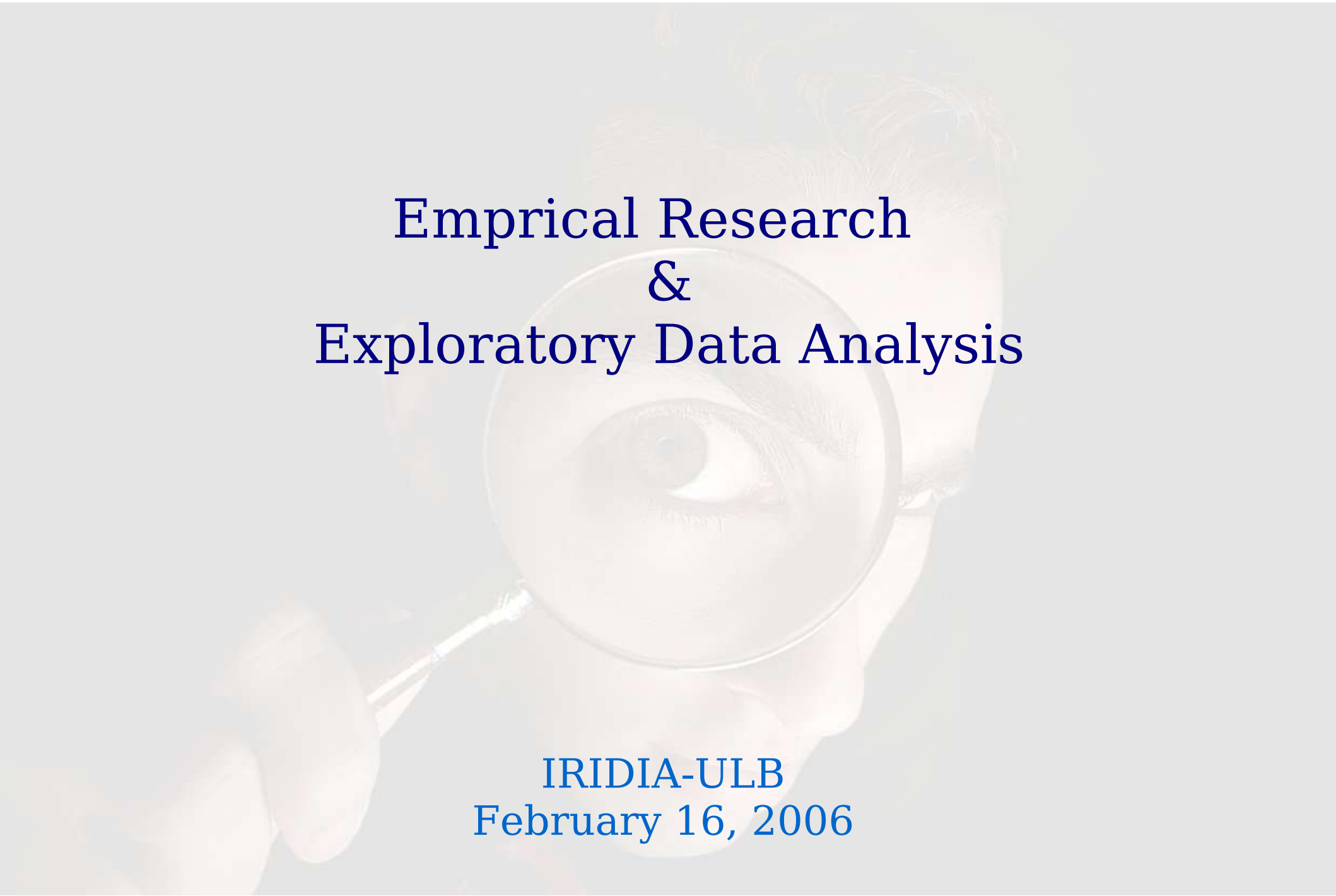


Emprical Research
&
Exploratory Data Analysis

IRIDIA-ULB
February 16, 2006

Outline
First Part

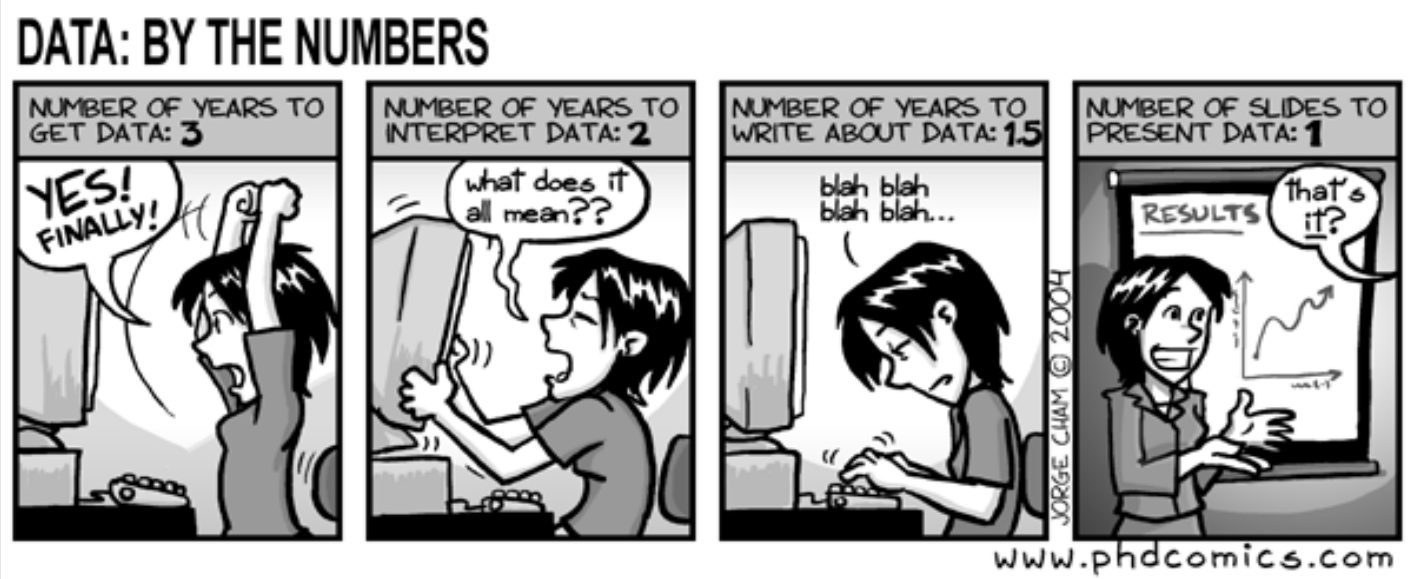
1)Empirical Research (from a CS perspective)

2)Computational entities as objects of empirical studies

3)Empirical generalization strategy

4)Kinds of empirical studies

Empirical Research Methods in Computer Science



Empirical research relies on direct or indirect observations to develop models that provide causal explanations to the phenomenon studied. It can be a tool for induction, i.e., deriving theories from observations.

Experiments can have two goals : Theory testing or exploration.

Exploratory experiments are part of what is called **exploratory data analysis**; whereas confirmatory experiments are part of what is known by **hypothesis tesing**.

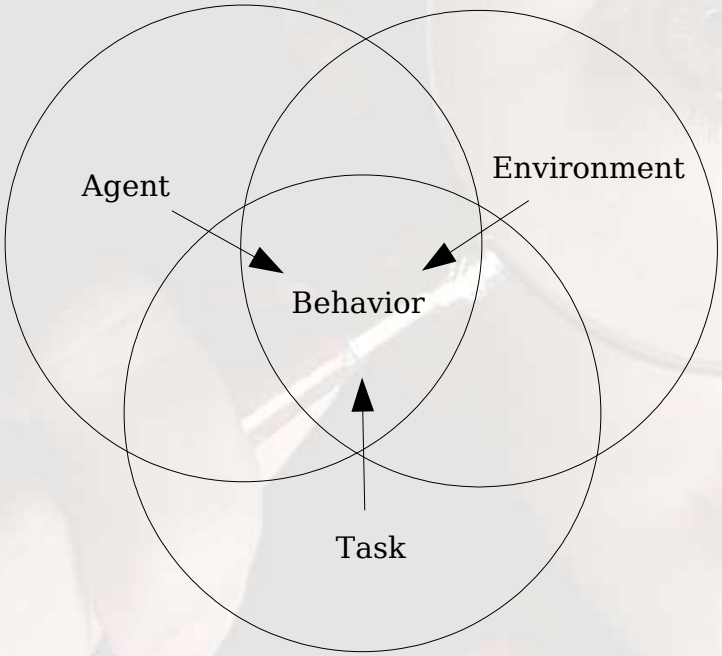
From a CS perspective, we want to characterize the behavior of computational entities (agents) for which we do not have a theoretically-derived model.

Computational entities as objects of empirical studies

Agents perform tasks under specific operating conditions. A slight change on any of these could result in a completely different outcome.

Being the author or creator of an agent is no guarantee that we will be able to predict how it will behave under certain circumstances. We need to actually let them run find it out.

If we are just curious then these runs are exploratory. If we have a hypothesis about the interaction of the components of our experiment (i.e., our agent, the environment and the task) then we are testing our hypothesis.



Basic questions:

Agent's structure change

Agent's task change → Behavior ?

Agent's environment change

Empirical generalization strategy

According to Cohen, developing theories from empirical studies involves the following steps:

0. **Agent design and construction**. Actually building the computational entity that we will study;

1. **Feature extraction.** Identify the features on the agent, its environment and tasks that may affect the behavior of interest;

2. **Modeling.** Develop a causal model that explains how these features affect the target behavior;

3. **Evaluation.** Test the devised model to see whether it accurately predicts the behavior exhibited by the agent.

4. **Generalization.** Generalize the model that accurately predicts the behavior of interest to other agents, tasks and environments and test whether this model accurately predicts the behavior of this larger set of agents, tasks and environments.

Kinds of empirical studies

Four classes of empirical studies can be identified:

1. **Exploratory studies.** These yield hypothesis that are later tested in observation or manipulation experiments. Usually they involve the collection of lots of data and their analysis to find regularities.

2. **Assessment studies.** They establish baselines and ranges and other assessments of the behaviors of a system or its environment.

3. **Manipulation experiments.** They test hypothesis avout causal influences of factors by manipulating them and noting effects on one or more measured variables.

4. **Observation experiments.** They disclose the effects of factors on measured variables by observing associations between the levels of the factors and values of the variables. These are also known as natural or quasi-experimental experiments.

Outline
Second Part

1)Data

2)Causal models

3)Visualizing univariate and joint distributions of data

4)Time series

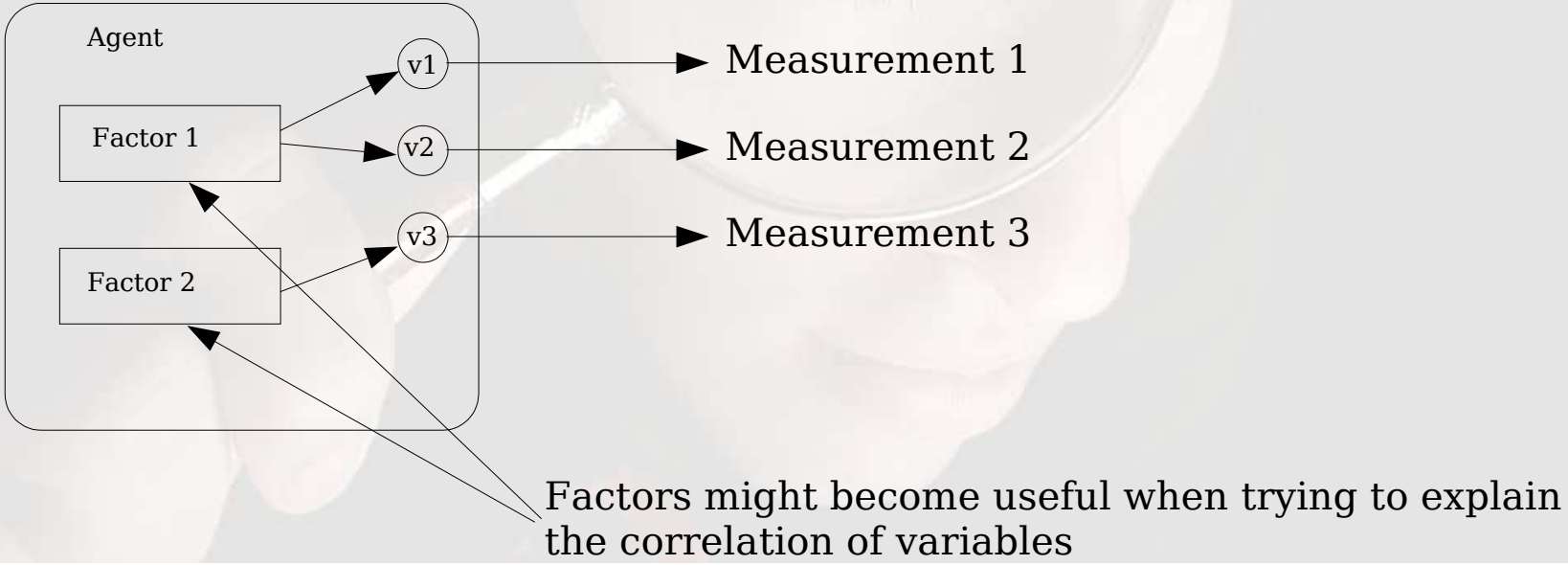
Data

A sample is a collection of individuals, **real** individuals. From a data analysis perspective, a sample is a collection of **measurements** on individuals.

For example, let's say that our sample will be the set of some variants of the PSO algorithm. We want to find out which one is the top performer variant. Our data sample could be the number of times the benchmark function is called by each of these variants before reaching the optimum with a certain precision.

The number of function evaluations (fes) is a function that maps individuals to data scales:

fes(PSOv1.0) = 3200



Data

Data classification is a crucial task to perform in the visualization process, as the type and structure of the data define the set of graphical mappings that can be performed on that data.

Data can have different scales. There are four common data scales:

1. **Nominal Data**
    * classification data, e.g. m/f
    * no ordering, e.g. it makes no sense to state that M > F
    * arbitrary labels, e.g., m/f, 0/1, etc
2. **Ordinal Data**
    * ordered but differences between values are not important
    * e.g., political parties on left to right spectrum given labels 0, 1, 2
    * e.g., Likert scales, rank on a scale of 1..5 your degree of satisfaction
    * e.g., restaurant ratings
3. **Interval Data**
    * ordered, constant scale, but no natural zero
    * differences make sense, but ratios do not (e.g., 30°-20°=20°-10°, but 20°/10° is not twice as hot!
    * e.g., temperature (C,F), dates
4. **Ratio Data**
    * ordered, constant scale, natural zero
    * e.g., height, weight, age, length

Data

Only certain operations can be performed on certain scales of measurement. The following list summarizes which operations are legitimate for each scale. Note that you can always apply operations from a 'lesser scale' to any particular data, e.g. you may apply nominal, ordinal, or interval operations to an interval scaled datum.

* **Nominal Scale.** You are only allowed to examine if a nominal scale datum is equal to some particular value or to count the number of occurrences of each value. For example, gender is a nominal scale variable. You can examine if the gender of a person is F or to count the number of males in a sample.

* **Ordinal Scale.** You are also allowed to examine if an ordinal scale datum is less than or greater than another value. Hence, you can 'rank' ordinal data, but you cannot 'quantify' differences between two ordinal values. For example, preference scores, e.g. ratings of eating establishments where 10=good, 1=poor, but the difference between an establishment with a 10 ranking and an 8 ranking can't be quantified.

* **Interval Scale.** You are also allowed to quantify the difference between two interval scale values but there is no natural zero. For example, temperature scales are interval data with 25C warmer than 20C and a 5C difference has some physical meaning. Note that 0C is arbitrary, so that it does not make sense to say that 20C is twice as hot as 10C.

* **Ratio Scale.** You are also allowed to take ratios among ratio scaled variables. Physical measurements of height, weight, length are typically ratio variables. It is now meaningful to say that 10 m is twice as long as 5 m. This ratio hold true regardless of which scale the object is being measured in (e.g. meters or yards). This is because there is a natural zero.

Causal Models

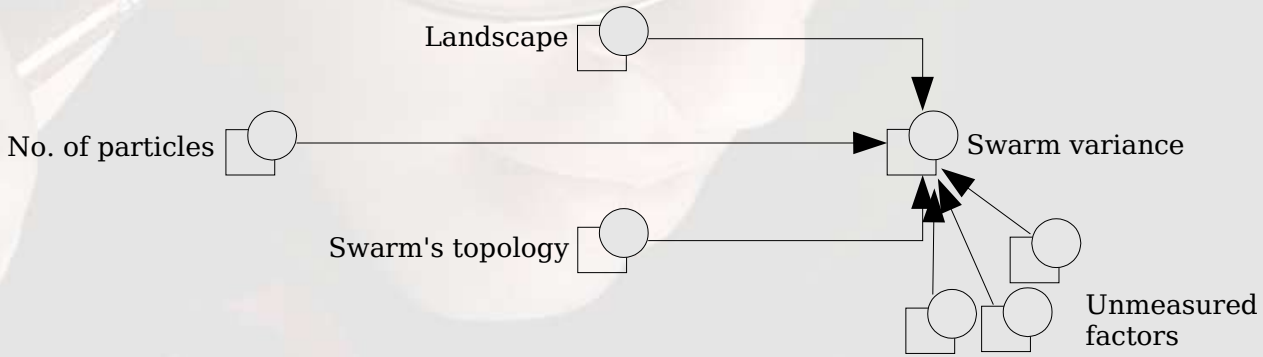
To collect data is the very first step in any exploratory data analysis. What we want to do is to explain the relationships among factors to predict the outputs of our agent.

The purpose of an exploratory data analysis is to develop and refine a causal model of data values.

Variables or factors might influence other variables or factors directly or indirectly. In a causal model we want to make these influences explicit. Causal models can even be sketched before any actual run and may help us identify what to measure.

For example, consider a subproblem of the PSO algorithm: How fast information about good positions (i.e., good solutions) spreads through the population.

In PSO, particles (solutions to an optimization problem) exchange information with their topological neighbors. The swarm's topology is set before the execution and remains fixed. Suppose that we measure how the swarm's population converges to a local optima (e.g, measuring the variance of the whole swarm). A possible causal model could be the following:



Visualizing univariate distributions

There are many ways of analizing data. A good practice is to look data in several ways. This will help us identify more/better patterns in data.

To look a univariate distribution you could use

a) a frequency histogram, which plots the relative frequencies of values in a distribution;
b) basic statistic measures.

Histograms.

General guidelines : Try to use different bin sizes since this affects the detail in the frequency histogram. We may draw erroneous conclusions about the distribution of the data or we could miss some important relationships among variables and/or factors.

Basic univariate statistics.

Sample Size. The number of data items in a sample.

Mean. The average value of a sample.

Median. If the values are sorted in nondecreasing order, the median is the value that splits the distribution in half.

Mode. The most common value in a distribution.

Visualizing univariate distributions

Skew. In a skewed distribution most of the data are at one end of the distribution. In skewed distributions, the median is preferred as a measure of central tendency. Another option is the trimmed mean.

Maximum, minimum and range. The maximum value, the minimum value of a distribution, and the difference between them respectively.

Interquartile range. Starting from a sorted distribution, divide the distribution into four groups each containing the same number of elements. Each group is a quartile. The difference between the highest value in the third quartile and the lowest value in the second quartile is the inter-quartile range. It is robust against outliers.

S.D. and variance. You already know it ;)

Variation coefficient. If s_x is the standard deviation of a set of samples x_i and x^ its mean, then V=(s_x)/(x^).

Using only means and S.D. or variances to analize our data, can lead us to wrong conclusions because of their sensitivity to outliers!

Visualizing joint distributions

Joint distributions help us discover whether a variable influences another or not.

Joint distributions of categorical and ordinal data.

Contingency tables.

A contingency table or cross-classification table represents the joint distribution of samples data according to their classification by two or more variables. Each cell in a contingency table corresponds to a unique combination of values for these variables. A pattern in a contingency table is often clearer if counts are expressed as proportions of the row marginal counts, also called row margins. One can also divide by column marginal counts.

Example :

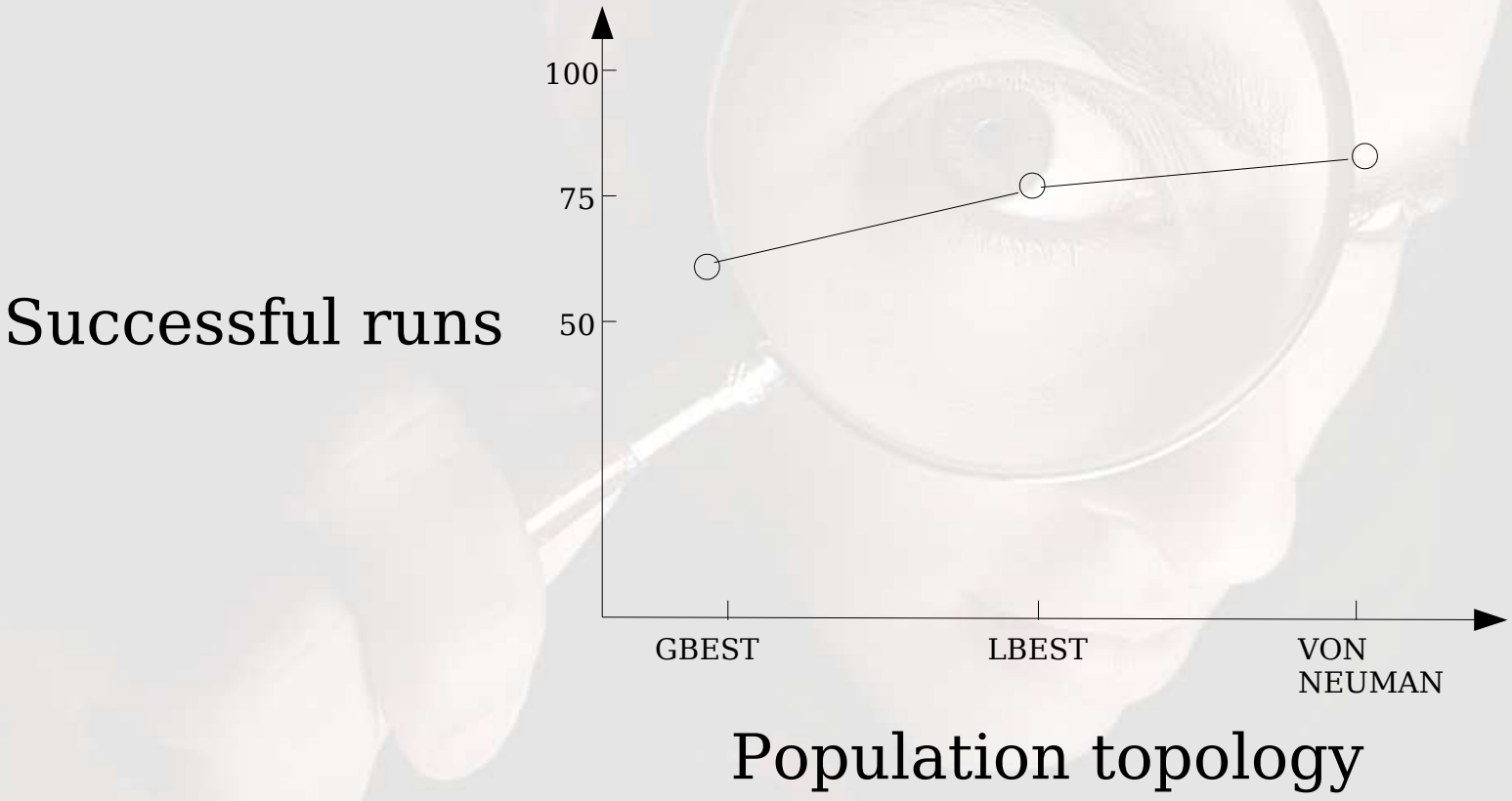
Suppose we have the following data:

Population topology	Successful runs	Unsuccessful runs	Total
GBEST	67	33	100
LBEST-2	76	24	100
VON NEUMAN	81	19	100
Total	224	76	300

Visualizing joint distributions

We can visuaize this data by letting the x axis to represent one of the variables (independent variables preferably). On the y axis we will plot the cell proportions with respect to the marginal frequencies of the variable on the x axis.

Following the example:



Visualizing joint distributions

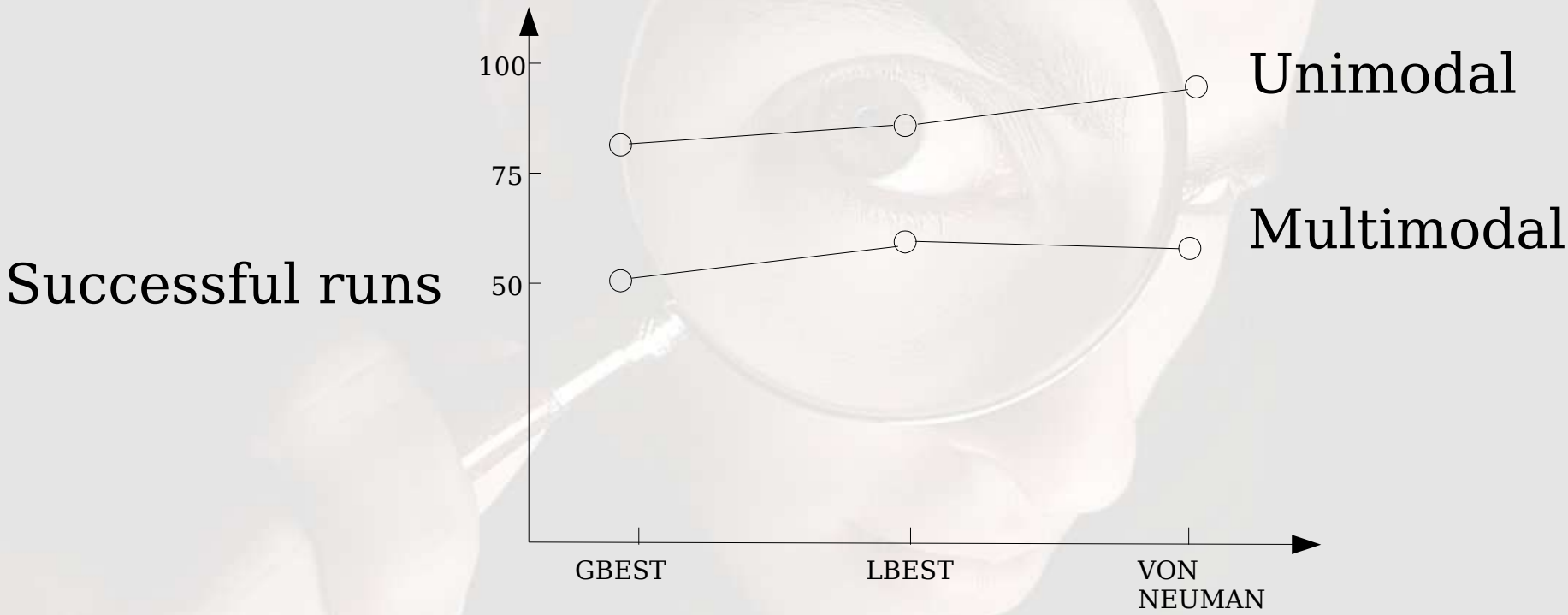
What if we now consider some features of the fitness landscape? (more than 2 variables)

The contingency table then might look like this:

Type of landscape	Population topology	Successful runs	Unsuccessful runs	Total
Unimodal	GBEST	45 (81)	11(19)	55
	LBEST-2	50 (83)	10(17)	60
	VON NEUMAN	60 (89)	7(11)	67
Multimodal	GBEST	22(50)	22(50)	44
	LBEST-2	26(65)	14(35)	40
	VON NEUMAN	21(63)	12(37)	33
	Total	224	76	300

Visualizing joint distributions

Three variables plot



Statistics for joint distributions of categorical values

The chi-squared statistic χ^2 summarizes the degree of dependence that holds between row and column variables in contingency tables.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_e is the expected frequency of a joint event assuming independence of the row and column variables.

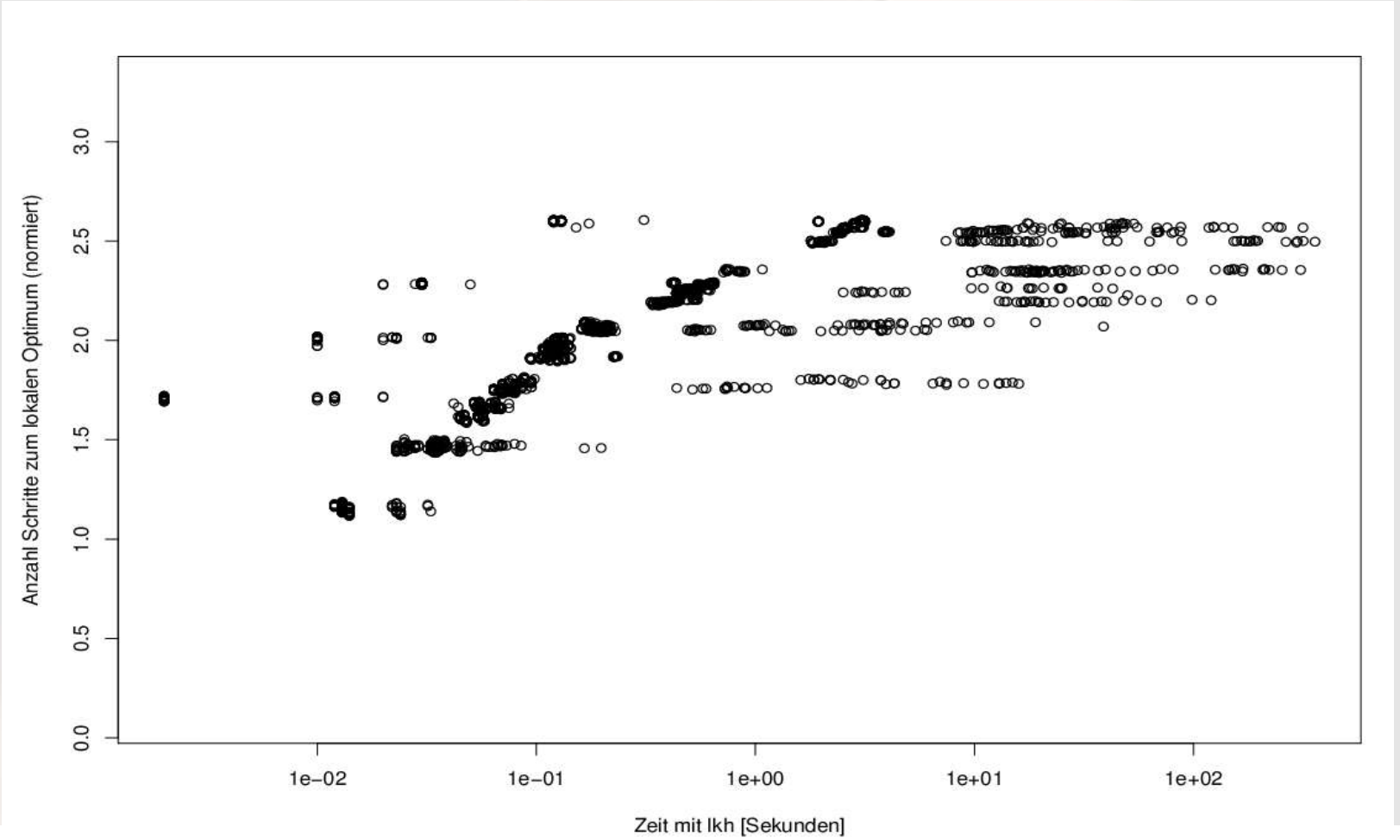
f_o Is the observed frequency of a joint event.

Therefore, larger values of χ^2 suggest that the row and column variables are not independent.

Visualizing joint distributions of two continuous variables

A joint distribution of two continuous variables is a distribution of pairs of measurements, each pair representing one individual.

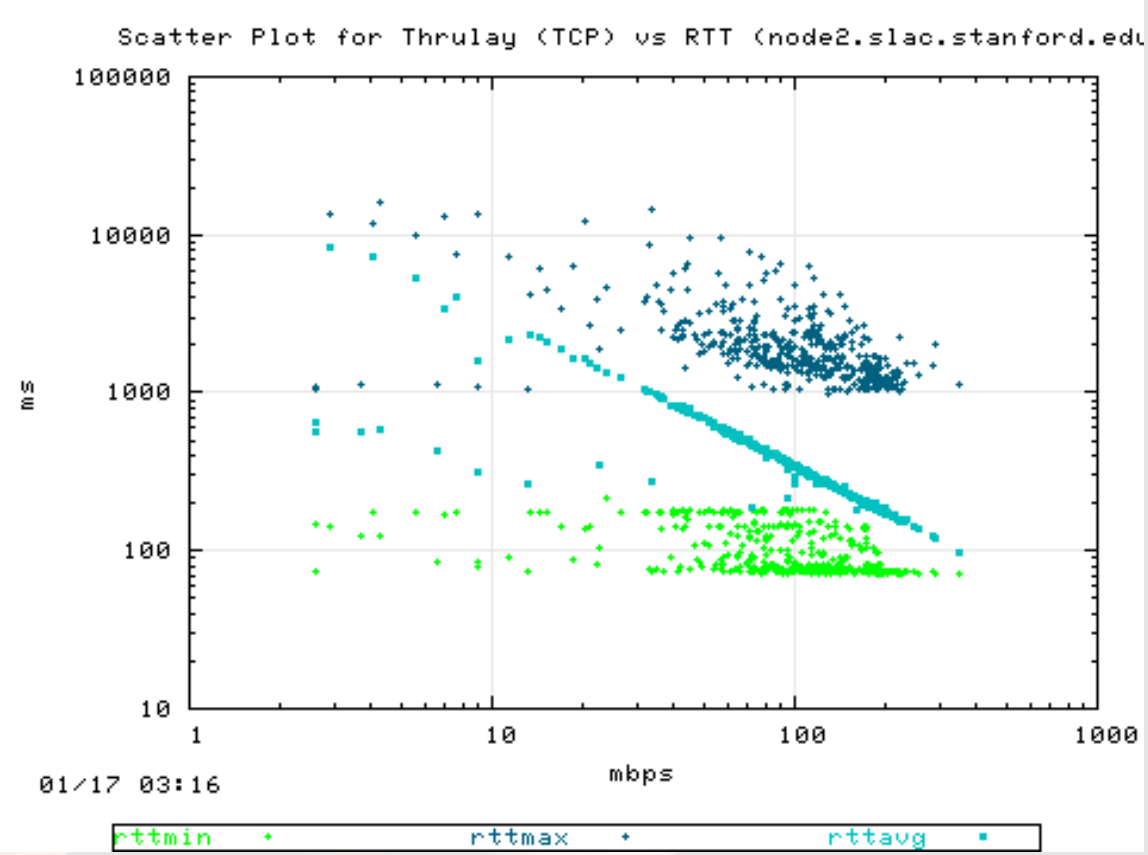
If these pairs are treated as coordinates, they can be plotted in two dimensions. These plots are known as scatterplots



Visualizing joint distributions of two continuous variables

A useful technique to uncover possible causal reationships in scatterplots is called point coloring.

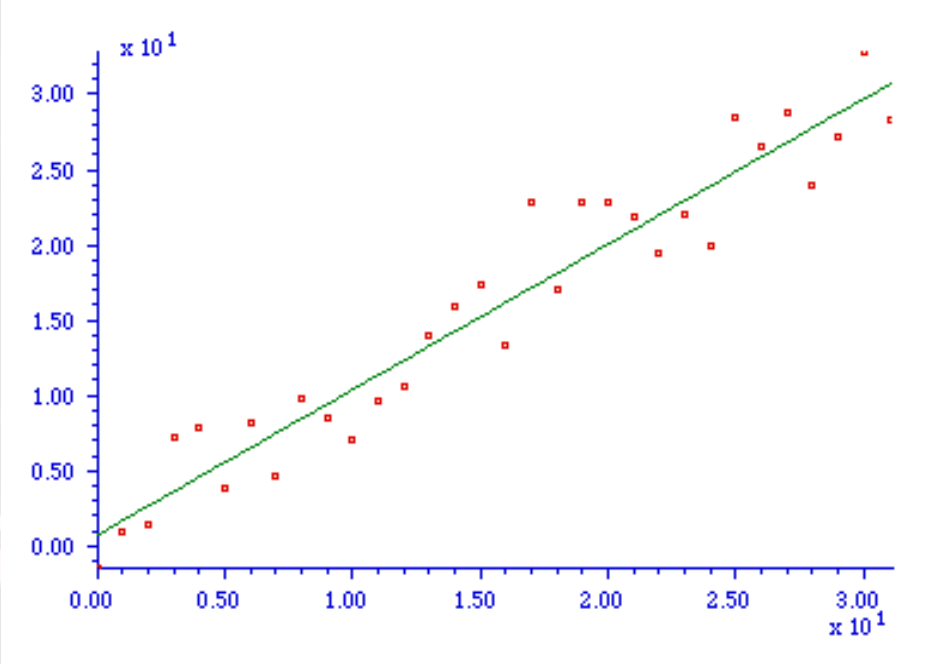
This technique allows the data analyst see three variables in a two-dimensional scatterplot and it sometimes explains patterns.



Visualizing joint distributions of two continuous variables

Sometimes we may want to have a compact and interpretable functional representation of data.

To do this, one might find useful to fit lines to data in scatterplots. This is not a trivial task and having knowledge about how data was generated might help to determine the best way to do it.



Statistics for joint distributions of two continuous variables

A statistic that captures the notion of association is the sample covariance of two variables.

$$cov(x,y)=\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}$$

The covariance measures linear association, so it will take large values (positive or negative) when the anaized data can be approximated by a straight line.

To standardize covariances and make them comparable, we have to divide it by the product of the standard deviations of x and y. This standardized covariance is known as the Pearson´s correlation coefficient.

$$r_{XY}=\frac{cov(x,y)}{s_x s_y}$$

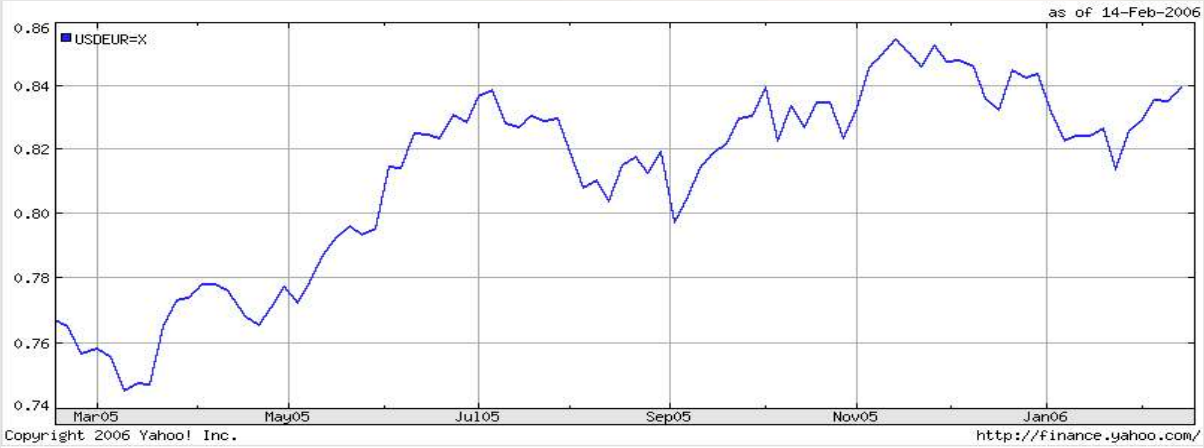
A low correlation value does not mean data are not associated, it just means that they are not linearly associated.

The Pearson´s correlation coefficient is sensitive to outliers. Alternatives are Spearman´s rank correlation and Kendal´s tau.

Time series

A time series is a joint distribution where one of the variables is time.

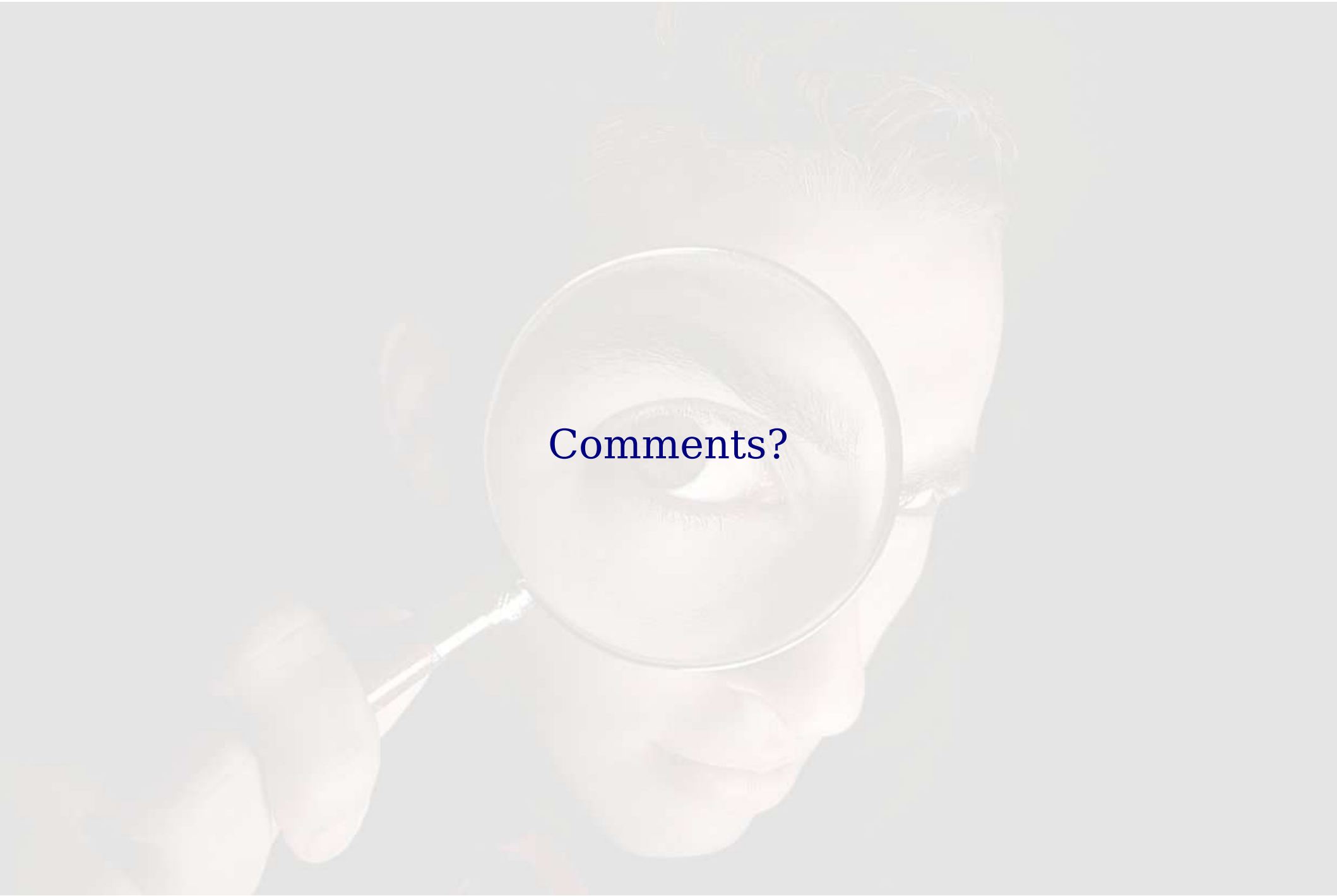
Since scatterplots of time series are diffcult to read, a smoothed lineplot is used instead.



In terms of statistics, a positive correlation between a variable and time, indicates a that this variable increases over time. Likewise, if the correlation is negative, the variable decreases over time. A way to compare two time series and see whether one predicts the other is to compute a cross-correlation with some lag. This is done by computing the correlation of two variables at a given time.

The correlation between two time series may be large only because they have similar trends.

Autorrelation is an statistic that allows us to determine the predictive power of a time series.


Comments?