Métodos Quantitativos para Ciência da Computação Experimental

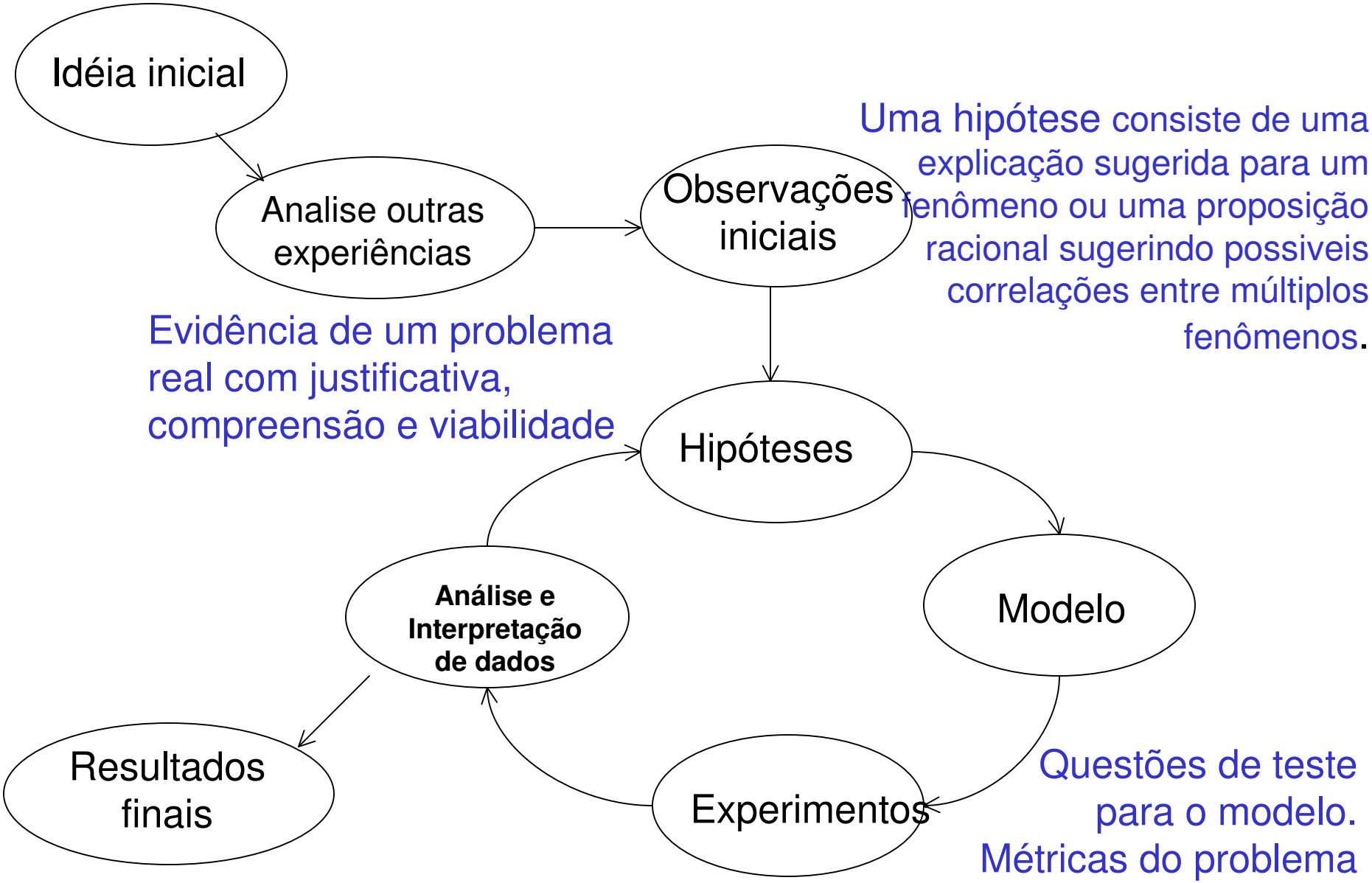
Aplicação do Método -Bloco #1c

Virgílio A. F. Almeida

Março 2008

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

Ciclo de Vida Experimental



Processo Experimental Sistemático

1. Entenda o problema, estabeleça as perguntas e defina os objetivos
2. Selecione métricas que ajudarão analisar as perguntas.
3. Identifique os parâmetros que afetam o comportamento
4. Decida quais parâmetros serão estudados, i.e., serão variados
5. Selecione a técnica, protótipos, simulação, medição de sistema real.
6. Selecione a carga de trabalho (workload)
7. Execute experimentos
8. Analise e interprete os resultados

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"
2. Selecione métricas que ajudarão analisar as perguntas.

Analyzing Web Robots and Their Impact on Caching

Virgilio Almeida
Rudolf Riedi
Rodrigo Fonseca

Daniel Menascé
Flávia Peligrinelli
Wagner Meira Jr.

Dept. Computer Science, Federal University of Minas Gerais, Brazil
Dept. Computer Science, George Mason University
Dept. of Elect. and Comp. Eng., Rice University

Web Caching Workshop and ACM Sigmetrics 2001
www.cs.bu.edu/techreports/ 2001-017-wcw01-proceedings/101_almeida.pdf

Onde e quais são as falhas do paper?

Analisar à luz de um método científico, como o apresentado

Previous Work

- **In Search of Invariants for E-Business Workloads,** Daniel Menascé, Virgilio Almeida, Rudolf Riedi, Wagner Meira Jr., Flavia Ribeiro and Rodrigo Fonseca, ACM-EC Conference, Minneapolis, 2000
- **A Hierarchical and Multiscale Analysis of E-Business Workloads,** Daniel Menascé, Virgílio Almeida, Rudolf Riedi, Flávia Ribeiro, Rodrigo Fonseca, Wagner Meira Jr., *Performance Evaluation* **54**(1), Sept 2003, pp 33--57.

Motivation
Non-Human Request Sources

- Vague idea: non-human requests are different from human generated Web requests.
- Robots are used frequently and for several tasks (crawling, prefetching, price comparison, …)
  - Crawlers
    - Search Engines, Resource Discovery Agents, DB Dumpers, Email Collectors, Site Maintenance
    - Automated, programmed behavior
  - Shopbots
    - Meta Searchers, Price Comparers
    - Human *Triggered*
  - Proxies
- Varying server QoS requirements

Initial Questions

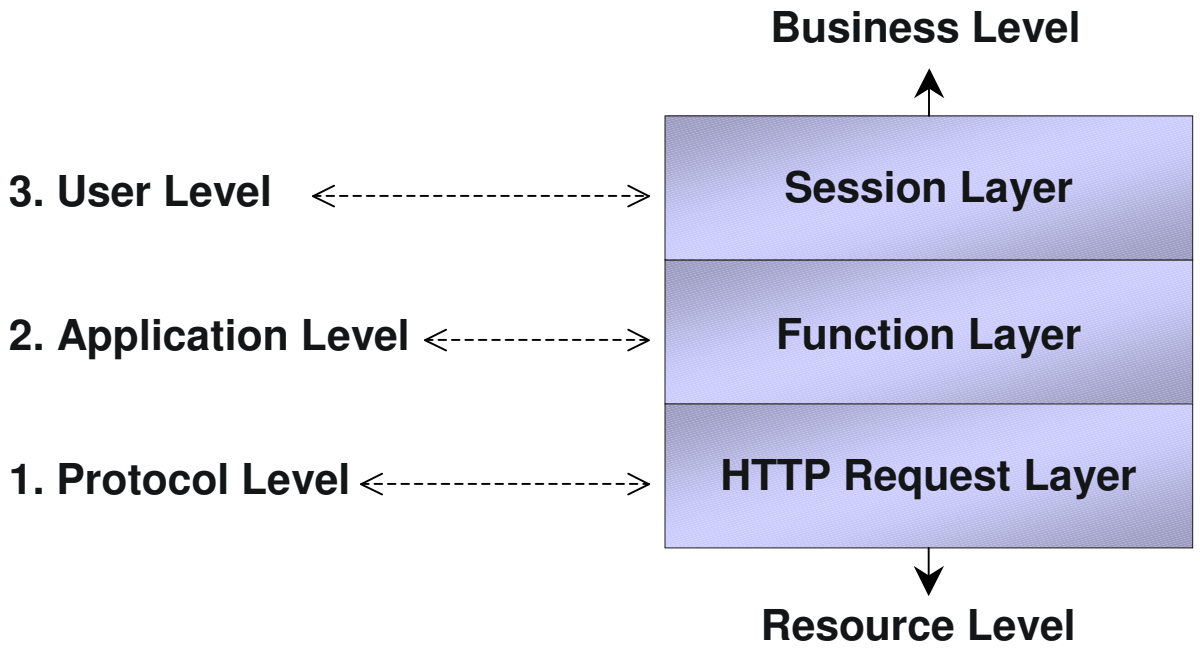
- How can we identify robots?
- How do robots affect reference locality?
- How should we handle robot's requests at both proxies and servers?
- What is the impact of robots on the system behavior?

Por que essas são questões relevantes?
Como você as faria hoje?

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"
2. Selecione métricas que ajudarão analisar as perguntas.

Multi-layered Hierarchical Workload Model and Metrics



Robot Identifying Metrics

- Session Layer
  - Session Length
  - Function Distribution
- Function Layer
  - Human Only Functions
  - Object Popularity
- Request Layer
  - Arrival process

Summary of the Identification Criteria

Source	Bookstore	Berkeley	WorldCup Site
Interval	01-15 Aug 1999	01-30 June 2000	23 May 1998
Number of requests	3,630,964	3,643,208	2,225,475
Percent of images	74%	44%	84%
Number of functions	955,818	2,038,249	340,719
% of robot's functions	33.51%	16.53%	6.46%
Number of sessions	130,314	371,242	33,995
Avg. robot's session length	2,409.60	1,324.93	1,398.16

Table 1: Characteristics of the Log Files

Robot Id	Session Length	Function	Human-Likely Function	Embedded Files	Self Identification	IAT Distribution
2	•	S		•		
6	•	S		•		
8	•	S		•		
25	•	S		•		
104	•	S		•		
3784	•	C		•		•
0	•	C				•
45282	•	C		•	•	•
584	•	C		•	•	•
47277	•	C		•		•

Table 2: Criteria used for the identification of the ten most important robots in the bookstore log

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)

4. Decida quais parâmetros serão estudados, i.e., serão variados.

Data Analyzed

- We analyzed three logs
  - Online Bookstore
  - UC Berkeley CS Department
  - 1998 World Cup Web Site

	Source		
	Bookstore	**Berkeley**	**World Cup Site**
Interval	01-15 Aug 1999	01-30 Jun 2000	23-May-98
Number of requests	3,630,964	3,643,208	2,225,475
Number of functions	955,818	2,038,249	340,719
% of Robot's Functions	33.51%	16.53%	6.46%
Number of Sessions	130,314	371,242	33,995
Avg. Robot Session Length	2,409.60	1,324.93	1,398.16

Processo Experimental Sistemático

5. Selecione a técnica:
   - Medição de uma implementação de protótipo
   - Quão invasivo? Podemos quantificar o "overhead"da monitoração? Podemos medir o que desejamos?
   - Simulação – quão 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

5. Selecione a técnica:
   - Medição de uma implementação de protótipo
   - Quão invasivo? Podemos quantificar o "overhead"da monitoração? Podemos medir o que desejamos?
   - Simulação – quão detalhada ? Como será a validação?
   - Repetibilidade
6. Selecione a carga de trabalho (workload)
   - Representativa?
   - É aceita pela comunidade científica?
   - Disponibilidade de dados?

Workload: Logs Analyzed

- We analyzed three logs
  - Online Bookstore
  - UC Berkeley CS Department
  - 1998 World Cup Web Site

	Source		
	Bookstore	**Berkeley**	**World Cup Site**
Interval	01-15 Aug 1999	01-30 Jun 2000	23-May-98
Number of requests	3,630,964	3,643,208	2,225,475
Number of functions	955,818	2,038,249	340,719
% of Robot's Functions	33.51%	16.53%	6.46%
Number of Sessions	130,314	371,242	33,995
Avg. Robot Session Length	2,409.60	1,324.93	1,398.16

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 Estatistica para analisar a variabilidade, “outliers”, etc.

Processo Experimental Sistemático

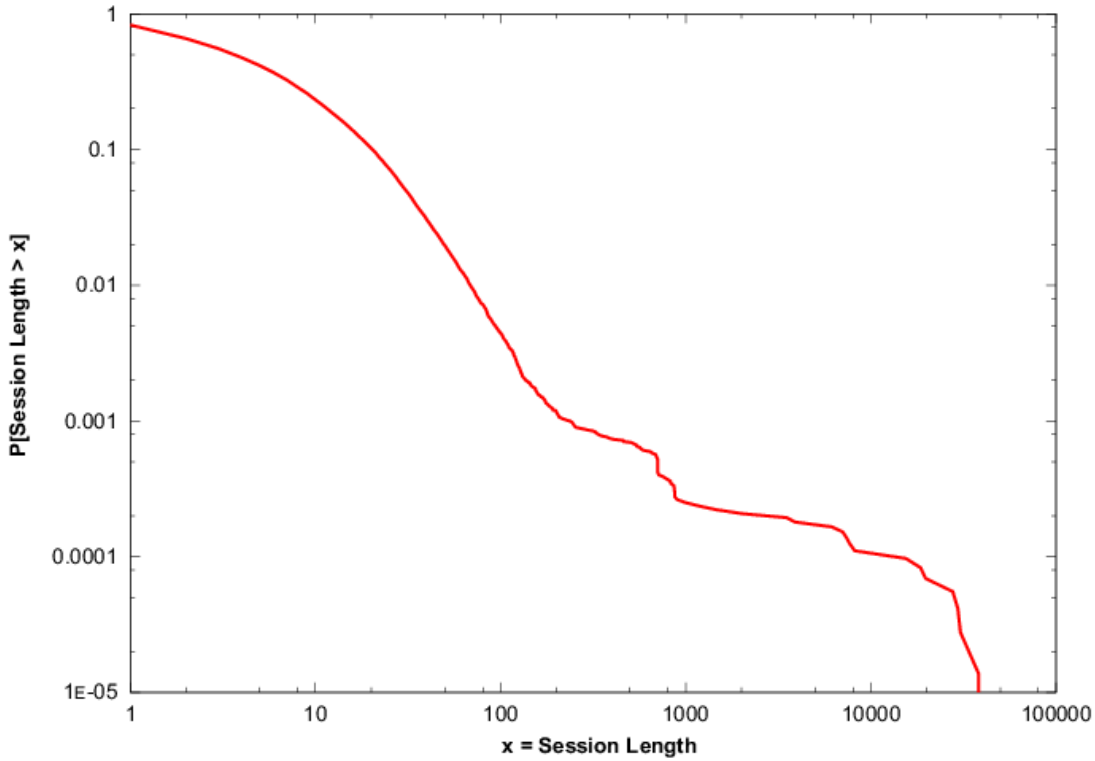
7. Execute experimentos
   - Quantos testes devem ser rodados? Quantas combinações dos parâmetros que formam o ambiente experimental?
   - Análise da sensibilidade dos outros parâmetros.
8. Analise e interprete os resultados
   - Use Estatistica para analisar a variabilidade, “outliers”, etc.

Processo Experimental Sistemático

7. Apresente adequadamente os resultados e dados do experimento
   - Gráficos: a questão da visualização dos resultados, distribuições estatísticas, etc.
8. Apresente conclusões
   - Para onde os resultados nos levam?
   - Quais os próximos passos
   - Novas hipóteses, novas questões, outros experimentos.

Session Layer
Session Length

- Sessions longer than 500 requests



Session Layer
Function Distribution

ShopBots

Function	Frequency	# Visits
Home	0.12%	36
Others	0.05%	14
Search	99.83%	30391

Crawlers

Function	Frequency	# Visits
View	38.30%	6221
Browse	36.50%	5926
Aux	13.60%	2208
Home	4.40%	716
Search	2.60%	415
Acc	2.20%	364
Add	2.10%	339
Others	0.21%	34
Robot	0.01%	1

Session Characterization: CBMG

	Entry	Exit	Home	Browse	Search	View	Add	Acc	Robo	Aux
Entry	0	0	33.333	33.333	0	33.333	0	0	0	0
Home	0	0.0196	16.606	24.29	0.031	8.794	30.332	4.061	0.016	15.844
Browse	0	0.006	8.088	68.817	0.098	2.783	0.053	0.002	0.006	20.146
Search	0	0	1.426	54.028	16.125	25.716	0	0	0	2.705
View	0	0.0020	1.240	4.182	0.783	92.601	0.124	0.015	0	1.052
Add	0	0	6.979	0	0	0	11.419	81.119	0	0.483
Acc	0	0	2.292	0.040	0.040	0	3.023	13.485	0	81.121
Robo	0	0	50.000	0	0	0	0	0	0	50.000
Aux	0	0	1.832	26.017	32.916	13.273	0.060	0	0	25.903

Table 6: Crawler CBMG.

Function	Function Distribution(%)	# of Visits
view	38.34	6221
browse	36.53	5926
aux	13.61	2208
home	4.41	716
search	2.56	415
acc	2.24	364
add	2.09	339
other	0.21	34
robot	0.01	1

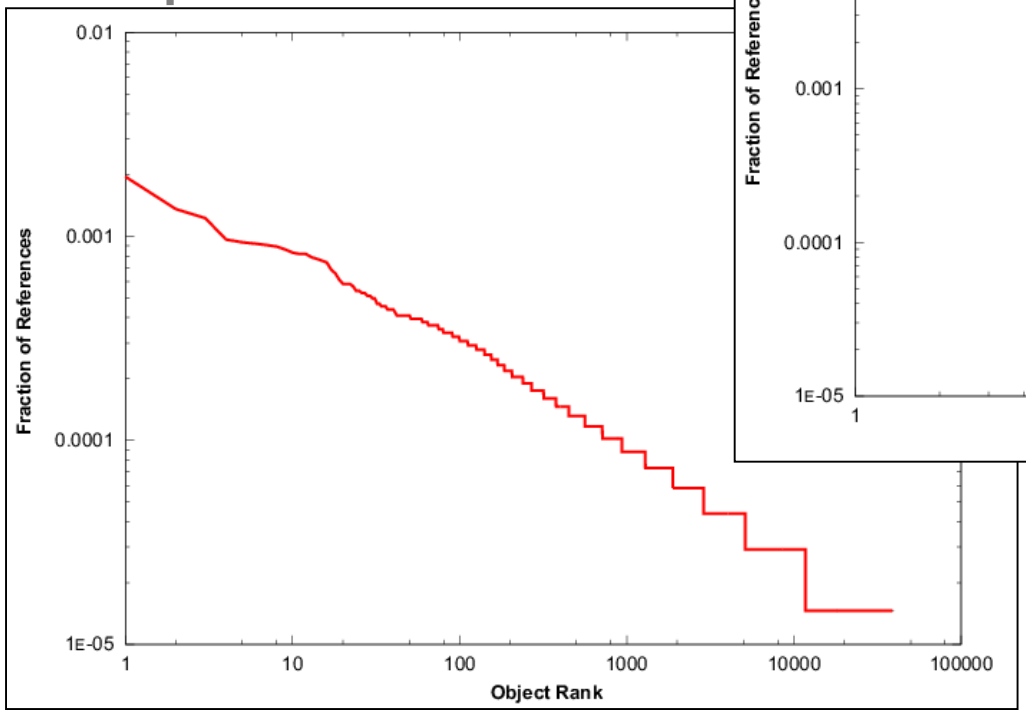
Table 7: Crawler Function Distribution and Visits

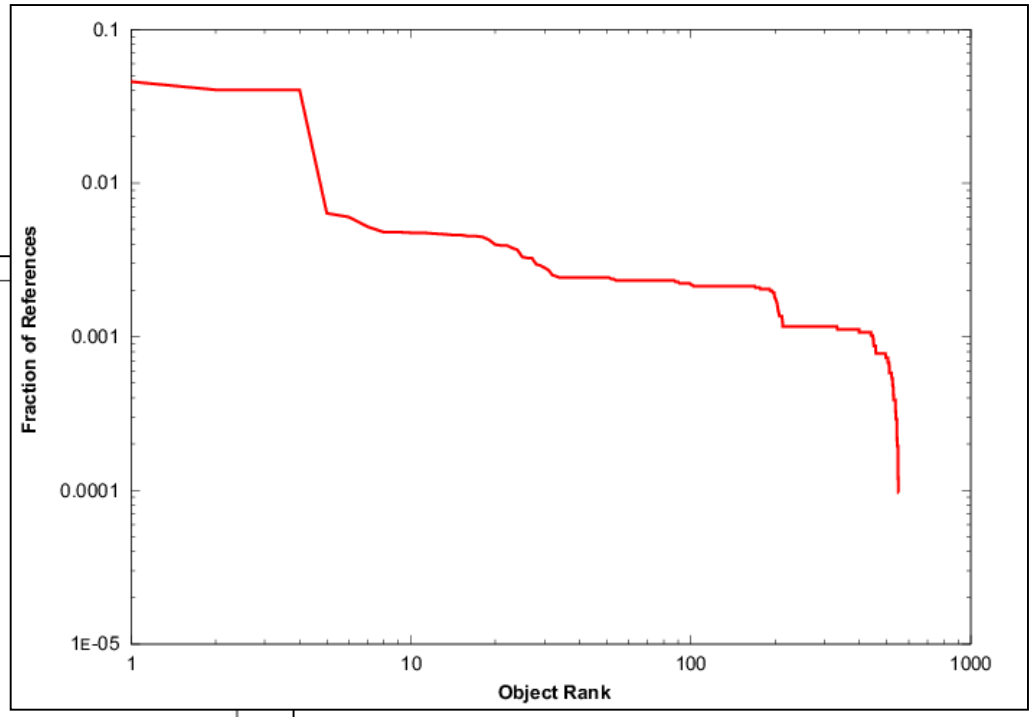
consider all objects to be equally sized. This mo
beit simplistic, can give us insight on the behavio
cache for the different streams. We use the margi
tribution of Least Recently Used (LRU) stack dista
determine the cache miss ratio [12, 14] under the LI
icy. If D is the random variable corresponding to th
distances and F_D is the cumulative distribution f
of D, then the miss ratio $m\ (x)$ for a cache of si
given by $P[D > x] = 1 - F_D(x) = m\ (x)$.

We start by noting the significant difference in the
totical miss ratio of the four streams, i.e., if the ca

Function Layer
Popularity of Objects

ShopBots

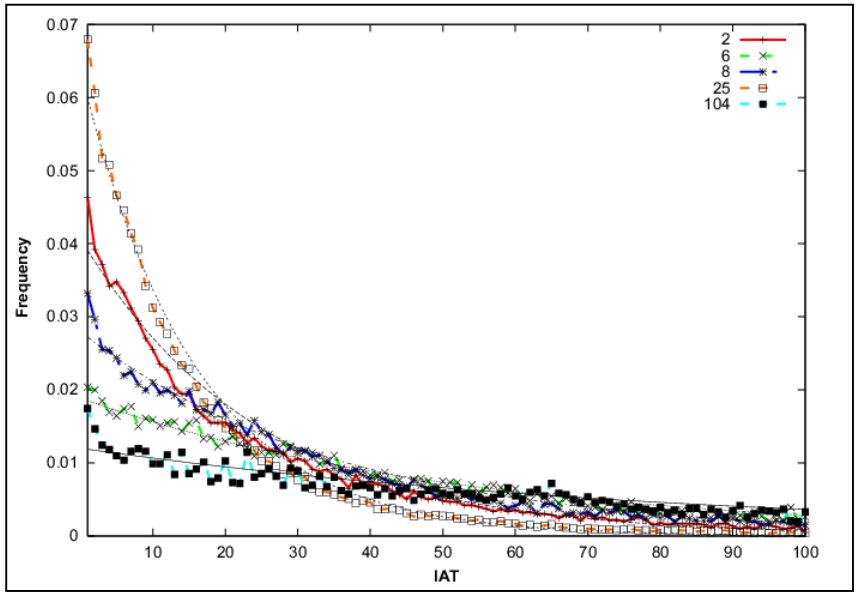




Crawlers

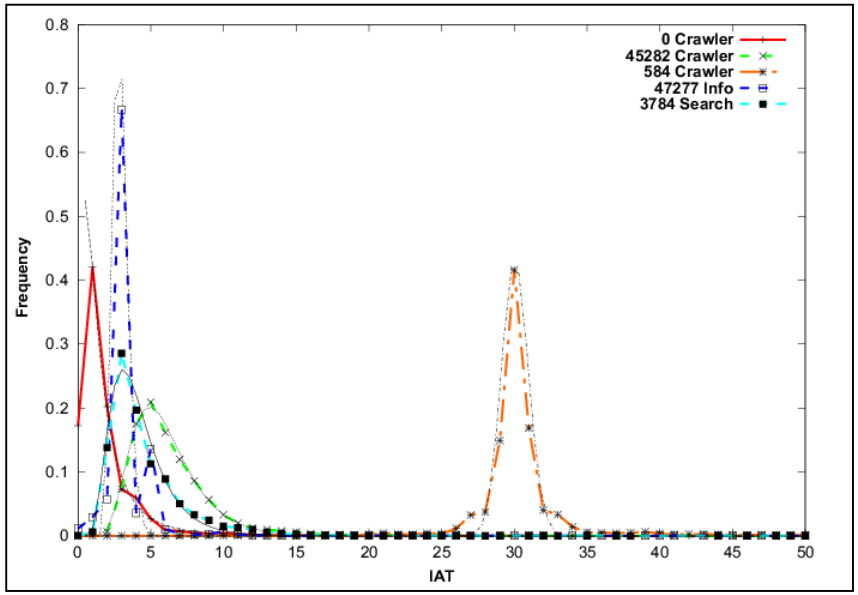
Request Layer
IAT Distribution

ShopBots



Approx. by an exponential-Poisson

Crawlers



Approx. by log-normal distributions

Request Layer
IAT Distribution

ShopBots				Crawlers				
ID	λ	mean	σ	ID	M	S	mean	σ
2	0.0407	24.57	24.57	3784	0.448	1.3280	3.78	8.31
6	0.0188	53.19	53.19	0	0.230	0.9120	1.90	2.17
8	0.0280	35.71	35.71	45282	1.733	0.3750	6.06	2.35
25	0.0639	15.64	15.64	584	3.403	0.0031	30.05	0.09
104	0.0120	83.33	83.33	47277	1.057	0.1764	2.92	0.51

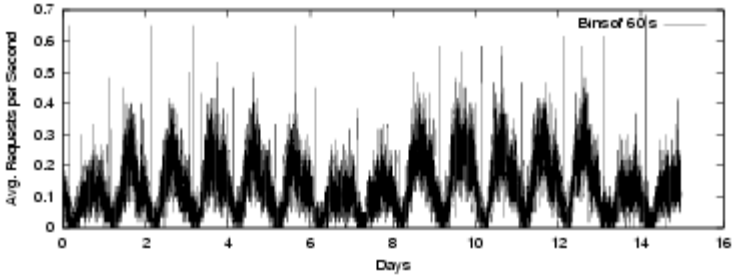
Table 3: Parameters for the fitted distribution of IAT's

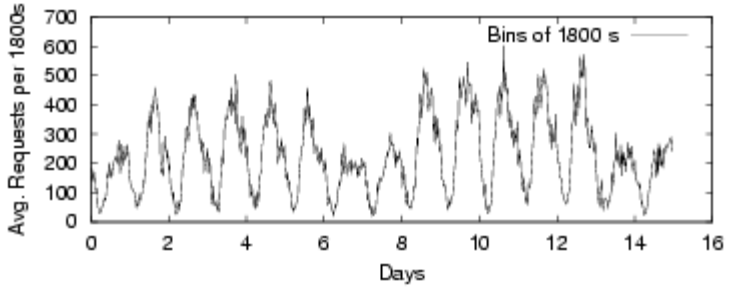
Approx. by an exponential-Poisson

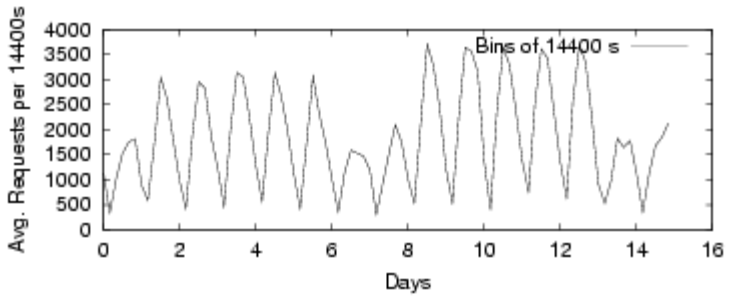
Approx. by log-normal distributions

Request Layer
Request Arrival Process

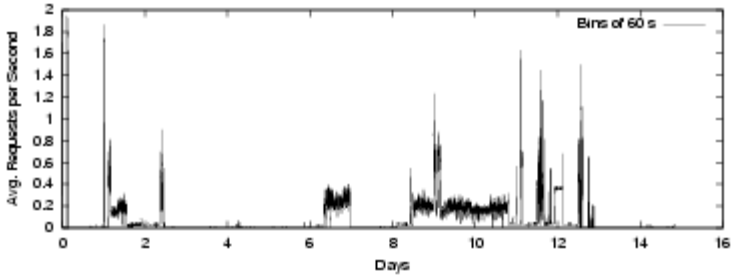
ShopBots

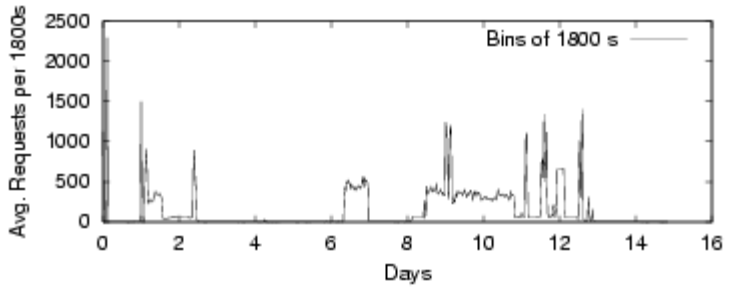


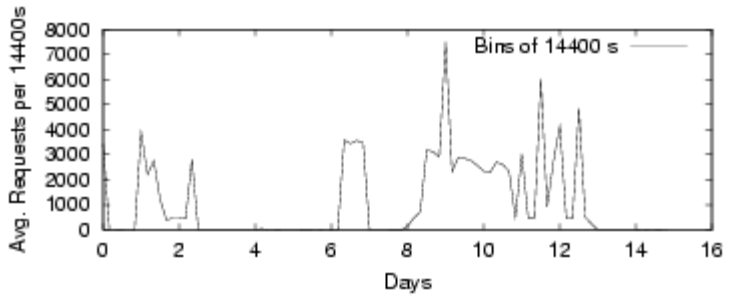




Crawlers







Processo Experimental Sistemático

7. Apresente adequadamente os resultados e dados do experimento
   - Gráficos: a questão da visualização dos resultados, distribuições estatísticas, etc.
8. Apresente conclusões
   - Para onde os resultados nos levam?
   - Quais os próximos passos
   - Novas hipóteses, novas questões, outros experimentos.

Evaluating Impact on Cache: a first cut

- Assumptions
- Server side caching system
- Used the Bookstore log
- Caching of ‘search’ and ‘info’ requests
  - Response to queries (possibly truncated)
  - Information records about books
- Uniformly sized objects
- LRU replacement policy
  - Simple
  - Baseline performance

LRU Stack Simulation

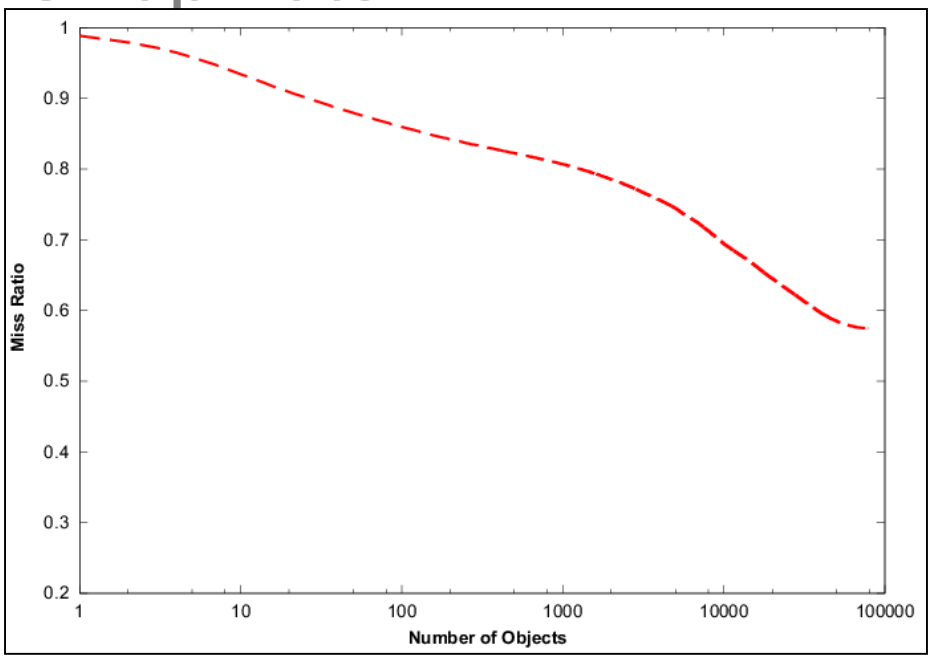
- Let S_t be the ordering by recency of reference of all objects seen up to time t
- The ***LRU stack distance*** of object o referenced at time t+1 is defined as
  - ∞ if it is the first reference to o
  - i if it is the i-th element on S_t
- Reference Stream → Distance Stream
- Miss Ratio
  - P[D>c] = 1 – CDF(D) = Miss ratio for cache of size c

Simulation Input

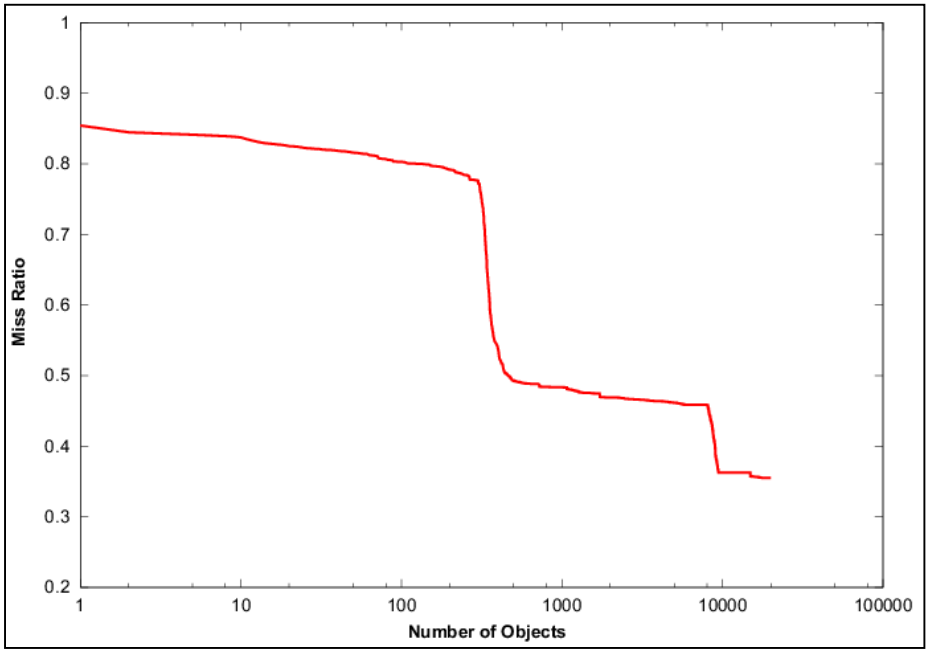
- Request streams
  - Full log (all requests)
  - Crawlers
  - Shopbots
  - No-robots

Miss Ratio x Cache Size
Robots

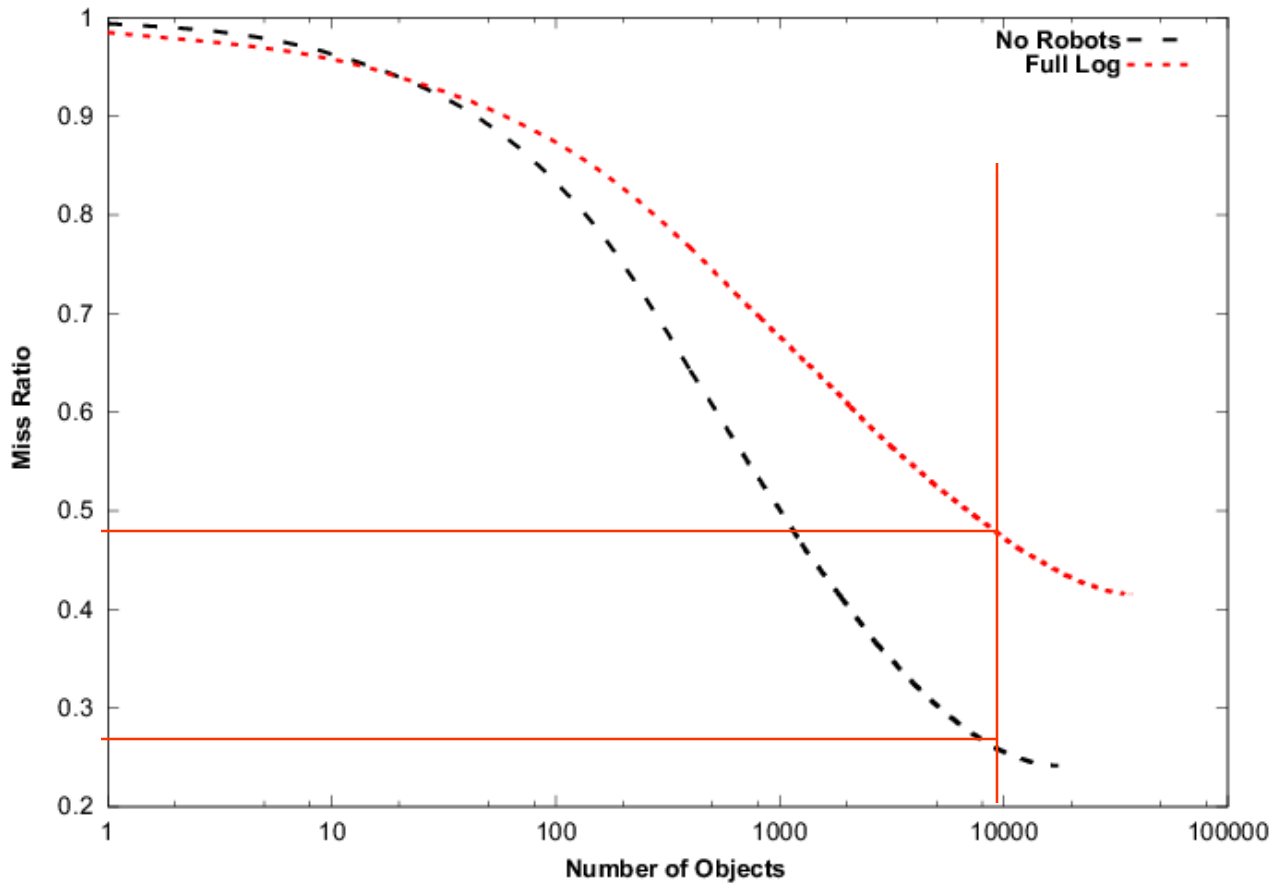
ShopBots



Crawlers



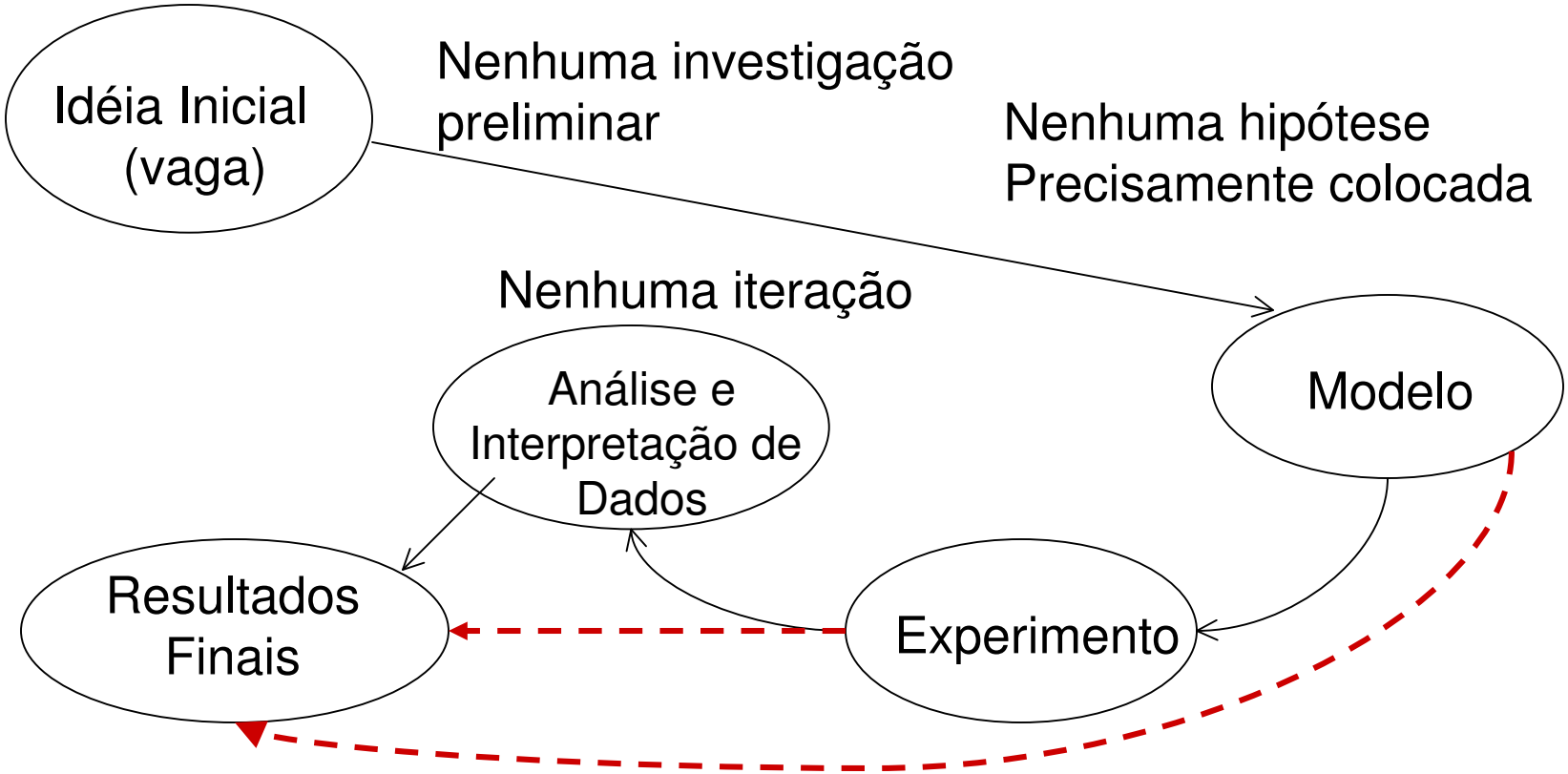
Miss Ratio x Cache Size
No-Robots log



Conclusions

- On the classification
  - Combination of several criteria necessary and effective
  - Multi-scale time analysis able to show short load peaks
- On the Impact of Robots
  - Significantly increase in miss ratio
  - Crawlers disrupt locality assumptions
  - Increase workload (globally and with bursts)

Prática Usual em Ciência da Computação



Próxima Aula

- Ler o artigo:
  - *Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?,* Bianca Schroeder and Garth A. Gibson (FAST 07)

- Voluntário para apresentar: o que entendeu, o que não entendeu e a ligação com os tópicos do curso