

The 6th Pharma Indochina Conference

Hue, Vietnam – Dec 16th, 2009

**EXTRACTION PROCESS DEVELOPMENT
ASSISTED BY INTELLIGENT SOFTWARE SYSTEMS**

Dang Van Giap

Faculty of Pharmacy, Univ. Med. & Pharm. - HCMC, Vietnam

Problems of interest



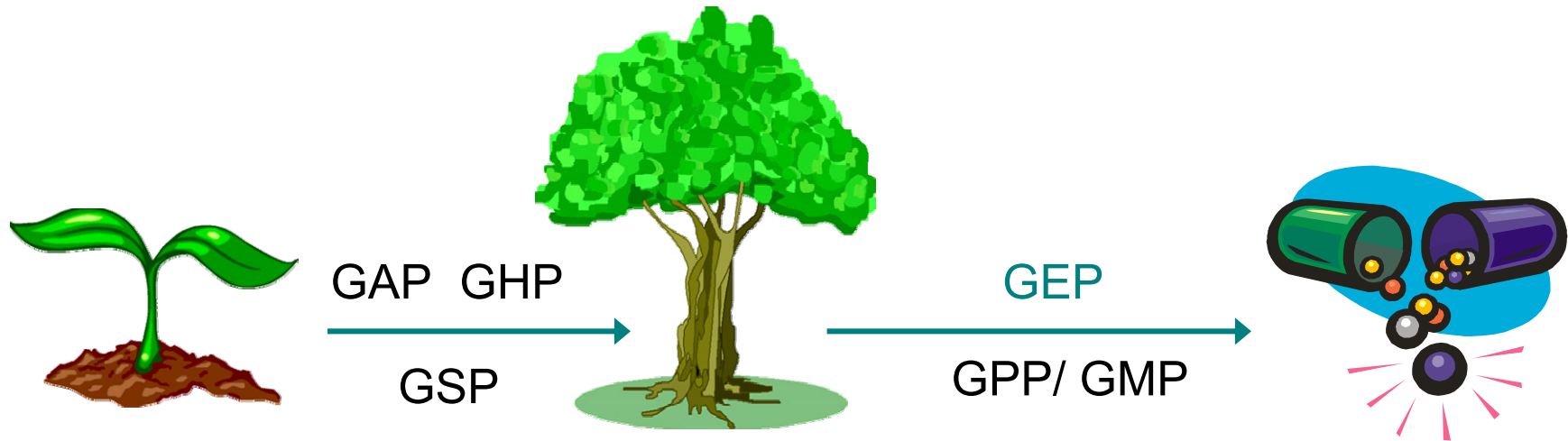
Design of experiments

Cause-effect relationships

Multivariate optimization

Practical applications

GP guidelines for medicinal plants



GAP: Good Agricultural Practice

GHP: Good Harvesting Practices

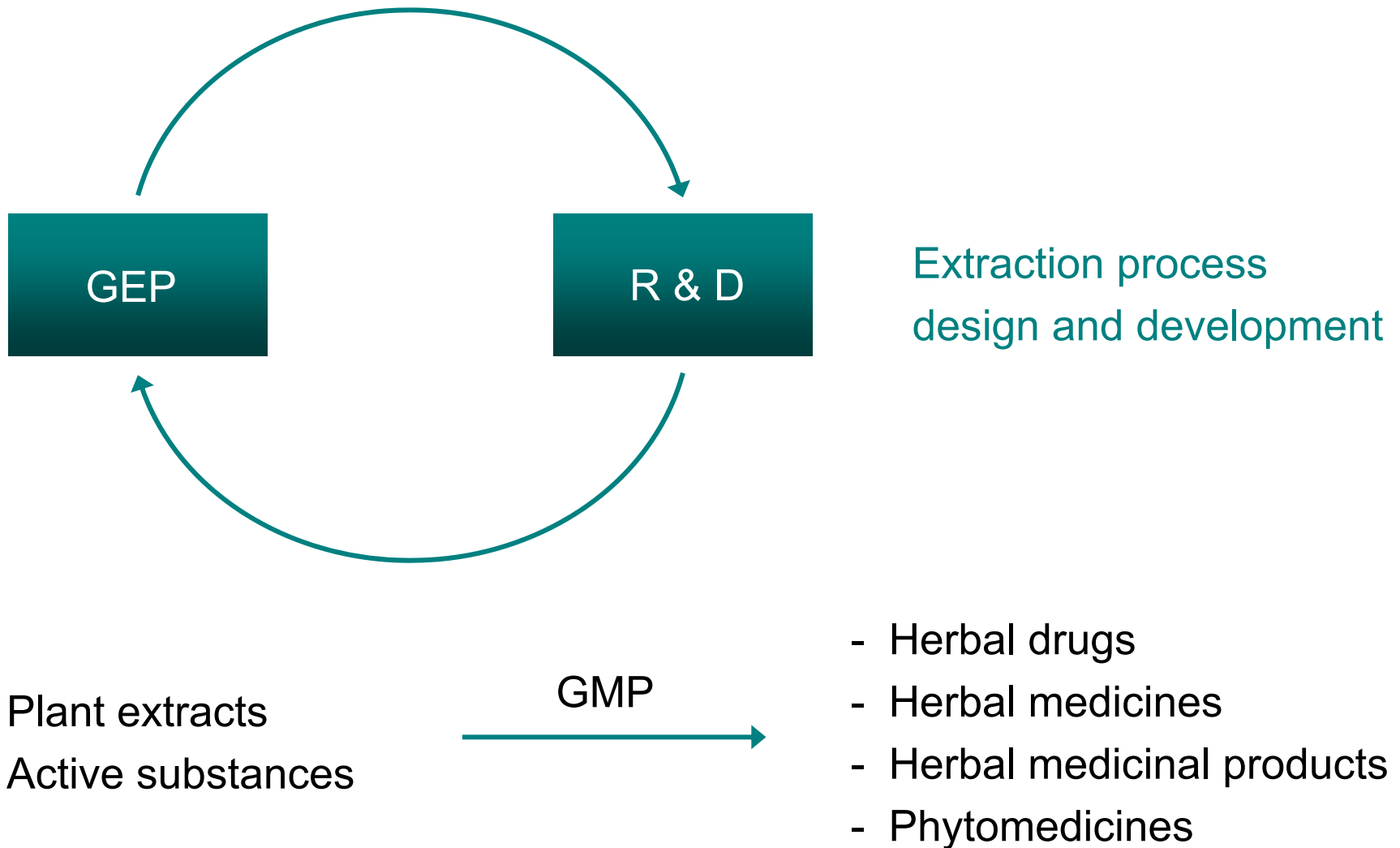
GSP: Good Sourcing Practices

GEP: Good Extraction Practices

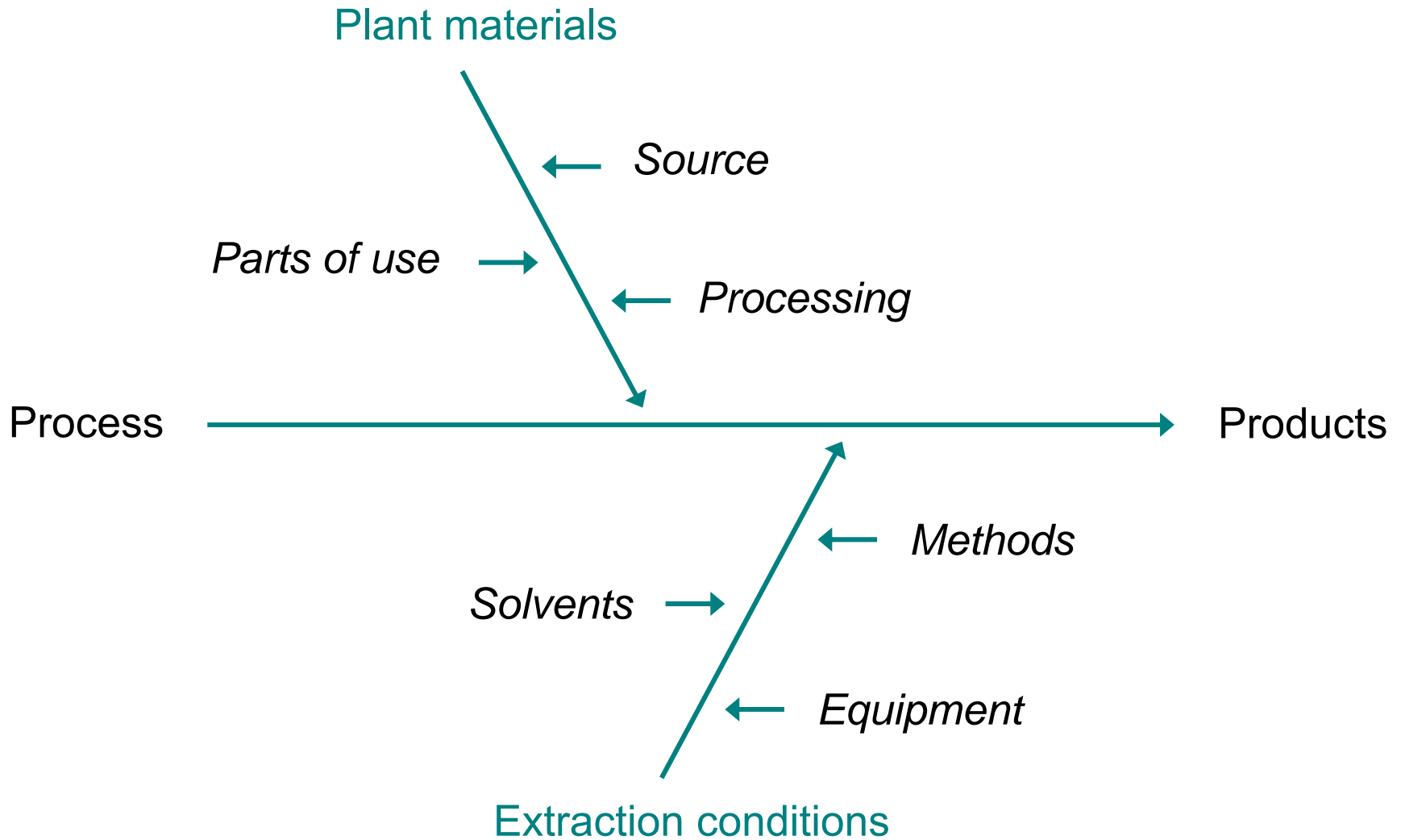
GPP: Good Processing Practices

GMP: Good Manufacturing Practices

The role of process design and development



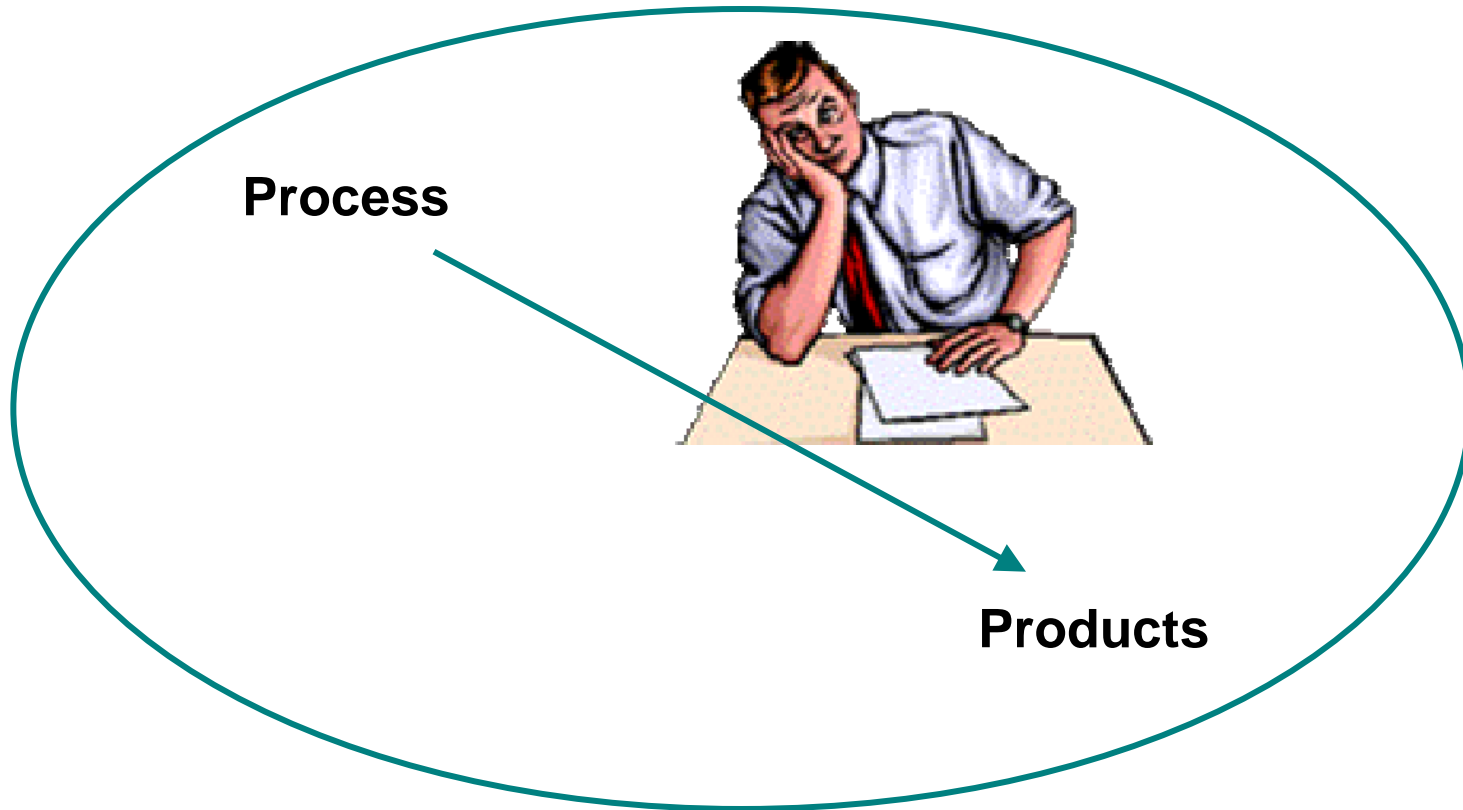
Factors affecting solvent extraction process



Information for extraction process development

Plant materials

Extraction facilities



Extract specifications

Commercial factors

Difficulties faced in extraction process development

Facing challenges

- Which variables affect each product property?
- What rules govern the extraction process?
- How to optimize process parameters in a right way?
- How changing conditions affect extract properties?



Unwanted mistakes

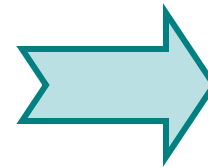
- Measuring unnecessary variables, or missing critical ones.
- Using solvents having little impact on product properties.
- Spending time on exploration of unfruitful ideas.

Artificial intelligence – an interdisciplinary field

Human brain



Artificial intelligence

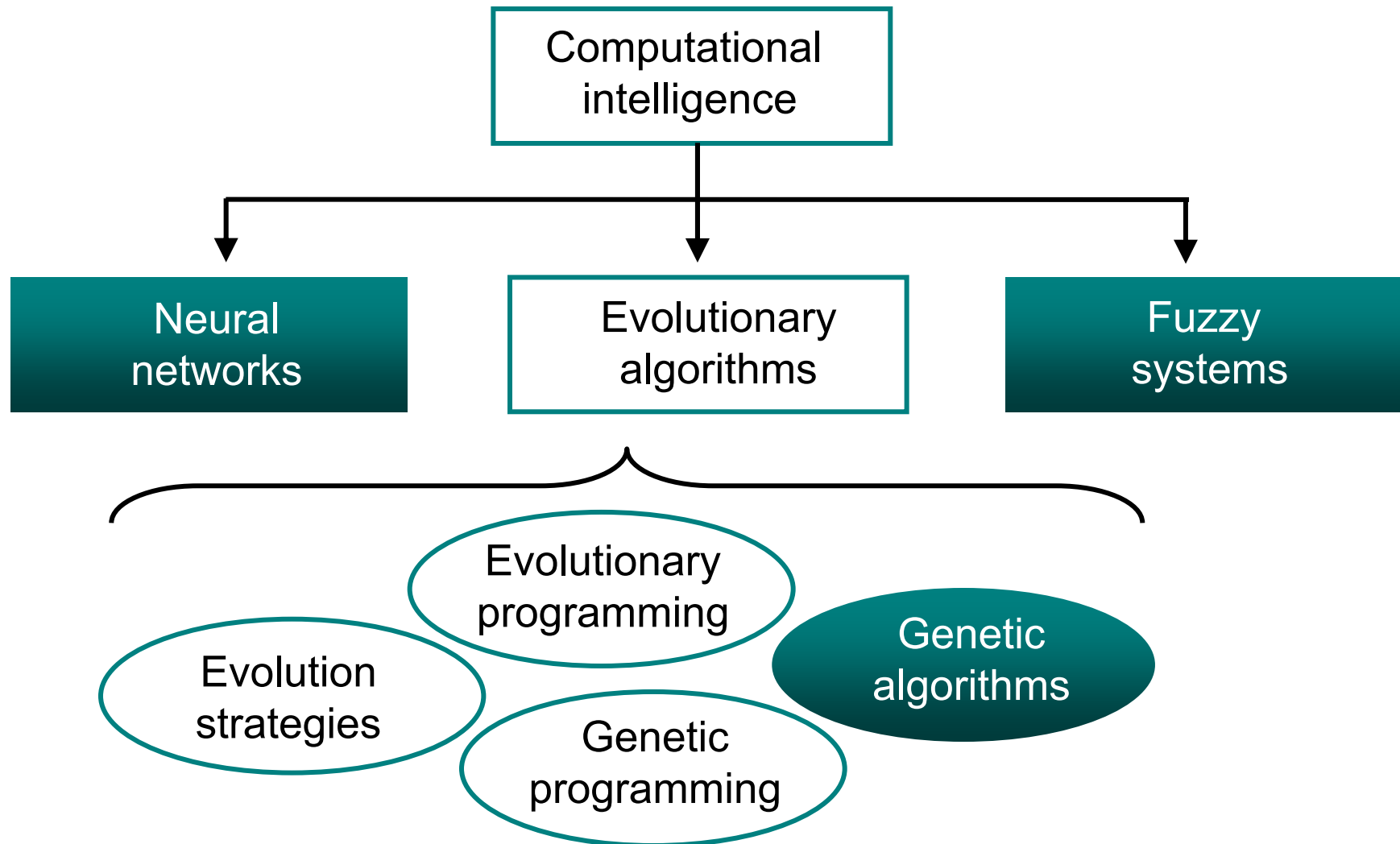


Data mining
Optimization
Machine learning
Neural computing
Problem solving
etc...

Linguistics, Physiology, Engineering,
Philosophy, Biology, Computer Science, etc...

“The capacity of a computer to perform tasks commonly associated with the higher intellectual processes characteristic of humans, such as the ability to reason, discover meanings, generalise or learn from past experience” (Encyclopaedia Britannica, 1995).

Domains of computational intelligence



Intelligent software systems

Artificial intelligence software

= Intelligent software systems

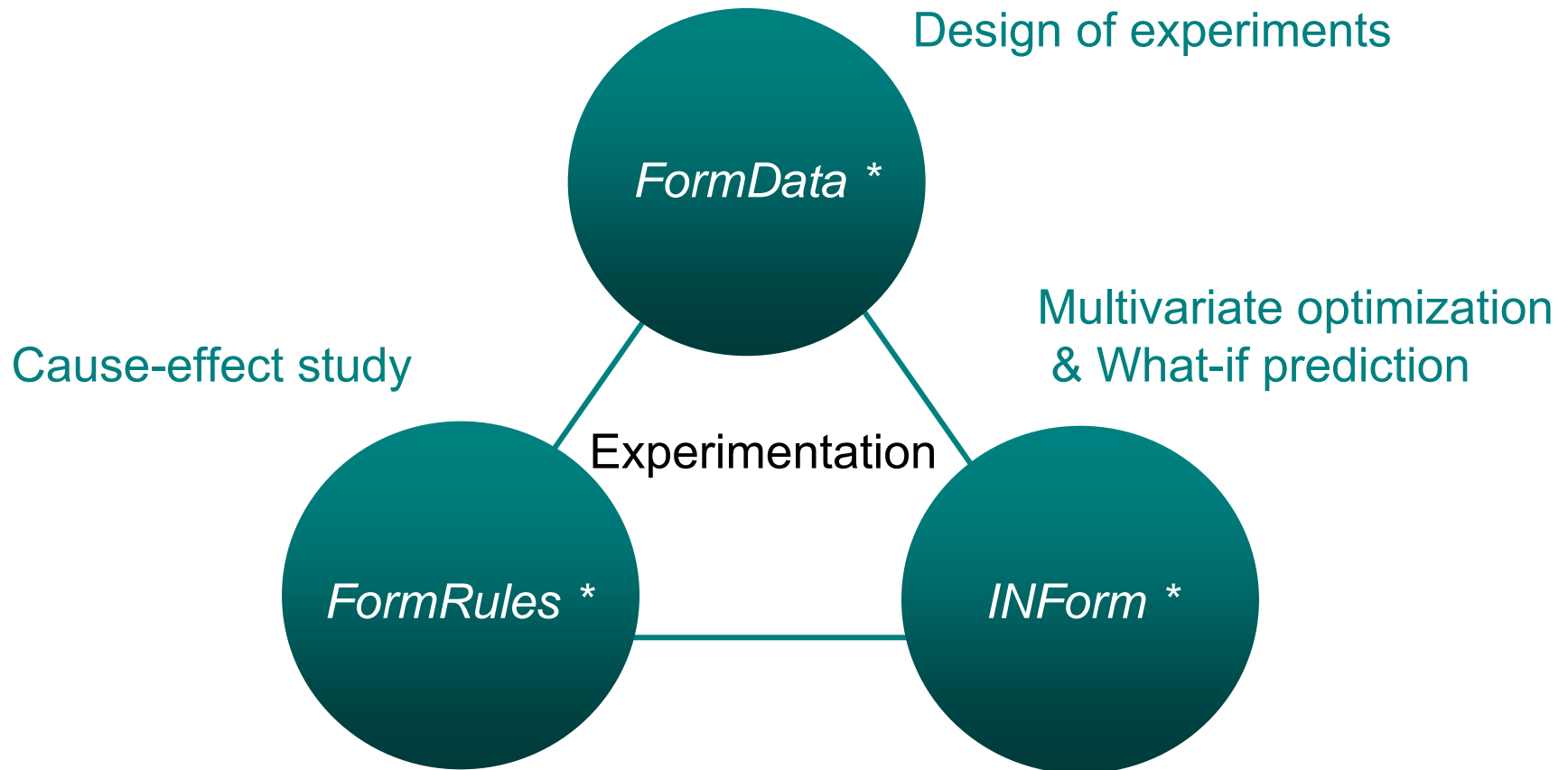
= Intelligent software

INForm intelligent software: Neural networks + Genetic algorithms

FormRules intelligent software: Neurofuzzy logic

- a. *Neural networks*: Computer-based systems discovering the cause-effect relationships buried in experimental data and also allowing "what if" possibilities to be investigated.
- b. *Genetic algorithms*: A sophisticated optimization technique, relying on the fact that a criterion of fitness can be defined.
- c. *Neurofuzzy logic*: A special combination of fuzzy logic with neural networks, capable of generating rules from numerical input data.

A powerful toolkit in process R & D



* *Intelligensys Ltd., Springboard Business Centre, Stokesley TS9 5JZ, United Kingdom.*

Problems of interest

Design of experiments

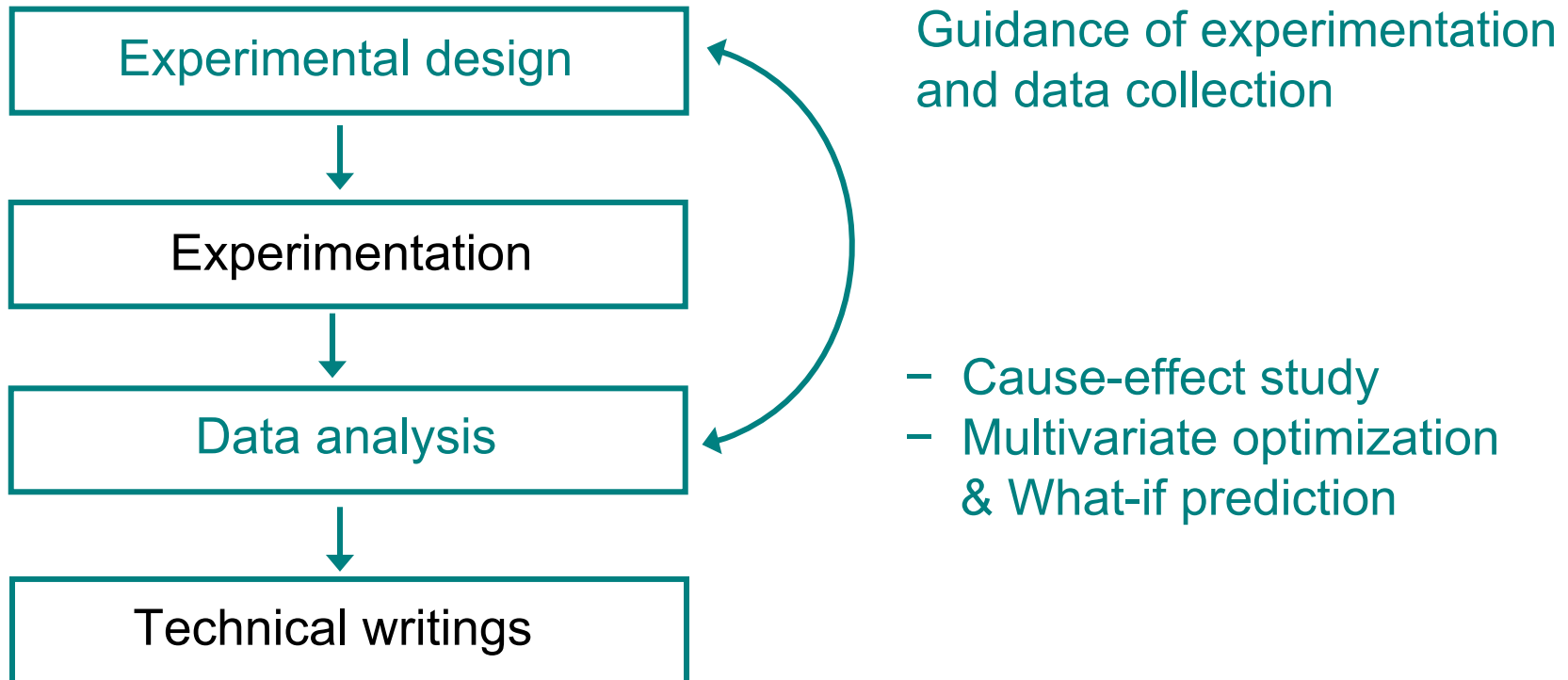


Cause-effect relationships

Multivariate optimization

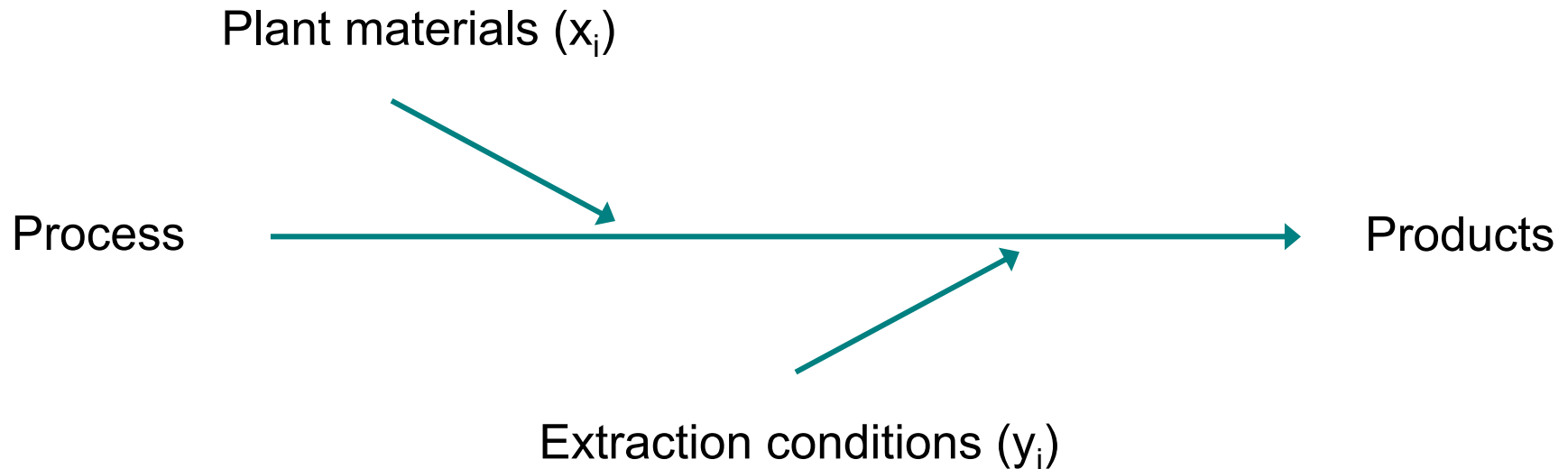
Practical applications

The purpose of experimental design



The **design of experiments** (DOE) is an efficient procedure for planning experiments so that the data obtained can be analyzed to yield valid and objective conclusions. DOE begins with determining the objectives of an experiment and selecting the process factors for the study.

Causality – the cause-effect relationship



Causality is the relationship between an event (the **cause**) that produces something new (the **effect**) or one that produces a change in an existing substance, where the effect is a direct consequence of the cause.

The cause must be prior to, or at least simultaneous with, the effect. And, that cause and effect must be in spatial contact or connected by a chain of intermediate things in contact.

Independent and dependent variables

Independent variables (x)  $X_i = X_1, X_2, X_3 \dots X_k$

The causes

- Plant materials: sources, parts of use, processing, ...
- Extraction conditions: solvents, methods, equipment, ...

Dependent variables (y)  $Y_j = Y_1, Y_2, Y_3, \dots Y_l$

The effects

- Extraction yield (%)
- Active substance(s)
- Unwanted substance(s)
- Extract cost

Definition of experimental designs

Mixture designs (or formulation designs)

A mixture design is an experimental design in which the independent factors are proportions of different components of a blend.

Factorial designs (or process designs)



A **full factorial design** is an experimental design that consists of ≥ 2 factors, each factor with discrete possible values or "levels". If the number of experiments in a full factorial design is too high to be logistically feasible, a **fractional factorial design** may be done, in which some of the possible experiments are omitted.

Combined designs

Mixture design + Factorial design

Software for designing experiments

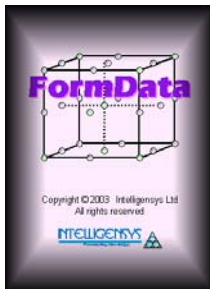
A model established with poor information is like a house built on sand.



DOE software:



Design-Expert 6.0.6
Stat-Ease Inc., Minneapolis (2002)



FormData v2
Intelligensys Ltd., UK (2003)

D-Optimal: a special factorial design

	x_1	x_2	x_3	y_1	y_2	y_3	...
1	c	c	c				
2	a	b	c				
3	a	a	a				
4	c	b	c				
5	b	c	a				
6	a	c	a				
7	c	a	a				
8	b	a	a				
9	b	a	c				
10	c	a	c				
11	b	b	c				
12	b	b	a				
13	a	b	a				
14	a	c	c				

n = 14 (~~18~~)

a = Low

b = Medium

c = High

Taguchi OA: a special factorial design

	x_1	x_2	x_3	x_4	x_5	y_1	y_2	y_3	...
1	3	3	1	2	4				
2	4	2	3	1	4				
3	1	2	2	2	2				
4	1	1	1	1	1				
5	3	1	3	4	2				
6	2	1	2	3	4				
7	4	1	4	2	3				
8	1	3	3	3	3				
9	3	4	2	1	3				
10	2	2	1	4	3				
11	1	4	4	4	4				
12	2	4	3	2	1				
13	2	3	4	1	2				
14	4	4	1	3	2				
15	4	3	2	4	1				
16	3	2	4	3	1				

$n = 16$ (~~1024~~)

1 = Low

2 = Medium-Low

3 = Medium-High

4 = High

An example of extraction process development *

Medicinal plant

Phyllanthus amarus Schum. and Thonn. (Euphorbiaceae) is an herbal plant which widely spread throughout the tropical and subtropical areas.

Extracts of the plant had been reported to have antibacterial activity, antiviral activity against hepatitis B antispasmodic properties... Several active compounds have been identified in this medicinal herb as *phyllanthin*, hypophyllanthin...



* *Duc-Hanh Nguyen, Minh-Duc Nguyen; Van-Giap Dang. Process development for Phyllanthus amarus Schum & Thonn. Journal of Medicine, HCM City, 13 (1), 263-267 (2009).*

Goals for extraction process development *

Research goals

- a. To study the cause-effect relationships (trends, strengths and rules) in the extraction process for *Phyllanthus amarus* Schum. & Thonn?
- b. To optimize the parameters including ethanol concentration, material-solvent ratio and extraction times for obtaining highest extraction yield and maximal phyllanthin content?

Variable selection

x_1 = Ethanol concentration (low, medium, high)

x_2 = Material-solvent ratio (1:9, 1:12, 1:15)

x_3 = Extraction times (2, 3)

y_1 = Extraction yield (%)

y_2 = Phyllanthin content



Experimental design (e.g. D-Optimal)

	x_1	x_2	x_3	y_1	y_2
1	Mid	1:15	3		
2	Mid	1:9	2		
3	Mid	1:15	2		
4	Low	1:9	2		
5	Low	1:12	3		
6	Low	1:9	3		
7	Mid	1:9	3		
8	High	1:15	2		
9	High	1:9	3		
10	Low	1:15	2		
11	Mid	1:12	2		
12	High	1:12	3		
13	Low	1:12	2		
14	High	1:15	3		

Experimental data (D-Optimal design)

	x_1	x_2	x_3	y_1	y_2
1	Mid	1:15	3	7.21	1.55
2	Mid	1:9	2	6.34	1.36
3	Mid	1:15	2	6.92	1.28
4	Low	1:9	2	5.50	0.67
5	Low	1:12	3	6.30	0.55
6	Low	1:9	3	6.06	0.63
7	Mid	1:9	3	6.89	1.07
8	High	1:15	2	6.02	4.01
9	High	1:9	3	5.89	3.04
10	Low	1:15	2	6.10	0.47
11	Mid	1:12	2	6.71	1.91
12	High	1:12	3	6.12	3.95
13	Low	1:12	2	5.88	0.52
14	High	1:15	3	6.32	2.66

Problems of interest

Design of experiments

Cause-effect relationships

Multivariate optimization

Practical applications



Knowledge discovery or data mining



Lack of
information

In the past



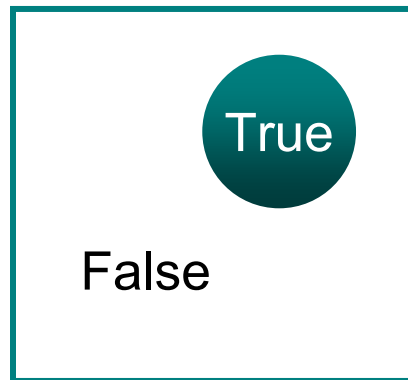
Dyspepsia due to
large databases

At the present

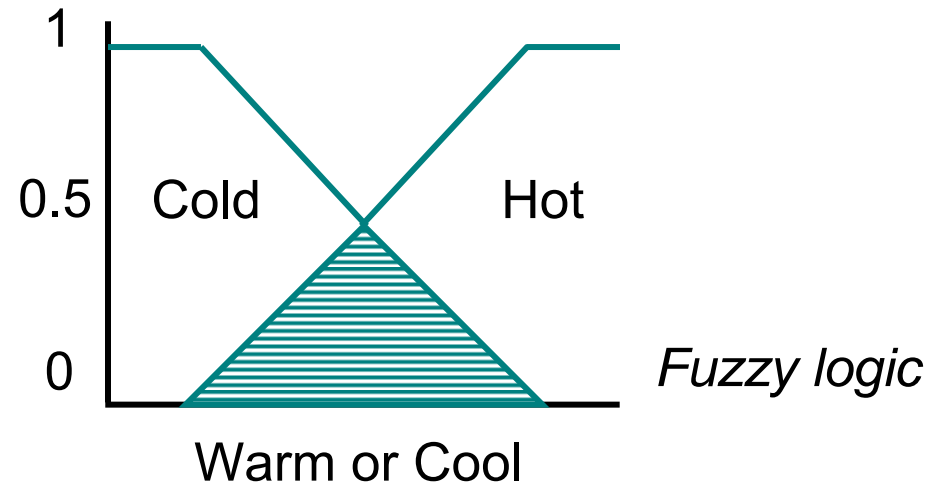
Knowledge discovery (or **data mining**) is a the process of automatically extracting knowledge from experimental data in market research, financial analysis, scientific R & D, ...

Neurofuzzy logic, which combines the strengths of neural networks with those of fuzzy logic, has been applied to knowledge discovery or data mining (**automatic rule induction**).

Fuzzy logic vs. crisp logic



Crisp logic

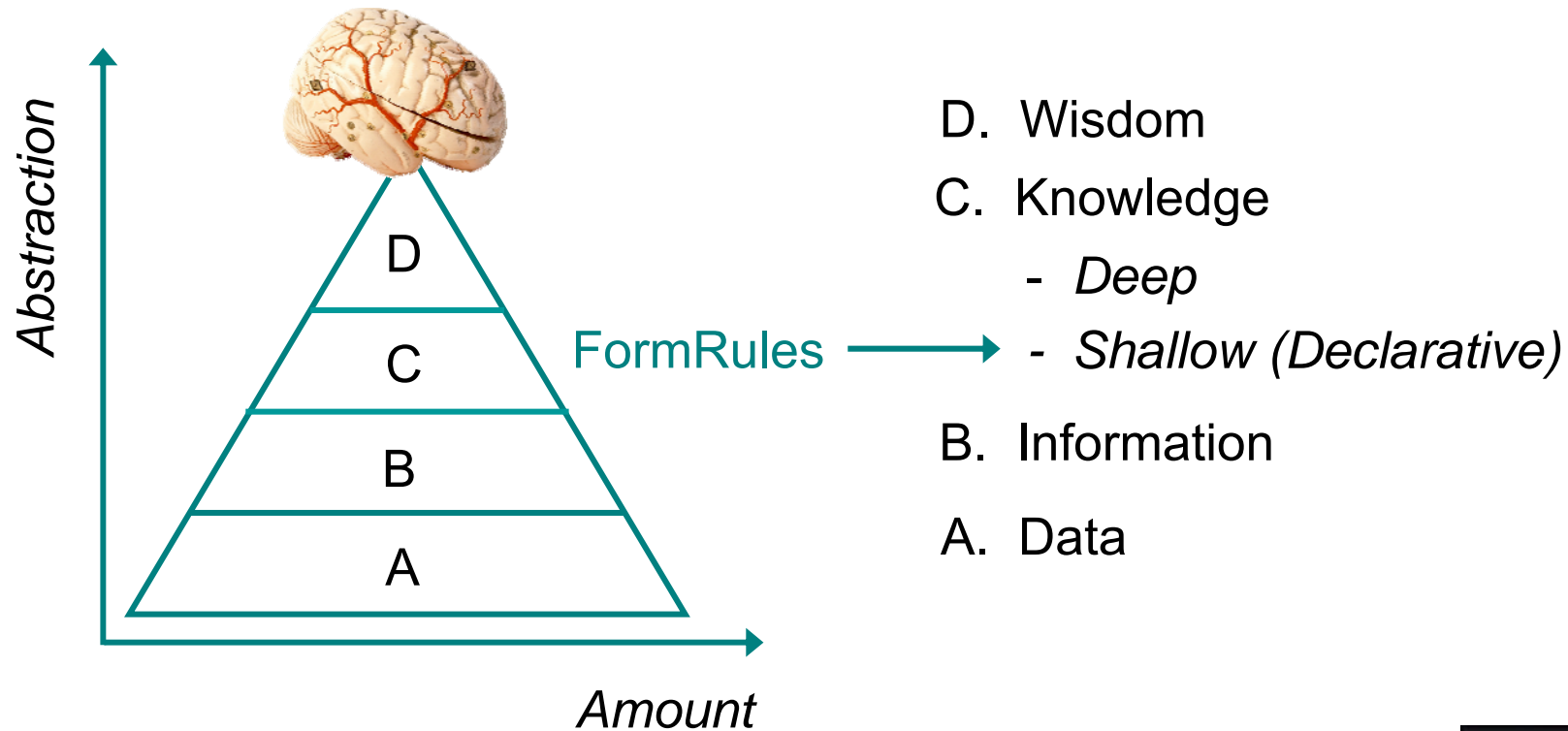


Fuzzy logic

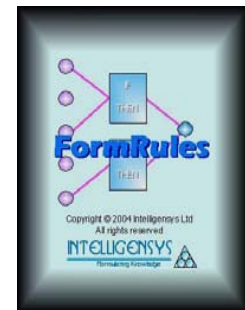
The fuzzy controller works in 3 steps:

- Fuzzification**, where a crisp input is translated into a fuzzy value.
- Rule evaluation**, where a calculation is performed to determine the relevant rules, and fuzzy output truth values are computed.
- Defuzzification**, where the fuzzy output is translated in to a crisp output value.

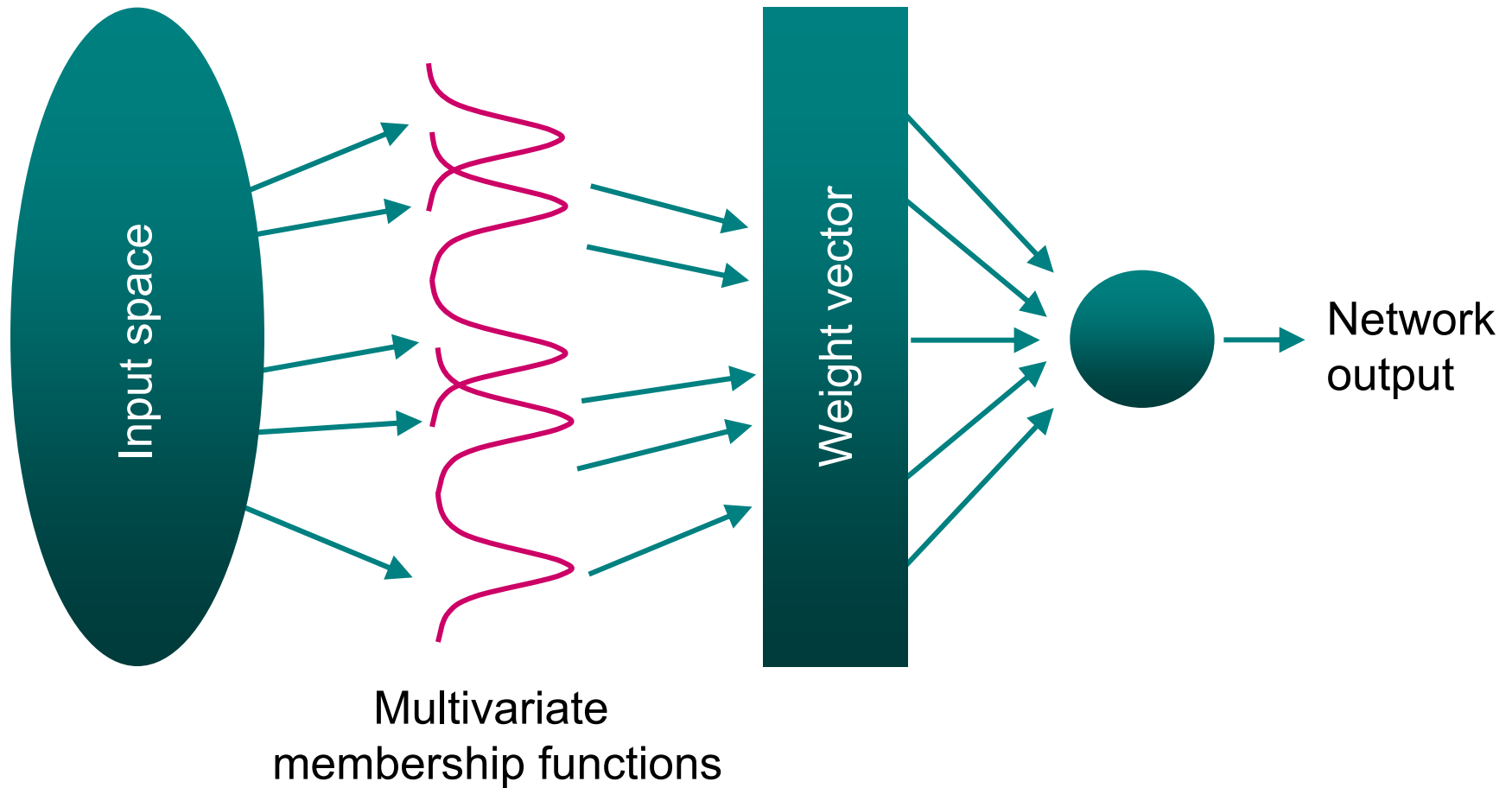
Intelligent software for knowledge discovery



FormRules intelligent software, using **neurofuzzy logic** as its underpinning technology, can quickly automate much of **knowledge discovery** in process development. It presents the information as easy-to-understand **rules** and 3D-graphs.



The basic structure of a neurofuzzy system



A grey box

Associative Memory Networks

Structure of generated rules from data

Deep knowledge = full understanding (exactly BECAUSE)

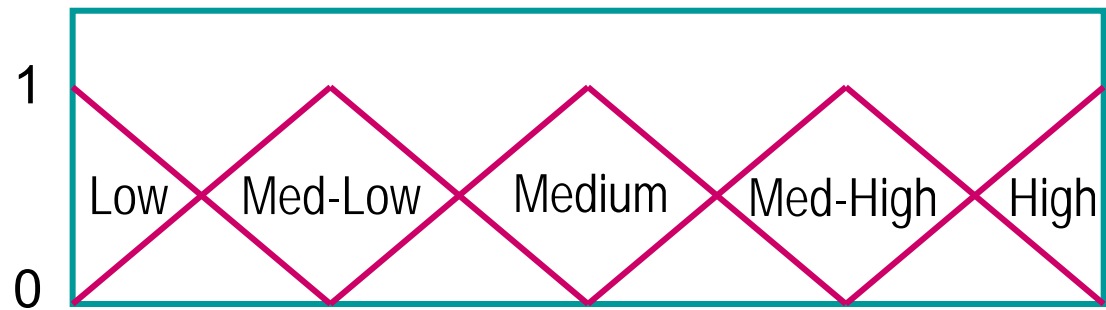
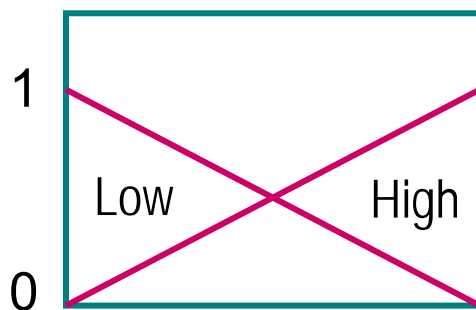
Shallow knowledge = rule expression (perhaps WHY)

The antecedent

IF (Condition 1) AND (Condition 2) THEN (Observed behaviour)

The consequent

The complexity of rule statements depends on the fuzzy set inputs:



Goal to study the cause-effect relationships

Research goal

To discover the **trends, strengths and rules** of the cause-effect relationships buried in the experimental data collected from the extraction process of *Phyllanthus amarus* Schum. & Thonn.?

Extraction process

For the dried leaves of *Phyllanthus amarus* Schum. & Thonn.

x_1 = Ethanol concentration (low, medium, high)

x_2 = Material-solvent ratio (1:9, 1:12 , 1:15)

x_3 = Extraction times (2, 3)

y_1 = Extraction yield (%)

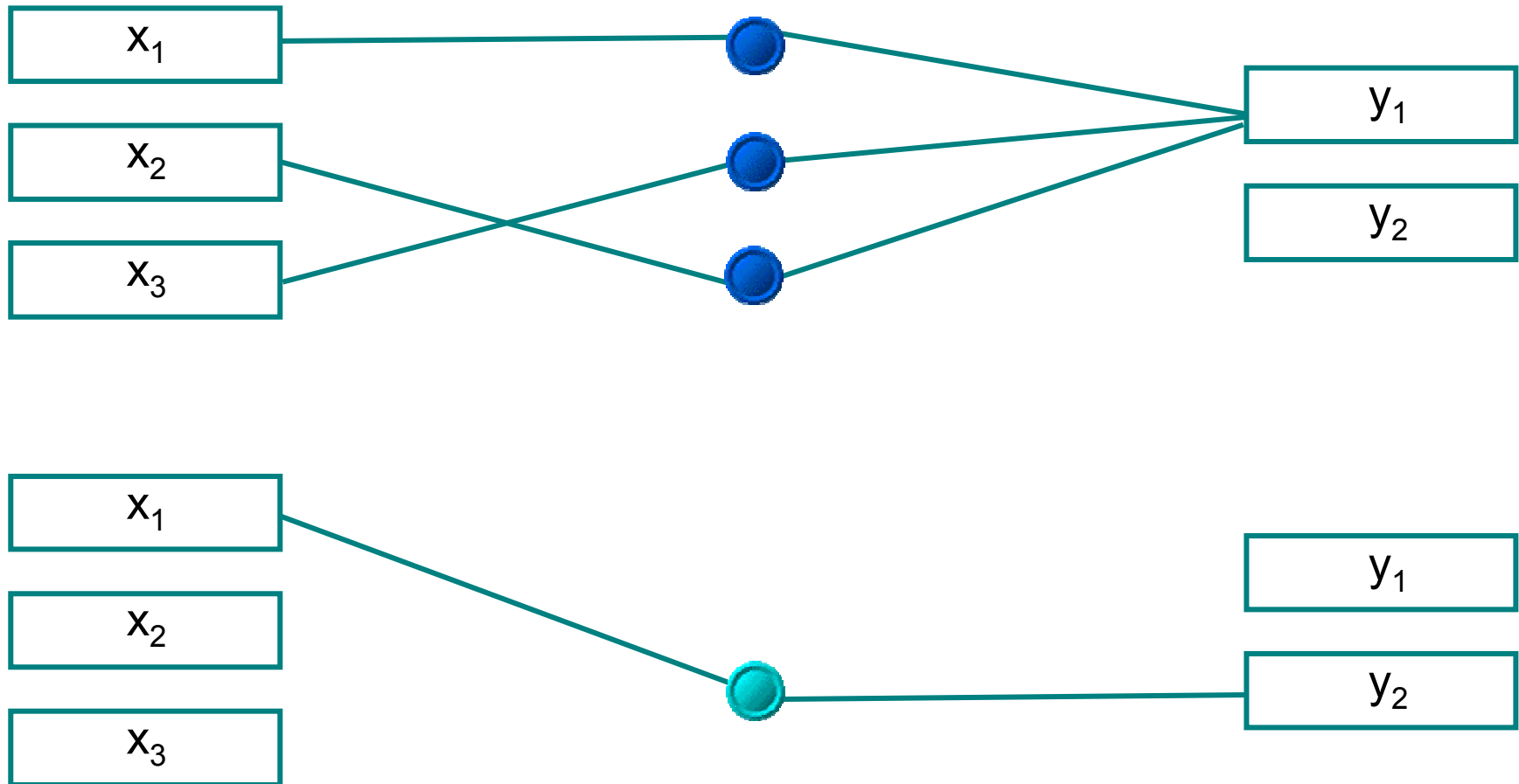
y_2 = Phyllanthin content



The data input for FormRules intelligent software

	x_1	x_2	x_3	y_1	y_2
1	Mid	1:15	3	7.21	1.55
2	Mid	1:9	2	6.34	1.36
3	Mid	1:15	2	6.92	1.28
4	Low	1:9	2	5.50	0.67
5	Low	1:12	3	6.30	0.55
6	Low	1:9	3	6.06	0.63
7	Mid	1:9	3	6.89	1.07
8	High	1:15	2	6.02	4.01
9	High	1:9	3	5.89	3.04
10	Low	1:15	2	6.10	0.47
11	Mid	1:12	2	6.71	1.91
12	High	1:12	3	6.12	3.95
13	Low	1:12	2	5.88	0.52
14	High	1:15	3	6.32	2.66

The trends of the cause-effect relationships *



* FormRules v3.3: Minimum Description Length (MDL)

The strengths of the cause-effect relationships

	x_1	x_2	x_3	Train R^2
y_1	+	+	+	98.6346
y_2	+	-	-	91.2376

In the extraction process of *Phyllanthus amarus* Schum. & Thonn.:

- The extraction yield (y_1) was **very highly related** ($R^2 = 98.6346\%$) to all investigated parameters such as ethanol concentration (x_1), material-solvent ratio (x_2) and extraction times (x_3).
- The amount of phyllanthin was **highly related** ($R^2 = 91.2376\%$) to ethanol concentration (x_1) only.

The rules of the cause-effect relationships

Related to the extraction yield

IF x_1 is LOW THEN y_2 is LOW (1.00)

IF x_1 is MID THEN y_2 is HIGH (1.00)

IF x_1 is HIGH THEN y_2 is LOW (1.00)

IF x_3 is LOW THEN y_2 is LOW (0.77)

IF x_3 is HIGH THEN y_2 is HIGH (0.95)

IF x_2 is 1:15 THEN y_2 is HIGH (0.79)

IF x_2 is 1:9 THEN y_2 is LOW (1.00)

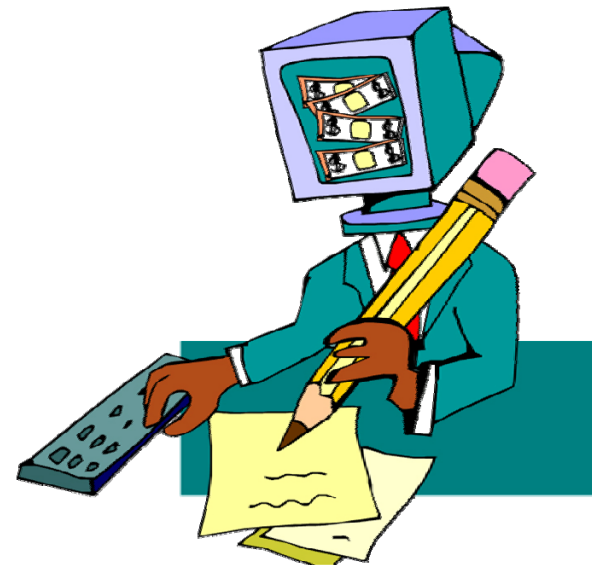
IF x_2 is 1:12 THEN y_2 is LOW (0.56)

Related to the phyllanthin content

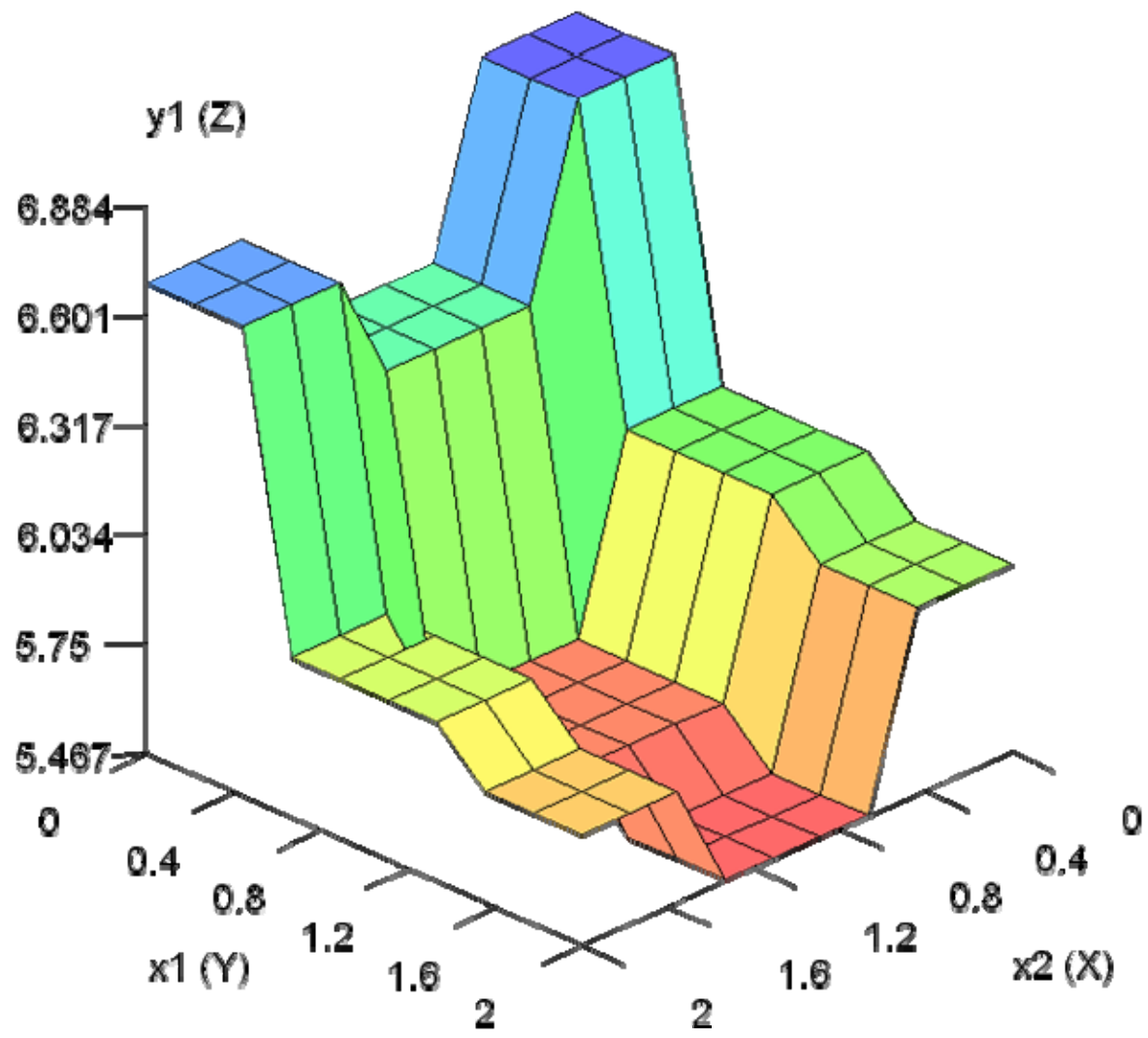
IF x_1 is LOW THEN y_1 is LOW (0.97)

IF x_1 is MID THEN y_1 is LOW (0.73)

IF x_1 is HIGH THEN y_1 is HIGH (0.83)



Visualization of the relationship $y_1 = f(x_1, x_2)$



Problems of interest
Design of experiments
Cause-effect relationships
Multivariate optimization
Practical applications



The general ideas of optimization

In the simplest case, optimization means solving problems in which one seeks a **minimum** or **maximum** value of some objective function.

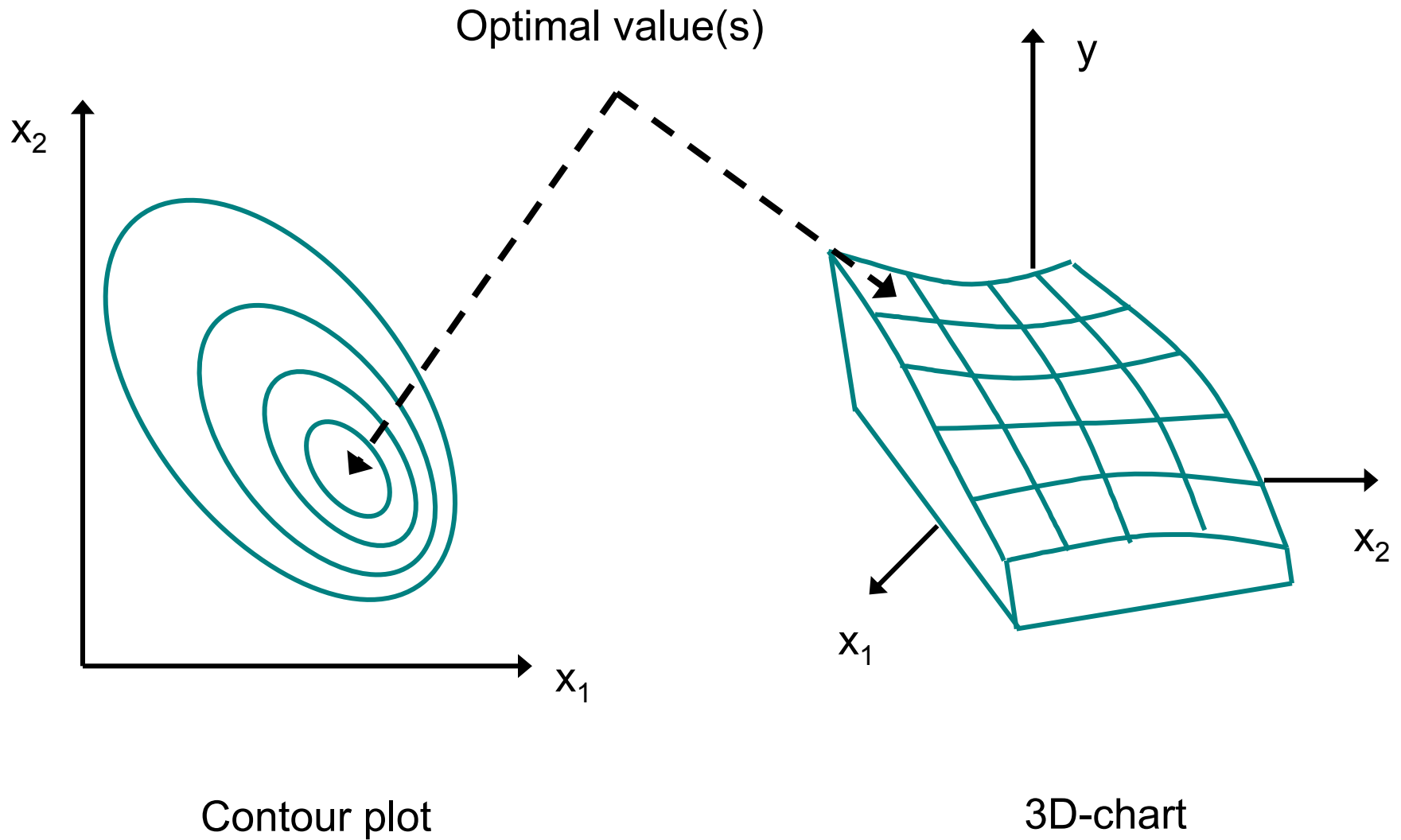
More generally, optimization means finding "best available = **optimal**" values of a number of objective functions in a defined domain.

To optimize an extraction process, one must find the best balance (or **multiple optimization**) of many dependent variables at the same time:

- Extraction yield: Maximum
- Active substance(s): Maximum
- Unwanted substance(s): Minimum
- Extract cost: Reasonable

The so-called multiple optimization can be professionally performed by intelligent software systems.

Graphic presentation of optimal values



Mathematical approaches for optimization

Mathematical optimization approaches have a number of limitations: **unsuitable** to many independent variables (x); using **only one** dependent variable (y) at a time; always requiring a **linear** mathematical equation, ...
For example, with 2 independent variables:

- Multiple linear regression:

$$y = b_0 + b_1x_1 + b_2x_2$$

- Multiple regression with the 1st order interaction:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$

- Multiple regression with the 1st and 2nd order interactions:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2$$

- Lagrangian-type mathematical equation:

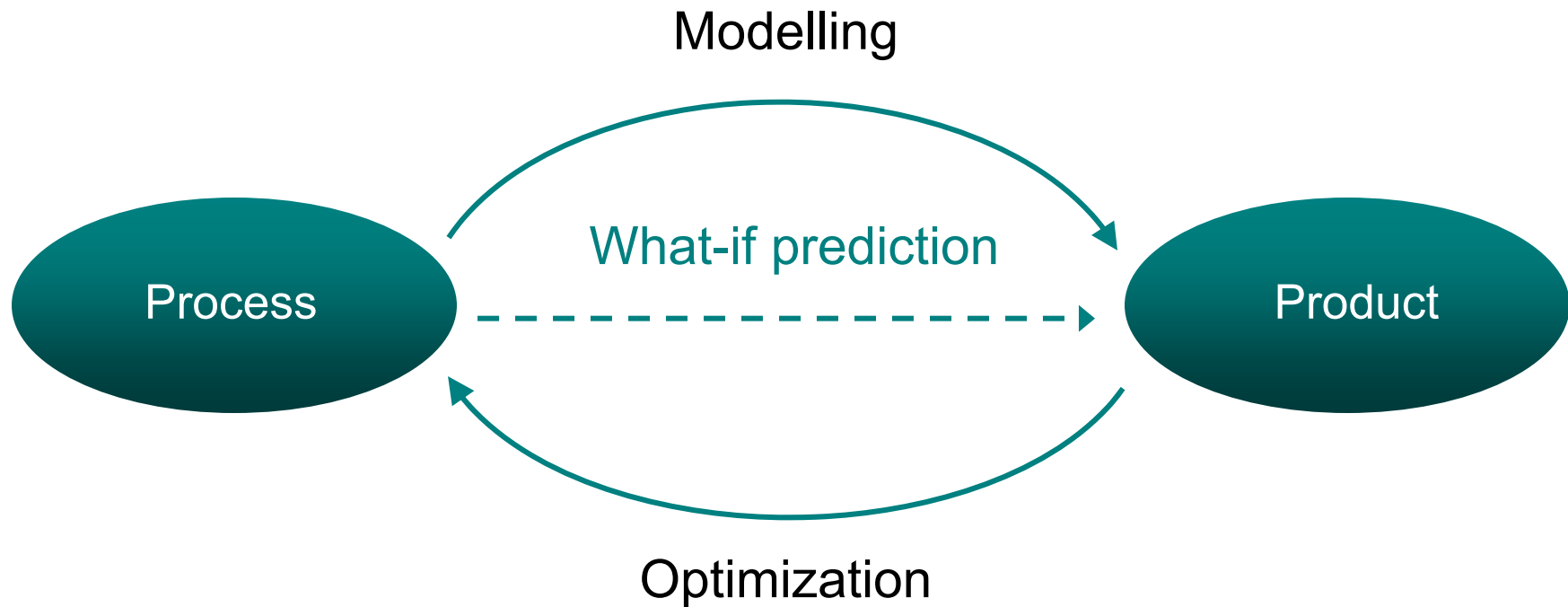
$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2 + b_6x_1^2x_2 + b_7x_1x_2^2 + b_8x_1^2x_2^2$$

Intelligent approaches for optimization

- a. Suitable to a large number of independent variables (x_i).
- b. Using many dependent variable (y_i) at the same time.
- c. Not requiring mathematical modelling of the domain.
- d. Capable of generating correct responses with training data.
- e. Able to deal with complex, incomplete or noisy data.
- f. Providing users with intuitive and convenient optimizers.

- | | | |
|--------------------------|------------|-------------------|
| - Extraction yield: | Maximum | Up |
| - Active substance(s): | Maximum | Up |
| - Unwanted substance(s): | Minimum | Down |
| - Cost: | Reasonable | Tent or Flat Tent |

Neural networks with genetic algorithms

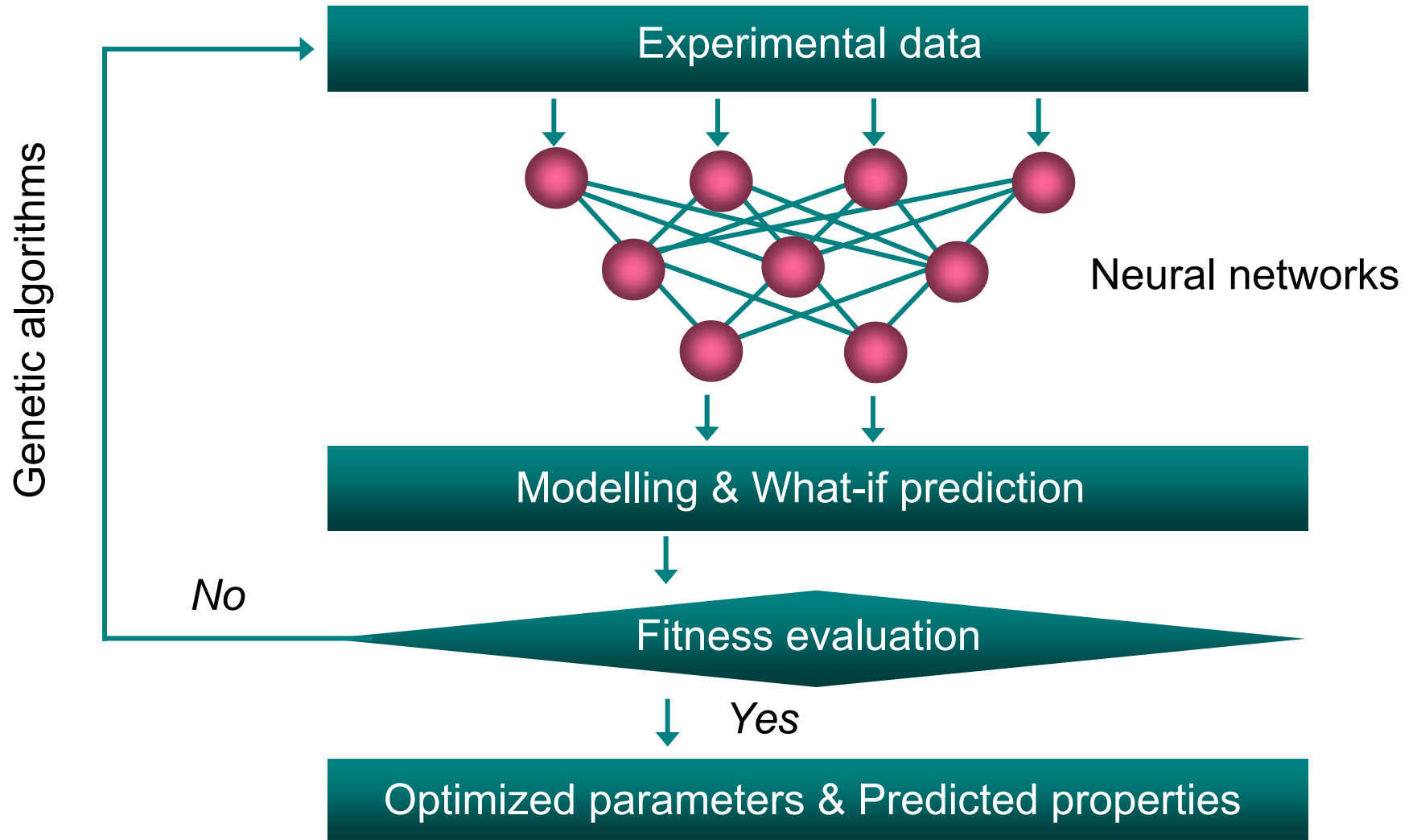


INForm intelligent software, employing neural networks in combination with genetic algorithms:

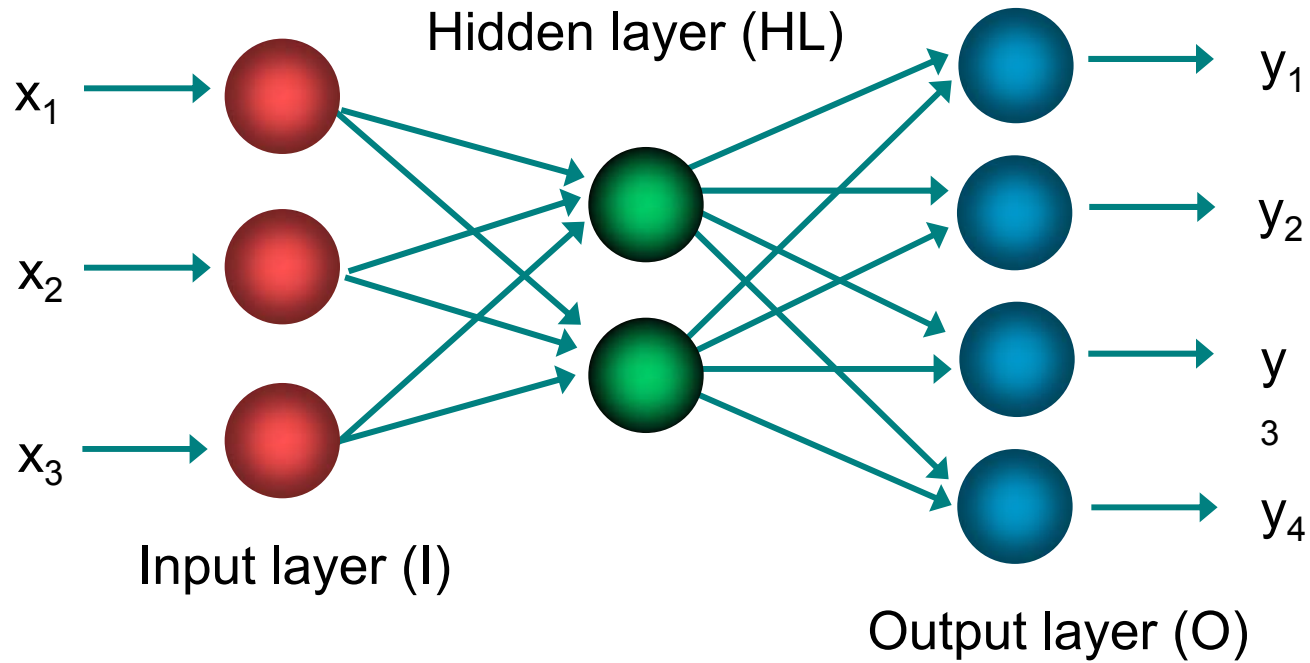
- neural networks: **modelling & what-if prediction**
- genetic algorithms: **multivariate optimization**



The principle of intelligent optimization



The multilayer perceptron with one hidden layer



I(a)-HL(b)-O(c):

a = Number of input nodes

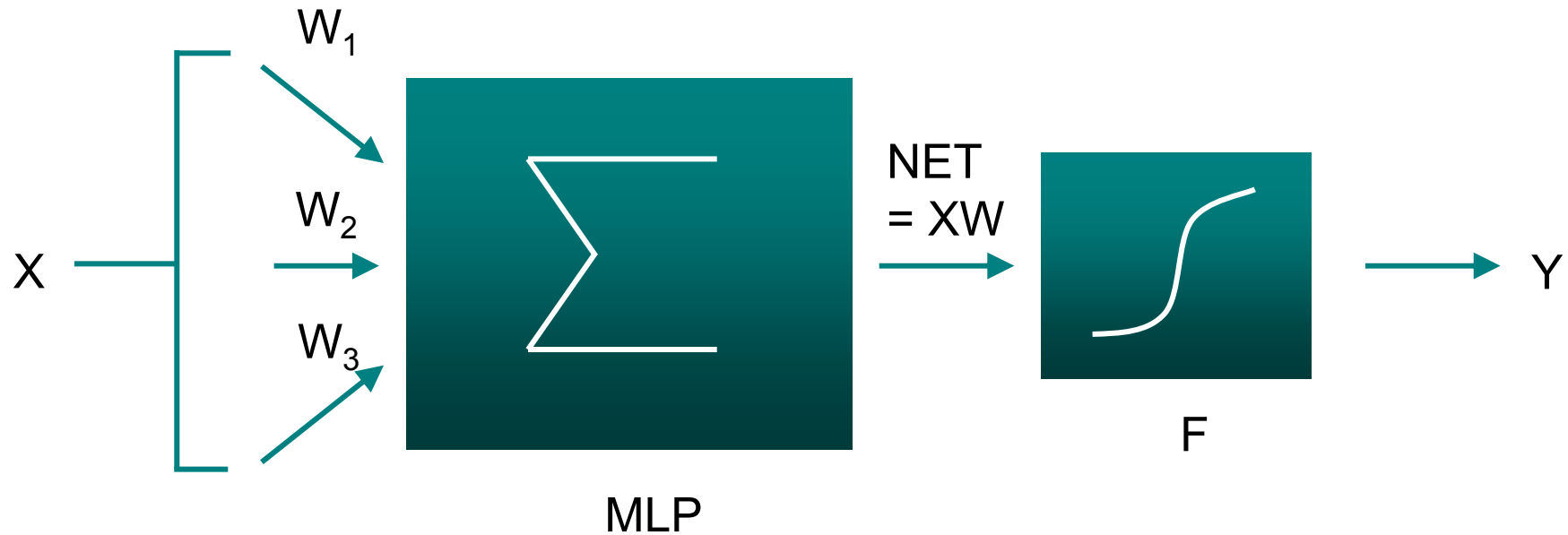
b = Number of hidden nodes

c = Number of output nodes

A black box

Multilayer Perceptron Networks (MLP)

The role of transfer functions



X = Input

Y = Output

W = Weight

F = Transfer functions:

Linear

Sigmoid (Symmetric or Asymmetric)

Tanh (Hyperbolic Tangent)

Neural network training (learning)

Back-propagation algorithms (BPA)

$$\Delta w_{ij}(n) = \varepsilon \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w_{ij}(n-1)$$

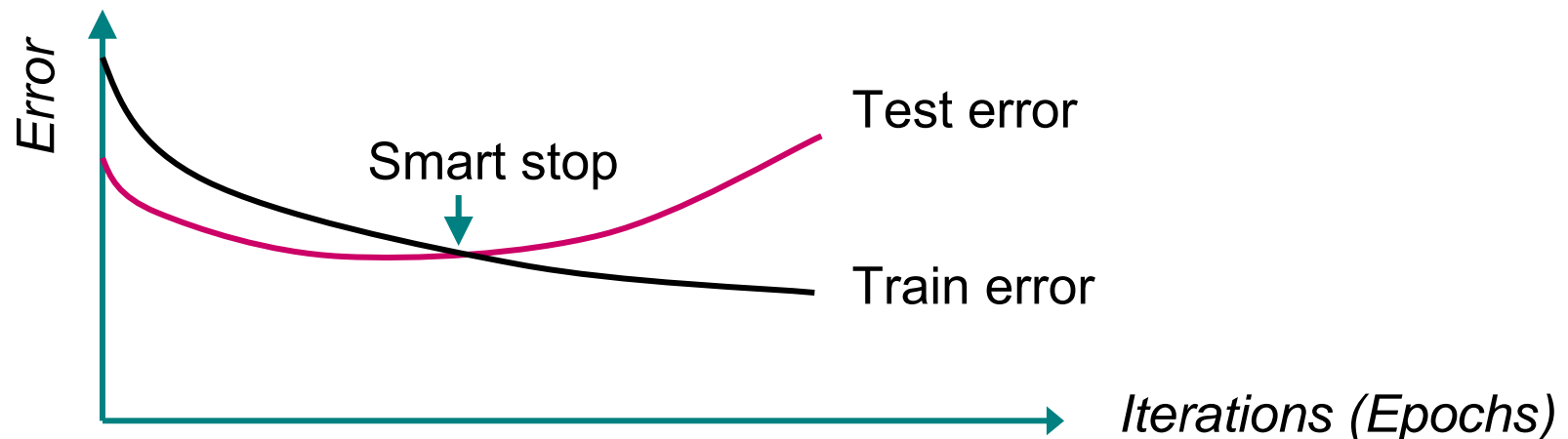
ε : Learning rate

α : Momentum

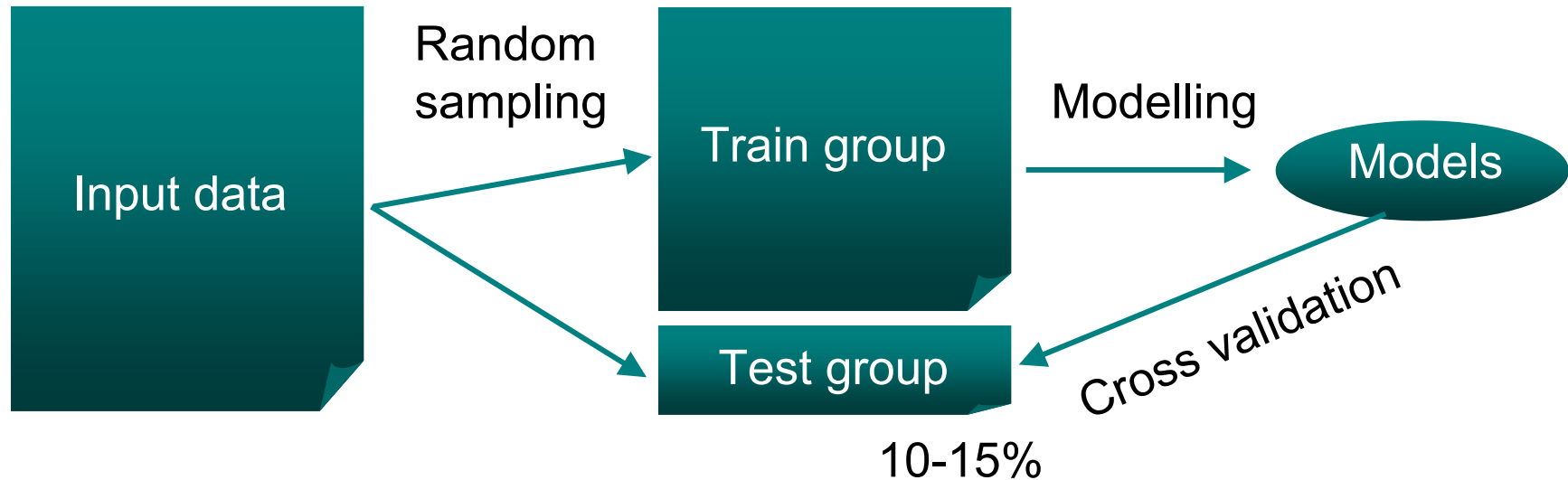
BPA in INForm:

- Standard Incremental
- Standard Batch
- RPROP
- Quickprop
- Angle Driven Learning

Smart stop functionality



Cross validation of the cause-effect models



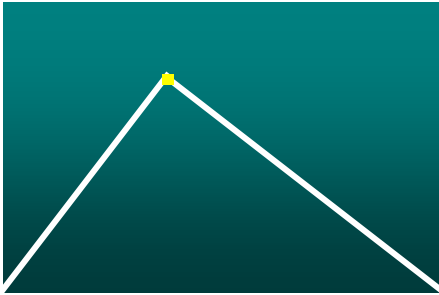
$$R^2 = \left(1 - \frac{ESS}{TSS}\right) 100 = \left(1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}\right) 100$$

y : Observed values

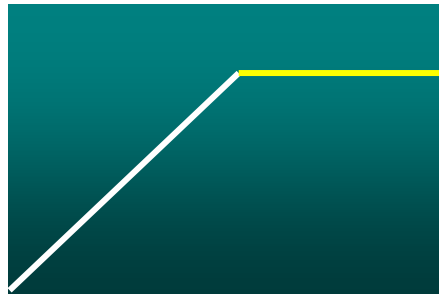
\hat{y} : Predicted values:

- Internal prediction (Train group): Train R^2 ($\geq 90\%$) \leftrightarrow Fitting goodness
- External prediction (Test group): Test R^2 ($\geq 70\%$) \leftrightarrow Predictability

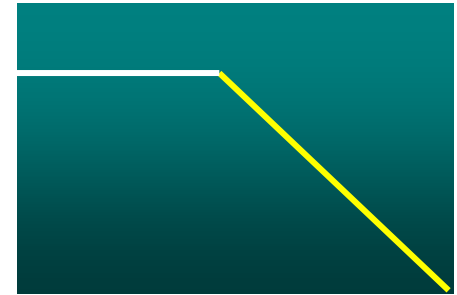
Types of desirability functions (in INForm)



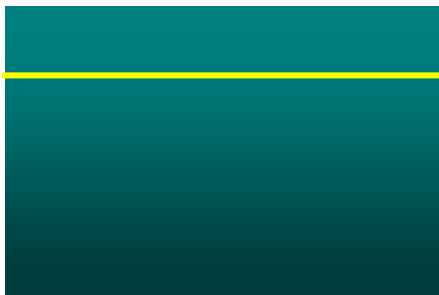
TENT



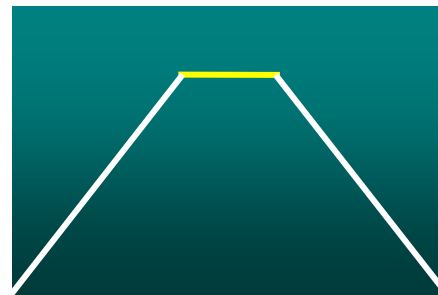
UP



DOWN



FLAT



FLAT TENT

Goal to optimize the extraction process

Research goal

To optimize the parameters including ethanol concentration, material-solvent ratio and extraction times for obtaining highest extraction yield and maximal phyllanthin content?

Extraction process

For the dried leaves of *Phyllanthus amarus* Schum. & Thonn.

x_1 = Ethanol concentration (low, medium, high) ?

x_2 = Material-solvent ratio (1:9, 1:12 , 1:15) ?

x_3 = Extraction times (2, 3) ?

y_1 = Extraction yield (%) maximum

y_2 = Phyllanthin content maximum

The data input for INForm intelligent software

	x_1	x_2	x_3	y_1	y_2
1	Mid	1:15	3	7.21	1.55
2	Mid	1:9	2	6.34	1.36
3	Mid	1:15	2	6.92	1.28
4	Low	1:9	2	5.50	0.67
5	Low	1:12	3	6.30	0.55
6	Low	1:9	3	6.06	0.63
7	Mid	1:9	3	6.89	1.07
8	High	1:15	2	6.02	4.01
9	High	1:9	3	5.89	3.04
10	Low	1:15	2	6.10	0.47
11	Mid	1:12	2	6.71	1.91
12	High	1:12	3	6.12	3.95
13	Low	1:12	2	5.88	0.52
14	High	1:15	3	6.32	2.66

The results of neural network training

Training parameters

- Test group: 3 and 9
- Back propagation algorithm: Standard Batch
- Transfer function: Asymmetric Sigmoid
- Output transfer function: Linear

Cross validation

R ² values (%)	y ₁	y ₂
Train R ²	99.7273	90.6542
Test R ²	97.6805	76.7828

Model y₁ proved to be very good in fitting and exact in prediction.

Model y₂ proved to be good in fitting and fairly exact in prediction.

Settings for the extraction process optimization

Constraints

None

Weight

Default (1)

Integer

x_2 = Material-solvent ratio = Integer

x_3 = Extraction times = Integer

Desirability functions

y_1 = Maximum $\Leftrightarrow y_1 = \text{Up}$

y_2 = Maximum $\Leftrightarrow y_2 = \text{Up}$



Results of the extraction process optimization

Optimization output

Optimized parameters:

x_1 = Ethanol concentration = Mid
 x_2 = Material-solvent ratio = 1:15
 x_3 = Extraction times = 2

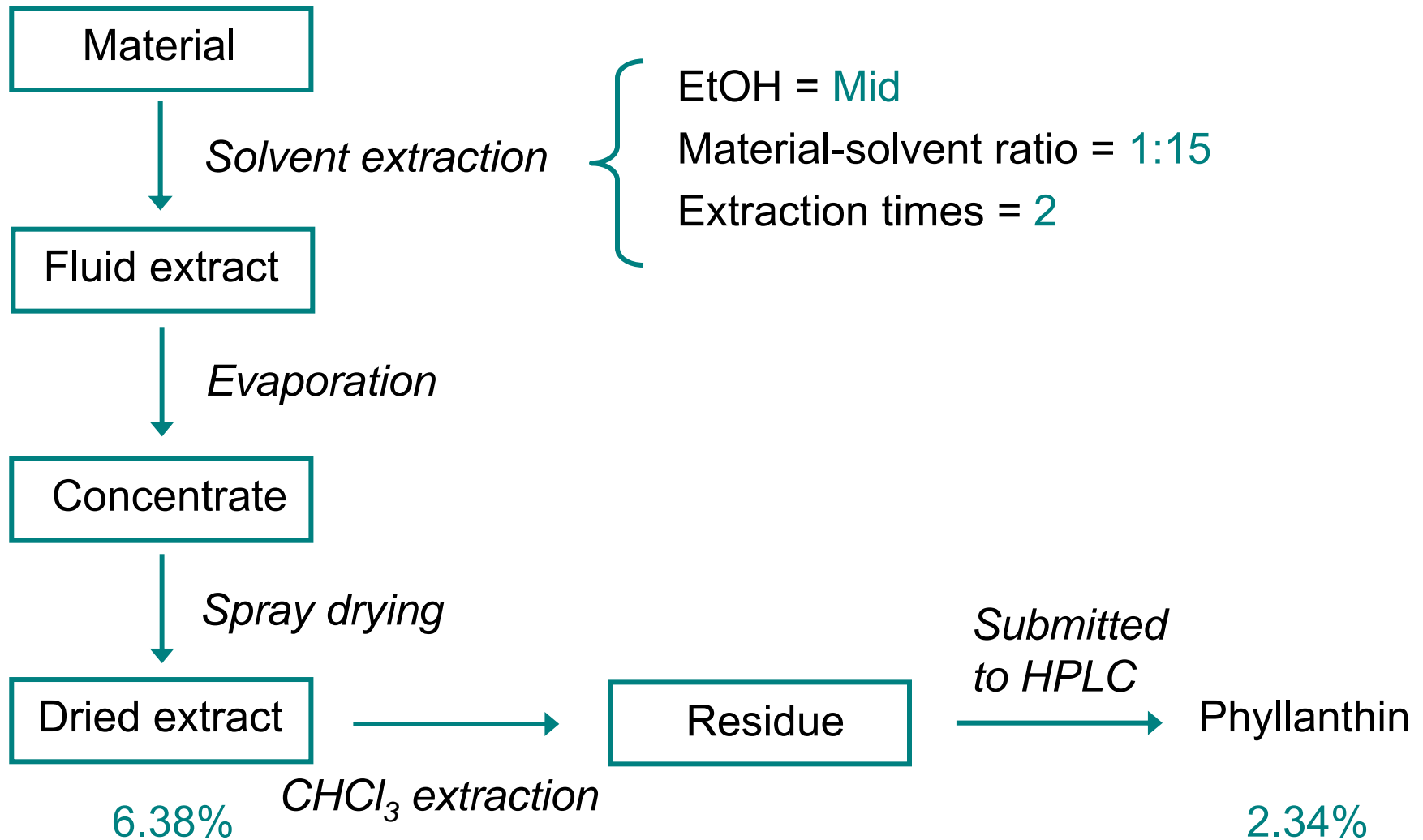
Predicted properties:

y_1 = Extraction yield = 6.38%
 y_2 = Phyllanthin content = 2.34%

Experimental validation

	Experimentation				Prediction
	1	2	3	Mean	
Yield (%)	6.18	5.98	6.09	6.09	6.38
Phyllanthin (%)	2.73	2.65	2.40	2.40	2.34

The extraction process with its optimized parameters



Problems of interest
Design of experiments
Cause-effect relationships
Multivariate optimization
Practical applications



Medicinal herbs mentioned in the examples



Phyllanthus amarus
Schum. & Thonn. (leaves)



Curcuma longa
L. (rhizomes)



Centella asiatica
(L.) Urb. (leaves)



Houttuynia cordata
Thunb. (leaves)



Morinda citrifolia
L. (fruits and roots)

Summary of the D-Optimal designs

Independent variables

- x_1 : Ethanol concentration (low, medium, high)
- x_2 : Material-solvent ratio (low, medium, high)
- x_3 : Extraction times (2 or 3 times)

Dependent variables

- y_1 : Extraction yield (%)
- y_2 : Marker content:

Phyllanthin

Phyllanthus amarus Schum. & Thonn. (leaves)

Curcumin I

Curcuma longa L. (rhizoma)

Asiatic acid

Centella asiatica (L.) Urb. (leaves)

Quercetin

Houttuynia cordata Thunb. (leaves)

Scopoletin

Morinda citrifolia L. (fruits)

Damnacanthal

Morinda citrifolia L. (roots)

Summary of the cause-effect relationships

Herbal extraction process	y_i	x_1	x_2	x_3	R^2
<i>Phyllanthus amarus</i> Schum. & Thonn., Euphorbiaceae (Leaves)	y_1	+	+	+	98.6346
	y_2	+	-	-	91.2376
<i>Curcuma longa</i> L., Zingiberaceae (Rhizoma)	y_1	+	+	+	93.5345
	y_2	+	-	-	91.2756
<i>Centella asiatica</i> (L.) Urb., Apiaceae (Leaves)	y_1	+	-	+	79.6059
	y_2	+	+	+	100
<i>Houttuynia cordata</i> Thunb., Saururaceae (Leaves)	y_1	+	+	+	100
	y_2	+	+	+	100
<i>Morinda citrifolia</i> L., Rubiaceae (Fruits)	y_1	+	+	+	100
	y_2	+	+	-	88.3219
<i>Morinda citrifolia</i> L., Rubiaceae (Roots)	y_1	+	+	+	99.2766
	y_2	+	-	-	90.8248

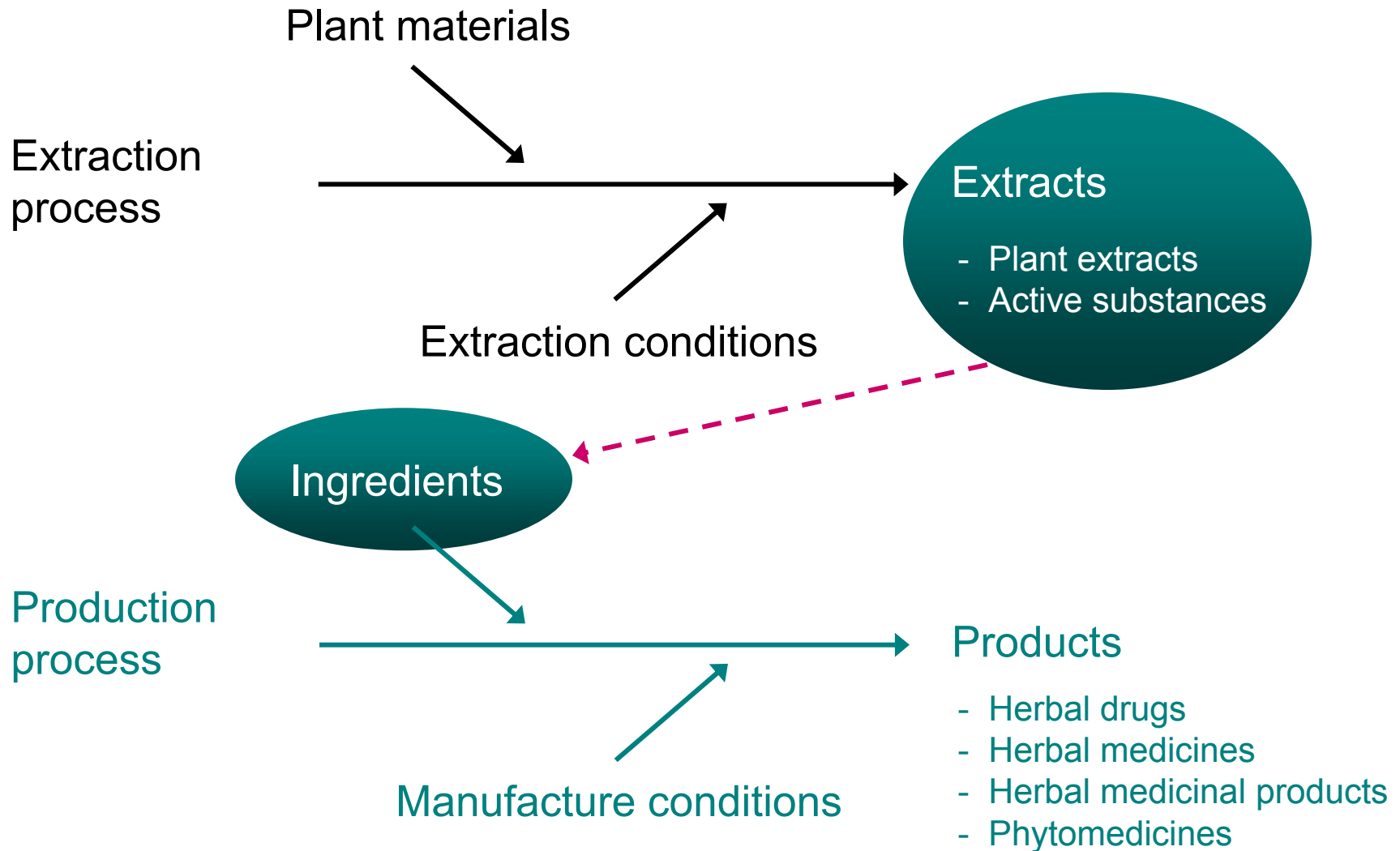
y_1 : Extraction yield (%); y_2 : Marker content (%)

Summary of the optimized extraction processes

Herbal extraction process	Properties	Predicted	Observed
<i>Phyllanthus amarus</i> Schum. & Thonn., Euphorbiaceae (Leaves)	y_1	6.38	6.08
	y_2	2.34	2.60
<i>Curcuma longa</i> L., Zingiberaceae (Rhizoma)	y_1	17.97	18.00
	y_2	8.75	8.82
<i>Centella asiatica</i> (L.) Urb., Apiaceae (Leaves)	y_1	9.65	11.09
	y_2	13.23	15.36
<i>Houttuynia cordata</i> Thunb., Saururaceae (Leaves)	y_1	7.63	7.14
	y_2	0.27	0.29
<i>Morinda citrifolia</i> L., Rubiaceae (Fruits)	y_1	11.76	10.98
	y_2	0.26	0.22
<i>Morinda citrifolia</i> L., Rubiaceae (Roots)	y_1	6.05	6.18
	y_2	2.74	2.69

y_1 : Extraction yield (%); y_2 : Marker content (%)

R&D applications: GEP and GMP



Acknowledgements

The examples in this presentation were derived from the Research Reports in 2009 of Prof. Dr. Nguyen Minh Duc and his MPharm students (Nguyen Duc Hanh, Nguyen Thi Linh Tuyen and Le Thi Hong Cuc), University of Medicine & Pharmacy, Ho Chi Minh City.

Links to Intelligensys Ltd. (UK)



Prof. Dr. Ray Rowe
Chief Scientist
Dr. Elizabeth Colbourn
Scientific Director



Prof. Dr. Peter York
Director

The author wish to thank [Prof. Dr. Ray Rowe](#) and [Prof. Dr. Peter York](#) for their kind arrangement of his visiting research at PROFITS group, Bradford University. [Dr. Elizabeth Colbourn](#)'s valuable advices for the usage of FormRules and INForm are gratefully acknowledged.