Automating Scientific Discovery

Anti-steams, Universal Similarity & Statistical Causality

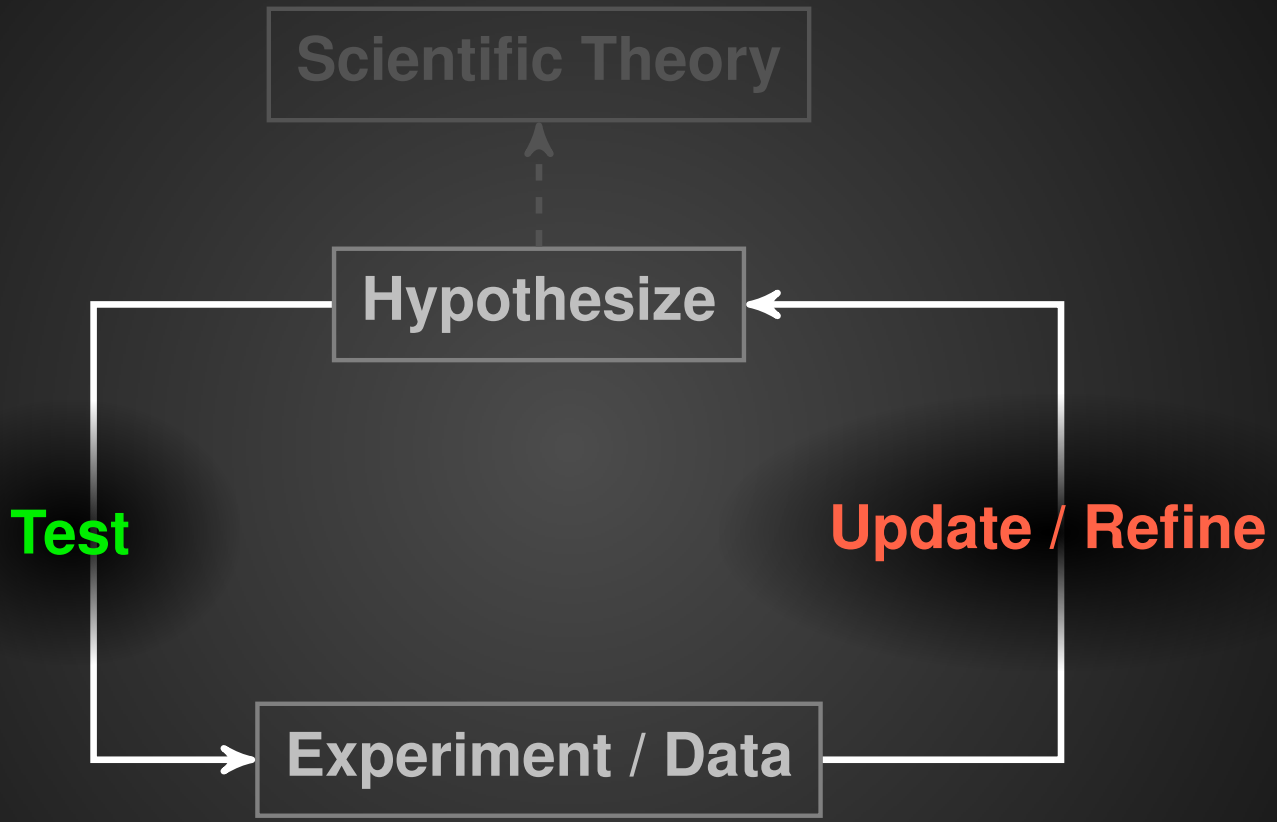
Ishanu Chattopadhyay
Research Scientist

Computation Institute
Institute for Genomics & Systems Biology

University of Chicago

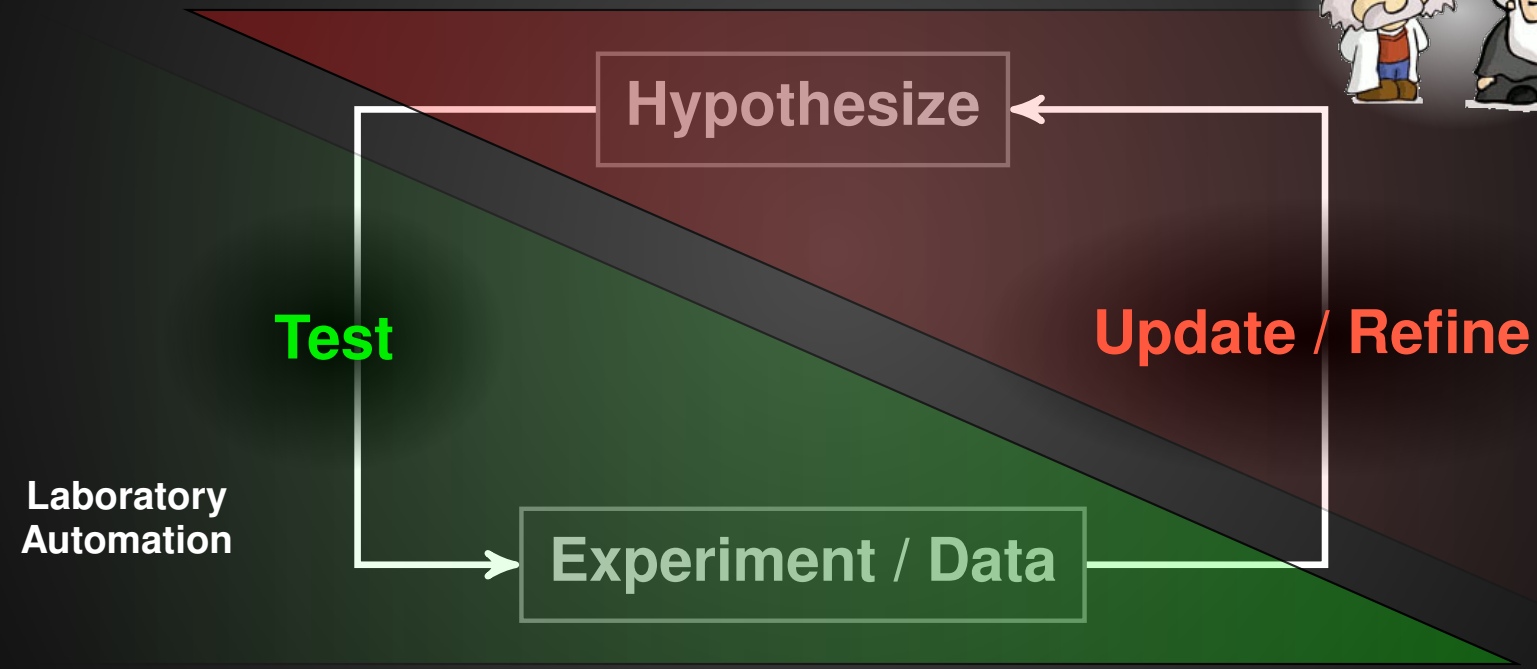
The Scientific Method

A la Francis Bacon



The Scientific Method

A la Francis Bacon



The Scientific Method

A la Francis Bacon

Hypothesize

Test

Update / Refine

Laboratory Automation

Experiment / Data

... computational approaches have been more successful in **small, well-defined systems than in larger, less studied, or more complex ones.** The explosion of data from high-throughput experiments, however, increasingly presents researchers with very complicated systems.

— *Machine Science*, James Evans and Andrey Rzhetsky, Science, 329 (5990), 399-400 (2010)

Automating Science

From Data to Explanation

Automating Science

From Data to Explanation

- **Universal Distance Metric**
- **Features**

Automating Science

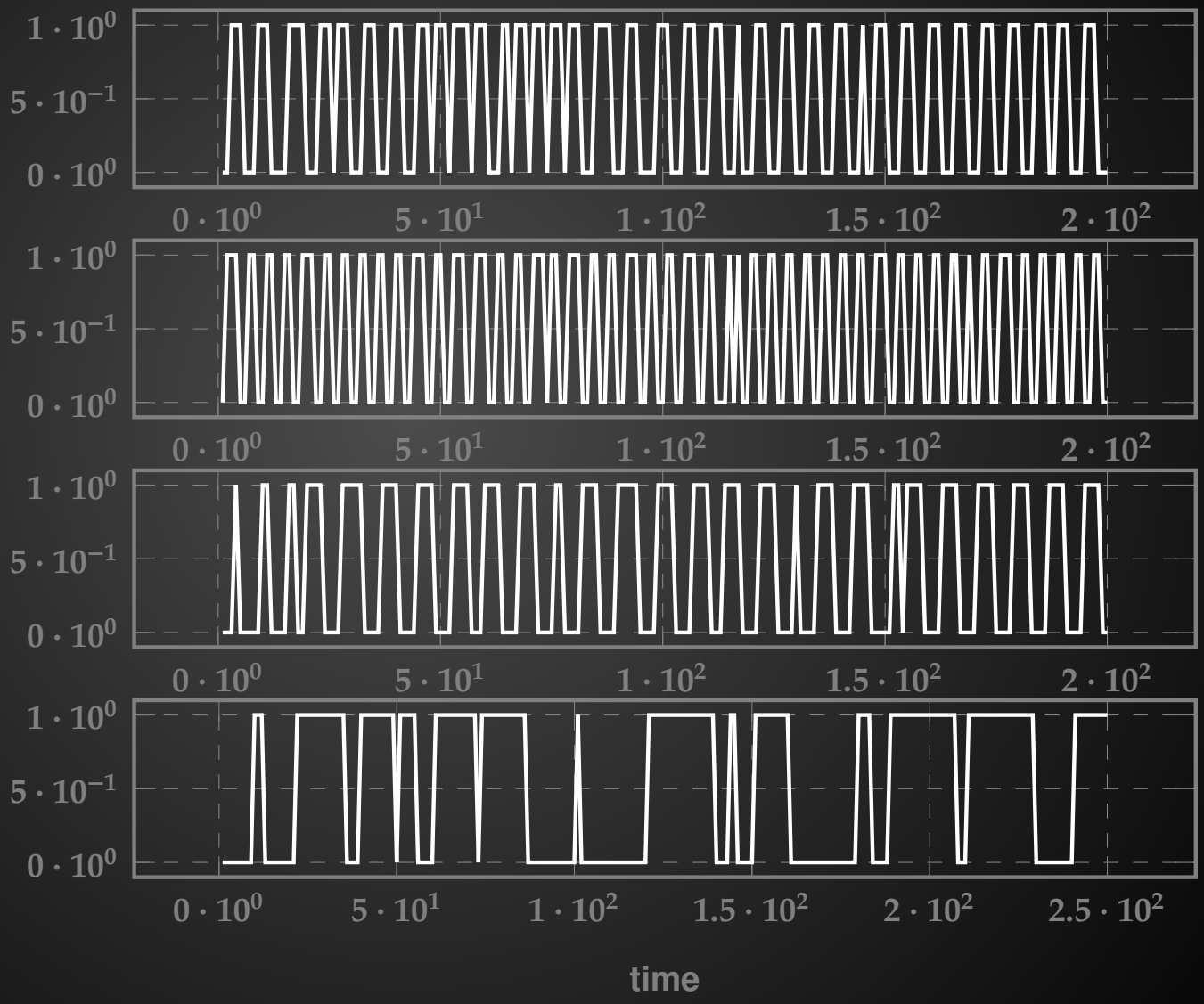
From Data to Explanation

- **Universal Distance Metric**
- **Features**

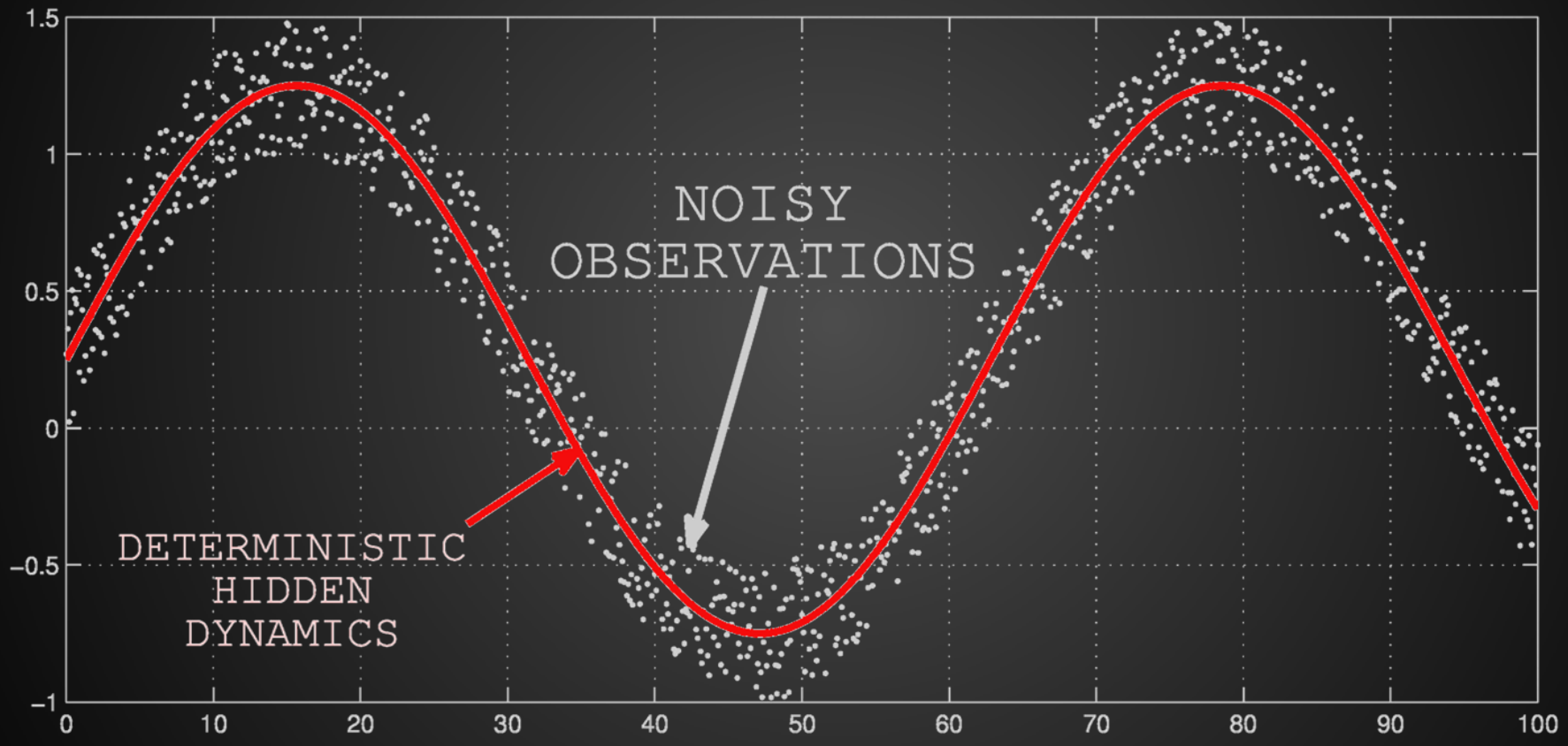
not similar

Similarity

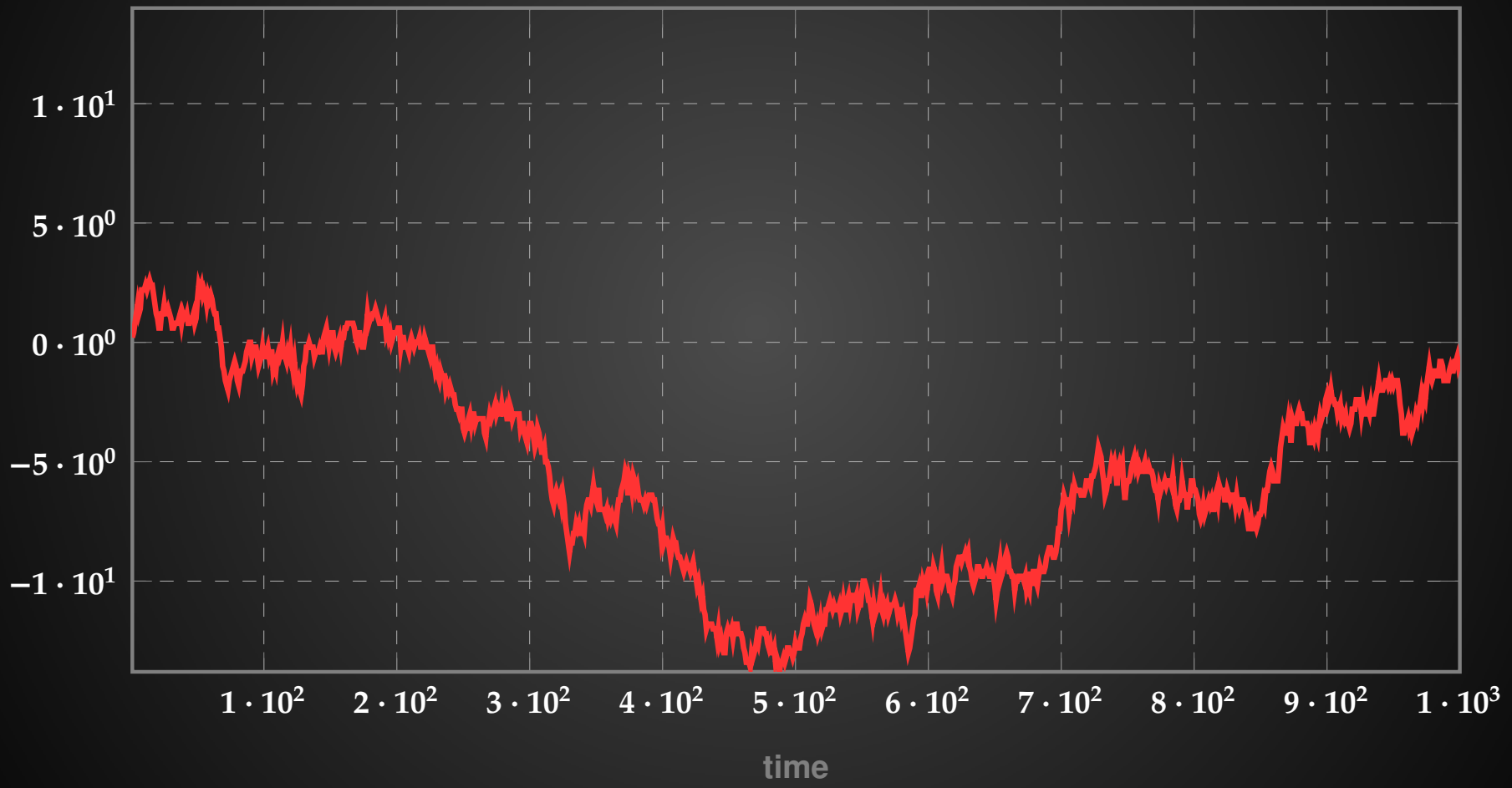
Pick the odd one out!



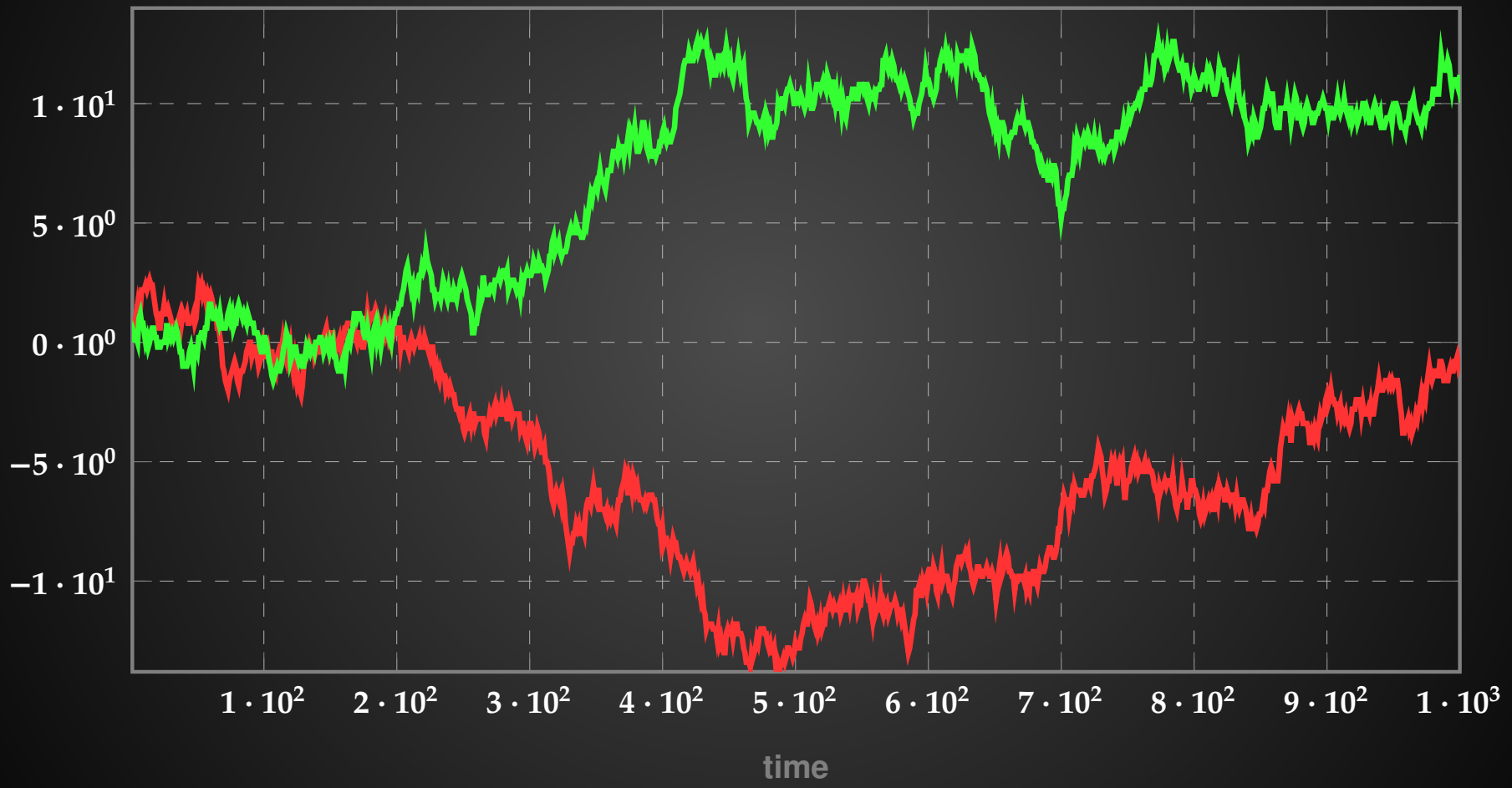
Stochastic Process ≠ Deterministic + Noise



Stochastic Process ≠ Deterministic + Noise



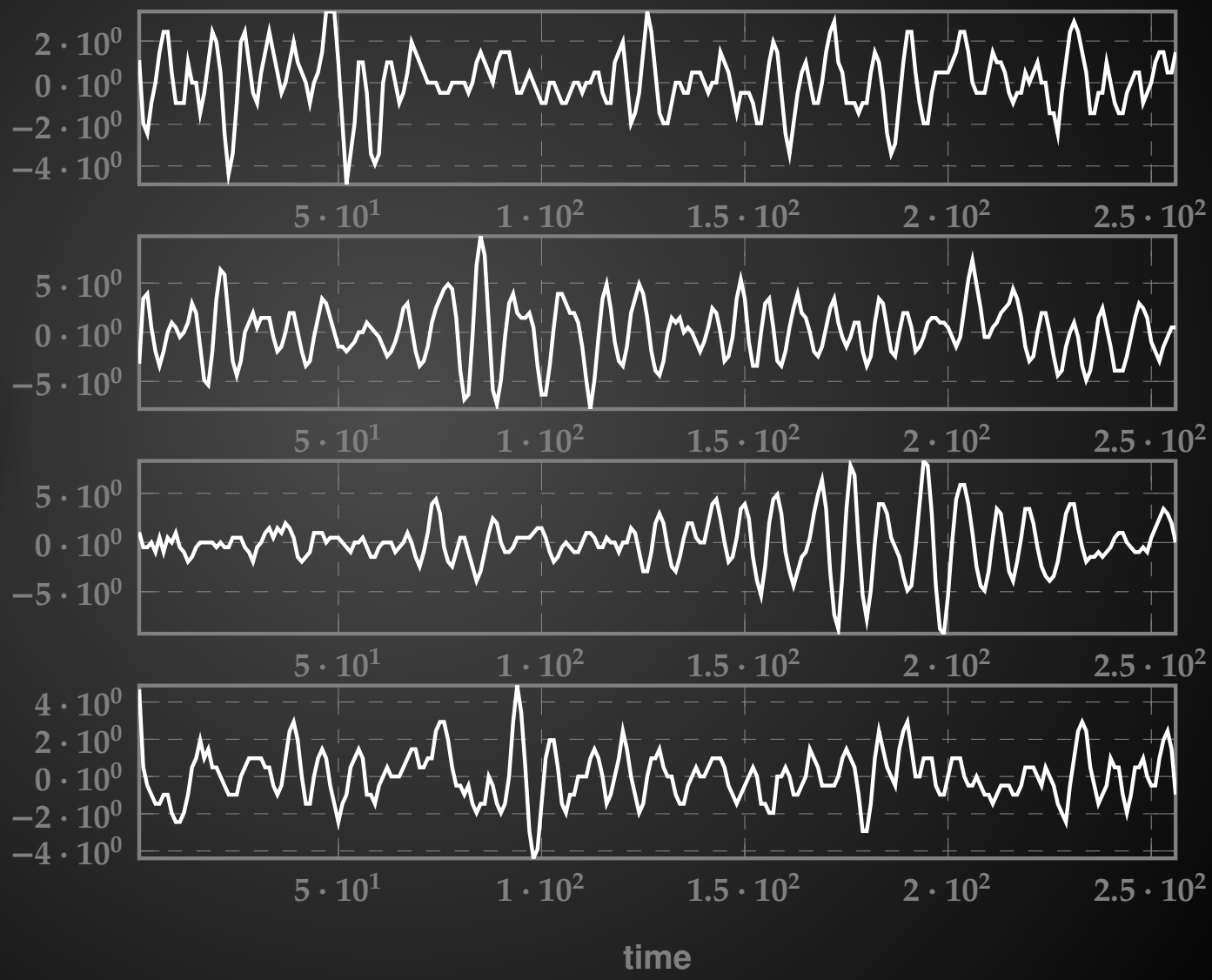
Stochastic Process ≠ Deterministic + Noise



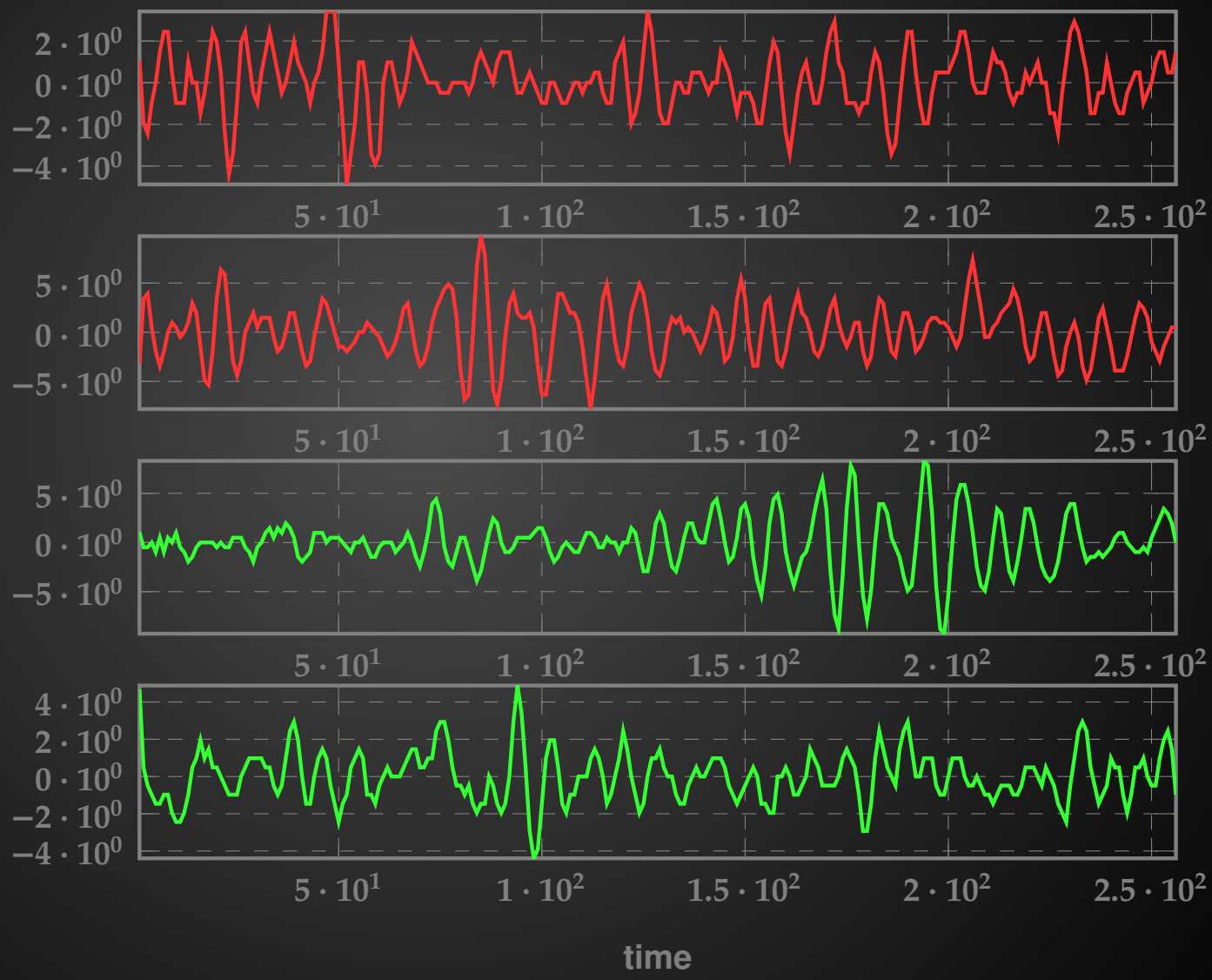
Stochastic Process ≠ Deterministic + Noise

Brainwaves: Identical Stimuli

Are they from the same individual?

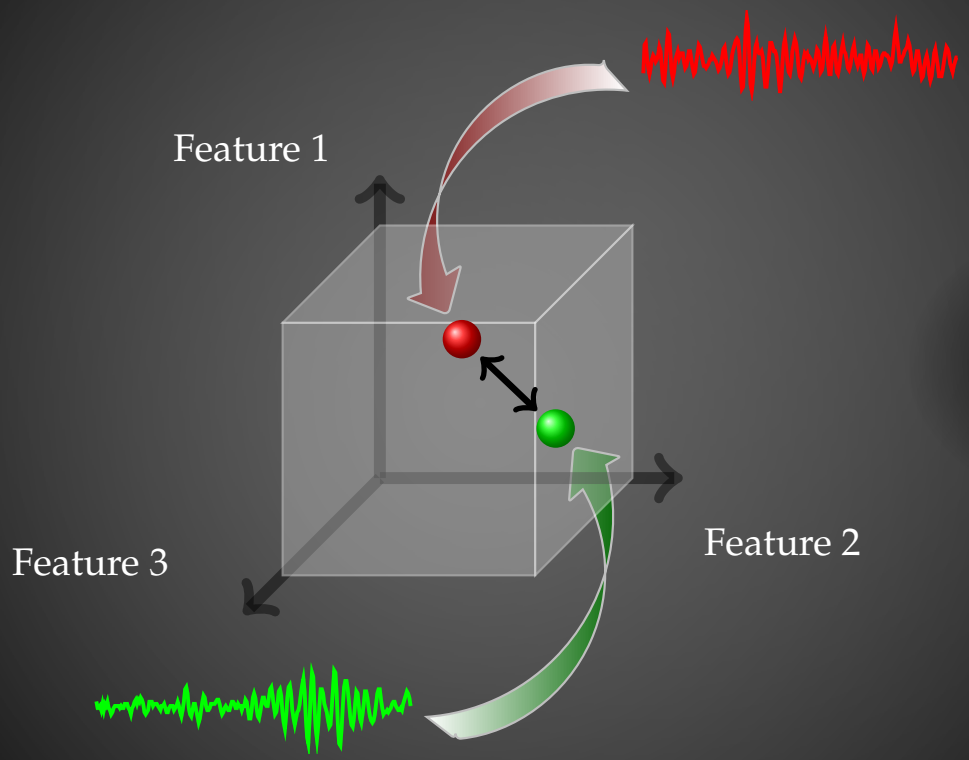


Brainwaves: Identical Stimuli



State of Art

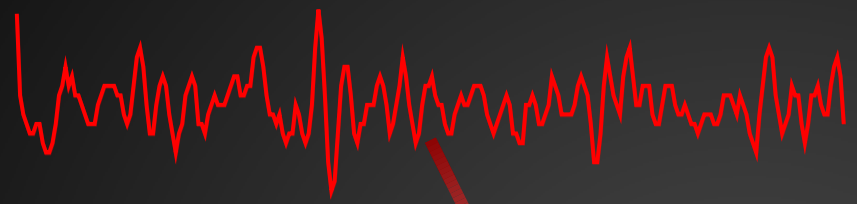
Requires Features



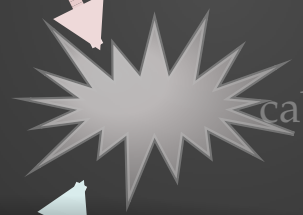
Similarity:

Distance between feature vectors

Data Smashing



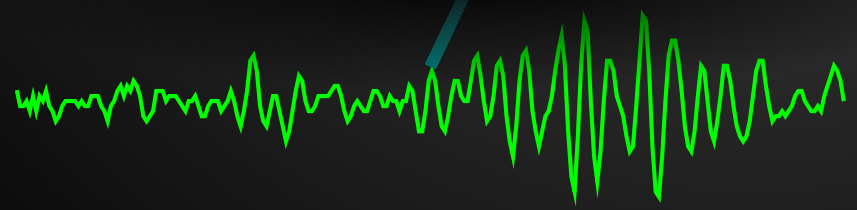
Quantize: bcbbbbacabbcbbbbabbbbba···

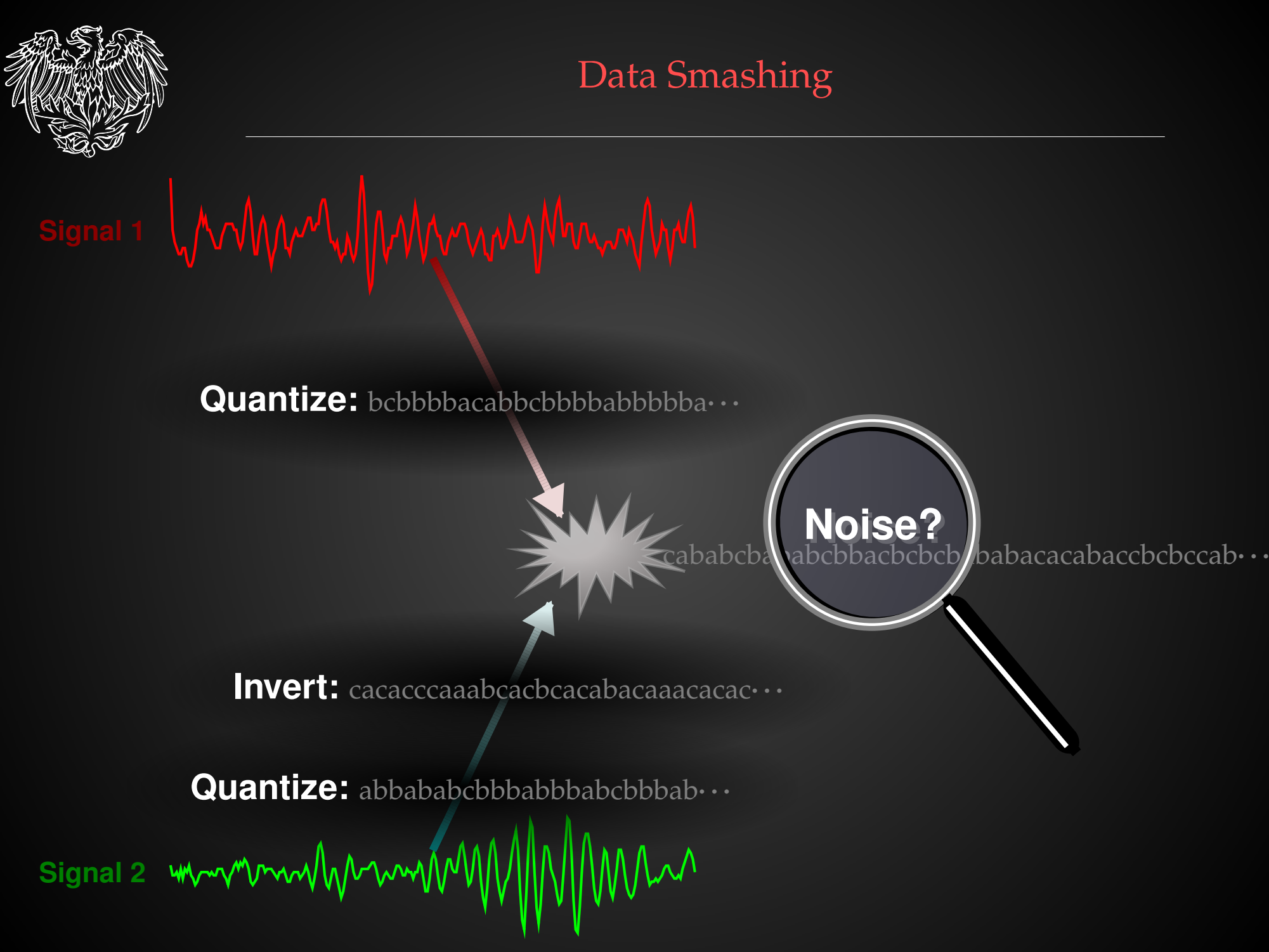


cababcbababcbbacbcbcbababacacabaccbcbccab···

Invert: cacacccaaabcacbcacabacaaacacac···

Quantize: abbababcbbbabbbabcbbbab···




Data Smashing
Signal 1
Quantize: bcbbbbacabbcbbbbabbbbba···
Noise?
cababcbabcbbacbcbcbbabacacabaccbcbccab···
Invert: cacacccaaabcacbcacabacaaacacac···
Quantize: abbababcbbbabbbabcbbbab···
Signal 2

Anti-streams

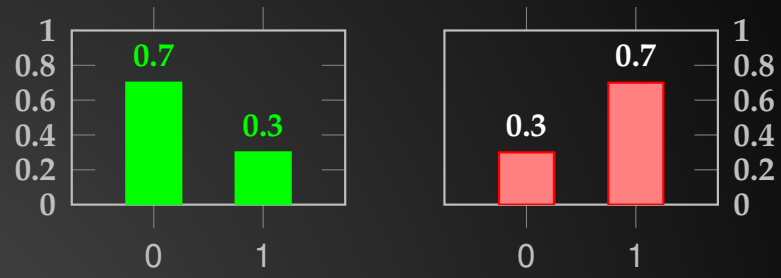
Intuitive Description

Stream $s \longrightarrow$ Anti-stream s'

Anti-streams

Intuitive Description

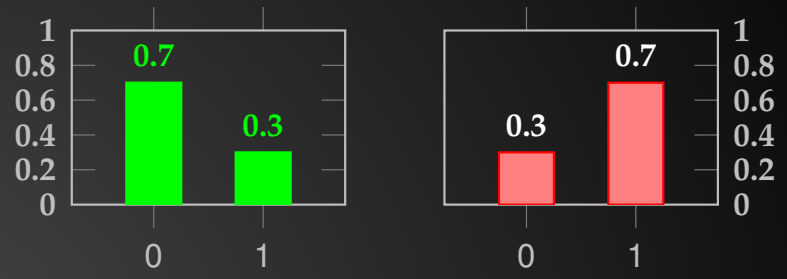
Stream $s \longrightarrow$ Anti-stream s'

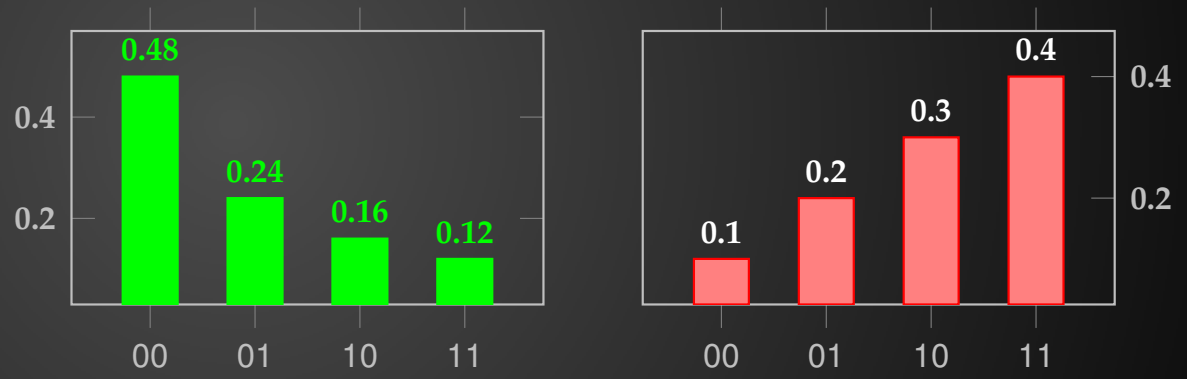


Anti-streams

Intuitive Description

Stream $s \longrightarrow$ Anti-stream s'

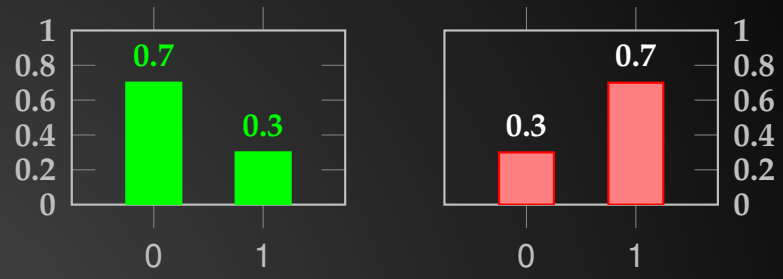


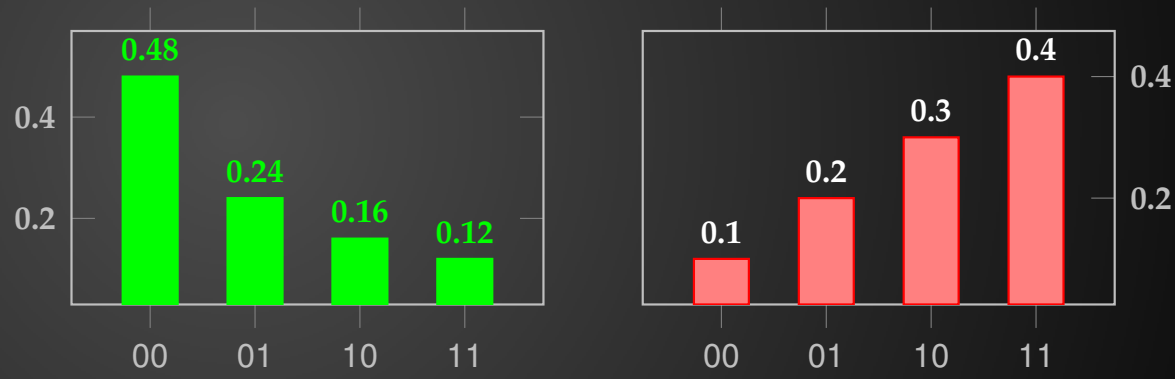


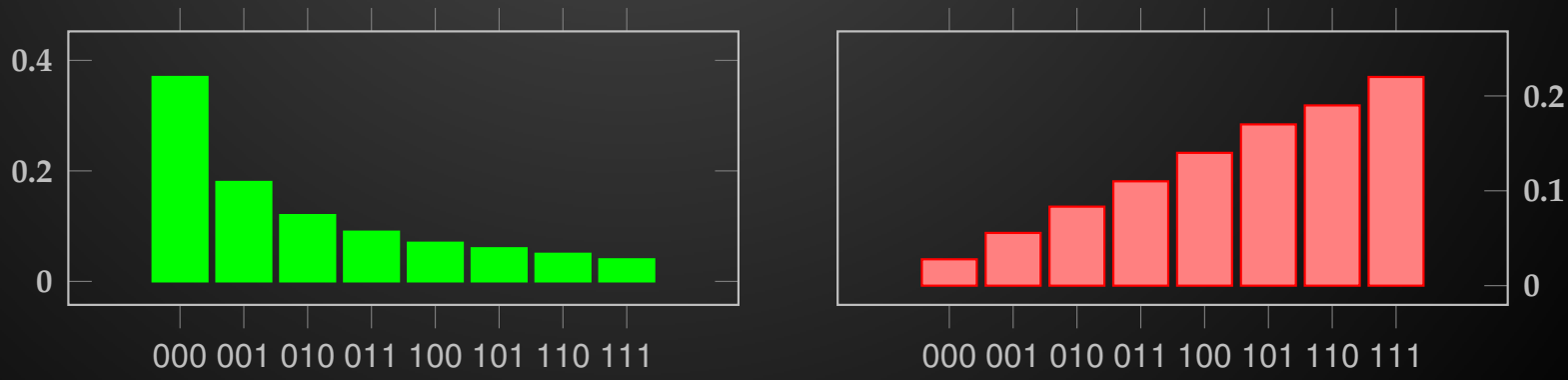
Anti-streams

Intuitive Description

Stream $s \longrightarrow$ Anti-stream s'







Quantization

Mapping Continuous Data to Symbol Stream

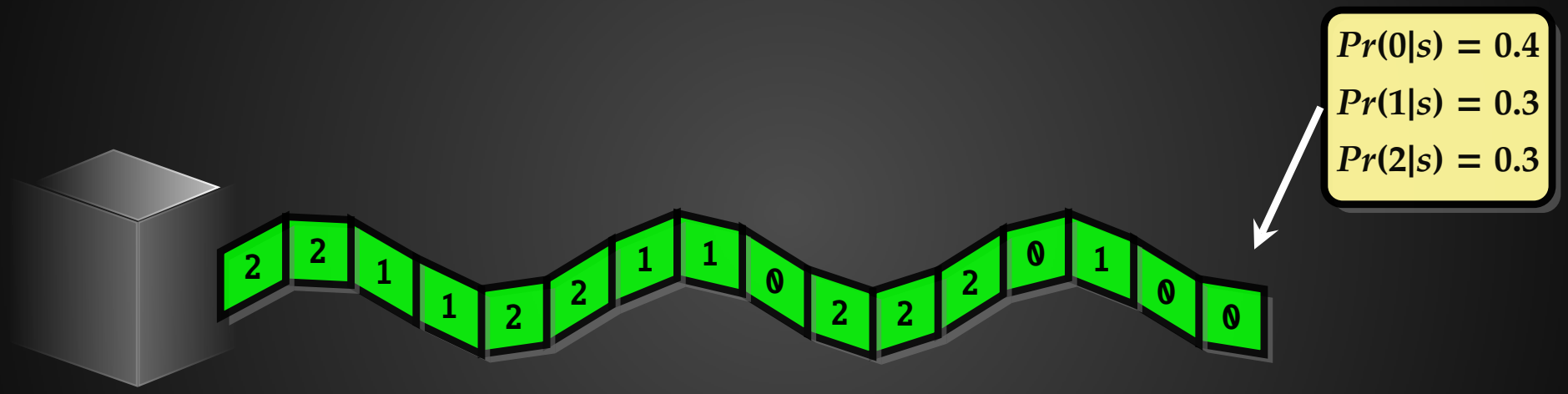
Quantization Alphabet $\Sigma = \{0, 1, 2, 3\}$

The Black Box Approach

Dynamical System As A Symbol Generator

The Black Box Approach

Dynamical System As A Symbol Generator



Ergodic Stationary Quantized Process $\Longleftrightarrow$ Probability Measure on Infinite Strings

System As A Symbol Generator

Probability Space $(\Sigma^\omega, \mathscr{B}, \mu)$

- Σ^ω : Set of strictly infinite strings on alphabet Σ
- $\mathscr{B}$: smallest σ-algebra generated by the sets $\left\{x\Sigma^\omega : x \in \Sigma^\star\right\}$
- μ : Probability measure on infinite strings:

$$\mu(x\Sigma^\omega) \mapsto [0,1]$$

$$\sum_{x \in \Sigma^\star} \mu(x\Sigma^\omega) = 1$$

System As A Symbol Generator

Probability Space $(\Sigma^\omega, \mathscr{B}, \mu)$

- Σ^ω : Set of strictly infinite strings on alphabet Σ
- $\mathscr{B}$: smallest σ-algebra generated by the sets $\left\{x\Sigma^\omega : x \in \Sigma^\star\right\}$
- μ : Probability measure on infinite strings:

$$\mu(x\Sigma^\omega) \mapsto [0,1]$$

$$\sum_{x \in \Sigma^\star} \mu(x\Sigma^\omega) = 1$$

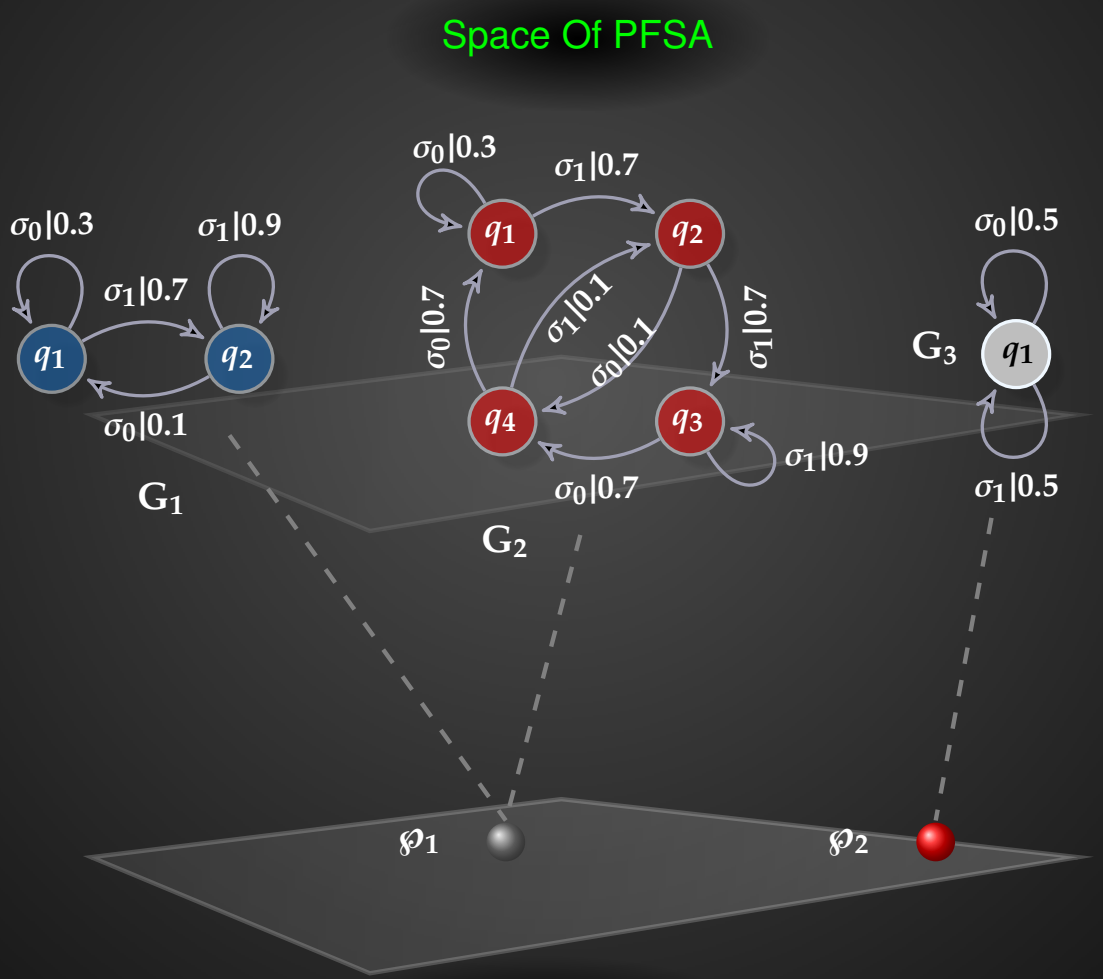
Probability Measure on Infinite Strings $\Longrightarrow$ Equivalence Relation on Finite Strings

$$\forall x_1, x_2 \in \Sigma^\star, x_1 \sim x_2$$
$$\text{if } \forall x \in \Sigma^\star, \mu(x_1 x\Sigma^\omega) = \mu(x_2 x\Sigma^\omega)$$

Equivalence Classes are causal states

Probabilistic Finite State Automata

Models For Quantized Stationary Ergodic Stochastic Processes



Adding Probability Measures

Consider two measures: $\begin{cases} \wp_1 : \mathscr{B} \rightarrow [0,1] \\ \wp_2 : \mathscr{B} \rightarrow [0,1] \end{cases}$

Define a binary operation:

$$\wp_1 \oplus \wp_2 \triangleq \wp_3$$

where

$$\wp_3(x\Sigma^\omega) = \wp_1(x\Sigma^\omega)\wp_2(x\Sigma^\omega) \times \textbf{Constant}$$

$$\sum_{x \in \Sigma^\star} \wp_3(x\Sigma^\omega) = 1$$

Adding Probability Measures

Consider two measures: $\begin{cases} \wp_1 : \mathscr{B} \rightarrow [0,1] \\ \wp_2 : \mathscr{B} \rightarrow [0,1] \end{cases}$

Define a binary operation:

$$\wp_1 \oplus \wp_2 \triangleq \wp_3$$

where

$$\wp_3(x\Sigma^\omega) = \wp_1(x\Sigma^\omega)\wp_2(x\Sigma^\omega) \times \textbf{Constant}$$

$$\sum_{x \in \Sigma^\star} \wp_3(x\Sigma^\omega) = 1$$

- **Commutative**
- **Closed**
- **Unique Inverse**
- **Unique Identity**

Abelian Group

Lifting Group Structure To PFSAs

Mathematical Structure Of Model Space

The Abelian Group

The Space of Models has the mathematical structure of an Abelian Group

Mathematical Structure Of Model Space

The Abelian Group

The Space of Models has the mathematical structure of an Abelian Group

$$1 + 2 = 3$$
$$2 - 2 = 0$$
$$3 + 0 = 3$$

Mathematical Structure Of Model Space

The Abelian Group

The Space of Models has the mathematical structure of an Abelian Group

$$1 + 2 = 3$$
$$2 - 2 = 0$$
$$3 + 0 = 3$$

We can “add and subtract” models:

$$G + H = J$$
$$G - H = K$$

Mathematical Structure Of Model Space

The Abelian Group

The Space of Models has the mathematical structure of an Abelian Group

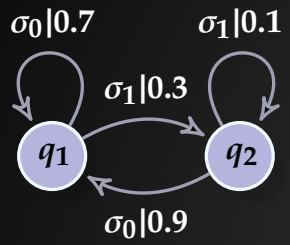
$$1 + 2 = 3$$
$$2 - 2 = 0$$
$$3 + 0 = 3$$

We can “add and subtract” models:

$$G + H = J$$
$$G - H = K$$

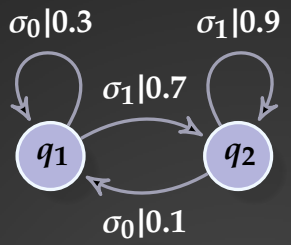
$$G - G = ?$$

The Zero Machine



G

+

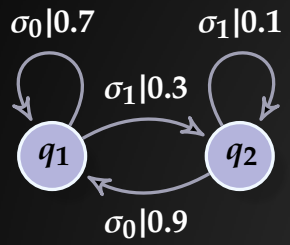


−G

=

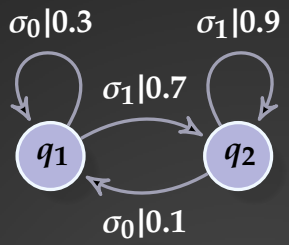
W
(Flat White Noise)

The Zero Machine



G

+

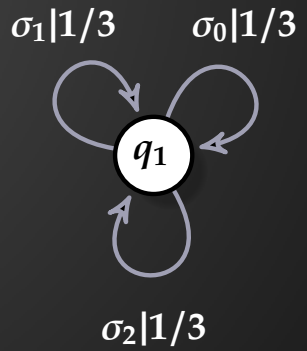


−G

=

W
(Flat White Noise)

Zero PFSA
for binary alphabet



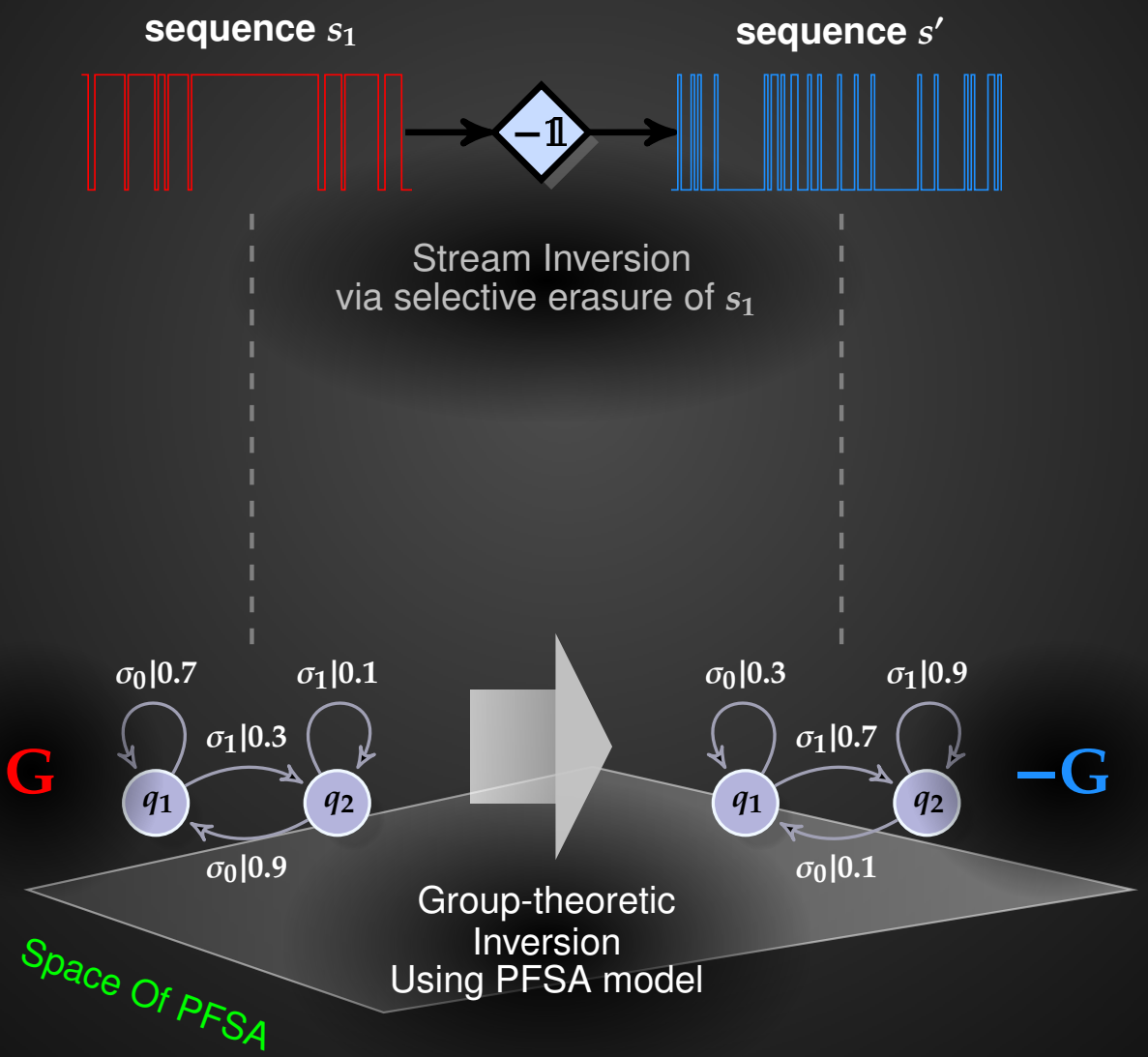
Zero PFSA
for trinary alphabet

The Zero Machine

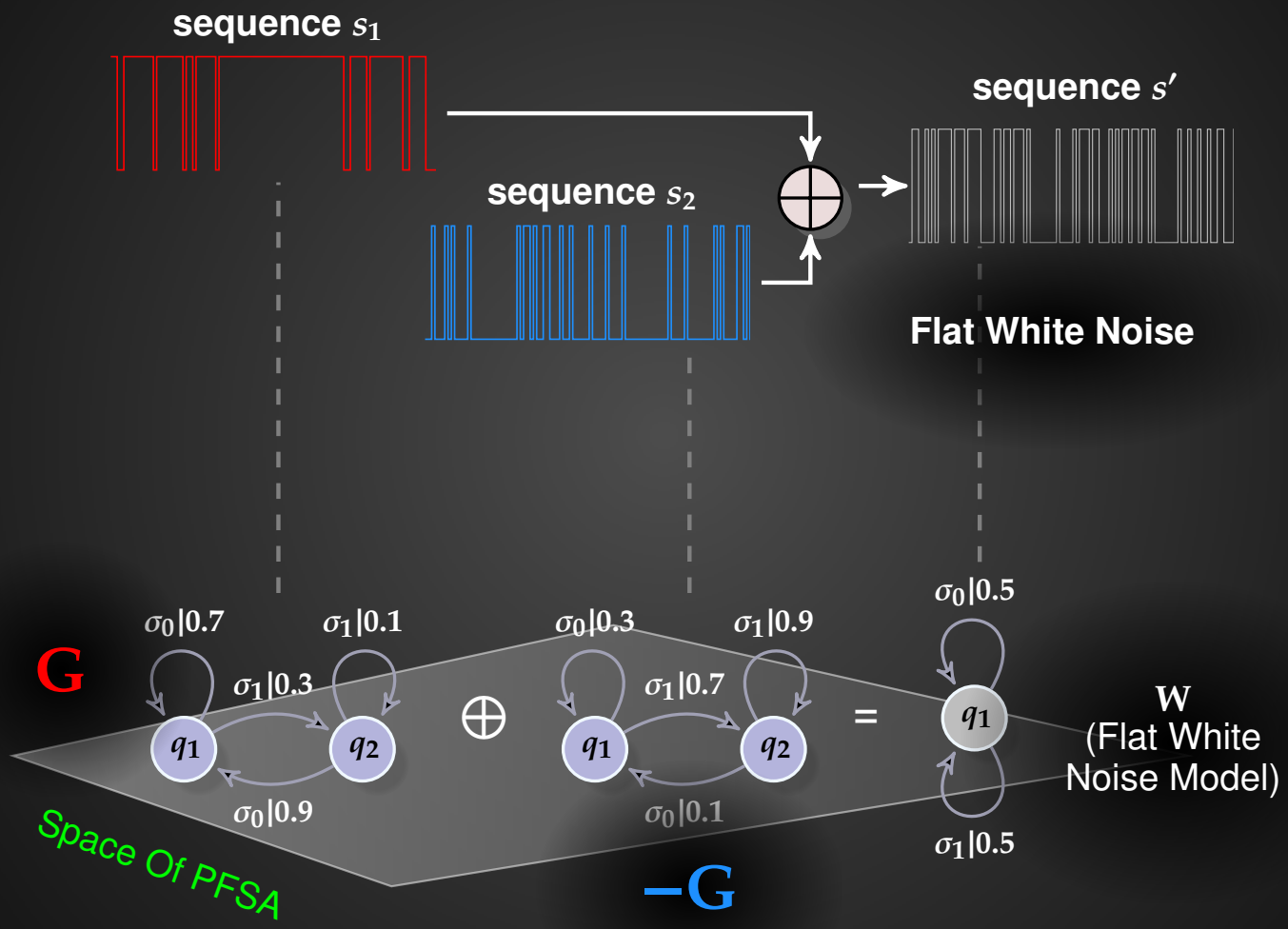
- Maximum entropy rate
- History is useless for prediction
- Encodes minimum information

Stream Inversion

Direct Generation of Anti-streams

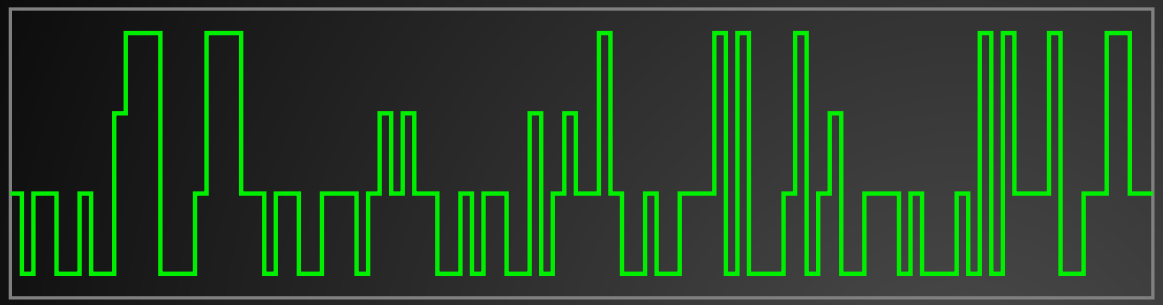


Annihilation Identity

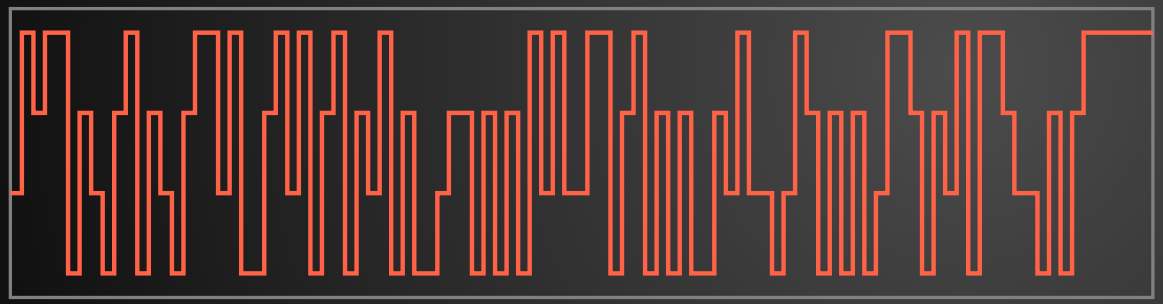


Example Of Stream, Anti-stream, & FWN

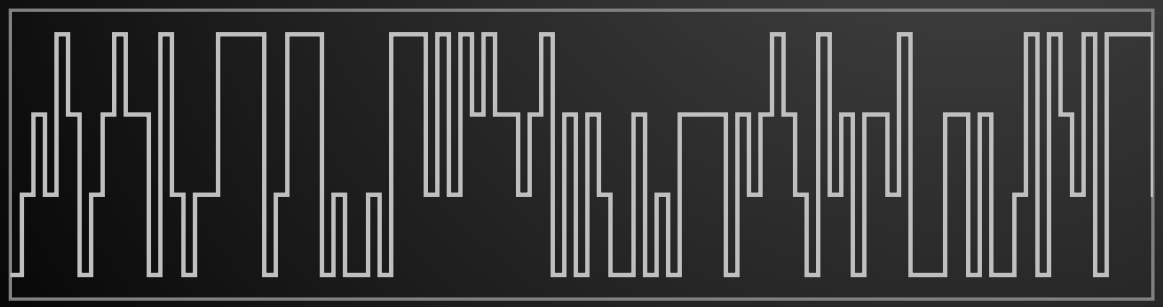
4 Letter Alphabet



Signal



Inverse Signal

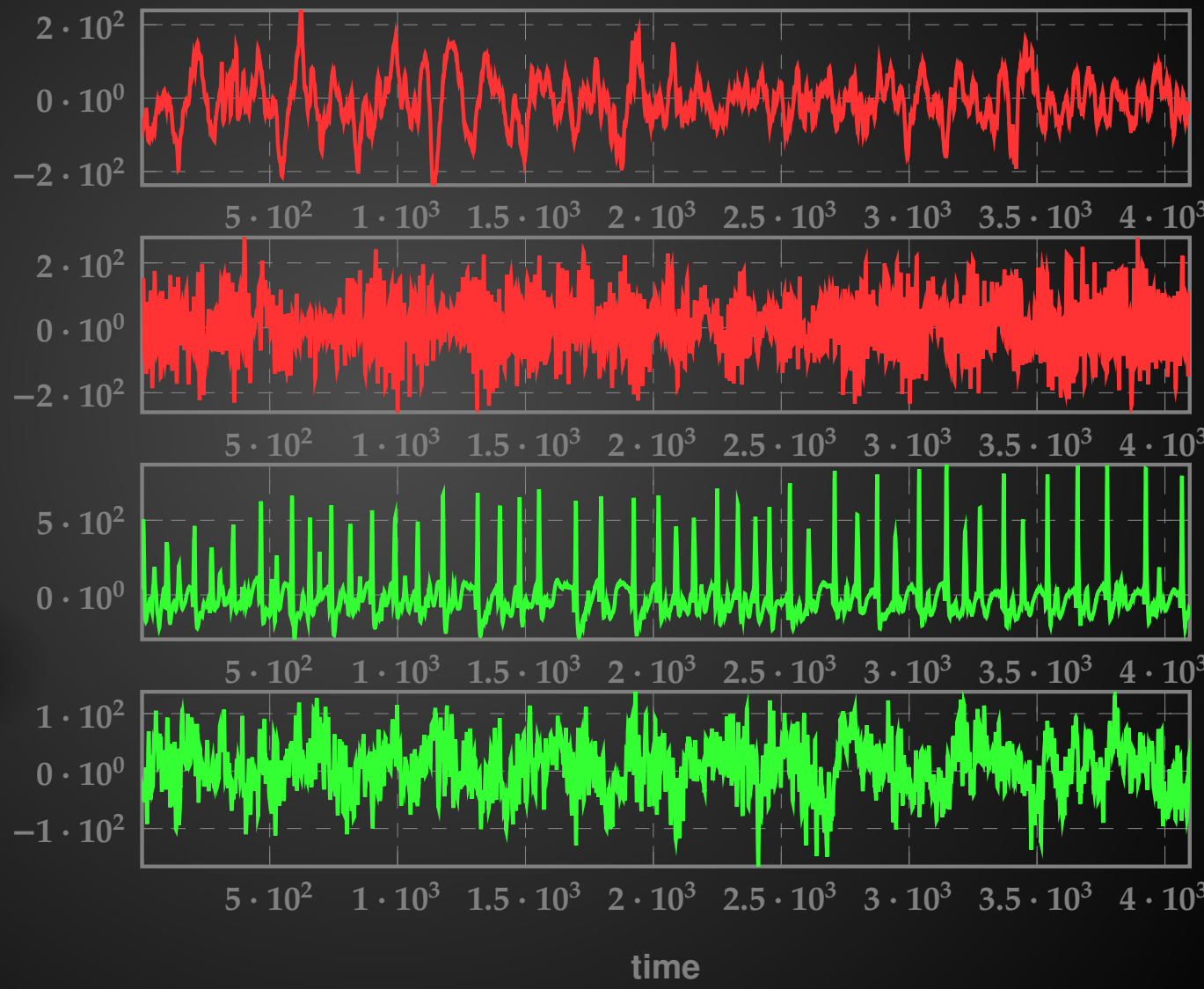


Flat White Noise

EEG - Epileptic Pathology

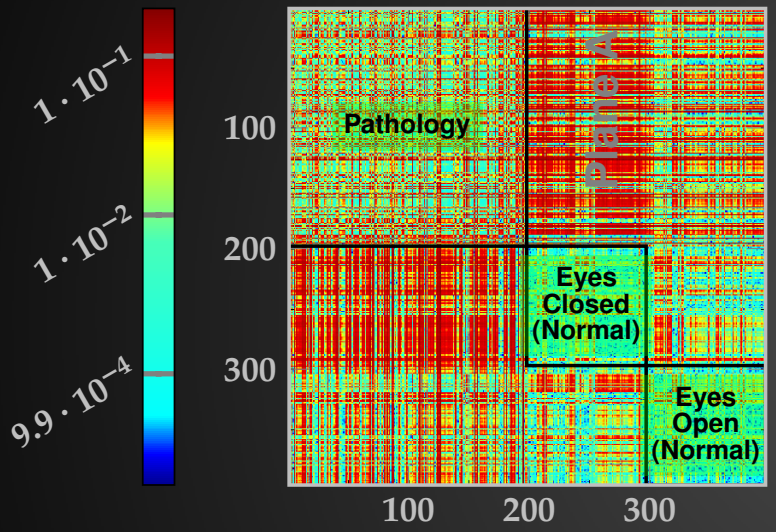
Eyes Open

Epileptic Pathology

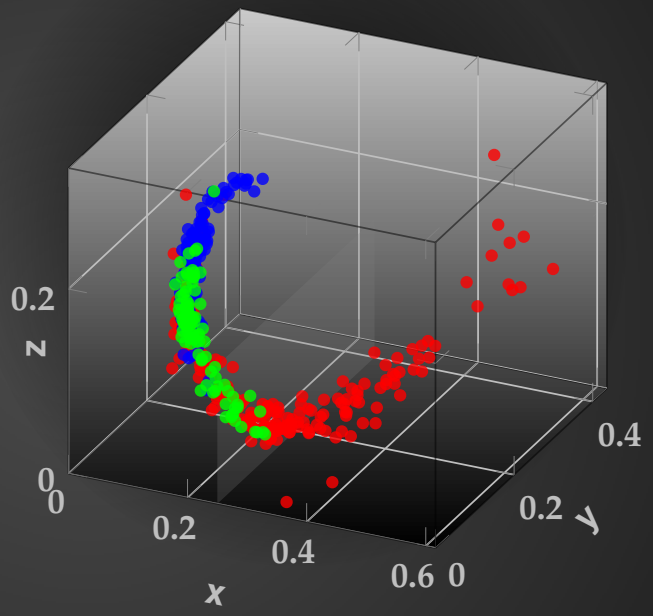


EEG - Epileptic Pathology

Pairwise Distance Matrix



3D Euclidean Embedding



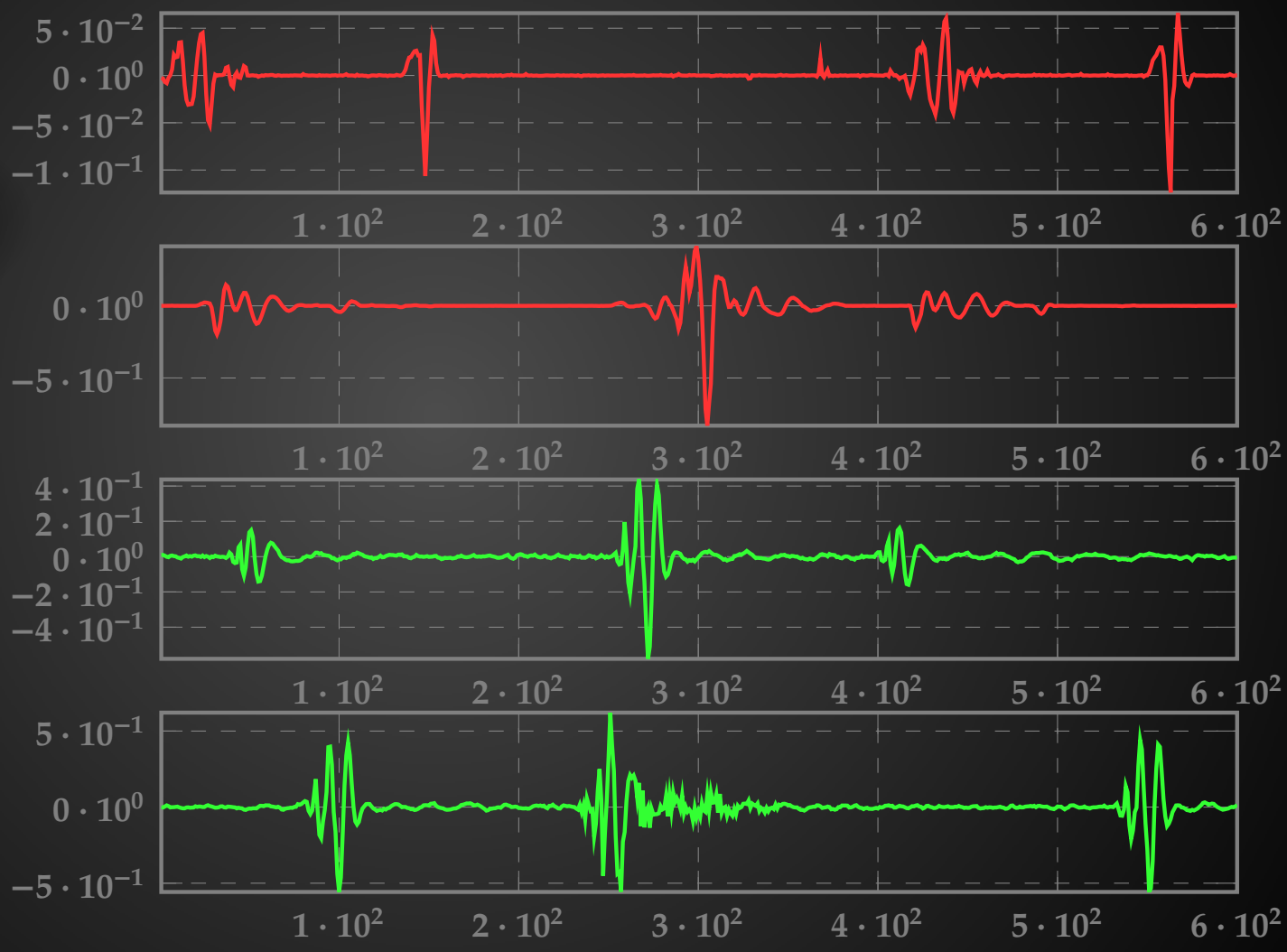
- Anomaly
- Normal (Eyes closed)
- Normal (Eyes open)

Cardiac Pathology

Disambiguate Normal Rhythm from Murmur

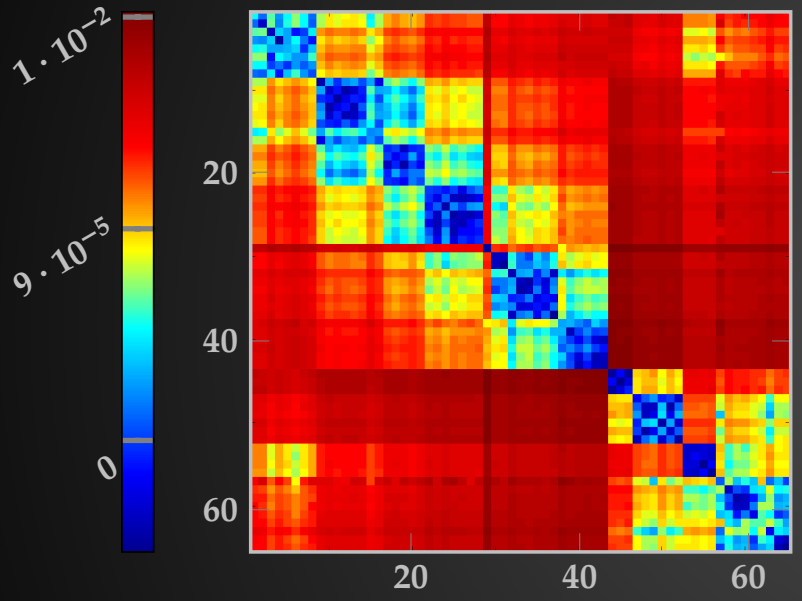
Normal Rhythm

Murmur

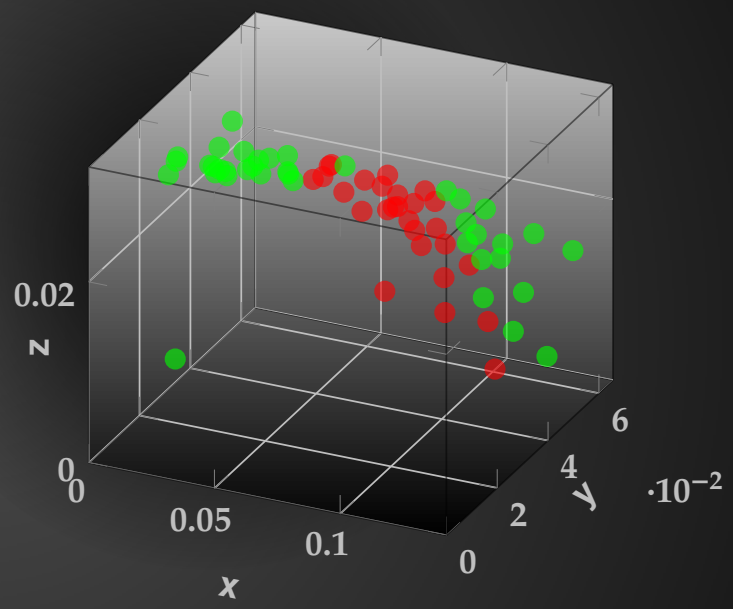


Cardiac Pathology

Distance Matrix



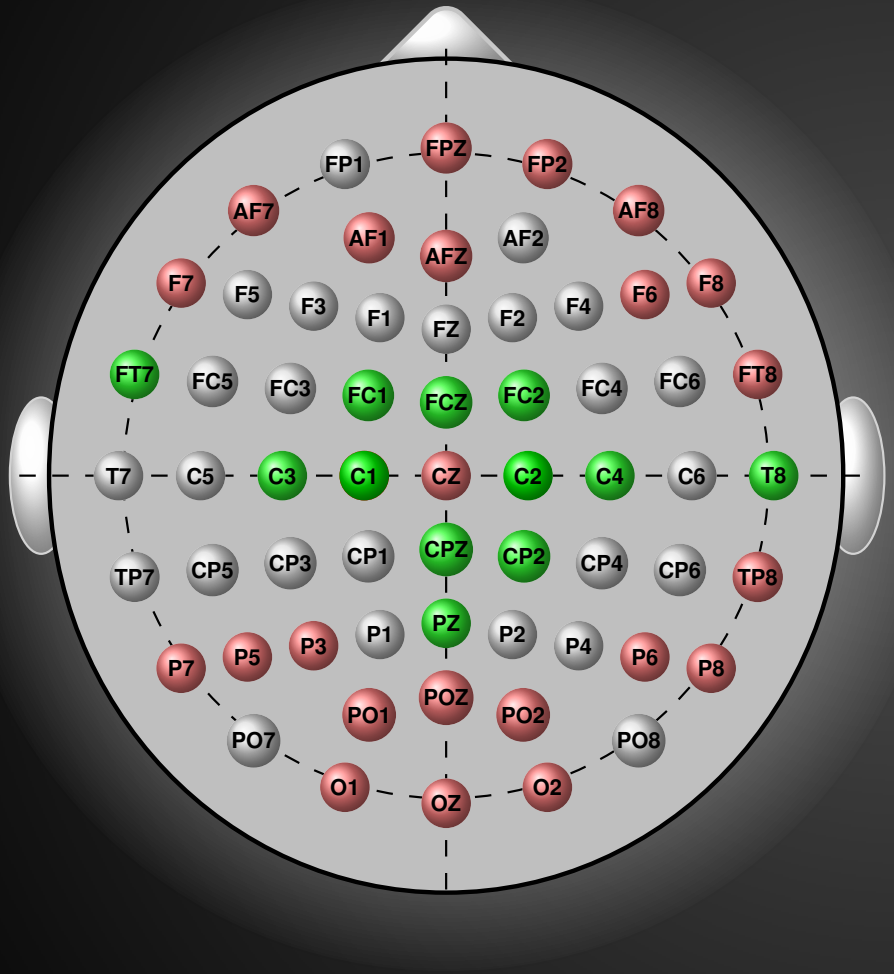
3D Euclidean Embeddin

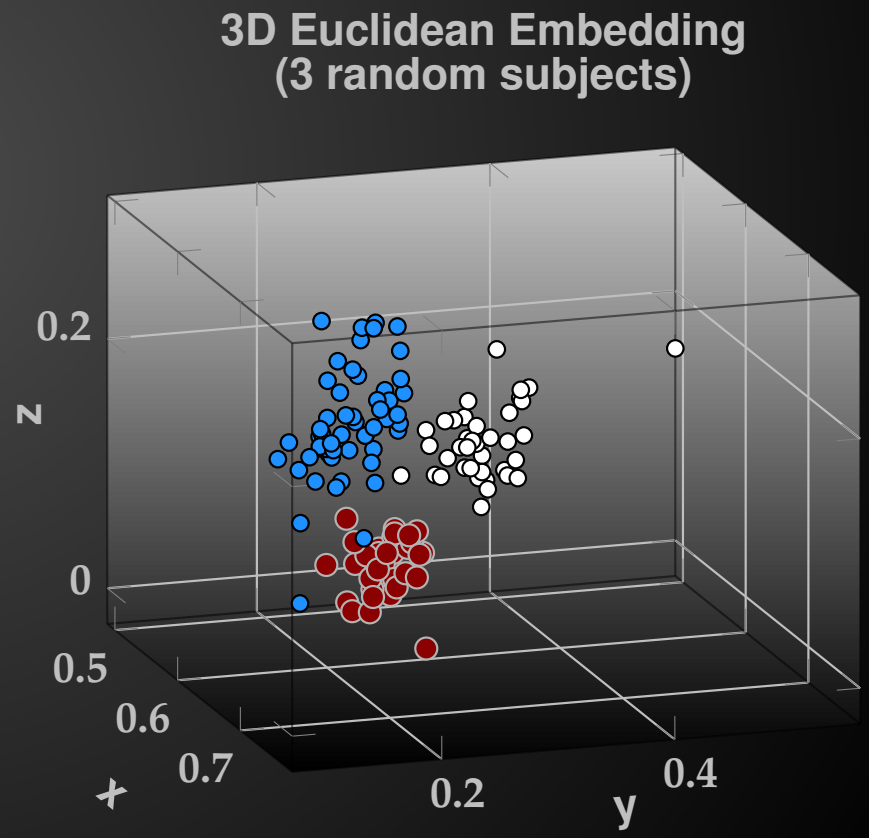


Murmur
Healthy

EEG Based Biometric Authentication

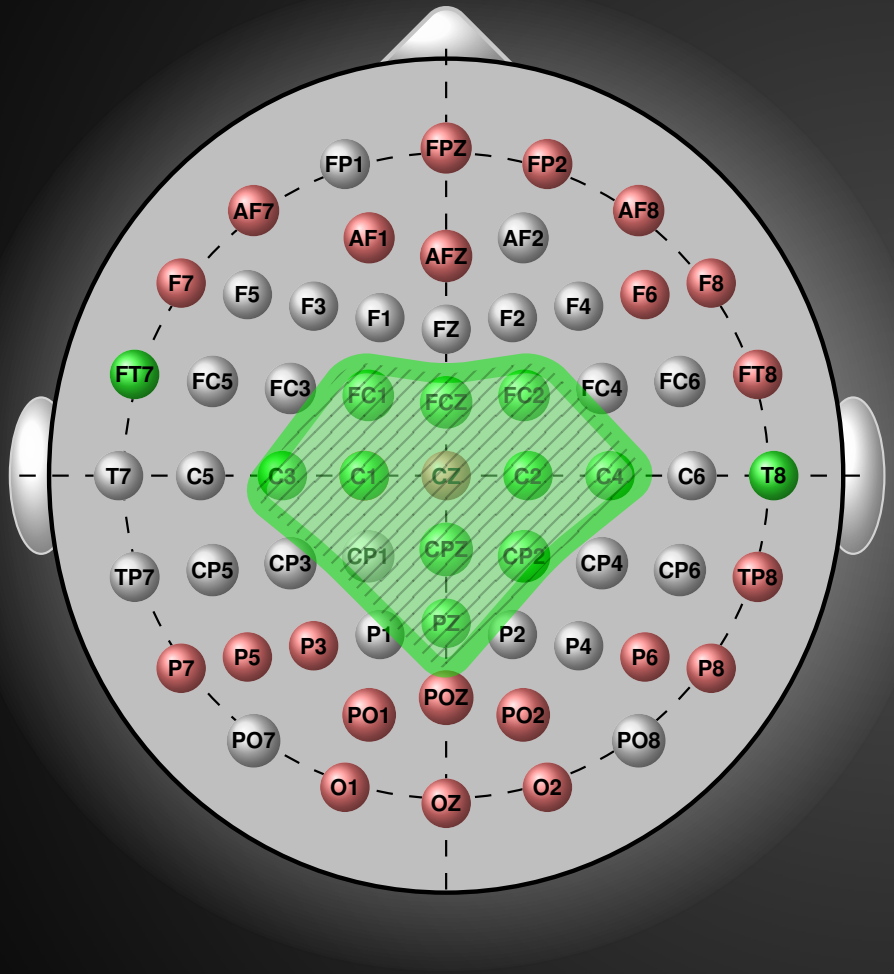
122 Subjects, 100% Accuracy

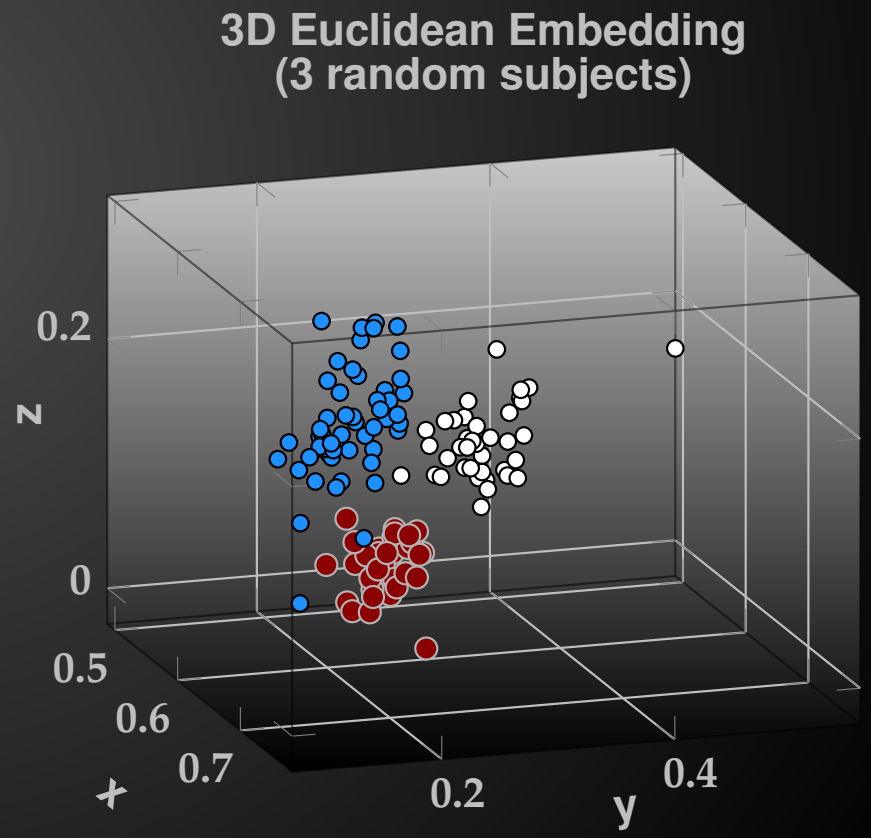




EEG Based Biometric Authentication

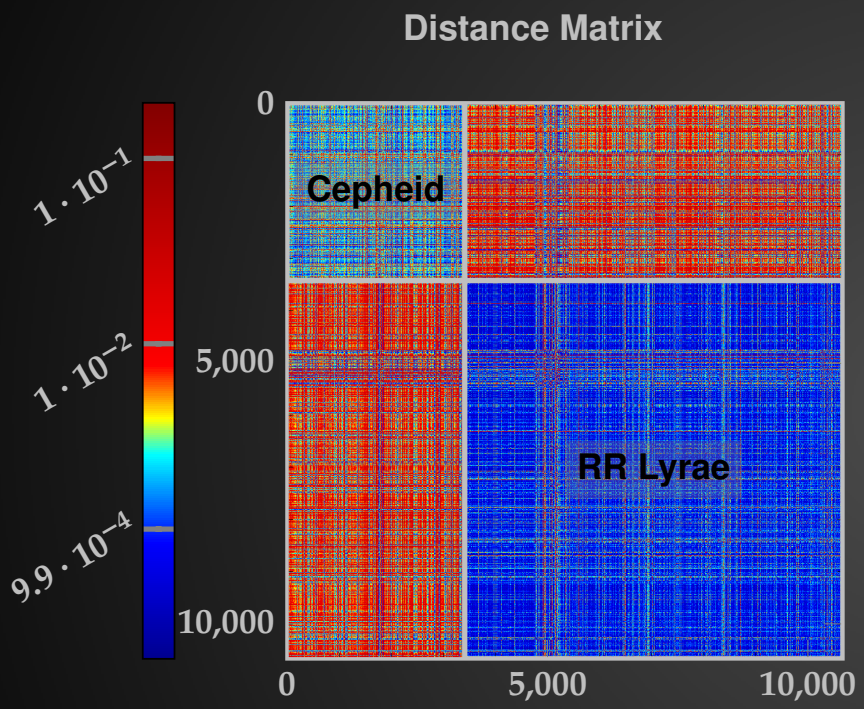
122 Subjects, 100% Accuracy

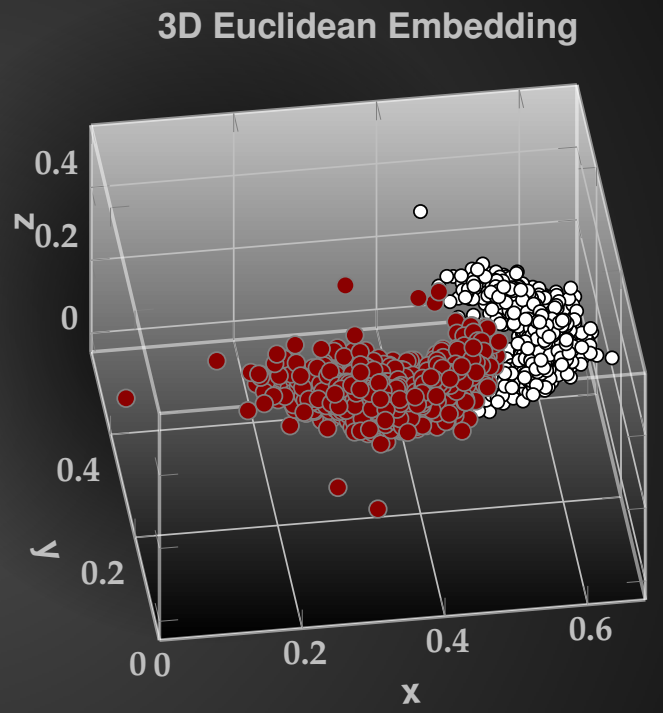




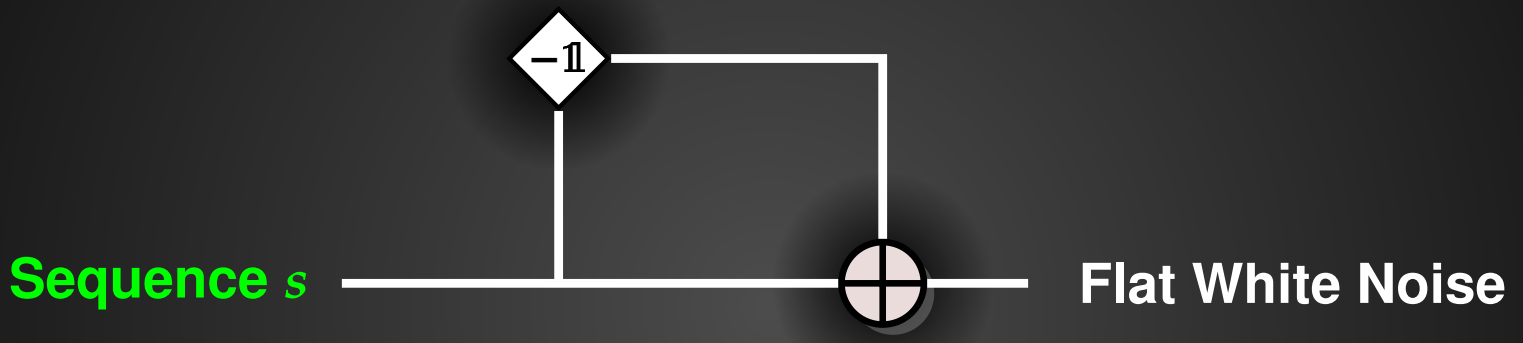
Classification of Variable Stars

Optical Gravitational Lensing Experiment (OGLE) database

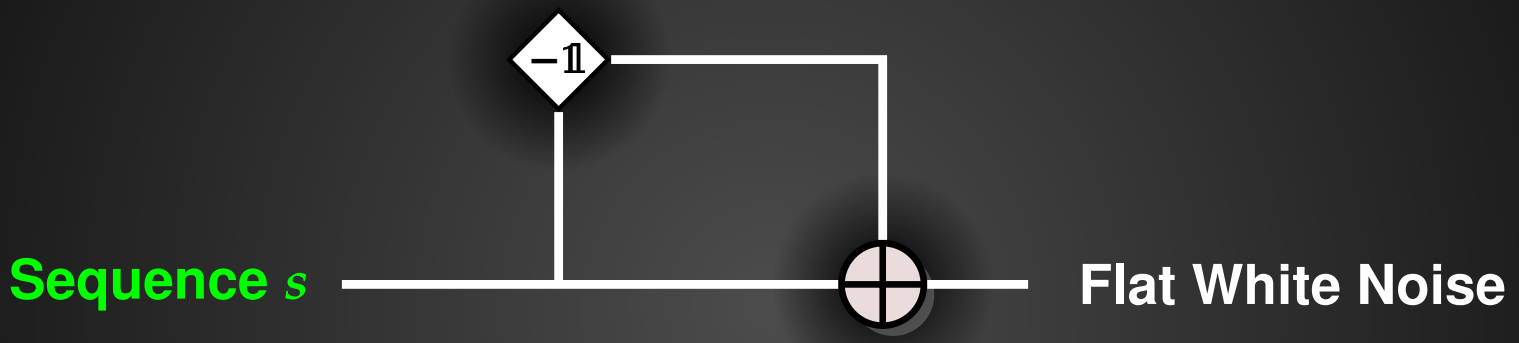




Self Annihilation Error

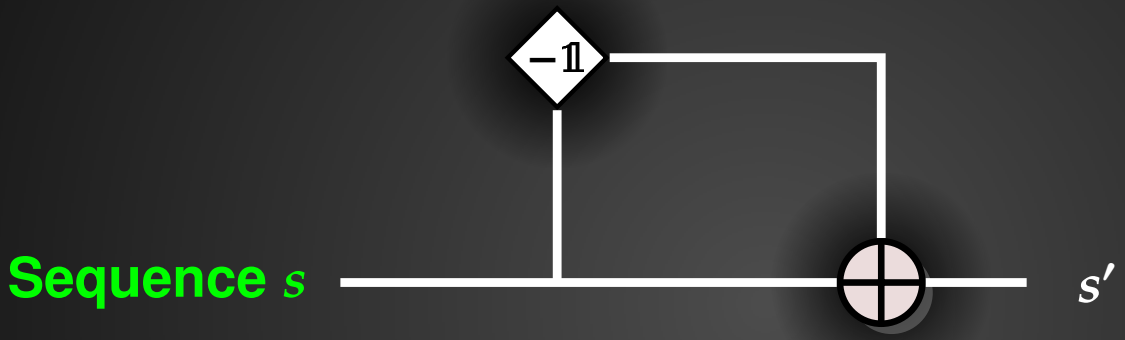


Self Annihilation Error



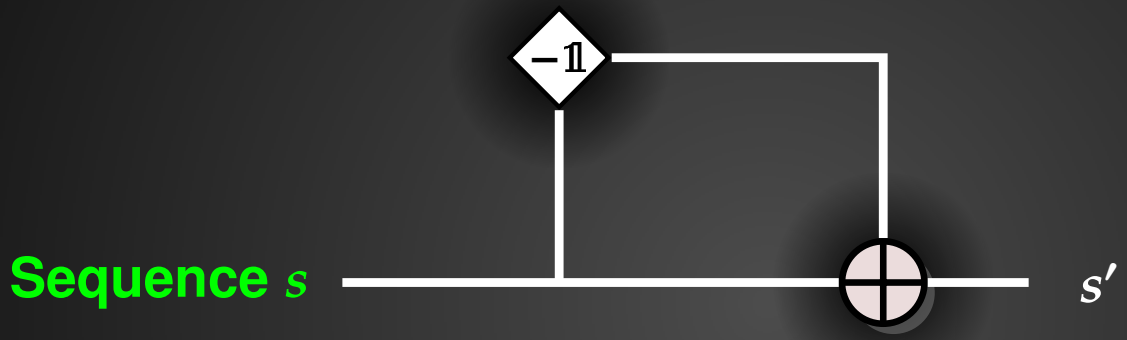
Only if $|s|$ is large enough !

Self Annihilation Error



$$\epsilon_s = \Theta(s', W)$$

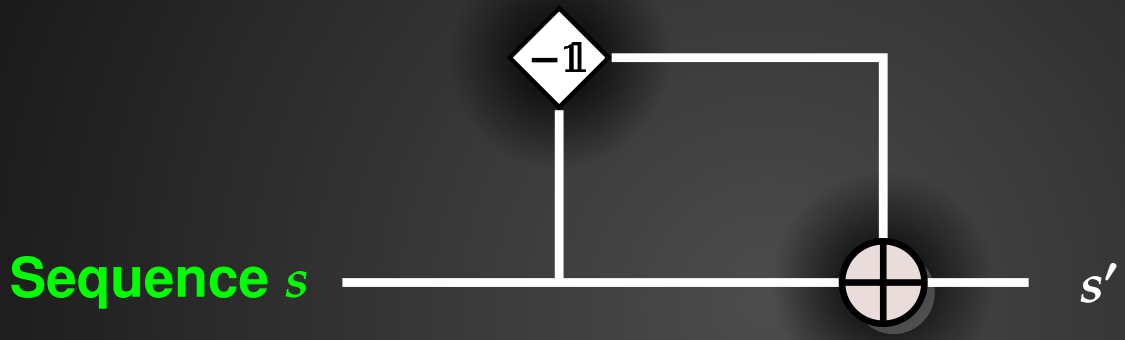
Self Annihilation Error



$$\epsilon_s = \Theta(s', W)$$

- $\epsilon_s \to 0$ exponentially fast with $|s|$
- Information content of stream

Self Annihilation Error



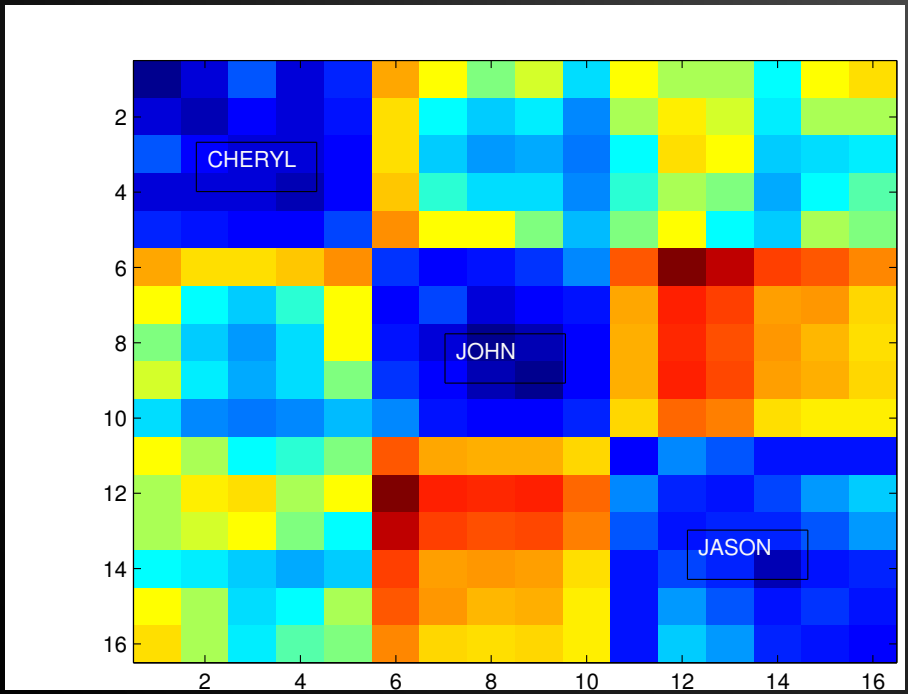
$$\epsilon_s = \Theta(s', W)$$

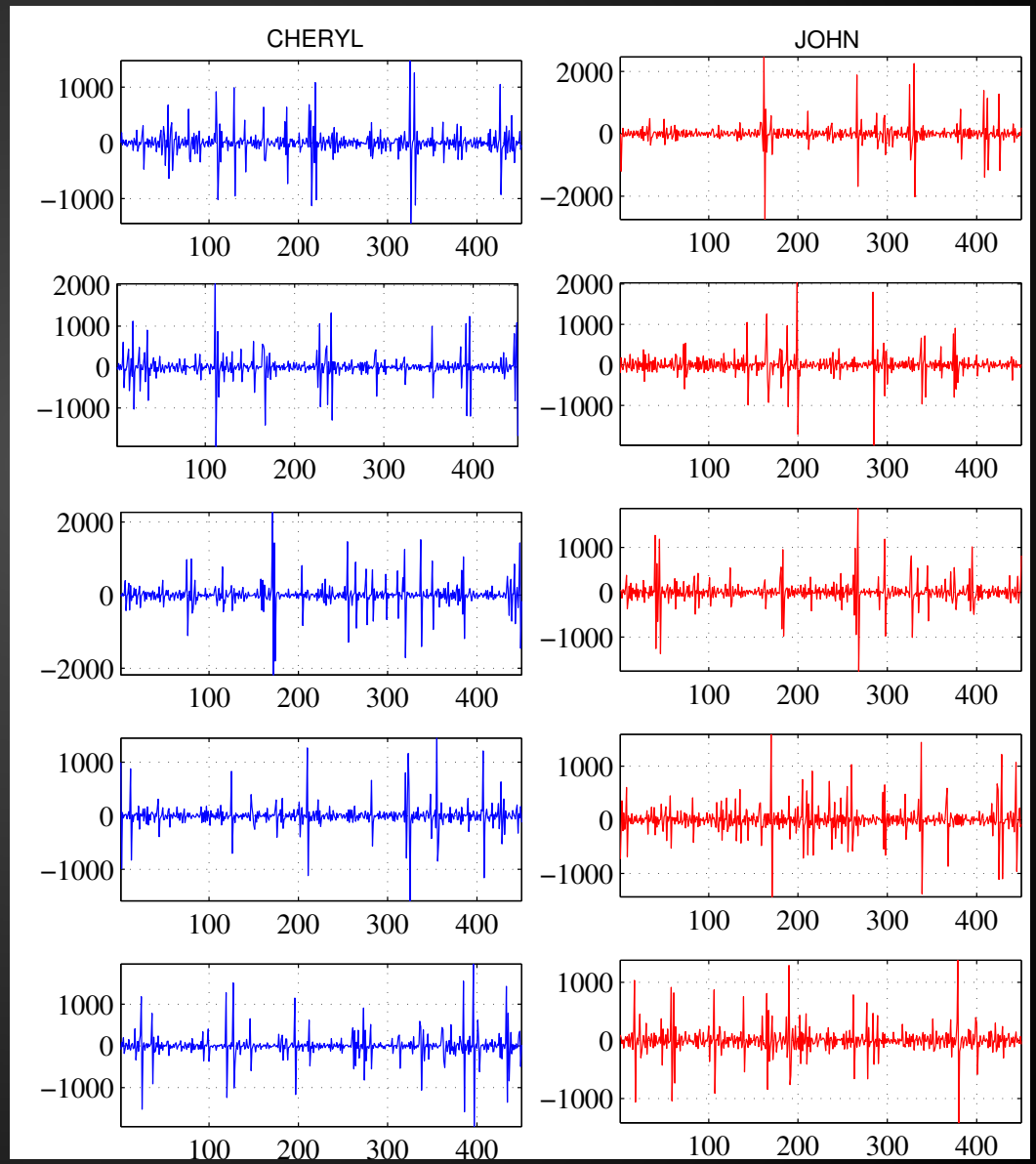
Auto-detect Data Sufficiency!

Cognitive Fingerprinting

User Authentication From Keypress Dynamics

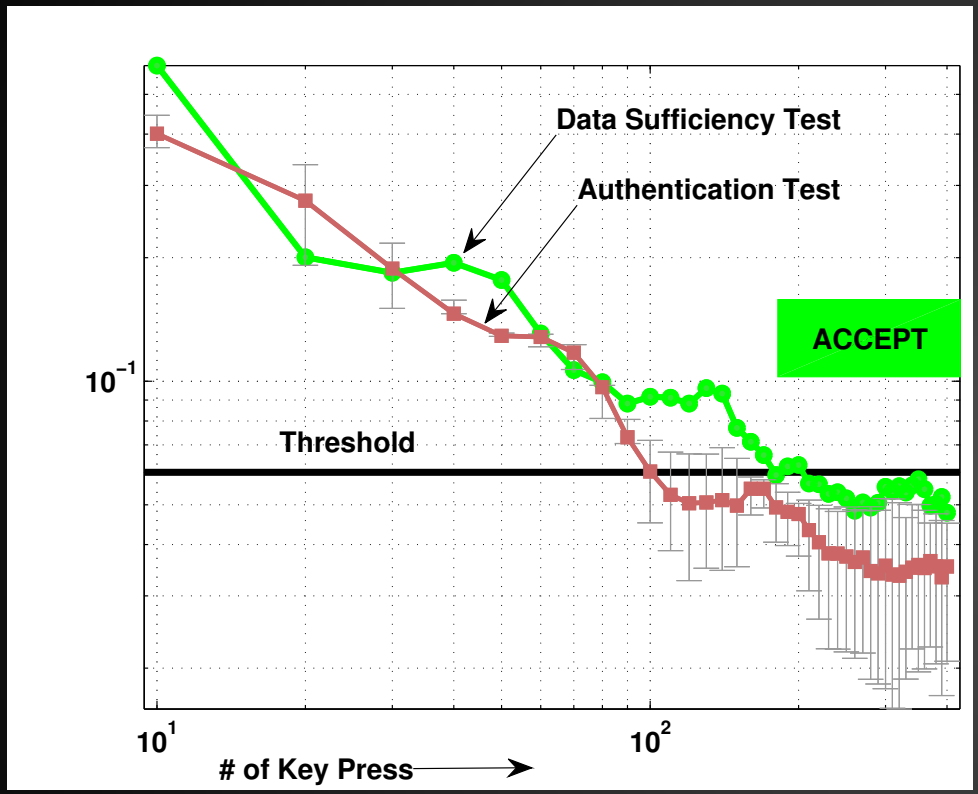
- **Time-series of keypress delays**
- **Random text**



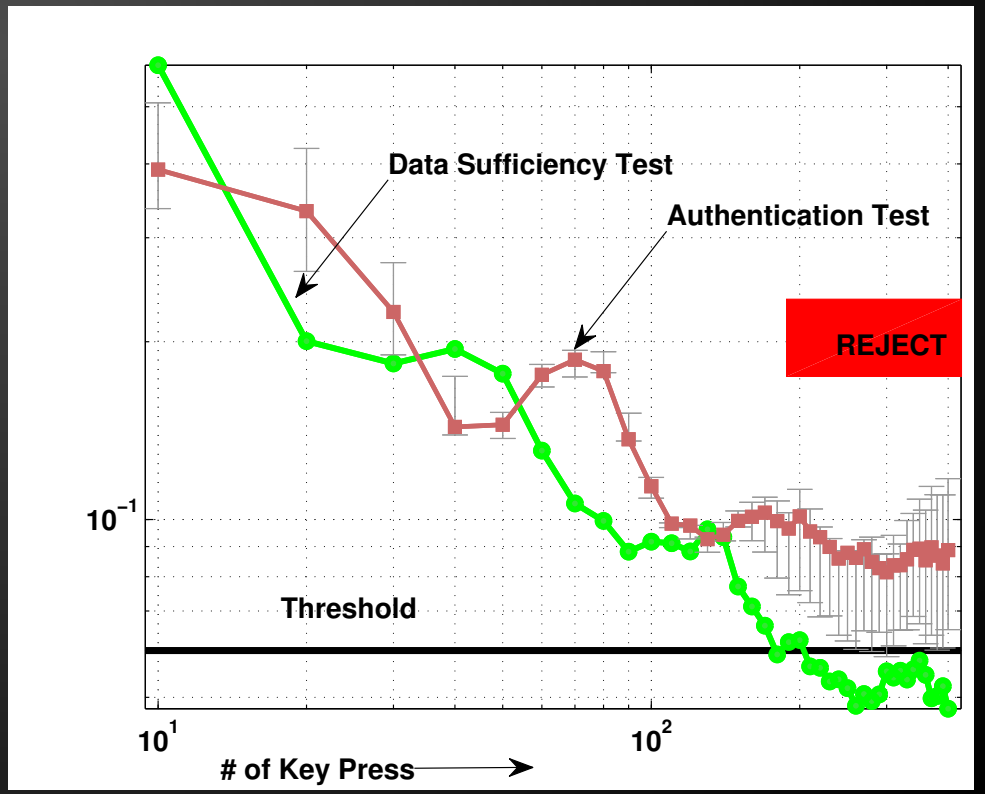


Cognitive Fingerprinting

User Authentication From Keypress Dynamics



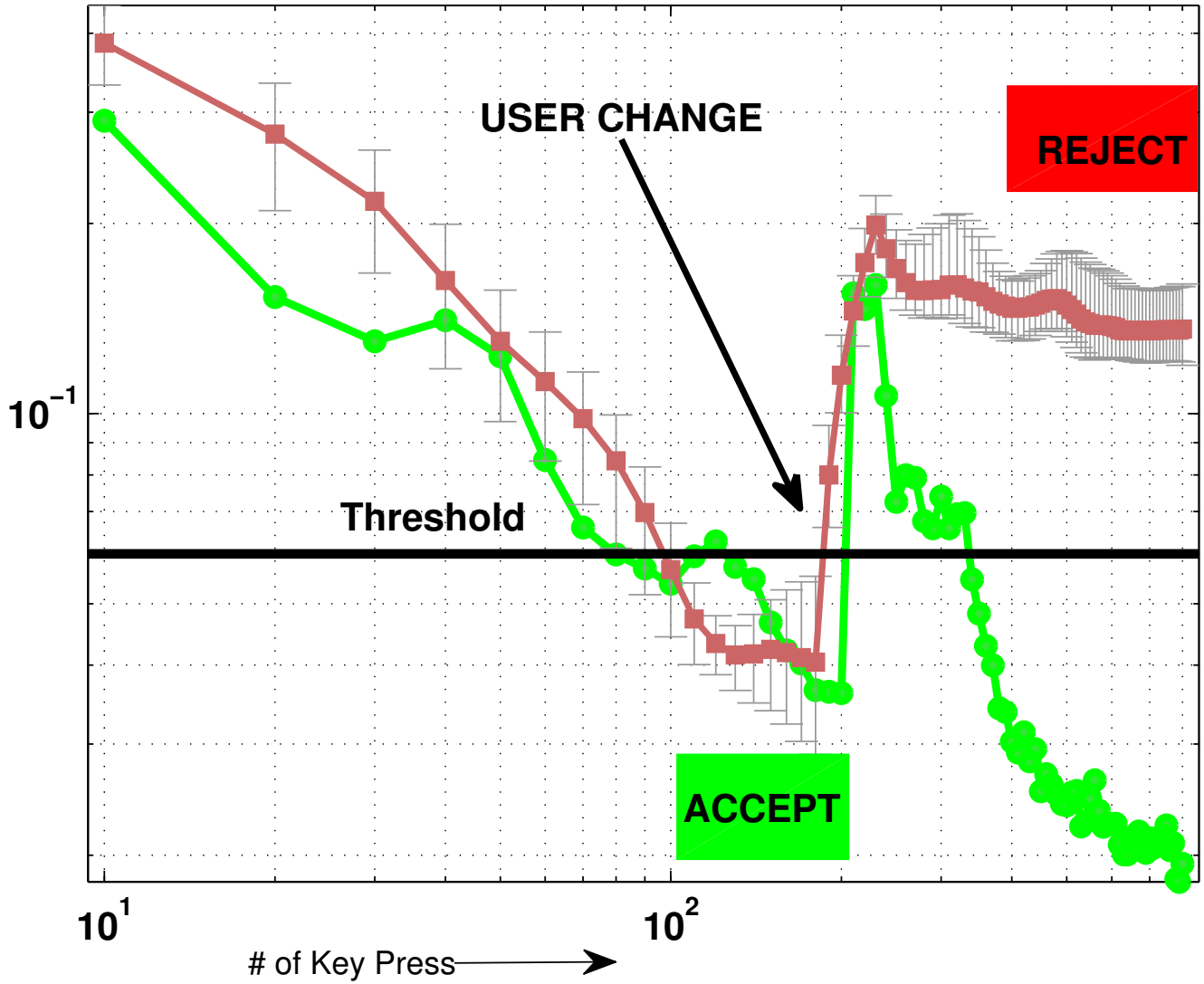
User Authorized



User Rejected

Cognitive Fingerprinting

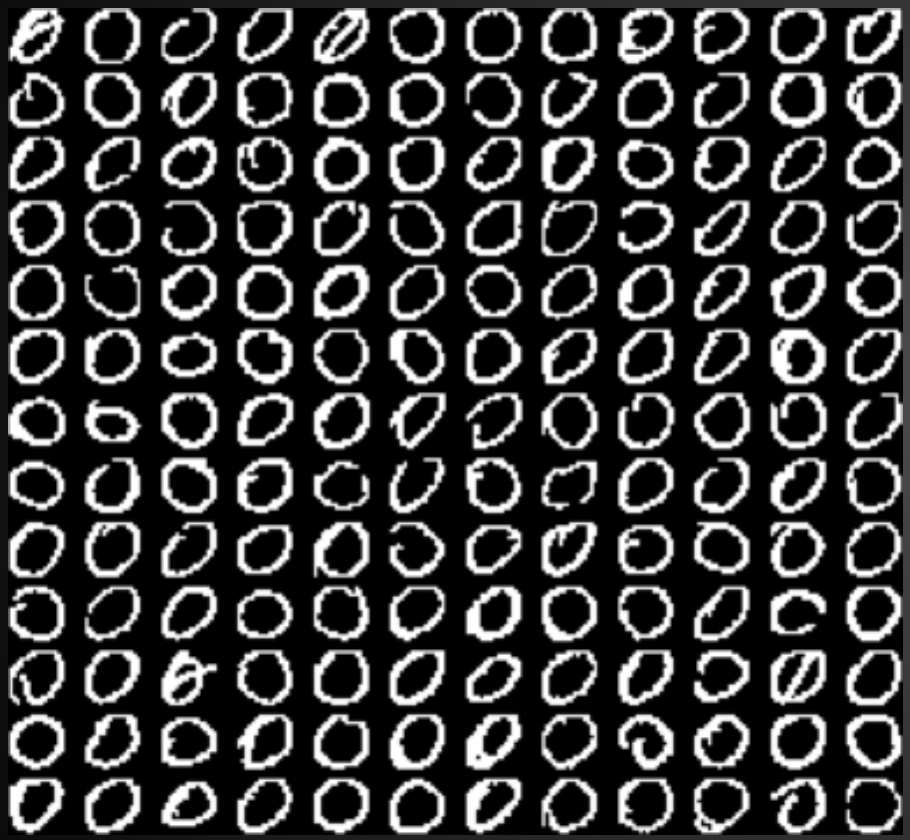
User Authentication From Keypress Dynamics

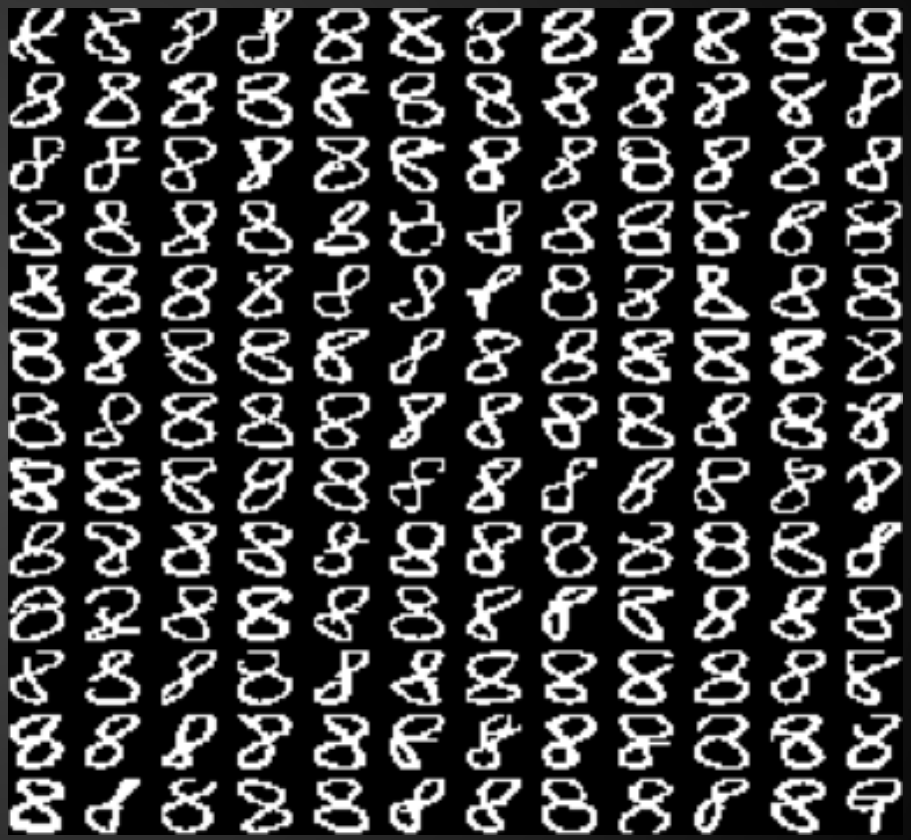


Authorization revoked on unexpected user change detection

Handwritten Digits

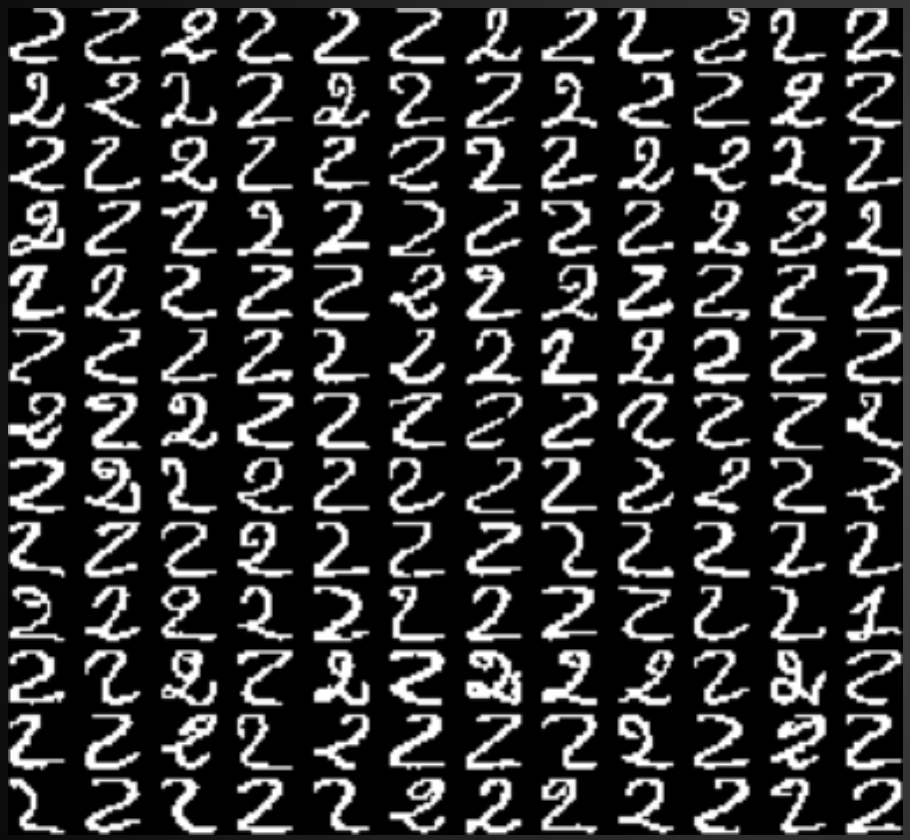
Recognition / Classification Problem

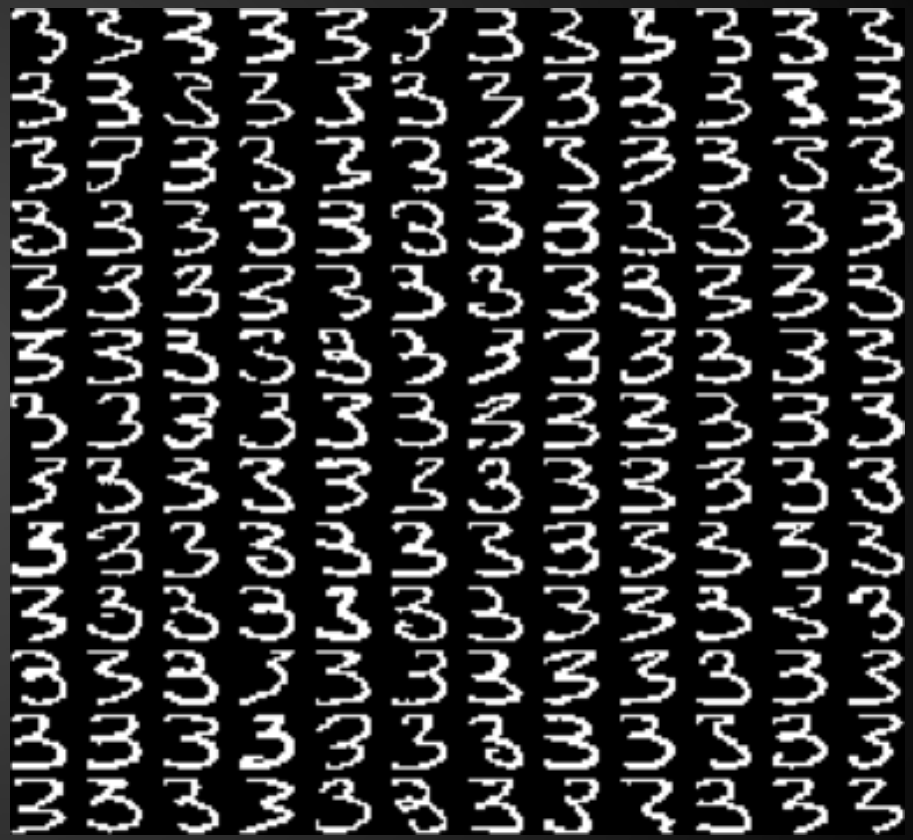




Handwritten Digits

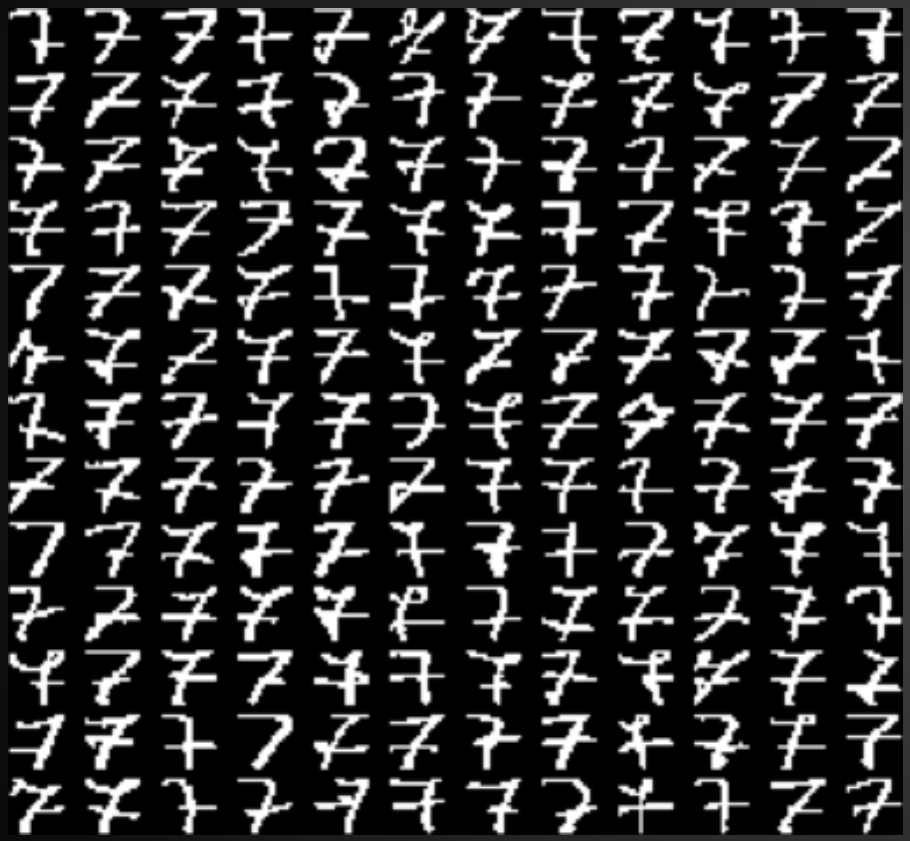
Recognition / Classification Problem

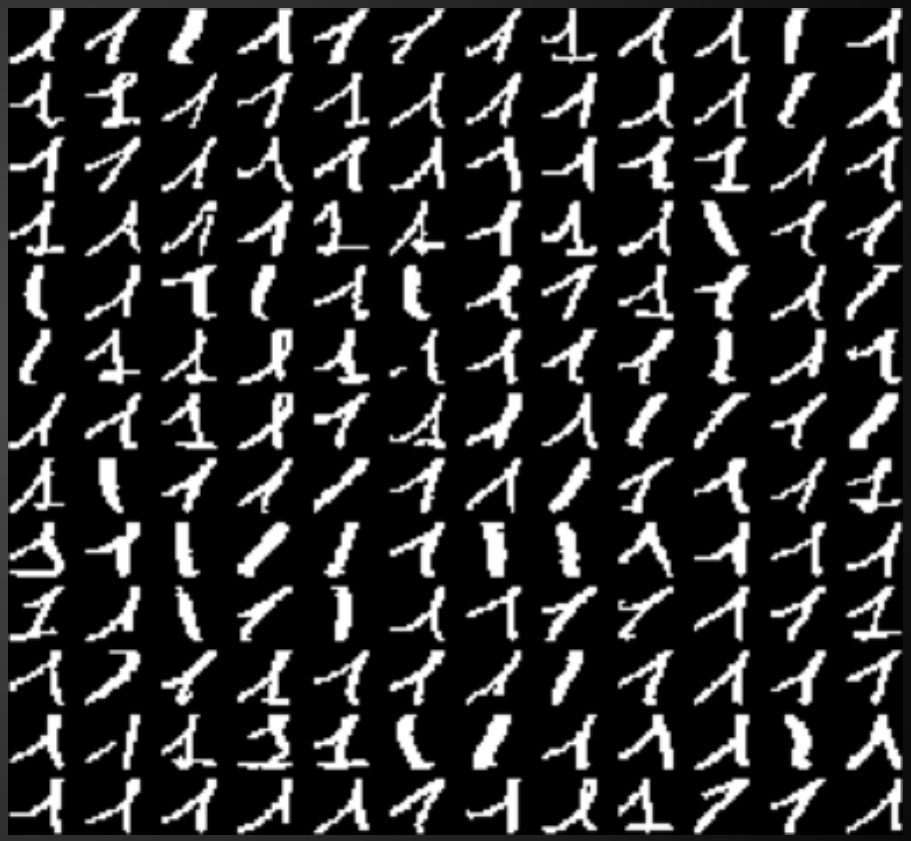




Handwritten Digits

Recognition / Classification Problem



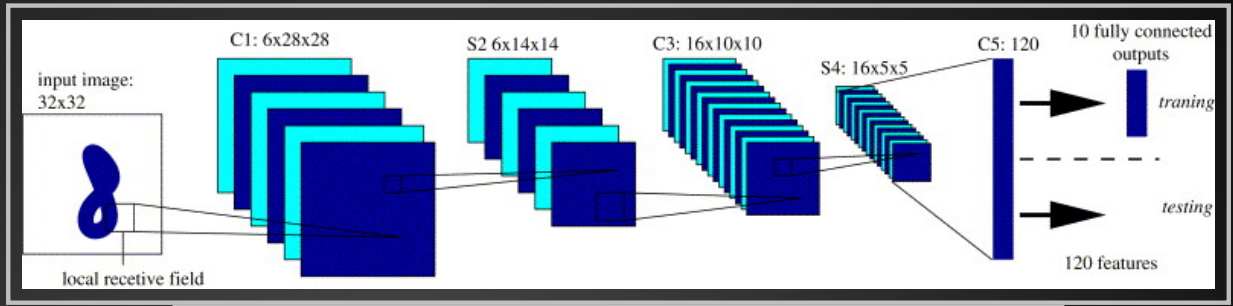


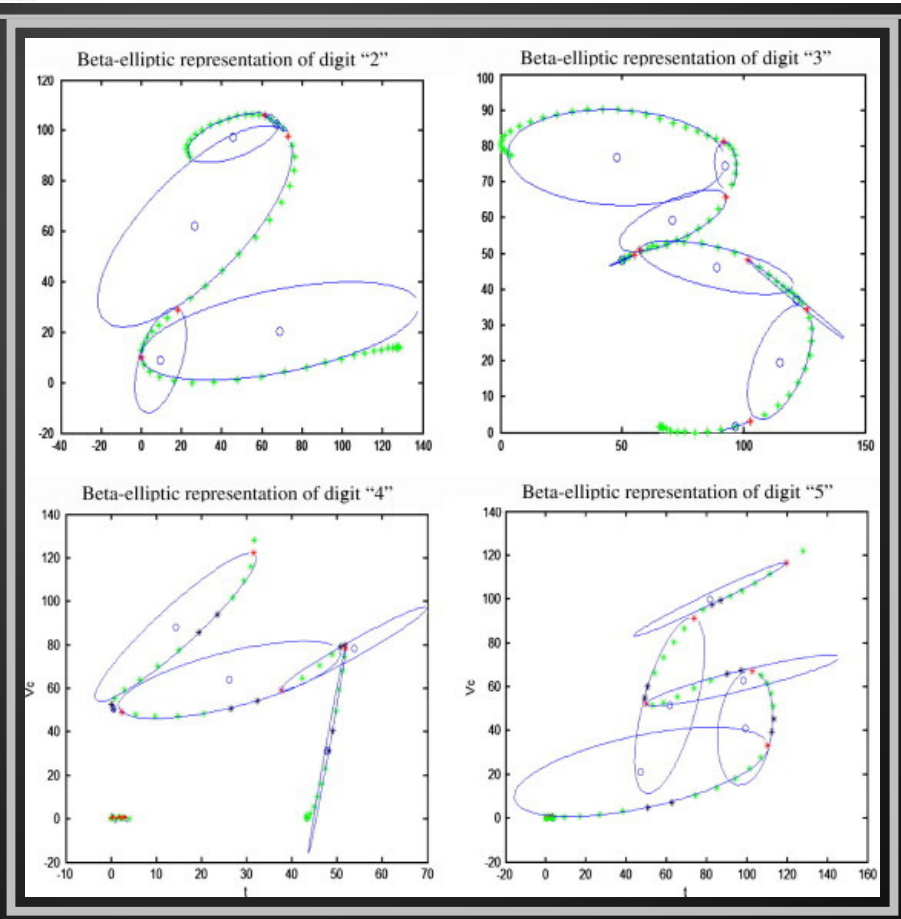
Handwritten Digits

Recognition / Classification Problem

State of Art Approaches

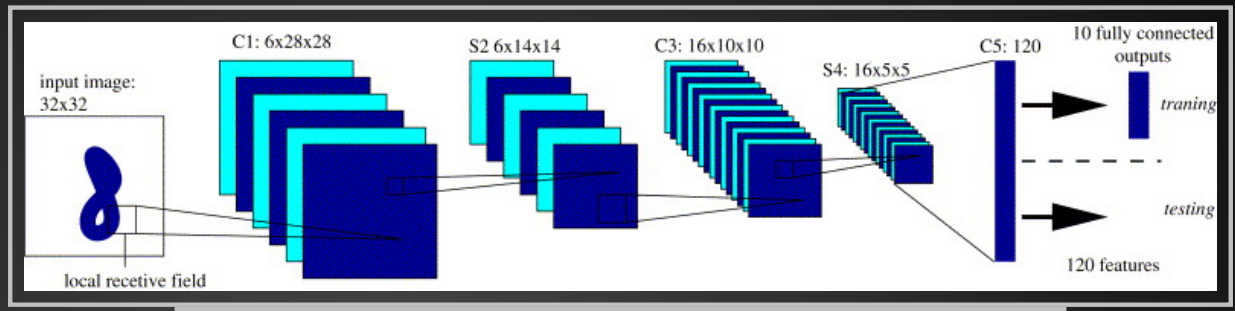
Find Clever Representations

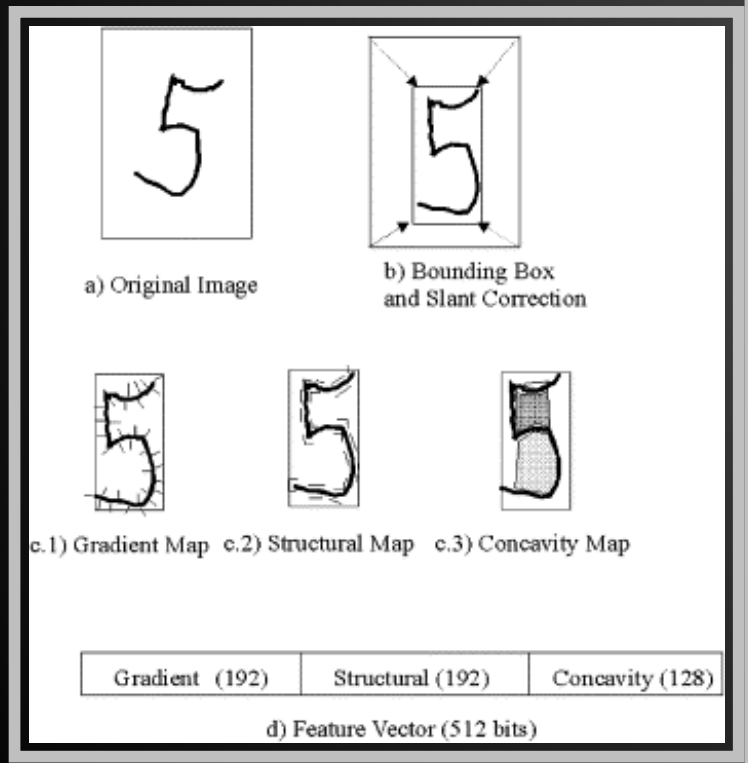


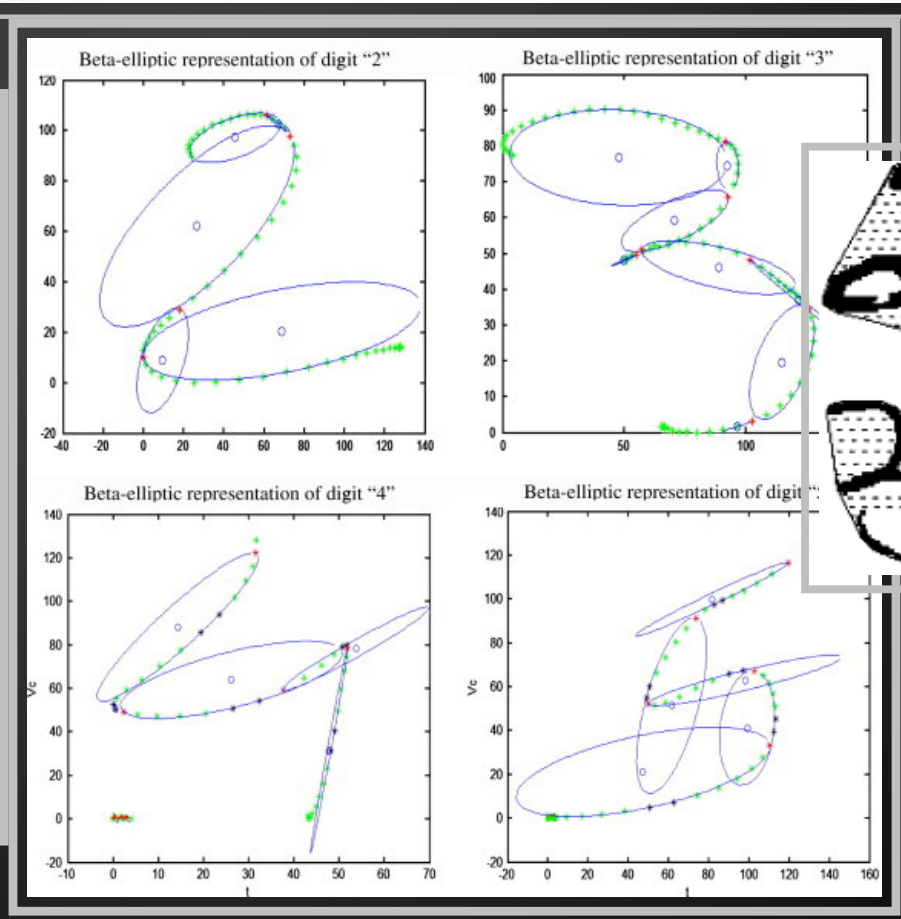


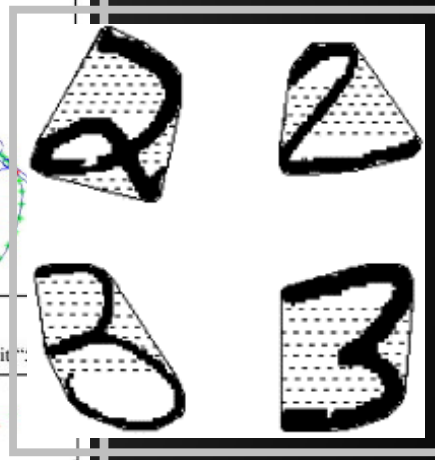
State of Art Approaches

Find Clever Representations



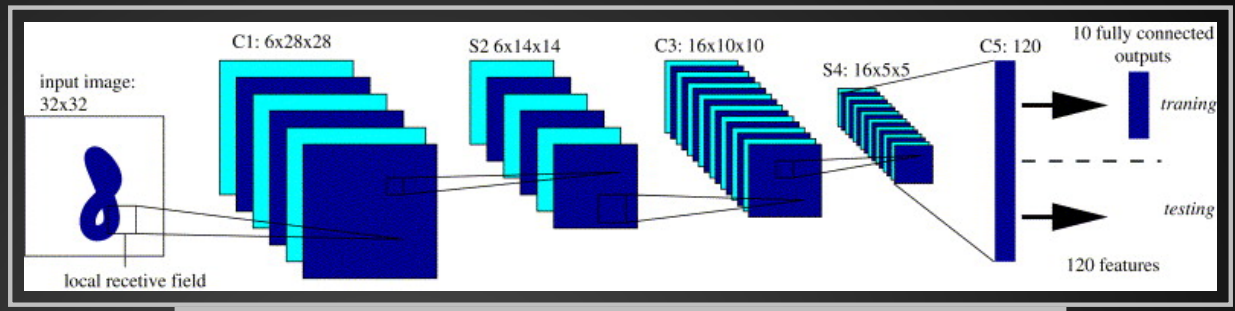


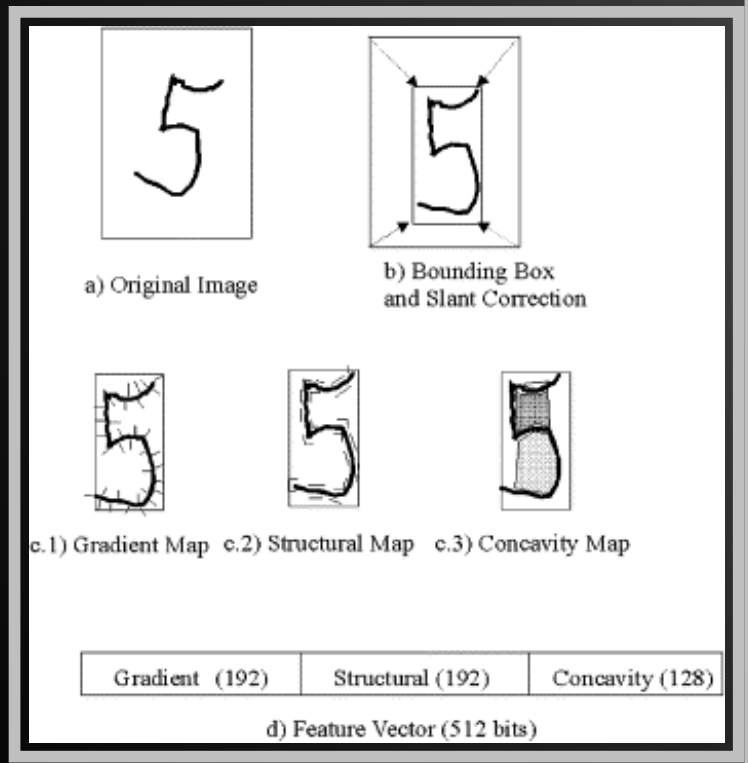


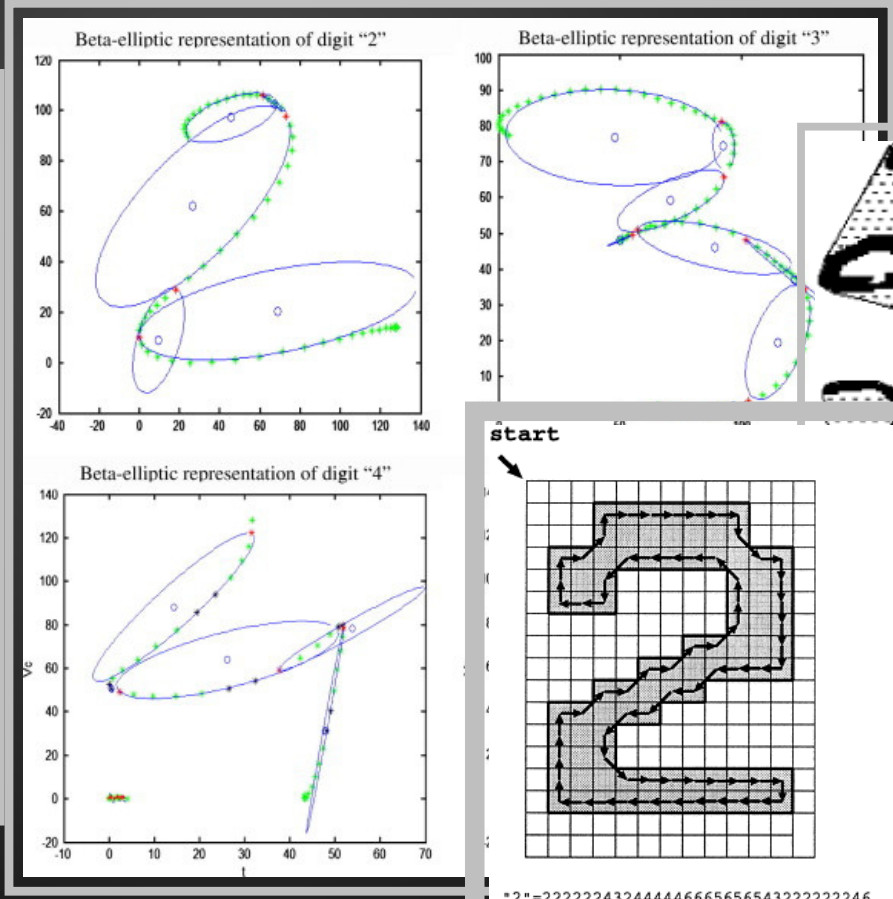


State of Art Approaches

Find Clever Representations

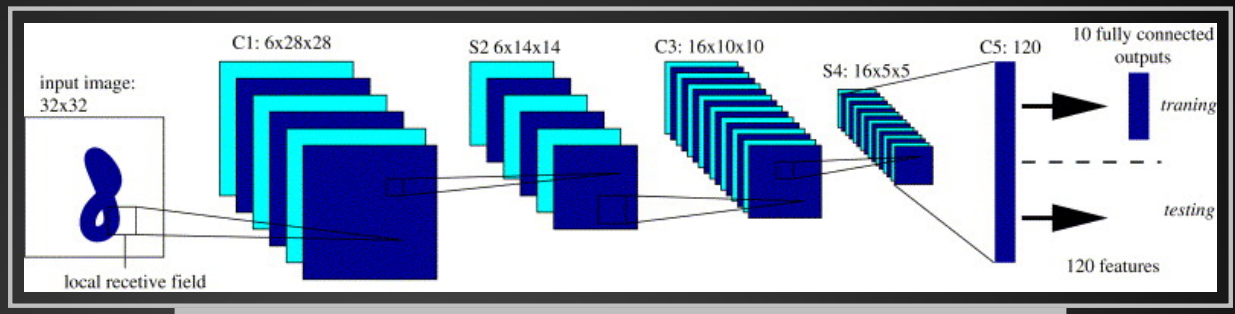


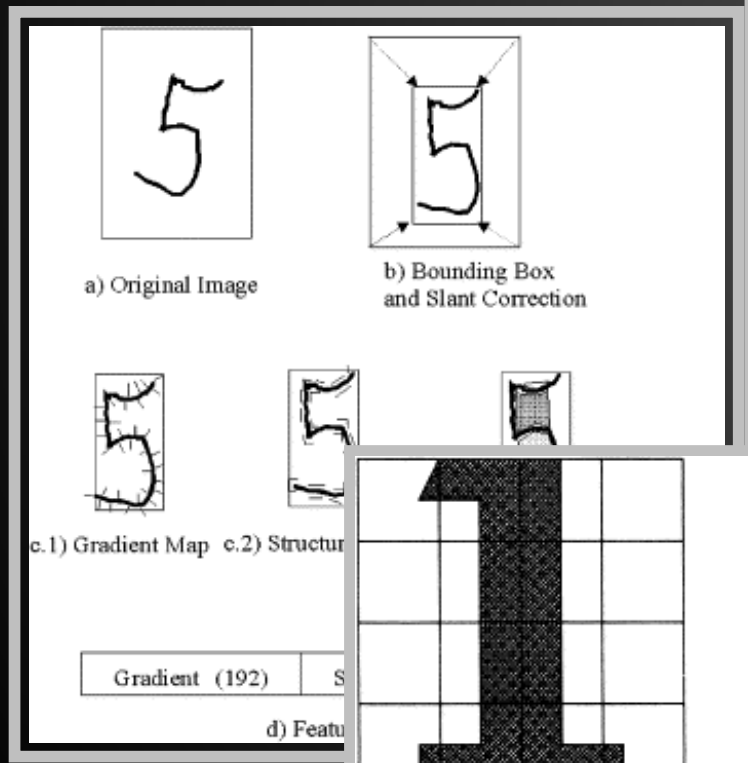


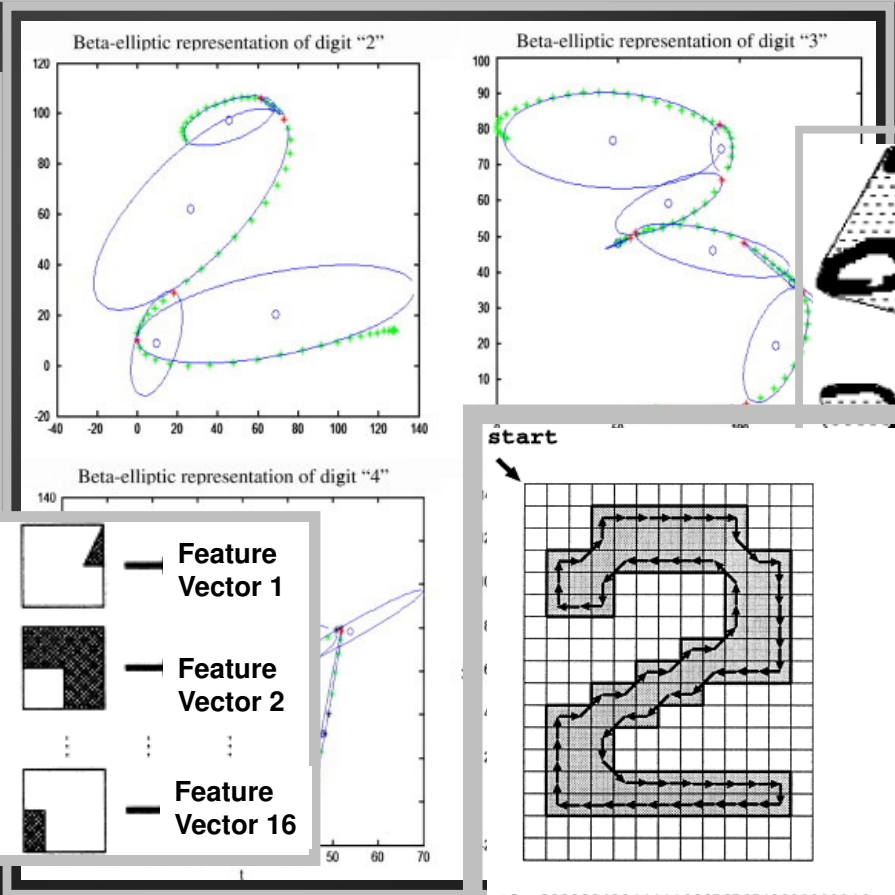


State of Art Approaches

Find Clever Representations







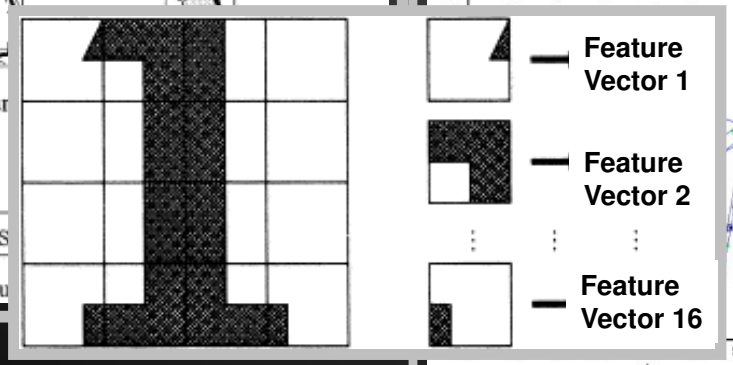
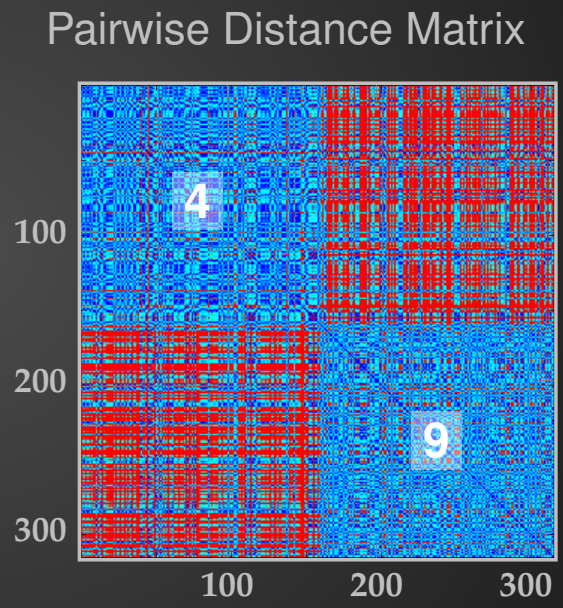


Image To Symbol Stream

Random Walk

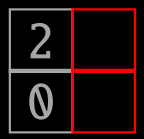
Handwritten Digits

Recognition / Classification Problem



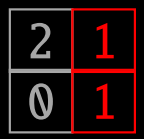
Memory-less Stream Manipulation

Generating First Independent Copy for String s_1:



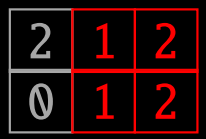
Memory-less Stream Manipulation

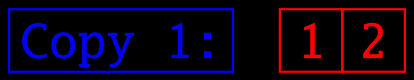
Generating First Independent Copy for String s_1:



Memory-less Stream Manipulation

Generating First Independent Copy for String s_1:





Memory-less Stream Manipulation

Generating First Independent Copy for String s_1:

2	1	2	0
0	1	2	2

Copy 1: | 1 | 2 |

Memory-less Stream Manipulation

Memory-less Stream Manipulation

Generating First Independent Copy for String s_1:

2	1	2	0	1	0	2	2	0	0	2	1	2	0	1	1	0	0	1	0	0	2	0	1	0	2
0	1	2	2	2	1	2	0	0	0	0	2	1	0	2	1	0	0	2	2	2	0	2	1	0	2

Copy 1:	1	2	2	0	0	0	1	0	0	1	0	2	1	0	2	2	2	1	2	2	1	2	2	1	2	1

Memory-less Stream Manipulation

Generating First Independent Copy for String s_1:

2	1	2	0	1	0	2	2	0	0	2	1	2	0	1	1	0	0	1	0	0	2	0	1	0	2
0	1	2	2	2	1	2	0	0	0	0	2	1	0	2	1	0	0	2	2	2	0	2	1	0	2

Copy 1:	1	2	2	0	0	0	1	0	0	1	0	2	1	0	2	2	2	1	2	2	1	2	2	1	2	1

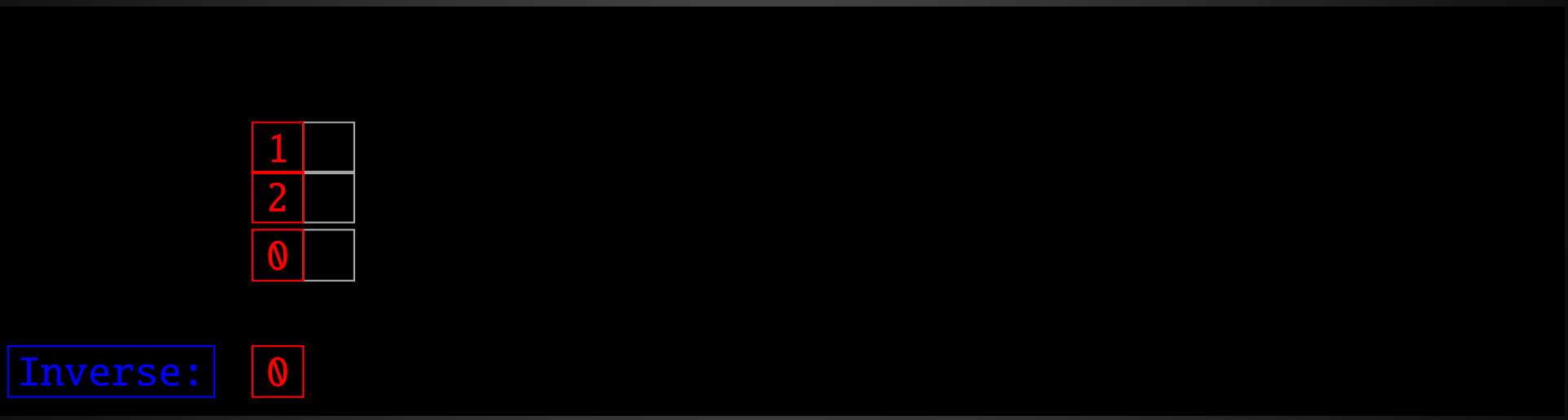
Generating Second Independent Copy for String s_1:

2	1	2	0	1	0	2	2	0	0	2	1	2	0	1	1	0	0	1	0	0	2	0	1	0	2
0	0	2	0	0	2	0	2	2	2	2	2	1	2	1	0	1	0	0	0	2	0	2	2	2	0

Copy 2:	2	0	2	2	1	0	0	2	2	2	1	2	2	0	1	2	2	1	2	1	2

Memory-less Stream Manipulation

Generating Inverse of String s_1:



Memory-less Stream Manipulation

Generating Inverse of String s_1:

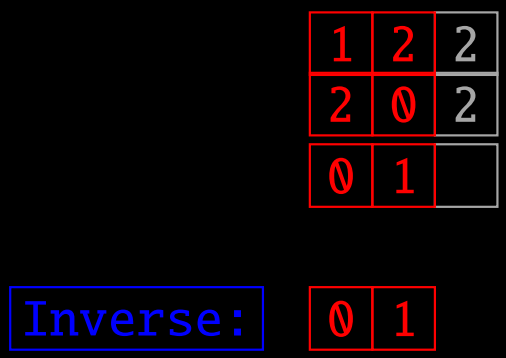
1	2
2	0
0	1

Inverse:

0	1

Memory-less Stream Manipulation

Generating Inverse of String s_1:



Memory-less Stream Manipulation

Memory-less Stream Manipulation

Generating Inverse of String s_1:

1	2	2	0	0	0	1	0	0	1	0	2	1	0	2	2	2	1	2	2	1	2	2	1	2	1
2	0	2	2	1	0	0	2	2	2	1	2	2	0	1	2	2	1	2	1	2					
0	1		1	2		2	1	1	0	2		0		0					0	0					

Inverse:

0	1	1	2	2	1	1	0	2	0	0	0	0

Summing s_2 with Inverse of s_1

2	2	1	1	2	2	1	1	0	2	2	2	0	1	0	0	2	1	2	2
0	1	1	2	2	1	1	0	2	0	0	0	0							

Result:

1	2	1	0