SCIENTIFIC PUBLICATION: Writing scientific papers, publication channels, evaluation procedures, citations

Universidade do Minho

Paulo Cortez pcortez@dsi.uminho.pt

http://www3.dsi.uminho.pt/pcortez

Departamento de Sistemas de Informação

Universidade do Minho

Guimarães

My CV: http://www3.dsi.uminho.pt/pcortez

Home CV R&D Projects Publications Working with me Downloads

Associate Professor
(Habilitation, PhD)

LinkedIn

ResearchGate

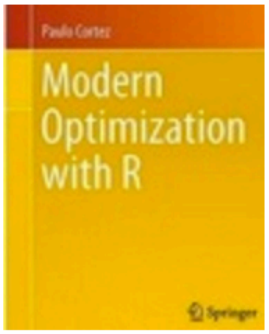
Paulo Cortez

Paulo Cortez is **Associate Professor** (with Habilitation) at the Department of Information Systems and **Coordinator** of the Information Systems and Technologies (IST) research group of ALGORITMI Research Centre, University of Minho (Short CV).

Current research interests:
* Business Intelligence: Decision Support Systems, Data Mining and Forecasting;
* Artificial Intelligence (AI): Neural Networks, Evolutionary Computation and AI Applications.

Co-author of more than 110 indexed (e.g., ISI, Scopus) publications in international scientific journals and conferences.

New/relevant items:
* **Research Grants: 2 PhD and 1 MSc grants.**
* **Best Paper Award** of the EPIA 2015 conference, 2015.
* Author of the **rminer** R package (version 1.4.2, 9/2016).
* Springer book: **Modern Optimization with R, 2014 (188 pages)**



Objectives of this lecture

To understand a simple set of rules that will guide in performing a better understanding of scientific publication and scholarly communication.


OK, I WANT TO WRITE MY THESIS, BUT WHERE DO I START??
remember, cecilia... every thesis begins with a single keystroke...
LAO-TZE!
JORGE CHAM © 2009
WAIT, WHICH K-
good luuuck!!
WWW.PHDCOMICS.COM

#1: Know what are the Scientific Publication Types

Q1: What are the Scientific Publication Types and their characteristics??

(In-Class Teams)

-Join in groups of +-3 elements.

-Find who within the group was born closer to this room → Team Leader (write and speak answers)

-Be fast (Total time:1 minute)

-Think in terms of Quantity

#1: Know what are the Scientific Publication Types

- **With** and Without **Peer Review (+)**;
- National vs. **International (+)**;
- Monograph: Final graduation report, MSc/PhD Thesis, ...
- Article in Proceedings (Conference) and **Journal (+)**;
- **Book (+)** or Book Chapter;
- Types of papers: Position paper, Theoretical Paper, **Research Paper**, Case Study paper, Industrial paper, Technical Paper, Invited Paper, Editorial paper, ..
- Others: Technical Report, Web Page, ...

http://en.wikipedia.org/wiki/Scientific_publishing#Types_of_scientific_publications

Productivity vs. Impact

To evaluate research publications (e.g., CV, R&D Unit, University, …) you can use:

- **Productivity** – Total number of papers, papers/year,…
- **Impact** – Total number of citations, citations/year, citations/papers, …
- Both – **H-index** (there is a number of H publications with at least H citations each)

#2: Know Where to Publish

- Ask your supervisor (or other experts) for advice;
- Be aware of **Write Only conferences**! (only productivity numbers, not impact)
- Search for scientific indexes and databases: ISI (web of science, JCR), SCImago SJR, DBLP, Scopus, ACM digital library, IEEE Xplore, ...
- **Other factors**: publisher, acceptance rate, program/technical committee, …

Final Goal: **match the quality of the publication target with the quality of the research.**

#2: Know Where to Publish

- http://isiknowledge.com
- http://www.scimagojr.com

JOURNALS

- http://www.scopus.com
- http://www.informatik.uni-trier.de/~ley/db/
- http://ieeexplore.ieee.org
- http://portal.acm.org/
- http://www.core.edu.au/
- http://academic.research.microsoft.com/

JOURNALS/
Conferences

- http://pdos.csail.mit.edu/scigen/

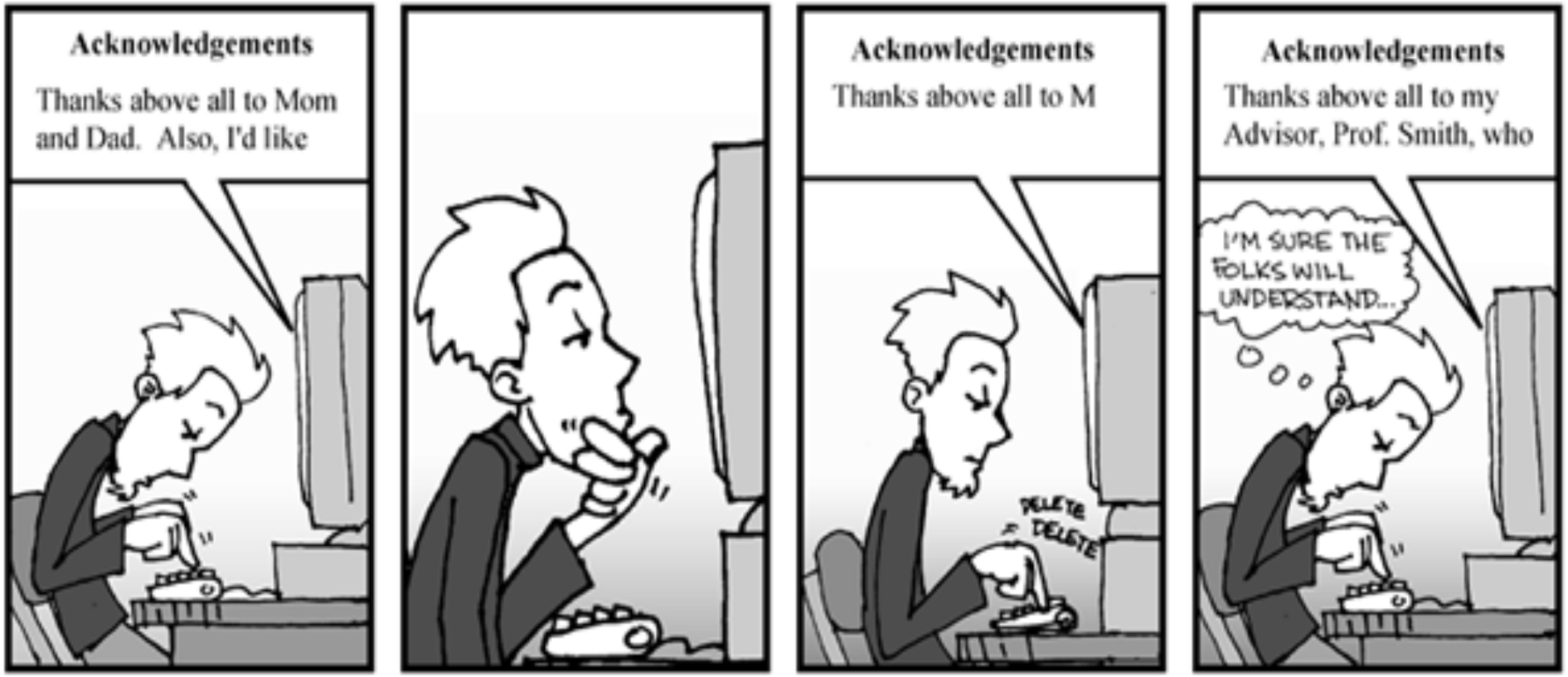
Random Paper Generator

Further reading:

Open Access Science Journals Accept Fake Papers:

http://www.thecrimson.com/article/2013/10/16/study-science-journals-fake-research/

“More than half of over three hundred fee-based, open access science journals accepted a bogus research paper for publication in a study conducted by John N. Bohannon, a visiting scholar at the Harvard Program in Ethics and Health.”



phd.stanford.edu/comics

#3: Be Aware of Ethics

Before starting your research, take care of ethical issues:

- If I publish anything, **who are the paper's authors**? Establish the publication authorship rules!
- What will appear in the "**Acknowledgments Section**"? (e.g. FCT project or PhD grant);
- Am I allowed to use this **research data/methods/software**? In which conditions?
- **Avoid plagiarism and do not fabricate data/results!**

More information at: **AIS code of research conduct**

http://start.aisnet.org/?CodeofResearch

Further reading:

How Science Goes Wrong:

http://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong

“Modern scientists are doing too much trusting and not enough verifying”

#4: Follow the Publication Submission Requirements

- Often, international conferences/journals offer **paper templates** (latex, word, ...);
- **Read carefully the paper preparation rules** (maximum number of pages, 1 or 2 columns, ...) **before starting to write**!

Never use your text editor as an excuse!

#4: Follow the Publication Submission Requirements

```
increase generations counter
establish the temporary population
as new current population
END WHILE
```

The pseudocode of a steady-state Genetic Algorithm would be similar to the above, except that the temporary population would be absent and we would have to use substitution algorithms.

One generation is created from a previous generation by means of two types of reproduction operators: cross-over and copy. Cross-over is a sexual reproduction that originates new descendants by exchanging the genetic information of the parents; copy consists in passing a certain number of individuals to the next generation without any variation. Once the new individuals are generated, mutation takes place with a *Pm* probability, and the errors of the genetic copy process are imitated.

The process finishes when there are sufficiently

the inconveniences of the repeated application of a Genetic Algorithm.

A pesar de que los Algoritmos Genéticos ofrecen soluciones óptimas a multitud de problemas, presentan limitaciones a la hora de utilizarlos para encontrar varias soluciones en escenarios que presentan múltiples puntos óptimos. Para tratar de paliar esta limitación, se buscaron otras soluciones. Entre estas soluciones destaca la agrupación en especies de la población de individuos.

Sin embargo, se puede comprobar en implementaciones de agrupación en especies llevadas a cabo que surgen algunas limitaciones en el método, debido a que tanto el número de especies como de los propios individuos tiende a seguir creciendo indefinidamente a lo largo de las distintas evoluciones [Seoa-06]. Tal aumento de individuos y especies provoca que los recursos computacionales vayan continuamente en aumento.

Spanish ???

#5: Start by the Paper Outline

•“Preparing an **outline** is the most important step in the process of producing a manuscript for publication in a journal.” – **SF Edit** (http://www.sfedit.net)

•A publication is made of: chapters, sections, subsections, etc...

Start first with your outline (or index) and then start writing....

#5: Start by the Paper Outline

Common structure examples:

Article: 1 Introduction, 2 Materials and Methods, 3 Results, 4 Conclusions (IMRC)

PhD/MSc Thesis:

- **1 – Introduction**
- **1.1 – Motivation**
- **1.2 – Objectives**
- **1.3 - Organization**
- **2 – State of the Art Chapter(s)**
- **…**
- **4 Conclusions**
- **4.1 Summary**
- **4.2 Discussion**
- **4.3 Future Work**
- **Appendix A …**

#6: Write about The State of The Art

- If you have not did before, research know about the state of the art, where you contextualize research in terms of:
  - What is similar, different and new?
  - How original is your work?
- http://www.misq.org/archivist/vol/no26/issue2/GuestEd.pdf
- The **Web** is a very good place to search for research:
  - **Wikipedia**: http://en.wikipedia.org/wiki/
  - **Google scholar**: **http://scholar.google.pt/**
  - **Semantic scholar:** https://www.semanticscholar.org/
  - **Video of how to use these tools: http://youtu.be/MrV59hMEy5o**

Be aware not to "reinvent the wheel"!

Citations vs. References

An article X can be **cited** by several articles.

An article X contains several **bibliographic references** (the works that are cited by X).


MORE AND MORE CURIOUS!
OOF!
WHAT ARE ALL THESE REFERENCES..?
READ ME
UH OH.
JORGE CHAM © 2010
WWW.PHDCOMICS.COM

#7: Know how to reference

- Put **integral text** (large portions should be avoided) within quotes: “Do not take work from another and pass it off as your own, i.e., plagiarize in any manner” **Code of Research conduct.**
- When you cite ideas/algorithms/opinions/results of other authors but in our own writing: use a **bibliographic reference** (often, at the end of a sentence).
- Also use references for: Algorithms, Figures, Tables (use “**in Figure 5**” and not “**in figure below**”);
- **There are several bibliography styles**: APA style, MLA style, Harvard, Numeric: http://www.cs.stir.ac.uk/~kjt/software/latex/showbst.html
- Use and abuse of **reference management software: bibtex**, Zotero, Mendeley, Endnotes, …

#7: Know how to reference

success [28]. Data mining (DM) techniques [33] aim at extracting high-level knowledge from raw data. There are several DM algorithms, each one with its own advantages. When modeling continuous data, the linear/multiple regression (MR) is the classic approach. The backpropagation algorithm was first introduced in 1974 [32] and later popularized in 1986 [23]. Since then, neural networks (NNs) have become increasingly used. More recently, support vector machines (SVMs) have also been proposed [4,26]. Due to their higher flexibility and nonlinear learning capabilities, both NNs and SVMs are gaining an attention within the DM field, often attaining high predictive performances [16,17]. SVMs present theoretical advantages over NNs, such as the absence of local minima in the learning phase. In effect, the SVM was recently considered one of the most influential DM algorithms [34]. While the MR model is easier to interpret, it is still possible to extract knowledge from NNs and SVMs, given in terms of input variable importance [18,7].

#7: Know how to reference

$$\hat{p}_j = \frac{\exp(y_j)}{\sum_{k=1}^{C}\exp(y_k)} \quad \text{(softmax function)}$$

$$y_i = w_{i,0} + \sum_{m=I+1}^{I+H} f\left(\sum_{n=1}^{I} x_n w_{m,n} + w_{m,0}\right) w_{i,n} \qquad (2)$$

where y_i is the output of the network for the node i; $f = 1/(1+\exp(-x))$ is the logistic function; I represents the number of input neurons; $w_{d,s}$ the weight of the connection between nodes s and d; and $w_{d,0}$ is a constant called bias. The first equation, known as the *softmax* function, warranties that $\hat{p}_j \in [0,1]$ and $\sum_{j=1}^{C} \hat{p}_j = 1$. The simplest ANN (with $H = 0$) is equivalent to the MLR model and more complex discrimination functions can be learned with a higher number of hidden neurons (Fig. 3). Yet, a high value of H will induce generalization loss (i.e. overfitting).

The logistic model is easier to interpret than ANNs. Nevertheless, it is possible to gather knowledge about what the ANN has learned by measuring the relative importance of the inputs (Section 2.3) and extracting rules. The latter issue is still an active research domain [16]. In this work, the pedagogical technique presented in [9] will be adopted, where the direct relationships between the inputs and outputs of the ANN are extracted by using a decision tree [17].

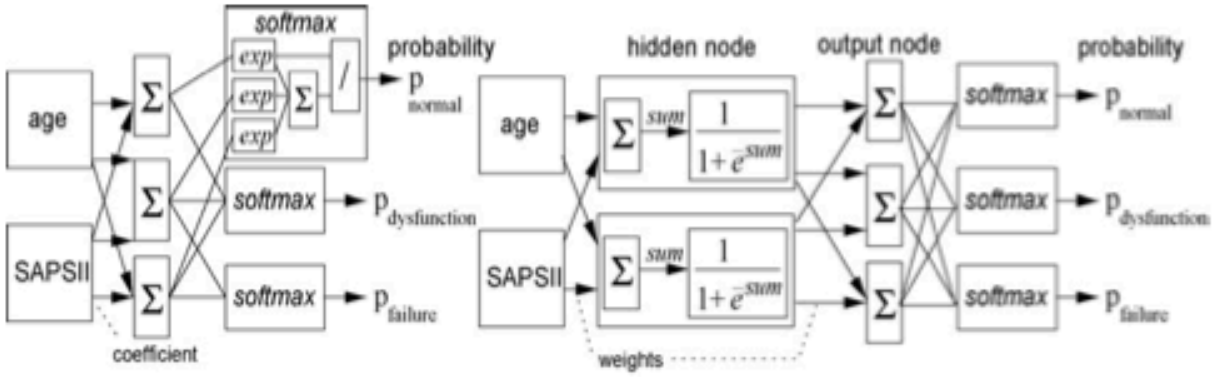


Figure 3 Example of a multinomial logistic regression (left) and artificial neural network with two hidden nodes (right).

#8: Build a proper Reference Bibliography Section

References

[1] Rosenberg A. Recent innovations in intensive care unit risk-prediction models. Curr Opin Crit Care 2002;8:321—30.

[2] Knaus W, Wagner D, Draper E, Zimmerman J, Bergner M, Bastos P, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest 1991;100:1619—36.

[3] Le Gall J, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 1993;270:2957—63.

References

Agapie, A. and A. Agapie. (1997). "Forecasting the Economic Cycles Based on an Extension of the Holt-Winters Model. A Genetic Algorithms Approach." In *Proc. of the IEEE Int. Conf. On Computational Intelligence for Financial Forecasting Engineering (CIFEr'97)*, New York, pp. 96–99.

Bäck, T. (1996). *Evolutionary Algorithms in Theory and Practice*. Oxford University Press.

Box, G. and G. Jenkins. (1976). *Time Series Analysis: Forecasting and Control*. Holden Day, San Francisco, USA.

#9: Know How to Build Tables

Table 2
The wine modeling results (test set errors and selected models; best values in bold).

	Red wine		
	MR	NN	SVM
MAD	0.50 ± 0.00	0.51 ± 0.00	**0.46** $\pm 0.00^{a}$
$\text{Accuracy}_{T=0.25}$ (%)	31.2 ± 0.2	31.1 ± 0.7	**43.2** $\pm 0.6^{a}$
$\text{Accuracy}_{T=0.50}$ (%)	59.1 ± 0.1	59.1 ± 0.3	**62.4** $\pm 0.4^{a}$
$\text{Accuracy}_{T=1.00}$ (%)	88.6 ± 0.1	88.8 ± 0.2	**89.0** $\pm 0.2^{b}$
$\text{Kappa}_{T=0.5}$ (%)	32.2 ± 0.3	32.5 ± 0.6	**38.7** $\pm 0.7^{a}$
Inputs ($\bar{I}$)	9.2	9.3	9.8
Model	–	$\bar{H}=1$	$\bar{\gamma}=2^{0.19}$
Time (s)	**518**	847	5589

[a] Statistically significant under a pairwise comparison with MR and NN.
[b] Statistically significant under a pairwise comparison with MR.

Q2: What rules were followed when building this Table?

(In-Class Teams, 1min.)

-Join in groups of 3 elements

-Who was born far away from this room → Team Leader (pen, ...)

#9: Know How to Build Tables

Table 4 The discrimination power (mean AUC value of the 20 runs, in %) for each organ, condition and method (values of AUC $>70\%$ are in bold)

	Normal		Dysfunction		Failure		Global	
Organ	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN
Respiratory	67.2	69.5	59.2	61.0	65.6	68.9	63.6	66.0
Coagulation	63.6	65.5	60.1	62.0	**72.6**	**73.9**	63.3	65.1
Hepatic	64.7	66.7	62.5	64.2	**72.6**	**76.0**	64.6	66.6
Cardiovascular	67.9	**71.2**	63.8	65.6	67.3	**71.0**	67.1	**70.2**
Neurological	**70.0**	**72.1**	58.8	61.2	**74.7**	**76.7**	68.8	**70.9**
Renal	69.4	**70.7**	66.0	66.8	**73.5**	**76.1**	69.1	**70.4**
Average	67.1	69.3	61.7	63.5	**71.0**	**73.8**	66.1	68.2

#10: Know How to Build a Figure

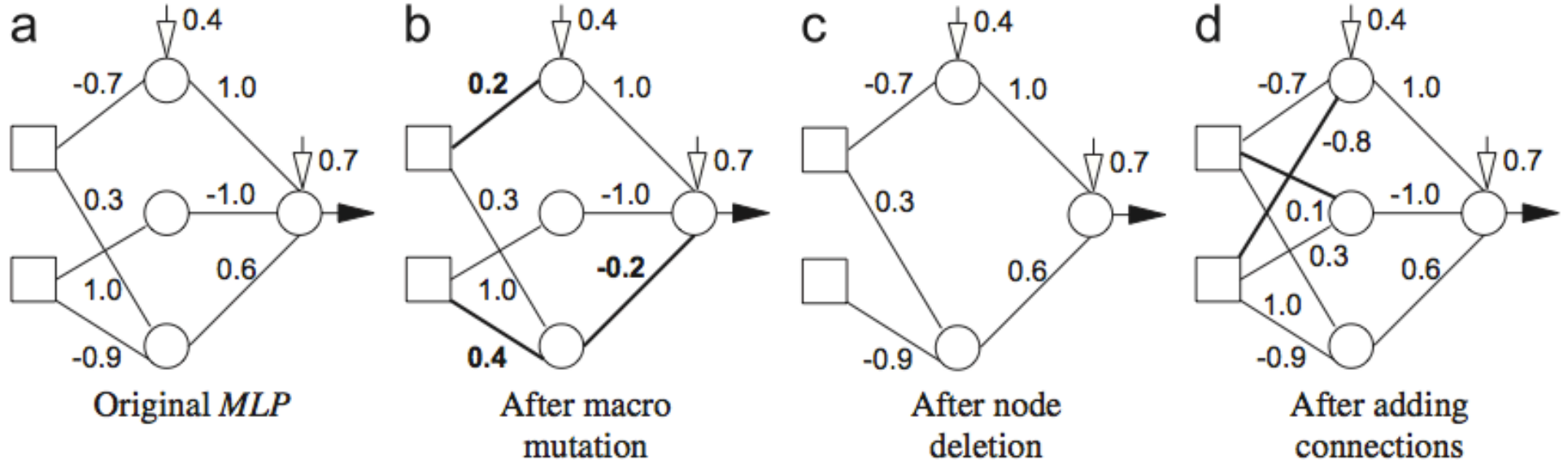
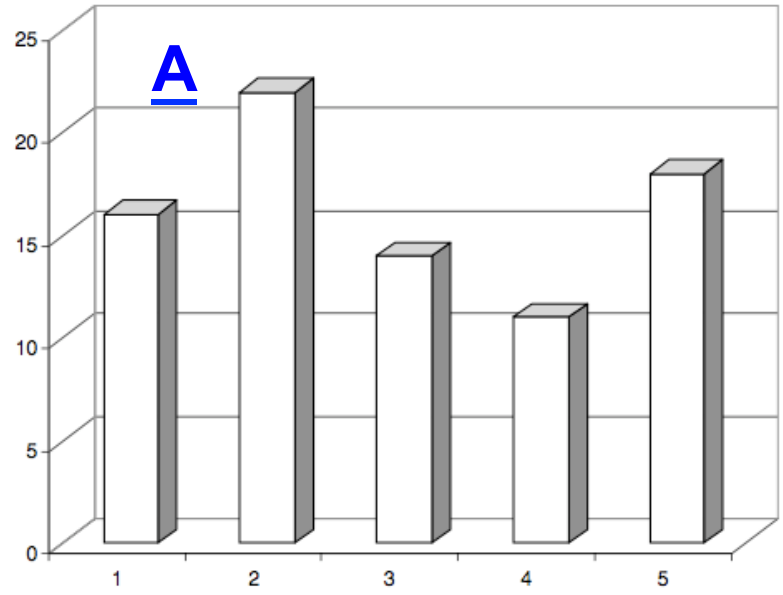
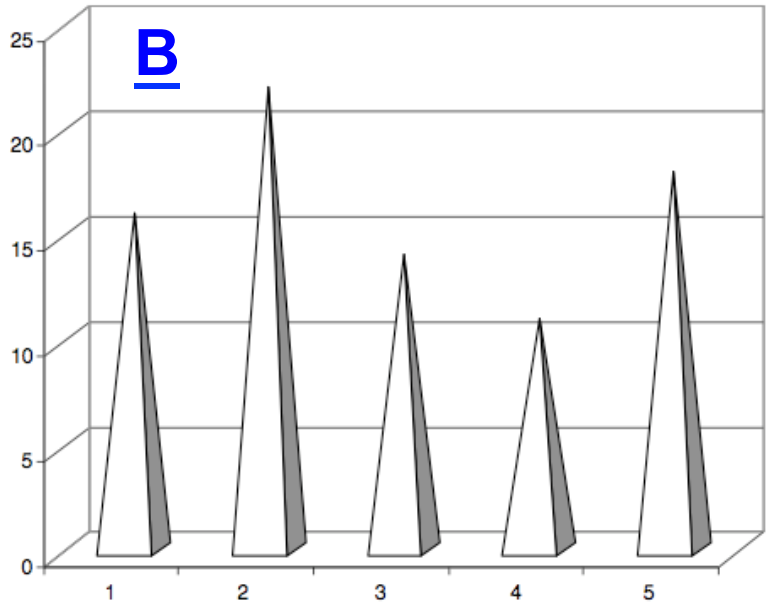
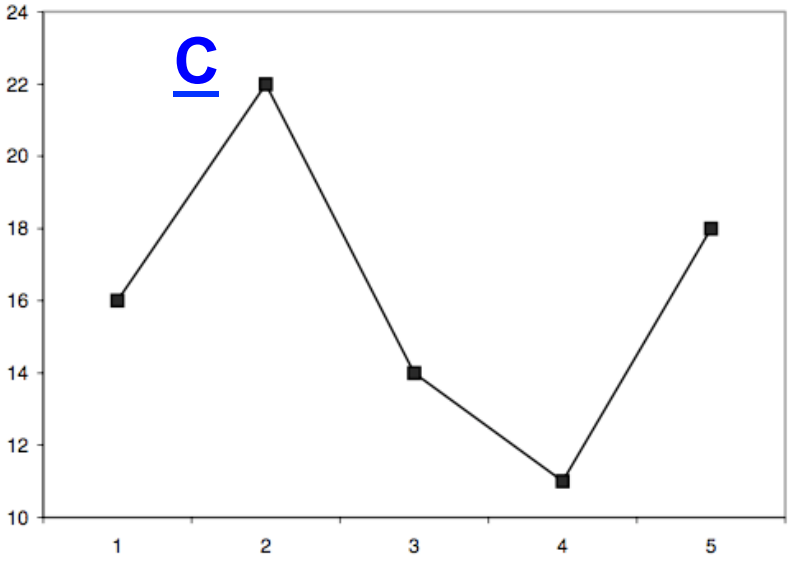


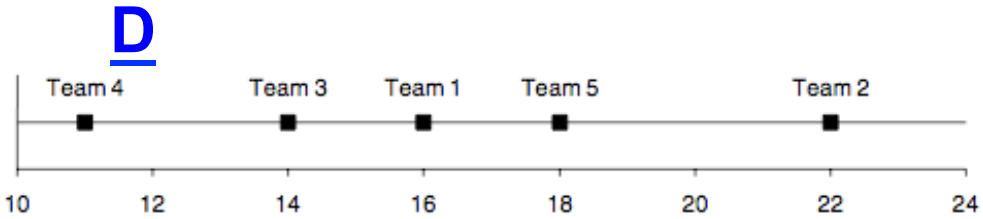
Fig. 2. Example of the application of the mutation operators.

#11: “KISS—Keep It Simple, but Scientific”







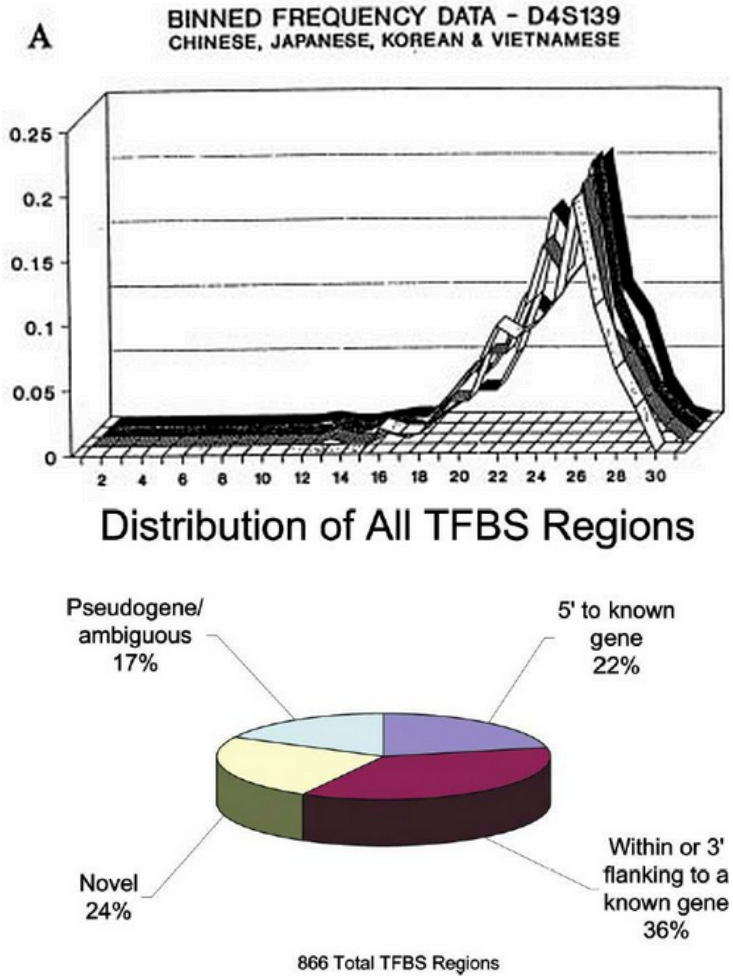


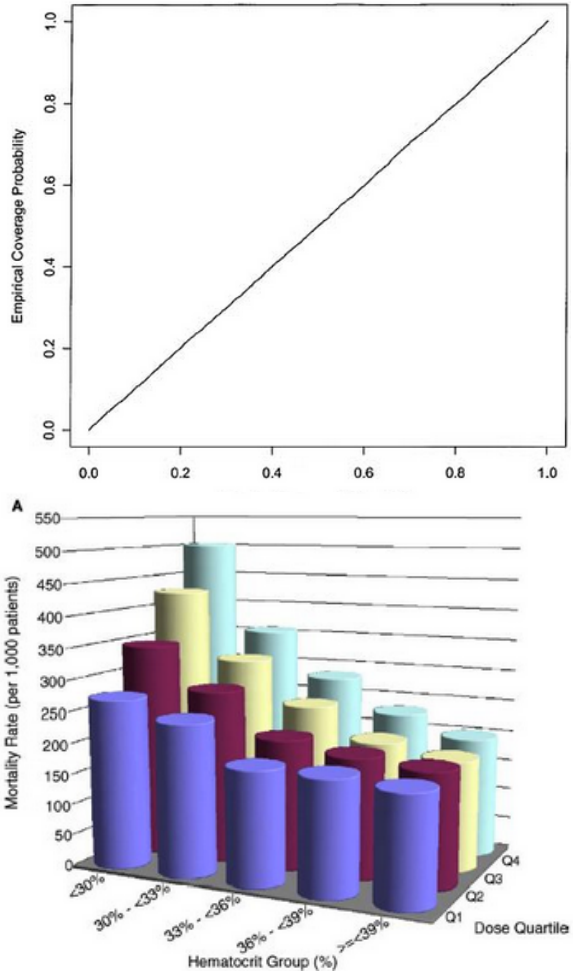
Q3: What is the best way to show the total number of scored goals by 5 soccer teams? (In-Class Teams, 1min.)

-Join in groups of 3. Leader: oldest member.

#11: “KISS—Keep It Simple, but Scientific”

Top ten worst graphs:
http://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/





Further reading:

Creating More Effective Graphs:

http://assets.en.oreilly.com/1/event/55/Communicating%20Data%20Clearly%20Presentation.pdf

#11: “KISS—Keep It Simple, but Scientific”

Online newspapers are blooming because the online medium offers advantages that are simply not available in the printed medium. Online news can be delivered in real time. Online news can be paid individually with micro payments. Online news can be personalized.

vs

Online newspapers are increasing and present advantages when compared with printed news: real-time delivery, individual (micro) payments and personalization.

#12: Use numbered equations

labels. The global AUC can then be computed by summing the AUCs weighted by the prevalence of c_i in the data, using [22]:

$$\mathrm{AUC}_{\mathrm{Global}} = \sum_{c_i \in C} \mathrm{AUC}(c_i) \cdot \mathrm{prev}(c_i) \qquad (4)$$
$$\mathrm{prev}(c_i) = c_i / N$$

where $\mathrm{AUC}(c_i)$ denotes the AUC for class reference c_i, c_i the number of patients with condition c_i and N is the total number of patients.

As an example, the next two rules for renal failure prediction can be extracted from the tree:

$$\begin{aligned} &\text{IF TCRUR} \geq 13.8 \text{ AND NUR} \geq 15 \text{ THEN } \textit{failure} \\ &\text{IF TCRUR} < 13.8 \text{ AND admfrom} \notin \{5, 6\} \\ &\quad \text{AND NCRHR} = 0 \text{ AND SAPSII} \geq 93 \text{ THEN } \textit{failure} \end{aligned} \qquad (8)$$

#13: Know How to Build Algorithms

Algorithm 5 *Steepest Ascent Hill-Climbing*

```
1: n ← number of tweaks desired to sample the gradient
2: S ← some initial candidate solution
3: repeat
4:     R ← Tweak(Copy(S))
5:     for n − 1 times do
6:         W ← Tweak(Copy(S))
7:         if Quality(W) > Quality(R) then
8:             R ← W
9:     if Quality(R) > Quality(S) then
10:        S ← R
11: until S is the ideal solution or we have run out of time
12: return S
```

Algorithm 4. Threshold averaging of ROC curves

Inputs: *samples*, the number of threshold samples; *nrocs*, the number of ROC curves to be sampled; *ROCS*[*nrocs*], an array of *nrocs* ROC curves sorted by score; *npts*[*m*], the number of points in ROC curve *m*. Each ROC point is a structure of three members, *fpr*, *tpr* and score.

Output: *Avg*[*samples* + 1], an array of (*X*, *Y*) points constituting the average ROC curve.

Require: *samples* > 1

```
1: initialize array T to contain all scores of all ROC points
2: sort T in descending order
3: s ← 1
4: for tidx = 1 to length(T) by int(length(T)/samples) do
5:     fprsum ← 0
6:     tprsum ← 0
7:     for i = 1 to nrocs do
8:         p ← ROC_POINT_AT_THRESHOLD(ROCS[i], npts[i], T[tidx])
9:         fprsum ← fprsum + p.fpr
10:        tprsum ← tprsum + p.tpr
11:    end for
12:    Avg[s] ← (fprsum/nrocs, tprsum/nrocs)
13:    s ← s + 1
14: end for
15: end
1: function ROC_POINT_AT_THRESHOLD(ROC, npts, thresh)
2: i ← 1
3: while i ⩽ npts and ROC[i]. score > thresh do
4:     i ← i + 1
5: end while
6: return ROC[i]
7: end function
```

#14: Use Acronyms Properly

dated on a daily basis. The most used scores include [5]: the sequential organ failure assessment (SOFA), multiple organs dysfunction score (MODS) and logistic organ dysfunction (LOD). Our focus is on the SOFA

In this work, the first ENN to be presented is the *Topology-optimization Evolutionary Neural Network* (TENN), where the aim is to evolve the optimal structure of a MLP to a given problem. Under this approach, a

Extra Tip: Give a (new) name to your system/algorithm/method, so other researchers can mention it!
Example: Variable Effect Characteristic (VEC) curve

#15: Dimension well your text size

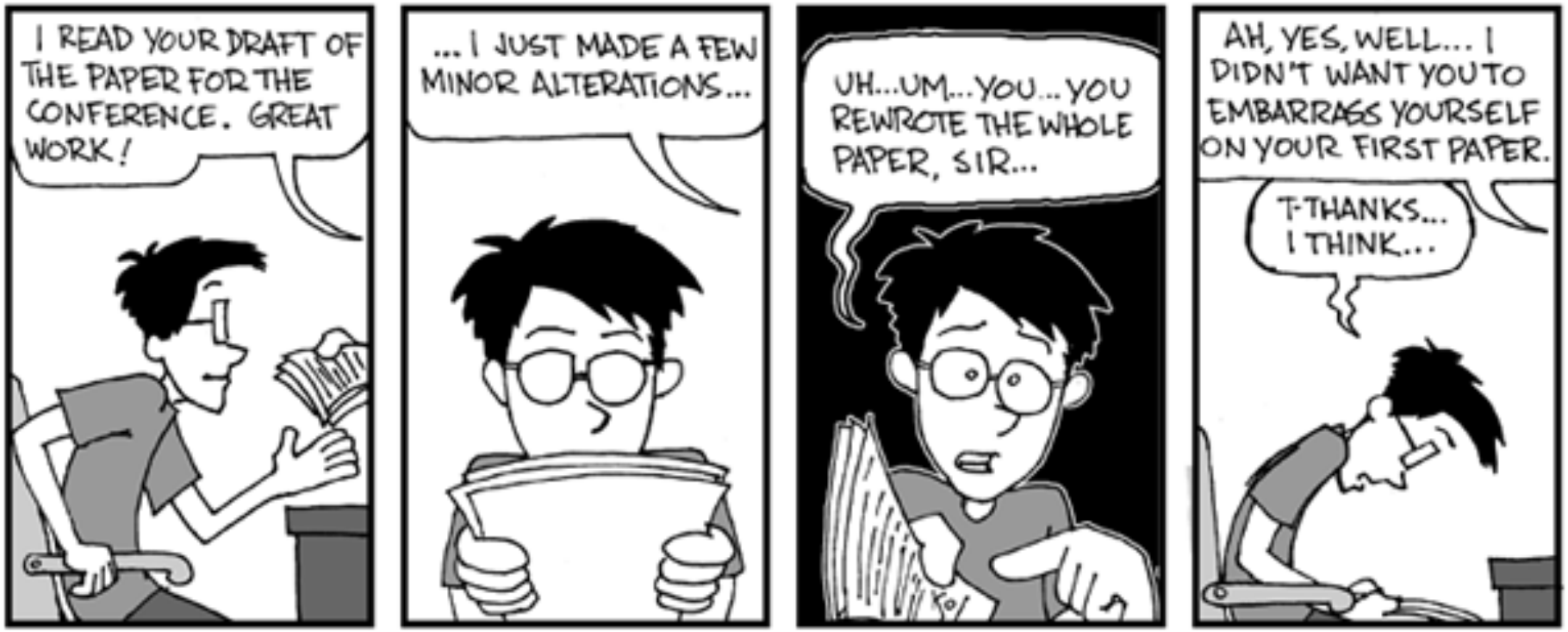
- Chapters and sections should have a reasonable size. If possible, all chapters/sections should have around the same size (exceptions: Introduction and Conclusions).
- Never use one isolated subsection (or one bullet item, etc.);
- Use paragraphs with a reasonable size (from 3 to 6 sentences). Very short paragraphs should be exceptions.
- Do not use too long or too short sentences.


CLICK
YES!!
FINISHED ANOTHER PARAGRAPH, DID WE?
I AM A WRITING GOD!
JORGE CHAM © 2003
www.phdcomics.com

#16: Use a Writing Method

- Get first all necessary documentation (data, code, tables, ...);
- Write fast (the intention is to put all important ideas);
- Write without editing, to let the writing flow;
- Follow your outline but write by sections/parts;
- Keep your first draft away (at least during 1 day);
- Then, review your draft (review, review and review)!
- Some errors are only detected in a printed version.
- Be coherent! (use of capital letters, notation, letter fonts, use of *italic* and **bold**, etc.)

(based on **SFEdit** tips)



#17: Avoid Writing Errors

- At the very least, use a spell checker;
- If writing in Portuguese, be aware of **accents**!

Q4: Detect what are the errors in the text below?
(In-Class Teams, 1 min.)

-Groups of 3 elements, Leader= youngest member

1. There are 3 main solutions: SCIS, RCIS and JCIS. The later is the best system in terms of security.
2. We adopted a integrated approach, that delivers an high performance.

See Also:
http://www.wsu.edu/~brians/errors/errors.html
http://www.serendipity.li/errors.html
http://www.googlebattle.com/

#18: Be persistent!!!

-“**Resiliency** is the key to publication success”
-“Those who publish a lot – submit a lot”
Richard T. Watson

- Probability for publishing an article:
  - Without submission = 0
  - **With Submission >= 0** ☺

Further reading:

P. Cortez. **Some Scholarly Communication Guidelines**. Teaching Report, Department of Information Systems, Engineering School, University of Minho, Guimarães, Portugal, January 2011.
http://hdl.handle.net/1822/11599

Not so serious reading:

http://www.phdcomics.com/
http://phdmovie.com/