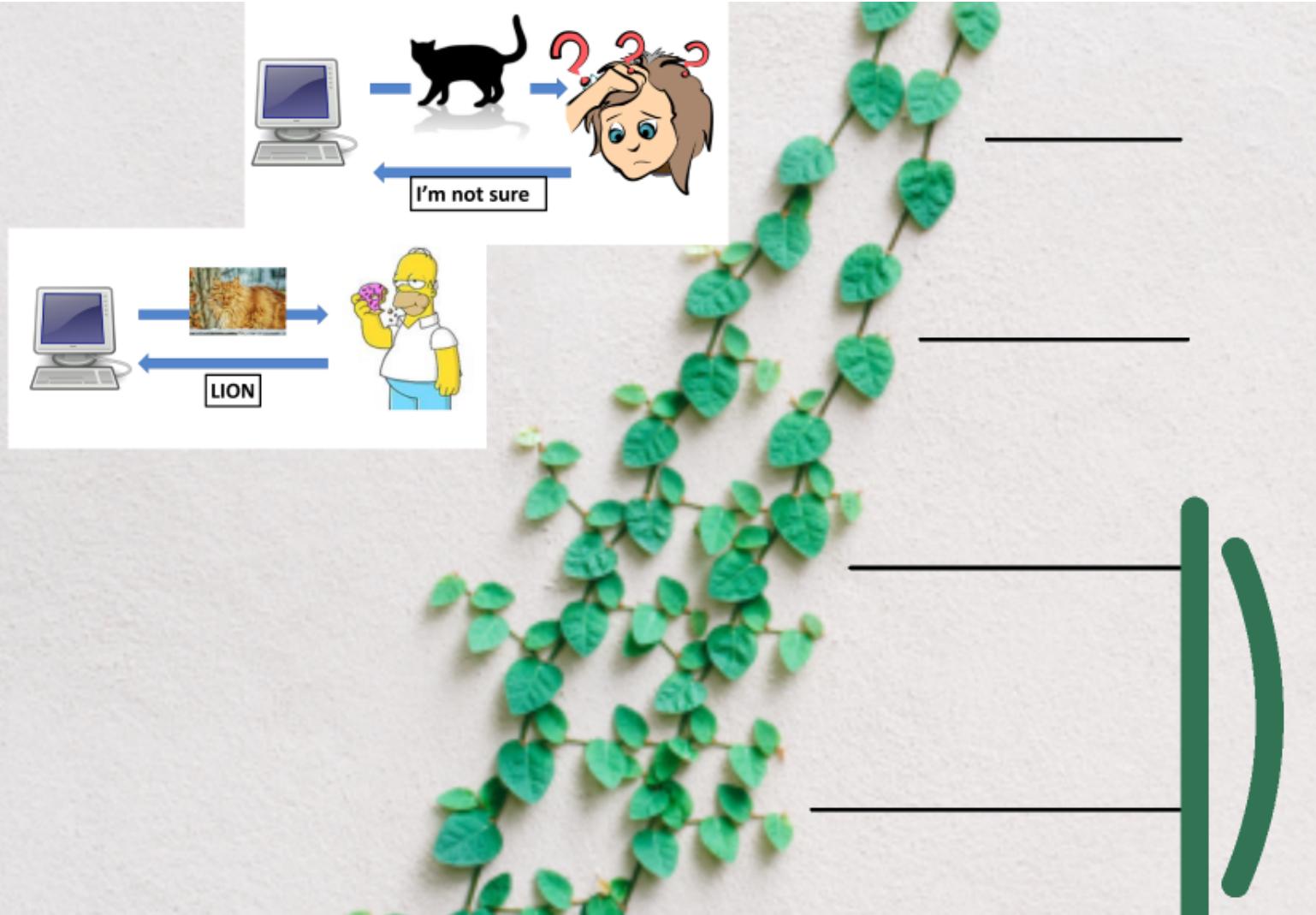
**Actively collecting data
and
Controlling sensing modality**

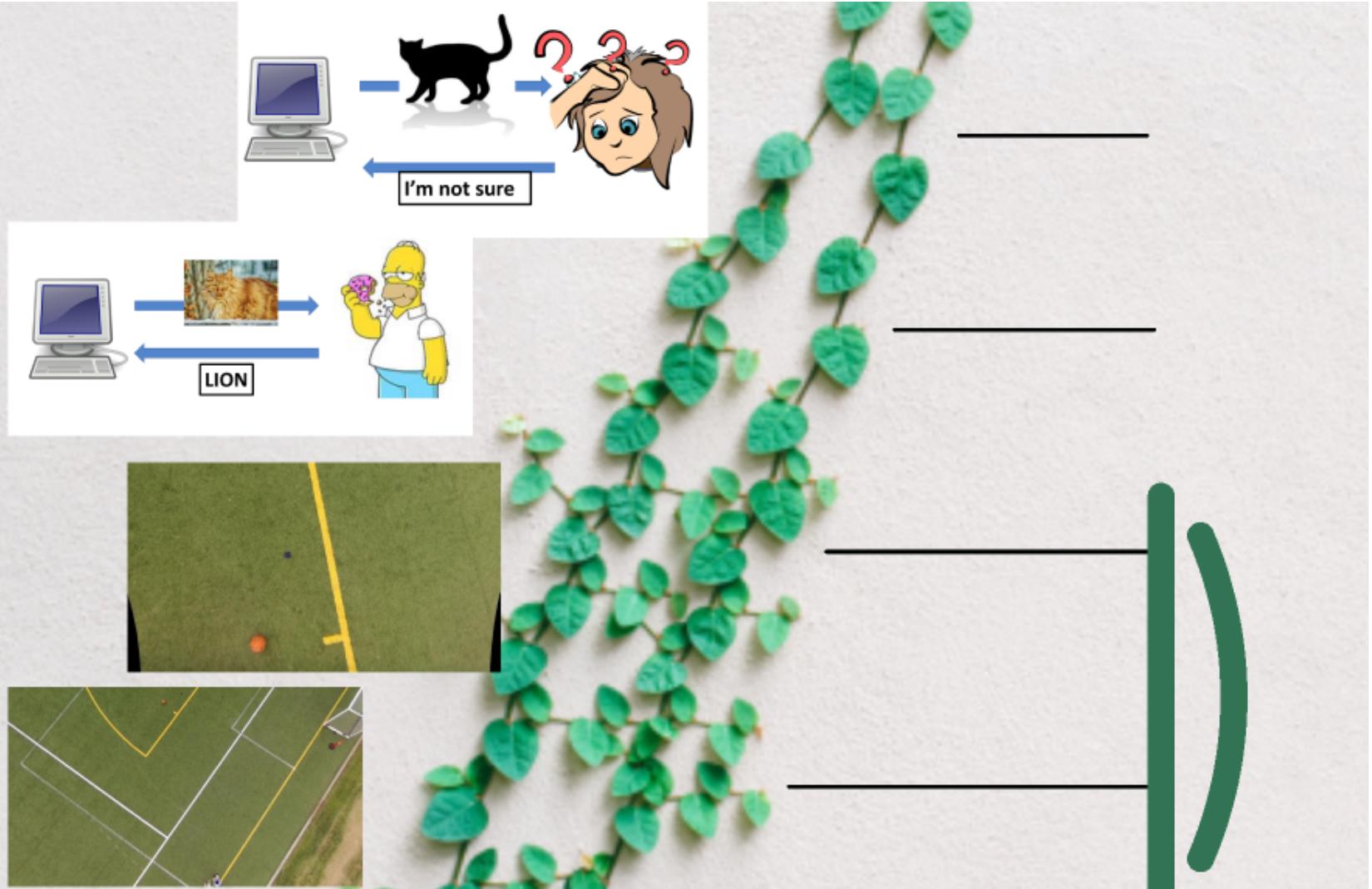


Actively collecting data and Controlling sensing modality

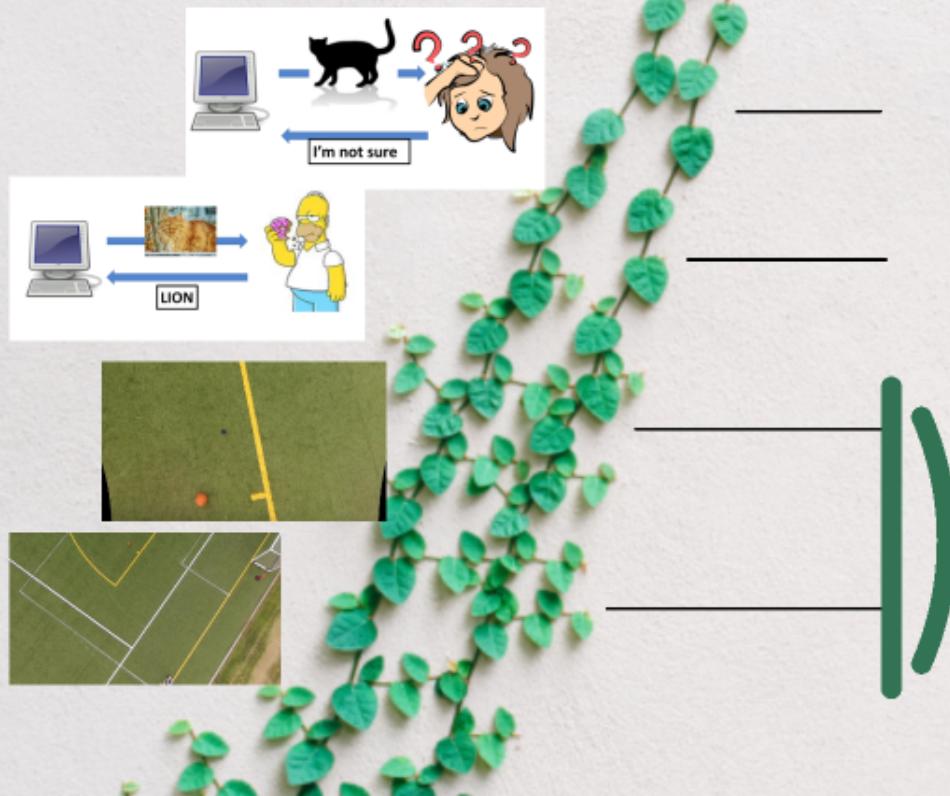




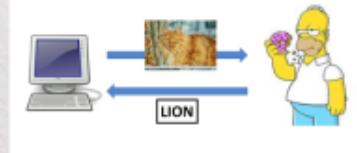
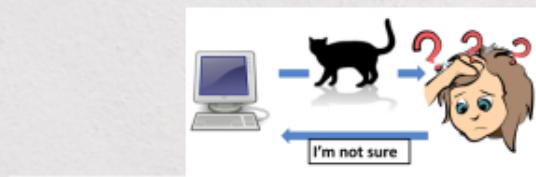




Actively collecting data and Controlling sensing modality

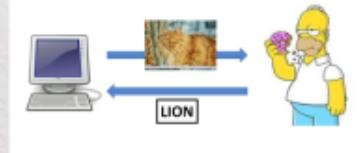
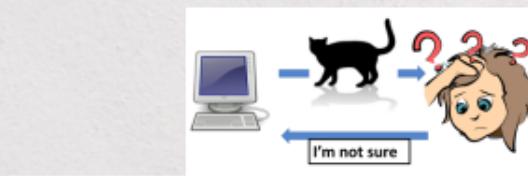


Actively collecting data and Controlling sensing modality



informative sample must maximally
reduce uncertainty

Actively collecting data and Controlling sensing modality

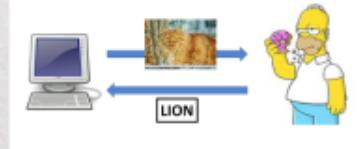
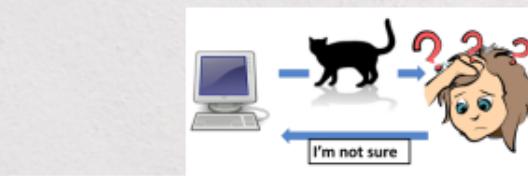


informative sample must maximally reduce uncertainty

identify appropriate notion of uncertainty



Actively collecting data and Controlling sensing modality



informative sample must maximally reduce uncertainty

identify appropriate notion of uncertainty

converses: fundamental limits
achievability: algorithms

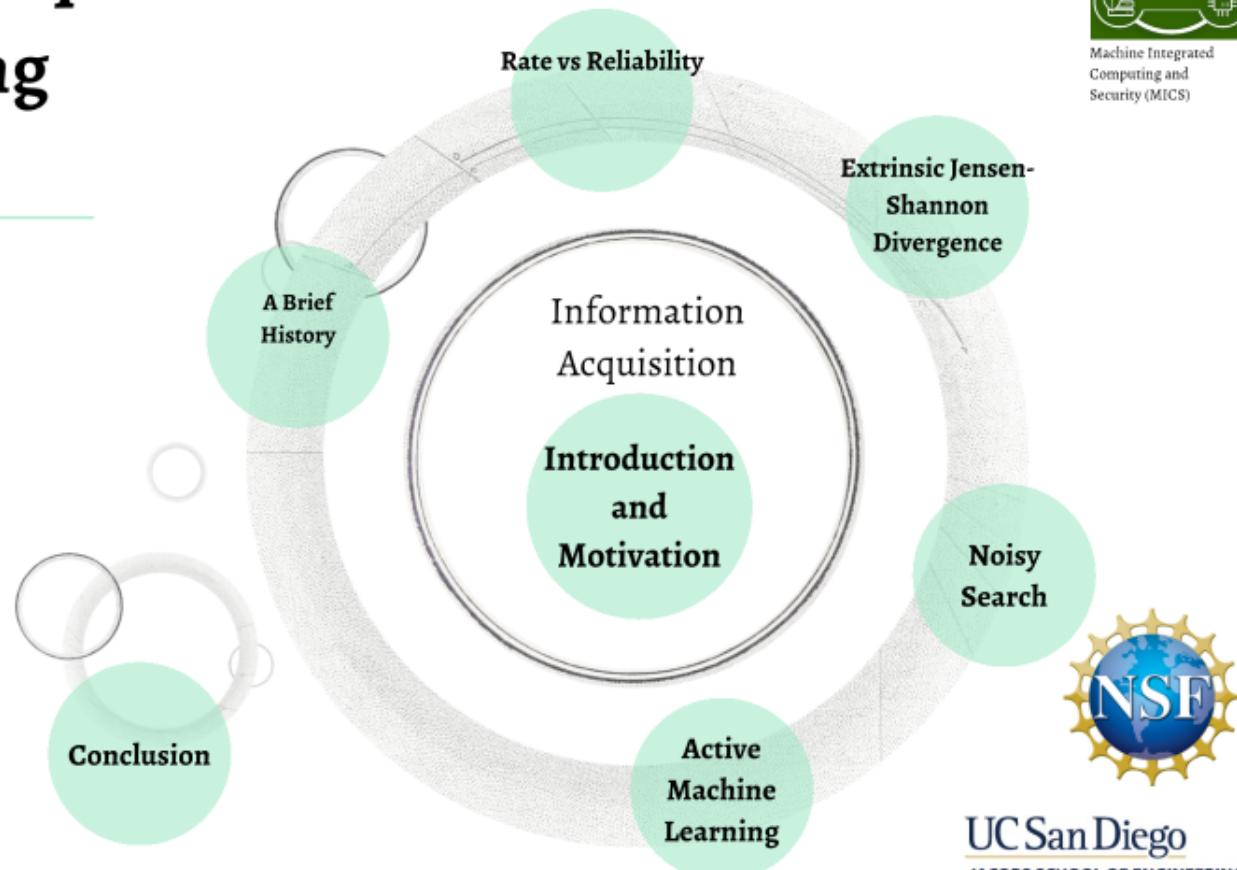
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz



Machine Integrated
Computing and
Security (MICS)



UC San Diego
JACOBS SCHOOL OF ENGINEERING
Center for Wireless Communications

Information Acquisition

- Stochastically varying state/parameter
- Tracked via partial/noisy yet controlled observations
- Controlled sequence of observations
- Generalizes:

time	1	2	...	T
state	X_1	X_2	...	X_T
acquisition action	A_1	A_2	...	A_T
observation	Y_1	Y_2	...	Y_T
utilization action	U_1	U_2	...	U_T

• Distribution of Y_t is determined by X_t and A_t

• Stochastic dynamic given as $P(X_{t+1}|X_t)$

Objective:

$$\text{Maximize } \mathbb{E} \left[\sum_t R_u(U(t), X(t)) - C_u(A(t)) \right]$$



Comparison of Experiments



Different Conclusions!



Information Acquisition

- Stochastically varying state/parameter
- Tracked via partial/noisy yet controlled observations
- Controlled sequence of observations
- Generalizes:



Comparison of Experiments



Different Conclusions!



time	1	2	...	T
state	X_1	X_2	...	X_T
acquisition action	A_1	A_2	...	A_T
observation	Y_1	Y_2	...	Y_T
utilization action	U_1	U_2	...	U_T

• Distribution of Y_t is determined by X_t and A_t

• Stochastic dynamic given as $P(X_{t+1}|X_t)$

Objective:

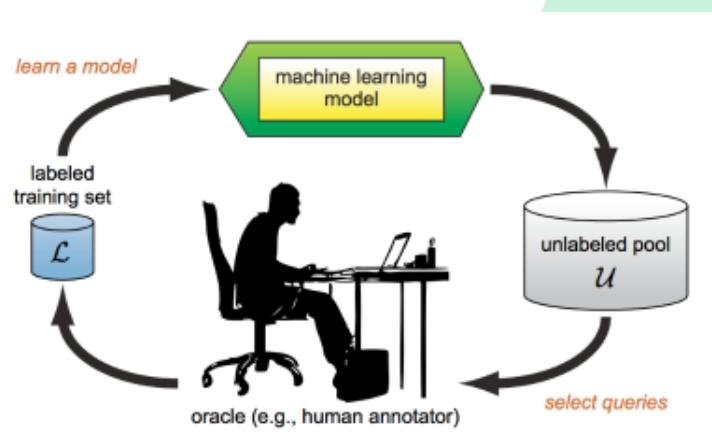
$$\text{Maximize } \mathbb{E} \left[\sum_t R_u(U(t), X(t)) - C_u(A(t)) \right]$$

time	1	2	...	T
state	X_1	X_2	...	X_T
acquisition action	A_1	A_2	...	A_T
observation	Y_1	Y_2	...	Y_T
utilization action	U_1	U_2	...	U_T

- Distribution of Y_t is determined by X_t and A_t
- Stochastic dynamic given as $\mathbb{P}(X_{t+1}|X_{1:t})$

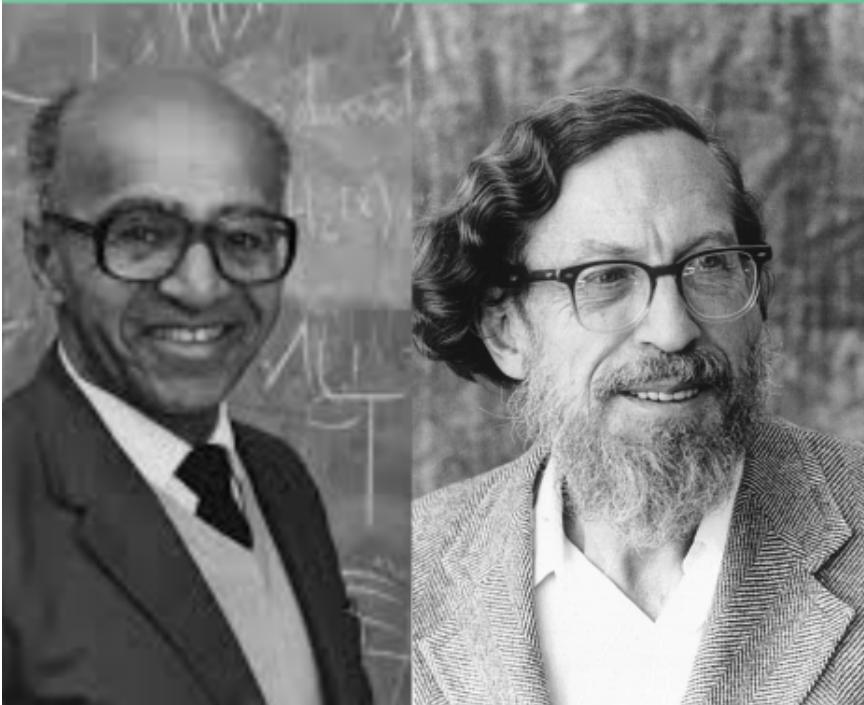
Objective:

$$\text{Maximize } \mathbb{E} \left[\sum_t R_u(U(t), X(t)) - C_a(A(t)) \right]$$



Information

- Stochastically

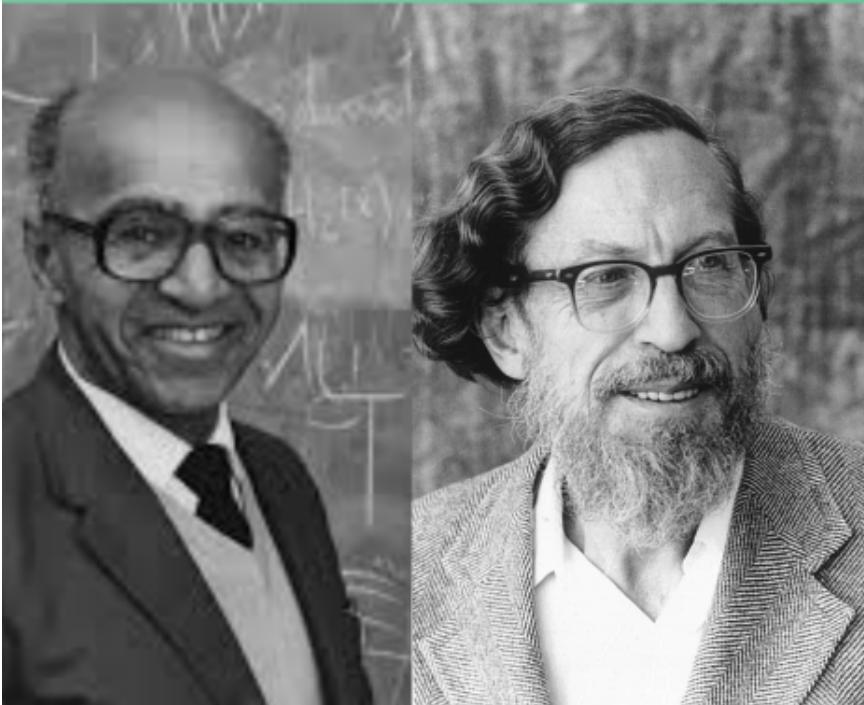


Experiment Design

Introduced by Blackwell & Stein in 1952

Single-shot Problem

- Consider a single experiment $a \in \mathcal{A}$
 - M mutually exclusive hypotheses: $H_i \Leftrightarrow \{\theta = i\}$,
 - Noisy observations subject to $\{q_i^a(\cdot)\}_{i,a}$
- What should a be? Compare experiment a with a' ?
 - Stochastically degraded case [Blackwell '53], [Stein '53]

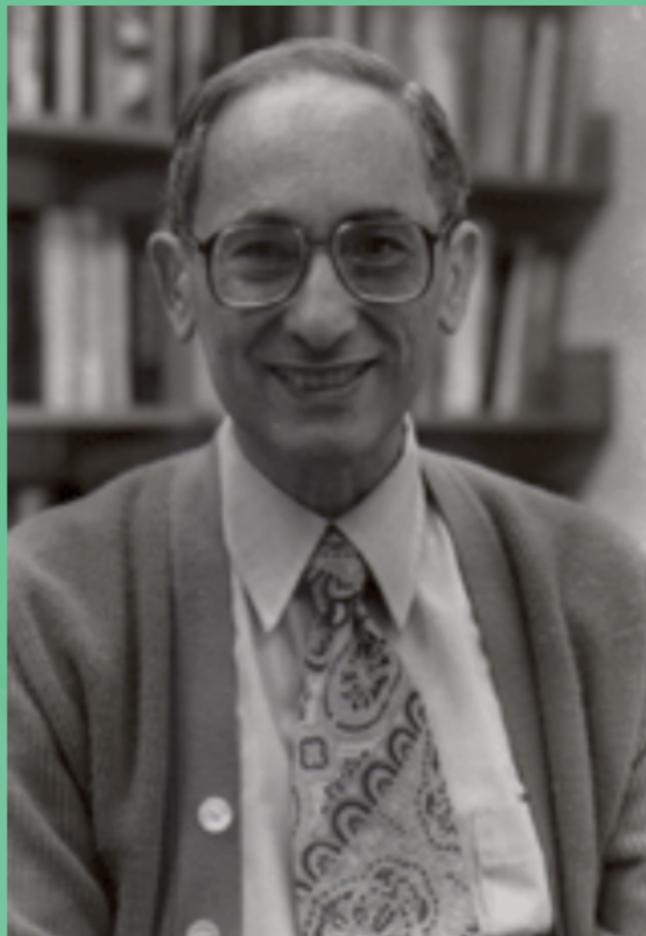


Experiment Design

Introduced by Blackwell & Stein in 1952

Single-shot Problem

- Given experiment a :
 - True hypothesis $\theta = i$ with probability ρ_i
 - Output distribution $Z^a \sim \sum_{i=1}^M \rho_i q_i^a(\cdot)$
 - Posterior upon observation $\theta|Z^a \sim \underbrace{\Phi^a(\rho, Z^a)}_{\text{Bayes operator}}$
 - How does $\Phi^a(\cdot, \cdot)$ compare with $\Phi^{a'}(\cdot, \cdot)$



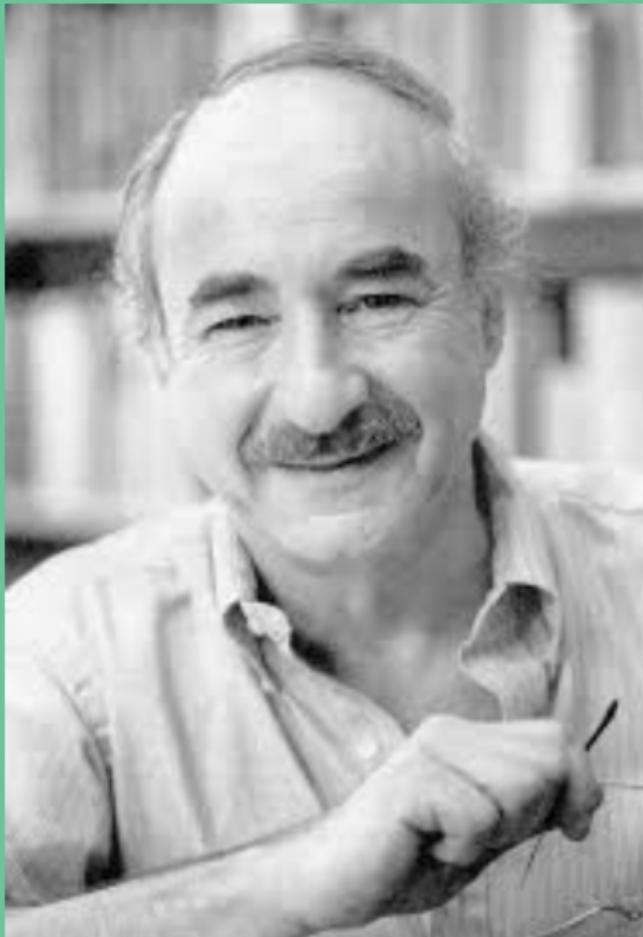
Active Hypothesis Testing

Introduced by Chernoff in 1953

- M mutually exclusive hypotheses: $H_i \Leftrightarrow \{\theta = i\}$, $i = 1, 2, \dots, M$
- Experiments $A(t) \in \mathcal{A}$ chosen sequentially

time	0	1	...	$\tau - 1$	τ
sample	$A(0)$	$A(1)$...	$A(\tau - 1)$	
observation	$Z(0)$	$Z(1)$...	$Z(\tau - 1)$	
declaration					$\hat{\theta} = d(Z^{\tau-1}, x^{\tau-1})$
error					$1_{\{\hat{\theta} \neq \theta\}}$

- $Z|_{\{\theta=i, A=a\}} \sim q_i^a(\cdot)$: observation density given $a \in \mathcal{A}$ and H_i
- Uniform Prior $\rho(0) = [\frac{1}{M} \frac{1}{M} \dots \frac{1}{M}]$



Information Utility

Introduced by De Groot in 1963

Information Utility Heuristics:

- Measure of uncertainty V [DeGroot 1962]
- Information utility associated with V

$$\mathcal{IU}(a, \rho, V) = V(\rho) - \mathbb{E}[V(\underbrace{\Phi^a(\rho, Z)}_{\text{Bayes operator}})]$$

- Most informative action $\arg \max_a \mathcal{IU}(a, \rho, V)$

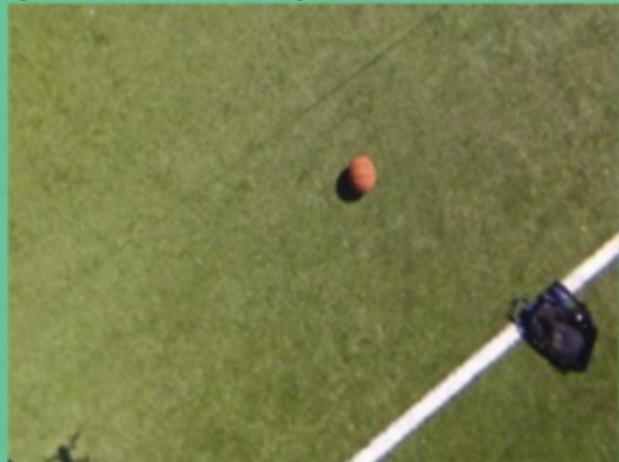
Which intuition works better?



Which intuition works better?



Asymptotic Optimal
(Chernoff's):



Information Utility
(DeGroot):



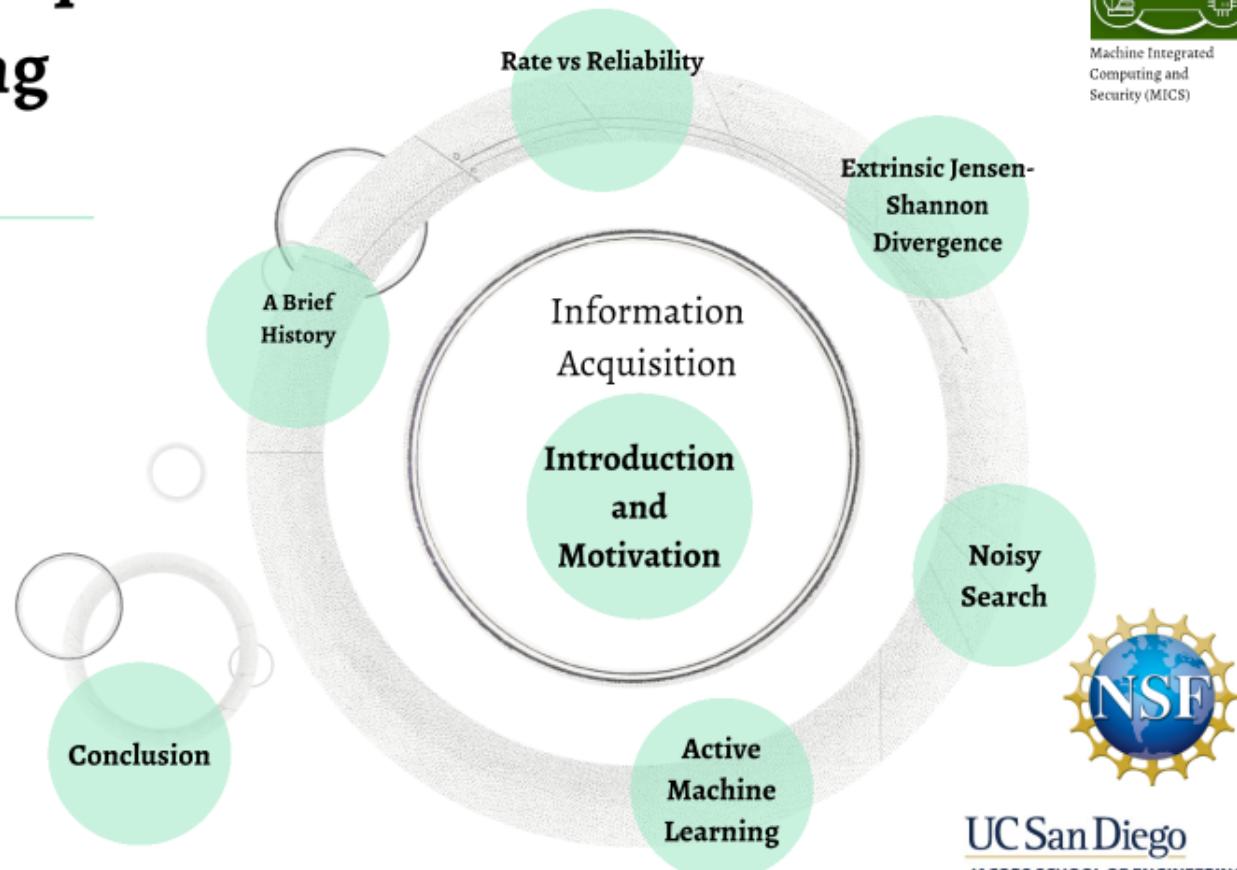
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

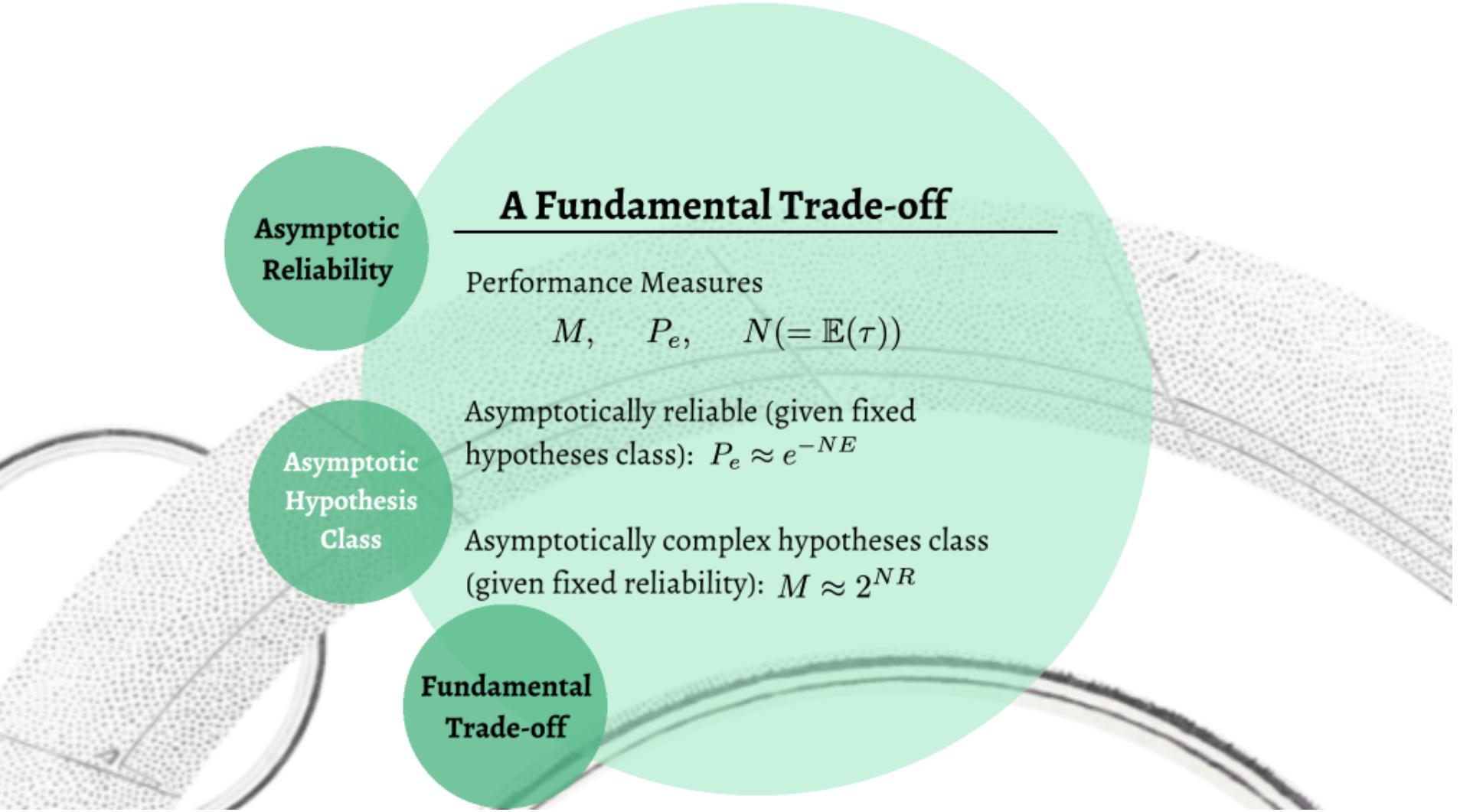
Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz



Machine Integrated
Computing and
Security (MICS)



UC San Diego
JACOBS SCHOOL OF ENGINEERING
Center for Wireless Communications



**Asymptotic
Reliability**

**Asymptotic
Hypothesis
Class**

**Fundamental
Trade-off**

A Fundamental Trade-off

Performance Measures

$$M, \quad P_e, \quad N(= \mathbb{E}(\tau))$$

Asymptotically reliable (given fixed hypotheses class): $P_e \approx e^{-NE}$

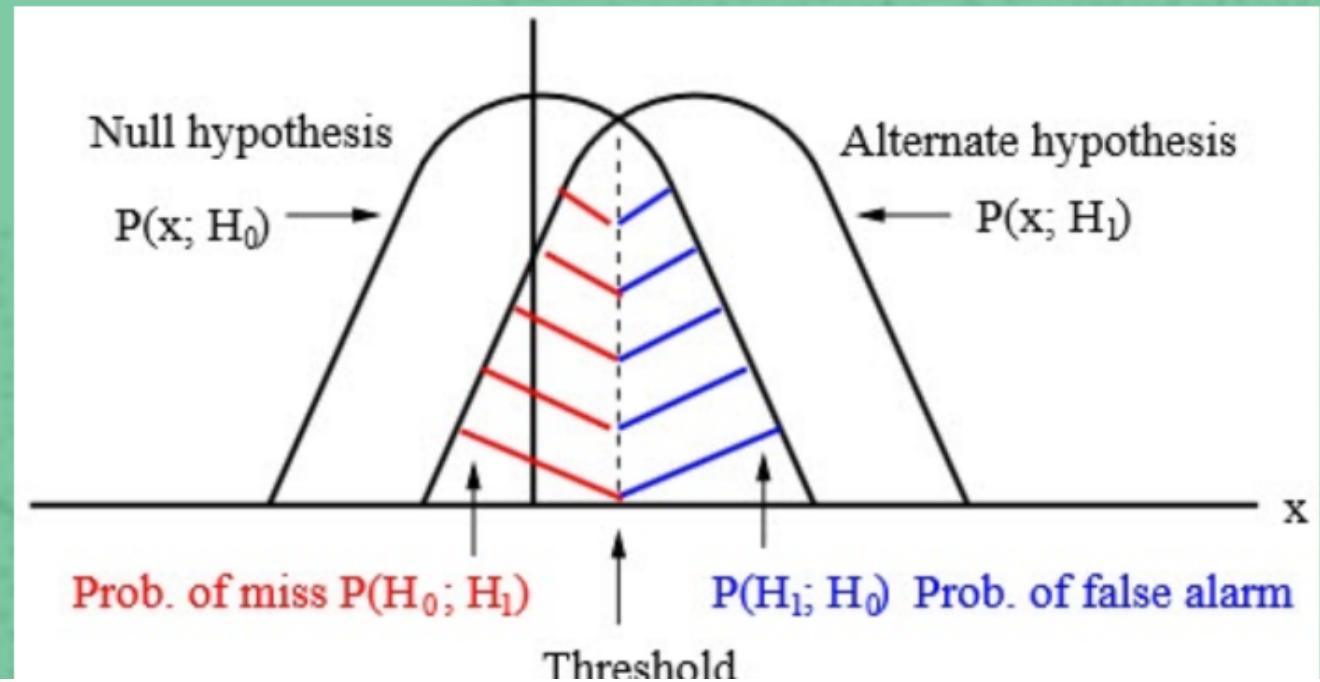
Asymptotically complex hypotheses class (given fixed reliability): $M \approx 2^{NR}$

Asymptotic reliability (Chernoff's)

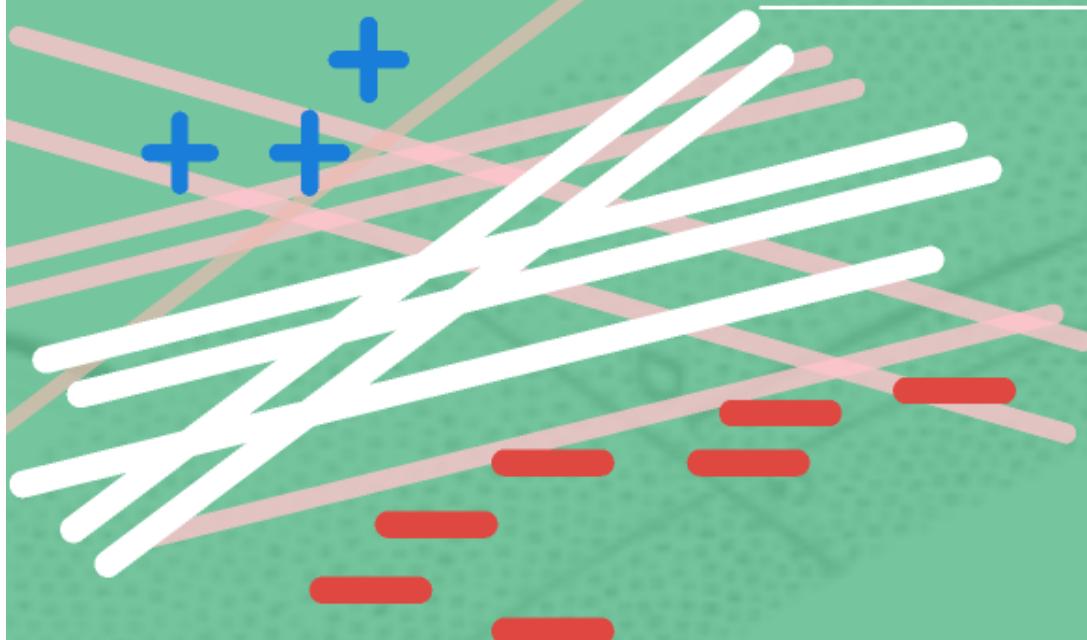
Classical Statistics

max-min statistical
discremimation

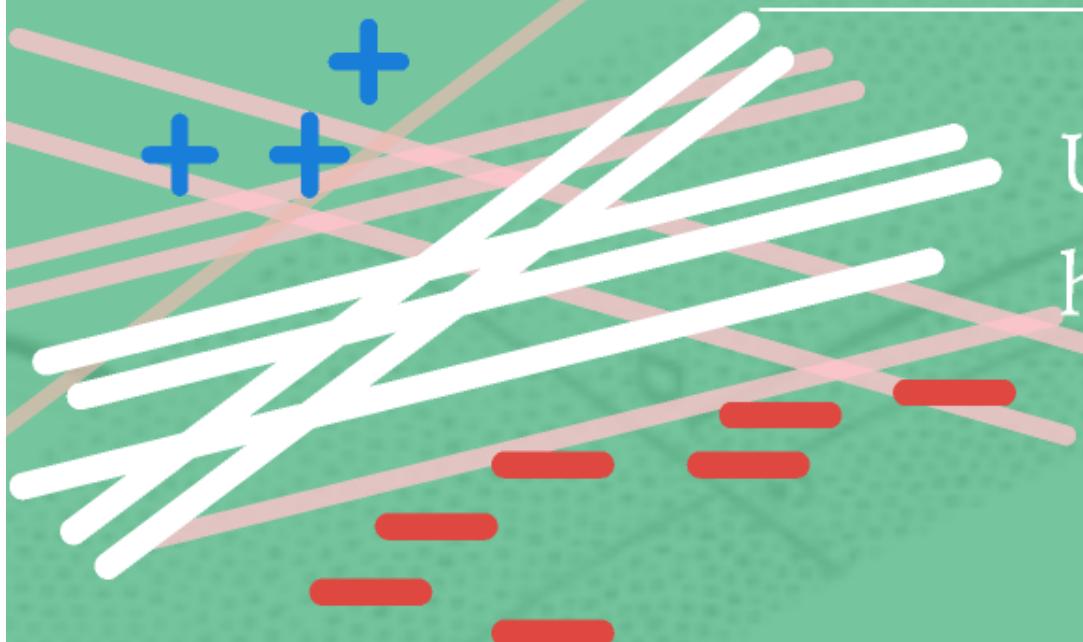
(KL-divergence)



Complex Hypothesis Class (Small Sample Regime)

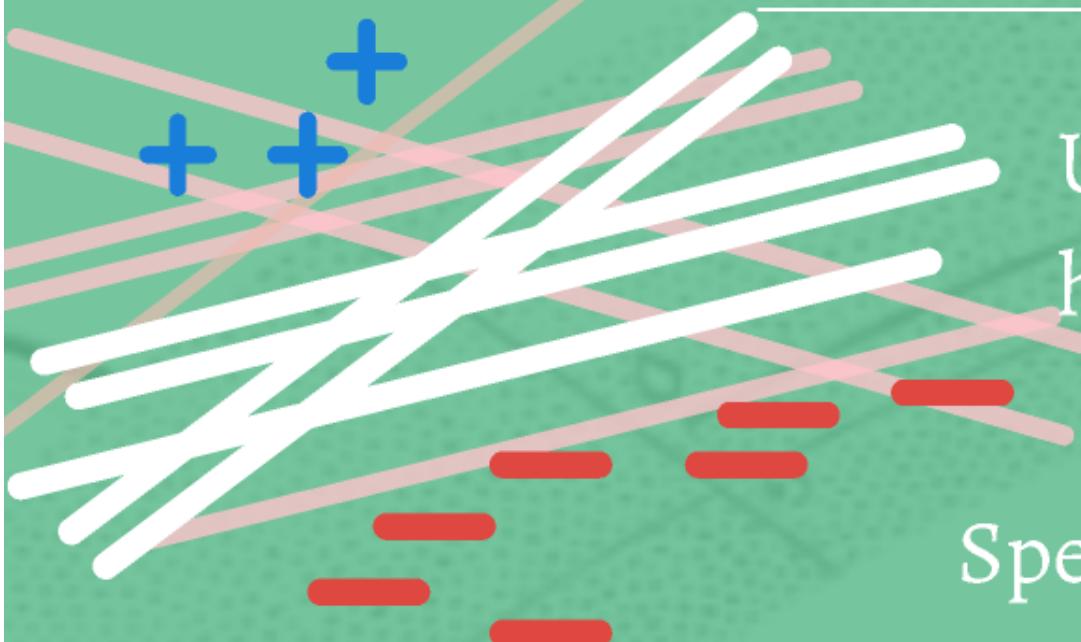


Complex Hypothesis Class (Small Sample Regime)



Uncountable
hypothesis classs

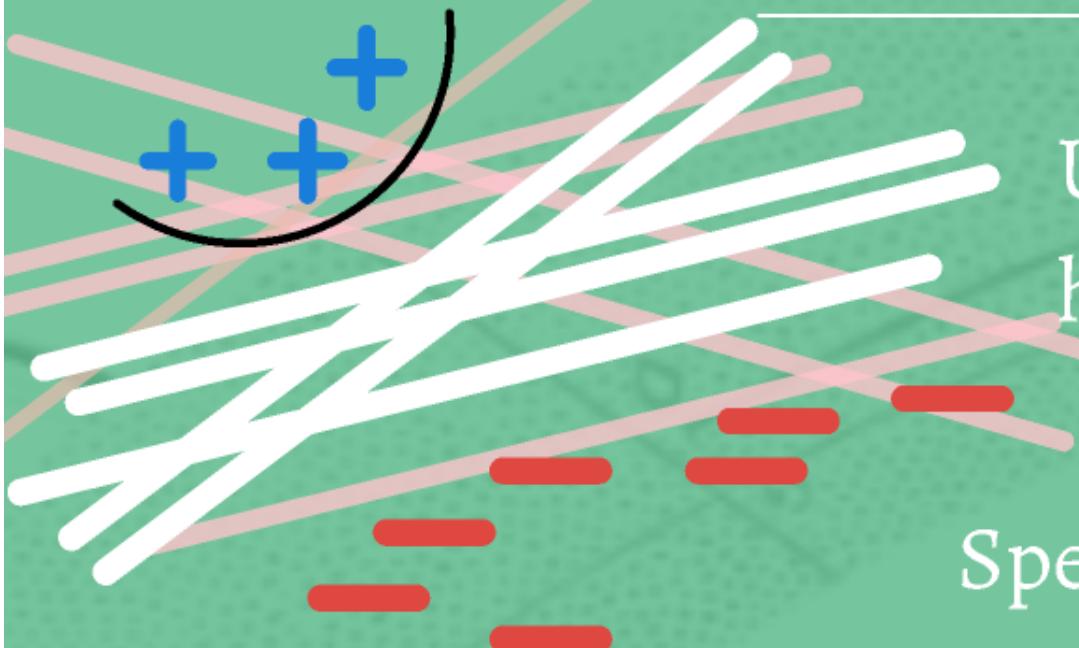
Complex Hypothesis Class (Small Sample Regime)



Uncountable
hypothesis classss

Speedy elimination

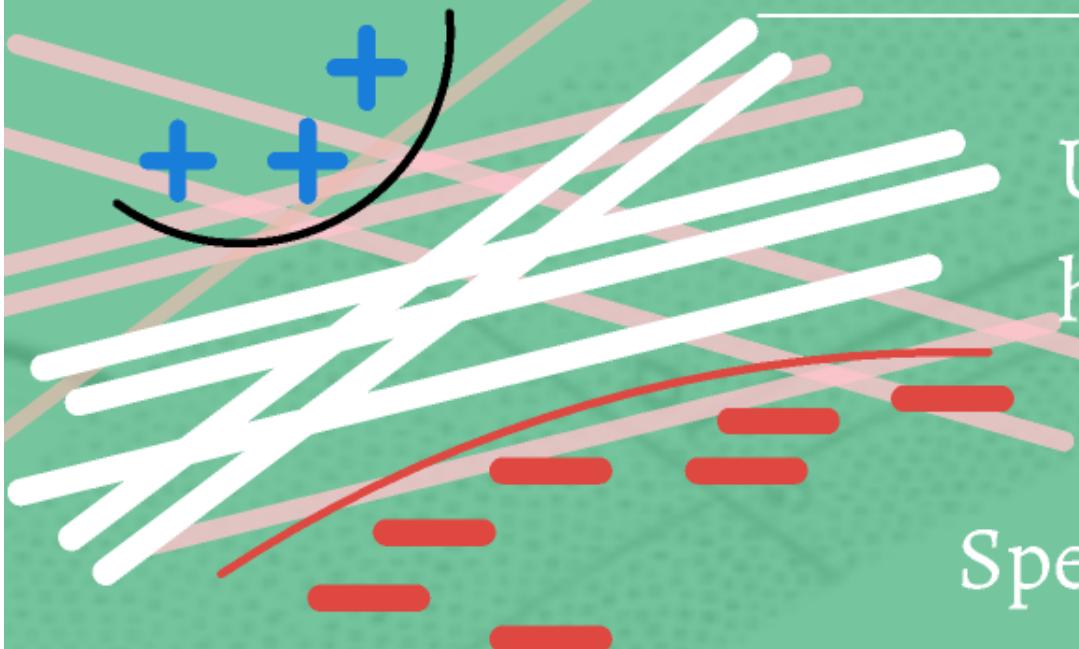
Complex Hypothesis Class (Small Sample Regime)



Uncountable
hypothesis classs

Speedy elimination

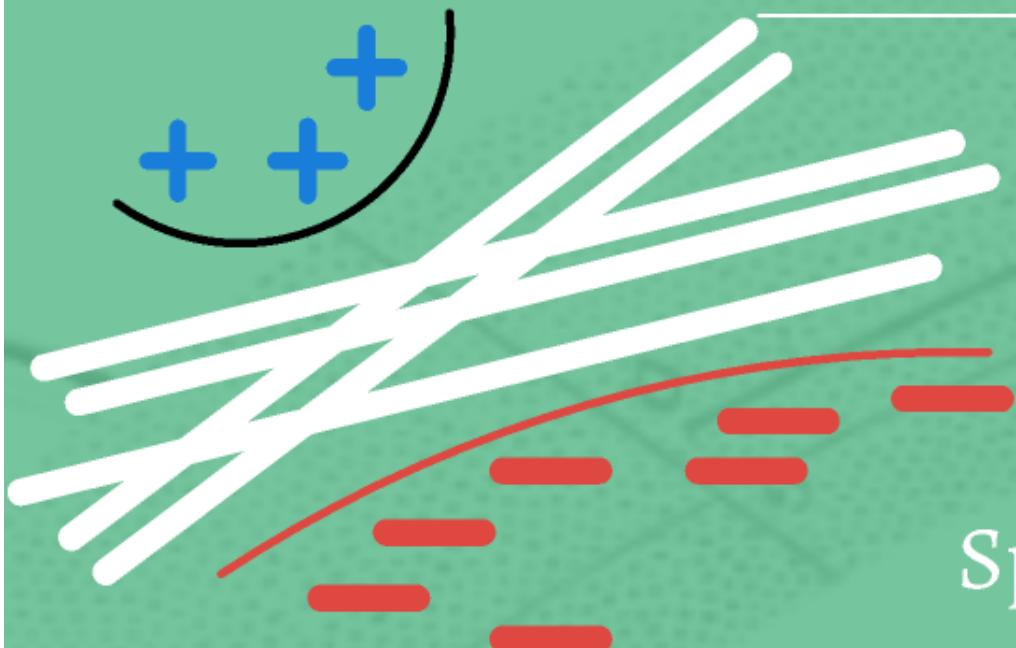
Complex Hypothesis Class (Small Sample Regime)



Uncountable
hypothesis classss

Speedy elimination

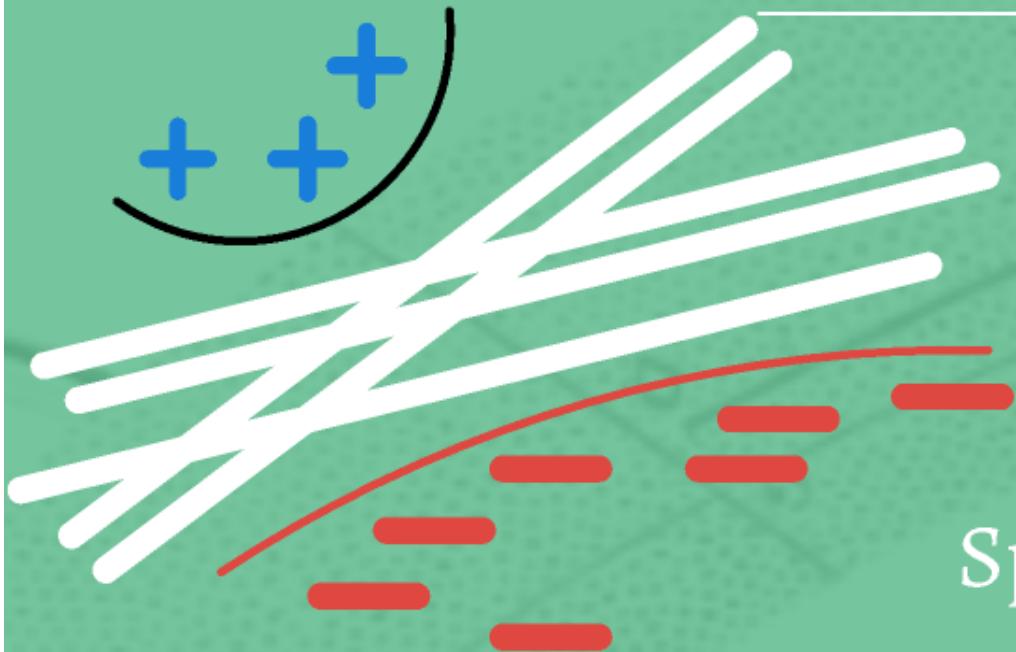
Complex Hypothesis Class (Small Sample Regime)



Uncountable
hypothesis classs

Speedy elimination

Complex Hypothesis Class (Small Sample Regime)



Uncountable
hypothesis classs

Speedy elimination

Information Acquisition Rate

Performance Measures

$$M, \quad P_e, \quad N(= \mathbb{E}(\tau))$$

Asymptotically reliable (given fixed hypotheses class): $P_e \approx e^{-NE}$

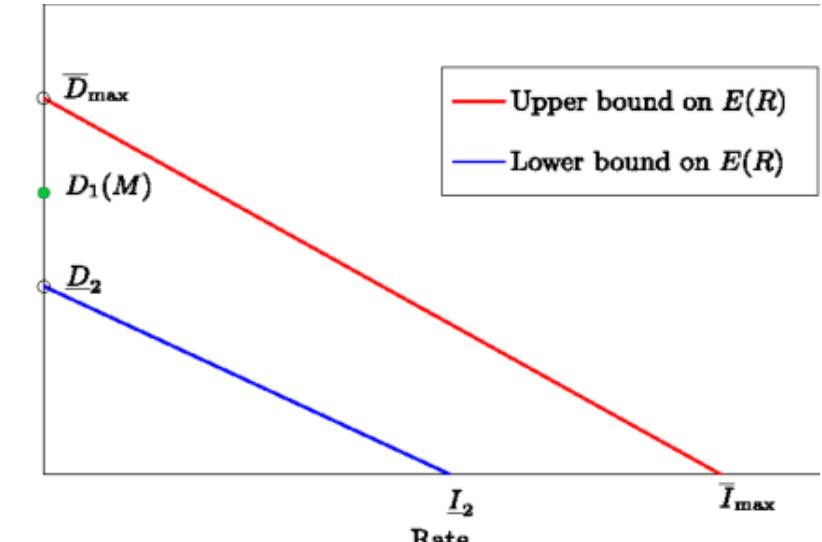
Asymptotically complex hypotheses class (given fixed reliability): $M \approx 2^{NR}$

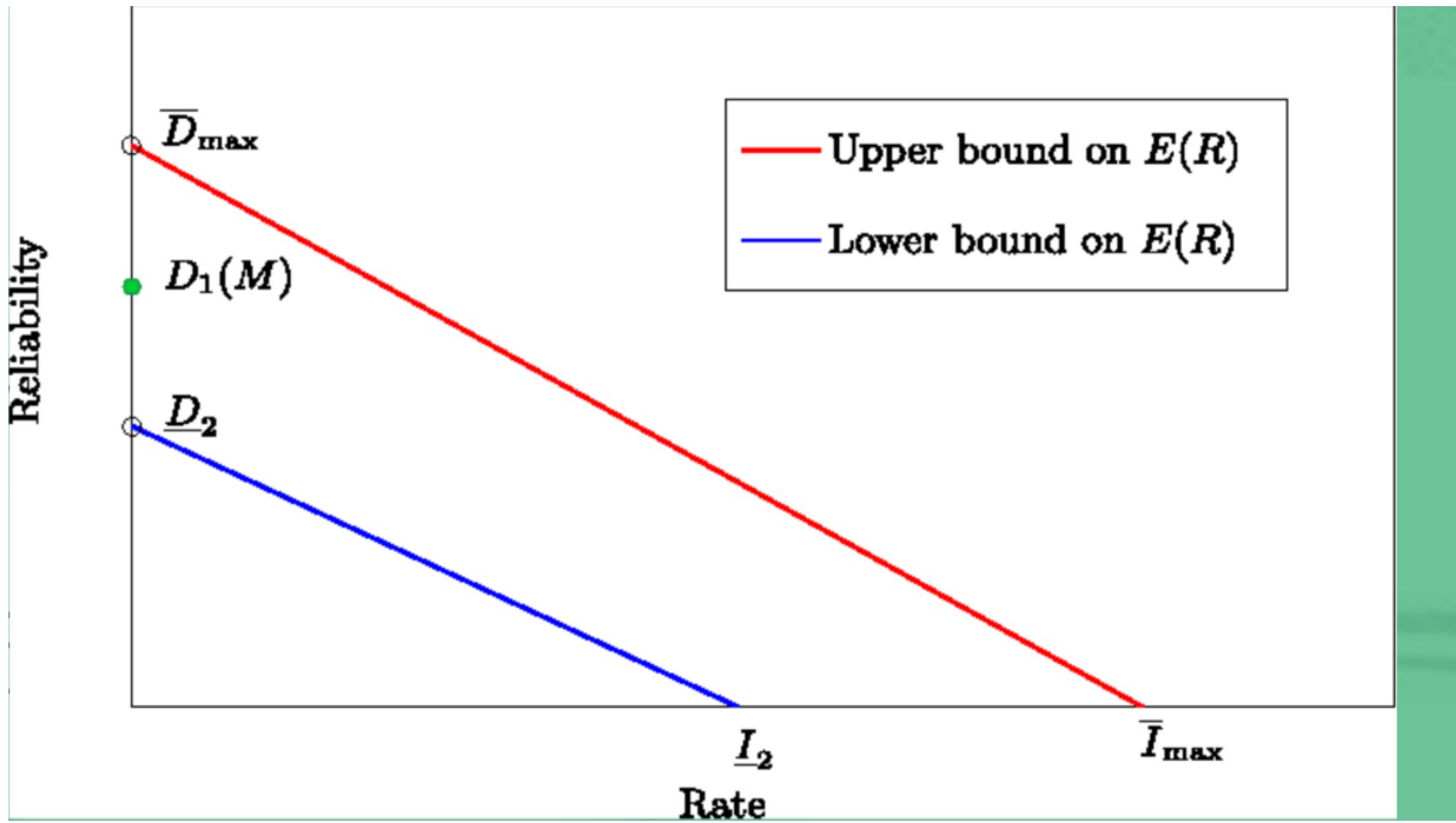
Performance Measures

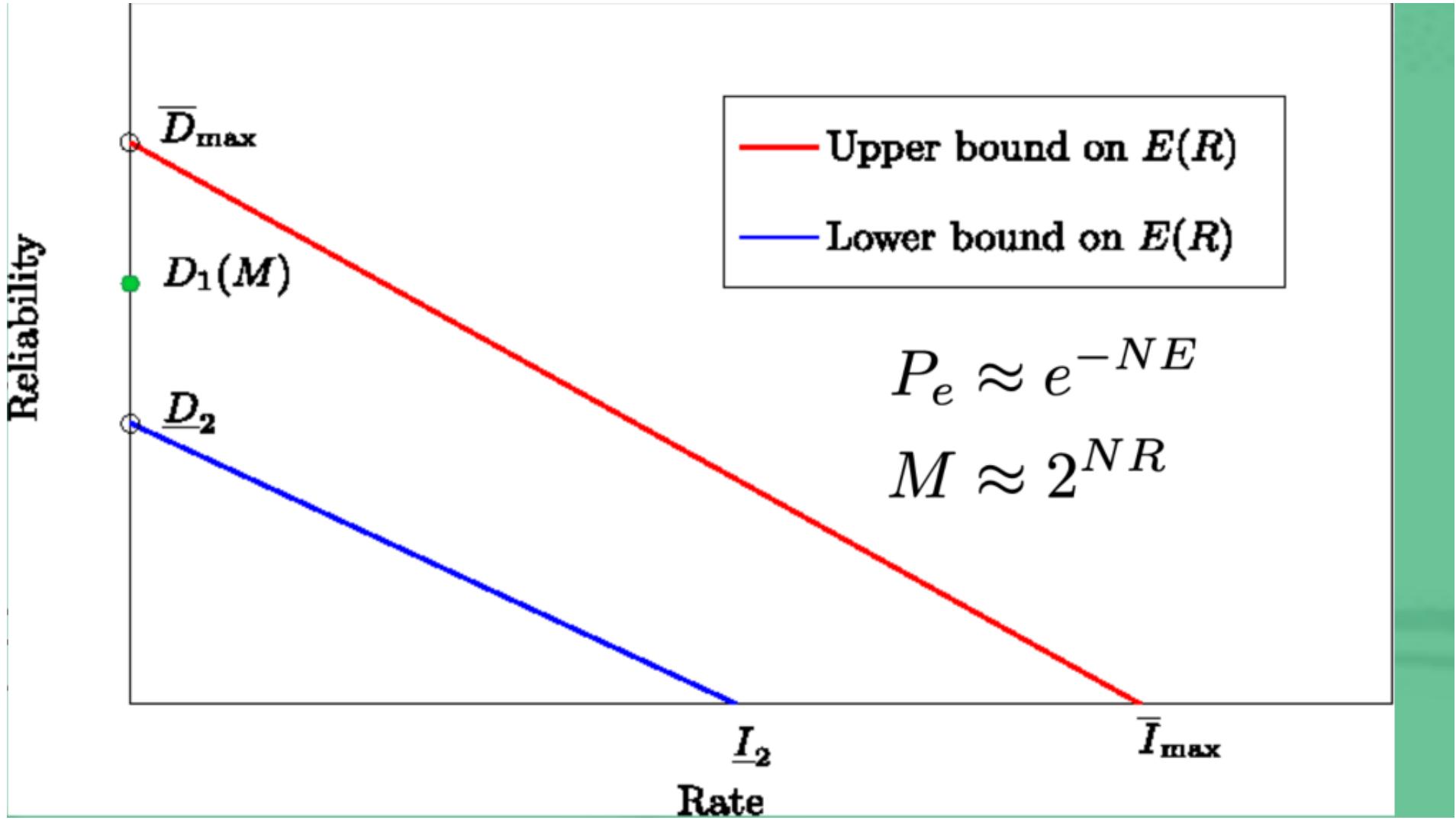
$$M, \quad P_e, \quad N(=$$

Asymptotically reliable (given fixed hypotheses class): $P_e \approx e^{-NE}$

Asymptotically complex hypotheses class
(given fixed reliability): $M \approx 2^{NR}$







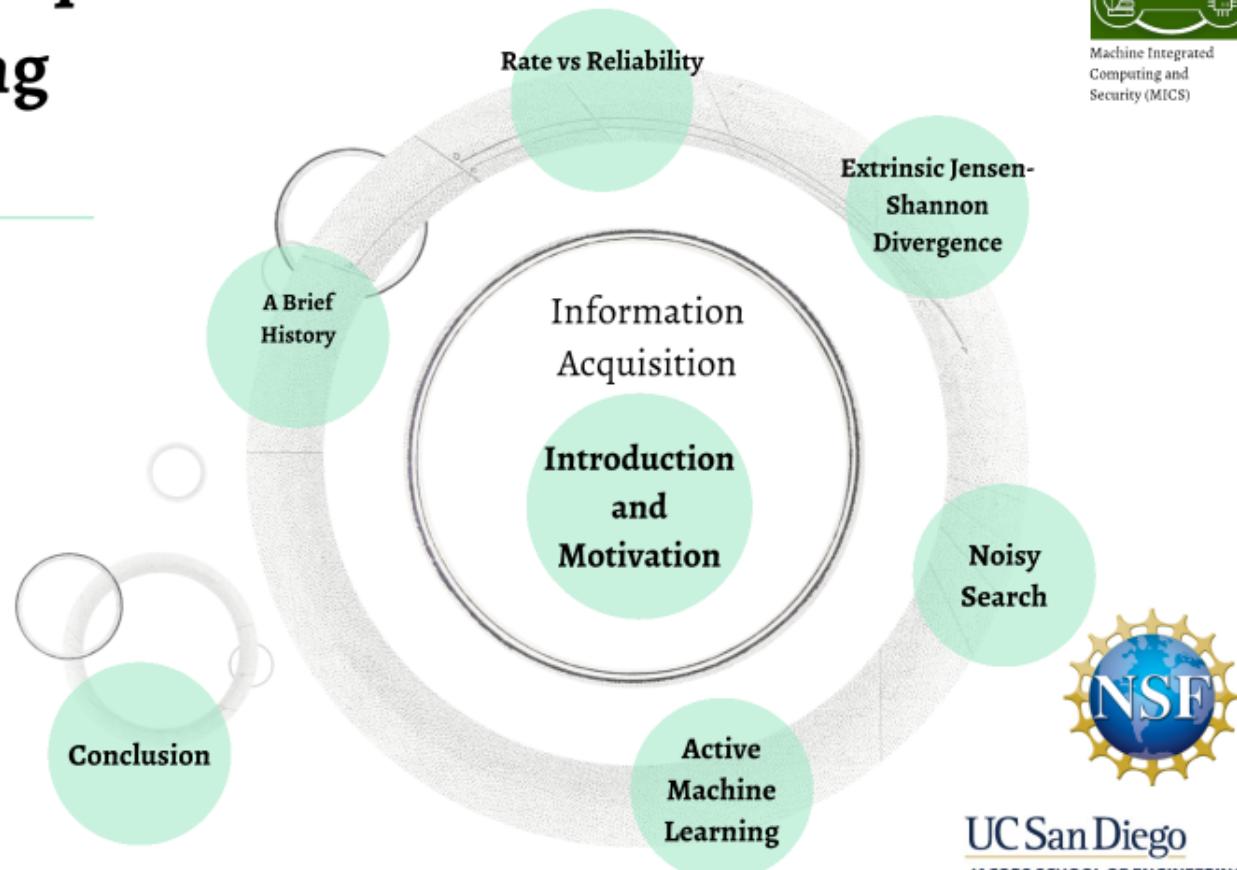
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

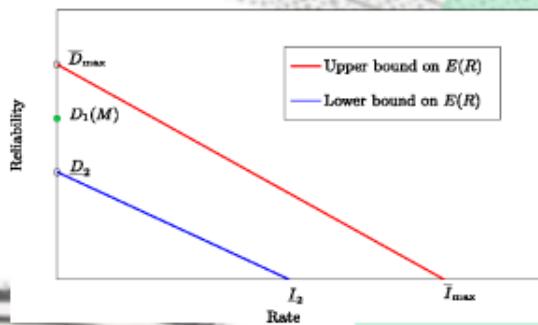
Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz



Machine Integrated
Computing and
Security (MICS)



UC San Diego
JACOBS SCHOOL OF ENGINEERING
Center for Wireless Communications



Extrinsic Jensen-Shannon

Analysis is based on geometry
of belief update (filtering)

**Dynamical
System View**

Information Utility
Divergence

Extrinsic
Jensen-Shannon

Information
Utility

Symmetrized
Divergence

Recall

Information Utility Heuristics:

- Measure of uncertainty V [DeGroot 1962]
- Information utility associated with V

$$\mathcal{IU}(a, \rho, V) = V(\rho) - \mathbb{E}[V(\Phi^a(\rho, Z))]$$

Bayes operator

- Most informative action $\arg \max_a \mathcal{IU}(a, \rho, V)$

Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized divergence” among $q_1^a, q_2^a, \dots, q_M^a$
- Best action must maximize the divergence
 - maximize discrimination among H_1, H_2, \dots, H_M

Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized divergence” among $q_1^a, q_2^a, \dots, q_M^a$
- Best action must maximize the divergence
 - maximize discrimination among H_1, H_2, \dots, H_M

Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized divergence” among $q_1^a, q_2^a, \dots, q_M^a$
- Best action must maximize the divergence
 - maximize discrimination among H_1, H_2, \dots, H_M

The Kullback-Leibler (KL) divergence between $p(\cdot)$ and $q(\cdot)$:

$$D(p||q) = \sum_z p(z) \log \frac{p(z)}{q(z)}$$

Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized divergence” among $q_1^a, q_2^a, \dots, q_M^a$
- Best action must maximize the divergence
 - maximize discrimination among H_1, H_2, \dots, H_M

The Kullback-Leibler (KL) divergence between $p(\cdot)$ and $q(\cdot)$:

$$D(p||q) = \sum_z p(z) \log \frac{p(z)}{q(z)}$$

Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized divergence” among $q_1^a, q_2^a, \dots, q_M^a$
- Best action must maximize the divergence
 - maximize discrimination among H_1, H_2, \dots, H_M

The whiteboard shows two formulas for divergence:

J-divergence [Jefferys 73]

$$D_J(f, g) = \frac{1}{2} D(f||g) + \frac{1}{2} D(g||f)$$

L-divergence [Lin 91]

$$D_L(f, g) = \frac{1}{2} D(f||\frac{f+g}{2}) + \frac{1}{2} D(g||\frac{f+g}{2})$$

The whiteboard shows the Kullback-Leibler (KL) divergence formula:

The Kullback-Leibler (KL) divergence between $p(\cdot)$ and $q(\cdot)$:

$$D(p||q) = \sum_z p(z) \log \frac{p(z)}{q(z)}$$

Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized” divergence: $I(\theta; Z^a) = \sum_{i=1}^M \rho_i D(q_i^a || \sum_{i=1}^M \rho_i q_i^a)$
- Best action must maximize:
 - Jensen-Shannon divergence [Lin 1991]
 - Generalizing L divergence: $D_L(f, g) = \frac{1}{2}D(f||\frac{f+g}{2}) + \frac{1}{2}D(g||\frac{f+g}{2})$

J-divergence [Jefferys 73]

L-divergence [Lin 91]

$$D_J(f, g) = \frac{1}{2}D(f||g) + \frac{1}{2}D(g||f)$$

$$D_L(f, g) = \frac{1}{2}D(f||\frac{f+g}{2}) + \frac{1}{2}D(g||\frac{f+g}{2})$$

The Kullback-Leibler (KL) divergence between $p(\cdot)$ and $q(\cdot)$:
$$D(p||q) = \sum_z p(z) \log \frac{p(z)}{q(z)}$$

Proposed Divergence

Extrinsic Jensen-Shannon Divergence [Naghshvar, J. ISIT'12]

The *Extrinsic Jensen-Shannon (EJS) divergence* among densities q_1, q_2, \dots, q_M with respect to $\rho = [\rho_1, \rho_2, \dots, \rho_M]$ is defined as

$$EJS(\rho; q_1, q_2, \dots, q_M) = \sum_{i=1}^M \rho_i D(q_i || \sum_{k \neq i} \frac{\rho_k}{1-\rho_i} q_k).$$

Proposed Divergence

Extrinsic Jensen-Shannon Divergence [Naghshvar, J. ISIT'12]

$$I(\theta; Z^a) = \sum_{i=1}^M \rho_i D(q_i^a || \sum_{i=1}^M \rho_i q_i^a)$$

Jensen-Shannon divergence [Lin 1991]

Generalizing L divergence: $D_L(f, g) = \frac{1}{2}D(f||\frac{f+g}{2}) + \frac{1}{2}D(g||\frac{f+g}{2})$

'trinsic Jensen-Shannon (EJS) divergence among densities q_1, q_2, \dots, q_M with respect to $\rho = [\rho_1, \rho_2, \dots, \rho_M]$ is defined as

$$EJS(\rho; q_1, q_2, \dots, q_M) = \sum_{i=1}^M \rho_i D(q_i || \sum_{k \neq i} \frac{\rho_k}{1-\rho_i} q_k).$$

Proposed Divergence

Extrinsic Jensen-Shannon Divergence [Naghshvar, J. ISIT'12]

$$I(\theta; Z^a) = \sum_{i=1}^M \rho_i D(q_i^a || \sum_{i=1}^M \rho_i q_i^a)$$

nsen-Shannon divergence [Lin 1991]

Generalizing L divergence: $D_L(f, g) = \frac{1}{2}D(f||\frac{f+g}{2}) + \frac{1}{2}D(g||\frac{f+g}{2})$ Extrinsic Jensen-Shannon (EJS) divergence among densities q_1, q_2, \dots, q_M with respect to $\rho = [\rho_1, \rho_2, \dots, \rho_M]$ is defined as

$$EJS(\rho; q_1, q_2, \dots, q_M) = \sum_{i=1}^M \rho_i D(q_i || \sum_{k \neq i} \frac{\rho_k}{1-\rho_i} q_k).$$

Proposition

EJS is the information utility associated with the average likelihood function $U(\rho) = \sum_{i=1}^M \rho_i \log \frac{1-\rho_i}{\rho_i}$, i.e.

$$EJS(\rho; q_1^a, \dots, q_M^a) = \mathcal{IU}(a, \rho, U)$$

Proposed Divergence

Extrinsic Jensen-Shannon Divergence [Naghshvar, J. ISIT'12]

$$I(\theta; Z^a) = \sum_{i=1}^M \rho_i D(q_i^a || \sum_{i=1}^M \rho_i q_i^a)$$

nsen-Shannon divergence [Lin 1991]

Generalizing L divergence: $D_L(f, g) = \frac{1}{2}D(f||\frac{f+g}{2}) + \frac{1}{2}D(g||\frac{f+g}{2})$ Extrinsic Jensen-Shannon (EJS) divergence among densities q_1, q_2, \dots, q_M with respect to $\rho = [\rho_1, \rho_2, \dots, \rho_M]$ is defined as

$$EJS(\rho; q_1, q_2, \dots, q_M) = \sum_{i=1}^M \rho_i D(q_i || \sum_{k \neq i} \frac{\rho_k}{1-\rho_i} q_k).$$

$$\begin{aligned} I(\theta; Z^a) &= H(\rho) - \mathbb{E}(H(\Phi^a(\rho, Z^a))) \\ &= \mathcal{IU}(a, \rho, H) \end{aligned}$$

Proposition

EJS is the information utility associated with the average likelihood function $U(\rho) = \sum_{i=1}^M \rho_i \log \frac{1-\rho_i}{\rho_i}$, i.e.

$$EJS(\rho; q_1^a, \dots, q_M^a) = \mathcal{IU}(a, \rho, U)$$

Dynamical System View

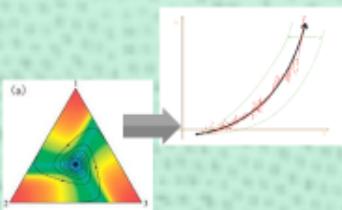
Flattened geometry of Information Utility
e.g. in hypothesis testing:

- Consider (suboptimal) $\tau = \min\{t : \max_i \rho_i(t) \geq 1 - \epsilon\}$
 - Stop search and declare $\hat{\theta} = i$ if $\rho_i(t) \geq 1 - \epsilon$ (satisfies $\text{Pe} \leq \epsilon$)

instead work with:

- Take concave functional U (bounded $|U(\rho(t+1)) - U(\rho(t))| \leq \Delta$)

Lyapunov-type
Analysis



Dynamical System View

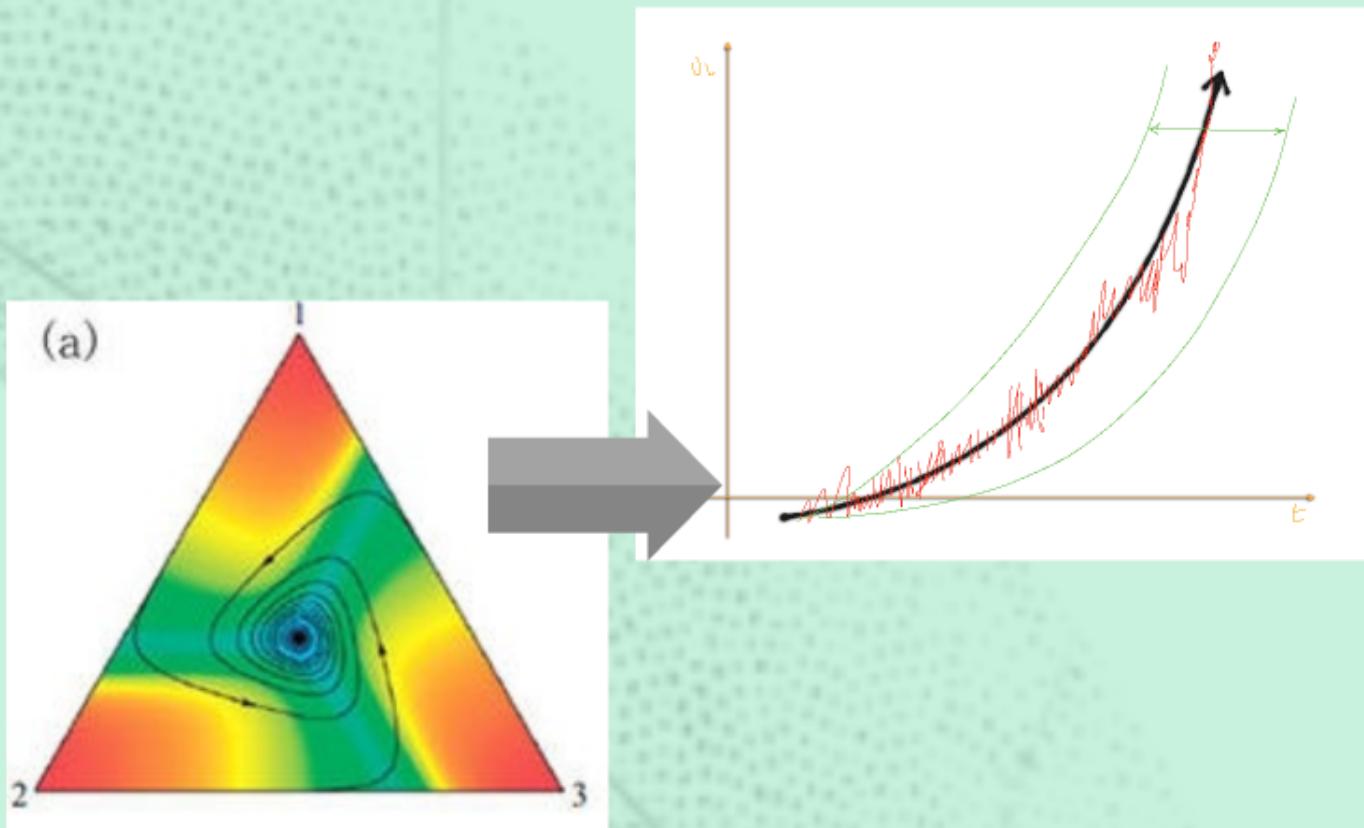
Flattened geometry of Information Utility
e.g. in hypothesis testing:

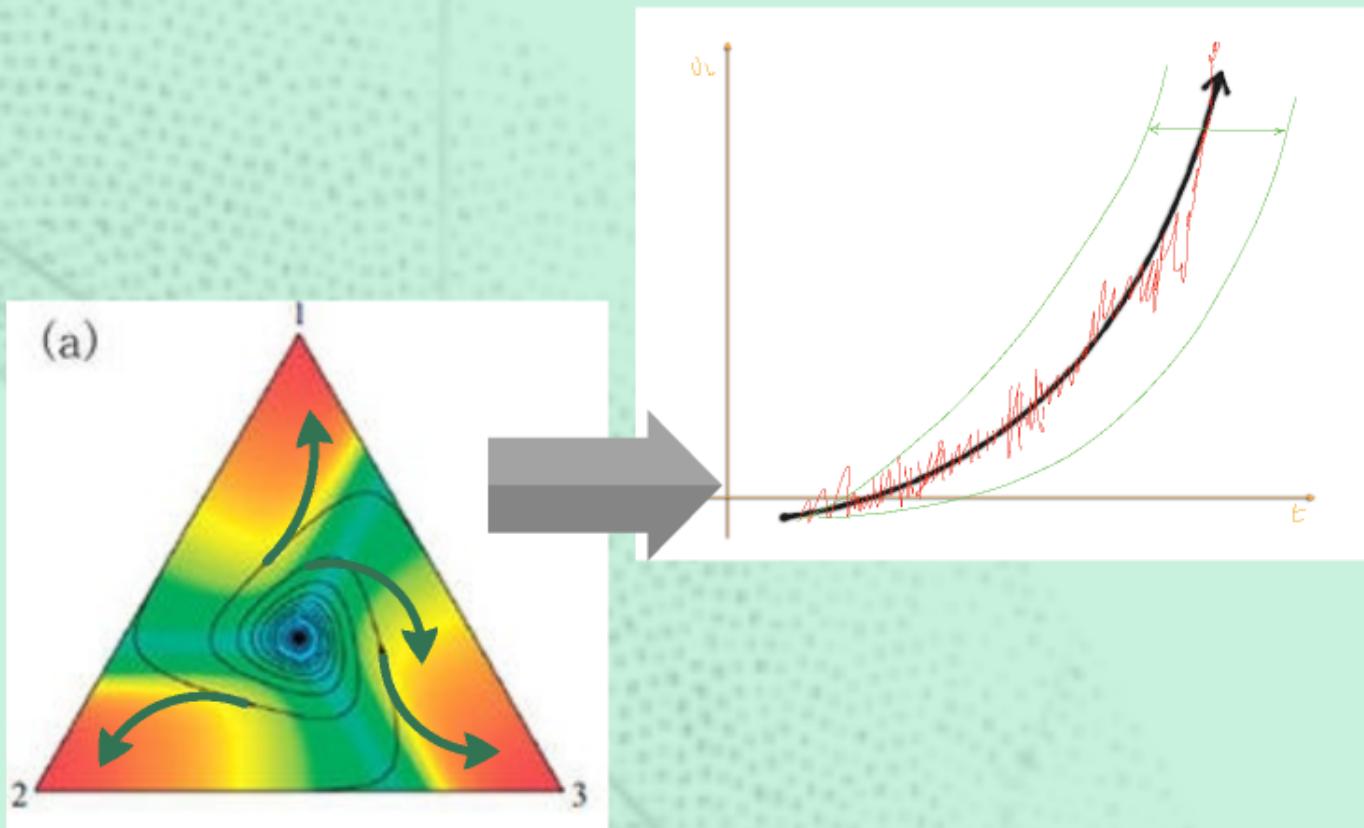
- Consider (suboptimal) $\tau = \min\{t : \max_i \rho_i(t) \geq 1 - \epsilon\}$
 - Stop search and declare $\hat{\theta} = i$ if $\rho_i(t) \geq 1 - \epsilon$ (satisfies $\text{Pe} \leq \epsilon$)

instead work with:

- Take concave functional U (bounded $|U(\rho(t+1)) - U(\rho(t))| \leq \Delta$)

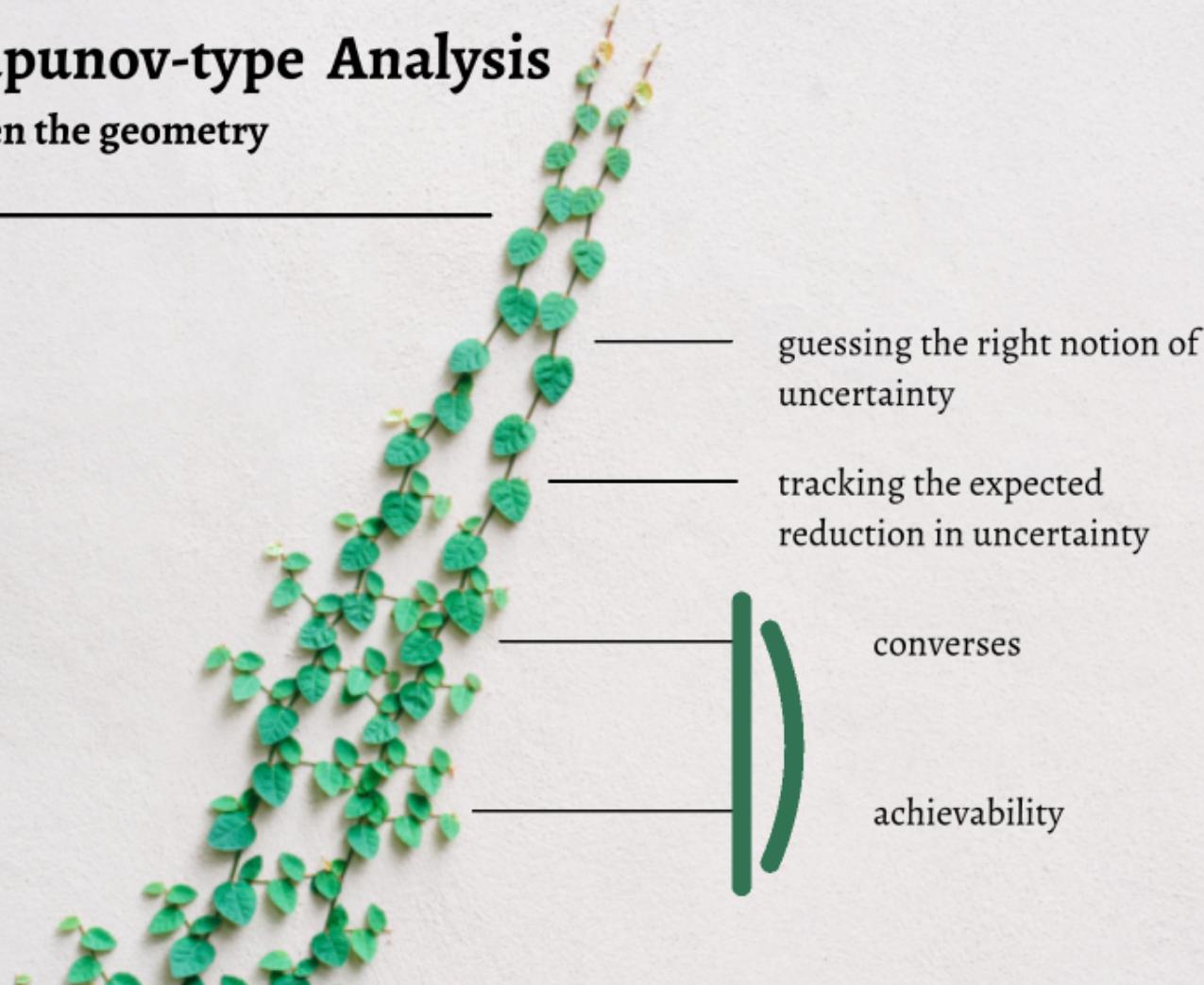
**Lyapunov-type
Analysis**





Lyapunov-type Analysis

flatten the geometry



Lyapunov-type Analysis

flatten the geometry

Information Utility Based Analysis: Converse

- Consider (suboptimal) $\tau = \min\{t : \max_i \rho_i(t) \geq 1 - \epsilon\}$
 - Stop search and declare $\hat{\theta} = i$ if $\rho_i(t) \geq 1 - \epsilon$ (satisfies $P_e \leq \epsilon$)
- Take concave function U (bounded $|U(\rho(t+1)) - U(\rho(t))| \leq \Delta$).
- Suppose for any policy c selecting action a

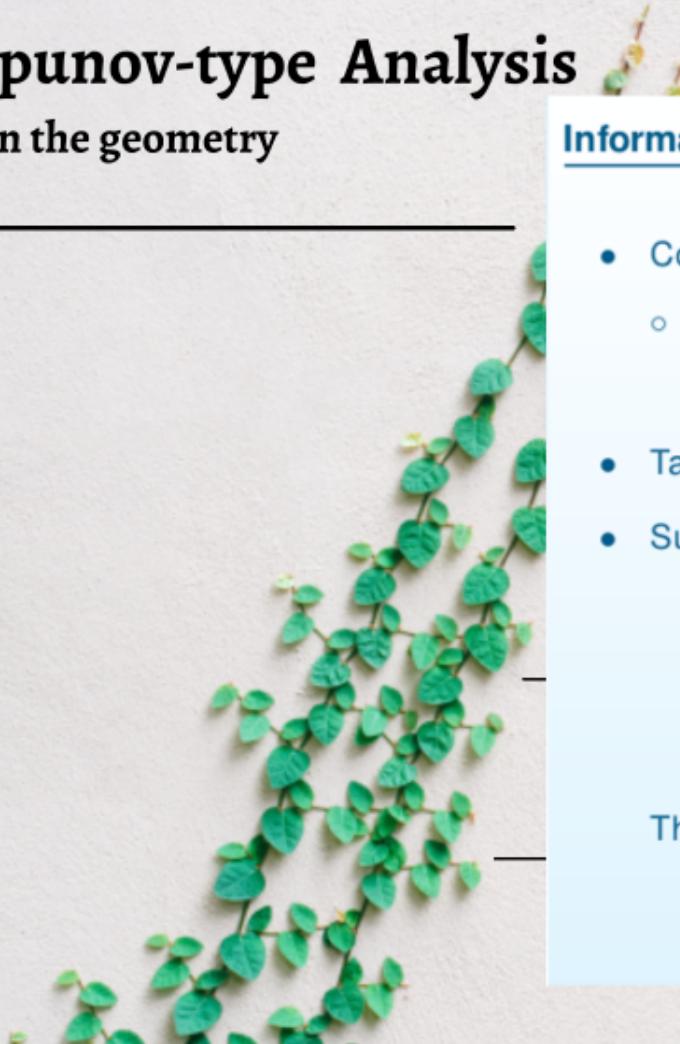
$$\mathcal{I}(a, \rho, U) \leq \bar{\alpha}, \text{ for some positive } \bar{\alpha}.$$

Then,

$$\mathbb{E}[\tau^*] \gtrsim \frac{U(\rho) - U([1 - \epsilon, \epsilon])}{\bar{\alpha}} + \frac{\Delta}{\bar{\alpha}}.$$

Lyapunov-type Analysis

flatten the geometry



Information Utility Based Analysis: Achievability

- Consider (suboptimal) $\tau = \min\{t : \max_i \rho_i(t) \geq 1 - \epsilon\}$
 - Stop search and declare $\hat{\theta} = i$ if $\rho_i(t) \geq 1 - \epsilon$ (satisfies $\text{Pe} \leq \epsilon$)
- Take concave function W (bounded $|W(\rho(t+1)) - W(\rho(t))| \leq \Delta$).
- Suppose policy c selects action a such that

$$\mathcal{I}(a, \rho, W) \geq \underline{\alpha}, \text{ for some positive } \underline{\alpha}.$$

Then,

$$\mathbb{E}[\tau^*] \lesssim \frac{W(\rho) - W([1 - \epsilon, \epsilon])}{\underline{\alpha}} + \frac{\Delta}{\underline{\alpha}}.$$

Lyapunov-type Analysis

flatten the geometry

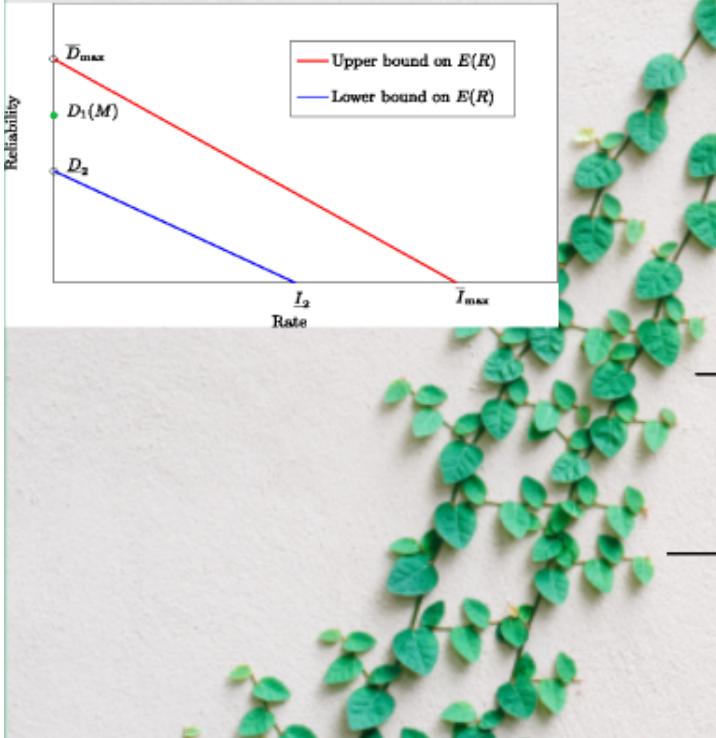
Information Utility Based Analysis: Achievability

- Consider (suboptimal) $\tau = \min\{t : \max_i \rho_i(t) \geq 1 - \epsilon\}$
 - Stop search and declare $\hat{\theta} = i$ if $\rho_i(t) \geq 1 - \epsilon$ (satisfies $\text{Pe} \leq \epsilon$)
- Take concave function W (bounded $|W(\rho(t+1)) - W(\rho(t))| \leq \Delta$).
- Suppose policy c selects action a such that

$$\mathcal{I}(a, \rho, W) \geq \underline{\alpha}, \text{ for some positive } \underline{\alpha}.$$

Then,

$$\mathbb{E}[\tau^*] \lesssim \frac{W(\rho) - W([1 - \epsilon, \epsilon])}{\underline{\alpha}} + \frac{\Delta}{\underline{\alpha}}.$$



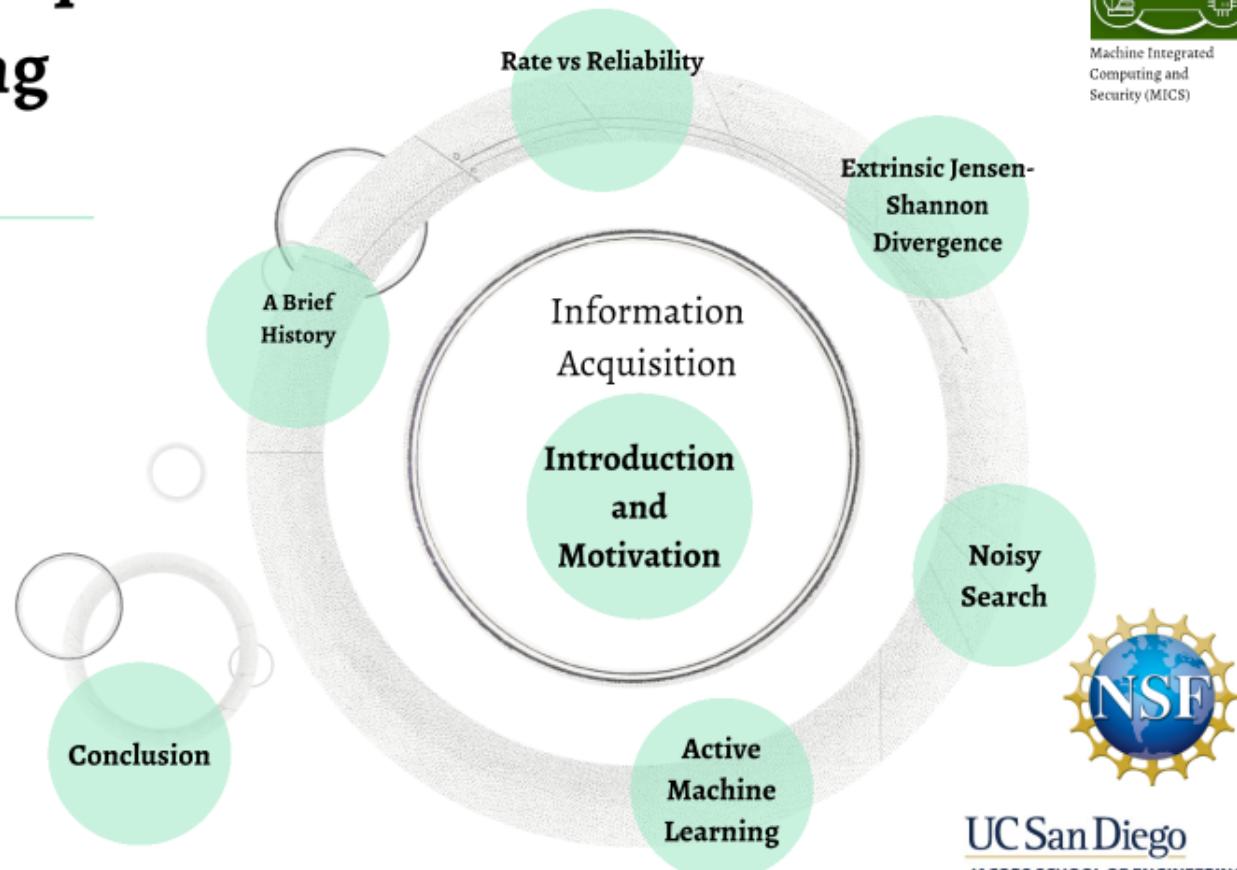
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz



Machine Integrated
Computing and
Security (MICS)



UC San Diego
JACOBS SCHOOL OF ENGINEERING
Center for Wireless Communications

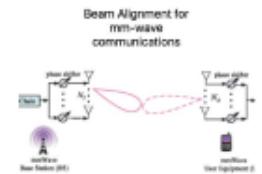
Case Study I

Measurement-dependent Noisy Search

Motivation: UAVs for object search in rescue/survey

Challenge: Trade-off between elevation and coverage

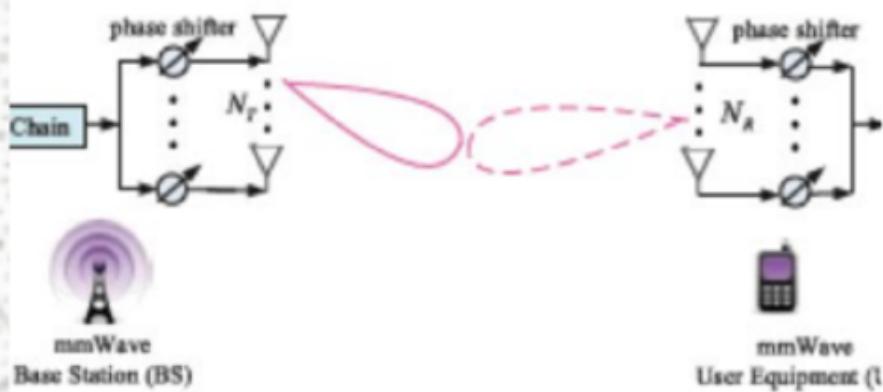
- CV-based classifiers (convolutional neural net) fail when flying high
- Flying close inherently inefficient especially in large search scenarios



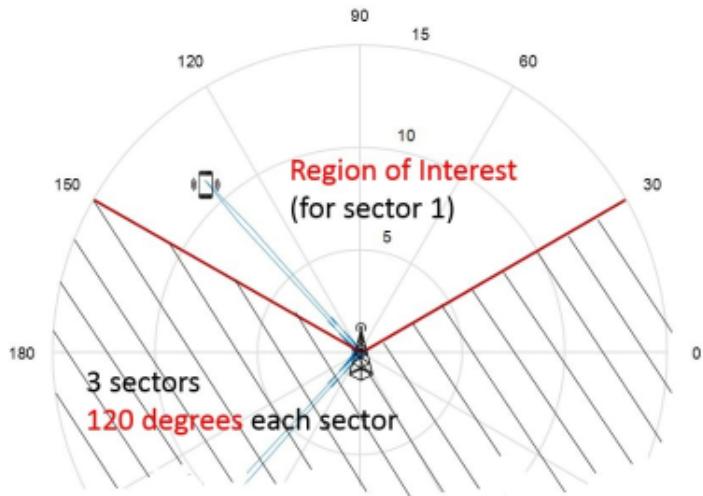
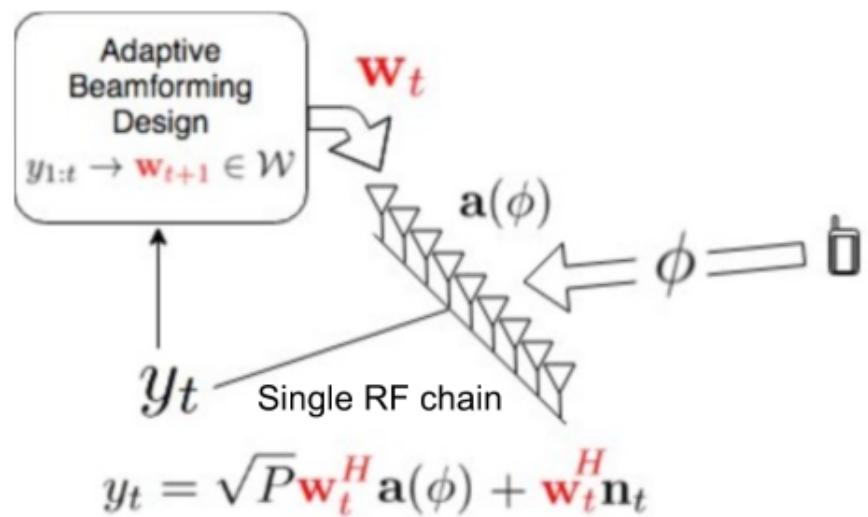
Multi-Resolution Adaptive

In Practice...

Beam Alignment for mm-wave communications



Measurement-Dependent Noisy Search



$$\mathbf{a}(\phi) := \alpha [1, e^{j \frac{2\pi d}{\lambda} \sin \phi}, \dots, e^{j(N-1) \frac{2\pi d}{\lambda} \sin \phi}]$$

Problem
Formulation

Multi-Resolution
Adaptive

In Practice...

Case Study I

Measurement-dependent Noisy Search

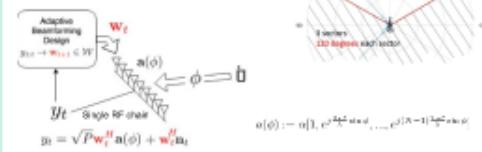
Motivation: UAVs for object search in rescue/survey

Challenge: Trade-off between elevation and coverage

- CV-based classifiers (convolutional neural net) fail when flying high
- Flying close inherently inefficient especially in large search scenarios



Measurement-Dependent Noisy Search



Measurement-Dependent Noisy Search

- Unknown parameter in an interval size B with resolution δ , $\theta \in \{0,1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$
 - Noise variance increases w/ $|A(t)|$ e.g. $Y(t) = A(t)(\theta + \mathbf{Z}_t)$, $\mathbf{Z}_t \sim \mathcal{N}(0, \delta\sigma I)$

time	1	...	$\tau - 1$	τ
sample	$A(1)$...	$A(\tau - 1)$	
observation	$Y(1)$...	$Y(\tau - 1)$	
declaration				$\hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1})$
error				$\mathbf{1}_{\{\hat{\theta} \neq \theta\}}$

Measurement-Dependent Noisy Search

- Unknown parameter in an interval size B with resolution δ , $\theta \in \{0,1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$
 - Noise variance increases w/ $|A(t)|$ e.g. $Y(t) = A(t)(\theta + \mathbf{Z}_t)$, $\mathbf{Z}_t \sim \mathcal{N}(0, \delta\sigma I)$

time	1	...	$\tau - 1$	τ
sample	$A(1)$...	$A(\tau - 1)$	
observation	$Y(1)$...	$Y(\tau - 1)$	
declaration				$\hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1})$
error				$\mathbf{1}_{\{\hat{\theta} \neq \theta\}}$

Measurement-Dependent Noisy Search

- Unknown parameter in an interval size B with resolution δ , $\theta \in \{0,1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$
 - Noise variance increases w/ $|A(t)|$ e.g. $Y(t) = A(t)(\theta + \mathbf{Z}_t)$, $\mathbf{Z}_t \sim \mathcal{N}(0, \delta\sigma I)$

time	1	...	$\tau - 1$	τ
sample	$A(1)$...	$A(\tau - 1)$	
observation	$Y(1)$...	$Y(\tau - 1)$	
declaration				$\hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1})$
error				$\mathbf{1}_{\{\hat{\theta} \neq \theta\}}$

Measurement-Dependent Noisy Search

- Unknown parameter in an interval size B with resolution δ , $\theta \in \{0,1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$
 - Noise variance increases w/ $|A(t)|$ e.g. $Y(t) = A(t)(\theta + \mathbf{Z}_t)$, $\mathbf{Z}_t \sim \mathcal{N}(0, \delta\sigma I)$

time	1	...	$\tau - 1$	τ
sample	$A(1)$...	$A(\tau - 1)$	
observation	$Y(1)$...	$Y(\tau - 1)$	
declaration				$\hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1})$
error				$\mathbf{1}_{\{\hat{\theta} \neq \theta\}}$

Measurement-Dependent Noisy Search

- Unknown parameter in an interval size B with resolution δ , $\theta \in \{0,1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$
 - Noise variance increases w/ $|A(t)|$ e.g. $Y(t) = A(t)(\theta + \mathbf{Z}_t)$, $\mathbf{Z}_t \sim \mathcal{N}(0, \delta\sigma I)$

time	1	...	$\tau - 1$	τ
sample	$A(1)$...	$A(\tau - 1)$	
observation	$Y(1)$...	$Y(\tau - 1)$	
declaration				$\hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1})$
error				$\mathbf{1}_{\{\hat{\theta} \neq \theta\}}$

Measurement-Dependent Noisy Search

- Unknown parameter in an interval size B with resolution δ , $\theta \in \{0,1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$
 - Noise variance increases w/ $|A(t)|$ e.g. $Y(t) = A(t)(\theta + \mathbf{Z}_t)$, $\mathbf{Z}_t \sim \mathcal{N}(0, \delta\sigma I)$

time	1	...	$\tau - 1$	τ
sample	$A(1)$...	$A(\tau - 1)$	
observation	$Y(1)$...	$Y(\tau - 1)$	
declaration				$\hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1})$
error				$\mathbf{1}_{\{\hat{\theta} \neq \theta\}}$

Two Important Findings:

Advantage of allowing for wide-area Searches

Advantage of adaptive search strategies over open-loop

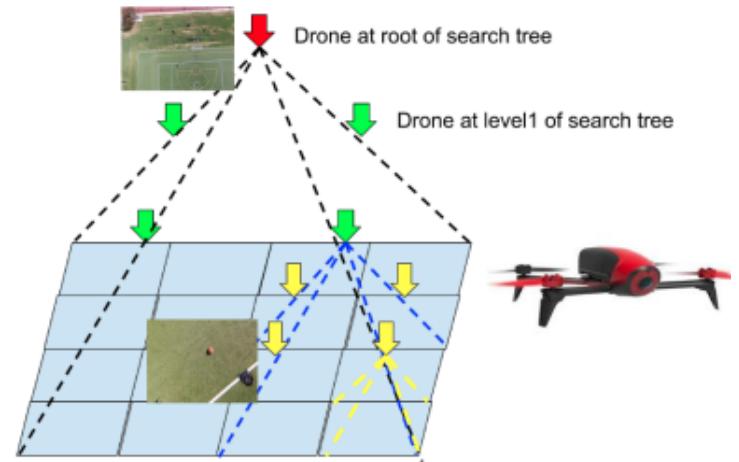
A Lalitha, N Ronquillo, T Javidi. Improved Target Acquisition Rates With Feedback Codes. IEEE Journal of Selected Topics in Signal Processing 12 (5), 871-885

Two Important Findings:

Advantage of allowing for wide-area Searches

Advantage of adaptive search strategies over open-loop

System Design:



A Lalitha, N Ronquillo, T Javidi. Improved Target Acquisition Rates With Feedback Codes. IEEE Journal of Selected Topics in Signal Processing 12 (5), 871-885

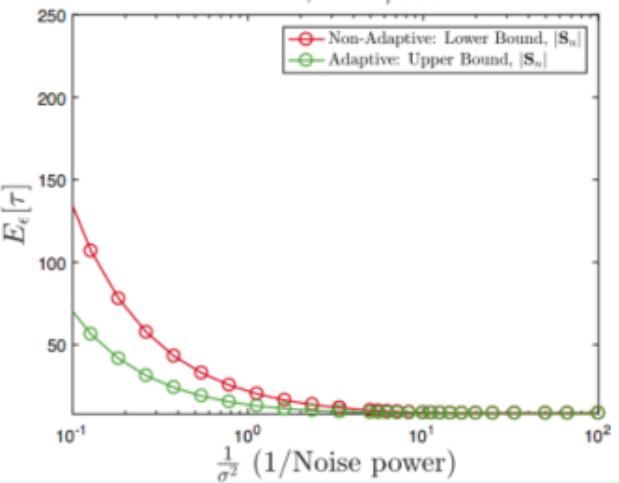
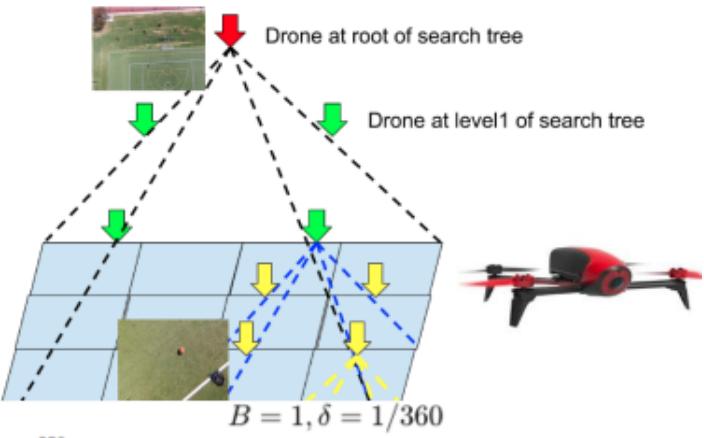
Two Important Findings:

Advantage of allowing for wide-area Searches

Advantage of adaptive search strategies over open-loop

A Lalitha, N Ronquillo, T Javidi. Improved Target Acquisition Rates With Feedback Codes. IEEE Journal of Selected Topics in Signal Processing 12 (5), 871-885

System Design:



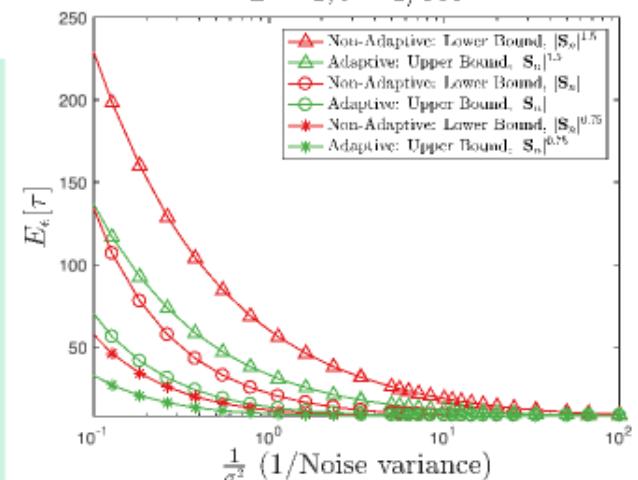
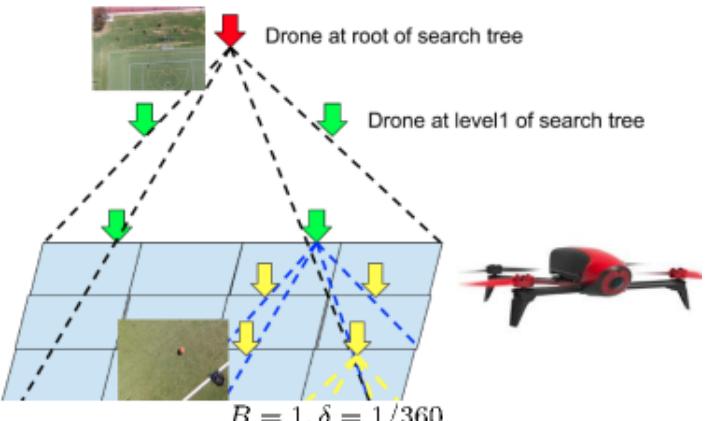
Two Important Findings:

Advantage of allowing for wide-area Searches

Advantage of adaptive search strategies over open-loop

A Lalitha, N Ronquillo, T Javidi. Improved Target Acquisition Rates With Feedback Codes. IEEE Journal of Selected Topics in Signal Processing 12 (5), 871-885

System Design:



Main Analytical (Information Theoretic) Insight

Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{X^A}^{\sup(A) \cap \sup(\theta) \neq 0} + \hat{Z}^A$

Main Analytical (Information Theoretic) Insight

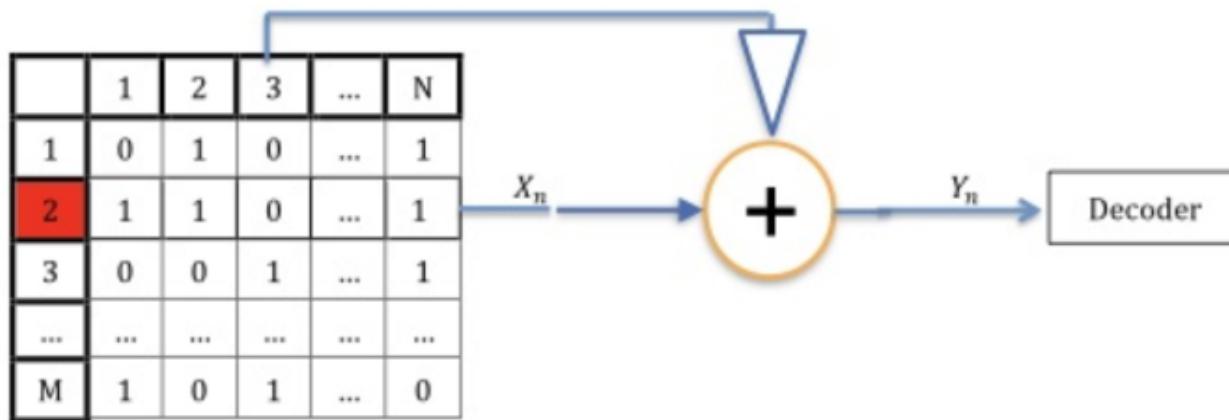
$$\textbf{Observation: } Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$$

Binary input additive (Gaussian) channel

Main Analytical (Information Theoretic) Insight

$$\text{Observation: } Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$$

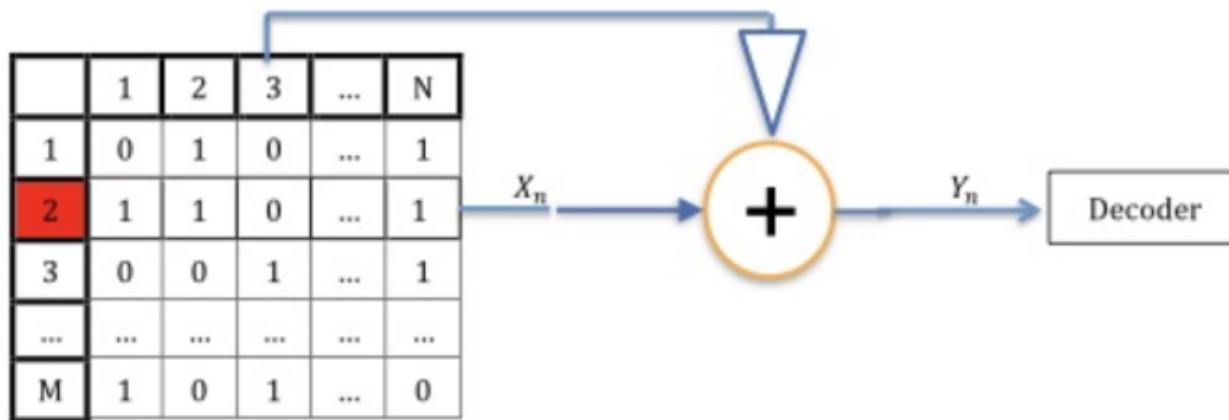
Binary input additive (Gaussian) channel



Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

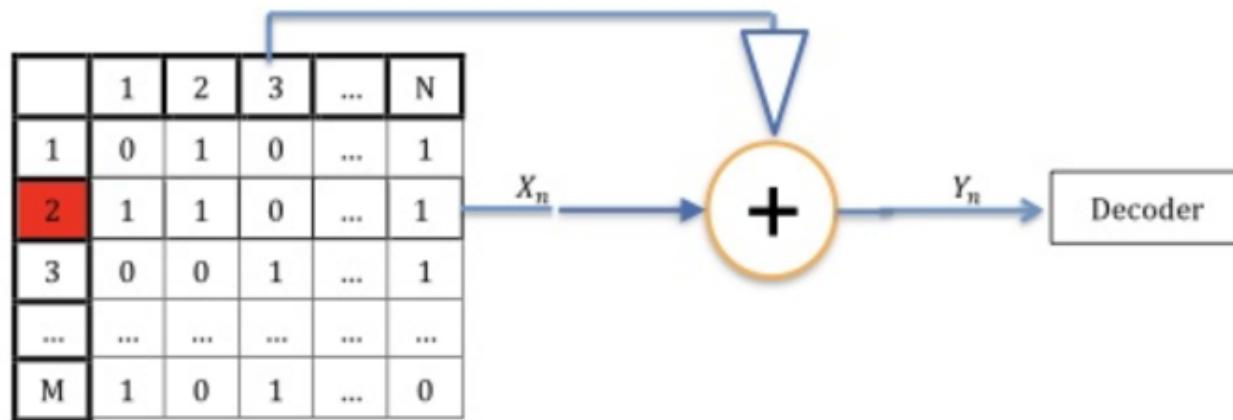
Binary input additive (Gaussian) channel $\Rightarrow \mathbb{E}[\tau] \approx \frac{\log B/\delta\epsilon}{I(X, Y^A)}$ is sufficient



Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

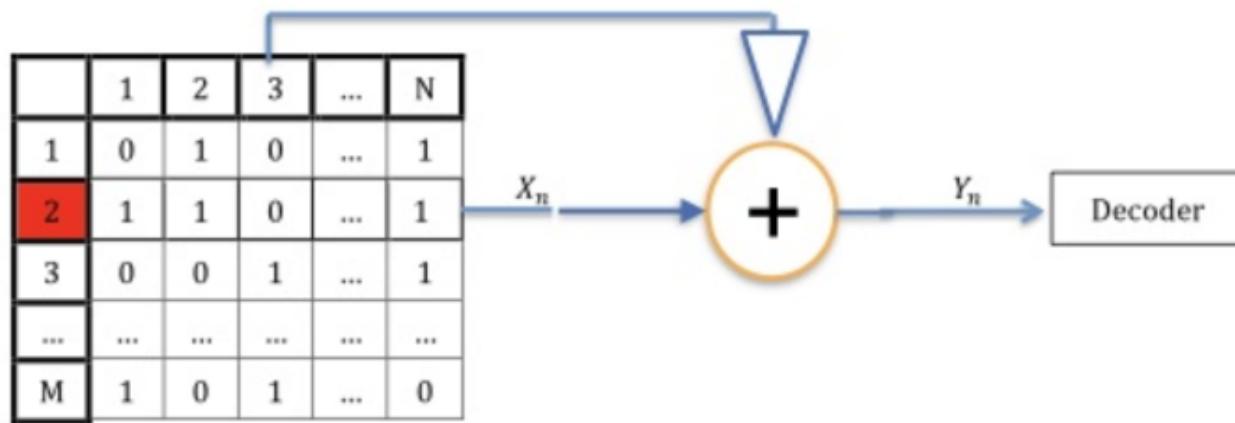
$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q, \sigma\sqrt{qB})}$$



Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})}$$

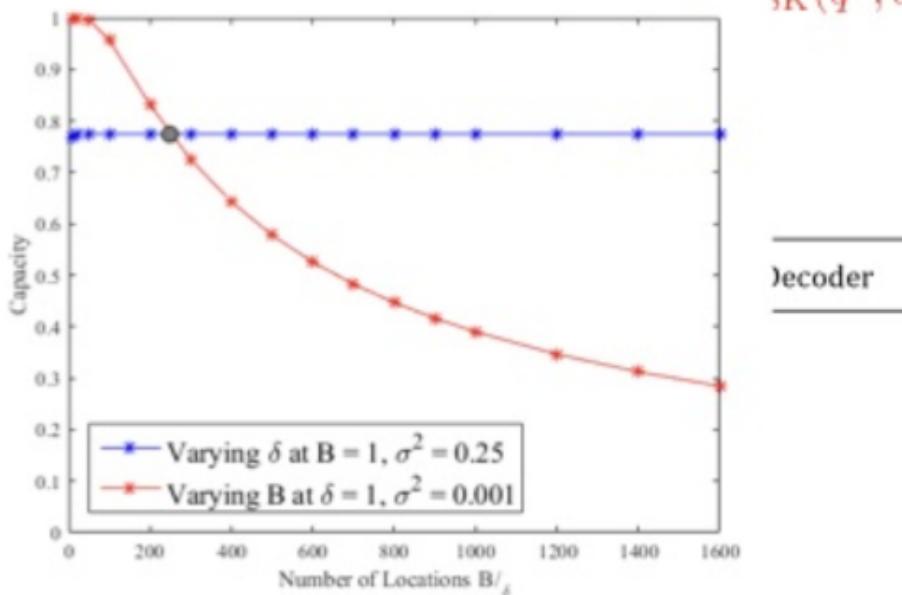


Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \quad Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_{\epsilon}^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{\text{SIK}(q^*, \sigma\sqrt{q^*B})}$$

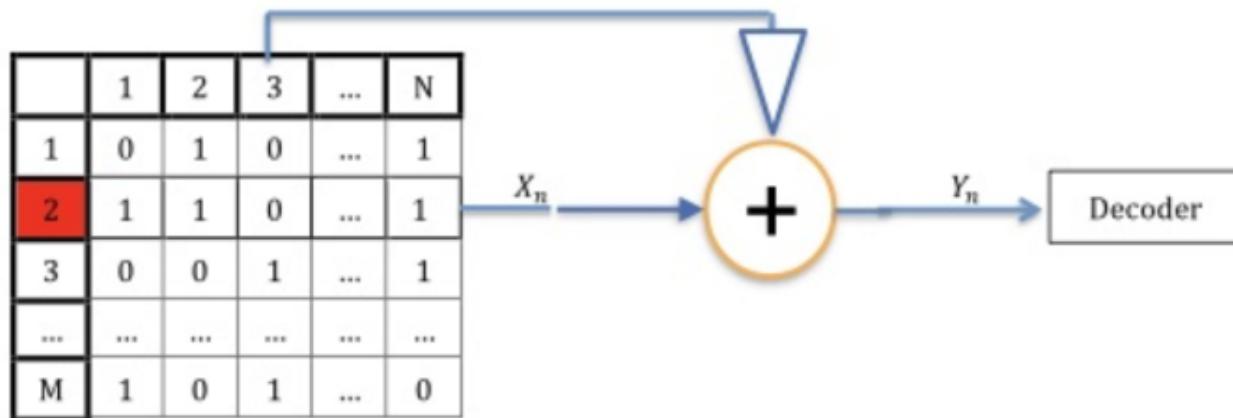
	1	2	3
1	0	1	0
2	1	1	0
3	0	0	1
...
M	1	0	1



Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

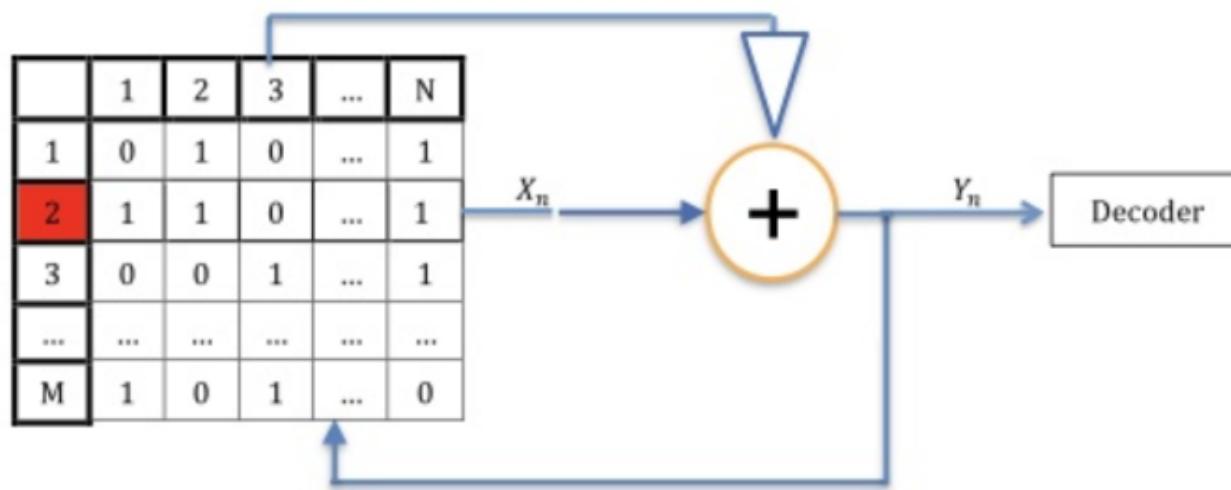
$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})}$$



Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

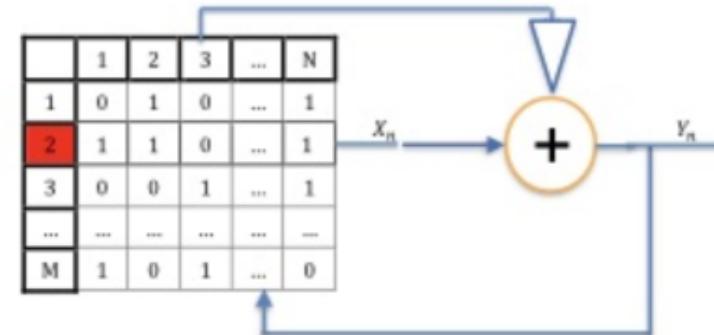
$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})}$$



Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})}$$



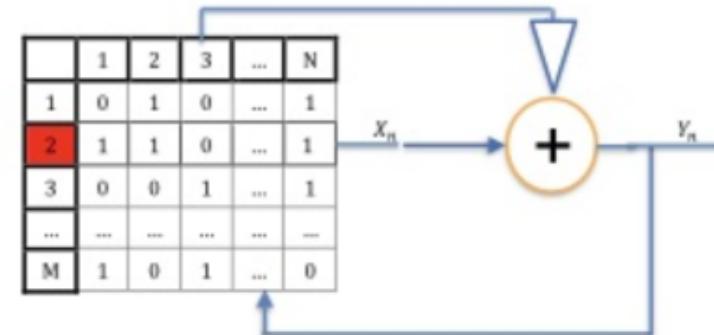
Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})}$$

- Adaptive strategy builds on posterior matching

- Ensures high $I(\theta, Y(t))$



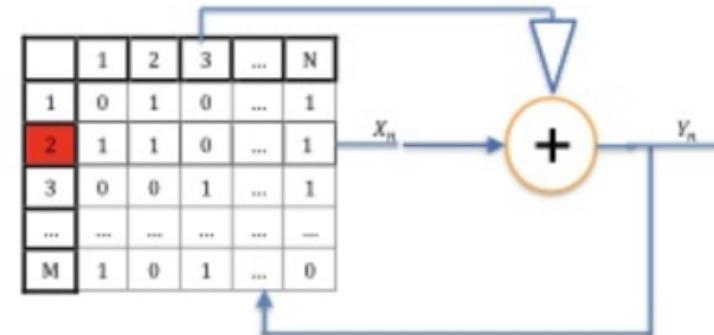
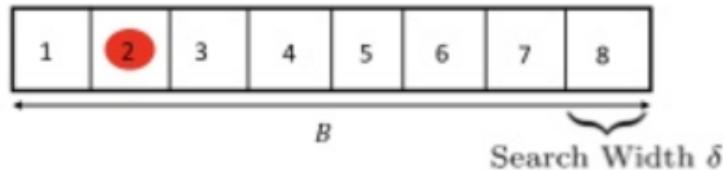
Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})}$$

- Adaptive strategy builds on posterior matching

- Ensures high $I(\theta, Y(t))$



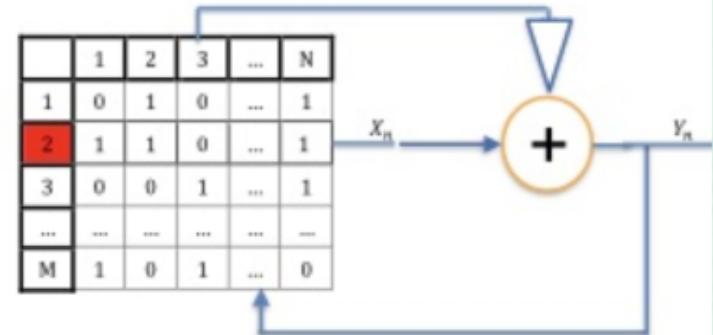
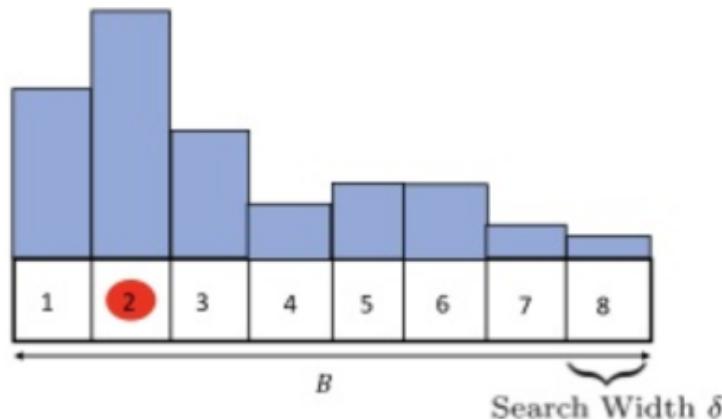
Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})}$$

- Adaptive strategy builds on posterior matching

- Ensures high $I(\theta, Y(t))$



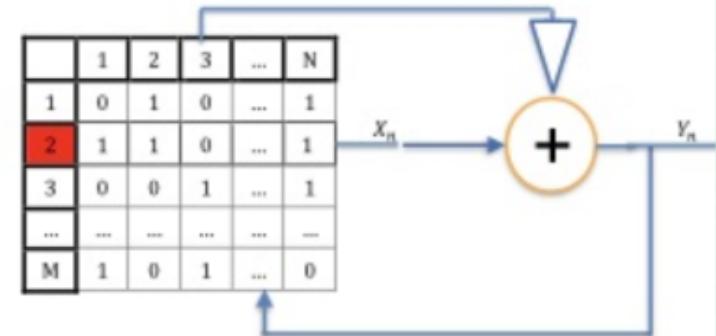
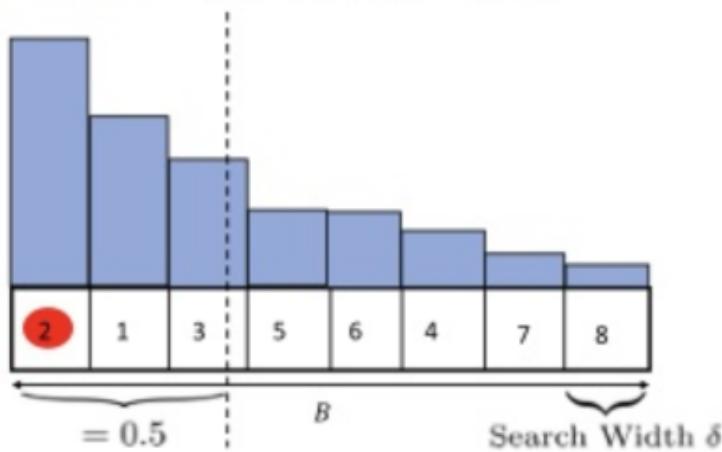
Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})}$$

- Adaptive strategy builds on posterior matching

- Ensures high $I(\theta, Y(t))$



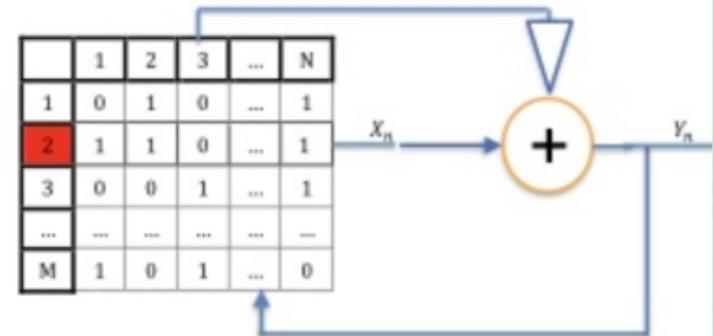
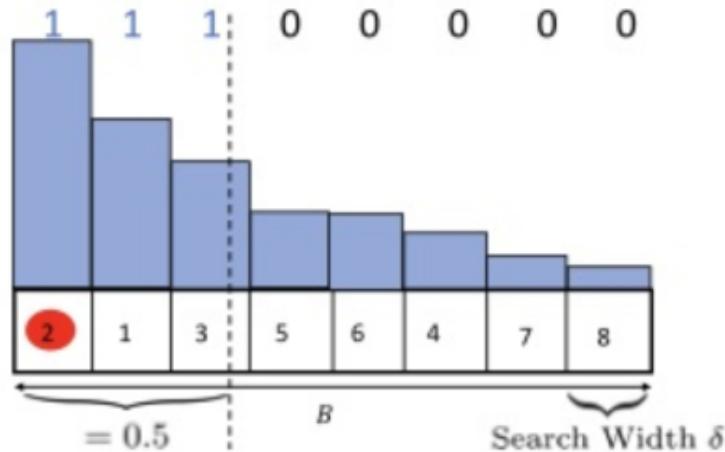
Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \overbrace{\mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0}}^{X^A} + \hat{Z}^A$

$$Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau_\epsilon^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})}$$

- Adaptive strategy builds on posterior matching

- Ensures high $I(\theta, Y(t))$



Summary: Two Important Questions Answered

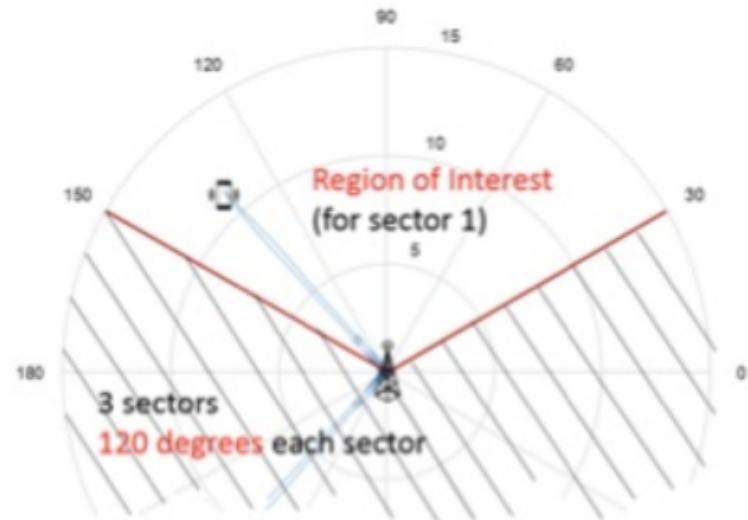
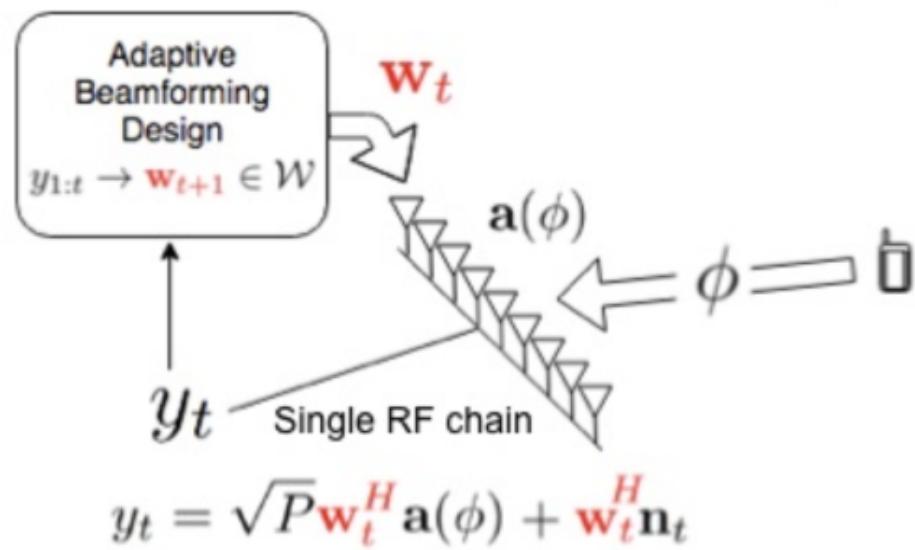
- Role of allowable actions set \mathcal{A}
 - Designing \mathcal{A} can significantly reduce the overhead
 - ▶ Even when noise variance increases w $|A(t)|$ (linearly)!
- Adapt $A(t)$ to past observations (feedback) or not?
 - Adaptive policies are computationally expensive but significant adaptivity gain in low SNR regimes

In theory, there is no difference
between theory and practice...

In practice, there is!

Measurement-Dependent Noisy Search

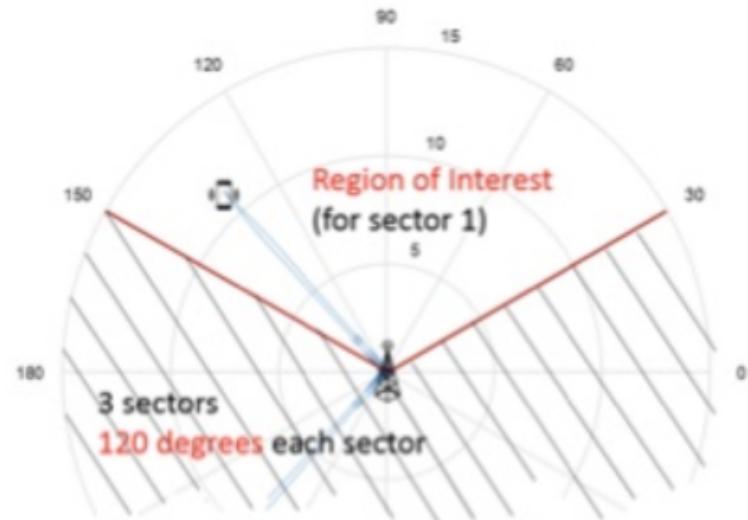
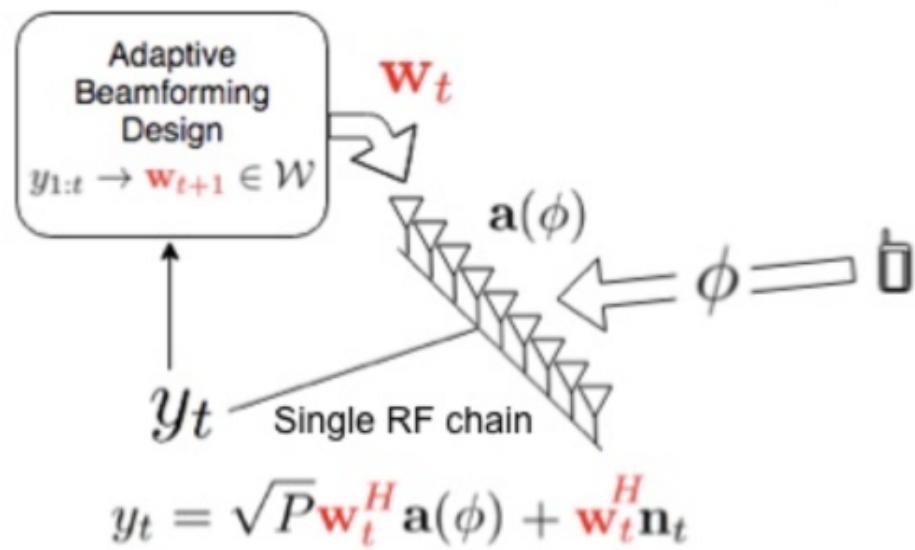
time	1	\dots	$\tau - 1$	τ
beam	W_1	\dots	$W_{\tau-1}$	
observe	y_1	\dots	$y_{\tau-1}$	
detect				$\hat{\phi} = d(y_{1:\tau-1}, W_{1:\tau-1})$
error				$1_{\{\hat{W} \neq W^*\}}$



$$a(\phi) := \alpha [1, e^{j \frac{2\pi d}{\lambda} \sin \phi}, \dots, e^{j(N-1) \frac{2\pi d}{\lambda} \sin \phi}]$$

Measurement-Dependent Noisy Search

time	1	\dots	$\tau - 1$	τ
beam	W_1	\dots	$W_{\tau-1}$	
observe	y_1	\dots	$y_{\tau-1}$	
detect				$\hat{\phi} = d(y_{1:\tau-1}, W_{1:\tau-1})$
error				$1_{\{\hat{W} \neq W^*\}}$

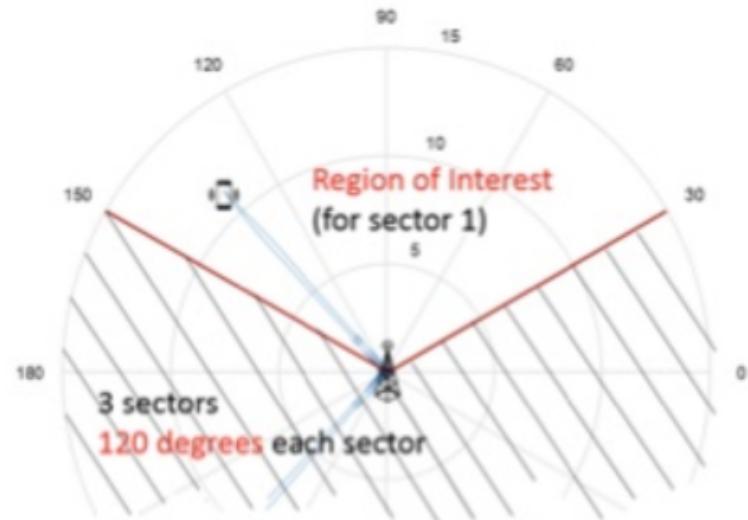
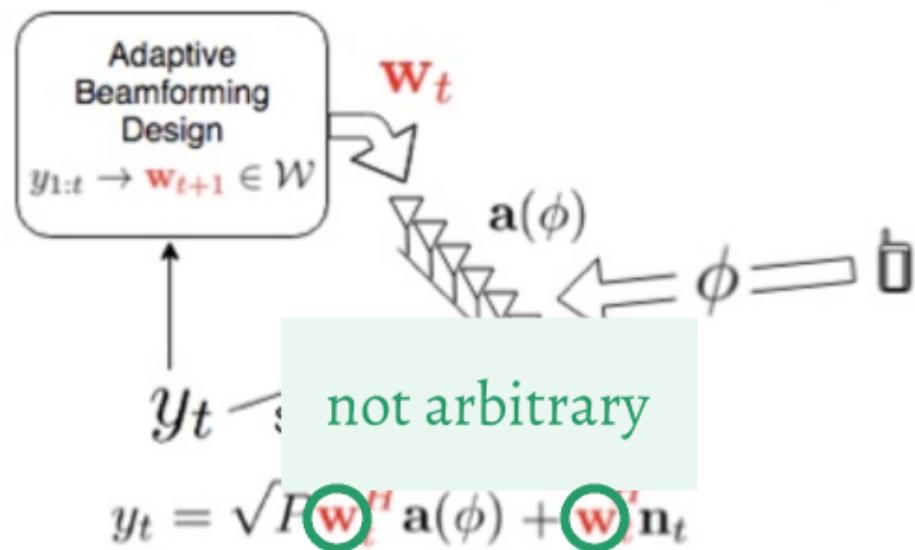


not known

$$\mathbf{a}(\phi) := [\alpha 1, e^{j \frac{2\pi d}{\lambda} \sin \phi}, \dots, e^{j(N-1) \frac{2\pi d}{\lambda} \sin \phi}]$$

Measurement-Dependent Noisy Search

time	1	...	$\tau - 1$	τ
beam	W_1	...	$W_{\tau-1}$	
observe	y_1	...	$y_{\tau-1}$	
detect				$\hat{\phi} = d(y_{1:\tau-1}, W_{1:\tau-1})$
error				$1_{\{\hat{W} \neq W^*\}}$



not known

$$a(\phi) := [\alpha 1, e^{j \frac{2\pi d}{\lambda} \sin \phi}, \dots, e^{j(N-1) \frac{2\pi d}{\lambda} \sin \phi}]$$

Measurement-Dependent Noisy Search

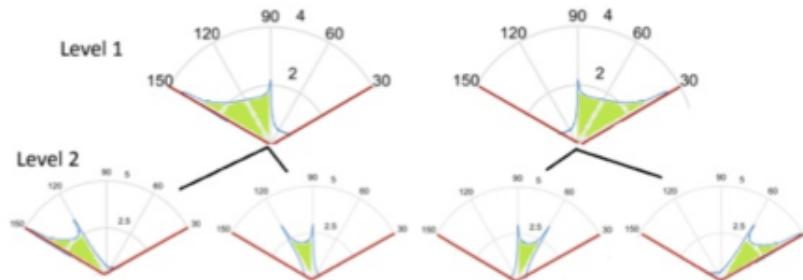
time 1
beam W_1
observe y_1
detect
error

Adaptive Beamforming Design
 $y_{1:t} \rightarrow \mathbf{w}_{t+1} \in \mathcal{W}$

$$y_t \leftarrow \text{not arbitrary}$$

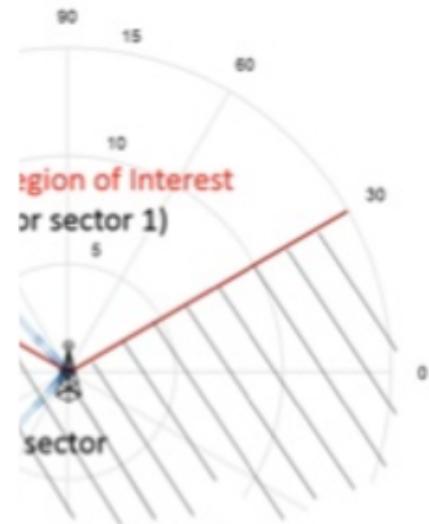
$$y_t = \sqrt{P} \mathbf{w}^H \mathbf{a}(\phi) + \mathbf{w}^H \mathbf{n}_t$$

Hierarchical Beam Patterns



Binary search
Repeat to increase SNR (linear in beam width)

[1] Alkhateeb, A., Leus, G., Heath, R. "Multi-Layer Precoding: A Potential Solution for Full-Dimensional Massive MIMO Systems." on arXiv

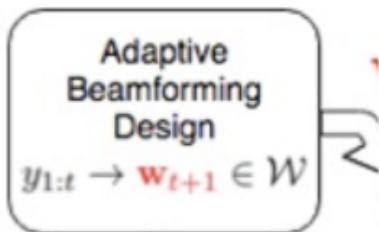


not known

$$\mathbf{a}(\phi) := [\alpha 1, e^{j \frac{2\pi d}{\lambda} \sin \phi}, \dots, e^{j(N-1) \frac{2\pi d}{\lambda} \sin \phi}]$$

Measurement-Dependent Noisy Search

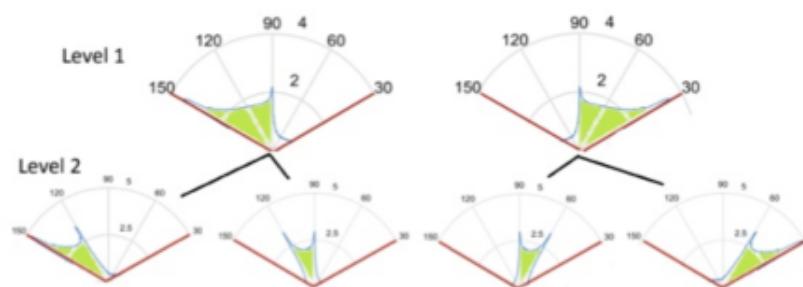
time 1
 beam W_1
 observe y_1
 detect
 error



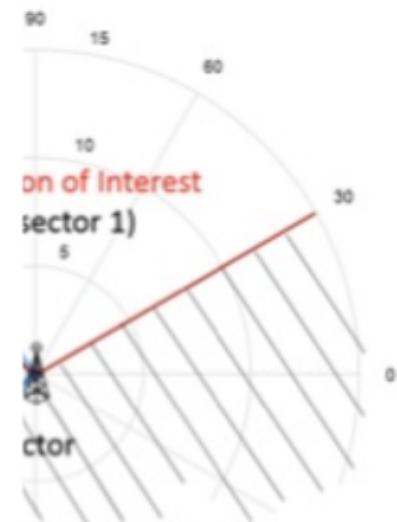
$$y_t \leftarrow \text{not arbitrary}$$

$$y_t = \sqrt{P} \mathbf{w}^H \mathbf{a}(\phi) + \mathbf{w}^H \mathbf{n}_t$$

Hierarchical Beam Patterns

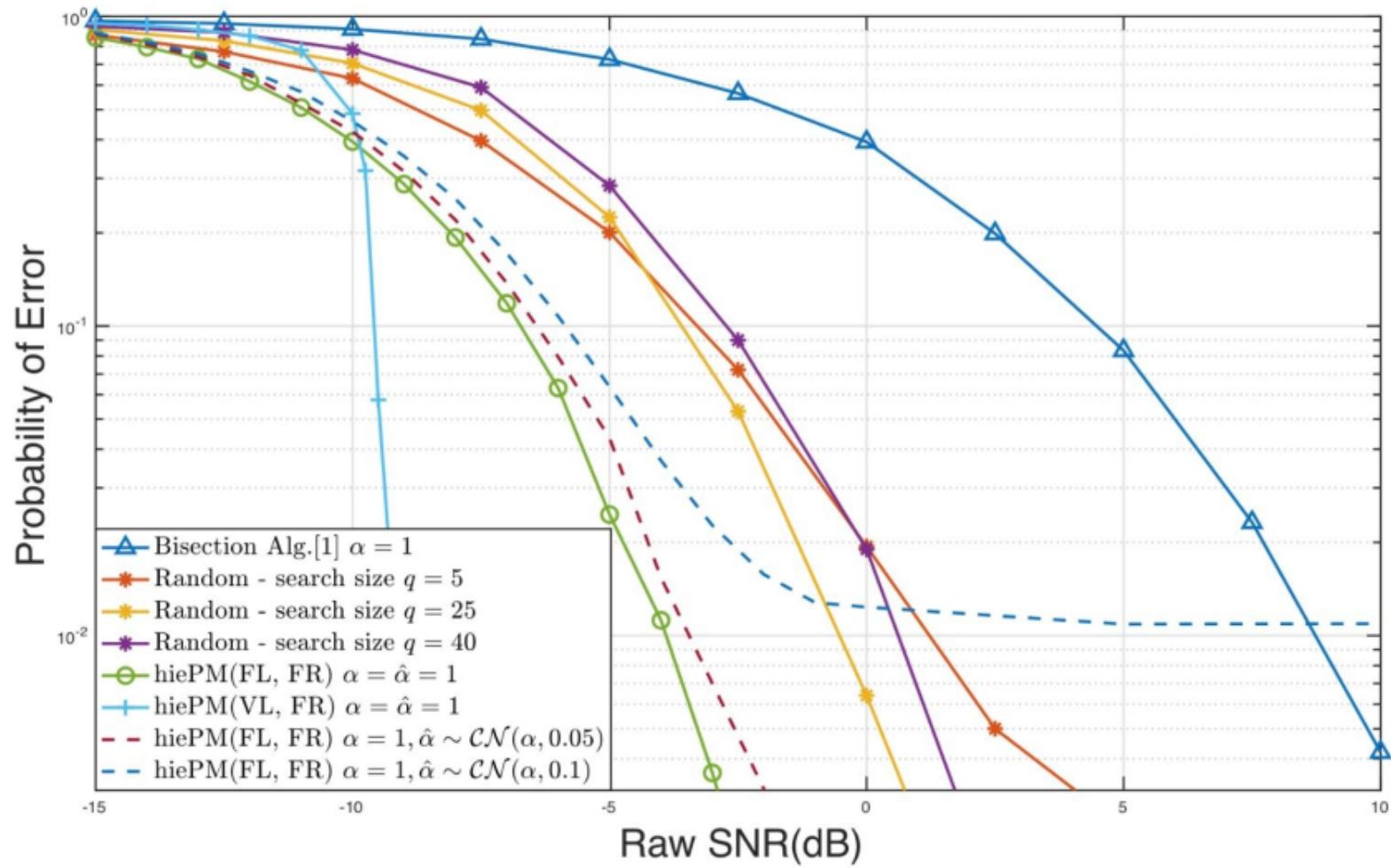


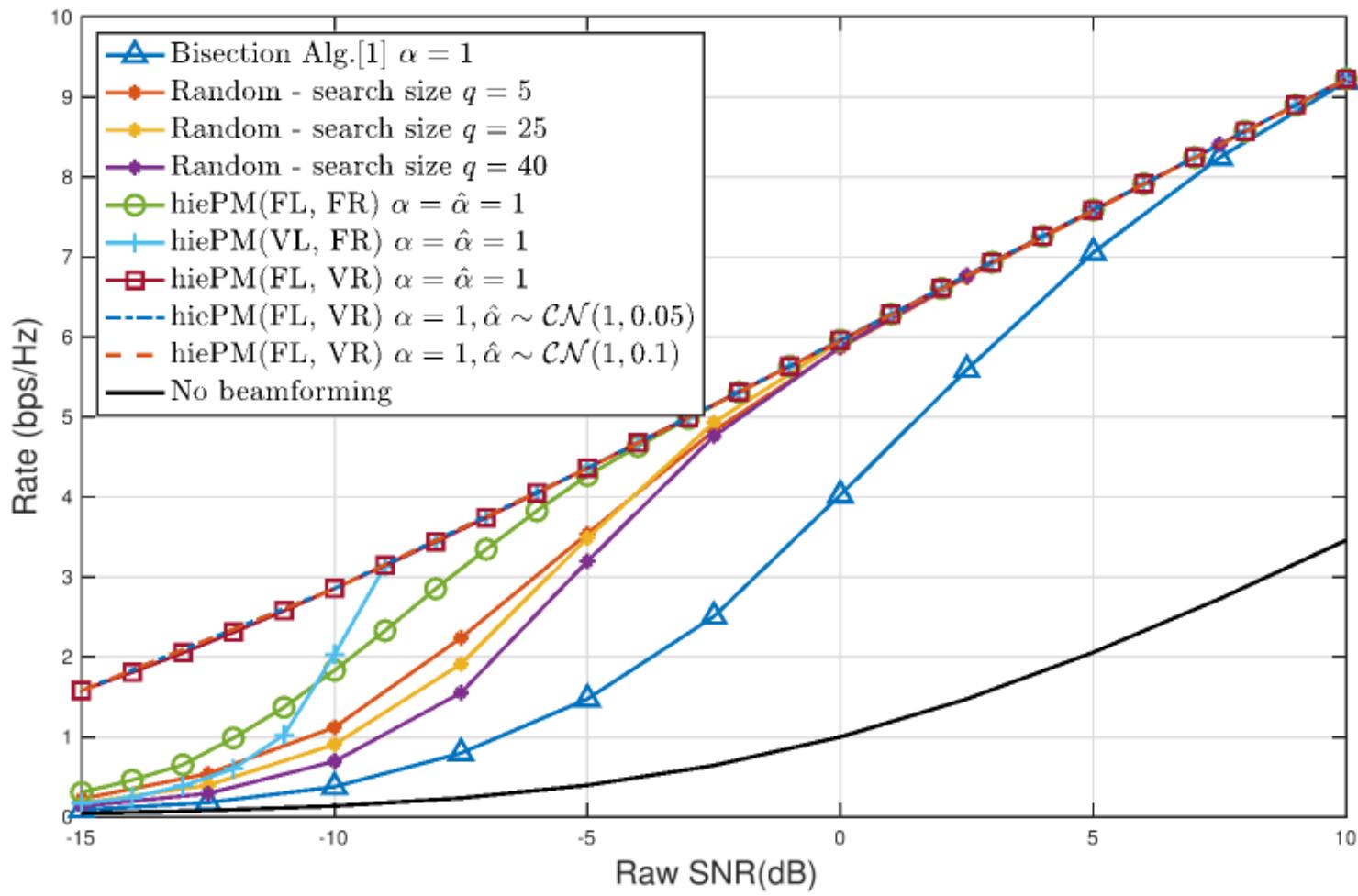
[2] Chiu, S., Ronquillo, N., Javidi, T. "Sequential Beam Alignment in mmWave Communication." on arXiv



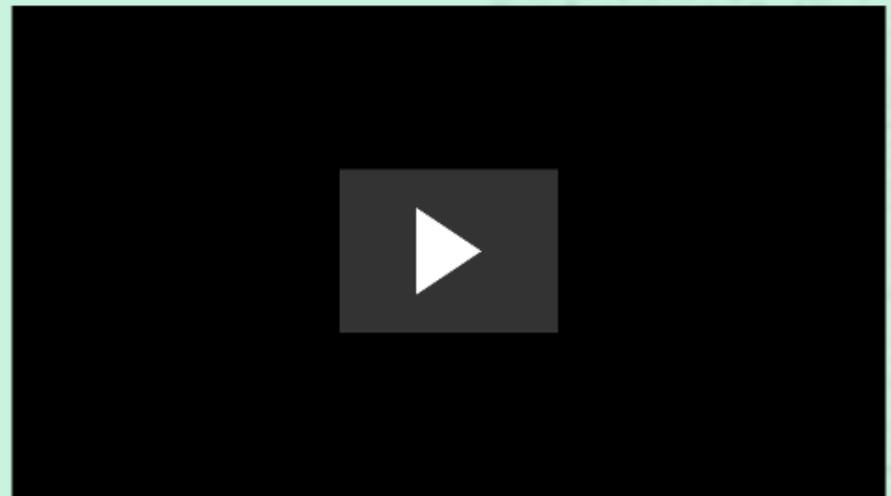
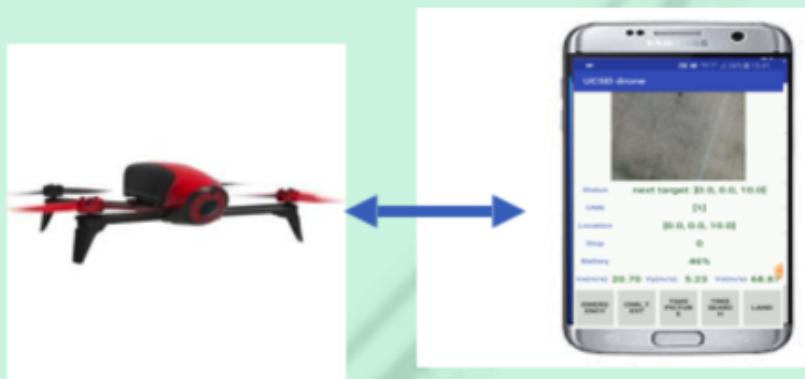
not known

$$\mathbf{a}(\phi) := [\alpha, 1, e^{j\frac{2\pi d}{\lambda} \sin \phi}, \dots, e^{j(N-1)\frac{2\pi d}{\lambda} \sin \phi}]$$

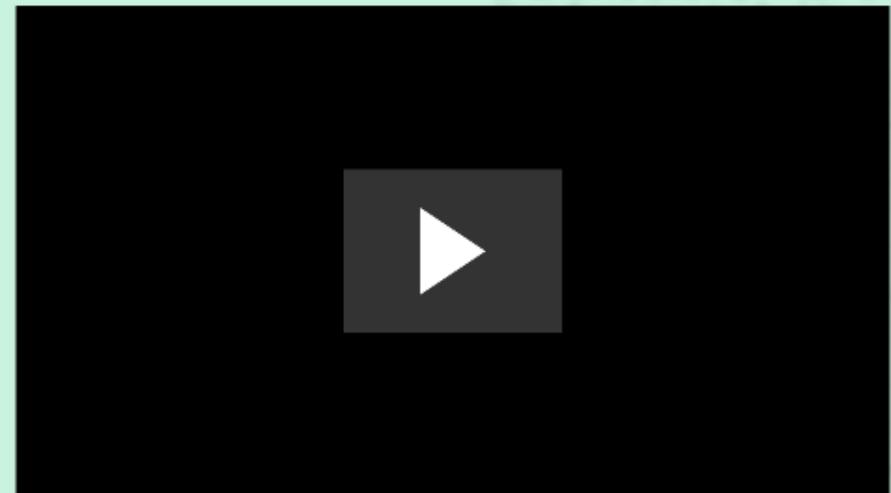
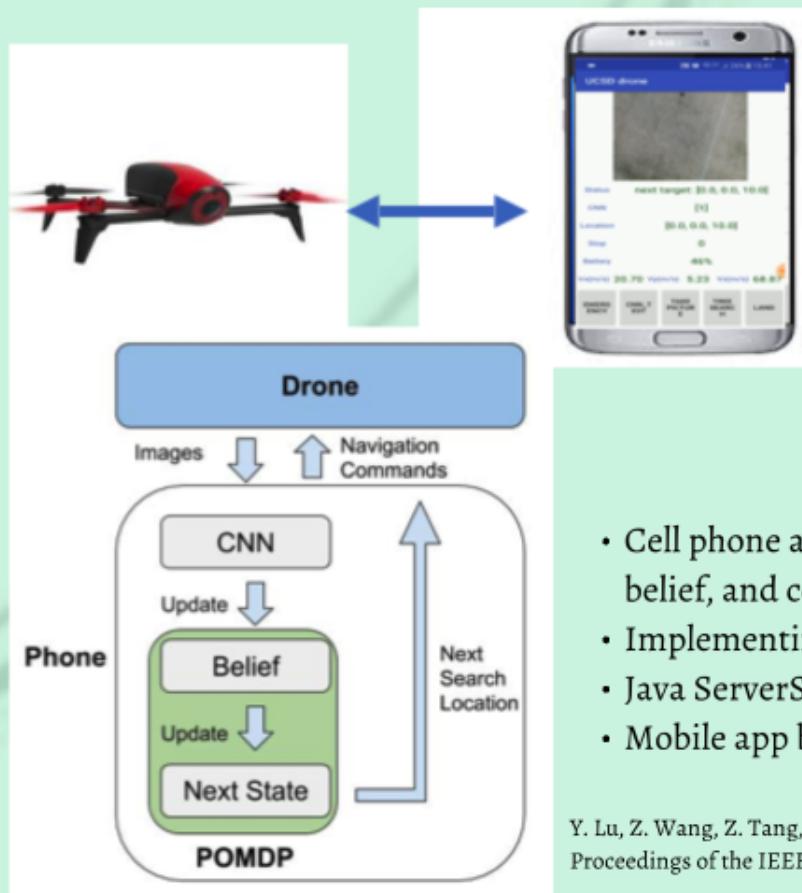




Demo: Parrot Platform

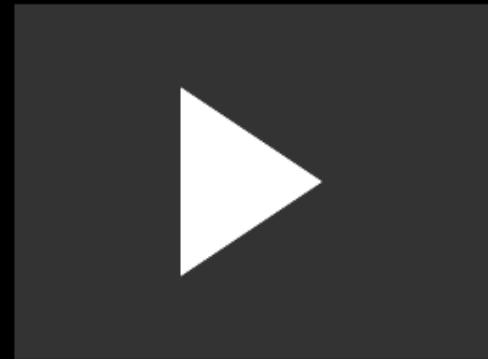


Demo: Parrot Platform



- Cell phone acts as a server for downloading images from, classify/compute belief, and control commands to drone
- Implementing/tuning MobileNet CNN
- Java ServerSocket for receiving/sending messages on drone
- Mobile app built with Android Studio + Parrot Software Development Kit

Y. Lu, Z. Wang, Z. Tang, and T. Javidi. Target Localization with Drones using Mobile CNNs. to be presented and appear in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2018



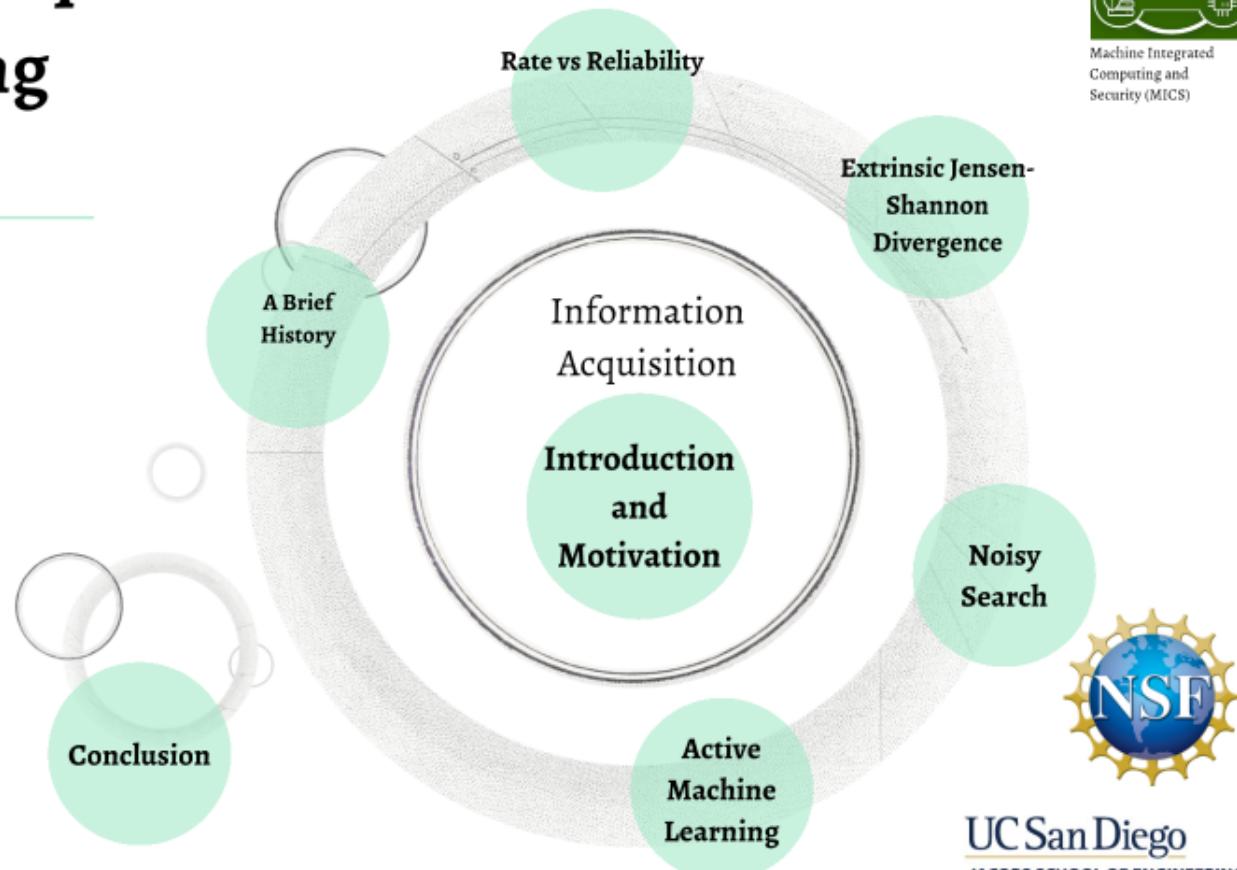
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

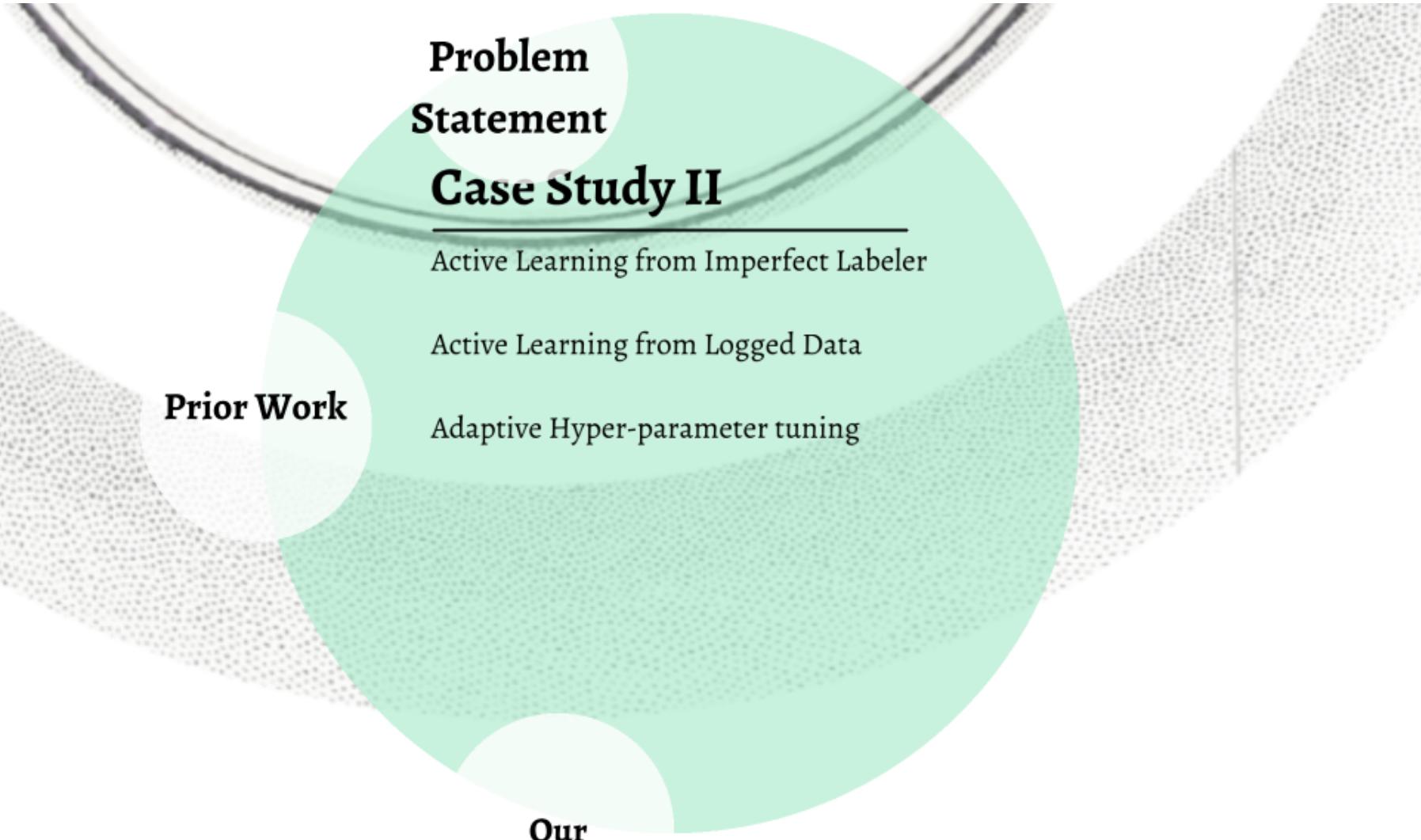
Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz



Machine Integrated
Computing and
Security (MICS)



UC San Diego
JACOBS SCHOOL OF ENGINEERING
Center for Wireless Communications



**Problem
Statement**

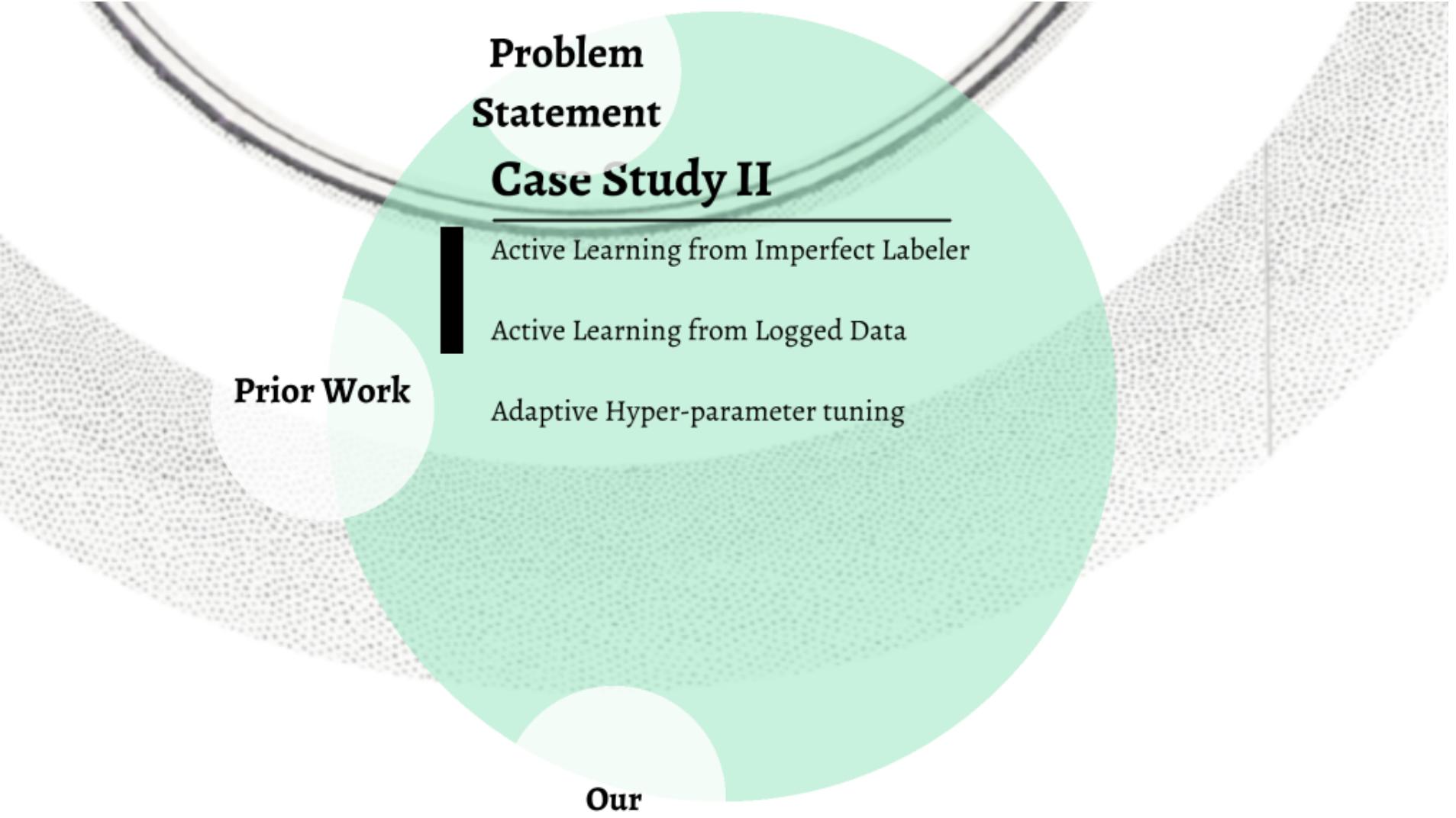
Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning

Our



Problem Statement

Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning

Prior Work

Our

Prior Work

Problem Statement

Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning

- Classification
- Classical Approach: Passive Learning
- Problem: Largely Redundant Labels
- Active Learning



Our

Prior Work

Problem Statement

Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning

- Classification
- Classical Approach: Passive Learning
- Problem: Largely Redundant Labels
- Active Learning



Our

Problem Statement

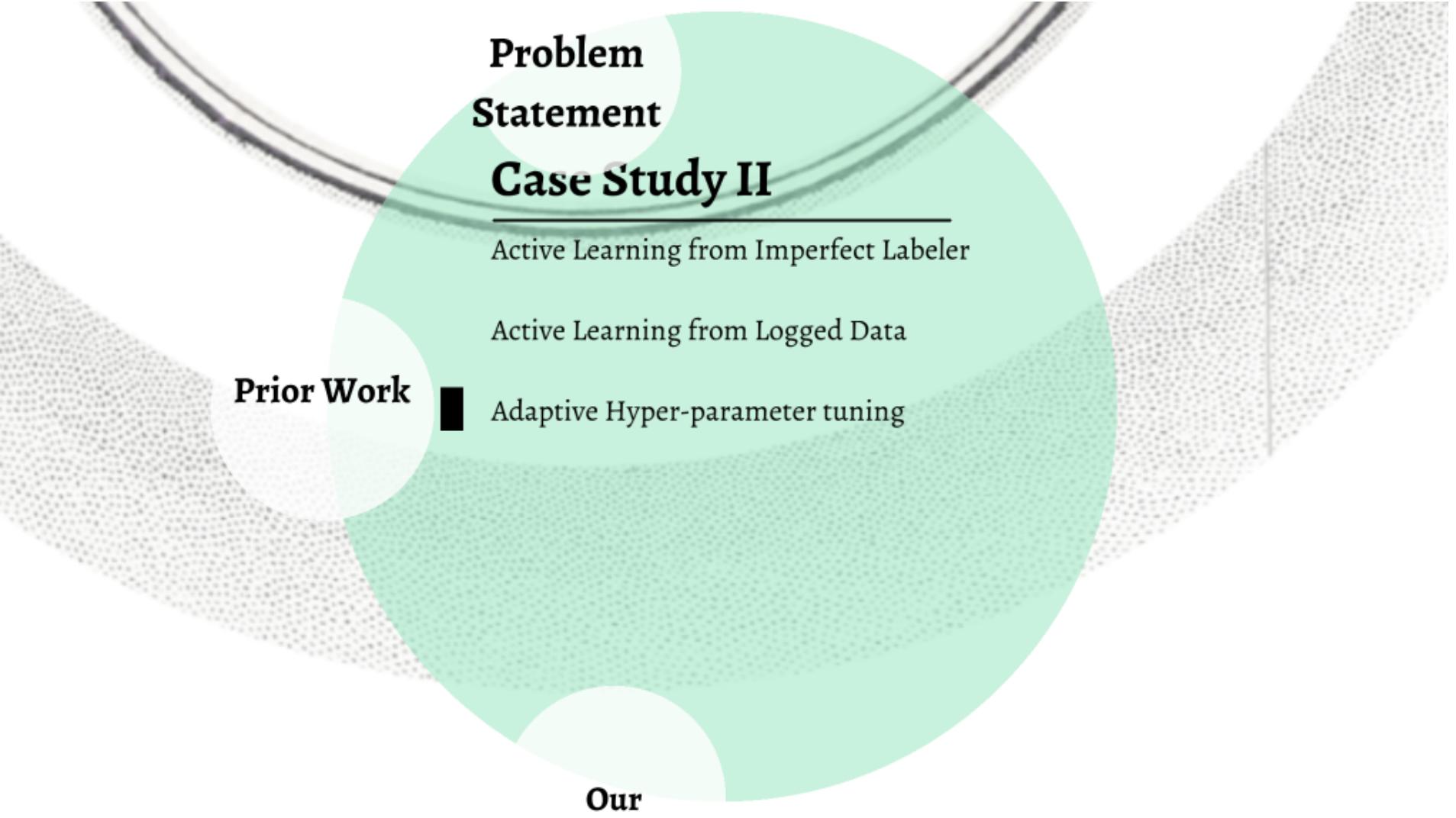
Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning

Our



Problem Statement

Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning

Prior Work

Our

Prior Work

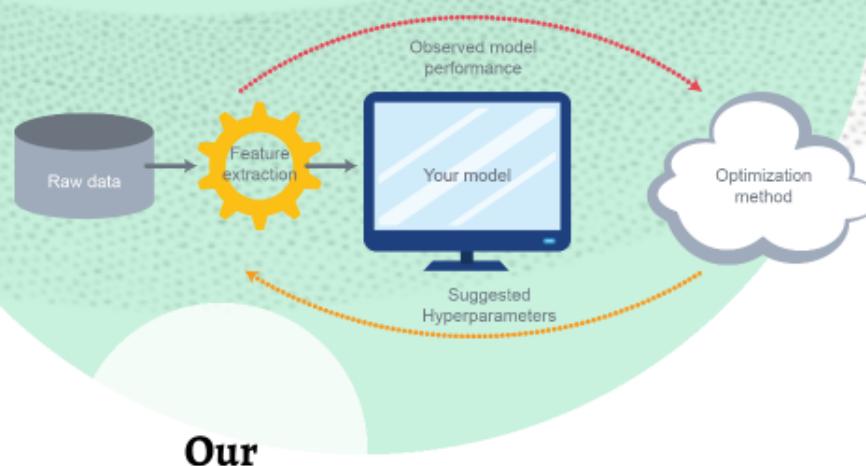
Problem Statement

Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning



Our

Black-box optimization

Consider optimizing a function $f : \mathcal{X} \rightarrow \mathbb{R}$

Caveat:

- ▶ f is **not known explicitly**: accessed only through **noisy** and **expensive** evaluation queries

Black-box optimization

Consider optimizing a function $f : \mathcal{X} \rightarrow \mathbb{R}$

Caveat:

- ▶ f is **not known explicitly**: accessed only through **noisy** and **expensive** evaluation queries

Goal: Design a sequential strategy of selecting n query points x_1, \dots, x_n to *efficiently* optimize f over the horizon n

Black-box optimization

Consider optimizing a function $f : \mathcal{X} \rightarrow \mathbb{R}$

Caveat:

- ▶ f is **not known explicitly**: accessed only through **noisy** and **expensive** evaluation queries

Goal: Design a sequential strategy of selecting n query points x_1, \dots, x_n to *efficiently* optimize f over the horizon n

- ▶ Performance measures:
 - ▶ Simple regret: $\mathcal{S}_n = f(x^*) - f(x_n)$

Black-box optimization

Consider optimizing a function $f : \mathcal{X} \rightarrow \mathbb{R}$

Caveat:

- ▶ f is **not known explicitly**: accessed only through **noisy** and **expensive** evaluation queries

Goal: Design a sequential strategy of selecting n query points x_1, \dots, x_n to *efficiently* optimize f over the horizon n

- ▶ Performance measures:
 - ▶ Simple regret: $\mathcal{S}_n = f(x^*) - f(x_n)$
 - ▶ Cumulative regret: $\mathcal{R}_n = \sum_{t=1}^n f(x^*) - f(x_t)$

Black-box optimization

Consider optimizing a function $f : \mathcal{X} \rightarrow \mathbb{R}$ (consider $\mathcal{X} \subset \mathbb{R}^D$)

Caveat:

- ▶ f is **not known explicitly**: accessed only through **noisy** and **expensive** evaluation queries

Goal: Design a sequential strategy of selecting n query points x_1, \dots, x_n to *efficiently* optimize f over the horizon n

- ▶ Performance measures:
 - ▶ Simple regret: $\mathcal{S}_n = f(x^*) - f(x_n)$
 - ▶ Cumulative regret: $\mathcal{R}_n = \sum_{t=1}^n f(x^*) - f(x_t)$

Black-box optimization

Consider optimizing a function $f : \mathcal{X} \rightarrow \mathbb{R}$

Caveat:

- ▶ f is **not known explicitly**: accessed only through **noisy** and **expensive** evaluation queries

Goal: Design a sequential strategy of selecting n query points x_1, \dots, x_n to *efficiently* optimize f over the horizon n

- ▶ Performance measures:
 - ▶ Simple regret: $\mathcal{S}_n = f(x^*) - f(x_n)$
 - ▶ Cumulative regret: $\mathcal{R}_n = \sum_{t=1}^n f(x^*) - f(x_t)$
- ▶ Ill-posed unless learning $f(x)$ gives information about $f(x')$

Bayesian Setting: Gaussian Prior and Additive Noise

- ▶ f is a sample drawn from a zero mean Gaussian process with covariance function $K(x, x') = \mathbb{E}[f(x)f(x')]$
- ▶ Observation model: $y = f(x) + \eta$ with $\eta \sim N(0, \sigma^2)$
- ▶ Gaussian posterior
 - ▶ Posterior mean and variance at x :

$$\mu_t(x) = k_t(x)^T(K_t + \sigma^2 I)^{-1}y_{1:t-1}$$

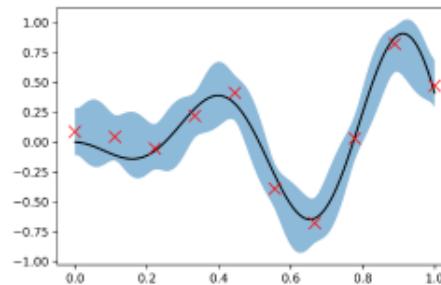
$$\sigma_t^2(x) = k(x, x) - k_t(x)^T(K_t + \sigma^2 I)^{-1}k_t(x)$$

where

$$k_t(x) = \text{cov}(f(x), \underline{f}(x_{[1:t-1]}))$$

and

$$K_t = \text{cov}(\underline{f}(x_{[1:t-1]}), \underline{f}(x_{[1:t-1]}))$$



3/18

Application 1: Hyper-parameter tuning in ML models

- ▶ Training with hyperparameters θ outputs classifier $A(\theta)$
 - ▶ \mathcal{X} = space of hyperparameters.
 - ▶ $f(\theta)$ = performance of $A(\theta)$ on some test set.
 - ▶ Finite n trials \Rightarrow good θ_n (pure exploration)

Goal: After n rounds, output θ_n^* to minimize simple regret:

$$\mathcal{S}_n = f(\theta^*) - f(\theta_n^*)$$

Application 1: Hyper-parameter tuning in ML models

- ▶ Training with hyperparameters θ outputs classifier $A(\theta)$
 - ▶ \mathcal{X} = space of hyperparameters.
 - ▶ $f(\theta)$ = performance of $A(\theta)$ on some test set.
 - ▶ Finite n trials \Rightarrow good θ_n (pure exploration)

Goal: After n rounds, output θ_n^* to minimize simple regret:

$$\mathcal{S}_n = f(\theta^*) - f(\theta_n^*)$$

- ▶ Practically formulated as GP w Matérn family:

$$K_{\nu}^{\text{Matérn}}(x, x') = K(0)\left(1 + \sum_{i=1}^m a_i \|x - x'\|^i\right) e^{-c_1 \sqrt{\nu} \|x - x'\|},$$

(for half integer values of $\nu = m + 1/2$ and some $a_i > 0$)

Application 1: Hyper-parameter tuning in ML models

- ▶ Training with hyperparameters θ outputs classifier $A(\theta)$
 - ▶ \mathcal{X} = space of hyperparameters.
 - ▶ $f(\theta)$ = performance of $A(\theta)$ on some test set.
 - ▶ Finite n trials \Rightarrow good θ_n (pure exploration)

Goal: After n rounds, output θ_n^* to minimize simple regret:

$$\mathcal{S}_n = f(\theta^*) - f(\theta_n^*)$$

- ▶ Practically formulated as GP w Matérn family:

$$K_\nu^{\text{Matérn}}(x, x') = K(0)(1 + \sum_{i=1}^m a_i \|x - x'\|^i) e^{-c_1 \sqrt{\nu} \|x - x'\|}, \quad \left\{ \begin{array}{l} \nu = 1/2 \\ \nu = 3/2 \\ \nu = 5/2 \\ \nu \rightarrow \infty \end{array} \right.$$

(for half integer values of $\nu = m + 1/2$ and some $a_i > 0$)

Gaussian Process Optimization (GP): Prior Work

- ▶ Zero mean Gaussian prior with known covariance function
- ▶ Update the posterior \mathbb{P}_t based on $x_{1:t-1}, y_{1:t-1}$
- ▶ Query point according to acquisition rule:

$$x_t = \arg \max_{x \in \mathcal{X}} \alpha(x)$$

Gaussian Process Optimization (GP): Prior Work

- ▶ Zero mean Gaussian prior with known covariance function
- ▶ Update the posterior \mathbb{P}_t based on $x_{1:t-1}, y_{1:t-1}$
- ▶ Query point according to acquisition rule:

$$x_t = \arg \max_{x \in \mathcal{X}} \alpha(x)$$

- ▶ $\alpha(x)$ is the utility of querying x balances *exploration* and *exploitation*
- ▶ Commonly used $\alpha(\cdot)$:
 - ▶ *Probability of Improvement*: $\alpha_{PI}(x) = \mathbb{P}_t(f(x) > \tau)$
 - ▶ *Expected Improvement*: $\alpha_{EI}(x) = \mathbb{E}_t[(f(x) - \tau)\mathbf{1}_{\{f(x)>\tau\}}]$
 - ▶ *Upper Confidence Bound*: $\alpha_{UCB}(x) = \mu_t(x) + \beta_n \sigma_t(x)$

6/18

Gaussian Process Optimization (GP): Prior Work

- ▶ Zero mean Gaussian prior with known covariance function
- ▶ Update the posterior \mathbb{P}_t based on $x_{1:t-1}, y_{1:t-1}$
- ▶ Query point according to acquisition rule:

$$x_t = \arg \max_{x \in \mathcal{X}} \alpha(x)$$

- ▶ $\alpha(x)$ is the utility of querying x balances *exploration* and *exploitation*
- ▶ Commonly used $\alpha(\cdot)$:
 - ▶ *Probability of Improvement*: $\alpha_{PI}(x) = \mathbb{P}_t(f(x) > \tau)$
 - ▶ *Expected Improvement*: $\alpha_{EI}(x) = \mathbb{E}_t[(f(x) - \tau)\mathbf{1}_{\{f(x)>\tau\}}]$
 - ▶ *Upper Confidence Bound*: $\alpha_{UCB}(x) = \mu_t(x) + \beta_n \sigma_t(x)$
- ▶ Non-convex optimization w many local maximas

6/18

Gaussian Process Optimization (GP): Prior Work

- ▶ Zero mean Gaussian prior with known covariance function
- ▶ Update the posterior \mathbb{P}_t based on $x_{1:t-1}, y_{1:t-1}$
- ▶ Query point according to acquisition rule:

$$x_t = \arg \max_{x \in \mathcal{X}} \alpha(x)$$

- ▶ $\alpha(x)$ is the utility of querying x balances *exploration* and *exploitation*
- ▶ Practically we rely on a discretization \mathcal{X}_t

$$x_t = \arg \max_{x \in \mathcal{X}} \alpha(x)$$

6/18

Gaussian Process Optimization (GP): Prior Work

- ▶ Zero mean Gaussian prior with known covariance function
- ▶ Update the posterior \mathbb{P}_t based on $x_{1:t-1}, y_{1:t-1}$
- ▶ Query point according to acquisition rule:

$$x_t = \arg \max_{x \in \mathcal{X}} \alpha(x)$$

- ▶ $\alpha(x)$ is the utility of querying x balances *exploration* and *exploitation*
- ▶ Practically, we rely on a discretization $\{\mathcal{X}_t\}_t$

$$x_t = \arg \max_{x \in \mathcal{X}_t} \alpha(x)$$

- ▶ Prior work: off-line discretization $\{\mathcal{X}_t\}_t$ with $|\mathcal{X}_t| = \mathcal{O}(t^D)$

6/18

Prior Work: Information-type Regret Bounds

- ▶ Existing bounds on \mathcal{R}_n have the general form:

$$\mathcal{R}_n \leq \mathcal{O}(\sqrt{n\gamma_n \log n}) \quad (1)$$

- ▶ Here γ_n is the maximum *information gain* from n observations

$$\gamma_n = \sup_{S \subset \mathcal{X}: |S|=n} I(y_S; f) \quad (2)$$

- ▶ For specific kernels, bounds on γ_n can be obtained

$$\gamma_n^{\text{Matérn}}(\nu) = \begin{cases} \mathcal{O}(n^{\frac{D(D+1)}{D(D+1)+2\nu}} \log n) & \nu > 1 \\ \mathcal{O}((\log n)^{D+1}) & \nu = \infty \end{cases}$$

7/18

Prior Work: Information-type Regret Bounds

- ▶ Existing bounds on \mathcal{R}_n have the general form:

$$\mathcal{R}_n \leq \mathcal{O}(\sqrt{n\gamma_n \log n}) \quad (1)$$

- ▶ Here γ_n is the maximum *information gain* from n observations

$$\gamma_n = \sup_{S \subset \mathcal{X}: |S|=n} I(y_S; f) \quad (2)$$

- ▶ γ_n : maximum information about f , and not necessarily x^* .
- ▶ For specific kernels, bounds on γ_n can be obtained

$$\gamma_n^{\text{Matérn}}(\nu) = \begin{cases} \mathcal{O}(n^{\frac{D(D+1)}{D(D+1)+2\nu}} \log n) & \nu > 1 \\ \mathcal{O}((\log n)^{D+1}) & \nu = \infty \end{cases}$$

7/18

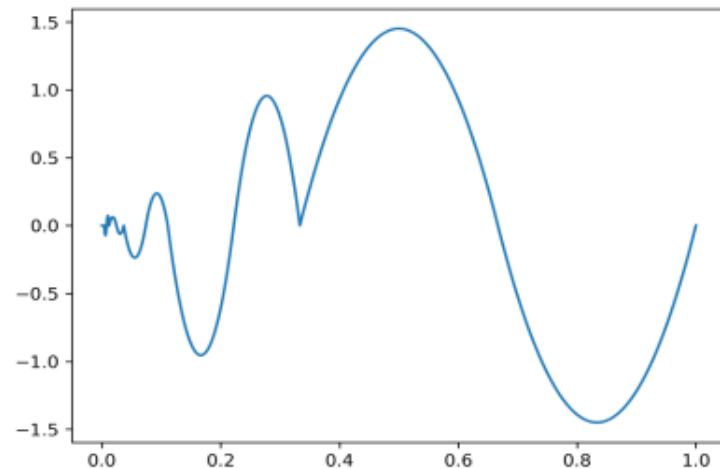
A toy example

Suppose $\mathcal{X} = [0, 1]$ and let $f : [0, 1] \rightarrow \mathbb{R}$ be a sample from:

$$f(x) = \sum_{i=1}^{\infty} a_i X_i (\psi(3^i x - 1) - \psi(3^i x - 2))$$

$$\psi(x) = 1 - 4(x - 0.5)^2$$

where $(a_i)_{i \geq 1}$ non-increasing positive constants and $X_i \sim \mathcal{N}(0, 1)$.



8/18

1. Algorithmic Improvement: Adaptive discretization for GP

- ▶ Opportunistically adapts to the (simple) structure of f
- ▶ Strictly lower complexity ($\mathcal{O}(D)$) for $\mathcal{X} \subset \mathbb{R}^D$

2. Analytic Improvement: Dimensional-type regret bound

- ▶ As good or better regret bound than prior work
- ▶ First sublinear bound for exponential kernels (Matérn- $\nu = 1/2$)
- ▶ Strictly tighter bounds for Matérn kernels if $D > \nu - 1$

S. Shekhar, and T. Javidi. Gaussian Process Bandits with Adaptive Discretization. To appear in Electronic Journal of Statistics

GP Optimization with Adaptive Discretization

Idea: Piecewise constant upper-bound confidence function

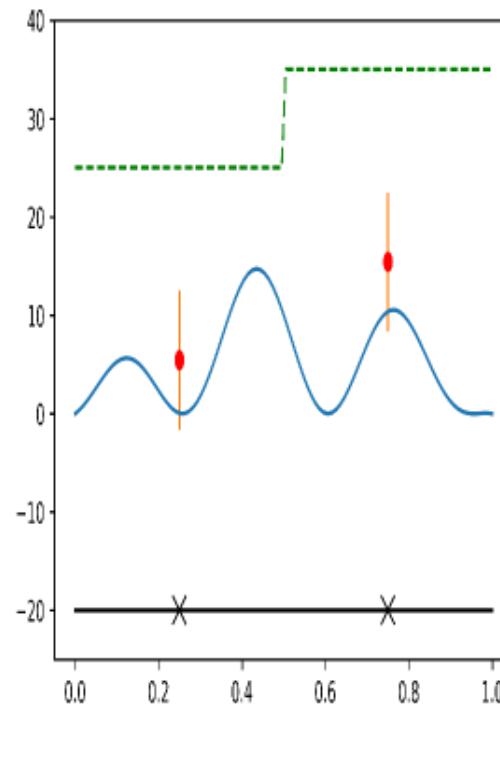


10/18

GP Optimization with Adaptive Discretization

Idea: Piecewise constant upper-bound confidence function

- ▶ Rely on two bounds:
 - ▶ Compute UCB on the function value at a point.
 - ▶ Upper bound the variation of the function in a region.
- ▶ Maximally select the best region (piecewise constant UCB)

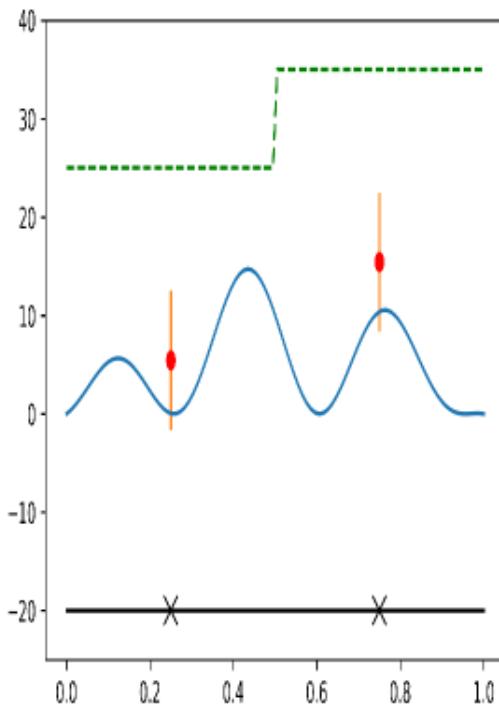


B

GP Optimization with Adaptive Discretization

Idea: Piecewise constant upper-bound confidence function

- ▶ Rely on two bounds:
 - ▶ Compute UCB on the function value at a point.
 - ▶ Upper bound the variation of the function in a region.

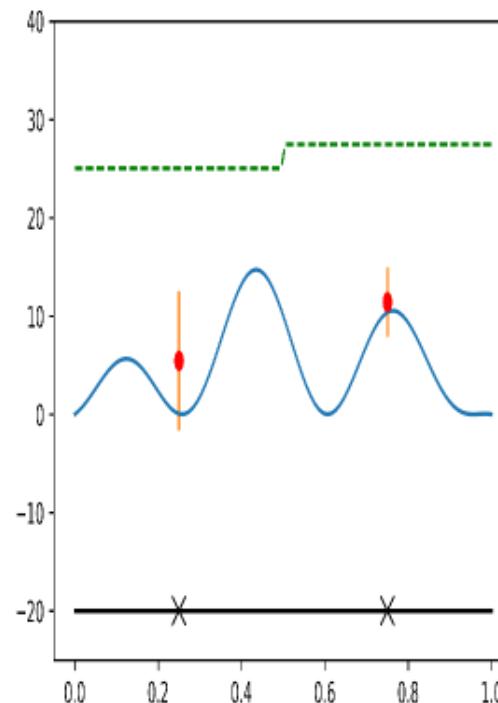


B

GP Optimization with Adaptive Discretization

Idea: Piecewise constant upper-bound confidence function

- ▶ Rely on two bounds:
 - ▶ Compute UCB on the function value at a point.
 - ▶ Upper bound the variation of the function in a region.
- ▶ Maximally select the best region (piecewise constant UCB)

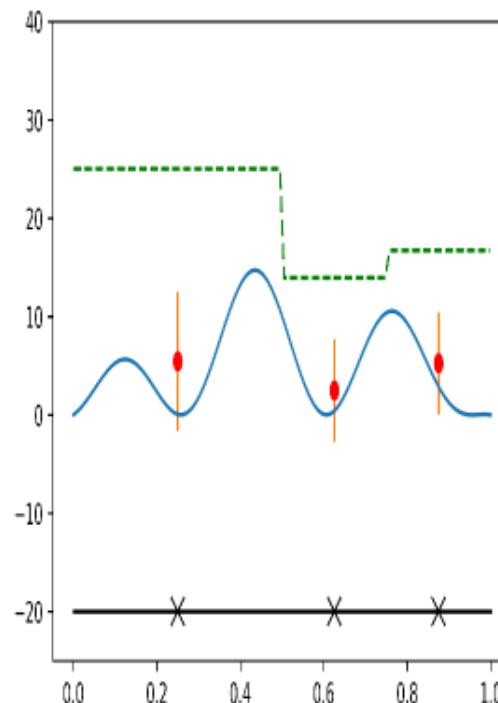


8

GP Optimization with Adaptive Discretization

Idea: Piecewise constant upper-bound confidence function

- ▶ Rely on two bounds:
 - ▶ Compute UCB on the function value at a point.
 - ▶ Upper bound the variation of the function in a region.
- ▶ Maximally select the best region (piecewise constant UCB)
- ▶ Refine the discretization after sufficient number of observations in a region.

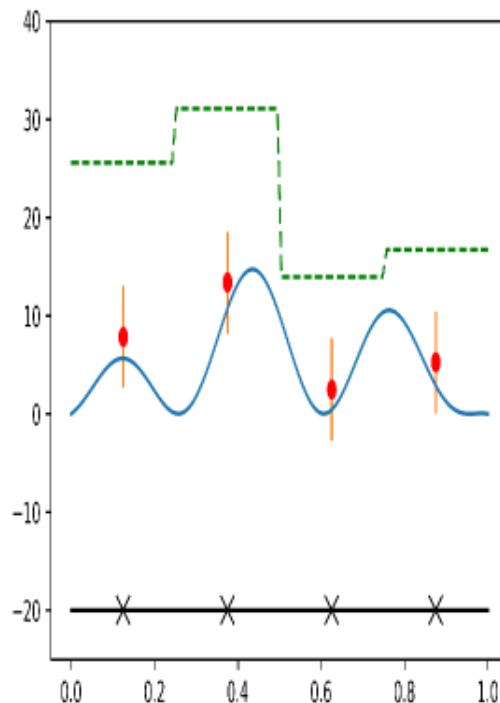


8

GP Optimization with Adaptive Discretization

Idea: Piecewise constant upper-bound confidence function

- ▶ Rely on two bounds:
 - ▶ Compute UCB on the function value at a point.
 - ▶ Upper bound the variation of the function in a region.
- ▶ Maximally select the best region (piecewise constant UCB)
- ▶ Refine the discretization after sufficient number of observations in a region.



8

Algorithm-1: Tree-based Adaptive Discretization

- ▶ Work with a fixed tree of partitions (with fan-out N):

- ▶ Increasingly refined subsets $\mathcal{X}_h = \{x_{h,i} : 1 \leq i \leq N^h\}$
 - ▶ for each $x_{h,i}$, we have a cell

$$\mathcal{X}_{h,i} = \{x \in \mathcal{X} : l(x, x_{h,i}) \leq l(x, x_{h,j}) \quad \forall j \neq i\}$$

- ▶ Assume cells $\mathcal{X}_{h,i}$ satisfy:

- ▶ The radius of $\mathcal{X}_{h,i}$ geometrically decaying with h

$$B(x_{h,i}, \gamma \rho^h) \subset \mathcal{X}_{h,i} \subset B(x_{h,i}, \gamma^{-1} \rho^h)$$

- ▶ For a fixed h , $\cup_i \mathcal{X}_{h,i} = \mathcal{X}$

11/18

Algorithm-1: Tree-based Adaptive Discretization

- ▶ Work with a fixed tree of partitions (with fan-out N):

- ▶ Increasingly refined subsets $\mathcal{X}_h = \{x_{h,i} : 1 \leq i \leq N^h\}$
 - ▶ for each $x_{h,i}$, we have a cell

$$\mathcal{X}_{h,i} = \{x \in \mathcal{X} : l(x, x_{h,i}) \leq l(x, x_{h,j}) \quad \forall j \neq i\}$$

- ▶ Assume cells $\mathcal{X}_{h,i}$ satisfy:

- ▶ The radius of $\mathcal{X}_{h,i}$ geometrically decaying with h

$$B(x_{h,i}, \gamma \rho^h) \subset \mathcal{X}_{h,i} \subset B(x_{h,i}, \gamma^{-1} \rho^h)$$

- ▶ For a fixed h , $\cup_i \mathcal{X}_{h,i} = \mathcal{X}$

- ▶ Can be constructed if $\mathcal{X} = [a, b]^D$



11/18

Algorithm-1: Tree-based Adaptive Discretization

- ▶ In round t , maintain a set of leaf nodes \mathcal{L}_t partitioning \mathcal{X}



12/18

Algorithm-1: Tree-based Adaptive Discretization

- ▶ In round t , maintain a set of leaf nodes \mathcal{L}_t partitioning \mathcal{X}
- ▶ Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and V_h is a h.p.u.b on the maximum function variations



12/18

Algorithm-1: Tree-based Adaptive Discretization

- ▶ In round t , maintain a set of leaf nodes \mathcal{L}_t partitioning \mathcal{X}
- ▶ Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and V_h is a h.p.u.b on the maximum function variations
- ▶ Select a node/point from \mathcal{L}_t by maximizing index

$$I_t(x_{h,i}) = \bar{U}_t(x_{h,i}) + V_h,$$

12/18

Algorithm-1: Tree-based Adaptive Discretization

- ▶ In round t , maintain a set of leaf nodes \mathcal{L}_t partitioning \mathcal{X}
- ▶ Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and V_h is a h.p.u.b on the maximum function variations
- ▶ Select a node/point from \mathcal{L}_t by maximizing index

$$I_t(x_{h,i}) = \bar{U}_t(x_{h,i}) + V_h,$$

- ▶ *Refine:* If $\beta_n \sigma_{t-1}(x_{h_t,i_t}) \leq V_h$, then the node \mathcal{X}_{h_t,i_t} is expanded into and replaced by its N children nodes

12/18

Algorithm-1: Tree-based Adaptive Discretization

- ▶ In round t , maintain a set of leaf nodes \mathcal{L}_t partitioning \mathcal{X}
- ▶ Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and V_h is a h.p.u.b on the maximum function variations
- ▶ Select a node/point from \mathcal{L}_t by maximizing index

$$I_t(x_{h,i}) = \bar{U}_t(x_{h,i}) + V_h,$$

- ▶ *Refine:* If $\beta_n \sigma_{t-1}(x_{h_t,i_t}) \leq V_h$, then the node \mathcal{X}_{h_t,i_t} is expanded into and replaced by its N children nodes
- ▶ *Evaluate:* Otherwise, observe the noisy function value $y_t = f(x_{h_t,i_t}) + \eta_t$ and update the posterior distribution of f

Algorithm-1: Tree-based Adaptive Discretization

- ▶ In round t , maintain a set of leaf nodes \mathcal{L}_t partitioning \mathcal{X}
- ▶ Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and V_h is a h.p.u.b on the maximum function variations
- ▶ Select a node/point from \mathcal{L}_t by maximizing index

$$I_t(x_{h,i}) = \bar{U}_t(x_{h,i}) + V_h,$$

- ▶ *Refine*: If $\beta_n \sigma_{t-1}(x_{h_t,i_t}) \leq V_h$, then the node \mathcal{X}_{h_t,i_t} is expanded into and replaced by its N children nodes
- ▶ *Evaluate*: Otherwise, observe the noisy function value $y_t = f(x_{h_t,i_t}) + \eta_t$ and update the posterior distribution of f
- ▶ Complexity: $\mathcal{O}(Nn^4 \log n + NDn \log n)$

12/18

Algorithm-1: Regret Bounds

Theorem-1

Under mild technical conditions on $K(\cdot, \cdot)$, with high probability we have that

$$\begin{aligned}\mathcal{R}_n &= \sum_{t \leq n} f(x^*) - f(x_t) \leq \min\{\mathcal{O}(\sqrt{n\gamma_n \log n}), \tilde{\mathcal{O}}(n^{1-\frac{\alpha}{\tilde{D}+2\alpha}})\} \\ \mathcal{S}_n &= f(x^*) - f(x(n)) \leq \tilde{\mathcal{O}}(n^{-\alpha/(\tilde{D}+2\alpha)})\end{aligned}$$

- ▶ \tilde{D} is a notion of dimension of the near optimal regions of f
 - ▶ Regret bound is a random variable, a function of dimensionality of f around its maxima
 - ▶ For almost all realization of f , $\tilde{D} \leq D$

13/18

Improved bounds for Matérn kernels

- ▶ Matérn kernels parameterized by $\nu = m + 1/2$:

$$K_{\nu}^{\text{Matérn}}(x, x') = K(0)(1 + \sum_{i=1}^m a_i \|x - x'\|^i) e^{-c_1 \sqrt{\nu} \|x - x'\|},$$

- ▶ Improved bound for all when $D \geq \nu - 1$

14/18

Improved bounds for Matérn kernels

- Matérn kernels parameterized by $\nu = m + 1/2$:

$$K_{\nu}^{\text{Matérn}}(x, x') = K(0)(1 + \sum_{i=1}^m a_i \|x - x'\|^i) e^{-c_1 \sqrt{\nu} \|x - x'\|}, \quad \left\{ \begin{array}{l} \nu = 1/2 \\ \nu = 3/2 \\ \nu = 5/2 \end{array} \right.$$



- Improved bound for all when $D \geq \nu - 1$

14/18

Improved bounds for Matérn kernels

- ▶ Matérn kernels parameterized by $\nu = m + 1/2$:

$$K_{\nu}^{\text{Matérn}}(x, x') = K(0)(1 + \sum_{i=1}^m a_i \|x - x'\|^i) e^{-c_1 \sqrt{\nu} \|x - x'\|},$$

When $\mathcal{X} \subset [0, 1]^D$ and $\nu > 2$, $\tilde{D} \leq 3D/4$ with high prob

- ▶ Improved bound for all when $D \geq \nu - 1$

14/18

Improved bounds for Matérn kernels

- ▶ Matérn kernels parameterized by $\nu = m + 1/2$:

$$K_{\nu}^{\text{Matérn}}(x, x') = K(0)(1 + \sum_{i=1}^m a_i \|x - x'\|^i) e^{-c_1 \sqrt{\nu} \|x - x'\|},$$

- ▶ Our bounds improve on the existing bounds in two ways:
 - ▶ For $\nu = 1/2$, we provide the first explicit sublinear bounds on cumulative regret.
 - ▶ For $\nu = 3/2$ and $\nu = 5/2$ our bounds are tighter for $D \geq 2$.
- ▶ Improved bound for all when $D \geq \nu - 1$

14/18

Algorithm-1: Regret Bounds for Noiseless Observations

Theorem-2

If in addition to the assumptions of Theorem-1, we further assume $\sigma = 0$. With high probability,

$$\mathcal{R}_n \leq \tilde{\mathcal{O}}(n^{1-\frac{\alpha}{\tilde{D}}}) \quad (3)$$

$$\mathcal{S}_n \leq \tilde{\mathcal{O}}(n^{-\alpha/\tilde{D}}) \quad (4)$$

if $\tilde{D} > 0$, and

$$\mathcal{R}_n \leq \tilde{\mathcal{O}}(1) \quad (5)$$

$$\mathcal{S}_n \leq \tilde{\mathcal{O}}(e^{-c_1 \log(1/\rho)n}) \quad (6)$$

if $\tilde{D} = 0$ and $h_{\max} = \Omega(n)$, for some constant $c_1 > 0$.

Comparison with BaMSOO

- ▶ Our tree algorithm motivated by Bayesian Multi-Scale Optimistic Optimization (BaMSOO)
 - ▶ BaMSOO relies on an adaptive construction of a partition tree
 - ▶ BaMSOO only works with noiseless observations

Our method has some advantages:

- ▶ BaMSOO's regret analysis only under very restrictive conditions on K , e.g. excludes Matérn $\nu = 1/2$
- ▶ BaMSOO has strictly worse simple regret, S_n , for some fairly practical cases

Summary of Results

1. Algorithmic Contribution: Adaptive discretization for GP

- ▶ Opportunistically adapts to the (simple) structure of f
- ▶ Strictly lower complexity ($\mathcal{O}(D)$) for $\mathcal{X} \subset \mathbb{R}^D$

2. Analytic Contribution: Dimensional-type regret bound

- ▶ As good or better regret bound than prior work
- ▶ First sublinear bound for exponential kernels (Matérn- $\nu = 1/2$)
- ▶ Strictly tighter bounds for Matérn kernels if $D > \nu - 1$

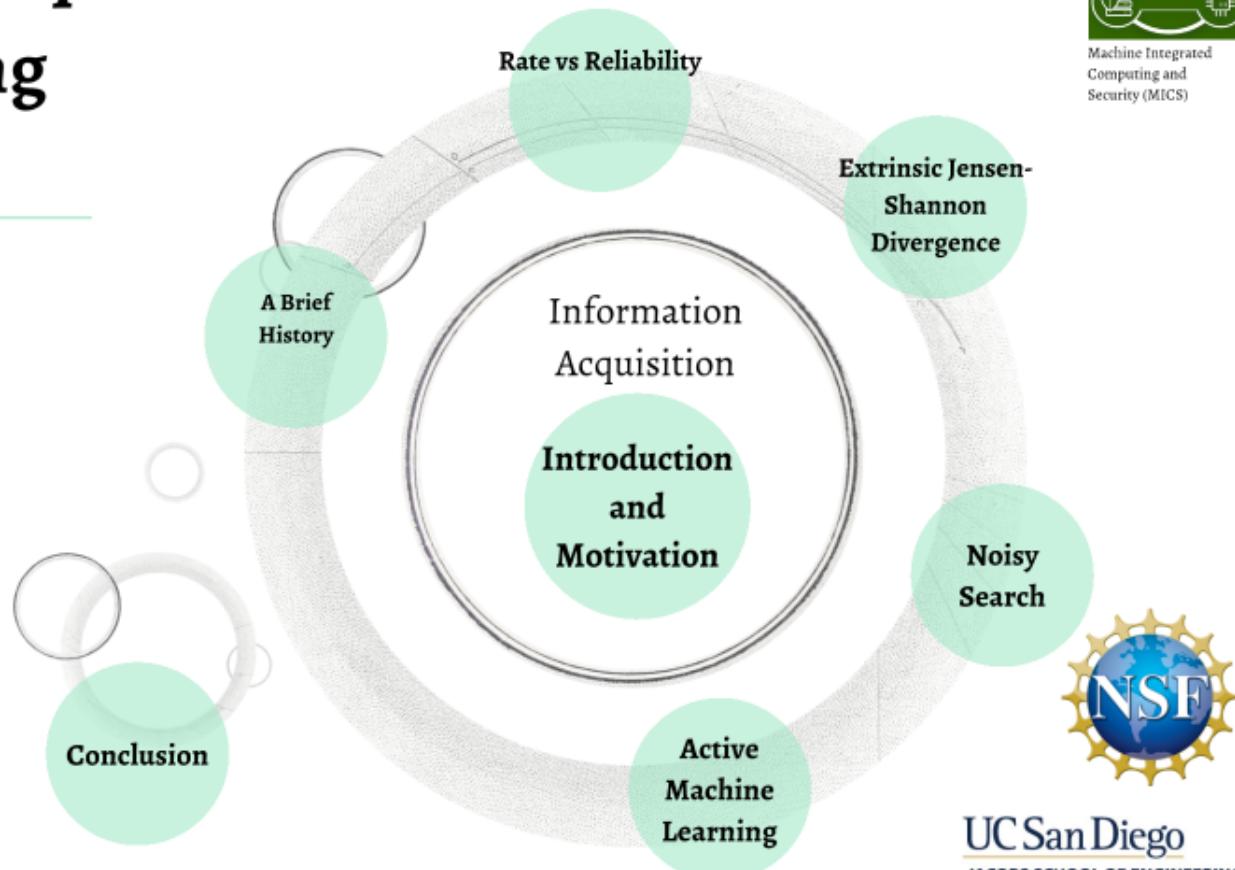
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz



Machine Integrated
Computing and
Security (MICS)



UC San Diego
JACOBS SCHOOL OF ENGINEERING
Center for Wireless Communications

Summary & Extensions

Information Acquisition
and Evolution of Belief
Vector

Generalized notions of rate
and reliability to acquisition
rate--reliability

Establishing acquisition
rate-- reliability trade-off
reminiscent of that of codes

**Uncertainty measure
beyond entropy**

**More dynamic
notion of uncertainty**

**Converses that
account for the
unpredictable
component of the
state**



DetecDrone

Drones that Actively Seek Information and Learn



Intelli-Ranch

Camera Enabled Drones
Monitoring Livestock
Alarm and Rescue
Cross-referencing



Wide Area Object Tracking

Camera Enabled Drones
Flight Path Optimization for Maximum Battery Life
Multi-resolution Mapping



Assisted Living

Camera Enabled Drones
Voice Activation
Augmented Sensing with Easy Voice Control

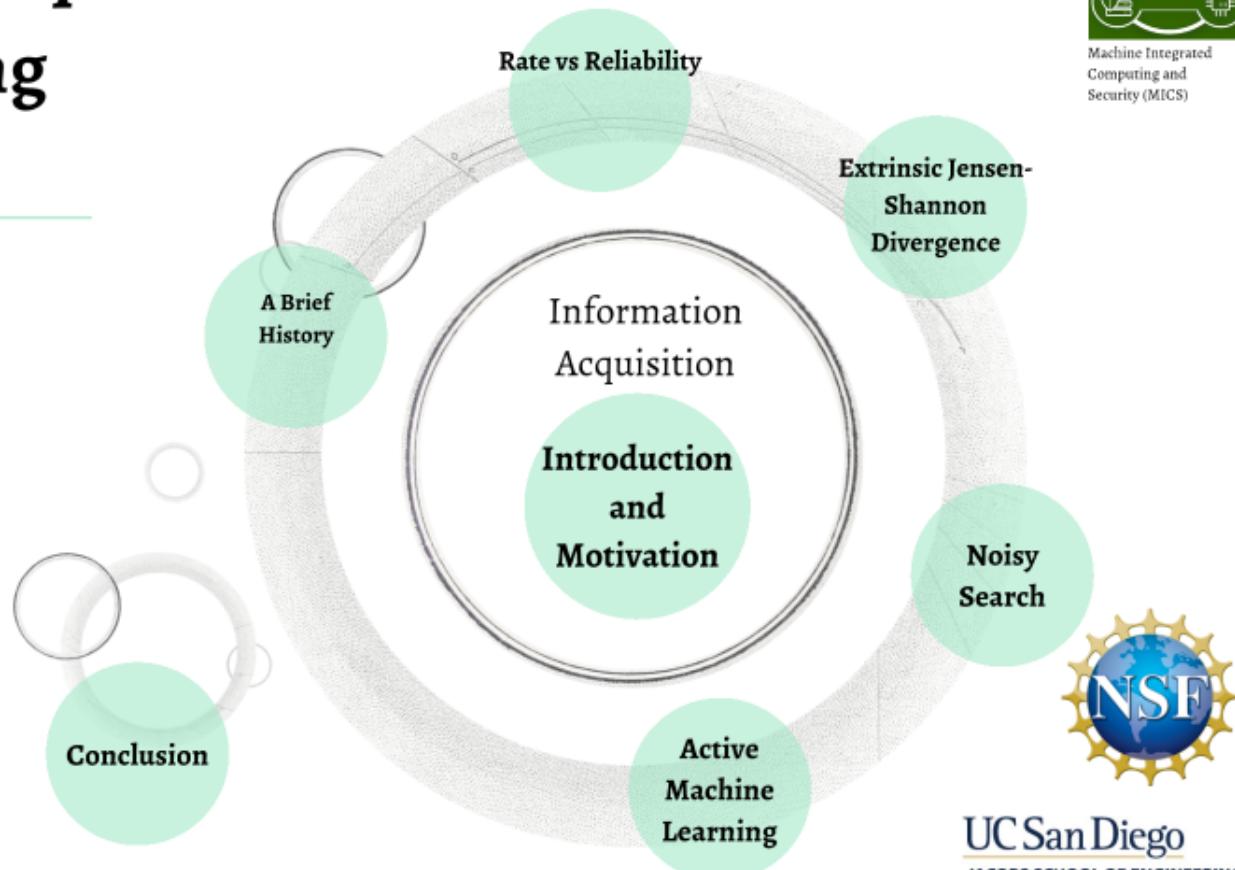
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz



Machine Integrated
Computing and
Security (MICS)



UC San Diego
JACOBS SCHOOL OF ENGINEERING
Center for Wireless Communications