Walla Walla University

Retired

North American Simulation and Gaming Association

NASAGA

www.Mihail.Motzev.com

University of National and World Economy 1920 Sofia

Bulgarian Academic Simulation And Gaming Association

ISAGA

The International Simulation and Gaming Association

ABSEL

Association for Business Simulation and Experiential Learning

VANGUARD SCIENTIFIC INSTRUMENTS IN MANAGEMENT VSIM'2024:

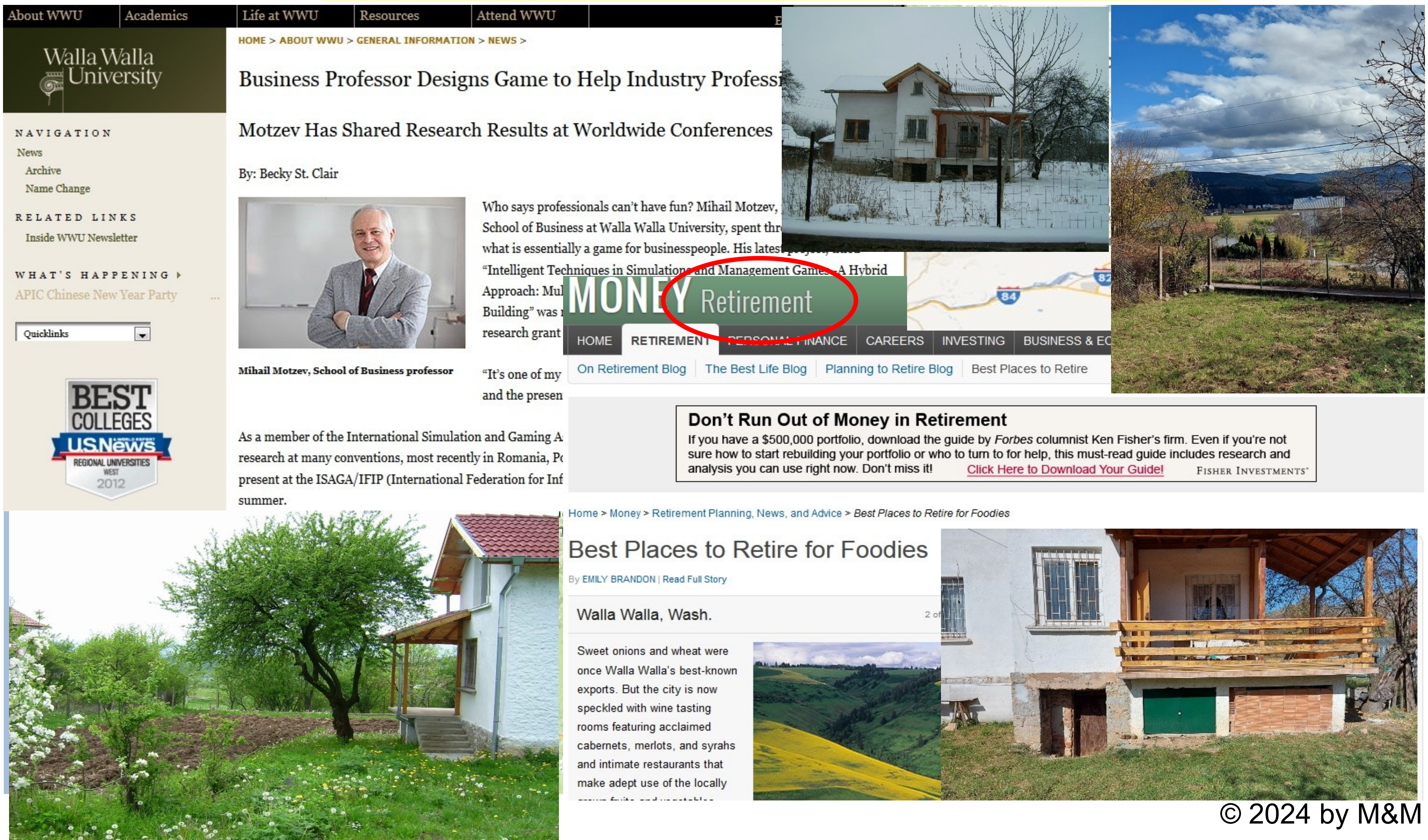
"COMPOSITE CRITERIA FOR MODEL SELECTION AND VALIDATION IN STATISTICAL LEARNING NETWORKS"

Mihail Motzev

Ph.D,M.Sc,P.D.D

(MRMotzev@yahoo.com)

The Best Place to Retire - for Me …



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

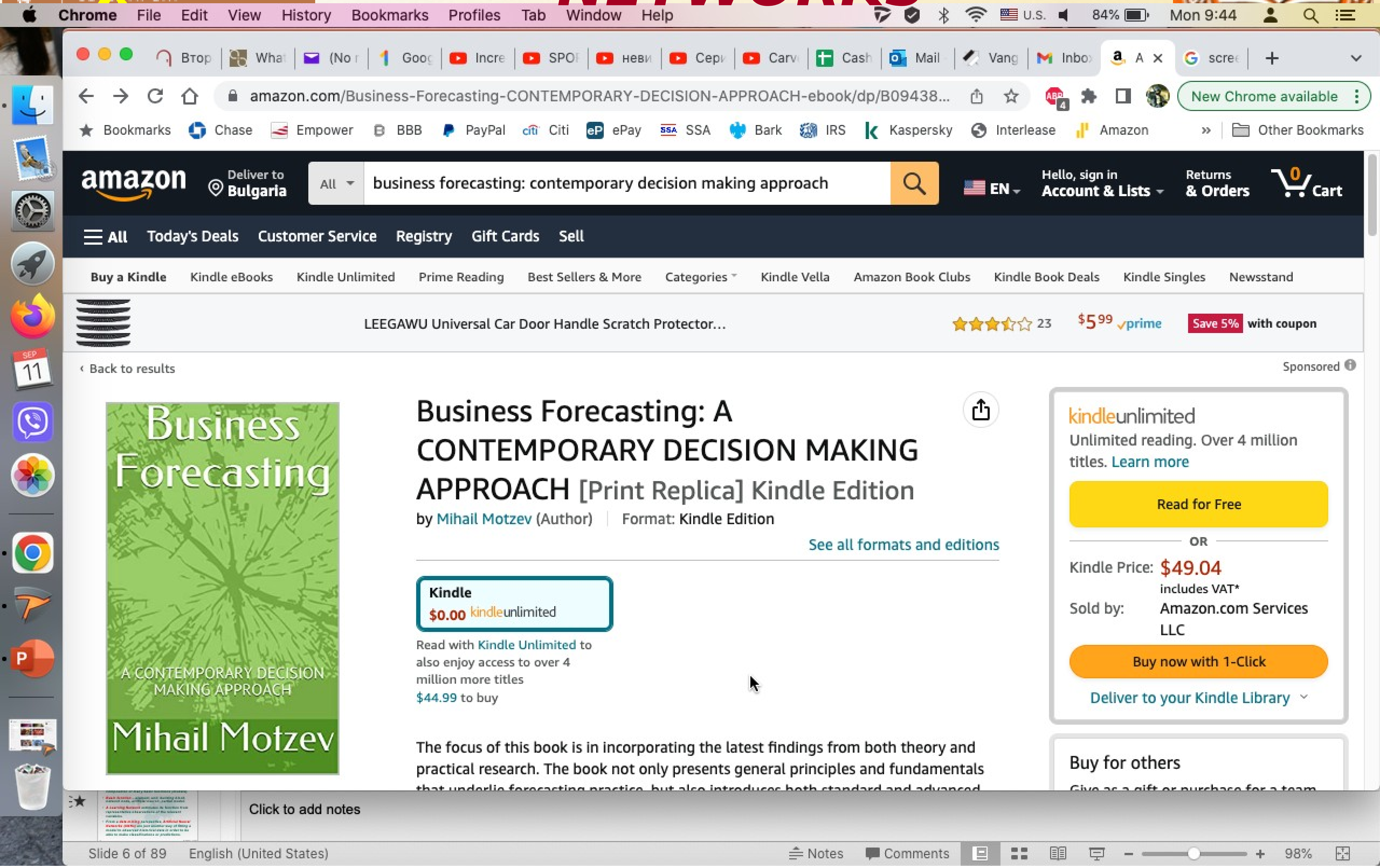
Fundamentals:

- *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
Fundamentals:

- ***Machine Learning (ML)*** *is an umbrella term for solving problems for which development of algorithms by human programmers would be cost-prohibitive, and instead the problems are solved by helping machines 'discover' their 'own' algorithms, without needing to be explicitly told what to do by any human-developed algorithms.*
- ***ML*** *is also known in its application across business problems as* ***predictive analytics****. Although not all* ***ML*** *is statistically-based, computational statistics is an important source of the field's methods.*
- *The term was coined in 1959 by Arthur Samuel (IBM)*

STATISTICAL LEARNING NETWORKS



STATISTICAL LEARNING NETWORKS

Predictive analytics and Statistical learning

- ***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** techniques such as **hidden Markov models** 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.*

Fundamentals:

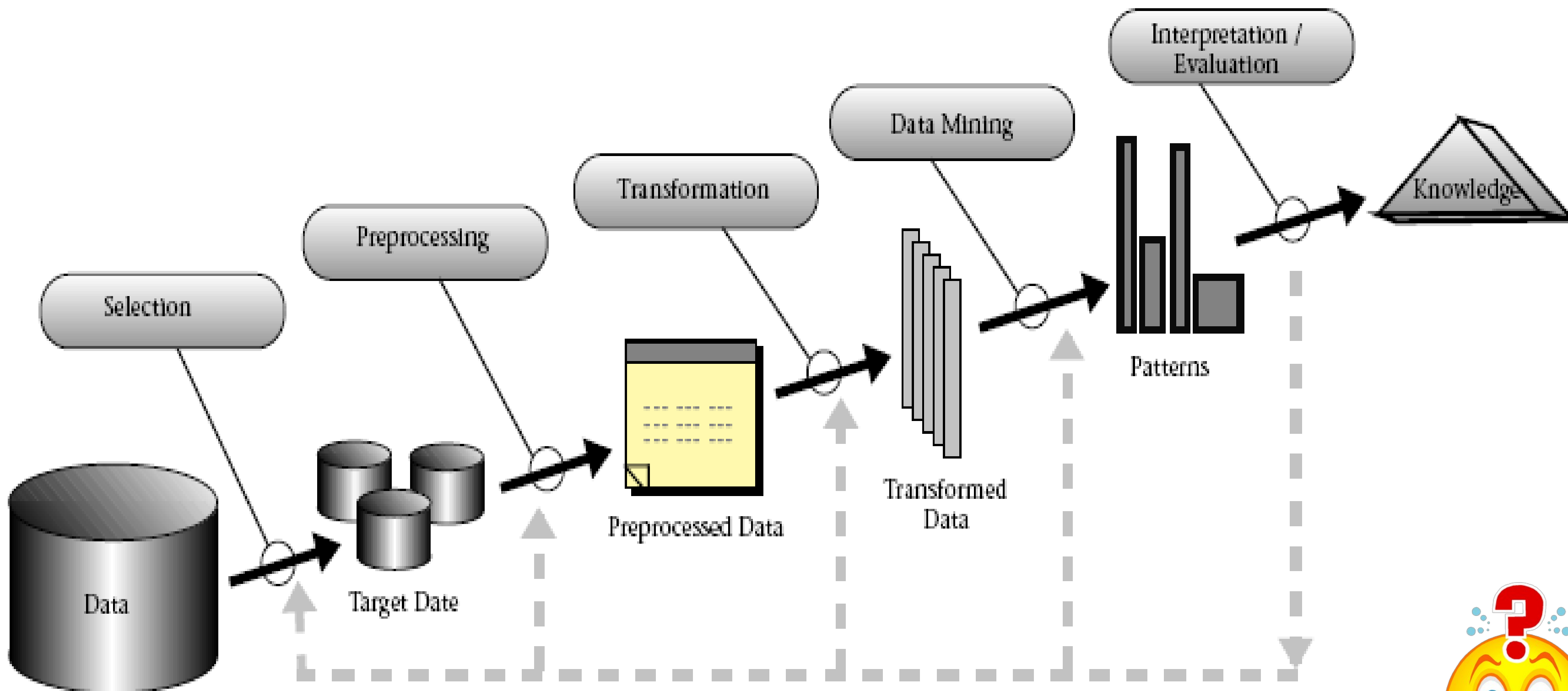
- ***Network* – a function (model) represented by the composition of many basic functions (models).**
- ***Basic function* – element, unit, building block, network node, artificial neuron, partial model.**
- ***A Learning Network* estimates its function from representative observations of the relevant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

- ***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.**

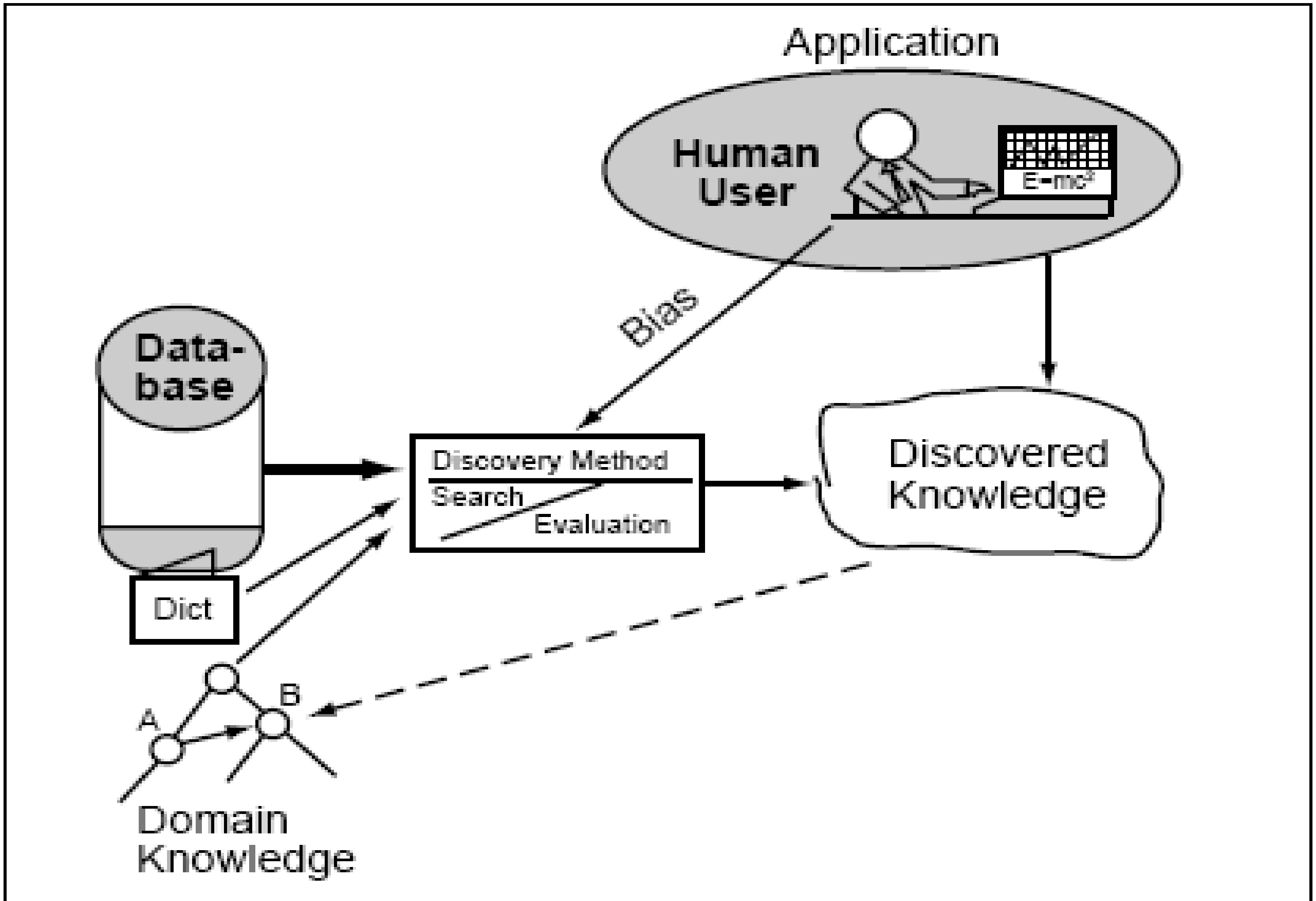
Knowledge Discovery in Databases –

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

An overview of KDD process



A Framework for Knowledge Discovery in Databases



Data Mining

- ***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.

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.

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.

- ***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.**

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:

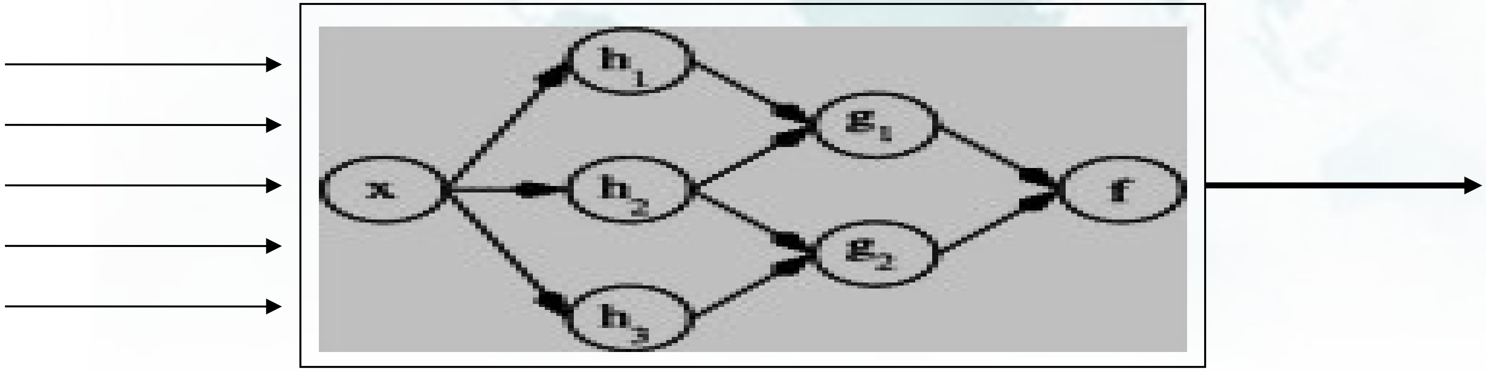
- ***Affinity grouping or association rules***: finding a model that describes significant dependencies between variables;
- ***Clustering***: identifying a finite set of categories or clusters to describe the data;
- ***Description and visualization (summarization)***: finding a compact description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

Inputs **Output**



Data Mining Process

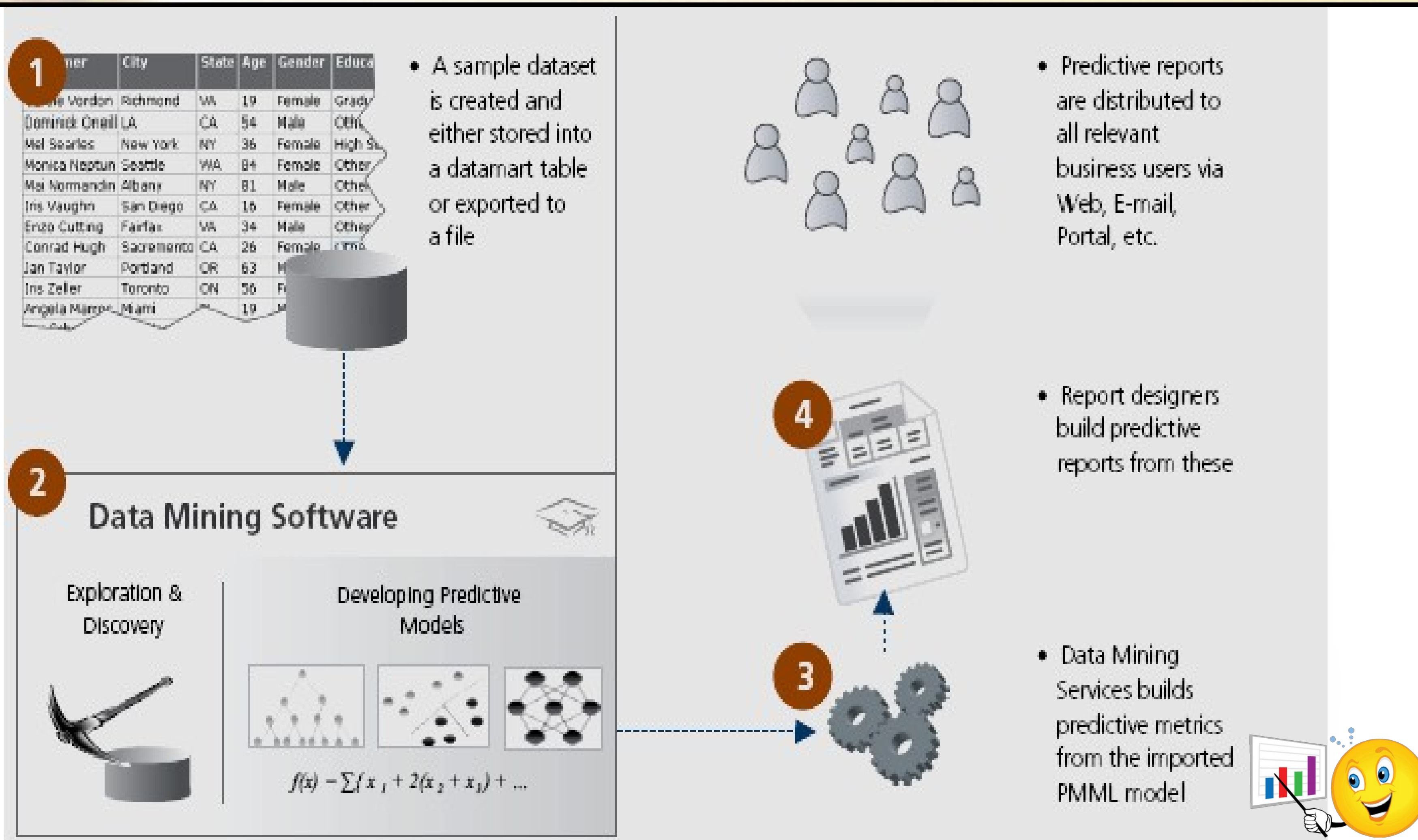
1. **Create a predictive model from a data sample**
2. **Train the model against datasets with known results**
3. **Apply the model against a new dataset with an unknown outcome (*cross-validation*)**

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

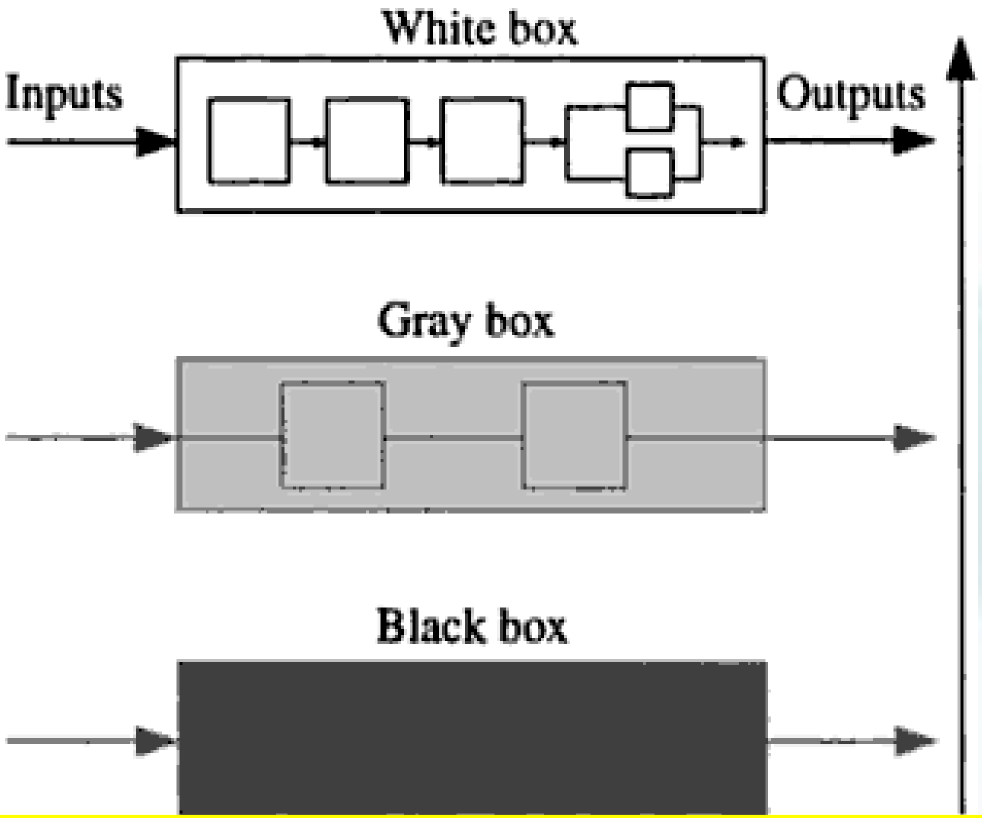
IBM Corp. has a slightly different interpretation of the data mining process and other companies may have their own view as well.

STATISTICAL LEARNING NETWORKS

DM Workflow in MicroStrategy platform



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

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.

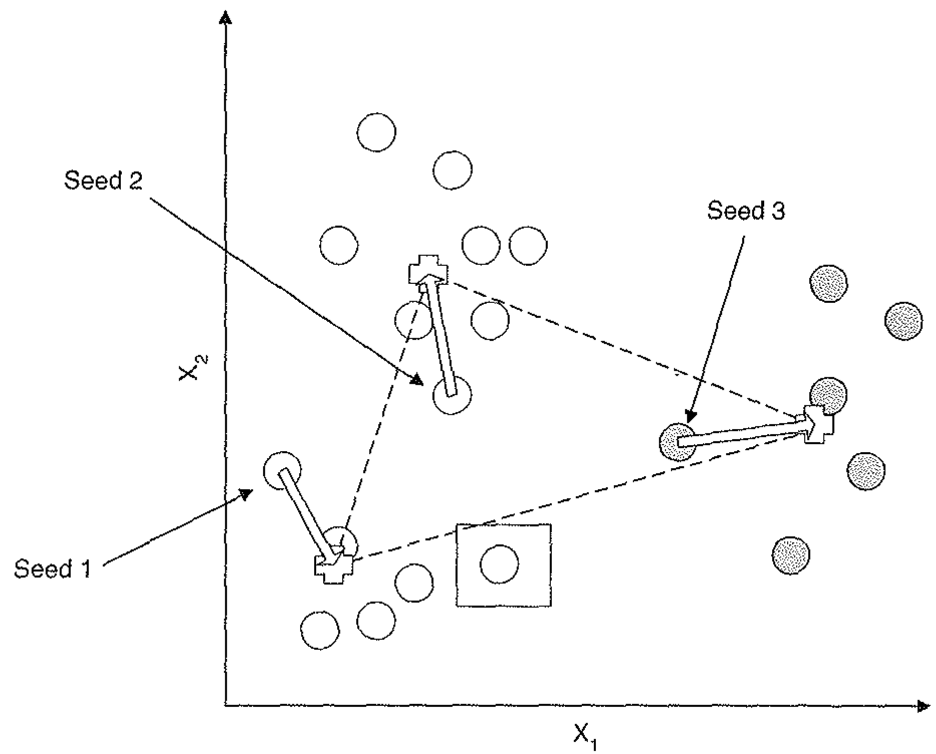
3. ***Data Mining Techniques*** – *clustering, decision trees and neural networks.*

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.

- ***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.**

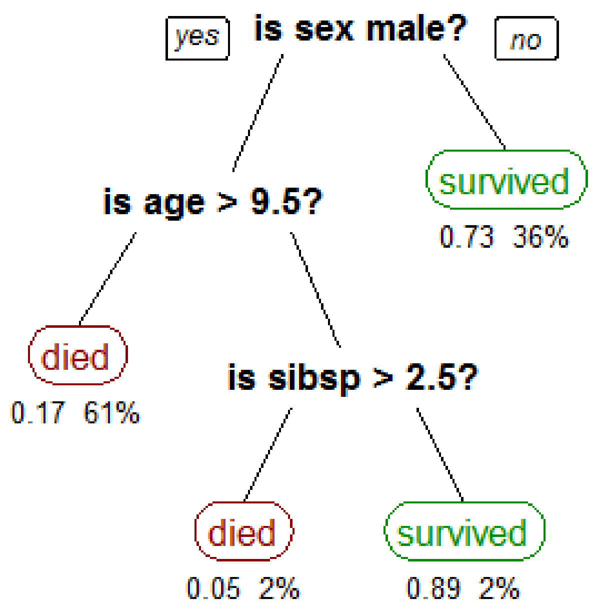
Automatic Cluster Detection



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

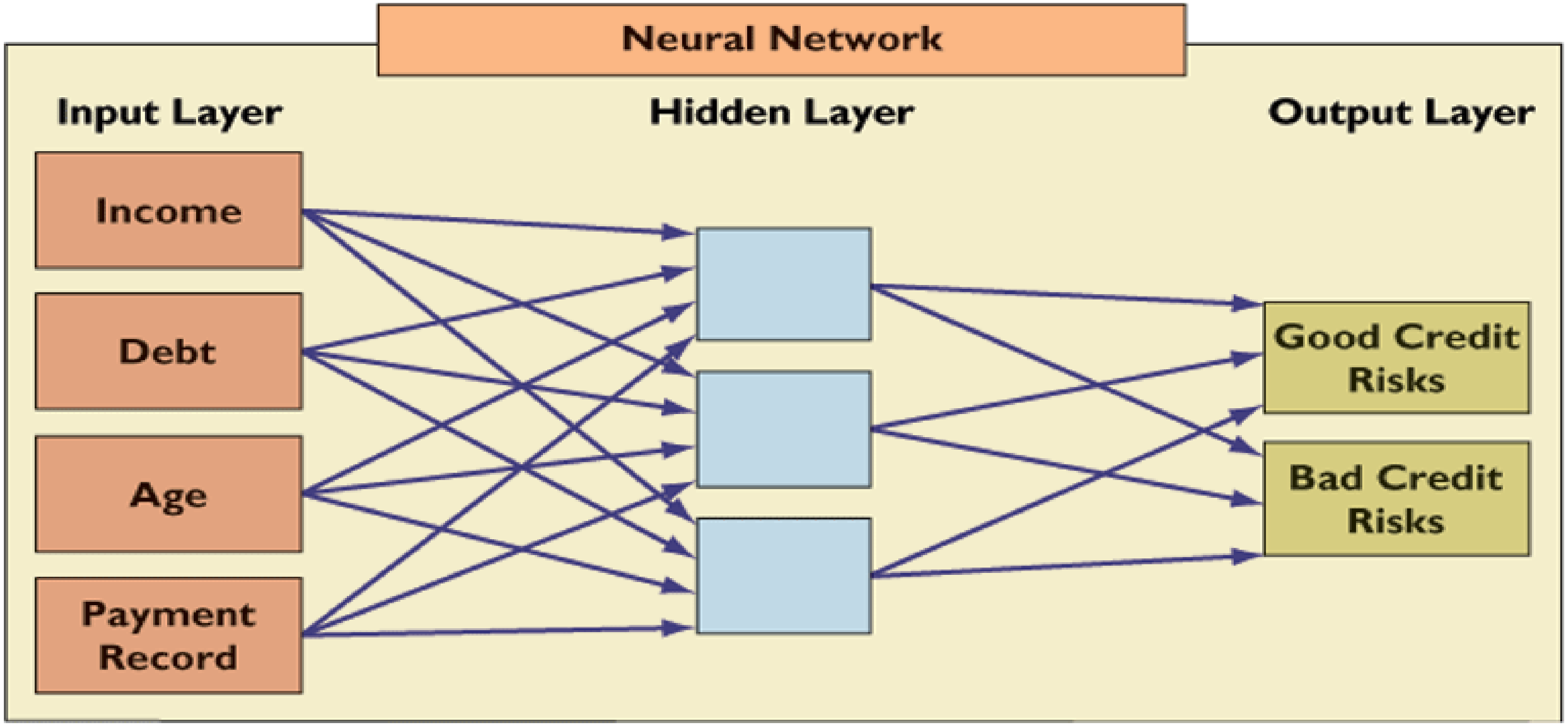


A tree showing survival of passengers on the Titanic ("sibsp" is the number of spouses or siblings aboard). The figures under the leaves show the probability of survival and the percentage of observations in the leaf

Artificial Neural Networks

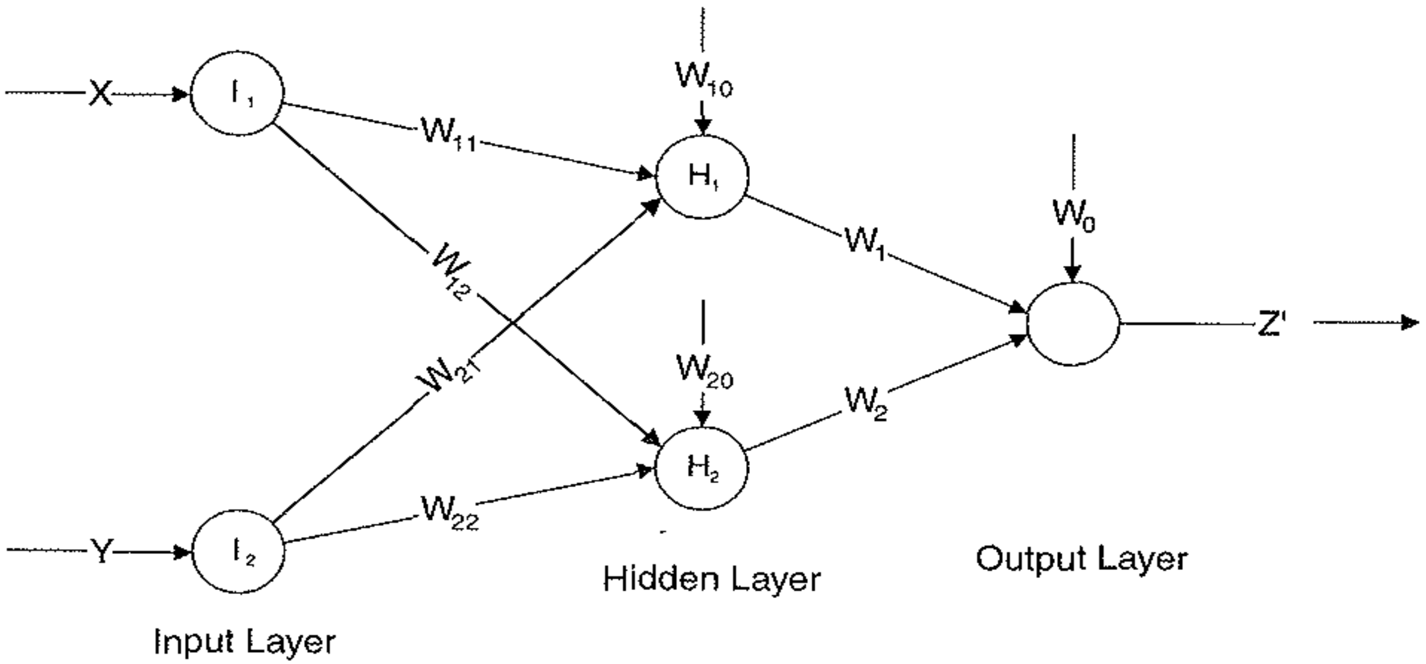
Data Mining Techniques

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.

DM Techniques - ANNs

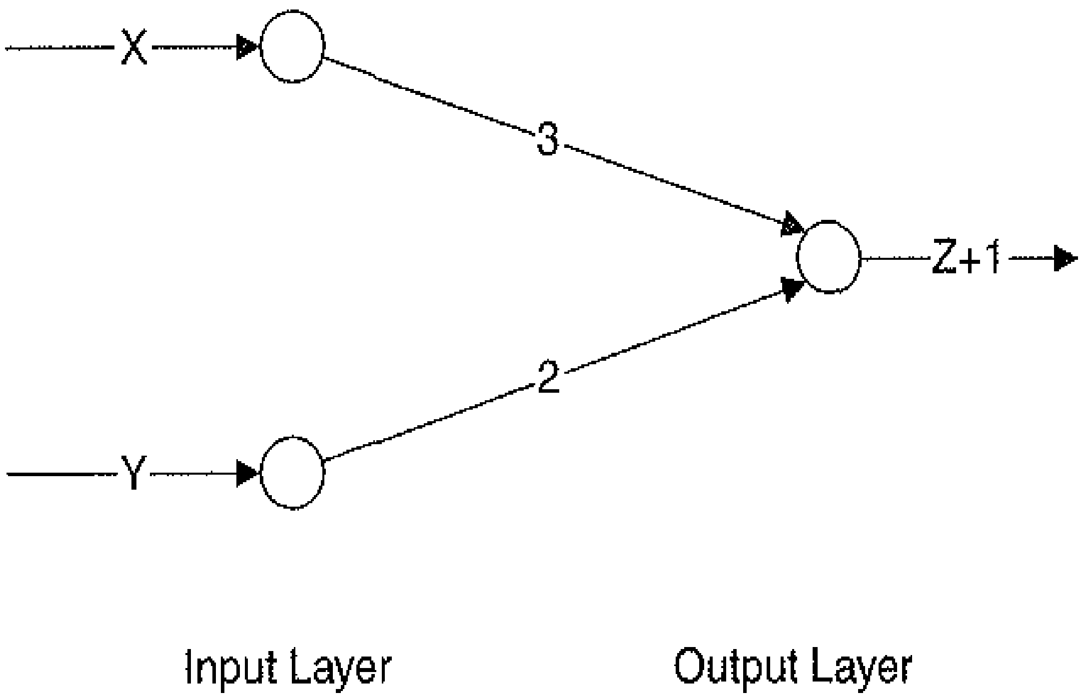


A neural network with a hidden layer.

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

How a Neural Network Works

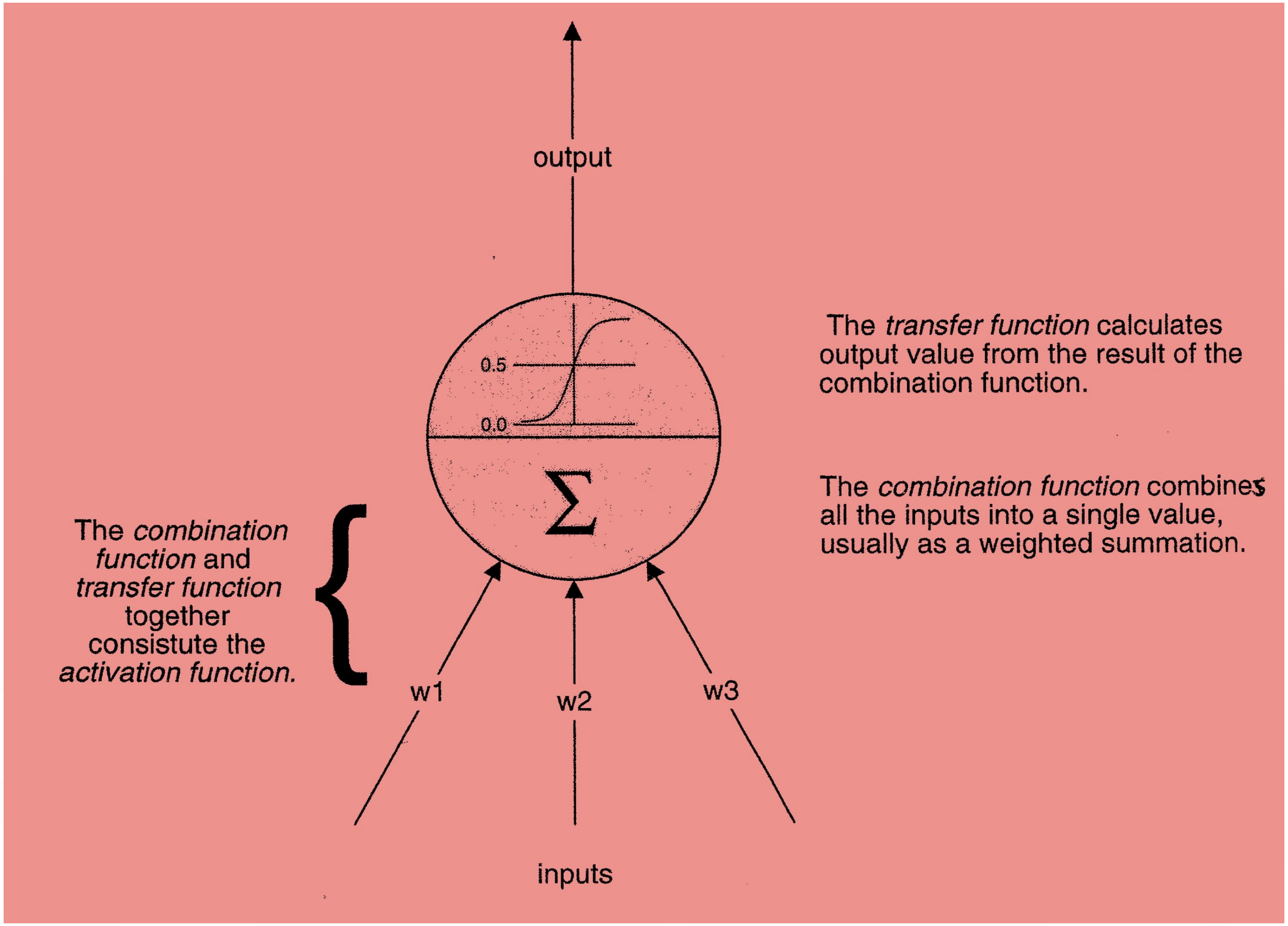
Simple ANN – one hidden layer



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

How a Neural Network Works

Linear Transfer Function

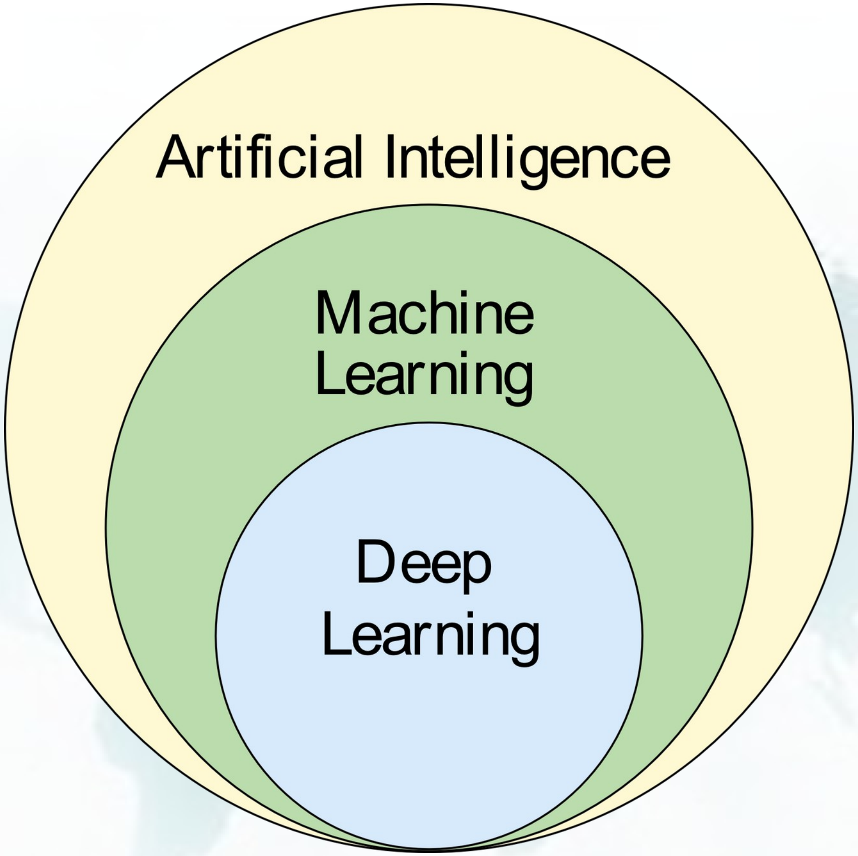


How a Neural Network Works

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.

Deep Learning



Source: https://en.wikipedia.org/wiki/Machine_learning#cite_note-journalimcms.org-22

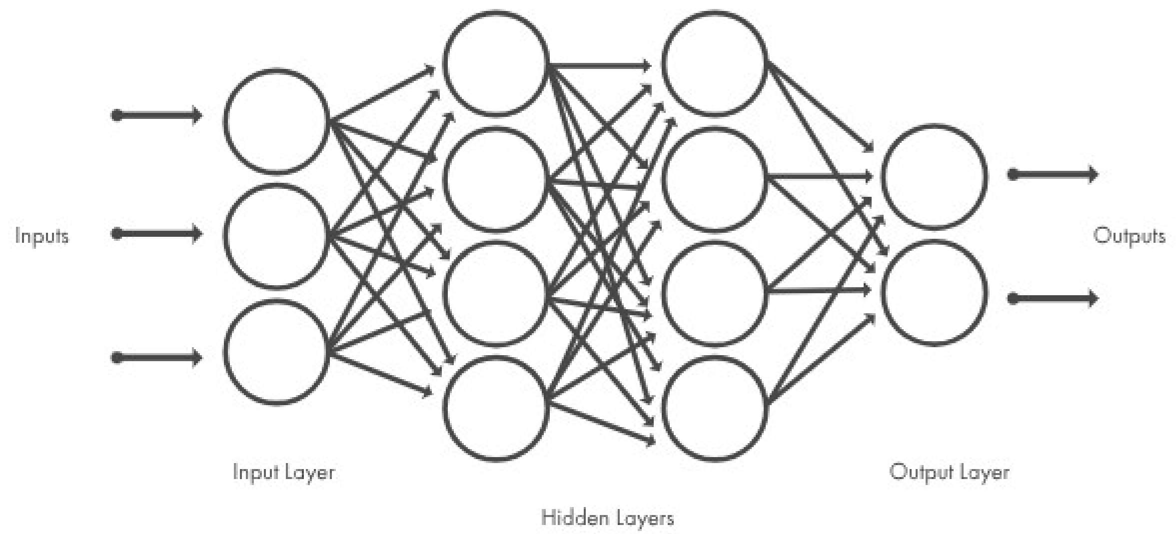
Deep Learning Networks

- ***Deep Learning*** *is a part of a broader family of ML methods, which is based on artificial neural networks with representation (feature) learning. The adjective "deep" in deep learning refers to the use of multiple layers in the network. Methods used can be either supervised, semi-supervised or unsupervised.*
- ***Supervised feature learning*** *- set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data. Examples include* ***supervised neural networks, multilayered perceptron*** *etc.*

	ANNs	Statistical Learning Networks
Data analysis	universal approximator	structure identificator
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

Deep Learning and DNN – many hidden layers

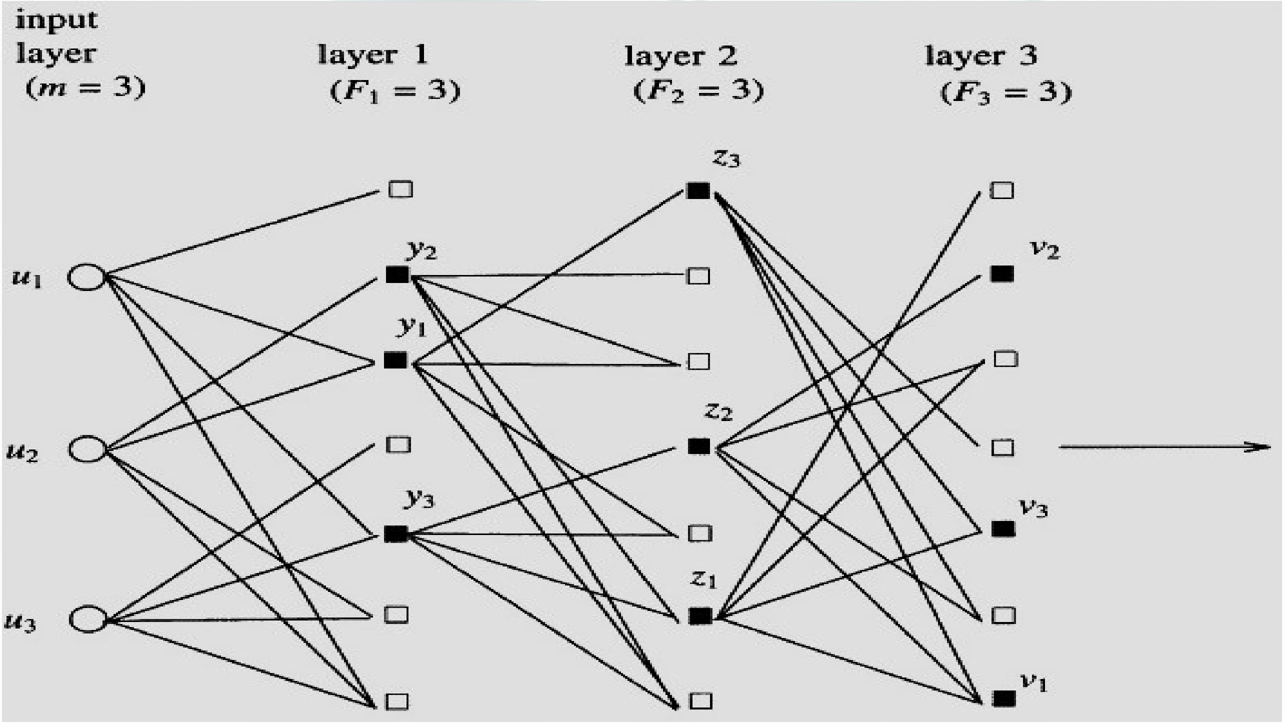


Deep learning attempts to model high-level abstractions in data by using model architectures composed of multiple non-linear transformations. Many of the most successful deep learning methods involve the ANNs where a ***Deep Neural Network (DNN)*** is defined to be an artificial neural network with multiple hidden layers of units between the input and output layers.

STATISTICAL LEARNING NETWORKS

Deep Neural Network (DNN)

"The first general, working learning algorithm for supervised, deep, feedforward, multilayer perceptron(s) was published by Alexey Ivakhnenko and Lapa in 1967"
(https://en.wikipedia.org/wiki/Deep_learning >History)

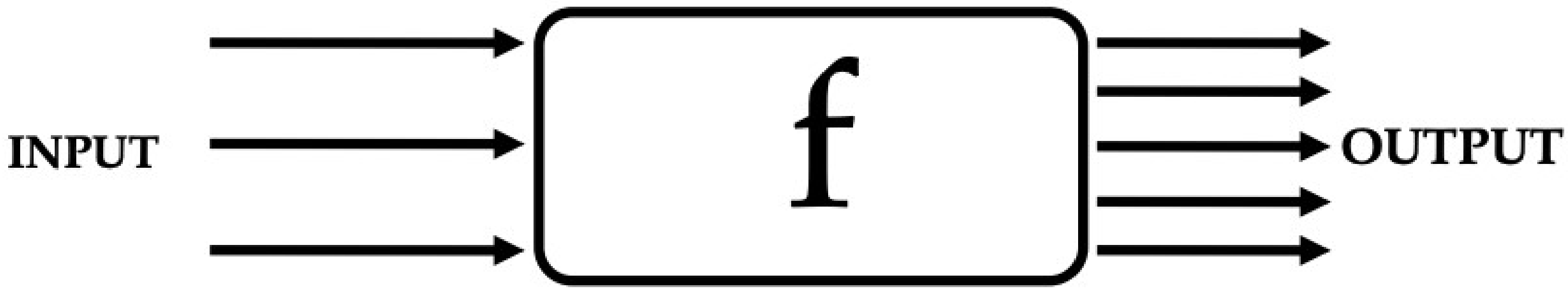


 Source: Madala & Ivakhnenko, 1994, p.39

Deep Learning and Statistical Learning Networks

- The main advantages of ***DNNs*** are that they make it possible to build faster and more accurate simulation models but at the same time DNNs are difficult to develop and hard to understand.
- ***Statistical Learning Networks (SLNs)*** can address the common problems of DNNs such as:
  - difficulties in interpretation of the results (DNNs are implicit models with no explanation component by default),
  - the problem of overfitting, designing DNN topology – it is in general a trial-and-error process, and
  - there are no rules how to use the theoretical a priori knowledge in DNN design, etc.

Statistical Learning Theory: supervised learning



Given a set of l examples (data)

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

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.

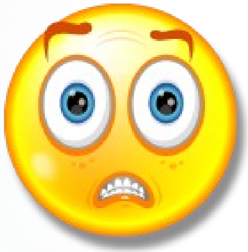
General Prediction Model

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

Where:

$X(x_1, x_2, \ldots, x_M)$ - input variables vector;

$A(a_1, a_2, \ldots, a_M)$ - 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).***

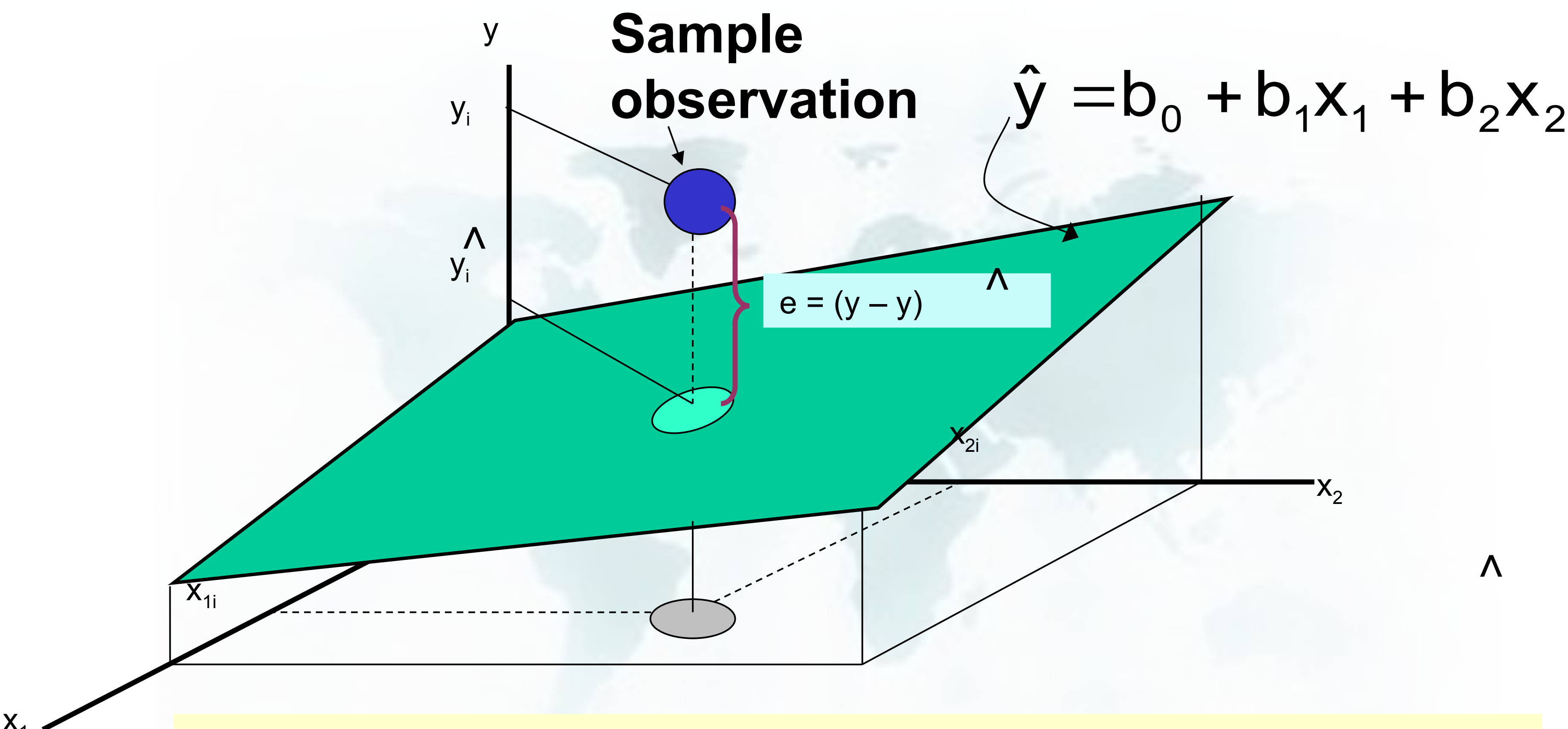
Model Building

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

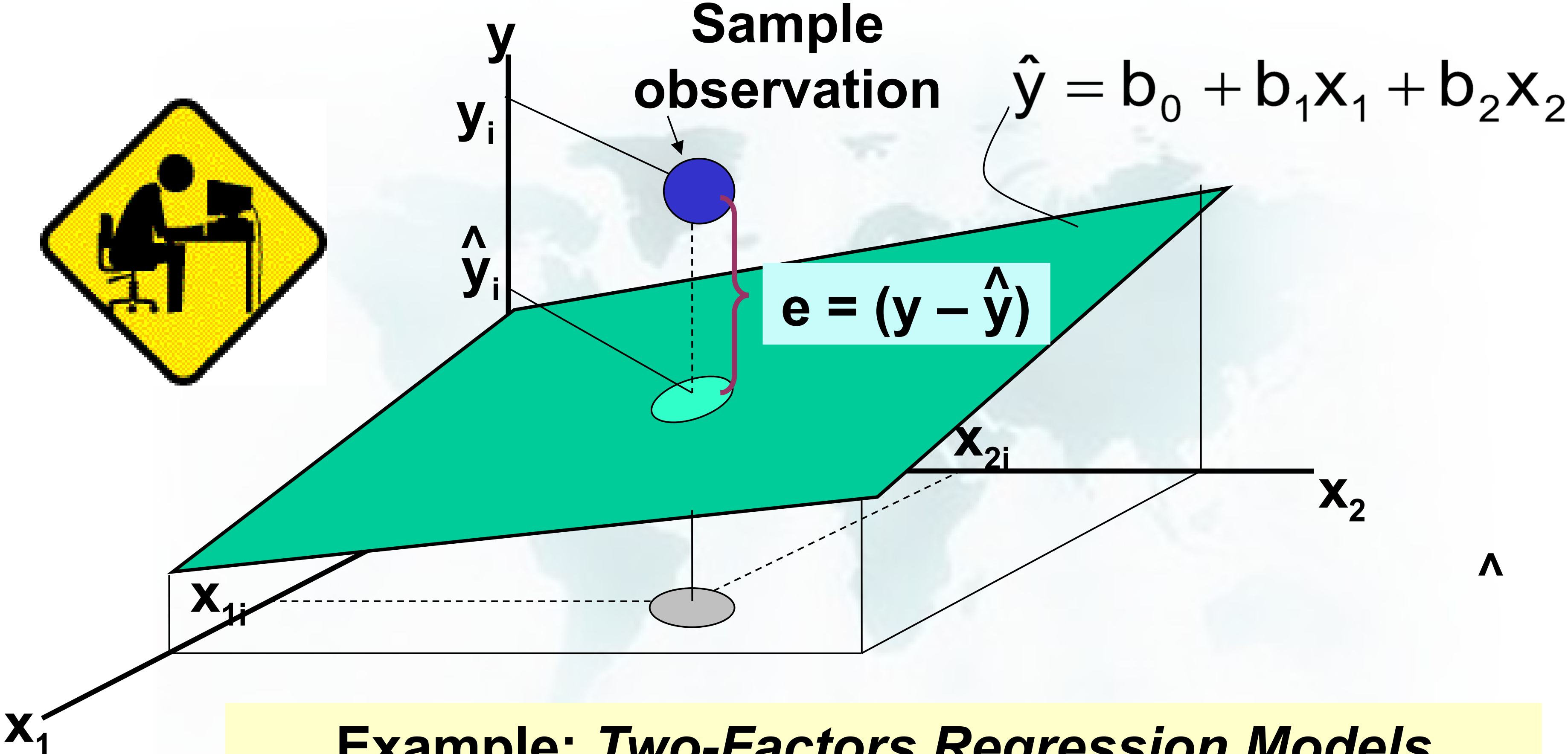
Regression Models



Ordinary Least Squares: ***Minimizing the MSE (e^2)***

Model Building

Regression Analysis



Example: *Two-Factors Regression Models*

Ordinary Least Squares:

Minimizing the MSE (e^2)

- The OLS estimator is consistent when the regressors are exogenous, and
- by the Gauss–Markov theorem, optimal in the class of linear unbiased estimators when the errors are homoscedastic and serially uncorrelated.
- Under these conditions, the method of OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances.
- Under the additional assumption that the errors are normally distributed, OLS is the maximum likelihood estimator.

Model Building

Regression Models – Problems:

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

APPENDICES

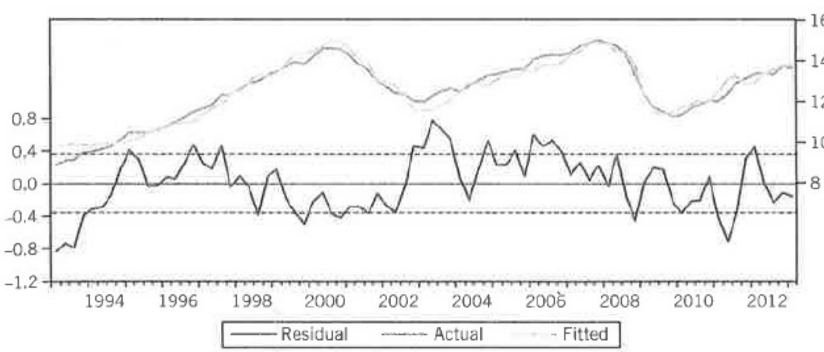
Exhibit 4.7

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 covarian

Exhibit 4.6

Dependent Variable (Time Period: Q1 1970–Q4 2012, 172 obs.)
ln [Real GDP / Real GDP (4 quarters ago)]

Independent Variable(s)	Coefficient	t-Statistic*
ln [**Corp & Home Equity, Period Avg (1 quarter ago) / **Corp & Home Equity, Period Avg (5 quarters ago)]	0.127	9.691

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.

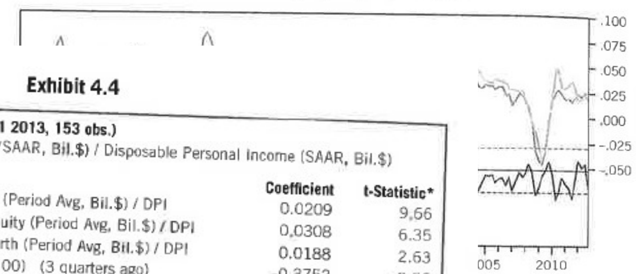


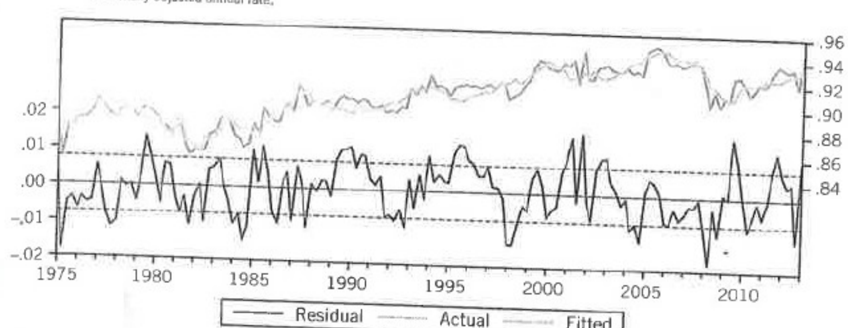
Exhibit 4.4

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.$) / DPI	0.0209	9.66
Household (incl. NPOs) Homeowners' Equity (Period Avg, Bil.$) / 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)	0.2666	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) + (0.5*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

y = 0.00001 + b + c

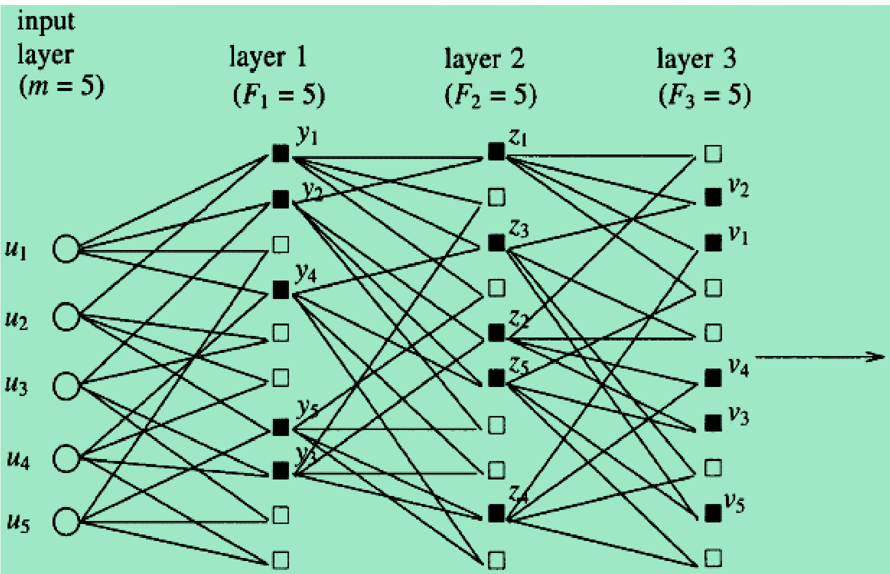
y = 9- 0.0319a+ 0.0319c

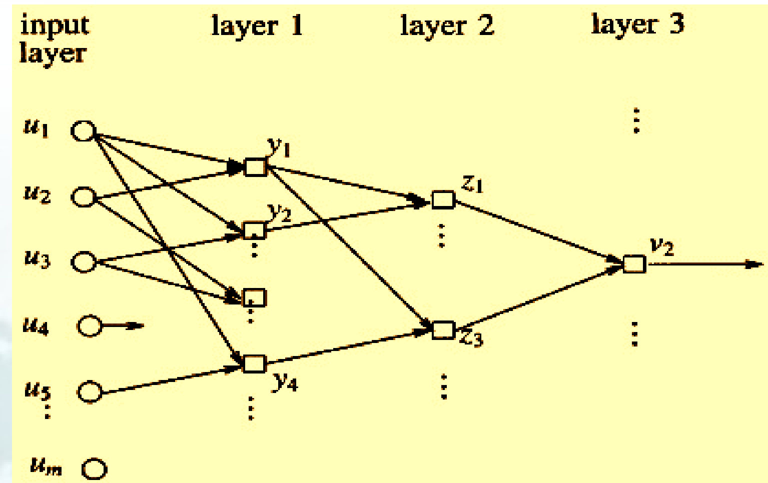
Source: Mueller and Lemke (2003 p. 18)

Model Building

Multilayered Nets of Active Neurons

Multilayered network structure with five input arguments and selected nodes:

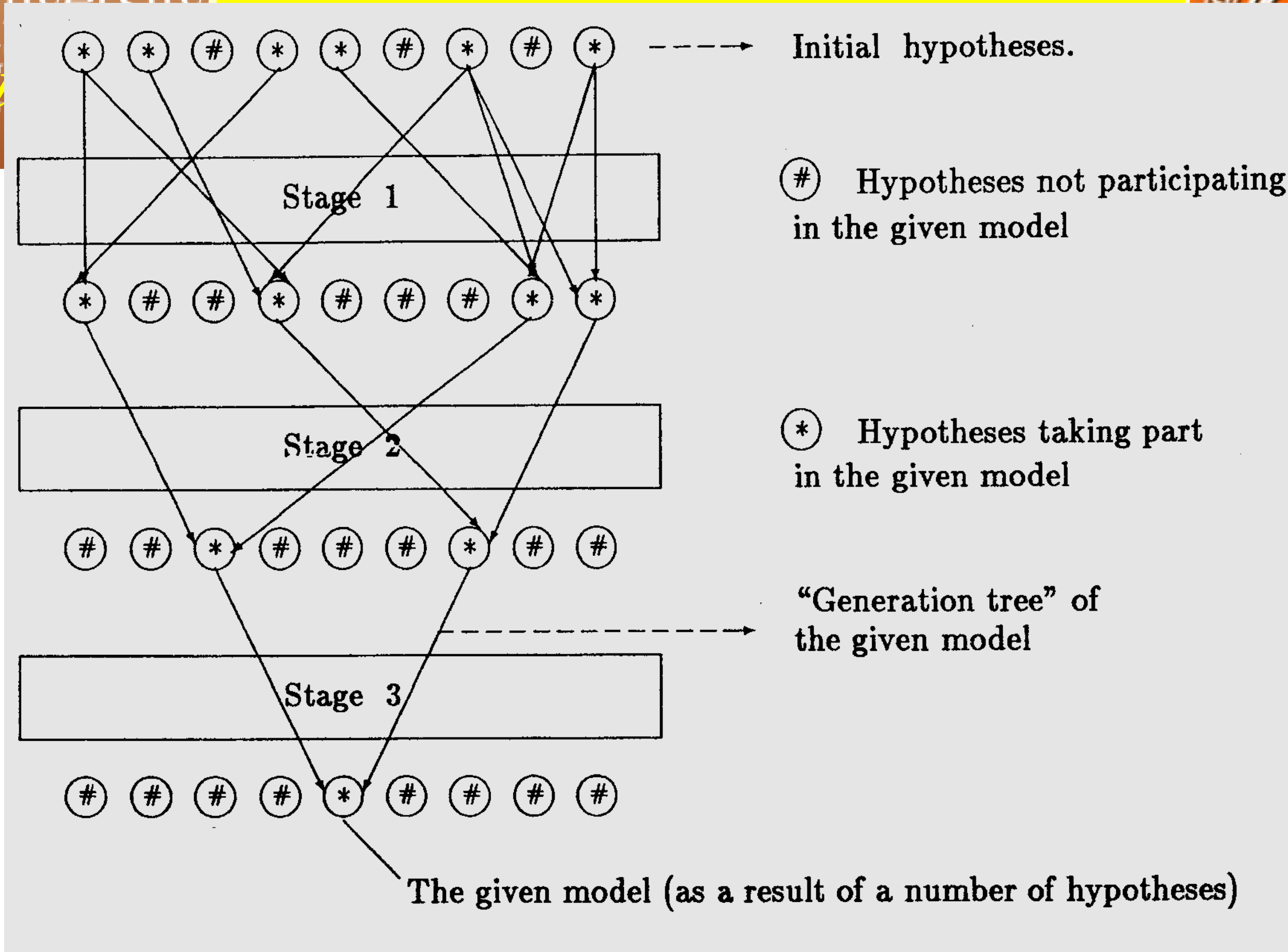




Multilayered network structure representing the output flow to unit 2 of layer 3

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

Multi-Stage Selection Algorithm



Source: *1. IIASA International Workshop on Methodology and Software for Interactive Decision Support, Albena, Bulgaria, October, 1987;*
2. XII IMACS World Congress, Paris, France, July, 1988.

STATISTICAL LEARNING NETWORKS

Multilayered Nets of Active Neurons – Main Pillars

GMDH

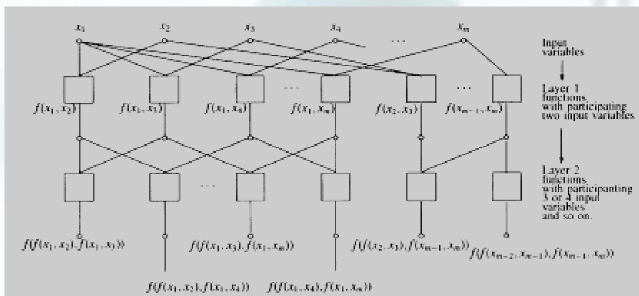
Gödel's incompleteness theorems

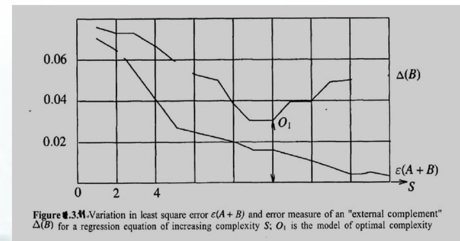
Genetic selection of pairwise features

Alexey G. Ivakhnenko.

(1913-2007)

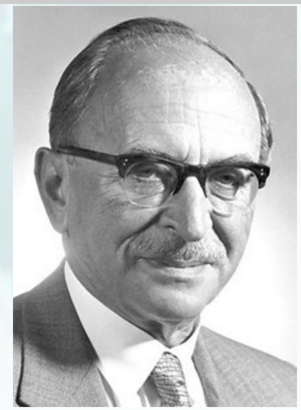
Two State Prizes of the USSR, Medal “For Labor”, Order of Friendship of Peoples …





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)

Kurt Gödel

(1906-1978)

Notable awards:

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

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**

Walla Walla University Seventh-day Adventist Higher Education Retired

Gödel's Incompleteness Theorems: Two theorems of mathematical logic that are concerned with the limits of provability in formal axiomatic theories

University of National and World Economy 1920 Sofia

- **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**

Gödel's Incompleteness Theorems: Two theorems of mathematical logic that are concerned with the limits of provability in formal axiomatic theories

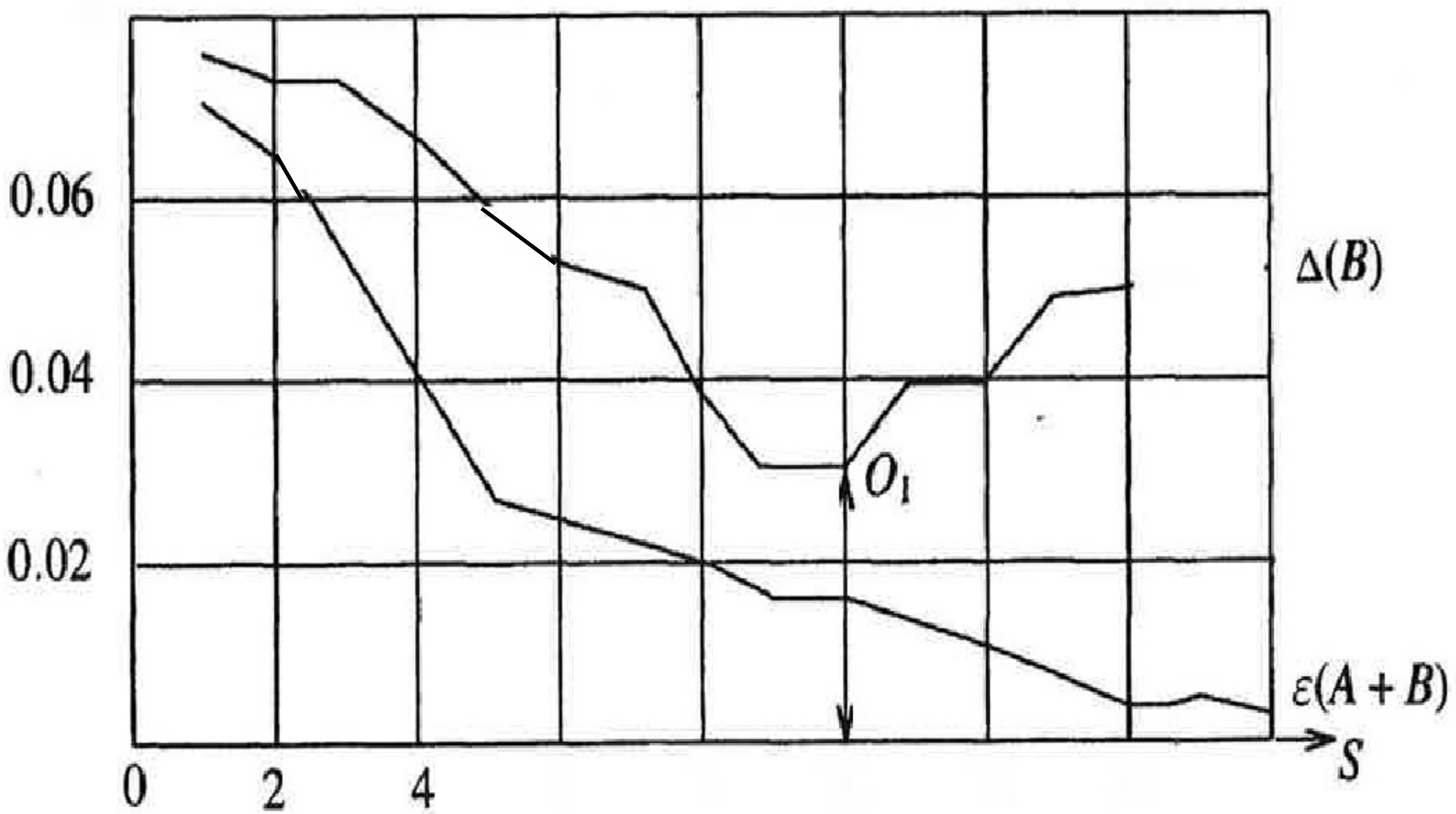
- ***The first incompleteness theorem*** **states that no consistent system of axioms whose theorems can be listed by an effective procedure (i.e., an algorithm) is capable of proving all truths about the arithmetic of natural numbers. For any such consistent formal system, there will always be statements 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 {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 Cons(F) expressing the consistency of F. Gödel's second incompleteness theorem shows that, under**

- **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 ε0 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 ε0", 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

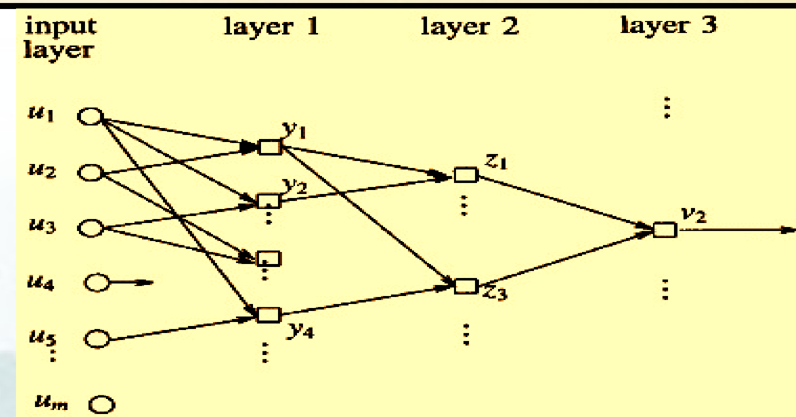


Variation in least square error $\varepsilon(A+B)$ and error measure of an "external complement" $\Delta(B)$ for a regression equation of increasing complexity S; O_1 is the model of optimal complexity

Statistical Learning
Networks of Active Neurons

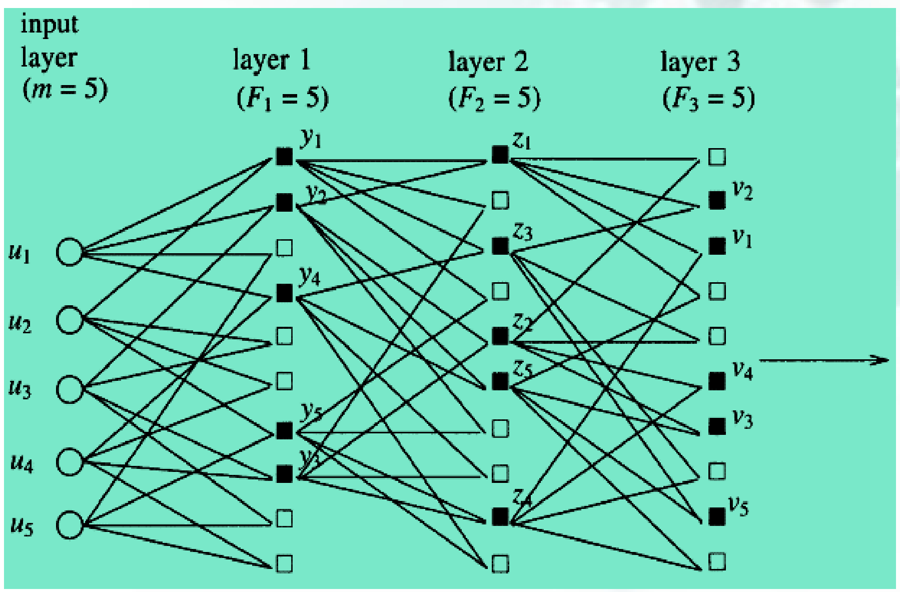
Multilayered Net of Active Neurons (MLNAN)

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.



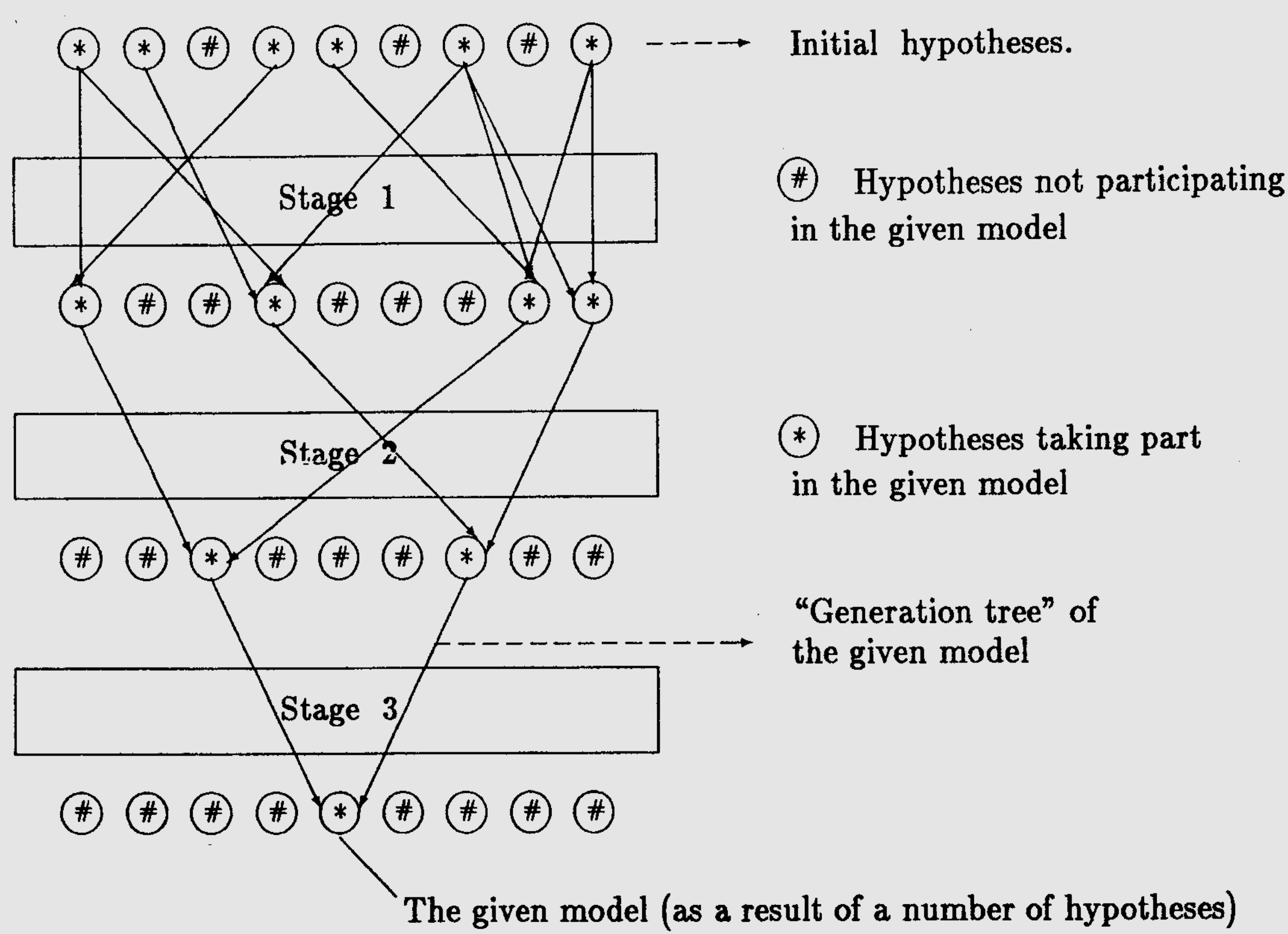
Multilayer network structure representing the output flow to unit 2 of layer 3

This method grows a tree-like network out of data of input and output variables in a pair-wise 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.

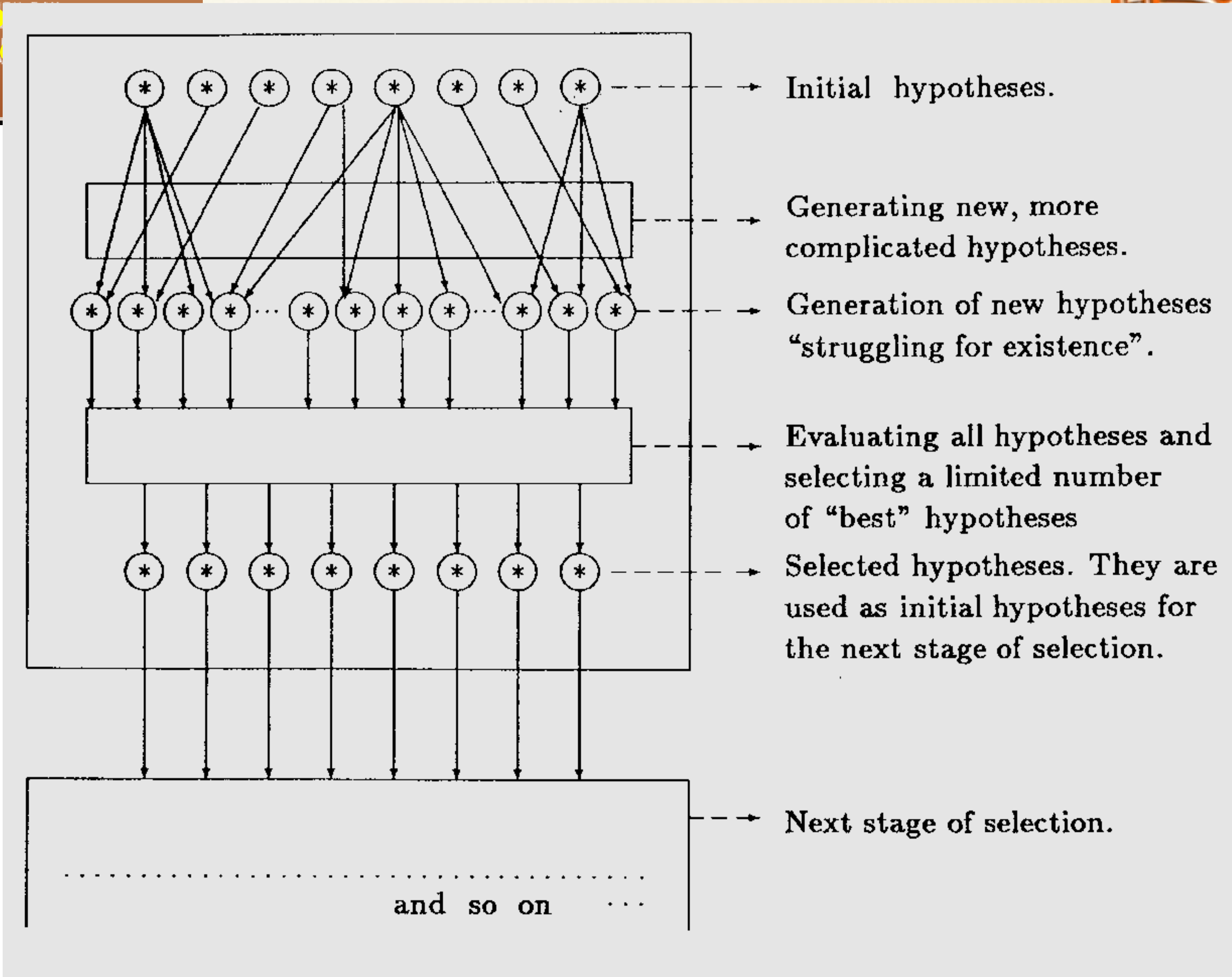


Multilayer network structure with five input arguments and selected nodes:

Multi-Stage Selection Algorithm

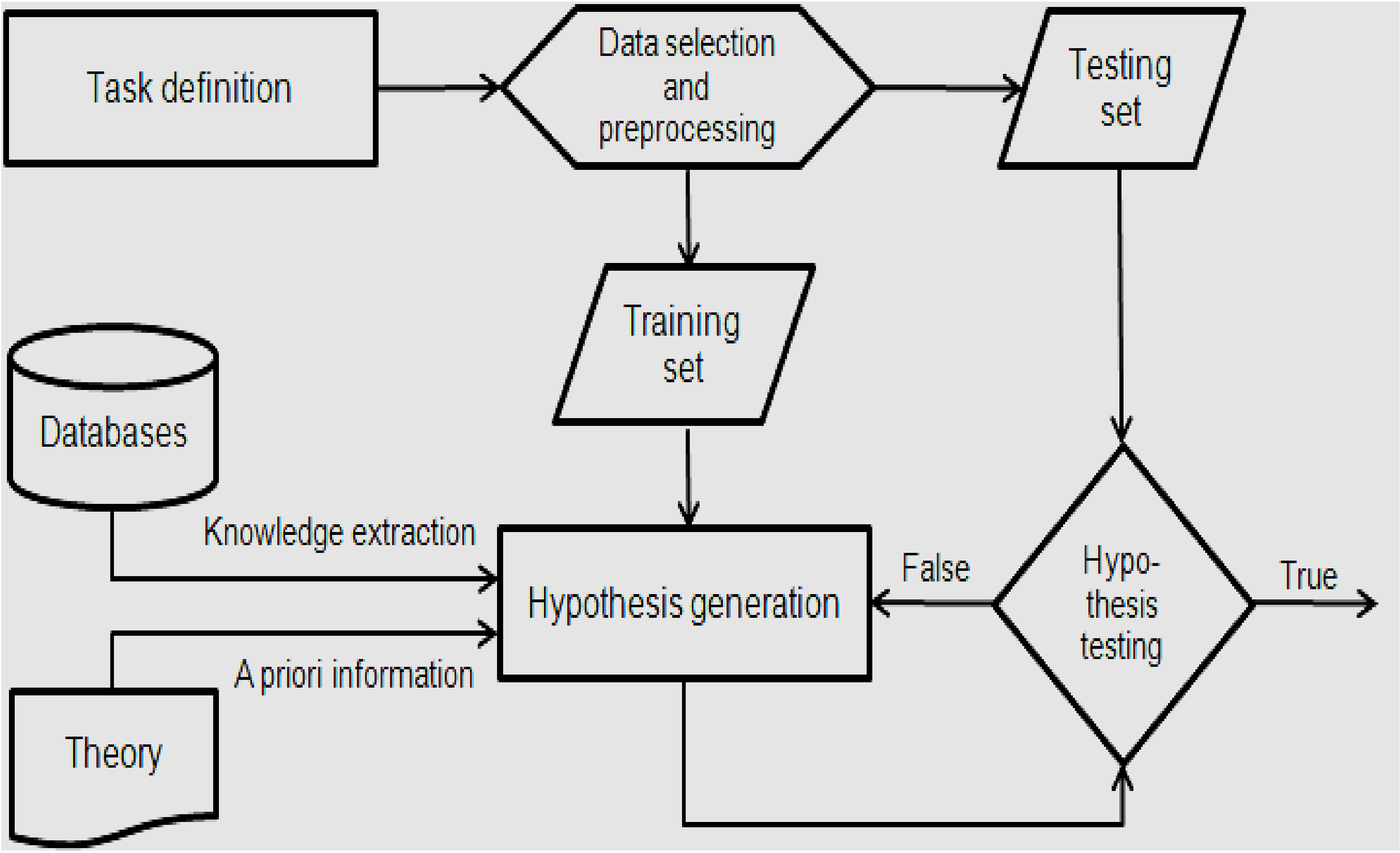


Pair-Wise Selection Using External Criteria



Model Selection

Cross Validation and a-priori information



- ***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.

This PC > Local Disk (C:) > MRM > Research > Books&Papers

- Quick access
  - Desktop
  - Downloads
  - Documents
  - Pictures
  - 2022
  - Books&Papers
  - Pictures
  - Summer
- OneDrive - Personal
  - Case
  - Desktop
  - Documents
  - Email attachments
  - Forecasting
  - P-Drive
  - Pictures
  - TextBook
- This PC
- Network

Name	Date modified	Type	Size
2Step CV	9/7/2022 4:20 PM	Microsoft Edge P...	1,278 KB
A Predictive Approach to Model Selection	3/11/2010 2:31 AM	Microsoft Edge P...	995 KB
Asympt Optim for CV	9/7/2022 4:24 PM	Microsoft Edge P...	1,504 KB
Asymptotic Properties	3/11/2010 2:49 AM	Microsoft Edge P...	193 KB
Bias Correction in CV	9/7/2022 4:19 PM	Microsoft Edge P...	403 KB
Comp Study	9/7/2022 4:22 PM	Microsoft Edge P...	1,278 KB
Comparisson	9/7/2022 4:06 PM	Microsoft Edge P...	1,915 KB
Consist CV	9/7/2022 4:21 PM	Microsoft Edge P...	842 KB
Cross-Val-Correction	9/7/2022 4:08 PM	Microsoft Edge P...	1,364 KB
Cross-Validation of Regression Models	3/11/2010 2:51 AM	Microsoft Edge P...	1,211 KB
Cross-Validatory Choice and Assessment ...	9/7/2022 3:44 PM	Microsoft Edge P...	3,947 KB
Dependent Data CV	9/7/2022 4:22 PM	Microsoft Edge P...	1,061 KB
Godel's Theorem	9/7/2022 8:22 PM	Rich Text Format	50 KB
Improv of Cross-Val	9/7/2022 4:07 PM	Microsoft Edge P...	2,517 KB
Ivakhnenko_Lapa_Table of contents 1	9/2/2022 11:21 AM	JPG File	319 KB
Ivakhnenko_Lapa_Table of contents 2	9/2/2022 11:21 AM	JPG File	286 KB
Ivaknenko_polynomial	9/2/2022 11:02 AM	Microsoft Edge P...	746 KB
Lasso CV	9/7/2022 4:27 PM	Microsoft Edge P...	886 KB
Model Selection	9/7/2022 4:17 PM	Microsoft Edge P...	1,565 KB
Multifold CV	9/7/2022 4:18 PM	Microsoft Edge P...	1,072 KB
Note on General CV	9/7/2022 4:24 PM	Microsoft Edge P...	511 KB
Prediction Error	9/7/2022 4:05 PM	Microsoft Edge P...	4,904 KB
Pred-Interval	9/7/2022 4:09 PM	Microsoft Edge P...	1,413 KB
Sources in Cross Validation	6/26/2023 6:24 PM	Rich Text Format	51 KB
Stone_atall-Corrigendum	9/7/2022 4:15 PM	Microsoft Edge P...	331 KB
Stone1page1977Biom	3/11/2010 1:47 AM	GIF File	43 KB
Stone1974	9/7/2022 3:46 PM	JPG File	228 KB
Stone1974_Biometrica	9/7/2022 4:00 PM	Microsoft Edge P...	701 KB
Stone1974_Oxford	9/7/2022 4:05 PM	Microsoft Edge P...	710 KB
Stone1974_Wiley	9/7/2022 4:01 PM	Microsoft Edge P...	3,869 KB
Stone1977	9/7/2022 3:48 PM	JPG File	176 KB
Stone1977_Biometrics	9/7/2022 3:57 PM	Microsoft Edge P...	710 KB
Stone1977Biometrica	9/7/2022 3:54 PM	JPG File	233 KB
Stone2020_meryvn-stone-obituary	5/22/2023 10:10 AM	Microsoft Edge P...	112 KB
Stone-Akaike	9/7/2022 4:14 PM	Microsoft Edge P...	461 KB
Stone-Corrigenda	9/7/2022 4:12 PM	Microsoft Edge P...	247 KB

- Horst, P. (1941). Prediction of Personal Adjustment. New York: Social Science Research Council (Bulletin 48), found a "drop in predictability" between an "original" sample and a "check" sample that depended strongly on the method of construction of the predictor.
- Herzberg, P. A. (1969). The parameters of cross-validation. Monograph Supplement to Psychometrika, 34, made a detailed theoretical and numerical study of predictor construction methods, using cross-validatory assessment.

STATISTICAL LEARNING NETWORKS

Cross Validation

- Ivakhnenko, A.G. (1971) Polynomial Theory of Complex Systems, IEEE (Institute of Electrical and Electronics Engineers, Inc.) TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS Vol. SMC-1, No. 4, October 1971, pp. 364-378.
- Stone, M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions, Cross-Validation and Multinomial Prediction, (1977) An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion, Asymptotics For and Against Cross-Validation. Journal of the Royal Statistical Society, pp. 111-147, 44-47; Biometrika, pp. 509-515, 29-35.

- ***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.

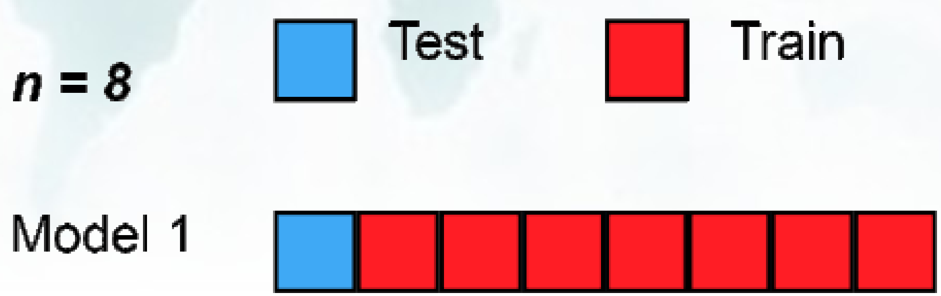
- ***Leave-one-out cross-validation:***

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, … 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:



Model Selection & Validation

- ***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

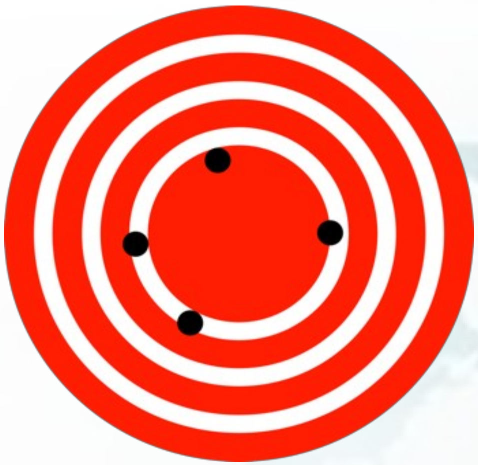
- ***Rolling forecasting origin* -** since it is not possible to get a reliable forecast based on a very small training set, the earliest observations ***n*** are not considered as testing sets.

1. We select the observation at time (n+i) for the testing set and use the observations at times t= {1, 2, … (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, … (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

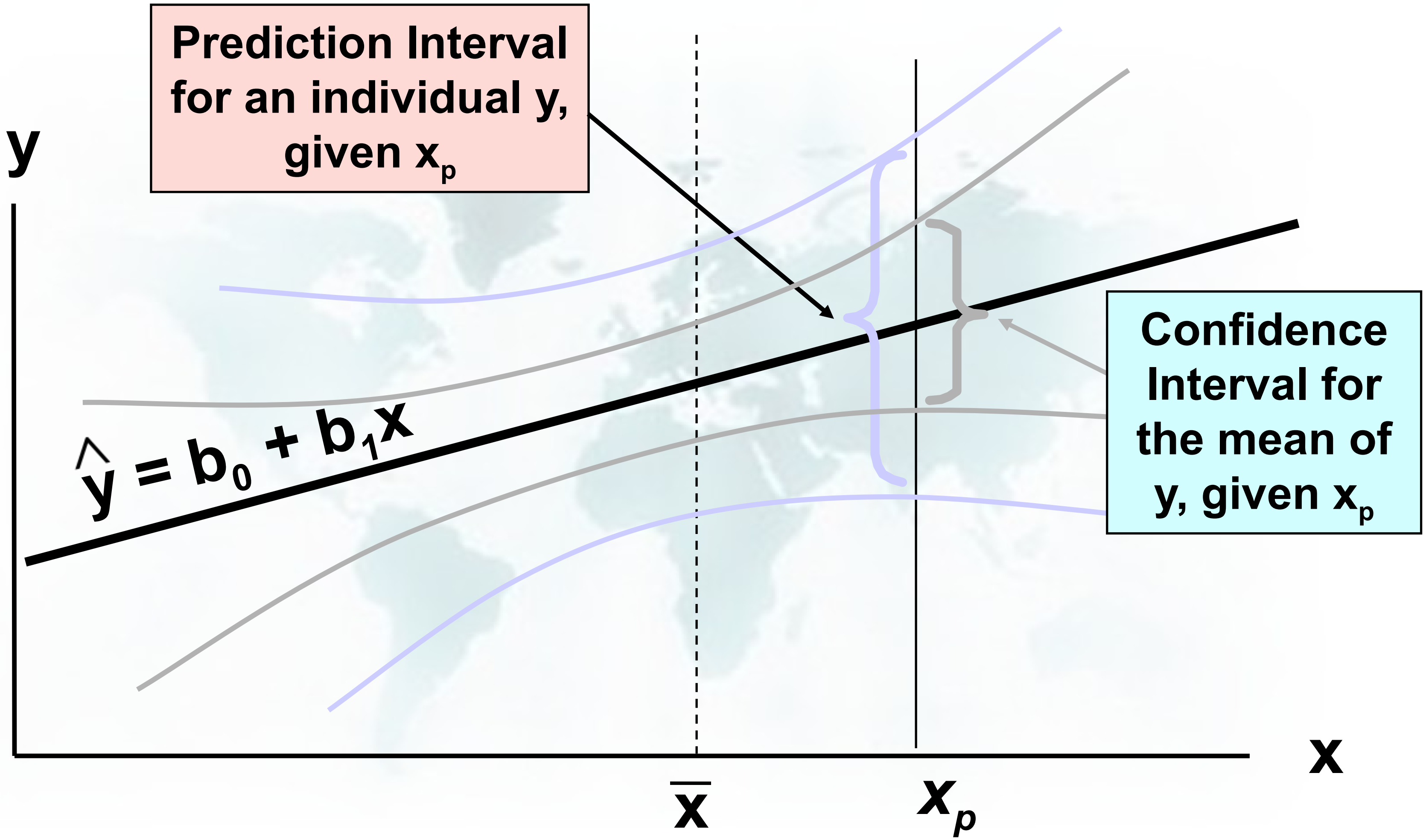


A) Low accuracy due to poor precision.

B) Low accuracy due to poor trueness.

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

Model Selection & Validation - Accuracy



Prediction (simulation) error:

$$e_t = y_t - F_t$$

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

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

$\mathbf{y_t}$ is the actual value at period ***t*** and

$\mathbf{F_t}$ is the forecast for period ***t***.

Mean Forecast Error (forecast bias):

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

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

MAD

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

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

MSE

- 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

MPE

- 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.

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

CV(RMSE)

- 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.

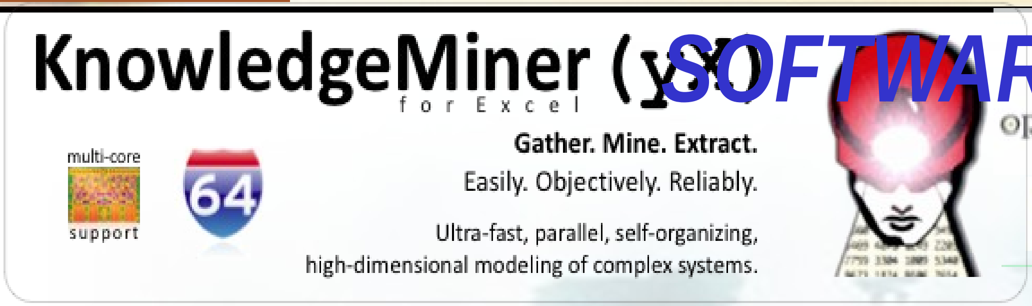
$$\mathrm{CV(RMSE)} = \mathrm{RMSE}/\bar{\mathrm{y}} \qquad RMSE = \sqrt{\mathrm{MSE}}$$

MAPE

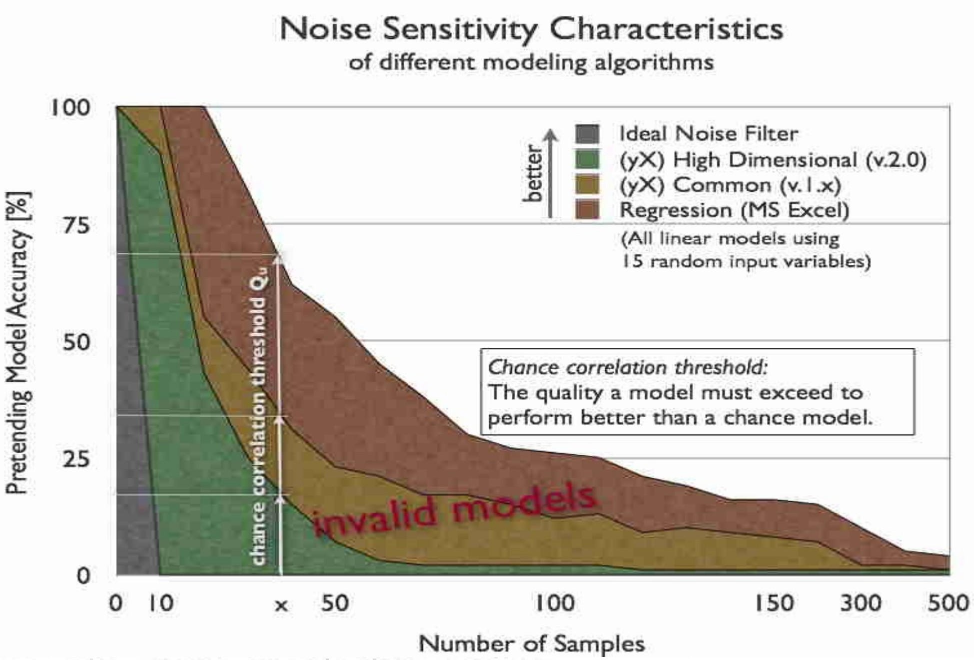
- Mean Absolute Percent Error (MAPE) - Puts errors in perspective:
  - – Average absolute percent error – useful when the size of the forecast variable is important in evaluating. It provides an indication of how large the forecast errors are in comparison to the actual values of the series. It is also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

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

STATISTICAL LEARNING NETWORKS KNOWLEDGEMINER SOFTWARE



	A	B	C	
1	XL Source:	COD concentration.xlsx		
2	MONTH	Transparency	Filtered COD	MODE
3	Sept.	2,5	2	
4	Oct.	2,5	1,6	3,
5	Nov.	3	1,7	2,7
6	Dec.	3	2,1	2,9
7	Jan.	2,5	1,8	3,
8	Feb.	2,4	2,2	3,51
9	March	2,5	2,1	3,
10	April	2,5	2,1	3,
11	May	2,1	2,9	4,30
12	June	1,6	2	= 1,6
13	July	3	2	2,
14				
15	Minimum of le	1,1	0,8	
16	Maximum of le	3,8	3,8	



Analytical model implemented in a new Excel worksheet by (yX) for Excel.

Model Selection

Supposition One **– it is important to consider more than one selection/evaluation criteria – this will help to obtain a reasonable knowledge about the amount, magnitude, and direction of the overall model error. Experienced researchers normally use the criteria MPE, MAPE, RMSE, and CV(RMSE) together:**

- **Measures of Trueness (Systematic error, Statistical Bias) - RMSE and MPE;**
- **Measures for model's precision (i.e. its random error) - use MAPE and CV(RMSPE) in tandem.**

Measures of Trueness (Systematic error, Statistical Bias):

- **Mean Percentage Error (MPE)**

$$\mathrm{MPE}\,(\%) = \frac{1}{\mathrm{N}}\sum_{t=1}^{N}(e_t / y_t)\mathrm{x}100$$

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

$$RMSE = \sqrt{\mathrm{MSE}} \qquad \mathrm{MSE} = \sum\left(e_t\right)^2/(n-1)$$

When selecting a good model based on a testing dataset, it is desirable that both criteria should be as close to zero as possible.

Measures of Precision (Random Error):

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

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

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

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

CV(RMSE) **penalizes extreme errors and** ***MAPE*** **does not, i.e. first goal should be to select a model where the calculated values of both criteria are very close meaning there are no extreme error values. The second goal is that both criteria values are as close to zero as possible.**

Prediction Accuracy

Experimental Test Results - 2018

Best Model	Second Best	Third Best
MLNAN: MASE: 0.0414 MPE = 1.42% MAPE = 1.42% CV(RMSE) = 1.56%	Triple Exponential MASE = 0.0627 MPE = -0.57% MAPE = 1.76% CV(RMSE) = 2.45%	Multiple Autoregression MASE = 0.0908 MPE = 2.03% MAPE = 2.58% CV(RMSE) = 3.17%

Prediction Accuracy

Experimental Test Results - 2019

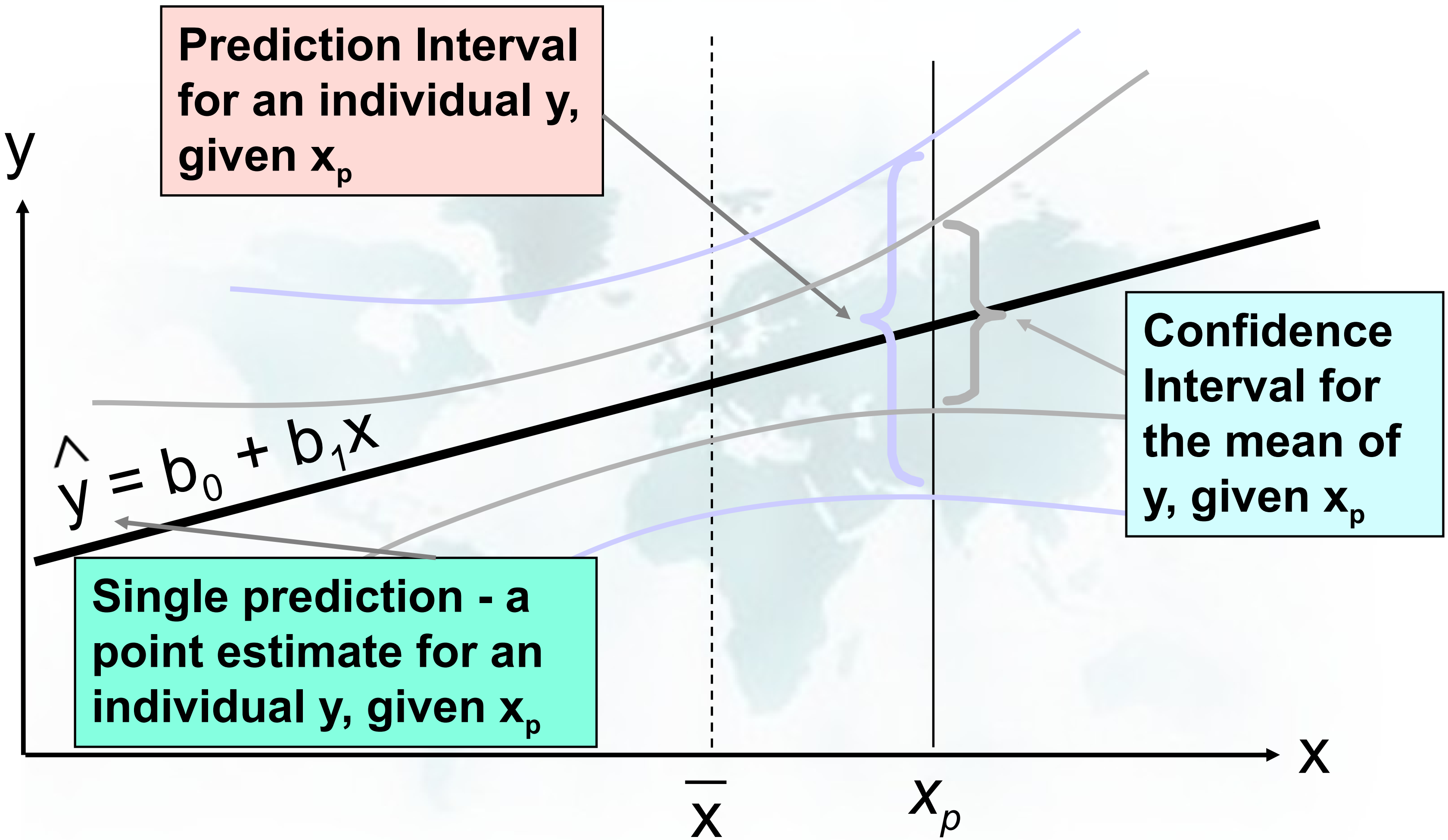
Best Model	Second Best	Third Best
MLNAN: MASE: 0.0446 MPE = 1.55% MAPE = 1.55% CV(RMSE) = 1.56%	Multiple Regression with Time and Dummy Seasonal MASE = 0.0508 MPE = -1.09% MAPE = 1.59% CV(RMSE) = 1.56%	Triple Exponential MASE = 0.0627 MPE = -0.57% MAPE = 1.76% CV(RMSE) = 2.45%

Model Selection

Supposition Two **– predictions are more likely to be closer to intervals rather than to a single point, i.e. predictions are not perfect, and their results usually differ from the real-life values. Consequently, it is better to consider the calculated values as intervals rather than point estimates.**

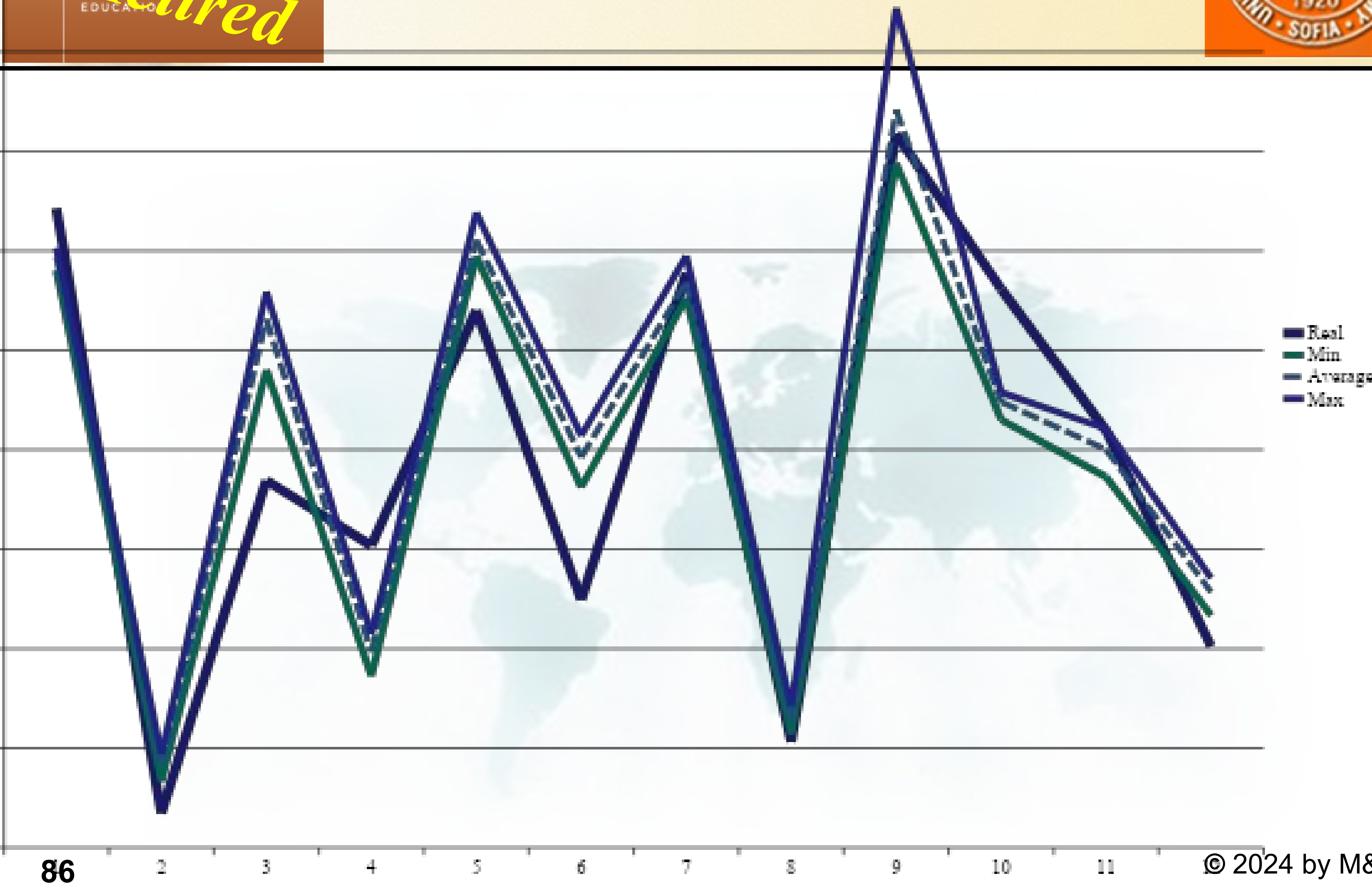
To construct a *prediction interval,* we can calculate the upper and lower limits from the given data using the *RMSE*. This estimation provides a range of values where the parameter is expected to lie. It gives more information than point estimates and is preferred when making inferences. Often, the upper limit of the interval is called optimistic (or *Maximum*) prediction and the lower limit pessimistic (or *Minimum*) prediction.

Model Selection & Validation - Accuracy



Walla Walla University
Retired

Multiple Criteria Approach (Prediction Intervals)



Prediction Accuracy

Experimental Test Results – 2020 Case 1

Best Model	Second Best	Third Best
MLNAN Optimistic MPE = -0.08% MAPE = 0.82% CV(RMSE) = 0.93%	VARMAX MPE = -0.37% MAPE = 0.73% CV(RMSE) = 1.02%	MLNAN Nonlin MPE = 0.06% MAPE = 1.17% CV(RMSE) = 1.44%

Prediction Intervals

Experimental Test Results – 2020 Case 2

Best Model	Second Best	Third Best
Pessimistic - Min MASE: 6.0075 MPE = -0.07% MAPE = 7.40% CV(RMSE) = 8.67%	Composite - Average MASE = 6.3350 MPE = -2.94% MAPE = 8.13% CV(RMSE) = 9.42%	Optimistic-Max MASE = 7.6934 MPE = -5.65% MAPE = 9.70% CV(RMSE) = 11.13%

STATISTICAL LEARNING NETWORKS

Model Selection

***Supposition Two* – inferences:**

The *MPE* which shows the direction of the *Systematic error* can be used to make more precise decisions. If its value is very close to zero, we should select the *average* prediction model. When *MPE* is negative, we should select the *minimum* prediction model and finally, when it is positive, we should select the *maximum* prediction model.

The good models are first evaluated using multiple criteria and among their *Max*, *Min*, and *Average* versions the best model is selected using the prediction bias (the systematic error measured with *MPE*) and the model's precision (i.e. its random error) presented by *MAPE* and *CV(RMSPE)*.

STATISTICAL LEARNING NETWORKS

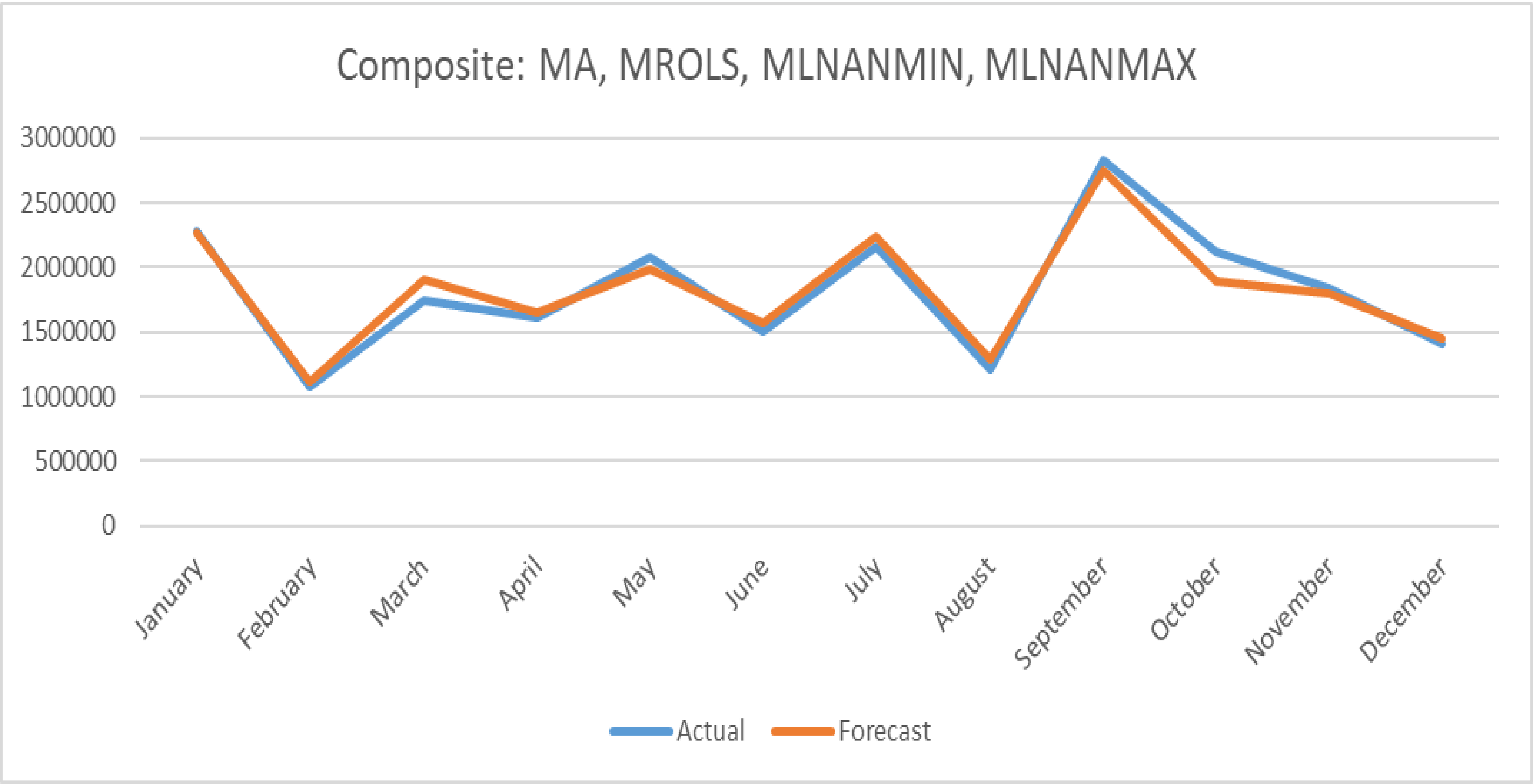
Model Selection

Supposition Three **– the best model is a *composite model*. There is no single universal model that works in all possible cases.**

In Suppositions one and two, we achieved many positive results using model errors, measured by multiple criteria, and prediction intervals. The goal in Supposition three is slightly different. Here, we try to improve predictions accuracy with a procedure which creates composite models of a different type.

Best Model Selection

Experimental Test Results – 2021, WWU



Best Model Selection

Prediction Intervals

Experimental Test Results – 2021 Case 1

Error Summary Chart				
	Linear Average	Linear Min	Multi-Linear Average	Multi-Linear Min
Bias:	-415040.33	326573.05	-628719.09	-211879.87
MPE(%)	-3.35%	0.50%	-4.85%	-2.76%
MAPE (%)	10.59%	8.64%	11.77%	10.82%
MAD	179318.36	156675.29	191147.22	179976.37
MSE	46153196596.12	38895823660.74	48290483323.89	41576923609.48
MASE	8.33	7.28	8.88	8.36
RMSE	214832.95	197220.24	219750.96	203904.20
CV(RMSE)=	11.80%	10.83%	12.07%	11.20%

MLNAN AR Forecasts Error Summary Chart

Best Model Selection

Prediction Intervals

Experimental Test Results – 2021 Case 1B

Error Summary Chart		
	MLNAN Average	MLNAN Min
Bias:	-43315.28	5715.74
MPE(%)	-3.74%	-0.70%
MAPE (%)	10.38%	8.28%
MAD	172390.82	144589.76
MSE	38911149258.38	30815035888.85
MASE	1.47	1.24
RMSE	197259.09	175542.12
CV(RMSE)=	10.83%	9.64%

MLNAN ARMAX Forecasts Error Summary Chart

Walla Walla University
Seventh-day Adventist Higher Education
Retired

Best Model Selection

Prediction Intervals

Experimental Test Results – 2021 Case 2B

Bias:	-4575.996251
MPE(%)	-1.05%
MAPE (%)	4.63%
MAD	83008.18999
MSE	10342417863
MASE	1.073967405
RMSE	101697.6788
Coeff. Var (RMSE)	5.59%

Conclusions:

SLNs help increasing models' accuracy, which:

- helps researchers to analyze problems more precisely, which
- leads to deeper and better understanding of the case;
- helps to generate better predictions, which
- supports managers in making better decisions that relate more closely to real-life business problems.

Conclusions:

SLNs help increasing models' accuracy, which:

- provides more reliable bases for simulations and what-if analysis;
- makes it possible to analyze more precisely the problem in consideration;
- provides more realistic predictions;
- helps managers make better and more cost-effective decisions.

Thank You!

Questions?

Thank You!

and I'll

See You again…