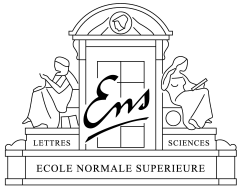
Sensory Memory
as
Scalable Statistics

Alain de Cheveigné

CNRS / Ecole Normale Supérieure / UCL



Scale,

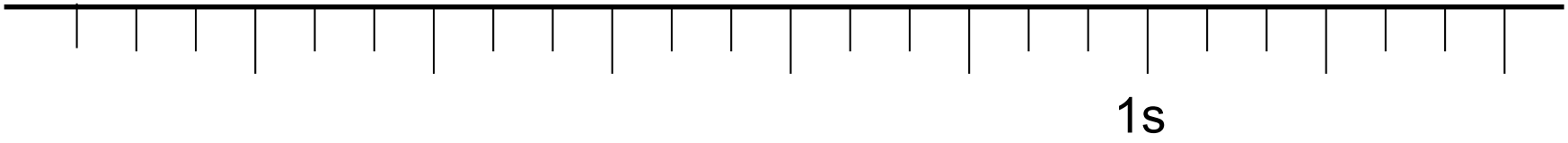
Time,

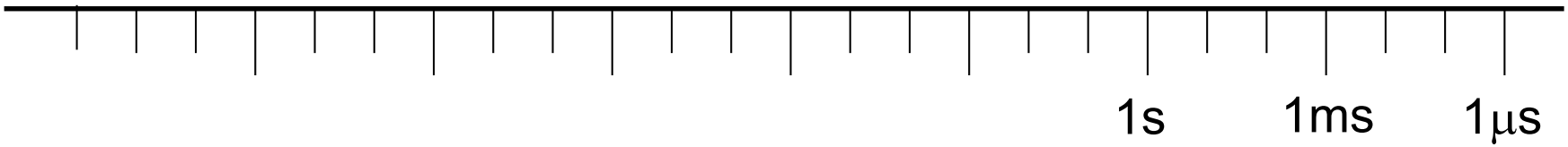
Memory,

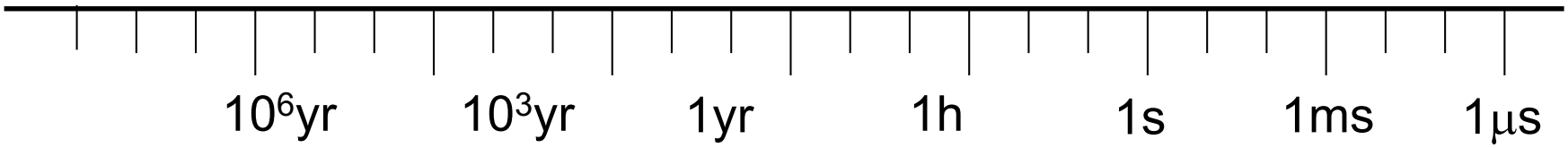
Modulation,

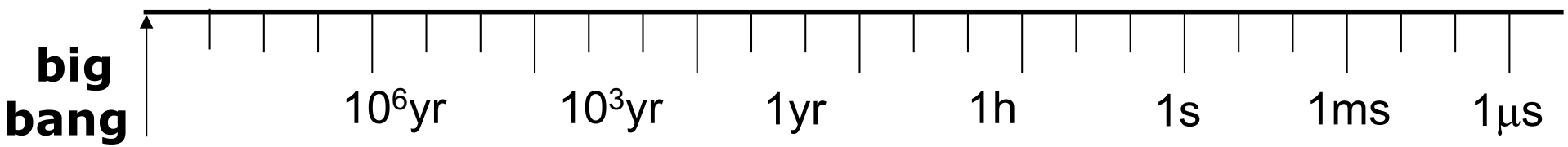
Scalable Statistics

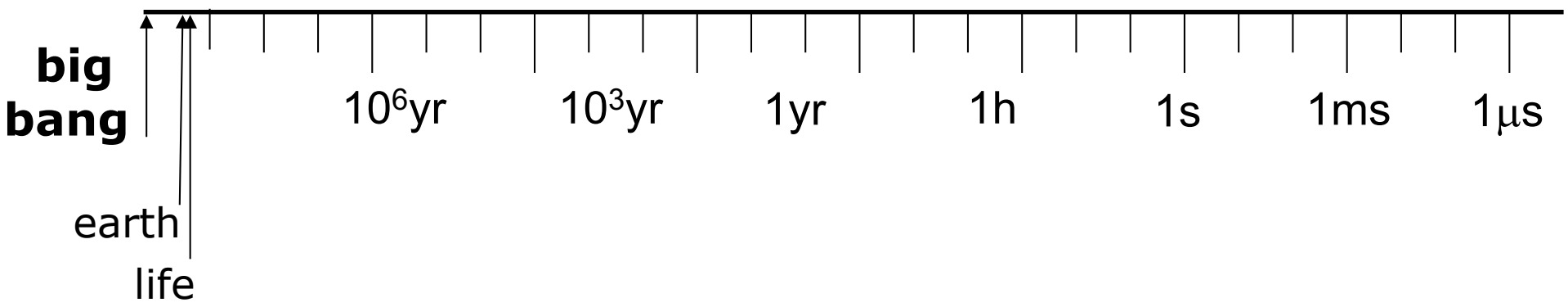
Scale

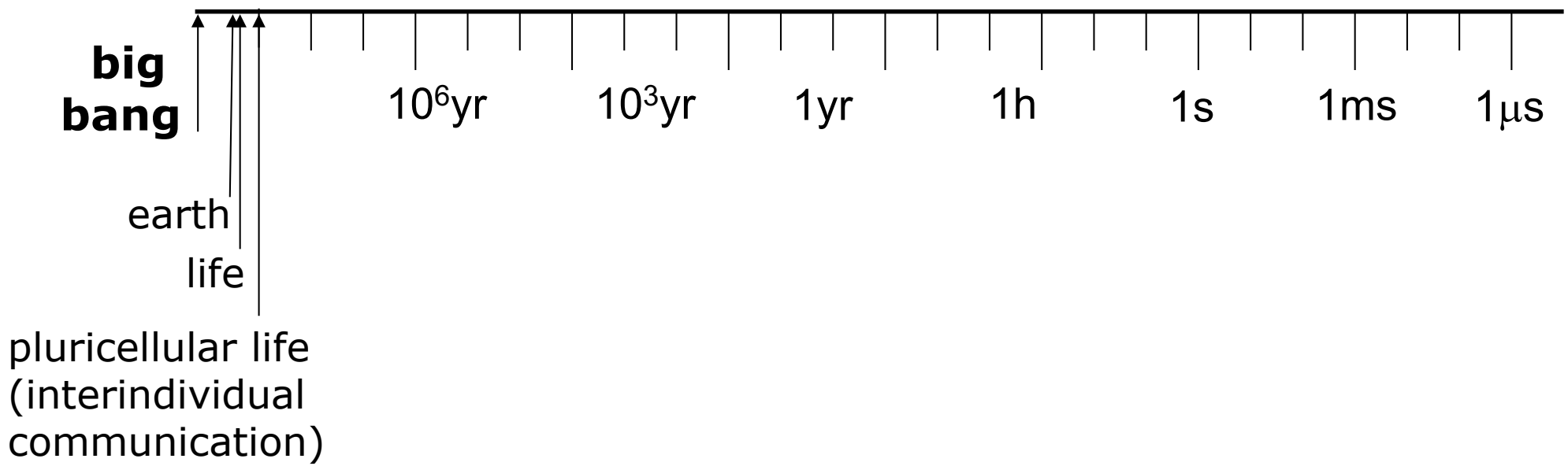

1s

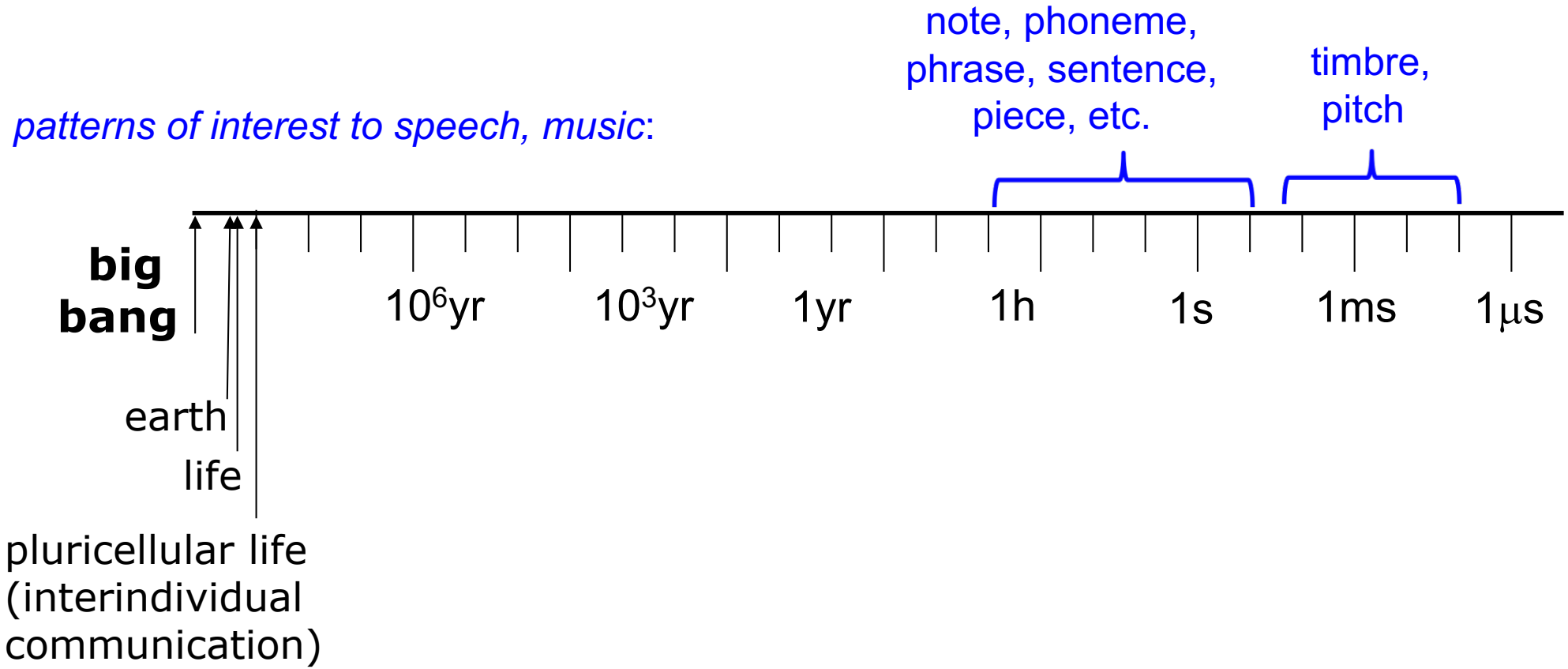

1s
1ms
1μs

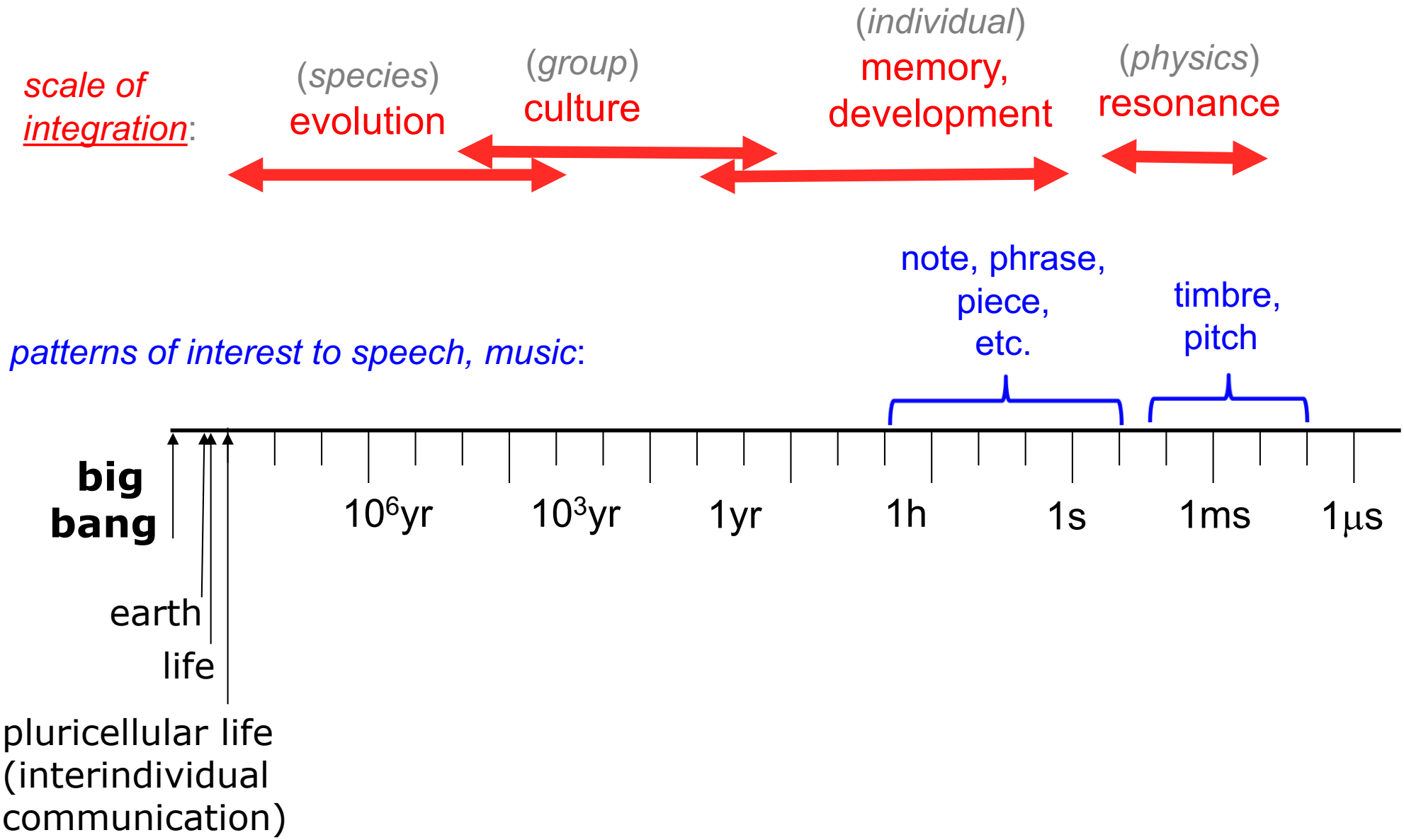

10^6yr
10^3yr
1yr
1h
1s
1ms
1μs


big
bang
10^6yr
10^3yr
1yr
1h
1s
1ms
1μs


big
bang
earth
life
10^6yr
10^3yr
1yr
1h
1s
1ms
1μs


big bang
earth
life
pluricellular life (interindividual communication)
10^6yr
10^3yr
1yr
1h
1s
1ms
1μs


note, phoneme,
phrase, sentence,
piece, etc.
timbre,
pitch
patterns of interest to speech, music:
big
bang
10^6yr
10^3yr
1yr
1h
1s
1ms
1μs
earth
life
pluricellular life
(interindividual
communication)


scale of integration:
(species)
evolution
(group)
culture
(individual)
memory, development
(physics)
resonance
patterns of interest to speech, music:
note, phrase, piece, etc.
timbre, pitch
big bang
10^6yr
10^3yr
1yr
1h
1s
1ms
1μs
earth
life
pluricellular life (interindividual communication)

integration window

$$R_t(\tau) = \int_t^{t+W} x(t')x(t'-\tau)dt'$$

short-term autocorrelation function

absolute time

span of pattern

integration window

$$S_t(\omega) = \int_t^{t+W} x(t')e^{j\omega t'}dt'$$

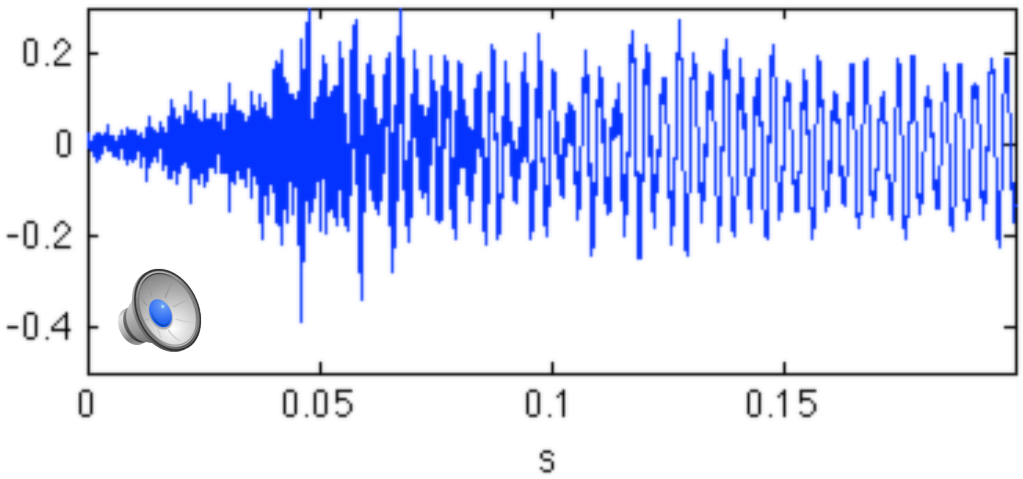
short-term Fourier transform

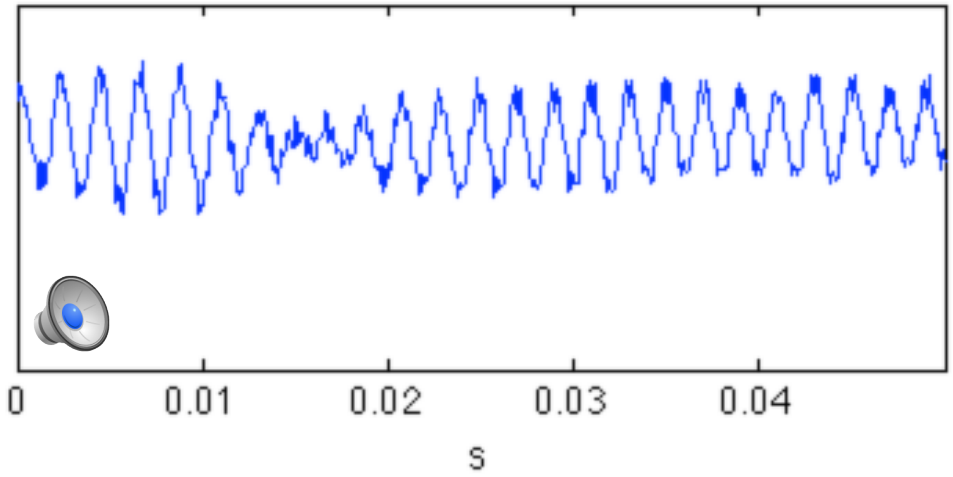
absolute time

span of pattern

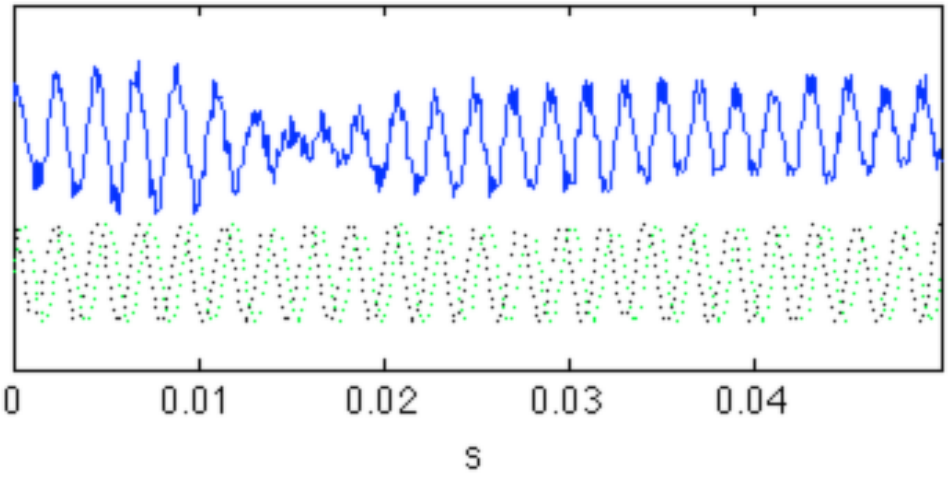
the integration window is usually *larger* than the span of the pattern

Time



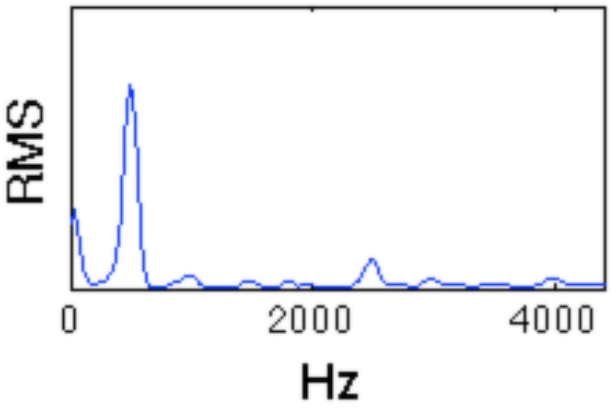


sound is change

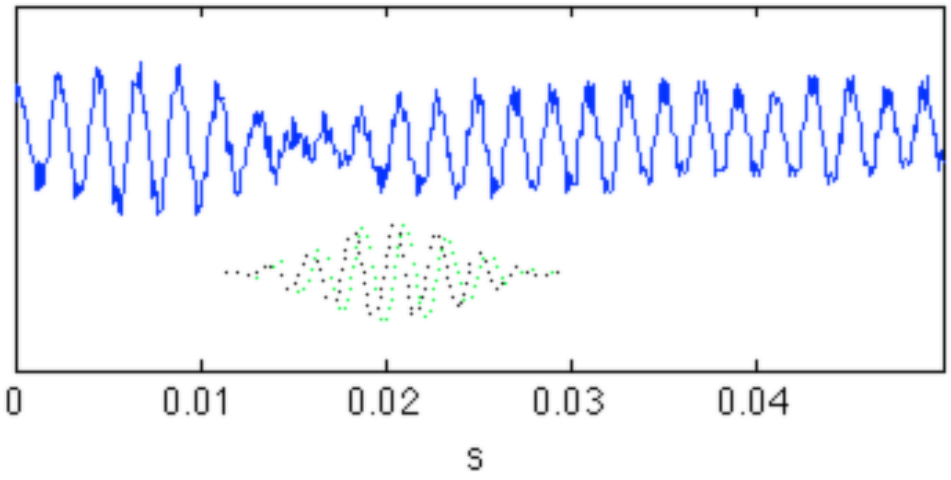


sinusoids = "yardstick"

$$s(\omega) = \int_{-\infty}^{+\infty} x(t')e^{j\omega t'}dt'$$



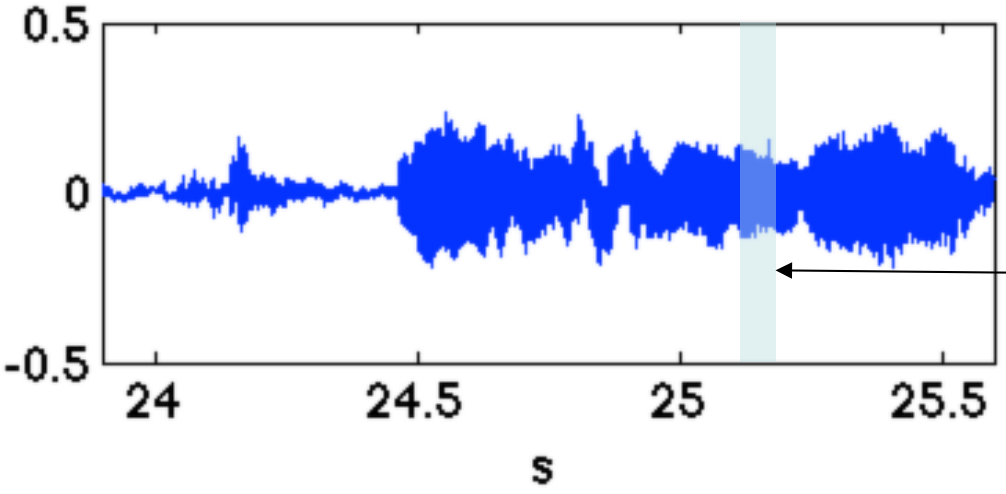
Fourier transform



window

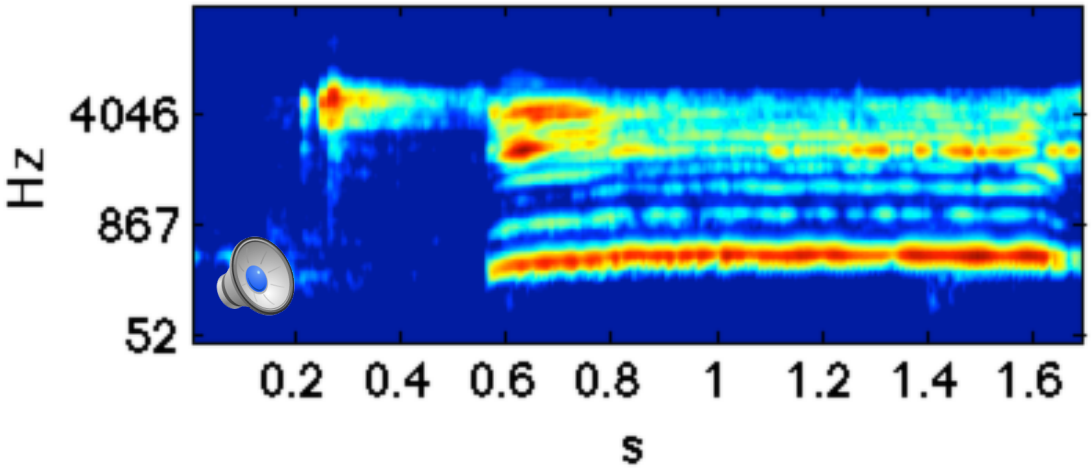
$$s(\omega) = \int_{-\infty}^{+\infty} x(t')W_t e^{j\omega t'} dt'$$

Short-Term Fourier transform



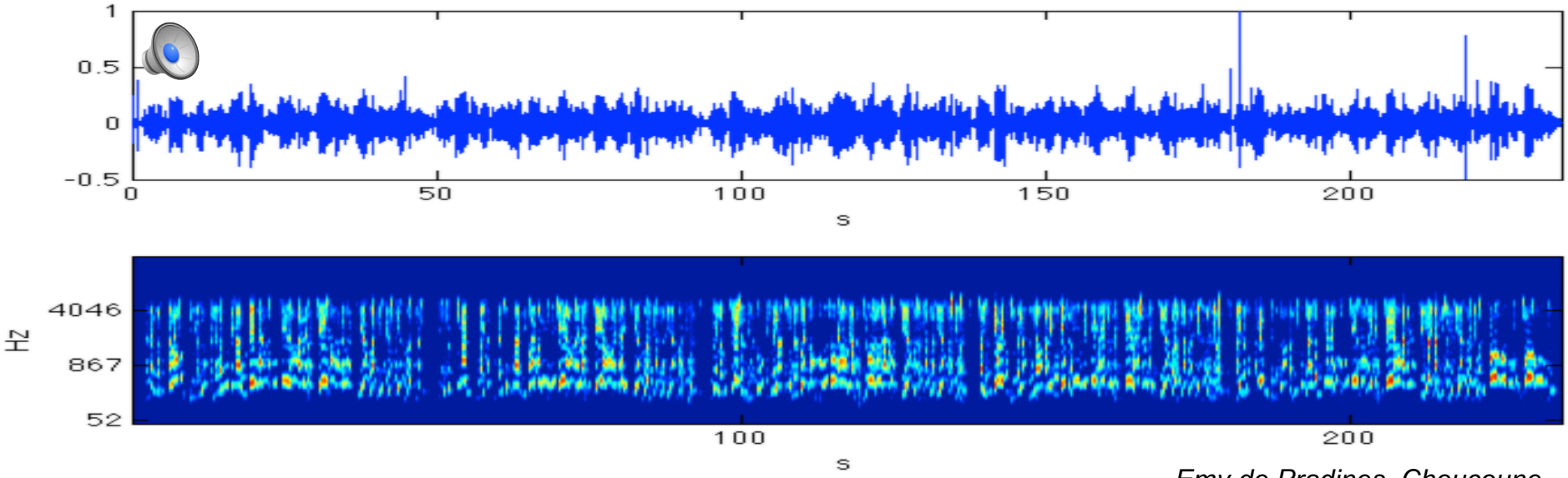
window moves in time

$$s(\omega) = \int_{-\infty}^{+\infty} x(t')W_t e^{j\omega t'} dt'$$

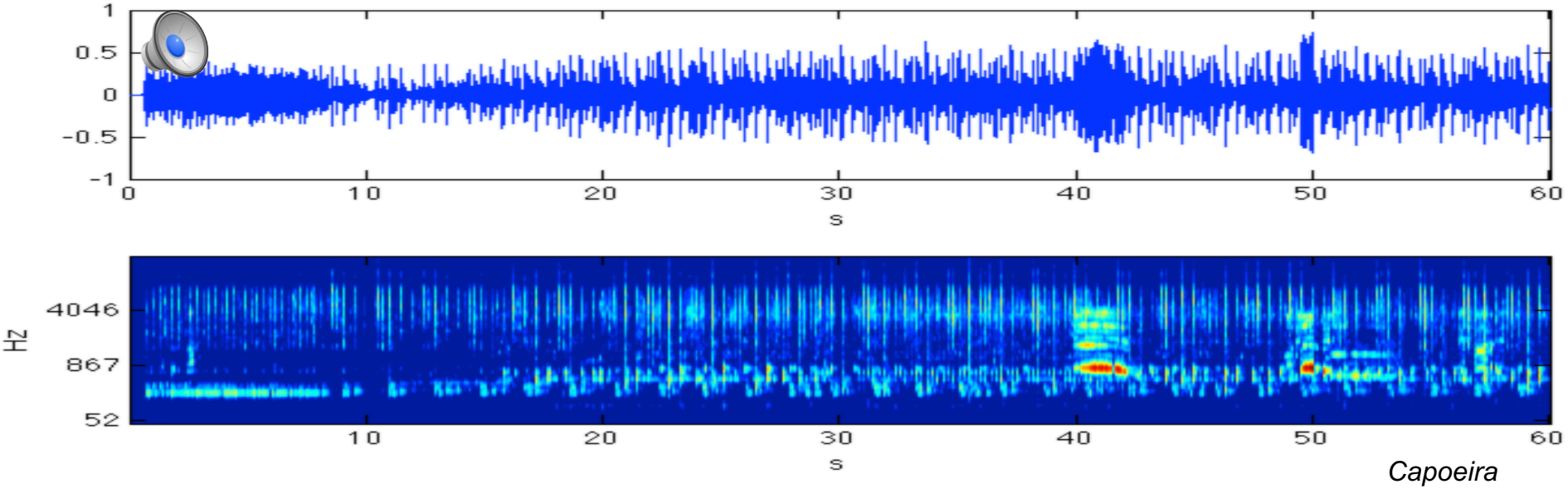


spectrogram

structure over longer time scales is not well captured by spectral analysis



Emy de Pradines, Choucoune



Capoeira

Memory

several facets:

(1) *what is there to remember?*

(2) *what should we remember?*

(3) *what do we remember?*

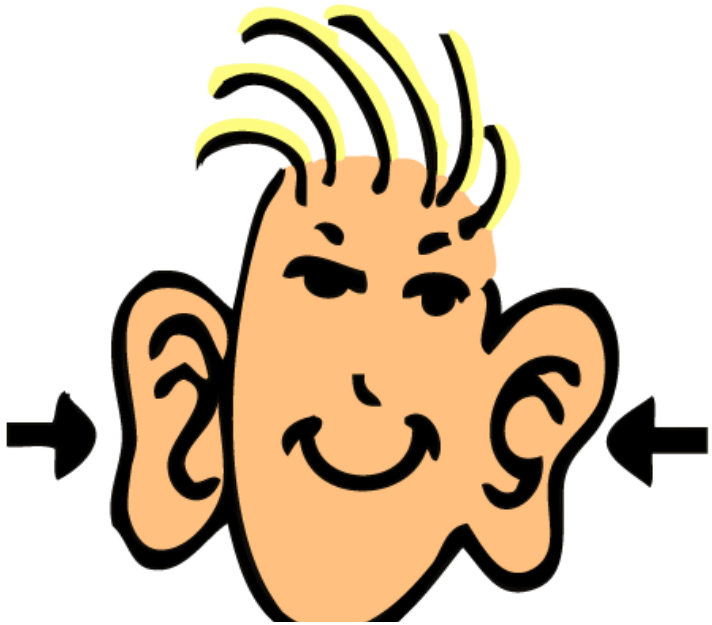
(4) *how do we remember it?*

What is there to remember?

potentially a lot...
- 32 to 320 kbits/s (MP3 bit rate) → 20 tbytes over 50 years

- for auditory system estimates vary widely, from 150 kbits/s in cochlear nerve (9000 kbits/s in optic nerve), to ~50 bits/s information rate of speech

actual rate depends on:
- bandwidth
- dynamic range
- redundancy



Similar situation faces science and engineering:

- LHC sensors produce ~20000 exabytes/year of raw data (2 x 10^22 bytes, of which 0.3 ppm are actually stored)
- NCCS stores 32 petabytes of climate data (Wikipedia)
- Level 3+ (internet & telecom provider) carries ~10 exabytes/year
- Google's network carries ~ 4 exabytes/year
- Facebook...
- Twitter...
- etc.

exponential trend is predicted by
Moore's law, Parkinson's law, Keck's law, Nielsen's law, Carlson's law, etc.

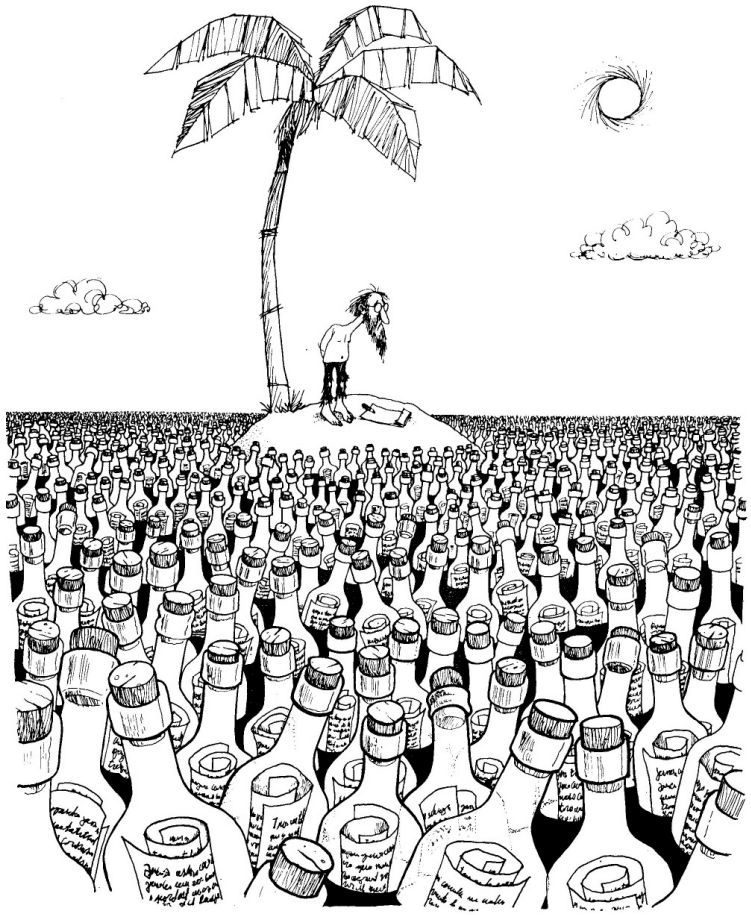
What *should* we remember?

naive answer:

- everything! (might come in handy, some day...)

less naive:

- costly to store → *need to compress*
- costly to search → *need to structure*
- details not useful... → *need to abstract*



Jorge Luis Borges

FUNES THE MEMORIOUS

Funes remembered not only every leaf of every tree of every wood, but also every one of the times he had perceived or imagined it.

Two or three times he had reconstructed a whole day; he never hesitated, but each reconstruction had required a whole day

it

bothered him that the dog at three fourteen (seen from the side) should have the same name as the dog at three fifteen (seen from the front)

To think is to forget differences, generalize, make abstractions. In the teeming world of Funes, there were only details, almost immediate in their presence.

Predictability, Complexity, and Learning

William Bialek
NEC Research Institute, Princeton, NJ 08540, U.S.A.

Ilya Nemenman
NEC Research Institute, Princeton, New Jersey 08540, U.S.A., and Department of Physics, Princeton University, Princeton, NJ 08544, U.S.A.

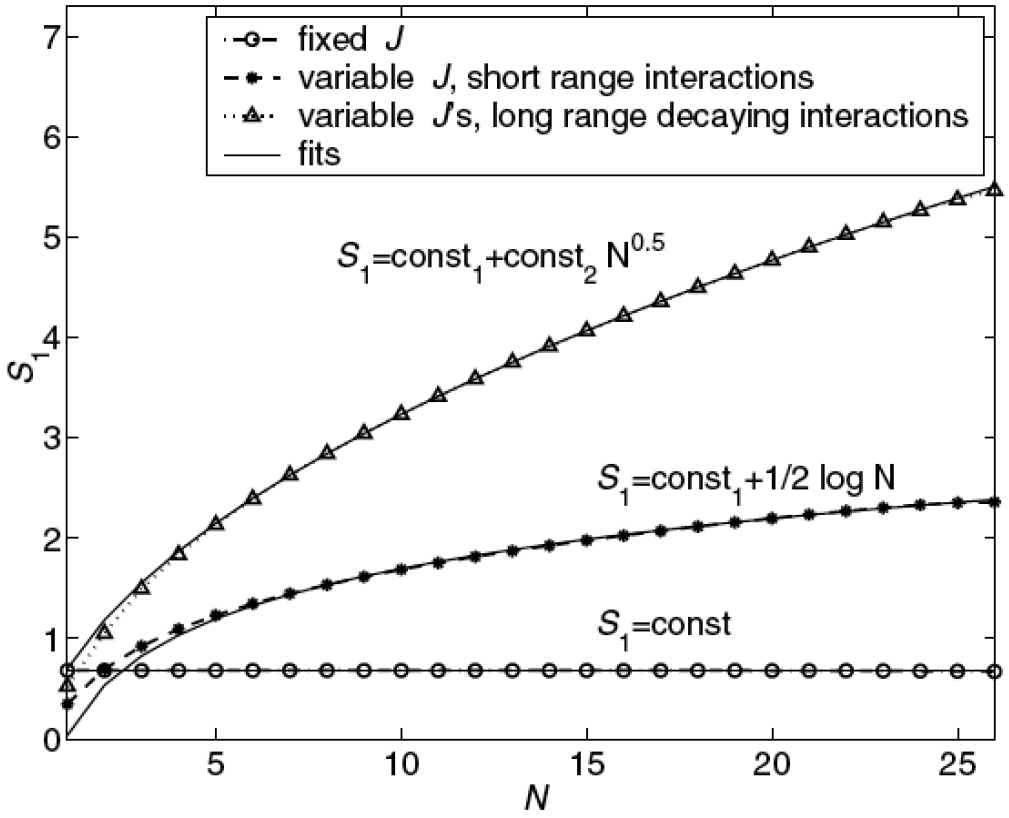
Naftali Tishby
NEC Research Institute, Princeton, NJ 08540, U.S.A., and School of Computer Science and Engineering and Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel

We define *predictive information* $I_{\rm pred}(T)$ as the mutual information between the past and the future of a time series.

Put bluntly, *nonpredictive information is useless to the organism*

. It will turn out that most of the information we collect over a long period of time is nonpredictive, so that isolating the predictive information must go a long way toward separating out those features of the sensory world that are relevant for behavior.

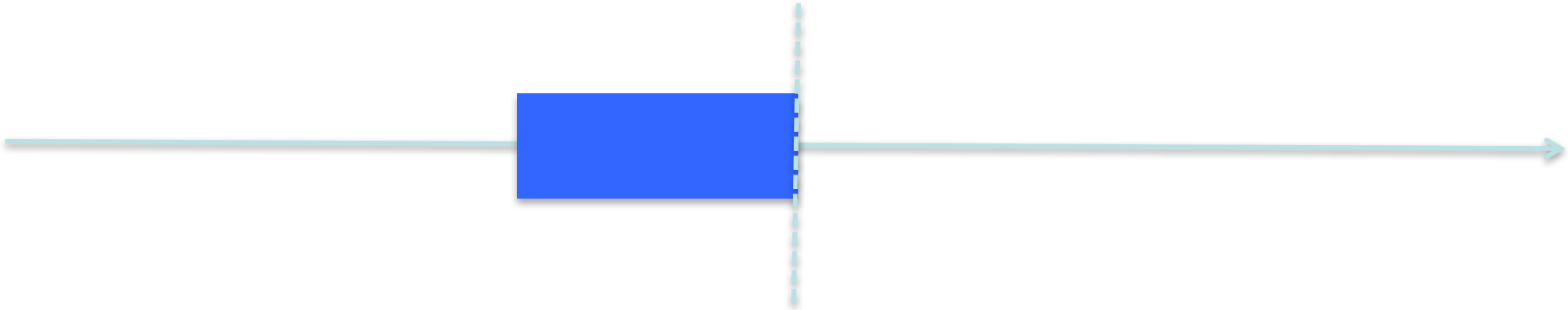
Neural Computation **13, 2409–2463** (2001)



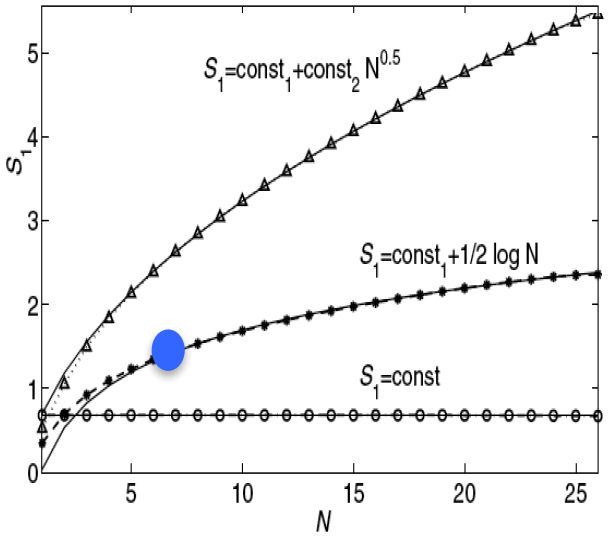
growth of predictive information with observation interval size is:

- *logarithmic* if finite parameters
- *power law* if infinite parameters
- in any case *less than linear*

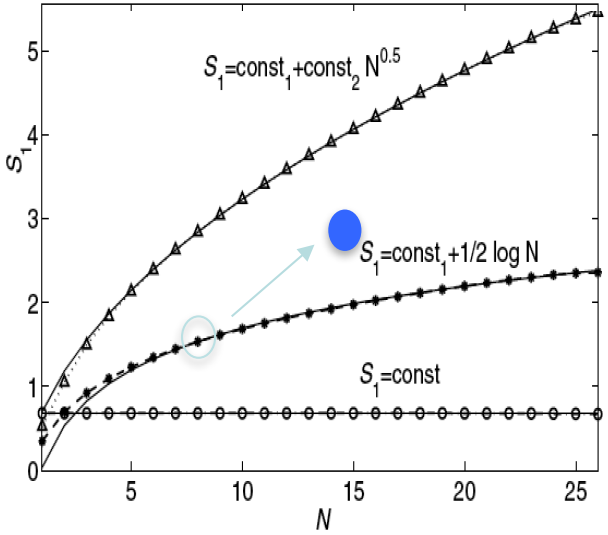
==> *no need to keep it all!*



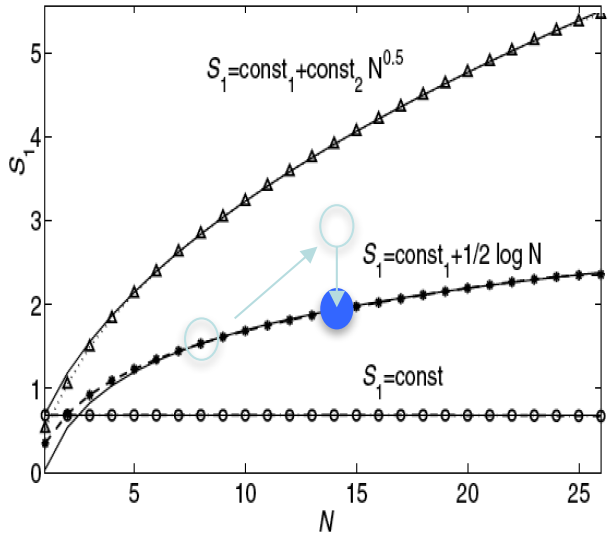
now


S_1=const$_1$+const$_2$ N$^{0.5}$
S_1=const$_1$+1/2 log N
S_1=const
S_1
N
0
1
2
3
4
5
5
10
15
20
25

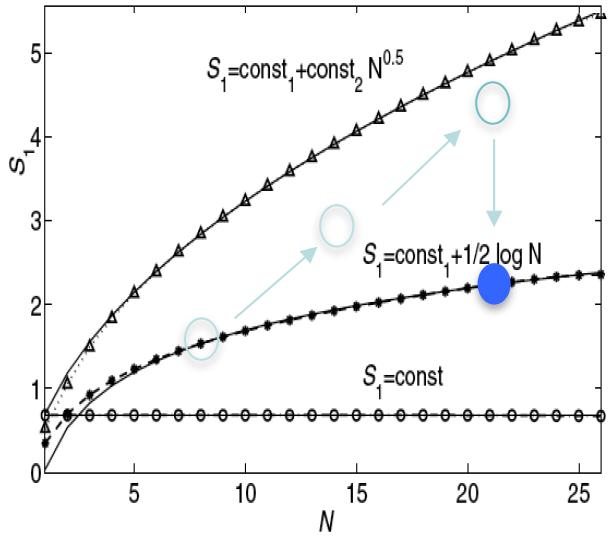
now

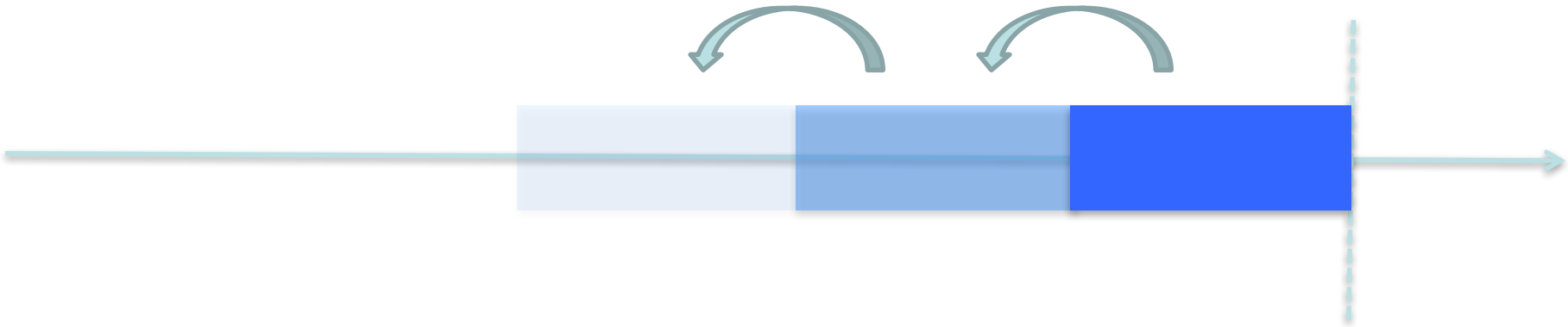

S_1=const$_1$+const$_2$ N$^{0.5}$
S_1=const$_1$+1/2 log N
S_1=const
S_1
N
0
1
2
3
4
5
5
10
15
20
25


now

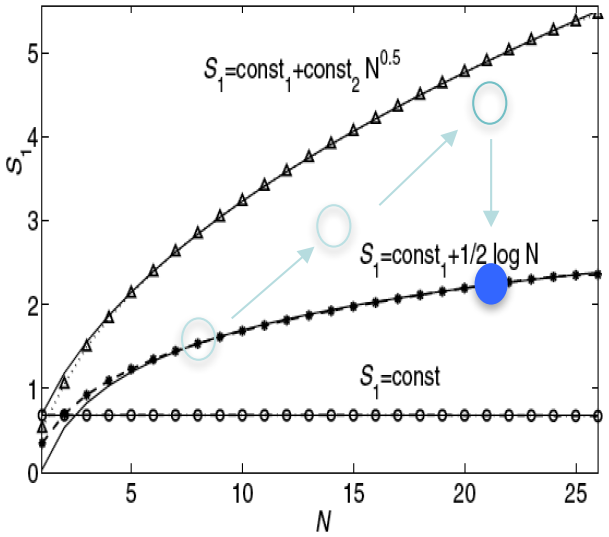

S_1=const$_1$+const$_2$ N$^{0.5}$
S_1=const$_1$+1/2 log N
S_1=const
S_1
N
0
1
2
3
4
5
5
10
15
20
25


now


S_1=const$_1$+const$_2$ N$^{0.5}$
S_1=const$_1$+1/2 log N
S_1=const
S_1
N
0
1
2
3
4
5
5
10
15
20
25



now



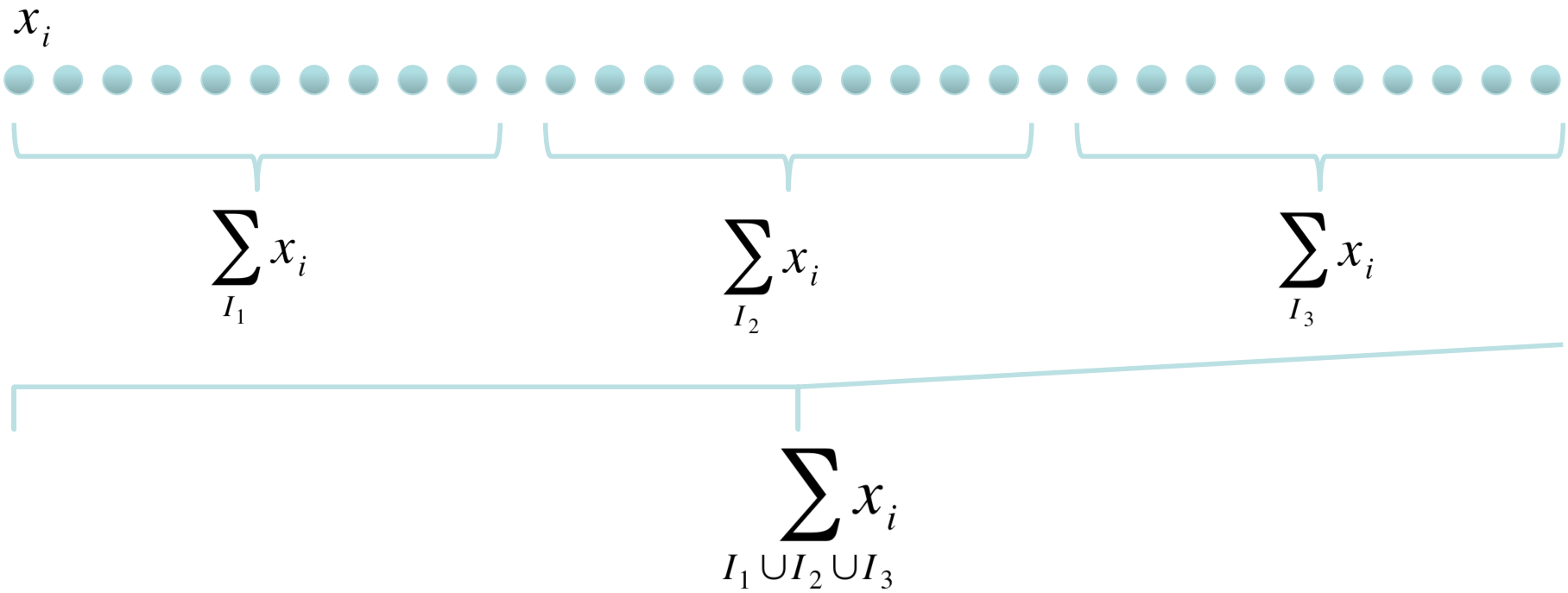
memory must fade

i.e. existing memory traces must be recoded to a coarser representation

Scalable statistics

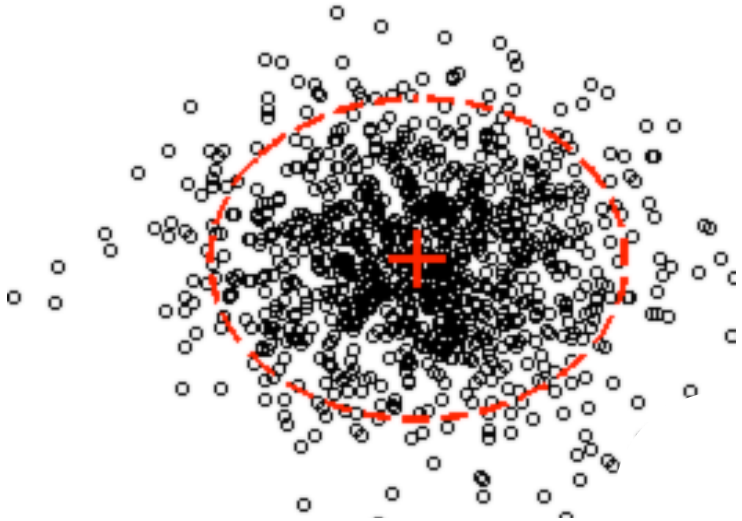
Scalable statistics are *summary statistics* that:

- can be instantiated at any scale
- can be rescaled from fine resolution to coarse.



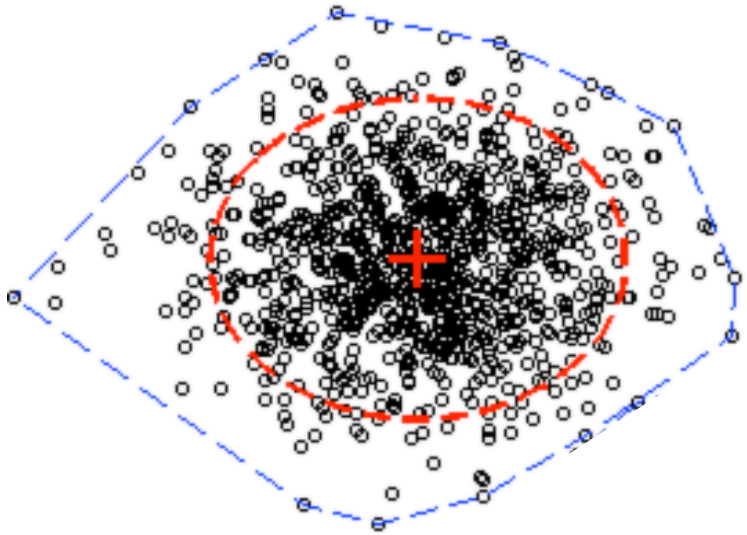
examples: *mean, variance, extrema, histogram, cardinality*, etc

1000 points



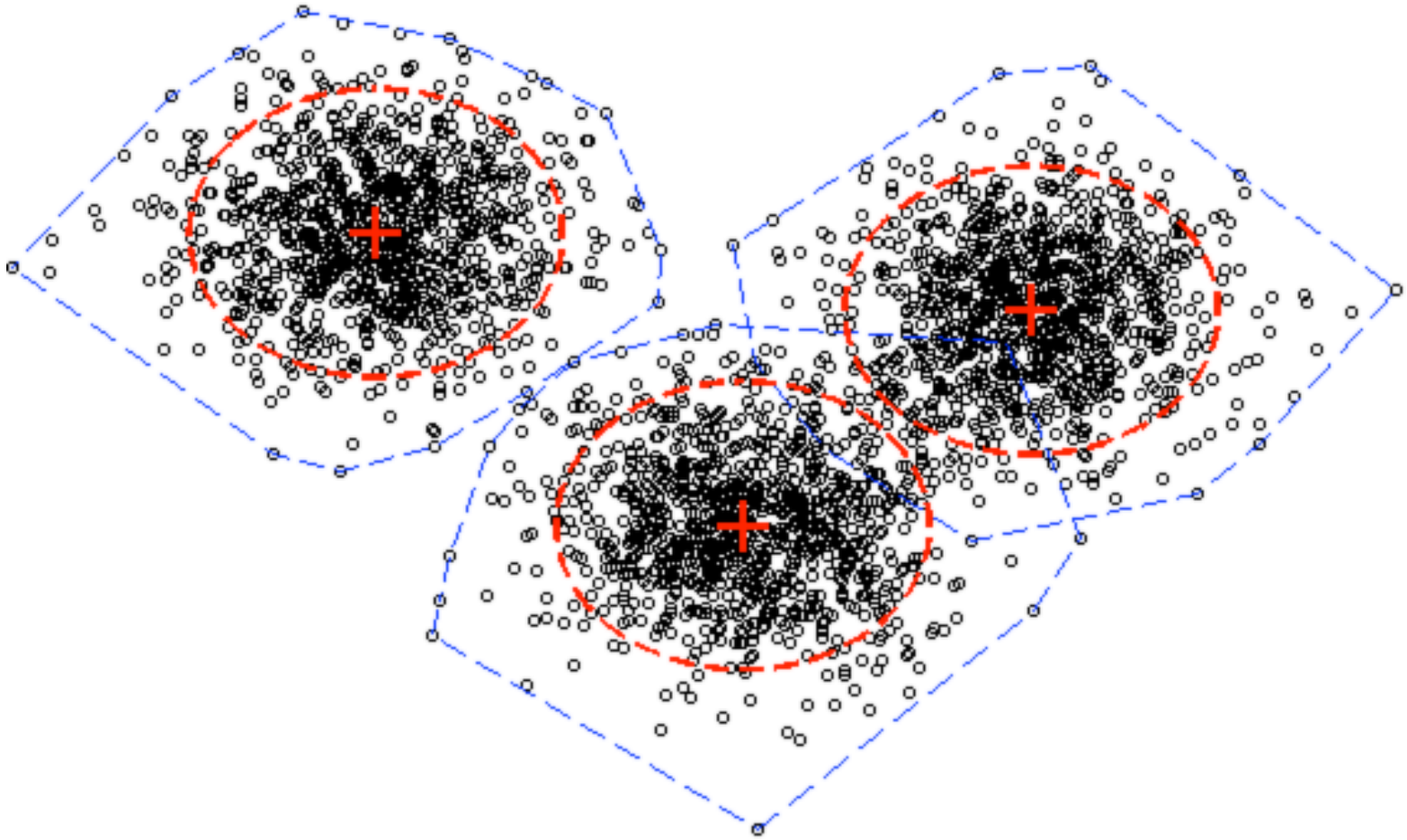
→ mean (2 numbers), covariance (3 numbers)

1000 points



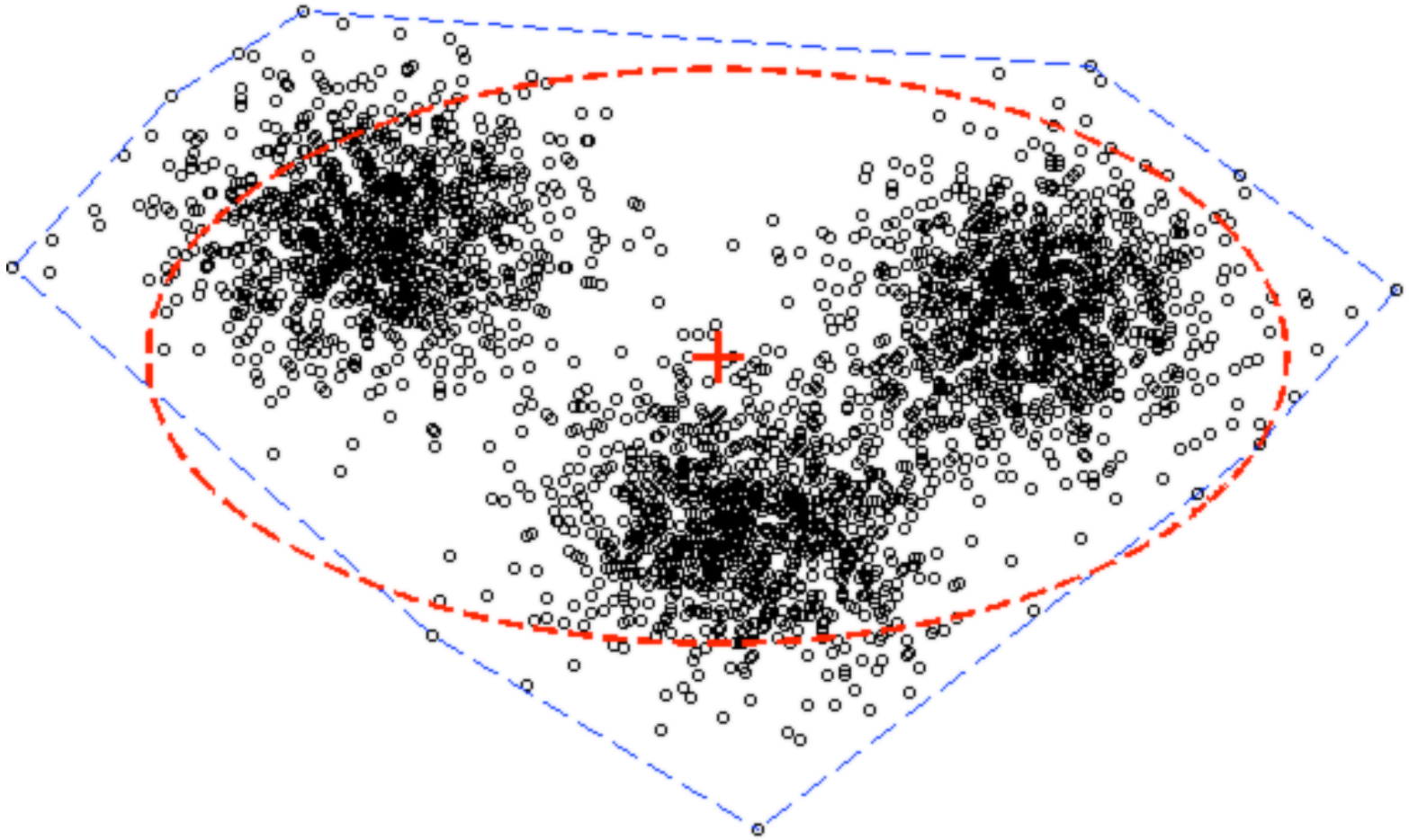
→ mean (2 numbers), covariance (3 numbers), convex hull (~20 numbers)

3000 points



→ mean (6 numbers), covariance (9 numbers), convex hull (~60 numbers)

3000 points



→ mean (6 numbers), covariance (9 numbers), convex hull (~60 numbers)

Some scalable statistics:

$n_I = |I|$ cardinality

$s_I = \sum_{i \in I} x_i$ sum

$m_I = s_I / n_I$ mean

$\wedge_I = \max_{i \in I} x_i$ max (min)

$v_I = (\frac{1}{n_I}) \sum_{i \in I} (x_i - m_I)^2$ variance

Some scalable statistics:

$n_I = |I|$ cardinality (histogram)

$s_I = \sum_{i \in I} x_i$ sum

$m_I = s_I / n_I$ mean

$\wedge_I = \max_{i \in I} x_i$ max (min) (convex hull)

$v_I = (\frac{1}{n_I}) \sum_{i \in I} (x_i - m_I)^2$ variance (covariance)

$s_I = \left(\frac{1}{w_I}\right) \sum_{i \in I} w_i x_i$ weighted statistics

$w_I = (\frac{1}{n_I}) \sum_{i \in I} w_i$ weight

rescaling formulae:

$\left(J = \bigcup I\right)$ $\quad n_J = \sum_I n_I$ $\quad s_J = \sum_I s_I$ $\quad m_J = (\frac{1}{n_J}) \sum_I n_I m_I$ $\quad \wedge_J = \max_I \wedge_I$ etc.

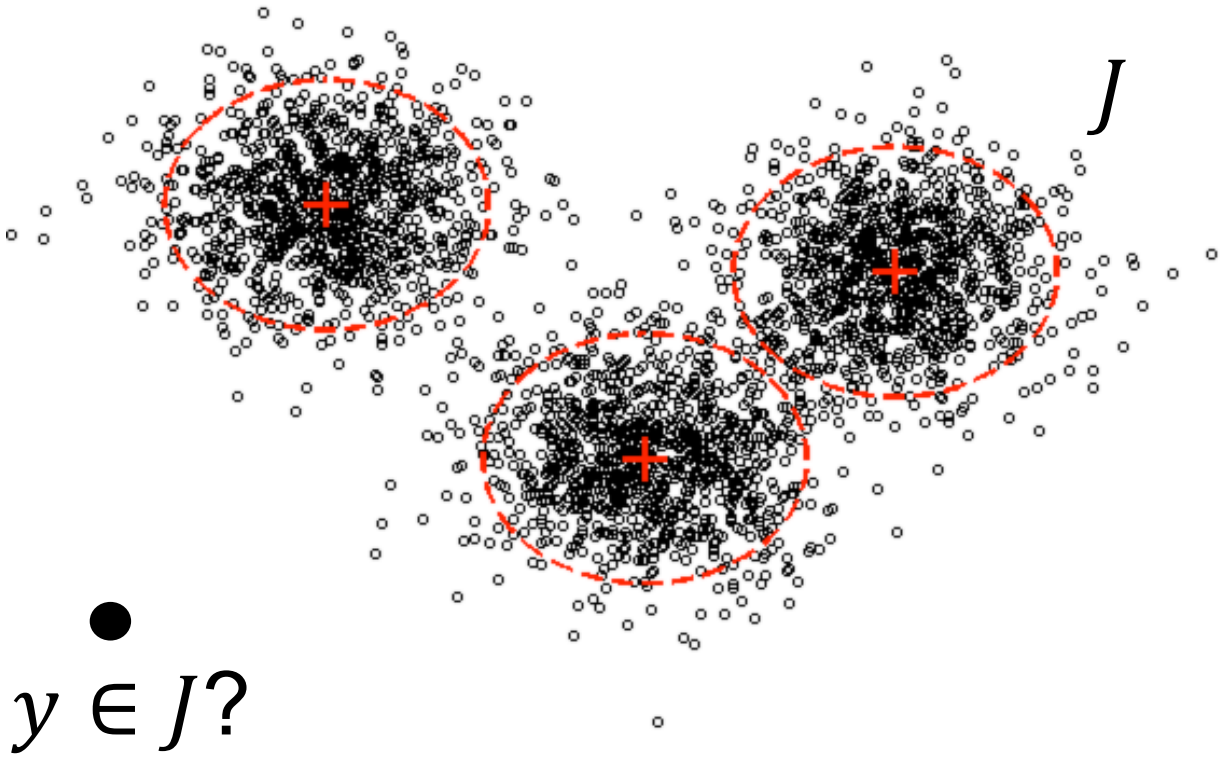
cardinality sum mean max

Non-scalable:

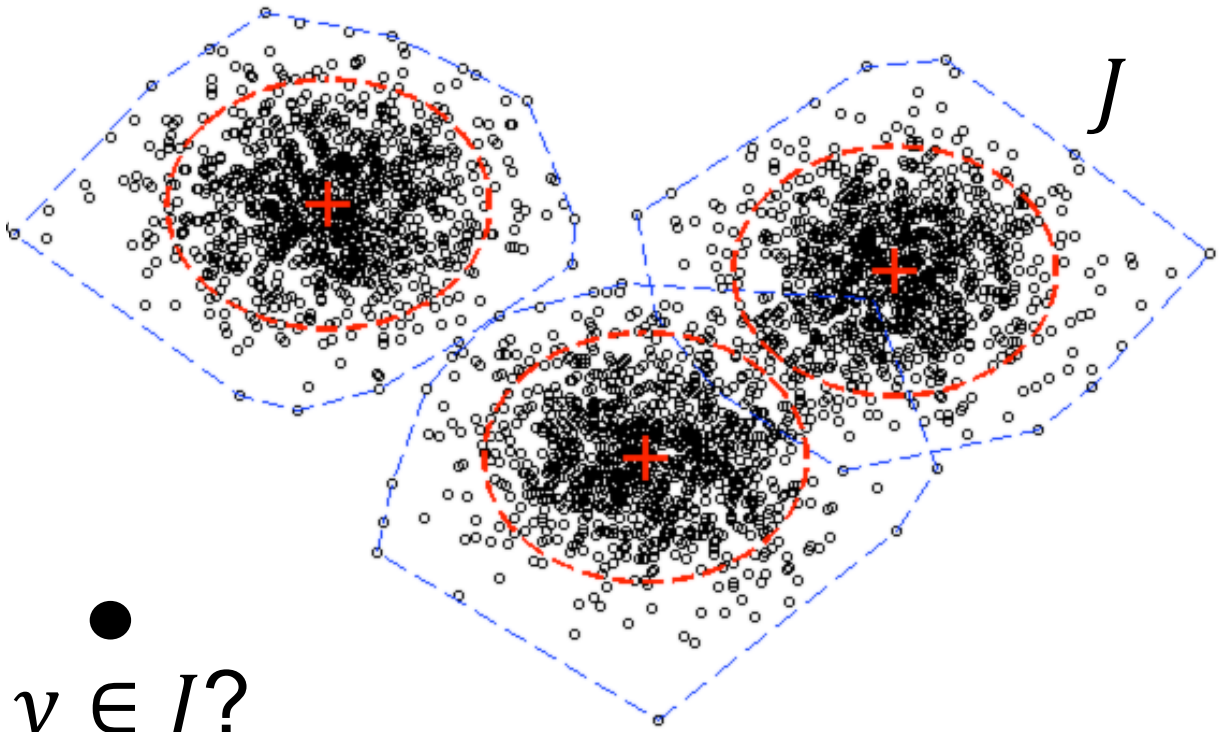
median, quartile, mean (in the absence of cardinality), etc.

What is this useful for?

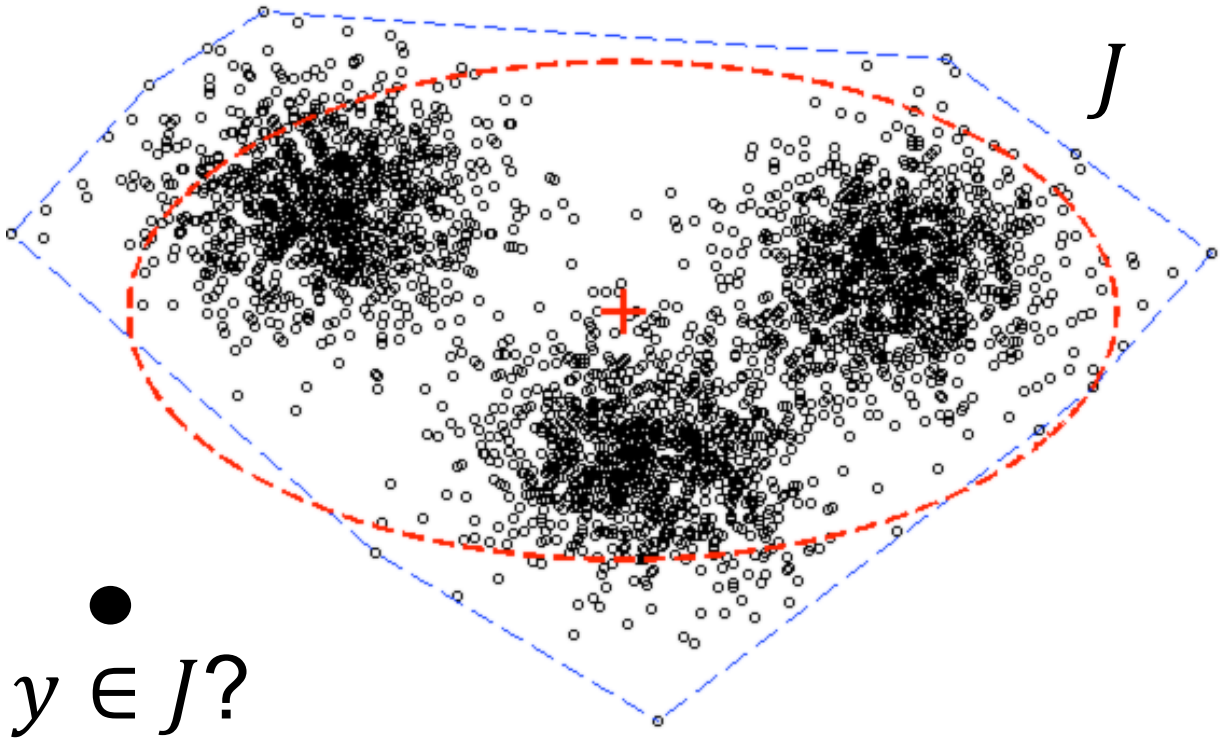
Search



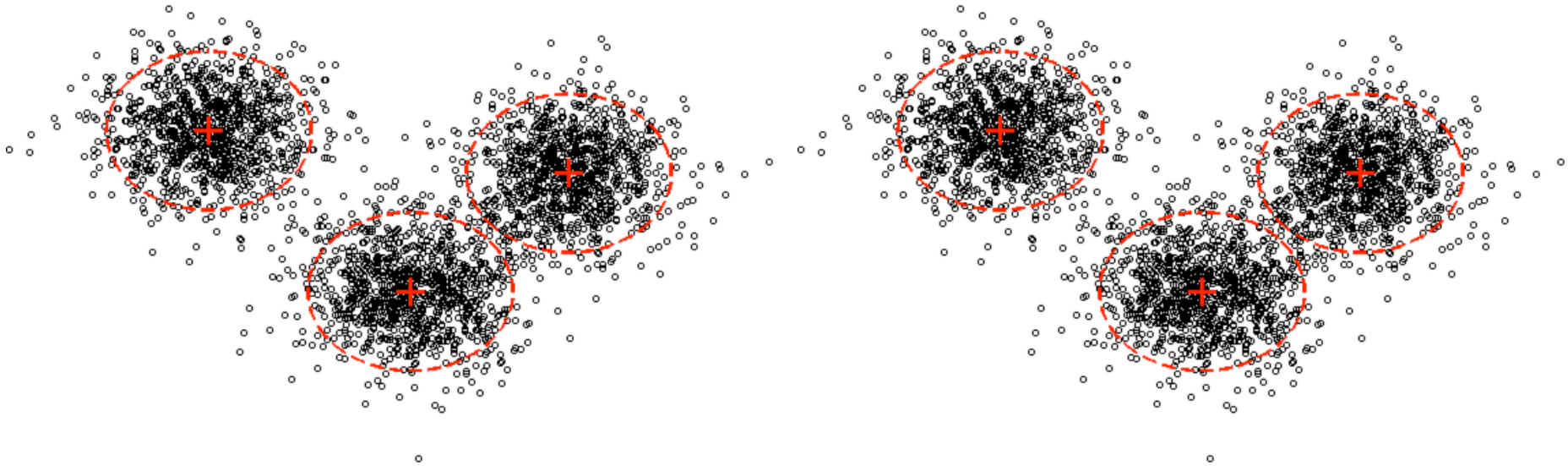
Search



Search

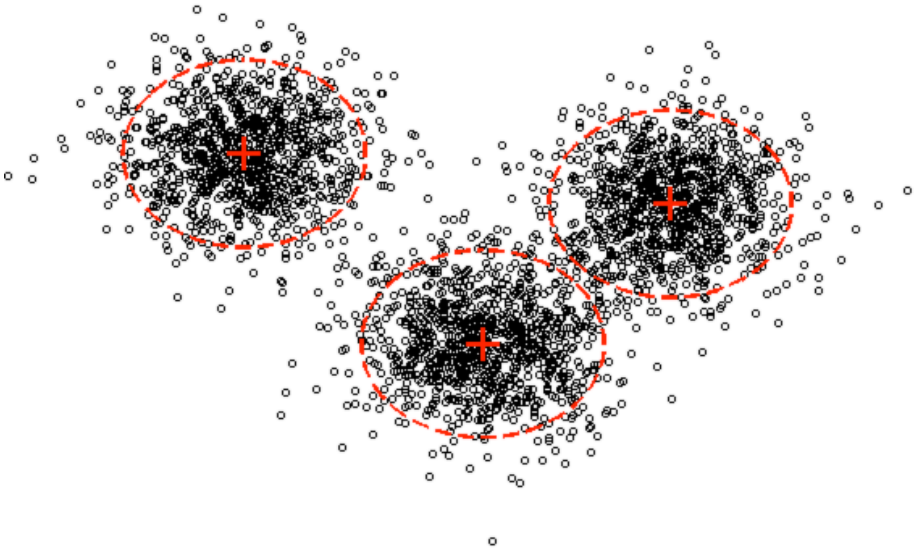


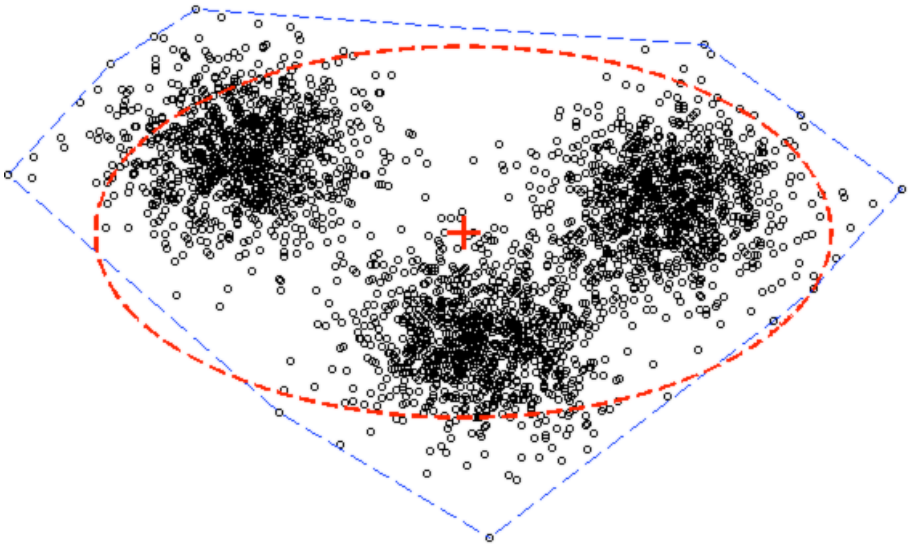
Comparison, discrimination

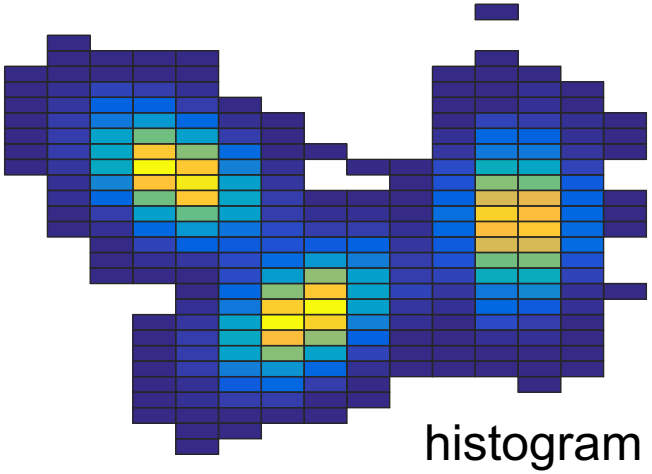


- do these *data* differ?
- do these *distributions* differ?
- do these *patterns* belong to a different class?

comparison / discrimination

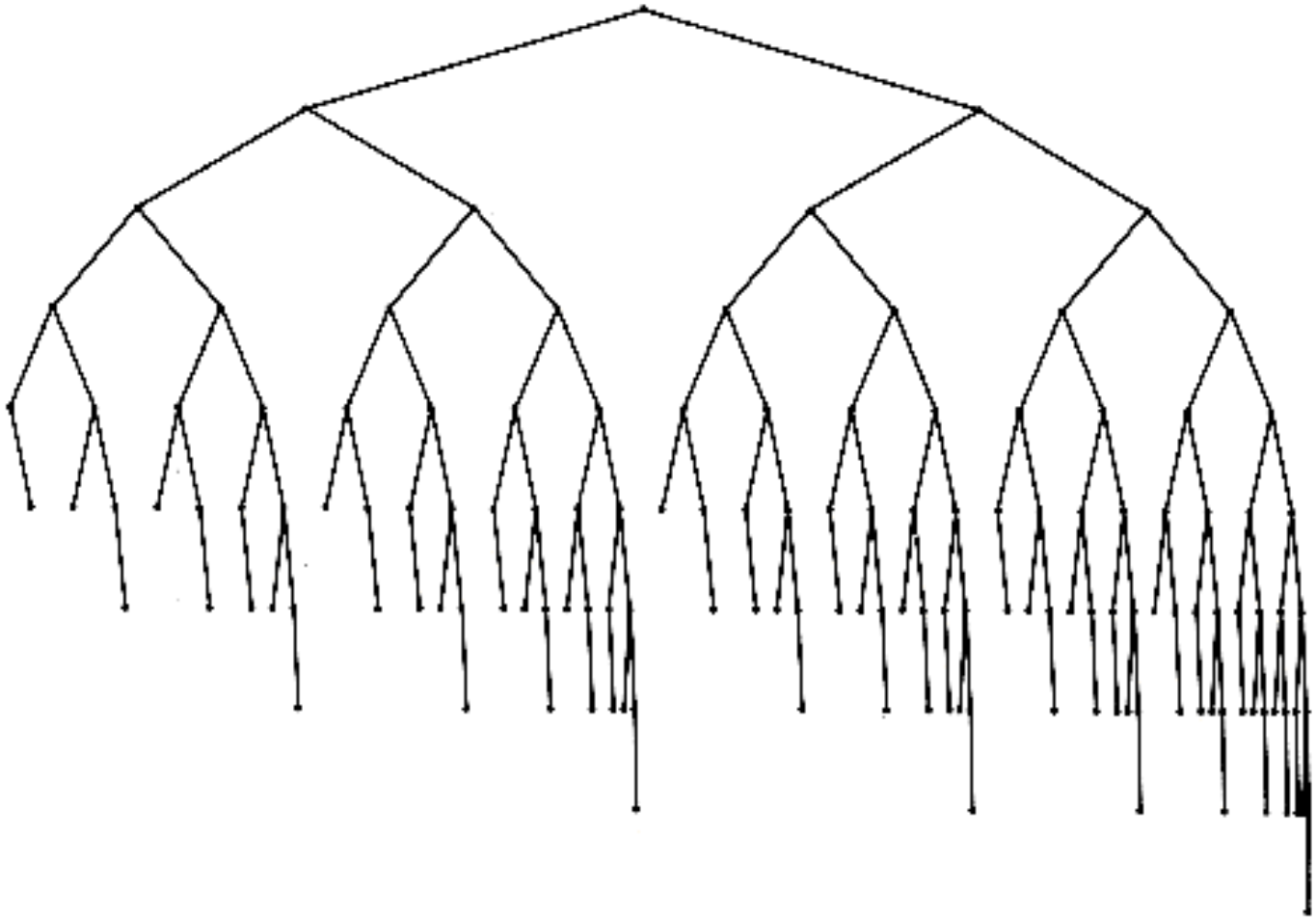


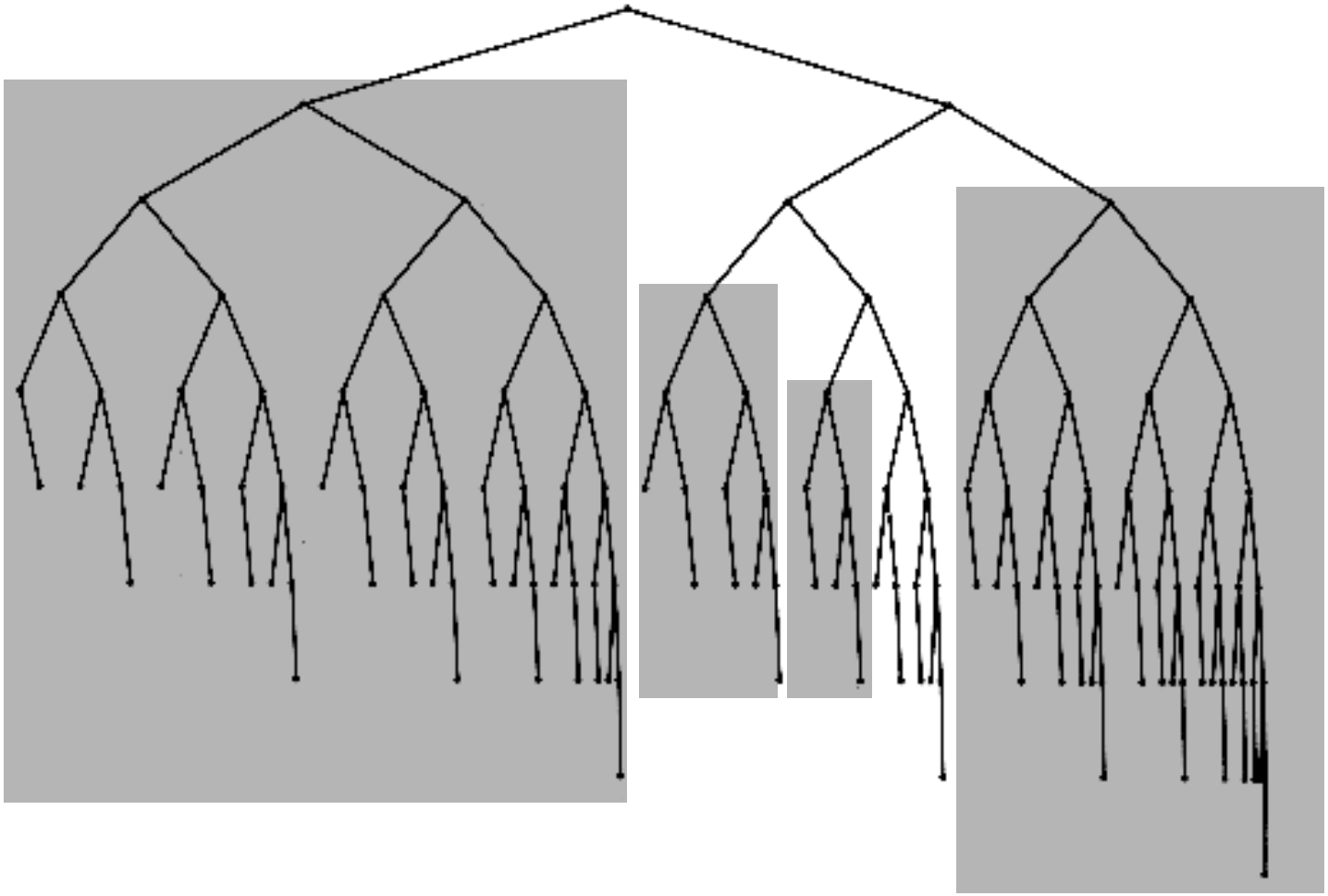




histogram

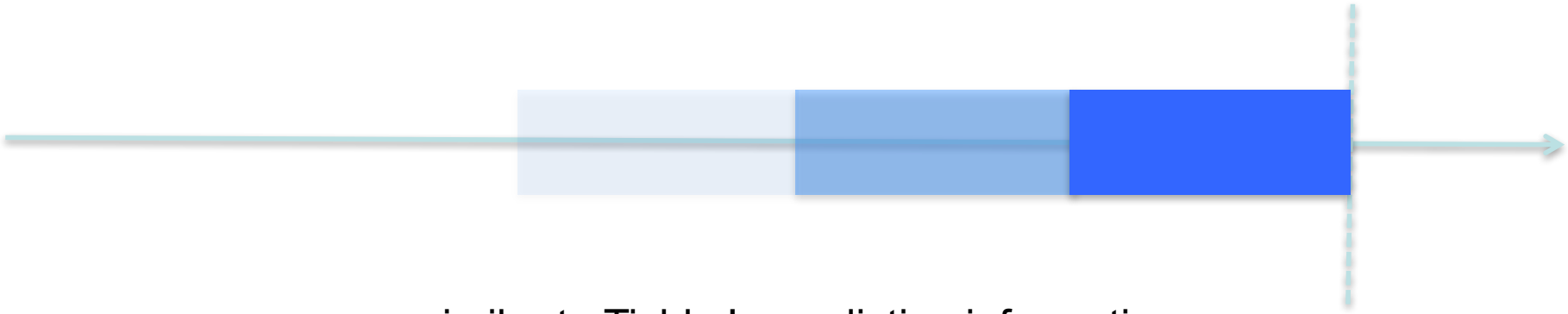
- the statistics don't need to be the same
- distributions can be compared e.g. using Kullback-Leibler divergence
- statistics allow pruning (→ fast search)





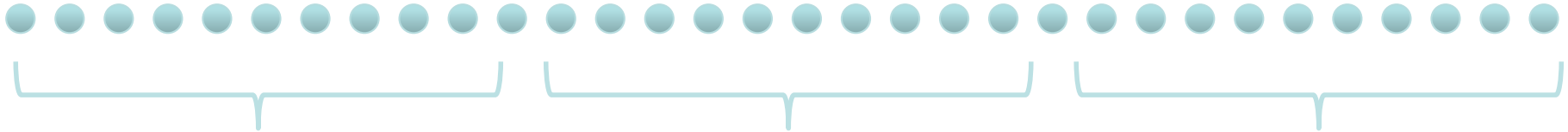
pruning → fast search

- The representation can be arbitrarily compact.
- An existing representation can always be made *more* compact, and still retain useful information
- Rescaling throws away information (hopefully the least useful)



How does this work for a time series like sound?

time series of data:



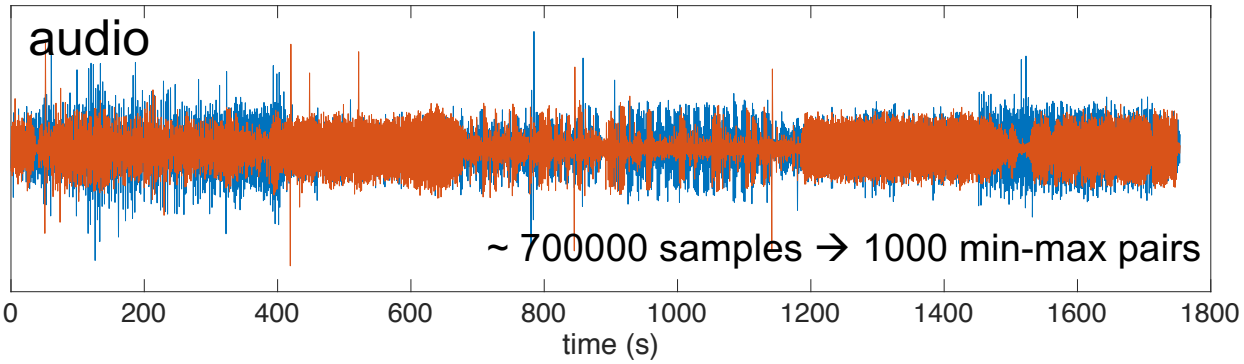
cardinality	n_1	n_2	n_3
mean	m_1	m_2	m_3
variance	v_1	v_2	v_3
etc.	⋮	⋮	⋮

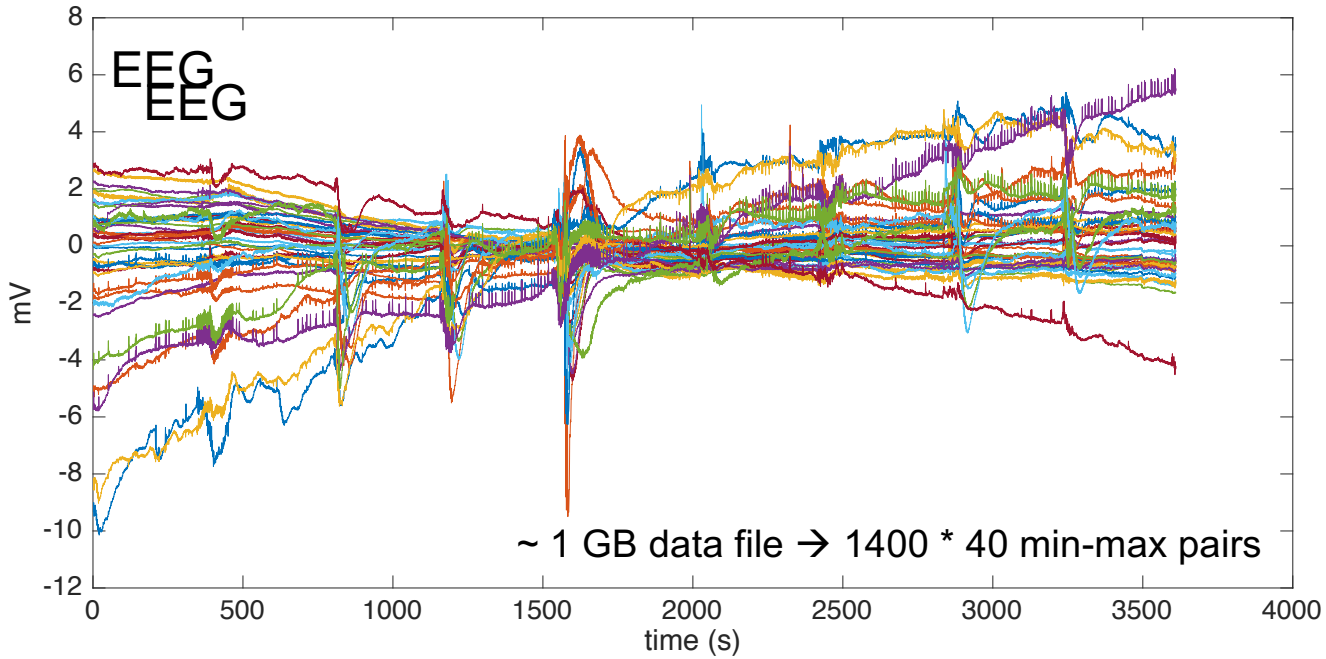
→ *time series of statistics*

other useful statistics:

- min & max,
- autocorrelation (or power spectrum),
- covariance (multichannel),
- histogram, etc.

simple application: big data display

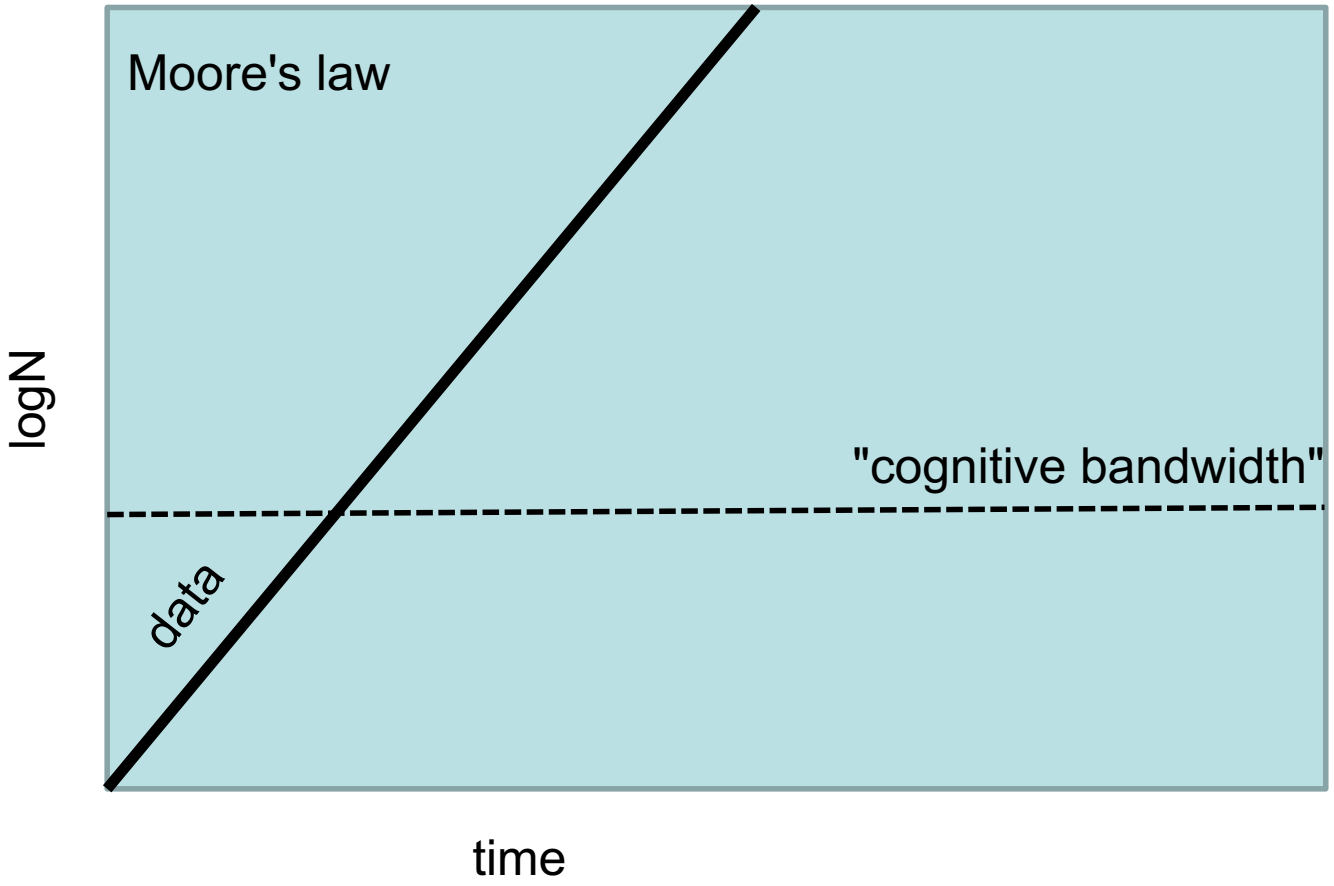




- visually indistinguishable from full data
- works the same whatever the size (e.g. lifetime audio recording)

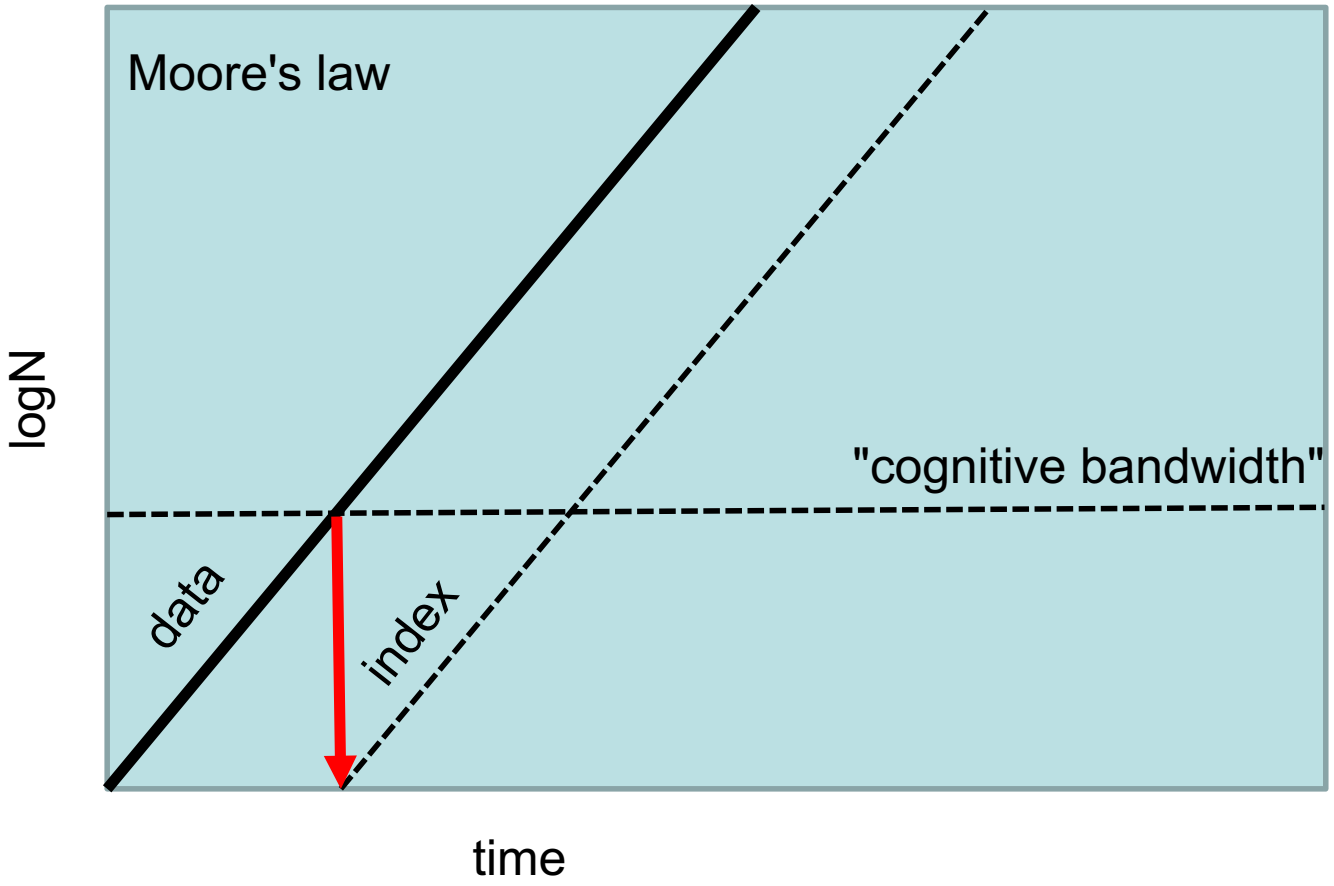
more generally: audio/multimedia/data indexing & classification

search by content, find duplicates, cluster, etc.



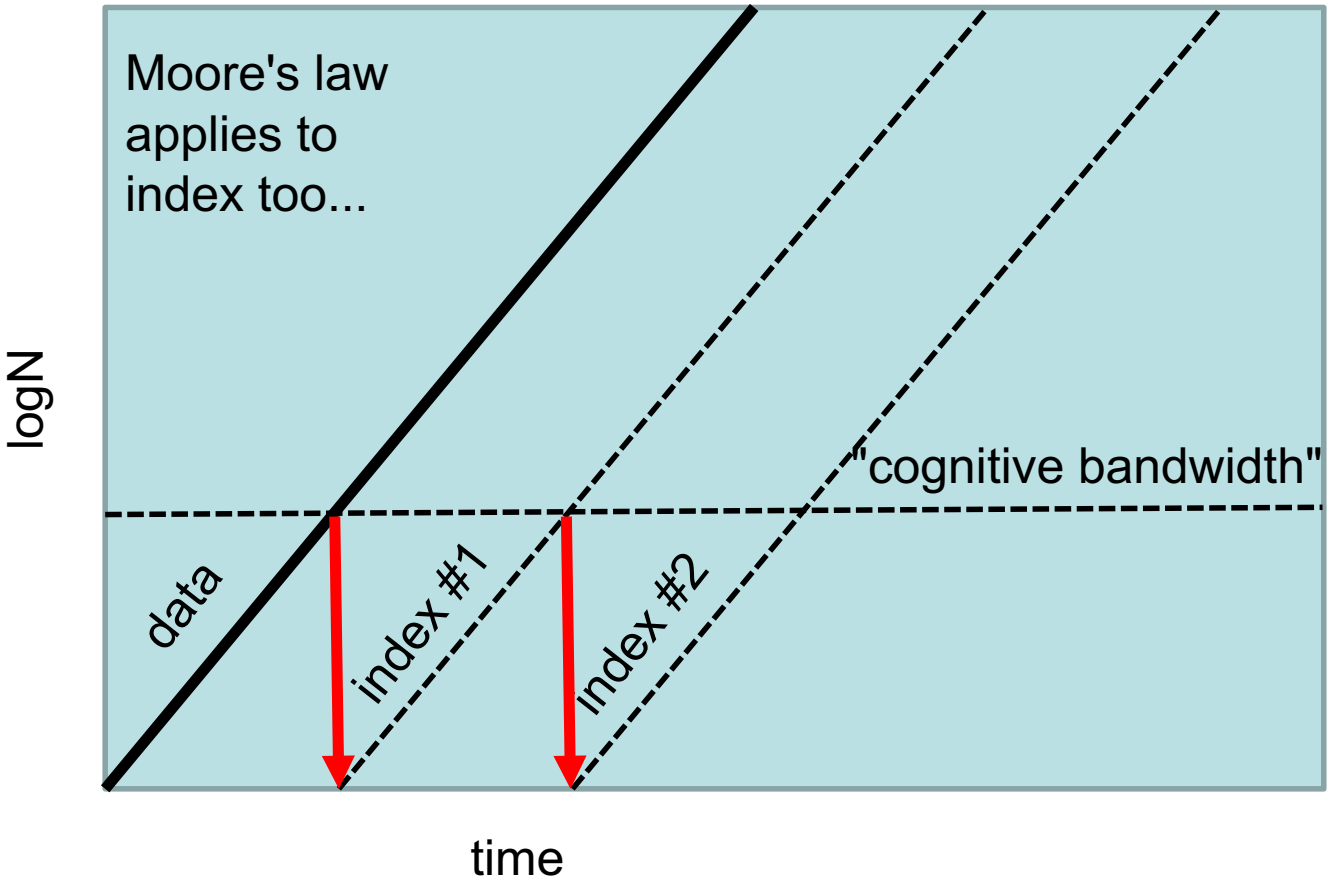
more generally: audio/multimedia/data indexing & classification

search by content, find duplicates, cluster, etc.



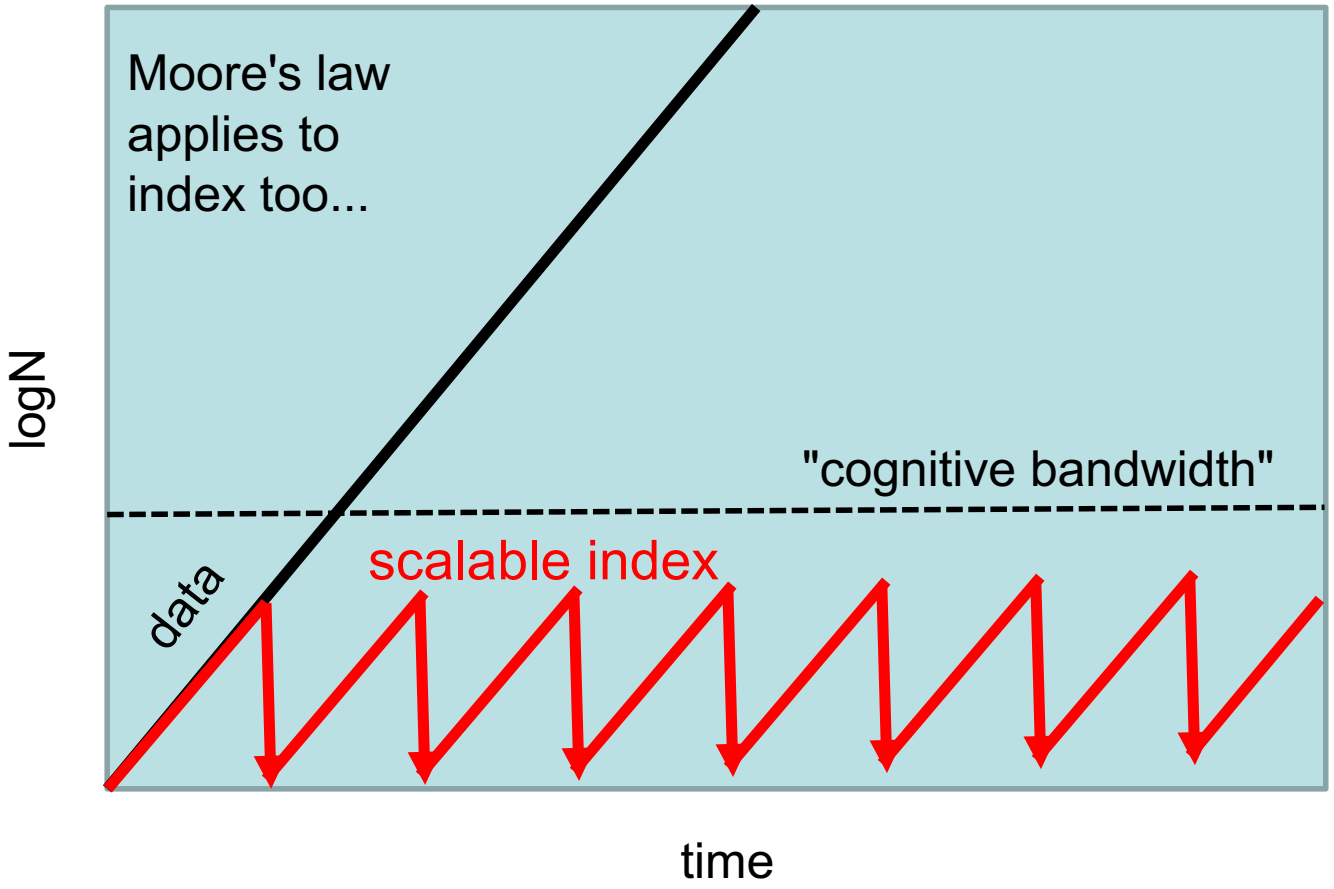
more generally: audio/multimedia/data indexing & classification

search by content, find duplicates, cluster, etc.



more generally: audio/multimedia/data indexing & classification

search by content, find duplicates, cluster, etc.

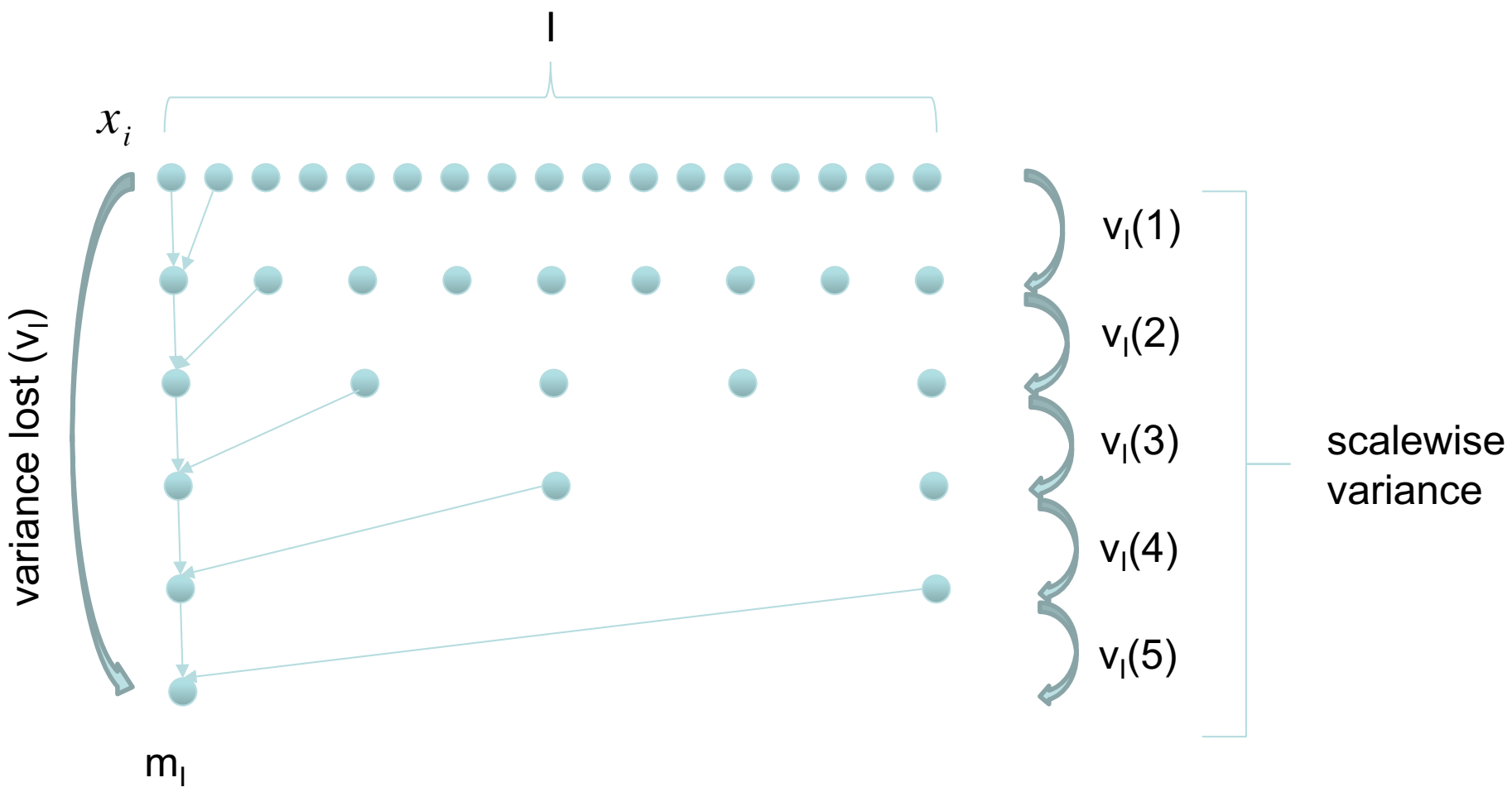


a scalable index:

- adapts to arbitrary data size
- can include a variety of useful statistics
- supports hierarchical "statistics of statistics"
- supports rescaling at the system level
- frees us from the need to worry about index storage cost

The designer of an indexing scheme can include the craziest of statistics, as long as they are scalable, knowing that their cost can be kept under control by rescaling.

Scalewise transform

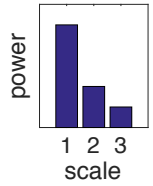

I
x_i
variance lost (v_l)
m_l
$v_l(1)$
$v_l(2)$
$v_l(3)$
$v_l(4)$
$v_l(5)$
scalewise
variance

interpretation of scalewise variance:

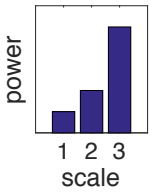
- logarithmic power spectrum:

$$\sum_{i \in I} x_i^2 = \sum_k v_I(k)$$

Parseval's relation

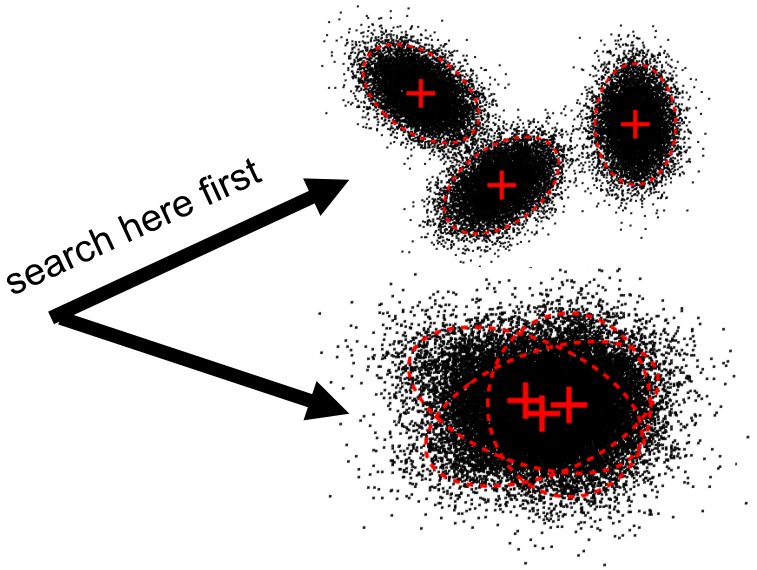


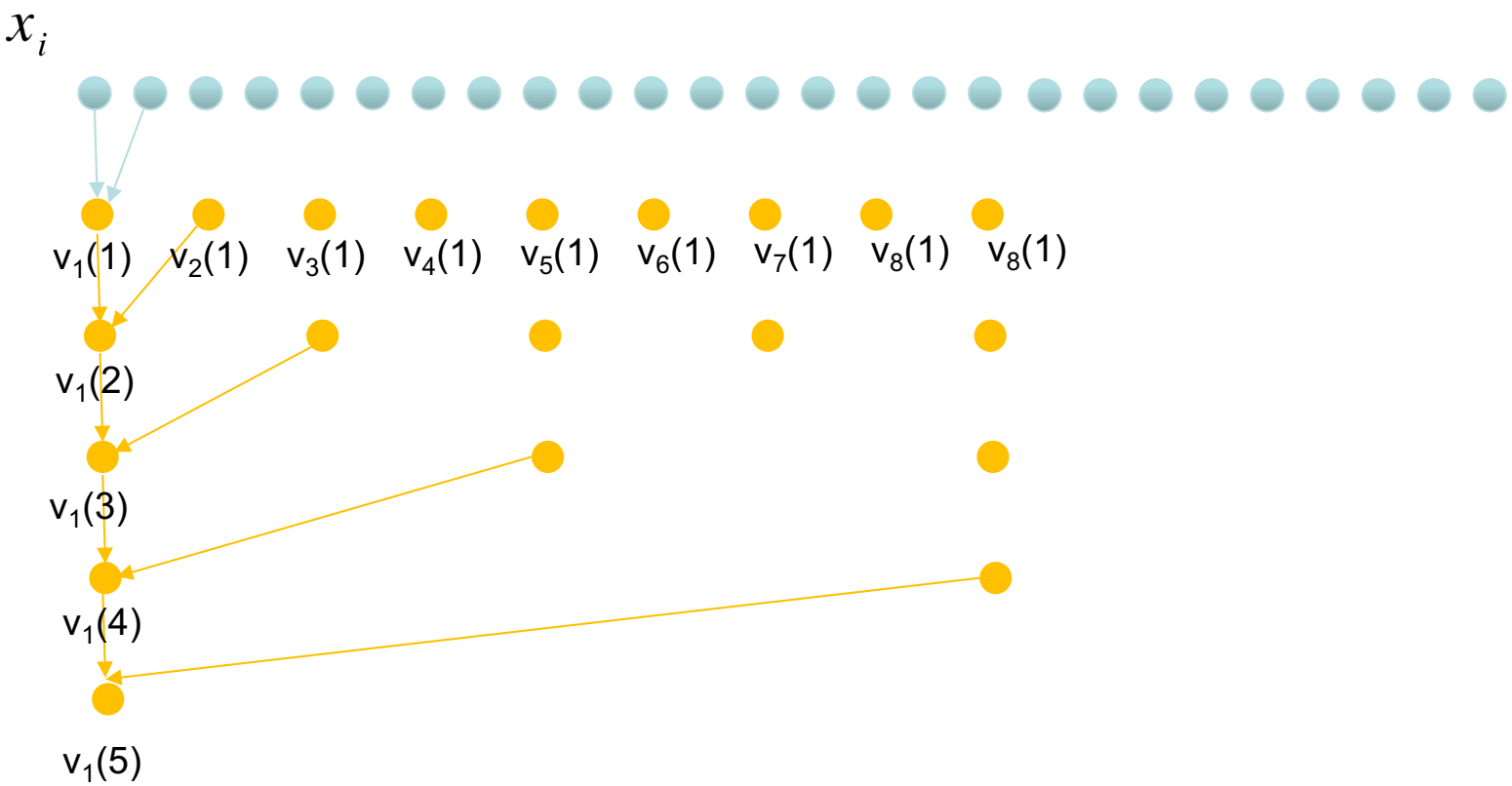
lowpass



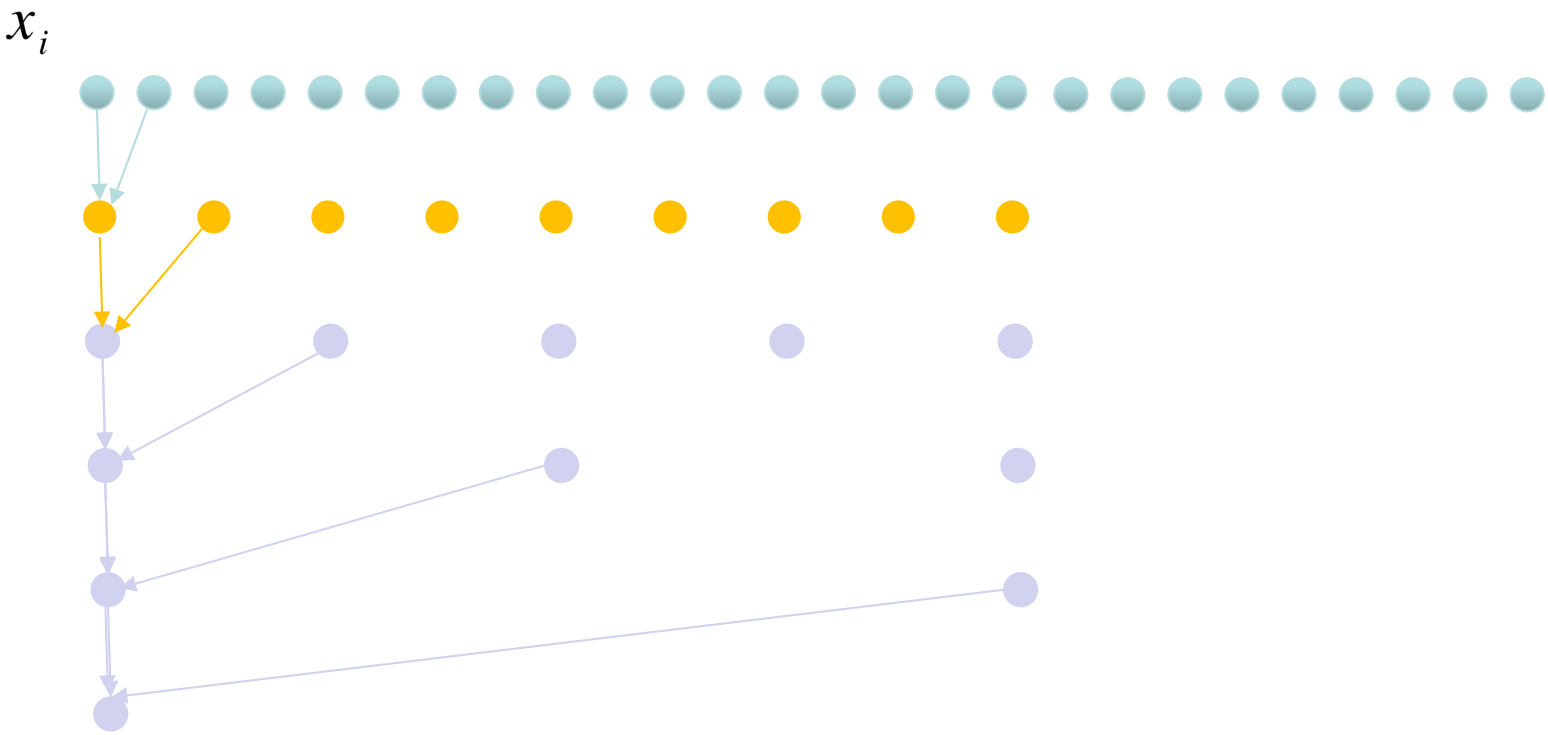
highpass

- "granularity" of distribution:

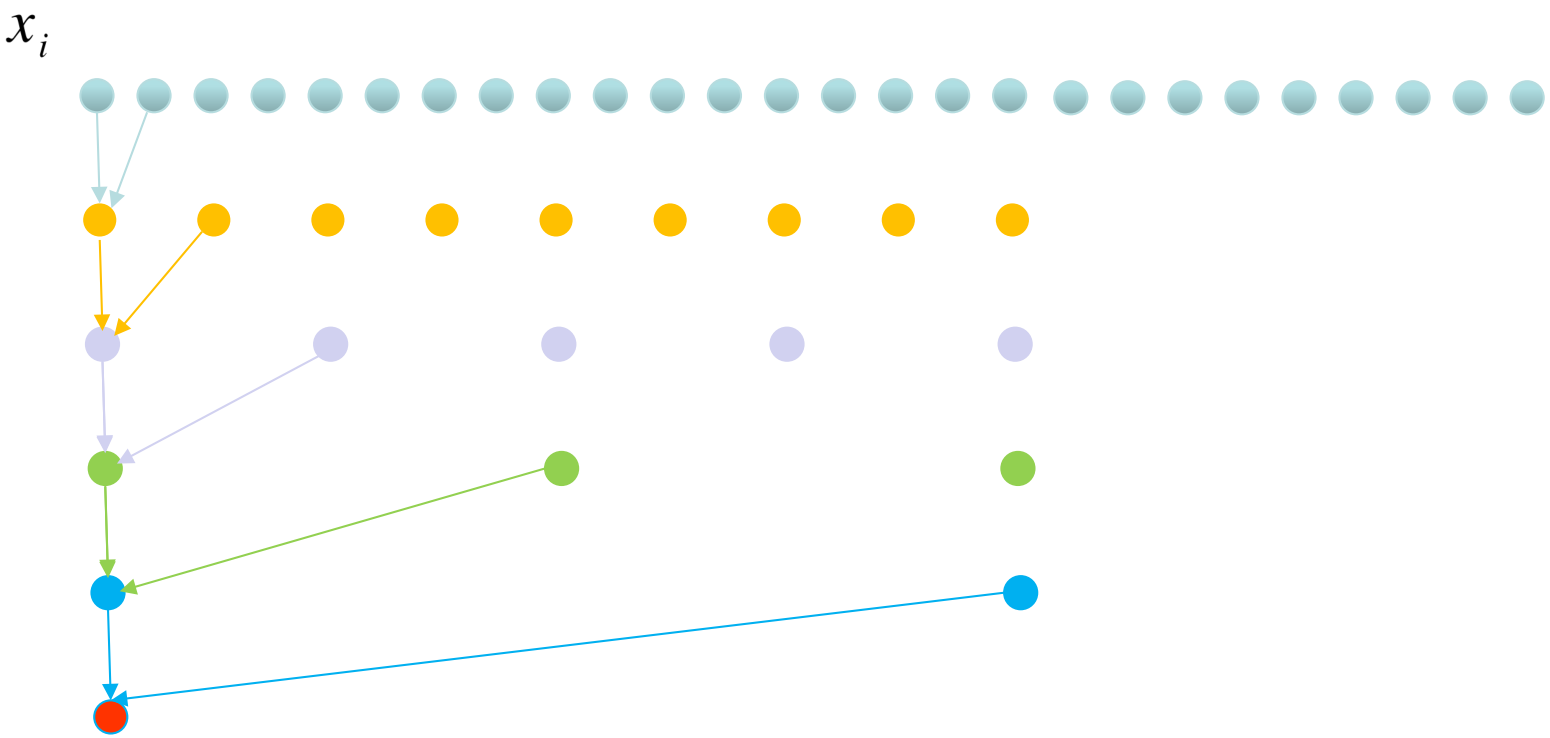




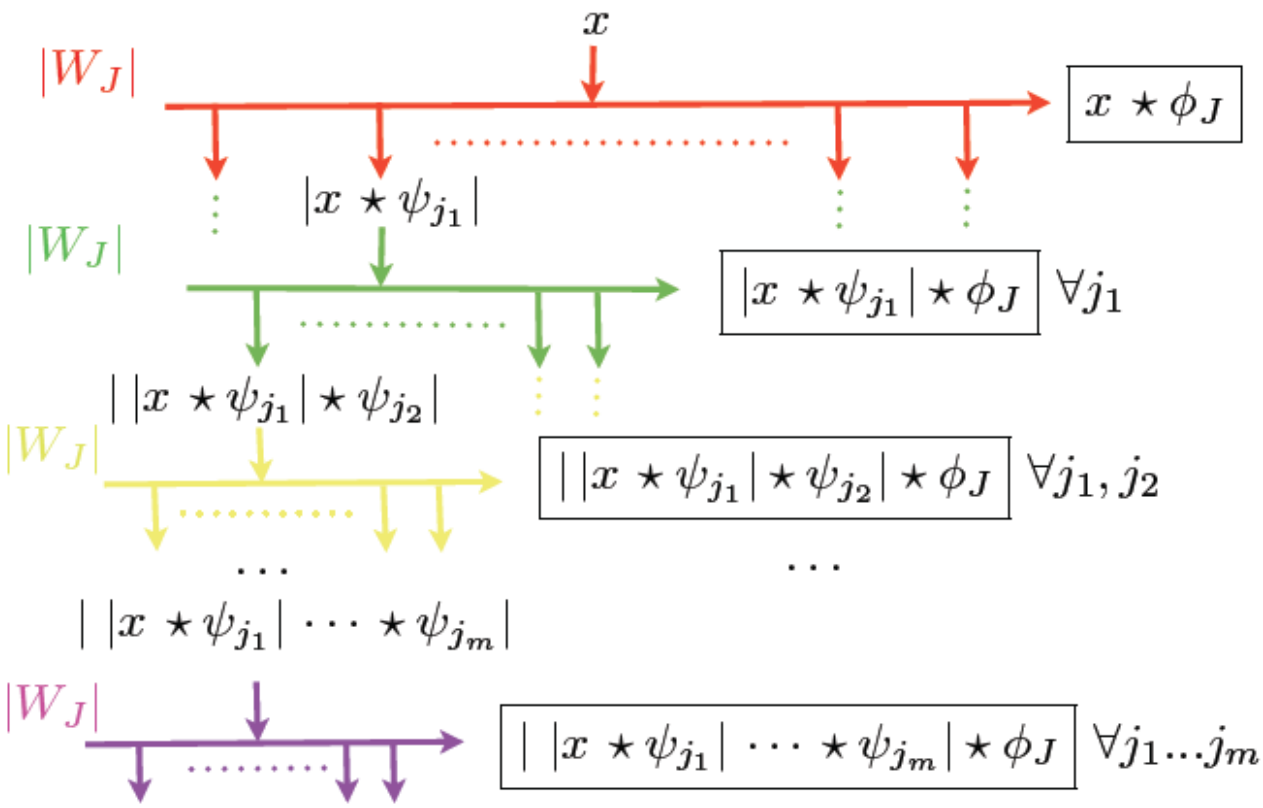
- scalewise variance coeffients form a time series
- they scale according to the "sum" rule



we can calculate *scalewise variance* of *scalewise variance*



and *scalewise variance* of *scalewise variance* of *scalewise variance,*
and so on...



similar to Stéphane Mallat's scattering transform

Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers[a)]

Torsten Dau[b)] and Birger Kollmeier
Carl von Ossietzky Univer
D-26111 Oldenburg, Ger

Armin Kohlrausch
IPO Center for Research

HISTORY OF MODULATION SPECTRUM IN ASR

Hynek Hermansky

COMPARISON OF MODULATION FEATURES FOR PHONEME RECOGNITION

Sriram Ganapathy[1], *Samuel Thomas*[1], *Hynek Hermansky*[1,2]

Second-order temporal modulation transfer functions

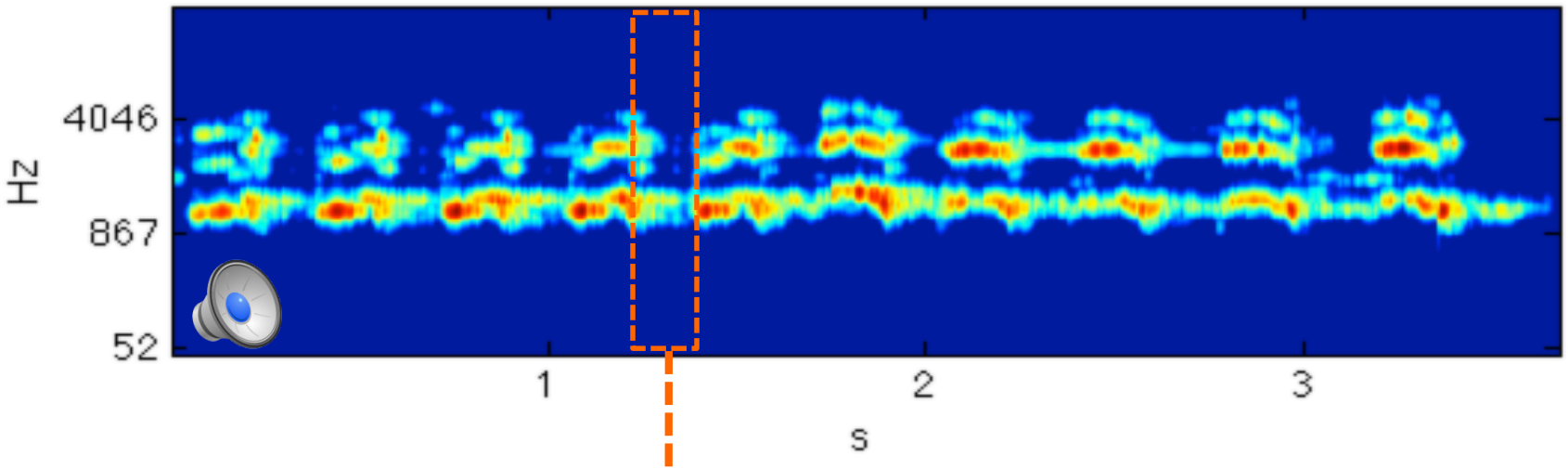
Christian Lorenzi,[a)] Catherine Soares, and Thomas Vonner

Neural Processing of Amplitude-Modulated Sounds

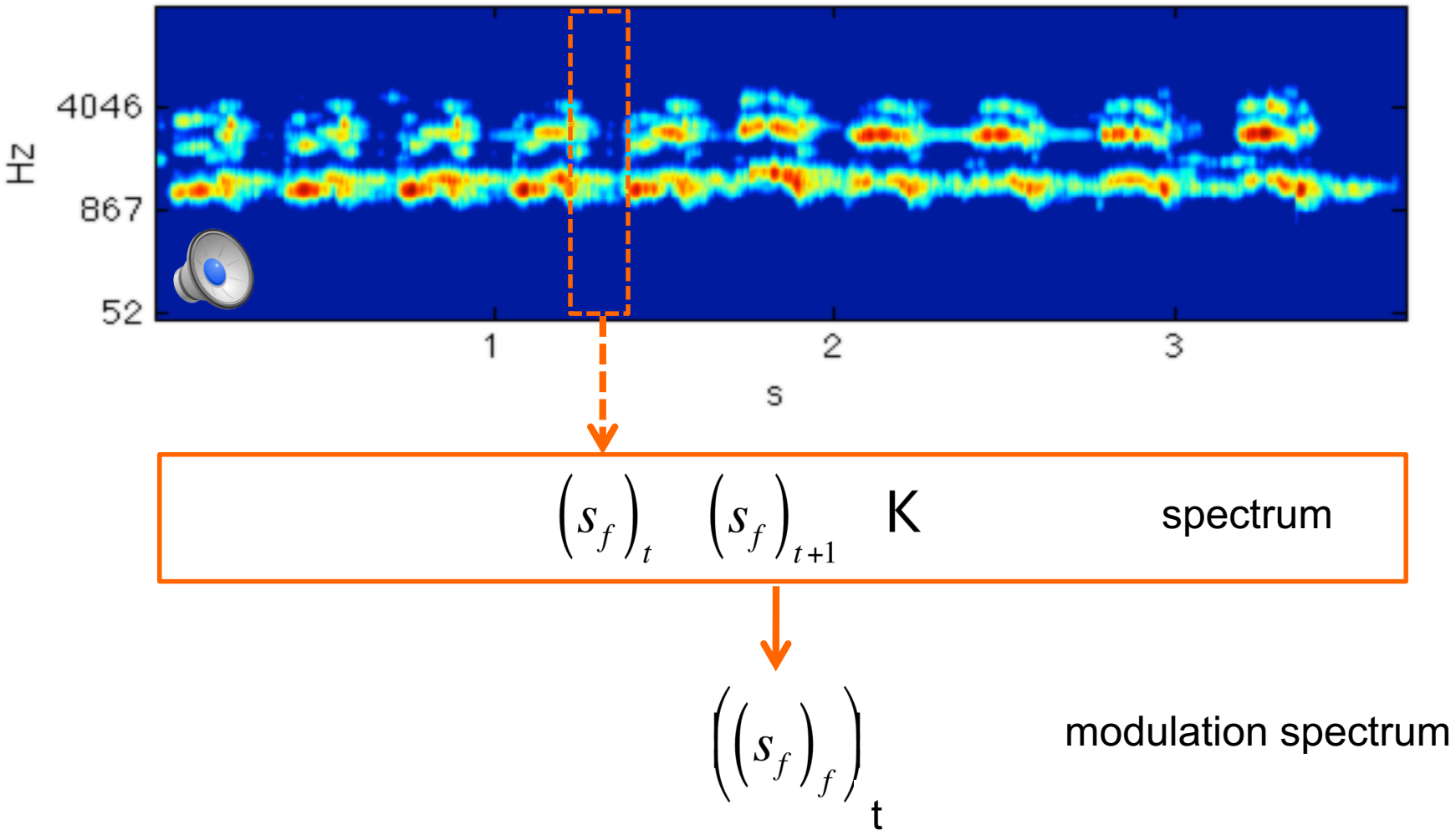
P. X. JORIS, C. E. SCHREINER, AND A. REES

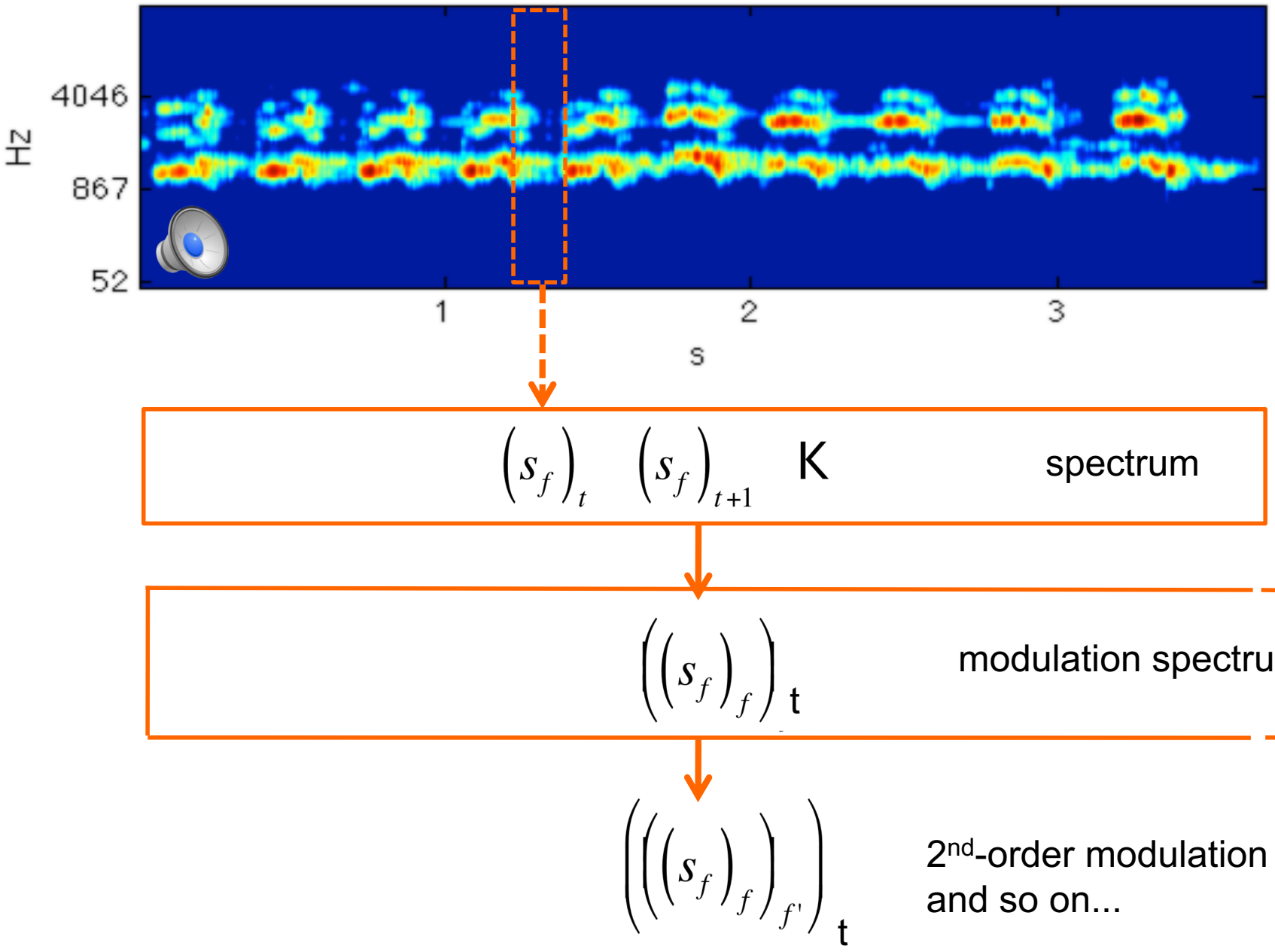
FEASIBILITY OF SINGLE CHANNEL SPEAKER SEPARATION BASED ON MODULATION FREQUENCY ANALYSIS

Steven M. Schimmel, Les E. Atlas* and Kaibao Nie***



$\left(s_f\right)_t \quad \left(s_f\right)_{t+1}$ K spectrum

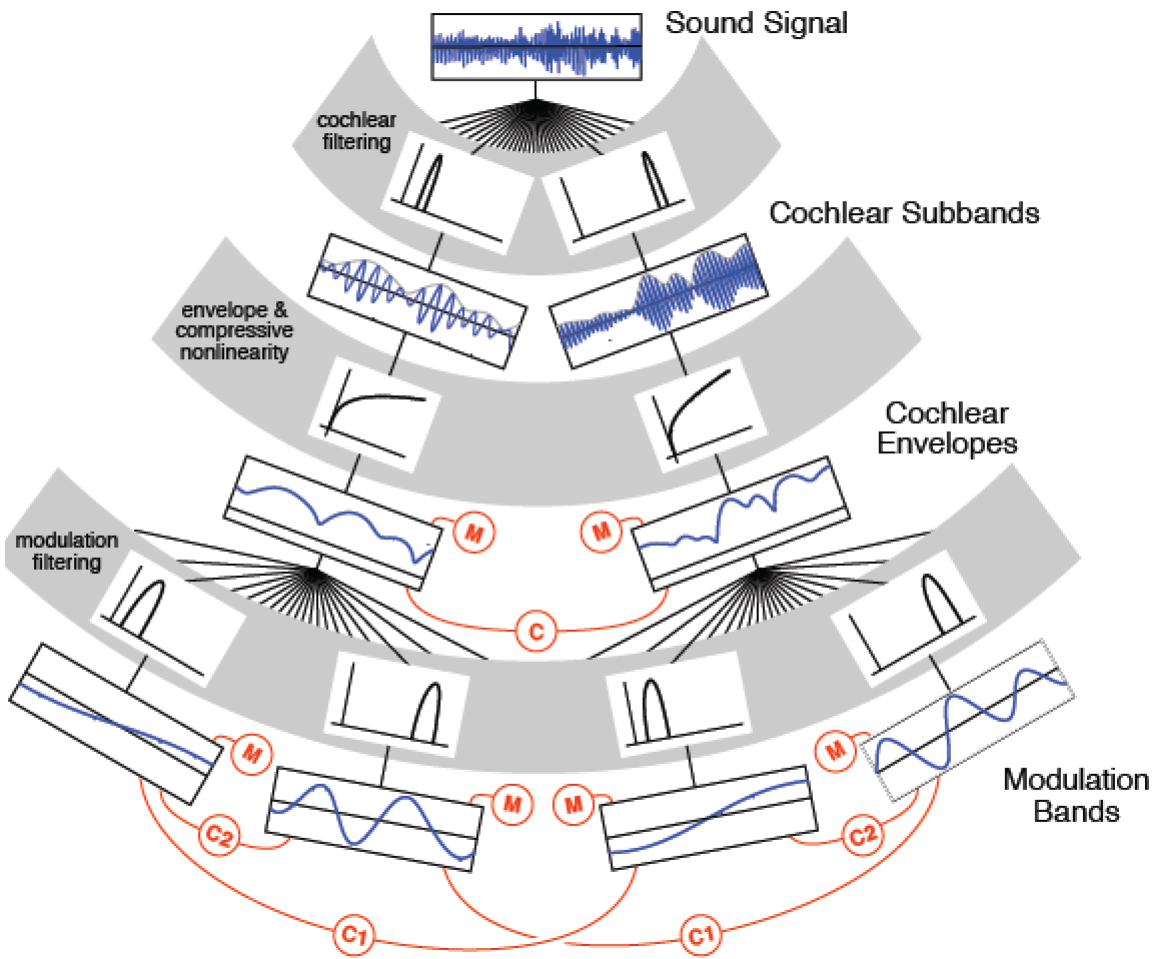

4046
867
52
Hz
1
2
3
s
$(s_f)_t$ $(s_f)_{t+1}$ K
spectrum
$((s_f)_f)_t$
modulation spectrum


4046
867
52
Hz
1
2
3
s
$\left(s_f\right)_t \quad \left(s_f\right)_{t+1}$ K
spectrum
$\left(\left(s_f\right)_f\right)_t$
modulation spectrum
$\left(\left(\left(s_f\right)_f\right)_{f'}\right)_t$
2nd-order modulation spectrum
and so on...

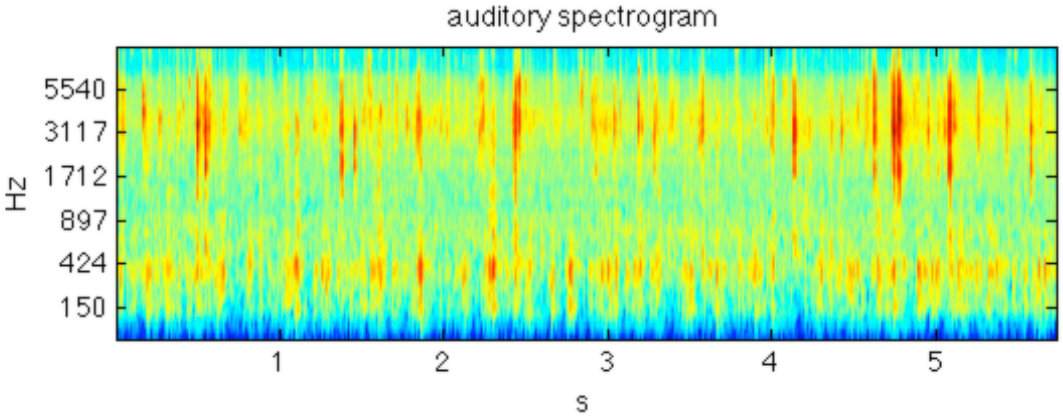
Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis

Josh H. McDermott[1,2,*] and Eero P. Simoncelli[1]

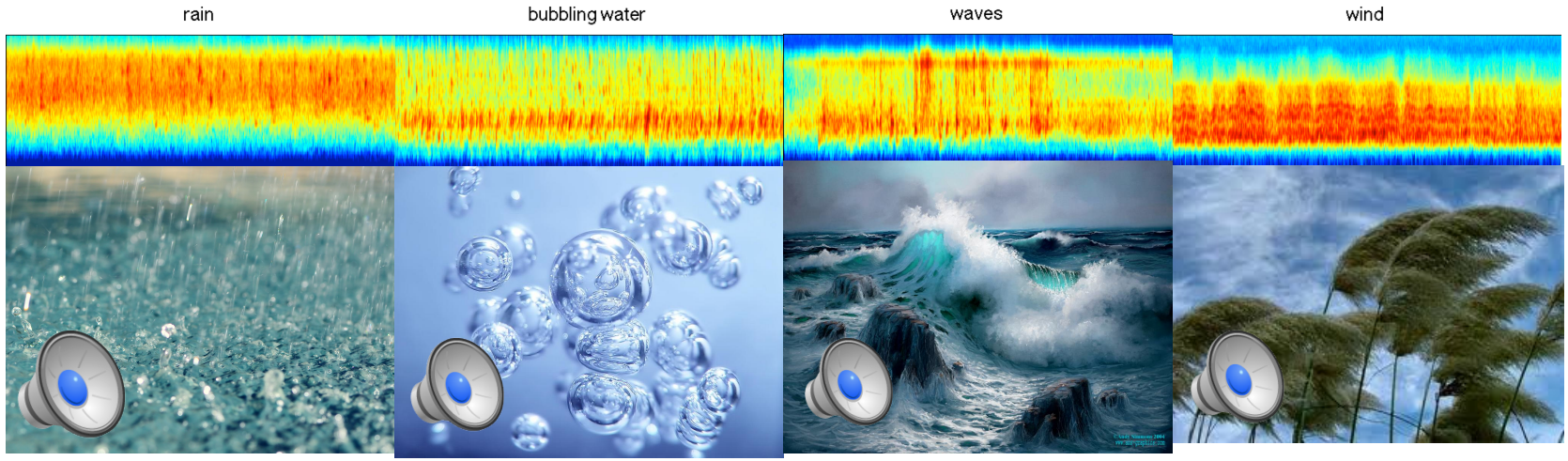
summary statistics:
- *spectrogram* (mean, variance, kurtosis)
- *modulation spectrogram*
- cross-channel *correlation*



Neuron 71, 926–940, September 8, 2011

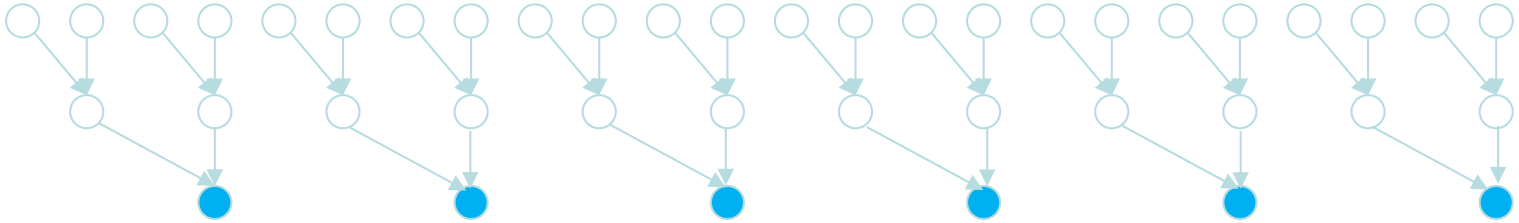


small number of statistics
(water: ~90 numbers, fire: ~900)

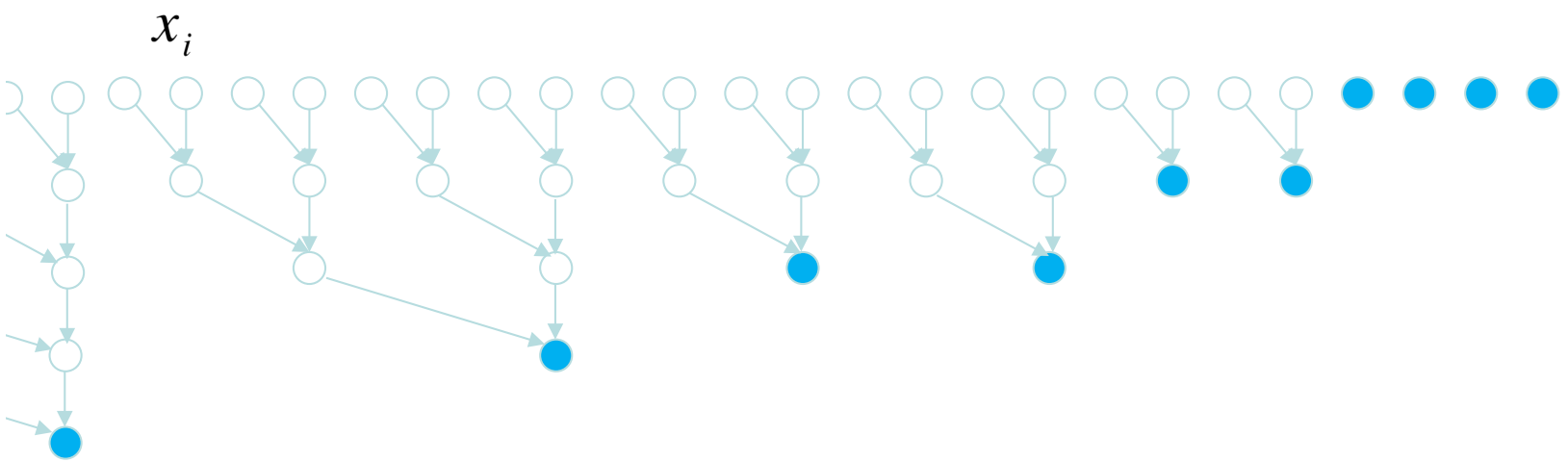


Non-uniform sampling

x_i

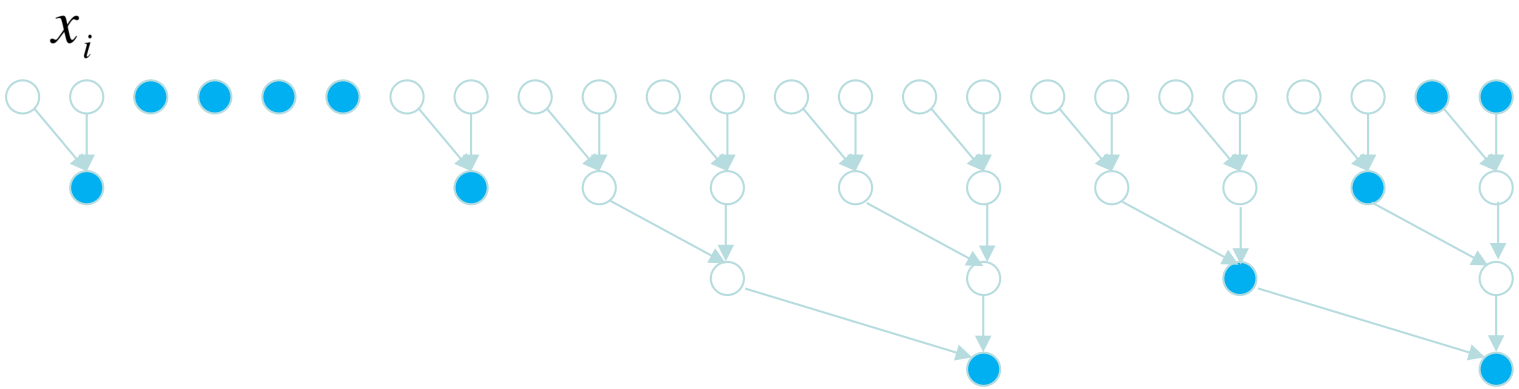


uniform sampling

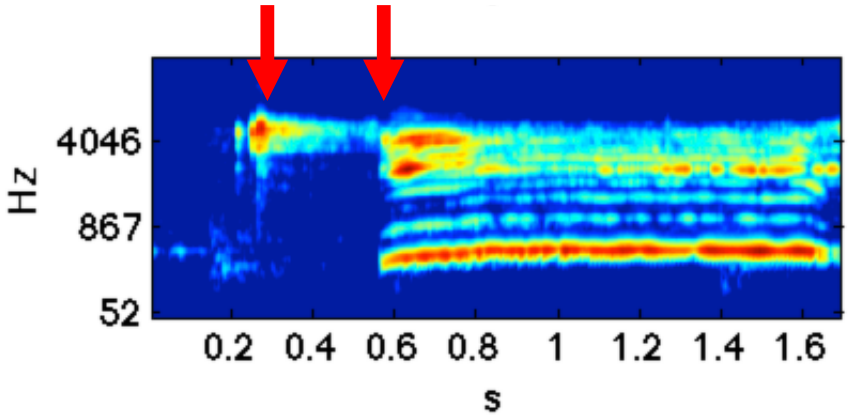


decaying sampling

similar to Tishby's predictive information



non-uniform sampling



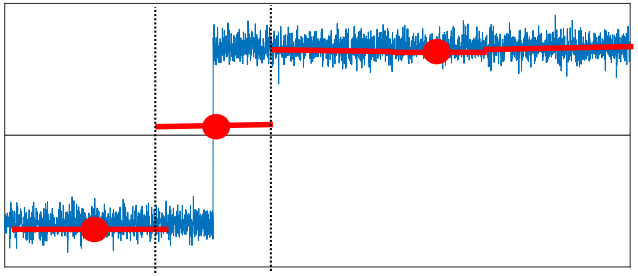
How to choose a sampling schedule?

- Extrinsic criteria (correlation with events in the world, other senses, etc.)
    - organism or robot: higher resolution for features associated with critical choices, errors, or events
    - many of us keep detailed memories of events circa 11 Sept 2001
    - sound database: higher resolution for features predictive of class, or speech recognition
    - multimedia database: higher resolution for features correlated with other modalities

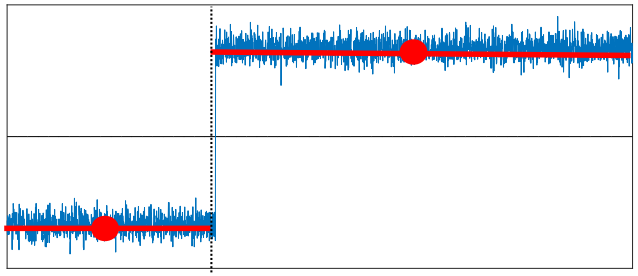
- Intrinsic criteria (depend on data)

Intrinsic criteria (signal-dependent):

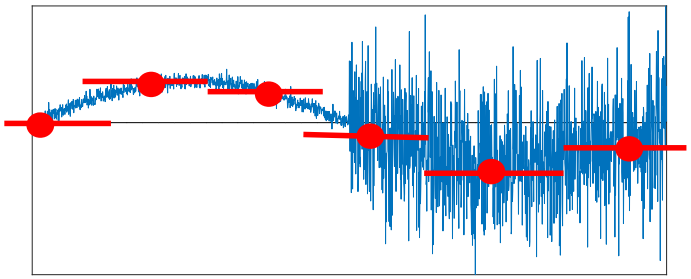
suboptimal

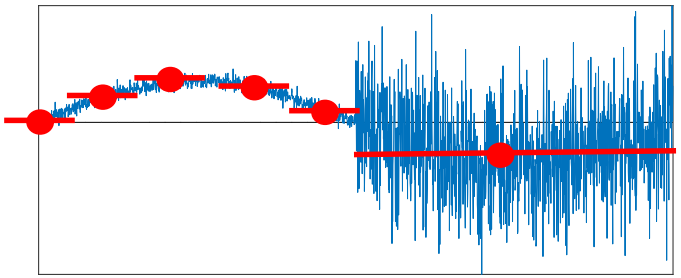


optimal



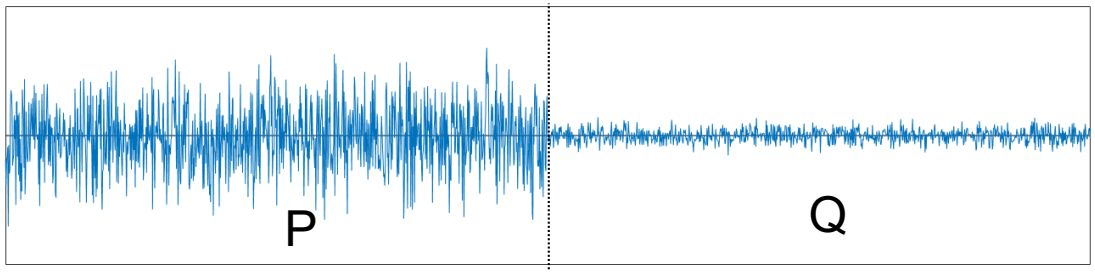
minimum total variance $\sum v_I$



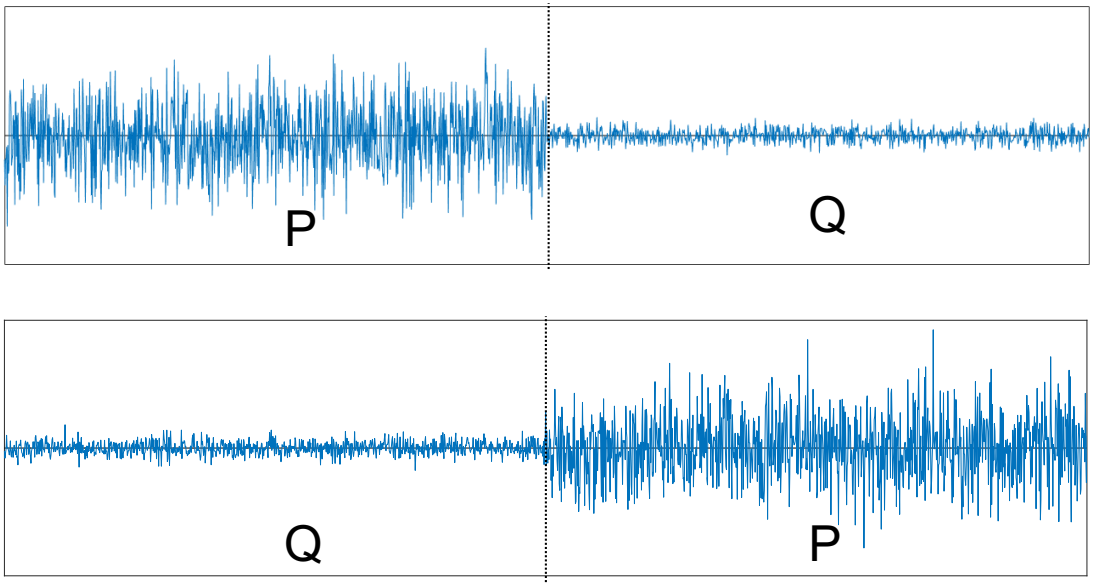


More generally: maximize the difference between the distributions of values in adjacent intervals (e.g. Kullback-Leibler divergence)

$$D_{KL}(P||Q) = \sum_i P(i)\log(P(i)/Q(i)$$



The asymmetry of time



$D_{KL}(P||Q) = 16$

$D_{KL}(Q||P) = 106$

A transition from low to high variance (e.g. Q→P) is more likely to justify a new boundary than the opposite (P→Q).

==> onsets better represented than offsets!

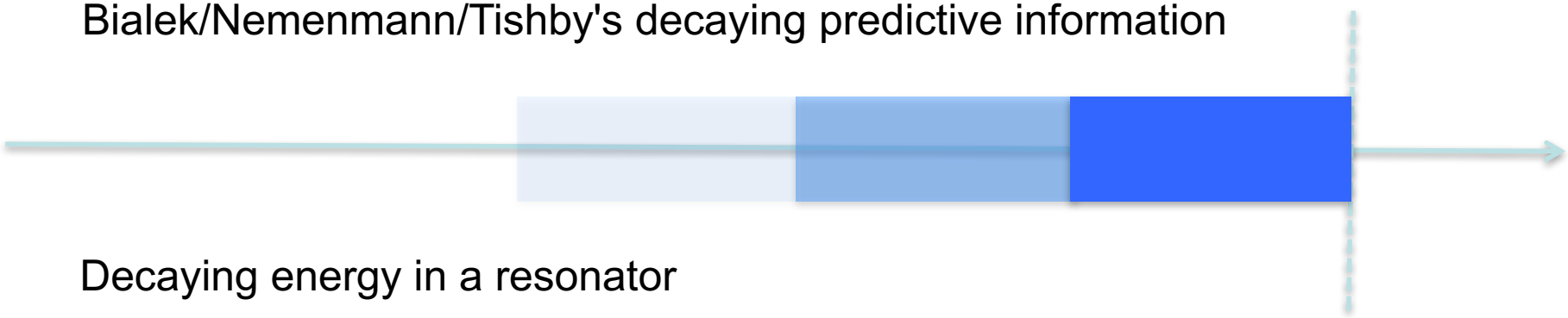
The Journal of Neuroscience, May 9, 2007 • 27(19):5207–5214

Behavioral/Systems/Cognitive

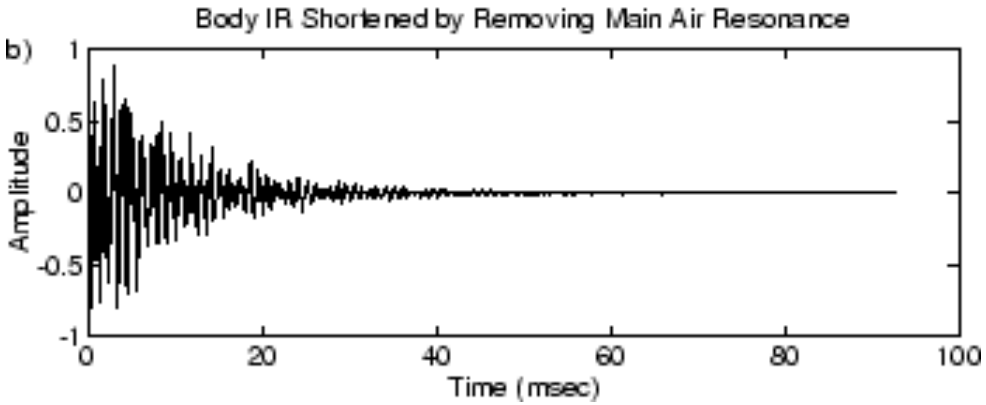
Processing Asymmetry of Transitions between Order and Disorder in Human Auditory Cortex

Maria Chait,[1] David Poeppel,[1,2,3] Alain de Cheveigné,[5] and Jonathan Z. Simon[1,3,4]

Bialek/Nemenmann/Tishby's decaying predictive information



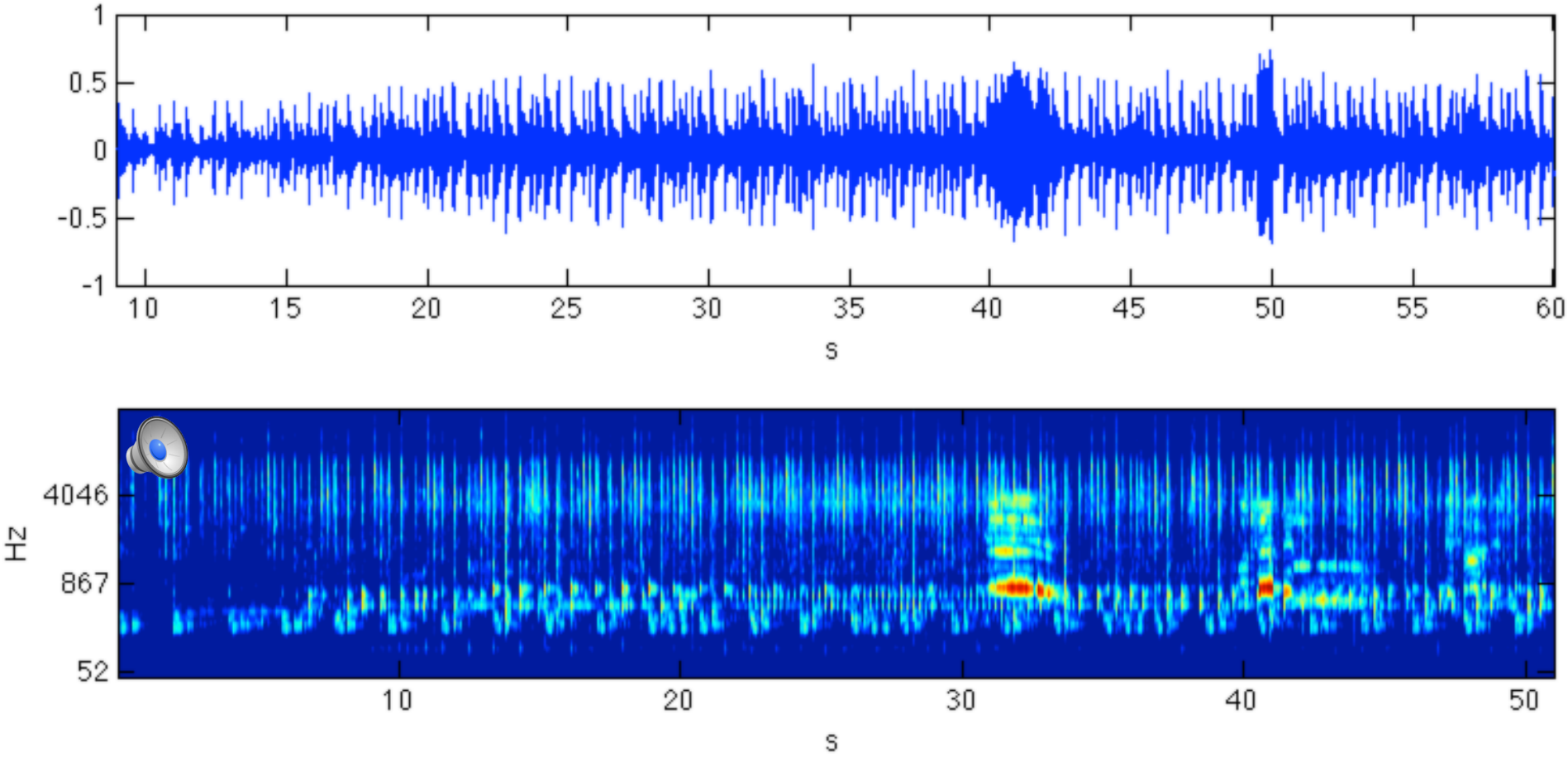
Decaying energy in a resonator



now

Time is asymmetric!

The asymmetry of time



easy to hear appearance...

The asymmetry of time

Haydn's Farewell symphony (Mariinsky / Igor Gruppman, Youtube)

The asymmetry of time

harder to hear disappearance...

Connection to deep learning:

The success of convolutional networks is due in part to the *dimensionality reduction* that results from *tied weights*. This enforces a reasonable constraint (e.g. shift invariance) that does not need to be learned.

Scalability enforces similarly reasonable constraints, and also results in dimensionality reduction. The basic scalability constraint is generic (independent from any application). Details (e.g. non-uniform sampling) can be tailored to optimize for a particular task.

This idea is reinforced by the success of scattering transform approaches that are in some respect similar to scalable statistics.

To summarize:

- Lots to remember, can't remember it all
- Forgetting is necessary *(Borges)*, the severity of the loss is limited *(Tishby)*
- Memory is not uniform:
    - more detail for recent, less for old *(Tishby's fisheye)*
    - more for "interesting" regions of time, less for "dull"
- Must be scalable (memory trace morphs from detailed to coarse)
- This might help us undersand perceptual representations
- The idea may also be useful for engineering applications (indexing, search, classification)