

Theory (*and practice*) of  
measurement

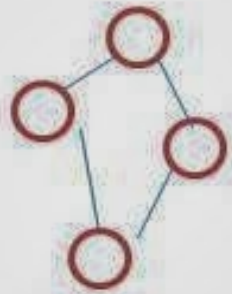
"A human being should be able to change a diaper, plan an invasion, butcher a hog, conn a ship, design a building, write a sonnet, balance accounts, build a wall, set a bone, comfort the dying, take orders, give orders, cooperate, act alone, solve equations, analyze a new problem, pitch manure, program a computer, cook a tasty meal, fight efficiently, die gallantly. Specialization is for insects." -Robert A. Heinlein

Statistical appreciation of the data is never neutral with respect to the studied phenomenon and implies the conscious acquiring of a specific perspective necessitating both a global attitude and the humility to look at the details.

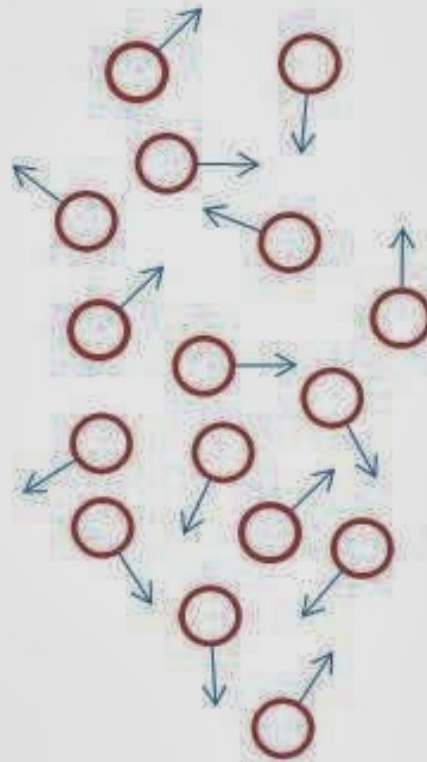
We define as emergent, a property that can be observed even by an erroneous mathematical model.

R.Laughlin

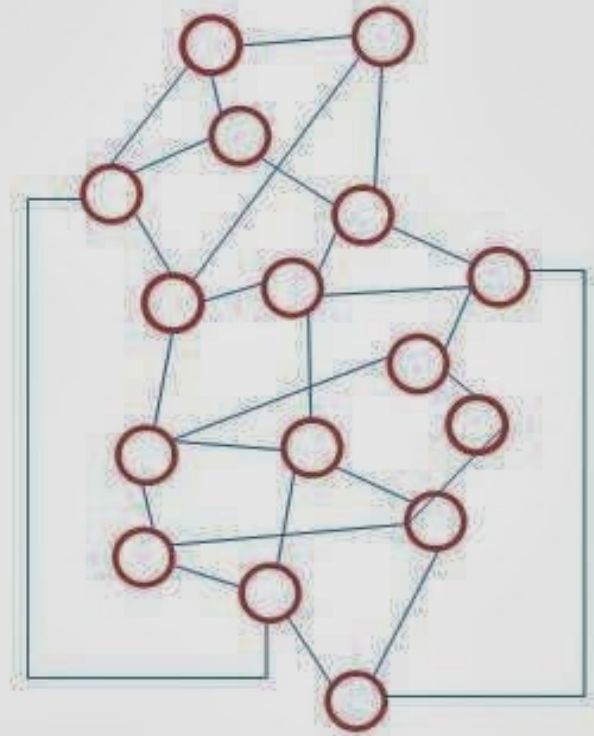
Weaver W. (1948) Science and Complexity. *Am.Scientist.* 36, 536-549



Organized  
Simplicity

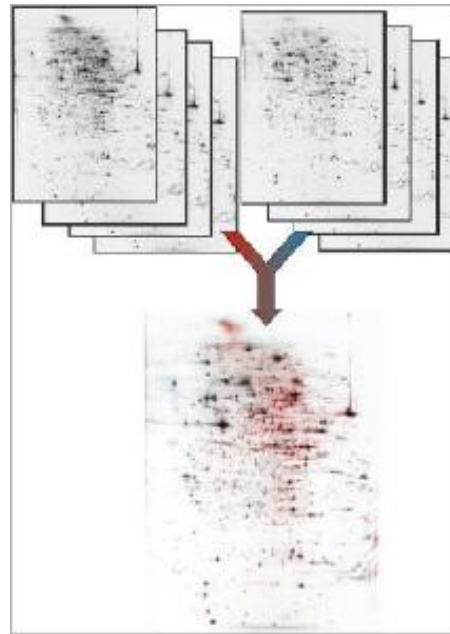


Disorganized  
Complexity



Organized  
Complexity

It is important to keep in mind that a complex system (if it is stable) allows for a level of analysis in which it displays very simple (and repeatable) behaviour. This is why medical diagnoses are much more reliable than Molecular Biology.

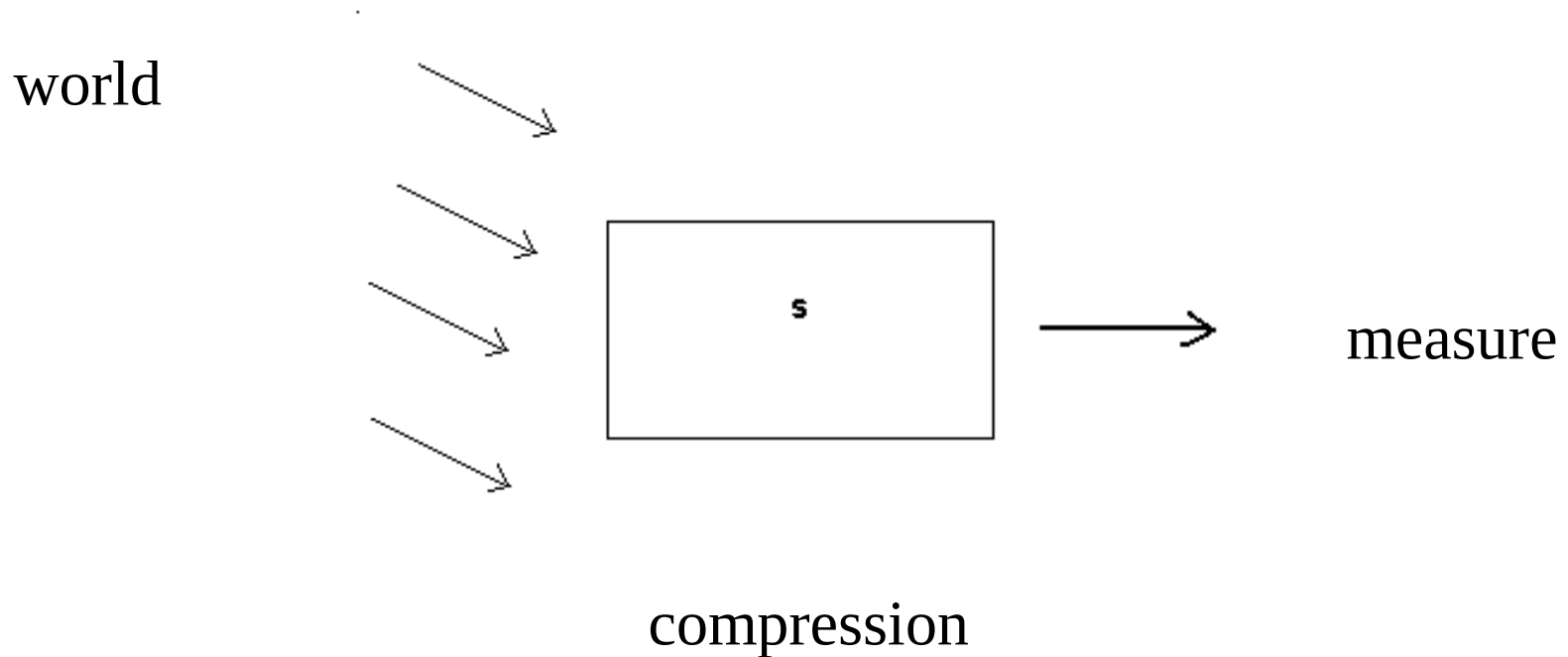


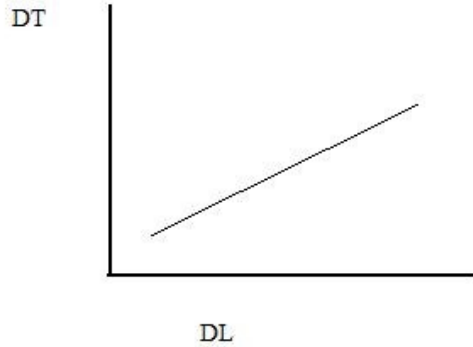
A rabbit is infinitely more complex than its proteome while having a much more predictable behaviour.



Organized Complexity does not need complicated math !!!

Any measurement implies a well defined and unique choice of perspective . We concentrate only on specific features of the system discarding the others. This provokes a compression of the original information carried by the world: different entities become indistinguishable after the measurement.





Each measurement refers to ‘something else’ that ‘as such’ is not reachable by our direct investigation. Measures are ‘proxies’ of an underlying reality, a classical case is the link between temperature (that per se corresponds to the Maxwell-Boltzmann distribution of kinetic energies of the particles of a system) and the elongation of thin columns of a suited metal in a thermometer. The link between the measure (observable) and the underlying reality (not observable) holds only into a specific domain.

**Apply Payroll/Payment**

Save   Open   Copy   Clear   Close   By Employee #

Employee #  Name  Dept #

Earnings/Deductions   Tax Withholdings   Worker's Comp   GL Distribution

Seq	Earning Code	Base On	Hours	Rate	Amount	Taxable
10	SALARY		80.00	56.54	4,447.68	<input checked="" type="checkbox"/>
20	COMMISSION		0.00	0.00	500.00	<input checked="" type="checkbox"/>
30	HOLIDAY	SALARY	8.00	56.54	452.32	<input checked="" type="checkbox"/>
40	OVERTIME	SALARY	4.00	84.81	339.24	<input checked="" type="checkbox"/>
50	OTHERNONTX		0.00	0.00	150.00	<input type="checkbox"/>

Deduction	Type	Amt/Pct/Rate	Deduction Amt	M/C Amt
401K	%	15.00	860.89	186.53
CAFE	%	12.00	688.71	660.02
CREDITUN	\$	25.00	25.00	0.00
MEDICAL	\$	33.33	33.33	0.00
UNION	\$/Hr	0.05	4.40	4.60

Pay Period Range: 06/17/09-09/01/08

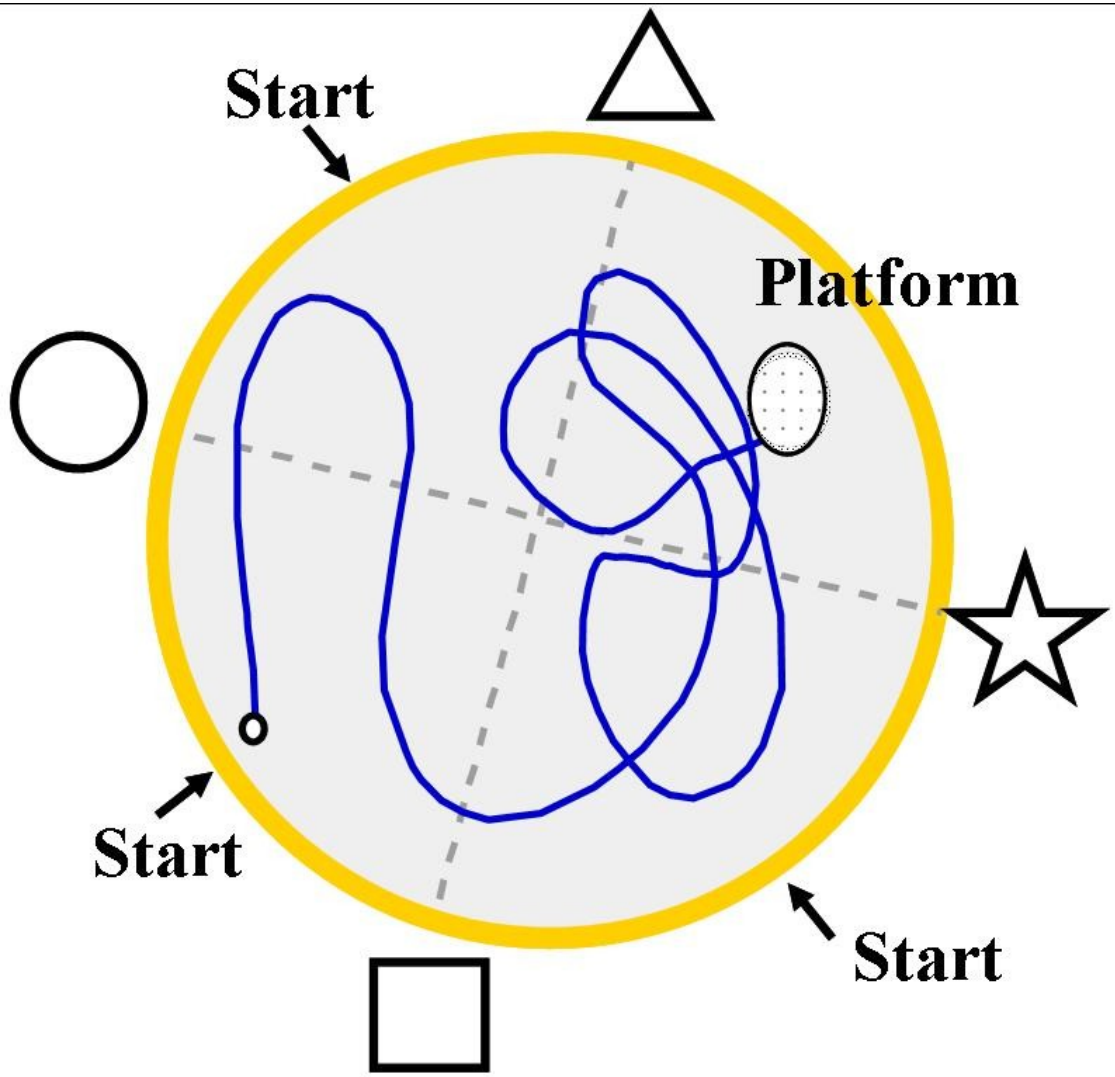
Earnings	5,889.24
Total Taxable	5,739.24
Total Non-Taxable	150.00
Total Wages	5,889.24
Total Deductions	1,612.33
Federal Tax	1,245.02
State Tax	326.52
Local Tax	0.00
Prepaid	0.00
Net Amount	2,705.37

Accrue Leaves

Tax Payment modules return a proxy of a per se not measurable entity: Wealth.



Morris Water Maze



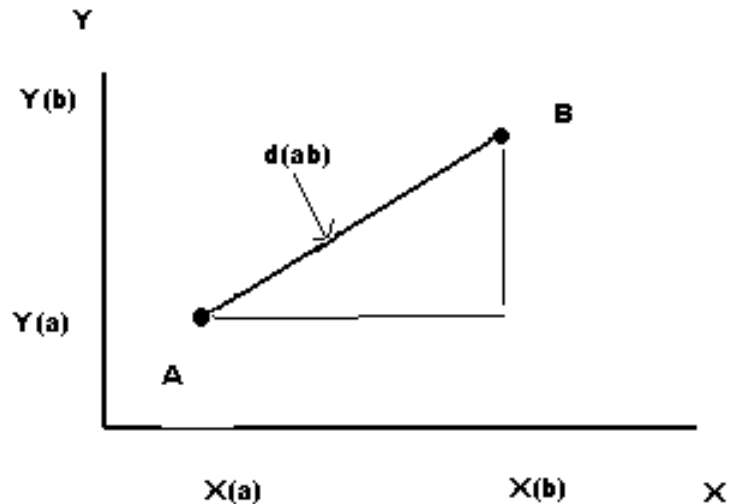
Some measures rely on physical laws, like usual thermometers based on the linear relation holding between the length of a solid (or a confined viscous liquid) and temperature, or gravity general constant like balances.

..some other correspond to a score resulting from the answer to a set of questions (e.g. fiscal modules, psychological tests).

In any case measurements will never be the 'thing-as-it-is' but proxies, something related to an 'hidden' reality behind a curtain. If the link between this 'hidden reality' and our measures changes, the sense of what we are observing changes abruptly. This is why is much better to rely on the correlation of many different measures, a change in their correlation structure tells us something is happened.

A measure is a set of rules allowing us to assign to a given event (blood sample, rat, air volume) a value.

This value must be chosen in a way suitable for a metrics to be established, i.e. it must be Possible to unequivocally say that event A is more 'similar' to B than to C.



$$D(a,b) = \text{SQRT} (X(a)-X(b))^2 + (Y(a) - Y(b))^2$$

Person 1	Person 2	Person 3	Person 4
██████	— —	— —	1
██████ —	██████ —	██████	2
██████	██████	— —	3
— —	— —	— —	4
— —	— —	— —	5
— —	— —	— —	6
— —	— —	— —	7

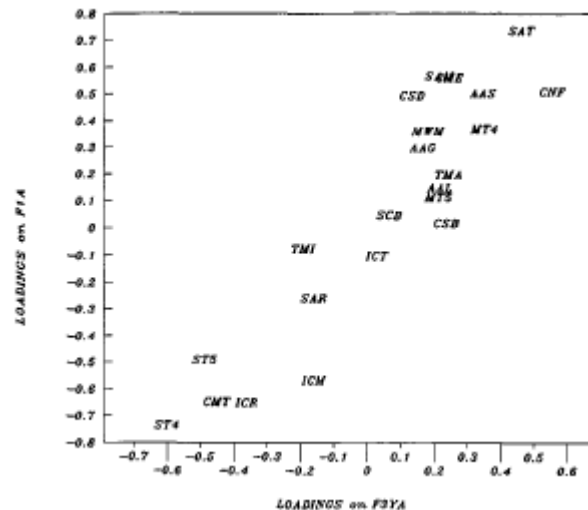
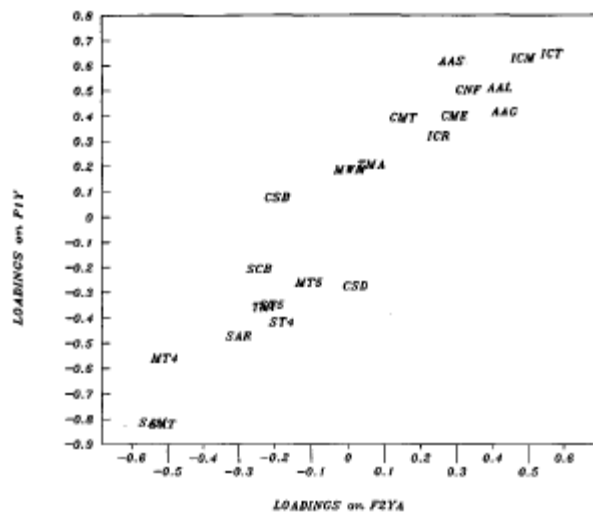
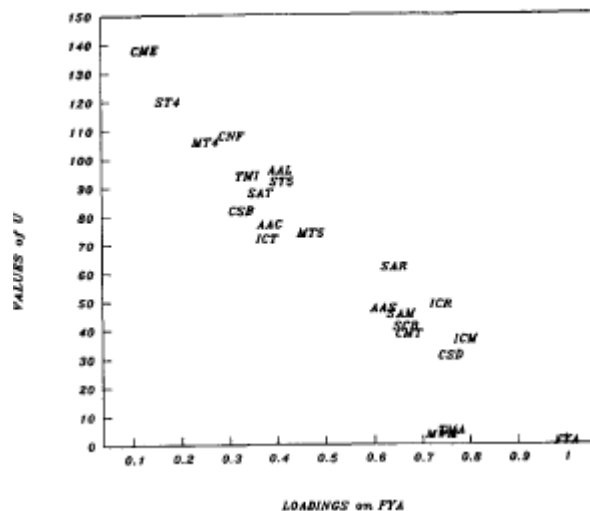
The distances between the different persons correspond to the so called Hamming distance, that in turn is the number of times, in a given position, the two samples differ as for the presence of a band.

DNA fingerprint is a qualitative feature that becomes quantitative so allowing to establish a metric space thanks to a distance operator

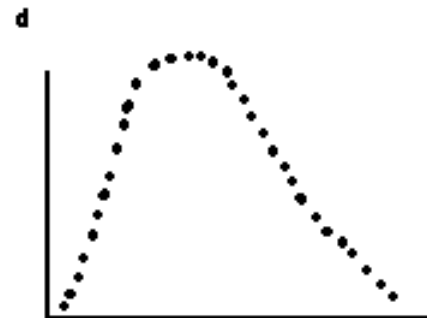
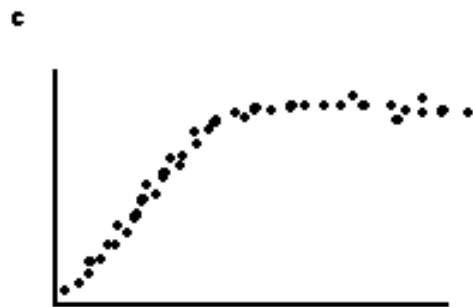
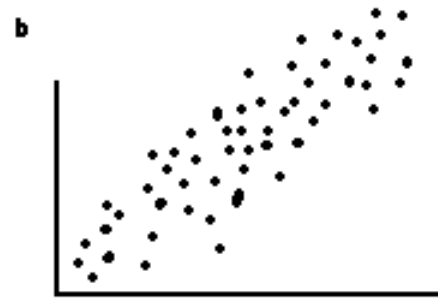
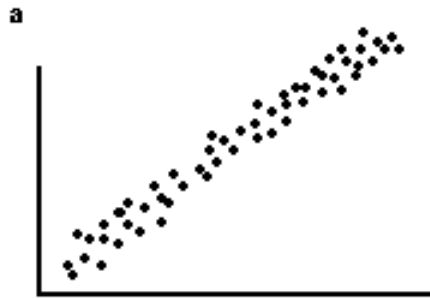
## Multivariate Analysis of Behavioral Aging Highlights Some Unexpected Features of Complex Systems Organization

ALESSANDRO GIULIANI, ORLANDO GHIRARDI, ANTONIO CAPRIOLI, STEFANO DI SERIO, MARIA TERESA RAMACCI, AND LUCIANO ANGELUCCI\*<sup>1</sup>

*Institute for Research on Senescence, Sigma-Tau S.p.A., Via Pontina km. 30.400, 00040 Pomezia, Rome, Italy; and \*Institute of Pharmacology II, School of Medicine, "La Sapienza" University of Rome, 00187 Rome, Italy*



# Different patterns of measurement



# Measurement Scales

- *Interval Scale* : The differences have a quantitative invariant meaning.
- *Ordinal Scale*: Only rank is invariant, not the actual differences.
- *Qualitative Scale*: Categories, only class allocation is reliable.

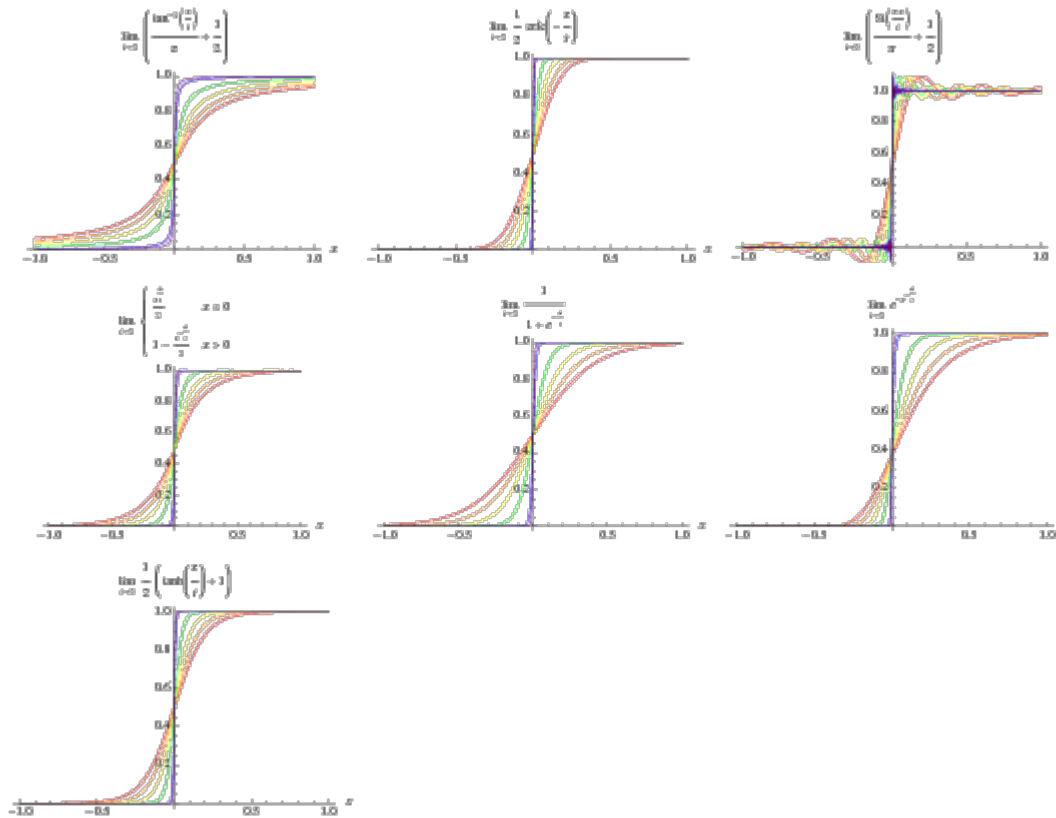
*Interval Scale*: temperature, pressure, weight : All the arithmetic operations allowed

*Ordinal Scale*: school grades, arrival order of a race: Order ( $>$  =  $<$ ) operations allowed

*Qualitative Scale*: hair colour, sex, genotype: Only logical ( $=$ ) operations allowed

We can deliberately downsizing the definition level of our measurement if this allows to get a better information quality

If the departure from linearity, in the range of interest, is very marked, a continuous, interval, measure can profitably be considered as a qualitative YES/NO measure. This corresponds to a filter maintaining the signal portion of the information and eliminating noise.





# Coding means choosing a privileged view on the data

Chem. Rev. 2002, 102, 1471–1491

1471

## Nonlinear Signal Analysis Methods in the Elucidation of Protein Sequence–Structure Relationships

Alessandro Giuliani,<sup>\*†</sup> Romualdo Benigni,<sup>†</sup> Joseph P. Zbitun,<sup>‡</sup> Charles L. Webber, Jr.,<sup>§</sup> Paolo Sirabella,<sup>||</sup> and Alfredo Colosimo<sup>||</sup>

*Istituto Superiore di Sanità, TCE Laboratory, V.le R. Elena 299, 00161 Roma, Italy, Department of Molecular Biophysics and Physiology, Rush University, 1653 West Congress Parkway, Chicago, Illinois 60612, Department of Physiology, Loyola University Medical Center, 2160 South First Avenue, Maywood, Illinois 60153, and Department of Biochemical Sciences, University of Rome, "La Sapienza", P.le A. Moro, 5-00185 Roma, Italy*

Aminoacid sequence (monoletter symbols)

VLSGADKTNVKAAWPKVPAHAPEYPAEALERMFLSFGTTKTYFGHF  
DLSHPSAQVKPHPKVADALTNVAHAVDDMGNALSALSSDLHAH  
KLRVDGVNFKLHSHCLLVTLAAHLGAEFTGAVHASLKDKFLASVSTVLTSKYR

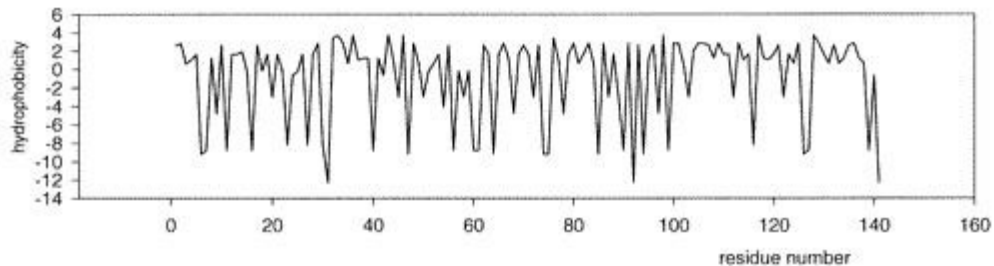


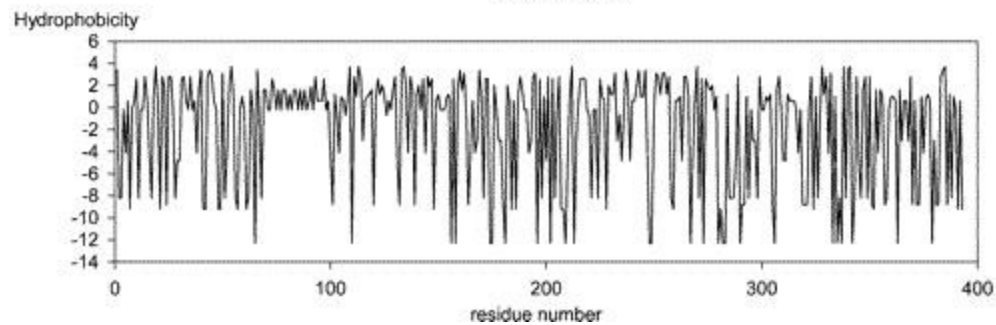
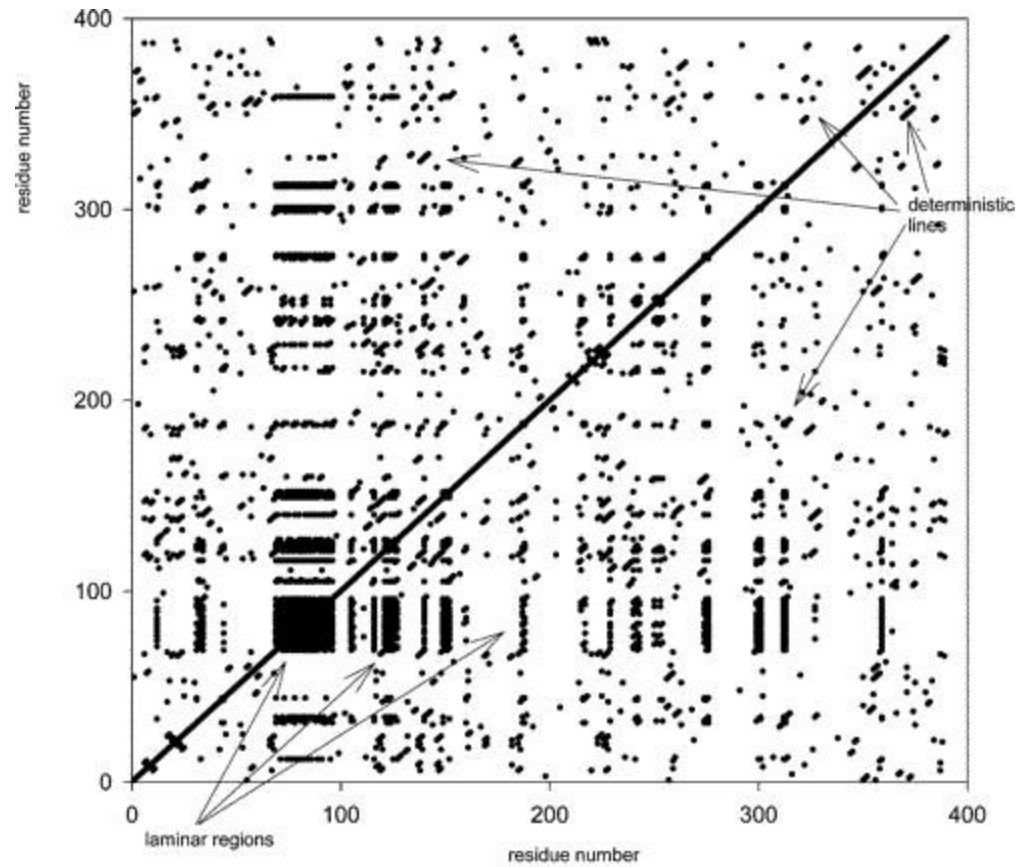
Hydrophobicity scale

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1.6	-12.3	-4.8	-9.2	2	-4.1	-8.2	1	-3	3.1	2.8	-8.8	3.4	3.7	-0.2	0.6	1.2	1.9	-0.7	2.6

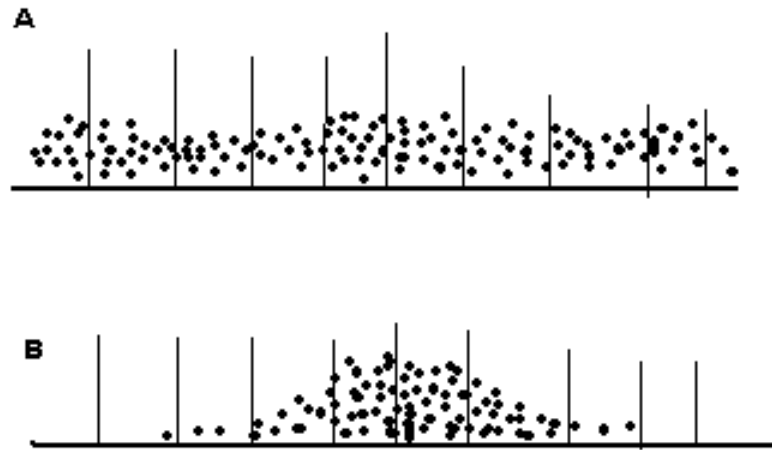


Aminoacid sequence (numerical series)





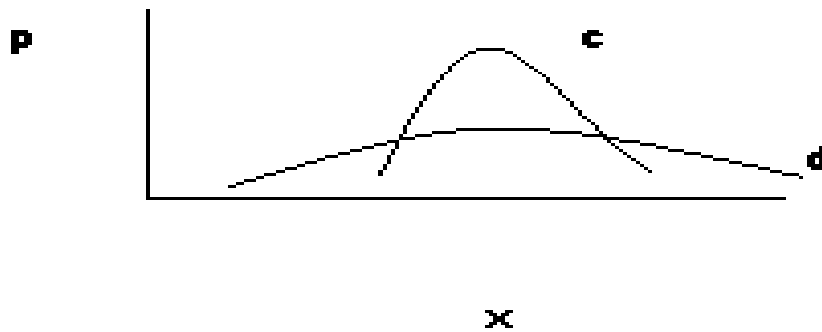
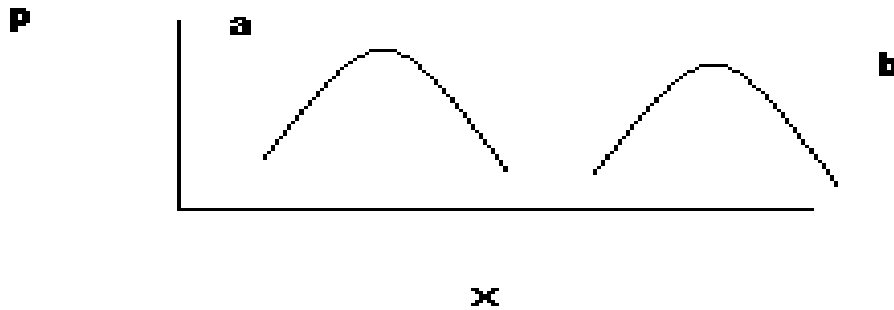
The amount of information we can derive from a given measurement depends on its frequency distribution.



$$\text{Shannon's Entropy} = -\sum p(i) \lg(p(i))$$

$E(X) = \sum (X(i))/N$  ...for rank it becomes the Median, for qualitative the Mode

Std. Dev.  $(X) = \sqrt{\sum (X(i) - E(X))^2 / N}$ ...for rank it becomes the interquartile range, for qualitative the Entropy.



Normalization allows for judging about the order of magnitude of a measurement value

**20 is big or small ?**

*Two common normalizations*

- 1) Dividing for the physical maximum (it is OK for positive numbers)
- 2) Subtracting the mean and dividing for SD (context dependent)

## Data Matrix

nome	Eta'	Abitazione	Reddito	Sesso	Lavoro
Mario	23	120	A	M	Lib. Prof.
Vanda	56	80	B	F	Cas.
Pietro	72	100	M	M	Pens.
Luca	38	130	M	M	Imp.
Pina	18	60	B	F	Stud
Lucia	25	75	M	F	Imp.
Tonino	42	62	B	M	Op.
Andrea	58	100	B	M	Contad.
Virginia	34	80	A	F	Lib. Prof.

## Perspective: Sloppiness and emergent theories in physics, biology, and beyond

Mark K. Transtrum,<sup>1</sup> Benjamin B. Machta,<sup>2</sup> Kevin S. Brown,<sup>3,4</sup> Bryan C. Daniels,<sup>5</sup>  
Christopher R. Mvers,<sup>6,7</sup> and James P. Sethna<sup>6</sup>

As a young physicist, Dyson paid a visit to Enrico Fermi<sup>1</sup> (recounted in Ditley, Mayer, and Loew<sup>2</sup>). Dyson wanted to tell Fermi about a set of calculations that he was quite excited about. Fermi asked Dyson how many parameters needed to be tuned in the theory to match experimental data. When Dyson replied there were four, Fermi shared with Dyson a favorite adage of his that he had learned from Von Neumann: “with four parameters I can fit an elephant, and with five I can make him wiggle his trunk.” Dejected, Dyson took the next bus back to Ithaca.

# Summarizing Data Sets

Main topic: How to get an immediate (albeit very rough) picture of a large set of observations.

Ancillary topics: Location, Variability and Shape descriptors;  
Graphical Methods



# What is Statistics

---

Definition: Science of collection, presentation, analysis, and reasonable interpretation of data.

Statistics presents a rigorous scientific method for gaining insight into data. For example, suppose we measure the weight of 100 patients in a study. With so many measurements, simply looking at the data fails to provide an informative account. However statistics can give an instant overall picture of data based on graphical presentation or numerical summarization irrespective to the number of data points. Besides data summarization, another important task of statistics is to make inference and predict relations of variables.

# Statistical Description of Data

- Statistics describes a numeric set of data by its
  - Center
  - Variability
  - Shape
- Statistics describes a categorical set of data by
  - Frequency, percentage or proportion of each category

# Some Definitions

*Variable* - any characteristic of an individual or entity. A variable can take different values for different individuals. Variables can be *categorical* or *quantitative*. Per S. S. Stevens...

- Nominal - Categorical variables with no inherent order or ranking sequence such as names or classes (e.g., gender). Value may be a numerical, but without numerical value (e.g., I, II, III). The only operation that can be applied to Nominal variables is enumeration.
- Ordinal - Variables with an inherent rank or order, e.g. mild, moderate, severe. Can be compared for equality, or greater or less, but not *how much* greater or less.
- Interval - Values of the variable are ordered as in Ordinal, and additionally, differences between values are meaningful, however, the scale is not absolutely anchored. Calendar dates and temperatures on the Fahrenheit scale are examples. Addition and subtraction, but not multiplication and division are meaningful operations.
- Ratio - Variables with all properties of Interval plus an absolute, non-arbitrary zero point, e.g. age, weight, temperature (Kelvin). Addition, subtraction, multiplication, and division are all meaningful operations.

# Some Definitions

***Distribution*** - (of a variable) tells us what values the variable takes and how often it takes these values.

- Unimodal - having a single peak
- Bimodal - having two distinct peaks
- Symmetric - left and right half are mirror images.

# Frequency Distribution

Consider a data set of 26 children of ages 1-6 years. Then the frequency distribution of variable 'age' can be tabulated as follows:

## Frequency Distribution of Age

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2

## Grouped Frequency Distribution of Age:

Age Group	1-2	3-4	5-6
Frequency	8	12	6

# Cumulative Frequency

Cumulative frequency of data in previous page

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2
Cumulative Frequency	5	8	15	20	24	26

Age Group	1-2	3-4	5-6
Frequency	8	12	6
Cumulative Frequency	8	20	26

# Data Presentation

Two types of statistical presentation of data - graphical and numerical.

Graphical Presentation: We look for the overall pattern and for striking deviations from that pattern. Overall pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an *outlier*.

Bar diagram and Pie charts are used for categorical variables.

Histogram, stem and leaf and Box-plot are used for numerical variable.

# Data Presentation –Categorical Variable

Bar Diagram: Lists the categories and presents the percent or count of individuals who fall in each category.



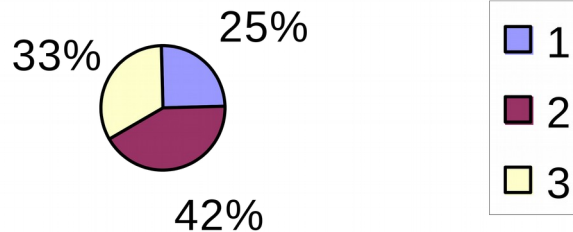
Treatment Group	Frequency	Proportion	Percent (%)
1	15	$(15/60)=0.25$	25.0
2	25	$(25/60)=0.333$	41.7
3	20	$(20/60)=0.417$	33.3
Total	60	1.00	100



# Data Presentation –Categorical Variable

Pie Chart: Lists the categories and presents the percent or count of individuals who fall in each category.

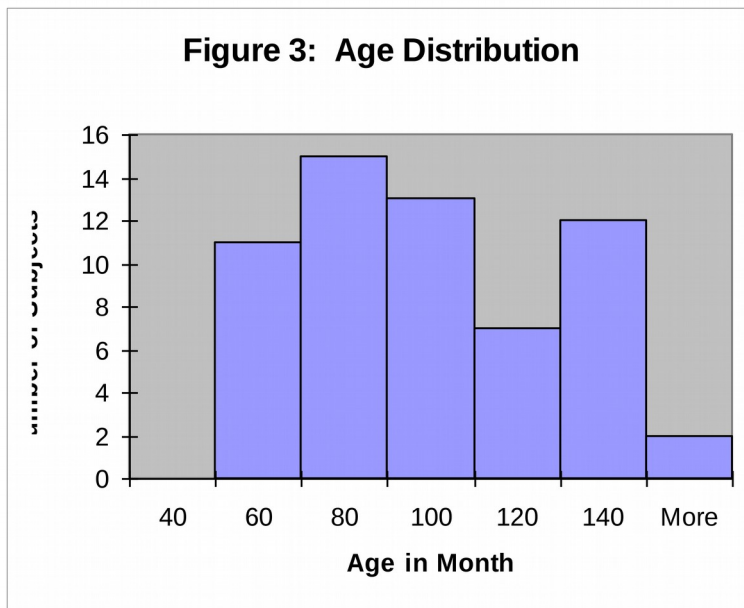
Figure 2: Pie Chart of Subjects in Treatment Groups



Treatment Group	Frequency	Proportion	Percent (%)
1	15	$(15/60)=0.25$	25.0
2	25	$(25/60)=0.333$	41.7
3	20	$(20/60)=0.417$	33.3
Total	60	1.00	100

# Graphical Presentation – Numerical Variable

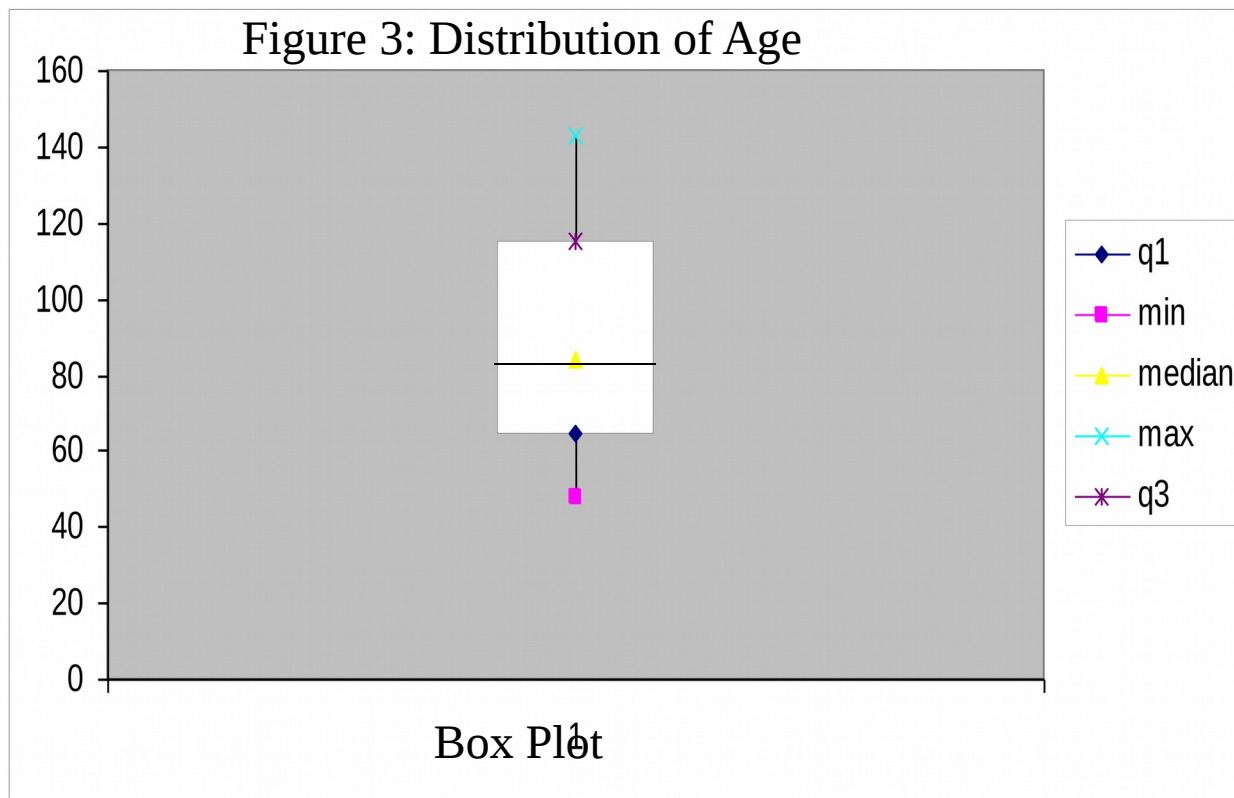
Histogram: Overall pattern can be described by its shape, center, and spread. The following age distribution is right skewed. The center lies between 80 to 100. No outliers.



Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60

# Graphical Presentation – Numerical Variable

Box-Plot: Describes the five-number summary



# Numerical Presentation

A fundamental concept in summary statistics is that of a *central value* for a set of observations and the extent to which the central value characterizes the whole set of data. Measures of central value such as the mean or median must be coupled with measures of data dispersion (e.g., average distance from the mean) to indicate how well the central value characterizes the data as a whole.

To understand how well a central value characterizes a set of observations, let us consider the following two sets of data:

A: 30, 50, 70

B: 40, 50, 60

The mean of both two data sets is 50. But, the distance of the observations from the mean in data set A is larger than in the data set B. Thus, the mean of data set B is a better representation of the data set than is the case for set A.

# Methods of Center Measurement

Center measurement is a summary measure of the overall level of a dataset

Commonly used methods are mean, median, mode, geometric mean etc.

Mean: Summing up all the observation and dividing by number of observations. Mean of 20, 30, 40 is  $(20+30+40)/3 = 30$ .

Notation : Let  $x_1, x_2, \dots, x_n$  are  $n$  observations of a variable  $x$ . Then the mean of this variable,

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

# Methods of Center Measurement

**Median:** The middle value in an ordered sequence of observations. That is, to find the median we need to order the data set and then find the middle value. In case of an even number of observations the average of the two middle most values is the median. For example, to find the median of {9, 3, 6, 7, 5}, we first sort the data giving {3, 5, 6, 7, 9}, then choose the middle value 6. If the number of observations is even, e.g., {9, 3, 6, 7, 5, 2}, then the median is the average of the two middle values from the sorted sequence, in this case,  $(5 + 6) / 2 = 5.5$ .

**Mode:** The value that is observed most frequently. The mode is undefined for sequences in which no observation is repeated.

# Mean or Median

The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income. For example mean of 20, 30, 40, and 990 is  $(20+30+40+990)/4 = 270$ . The median of these four observations is  $(30+40)/2 = 35$ . Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

# Methods of Variability Measurement

Variability (or dispersion) measures the amount of scatter in a dataset.

Commonly used methods: *range, variance, standard deviation, interquartile range, coefficient of variation etc.*

Range: The difference between the largest and the smallest observations. The range of 10, 5, 2, 100 is  $(100-2)=98$ . It's a crude measure of variability.



# Methods of Variability Measurement

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of the  $n$  observations  $x_1, x_2, \dots, x_n$  is

$$S^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$

Variance of 5, 7, 3? Mean is  $(5+7+3)/3 = 5$  and the variance is

$$\frac{(5 - 5)^2 + (3 - 5)^2 + (7 - 5)^2}{3 - 1} = 4$$

Standard Deviation: Square root of the variance. The standard deviation of the above example is 2.

# Methods of Variability Measurement

Quartiles: Data can be divided into four regions that cover the total range of observed values. Cut points for these regions are known as quartiles.

In notations, quartiles of a data is the  $((n+1)/4)q^{\text{th}}$  observation of the data, where  $q$  is the desired quartile and  $n$  is the number of observations of data.

The first quartile (Q1) is the first 25% of the data. The second quartile (Q2) is between the 25<sup>th</sup> and 50<sup>th</sup> percentage points in the data. The upper bound of Q2 is the median. The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and Q3 is the median of the second half of the ordered observations.

# Methods of Variability Measurement

In the following example  $Q1 = ((15+1)/4)1 = 4^{\text{th}}$  observation of the data. The 4<sup>th</sup> observation is 11. So Q1 is of this data is 11.

An example with 15 numbers

3 6 7 11 13 22 30 40 44 50 52 61 68 80 94

Q1            Q2            Q3

The first quartile is  $Q1=11$ . The second quartile is  $Q2=40$  (This is also the Median.)

The third quartile is  $Q3=61$ .

Inter-quartile Range: Difference between Q3 and Q1. Inter-quartile range of the previous example is  $61 - 40 = 21$ . The middle half of the ordered data lie between 40 and 61.

# Deciles and Percentiles

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25<sup>th</sup> percentile is the Q1, 50<sup>th</sup> percentile is the Median (Q2) and the 75<sup>th</sup> percentile of the data is Q3.

In notations, percentiles of a data is the  $((n+1)/100)p$  th observation of the data, where p is the desired percentile and n is the number of observations of data.

Coefficient of Variation: The standard deviation of data divided by it's mean. It is usually expressed in percent.

$$\text{Coefficient of Variation} = \frac{\sigma}{\bar{x}} \times 100$$

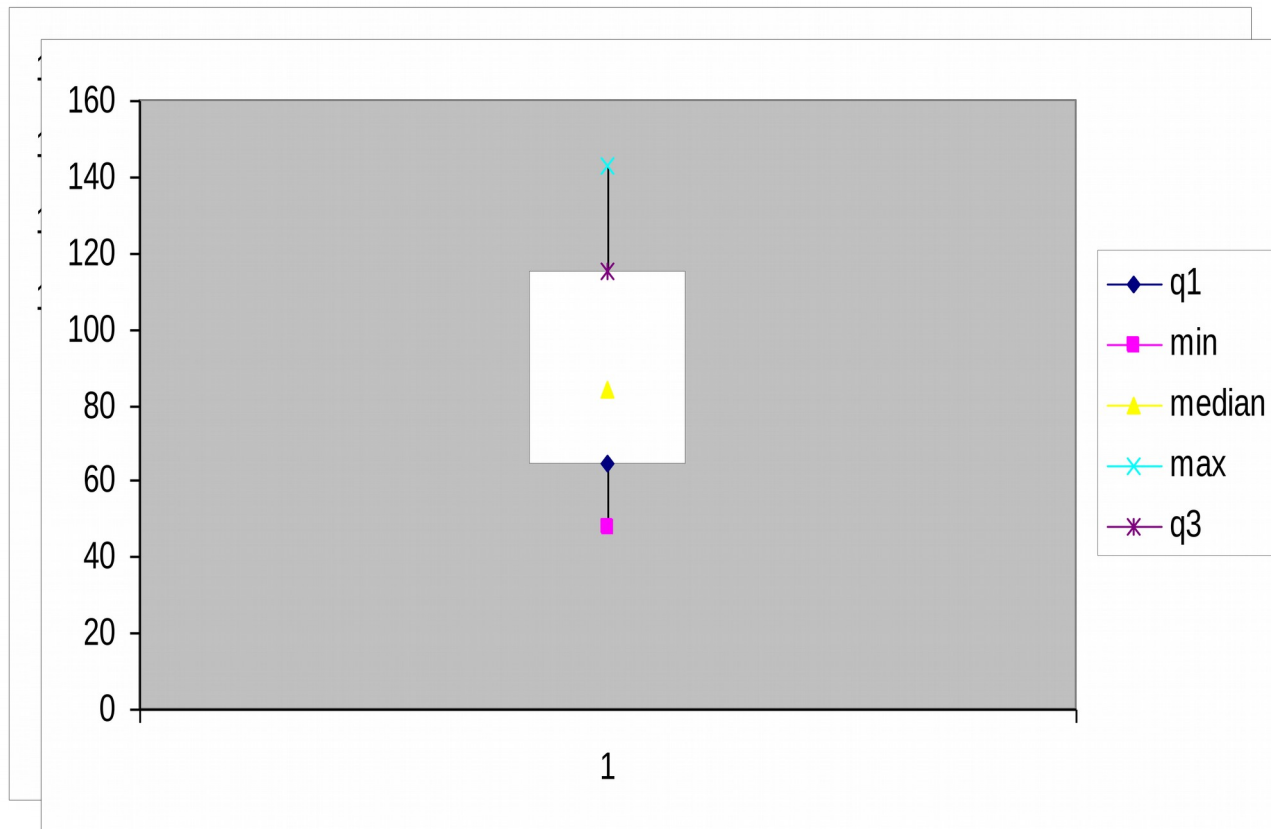
# Five Number Summary

Five Number Summary: The five number summary of a distribution consists of the smallest (Minimum) observation, the first quartile (Q1), The median(Q2), the third quartile, and the largest (Maximum) observation written in order from smallest to largest.

Box Plot: A box plot is a graph of the five number summary. The central box spans the quartiles. A line within the box marks the median. Lines extending above and below the box mark the smallest and the largest observations (i.e., the range). Outlying samples may be additionally plotted outside the range.

# Boxplot

Distribution of Age in Month



# Choosing a Summary

The five number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with extreme outliers. The mean and standard deviation are reasonable for symmetric distributions that are free of outliers.

In real life we can't always expect symmetry of the data. It's a common practice to include number of observations ( $n$ ), mean, median, standard deviation, and range as common for data summarization purpose. We can include other summary statistics like Q1, Q3, Coefficient of variation if it is considered to be important for describing data.

# Shape of Data

- Shape of data is measured by
  - Skewness
  - Kurtosis



# Skewness

- Measures asymmetry of data
  - Positive or right skewed: Longer right tail
  - Negative or left skewed: Longer left tail

Let  $x_1, x_2, \dots, x_n$  be  $n$  observations. Then,

$$\text{Skewness} = \frac{\sqrt{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

# Kurtosis

- Measures peakedness of the distribution of data. The kurtosis of normal distribution is 0.

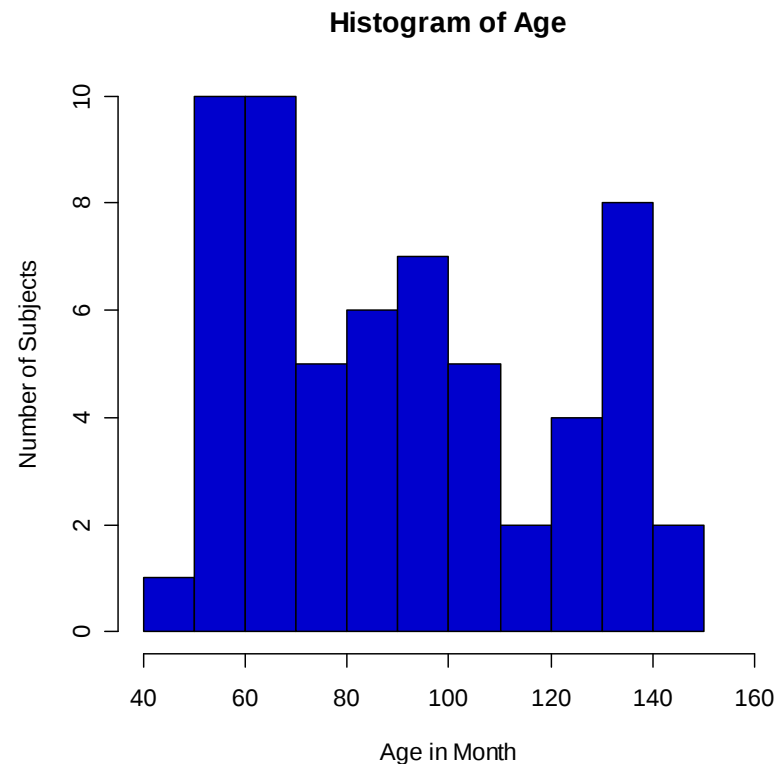
Let  $x_1, x_2, \dots, x_n$  be  $n$  observations. Then,

$$\text{Kurtosis} = \frac{n \sum_{i=1}^n (x_i - \bar{x})^4}{\left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} - 3$$

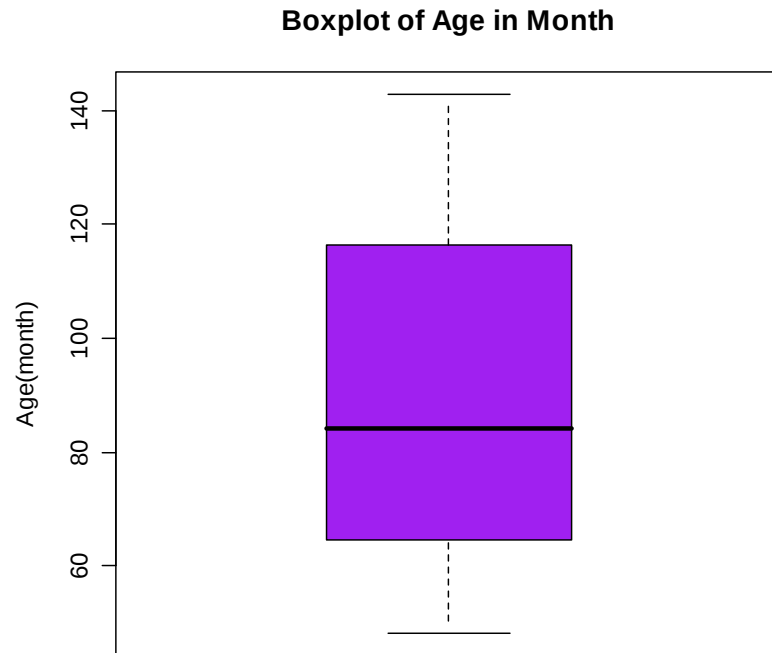
# Summary of the Variable 'Age' in the given data set

Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60

---



# Summary of the Variable 'Age' in the given data set



# Some applications

## 1. Control Charts

# Statistics is strictly linked to Quality Control



William Gossett  
(Guinness Brewery)



Walter Shewart  
(Bell Labs)

Student's t test publication: 1907

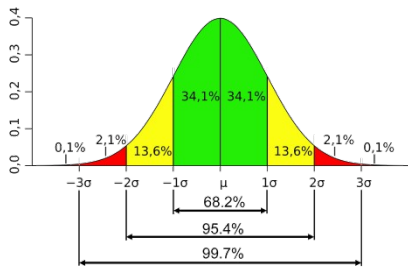
Introduction of control charts: 1920

Longitudinal evaluation of Scanner Performance for fMRI studies at 3T: a comparison of quality parameters across 8 years

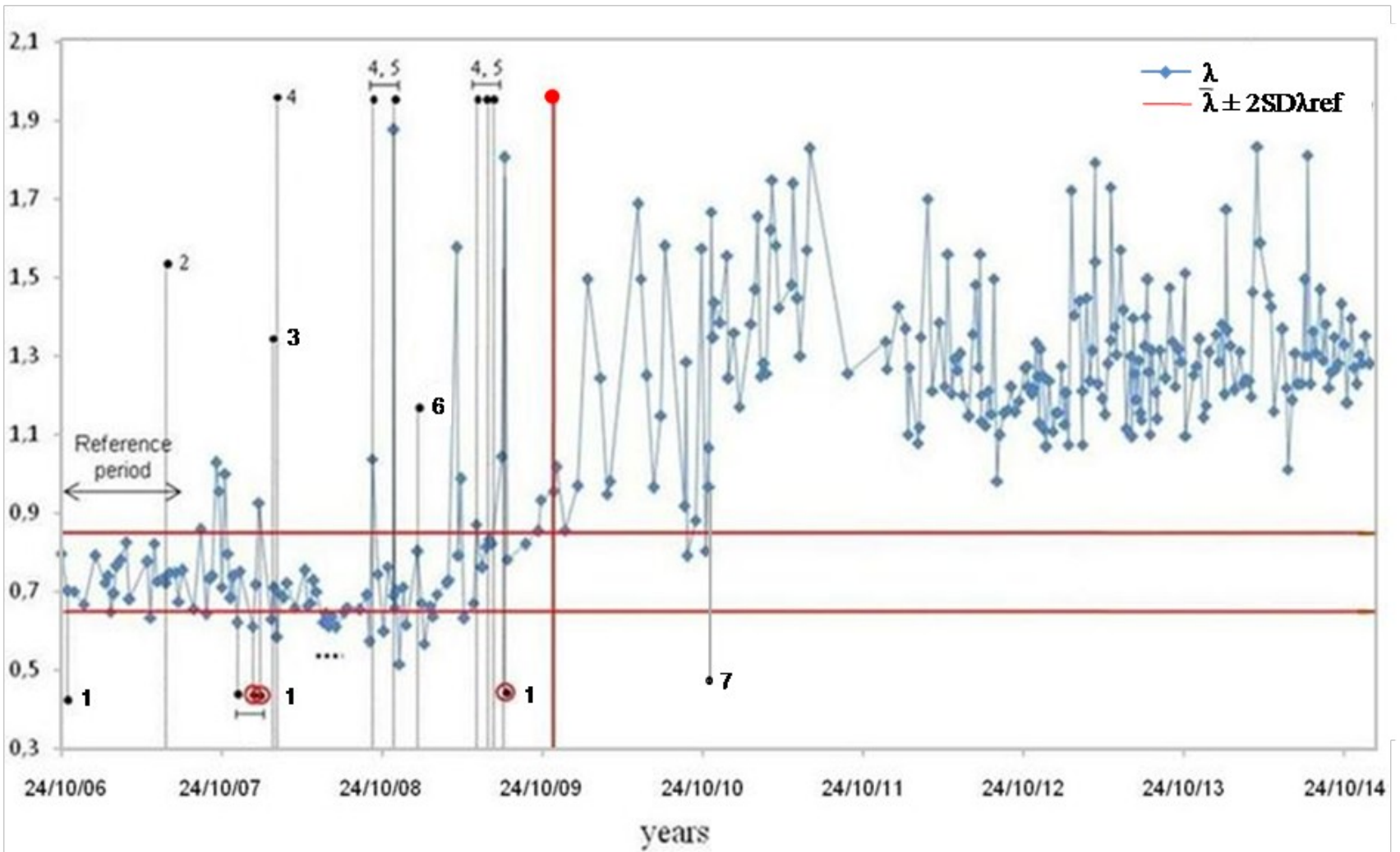
Elisa Tuzzi\*<sup>1</sup>, Fabrizio Fasano<sup>2</sup>, Valentina Brainovich<sup>4</sup>, Danilo Aragno<sup>3</sup>, Carlo Caltagirone<sup>1</sup>, Gisela E. Hagberg<sup>1,5,6</sup>

<sup>1</sup> Santa Lucia Foundation, I.R.C.C.S, Rome, (Italy); <sup>2</sup> Neuroscience Department, Parma University, Parma (Italy); <sup>3</sup> San Camillo-Forlanini Hospital, Rome, (Italy); <sup>4</sup> Grosseto Hospital U.S.L. 9; Grosseto, (Italy); <sup>5</sup> Biomedical Magnetic Resonance, University Hospital, Tübingen, (Germany); <sup>6</sup> Max-Planck Institute for Biological Cybernetics, High Field Magnetic Resonance, Tübingen, (Germany).

We assessed the parameter  $\lambda$  which represents a physical measure of the spatial degradation of the temporal SNR within each run due to fluctuations related to scanner instabilities.



Legend: <sup>1</sup> RF coil problems; <sup>2</sup> Receive path calibration failed; <sup>3</sup> Gradient water replacement with MNSO4; <sup>4</sup> Problems with external cooling system; <sup>5</sup> Problem with Helium pump; <sup>6</sup> Cold head replacement; <sup>7</sup> Gradient coil replacement; <sup>8</sup> Start of systemic cooling system problems.





# Control charts

A control chart consists of:

- Points representing a statistic (e.g., a mean, range, proportion) of measurements of a quality characteristic in samples taken from the process at different times (i.e., the data)
- The mean of this statistic using all the samples is calculated (e.g., the mean of the means, mean of the ranges, mean of the proportions)
- A centre line is drawn at the value of the mean of the statistic
- The standard error (e.g., standard deviation/sqrt(n) for the mean) of the statistic is also calculated using all the samples
- Upper and lower control limits (sometimes called "natural process limits") that indicate the threshold at which the process output is considered statistically 'unlikely' and are drawn typically at 3 standard errors from the centre line

1. Any point outside of the control limits (3 standard errors)



**Anomalies**

2. A Run of 7 Points all above or All below the central line - Stop the production

# 10.3 - Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value

In this example, two columns indicate the actual condition of the subjects, diseased or non-diseased. The rows indicate the results of the test, positive or negative.

Cell A contains true positives, subjects with the disease and positive test results. Cell D subjects do not have the disease and the test agrees.

A good test will have minimal numbers in cells B and C. Cell B identifies individuals without disease but for whom the test indicates 'disease'. These are false positives. Cell C has the false negatives.

If these results are from a population-based study, prevalence can be calculated as follows:

- **Prevalence of Disease**=  $T_{\text{disease}} / \text{Total} \times 100$

The population used for the study influences the prevalence calculation.

Sensitivity is the probability that a test will indicate 'disease' among those with the disease:

- **Sensitivity**:  $A / (A+C) \times 100$

Specificity is the fraction of those without disease who will have a negative test result:

- **Specificity**:  $D / (D+B) \times 100$

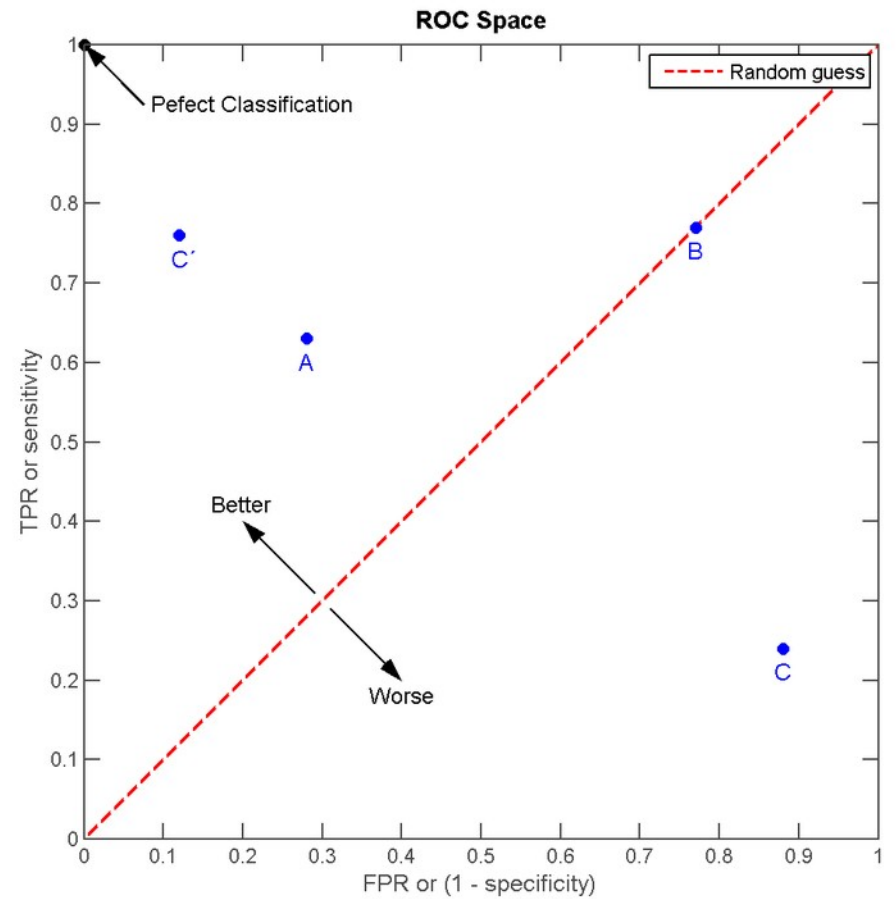
*Sensitivity and specificity are characteristics of the test. The population does not affect the results.*

A clinician and a patient have a different question: what is the chance that a person with a positive test truly has the disease? If the subject is in the first row in the table above, what is the probability of being in cell A as compared to cell B? A clinician calculates across the row as follows:

- **Positive Predictive Value**:  $A / (A+B) \times 100$
- **Negative Predictive Value**:  $D / (D+C) \times 100$

*Positive and negative predictive values are influenced by the prevalence of disease in the population that is being tested. If we test in a high prevalence setting, it is more likely that persons who test positive truly have disease than if the test is performed in a population with low prevalence..*

		Truth		
		Disease (number)	Non Disease (number)	Total (number)
Test Result	Positive (number)	<b>A</b> <i>(True Positive)</i>	<b>B</b> <i>(False Positive)</i>	$T_{\text{Test Positive}}$
	Negative (number)	<b>C</b> <i>(False Negative)</i>	<b>D</b> <i>(True Negative)</i>	$T_{\text{Test Negative}}$
		$T_{\text{Disease}}$	$T_{\text{Non Disease}}$	<b>Total</b>



Specificity = True Negative / (True Negative + False Positive)

Sensitivity = True Positive / (True Positive + False Negative)

Accuracy = (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative)

Some applications


2. Construction of a statistical index

# Construction of a statistical index

BIOLOGICAL AGRICULTURE & HORTICULTURE, 2018  
<https://doi.org/10.1080/01448765.2018.1434832>



## Assessing naturalness of arable weed communities: a new index applied to a case study in central Italy

E. Fanfarillo<sup>a</sup> , A. Kasperski<sup>b</sup>, A. Giuliani<sup>c</sup>, E. Cicinelli<sup>d</sup>, M. Latini<sup>a</sup> and G. Abbate<sup>a</sup>

<sup>a</sup>Department of Environmental Biology, "Sapienza" University of Rome, Rome, Italy; <sup>b</sup>Faculty of Biological Sciences, Department of Biotechnology, University of Zielona Góra, Zielona Góra, Poland; <sup>c</sup>Istituto Superiore di Sanità (ISS), Rome, Italy; <sup>d</sup>Department of Sciences, University of Roma Tre, Rome, Italy

# Construction of a statistical index

**Table 1.** Floristic descriptors used to characterize segetal species: binary options, their associated code, the information they provide on naturalness, and the bibliographic source data were taken from.

Floristic descriptor	Options	Code	Indicated naturalness	Bibliographic source
Life form	Perennial	P	Higher	Pignatti 2005;
	Annual	A	Lower	
Status according to native range	Native	N	Higher	Pignatti 2005;
	Exotic	E	Lower	
Introduction moment for exotic (E) species	Archaeophyte	a	Higher	Celesti-Grpow et al. 2010;
	Neophyte	n	Lower	
Edaphic preference	Non-nitrophilous	ni	Higher	Pignatti 2005
	Nitrophilous	NI	Lower	

# Construction of a statistical index

**Table 2.** Species types deriving from the 12 possible combinations of the binary options of the four descriptors and their associated weights, in a decreasing natural value order (codes as in Table 1).

<i>n</i>	Species type	Weight
1	P,N,-,ni	1
2	A,N,-,ni	1/2
3	P,N,-,NI	1/4
4	A,N,-,NI	1/8
5	A,E,a,ni	1/16
6	A,E,a,NI	1/32
7	P,E,a,ni	1/64
8	P,E,a,NI	1/128
9	A,E,n,ni	1/256
10	A,E,n,NI	1/512
11	P,E,n,ni	1/1024
12	P,E,n,NI	1/2048

# Construction of a statistical index

**Table 3.** The established four ALNI ranges, the respective naturalness classes, and the supposed management type. Abbreviations are reported.

ALNI value	Naturalness	Management type
$ALNI \geq 12$	High (H)	Extensive/traditional (E/T)
$8 \leq ALNI < 12$	Good (G)	Extensive/traditional (E/T)
$4 \leq ALNI < 8$	Average (A)	Intermediate/uncertain(I/U)
$0 \leq ALNI < 4$	Low (L)	Intensive (I)

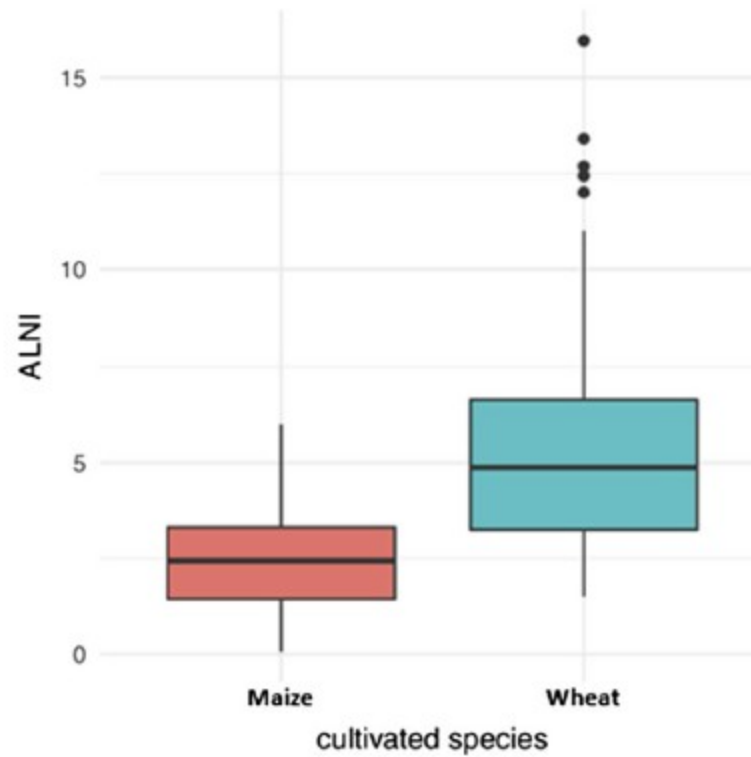


**Table 4.** ALNI value, related naturalness level and predicted management type of wheat weed communities.

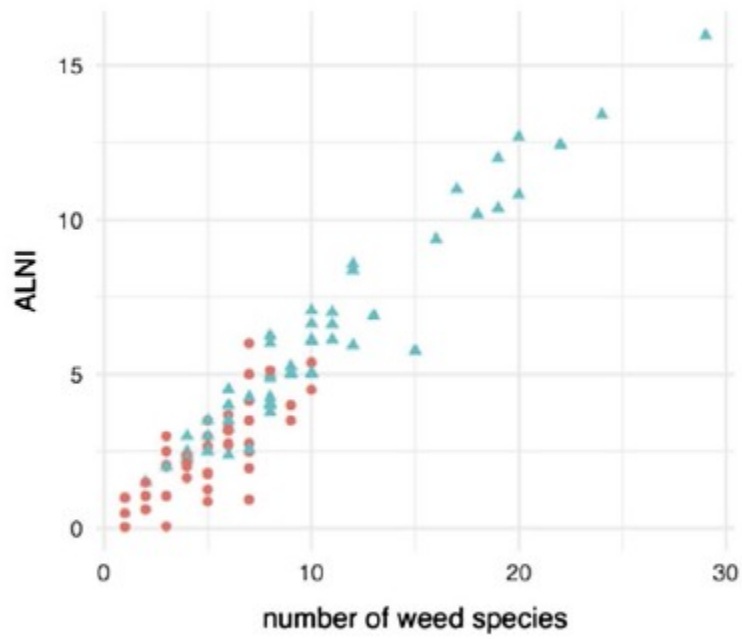
Field code	Field #	ALNI	Natur.	Manag.	Lat.	Long.	Elev.	Wheat height (cm)	N sp.
F60	1	15.97	H	E/T	41.758	13.272	562	60	29
F75	2	13.41	H	E/T	41.696	13.201	226	60	24
F20	3	12.69	H	E/T	41.701	13.328	283	70	20
CF1	4	12.44	H	E/T	41.759	12.924	243	100	22
F32	5	12.00	H	E/T	41.736	13.319	470	100	19
CF15	6	11.00	G	E/T	41.738	12.878	300	70	17
F27	7	10.81	G	E/T	41.732	13.288	502	110	20
F52	9	10.38	G	E/T	41.746	13.258	551	60	19
F38	8	10.19	G	E/T	41.744	13.314	410	100	18
F67	10	9.38	G	E/T	41.736	13.247	690	70	16
N8	11	8.56	G	E/T	42.154	12.644	30	80	12
CF28	12	8.38	G	E/T	41.532	13.438	150	65	12
N11	13	7.06	G	E/T	42.184	12.663	123	60	10
CF36	14	7.02	G	E/T	41.794	12.566	150	75	11
N2	15	6.89	A	I/U	42.178	12.670	65	80	13
N10	16	6.64	A	I/U	42.182	12.645	108	60	10
CF29	17	6.63	A	I/U	41.532	13.438	150	60	11
N7	18	6.25	A	I/U	42.186	12.655	100	70	8
CF30	19	6.13	A	I/U	41.532	13.438	150	60	10
CF23	23	6.13	A	I/U	41.744	12.920	300	80	11
CF31	20	6.06	A	I/U	41.532	13.438	100	70	10
CF26	21	6.00	A	I/U	41.665	13.192	150	115	8
CF6	22	5.94	A	I/U	42.017	12.474	80	55	12
CF3	24	5.75	A	I/U	42.083	12.405	250	80	15
CF45	25	5.25	A	I/U	41.740	12.852	300	80	9
CF25	26	5.06	A	I/U	41.731	12.937	350	75	10
CF12	28	5.06	A	I/U	41.750	12.871	300	115	9
N9	27	5.02	A	I/U	42.185	12.637	65	70	10
CF22	29	5.00	A	I/U	41.737	12.940	300	100	9
CF35	30	5.00	A	I/U	41.519	13.452	100	75	8
N6	31	4.88	A	I/U	42.172	12.654	70	60	8
CF33	32	4.50	A	I/U	41.513	13.410	150	70	6
CF16	33	4.25	A	I/U	41.750	12.924	300	60	8
CF19	34	4.25	A	I/U	41.744	12.926	300	60	7
CF2	35	4.06	A	I/U	42.033	12.426	100	55	8
CF24	38	4.06	A	I/U	41.738	12.942	300	120	8
CF42	36	4.00	A	I/U	41.761	12.516	100	65	8
CF13	37	4.00	A	I/U	41.747	12.869	300	130	8
CF37	39	4.00	A	I/U	41.768	12.528	100	60	6
CF44	60	3.77	L	I	41.739	12.859	300	75	8
CF14	40	3.50	L	I	41.763	12.874	350	85	6
CF39	41	3.50	L	I	41.760	12.523	100	80	6
CF34	42	3.50	L	I	41.519	13.452	100	75	5
CF32	45	3.50	L	I	41.507	13.439	200	65	6
N4	43	3.27	L	I	42.157	12.660	37	50	6
CF11	44	3.25	L	I	42.256	12.465	180	80	6
CF41	46	3.00	L	I	41.760	12.498	100	60	5
CF40	47	3.00	L	I	41.764	12.506	100	75	4
CF10	48	2.56	L	I	42.250	12.355	200	100	7
CF21	49	2.50	L	I	41.741	12.933	300	60	5
CF38	50	2.50	L	I	41.760	12.523	100	70	5
CF4	51	2.50	L	I	42.064	12.317	160	80	5
CF5	52	2.50	L	I	42.026	12.468	80	85	5
CF46	53	2.50	L	I	41.763	12.939	250	80	4
CF43	54	2.38	L	I	41.760	12.831	300	120	6
CF27	55	2.38	L	I	41.532	13.438	150	65	4
CF18	56	2.25	L	I	41.748	12.926	300	60	4
CF8	57	2.25	L	I	42.181	12.366	200	65	4
CF17	58	2.00	L	I	41.747	12.931	300	60	3
CF20	59	2.00	L	I	41.741	12.932	300	70	3
CF9	61	1.50	L	I	42.161	12.336	200	80	2

Notes: Natur. = Naturalness; Manag. = Management type; Lat. = Latitude; Long. = longitude; Elev. = Elevation in metres a.s.l.; N sp. = Number of species; other abbreviations as reported in Table 3.

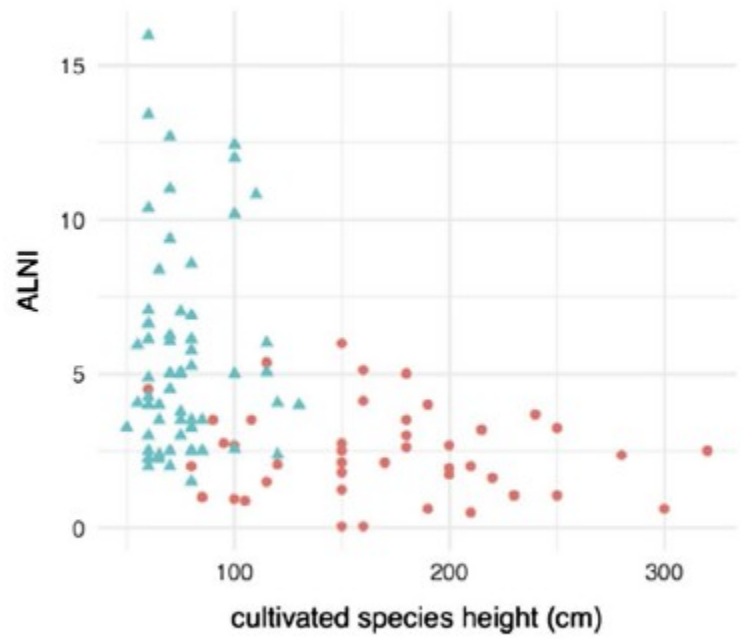
# Construction of a statistical index



**Figure 2.** Boxplot for the ALNI values of maize and wheat weed communities.  $p < 0.001$  according to both Student's t-test and Mann-Whitney U.



cultivated species ● maize ▲ wheat



cultivated species ● maize ▲ wheat

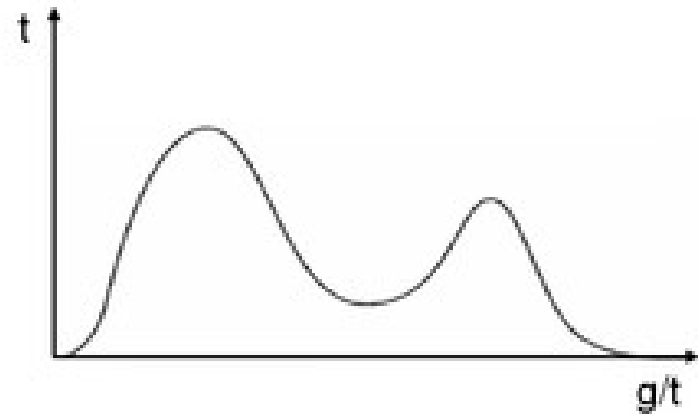
Some applications

3. Bimodality and cell fate

# Bimodality and cell fate

## Sarle's Bimodality Index

$$B = (\text{Skew}^2 + 1) / \text{Kurt}$$





The logic behind this index is that a bimodal distribution will have a very low kurtosis, an asymmetric character or both. All of these features increase B that in turn varies between 0 and 1.


# Bimodality and cell fate

RESEARCH ARTICLE

## Cell Fate Decision as High-Dimensional Critical State Transition

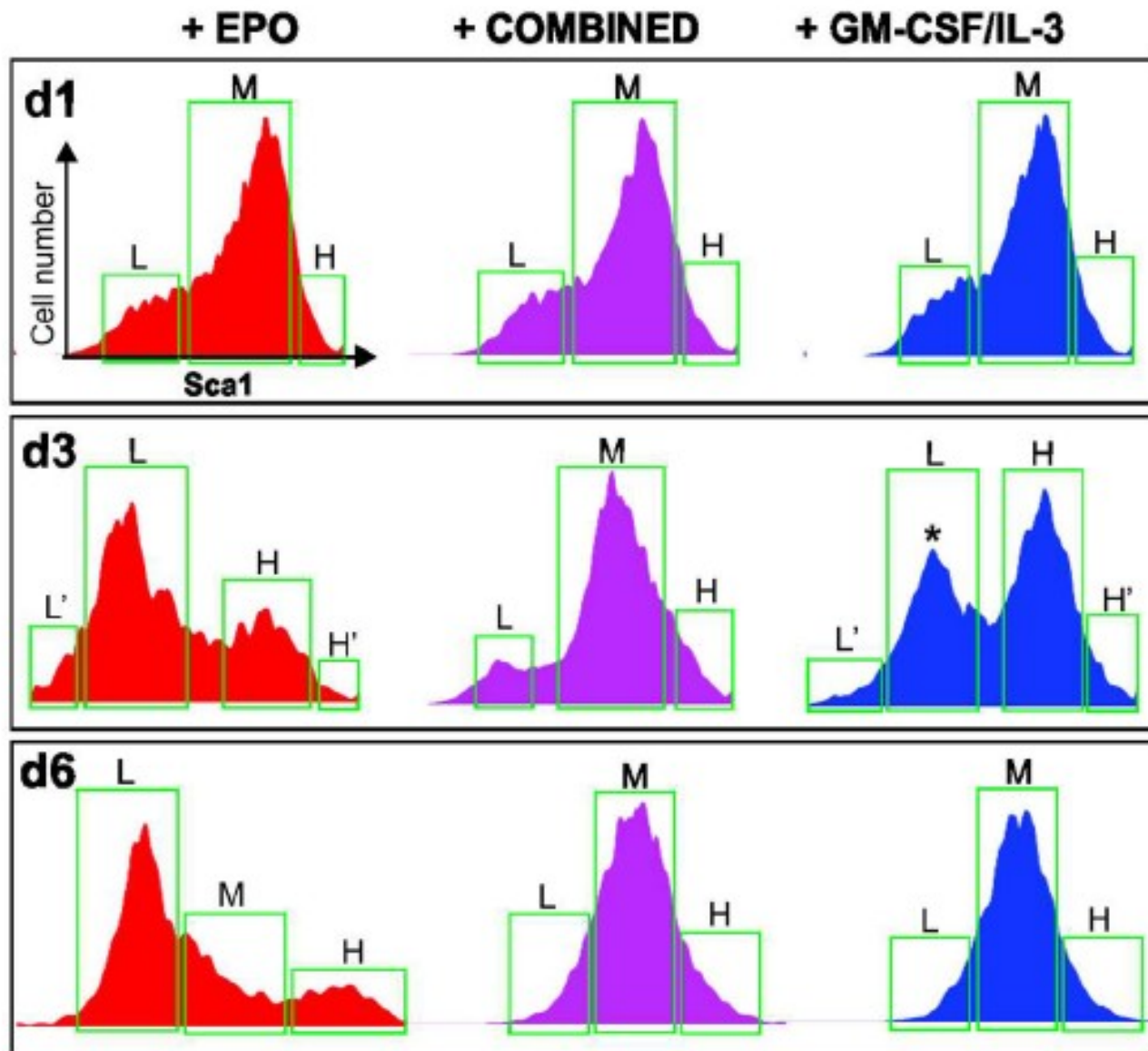
Mitra Mojtahedi<sup>1,2</sup>, Alexander Skupin<sup>2,3</sup>, Joseph Zhou<sup>2</sup>, Ivan G. Castaño<sup>1,4</sup>, Rebecca Y. Y. Leong-Quong<sup>1</sup>, Hannah Chang<sup>5</sup>, Kalliopi Trachana<sup>2</sup>, Alessandro Giuliani<sup>6</sup>, Sui Huang<sup>1,2\*</sup>

1 Department of Biological Sciences, University of Calgary, Calgary, Alberta, Canada, 2 Institute for Systems Biology, Seattle, Washington, United States of America, 3 Luxembourg Centre for Systems Biomedicine, Esch-sur Alzette, Luxembourg, 4 Corporación Parque Explora, Department of innovation and design, Medellin, Colombia, 5 5AM Ventures, Menlo Park, California, United States of America, 6 Environment and Health Department, Istituto Superiore di Sanità, Roma, Italy

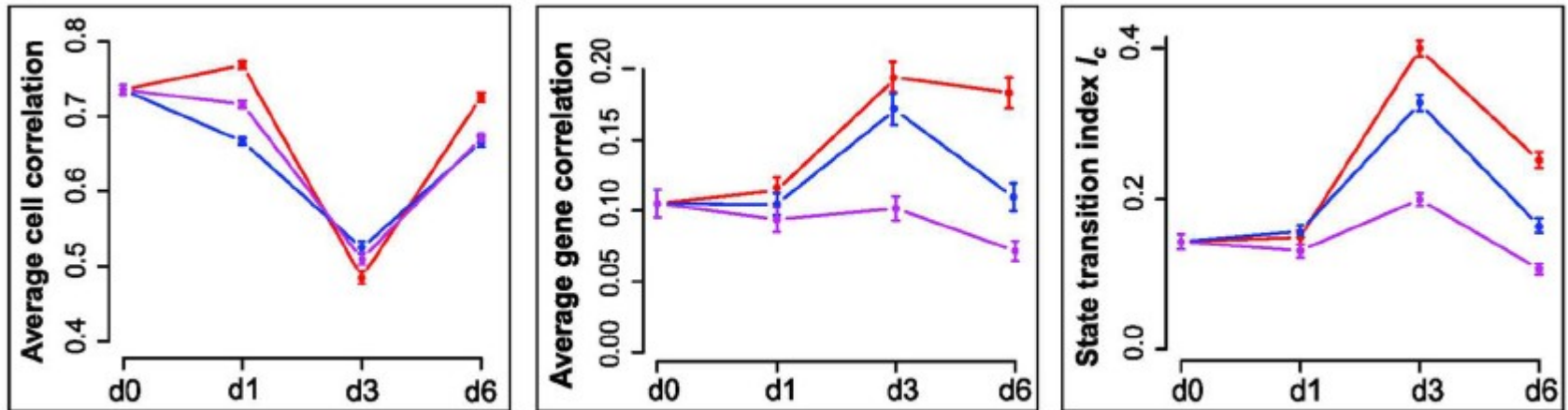
 These authors contributed equally to this work.

\* [sui.huang@systemsbiology.org](mailto:sui.huang@systemsbiology.org)

# Bimodality and cell fate



# Bimodality and cell fate



$$I_c(t) = \frac{\langle |R(\mathbf{g}_i, \mathbf{g}_j)| \rangle}{\langle R(S^k, S^l) \rangle},$$