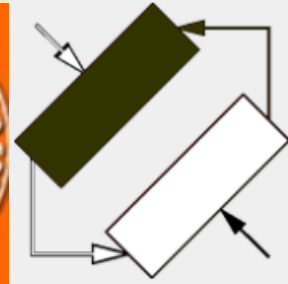




www.Mihail.Motzev.com



The International Simulation and Gaming Association



Summer School On Modelling And Complex Systems '2023:

**“STATISTICAL LEARNING
NETWORKS”**

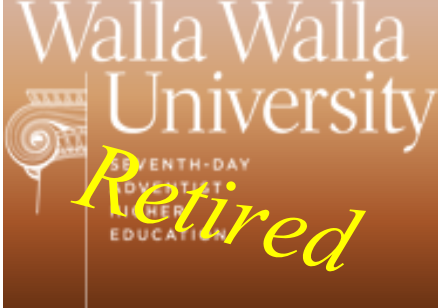


**Mihail
Motzev**

Ph.D, M.Sc, P.D.D
(MRMotzev@yahoo.com)

Proud Nerd (Zubar) Generation 1





The Best Place to Retire - for Me ...



About WWUAcademicsLife at WWUResourcesAttend WWU

Walla Walla University

NAVIGATION

NewsArchiveName Change

RELATED LINKS

Inside WWU Newsletter

WHAT'S HAPPENING ▶

APIC Chinese New Year Party

Quicklinks


BEST COLLEGES USNews REGIONAL UNIVERSITIES WEST 2012

HOME > ABOUT WWU > GENERAL INFORMATION > NEWS >

Business Professor Designs Game to Help Industry Professors

Motzev Has Shared Research Results at Worldwide Conferences

By: Becky St. Clair



Mihail Motzev, School of Business professor

Who says professionals can't have fun? Mihail Motzev, School of Business at Walla Walla University, spent the summer of 2022 in Romania, where he was awarded a research grant to develop a game for businesspeople. His latest research, "Intelligent Techniques in Simulation and Management Games: A Hybrid Approach: Multi-Agent Simulation in a Hybrid Building" was presented at the ISAGA/IFIP (International Federation for Information Systems) conference in Romania, where he was awarded a research grant.

It's one of my favorite games and the present

As a member of the International Simulation and Gaming Association, Motzev has shared his research at many conventions, most recently in Romania. He was present at the ISAGA/IFIP (International Federation for Information Systems) conference in Romania, where he was awarded a research grant.

HOME

RETIREMENT

PERSONAL FINANCE

CAREERS

INVESTING

BUSINESS & ECONOMICS

On Retirement Blog

The Best Life Blog

Planning to Retire Blog

Best Places to Retire

Don't Run Out of Money in Retirement

If you have a \$500,000 portfolio, download the guide by Forbes columnist Ken Fisher's firm. Even if you're not sure how to start rebuilding your portfolio or who to turn to for help, this must-read guide includes research and analysis you can use right now. Don't miss it! [Click Here to Download Your Guide!](#)

FISHER INVESTMENTS

A photograph of a large, leafy tree in front of a white house with a red roof. The house has a small porch and is surrounded by green grass and flowers.



Home > Money > Retirement Planning, News, and Advice > Best Places to Retire for Foodies

Best Places to Retire for Foodies

By EMILY BRANDON | Read Full Story

Walla Walla, Wash.

Sweet onions and wheat were once Walla Walla's best-known exports. But the city is now speckled with wine tasting rooms featuring acclaimed cabernets, merlots, and syrahs and intimate restaurants that make adept use of the locally sourced fruits and vegetables.



© 2023 by M&M

STATISTICAL LEARNING NETWORKS

Fundamentals:



- **"Artificial Intelligence"** was coined by **John McCarthy** (Dartmouth College - 1956) to distinguish the field from cybernetics and escape the influence of the cyberneticist **Norbert Wiener**.
- **Artificial general intelligence** (AGI) studies GI (the ability to take on any arbitrary problem) exclusively. Most AI research usually produced programs that can solve only one problem (**narrow AI**).
- **"Statistical learning"** techniques such as HMM and neural networks gain higher levels of accuracy in many practical domains such as data mining, without necessarily acquiring a semantic understanding of the datasets.

STATISTICAL LEARNING NETWORKS

Fundamentals:



- **Artificial general intelligence** (AGI, strong AI, full AI etc.) is the hypothetical ability of an intelligent agent to understand or learn any intellectual task that a human being can.
- **Narrow AI** (weak AI) is limited to the use of software to study or accomplish specific pre-learned problem solving or reasoning tasks (expert systems).
- In the 1990s and early 21st century, mainstream AI achieved great commercial success and academic respectability by focusing on specific sub-problems where they can produce verifiable results and commercial applications, such as **artificial neural networks** and **statistical machine learning**.

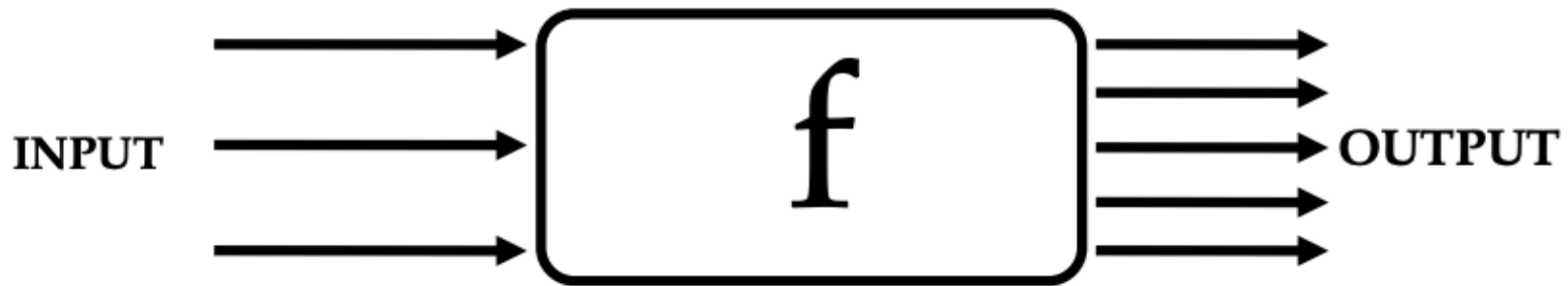
STATISTICAL LEARNING NETWORKS

Learning Models & Approaches



- ***Supervised learning*** - is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
- ***Unsupervised learning*** - looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision, also known as self-organization.
- ***Semi-supervised learning*** - an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training.

Statistical Learning Theory: supervised learning



Given a set of l examples (data)

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_\ell, y_\ell)\}$$

Question: find function f such that

$$f(x) = \hat{y}$$

is a **good predictor** of y for a **future** input x (fitting the data is **not** enough!)

*A framework for machine learning drawing from the fields
of statistics and functional analysis.*

STATISTICAL LEARNING NETWORKS

Fundamentals:



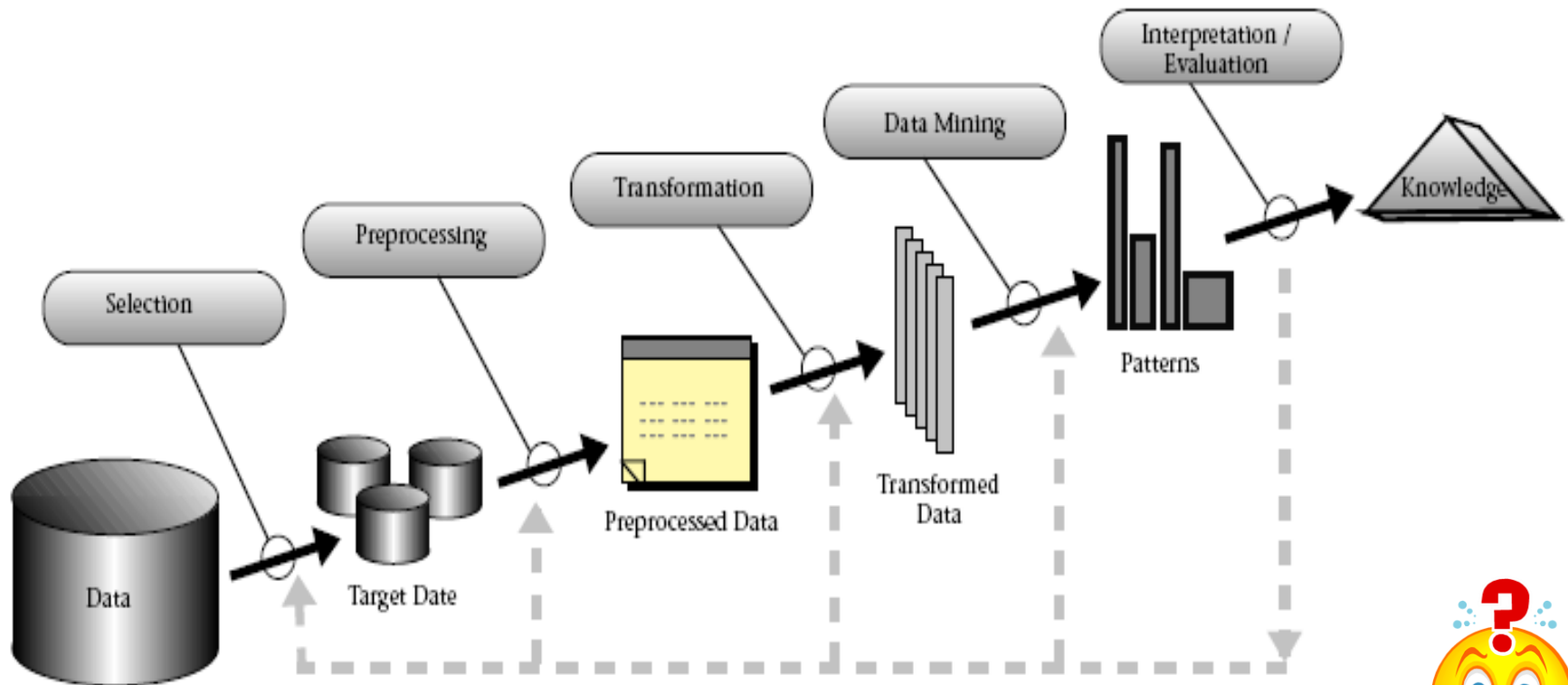
- **Network** – a function (model) represented by the composition of many basic functions (models).
- **Basic function** – element, unit, building block, network node, artificial neuron, partial model.
- **A Learning Network** estimates its function from representative observations of the relevant variables.
- From a data mining perspective, ANNs are just another way of fitting a model to observed historical data in order to be able to make classifications or predictions.

Knowledge Discovery in Databases –

“Identification of underlying patterns, categories, and behaviors in large data sets using techniques such as *neural networks* and *DM*”

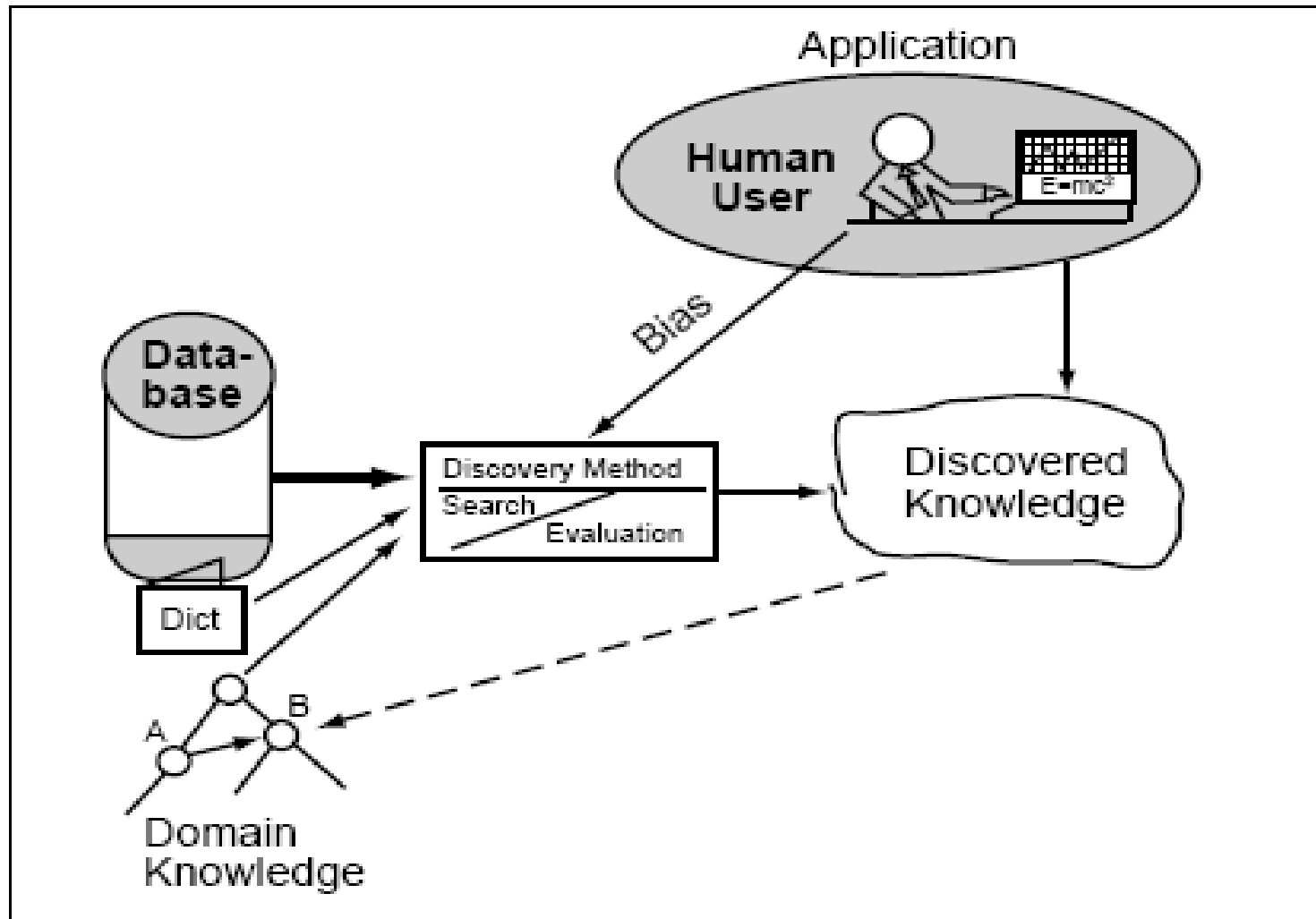


An overview of KDD process

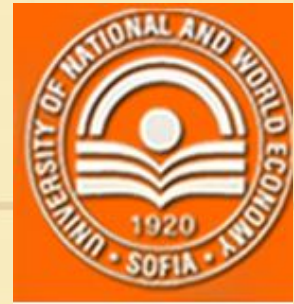


Knowledge Discovery from Data

A Framework for Knowledge Discovery in Databases



STATISTICAL LEARNING NETWORKS



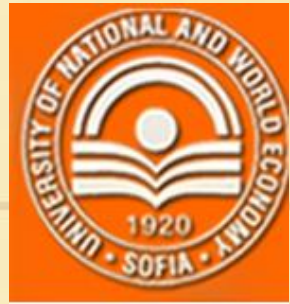
DM and Predictive analytics

- **Data mining** is the process of exploration and analysis (by automatic or semi- automatic means) of large quantities of data in order to discover meaningful patterns and rules.
- **Predictive analytics** encompasses a variety of techniques from *statistics*, *machine learning* and *data mining* that analyze current and historical facts to make predictions about future or otherwise unknown events - technically, predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns.



STATISTICAL LEARNING NETWORKS

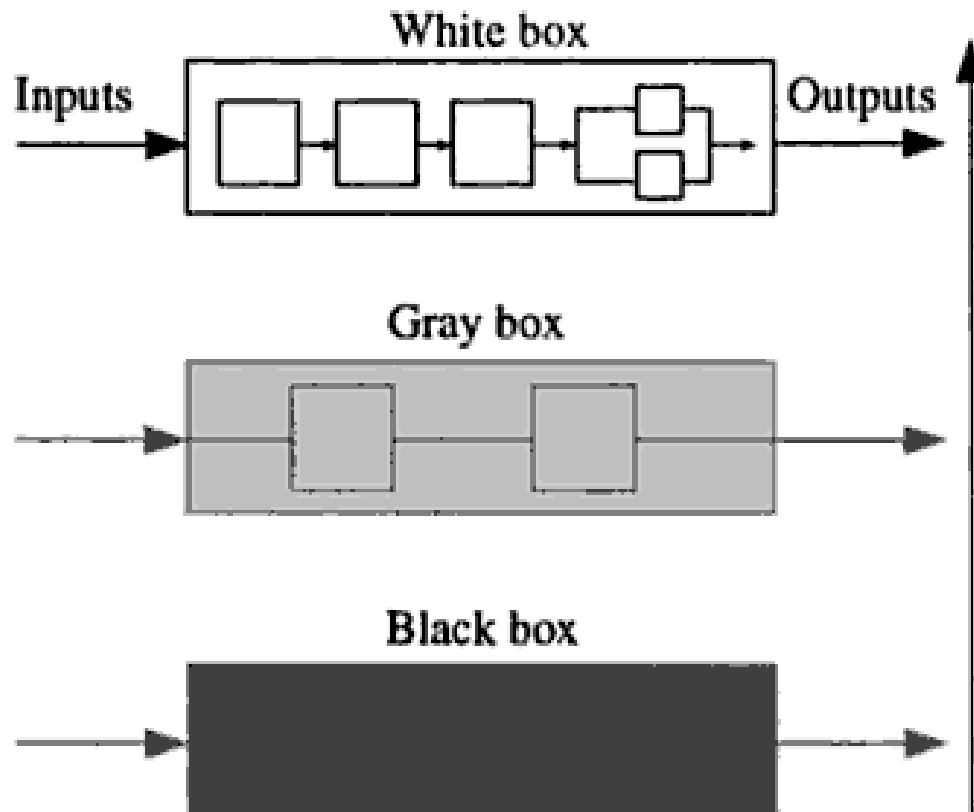
Data mining activities:



- **Classification:** learning a function that maps (classifies) a data item into one of several predefined classes;
- **Estimation (regression):** learning a function that maps a data item into a real-valued prediction variable, building a model;
- **Prediction (predictive modeling):** building a model which can be used to make reliable forecasts;
- **Affinity grouping or association rules:** finding a model that describes significant dependencies between variables;
- **Clustering:** identifying a finite set of categories or clusters to describe the data;
- **Description and visualization (summarization):** finding a compact description for a subset of data.

STATISTICAL LEARNING NETWORKS

Model Identification



Increasing
internal
knowledge



Data Mining

Computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems

STATISTICAL LEARNING NETWORKS



DIRECTED DATA MINING

The goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. A top-down approach, used when we know what we are looking for. It often takes the form of predictive modeling. The model is considered as a **black box**.

Data mining activities:

- **Classification:** learning a function that maps (classifies) a data item into one of several predefined classes;
- **Estimation (regression):** learning a function that maps a data item into a real-valued prediction variable, building a model;
- **Prediction (predictive modeling):** building a model which can be used to make reliable forecasts.

STATISTICAL LEARNING NETWORKS



DIRECTED DATA MINING

- *A top-down approach* – often takes the form of *predictive modeling* where we know exactly what we want to predict. In this case the model is considered as a *black box*, i.e., it is not important what the model is doing, we just want the most accurate result possible.



STATISTICAL LEARNING NETWORKS



UNDIRECTED DATA MINING

A bottom-up approach that finds patterns in the data and leaves it up to the user to determine whether or not these patterns are important, i.e., it is about discovering new patterns inside the data. The goal is to establish some relationship among all the variables (represented with *semitransparent boxes*).

Data mining activities:

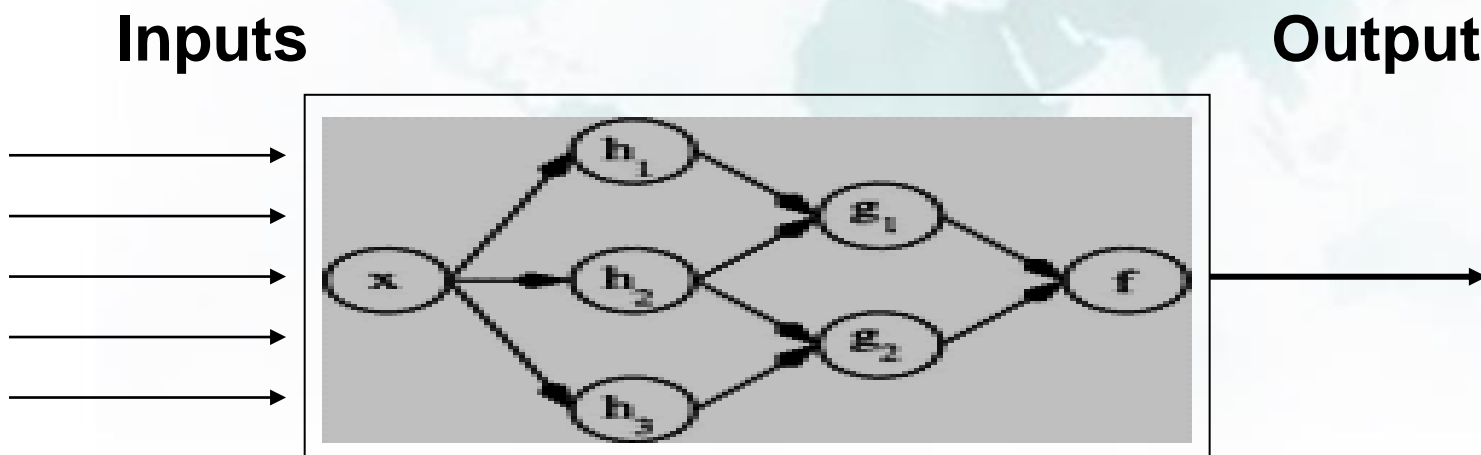
- ***Affinity grouping or association rules***: finding a model that describes significant dependencies between variables;
- ***Clustering***: identifying a finite set of categories or clusters to describe the data;
- ***Description and visualization (summarization)***: finding a compact description for a subset of data.

STATISTICAL LEARNING NETWORKS



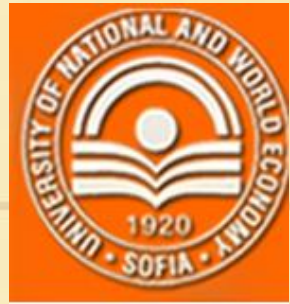
UNDIRECTED DATA MINING

- A *bottom-up approach* that finds patterns in the data which provide insights. This form of data mining is represented with *semitransparent boxes* and unlike directed *DM*, here users want to know what is going on, how the model is coming up with an answer.



STATISTICAL LEARNING NETWORKS

Data Mining Process



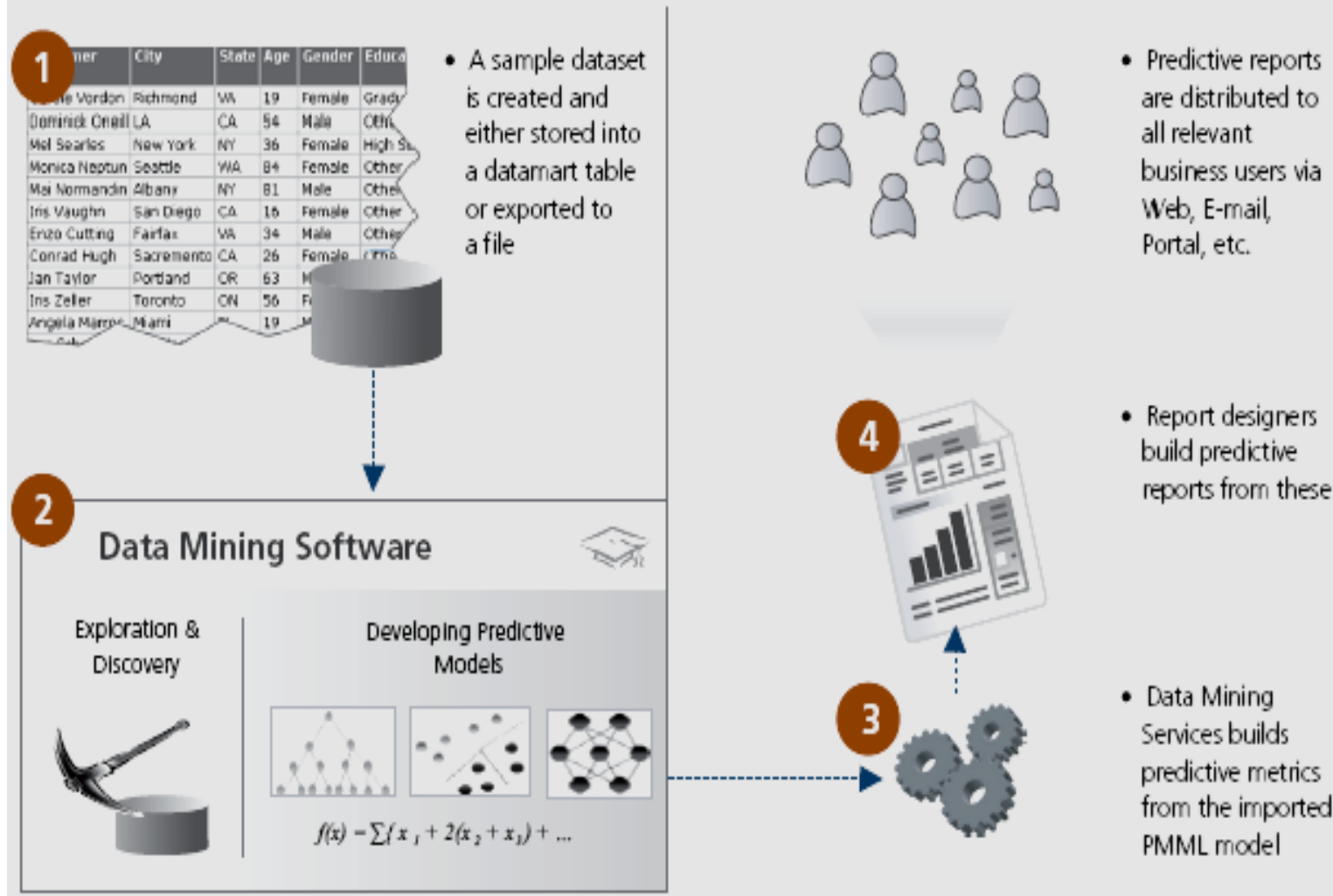
1. Create a predictive model from a data sample
2. Train the model against datasets with known results
3. Apply the model against a new dataset with an unknown outcome (*cross-validation*)

Notes: SAS Institute Inc. developed a five-step data mining cycle process known as **SEMMA**: Sample, explore, modify, model, and assess.

IBM Corp. has a slightly different interpretation of the data mining process and other companies may have their own view as well.

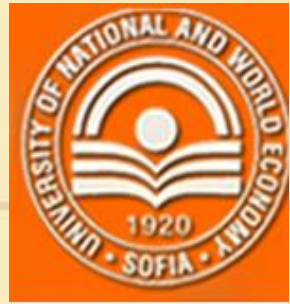
STATISTICAL LEARNING NETWORKS

DM Workflow in MicroStrategy platform



STATISTICAL LEARNING NETWORKS

DM Process - the Three Pillars of Data Mining



Three main components in Data Mining process:

1. *Data* - The power of data mining is leveraging the data that a company collects to make better informed business decisions.
2. *Modeling Skills* - The set of *modeling skills* needed to build predictive models in data mining in general is the same as in business forecasting process and which is working well for both directed and undirected data mining.
3. *Data Mining Techniques* – *clustering, decision trees and neural networks.*

STATISTICAL LEARNING NETWORKS

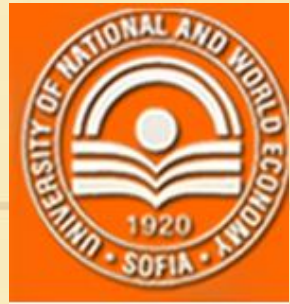


Data mining tasks:

- **classification:** learning a function that maps (classifies) a data item into one of several predefined classes;
- **regression:** learning a function that maps a data item into a real-valued prediction variable;
- **clustering:** identifying a finite set of categories or clusters to describe the data;
- **summarization:** finding a compact description for a subset of data;
- **dependency modeling:** finding a model that describes significant dependencies between variables;
- **change and deviation detection:** discovering the most significant changes in the data from previously measured or normative values.

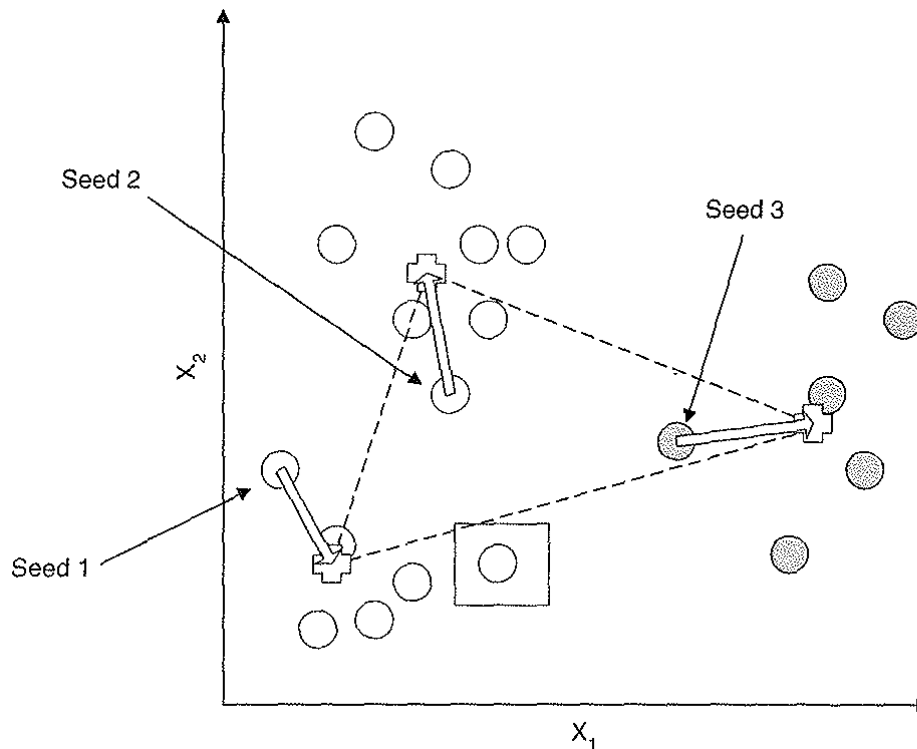
STATISTICAL LEARNING NETWORKS

Data Mining Techniques



- ***Automatic Cluster Detection*** - use cluster detection when we suspect that there are natural groupings that may represent groups of customers or products that have a lot in common with each other.
- ***Decision Trees (Classification & Regression)*** - a good choice when the data mining task is classification of records or prediction of outcomes. We should use decision trees when the goal is to assign each record to one of a few broad categories.
- ***Artificial Neural Networks (the most widely known and the least understood of the major data mining techniques)*** - a good choice for most classification and prediction tasks when the results of the model are more important than understanding how the model works. ANNs represent complex mathematical equations, with lots of summations, exponential functions, and many parameters.

Data Mining Techniques

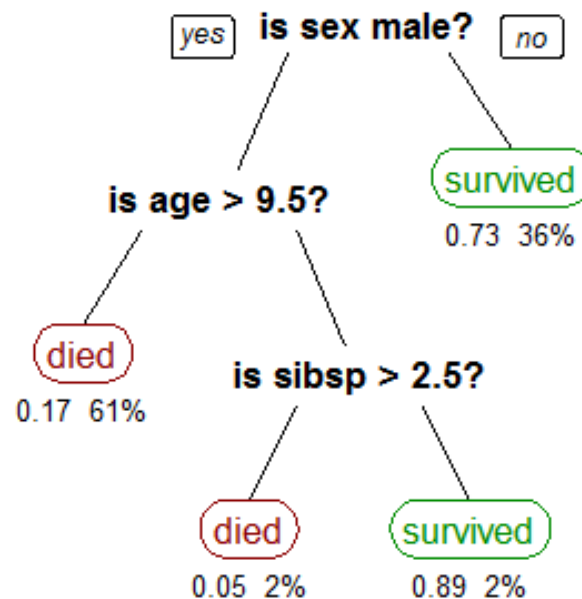


Grouping a set of objects in such a way that objects in the same group (cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters)

Decision Trees



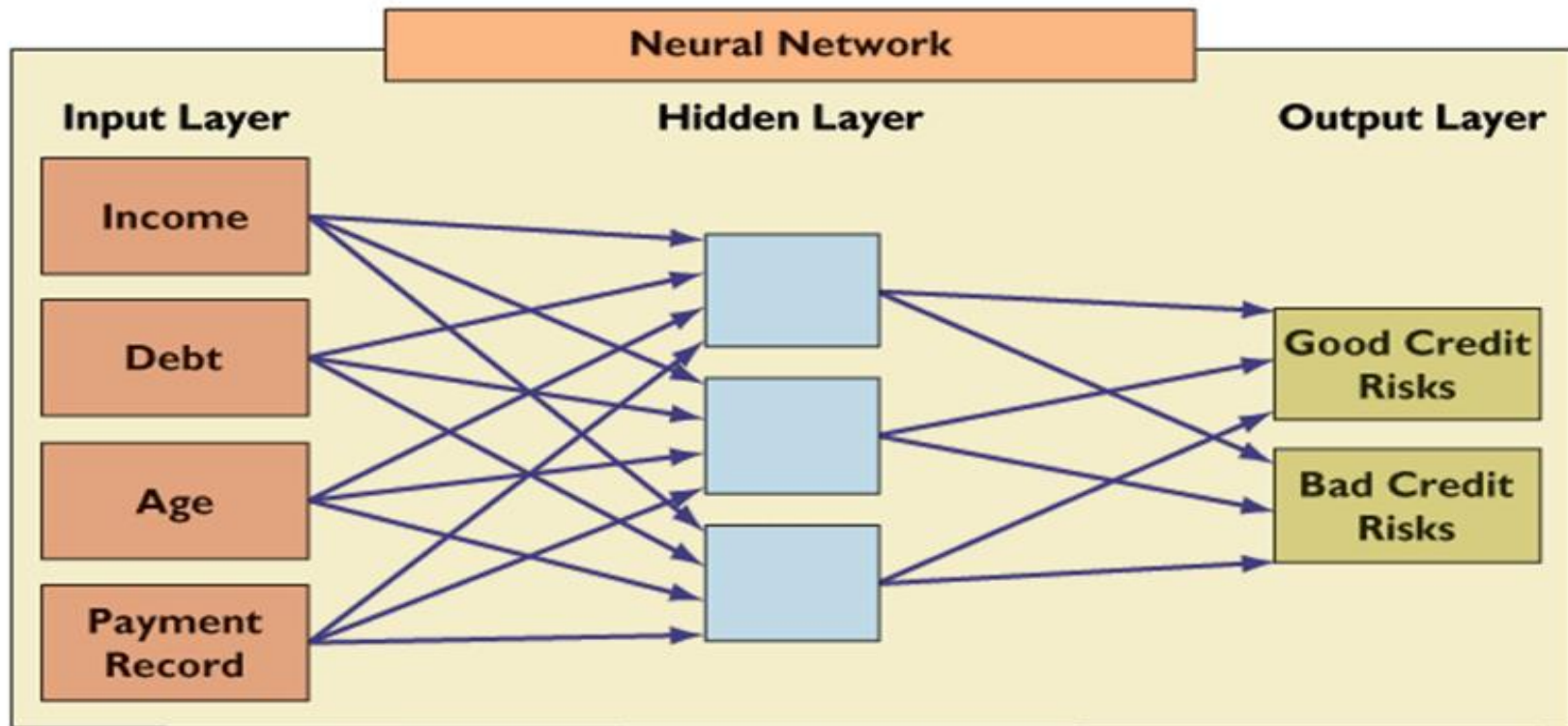
Data Mining Techniques



A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf

Data Mining Techniques

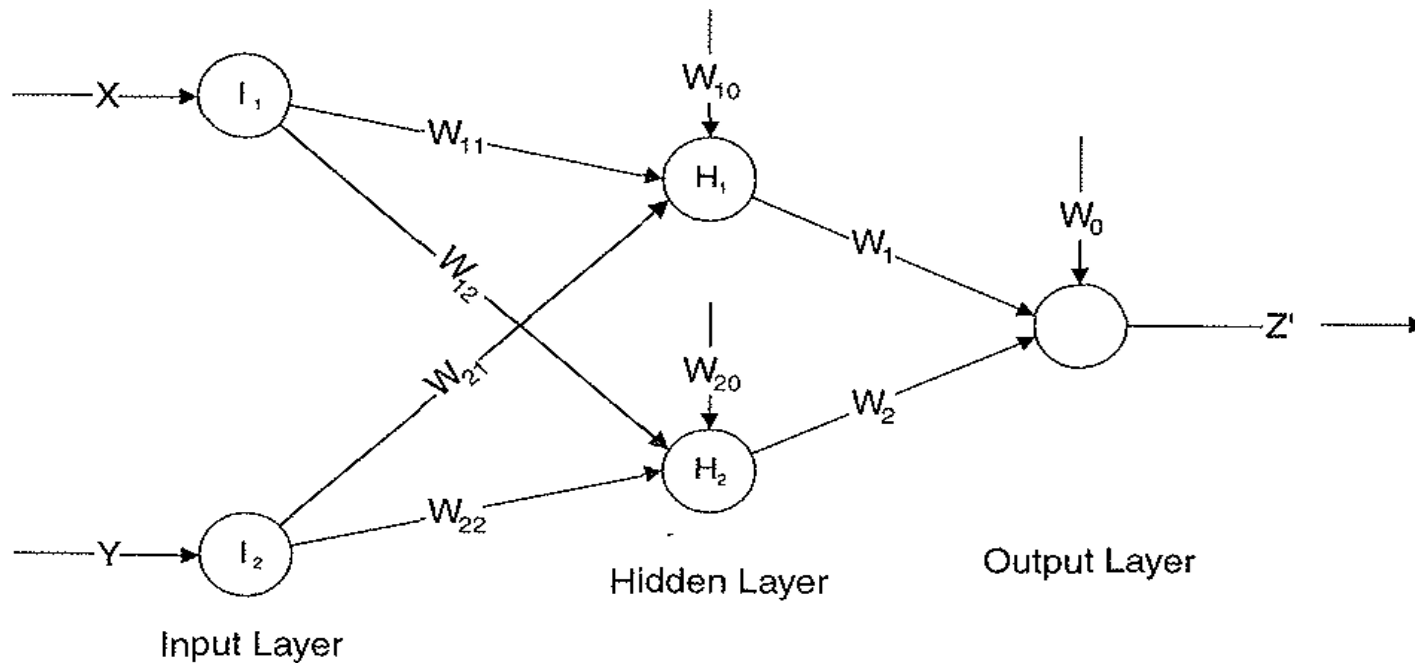
ANN – artificial systems which emulate the processing patterns of the biological brain to discover patterns and relationships in massive amounts of data (“Perceptron” - Ph. Rozenblat)



Source: Herb Edelstein, "Technology How-To: Mining Data Warehouses," *InformationWeek*, January 8, 1996.
Copyright © 1996 CMP Media, Inc., 600 Community Drive, Manhasset, NY 11030. Reprinted with permission.

STATISTICAL LEARNING NETWORKS

DM Techniques - ANNs



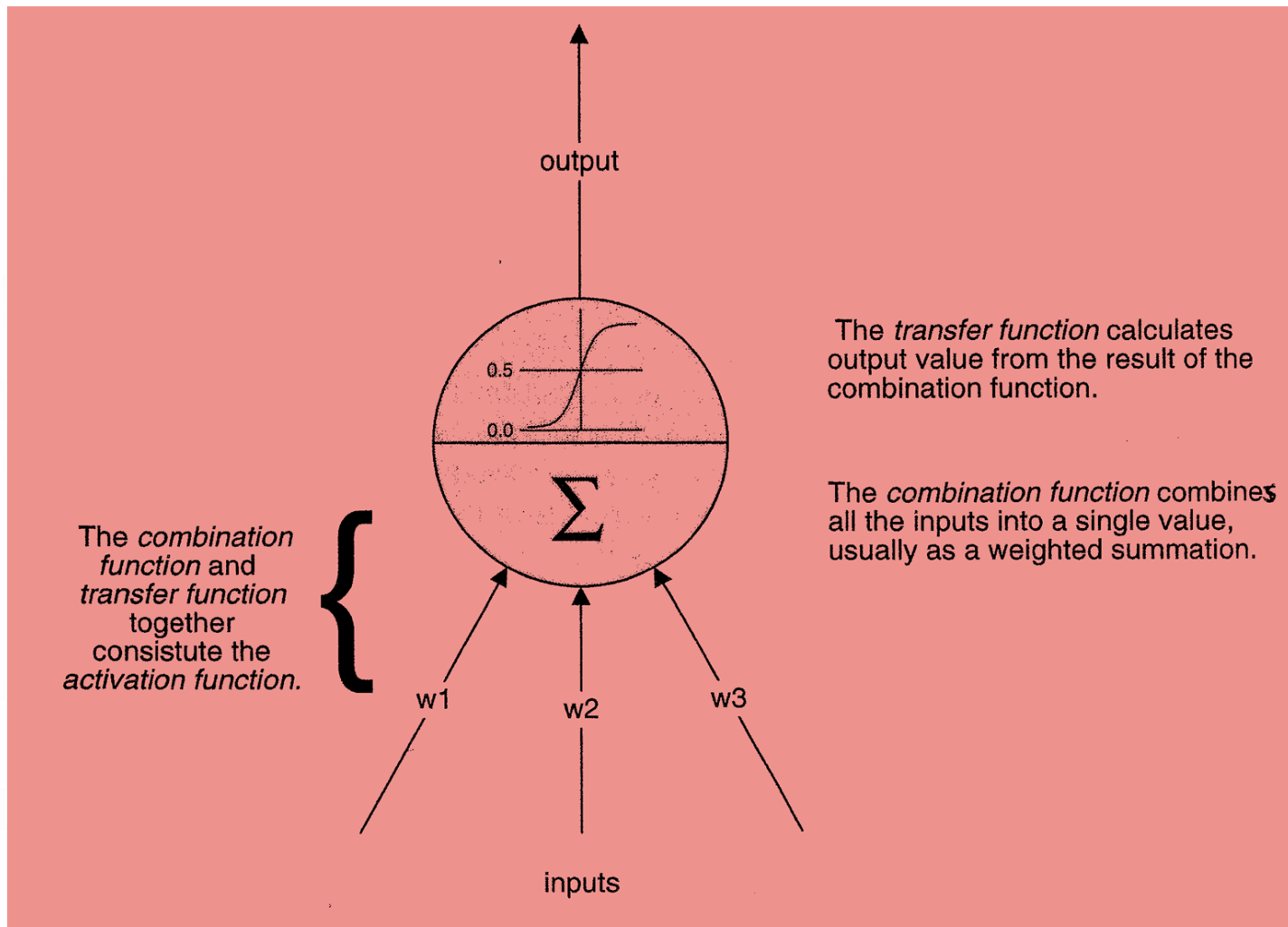
A neural network with a hidden layer.

“The most widely known and the least understood of the major data mining techniques.”

How a Neural Network Works



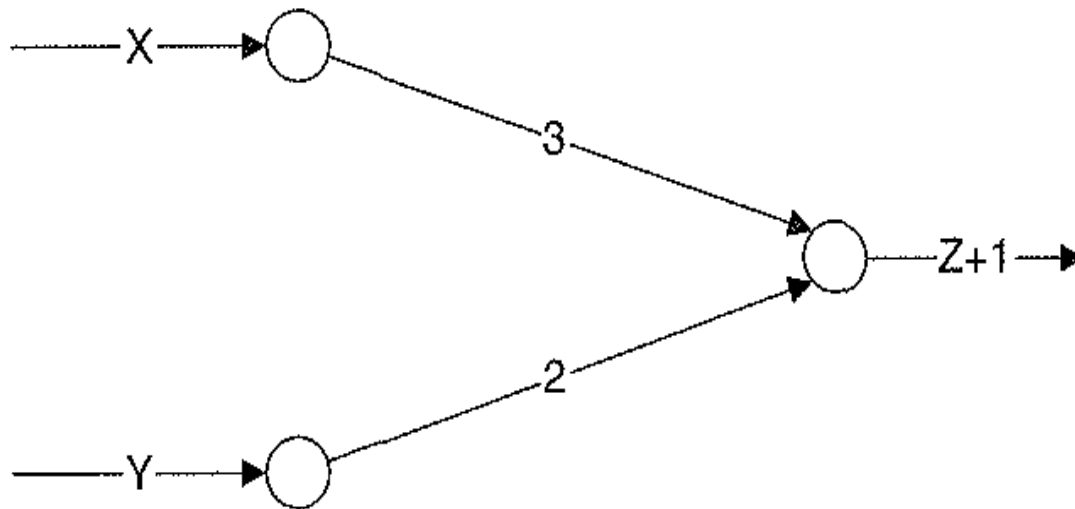
Linear Transfer Function



How a Neural Network Works



Simple ANN – one hidden layer



Input Layer

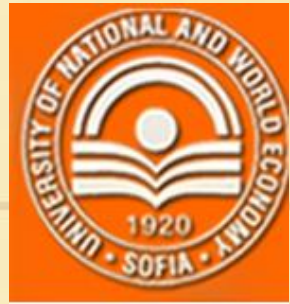
Output Layer

Neural network representation of $z=3x+2y-1$.



How a Neural Network Works

When to use Artificial Neural Networks



ANNs are a good choice for most classification and prediction tasks when the results of the model are more important than understanding how the model works. ANN represent complex mathematical equations, with lots of summations, exponential functions, and many parameters. The equations are the rule of the network but are useless for our understanding. Also, ANN does not work well when there is large number of inputs. This makes it more difficult for the network to find patterns and can result in long training phases that never converge to a good solution.

	ANNs	Statistical Learning Networks
Data analysis	universal approximator	structure identifier
Analytical model	indirect by approximation	direct
Architecture	unbounded network structure; experimental selection of adequate architecture demands time and experience	bounded network structure [1]; adaptively synthesised structure
A-priori-Information	without transformation in the world of ANNs not usable	can be used directly to select the reference functions and criteria
Self-organisation	deductive, given number of layers and number of nodes (subjective choice)	inductive, number of layers and of nodes estimated by minimum of external criterion (objective choice)
Parameter estimation	in a recursive way; demands long samples	estimation on training set by means of maximum likelihood techniques, selection on testing set (extremely short)
Feature	result depends on initial solution, time-consuming technique, necessary knowledge about the theory of neural networks	existence of a model of optimal complexity, not time-consuming technique, necessary knowledge about the task (criteria) and class of system (linear, non-linear)

STATISTICAL LEARNING NETWORKS



General Prediction Model

$$y = a_0 + \sum_{i=1}^M a_i x_i + \sum_{i=1}^M \sum_{j=1}^M a_{ij} x_i x_j + \sum_{i=1}^M \sum_{j=1}^M \sum_{k=1}^M a_{ijk} x_i x_j x_k$$

where

- input variables vector;

$X(x_1, x_2, \dots, x_M)$

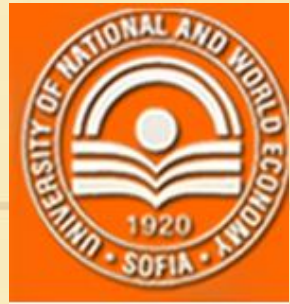
$A(a_1, a_2, \dots, a_M)$ - vector of coefficients or weights.

$$Y = F(X, e)$$



where F can be any mathematical function describing the variable Y (*the output*) as a function of *input variables* X and the stochastic component e (*model error*).

Model Building Problems

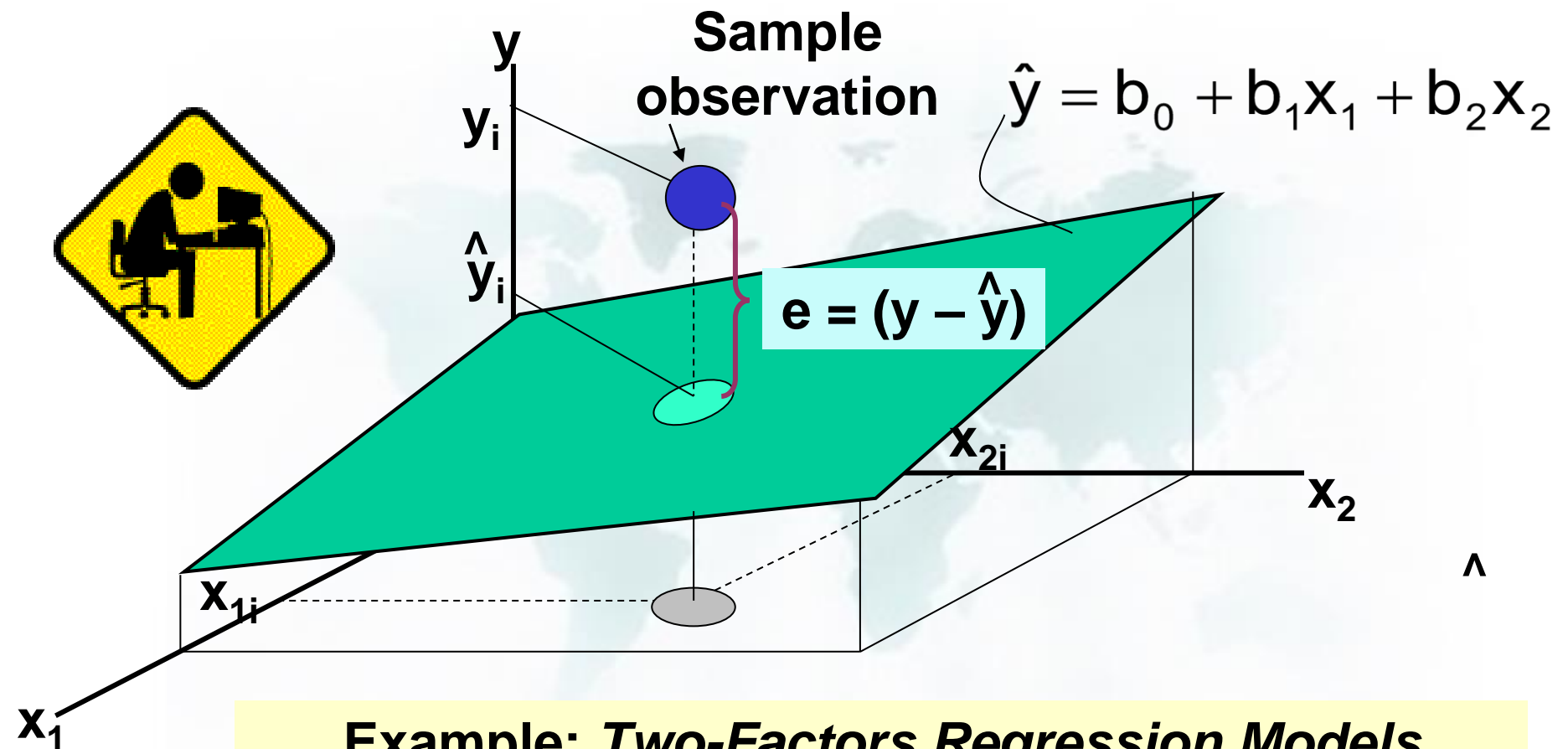


- *Model specification;*
- *Overfitting;*
- *Autocorrelation;*
- *Multicollinearity*
- *ANNs:*
 - *number of layers;*
 - *how many input nodes;*
 - *best activation function;*
 - *ANN training;*
 - *lack of transparency (interpretation), etc.*



Model Building

Regression Analysis



Example: Two-Factors Regression Models

Model Building



Regression Models – Problems:

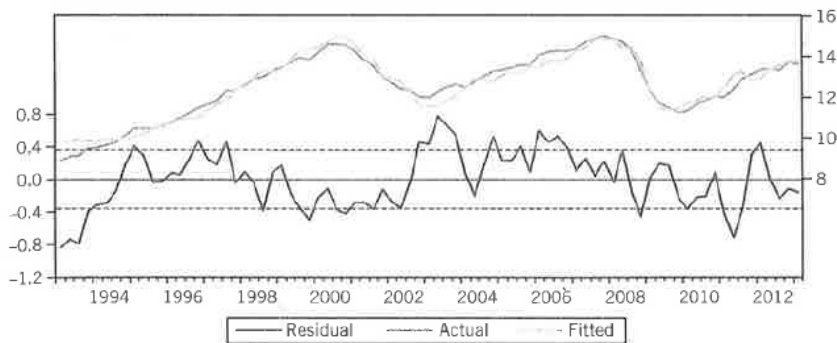
Alan Greenspan (The Map and the Territory: Risk, Human Nature, and the Future of Forecasting):

APPENDICES

Exhibit 4.7

Dependent Variable (Time Period: Q1 1993–Q1 2013, 81 obs.) Real Pvt Nonres Fixed Invst (SAAR, Bil.Chn.2005\$) / Pvt Nonres Fixed Assets (2005 = 100)		
Independent Variable(s)	Coefficient	t-Statistic*
S&P 500 (1941-43=10) / Pvt Nonres Fixed Invst Price (SA, 2005 = 100) (1 quarter ago)	0.473	19.044
Nonfarm Operating Rate (SA, % of capacity) (3 quarters ago)	0.165	6.118
Structures' share of nominal Pvt Nonres Fixed Invst	6.332	4.517
Adjusted R-sq	Durbin-Watson	
0.946	0.585	

*t-statistic calculated using Newey-West HAC standard errors and covariance.



Source: U.S. Department of Commerce; Standard and Poor's; Federal Reserve Board; author's calculations.

Exhibit 3.3

Dependent Variable (Time Period: Jan. 1991–Dec. 2005, 180 obs.) m/m % Δ in CoreLogic Home Price Index (Seasonally adjusted)	
Independent Variable(s) Freddie Mac 30yr Fixed-Rate Mortgage Rate, % p.a. (3 mo)	
Adjusted R-sq	Durbin-Watson
0.604	0.159

*t-statistic calculated using Newey-West HAC standard errors and covaria

Exhibit 4.6

Dependent Variable (Time Period: Q1 1970–Q4 2012, 172 obs.) ln [Real GDP / Real GDP (4 quarters ago)]		
Independent Variable(s) ln [**Corp & Home Equity, Period Avg (1 quarter ago) / **Corp & Home Equity, Period Avg (5 quarters ago)]	Coefficient	t-Statistic*
Adjusted R-sq	Durbin-Watson	
0.419	0.364	

*t-statistic calculated using Newey-West HAC standard errors and covariance.

**Domestic holdings of domestic corporate equities and foreign corporate equities, at market value.

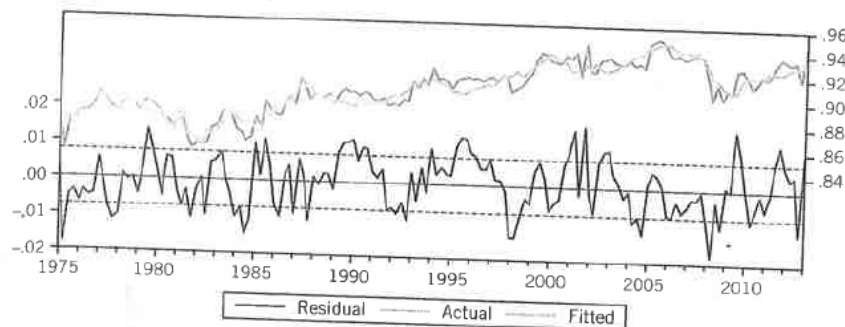
Exhibit 4.4

Dependent Variable (Time Period: Q1 1975–Q1 2013, 153 obs.) Personal Consumption Expenditures (**SAAR, Bil.\$) / Disposable Personal Income (SAAR, Bil.\$)		
Independent Variable(s)	Coefficient	t-Statistic*
Household (incl. NPOs) Stock Net Worth (Period Avg, Bil.\$) / DPI	0.0209	9.56
Household (incl. NPOs) Homeowners' Equity (Period Avg, Bil.\$) / DPI	0.0308	6.35
Household (incl. NPOs) All Other Net Worth (Period Avg, Bil.\$) / DPI	0.0188	2.63
6-Month Certificates of Deposit (% p.a./100) (3 quarters ago)	-0.3752	-9.56
[**Adjusted PI / DPI] (2 quarters ago)	0.2656	2.30
Adjusted R-sq	Durbin-Watson	
0.903	1.089	

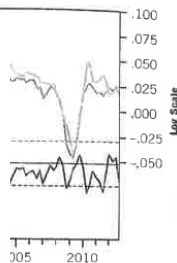
*t-statistic calculated using Newey-West HAC standard errors and covariance.

**Adjusted PI = (0.9*Wages and Salary Disbursements) + (1.0*Personal Current Transfer Receipts) + (0.6*All Other Personal Income).

***Seasonally adjusted annual rate.



Source: Federal Reserve Board; U.S. Department of Commerce.



Simple numerical example

Consider the following data set :

y	a	b	c
9	1	8	1
9	2	7	2
9	3	6	3
9	4	5	4
9	5	4	5
9	6	3	6
9	7	2	7
6	99	1	5

Model:

$$Y = F(a, b, c)$$

Solutions:

$$y = 9.3 - 0.033a -$$

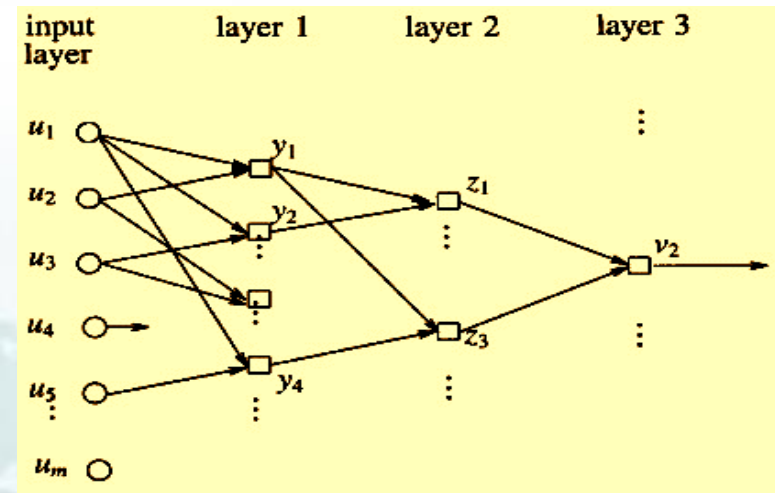
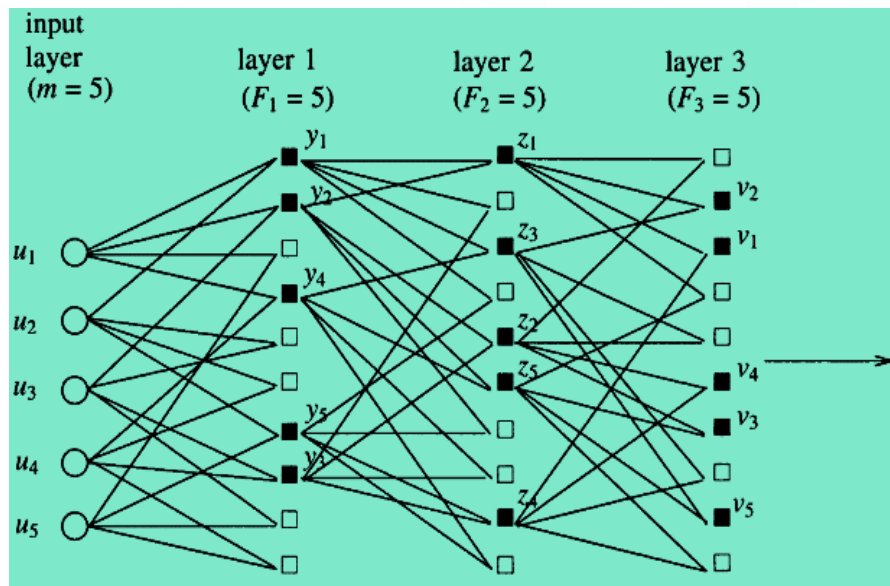
$$0.033b + 0.00001 + b + c$$

$$y = 9 - 0.0319a + 0.0319c$$

Model Building

Multilayered Nets of Active Neurons

Multilayered network structure with five input arguments and selected nodes:



Multilayered network structure representing the output flow to unit 2 of layer 3

Source: ISAGA 2014 - Predictive Analytics in Business Games and Simulations

“The first general, working learning algorithm for supervised, deep, feedforward, multilayer perceptron(s) was published by Alexey Ivakhnenko and Lapa in 1967” (Wikipedia).

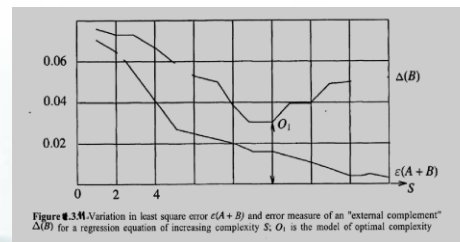
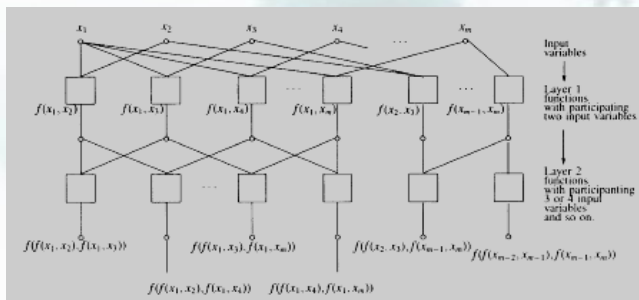
GMDH



Alexey G. Ivakhnenko.
(1913-2007)

Two State Prizes of the USSR, Medal “For Labor”, Order of Friendship of Peoples ...

Genetic selection of pairwise features



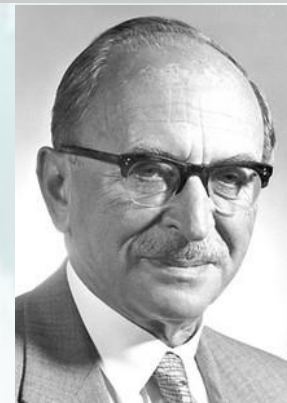
Kurt Gödel
(1906-1978)

Notable awards:

- Albert Einstein Award (1951)
- National Medal of Science (USA) in Mathematical, Statistical, and Computational Sciences (1974)

Gabor's principle of "freedom of decisions choice"

Knowledge extraction from experimental data, Self-Organization etc...



Dennis Gabor (1900-1978)

Numerous (>20) awards:

- Nobel Prize in Physics (1971)
- Honorary Doctorate, Delft University of Technology (1971)

Gödel's Incompleteness Theorems:

Two theorems of mathematical logic that are concerned with the limits of provability in formal axiomatic theories



- **First Incompleteness Theorem:** "Any consistent formal system F within which a certain amount of elementary arithmetic can be carried out is incomplete; i.e., there are statements of the language of F which can neither be proved nor disproved in F ."
- The unprovable statement $G(F)$ referred to by the theorem is often referred to as "the Gödel sentence" for the system F . The proof constructs a particular Gödel sentence for the system F , but there are infinitely many statements in the language of the system that share the same properties.
- Each effectively generated system has its own Gödel sentence. It is possible to define a larger system F' that contains the whole of F plus GF as an additional axiom.
- This will not result in a complete system, because Gödel's theorem will also apply to F' , and thus F' also cannot be complete. In this case, GF is indeed a theorem in F' , because it is an axiom. Because GF states only that it is not provable in F , no contradiction is presented by its provability within F' . However, because the incompleteness theorem applies to F' , there will be a new Gödel statement GF' for F' , showing that F' is also incomplete. GF' will differ from GF in that GF' will refer to F' , rather than F .

Gödel's Incompleteness Theorems:

Two theorems of mathematical logic that are concerned with the limits of provability in formal axiomatic theories



- The first incompleteness theorem shows that the Gödel sentence GF of an appropriate formal theory F is unprovable in F . Because, when interpreted as a statement about arithmetic, this unprovability is exactly what the sentence (indirectly) asserts, the Gödel sentence is, in fact, true. For this reason, the sentence GF is often said to be "true but unprovable." However, since the Gödel sentence cannot itself formally specify its intended interpretation, the truth of the sentence GF may only be arrived at via a meta-analysis from outside the system.
- Compared to the theorems stated in Gödel's 1931 paper, many contemporary statements of the incompleteness theorems are more general in two ways. These generalized statements are phrased to apply to a broader class of systems, and they are phrased to incorporate weaker consistency assumptions.
- Gödel demonstrated the incompleteness of the system of Principia Mathematica (particular system of arithmetic) but a parallel demonstration could be given for any effective system of a certain expressiveness. Gödel commented on this fact in the introduction to his paper but restricted the proof to one system for concreteness. *In modern statements of the theorem, it is common to state the effectiveness and expressiveness conditions as hypotheses for the incompleteness theorem, so that it is not limited to any particular formal system.*

Gödel's Incompleteness Theorems:

Two theorems of mathematical logic that are concerned with the limits of provability in formal axiomatic theories



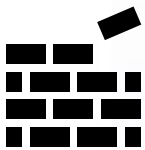
- *The first incompleteness theorem* states that no consistent system of axioms whose theorems can be listed by an effective procedure (i.e., an algorithm) is capable of proving all truths about the arithmetic of natural numbers. For any such consistent formal system, there will always be statements about natural numbers that are true, but that are unprovable within the system. *The second incompleteness theorem, an extension of the first, shows that the system cannot demonstrate its own consistency.* A consistent theory is one that does not lead to a logical contradiction.
- The semantic definition states that a theory is consistent if it has a model, i.e., there exists an interpretation under which all formulas in the theory are true. The syntactic definition states a theory $\{T\}$ is consistent if there is no formula (f) and its negation $\{\text{not } f\}$ are elements of the set of consequences of $\{T\}$.
- For each formal system F containing basic arithmetic, it is possible to canonically define a formula $\text{Cons}(F)$ expressing the consistency of F . Gödel's second incompleteness theorem shows that, under general assumptions, this canonical consistency statement $\text{Cons}(F)$ will not be provable in F .

Gödel's Incompleteness Theorems:

Two theorems of mathematical logic that are concerned with the limits of provability in formal axiomatic theories



- The second incompleteness theorem does not rule out altogether the possibility of proving the consistency of some theory T , only doing so in a theory that T itself can prove to be consistent. For example, Gerhard Gentzen proved the consistency of Peano arithmetic in a different system that includes an axiom asserting that the ordinal called ϵ_0 is wellfounded.
- Gentzen's consistency proof is a result of proof theory in mathematical logic, published by Gerhard Gentzen in 1936. It shows that the Peano axioms of first-order arithmetic do not contain a contradiction (i.e., are "consistent"), if a certain other system used in the proof does not contain any contradictions either. This other system, today called "primitive recursive arithmetic with the additional principle of quantifier-free transfinite induction up to the ordinal ϵ_0 ", is neither weaker nor stronger than the system of Peano axioms. Gentzen argued that it avoids the questionable modes of inference contained in Peano arithmetic and that its consistency is therefore less controversial.

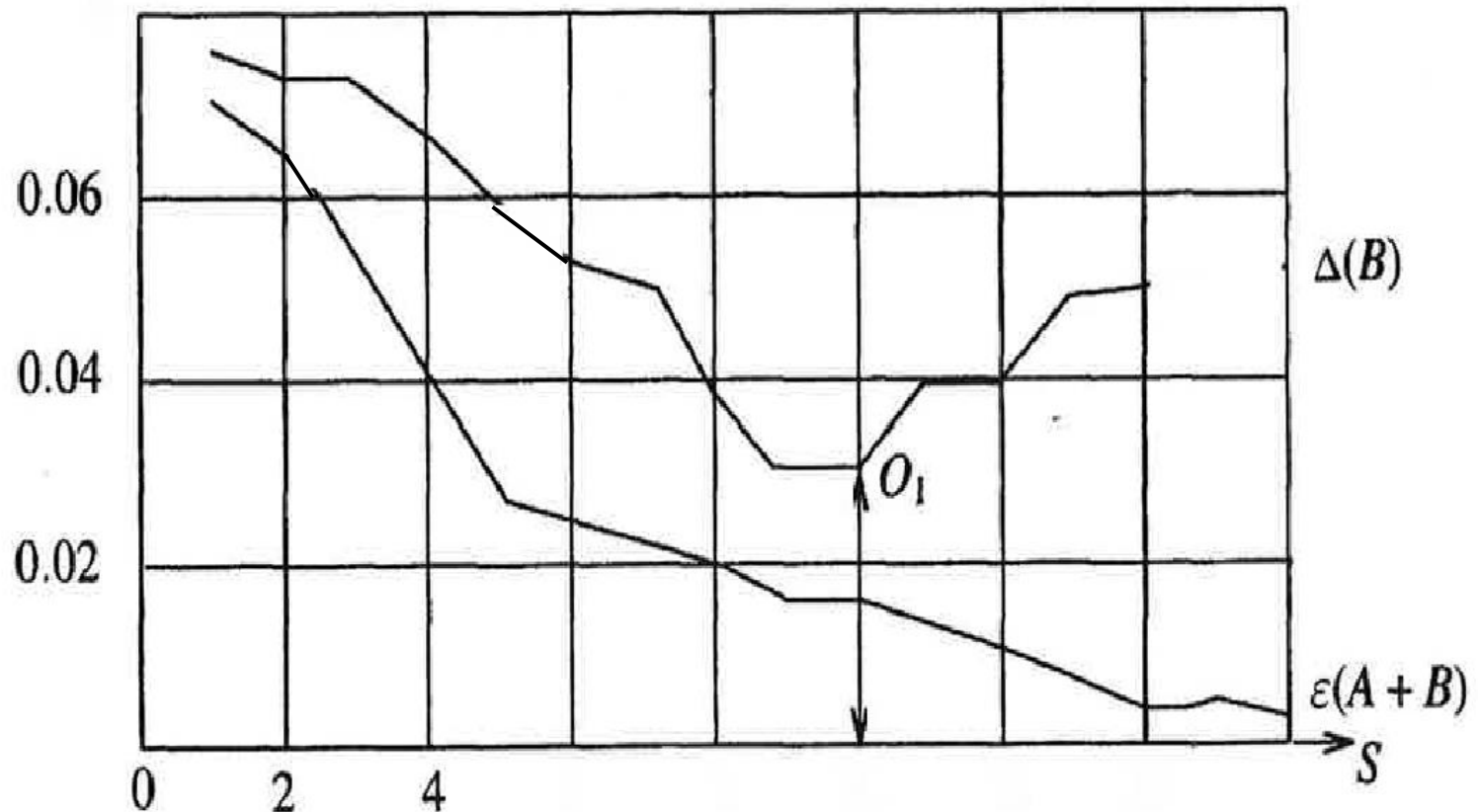


Artificial neural networks (ANNs): **Over-training arises in over-specified systems when the network capacity exceeds the needed free parameters.**



- The first approach to address this is to use *cross-validation* to check for the presence of over-training and to select hyperparameters to minimize the generalization error.
- The second is to use some form of *regularization*. This concept emerges in a probabilistic (Bayesian) framework but also in statistical learning theory, where the goal is to minimize over two quantities: the 'empirical risk' and the 'structural risk', which roughly corresponds to the error over the training set and the predicted error in unseen data due to overfitting.
- Supervised ANNs that use a mean squared error (MSE) cost function can use formal statistical methods to determine the confidence of the trained model. The MSE on a validation set can be used as an estimate for variance. This value can then be used to calculate the confidence interval of network output, assuming a normal distribution.
- By assigning a softmax activation function, a generalization of the logistic function, on the output layer of the neural network for categorical target variables, the outputs can be interpreted as posterior probabilities.

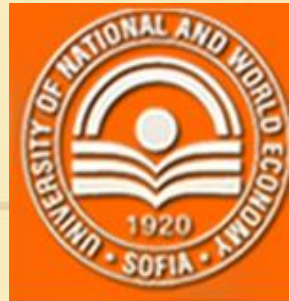
Overfitting – Internal vs External (Cross) Validation



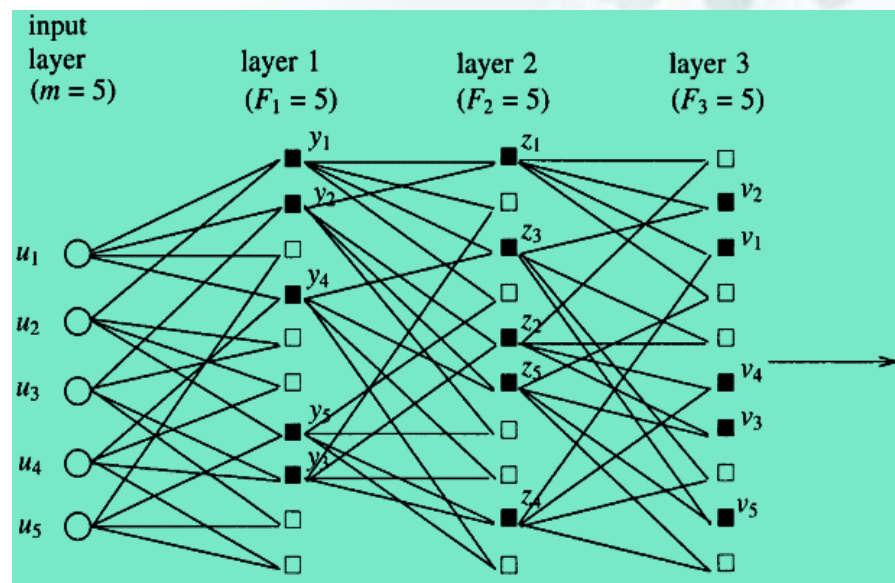
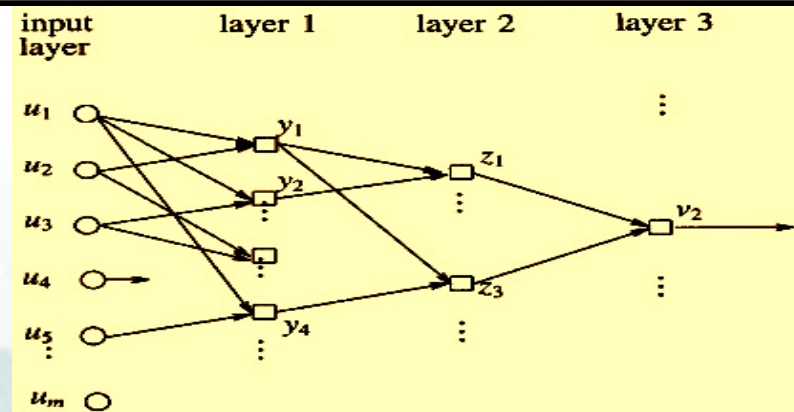
Variation in least square error $\epsilon(A+B)$ and error measure of an "external complement" $\Delta(B)$ for a regression equation of increasing complexity S ; O_1 is the model of optimal complexity

Statistical Learning Networks of Active Neurons

Multilayered Net of Active Neurons (MLNAN)



In this approach, neither the number of neurons and the number of layers in the network, nor the actual behavior of each created neuron is predefined. The modeling process is self-organizing because all of them (the number of neurons, the number of layers, and the actual behavior of each created neuron) are adjusting during the process of self-organization.

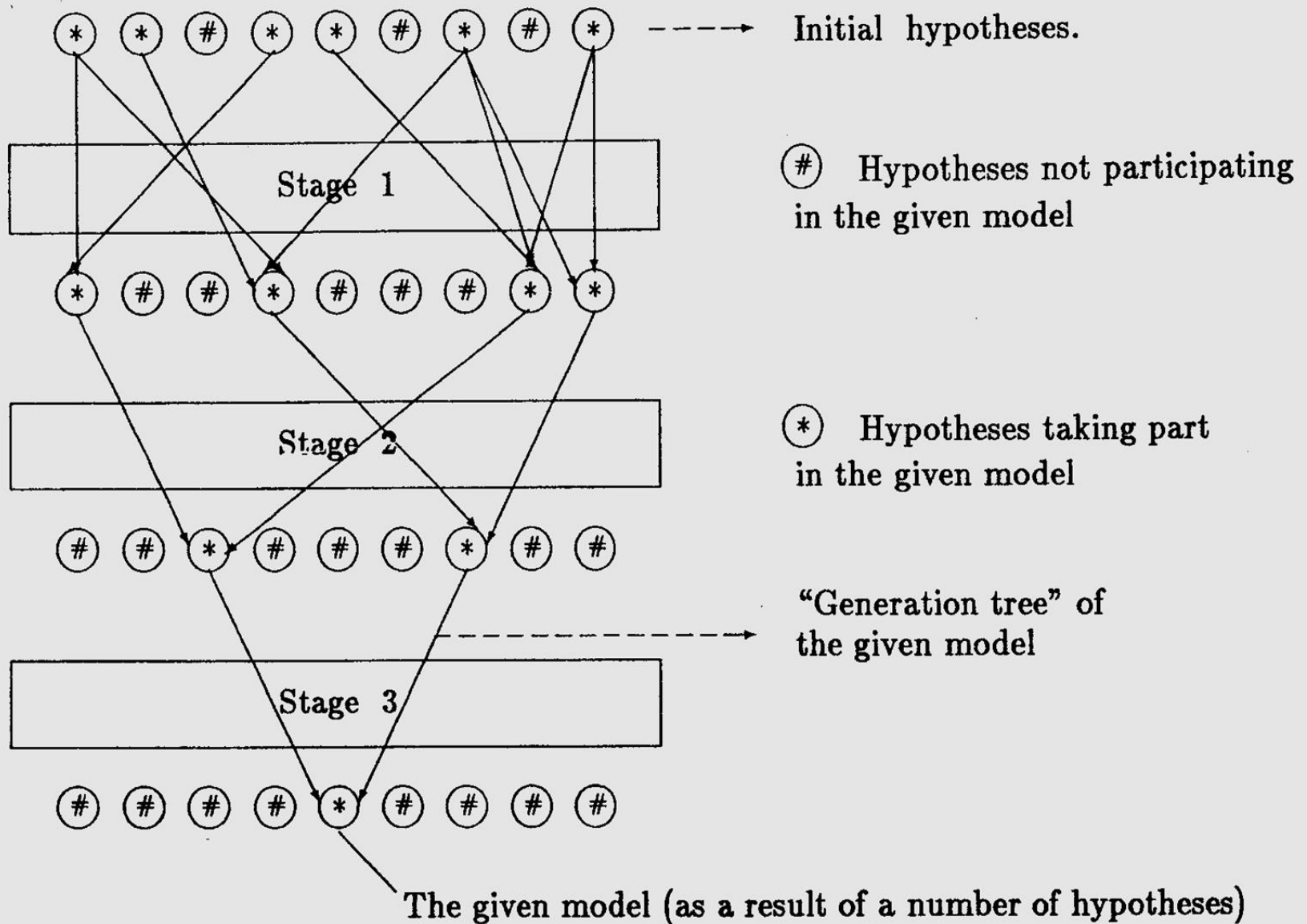


Multilayer network structure representing the output flow to unit 2 of layer 3

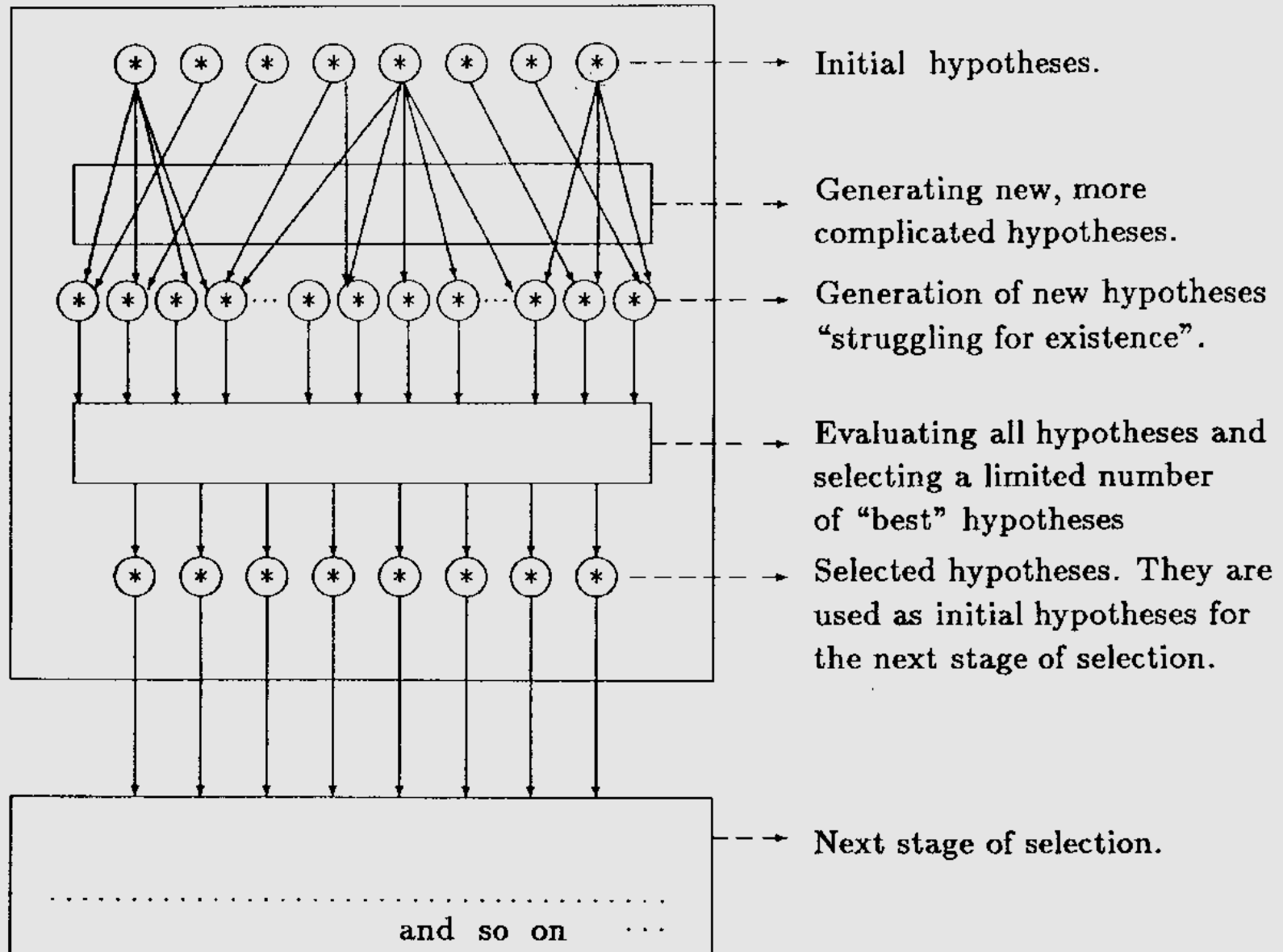
This method grows a tree-like network out of data of input and output variables in a pairwise combination and competitive selection from a simple single unit to a desired final solution that does not have a predefined model. The basic idea is that first the elements on a lower level are estimated and the corresponding intermediate outputs are computed and then the parameters of the elements of the next level are estimated.

Multilayer network structure with five input arguments and selected nodes:

Multi-Stage Selection Algorithm

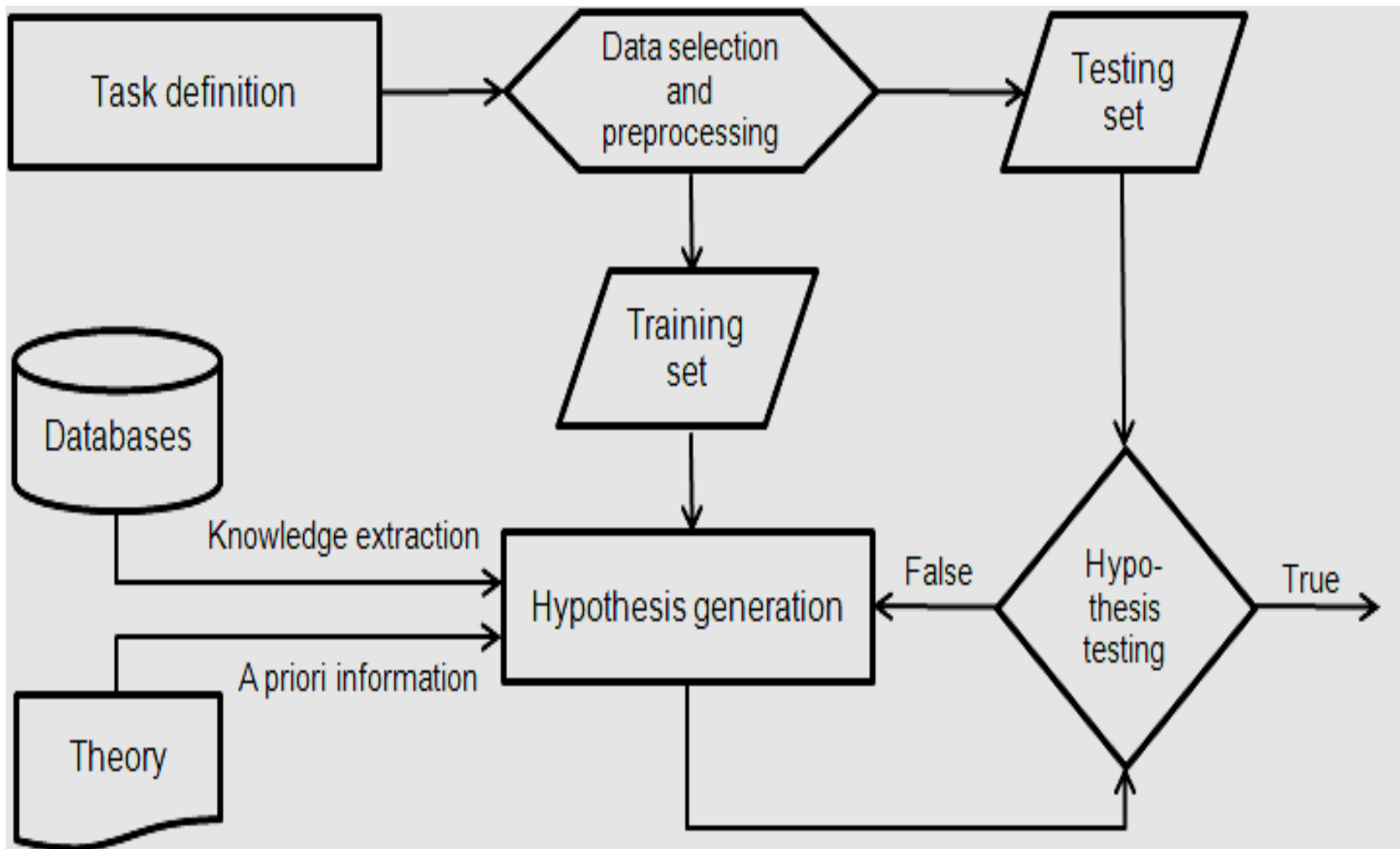


Pair-Wise Selection Using External Criteria



Model Selection

Cross Validation and a-priori information



STATISTICAL LEARNING NETWORKS

Model Selection & Validation



- ***The concept of Cross Validation, also called **rotation estimation, out-of-sample testing, predictive sample reuse, reuse of the sample** etc. is an old one:***
 - (1951). Symposium: The need and means of cross-validation:
 - I. Problem and designs of cross-validation.
 - II. Approximate linear restraints and best predictor weights.
 - III. Cross-validation of item analyses.

STATISTICAL LEARNING NETWORKS

Model Selection & Validation



- **Cross Validation** - also called *rotation estimation* or *out-of-sample testing*, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.
- Involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the *training set*), and validating the analysis on the other subset (called the *validation set* or *testing set*).
- Two types of cross-validation can be distinguished: *exhaustive* and *non-exhaustive cross-validation*.

- ***Exhaustive cross-validation*** - learn and test on all possible ways to divide the original sample into a training and a validation set.
 - ***Leave-p-out cross-validation*** - involves using p observations as the validation set and the remaining observations as the training set. This is repeated on all ways to cut the original sample on a validation set of p observations and a training set.
 - ***Leave-one-out cross-validation*** - a particular case of ***leave-p-out cross-validation*** with $p = 1$.

- **Leave-one-out cross-validation:**

1. Select (it could be random) observation i for the testing set and use the remaining observations in the training set. Compute the error on the test observation.
2. Repeat the above step for $i = 1, 2, \dots, N-1$, where N is the total number of observations.
3. Compute the forecast accuracy measures based on all errors obtained.

A total of 8 models
will be trained and
tested:

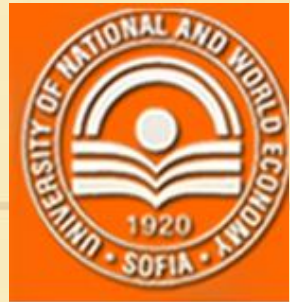


- ***Non-exhaustive cross-validation*** - do not compute all ways of splitting the original sample. Those methods are approximations of ***leave-p-out cross-validation***.
 - ***k-fold cross-validation*** - the sample is randomly partitioned into k equal sized subsamples. When $k = n$ (the number of observations), *k-fold cross-validation* is equivalent to *leave-one-out cross-validation*.
 - ***holdout method*** - randomly assign data points to two sets A and B (training set and test set).
 - ***repeated random sub-sampling validation*** or ***Monte Carlo cross-validation*** creates multiple random splits of the dataset into training and validation data

- ***Nested cross-validation*** - cross-validation is used simultaneously for selection of the best set of hyperparameters and for error estimation.
 - ***k*l-fold cross-validation*** - contains an outer loop of k folds and an inner loop of l folds. One by one, a set is selected as (outer) test set and the $k - 1$ other sets are combined into the corresponding outer training set.
 - ***k-fold cross-validation with validation and test set*** - $k*l$ -fold cross-validation when $l = k - 1$. One by one, a set is selected as a test set. Then, one by one, one of the remaining sets is used as a validation set and the other $k - 2$ sets are used as training sets until all possible combinations have been evaluated.

STATISTICAL LEARNING NETWORKS

Cross Validation with Time Series data



- **Rolling forecasting origin** - since it is not possible to get a reliable forecast based on a very small training set, the earliest observations n are not considered as testing sets.
 1. We select the observation at time $(n+i)$ for the testing set and use the observations at times $t = \{1, 2, \dots, (n+i-1)\}$ to estimate the forecasting model. Then we compute the error on the forecast for the time $(n+i)$.
 2. The above step should be done for all $i = \{1, 2, \dots, (T-n)\}$, where T is the total number of observations and the forecast error should be measured on each $(n+i)$ period accordingly.
 3. In the end, we compute the forecast accuracy measures based on all errors obtained.

Accuracy, Trueness and Precision



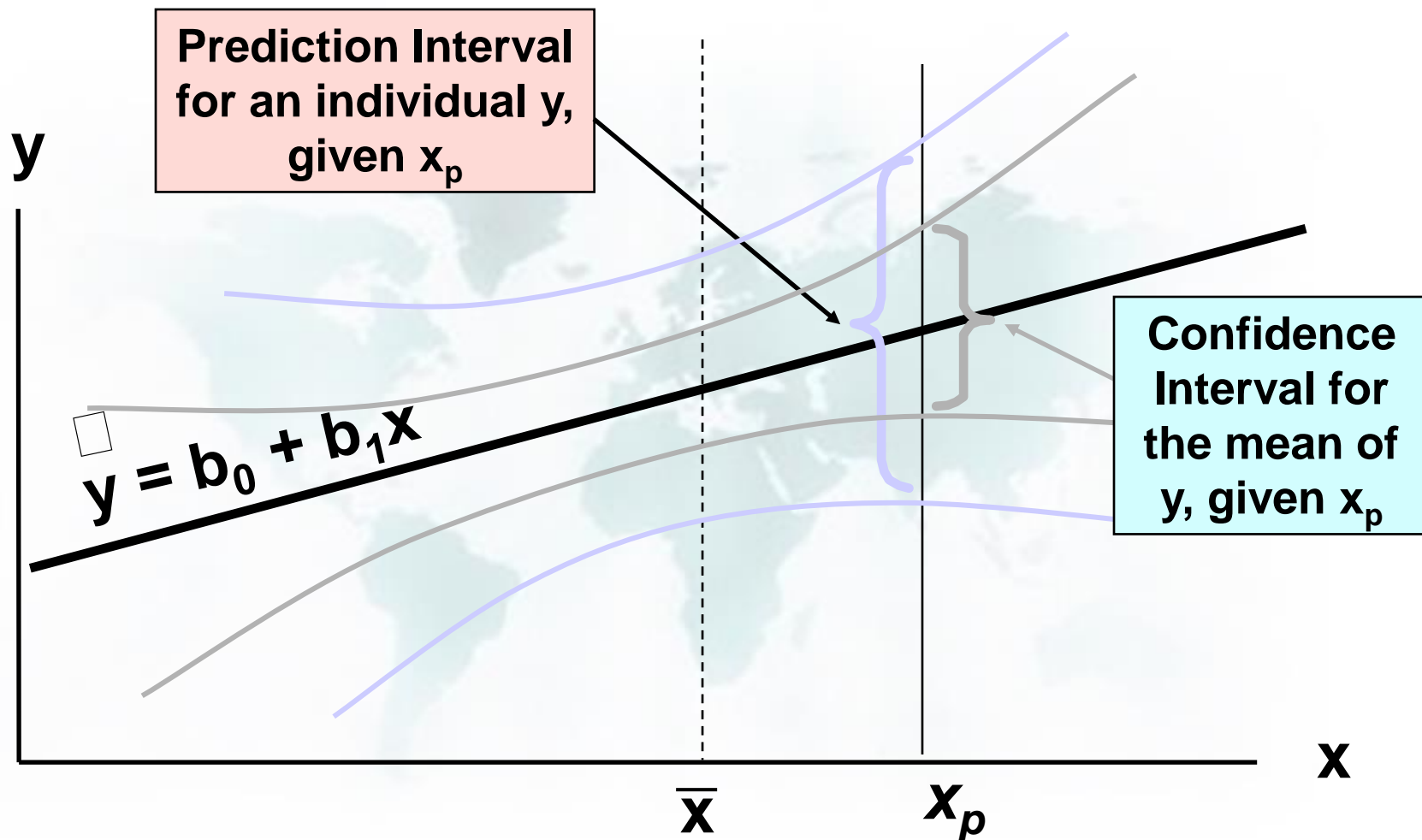
A) Low accuracy due to poor precision.



B) Low accuracy due to poor trueness.

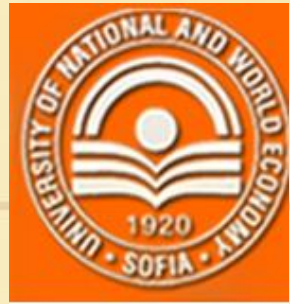
ISO 5725 (1994) *Accuracy – trueness and precision*

Model Selection & Validation - Accuracy



STATISTICAL LEARNING NETWORKS

Model Selection & Accuracy



Prediction (simulation) error:

$$e_t = y_t - F_t$$

where e_t is the error at period t ($t=\{1, 2, 3...N\}$);

N is the prediction interval (or the size of the dataset);

y_t is the actual value at period t and

F_t is the forecast for period t .

Mean Forecast Error (forecast bias):

$$\text{MFE} = \frac{1}{N} \sum_{t=1}^N e_t$$

Two common Measures of Fit

- Measures of fit are used to gauge how well the forecasts match the actual values

MSE (mean squared error)

- Average **squared** difference between y_t and F_t

MAD (mean absolute deviation)

- Average **absolute value** of difference between y_t and F_t
- Less sensitive to extreme values

- Mean Absolute Deviation (MAD)
 - Average absolute error – most useful to measure the forecast error in the same units as the original series.

$$\text{MAD} = \frac{\sum |\text{Actual} - \text{Forecast}|}{n} = \frac{\sum |e(t)|}{n}$$

- Mean Squared Error (MSE)
 - Average of squared error – provides a penalty for large forecasting errors (it squares each)

$$\text{MSE} = \frac{\sum (\text{Actual} - \text{forecast})^2}{n - 1}$$

STATISTICAL LEARNING NETWORKS

MSE vs. MAD

Mean Squared Error

$$MSE = \frac{\sum (y_t - F_t)^2}{n - 1}$$

Mean Absolute Deviation

$$MAD = \frac{\sum |y_t - F_t|}{n}$$

where:

y_t = Actual value at time t

F_t = Predicted value at time t

n = Number of time periods

■ MSE

- Squares errors
- More weight to large errors

MAD

- Easy to compute
- Weights errors linearly

- Mean Percentage Error (MPE)
 - Average percentage error – useful when it is necessary to determine whether a forecasting method is biased. If the forecast is unbiased MPE will produce a % that is close to 0. Large –% means overestimating. Large +% - the method is consistently underestimating.

$$\text{MPE} = \frac{\sum (\text{Actual} - \text{Forecast}) / \text{Actual}}{n} \times 100$$

- Coefficient of variation of the Root Mean Squared Error, $CV(RMSE)$: The RMSE serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power and $CV(RMSE)$ helps to compare forecasting errors of different models.

$$CV(RMSE) = RMSE/\bar{y} \quad RMSE = \sqrt{MSE}$$

STATISTICAL LEARNING NETWORKS

Model Selection



Measures of Trueness (Systematic error, Statistical Bias):

- **Mean Percentage Error (MPE)**

$$\text{MPE (\%)} = \frac{1}{N} \sum_{t=1}^N (e_t / y_t) \times 100$$

- **Root Mean Squared Error (RMSE)**

$$\text{RMSE} = \sqrt{\text{MSE}}$$

$$\text{MSE} = \sum (e_t)^2 / (n - 1)$$

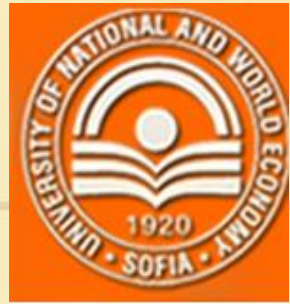
Model Selection

- Mean Absolute Percent Error (MAPE) - Puts errors in perspective:
 - Average absolute percent error – useful when the size of the forecast variable is important in evaluating. It provides an indication of how large the forecast errors are in comparison to the actual values of the series. It is also useful to compare the accuracy of different techniques on same/different series.

$$\text{MAPE} = \frac{\sum (|\text{Actual} - \text{forecast}|) / \text{Actual} * 100}{n}$$

STATISTICAL LEARNING NETWORKS

Model Selection



Measures of Precision (Random Error):

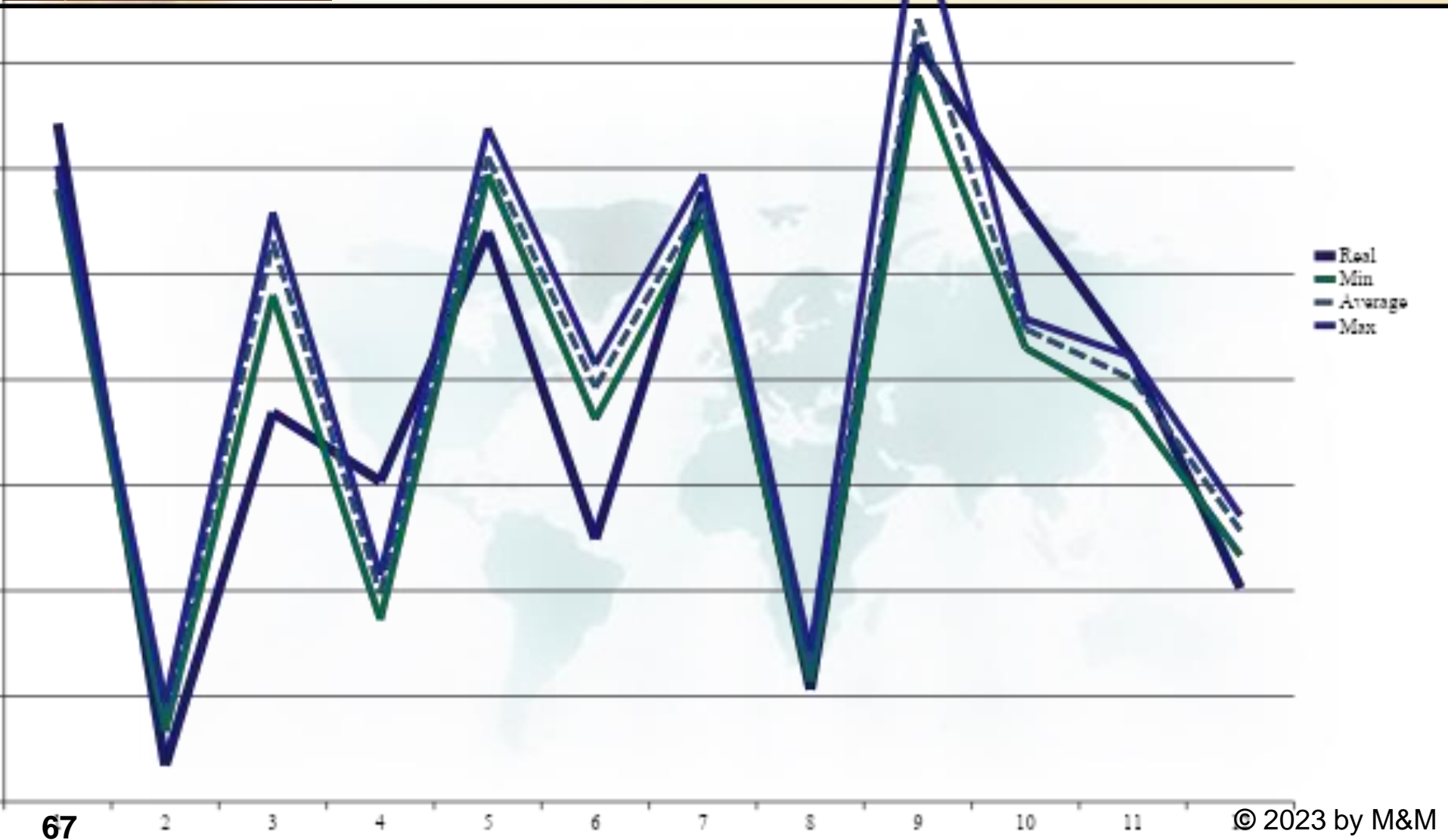
- **Mean Absolute Percentage Error (MAPE)**

$$\text{MAPE (\%)} = \frac{1}{N} \sum_{t=1}^N (|e_t| / y_t) \times 100$$

- **Coefficient of Variation of the RMSE, CV(RMSE)**

$$\text{CV(RMSE)} = \text{RMSE} / \bar{y}$$

Model Selection – a Multiple Criteria Approach (Prediction Intervals)

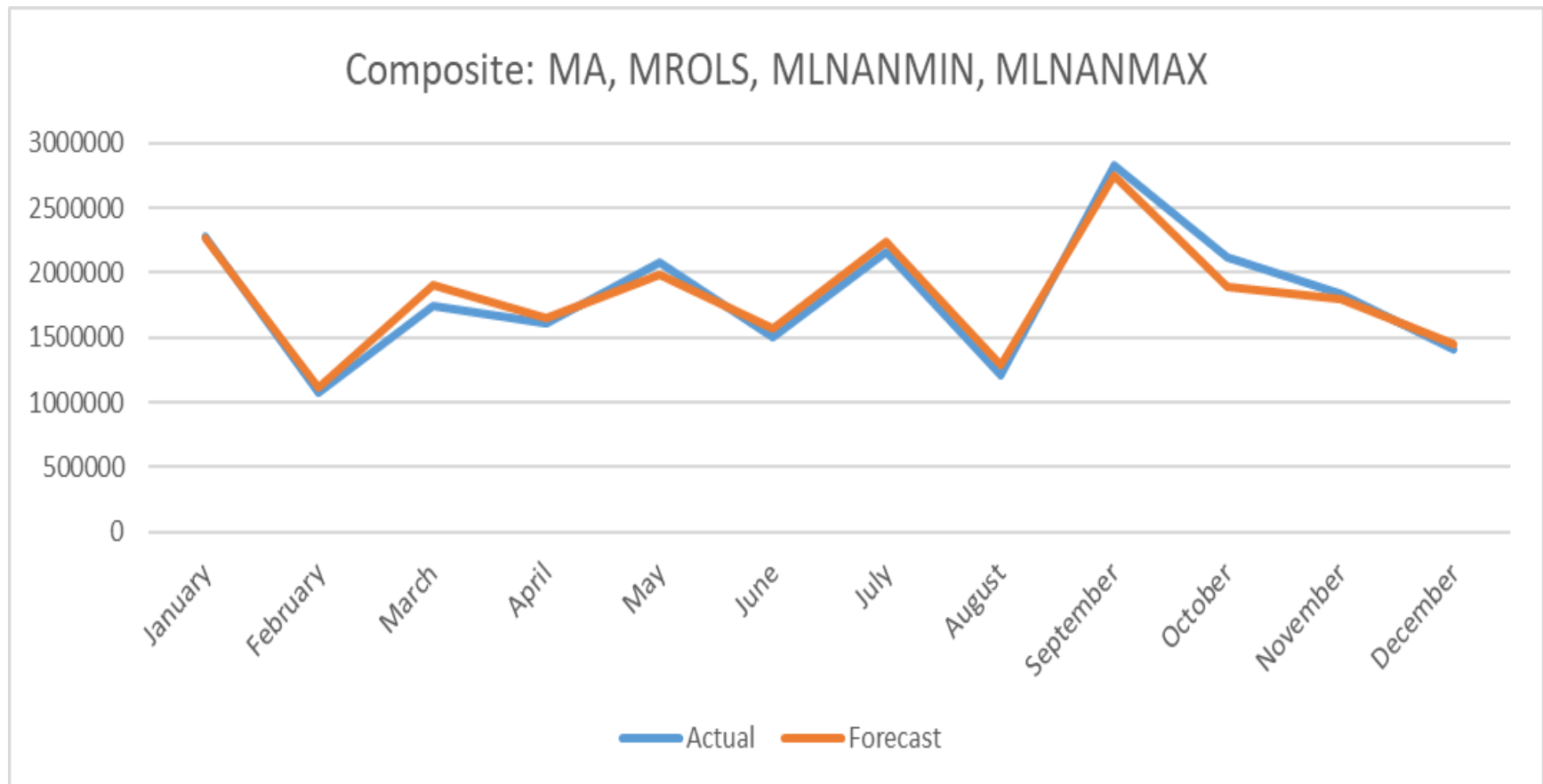


STATISTICAL LEARNING NETWORKS

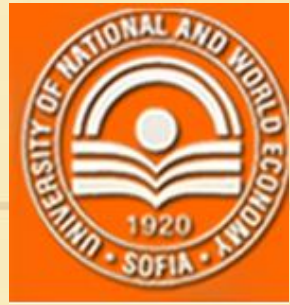
Best Model Selection



Experimental Test Results – 2021, WWU



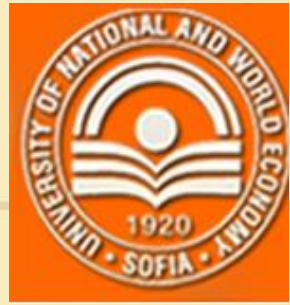
Thank You!



Questions?



Thank You!



and I'll

See You again...

