Reasoning Experiments

Bijan Parsia
<bijan.parsia@manchester.ac.uk>

If it has “science” in the name...

Tichy et al’s schema	1993 (Tichy et al)	2005 (Wainer et al)
Formal theory	12%	4%
Design and modeling (need eval)	70%	70%
Empirical work (all eval)	2%	17%
Hypothesis testing	2%	5%
Others	14%	3%

http://www.ipd.uka.de/Tichy/uploads/publikationen/156/1994-17.pdf
http://www.dcc.unicamp.br/~wainer/papers/empirical-acm.pdf

If it has “science” in the name...

Table 3
Ninety percentage confidence interval (using the adjusted Wald method) for the proportions (in percentages) for each class, for 1993 and 2005.

Class	1993	2005
Theory	6.1–21.8	2.0–7.8
Empirical	0–9.1	13.1–23.5
Hypothesis	0–9.1	2.4–8.7
Other	7.6–24.1	1.5–6.9
Design total	58.5–79.4	63.5–75.8
0%	20.5–41.5	17.9–29.3
0–10%	2.0–15.5	4.0–11.1
10–20%	6.1–21.8	10.7–20.5
20–50%	13.8–33.0	16.1–27.1
>50%	0–6.1	2.0–7.8

http://www.ipd.uka.de/Tichy/uploads/publikationen/156/1994-17.pdf
http://www.dcc.unicamp.br/~wainer/papers/empirical-acm.pdf

90% confidence interval

....it's (probably) not

- Three (or more!) paradigms
  - Math! (Prove a theorem)
  - Science! (Experiment! Prove a hypothesis)
  - Engineering! (Build something. It doesn't fall down.)
- The latter two require some empirical work!
  - And we're not doing it
    - Often scorning it
  - Lots of quality problems
    - Even in publishing
    - Reproducibility
    - Validity

What do I care about?

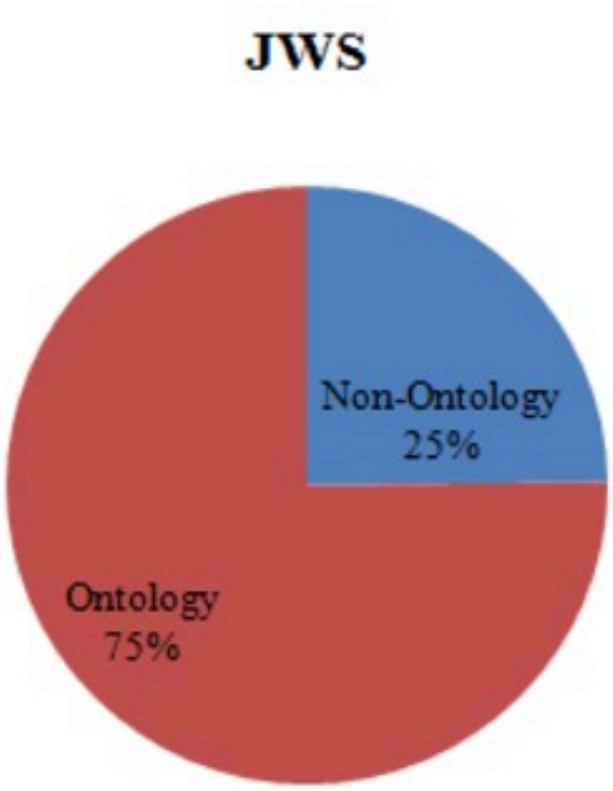
- Ontology engineering!
  - The construction, maintenance, and exploitation of computational artifacts that encode a cognitive model of a domain (typically, the conceptual aspects)
    - Y'know, classes and properties and stuff
- Specifically
  - What formalisms are suitable
    - I.e., are representationally, computationally, and cognitively adequate
      - I don't strive for excellence. Adequacy would suffice.
  - What are reasonable methodologies?
  - What are the useful and interesting services?
    - And how do we realize them
  - What's the ROI, ceterius paribus, of ontologies?
    - Not just ceterius paribus, but all things considered too
    - Time to develop? Total cost of ownership? Error rates?
- This clearly requires a lot of a lot of empirical work

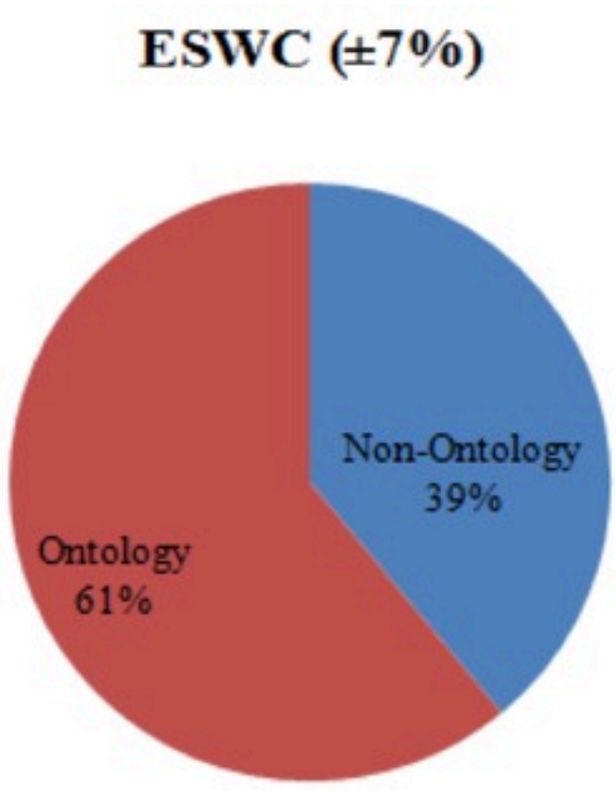
People agree!

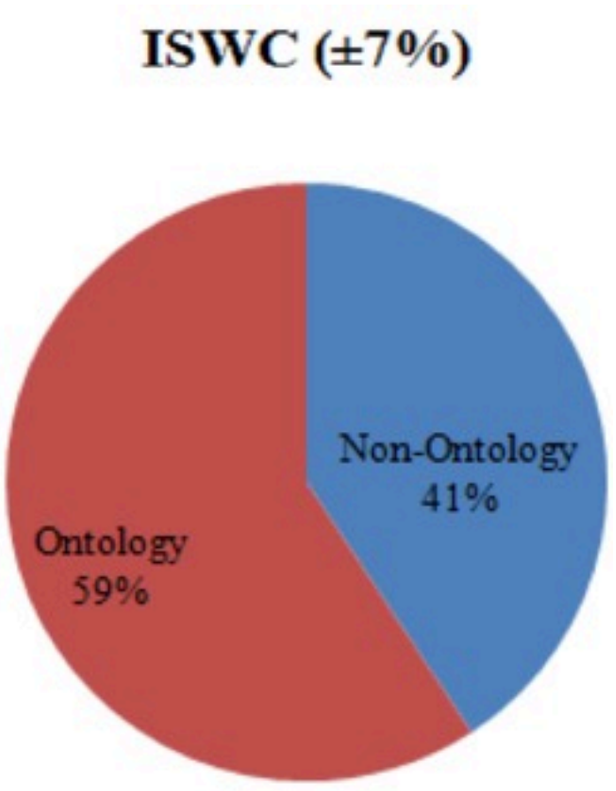
- Systematic review of Empirical Ontology Engineering
  - 3rd year students; second year running
  - All of JWS
  - Year stratified sample of ISWC and ESWC
- How are we doing?

Ontology papers!

- We rule
  - (Fairly generous notion here)



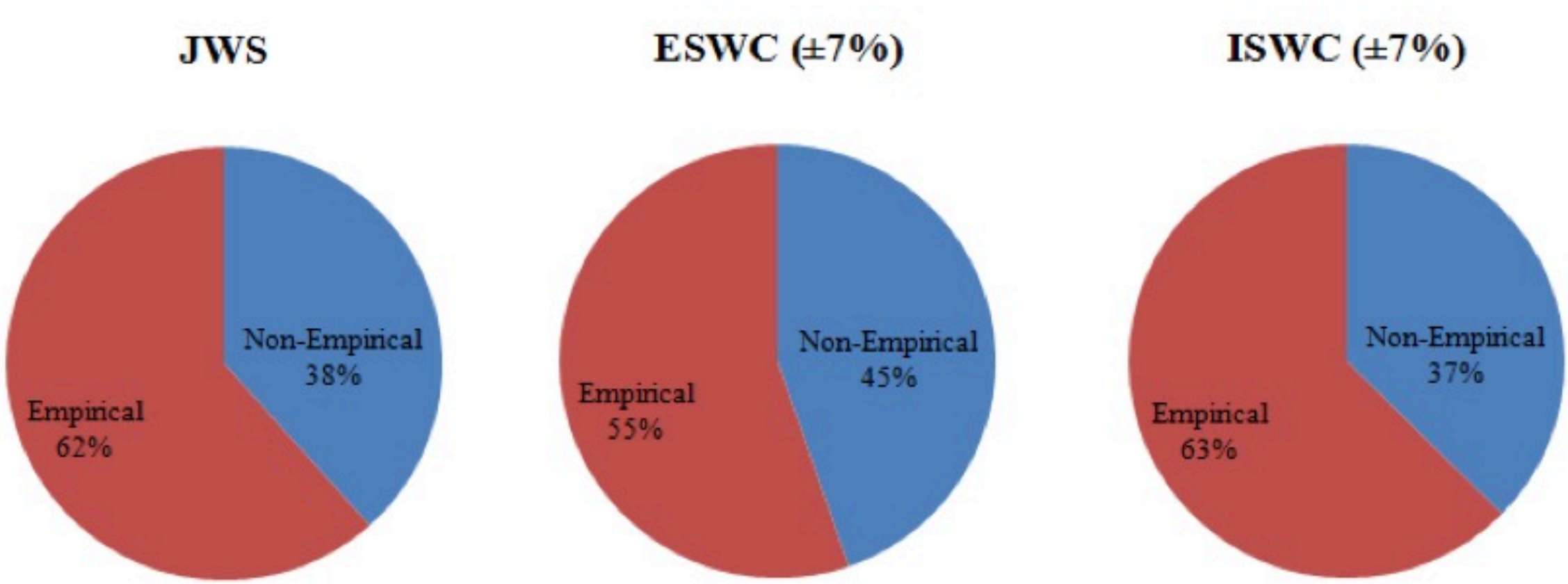




Thanks Stephan Hall

Empirical Ontology Papers

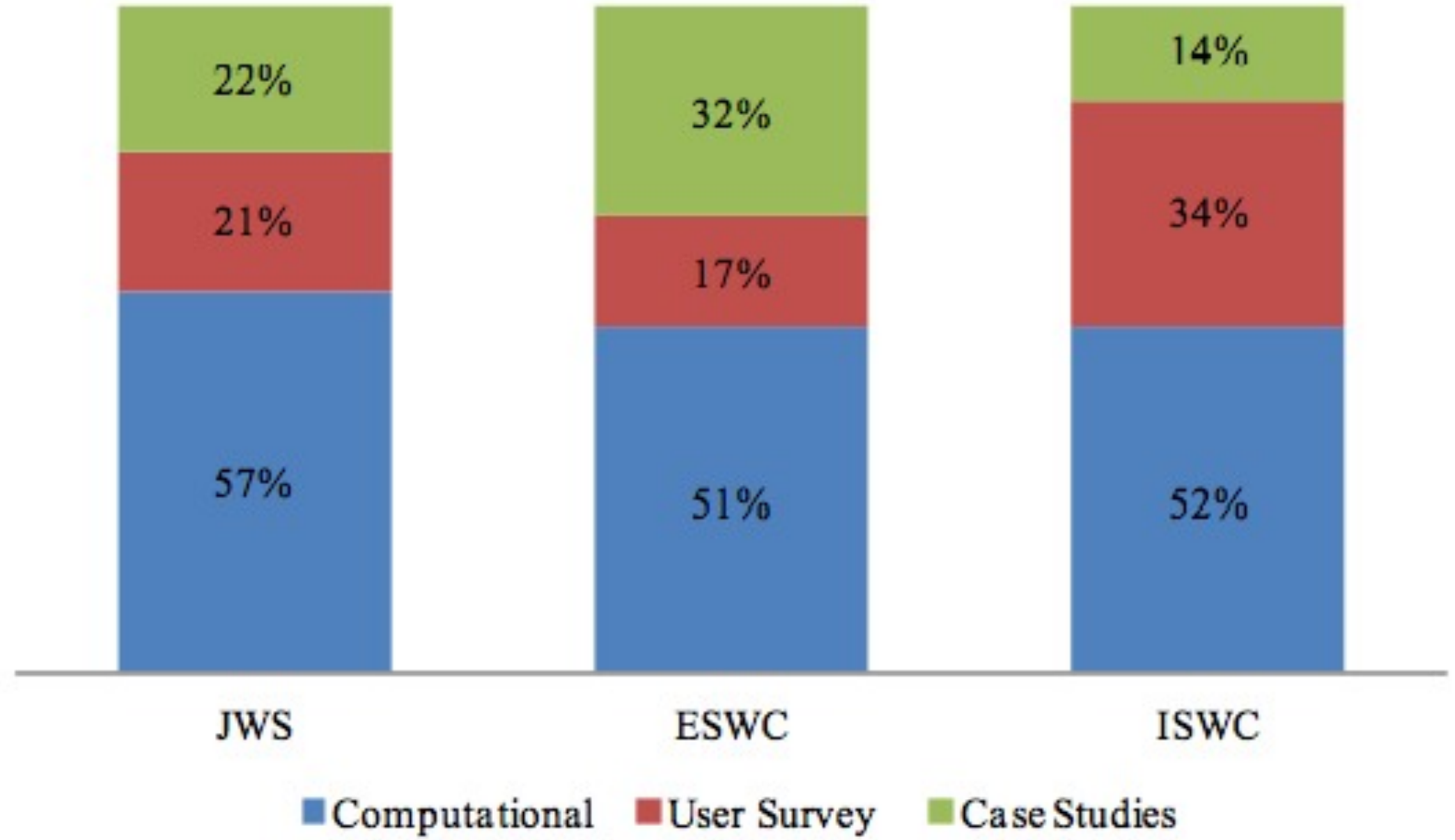
- Also good
  - (Unfortunately, don't have design vs hypothesis etc.)
  - (Most are design)



Thanks Stephan Hall

What kind of Empirical Work?

- Promising!



Thanks Stephan Hall

Quality

- Well, we have some indirect metrics
- Machine setup
  - 23% stated processor OR memory
  - 10% stated both
  - 67% stated neither
  - (33% is lower than the % of performance tests!)
- Ontologies used
  - 31% mentioned what they used AT ALL
    - Mostly "number of"
    - Occasionally vague
      - "about 100"
    - Most of these didn't name names

Thanks Stephan Hall

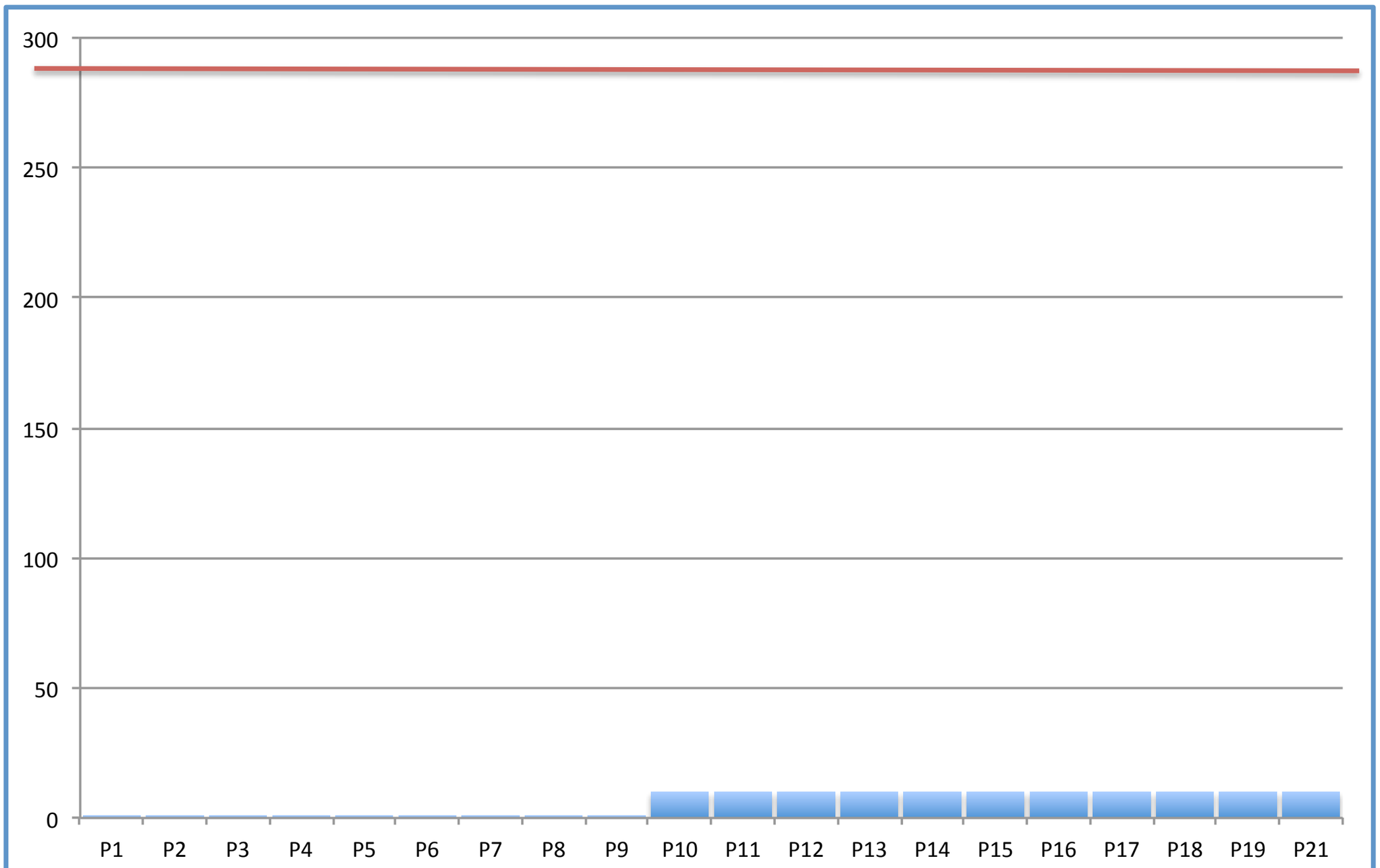
Practical Classification

A Question

- Is $\mathcal{ALC}$ (or beyond!) Classification “Practical”?
  - Quadratic + NEXPTIME = AIEEE?
  - How good are the optimisations?
  - Recall from last night....

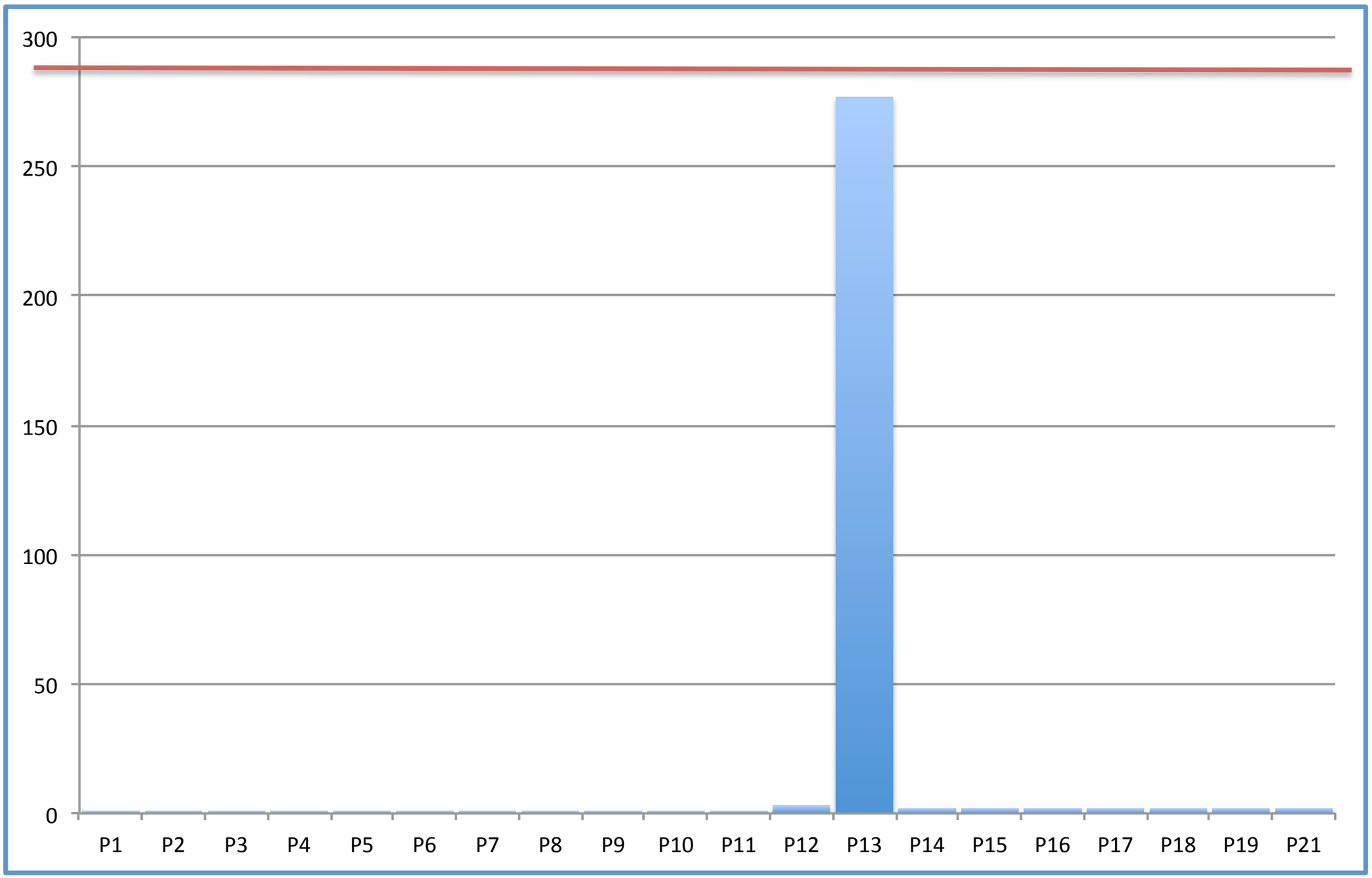
Worst Case Complexity

- if we know that *RTime(x ≤ 7)* ≤ *285*, do we expect



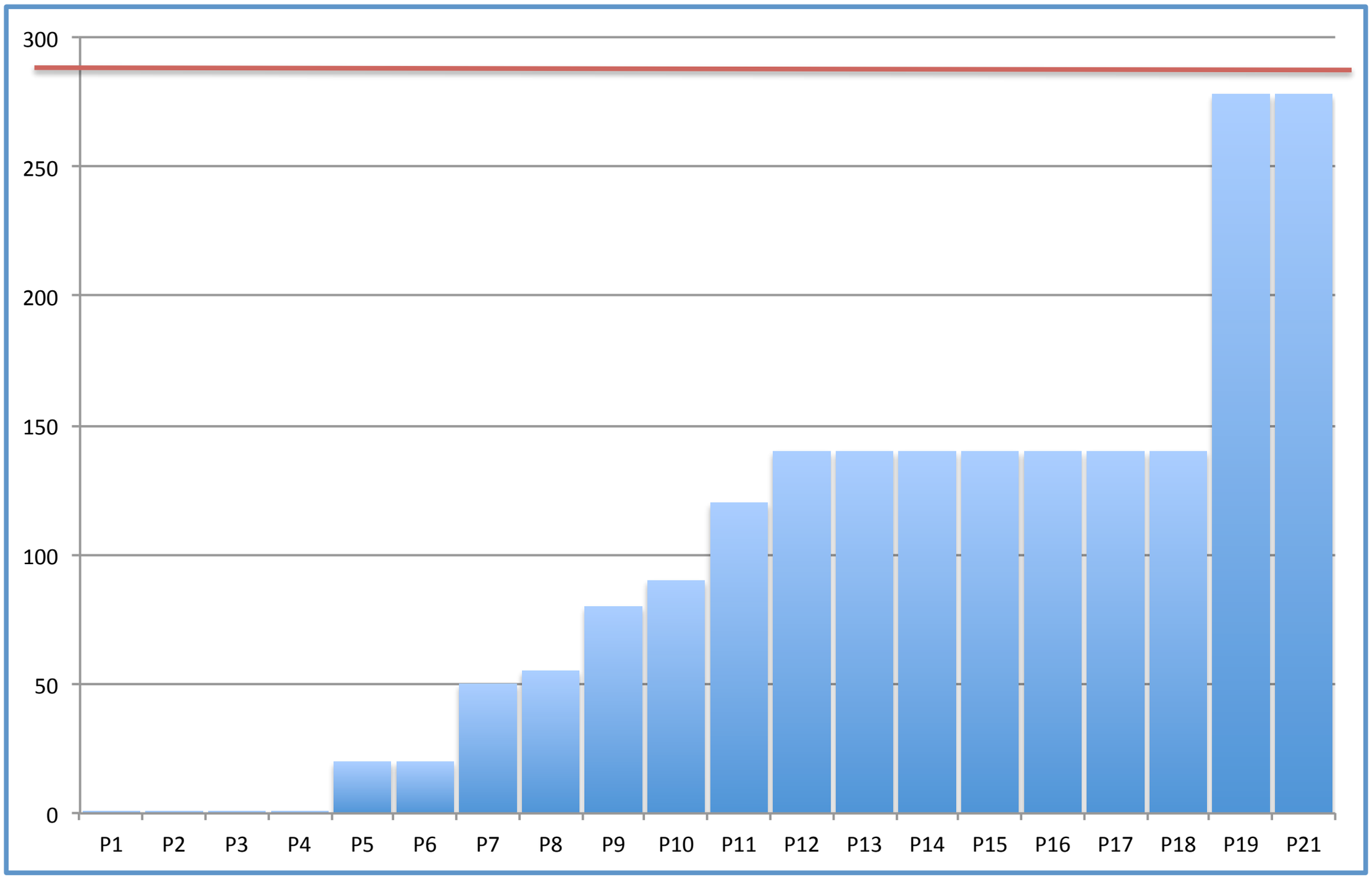
Worst Case Complexity

- if we know that *RTime(x ≤ 7)* ≤ *285*, do we expect



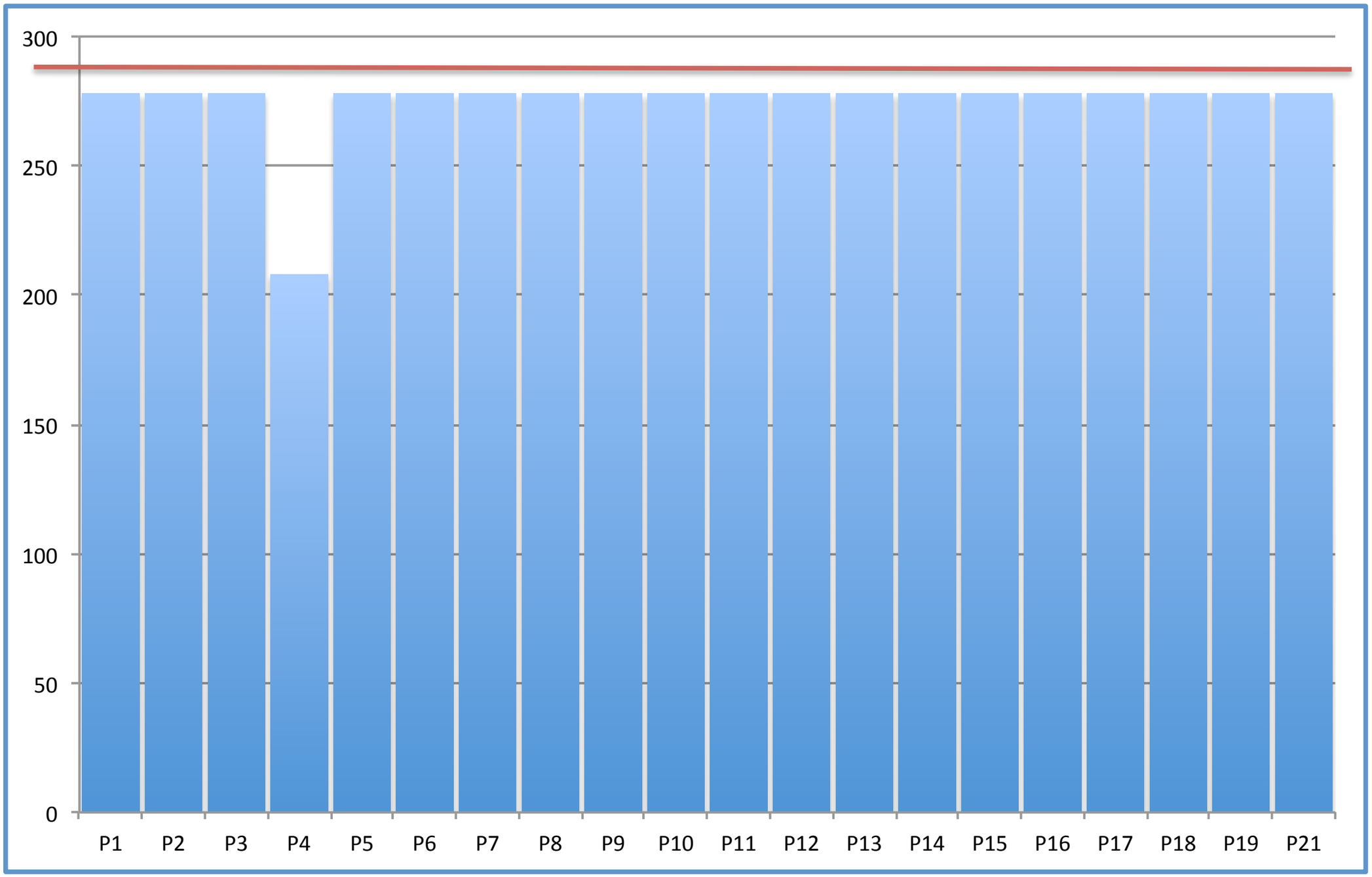
Worst Case Complexity

- if we know that *RTime(x ≤ 7) ≤ 285*, do we expect



Worst Case Complexity

- if we know that *RTime(x ≤ 7)* ≤ *285*, do we expect



Practical? What's the theory?

- We focus on "robustness"
  - Intuition: resistant to failure in the face of a range of input
    - What range of input?
    - Which properties (either functional or non-functional)?
    - What counts as failure?
- Key scenario
  - An ontology engineer downloads an OWL ontology from the Web and wants to evaluate it for reuse. They classify the ontology to check the class hierarchy.
- Robustness concretified
  - Input: ontologies from the Web
  - Properties of interest
    - Functional: process the ontologies; classification correctness
    - Non-functional: Time (primarily)
  - Failure: Fails to classify in 2 hours
    - Or <100 seconds for the impatient

Materials & Methods

- 3 distinct corpora
  - NCIt (through 12.11d)
  - BioPortal snapshot (2012)
  - Our own Web Crawl (sample)
  - Download! https://sites.google.com/site/reasonerbenchmark/
- 2 versions of 4 reasoners
  - 2011 (Pellet v2.2.2, HermiT v1.3.3, FaCT++ v1.5.3 and JFact v0.2)
  - 2013 (Pellet v2.3.0, HermiT v1.3.6, FaCT++ v1.6.1 and JFact v1.0)
- 2 machines
  - Intel Quad-Core Xeon 3.2GHz processor;32GB DDR3 RAM.
  - NCIt test: Intel Dual-Core i7 2.7GHz processor; 16GB DDR3 RAM
    - Reusing past results!
  - Mac OS X 10.7.5; Java v1.7; OWL API v3.4.1

Robustness operationalised

- A reasoner is
  - robust for a corpus if successful for 90% of the corpus
  - strongly robust if successful for 95%
  - extremely robust if successful for 99%
- Our ex ante predications for current reasoners:
  - NCIt
    - Current reasoners are strongly to extremely robust; 1 is extremely
    - Reasoners (aside from JFaCT) didn't change a lot between 2011-2013
    - Best union reasoner is extremely robust
  - Bioportal
    - 2 reasoners are robust; one reasoner may be extremely robust
    - 1 might not be
    - Best union reasoner is very robust
  - Web crawl
    - At least 1 reasoner is robust (maybe)
    - Best union reasoner is robust

Quick Corpora Discussion: NCIt

- Centralized authoring; monthly releases; used experimentally
- Some optimisations solely or prominently for NCIt
  - "For example, the deterministic treatment of GCIs significantly reduces the classification time for the NCI ontology." --Boris
  - "Using CD speeds up the classification of NCI by a factor of more than 20. In both cases, all subsumption tests are solved cheaply using cached models, but more than ten million tests are performed when CD is not used; employing CD reduces this number to less than one million." --Dmitry
- They use reasoners
  - FaCT++ and now Pellet; They have paid for development
- Willing (but not happy) to wait overnight
- 106 versions parseable by the OWL API
  - 02.00 (October 2003) to 12.11d (November 2012)
  - Size: from 49,475 to 133,900 logical axioms
  - Expressivity: from ALE to SH(D)

Results

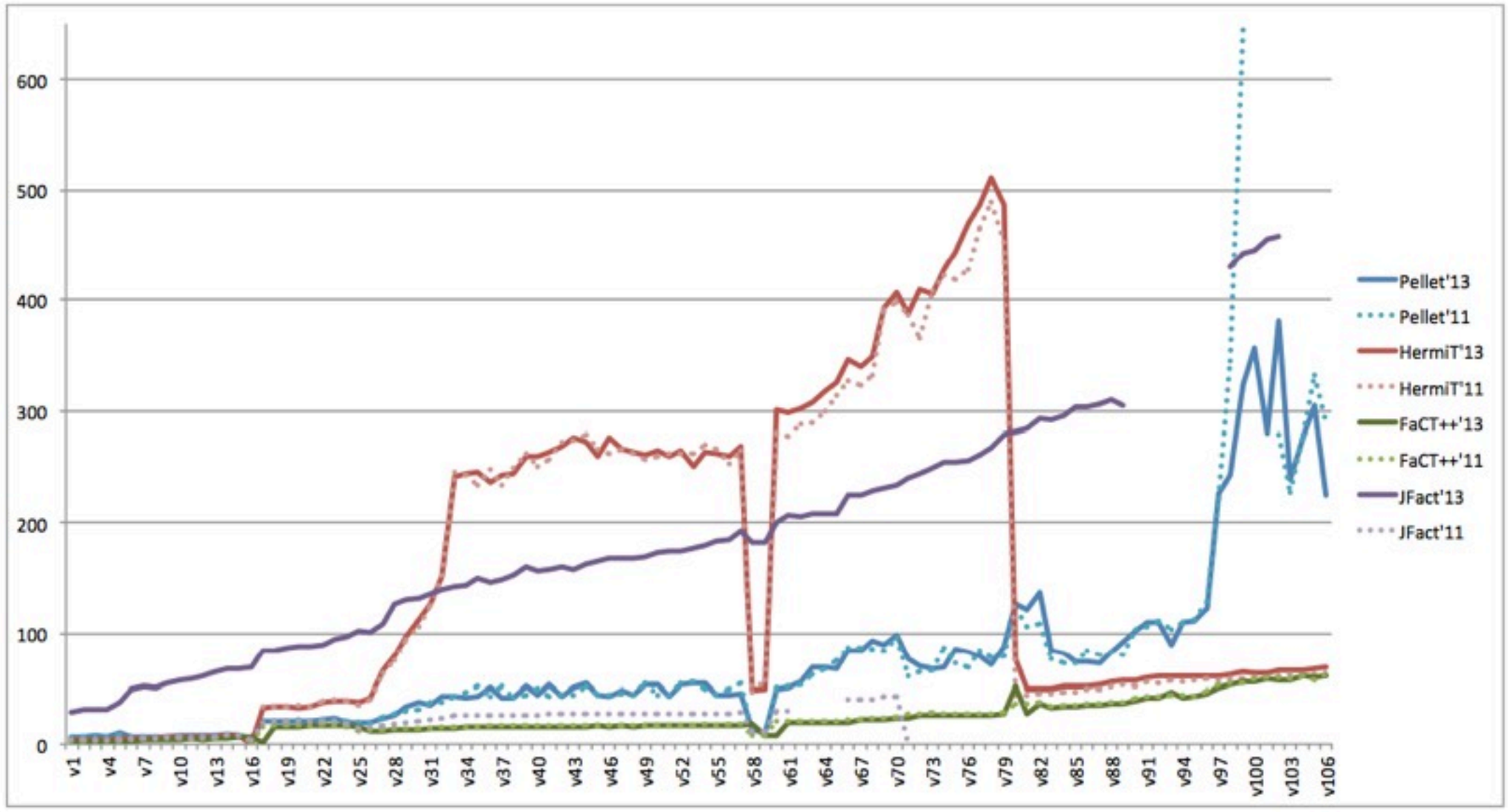


Fig. 1: Comparison of classification times between the 2011 reasoner version set (suffixed '11) and the 2013 set (suffixed '13) over the NCIt corpus (*y*-axis: time in seconds, *x*-axis: version number).

Performance Binning

- Very Easy (≤ 1 second)
- Easy (1-10 seconds)
- Medium (10-100 seconds)
- Hard (100- 1000 seconds)
- Very Hard (>1000 seconds but under 2 hrs (7200 sec))
- Bins give a better sense of the “weight” of the times

2011 NCIt results

	Pellet	HermiT	JFact	FaCT++	Best Combo	Worst Combo
Very Easy	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Easy	16 (15%)	15 (14%)	16 (15%)	18 (17%)	18 (17%)	15 (14%)
Medium	70 (66%)	42 (40%)	52 (49%)	88 (83%)	88 (83%)	15 (14%)
Hard	18 (17%)	48 (45%)	0 (0%)	0 (0%)	0 (0%)	37 (35%)
Very Hard	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Timeout	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Errors	2 (2%)	1 (1%)	38 (36%)	0 (0%)	0 (0%)	39 (37%)
Impatient Robustness	81%	54%	64%	100%	100%	28%
Overall Robustness	98%	99%	64%	100%	100%	63%

Table 1: Binning of the NCIt corpus according to performance, using the 2011 reasoner versions set.

2013 NCIt results

	Pellet	HermiT	JFact	FaCT++	Best Combo	Worst Combo
Very Easy	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Easy	16 (15%)	15 (14%)	0 (0%)	19 (18%)	19 (18%)	0 (0%)
Medium	71 (67%)	42 (40%)	24 (23%)	87 (82%)	87 (82%)	23 (22%)
Hard	19 (18%)	48 (45%)	70 (66%)	0 (0%)	0 (0%)	70 (66%)
Very Hard	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Timeout	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Errors	0 (0%)	1 (1%)	12 (11%)	0 (0%)	0 (0%)	13 (12%)
Impatient Robustness	82%	54%	23%	100%	100%	22%
Overall Robustness	100%	99%	89%	100%	100%	88%

Table 2: Binning of the NCIt corpus according to performance, using the 2013 reasoner versions set.

3 out of 3 predictions!
JFaCT progress is dramatic

Remember the graph

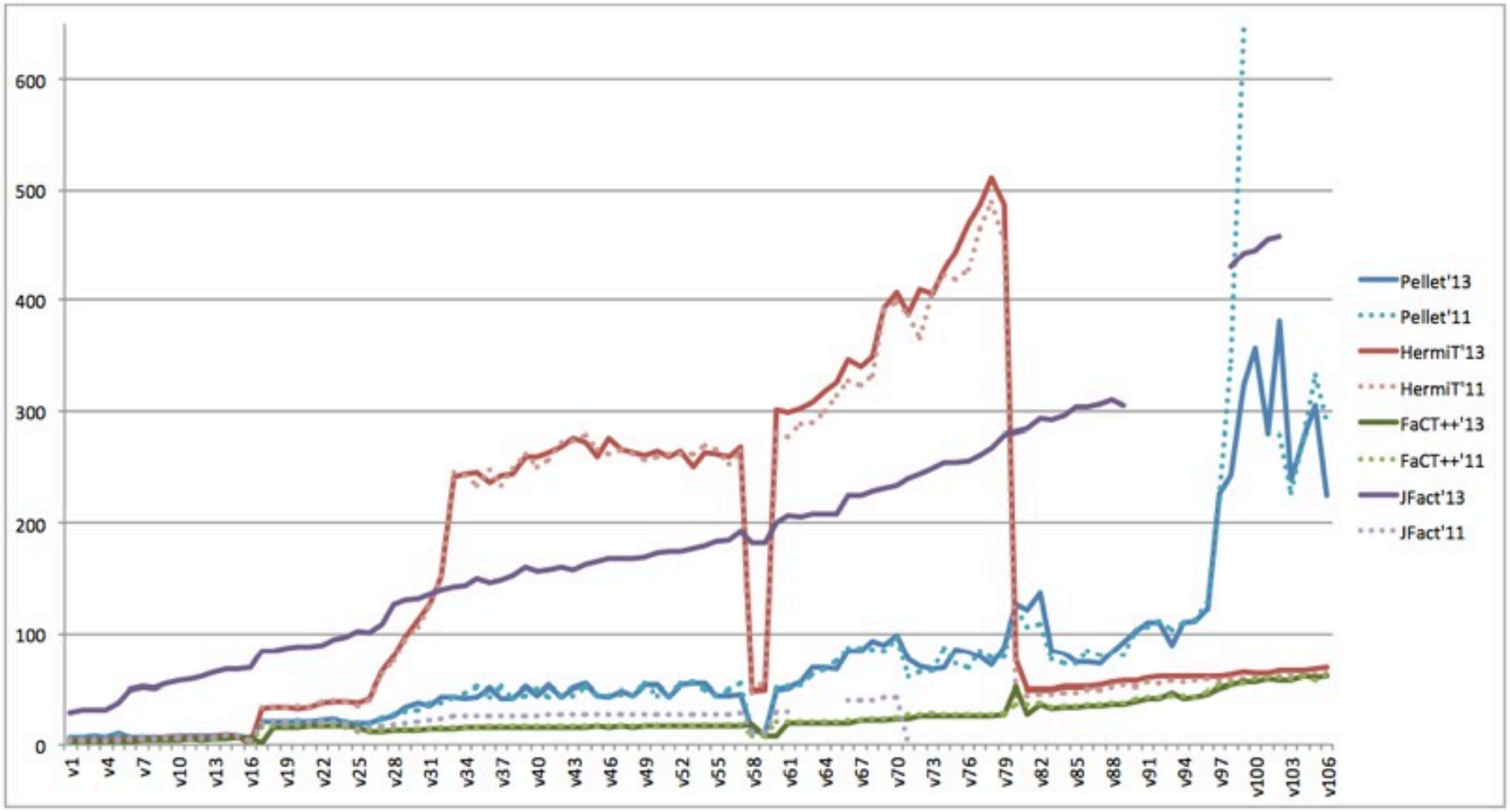


Fig. 1: Comparison of classification times between the 2011 reasoner version set (suffixed '11) and the 2013 set (suffixed '13) over the NCIt corpus (y-axis: time in seconds, x-axis: version number).

Corpora Discussion: Bioportal

- Biomedically oriented
- Community driven
  - People submit to Bioportal
  - Lots of versions
  - Lots of diverse ontologies and applications
- Our go-to corpus
  - Independent
  - Unedited
  - Updating
  - Rich
  - A community that cares about our services
- Gathered November 2012
  - 292 OWL and OBO parseable ontologies.
  - Average size: 28,439 (total: 8,190,504 and median: 979 axioms)
  - 89 of these ontologies contain named individuals
  - 4 ontologies with no logical axioms (discarded)
  - Expressivity: from AL to SROIQ.

Bioportal 2012 Results

	Pellet	HermiT	JFact	FaCT++	Best Combo	Worst Combo
Very Easy	190 (66%)	170 (59%)	184 (64%)	218 (76%)	236 (82%)	152 (53%)
Easy	56 (19%)	61 (21%)	58 (20%)	24 (8%)	28 (10%)	58 (20%)
Medium	10 (3%)	15 (5%)	8 (3%)	7 (2%)	11 (4%)	10 (3%)
Hard	4 (1%)	4 (1%)	2 (1%)	2 (1%)	4 (1%)	2 (1%)
Very Hard	6 (2%)	3 (1%)	0 (0%)	3 (1%)	4 (1%)	2 (1%)
Timeout	13 (5%)	8 (3%)	11 (4%)	10 (3%)	5 (2%)	15 (5%)
Errors	9 (3%)	27 (9%)	25 (9%)	24 (8%)	0 (0%)	49 (17%)
Impatient Robustness	89%	85%	87%	86%	95%	76%
Overall Robustness	92%	88%	88%	88%	98%	78%

Table 4: Binning of the BioPortal corpus according to performance.

- 288 ontologies; 234 processed by all reasoners
  - Inside avg: FaCT++ (2.9s), JFact (5.9s), HermiT (9.8s), Pellet (16.7s)
  - What?!
    - Ooo, timeouts!
- Bioportal is tough!
  - 2.5 predictions correct!

Bioportal Errors

Error	Pellet	HermiT	JFact	FaCT++
StackOverflow	2	0	1	0
OutOfMemory	1	1	2	0
UnsupportedDatatype	0	13	4	14
InternalReasoner	2	0	1	0
IllegalArgument	0	12	16	6
MalformedLiteral	0	1	0	0
ConcurrentModification	3	0	0	0
Reasoner crashed	0	0	0	4
IndexOutOfBounds	1	0	1	0
Total Errors	9	27	25	24

Corpora Discussion: Crawl

- Gathered from the Web
  - Our own crawler
  - A “short” run
  - Seeded from Swoogle, Google, and various repos
  - Somewhat curated (see next case study!)
- Arbitrary subjects, origins, and uses
- Filtered for “OWL ontologies”
- A sample of 822 ontologies (from around 4.5K)
  - 145 with no logical axioms (discarded), leaving 677 ontologies
    Size: average 2,405 (total: 1,628,207 and median: 57)
  - Expressivity: AL to SHOIQ (and an SRI)

Web Crawl

	Pellet	HermiT	JFact	FaCT++	Best Combo	Worst Combo
Very Easy	597 (88%)	536 (79%)	557 (82%)	566 (84%)	642 (95%)	493 (73%)
Easy	44 (6%)	36 (5%)	45 (7%)	12 (2%)	26 (4%)	44 (6%)
Medium	2 (0%)	8 (1%)	11 (2%)	0 (0%)	3 (0%)	12 (2%)
Hard	1 (0%)	1 (0%)	4 (1%)	5 (1%)	2 (0%)	3 (0%)
Very Hard	0 (0%)	1 (0%)	1 (0%)	1 (0%)	0 (0%)	1 (0%)
Timeout	16 (2%)	6 (1%)	5 (1%)	5 (1%)	4 (1%)	10 (1%)
Reasoner Errors	17 (3%)	89 (13%)	54 (8%)	88 (13%)	0 (0%)	114 (17%)
Impatient Robustness	95%	86%	91%	85%	99%	81%
Overall Robustness	95%	86%	91%	86%	99%	82%

Table 6: Binning of the Web crawl corpus according to performance.

- 560/677 completed by all
  - Pellet (avg 0.5s), FaCT++ (1.5s), HermiT (3.1), JFact (6.2s)
  - Pellet win due to JNI? (Dmitry likes to think so!)
- Both (weak) predictions correct!

Web Crawl Errors

Error	Pellet	HermiT	JFact	FaCT++
StackOverflow	13	0	0	0
OutOfMemory	2	0	2	0
NullPointer	0	0	36	0
UnloadableImport	0	1	1	1
ClassCast	0	0	1	0
UnsupportedDatatype	0	81	1	86
Datatype constraint	2	0	0	0
IllegalArgument	0	3	5	0
MalformedLiteral	0	2	0	0
ReasonerInternal	0	0	8	1
UnsupportedFacet	0	2	0	0
Total	17	89	54	88

Over all Corpora

	Pellet	HermiT	JFact	FaCT++	Best Combo	Worst Combo
Very Easy	787 (73%)	706 (66%)	741 (69%)	784 (73%)	878 (82%)	645 (60.2%)
Easy	116 (11%)	112 (10%)	103 (10%)	55 (5%)	73 (7%)	102 (9.5%)
Medium	83 (8%)	65 (6%)	43 (4%)	94 (9%)	101 (9%)	45 (4.2%)
Hard	24 (2%)	53 (5%)	76 (7%)	7 (1%)	6 (1%)	75 (7.0%)
Very Hard	6 (1%)	4 (0%)	1 (0%)	4 (0%)	4 (0%)	3 (0.3%)
Timeout	29 (3%)	14 (1%)	16 (1%)	15 (1%)	9 (1%)	25 (2.3%)
Errors	26 (2%)	117 (11%)	91 (8%)	112 (10%)	0 (0%)	176 (16.4%)
Total (excl. Errors)	1016	940	964	944	1062	870
Total (incl. Errors)	1071	1071	1071	1071	1071	1071
Impatient Robustness	92%	82% (90%)	83%	87% (96%)	98%	74% (87%)
Overall Robustness	95%	88% (96%)	90%	88% (97%)	99%	81% (96%)

Table 8: Binning of all three corpora: BioPortal, NCIt (2013), and Web crawl. Under robustness rows, bracketed values indicate robustness w.r.t. OWL 2 datatype map.

Maybe Easy Corpora?

- Real complaint:
  - Maybe everything is EL so easy.
  - Wimps!
- Remember our scenario
  - This would be striking if true!
- But it's not true
  - Nor are the PTime Logics always easy

Threats to Validity (Limitations)

- Internal Validity
  - Lots of issues
    - Time outs alone
    - Expressivity is nonsense
      - Some bits we know how to do better, others not
    - Bugs and implementation obscure algorithms
    - Do you have all day?
- External Validity
  - The biggest one I care about is unpublished ontologies
    - Are the hard ones hidden?
  - Change over time
  - Tool effects
    - Maybe the reasoners aren't getting better but people are getting better at yielding to their foibles
  - Other reasoning problems
    - Classification isn't the worst proxy

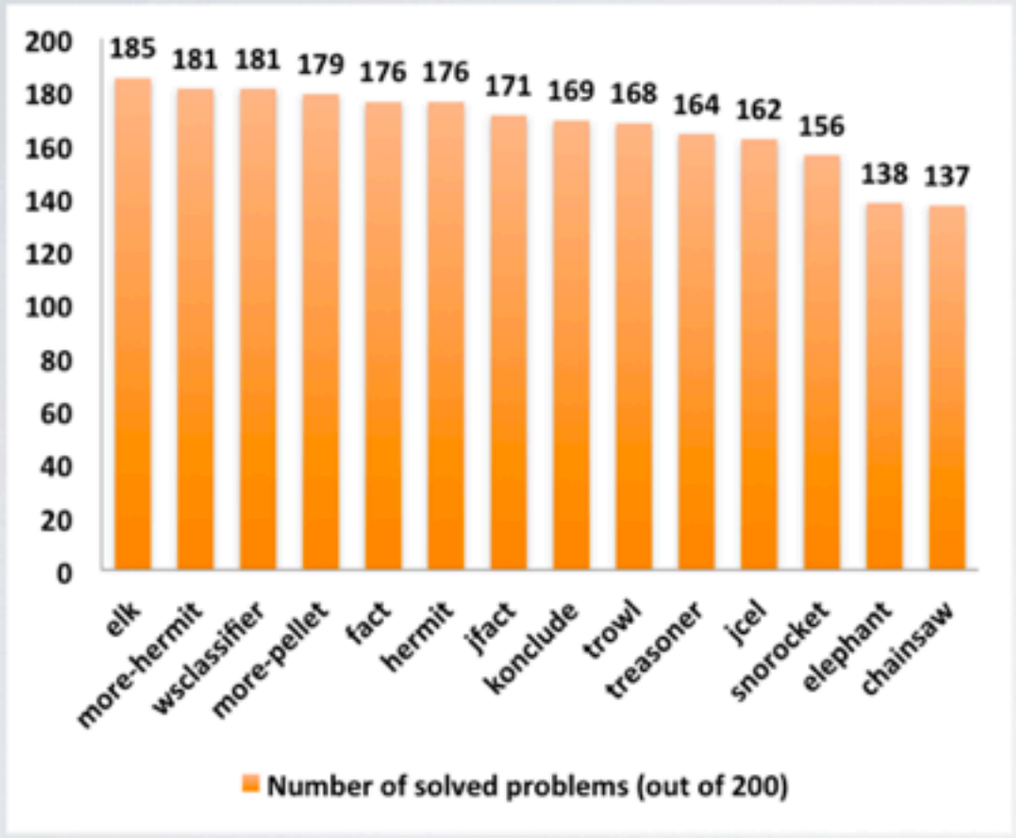
Problems with the Analysis

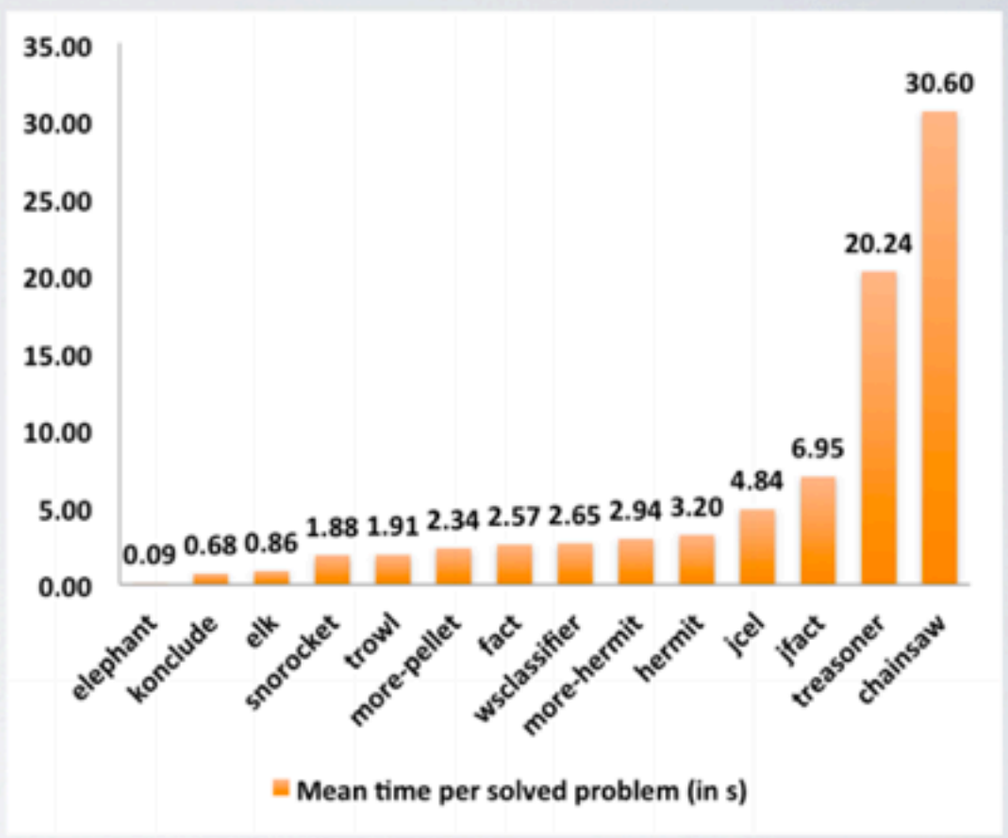
- We need to break out more factors
  - Size, axiom size, axiom complexity
  - Signature issues, modular structure
  - Some sort of normalization
    - To account for time outs and errors
- Data gathering missed a lot
  - No telemetry
    - SAT hardness or avoided tests?
  - Still never quite sure what HermiT is doing
    - We suspect more than it “should”
- “Contributions”
  - Need to be more fine grained

Problems with the Experiments

- We ran a reasoner competition
  - (Somewhat) different set of inputs
  - Rather different set of reasoners

RESULTS: CLASSIFICATION EL





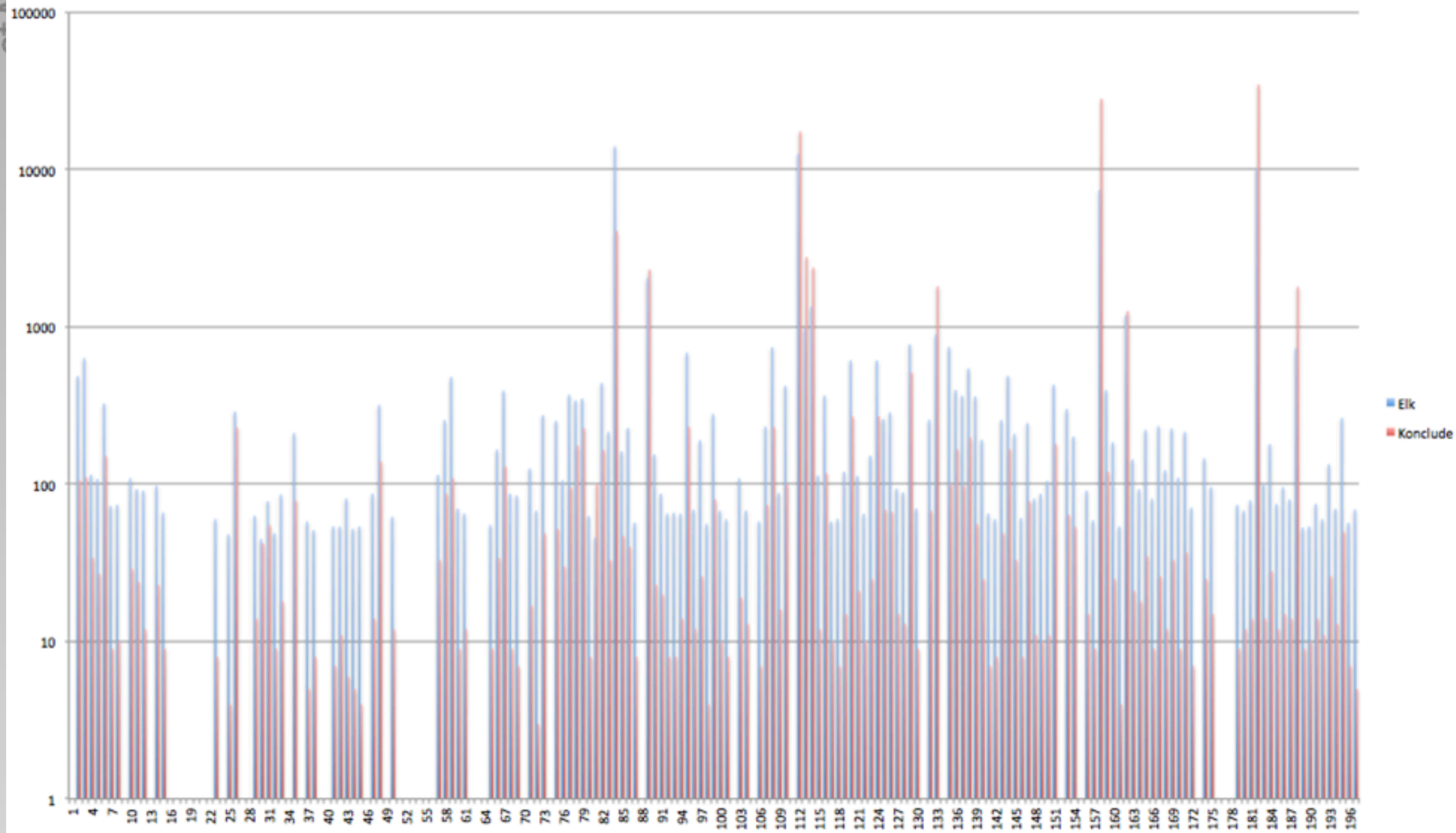
BUT WAIT!!!!

Within the commonly solved 155:

	ELK	Konclude
Total	77,740 ms	104,210 ms
Average	502 ms	672ms

Does Pavel deserve his trophy?

Interesting question!



Understanding

- Empirical work is difficult
  - The world is a funny place
  - Designing experiments is hard
  - Executing experiments is hard!
  - Interpreting result is hard!!
  - We should do a lot more of it
    - but well
- Consult with other people
  - Always feel free to send me an email
- Be meticulous
  - Record every step
  - Try to think through various outcomes
  - Use good data management tools
- Be bold!