

Métodos Quantitativos para Ciência da Computação Experimental

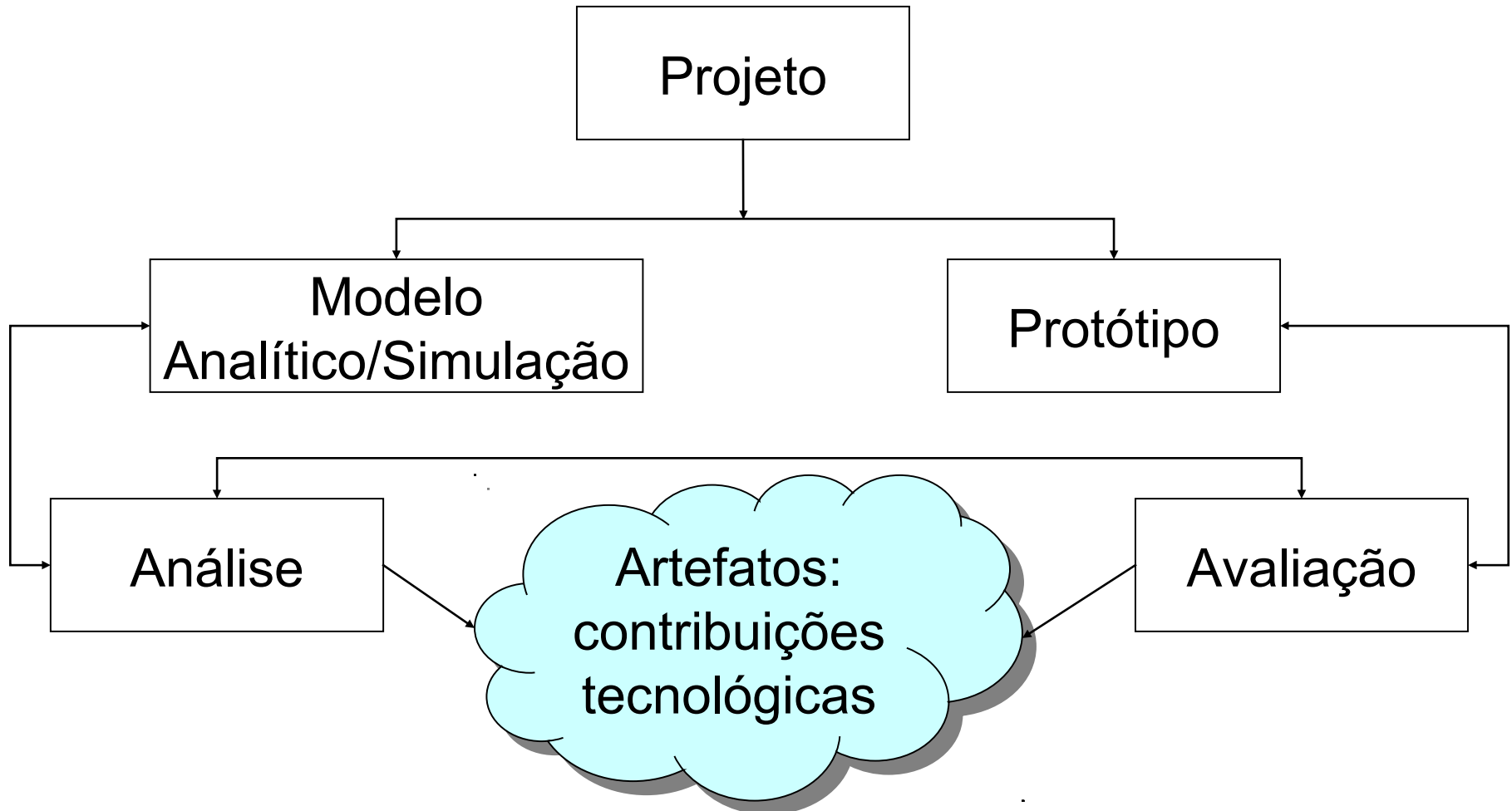
Programa

- Ciência da Computação Experimental e o Método Científico
- Monitoração
- Conceitos de Probabilidade e Sumarização de Dados Medidos
- Comparando Sistemas Usando Dados de Amostragem e Intervalos de Confiança
- Projeto Experimental (Design of Experiments)
- Modelos de Regressão Linear Simples e outros modelos
- Apresentação Gráfica de Dados Experimentais

O que significa ciência experimental?

- Baseado em observações, dados, experimentos
- *Trabalho experimental* deve complementar a pesquisa teórica
 - Teorias podem ter incertezas.
 - Teorias podem resultar de observações.
 - Teorias podem ser testadas por observações.
 - De forma oposta, teorias podem guiar a pesquisa experimental.
 - Nem toda pesquisa em ciência da computação pode ser resolvida teoricamente (ex.: IHC, uso malicioso)
- Em resumo, podemos dizer que a pesquisa experimental “quer entender o comportamento de sistemas complexos em computação.”

Ciência da Computação Experimental



Experimentação em Sistemas Computacionais

Por quê?:

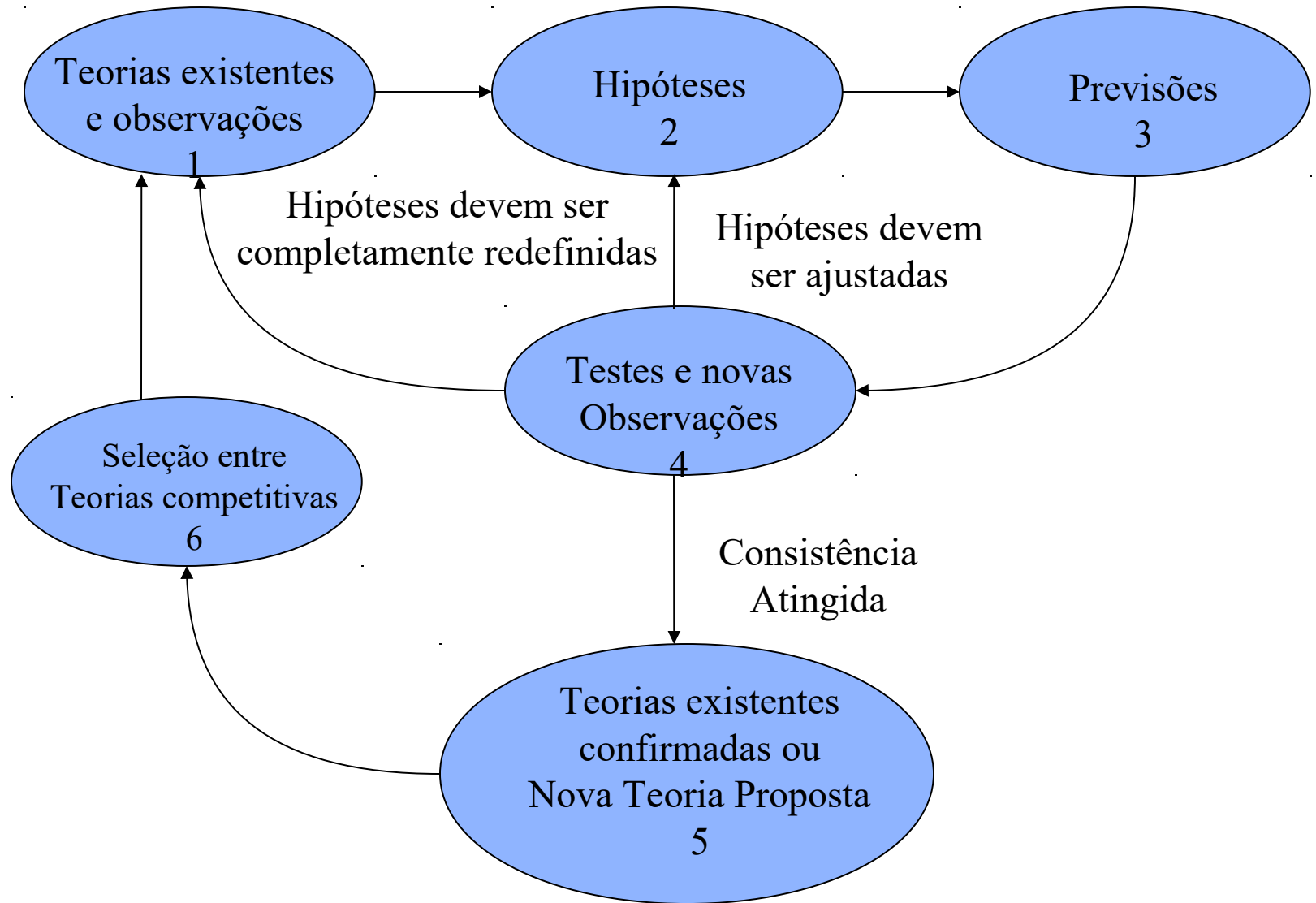
“It doesn’t matter how beautiful your theory is, it doesn’t matter how smart you are – if it doesn’t agree with the experiment, it’s wrong.”

Richard Feynman, físico
Prêmio Nobel 1965

O Processo Experimental em Ciência da Computação e o Método Científico

Alguns tipos de “papers” em Ciência da Computação

- Três tipos de artigos que descrevem a implementação de um algoritmo:
 - Artigo de aplicação
“Aqui está um bom algoritmo para a solução X”
 - Artigo de “marketing”
“Aqui está um algoritmo novo e interessante”
 - Artigo experimental
“Aqui está como o algoritmo comporta-se com dados reais”
- O método experimental deve se aplicar aos três tipos de artigos.



O Método Científico usado por cientistas para buscar respostas para as questões das ciências.

Método Científico: história

- Sir Francis Bacon (1561-1626) –
 - Paradigma do método científico:
 - “limpe a mente” de pre-conceitos
 - Colete fatos (não ao acaso....)
 - As conclusões iniciais podem estar incorretas.
 - “Verdade” vem de erros e não de confusões
 - “Obscuridade” de algumas partes faltantes pode ocorrer, mas não a obscuridade da distorção.

"There are and can be only two ways of searching into and discovering truth. The one flies from the senses and particulars to the most general axioms, and from these principles, the truth of which it takes for settled and immovable, proceeds to judgement and to the discovery of middle axioms. And this way is now in fashion.

The other derives axioms from the senses and particulars, rising by a gradual and unbroken ascent, so that it arrives at the most general axioms last of all. This is the true way, but as yet untried."

Método Científico: história

- Rene Descartes (1595-1650) –
 - experimentação acoplada a análise matemática
 - Experimentação permite a análise de detalhes onde várias alternativas são possíveis.

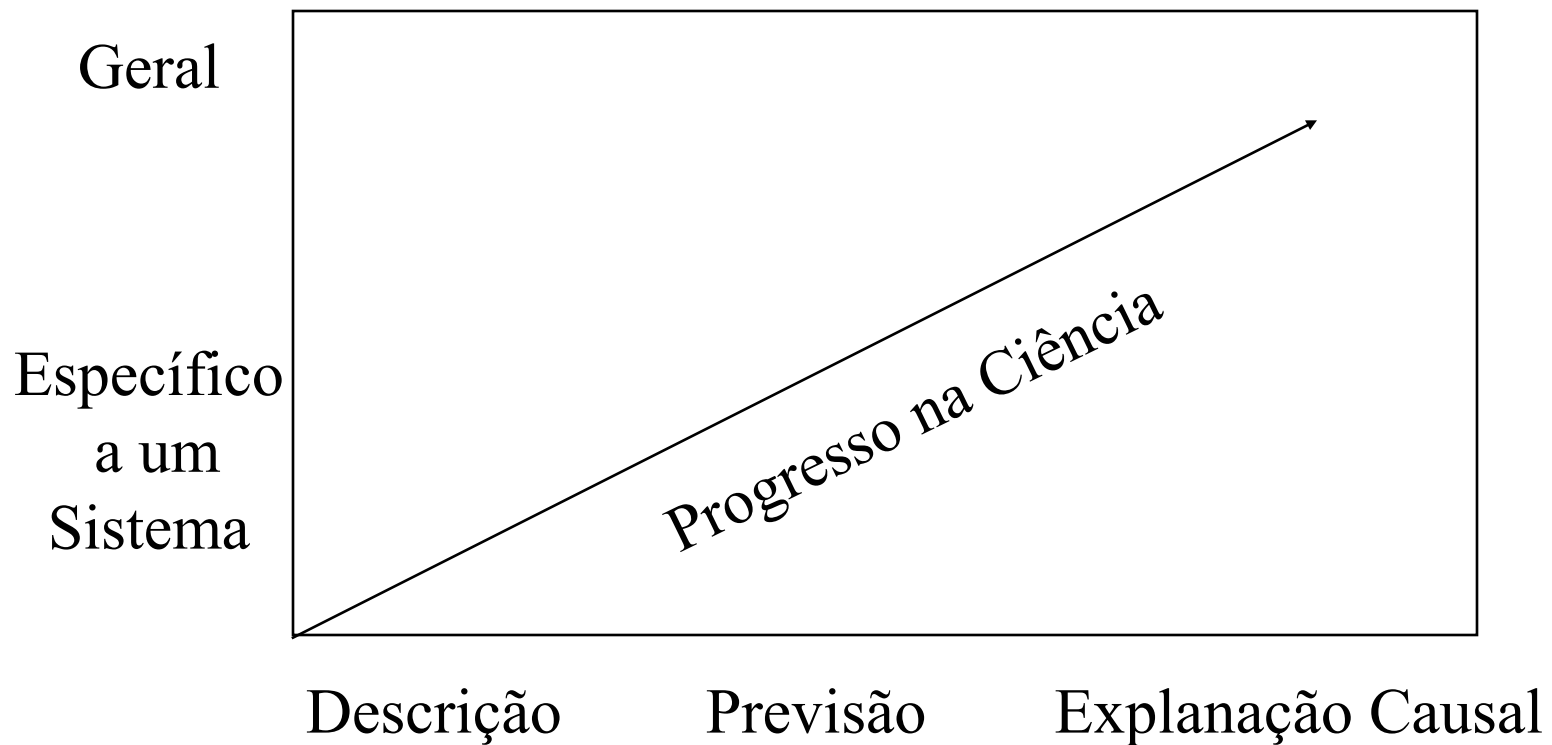
"The first was never to accept anything for true which I did not clearly know to be such; that is to say, carefully to avoid precipitancy and prejudice, and to comprise nothing more in my judgement than what was presented to my mind so clearly and distinctly as to exclude all ground of methodic doubt.

The second, to divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution.

The third, to conduct my thoughts in such order that, by commencing with objects the simplest and easiest to know, I might ascend by little and little, and, as it were, step by step, to the knowledge of the more complex; assigning in thought a certain order even to those objects which in their own nature do not stand in a relation of antecedence and sequence.

And the last, in every case to make enumerations so complete, and reviews so general, that I might be assured that nothing was omitted."

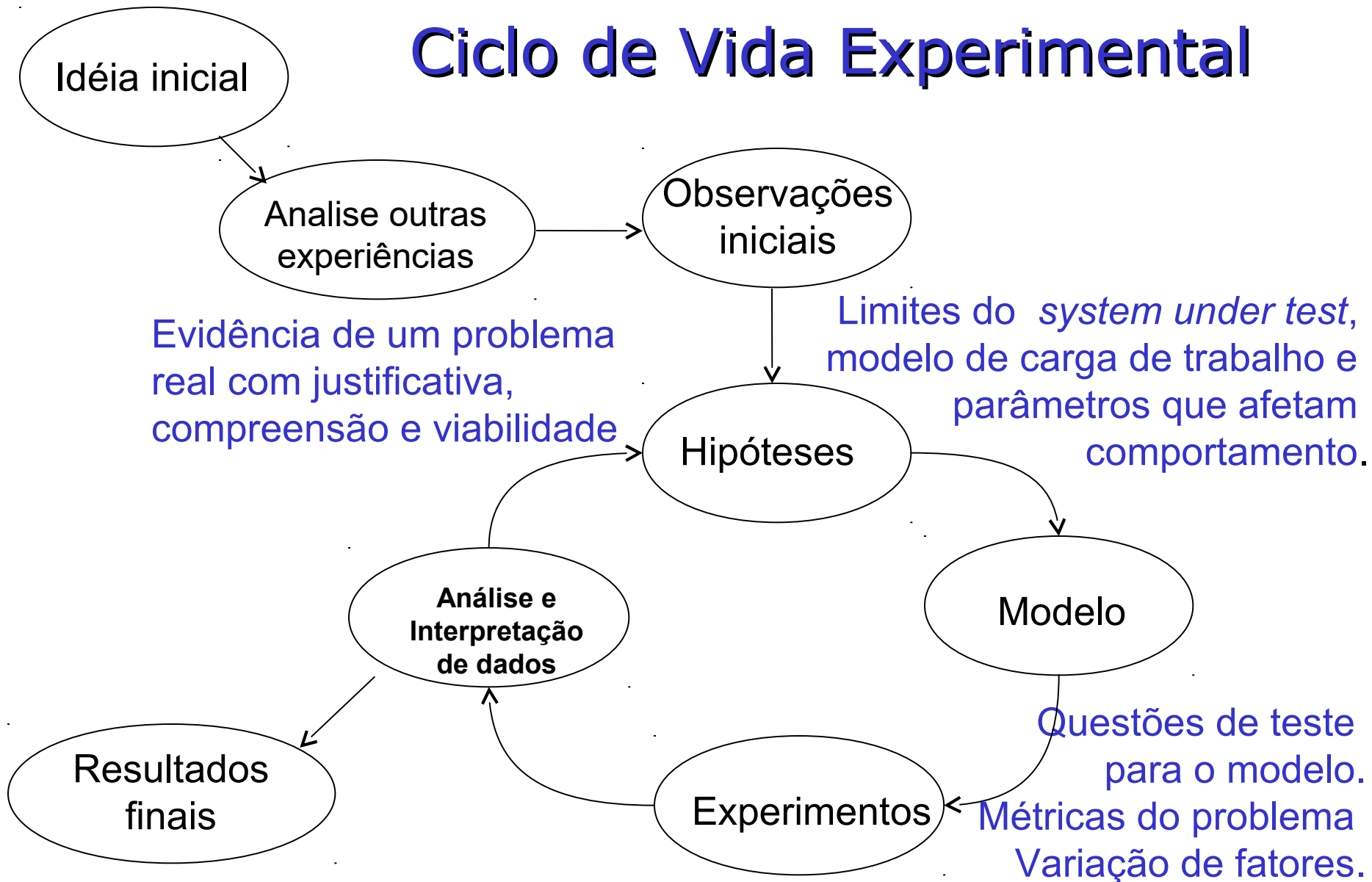
Espaço das Questões da Pesquisa Básica



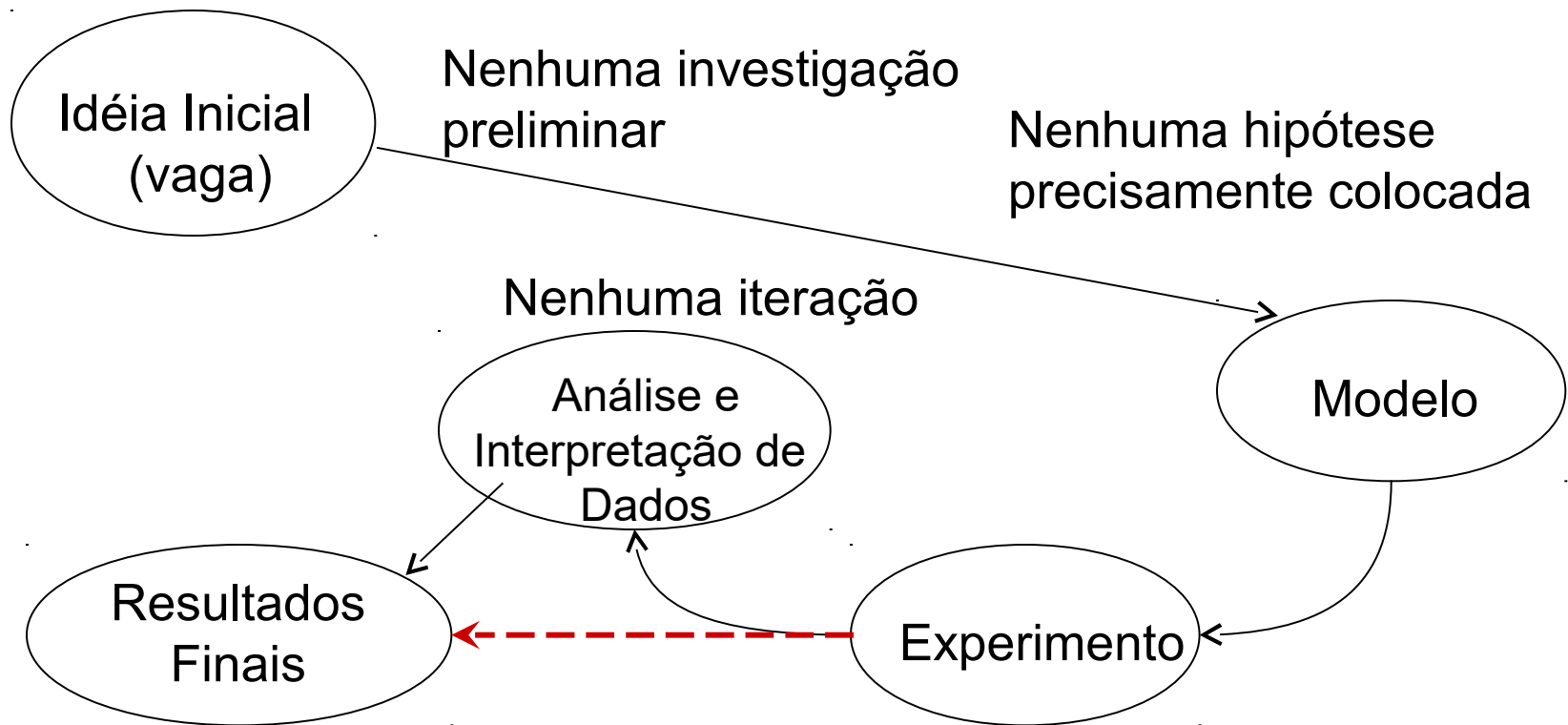
Por que Ciência da Computação Experimental?

- Experimentação: Modelo-teste da teoria
 - Feynman: um experimento pode sempre identificar falhas na teoria (ex.: hipótese/suposições violadas pela realidade)
- Artigos para discutir:
 - D. Feitelson in “Experimental Computer Science: The Need for a Cultural Change”,
<http://www.cs.huji.ac.il/~feit/pub.html>
 - T. Mytkowicz et al., “Producing wrong data without doing anything obviously wrong”, Proc. ASPLOS 2010
 - W. Tichy in “Should Computer Scientists Experiment More?”, Computer 31(5), 1998
 - Vide referências no site: <http://www.cs.huji.il/~feit/exp/related.html>
e na página do curso <http://www.dcc.ufmg.br/~jussara/metq>

Ciclo de Vida Experimental



Prática Usual em Ciência da Computação



Várias maneiras questionáveis de atacar o processo experimental

- Específico, que se aplica somente ao “system under test”.
- Testes que não enxergam o futuro – motivações e observações baseadas apenas no passado.
- Ausência de cargas de trabalho representativas, falta de benchmarks confiáveis, etc.
- Ausência de uma cultura de replicação independente de outros experimentos.
- Dados reais, que são confusos e não confiáveis.

Aprenda o método correto

Processo Experimental Sistemático

1. Entenda o problema, estabeleça as perguntas e defina os objetivos
2. Selecione métricas
3. Identifique os parâmetros
4. Decida quais parâmetros serão estudados, i.e., serão variados (fatores)
5. Selecione a técnica
6. Selecione a carga de trabalho (workload)
7. Execute experimentos
8. Analise e interprete os resultados
9. Apresente os resultados e dados do experimento
10. Apresente conclusões

Processo Experimental Sistemático

1. Entenda o problema, estabeleça as perguntas e defina os objetivos: “*A problem well-stated is half-solved*”.
 - Deve-se ser objetivo
 - Seja capaz de responder “por que”, e também “como”
 - Delimite o escopo
 - Defina as perguntas que pretende responder
2. Selecione métricas que ajudarão analisar as perguntas.

Processo Experimental Sistemático

3. Identifique os parâmetros que afetam o comportamento
 - Parâmetros do sistema (ex.: configuração de hardware)
 - Parâmetros da carga (workload, ex.: padrões de chegada de requisições)
3. Decida quais parâmetros serão estudados, i.e., serão variados (fatores)
 - Normalização
 - Comece com lista pequena

Processo Experimental Sistemático

5. Selecione a técnica:

- Medição de uma implementação de protótipo
 - Quanto invasivo? Podemos quantificar o “overhead” da monitoração? Podemos medir o que desejamos?
- Simulação – quanto detalhada ? Como será a validação?
- Repetibilidade

6. Selecione a carga de trabalho (workload)

- Representativa?
- É aceita pela comunidade científica?
- Disponibilidade de dados?

Processo Experimental Sistemático

7. Execute experimentos

- Quantos testes devem ser rodados?
Quantas combinações dos parâmetros que formam o ambiente experimental?
- Análise da sensibilidade dos outros parâmetros.

8. Analise e interprete os resultados

- Use Estatística para analisar a variabilidade, “outliers”, etc.

Processo Experimental Sistemático

9. Apresente adequadamente os resultados e dados do experimento

- Gráficos: a questão da visualização dos resultados, distribuições estatísticas, etc.

10. Apresente conclusões

- Para onde os resultados nos levam?
- Quais os próximos passos?
- Novas hipóteses, novas questões, outros experimentos.

Processo Experimental Sistemático

1. Entenda o problema, estabeleça as perguntas e defina os objetivos
2. Selecione métricas
3. Identifique os parâmetros
4. Decida quais parâmetros serão estudados, i.e., serão variados (fatores)
5. Selecione a técnica
6. Selecione a carga de trabalho (workload)
7. Execute experimentos
8. Analise e interprete os resultados
9. Apresente os resultados e dados do experimento
10. Apresente conclusões

Discussão

- Apresentação de cada aluno
- Quais são interesses de pesquisa?
 - Quais conferências representam os problemas de interesse de sua comunidade de pesquisa?
- Você tem algum projeto de pesquisa em andamento?
- Quais as razões que o levaram a fazer este curso? (ex: *Meu orientador obrigou-me!*)
- Você tem alguma “expertise” anterior no tema que pode ser útil para a turma?

Proposta de Projeto

Deve apresentar os 5 primeiros passos:

1. Entenda o problema, estabeleça as perguntas e defina os objetivos
2. Selecione métricas
3. Identifique os parâmetros
4. Decida quais parâmetros serão estudados, i.e., serão variados (fatores)
5. Selecione a técnica

Projeto deve aderir fortemente à proposta. Logo esses passos devem ser feitos com cuidado.

Should Computer Scientists Experiment More?

Outline

- Is Computer Science a Science?
- Why should we experiment?
 - Eight fallacies exposed
- Inherent problems with experimentation

Is Computer Science a Science?

- Tradition: *science deals with fundamental laws of nature*
- No, an engineering discipline (Fred Brooks)
 - Computers and programs are human creations
 - Computer science is *not* a natural science
- Yes, much more than synthetic results
 - Study of information structures & processes
 - Synthetic results (computers & programs) are models
 - Difference from traditional science: work with information – neither energy nor matter

Is Computer Science a Science?

(P. Denning, CACM 4/2005)

- Science or technology (man-made objects)?
 - Scientific paradigm (Francis Bacon): process of forming hypotheses and testing them through experiments
 - Science means explaining, modeling and predicting phenomena in the world.
 - Computer Science: science of information processes and their interactions with the world.
- Computer Science:
 - Art, science, fundamental principles that are non-obvious and has a future (relations with other areas)
 - But: credibility problem
 - In a sample of 400 papers before 1995, Tichy found that about 50% of those proposing models or hypothesis did not test them. In other fields of science the fraction of papers with untested hypothesis was about 10%.

Futuro: Transdisciplinariedade



Why should we experiment?

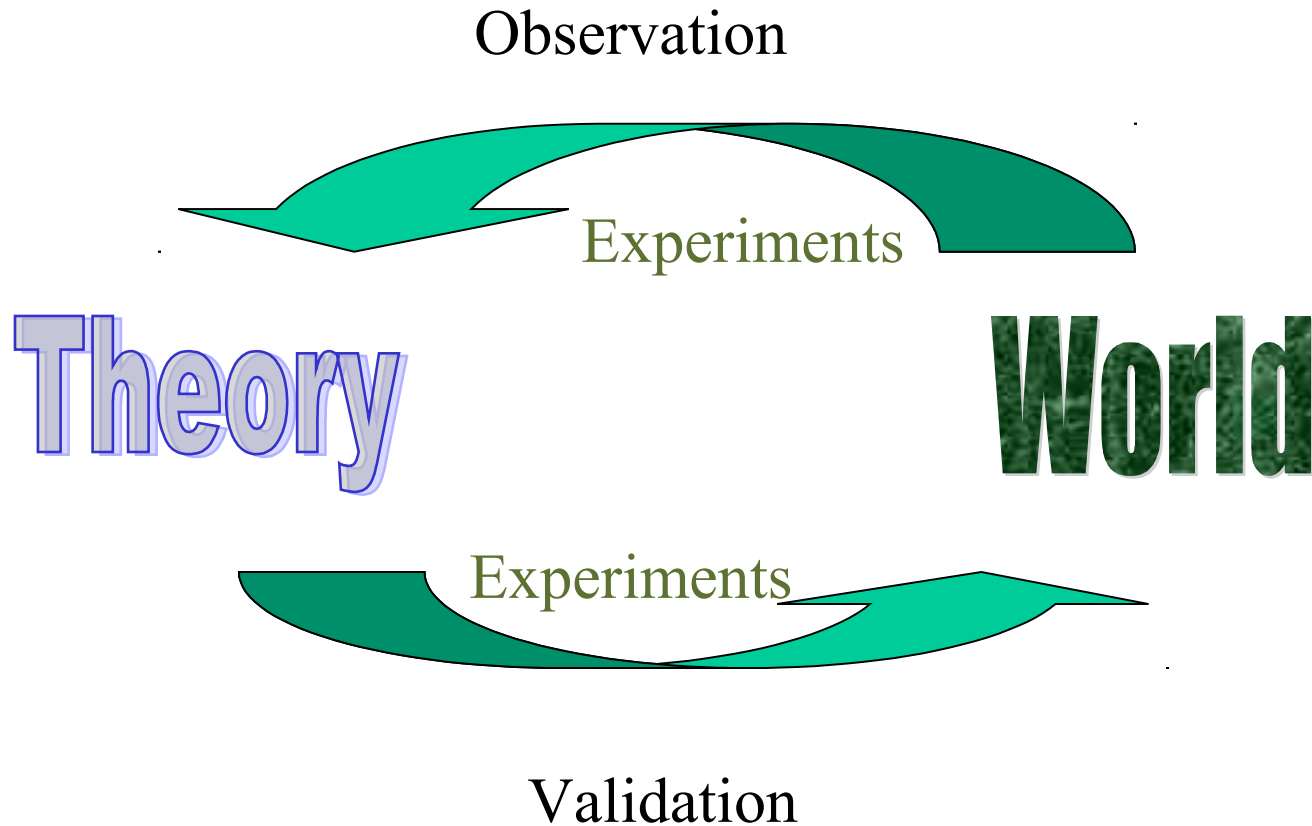
- Theory testing and exploration
 - Theory falsification: falsifiability (or refutability or testability) is the logical possibility that an assertion can be shown false by an observation or a physical experiment.
- Aid with induction or theory derivation
 - Exploration, deriving theories from observation

Eight Fallacies Exposed (#1)

Traditional scientific method isn't applicable

- *Rebuttal: To understand info processes, computer scientists must observe phenomena, formulate explanations, and test them. This is the scientific method.*
- Many CS theories have not been tested
 - OO programming improves programmer productivity, program quality.
- Repeatability is key requirement to any experiment


Eight Fallacies Exposed (#1)



Repeatability

ACM IMC 2016 Call for Papers

conferences.sigcomm.org/imc/2016/cfp.html



ACM IMC 2016
November 14-16
Santa Monica, California, USA

CFP in [PDF](#) and [plain text](#)

The Internet Measurement Conference (IMC) is a highly selective venue for the presentation of measurement-based research in data communications. The focus of IMC 2016 will be on research that either improves the practice of network measurement, or illuminates some facet of an operational network, or both. IMC takes a broad view of contributions that are considered in scope for improving the practice of network measurement, including, but not limited to:

- collection and analysis of data that yield new insight into network performance (e.g., traffic, topology, routing, energy consumption)
- collection and analysis of data that yield new insight into network economics, privacy, security, application interaction
- modeling of network internals and application behavior (e.g., performance bottlenecks)
- methods and tools to monitor and visualize network performance
- novel systems and algorithmic techniques that leverage network measurement
- advances in data collection, analysis, and storage (e.g., data management, data analysis)
- reappraisal of previous empirical findings

Operational networks of interest include:

- Short papers (up to 6 pages for text and figures + up to 1 page for references) that convey work that is less mature but shows promise, articulate a high-level vision, describe challenging future directions, critique current measurement wisdom, or offer results that do not merit a full submission.

Any submission exceeding short paper page-length limit will be evaluated as a full paper. Authors should only submit original work that has not been published before and is not under submission to any other venue. We will consider full paper submissions that extend previously published short, preliminary papers (including IMC short papers) following the model of the [ACM SIGCOMM policy](#).

IMC 2016 will bestow two awards. One award will recognize the outstanding paper at the conference, and all accepted papers are eligible for it. The other award will recognize a paper that contributes a novel dataset to the community. To be eligible for this award, the authors must make their dataset publicly available (e.g., through [DatCat](#) for Internet measurement data or [CRAWLAD](#) for wireless data) by the time of the camera-ready submission.

A few accepted papers may be forwarded for fast-track submission to [IEEE/ACM Transactions on Network Measurement](#).

Repeatability

Dear Jussara Almeida,

If you are receiving this email is because you don't have a conflict with the two finalists for the best dataset award of PAM 2009. I've asked all candidates to send me the link to their datasets; many responded that they can't publish their data, so I've removed them from the list. From the remaining five papers, I've selected the following two papers based on your previous votes and the reviews. Now, we need to pick the winner! :)

Please send me your vote by March 18th.

Repeatability

... I recommend the authors to make their code publicly available
(comment by reviewer of Elsevier Computer Communications)

Eight Fallacies Exposed (#2)

The current level of experimentation is good enough

- *Rebuttal: Relative to other sciences, the data shows that computer scientists validate a smaller percentage of their claims*
- 40-50% of software papers published in ACM journals were unvalidated (~ 15% in other areas)
 - Computer scientists publish a lot of untested ideas or that the ideas published are not worth testing
- Balancing theory and engineering with experiment
 - Build reliable base & reduce uncertainties
 - Lead to new areas of investigation
 - Accelerate progress by pruning fruitless approaches

Eight Fallacies Exposed (#3)

Experiments cost too much

- *Rebuttal: Meaningful experiments can fit into small budgets; expensive experiments can be worth more than their cost*
- Constrained by cost
 - Probe importance of research question
 - Plan appropriate research programs
 - Look for *affordable experimental techniques*
 - Intermediate steps with *partial results*
- Industry is beginning to value experiments
 - Three to five year lead over competition
- Experiments in other areas
 - Pharmaceuticals, aeronautics, biology

Eight Fallacies Exposed (#4)

Demonstrations will suffice

- *Rebuttal: Demos can provide incentives to study a question further. Too often, however, these demos merely illustrate a potential*
- Proof of concept
- No solid evidence
- Solid evidence requires careful analysis involving experiments, data and replication
 - Experiments require clear question, experimental apparatus to test the question, data collection, interpretation, sharing of results

Eight Fallacies Exposed (#5)

There's too much noise in the way

- *Rebuttal: Fortunately, benchmarking can be used to simplify variables and answer questions*
- Benchmarks
 - Task domain sample executed by a computer or by a human and a computer
 - allow repeatable and objective comparisons
 - aids in identifying promising approaches and discarding poor ones
 - Benchmark composition is the big challenge
- Experiments involving humans also repeatable (medicine & psychology)

Eight Fallacies Exposed (#6)

Experimentation will slow progress

- *Rebuttal: Increasing the ratio of papers with meaningful validation has a good chance of actually accelerating progress*
- Good conceptual papers will continue to be published
- Need to get beyond assertion
 - “it seems intuitively obvious”,
 - “it looks like a good idea”,
 - “I tried it on a small example and it worked”

Eight Fallacies Exposed (#7)

Technology changes too fast

- *Rebuttal: If a question becomes irrelevant quickly, it is too narrowly defined and not worth spending a lot of effort on.*
- Probe for fundamental and not the ephemeral
- Scientists should anticipate changes in assumptions and proactively employ experiments to explore consequence of changes

Eight Fallacies Exposed (#8)

You'll never get it published

- *Rebuttal: Smaller steps are still worth publishing because they improve our understanding and raise new questions*
- Non-theoretical journals and conferences accept papers on solid experimentation
- Respectable experimentalists articulate how their systems contribute to our knowledge
- “Systems come and go. We need insights about the concepts and phenomena underlying such systems”



41st International Conference on VERY LARGE DATA BASES

Hilton Waikoloa Hotel • Kohala Coast, Hawai'i
August 31 - September 4, 2015

Search

General Information

Conference Overview
Conference Officers
PVLDB Review Board
Acknowledgments

Program

Schedule at a Glance
Full Program
Talk by PC Chairs
Accepted Research Papers
Accepted Industrial Papers
Turing Award Lecture
Accepted Demos
Accepted Tutorials
Keynotes
Panels
Workshops
PhD Workshop
Social Events
Poster Sessions
• Poster Session I
• Poster Session II
Session Guidelines
Awards
Booklet

Experiments & Analysis Papers



Green turtle, *Chelonia mydas*, Hawaii. Photo by Brocken Inaglory. GNU Free Documentation/Creative Commons Attribution license, via Wikimedia

Experiments and Analysis Papers focus on the experimental evaluation of existing algorithms, data structures, and systems. Papers proposing new techniques should continue to be submitted to the regular research track. The primary contribution of Experiments and Analysis papers is performance evaluation through analytical modeling, simulation, and/or experiments. Suitable papers can fit in different categories:

1. **Experimental Surveys:** papers that compare a wide spectrum of approaches to a problem and, through extensive experiments, provide a comprehensive perspective on the results available and how they compare to each other.
2. **Result Verification:** papers that verify or refute results published in the past and that, through a renewed performance evaluation, help to advance the state of the art.



2015 VLDB Swag



Follow me on Twitter

Tweets by @VLDB2015



Inherent Problems with Experimentation

- Unrealistic assumptions
- Manipulated data
- Impossible to quantify key variable
- Few *Competing Theories* in CS
 - Rarely produce falsifiable theories: tend to pursue math theories that are disconnected from the real world

Similar assessment, 10 years later

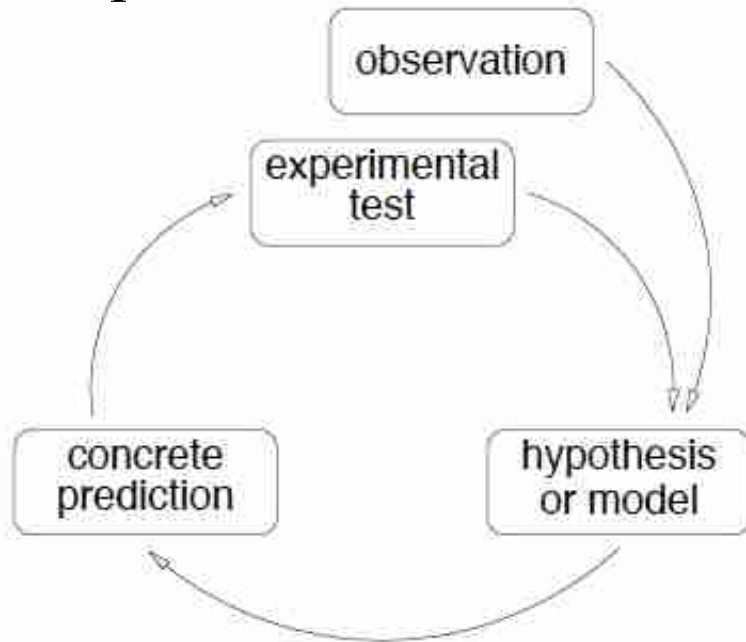
- Comm. of ACM Nov 2007
 - “In computer science, theory seems to play a more dominant role, typically with little if any direct connection to experimentation”
 - Experimentation:
 - Observe, measure under controlled conditions, replicate previous work
 - Systems are increasingly complex and require objective measurement to be studied and understood
 - “New system designs must be evaluated in the context of representative workloads”
 - “Evaluations under controlled conditions are critical for progress in the development of practical systems”

Experimental Computer Science: The Need for A Cultural Change

- Science
 - Observation + Hypothesis Testing + Reproducibility
- Research in CS:
 - typically part of
 - Engineering (building tools/systems) or
 - Math (building and studying abstract processes/structures)
 - Counter-example: Self-similarity in computer network traffic
 - Originated from pure observation - SCIENCE

The role of experimentation

Experimentation in science



Experimental feedback in engineering

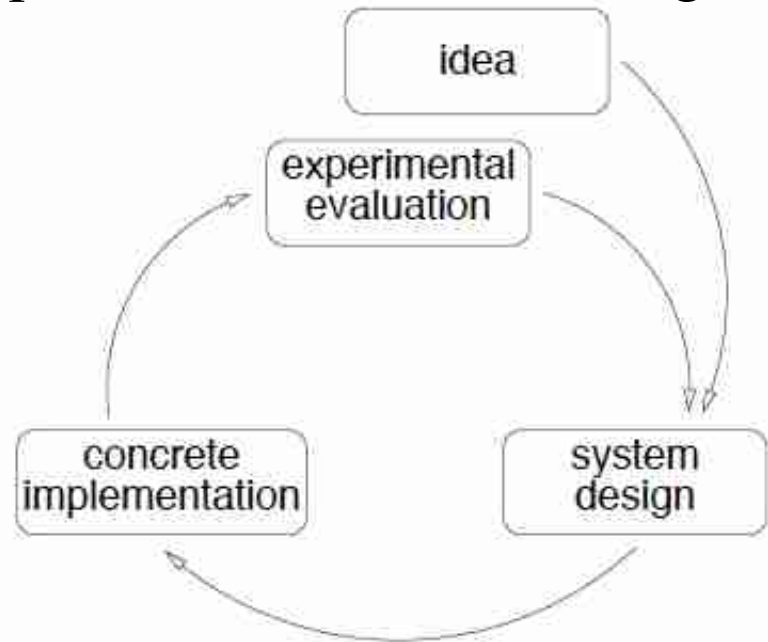


Figure 1: A comparison of the scientific method (on the left) with the role of experimentation in system design (right).

“In many cases system experiments are more demonstrations that the idea or system works than a real experiment

The role of experimentation

- Key questions that should be made:
 - Is it based on empirical evaluation and data?
 - Was the experiment designed correctly?
 - Is it based on a toy or a real situation?
 - Were the measurements used appropriate for the goals of the experiment?
 - Was the experiment run for a long enough time?
- Moreover, need to avoid (when comparing against alternative):
 - Bias in favor of own system
 - Tendency to compare against restricted, less optimized versions of the competition

Observation+Hypothesis+ Reproducibility

- Observation (measurement) leads to models and better understanding of process at hand
 - “Art and science have their meeting point in method”
- Need appropriate **metrics**
- Models turn measured data into **information**
 - Good models are simple (do we need a new model?)
 - Do not fit the measurements to the theory but rather fit the theory to the measurements
- Actual numbers may be of little use, the **understanding** derived from them has wider applicability

Processo Experimental Sistemático

1. Entenda o problema, estabeleça as perguntas e defina os objetivos
2. Selecione métricas
3. Identifique os parâmetros
4. Decida quais parâmetros serão estudados, i.e., serão variados (fatores)
5. **Selecione a técnica**
6. **Selecione a carga de trabalho (workload)**
7. Execute experimentos
8. Analise e interprete os resultados
9. Apresente os resultados e dados do experimento
10. Apresente conclusões

Visão Geral das Principais Técnicas que Iremos Estudar

1. Caracterização de Cargas
2. Sumarização de Dados
3. Comparando Quantitativamente Sistemas
4. Projeto Fatorial Completo:
 - a. Projetos fatoriais 2^k
 - b. Projetos fatoriais com replicação
 - c. Projetos fatoriais fracionários
 - d. Projeto de um fator
5. Regressão Linear Simples
6. Regressão Linear Múltipla
7. Regressão Curvilinear

Caracterização de Cargas

- Como descrever a carga do sistema alvo precisamente?
 - Qual importância?
 - Quais perguntas você deve tentar responder?
 - Qual propósito da caracterização?
 - Qual nível da descrição?
 - Quais componentes de carga queremos descrever?

Exemplo de um servidor Web

- Carga descrita por duplas (CPU time, I/O time)
- Dados disponíveis: logs de 10 requisições HTTP
- Caso 1: só 1 tamanho de documento (15KB)
 - 10 execuções ---> (0.013 sec, 0.09 sec)
- Caso 2 (mais realista):
 - documentos têm tamanhos diferentes.

Exemplo de um servidor Web

Request No.	CPU time	I/O time	Elapsed time
1	0.0095	0.04	0.071
2	0.0130	0.11	0.145
3	0.0155	0.12	0.156
4	0.0088	0.04	0.065
5	0.0111	0.09	0.114
6	0.0171	0.14	0.163
7	0.2170	1.20	4.380
8	0.0129	0.12	0.151
9	0.0091	0.05	0.063
10	0.0170	0.14	0.189
Average	0.0331	0.205	0.550

Como representar / sumarizar esta carga?

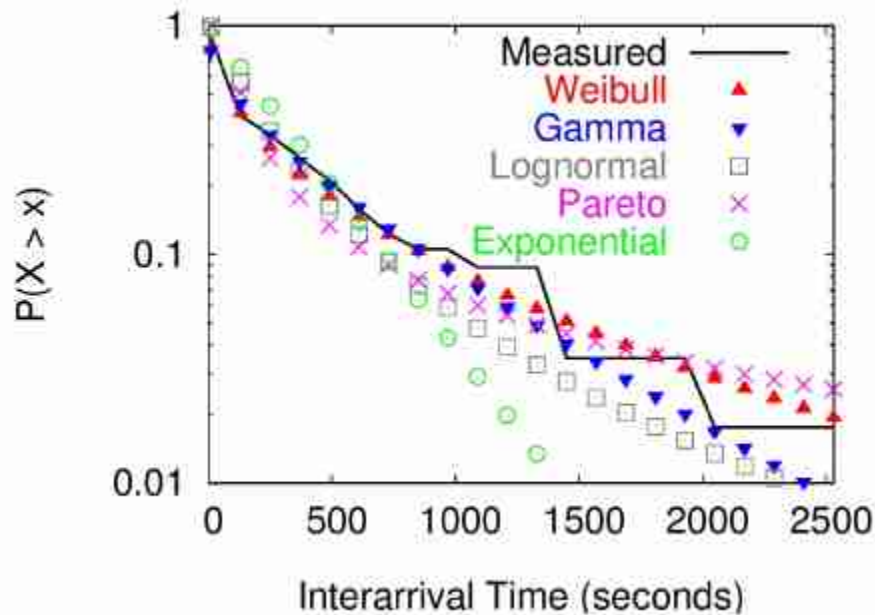
Somente uma média é suficiente (representativo)?

Caracterização de Cargas

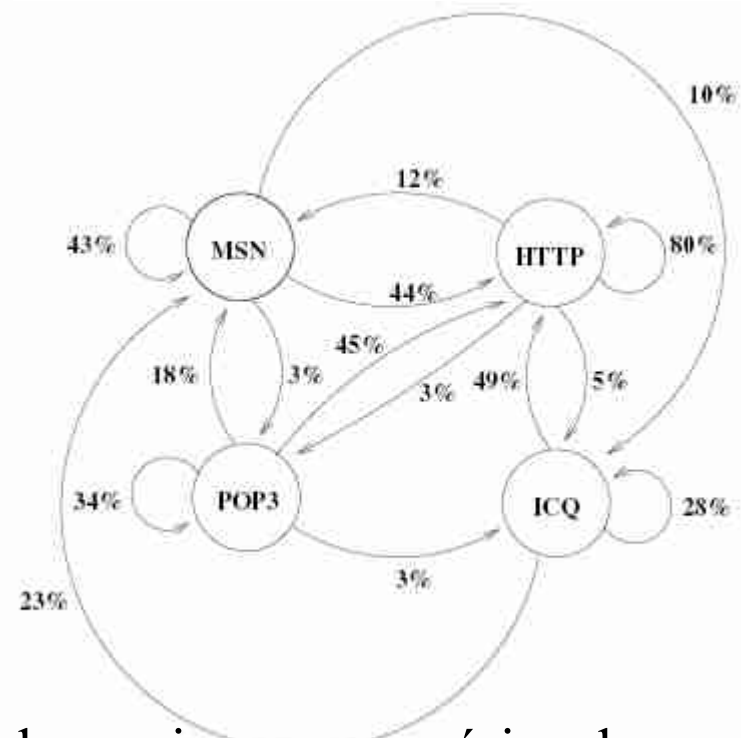
- Particionamento de cargas em sub-classes
 - Quando? Como?
- Representação de cargas em workload models
 - Distribuições estatísticas, geração de cargas sintéticas

Caracterização/Modelagem de Cargas

- Particionamento de cargas em sub-classes
 - Quando? Como?
- Representação de cargas em workload models
 - Distribuições estatísticas, geração de cargas sintéticas



eTeach

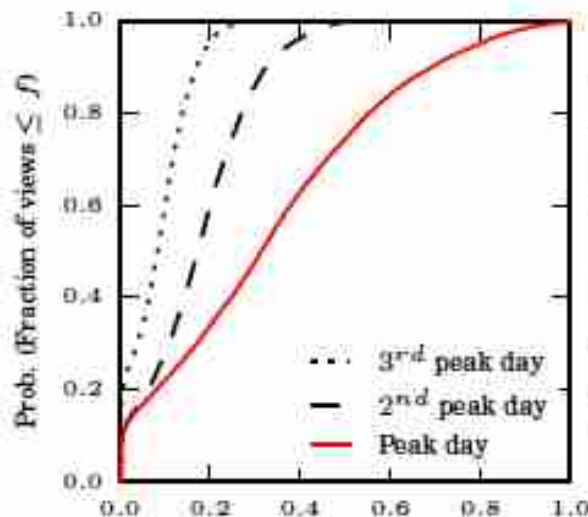


Tempos entre chegadas de requisições em um servidor de vídeo

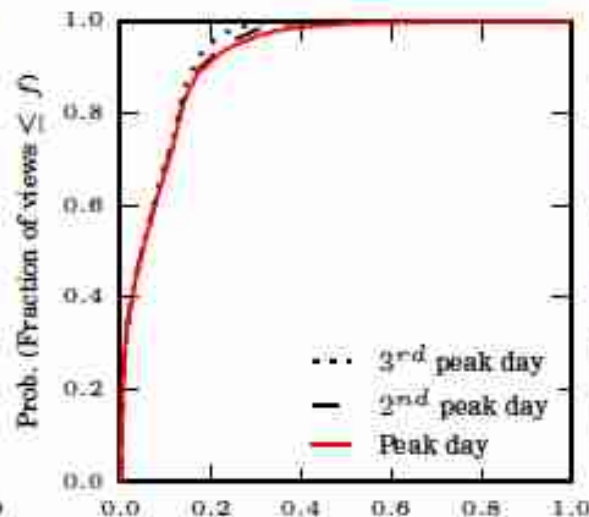
Uso de serviços por usuários de banda larga

Caracterização/Modelagem de Cargas

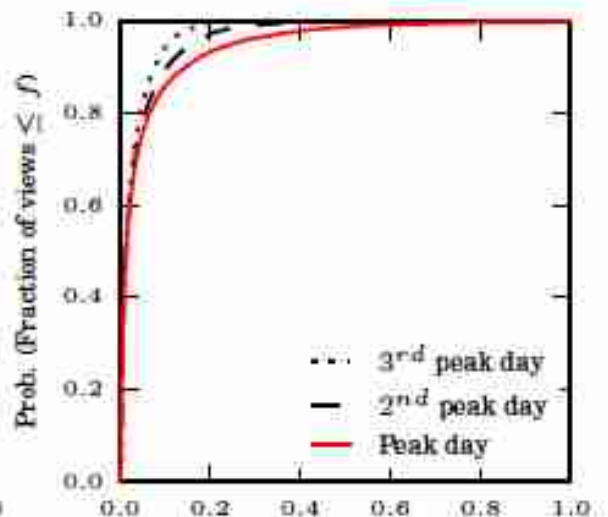
- Particionamento de cargas em sub-classes
 - Quando? Como?
- Representação de cargas em workload models
 - Distribuições estatísticas, geração de cargas sintéticas



(a) Top (Peak Day)



(b) YouTomb (Peak Day)

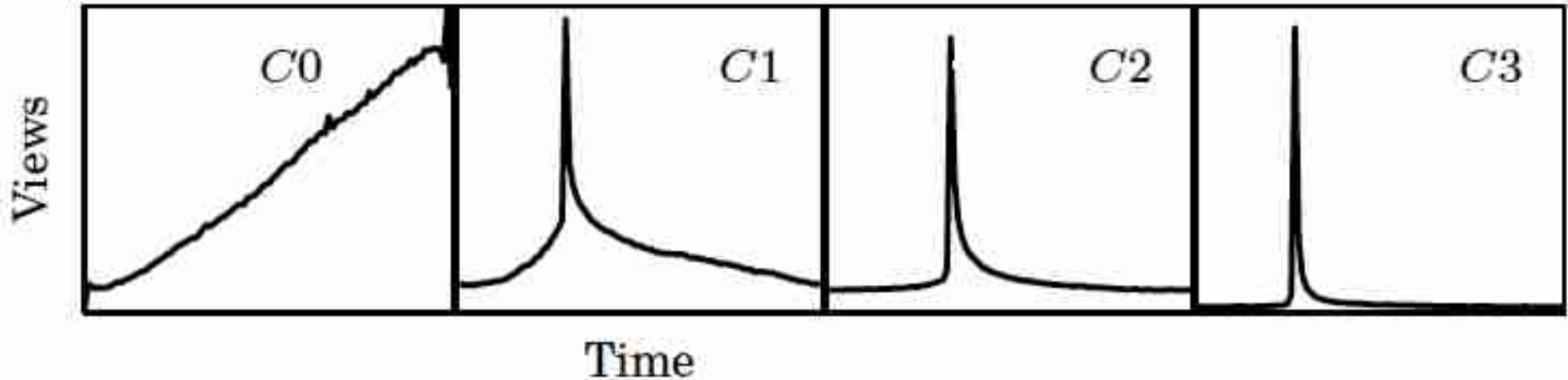


(c) Random (Peak Day)

Picos de popularidade em vídeos do YouTube

Caracterização/Modelagem de Cargas

- Particionamento de cargas em sub-classes
 - Quando? Como?
- Representação de cargas em workload models
 - Distribuições estatísticas, geração de cargas sintéticas



Perfis de popularidade em vídeos do YouTube

Caracterização/Modelagem de Cargas

- Particionamento de cargas em sub-classes
 - Quando? Como?

main-marcos.pdf - Adobe Reader

Arquivo Editar Visualizar Janela Ajuda

Abrir 10 / 16 125% Ferramentas Preencher e assinar Comentário

Table 3. Spearman's correlation coefficient between each productivity index and scholar popularity for individual scholars.

University	Course	Conference papers	Journal articles	Student supervisions
UFMG	Computer Science	0.616	0.612	0.437
	Electrical Engineering	0.450	0.670	0.481
	Mathematics	0.575	0.234	0.138
	Physics	0.174	0.645	0.308
	Animal Science	0.312	0.314	0.318
	Biochemistry and Immunology	0.124	0.526	0.403
	Bioinformatics	0.462	0.241	0.229
	Cell Biology	0.037	0.521	0.467
	Ecology	0.614	0.728	0.496
	Neurosciences	-0.370	0.263	0.147
	Molecular Medicine	-0.234	0.304	0.428
UNICAMP	Computer Science	0.500	0.645	0.525
	Chemical Engineering	0.417	0.752	0.403
	Electrical Engineering	0.197	0.452	0.076
	Mathematics	0.347	0.717	0.377
	Physics	0.069	0.462	0.173
	Medical Sciences	0.185	0.506	0.401
Stanford	Computer Science	0.396	0.384	-
	Medicine	0.480	0.672	-

Ask me anything!

Windows Taskbar: 09:11 21/03/2016

Sumarização de Dados

- Revisão:
 - probabilidade, independência, regra de Bayes
 - população X amostra
 - Média, variância, desvio padrão, CV, correlação, quantis
 - Distribuições estatísticas
 - Teorema Central do Limite:
 - “ a distribuição das médias amostrais será aproximadamente Normal”
- Estimando uma população a partir de uma amostra: tudo o que você quer fazer
 - Intervalo de Confiança : confiança X precisão

Exemplo

35 samples: 10 16 47 48 74 30 81 42 57 67 7
13 56 44 54 17 60 32 45 28 33 60 36 59 73
46 10 40 35 65 34 25 18 48 63

Sample mean $\bar{x} = 42.1$.

Standard deviation $s = 20.1$. $n = 35$

90% confidence interval is

$$42.1 \pm (1.645) \frac{20.1}{\sqrt{35}} = (36.5, 47.7)$$

De onde veio o 1.645?

Como interpretar o intervalo (36.5 , 47.7) para usá-lo na prática:
confiança X precisão da estimativa

Comparação de Sistemas

- Comparação Quantitativa de Sistemas: algoritmos, métodos, protótipos
 - Como usar Intervalos de Confiança
 - Teste com observações pareadas
 - Teste com observações não pareadas
 - Teste-t
 - Comparando valores médios X proporções
 - Escolha do tamanho da amostra

Exemplo 1

1. O tempo de processamento necessario para executar uma tarefa foi medido em dois sistemas.

Os tempos no sistema A foram:

$\{5.36, 16.57, 0.62, 1.41, 0.64\}$.

Os tempos no sistema B foram:

$\{19.12, 3.52, 3.38, 2.50, 3.60\}$.

Os dois sistemas são significativamente diferentes?

O sistema A é significativamente melhor/pior?

Exemplo 2

2. O tempo de processamento necessario para executar cinco consultas diferentes em dois sistemas foram medidos

Consulta	Sistema A	Sistema B
1	5.36	19.12
2	16.57	3.52
3	0.62	3.38
4	1.41	2.50
5	0.64	3.60

Os dois sistemas são significativamente diferentes?

O sistema A é significativamente melhor/pior?

Exemplos 2 e 3

2. Um experimento foi repetido em dois sistemas 40 vezes. O sistema A foi superior ao B em 26 repeticoes. Podemos dizer que, com confianca de 99%, o sistema A e superior ao sistema B? E com uma confianca de 90%?

2. Cinco execuções independentes de um sistema levaram a tempos de execução iguais a 22.5, 19.8, 21.1, 26.7, 20.2 segundos
Quantas vezes devo repetir o experimento para obter um intervalo de confiança com precisão de $\pm 5\%$ e um nível de confiança de 90%?

Projetos Fatoriais 2^k

- Usados para determinar os efeitos de k fatores
 - Cada um com duas alternativas ou níveis
 - Cada fator medido é representado por seu nível máximo e pelo seu nível mínimo.
- Em geral, são usados de maneira preliminar, antes de estudos mais detalhados
 - Responde a pergunta: qual(is) dos k fatores tem maior impacto no sistema alvo?
 - Pode oferecer algum “insight” sobre as interações entre os vários fatores.
 - Existem interações com impacto significativo?
 - Se sim, qual a consequencia para o projeto experimental?

Exemplo de Projeto Fatorial 2²

- Uma arquitetura de máquina de busca, composta por N servidores.
 - 1º fator é o número de servidores N ($N = 8$ ou 64)
- Pode-se usar vários esquemas de distribuição ou escalonamento de *queries* para os servidores, por exemplo, *round-robin*, *gang*, *random*, *priority*, etc
 - 2º fator é o escalonamento (random ou round-robin)
- O objetivo é completar as *queries* no menor tempo possível.

Logo, a métrica de interesse é o tempo de execução da *query* em microsegundos.

Exemplo de Projeto Fatorial 2²

- Execução única de uma carga benchmark de *queries* nas duas configurações resultou nos seguintes tempos de execucao:

	8 Serv. (-1)	64 Serv. (+1)
Rand. (-1)	820	217
RR (+1)	776	197

- Fator A = número de servidores
- Fator B = escalonamento

Exemplo de Projeto Fatorial 2²

- Objetivo: produzir um modelo (não linear) para estimativa do tempo de execução de uma query

$$y = 502.5 - 295.5x_A - 16x_B + 6x_Ax_B$$

$$q_0 = 502.5 = \text{tempo de execução médio}$$

Fator A - número de servidores -
faz uma diferença de $\pm 295,5$ no tempo de resposta

De fato A explica 97% da variação nos dados

Fator B – escalonamento – baixo impacto (3%)

Interferência entre A e B também baixo impacto (0.4%)

Proceder: refinar avaliação do impacto de A: fator único

Projetos Fatoriais

- Projetos Fatoriais 2^k
 - 2^k experimentos
 - Objetivo: selecionar parâmetros de maior impacto para avaliação mais refinada a seguir
- Projetos Fatoriais com replicação
 - Replicação de experimentos independentes para cada uma das 2^k configuração
 - Permite estimar erros de experimentação e avaliar qualidade do modelo
- Projetos Fatoriais Fracionários
 - E se for muito caro rodar todos os 2^k (com/sem replicação) experimentos?
- Projeto de Um Fator

Exemplo: CEC 2010

Tuning Genetic Programming Parameters with Factorial Designs

Elisa Boari de Lima, Gisele L. Pappa, Jussara Marques de Almeida,
Marcos A. Gonçalves and Wagner Meira Jr.

Abstract—Parameter setting of Evolutionary Algorithms is a time consuming task with two main approaches: parameter tuning and parameter control. In this work we describe a new methodology for tuning parameters of Genetic Programming algorithms using factorial designs, one-factor designs and multiple linear regression. Our experiments show that factorial designs can be used to determine which parameters have the largest effect on the algorithm's performance. This way, parameter setting efforts can focus on them, largely reducing the parameter search space. Two classical GP problems were studied, with six parameters for the first problem and seven for the second. The results show the maximum tree depth as the parameter with the largest effect on both problems. A one-factor design was performed to fine-tune tree depth on the first problem and a multiple linear regression to fine-tune tree depth and number of generations on the second.

parameters for genetic algorithms, and search the space of parameter values using some sort of heuristic.

Many disciplines in Computer Science and Engineering require parameter setting. Factorial experimental designs [8] are one of the most broadly applied techniques for this purpose. The idea behind a factorial design is to obtain the maximum information about the factors (i.e., parameters) with the minimum number of experiments. Furthermore, it allows us to determine whether a parameter's effect on the algorithm's output is significant compared to the experimental error inherent to the measurement process.

The proposed methodology for parameter tuning relies on a 2^k factorial design to estimate the influence of the parameters on the GP runs. Once the parameters with the

Exemplo: CEC 2010

TWO PARAMETER CONFIGURATIONS OF THE FACTORIAL DESIGNS FOR SYMBOLIC REGRESSION, WITH LOWER (LB) AND UPPER BOUNDS (UB).

Parameter	Code	Config. 1		Config. 2	
		LB	UB	LB	UB
Population Size	A	20	10000	350	650
Generations	B	5	250	35	65
Tree Depth	C	2	20	2	20
Crossover Rate	D	0.6	0.99	0.6	0.99
Mutation Rate	E	0	0.4	0	0.4
Reproduction Rate	F	0	0.2	0	0.2

Exemplo: CEC 2010

TABLE III
RESULTS OF THE 2^6 30 FACTORIAL DESIGN FOR SYMBOLIC
REGRESSION. SIGNIFICANT EFFECTS AND CORRESPONDING
PERCENTAGE OF EXPLAINED VARIATION ARE PRESENTED.

Factor	Config. 1 %	Effect	Config. 2 %	Effect
A	35.8	-0.35	0.53	-0.03
B	15.11	-0.23	2.28	-0.07
C	16.38	-0.24	69.00	-0.37
E	0.63	-0.05	0.42	0.03
F	-	-	0.19	-0.02
AB	0.45	-0.04	-	-
AC	3.88	-0.12	0.15	-0.02
AE	0.61	0.05	-	-
BC	10.37	-0.19	1.91	-0.06
BE	0.56	-0.04	-	-
CE	-	-	0.95	0.04
CF	-	-	0.22	-0.02
ABC	1.47	-0.07	-	-
ABE	0.53	0.04	-	-
ABCE	0.05	0.01	-	-
% Explained	86.07%		75.95%	
Std Deviation	0.005		0.005	

The influence of A and B in the first but not in the second parameter configuration is closely related to the lower and upper bounds chosen for these parameters: since the total number of fitness evaluations varies in orders of magnitude from the lower to the upper bound in the first configuration, the impact of these parameters will certainly be large.

Exemplo: CEC 2010

TABLE IV

ONE-FACTOR DESIGN RESULTS FOR SYMBOLIC REGRESSION. RESULTS PER TREE DEPTH: AVERAGE FITNESS, EFFECT AND 99% CONFIDENCE INTERVAL OF EACH EFFECT. $\mu = 5.84$. THE TREE DEPTH EXPLAINS 63.43% OF THE TOTAL FITNESS VARIATION.

	H2	H6	H8	H10	H12	H14	H16
\bar{y}_j	30.50	2.78	1.83	1.95	2.06	2.29	2.55
α_j	24.67	-3.06	-4.01	-3.89	-3.77	-3.55	-3.28
CI	(21.79, 27.53)	(-5.93, -0.19)	(-6.88, -1.14)	(-6.76, -1.02)	(-6.64, -0.91)	(-6.42, -0.68)	(-6.15, -0.42)

Exemplo: JASIST 2015

A quantitative analysis

online library wiley.com/enhanced/doi/10.1002/jasi.23452

Get access

JOURNAL OF THE ASSOCIATION FOR
INFORMATION SCIENCE AND TECHNOLOGY

Research Article

A quantitative analysis of the temporal effects on automatic text classification

Thiago Salles, Leonardo Rocha, Marcos André Gonçalves, Jussara M. Almeida, Fernando Mourão, Wagner Meira Jr., Felipe Vargas

First published: 7 August 2015 Full publication history

DOI: 10.1002/jasi.23452 View/save citation

Cited by: 0 articles Check for new citations

Abstract

Automatic text classification (TC) continues to be a relevant research topic and several TC algorithms have been proposed. However, the majority of TC algorithms assume that the underlying data distribution does not change over time. In this work, we are concerned with the challenges imposed by the temporal dynamics observed in textual data sets. We provide evidence

Early View

Read FREE 2016 Energy Sample Issues!

WILEY

Provide feedback or get help

Ask me anything

POR INTL 08:13 23/03/2016

Exemplo: JIDM 2014

Experimental Evaluation × +

← → ↻ | Certificate error - web.lcc.ufmg.br/index.php/jidm/article/view/297/567

JOURNAL OF INFORMATION AND DATA MANAGEMENT

HOME ABOUT LOGIN REGISTER SEARCH CURRENT ARCHIVES ANNOUNCEMENTS

Home > Vol 5, No 1 (2014) > Brandão

Download this PDF file

Experimental Evaluation of Academic Collaboration Recommendation Using Factorial Design

Michele A. Brandão, Mirilla M. Moro and Jussara M. Abreita

Universidade Federal de Minas Gerais, Brazil
(michelabrandao, mirilla, jussara)@dcc.ufmg.br

Abstract. Recommender systems have been used in e-commerce and online social networks. Among various challenges to construct such systems, how to parameterize them and their evaluation are two scarcely explored issues. Generally, each recommendation strategy has parameters and factors that can be varied. In this article, we propose to evaluate the impact of key parameters of two state-of-the-art functions that recommend academic collaborations. Our experimental results show that the factors affect recall, novelty, diversity and coverage of the recommendations in different ways. Finally, such evaluation shows the importance of studying the impact of the factors and factor interactions in the academic collaboration recommendations strategy.

Fullscreen

OPEN JOURNAL SYSTEMS

Journal help

USER

Username

Password

☐ Remember me

Login

JOURNAL CONTENT

Search

Search Scope

All

Search

Browse

By Issue

By Author

By Title

Other Journals

FONT SIZE

A⁺ A A⁻

INFORMATION

For Readers

For Authors

For Librarians

An official publication of the Brazilian Computer Society Special Interest Group on Databases.

Windows taskbar: Ask me anything, 08:14, 23/03/2016

Regressão Linear Simples

- Fazer uma regressão é uma maneira de criar um modelo de predição para uma variável resposta y utilizando observações anteriores (x, y)
 - Modelo útil para previsão e também para analisar impacto de fator x em vários níveis.
- Como determinar se o modelo criado é adequado / bom?
- Como estimar os parâmetros do modelo?
- Como utilizar o modelo para realizar predições?
 - Respostas : precisão X confiança
 - Intervalos de confiança de parâmetros, predições, etc
- Quais premissas têm que ser respeitadas?

Exemplo

- O numero de disk I/Os e o tempo de CPU foram medidos para sete programas
(14, 2) (16, 5) (27, 7) (42, 9) (39, 10) (50, 13)
(83, 20).

Desenvolva um modelo de predicao de tempo de CPU em funcao do numero de operacoes de disco.

Este modelo e razoavel?

Resolução

- Modelo: $\text{CPU Time} = -0.0083 + 0.2438 \text{ DiskTime}$
- Modelo é bom?
Regressão explica 97.15 % da variação nos valores de CPU Time: muuuuuiitttoooo bom!!!
- Existe modelo melhor?
Podemos mostrar que o parâmetro constante (-0.0083) não é significativo com 90% de confiança

Logo um modelo mais simples e tão bom quanto (em termos de significância estatística) é:

$$\text{CPU Time} = 0.2438 \text{ DiskTime}$$

Regressão Linear Múltipla

- Modelo de regressão linear para estimar variável y em função de mais de uma variável previsora
- Problemas na regressão: múltipla colinearidade
 - Teste F
- Regressão com previsores categóricos
 - Uma ou mais variáveis previsoras (mas não todas) não são numéricas
 $x_1 = \text{tipo A ou tipo B ou tipo C}$

Exemplo

- O desempenho de uma chamada de procedimento remota (RPC) foi comparada em um mesmo sistema operacional utilizando duas primitivas (A ou B) alternativas do mesmo. A métrica avaliada foi o tempo total para um tamanho fixo (mensagem) de dados. Qual o custo de processamento por mensagem para cada primitiva? Qual o custo fixo, comum às duas?

$y = \text{tempo de processamento} = f(x)$

$x = 1$ (primitiva A)

0 (primitiva B)

Regressão Curvilinear

- Como transformar para relações lineares?
- Quando?
- Qual transformação utilizar?

Exemplo

- Sete programas foram monitorados para observar suas demandas por recursos, em particular o numero de operacoes de discos, o tamanho da memoria (em KB) e o tempo de CPU (em ms). Utilize as observacoes (abaixo) para estimar o tempo de CPU de um programa em funcao de numero de operacoes de disco e do tamanho de memoria ocupados. Uma funcao linear e um bom modelo?

Tempo de CPU	Disk I/Os	Tamanho da memoria
Y_i	X_{1i}	X_{2i}
2	14	70
5	16	75
7	27	144
9	42	190
10	39	210
13	50	235
20	83	400

Ao fazermos uma regressão linear múltipla nestes dados, encontramos um modelo que explica 97% das variações observadas (muito bom) mas também encontramos que nenhum de seus parâmetros é significativo com uma confiança de 90% -> POR QUE e O QUE FAZER?

Exemplo: WWW 2014

Demographics, Weather and Online Reviews: A Study of Restaurant Recommendations

Saeideh Bakhshi
Georgia Tech
sbakhshi@cc.gatech.edu

Partha Kanuparth
Yahoo Labs*
parthak@yahoo-inc.com

Eric Gilbert
Georgia Tech
eegilbert@cc.gatech.edu

ABSTRACT

Online recommendation sites are valuable information sources that people contribute to, and often use to choose restaurants. However, little is known about the dynamics behind participation in these online communities and how the recommendations in these communities are formed. In this work, we take a first look at online restaurant recommendation communities to study what endogenous (i.e., related to entities being reviewed) and exogenous factors influence people's participation in the communities, and to what extent. We analyze an online community corpus of 840K restaurants and their 1.1M associated reviews from 2002 to 2011, spread across every U.S. state. We construct models for number of reviews and ratings by community members, based on several dimensions of endogenous and exogenous factors. We find that while endogenous factors such as restaurant attributes (e.g., meal, price, service) affect recommendations, surprisingly, exogenous factors such as demographics (e.g., neighborhood diversity, education) and weather (e.g., temperature, rain, snow, season) also exert a significant effect on reviews. We find that many of the effects in online communities can be explained using *offline* theories from experimental psychology. Our study is the first to look at exogenous factors and how it related to online online restaurant reviews. It has implications for designing online recommendation sites, and in general, social media and online communities.

In this paper, we ask the question: might these phenomena documented in psychology studies also affect¹ large-scale online behavior? Could weather and local demographics of restaurants drive how we rate them online?

Review and recommender sites are highly popular online resources. People contribute content in the form of recommendations to these communities. A recent work by Anderson and Magruder [2] found that an extra half star rating on Yelp causes restaurants to sell out 19% more frequently. People increasingly base their decisions on input from such online reviews and ratings. A recent survey found that 64% of consumers search for online reviews before spending on services [7], and 85% of them are more likely to purchase services when they can find online recommendations. The same study found that 87% of consumers say that positive online reviews reinforce their decisions, while 80% say that negative online reviews have led them to change their minds. These findings imply significant returns on an extra half-star rating or more number of reviews and suggest that restaurants have strong incentives to improve their online reviews and ratings.

Despite their widespread use, little is known about the *dynamics* of participation and of contributions by people in online recommendation sites. For example, would factors such as weather conditions, that are shown to influence mood and behavior, affect people's participation and recommendations? Are restaurants from neighborhoods with high education levels more likely to receive

Exemplo: WWW 2014









Type	Variable	min	mean	median	max	std.dev	Distribution
Restaurant	number of reviews*	0	1.23	0	595	5.19	
	restaurant rating	0	7.8	8	10	2.03	
Review	review rating	0	7.26	8	10	3.11	
	polarity	-1	0.26	0.25	1	0.22	
	subjectivity	0	0.57	0.57	1	0.13	
Demographics	median income*	0	52.0K	46.9K	278K	26.8K	
	population density*	0	3.9K	2.4K	69.4K	5.3K	
	diversity index	0	0.46	0.51	0.86	0.22	
	higher education	0	0.27	0.22	1	0.19	
Weather	mean temperature	0	29.43	23	104.5	30.48	
	precipitation*	0	5.68	0	823	27.49	
	snow*	0	0.46	0	640	5.57	

Table 2: Distributions of quantitative variables used in this paper. Variables marked with "*" are log transformed.

Exemplo: WWW 2014

The first model uses restaurant attributes (endogenous variables ed) as predictors of the number of reviews a restaurant receives.

$$\ln(y) = I + \sum_{x_i \in ed} \beta_i x_i \quad (1)$$

where I is the intercept for the model and the endogenous sum is computed using the following restaurant-related attributes:

$$\begin{aligned} \sum_{x_i \in ed} \beta_i x_i = & \beta_{price} x_{price} + \beta_{offers} * x_{offers} + \beta_{bar} * x_{bar} \\ & + \beta_{delivery} * x_{delivery} + \beta_{carryout} x_{carryout} \\ & + \beta_{meal} x_{meal} + \beta_{featured} * x_{featured} \end{aligned} \quad (2)$$

This model allows us to understand the effect on the number of reviews of endogenous variables alone.

We then model the impact of exogenous factors (local weather and demographics) on the number of reviews as follows. We construct a second model that includes both restaurant attributes and exogenous attributes as predictors. We also include predictors for the interaction between some pairs of the independent variables.

$$\ln(y) = I + \sum_{x_i \in ed} \beta_i x_i + \sum_{x_j \in ex} \beta_j x_j \quad (3)$$

where, the endogenous sum is taken from equation 2 and exogenous sum is computed using demographics variables and interaction between endogenous and demographic variables:

$$\begin{aligned} \sum_{x_j \in ex} \beta_j x_j = & \beta_{region} x_{region} + \beta_{pop} x_{pop} + \beta_{income} x_{income} \\ & + \beta_{edu} x_{edu} + \beta_{diversity} x_{diversity} \\ & + \beta_{carryout * pop} x_{carryout} x_{pop} + \beta_{delivery * pop} x_{delivery} x_{pop} \end{aligned} \quad (4)$$

Exemplo: WWW 2014

Predictor	ref. category	ed model		ed+ex model	
		β	p	β	p
(Intercept)		0.23	<e-15	-0.97	<e-15
β_{featured}	not featured	1.38	<e-15	1.33	<e-15
$\beta_{\text{price}=\$}$	price=\$	1.77	<e-15	1.59	<e-15
$\beta_{\text{price}=\$ \$}$	price=\$	2.19	<e-15	1.93	<e-15
$\beta_{\text{price}=\$ \$ \$}$	price=\$	2.30	<e-15	2.10	<e-15
β_{bar}	no bar	0.82	<e-15	0.72	<e-15
β_{carryout}	no carryout	-0.77	<e-15	-0.41	<e-15
β_{delivery}	no delivery	0.68	<e-15	0.28	<e-5
β_{offers}	no offers	1.69	<e-15	1.43	<e-15
$\beta_{\text{meal=breakfast}}$	meal=others	0.40	<e-15	0.33	<e-15
$\beta_{\text{meal=brunch}}$	meal=others	1.52	<e-15	1.32	<e-15
$\beta_{\text{meal=latenight}}$	meal=others	1.92	<e-15	1.55	<e-15
$\beta_{\text{meal=lunch}}$	meal=others	0.77	<e-15	0.58	<e-15
$\beta_{\text{region=mountain}}$	region=midwest			0.22	<e-15
$\beta_{\text{region=northeast}}$	region=midwest			0.23	<e-15
$\beta_{\text{region=pacific}}$	region=midwest			0.60	<e-15
$\beta_{\text{region=south}}$	region=midwest			-0.08	<e-15
$\beta_{\text{pop} \in (1000, 2500)}$	pop < 1000			0.20	<e-15
$\beta_{\text{pop} \in (2500, 5000)}$	pop < 1000			0.39	<e-15
$\beta_{\text{pop} \in (5000, 10000)}$	pop < 1000			0.32	<e-15
$\beta_{\text{pop} > 10000}$	pop < 1000			0.21	<e-15
$\beta_{\text{edu} \in (10\%, 25\%)}$	edu < 10%			0.31	<e-15
$\beta_{\text{edu} \in (25\%, 50\%)}$	edu < 10%			0.71	<e-15
$\beta_{\text{edu} > 50\%}$	edu < 10%			1.02	<e-15
$\beta_{\text{diversity} \in (0.3, 0.5)}$	diversity < 0.3			0.13	<e-15
$\beta_{\text{diversity} \in (0.5, 0.7)}$	diversity < 0.3			0.18	<e-15
$\beta_{\text{diversity} > 0.7}$	diversity < 0.3			-0.12	<e-15
$\beta_{\text{carryout} \bullet \text{pop} \in (1000, 2500)}$				-0.26	<e-15
$\beta_{\text{carryout} \bullet \text{pop} \in (2500, 5000)}$				-0.27	<e-15
$\beta_{\text{carryout} \bullet \text{pop} \in (5000, 10000)}$				-0.09	<e-3
$\beta_{\text{carryout} \bullet \text{pop} > 10000}$				0.03	0.42
$\beta_{\text{delivery} \bullet \text{pop} \in (1000, 2500)}$				0.13	0.07
$\beta_{\text{delivery} \bullet \text{pop} \in (2500, 5000)}$				0.28	<e-5
$\beta_{\text{delivery} \bullet \text{pop} \in (5000, 10000)}$				0.31	<e-3
$\beta_{\text{delivery} \bullet \text{pop} > 10000}$				0.70	<e-14
β_{polarity}				0.84	<e-15
$\beta_{\text{subjectivity}}$				0.13	<e-15

Table 5: Regression coefficients and their statistical significance for both models describing the number of reviews. The reference category refers to the category of the variable that each factor was compared against. For the interaction predictors, the reference can be determined by using the reference category of each of the variables.

Processo Experimental Sistemático

1. Entenda o problema, estabeleça as perguntas e defina os objetivos
2. Selecione métricas
3. Identifique os parâmetros
4. Decida quais parâmetros serão estudados, i.e., serão variados (fatores)
5. Selecione a técnica
6. Selecione a carga de trabalho (workload)
7. Execute experimentos
8. Analise e interprete os resultados
9. Apresente os resultados e dados do experimento
10. Apresente conclusões