Perpetual Motion, Evolutionary Computation in Industry and other Chimeras

Anna I Esparcia Alcázar
Software Production Methods Research Centre
Universitat Politècnica de València, Spain
The Chimera

My great-grandfather Juan
Oh ye seekers after perpetual motion, how many vain chimeras have you pursued? Go and take your place with the alchemists.

— Leonardo da Vinci, 1494
In theory, theory and same, but in practise

Attributed to:

Albert Einstein

Yogi Berra
Objectives of this talk

• To show that it is possible to do Computational Intelligence in an SME
• To show that it is impossible to do Computational Intelligence in an SME
• ...likely, to destroy the Universe in the process
The Engineer’s enemies

1. Money
2. Time
3. Space

0. People
People you should fear

• Your boss
• Your colleagues
• Yourself
Profiles of EC practitioners

- Edison:
  “If it’s technically good, people will buy it”

- Merlin:
  “Gi’es yer requirements, we’ll put them in a cauldron with all sorts of algorithms and...voilà!”

- Archimedes
  “Who cares if they buy it?”
6 popular academic misconceptions

1. Companies need what you’ve got
2. Companies want what you’ve got
3. Companies understand what you’ve got (and if they don’t they must be blind/stupid/not worth talking to)
4. You don’t need to do development, just show them proof of the pure research and they’ll love/buy it
5. Just give it to the PhD student/R.A., s/he’ll manage
But... EC has been applied in the REAL world”, hasn’t it?
Things people have done with EC

• Design of diesel engine cylinders for French car manufacturer Peugeot
• Chevron has a patent for predicting oil well production
• Find patterns in big data: Nutonian’s tool, Eureqa has been used to predict... basketball results
• ...plus other things they won’t tell you about
Meanwhile, in España
My History

Chapter 1
When I was in the “other side”
Success stories (i)

1. Logistics and transport
   - Minimization of transport and inventory costs
   - Allocation of products to shelves
2. Finance
   - Bankruptcy prediction
   - Portfolio optimisation
3. Modelling & optimisation
   - Placement of pheromone dispensers in agriculture
4. Bio-signal classification
   - Brain-computer interface for the disabled
   - Neuromarketing, usability
5. Bot evolution in computer games
Success story #1 - EVITA

An IRP aims to find the answer to three questions:

1. **When** to serve each customer?
2. **How much** to deliver to each customer each time it’s served?
3. **What routes** to use for the delivery?
EVITA’s Objective

**Minimise** global costs, $c_g$

$$c_g = c_h + c_t$$

where:

- $c_t$: transport costs (delivery)

  $$c_t = \sum_{H} \sum_{k \in K} \sum_{(i,j) \in E} c_{i,j,k}$$

- $c_h$: inventory holding costs

  $$c_h = \sum_{i \in V} h_i$$
My Inventory & Routing Problem

• Given a depot that supplies a retail chain, determine:
  – For each shop
    • Optimal delivery **frequency** (number of days a week).
    • Optimal delivery **pattern** (which days of the week)
  – Once both are known
    • Optimal set of **delivery routes** (for each day of the week)

![Diagram of delivery routes and shops]

---

Shop 9:
Delivery: Mon, Wed, Fri (pattern 21)
Holding cost: 968 €

---

Shop 10:
Delivery: Mon, Thurs (pattern 18)
Restrictions & simplifications

- Each shop must be served by only one vehicle.
- Maximum delivery time per vehicle = 8 hours.
- Each shop has its own set of admissible frequencies.
- The demand of each shop depends on the frequency it is supplied.

- *Time windows*: fixed time intervals in which shops can be served
- Unlimited number of vehicles
- Not all patterns are admissible for each frequency.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Frequency (days per week)</th>
<th>L</th>
<th>M</th>
<th>X</th>
<th>J</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>2</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>29</td>
<td>4</td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>31</td>
<td>5</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Inventory holding costs and demands

<table>
<thead>
<tr>
<th>Shop type</th>
<th>Inventory cost (€)</th>
<th>Delivery size (roll containers)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency (days)</td>
<td>Frequency (days)</td>
</tr>
<tr>
<td></td>
<td>2  3  4  5</td>
<td>2  3  4  5</td>
</tr>
<tr>
<td>A</td>
<td>-  -  336  325</td>
<td>-  -  2  2</td>
</tr>
<tr>
<td>B</td>
<td>-  -  327  317</td>
<td>-  -  2  2</td>
</tr>
<tr>
<td>C</td>
<td>330  311  303  301</td>
<td>4  2  2  1</td>
</tr>
<tr>
<td>D</td>
<td>310  292  285  283</td>
<td>3  2  2  1</td>
</tr>
<tr>
<td>E</td>
<td>293  276  269  267</td>
<td>3  2  2  1</td>
</tr>
<tr>
<td>F</td>
<td>277  261  255  -</td>
<td>2  2  1  -</td>
</tr>
<tr>
<td>G</td>
<td>268  253  -  -</td>
<td>2  1  -  -</td>
</tr>
</tbody>
</table>

Shop type codes
- A - Valencia city centre
- B - Valencia other + shopping malls
- C - Valencia suburbs
- D - Coastal villages
- E - La Ribera county
- F - Other villages
- G - Requena-Utiel
Solving with CI (in 2 levels)

• Higher level: finds the pattern for each shop
  ➢ Genetic algorithm

| $p_1$ | $p_2$ | $p_3$ | $p_i$ | $p_n$ |

Where $p_i$ = pattern for shop $i$

• Lower level: finds the routes for each day
  ➢ Vehicle Routing Problem (VRP)
The Druni network
Success story #2 – MELiSSA

• Given the description of a new shop:
  – Shelves placement
  – Capacity (in modules) of each shelf

• Determine:
  • **Optimal number of modules** to allocate to each product group.
  • **Optimal location** of the modules occupied for each product group
i. e. I have this:
And I want something like this:
Or this:

Breakfast

Complements
There can be as many distributions...
... as there are shops
“Hard” restrictions

Space and capacity of each shelf.
Given by a table of distances and the description of the relative positions between shelves.

Standard shop

Values of an “ideal” shop, as defined by the mgt of the chain
(it may not coincide with the shop at hand)
Each group has a Standard/minimum/maximum:
→ ideal/minimum /maximum number of modules to assign.

Affinities & Adversities

Some products must be placed near or far from others, or
near a reference point (oven, checkout line etc.)
    Adverse: baby products & pet food
    Indifferent: baby products & sauces
    Affine: snacks& beverages
    Affine to a reference point: biscuits& the oven
“Soft” restriction

Cohesion between groups
Multiobjective Evolutionary Algorithm (MOEA)

Optimal number of modules: **minimal deviation**
- If the size of the Standard Shop is less than the actual shop, all the groups must at least be in their standard
- If the size of the Standard Shop is more than the actual shop, adjust as best as possible.

Relative position of groups: **maximal affinity**
- Affine/adverse groups must be placed near/far.
- Groups affine to a reference point must be placed as near of it as possible

Maintain cohesion between groups: **minimal dispersion**
- Avoid “lost” modules
- Aim at making groups occupy full shelves
### The problem

<table>
<thead>
<tr>
<th>Group</th>
<th>Standard shop</th>
<th>Affine to</th>
<th>Adverse to</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>std</td>
<td>min</td>
<td>max</td>
</tr>
<tr>
<td>1 Food in general</td>
<td>14</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>2 Drugstore</td>
<td>40</td>
<td>34</td>
<td>52</td>
</tr>
<tr>
<td>3 Snacks</td>
<td>7</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>4 Sauces &amp; condiments</td>
<td>9</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>5 Bakery</td>
<td>21</td>
<td>15</td>
<td>28</td>
</tr>
<tr>
<td>6 Baby Products</td>
<td>13</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>7 Beverages</td>
<td>17</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>8 Pet products</td>
<td>9</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>Size of standard shop</td>
<td>130</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### A solution

![Diagram of a standard shop layout]

32
(In)Famous quotes

I don’t know how to sell.

A certain Commer
Success story # 3
Bankruptcy prediction

• Given a database of companies data for a given year $k$, predict:
  – If the company will go **bankrupt** in year $k+n$
  – If the company will **suffer loses** in year $k+n$

• Imbalanced classification problem: there are many more companies that go bankrupt than “healthy” ones

... **luckily!**
Why use genetic programming?

• It generates **comprehensible** & **analyzable** solutions.
• It allows to determine which data are relevant for the prediction.
• The reduction in the number of data brings about the generation of **simpler models**.
One solution

\[
y = \text{SI \ } \frac{x_0}{x_1 - x_2} \leq x_i \ \text{exp}(x_3)
\]

ENTONCES \ DEVUELVE \ 
\[
if_0 = \text{SI \ } x_2 \leq x_0
\]

ENTONCES \ DEVUELVE \ 
\[
if_1 = \text{SI \ } x_2 \leq x_3 + 43.45
\]

ENTONCES \ DEVUELVE \ 
\[
\text{SI \ } x_2 \leq x_3 + 43.45
\]

ENTONCES \ DEVUELVE \ 
\[
y > 0 \Rightarrow \text{THE COMPANY WILL GO BANKRUPT}
\]

\[
y \leq 0 \Rightarrow \text{THE COMPANY WILL NOT GO BANKRUPT}
\]
Famous quotes

Prediction is very difficult, especially of the future

Niels Bohr
(attributed also to Yogi Berra)
Success story #4
Pest control in agriculture

• **Sexual confusion** is a technique aimed at substituting pesticides.
• It consists on diffusing a high amount of female sexual pheromone in order to confuse the males and avoid coupling.
• How? → Using **pheromone dispensers**
Ecology in figures

• 1 kg of pheromone costs 1000 €
• 1 dispenser contains 200 mg of pheromone
  → One dispenser costs 20 cent. (+ manufacturing)
• In 1 Ha we need 500 or 1000 dispensers (depending on the pest)
  → The cost is 100 or 200 € / Ha (+ hand work)
    (real commercial price is 115 or 300 €/Ha)

On the other hand,

• Spraying a traditional pesticide (e.g. Malathion) costs 20-30 €/Ha
“Fields” of application

Vineyards, 500 dispenser per Ha.

Apples and pears, 1000 dispensers per Ha.

Photos courtesy of CEQA - IAM - UPV
Modeling & optimisation for sexual confusion systems

- The price of pheromone dispensers is a limitation for their massive application
  - Aim is to optimise their number and location to ensure maximal crop protection at an affordable price
  - Besides, models are required that allow making the costs of registration of the products cheaper
Aplication of Genetic Programming

Work carried out: Finding models of release kinetics

Let the residual $r$ be the percentage of product not released into the atmosphere
For a given dispenser, find a function $r = r(t)$ where $t = \text{time}$
Data: a sequence of points $(r,t)$ obtained in field conditions
Measuring $r$ is costly $\rightarrow$ few measures,
Measures unequispaced $\rightarrow$ more at the beginning, when the release is faster)
Resolution with MOGP
Cost function (“leave-one-out”):

$$\text{Obj}_i = |r_{\text{calculated}}(i) - r_{\text{measured}}(i)|$$

Further work:

Find models of pheromone distribution in the environment
Use these models to optimise the placement of dispensers and make cheaper the registration costs
Famous quotes

Us people from Madrid go to the countryside to ensure cows are not purple

J. Ignacio Hidalgo, 2009
Success story #5

Bio signal Classification

Applications in:
• BCI for communication & mobility of disabled persons,
• Neuromarketing
• Usability
Brain-computer interface (BCI)

- Objective: Control of devices by thought by means of the measurement and analysis of the electrical activity in the brain using electroencephalogram (EEG)
- Based on the fact that different thoughts originate different EEG patterns
Biosignal classification using CI

EEG data acquisition → Pre-processing → Feature Extraction

$\alpha(1) = \{a^1, b^1, \ldots\}$

$\alpha(2) = \{a^2, b^2, \ldots\}$

...$\alpha(n) = \{a^n, b^n, \ldots\}$

Validation → Classification → Data preparation

Repeat $n$ times
(In)Famous quotes

Who told you we wanted you to publish?

A Certain Manageress at a Certain Technological Institute, 2008
My History

Chapter n+1:

When I was in the "other" other side
Success stories (ii)

6. Evolving rules for a correlation system
7. Detection of Advanced Persistent Threats
A model company

• An SME specialising in cybersecurity that has developed its own product suite, which includes:
  – Sensors to measure physical, logical and business variables
  – A rule-based correlation platform, which triggers higher level events
  – An event management console, where technicians handle events
Event bus

Sensors

Processes of the organisation

Correlation platform

Event management console

Active dashboard
Beware of dangerous questions!

Is this an expert system?

Anna Esparcia to her boss, 2011
Expert systems have been obsolete for 10 years

Anna Esparcia to her boss, 2001
Problems with expert systems

• Technical Problems
  – Consistency: difficulty in detecting erroneous knowledge, handle differences in expert’s opinion, verify the correct interaction between rules
  – Scalability: small increases in system size make knowledge management and especially acquisition become unfeasible
  – Static, no learning
  – Subjective way of representing knowledge: difficult to back up expert-defined rules with physical or statistical data

• Structural Problems
  – Cost of acquisition and maintenance of knowledge

• Human Problems
  – Knowledge Extraction from experts
  – Distribution of incompetence (Peter Principle applied to pseudo-experts)
  – Legal issues
Success story #6: TritonLearn - Making a rule-based system “smart” using CI
TritonLearn

Capture .pcap

My Snort

RBS in a sandbox

Output .txt

Event Management System

Rule ¿set?

TritonLearn engine

56
Evolving population of rules

Evolutionary engine

New rule

Evolving population of rules

TritonLearn evaluator

TritonLearn

Representation

Evolutionary operators
  - Reproduction
  - Xover
  - Mutation

Evolutionary parameters

fi  eval(i)
TritonLearn problems

• Unconnected systems
• No logs
• Lack of understanding between parties
• Technicians do not have time for you
• Company changes business focus

Solution:
• Let’s do something else!
Success story #7: Advanced Persistent Threats, the new hype in security

A possible definition of APT:
A highly sophisticated targeted attack involving very skilled teams and significant financial resources

• One of the steps when they can be detected is called *exfiltration*

• During exfiltration it is assumed that the behaviour of the system will be anomalous

→ we’ll focus on *anomaly detection*
Example of http traffic

200 GET text/plain com 276 55490 notify9.dropbox.com
200 GET text/plain com 276 55246 notify5.dropbox.com
200 POST null com 197 307 twitterfall.com
200 GET application/octet-stream com 273 75962 su.ff.avast.com

... up to 5M instances
Available Data: HTTP session

- Characteristics:
  - 5 Million instances
  - 10 attributes/fields per instance
  - 4 types: categorical [c], numerical [n], string [s], timel [t].

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>http reply code</td>
<td>c</td>
</tr>
<tr>
<td>http method</td>
<td>c</td>
</tr>
<tr>
<td>Duration (ms)</td>
<td>n</td>
</tr>
<tr>
<td>Content type</td>
<td>c</td>
</tr>
<tr>
<td>Server IP</td>
<td>c</td>
</tr>
<tr>
<td>Time</td>
<td>t</td>
</tr>
<tr>
<td>Squid hierarchy</td>
<td>c</td>
</tr>
<tr>
<td>Bytes</td>
<td>n</td>
</tr>
<tr>
<td>Client IP</td>
<td>c</td>
</tr>
<tr>
<td>URL</td>
<td>s</td>
</tr>
</tbody>
</table>

- General data properties: multi-attribute, several types of attributes, sequential and data from different users.
## Data: Example

<table>
<thead>
<tr>
<th>http reply_code</th>
<th>http method</th>
<th>Duration (ms)</th>
<th>content_type</th>
<th>server address</th>
<th>Time</th>
<th>Squid hierarchy</th>
<th>bytes</th>
<th>Client address</th>
<th>url (reduced)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>GET</td>
<td>2014</td>
<td>application/javascript</td>
<td>192.168.4.4</td>
<td>05/03/2014 6:26:55</td>
<td>DEFAULT_PARENT</td>
<td>798</td>
<td>10.159.74.184</td>
<td>//syndication.twimg.com</td>
</tr>
<tr>
<td>200</td>
<td>GET</td>
<td>55581</td>
<td>text/plain</td>
<td>108.160.162.33</td>
<td>05/03/2014 6:26:57</td>
<td>DIRECT</td>
<td>276</td>
<td>192.168.4.4</td>
<td>//syndication.twimg.com</td>
</tr>
<tr>
<td>200</td>
<td>GET</td>
<td>55622</td>
<td>text/plain</td>
<td>192.168.4.4</td>
<td>05/03/2014 6:26:58</td>
<td>DEFAULT_PARENT</td>
<td>326</td>
<td>10.159.74.184</td>
<td>//notify5.dropbox.com</td>
</tr>
<tr>
<td>200</td>
<td>POST</td>
<td>135</td>
<td>application/vnd.google.safebrowsing-update</td>
<td>173.194.34.231</td>
<td>05/03/2014 6:27:02</td>
<td>DIRECT</td>
<td>885</td>
<td>192.168.4.4</td>
<td>//notify5.dropbox.com</td>
</tr>
<tr>
<td>200</td>
<td>POST</td>
<td>139</td>
<td>application/vnd.google.safebrowsing-update</td>
<td>192.168.4.4</td>
<td>05/03/2014 6:27:02</td>
<td>DEFAULT_PARENT</td>
<td>970</td>
<td>10.159.128.70</td>
<td>//safebrowsing.clients.google.com</td>
</tr>
<tr>
<td>200</td>
<td>GET</td>
<td>105</td>
<td>application/vnd.google.safebrowsing-chunk</td>
<td>173.194.34.232</td>
<td>05/03/2014 6:27:06</td>
<td>DIRECT</td>
<td>413</td>
<td>192.168.4.4</td>
<td>//safebrowsing.clients.google.com</td>
</tr>
<tr>
<td>200</td>
<td>GET</td>
<td>111</td>
<td>application/vnd.google.safebrowsing-chunk</td>
<td>192.168.4.4</td>
<td>05/03/2014 6:27:07</td>
<td>DEFAULT_PARENT</td>
<td>418</td>
<td>10.159.128.70</td>
<td>//safebrowsing-cache.google.com</td>
</tr>
</tbody>
</table>

**Note:** Content type and URL attributes can be divided in sub fields.
Unbalanced binary classification

- Two classes: suspicious / non-suspicious
- Labels are assigned by experts
- To obtain the training/testing sets
  - Step 1: Filter all data using anomaly score
  - Step 2: Expert classifies anomalous data
Abnormal attributes and combinations

• Detection of abnormal individual values:
  – *Infrequent values* in categorical attributes/characteristics.
  – *Values out of the normal range* distribution in num attributes/characteristics.
  – Logs in *abnormal time* interval.

• Detection of abnormal combinations:
  – Of code, method, content and tld attributes.
Semi-Supervised Learning

• Design method to automatically detect suspicious instances
  – Hypothesis: instances not labelled are non-suspicious.
  – 2/3 lab instances for training and 1/3 for test
  – Categorical attributes converted to integers

• Methods applied
  – SVM
  – GP
  – treefit (CART tree)
Problems

- Technicians still do not have time for you
- Company business focus still unclear
- Difficult to access real data
(In)Famous quotes

I just want a generic system that detects everything

A Certain Director of Development in a Certain Company, 2013
Famous quotes

Success is the ability of failure without losing
Chapter n+2: Full Wolfe
Success story #8: Automated software testing

TESTAR is a tool developed as part of European project FITTEST (2010-2013)
How Test works

START SUT

SCAN GUI + OBTAIN WIDGET TREE

DERIVE SET OF USER ACTIONS

optional instrumentation

more sequences?

STOP SUT

EXECUTE ACTION

SELECT ACTION

FAULT?

ORACLE

Replayable Erroneous Sequences

Domain Experts

Action Definitions

Oracle Definition

Yes

No

Yes

No

No

more actions?

Oracle Definition

Action Definitions
Test* in the “real world”

• In 2006 the UPV launched a programme called “Proof of concept” aimed at transferring results of research to companies
• As part of this programme, TESTAR has been deployed in 10 companies so far
• We have set up the Spanish Software Testing Innovation Alliance
Some possible keys to success

• Optimism
  – Believe in what you do
  – Deal with negativity and resistance

• Determination
  – Keep trying
  – Follow up with companies

• Visibility
  – Never miss a chance to be an evangelist
  – Use social media
In a good cause there are no failures, only delayed successes

Isaac Asimov, “The complete stories”
Take home ideas
Sometimes I have believed six impossible things before breakfast

L. Carroll, “Through the looking glass and what Alice found there”
Six impossible things

• There is life beyond the h-index
• You can apply EC in a company – even an SME, even in a “lesser” country
• You can understand what companies want
• You can convince them that EC is not just a PR stunt, or “vapourware”
• Customers will appreciate (perhaps even understand) what you do
• Your boss will be happy
So... perhaps the Chimera exists, and is as fierce as you want to make it
Thanks to:

• Lewis Carroll: “Alice in Wonderland” and “Through the looking glass and what Alice found there”
• Lewis Wolpert: “Six impossible things before breakfast”
• Arthur Kordon: “Applied Computational Intelligence: How to create value”
Any questions?