Are there physical limits on machine learning?

Dr. Fanglin Bao
baofanglin@westlake.edu.cn

Nov. 15, 2024, Beijing

Outline

Errors in AI

Applications of AI & Why Does Error Matter

Error bound: problem setup

Information Theory & Photon Statistics

Error bound: state of the art

Mutual Information & Quantum Chernoff Bound

Physical limits on hypothesis testing (classification)

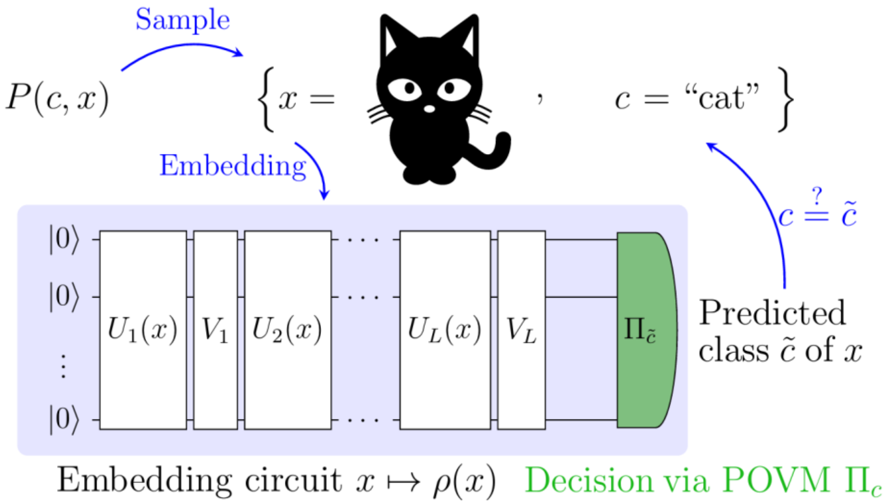
Exact Error Probability & Asymptotic Behavior & Symmetries

Physical limits on parameter estimation (regression)

Correspondence Problem & Position Estimation

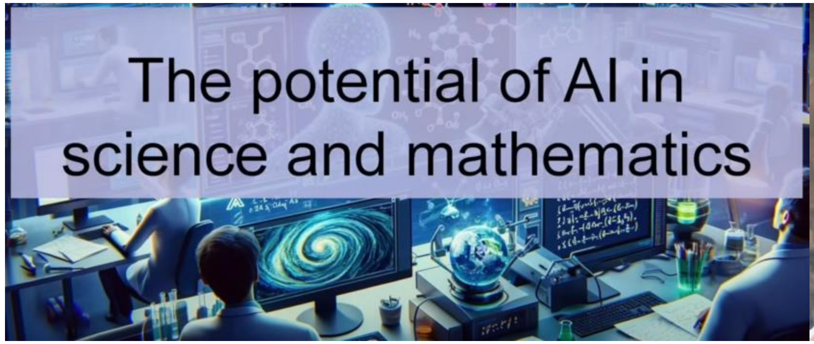
AI for Theoretical Sciences

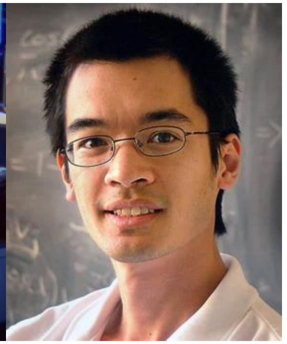
(a) Quantum state classification
(b) Classification of classical data
(c) Quantum channel discrimination
…



L. Banchi, et al, PRX QUANTUM **2**,040321(2021)

(d) Symbolic regression (AI physicist, Data -> Eq.)
(e) Symbolic mathematics (Eq. -> Eq.)
(f) Conjecture generation and formal proof
…





Terence Tao
Fields Medalist

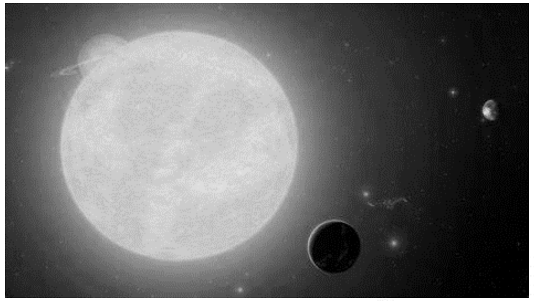
AI could be wrong!

AI could be wrong

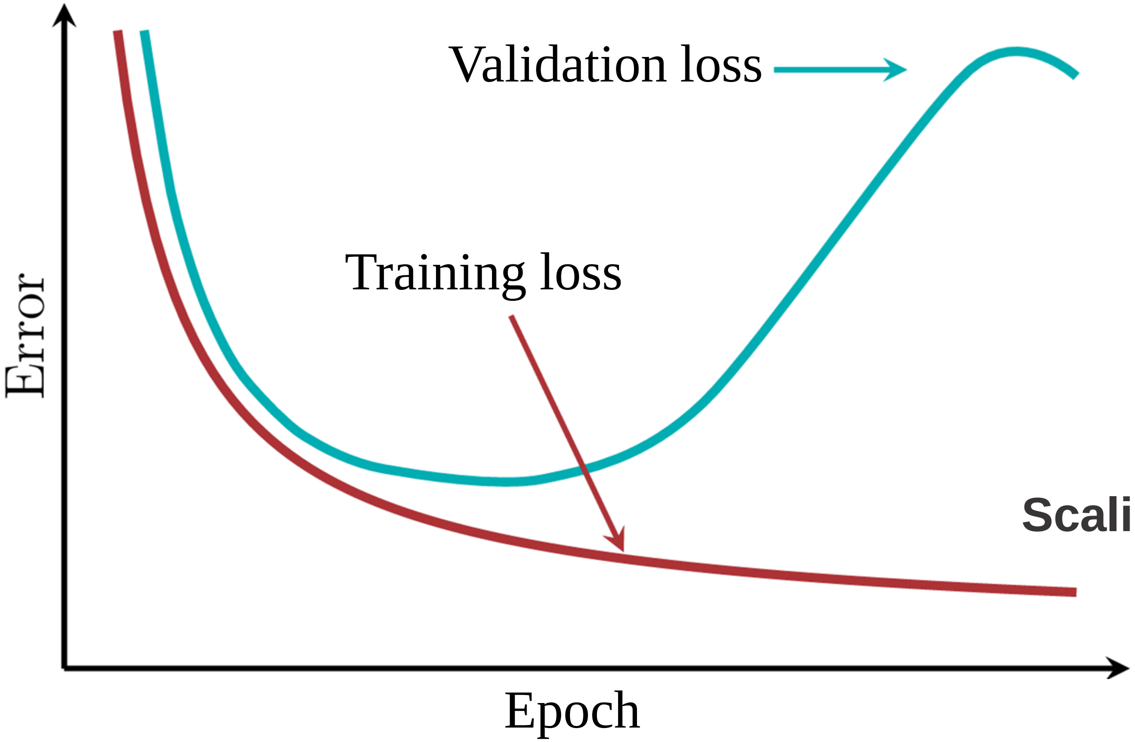
Muffin or Chihuahua?
(False vs. True theory)

Error bound

Exoplanet searching



Object classification
Vowel recognition
Electronic skin ...

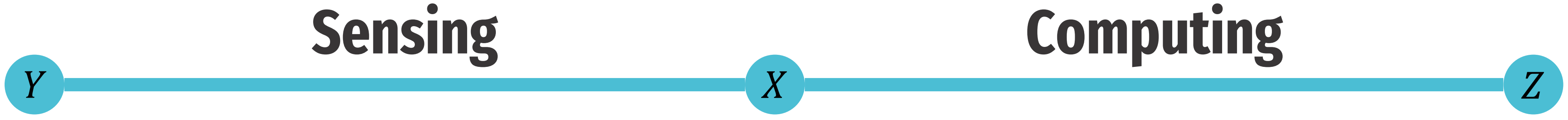


Scaling law?

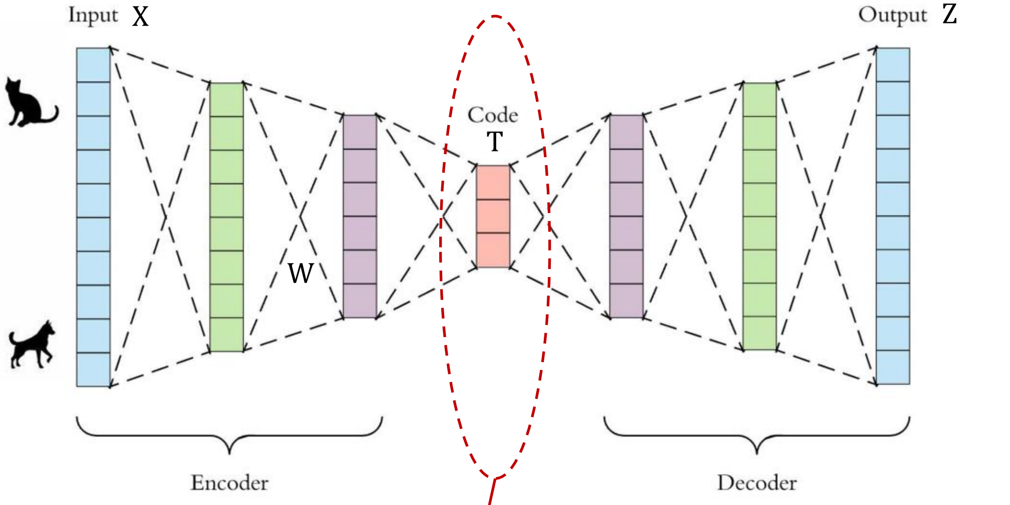
When errors occur, is it because the model is not well trained, or physical laws and resources set a fundamental limit?

Outline

Problem setup: information bottleneck



ρ_y $\quad p(x|y) = tr(\rho_y M_x)$



$$\min_{w} I(T;X), s.t.\ I(T;Z) > I(X;Z) - \epsilon$$

Information bottleneck $\sim \boldsymbol{\rho_y}$ & $\boldsymbol{M_x}$

G. E. Hinton, et al, *Science* **313**,504-507(2006)
for data representation

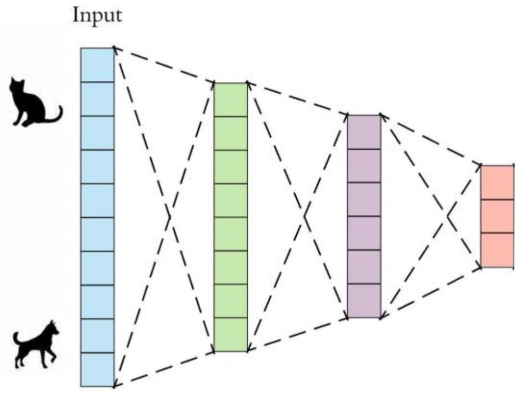
R. Shwartz, et al, *arXiv:*1703.00810 (2017)
for information bottleneck

Problem setup: data processing inequality

Sensing — **Computing**

Y — X — $\tilde{Y}$

ρ_y $\quad$ $p(x|y) = tr(\rho_y M_x)$ $\quad$ $p(y|x)$



What
&
Where

Computing (data processing) does not increase information.
(not just mutual information, but also Fisher information)

Problem setup: data processing inequality

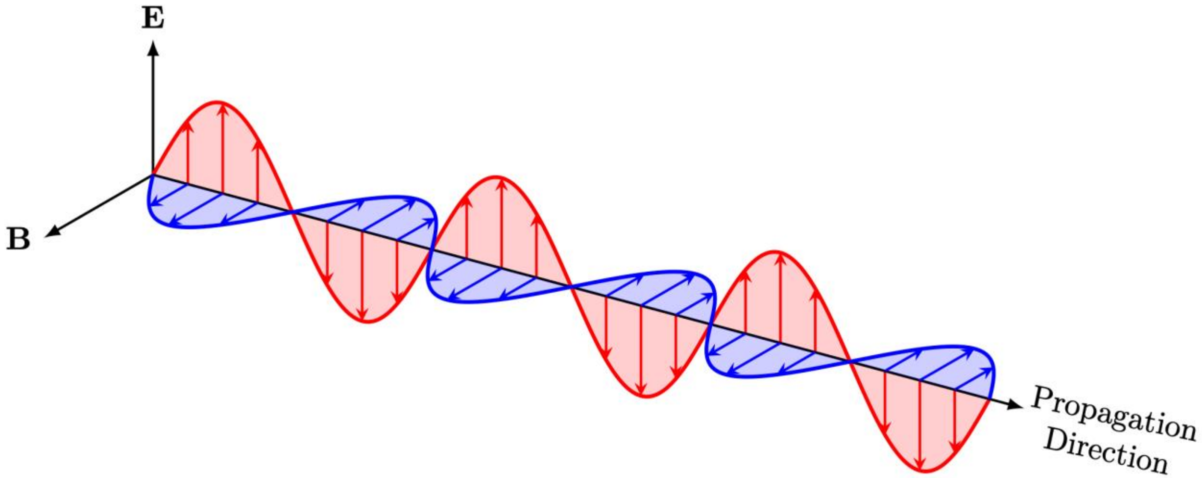
I. Shumailov, et al, *Nature* **631**, 755–759 (2024)

Garbage in, garbage out

- It highlights the importance of data quality against dataset size
- In machine perception, measurement optimization is essential.

Problem setup: wave-particle duality

Light source	Wave	Particle



Optica field interacts with a sensor

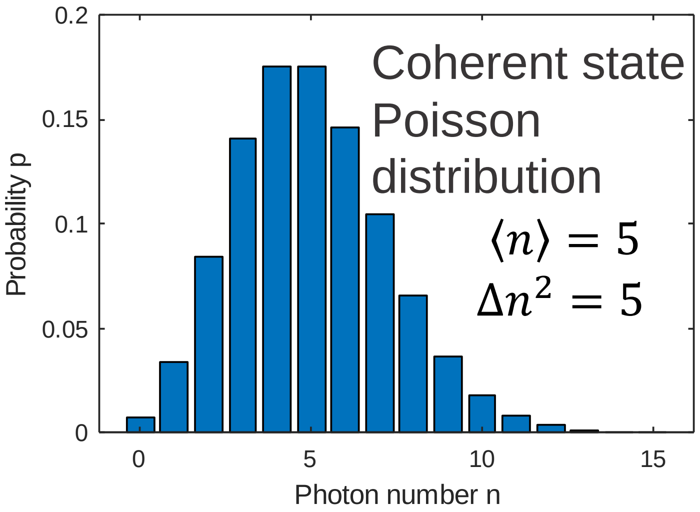
in the unit of 'photon' ...

Detector

Problem setup: photon statistics

Poissonian light

$$\Delta n^2 = \langle n \rangle$$

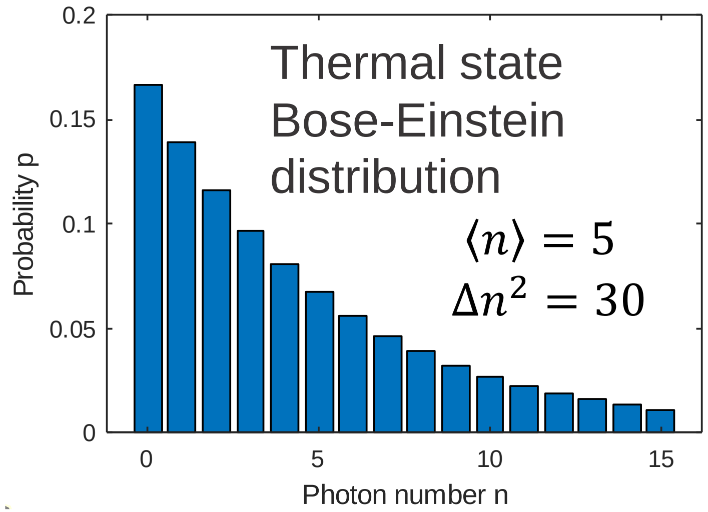


$$\hat{a}|\alpha\rangle = \alpha|\alpha\rangle, \langle n \rangle = |\alpha|^2$$

$$p(n) = |\langle n|\alpha\rangle|^2 = e^{-\langle n \rangle} \frac{\langle n \rangle^n}{n!}$$

Super-Poissonian light

$$\Delta n^2 > \langle n \rangle$$

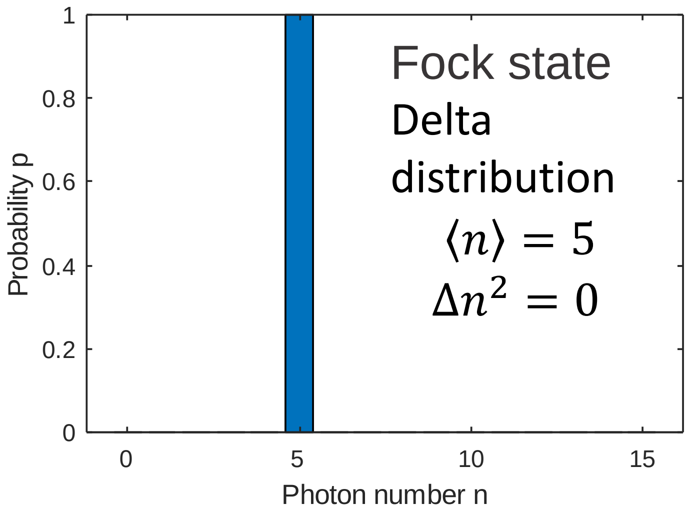


$$\rho = \int \frac{1}{\pi \langle n \rangle} e^{-|\alpha|^2/\langle n \rangle} |\alpha\rangle\langle\alpha| \mathrm{d}^2\alpha$$

$$p(n) = \langle n|\rho|n\rangle = \frac{\langle n \rangle^n}{(\langle n \rangle + 1)^{n+1}}$$

Sub-Poissonian light

$$\Delta n^2 < \langle n \rangle$$

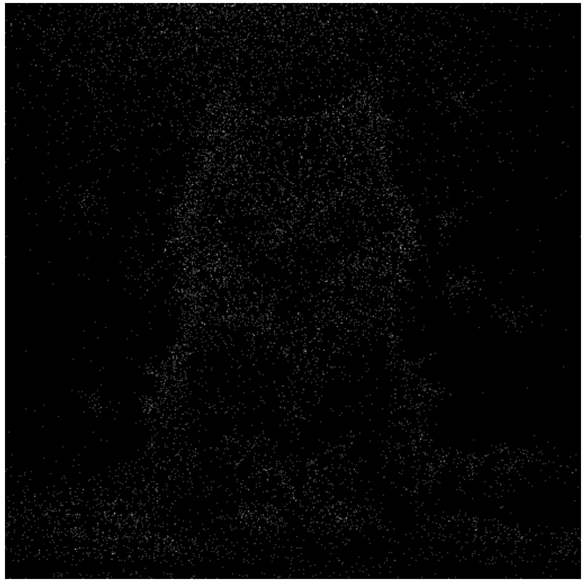


$$\hat{n}|n'\rangle = n'|n'\rangle$$

$$p(n) = \delta_{nn'}$$

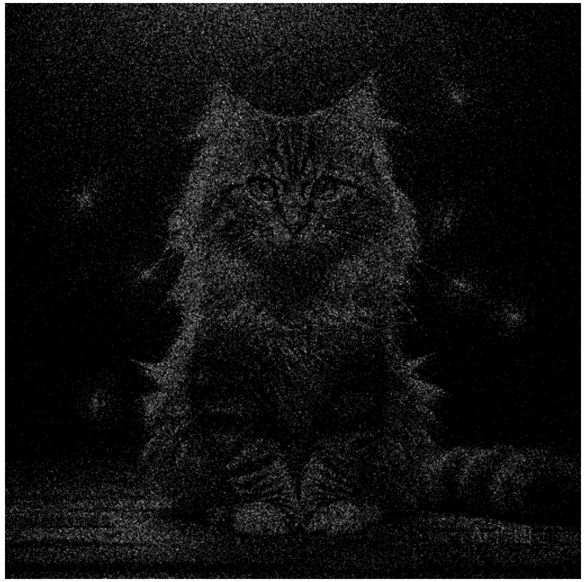
Problem setup: imaging

Ultrafast and ultrasensitive camera



10K photons

...

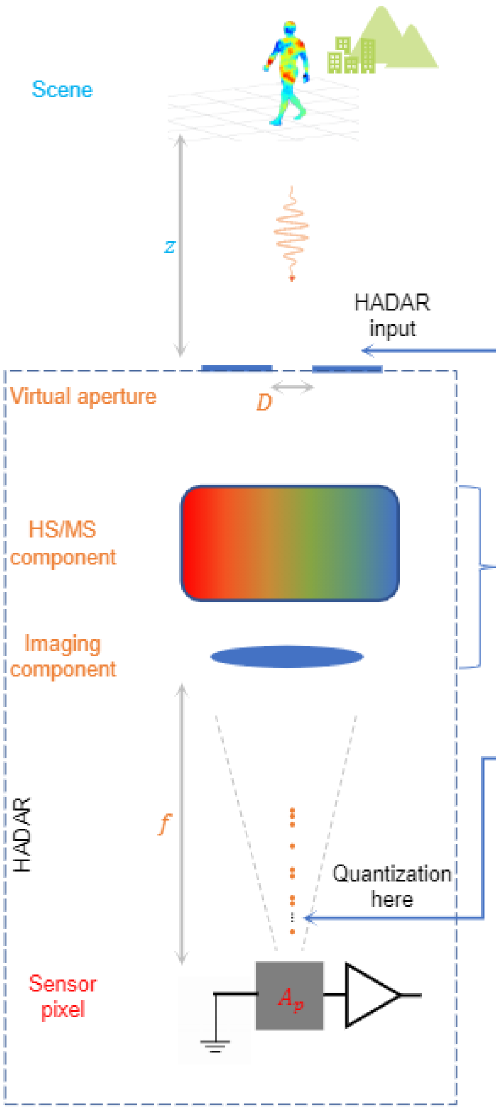


100K photons

1M photons

Each photon adds a bright dot in the images (512 * 512)

Problem setup: a comprehensive model of sensing



Heat signal

$$S_{\alpha\nu} = e_{\alpha\nu}B_\nu(T_\alpha) + (1 - e_{\alpha\nu})X_{\alpha\nu}$$

$$X_{\alpha\nu} = \sum_{\beta\neq\alpha} V_{\alpha\beta}S_{\beta\nu}$$

T : Temperature e : Emissivity X : Texture
η^o : Optics efficiency
W : Number of bands/filters
K_ν : Self emission spectrum of the sensor
η^Q : Quantum efficiency
R_ν : Spectral responsivity
ξ : Dark noise/count
ζ : Gain ϵ: Electronic charge
σ : std of electronic noise

Spectrum of mean incident photon flux to HADAR during time τ_c within bandwidth $\delta\nu$

$$\lambda_\nu = \phi_\nu \cdot \delta\nu\tau_c = \phi_\nu/c,$$

$$\tau_c \equiv 1/c\delta\nu, \quad a = t/\tau_c, \quad \phi_\nu = S_\nu \cdot \frac{\pi A_p}{4(f/D)^2 + (f/z)^2}$$

Dispersive/diffraction spectroscopy	Wavelength-multiplexing spectroscopy
Output spectrum for band q $\lambda_{q\nu} = \eta^o_\nu \mathcal{T}_{q\nu}\lambda_\nu$ for $\{q_1, q_2, \dots, q_W\}$ Each $\mathcal{T}_{q\nu}$ is a bandpass curve	Output spectrum for 'filter' q $\lambda_{q\nu} = \eta^o_\nu[\mathcal{T}_{q\nu}(\lambda_\nu - K_\nu) + K_\nu]$ for $\{q_1, q_2, \dots, q_W\}$ Each $\mathcal{T}_{q\nu}$ is a general transmittance curve

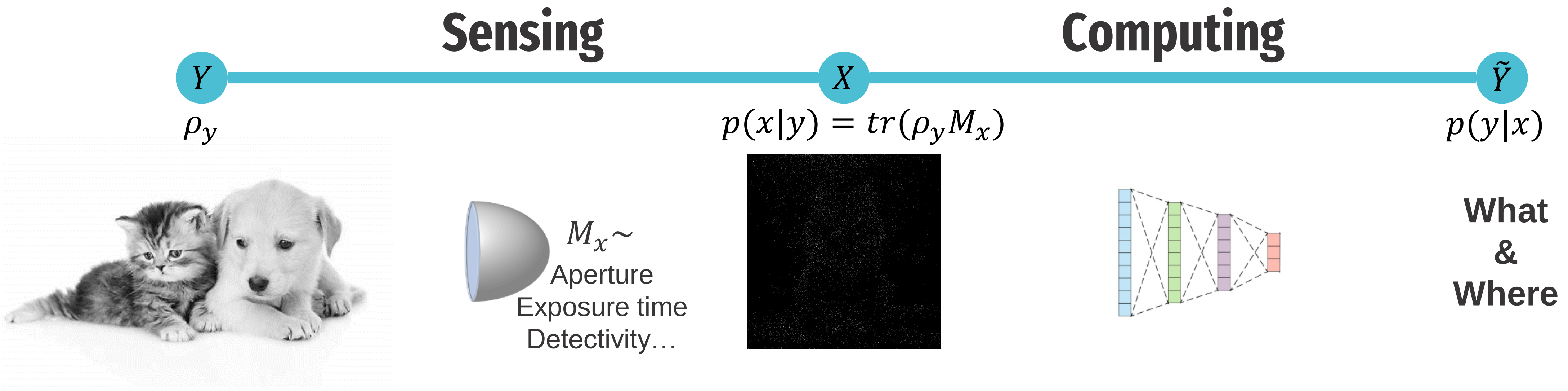
Actual photon number n onto the sensor during time t within bandwidth $\delta\nu$

$$\mathcal{P}_{q\nu}(n) = C^n_{n+a-1}\frac{\lambda^n_{q\nu}}{(\lambda_{q\nu}+1)^{n+a}}$$

Photoelectric detector	Thermal detector	SPD
Excited photocurrent i_q $i_q = \sum_{\nu=\nu_{min}}^{\nu_{max}} \eta^Q_\nu \cdot \epsilon \cdot n_{q\nu}/t + \xi$ Output signal current I_q $I_q = \mathcal{N}(\zeta i_q, \sigma)$	Induced voltage change $u_q = \sum_{\nu=\nu_{min}}^{\nu_{max}} R_\nu \cdot hc\nu \cdot n_{q\nu}/t + \xi$ Output $U_q = \mathcal{N}(\zeta u_q, \sigma)$	Output counts $c_q = \sum_{\nu=\nu_{min}}^{\nu_{max}} \eta^Q_\nu \cdot n_{q\nu} + \xi$

T : Temperature of the target α
e : Spectral emissivity of the target α
X : Geometric texture of the target α
z : Distance of the target α
S : Heat signal
ν : Wave number
B : Blackbody radiation
V : Thermal lighting factor
c : Speed of light in vacuum
λ: Mean photon number in coherent time
τ_c : Coherence time
$\delta\nu$: Ultra fine bandwidth
f : Focal length
D : Aperture diameter
A_p : Pixel area
t : Measurement time
η^o : Optics efficiency
W : Number of bands/filters
K_ν : Self emission of the sensor
might be zero if no back reflection
$\mathcal{T}_{q\nu}$: Transmittance curve
becomes $\delta_{q\nu}$ when using prisms etc.
R_ν : Responsivity/quantum efficiency
ξ : Electronic noise with mean $\bar{\xi}$ and std σ
$\mathcal{P}_{q\nu}$: Photon statistics
C^n_{n+a-1}: Binomial coefficient

Problem statement



For a given system ($\boldsymbol{\rho_y}$) and a given task ($\boldsymbol{y}$ = {'Cat' / 'Dog', distance}),

- What's the distance metric to depict the information bottleneck, subject to certain symmetries (translation, rotation, scaling, etc.)?
- What's the **structure of the distance metric** (how is it related to physical parameters)?
- How many photons are needed in the measurement with a given sensor?
- What's the **optimal measurement** requiring the least photons?

Physical limit matters

(1) With the physical limit, we can quantify/score a specific model for a given task.

Training a chatbot can use as much electricity as a neighborhood consumes in a year.

(2) We can optimize measurement to improve AI.

Photonic information vs. electronic information

(3) We can design public policies and industrial standards for machine perception.

My self-driving car can
be as fast as a rocket.

No bragging.
That's against (physics) laws!

Outline

Errors in AI

Applications of AI & Why Does Error Matter

Error bound: problem setup

Information Theory & Photon Statistics

Error bound: state of the art

Mutual Information & Quantum Chernoff Bound

Physical limits on hypothesis testing (classification)

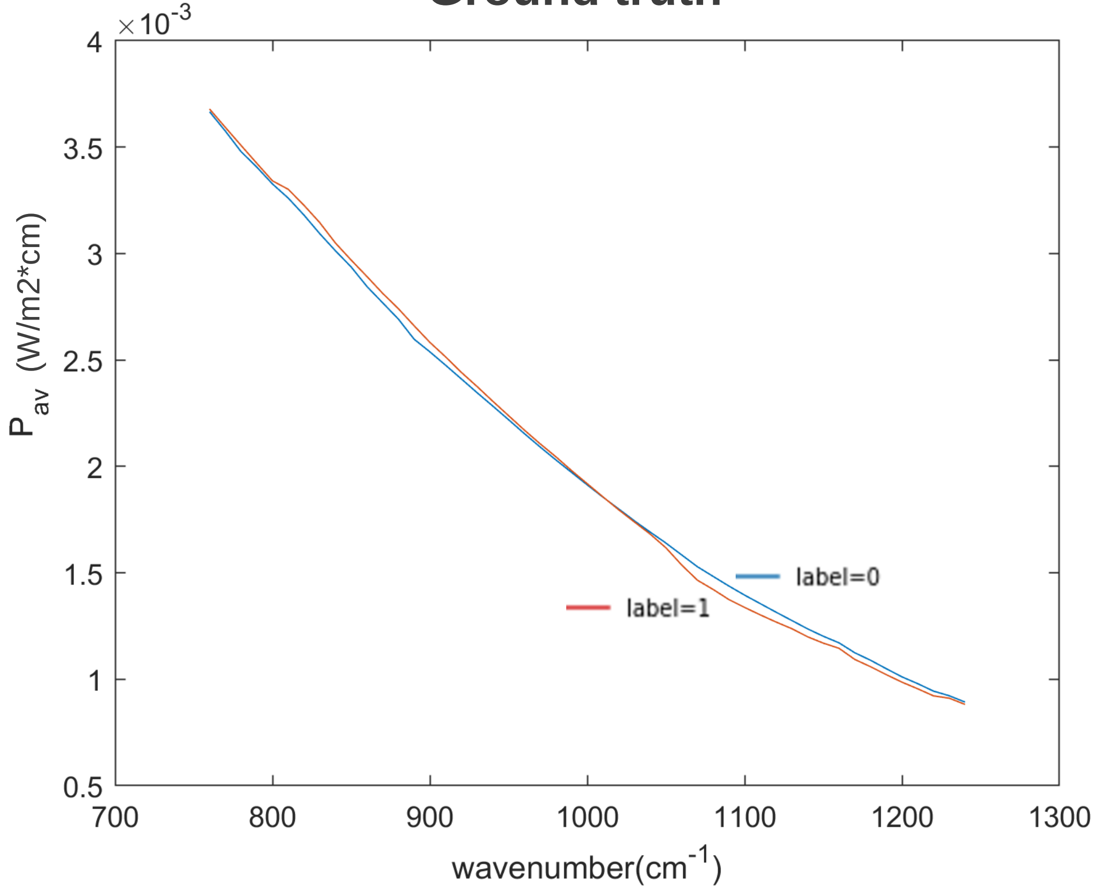
Exact Error Probability & Asymptotic Behavior & Symmetries

Physical limits on parameter estimation (regression)

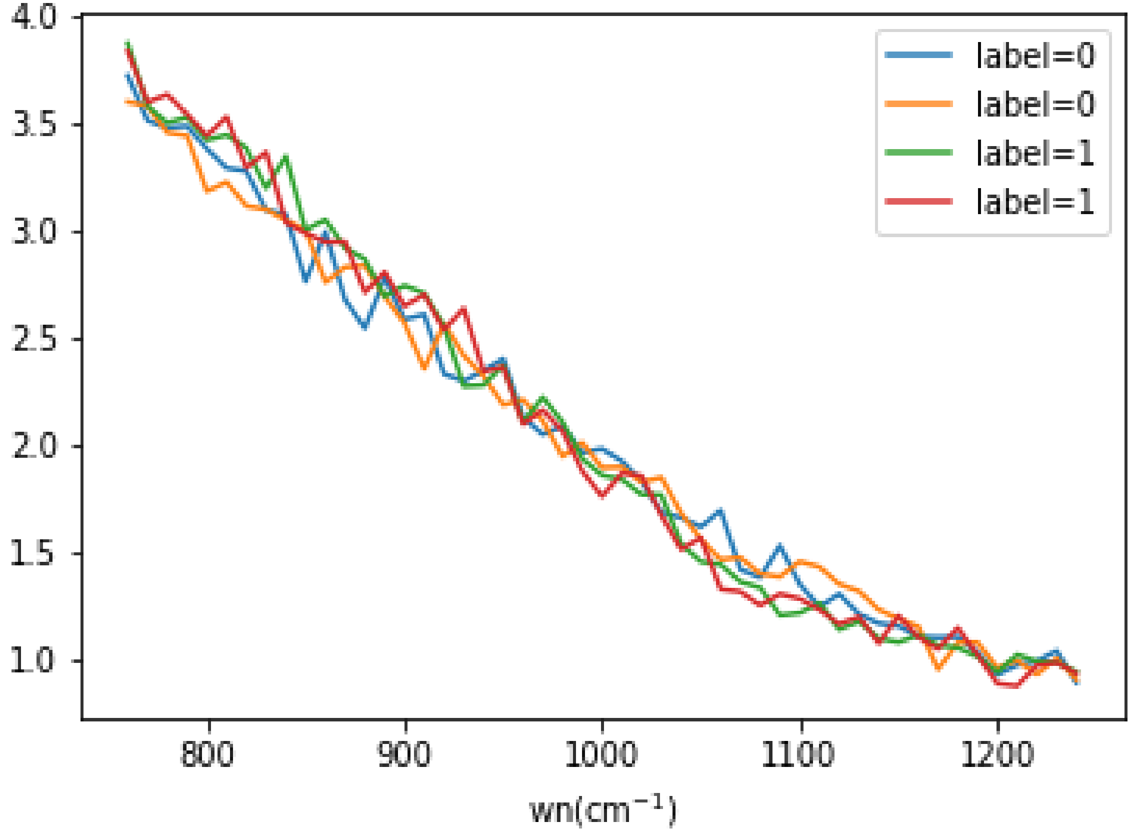
Correspondence Problem & Position Estimation

Experiment: classification of two spectra

Ground truth



Experimental data

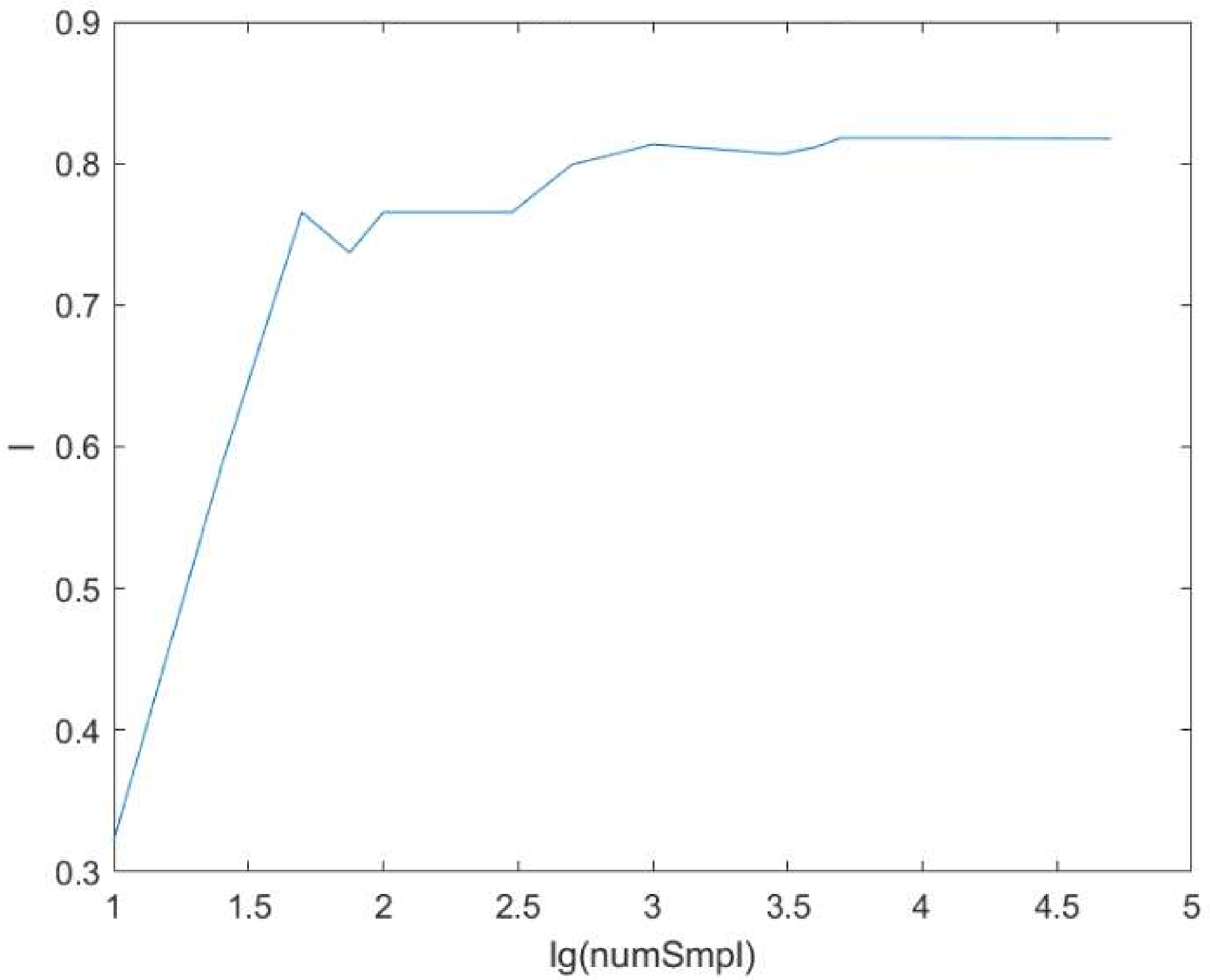


Experiment: convergent accuracy

Shannon's information about the object

$$I = \left(-\log\frac{1}{2}\right) - (-\log P_{acc})$$

$$P_{acc} = \frac{\Pr(0|0) + \Pr(1|1)}{2}$$

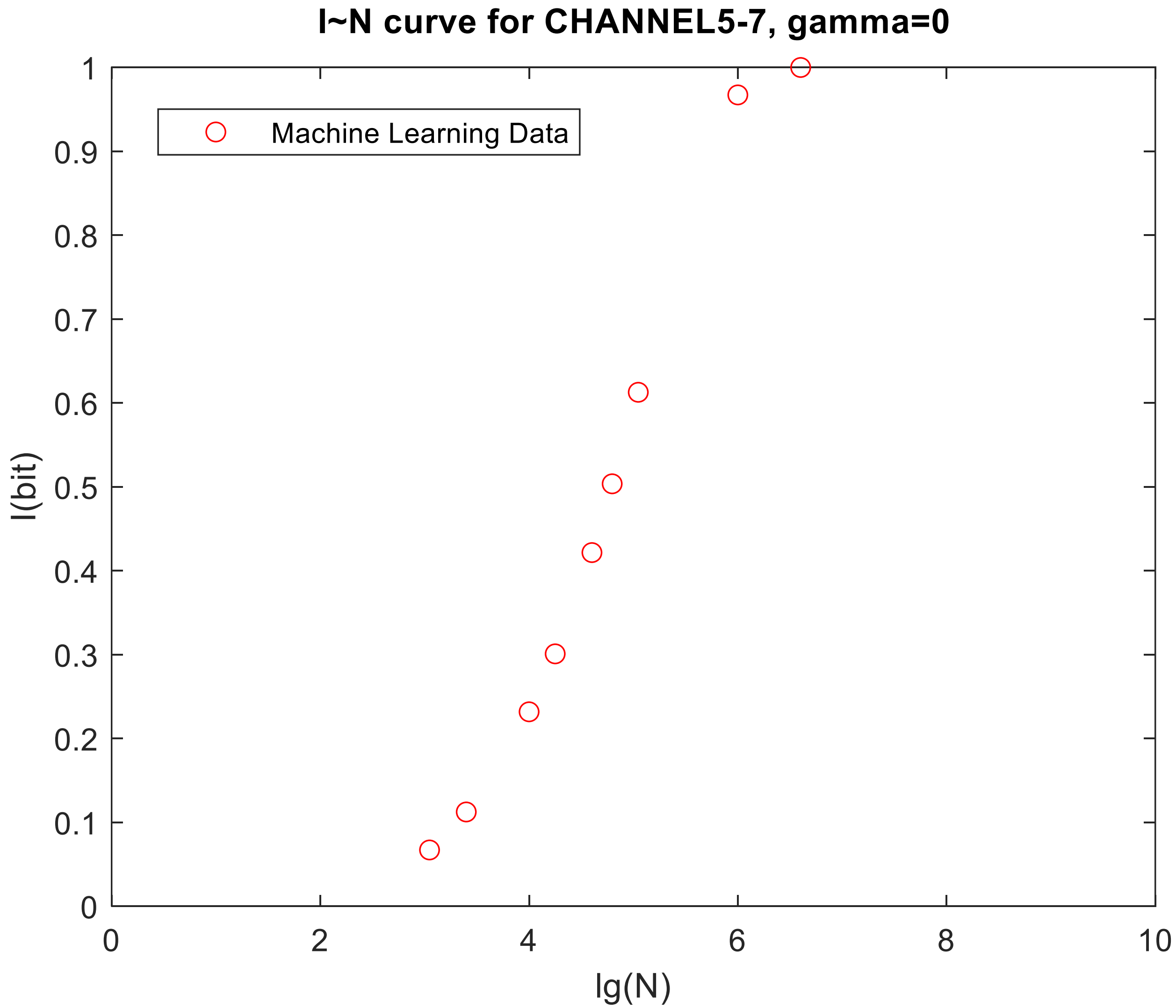


Experimental bound of accuracy

Shannon's information about the object

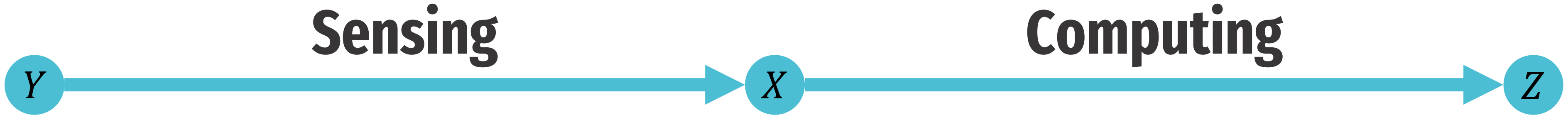
$$I = \left(-\log\frac{1}{2}\right) - (-\log P_{acc})$$

$$P_{acc} = \frac{\Pr(0|0) + \Pr(1|1)}{2}$$

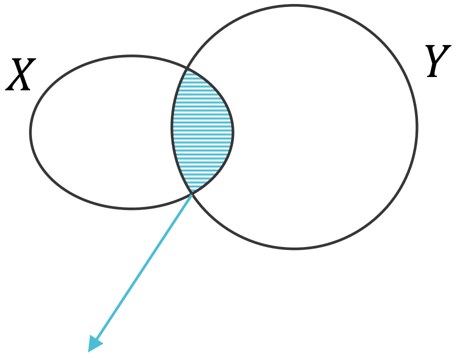


SOTA 1: Mutual information

- Markov chain

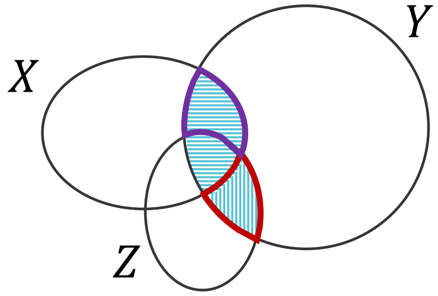


- Mutual information



$$I(X;Y) = \iint p(x,y) \log \frac{p(x,y)}{p(x) \cdot p(y)} \mathrm{d}x\mathrm{d}y \geq 0$$

- Data processing inequality



$$H[(X,Z);Y] = I(X;Y) + H[(Y;Z)|X] \quad = 0$$
$$= I(Z;Y) + H[(X;Y)|Z] \quad \geq 0$$

$$I(X;Y) \geq I(Z;Y)$$

Mutual information has no order information

Example

$Z = NOT(X)$

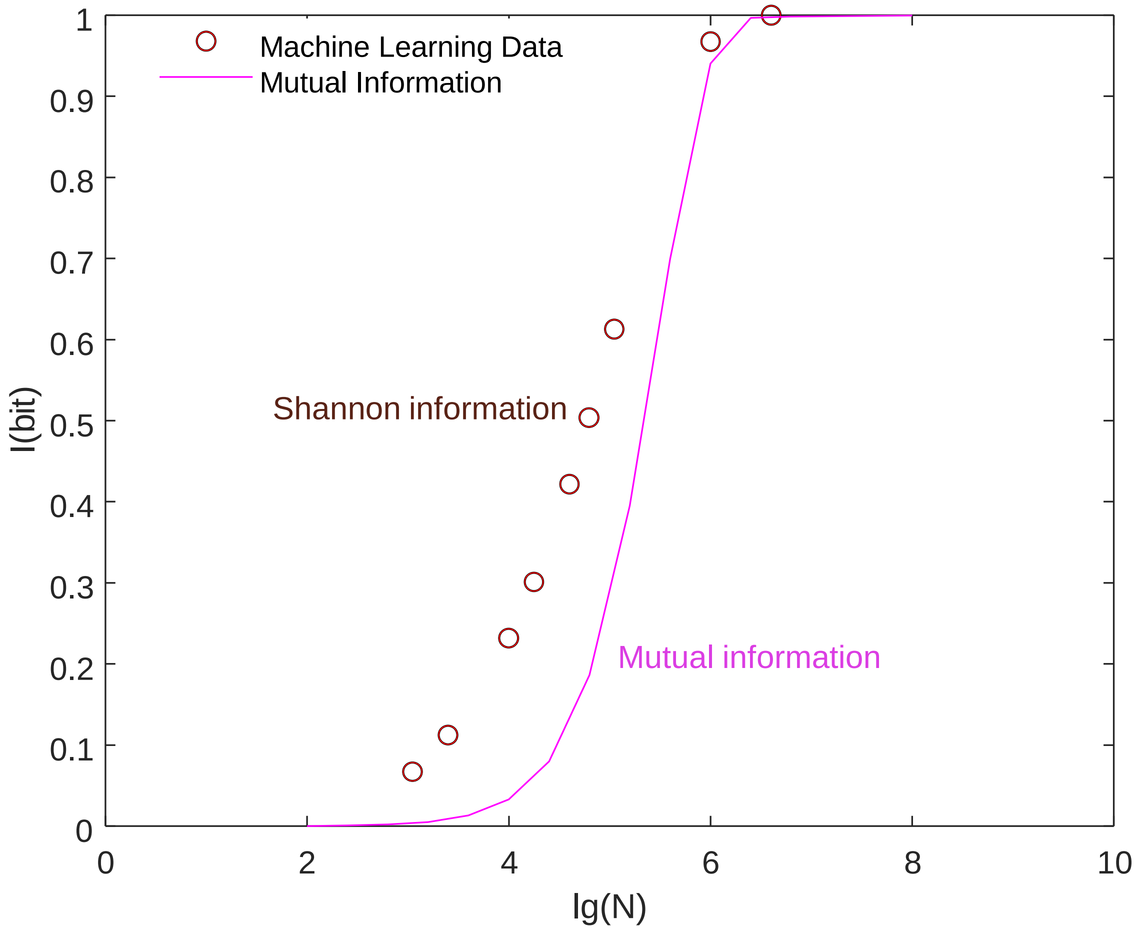
X	Z
0	1
1	0

$p(x,z)$

X \ Z	0	1
0	0	0.5
1	0.5	0

$H(X) = H(Z) = I(X;Z) = 1$ bit

Curve classification



SOTA 2: Quantum hypothesis testing

Hypothesis states

- Null hypothesis H0, $\boldsymbol{\rho_0^{\otimes N}}$
- Alternative hypothesis H1, $\boldsymbol{\rho_1^{\otimes N}}$

Test operator

$$\mathbf{\Pi}: \mathbb{C}^{N*W} \to \{\mathbf{0}, \mathbf{1}\}$$

Helstrom bound on error probability

$$P_{err} = \pi_0 * tr\left[\rho_0^{\otimes N} \cdot \Pi\right] + \pi_1 * tr\left[\rho_1^{\otimes N} \cdot (\mathbb{I} - \Pi)\right]$$

$$= \frac{1 - tr\sqrt{\rho^\dagger \rho}}{2}, \qquad \rho = \pi_0 \rho_0^{\otimes N} - \pi_1 \rho_1^{\otimes N}$$

Hilbert space $\sim \chi^{N*W}$

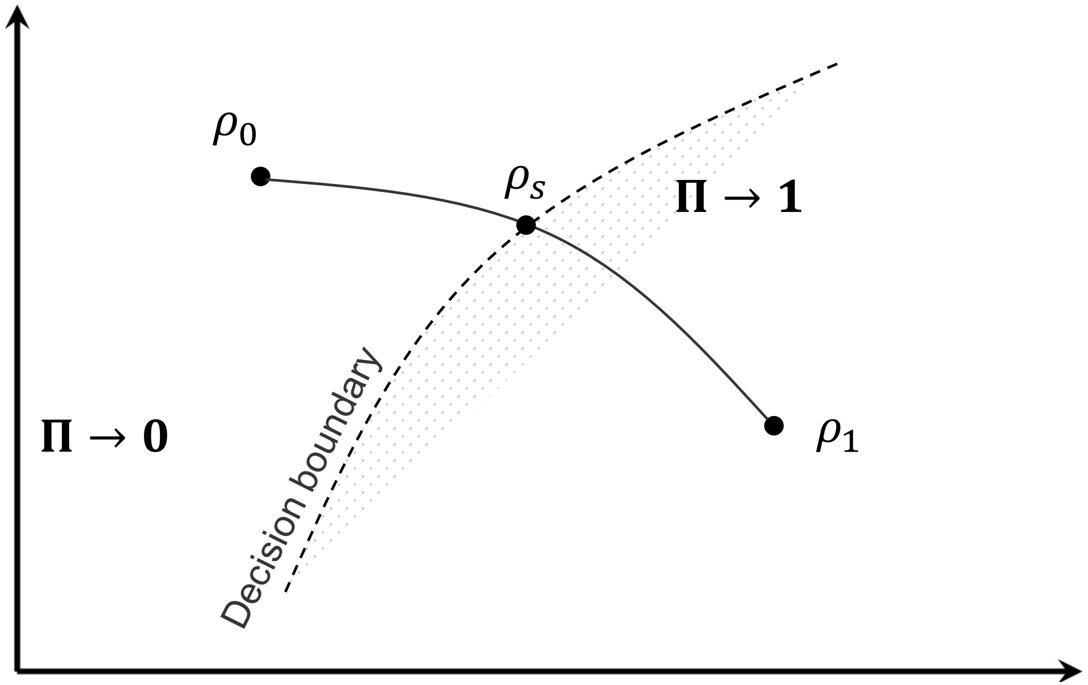
χ: number of possible states ~ 256
W: number of channels ~ 100-1M
N: number of measurements > 100

Quantum Chernoff bound

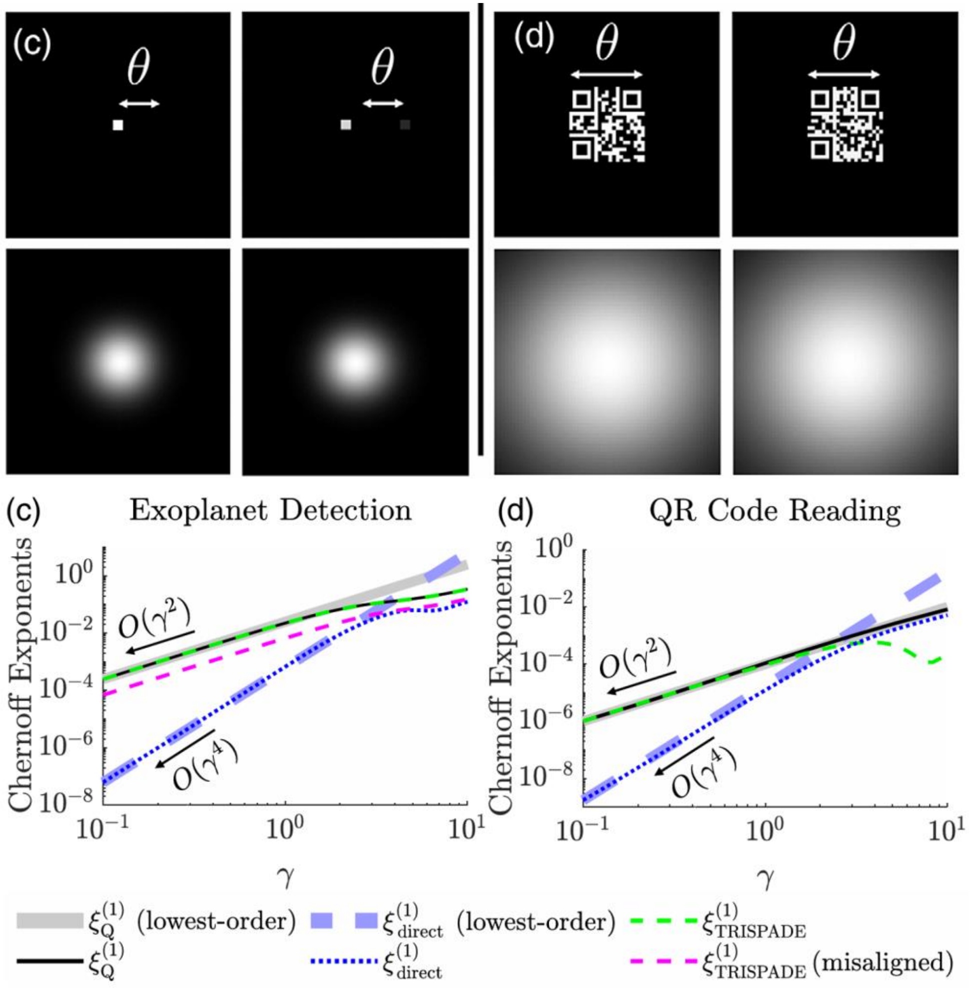
$$P_{err} \sim e^{-N \cdot \xi}, \qquad \xi = -\log\left[\min_{0 \le s \le 1} tr\left(\rho_0^{1-s} \rho_1^{s}\right)\right]$$

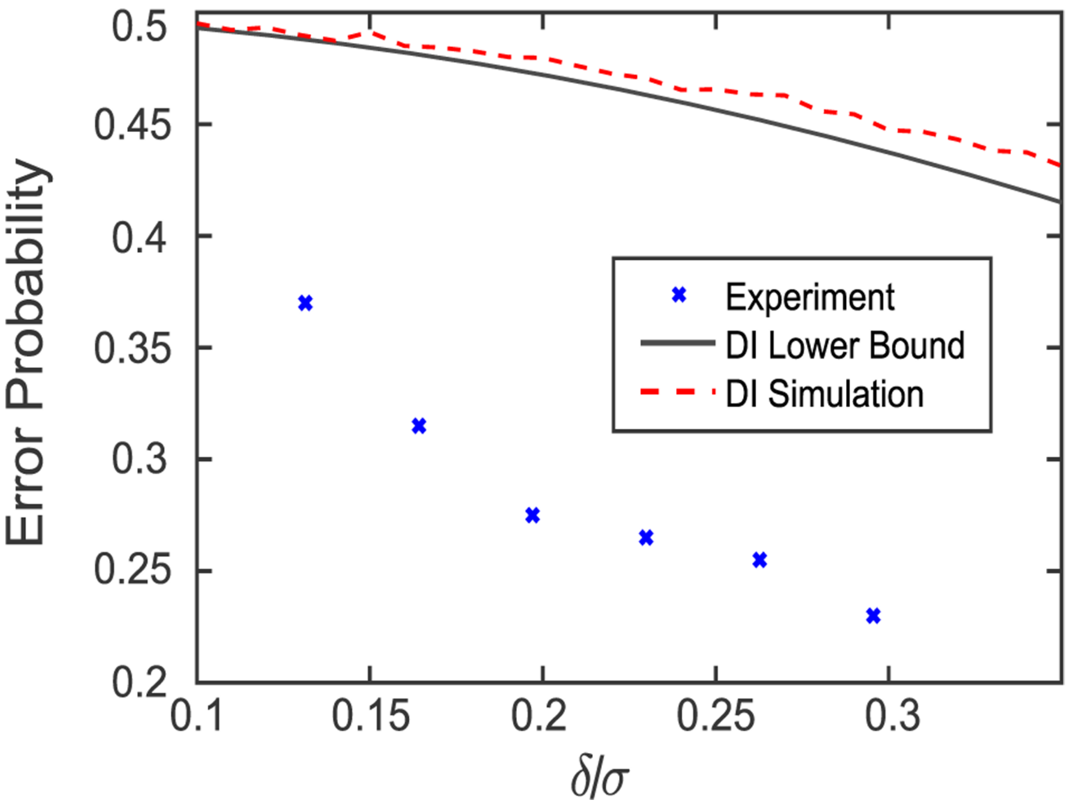
ρ_s

Hilbert space $\sim \chi^W$



Quantum Chernoff bound





Hilbert space $\sim \chi^W, W = 2$

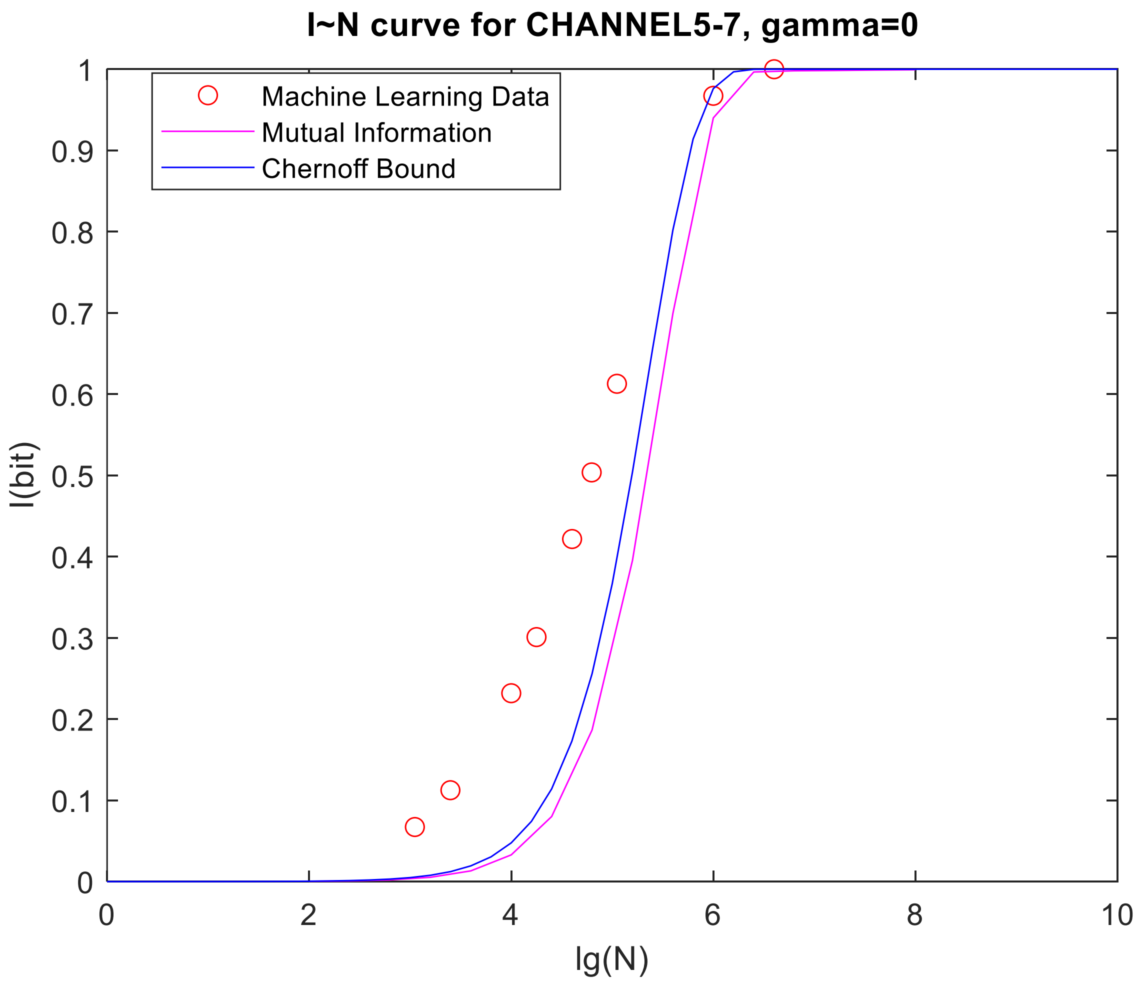
Identifying objects at the quantum limit for superresolution imaging.
Phys. Rev. Lett. **129** 180502 (2022)
Experimental demonstration of quantum-inspired optical symmetric hypothesis testing
Opt. Lett. **49**, 750-753 (2024)

Quantum Chernoff bound

Only asymptotically correct!

$$P_{err} \sim \frac{1}{2} e^{-N \cdot \xi_C}, \xi_Q = -\log\left[\min_{0 \le s \le 1} tr\left(\rho_0^s \rho_1^{1-s}\right)\right]$$

$$\xi_C = -\log\left[\min_{0 \le s \le 1} \sum_{k=1}^{W} p_{0k}^s p_{1k}^{1-s}\right]$$



Those metrics for the information bottleneck are:

- ➢ **Only asymptotically/qualitatively correct**
- ➢ **Not a distance metric invariant to symmetric transformations**
- ➢ **Lacking the structure of the distance metric**

  (no physics parameters about measurement)

An exact theory of the upper bound is critical

- For searching for the quantum-optimal measurement.
- To develop the information theory of machine learning.

Outline

Channel-phase uncorrelated (spectra, images, videos, etc.)

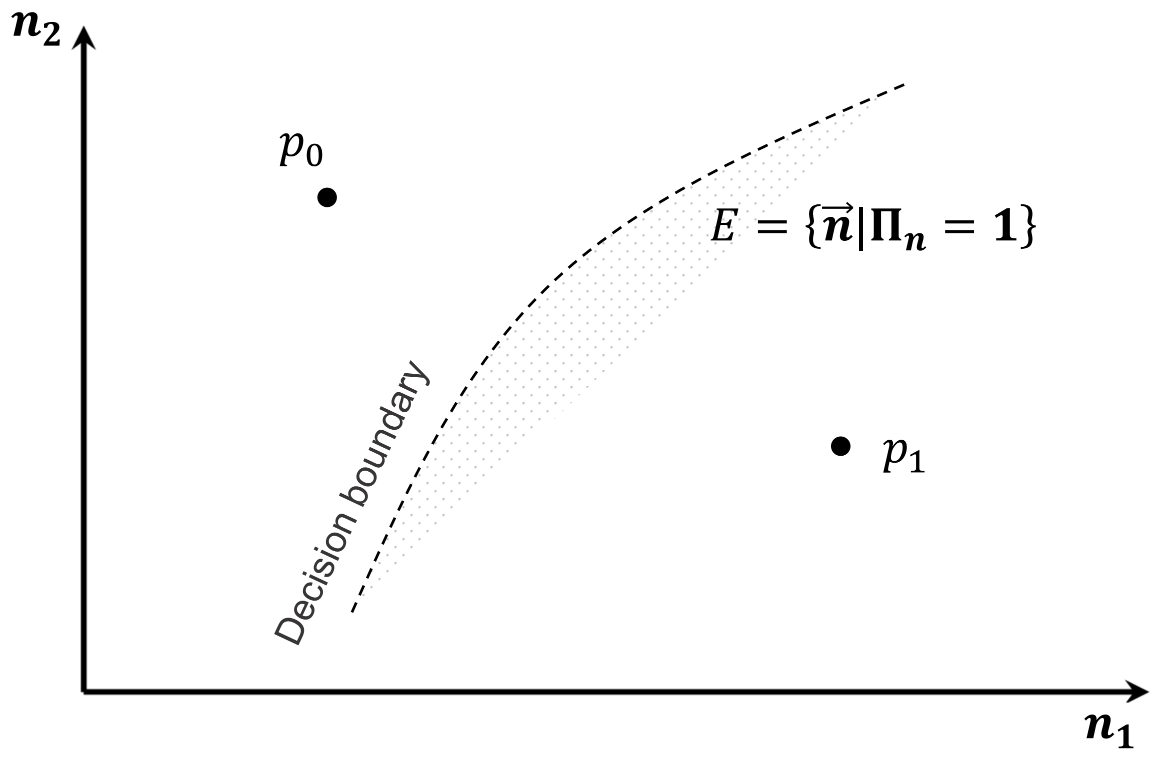
Fock space is the eigenspace for a given channel (e.g., pixel).

- Photon number is a variable but not fixed.
- Physics laws determine the distribution.

$$P_{err} = \pi_0 * \Pr(E|H_0) + \pi_1 * \Pr(E^c|H_1)$$
$$= \pi_1 + \pi_0 \Pr(E|H_0) - \pi_1 \Pr(E|H_1)$$

is minimized when

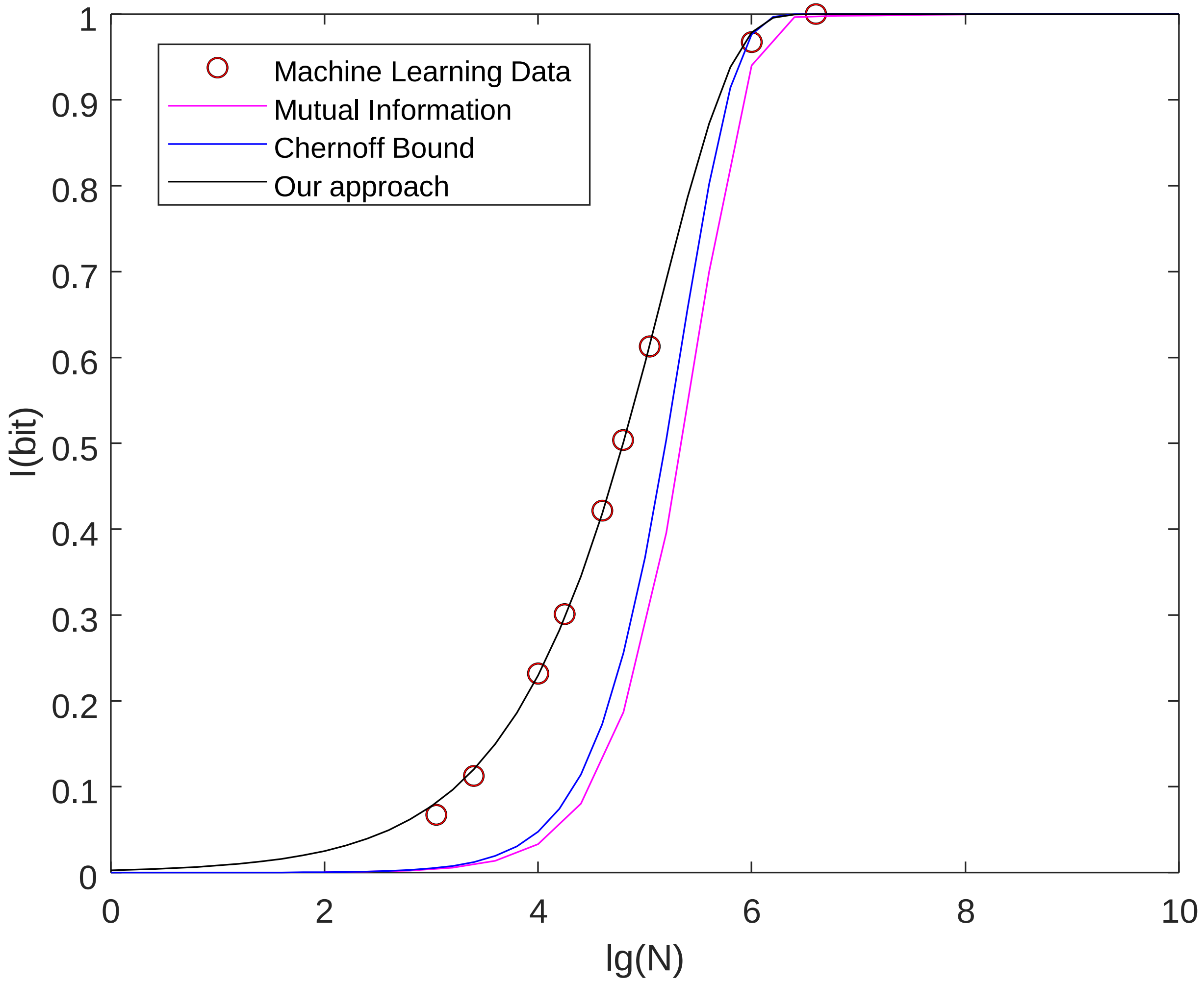
$$E = \{\vec{\boldsymbol{n}} | \pi_0 \Pr(\vec{\boldsymbol{n}}|H_0) - \pi_1 \Pr(\vec{\boldsymbol{n}}|H_1) < \boldsymbol{0}\}$$



Unpublished data

Unpublished data

Theoretical bound of accuracy

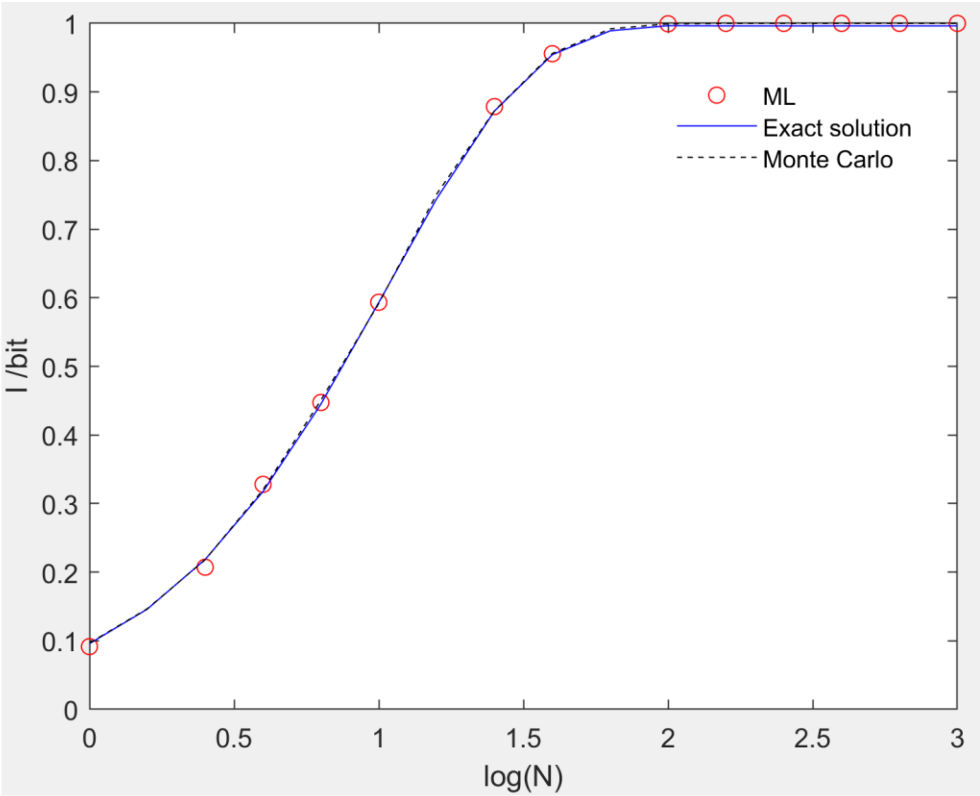


- An exact information-theoretic bound of ML
- Machine learning saturates the bound

Theoretical bound of accuracy

Hypothesis 0

Hypothesis 1



Unpublished data

Hypothesis test as parameter estimation

Unpublished data

- Mix the signals of two hypotheses with a mixture fraction, and then estimate the fraction.

$$\vec{\lambda} = (1-\eta)\vec{\lambda}_0 + \eta\vec{\lambda}_1$$

$$\vec{\lambda}_j = \vec{e}_j * \vec{B}(T)$$ symmetry

$$p(\vec{\boldsymbol{n}};\eta) = \prod_{k=1}^{W} \frac{\lambda_k^{n_k}}{(\lambda_k + 1)^{n_k+1}}$$ Physics parameters

- The uncertainty of estimating the Mean of a Gaussian distribution is given by its Variance (i.e., the inverse of its Fisher information matrix).

$$P_{acc} = \Pr(\eta < 1/2) = \int_{-\infty}^{1/2} \mathcal{N}(\eta; 0, 1/J_\eta)\mathrm{d}\eta$$

$$P_{err} = 1 - P_{acc}$$

Semantic distance for the information bottleneck

Semantic distance

$$d_0 \equiv 1/2\sigma_0 \text{ with } \sigma_0^2 \equiv [1/J^0]_{gg}$$

Single-photon Fisher information matrix:

$$J_{ij}^0 = \int \frac{(\partial_i p_{\alpha\nu})(\partial_j p_{\alpha\nu})}{p_{\alpha\nu}} \mathrm{d}\nu$$

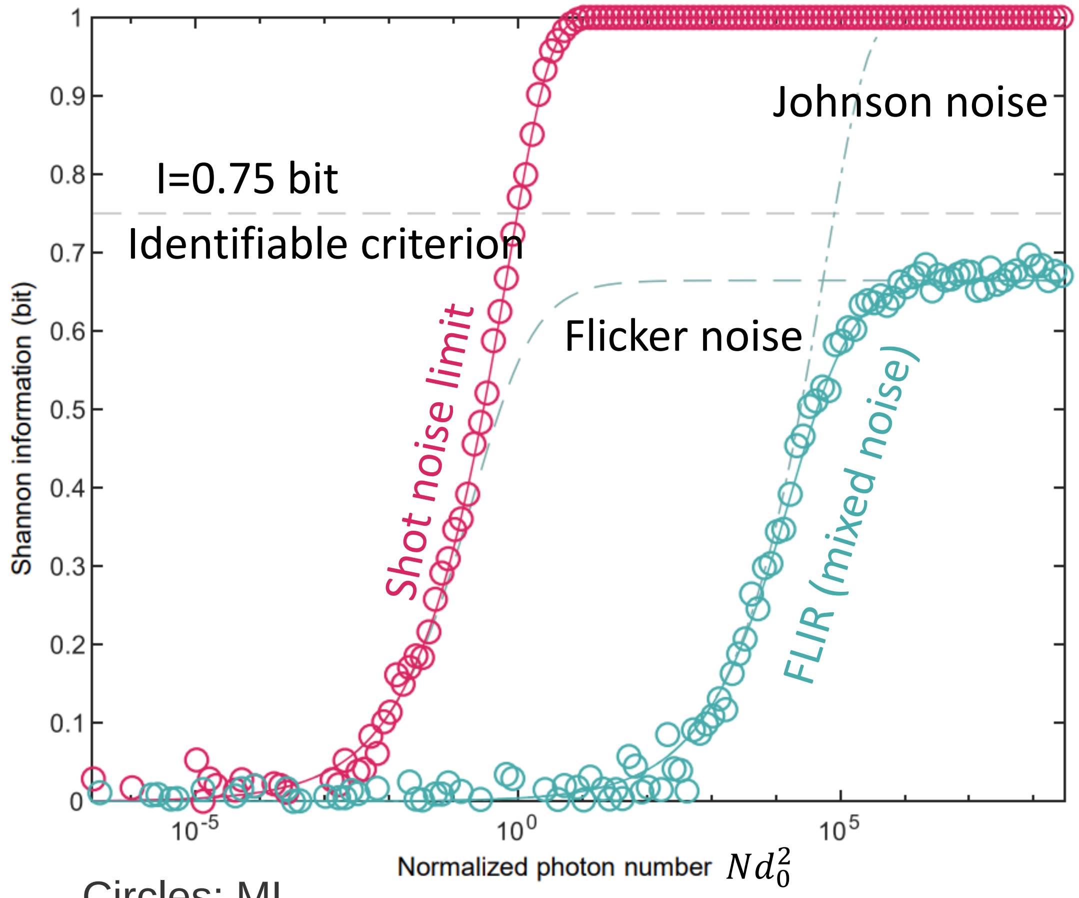
Fisher information matrix:

$$J_{ij} = N J_{ij}^0/(1+\gamma) \qquad i,j \in \{g,T\}$$

Shannon information:

$$I = \log_2\left[1 + \mathrm{erf}\left[\sqrt{\frac{Nd_0^2}{2(1+\gamma)}}\right]\right],$$

Statistical distance with structures



Circles: ML
Curves: Theory
Identifiable criterion:
$Nd_0^2 = 1$

F. Bao, et al, "Heat-assisted detection and ranging," *Nature* **619,** 743–748 (2023) (Cover article)

Data processing inequality: Fisher information

- Markov chain

$$p(z|y) = \int p(z|x) * p(x|y)\mathrm{d}x\,, \qquad \partial_y p(z|y) = \int p(z|x) * \partial_y p(x|y)\mathrm{d}x$$

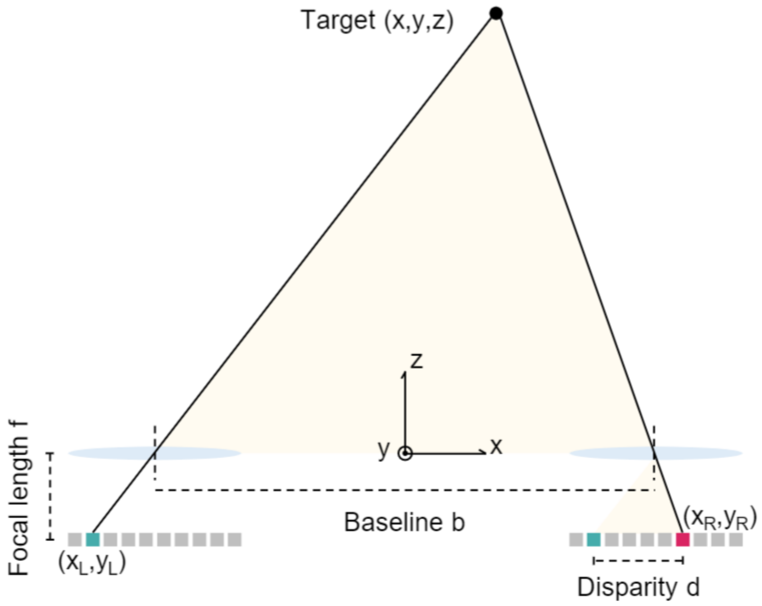
$$J_y^z = \int \frac{\left[\partial_y p(z|y)\right]^2}{p(z|y)}\mathrm{d}z = \int \frac{\left[\int p(z|x) * \partial_y p(x|y)\mathrm{d}x\right]^2}{\int p(z|x) * p(x|y)\mathrm{d}x}\mathrm{d}z$$

$$\leq \int\int p(z|x) * \frac{\left[\partial_y p(x|y)\right]^2}{p(x|y)}\mathrm{d}x\mathrm{d}z = \int \frac{\left[\partial_y p(x|y)\right]^2}{p(x|y)}\mathrm{d}x$$

$$= J_y^x$$

Outline

Binocular stereovision



Schematic of binocular stereo vision

$$x = -\frac{b}{d} \cdot \frac{x_L + x_R}{2}$$

$$y = -\frac{b}{d} \cdot \frac{y_L + y_R}{2}$$

$$z = \frac{bf}{d}$$

$$\delta z = \frac{z^2}{bf} \delta d$$

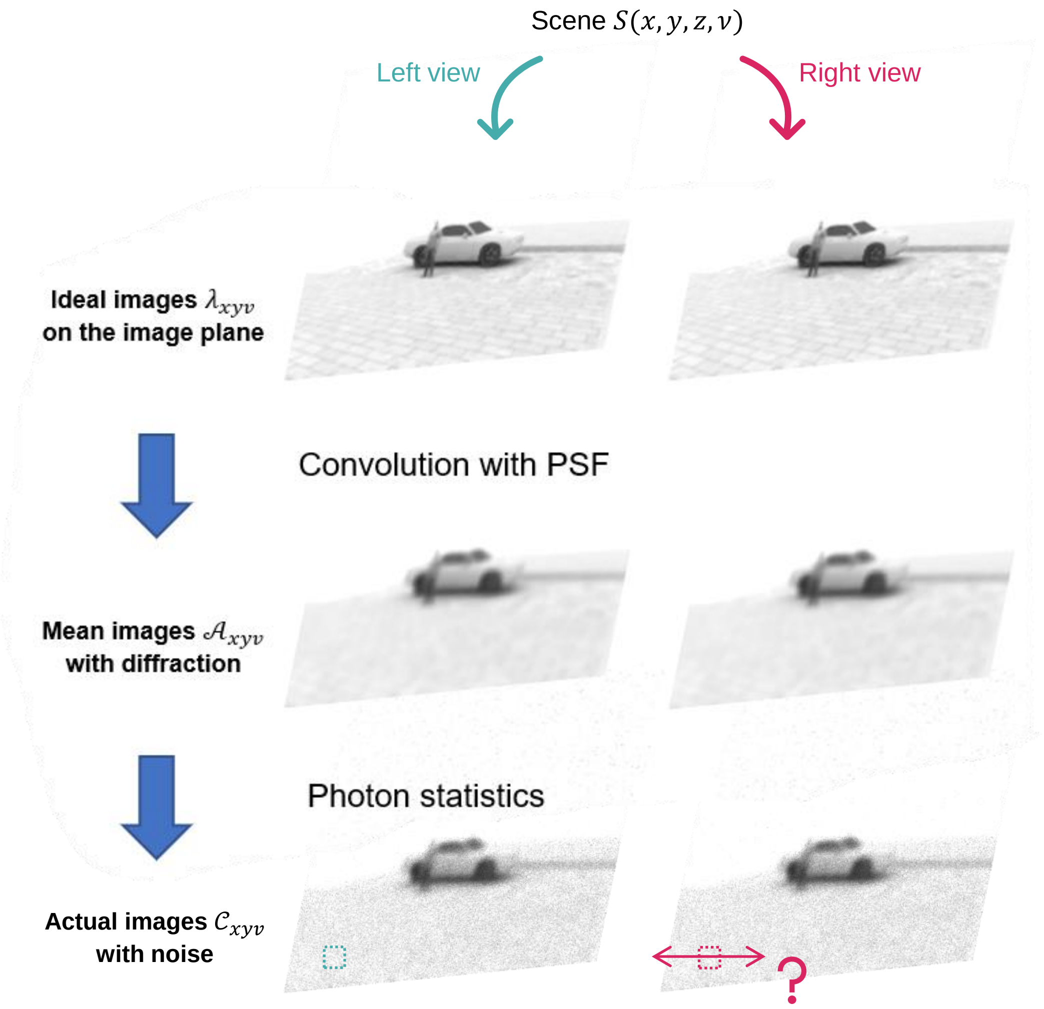
δz: ranging error

δd: disparity error

How many photons do we need to measure the distance within a given accuracy?...

(What is the fundamental limit of disparity error?)

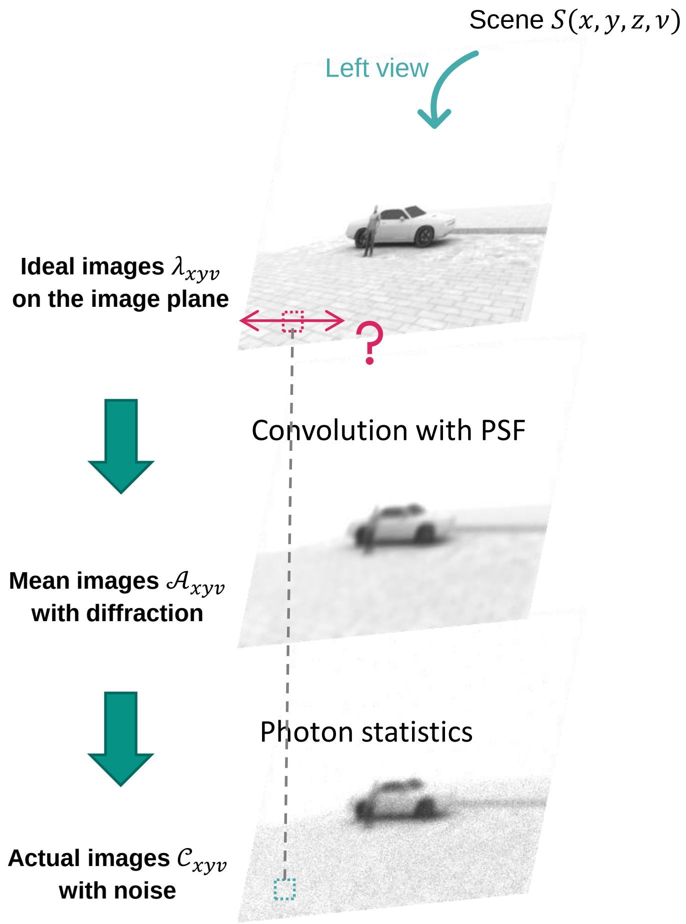
The correspondence problem



For a given point in the actual left image, find the corresponding point in the actual right image.

--- The correspondence problem in computer vision

Estimation theory



We re-interpret the correspondence problem in computer vision as a position-estimation problem in estimation theory:

For a given point in the actual image, find the corresponding point in the ideal image.

F. Bao, et al, "Heat-assisted detection and ranging," *Nature* **619,** 743–748 (2023) (Cover article)

Correspondence uncertainty

Depth resolution:

$$\sqrt{N}\delta z \geq \frac{z^2}{bf}\sqrt{2(1+\gamma)(\sigma_c^2+\sigma_d^2)},$$

δz: ranging error
N: photon number
z: distance
b: baseline
f: focal length
γ: Electronic noise power over shot-noise power
σ_d: photonic diffraction uncertainty (width of PSF)
σ_c: photonic correspondence uncertainty
$\sigma_c = \sqrt{1/J_x^0}$, J_x^0 is the single-photon FI

$$J_x^0 = \iint_\Omega \frac{(\partial_x p_{x\nu})^2}{p_{x\nu}}\,\mathrm{d}s\mathrm{d}\nu \qquad p_\lambda(x,\nu) \equiv \lambda_{x\nu} / \iint_\Omega \lambda_{x\nu}\,\mathrm{d}s\mathrm{d}\nu$$

x, y: Image coordinates
v: Wave number
γ: Normalized electronic noise power
δz: ranging error
δd: disparity error
Ω: window domain

F. Bao, et al, "Heat-assisted detection and ranging," *Nature* **619,** 743–748 (2023) (Cover article)

Physical limit on depth estimation

WESTLAKE UNIVERSITY 西湖大學

Machine perception obeys the information-theoretic bound…

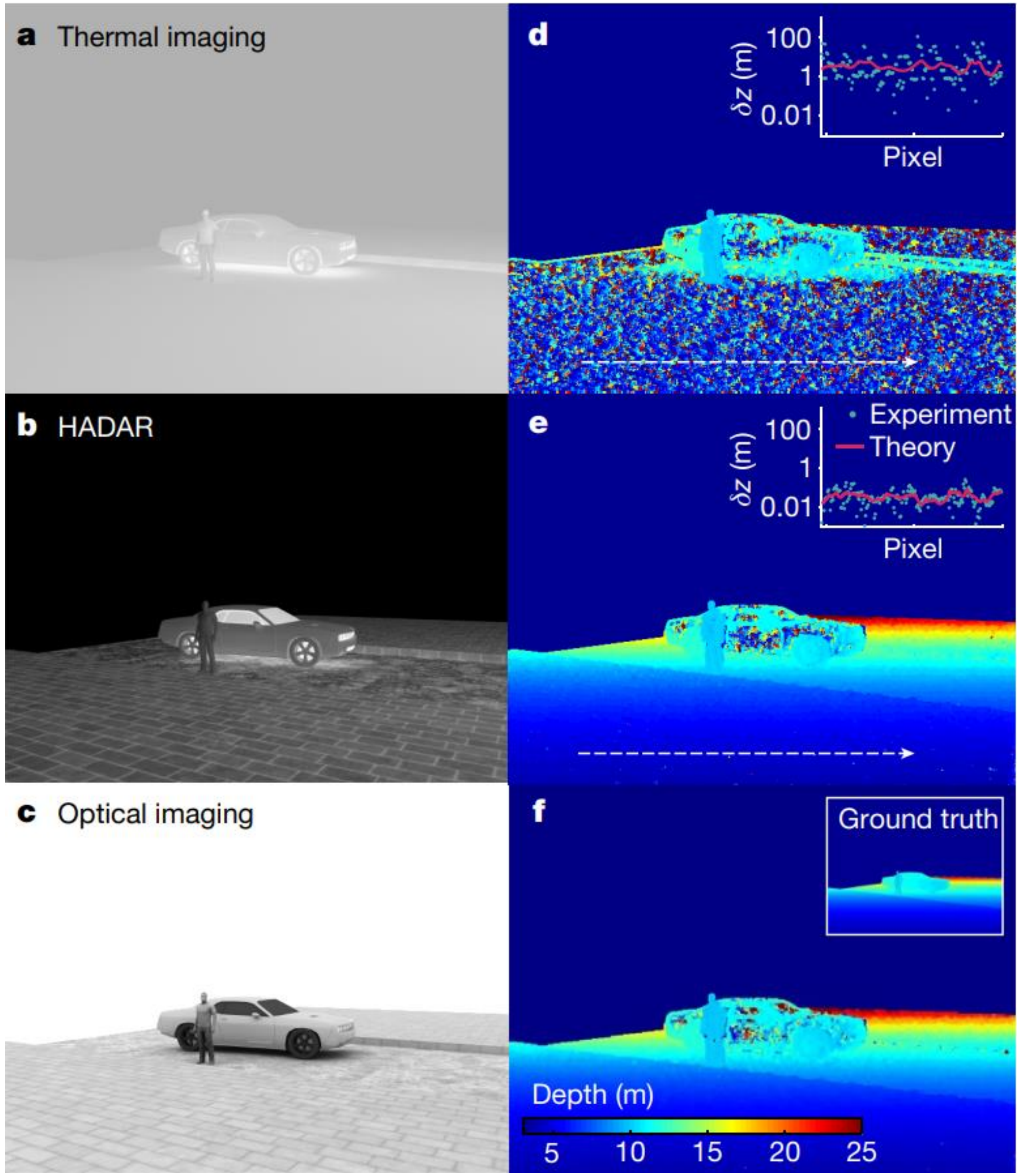
Machine perception experiment

1. Using stereo matching to derive the depth information (d, e, f).
2. Comparing with the ground truth to get error statistics along the white arrow (blue dots in insets)

Information theory

$$\sqrt{N}\delta z \geq \frac{z^2}{bf}\sqrt{2(1+\gamma)(\sigma_c^2+\sigma_d^2)},$$

(red curves in insets)



Summary & Outlook

- **Physical limits** Information bottleneck in physics parameters invariant to transformations
- **Physical limits on classification** Semantic distance
- **Physical limits on depth estimation** Correspondence uncertainty

- **Generalization to the quantum case,** (measurement optimization)
  Multi-label classification, SIFT operator, Language problem, ...