This work was partially supported from the European Community's FP7/2007-2013 under grant agreement number 257178 (project CHOReOS).

How is Life for a Table in an Evolving Relational Schema? **Birth, Death & Everything in Between**

Panos Vassiliadis, Apostolos Zarras, Ioannis Skoulis

Department of Computer Science and Engineering
University of Ioannina, Hellas

http://www.cs.uoi.gr/~pvassil/publications/2015_ER

WHAT ARE THE “LAWS” OF DATABASE SCHEMA EVOLUTION?

Imagine if we could predict how a schema will evolve over time…

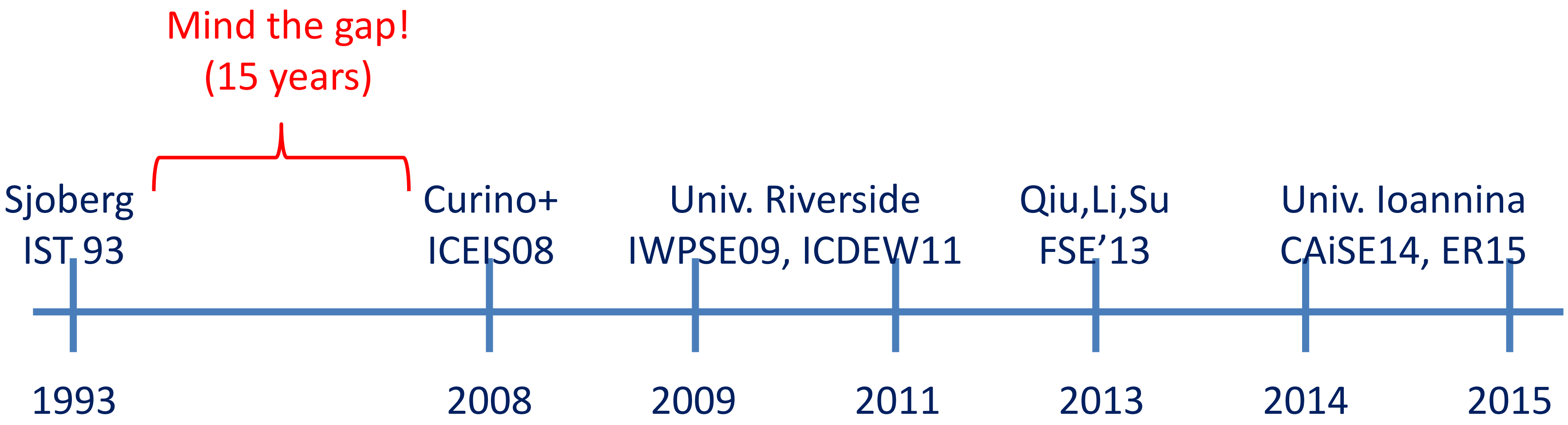
- … we would be able to **“design for evolution”** and **minimize the impact of evolution** to the surrounding applications
  - by applying design patterns
  - by avoiding anti-patterns & complexity increase

  … **in both the db and the code**
- … we would be able to **plan** administration and perfective maintenance tasks and resources, instead of responding to emergencies

Why aren’t we there yet?

- Historically, nobody from the research community had access + the right to publish to version histories of database schemata
- Open source tools internally hosting databases have changed this landscape:
  - not only is the code available, but also,
  - public repositories (git, svn, …) keep the entire history of revisions
- We are now presented with the opportunity to study the version histories of such “open source databases”

Timeline of empirical studies



Mind the gap!
(15 years)

Sjoberg
IST 93

Curino+
ICEIS08

Univ. Riverside
IWPSE09, ICDEW11

Qiu,Li,Su
FSE'13

Univ. Ioannina
CAiSE14, ER15

Our take on the problem

- Collected version histories for the schemata of 8 open-source projects
  - CMS's: MediaWiki, TYPO3, Coppermine, phpBB, OpenCart
  - Physics: ATLAS Trigger --- Bio: Ensemble, BioSQL
- Preprocessed them to be parsable by our **HECATE schema comparison tool** and exported the transitions between each two subsequent versions and measures for them (size, growth, changes)
- Exploratory search where we statistically studied / mined these measures, to **extract patterns & regularities for the lives of tables**
- **Available at:**

  https://github.com/DAINTINESS-Group/EvolutionDatasets

Scope of the study

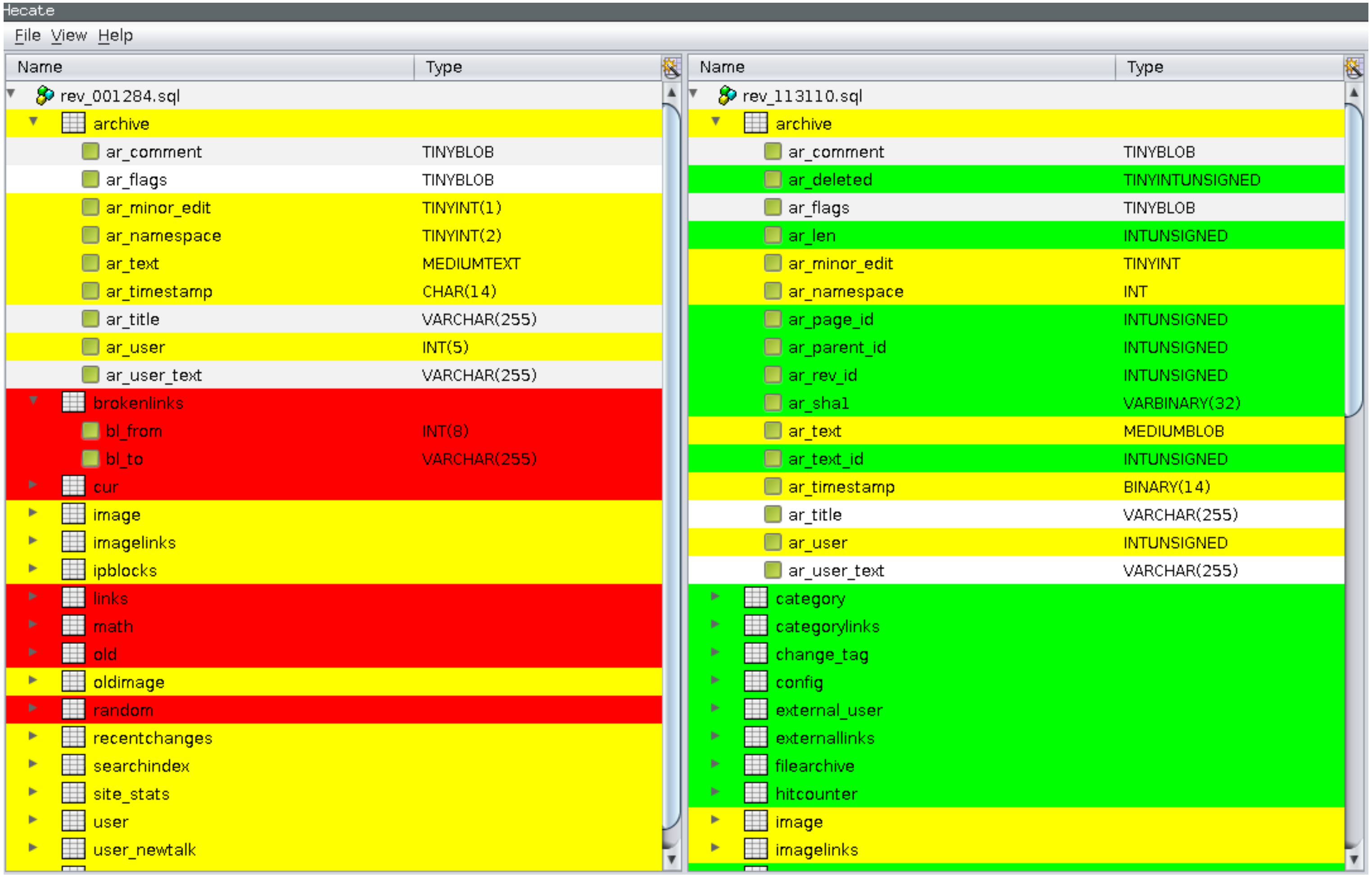
- **Scope**:
  - databases being part of open-source software (and not proprietary ones)
  - long history
  - we work only with changes at the logical schema level (and ignore physical-level changes like index creation or change of storage engine)

- We encompass datasets with different domains ([A]: physics, [B]: biomedical, [C]: CMS's), amount of growth (shade: high, med, low) & schema size

- We should be very careful to not overgeneralize findings to proprietary databases or physical schemata!

FoSS Dataset	Versions	Lifetime	Tables @ Start	Tables @ End
ATLAS Trigger [A]	84	2 Y, 7 M, 2 D	56	73
BioSQL [B]	46	10 Y, 6 M, 19 D	21	28
Coppermine [C]	117	8 Y, 6 M, 2 D	8	22
Ensembl [B]	528	13 Y, 3 M, 15 D	17	75
MediaWiki [C]	322	8 Y, 10 M, 6 D	17	50
OpenCart [C]	164	4 Y, 4 M, 3 D	46	114
phpBB [C]	133	6 Y, 7 M, 10 D	61	65
TYPO3 [C]	97	8 Y, 11 M, 0 D	10	23

Hecate: SQL schema diff extractor



Name	Type
rev_001284.sql	
archive	
ar_comment	TINYBLOB
ar_flags	TINYBLOB
ar_minor_edit	TINYINT(1)
ar_namespace	TINYINT(2)
ar_text	MEDIUMTEXT
ar_timestamp	CHAR(14)
ar_title	VARCHAR(255)
ar_user	INT(5)
ar_user_text	VARCHAR(255)
brokenlinks	
bl_from	INT(8)
bl_to	VARCHAR(255)
cur	
image	
imagelinks	
ipblocks	
links	
math	
old	
oldimage	
random	
recentchanges	
searchindex	
site_stats	
user	
user_newtalk	

Name	Type
rev_113110.sql	
archive	
ar_comment	TINYBLOB
ar_deleted	TINYINTUNSIGNED
ar_flags	TINYBLOB
ar_len	INTUNSIGNED
ar_minor_edit	TINYINT
ar_namespace	INT
ar_page_id	INTUNSIGNED
ar_parent_id	INTUNSIGNED
ar_rev_id	INTUNSIGNED
ar_sha1	VARBINARY(32)
ar_text	MEDIUMBLOB
ar_text_id	INTUNSIGNED
ar_timestamp	BINARY(14)
ar_title	VARCHAR(255)
ar_user	INTUNSIGNED
ar_user_text	VARCHAR(255)
category	
categorylinks	
change_tag	
config	
external_user	
externallinks	
filearchive	
hitcounter	
image	
imagelinks	

https://github.com/DAINTINESS-Group/Hecate

Exploratory search of the schema histories for patterns

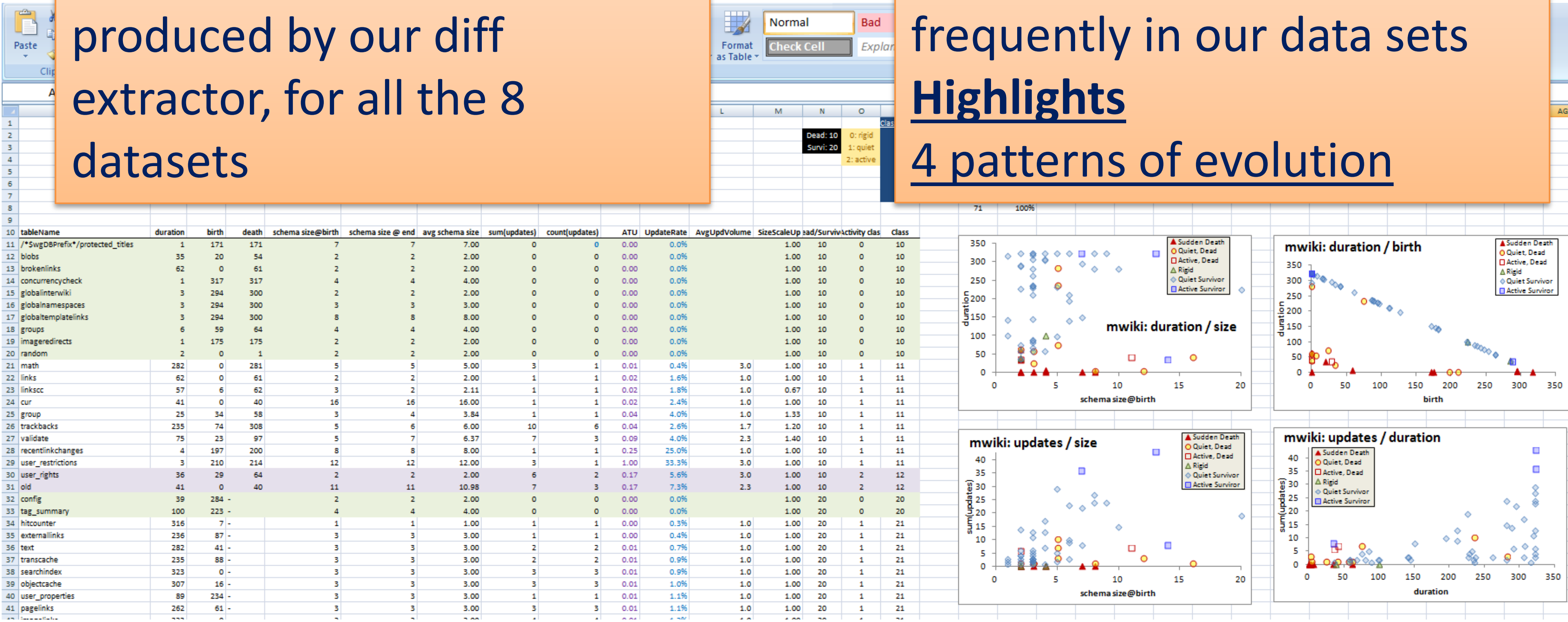
Input: schema histories from github/sourceforge/…

Raw material: details and stats on each table's life, as produced by our diff extractor, for all the 8 datasets

Output: properties & patterns on table properties (birth, duration, amt of change, …) that occur frequently in our data sets

Highlights

4 patterns of evolution



10	tableName	duration	birth	death	schema size@birth	schema size @ end	avg schema size	sum(updates)	count(updates)	ATU	UpdateRate	AvgUpdVolume	SizeScaleUp	ead/Surviv	ctivity clas	Class
11	/*$wgDBPrefix*/protected_titles	1	171	171	7	7	7.00	0	0	0.00	0.0%		1.00	10	0	10
12	blobs	35	20	54	2	2	2.00	0	0	0.00	0.0%		1.00	10	0	10
13	brokenlinks	62	0	61	2	2	2.00	0	0	0.00	0.0%		1.00	10	0	10
14	concurrencycheck	1	317	317	4	4	4.00	0	0	0.00	0.0%		1.00	10	0	10
15	globalinterwiki	3	294	300	2	2	2.00	0	0	0.00	0.0%		1.00	10	0	10
16	globalnamespaces	3	294	300	3	3	3.00	0	0	0.00	0.0%		1.00	10	0	10
17	globaltemplatelinks	3	294	300	8	8	8.00	0	0	0.00	0.0%		1.00	10	0	10
18	groups	6	59	64	4	4	4.00	0	0	0.00	0.0%		1.00	10	0	10
19	imageredirects	1	175	175	2	2	2.00	0	0	0.00	0.0%		1.00	10	0	10
20	random	2	0	1	2	2	2.00	0	0	0.00	0.0%		1.00	10	0	10
21	math	282	0	281	5	5	5.00	3	1	0.01	0.4%	3.0	1.00	10	1	11
22	links	62	0	61	2	2	2.00	1	1	0.02	1.6%	1.0	1.00	10	1	11
23	linkscc	57	6	62	3	2	2.11	1	1	0.02	1.8%	1.0	0.67	10	1	11
24	cur	41	0	40	16	16	16.00	1	1	0.02	2.4%	1.0	1.00	10	1	11
25	group	25	34	58	3	4	3.84	1	1	0.04	4.0%	1.0	1.33	10	1	11
26	trackbacks	235	74	308	5	6	6.00	10	6	0.04	2.6%	1.7	1.20	10	1	11
27	validate	75	23	97	5	7	6.37	7	3	0.09	4.0%	2.3	1.40	10	1	11
28	recentlinkchanges	4	197	200	8	8	8.00	1	1	0.25	25.0%	1.0	1.00	10	1	11
29	user_restrictions	3	210	214	12	12	12.00	3	1	1.00	33.3%	3.0	1.00	10	1	11
30	user_rights	36	29	64	2	2	2.00	6	2	0.17	5.6%	3.0	1.00	10	2	12
31	old	41	0	40	11	11	10.98	7	3	0.17	7.3%	2.3	1.00	10	2	12
32	config	39	284	-	2	2	2.00	0	0	0.00	0.0%		1.00	20	0	20
33	tag_summary	100	223	-	4	4	4.00	0	0	0.00	0.0%		1.00	20	0	20
34	hitcounter	316	7	-	1	1	1.00	1	1	0.00	0.3%	1.0	1.00	20	1	21
35	externallinks	236	87	-	3	3	3.00	1	1	0.00	0.4%	1.0	1.00	20	1	21
36	text	282	41	-	3	3	3.00	2	2	0.01	0.7%	1.0	1.00	20	1	21
37	transcache	235	88	-	3	3	3.00	2	2	0.01	0.9%	1.0	1.00	20	1	21
38	searchindex	323	0	-	3	3	3.00	3	3	0.01	0.9%	1.0	1.00	20	1	21
39	objectcache	307	16	-	3	3	3.00	3	3	0.01	1.0%	1.0	1.00	20	1	21
40	user_properties	89	234	-	3	3	3.00	1	1	0.01	1.1%	1.0	1.00	20	1	21
41	pagelinks	262	61	-	3	3	3.00	3	3	0.01	1.1%	1.0	1.00	20	1	21

-Statistical properties for schema size, change and duration of tables

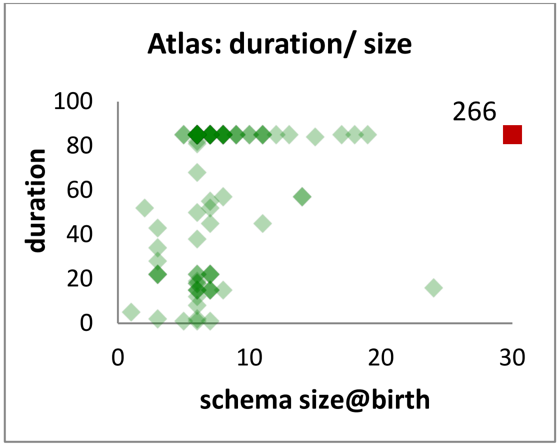
- How are these measures interrelated?

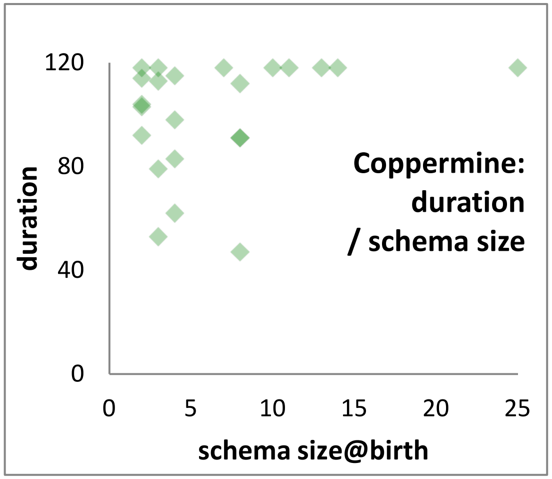
SCHEMA SIZE, CHANGE AND DURATION

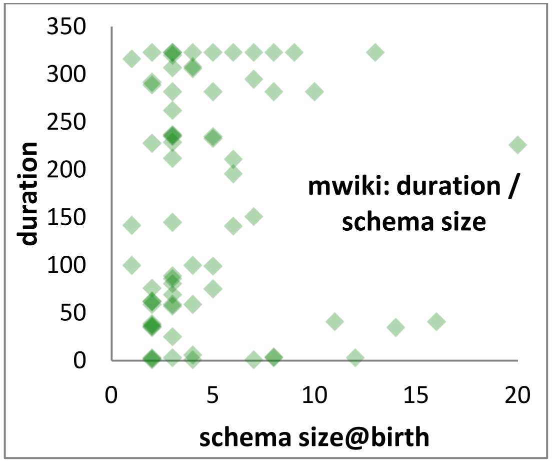
The Gamma Γ Pattern: "if you 're wide, you survive"

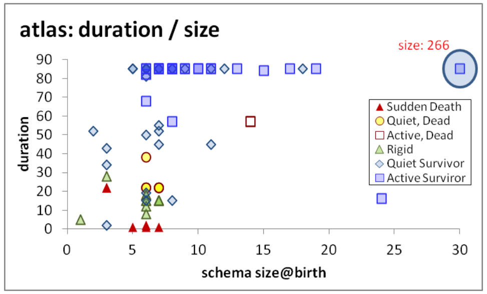
- The Gamma phenomenon:
  - tables with small schema sizes can have arbitrary durations, //small size does not determine duration
  - larger size tables last long

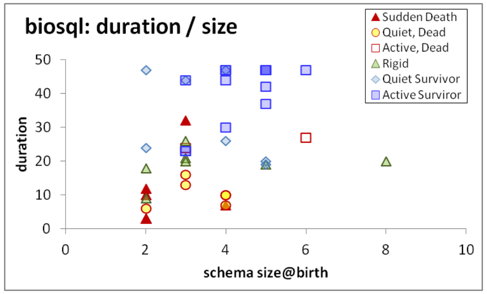
- Observations:
  - whenever a table exceeds the critical value of 10 attributes in its schema, its chances of surviving are high.
  - in most cases, the large tables are created early on and are not deleted afterwards.

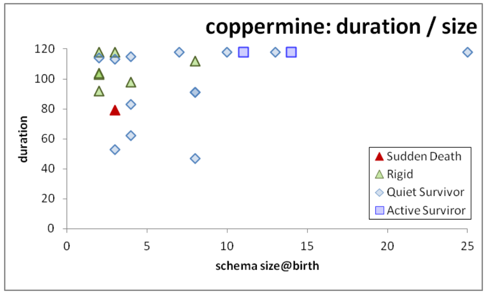


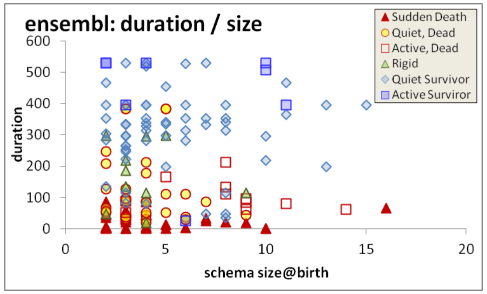


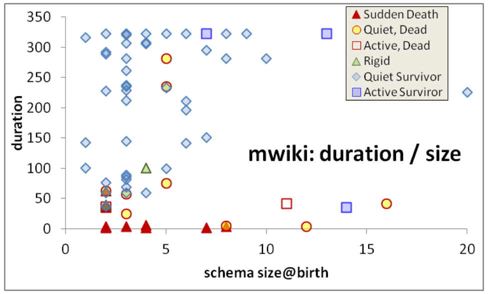


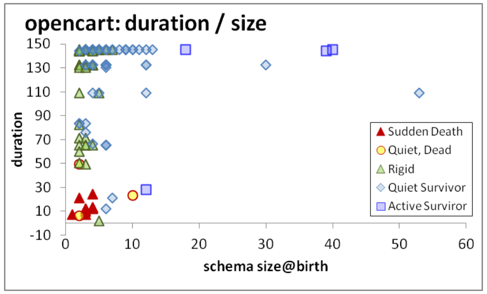


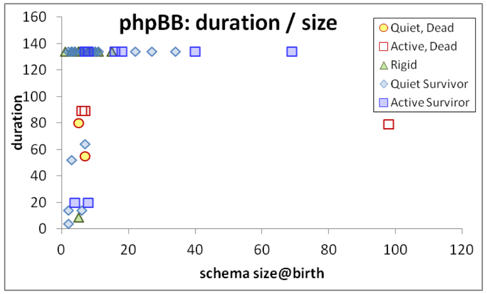


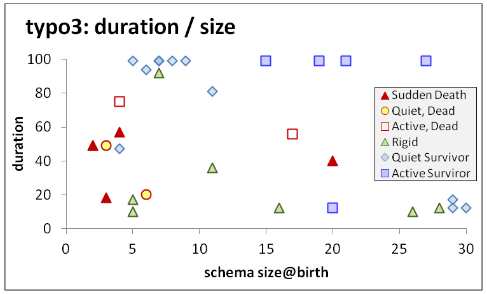












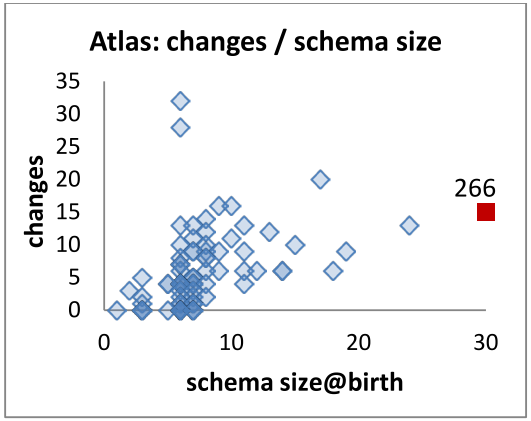
Exceptions

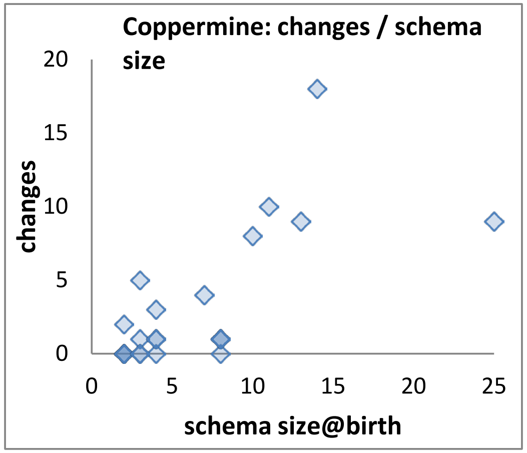
- Biosql: nobody exceeds 10 attributes
- Ensembl, mwiki: very few exceed 10 attributes, 3 of them died
- typo: has many late born survivors

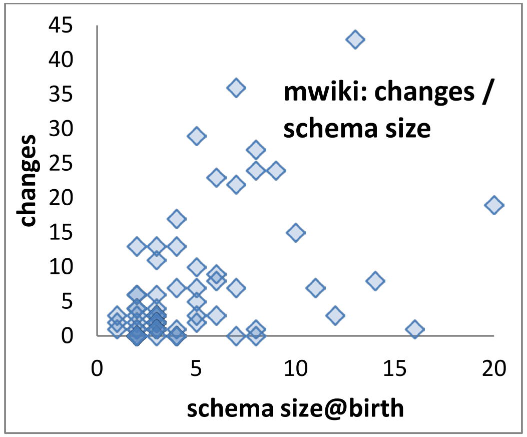
The Comet Pattern

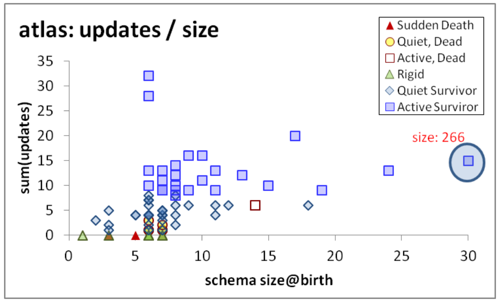
"Comet " for change over schema size with:

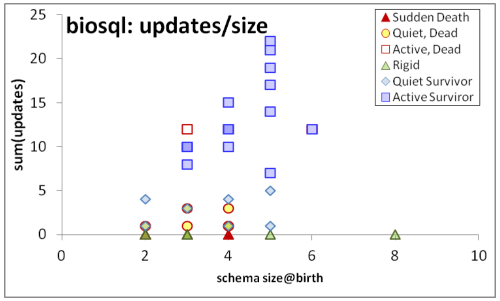
- a large, dense, nucleus cluster close to the beginning of the axes, denoting small size and small amount of change,
- medium schema size tables typically demonstrating medium to large change
  - The tables with the largest amount of change are typically tables whose schema is on average one standard deviation above the mean
- wide tables with large schema sizes demonstrating small to medium (typically around the middle of the y-axis) amount of change.

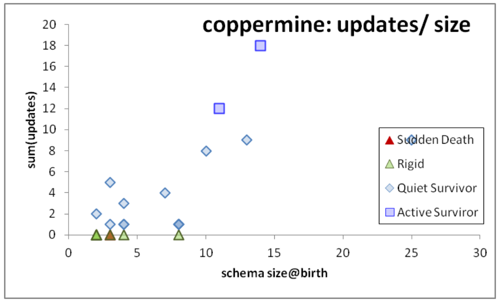


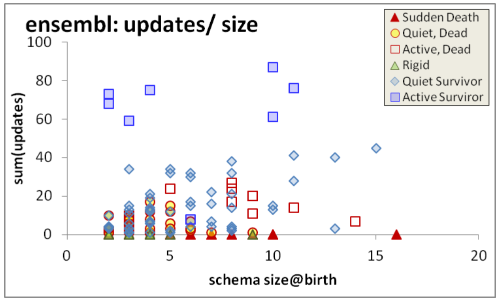


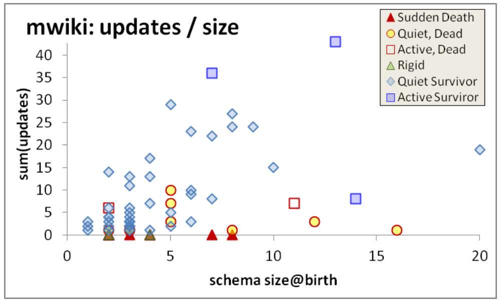


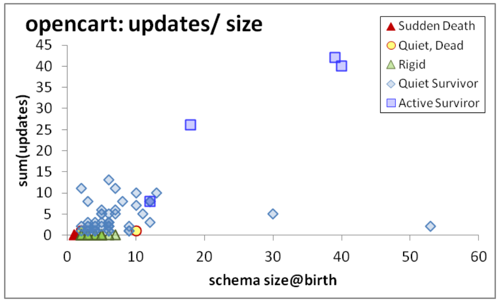

atlas: updates / size

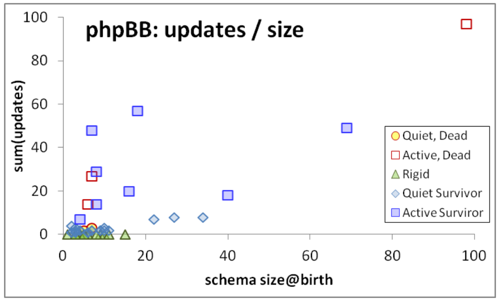

biosql: updates/size

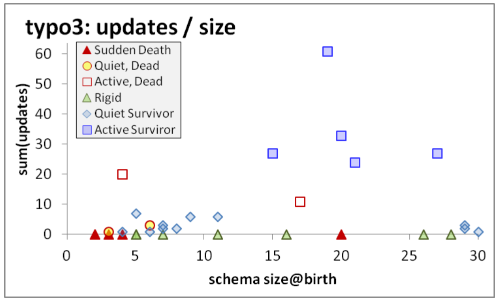

coppermine: updates/ size


ensembl: updates/ size

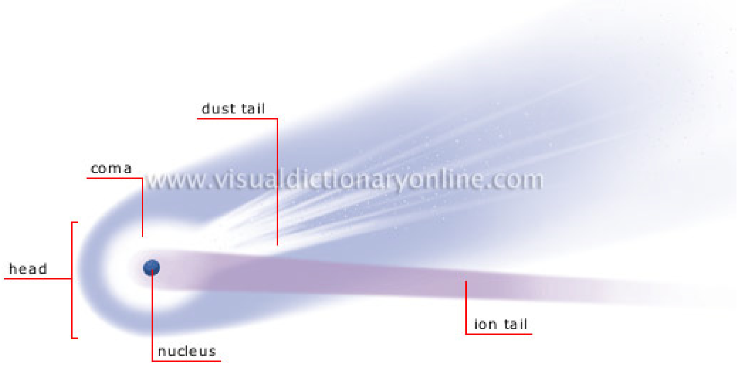

mwiki: updates / size


opencart: updates/ size


phpBB: updates / size


typo3: updates / size

http://visual.merriam-webster.com/astronomy/celestial-bodies/comet.php



Comets have two tails:
White one is made of comet dust particles. Blue one is made of electrically charged gas.

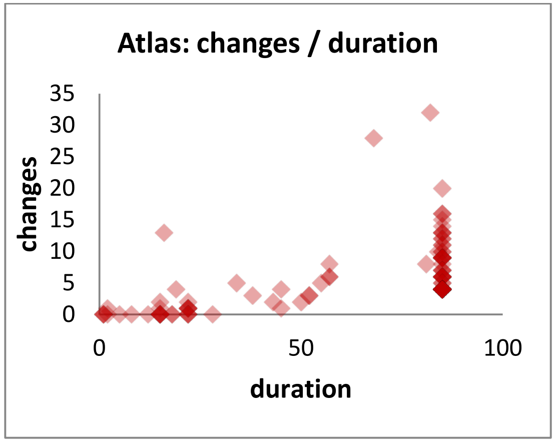
The coma is the cloud of comet dust particles surrounding the nucleus.

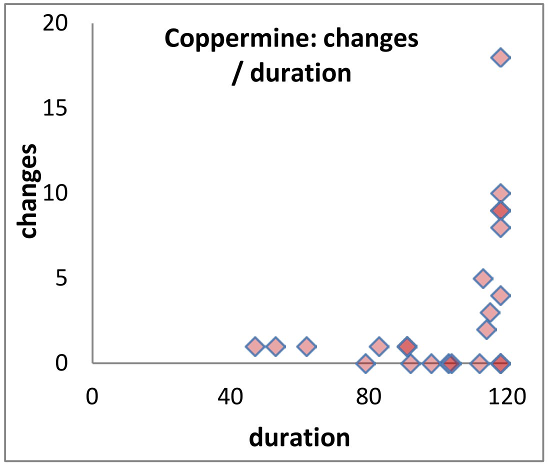
Nucleus is solid, icy heart of comet, inside the cloud of the coma

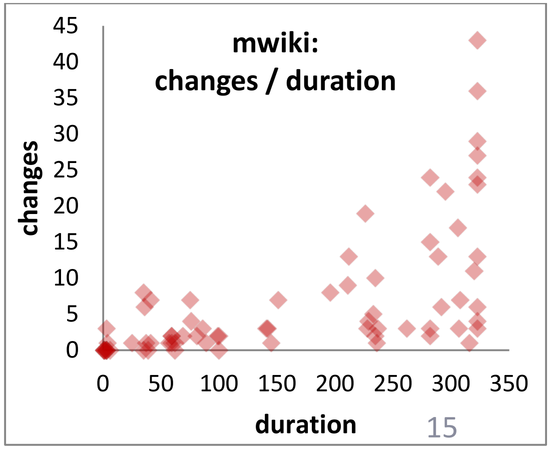
http://spaceplace.nasa.gov/comet-nucleus/en/

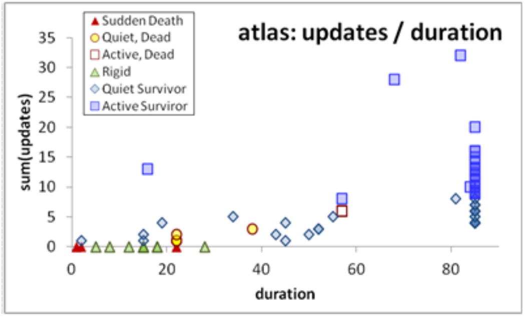
The inverse Gamma pattern

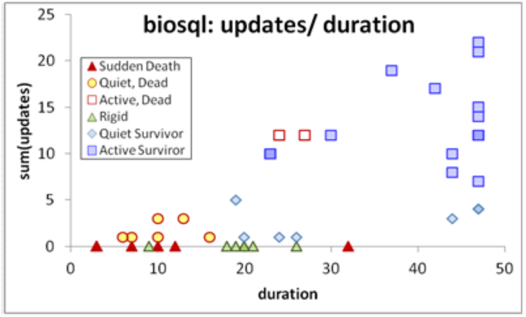
- The correlation of change and duration is as follows:
  - small durations come necessarily with small change,
  - large durations come with all kinds of change activity and
  - medium sized durations come mostly with small change activity (Inverse Gamma).

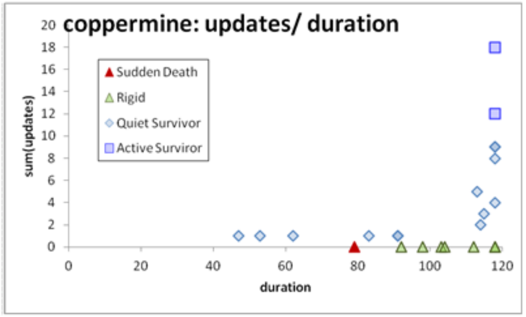


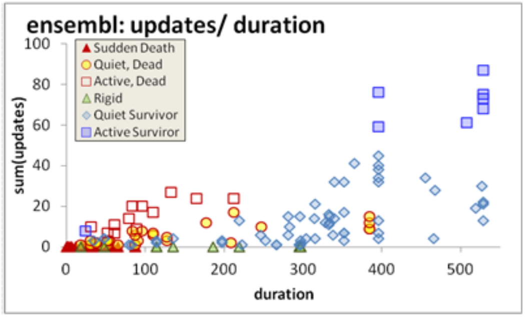


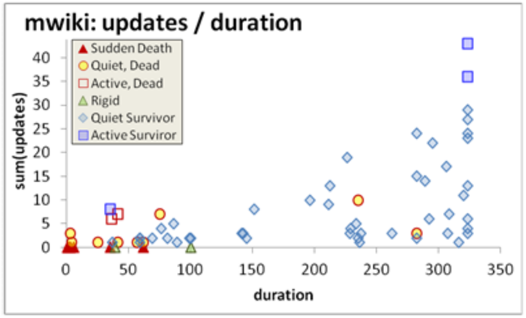


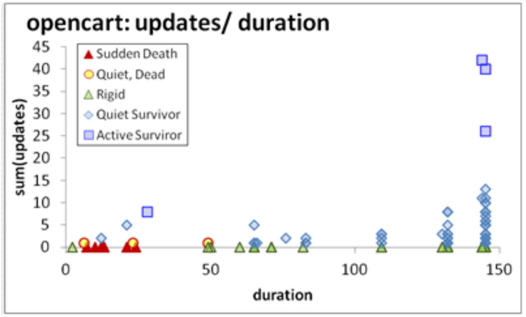

atlas: updates / duration
▲ Sudden Death
○ Quiet, Dead
□ Active, Dead
△ Rigid
◇ Quiet Survivor
□ Active Surviror
sum(updates)
0
5
10
15
20
25
30
35
0
20
40
60
80
duration

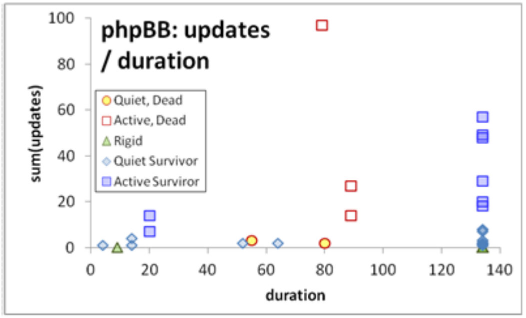

biosql: updates/ duration
▲ Sudden Death
○ Quiet, Dead
□ Active, Dead
△ Rigid
◇ Quiet Survivor
□ Active Surviror
sum(updates)
0
5
10
15
20
25
0
10
20
30
40
50
duration

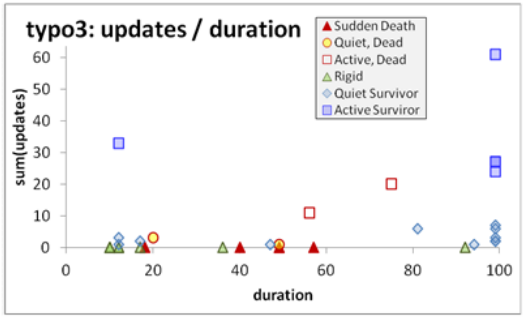

coppermine: updates/ duration
▲ Sudden Death
△ Rigid
◇ Quiet Survivor
□ Active Surviror
sum(updates)
0
2
4
6
8
10
12
14
16
18
20
0
20
40
60
80
100
120
duration


ensembl: updates/ duration
▲ Sudden Death
○ Quiet, Dead
□ Active, Dead
△ Rigid
◇ Quiet Survivor
□ Active Surviror
sum(updates)
0
20
40
60
80
100
0
100
200
300
400
500
duration


mwiki: updates / duration
▲ Sudden Death
○ Quiet, Dead
□ Active, Dead
△ Rigid
◇ Quiet Survivor
□ Active Surviror
sum(updates)
0
5
10
15
20
25
30
35
40
0
50
100
150
200
250
300
350
duration


opencart: updates/ duration
▲ Sudden Death
○ Quiet, Dead
△ Rigid
◇ Quiet Survivor
□ Active Surviror
sum(updates)
0
5
10
15
20
25
30
35
40
45
0
50
100
150
duration


phpBB: updates / duration
○ Quiet, Dead
□ Active, Dead
△ Rigid
◇ Quiet Survivor
□ Active Surviror
sum(updates)
0
20
40
60
80
100
0
20
40
60
80
100
120
140
duration


typo3: updates / duration
▲ Sudden Death
○ Quiet, Dead
□ Active, Dead
△ Rigid
◇ Quiet Survivor
□ Active Surviror
sum(updates)
0
10
20
30
40
50
60
0
20
40
60
80
100
duration

Who are the top changers?

Who are removed at some point of time?

How do removals take place?

BIRTHDAY & SCHEMA SIZE & MATTERS OF LIFE AND DEATH

Quiet tables rule, esp. for mature db's

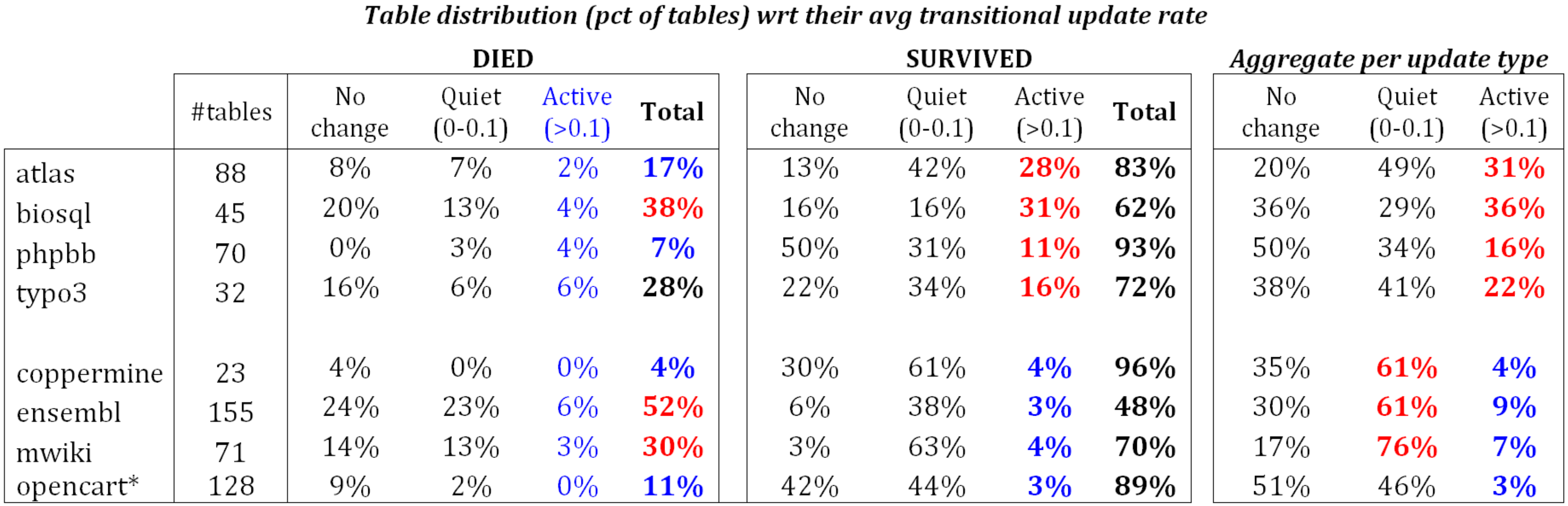


Table distribution (pct of tables) wrt their avg transitional update rate

		DIED				SURVIVED				*Aggregate per update type*		
	#tables	No change	Quiet (0-0.1)	Active (>0.1)	**Total**	No change	Quiet (0-0.1)	Active (>0.1)	**Total**	No change	Quiet (0-0.1)	Active (>0.1)
atlas	88	8%	7%	2%	**17%**	13%	42%	**28%**	**83%**	20%	49%	**31%**
biosql	45	20%	13%	4%	**38%**	16%	16%	**31%**	**62%**	36%	29%	**36%**
phpbb	70	0%	3%	4%	**7%**	50%	31%	**11%**	**93%**	50%	34%	**16%**
typo3	32	16%	6%	6%	**28%**	22%	34%	**16%**	**72%**	38%	41%	**22%**
coppermine	23	4%	0%	0%	**4%**	30%	61%	**4%**	**96%**	35%	**61%**	**4%**
ensembl	155	24%	23%	6%	**52%**	6%	38%	**3%**	**48%**	30%	**61%**	**9%**
mwiki	71	14%	13%	3%	**30%**	3%	63%	**4%**	**70%**	17%	**76%**	**7%**
opencart*	128	9%	2%	0%	**11%**	42%	44%	**3%**	**89%**	51%	46%	**3%**

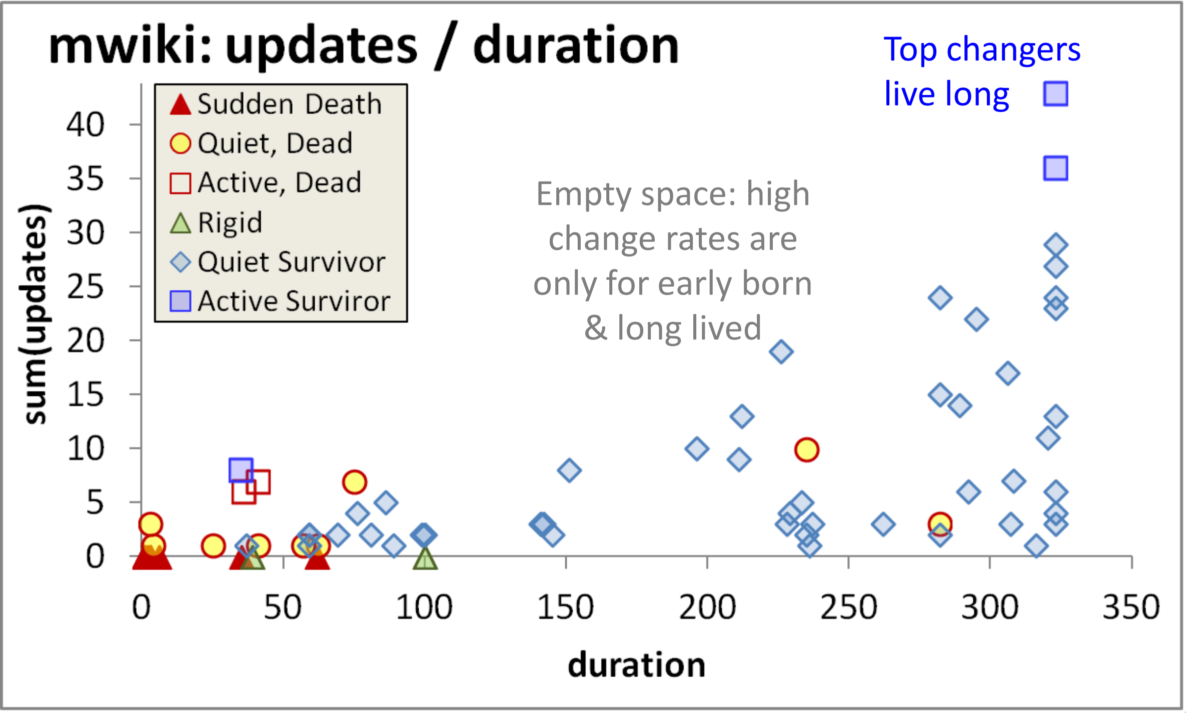
Non-survivors

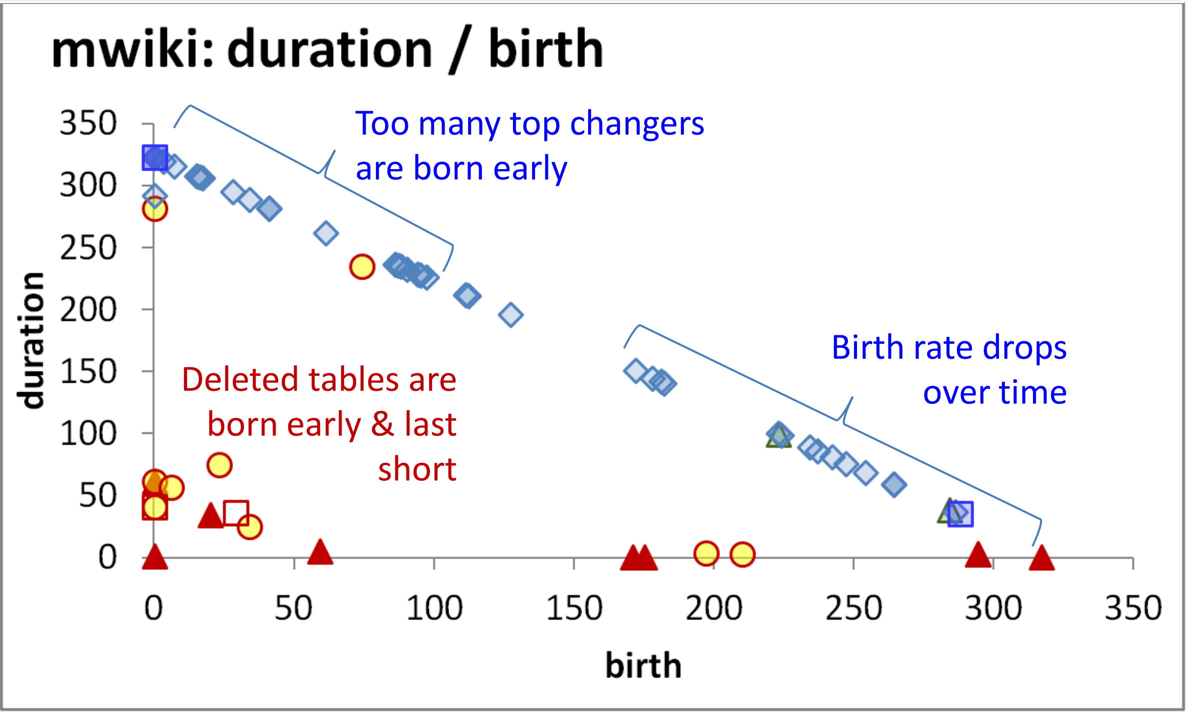
- Sudden deaths mostly
- Quiet come ~ close
- Too few active

Survivors

- Quiet tables rule
- Rigid and active then
- Active mostly in "new" db's

Mature DB's: the pct of active tables drops significantly

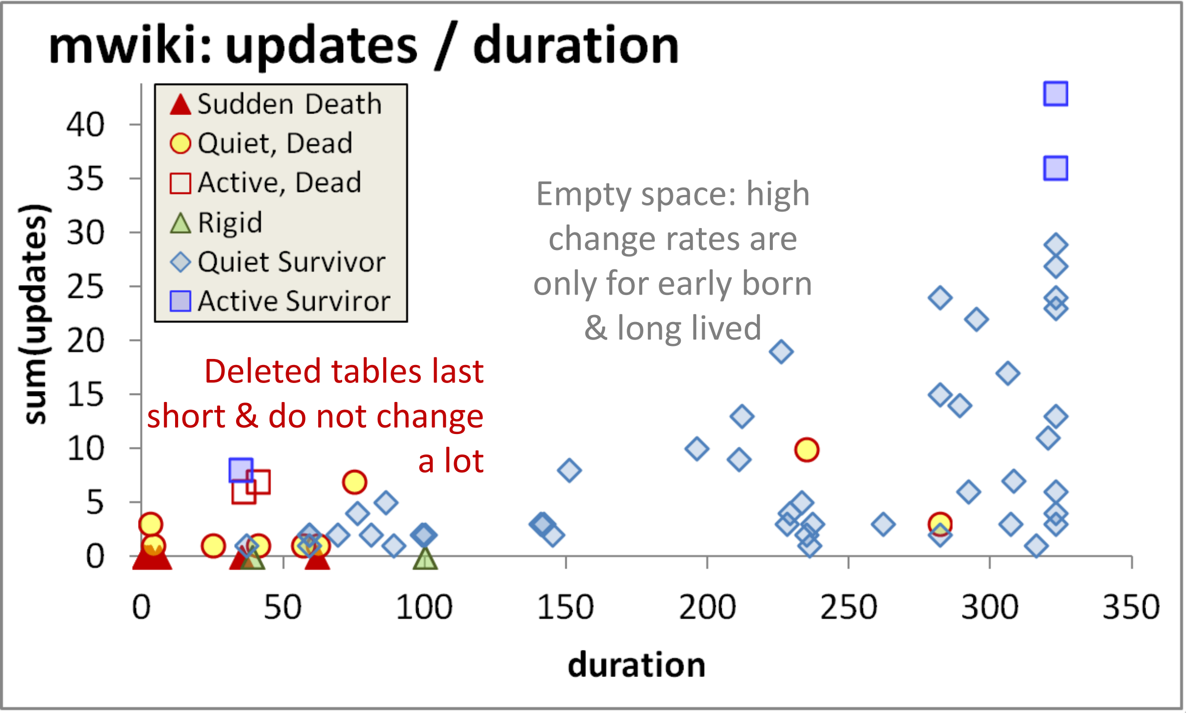




Longevity and update activity correlate !!

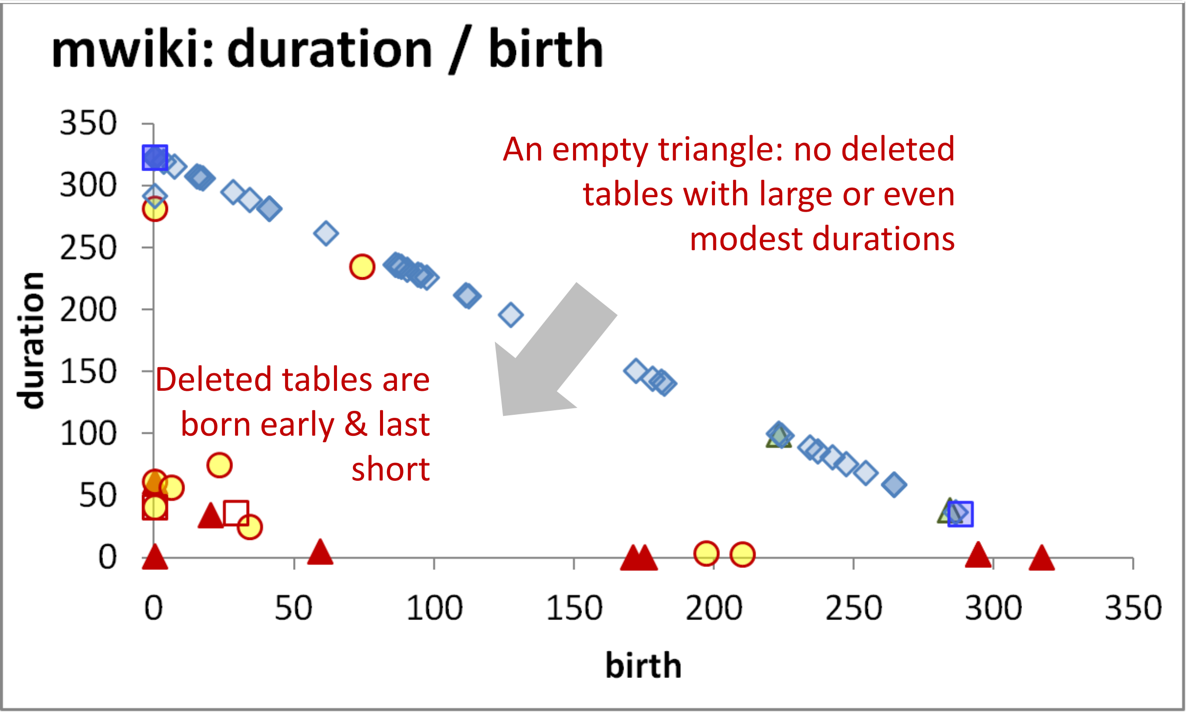
The few top-changers (in terms of avg trans. update – ATU)

- are long lived,
- typically come from the early versions of the database
- due to the combination of high ATU and duration => they have high total amount of updates, and,
- frequently survive!



Empty space: high change rates are only for early born & long lived

Deleted tables last short & do not change a lot

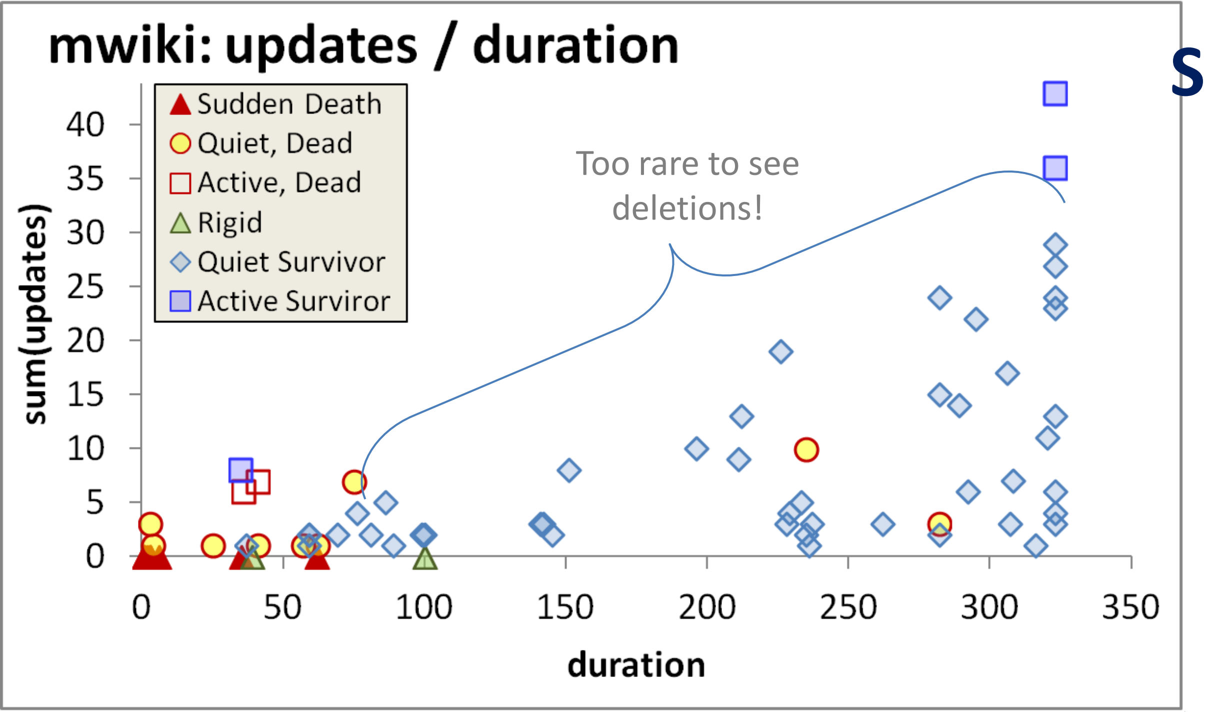


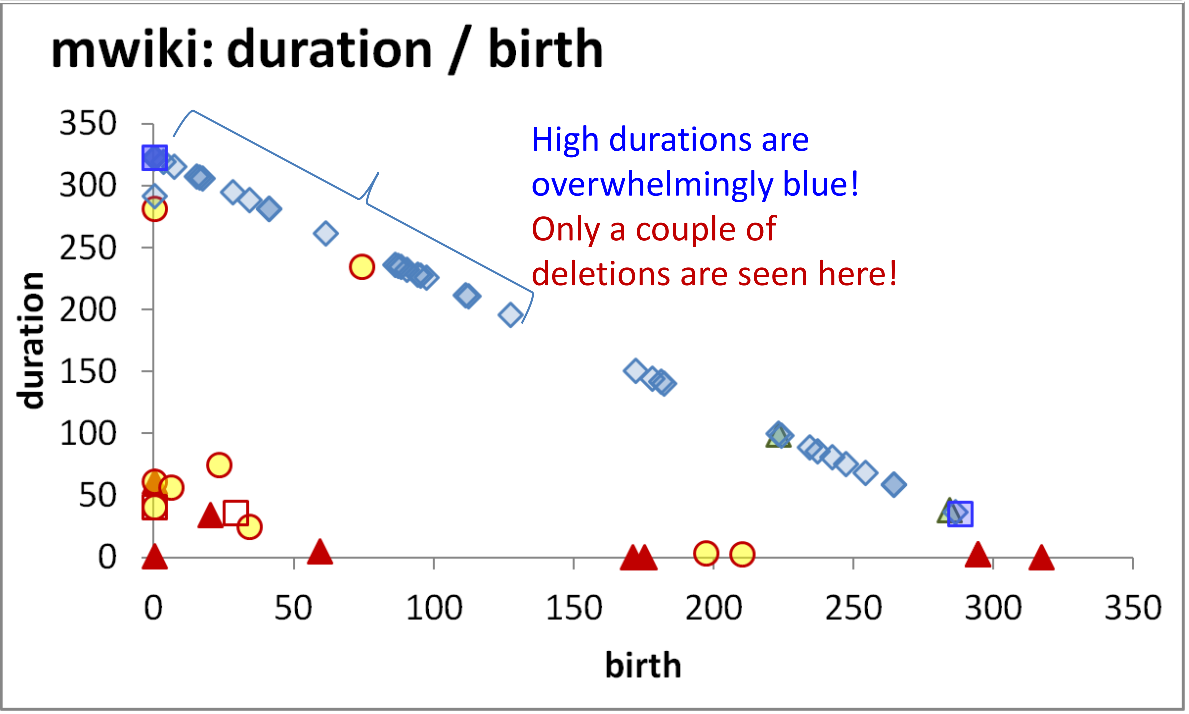
An empty triangle: no deleted tables with large or even modest durations

Deleted tables are born early & last short

Die young and suddenly

- There is a very large concentration of the deleted tables in a small range of newly born, quickly removed, with few or no updates…
- …. resulting in very low numbers of removed tables with medium or long durations (empty triangle).



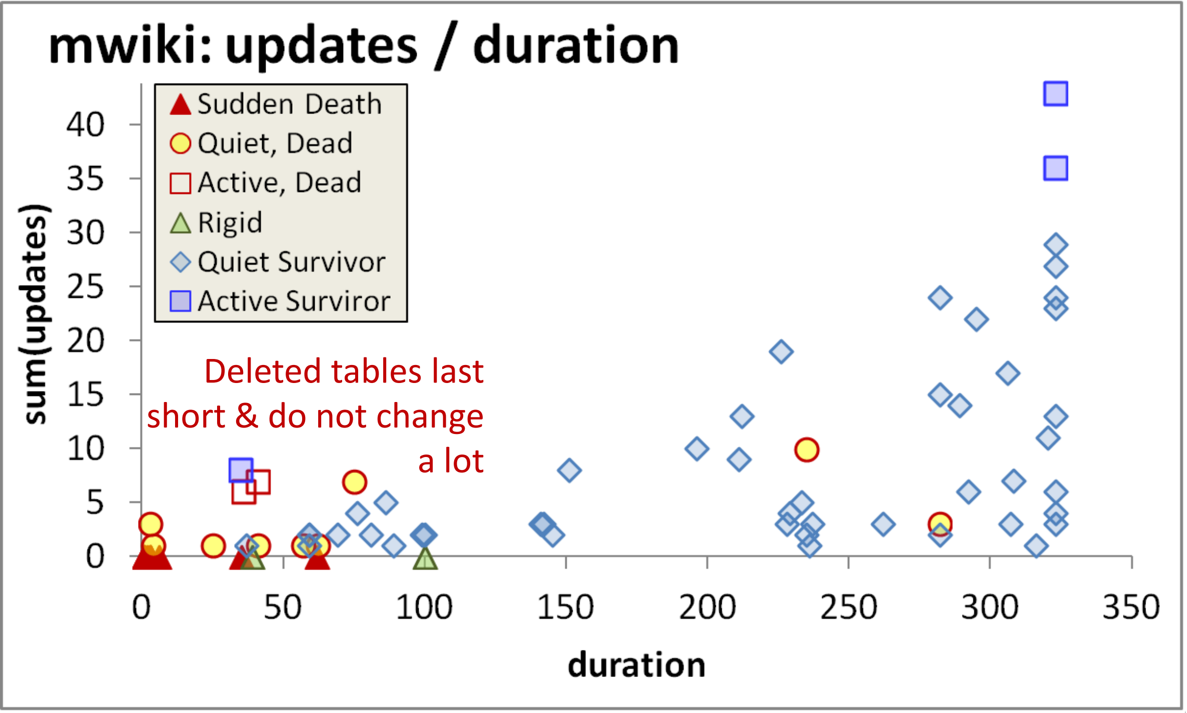


Survive long enough & you 're probably safe

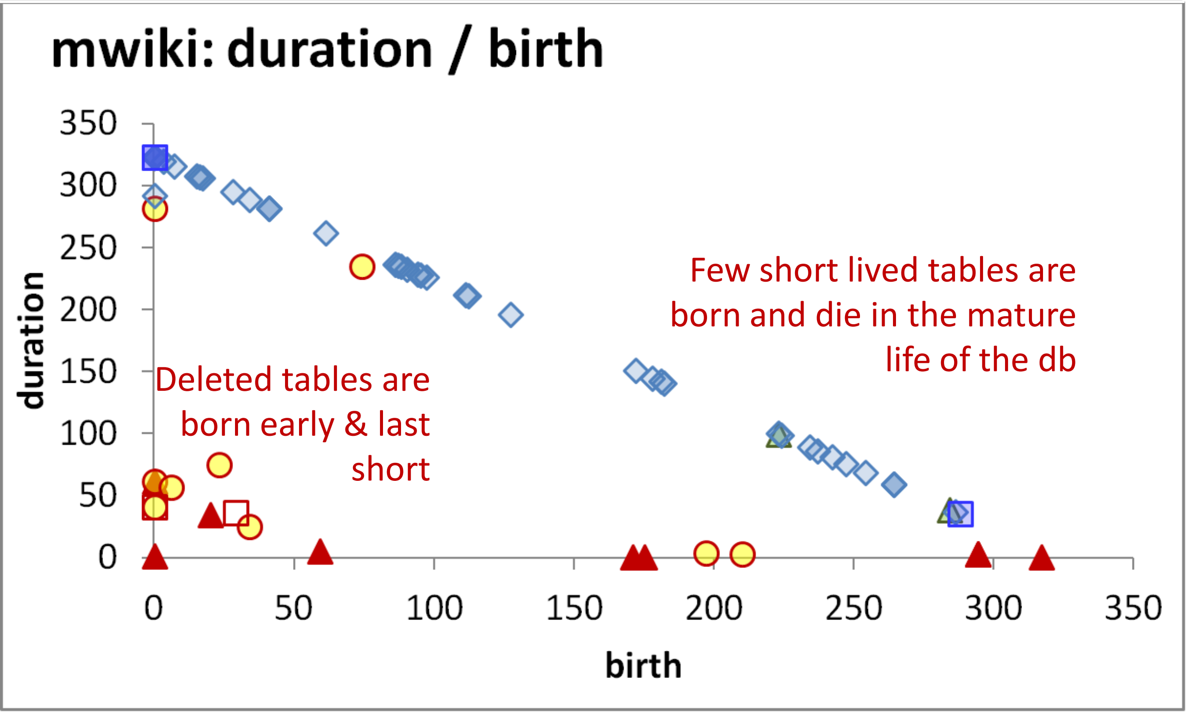
It is quite rare to see tables being removed at old age

Typically, the area of high duration is overwhelmingly inhabited by survivors (although each data set comes with a few such cases)!

Die young and suddenly


mwiki: updates / duration

Deleted tables last short & do not change a lot

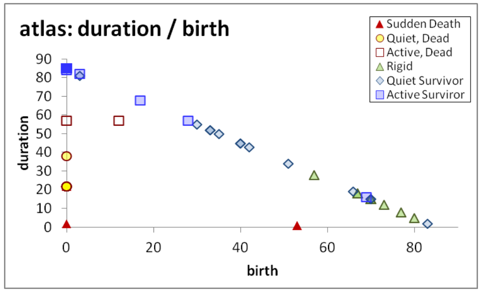

mwiki: duration / birth

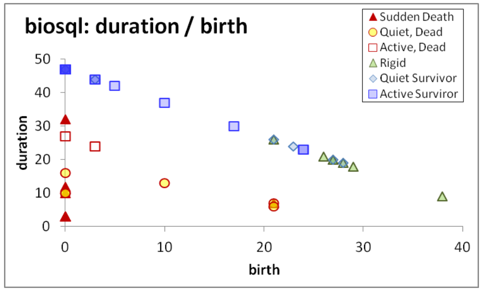
Few short lived tables are born and die in the mature life of the db

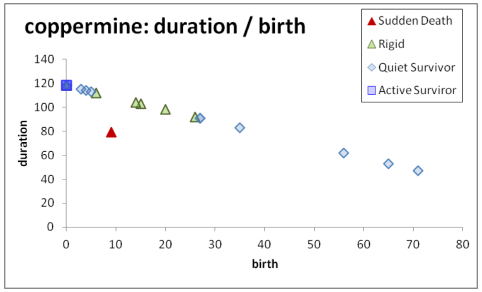
Deleted tables are born early & last short

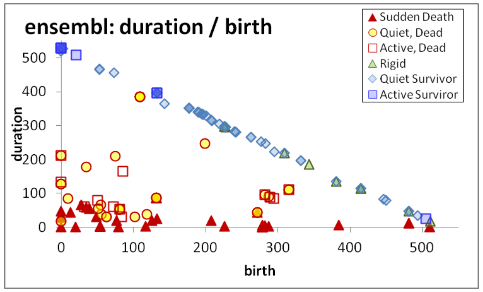
[Early life of the db] There is a very large concentration of the deleted tables in a small range of newly born, quickly removed, with few or no updates, resulting in very low numbers of removed tables with medium or long durations.

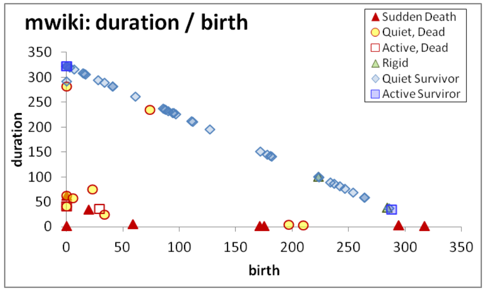
[**Mature db**] After the early stages of the databases, we see the birth of tables who eventually get deleted, but they mostly come with very small durations and sudden deaths.

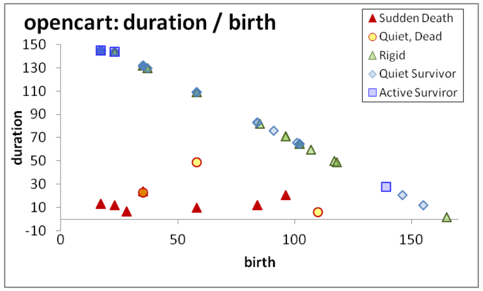

atlas: duration / birth

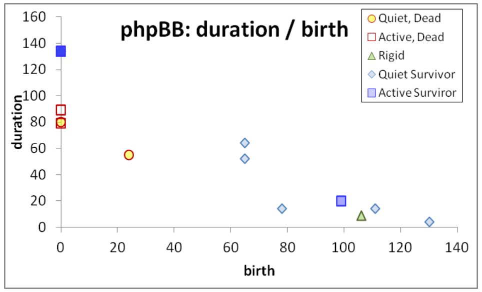

biosql: duration / birth

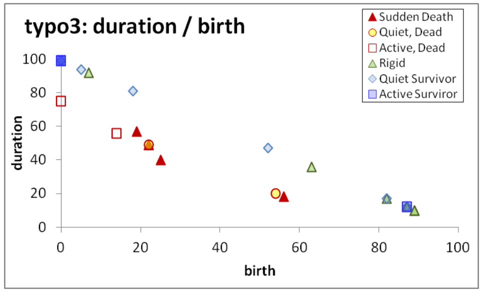

coppermine: duration / birth

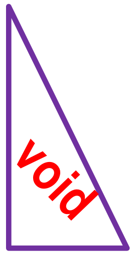

ensembl: duration / birth


mwiki: duration / birth


opencart: duration / birth


phpBB: duration / birth

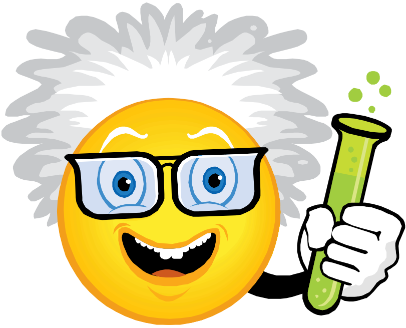

typo3: duration / birth



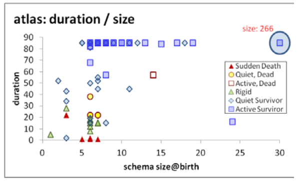
Main Findings

Open Issues

CONCLUSIONS & OPEN ISSUES

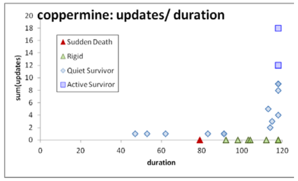


Regularities on table change do exist!



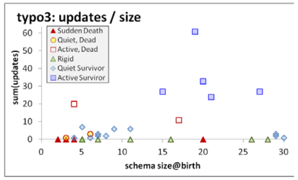
Γ

Only the thin die young, ~~all~~ the **wide ones seem to live forever**

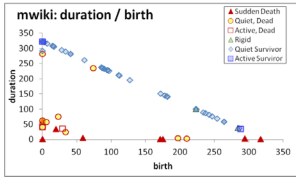


⅃

Top-changers typically live long, are early born, survive …



… and they are not necessarily the widest ones in terms of schema size



Progressive cooling: most change activity lies at the beginning of the db history

Void triangle: The few dead tables are typically quiet, early born, short lived, and quite often all three of them

Unexplored research territory (risky but possibly rewarding)

- Weather Forecast: given the history and the state of a database, predict subsequent events
  - Risky: frequently, changes come due to an external, changing world and have “thematic” affinity.
  - Big & small steps in many directions needed (more data sets, studies with high internal validity to find causations, more events to capture, …)
- Engineer for evolution: To absorb change gracefully we can try to (i) alter db design and DDL; (ii) encapsulate the database via a “stable” API; …

To probe further (code, data, details, presentations, …)

http://www.cs.uoi.gr/~pvassil/publications/2015_ER/

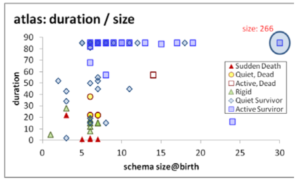
Q & A time…

Many thanks:

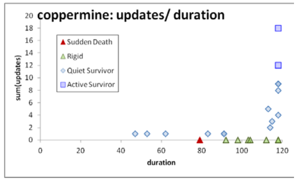
- to our hosts for all their efforts to organize ER 2015!
- to you for your attention!

Regularities on table change do exist!

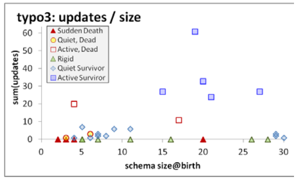
Tack!
Ett varmt tack till
våra värdar!



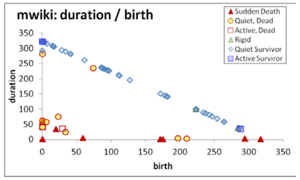
Γ Only the thin die young, ~~all~~ the **wide ones seem to live forever**



⅃ Top-changers typically live long, are early born, survive …



… and they are not necessarily the widest ones in terms of schema size



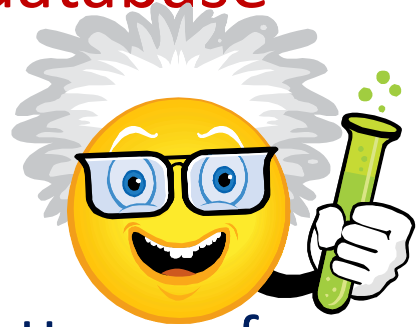
Progressive cooling: most change activity lies at the beginning of the db history

Void triangle: The few dead tables are typically quiet, early born, short lived, and quite often all three of them

AUXILIARY SLIDES

What are the “laws” of database (schema) evolution?

- How do databases change?
- In particular, how does the schema of a database evolve over time?



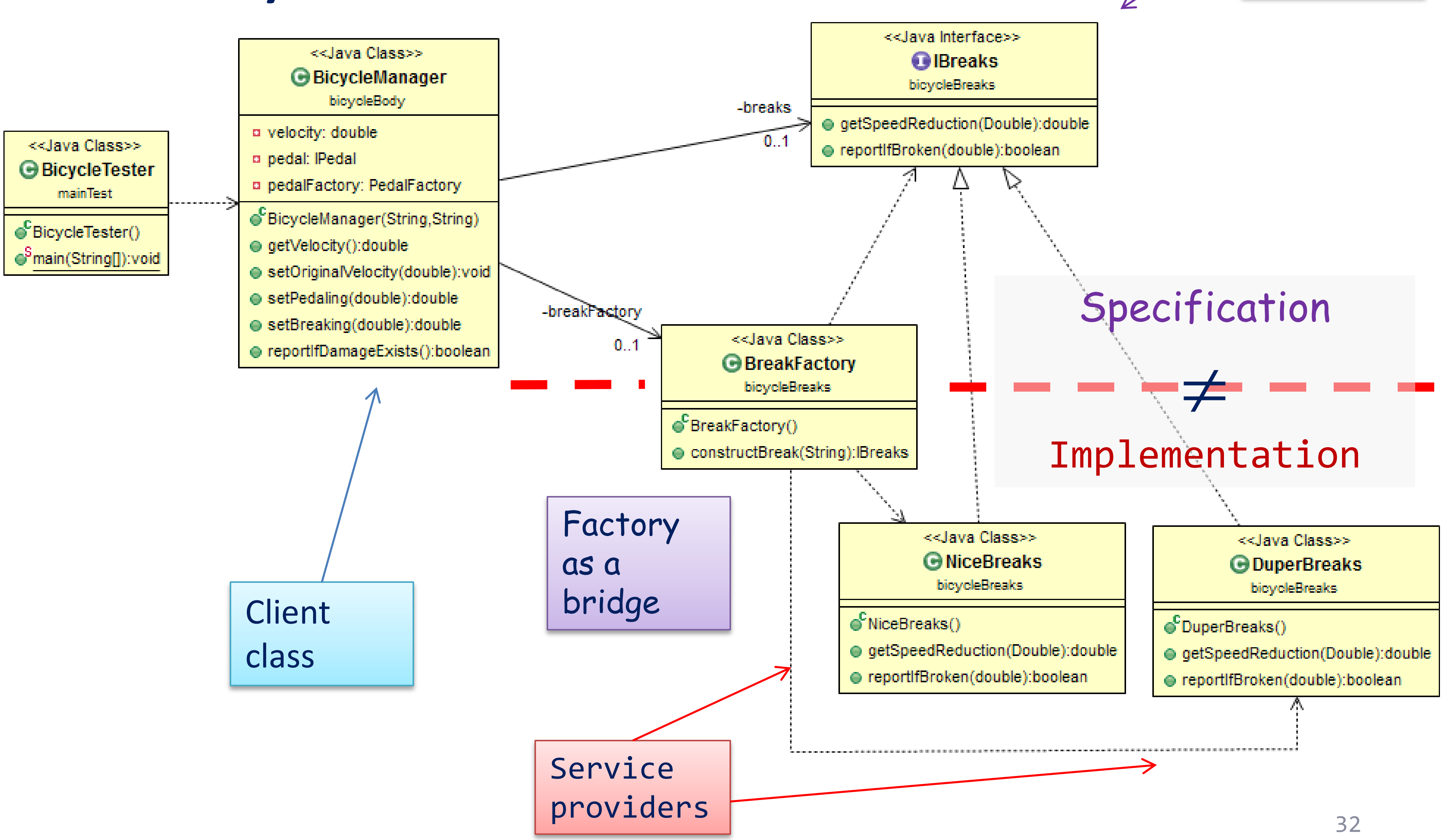
- Long term research goals:
  - Are there any “invariant properties” (e.g., patterns of repeating behavior) on the way database (schemata) change?
  - Is there a theory / model to explain them?
  - Can we exploit findings to engineer data-intensive ecosystems that withstand change gracefully?

Why care for the “laws”/patterns of schema evolution?

- Scientific curiosity!
- Practical Impact: DB’s are dependency magnets. Applications have to conform to the structure of the db…
  - typically, development waits till the “db backbone” is stable and applications are build on top of it
  - slight changes to the structure of a db can cause several (parts of) different applications to crash, causing the need for emergency repairing

Abstract coupling example from my SW Dev course

Interface as a contract



<<Java Class>>
BicycleTester
mainTest
BicycleTester()
main(String[]):void
<<Java Class>>
BicycleManager
bicycleBody
velocity: double
pedal: IPedal
pedalFactory: PedalFactory
BicycleManager(String,String)
getVelocity():double
setOriginalVelocity(double):void
setPedaling(double):double
setBreaking(double):double
reportIfDamageExists():boolean
-breaks
0..1
<<Java Interface>>
IBreaks
bicycleBreaks
getSpeedReduction(Double):double
reportIfBroken(double):boolean
-breakFactory
0..1
<<Java Class>>
BreakFactory
bicycleBreaks
BreakFactory()
constructBreak(String):IBreaks
<<Java Class>>
NiceBreaks
bicycleBreaks
NiceBreaks()
getSpeedReduction(Double):double
reportIfBroken(double):boolean
<<Java Class>>
DuperBreaks
bicycleBreaks
DuperBreaks()
getSpeedReduction(Double):double
reportIfBroken(double):boolean

Specification

≠

Implementation

Client class

Factory as a bridge

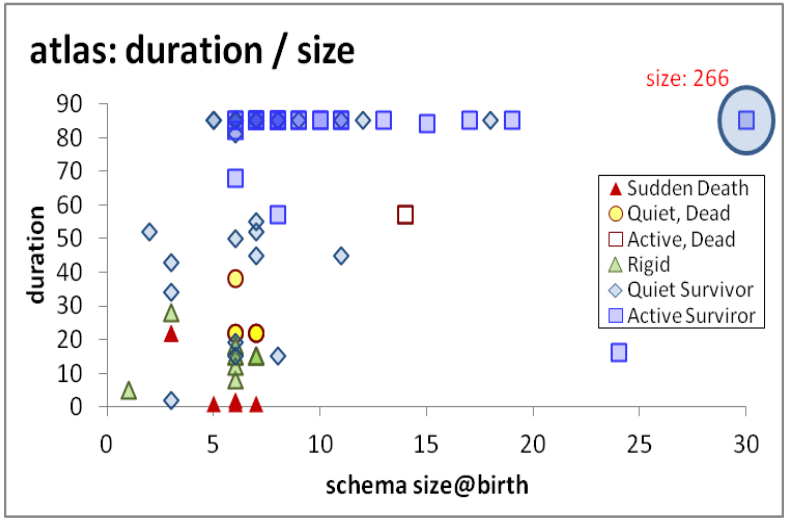
Service providers

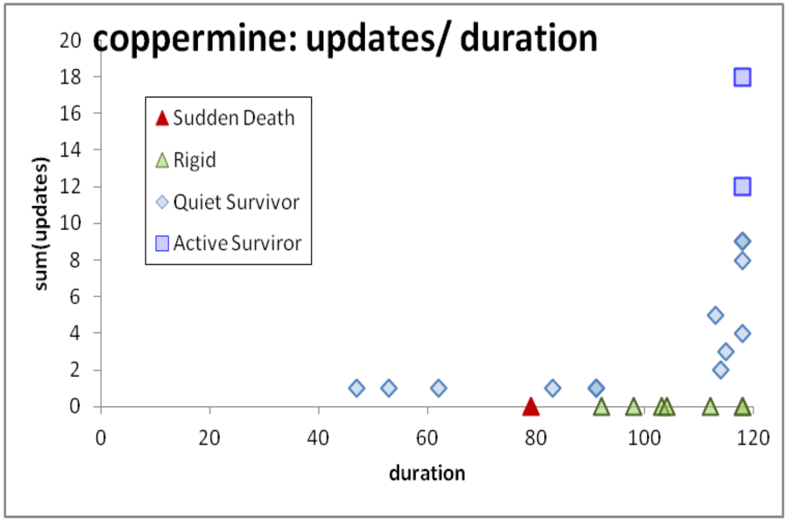
Hecate: SQL schema diff extractor

- Parses DDL files
- Creates a model for the parsed SQL elements
- Compares two versions of the same schema
- Reports on the diff performed with a variety of metrics
- Exports the transitions that occurred in XML format

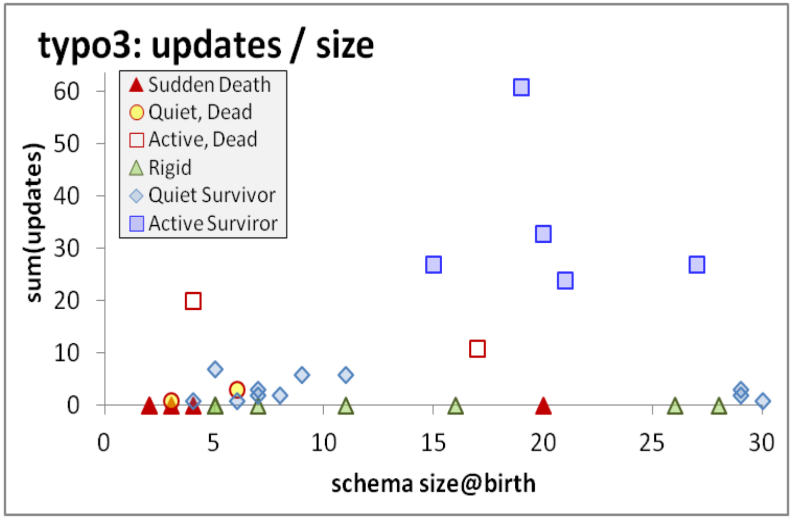
https://github.com/DAINTINESS-Group/Hecate

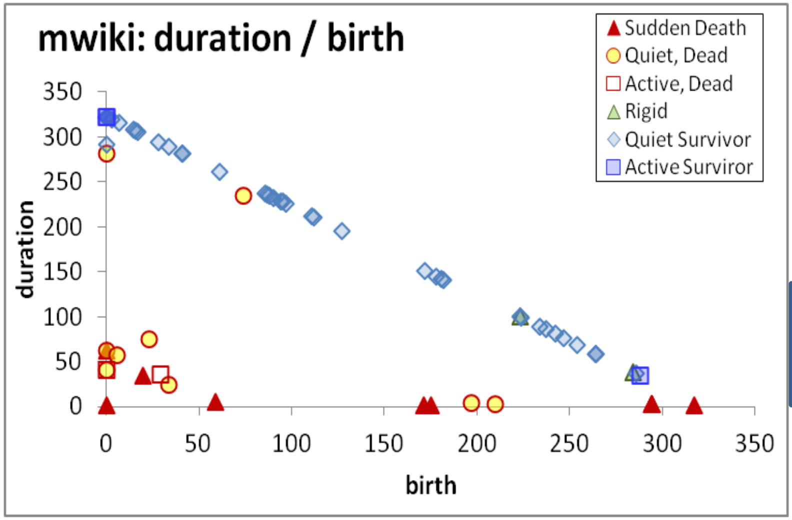
Γ

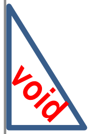

atlas: duration / size


coppermine: updates/ duration

˩


typo3: updates / size


mwiki: duration / birth



To probe further (code, data, details, presentations, …)

http://www.cs.uoi.gr/~pvassil/publications/2015_ER/

SCOPE OF THE STUDY && VALIDITY CONSIDERATIONS

Data sets

Dataset	Versions	Lifetime	Tables Start	Tables End	Attributes Start	Attributes End	Commits per Day	% commits with change	Repository URL
ATLAS Trigger	84	2 Y, 7 M, 2 D	56	73	709	858	0,089	82%	http://atdaq-sw.cern.ch/cgi-bin/viewcvs-atlas.cgi/offline/Trigger/TrigConfiguration/TrigDb/share/sql/combined_schema.sql
BioSQL	46	10 Y, 6 M, 19 D	21	28	74	129	0,012	63%	https://github.com/biosql/biosql/blob/master/sql/biosqldb-mysql.sql
Coppermine	117	8 Y, 6 M, 2 D	8	22	87	169	0,038	50%	http://sourceforge.net/p/coppermine/code/8581/tree/trunk/cpg1.5.x/sql/schema.sql
Ensembl	528	13 Y, 3 M, 15 D	17	75	75	486	0,109	60%	http://cvs.sanger.ac.uk/cgi-bin/viewvc.cgi/ensembl/sql/table.sql?root=ensembl&view=log
MediaWiki	322	8 Y, 10 M, 6 D	17	50	100	318	0,100	59%	https://svn.wikimedia.org/viewvc/mediawiki/trunk/phase3/maintenance/tables.sql?view=log
OpenCart	164	4 Y, 4 M, 3 D	46	114	292	731	0,104	47%	https://github.com/opencart/opencart/blob/master/upload/install/opencart.sql
phpBB	133	6 Y, 7 M, 10 D	61	65	611	565	0,055	82%	https://github.com/phpbb/phpbb3/blob/develop/phpBB/install/schemas/mysql_41_schema.sql
TYPO3	97	8 Y, 11 M, 0 D	10	23	122	414	0,030	76%	https://git.typo3.org/Packages/TYPO3.CMS.git/history/TYPO3_6-0:/t3lib/stddb/tables.sql

Scope of the study

- **Scope**:
  - databases being part of open-source software (and not proprietary ones)
  - long history
  - we work only with changes at the logical schema level (and ignore physical-level changes like index creation or change of storage engine)
- We encompass datasets with different domains ([A]: physics, [B]: biomedical, [C]: CMS's), amount of growth (shade: high, med, low) & schema size
- We should be very careful to not overgeneralize findings to proprietary databases or physical schemata!

FoSS Dataset	Versions	Lifetime	Tables @ Start	Tables @ End
ATLAS Trigger [A]	84	2 Y, 7 M, 2 D	56	73
BioSQL [B]	46	10 Y, 6 M, 19 D	21	28
Coppermine [C]	117	8 Y, 6 M, 2 D	8	22
Ensembl [B]	528	13 Y, 3 M, 15 D	17	75
MediaWiki [C]	322	8 Y, 10 M, 6 D	17	50
OpenCart [C]	164	4 Y, 4 M, 3 D	46	114
phpBB [C]	133	6 Y, 7 M, 10 D	61	65
TYPO3 [C]	97	8 Y, 11 M, 0 D	10	23

Can we generalize out findings broadly?

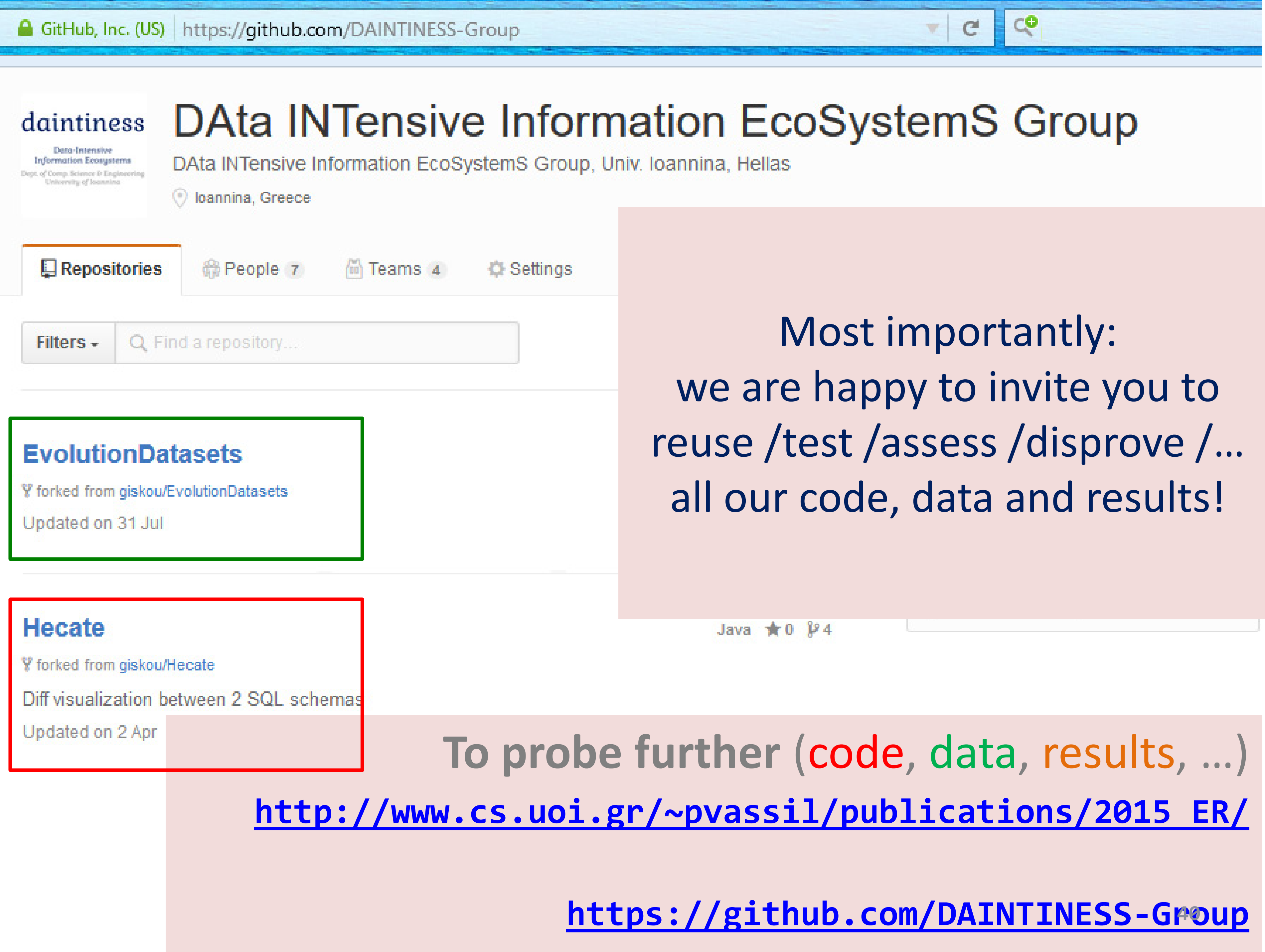
External validity

- We perform an **exploratory study to observe frequently occurring phenomena** within the scope of the aforementioned population
- **Are our data sets representative enough?** Is it possible that the observed behaviors are caused by sui-generis characteristics of the studied data sets?
  - Yes: we believe we have a good population definition & we abide by it
  - Yes: we believe we have a large number of databases, from a variety of domains with different profiles, that seem to give fairly consistent answers to our research questions (behavior deviations are mostly related to the maturity of the database and not to its application area).
  - Yes: we believe we have a good data extraction and measurement process without interference / selection / … of the input from our part
  - Maybe: unclear when the number of studied databases is large enough to declare the general application of a pattern as “universal”.

Can we generalize out findings broadly?

External validity

- Understanding the represented population
  - Precision: all our data sets belong to the specified population
  - Definition Completeness: no missing property that we knowledgably omit to report
  - FoSS has an inherent way of maintenance and evolution
- Representativeness of selected datasets
  - Data sets come from 3 categories of FoSS (CMS / Biomedical / Physics)
  - They have different size and growth volumes
  - Results are fairly consistent both in our ER’15 and our CAiSE’14 papers
- Treatment of data
  - We have tested our “Delta Extractor”, Hecate, to parse the input correctly & adapted it during its development; the parser is not a full-blown SQL parser, but robust to ignore parts unknown to it
  - A handful of cases where adapted in the Coppermine to avoid overcomplicating the parser; not a serious threat to validity ; other than that we have not interfered with the input
  - Fully automated counting for the measures via Hecate



GitHub, Inc. (US) https://github.com/DAINTINESS-Group

daintiness
Data-Intensive Information Ecosystems
Dept. of Comp. Science & Engineering University of Ioannina

DAta INTensive Information EcoSystemS Group

DAta INTensive Information EcoSystemS Group, Univ. Ioannina, Hellas

Ioannina, Greece

EvolutionDatasets

forked from giskou/EvolutionDatasets

Updated on 31 Jul

Hecate

forked from giskou/Hecate

Diff visualization between 2 SQL schemas

Updated on 2 Apr

Most importantly:
we are happy to invite you to
reuse /test /assess /disprove /…
all our code, data and results!

To probe further (code, data, results, …)

http://www.cs.uoi.gr/~pvassil/publications/2015_ER/

https://github.com/DAINTINESS-Group

Internal validity

- Can we confirm statements A=>B? No!
- Are there any spurious relationships? Maybe!

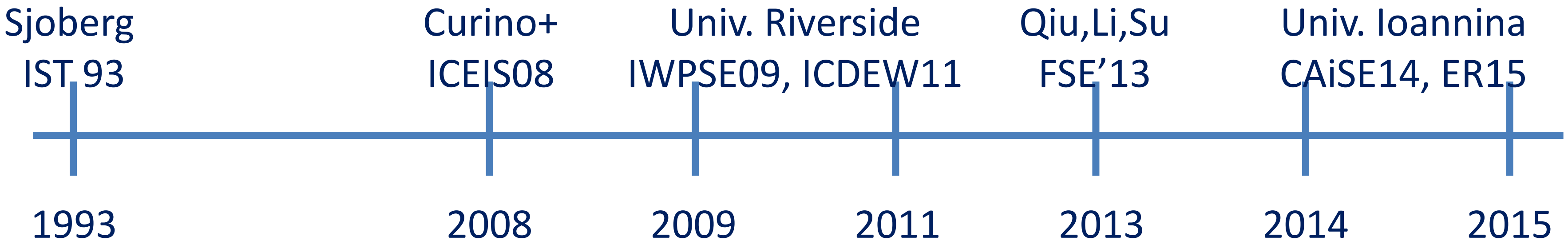
- Internal validity concerns the accuracy of cause-effect statements: “change in A => change in B”
- **We are very careful to avoid making strong causation statements!**
  - In some places, we just hint that we suspect the causes for a particular phenomenon, in some places in the text, but we have no data, yet, to verify our gut-feeling.
  - And yes, it is quite possible that our correlations hide cofounding variables.

Is there a theory?

- Our study should be regarded as a pattern observer, rather than as a collection of laws, coming with their internal mechanics and architecture.
- It will take too many studies (to enlarge the representativeness even more) and more controlled experiments (in-depth excavation of cause-effect relationships) to produce a solid theory.
- **It would be highly desirable if a clear set of requirements on the population definition, the breadth of study and the experimental protocol could be solidified by the scientific community (like e.g., the TREC benchmarks)**
- … and of course, there might be other suggestions on how to proceed…

RELATED WORK

Timeline of empirical studies



Sjoberg
IST 93
1993

Curino+
ICEIS08
2008

Univ. Riverside
IWPSE09, ICDEW11
2009
2011

Qiu,Li,Su
FSE'13
2013

Univ. Ioannina
CAiSE14, ER15
2014
2015

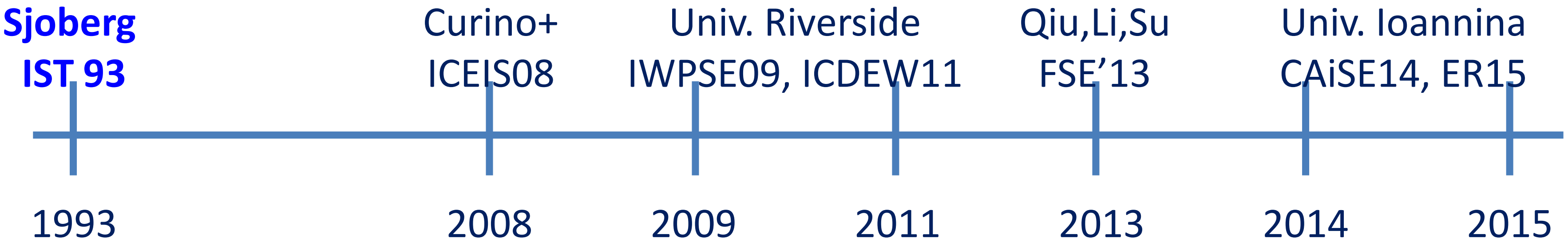
Timeline of empirical studies

Sjoberg @ IST 93: 18 months study of a health system.
139% increase of #tables ; 274% increase of the #attributes

Changes in the code (on avg):

- relation addition: 19 changes ; attribute additions: 2 changes
- relation deletion : 59.5 changes; attribute deletions: 3.25 changes

An **inflating period** during construction where almost all changes were additions, and a **subsequent period** where additions and deletions where balanced.



Sjoberg
IST 93
1993

Curino+
ICEIS08
2008

Univ. Riverside
IWPSE09, ICDEW11
2009
2011

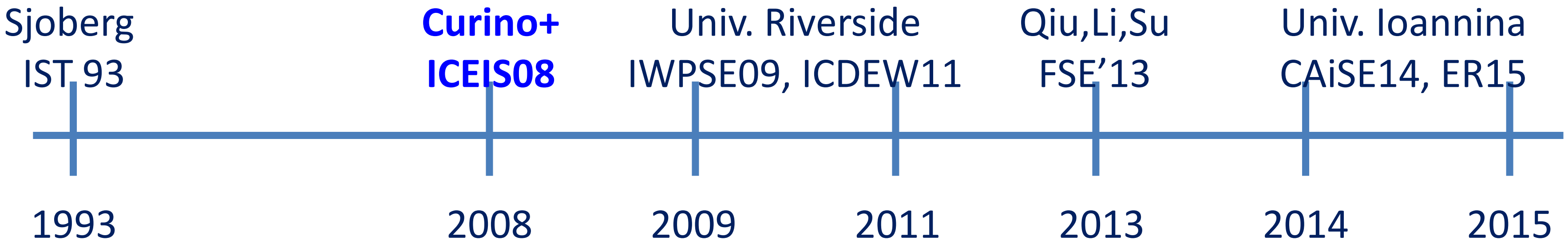
Qiu,Li,Su
FSE'13
2013

Univ. Ioannina
CAiSE14, ER15
2014
2015

Timeline of empirical studies

Curino+ @ ICEIS08: Mediawiki for 4.5 years
100% increase in the number of tables
142% in the number of attributes.

45% of changes do not affect the information capacity of the schema (but are rather index adjustments, documentation, etc)



Sjoberg
IST 93

Curino+
ICEIS08

Univ. Riverside
IWPSE09, ICDEW11

Qiu,Li,Su
FSE'13

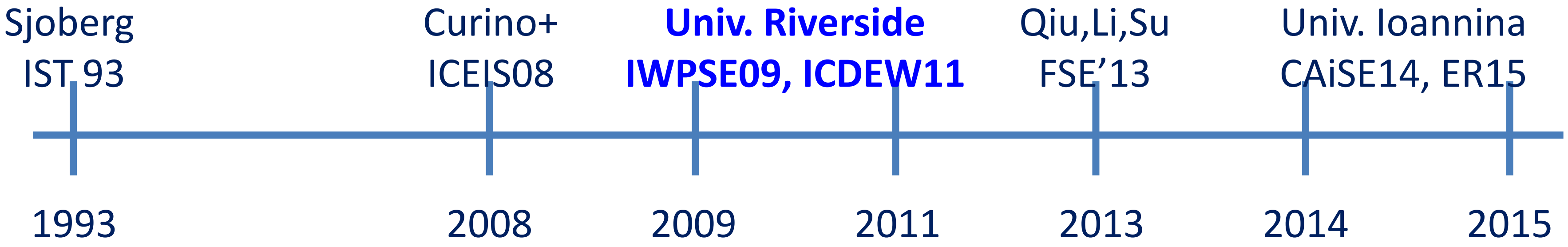
Univ. Ioannina
CAiSE14, ER15

1993 2008 2009 2011 2013 2014 2015

Timeline of empirical studies

IWPSE09: Mozilla and Monotone (a version control system)
Many ways to be out of synch between code and evolving db schema

ICDEW11: Firefox, Monotone , Biblioteq (catalogue man.) , Vienna (RSS)
Similar pct of changes with previous work
Frequency and timing analysis: **db schemata tend to stabilize over time**, as there is more change at the beginning of their history, but seem to converge to a relatively fixed structure later



Sjoberg
IST 93

Curino+
ICEIS08

Univ. Riverside
IWPSE09, ICDEW11

Qiu,Li,Su
FSE'13

Univ. Ioannina
CAiSE14, ER15

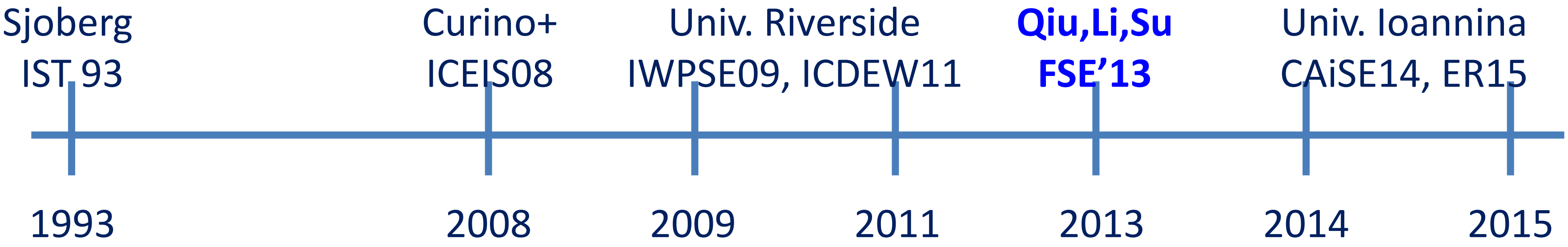
1993 2008 2009 2011 2013 2014 2015

Timeline of empirical studies

Qiu,Li,Su@ FSE 2013: 10 (!) database schemata studied.
Change is focused both (a) with respect to time and (b) with respect to the tables who change.

Timing: 7 out of 10 databases reached 60% of their schema size within 20% of their early lifetime.
Change is frequent in the early stages of the databases, with inflationary characteristics; then, the schema evolution process calms down.

Tables that change: 40% of tables do not undergo any change at all, and 60%-90% of changes pertain to 20% of the tables (in other words, 80% of the tables live quiet lives). The most frequently modified tables attract 80% of the changes.



Sjoberg
IST 93
1993

Curino+
ICEIS08
2008

Univ. Riverside
IWPSE09, ICDEW11
2009
2011

Qiu,Li,Su
FSE'13
2013

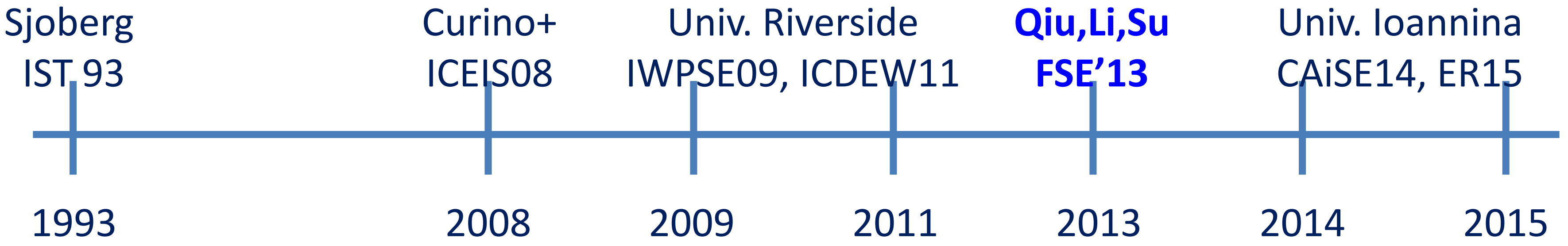
Univ. Ioannina
CAiSE14, ER15
2014
2015

Timeline of empirical studies

Qiu,Li,Su@ FSE 2013: **Code and db co-evolution, not always in synch**.

- Code and db changed in the same revision: 50.67% occasions
- Code change was in a previous/subsequent version than the one where the database schema change: 16.22% of occasions
- database changes not followed by code adaptation: 21.62% of occasions
- 11.49% of code changes were unrelated to the database evolution.

Each atomic change at the schema level is estimated to result in 10 -- 100 lines of application code been updated;
A valid db revision results in 100 -- 1000 lines of application code being updated



Sjoberg
IST 93

Curino+
ICEIS08

Univ. Riverside
IWPSE09, ICDEW11

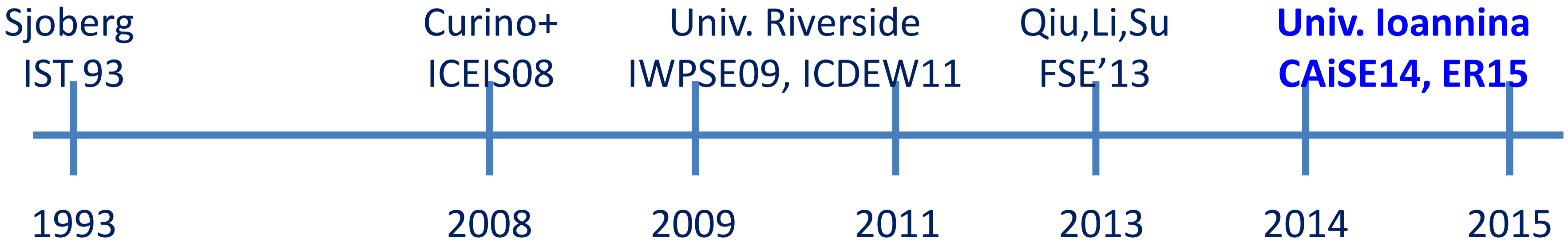
Qiu,Li,Su
FSE'13

Univ. Ioannina
CAiSE14, ER15

Timeline of empirical studies

CAiSE14: DB level

ER’15: Table level



Sjoberg
IST 93

Curino+
ICEIS08

Univ. Riverside
IWPSE09, ICDEW11

Qiu,Li,Su
FSE’13

Univ. Ioannina
CAiSE14, ER15

1993

2008

2009

2011

2013

2014

2015

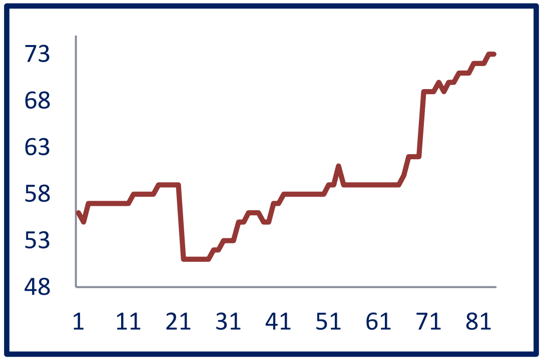
CAISE 14 / INF. SYSTEMS 15

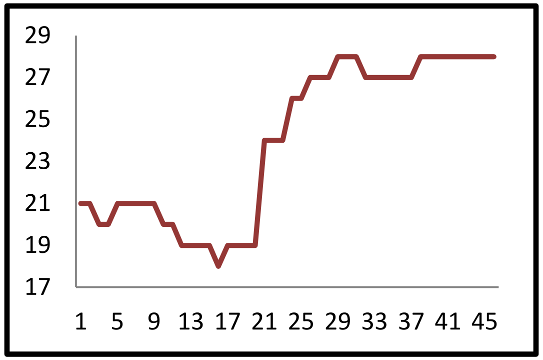
Datasets

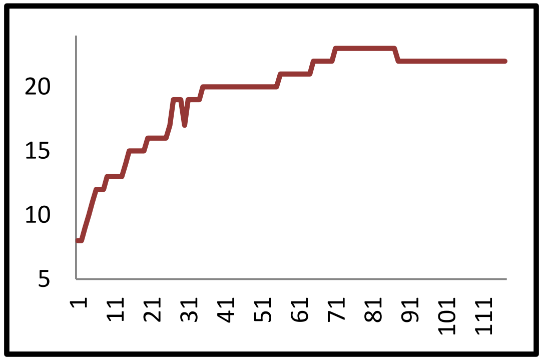
https://github.com/DAINTINESS-Group/EvolutionDatasets

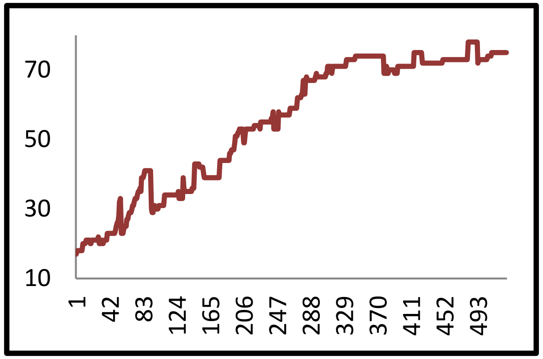
- Content management Systems
  - MediaWiki, TYPO3, Coppermine, phpBB, OpenCart
- Medical Databases
  - Ensemble, BioSQL
- Scientific
  - ATLAS Trigger

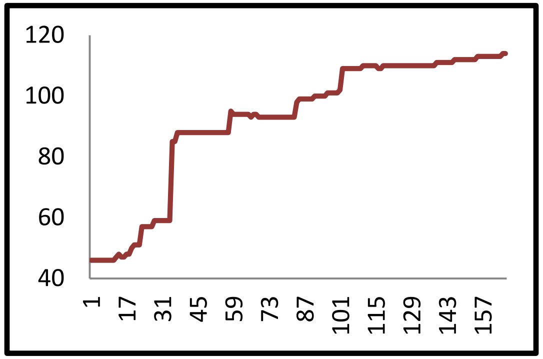
Schema Size (relations)

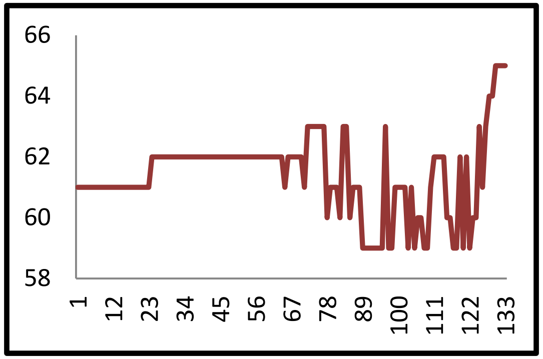


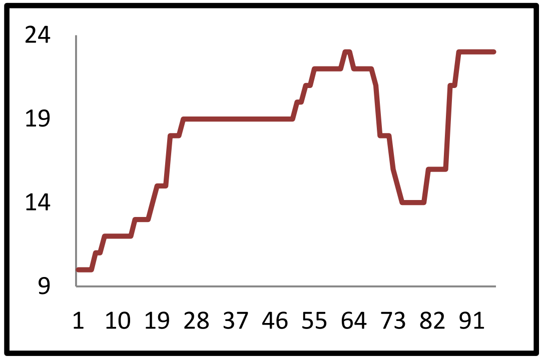


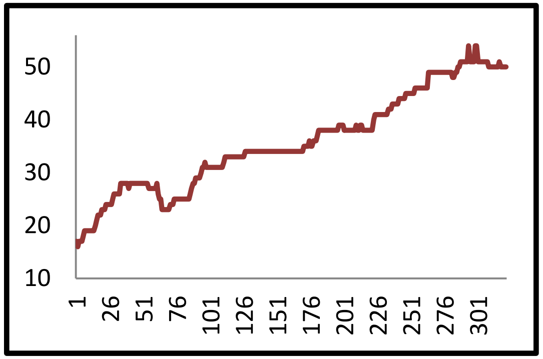












CaiSE’14: Main results

Schema size (#tables, #attributes) supports the assumption of a feedback mechanism

- Schema size grows over time; not continuously, but with bursts of concentrated effort
- Drops in schema size signifies the existence of perfective maintenance
- Regressive formula for size estimation holds, with a quite short memory

Schema Growth (diff in size between subsequent versions) is small!!

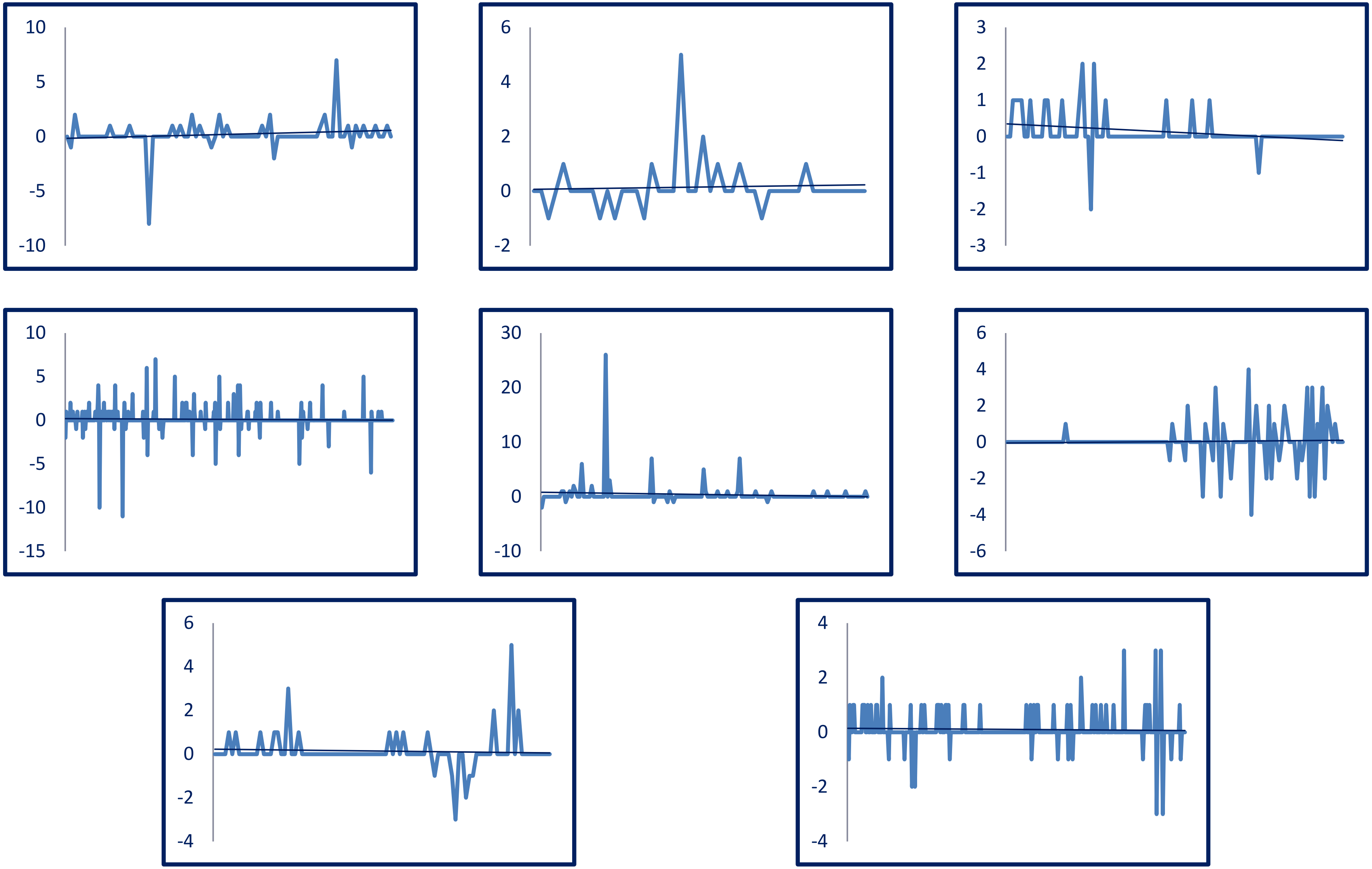
- Growth is small, smaller than in typical software
- The number of changes for each evolution step follows Zipf’s law around zero
- Average growth is close (slightly higher) to zero

Grey for results requiring further search

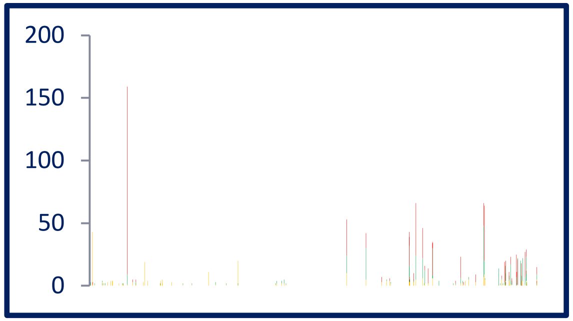
Patterns of change: no consistently constant behavior

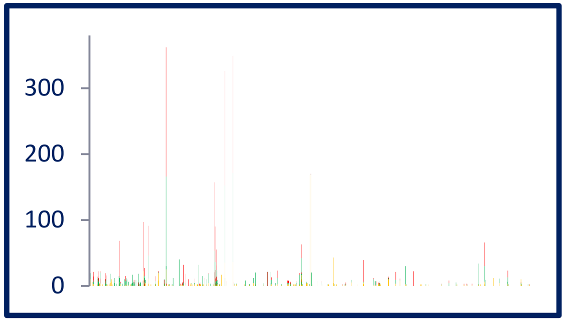
- Changes reduce in density as databases age
- Change follows three patterns: **Stillness**, **Abrupt change** (up or down), **Smooth growth upwards**
- Change frequently follows **spike** patterns
- **Complexity** does **not** increase with age

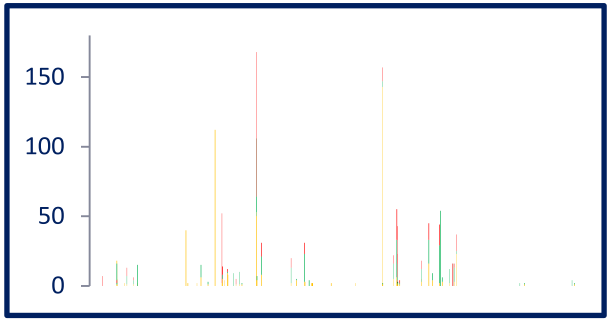
Schema Growth (diff in #tables)

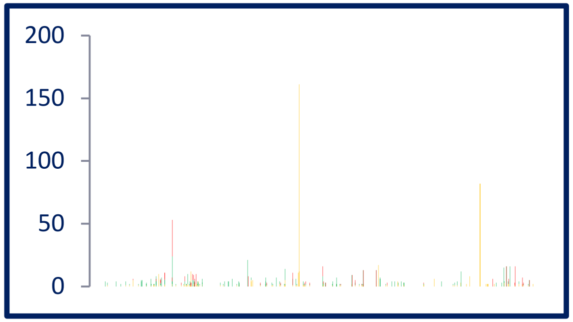


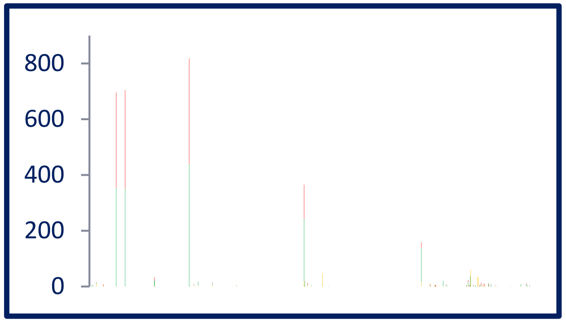
Change over time

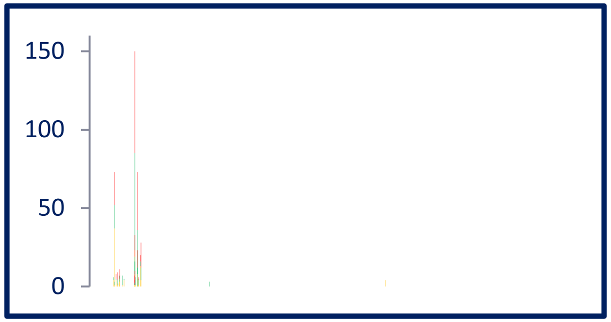


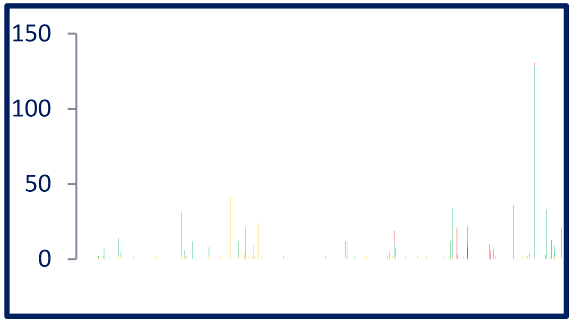


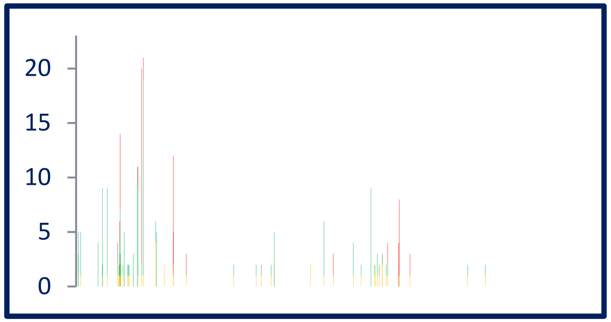








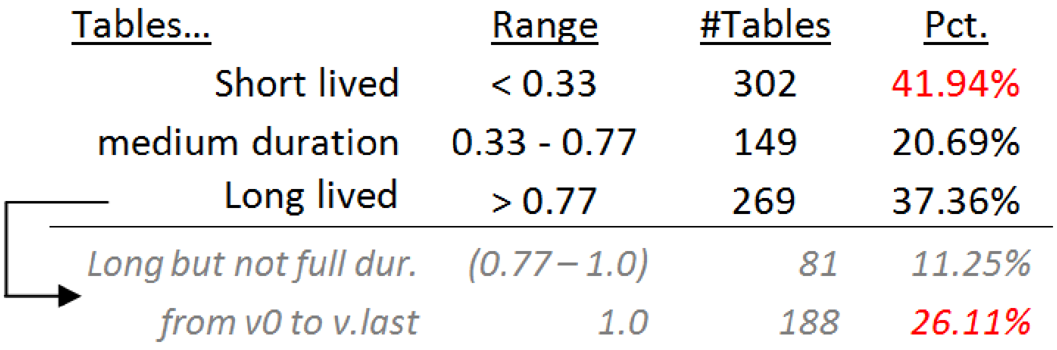




STATS

Statistical study of durations

- Short and long lived tables are practically equally proportioned
- Medium size durations are fewer than the rest!
- Long lived tables are mostly survivors (see on the right)



Tables…	Range	#Tables	Pct.
Short lived	< 0.33	302	41.94%
medium duration	0.33 - 0.77	149	20.69%
Long lived	> 0.77	269	37.36%
Long but not full dur.	*(0.77 – 1.0)*	*81*	*11.25%*
from v0 to v.last	*1.0*	*188*	*26.11%*

One of the fascinating revelations of this measurement was that there is a 26.11% fraction of tables that appeared in the beginning of the database and survived until the end.

In fact, if a table is long-lived there is a 70% chance (188 over 269 occasions) that it has appeared in the beginning of the database.

Tables are mostly thin

- On average, half of the tables (approx. 47%) are thin tables with less than 5 attributes.
- The tables with 5 to 10 attributes are approximately one third of the tables' population
- The large tables with more than 10 attributes are approximately 17% of the tables.

Pct of tables with num. of attributes …

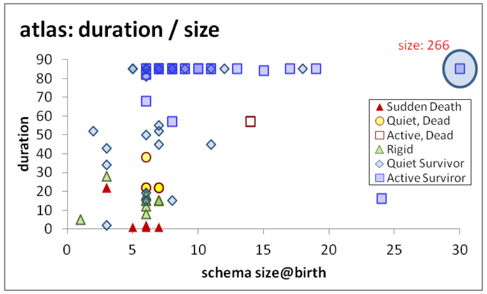
	<5	5-10	>10
atlas	10,23%	68,18%	21,59%
biosql	75,56%	24,44%	0,00%
coppermine	52,17%	30,43%	17,39%
ensembl	54,84%	38,06%	7,10%
mediawiki	61,97%	19,72%	18,31%
phpbb	40,00%	44,29%	15,71%
typo3	21,88%	31,25%	46,88%
opencart	57,20%	33,05%	9,75%
Average	46,73%	36,18%	17,09%

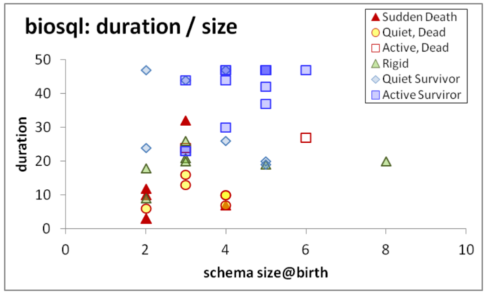
THE FOUR PATTERNS

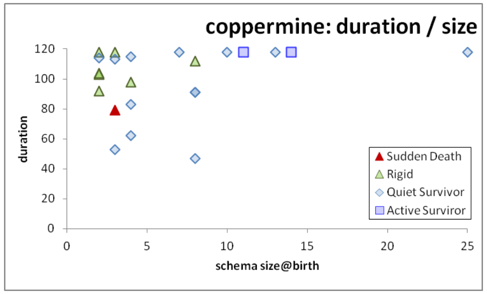
Schema size @ birth / duration

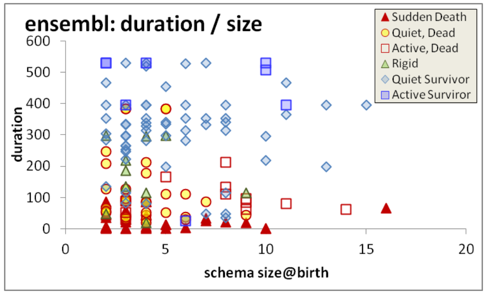
Only the thin die young, all the wide ones seem to live forever

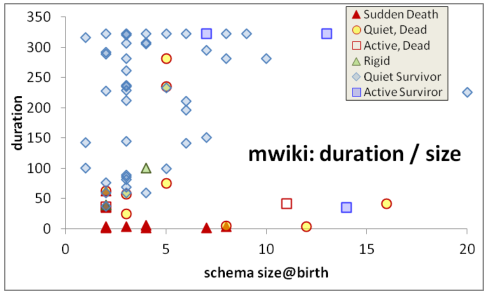
THE GAMMA PATTERN

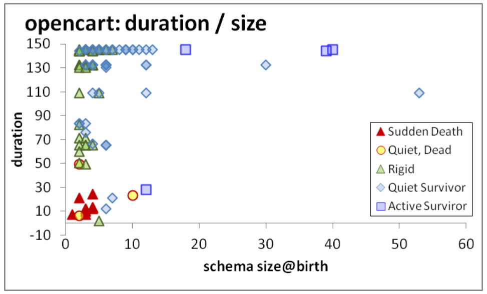


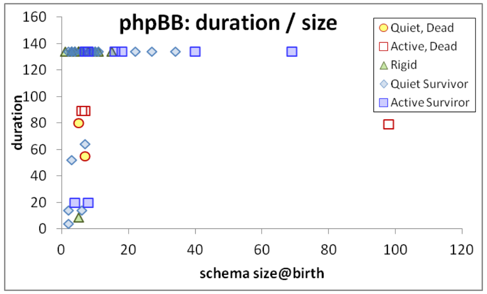


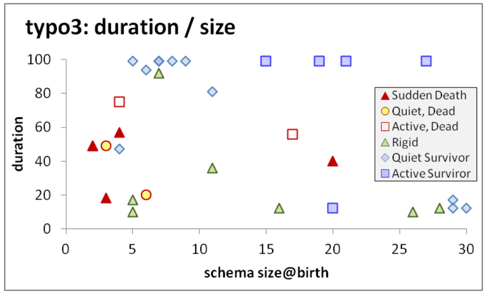












Exceptions

- Biosql: nobody exceeds 10 attributes
- Ensembl, mwiki: very few exceed 10 attributes, 3 of them died
- typo: has many late born survivors

Stats on wide tables and their survival

			As pct over #Tables…		*As pct over the set of Wide Tables …*		
	#	# Wide		…Wide of long	…	… Early Born	… of Long
	Tables	tables	…Wide	duration	Survivors	& Survivors	Duration
coppermine	23	4	17%	17%	100%	100%	100%
phpBB	70	11	16%	14%	91%	91%	91%
opencart*	128	12	9%	7%	100%	75%	75%
atlas	88	14	16%	11%	86%	71%	71%
typo3	32	15	47%	13%	87%	33%	27%
mwiki	71	6	8%	1%	50%	33%	17%
ensembl	155	9	6%	0%	67%	56%	0%
biosql	45	0	0%	0%	NA	NA	NA

Definitions:

Wide schema: strictly above 10 attributes.

The top band of durations (the upper part of the Gamma shape): the upper 10% of the values in the y-axis.

Early born table: ts birth version is in the lowest 33% of versions;

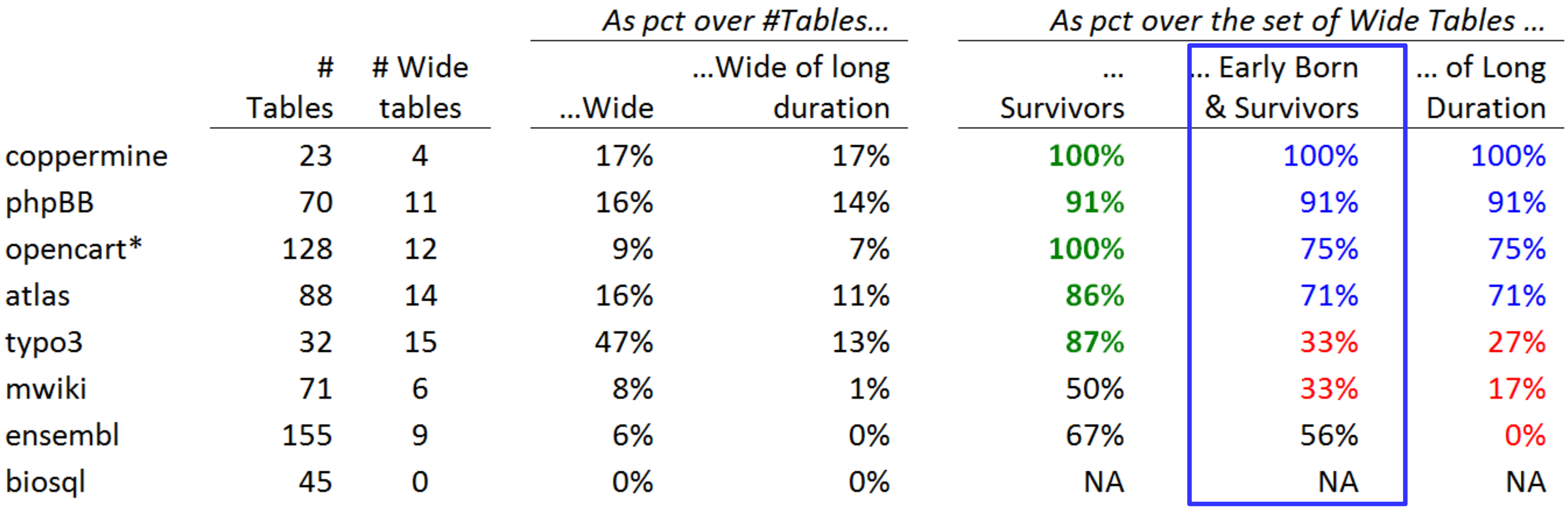
Late-comers: born after the 77% of the number of versions.

Whenever a table is wide, its chances of surviving are high

			As pct over #Tables...		*As pct over the set of Wide Tables ...*		
	# Tables	# Wide tables	...Wide	...Wide of long duration	... Survivors	... Early Born & Survivors	... of Long Duration
coppermine	23	4	17%	17%	100%	100%	100%
phpBB	70	11	16%	14%	91%	91%	91%
opencart*	128	12	9%	7%	100%	75%	75%
atlas	88	14	16%	11%	86%	71%	71%
typo3	32	15	47%	13%	87%	33%	27%
mwiki	71	6	8%	1%	50%	33%	17%
ensembl	155	9	6%	0%	67%	56%	0%
biosql	45	0	0%	0%	NA	NA	NA

Apart from mwiki and ensembl, all the rest of the data sets *confirm the hypothesis with a percentage higher than 85%*. The two exceptions are as high as 50% for their support to the hypothesis.

Wide tables are frequently created early on and are not deleted afterwards



			As pct over #Tables...		*As pct over the set of Wide Tables ...*		
	# Tables	# Wide tables	...Wide	...Wide of long duration	... Survivors	... Early Born & Survivors	... of Long Duration
coppermine	23	4	17%	17%	100%	100%	100%
phpBB	70	11	16%	14%	91%	91%	91%
opencart*	128	12	9%	7%	100%	75%	75%
atlas	88	14	16%	11%	86%	71%	71%
typo3	32	15	47%	13%	87%	33%	27%
mwiki	71	6	8%	1%	50%	33%	17%
ensembl	155	9	6%	0%	67%	56%	0%
biosql	45	0	0%	0%	NA	NA	NA

Early born, wide, survivor tables (as a percentage over the set of wide tables).

- in half the data sets the percentage is above 70%

- in two of them the percentage of these tables is one third of the wide tables.

Whenever a table is wide, its duration frequently lies within the top-band of durations (upper part of Gamma)

			As pct over #Tables...		*As pct over the set of Wide Tables ...*		
	# Tables	# Wide tables	...Wide	...Wide of long duration	... Survivors	... Early Born & Survivors	... of Long Duration
coppermine	23	4	17%	17%	100%	100%	100%
phpBB	70	11	16%	14%	91%	91%	91%
opencart*	128	12	9%	7%	100%	75%	75%
atlas	88	14	16%	11%	86%	71%	71%
typo3	32	15	47%	13%	87%	33%	27%
mwiki	71	6	8%	1%	50%	33%	17%
ensembl	155	9	6%	0%	67%	56%	0%
biosql	45	0	0%	0%	NA	NA	NA

What is probability that a wide table belongs to the upper part of the Gamma?

- there is a very strong correlation between the two last columns: the Pearson correlation is 88% overall; 100% for the datasets with high pct of early born wide tables.

-

- *Bipolarity on this pattern: half the cases support the pattern with support higher than 70%, whereas the rest of the cases clearly disprove it, with very low support values.*

Long-lived & wide => early born and survivor

			As pct over #Tables…		*As pct over the set of Wide Tables …*		
	# Tables	# Wide tables	…Wide	…Wide of long duration	… Survivors	… Early Born & Survivors	… of Long Duration
coppermine	23	4	17%	17%	100%	100%	100%
phpBB	70	11	16%	14%	91%	91%	91%
opencart*	128	12	9%	7%	100%	75%	75%
atlas	88	14	16%	11%	86%	71%	71%
typo3	32	15	47%	13%	87%	33%	27%
mwiki	71	6	8%	1%	50%	33%	17%
ensembl	155	9	6%	0%	67%	56%	0%
biosql	45	0	0%	0%	NA	NA	NA

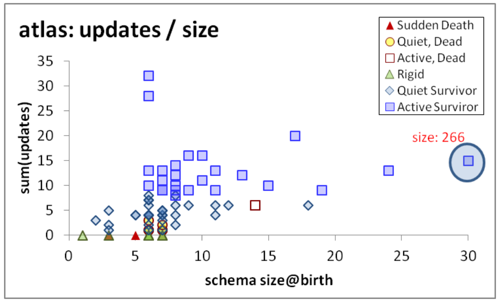
Subset relationship

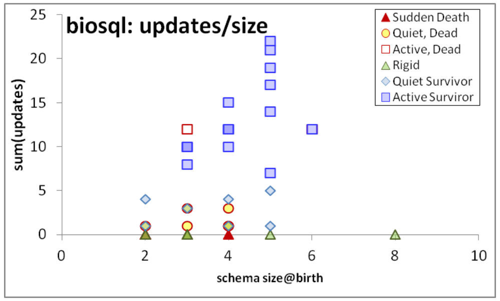
In all data sets, if a wide table has a long duration within the upper part of the Gamma, this deterministically (100% of all data sets) signifies that the table was also early born and survivor.

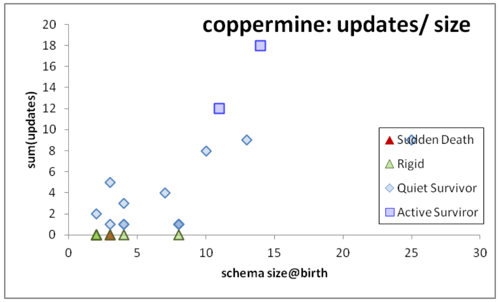
If a wide table is in the top of the Gamma line, it is deterministically an early born survivor.

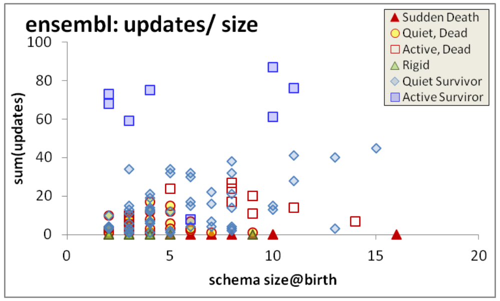
Schema size and updates

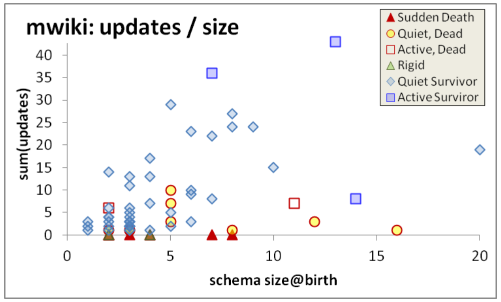
THE COMET PATTERN

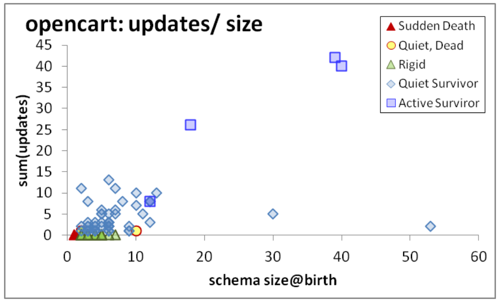

atlas: updates / size

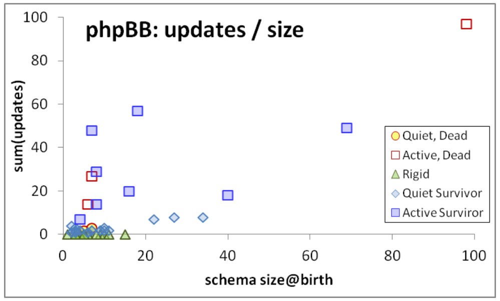

biosql: updates/size

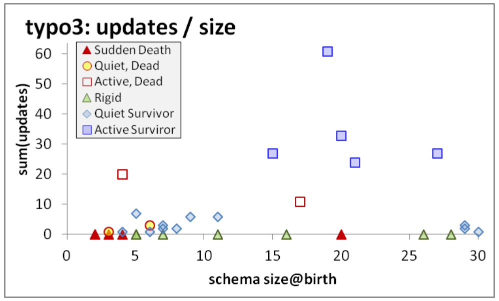

coppermine: updates/ size


ensembl: updates/ size

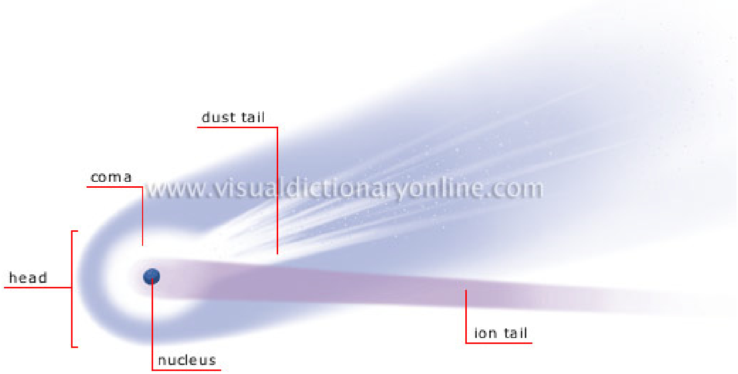

mwiki: updates / size


opencart: updates/ size


phpBB: updates / size


typo3: updates / size

http://visual.merriam-webster.com/astronomy/celestial-bodies/comet.php



Comets have two tails: White one is made of comet dust particles. Blue one is made of electrically charged gas.

The coma is the cloud of comet dust particles surrounding the nucleus.

Nucleus is solid, icy heart of comet, inside the cloud of the coma

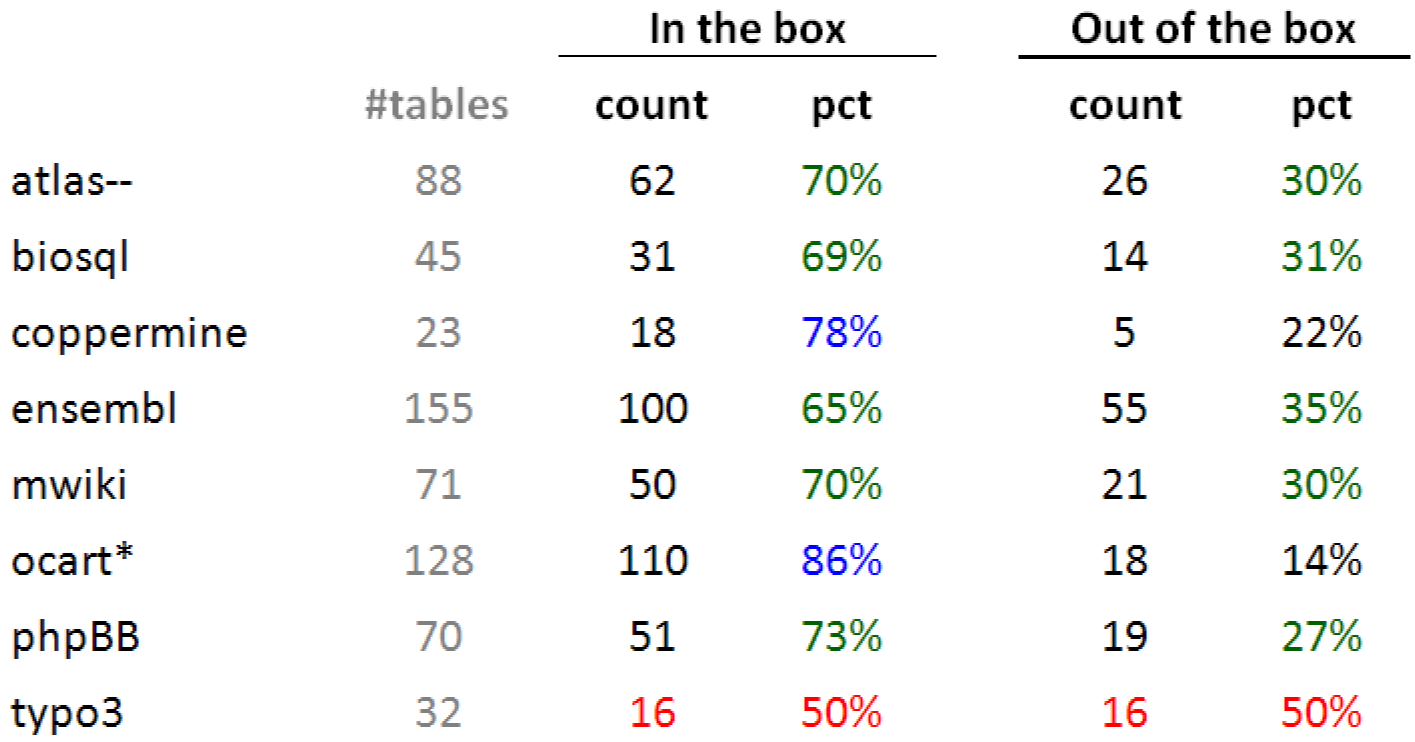
http://spaceplace.nasa.gov/comet-nucleus/en/

Statistics of schema size at birth and sum of updates

		Schema size at birth					Sum of updates				
	#tables	max	mean (μ)	stdev (σ)	median	mode	max	mean (μ)	stdev (σ)	median	mode
atlas--	87 / 88	24	7.53	3.67	7	6	32	5.86	11.81	4	0
biosql	45	8	3.6	1.37	3	2	22	5.38	11.91	1	0
coppermine	23	25	6.52	5.35	4	2	18	3.3	7.98	1	0
ensembl	155	16	4.98	2.98	4	3	87	10.38	27.05	3	0
mwiki	71	20	4.79	3.64	3	3	43	6.92	16.03	3	0
ocart*	128	53	5.73	7.02	4	3	42	2.56	8.56	0	0
phpBB	70	98	9.39	14.63	5	3	97	6.33	22.17	0.5	0
typo3	32	30	12.69	9.26	8.5	4	61	7.53	20.89	1.5	0

/ atlas: excluded table l1_prescale_set from the analysis (266 attributes; second largest value: 24); open cart: after version 22*/*

Typically: ~70% of tables inside the box

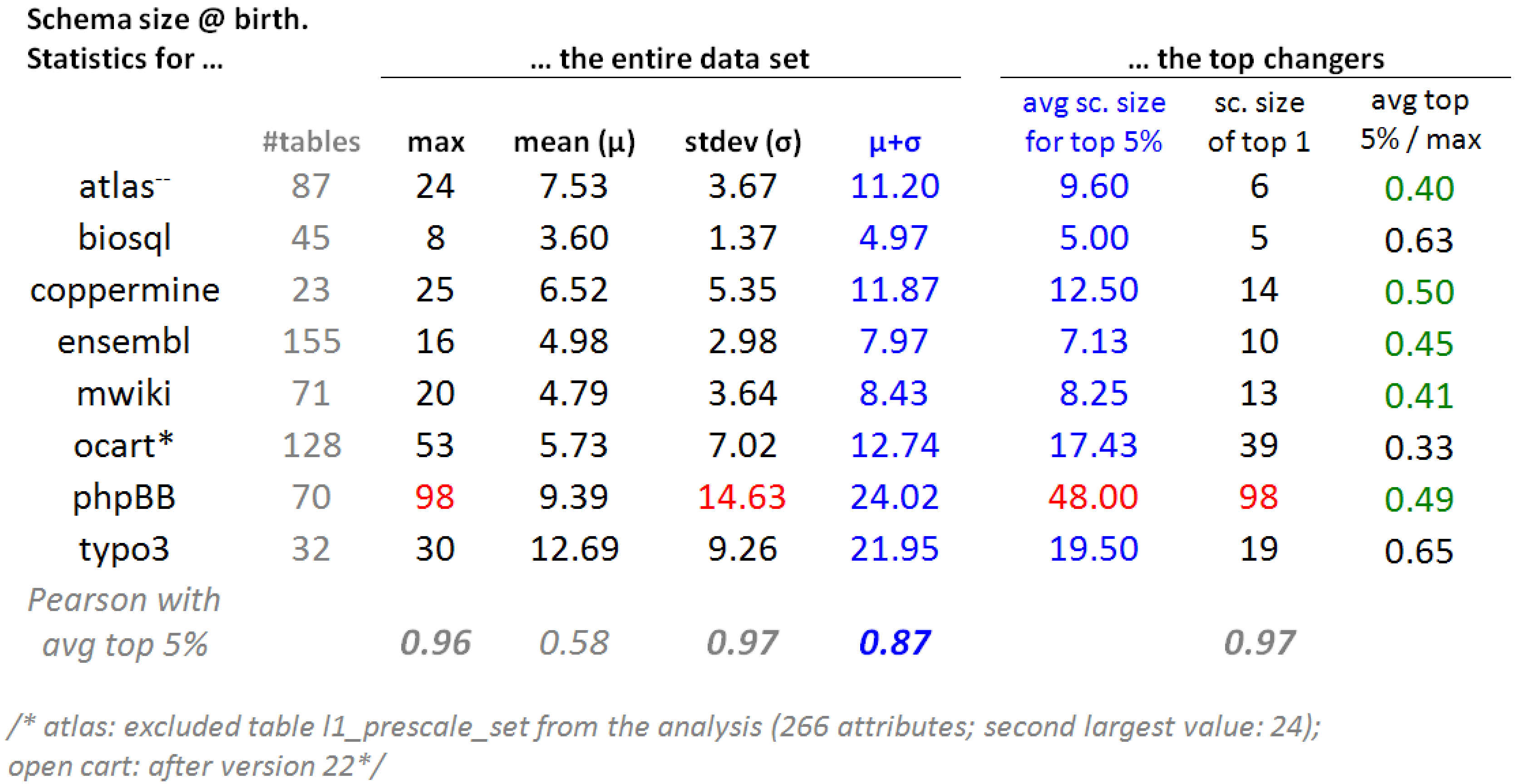


		In the box		Out of the box	
	#tables	count	pct	count	pct
atlas--	88	62	70%	26	30%
biosql	45	31	69%	14	31%
coppermine	23	18	78%	5	22%
ensembl	155	100	65%	55	35%
mwiki	71	50	70%	21	30%
ocart*	128	110	86%	18	14%
phpBB	70	51	73%	19	27%
typo3	32	16	50%	16	50%

/ atlas: excluded table l1_prescale_set from the analysis (266 attributes; second largest value: 24); open cart: after version 22*/*

Typically, around 70% of the tables of a database is found within the 10x10 box of *schemaSize@birth* x *sumOfUpdates* (10 excluded in both axes).

Top changers tend to have medium schema sizes

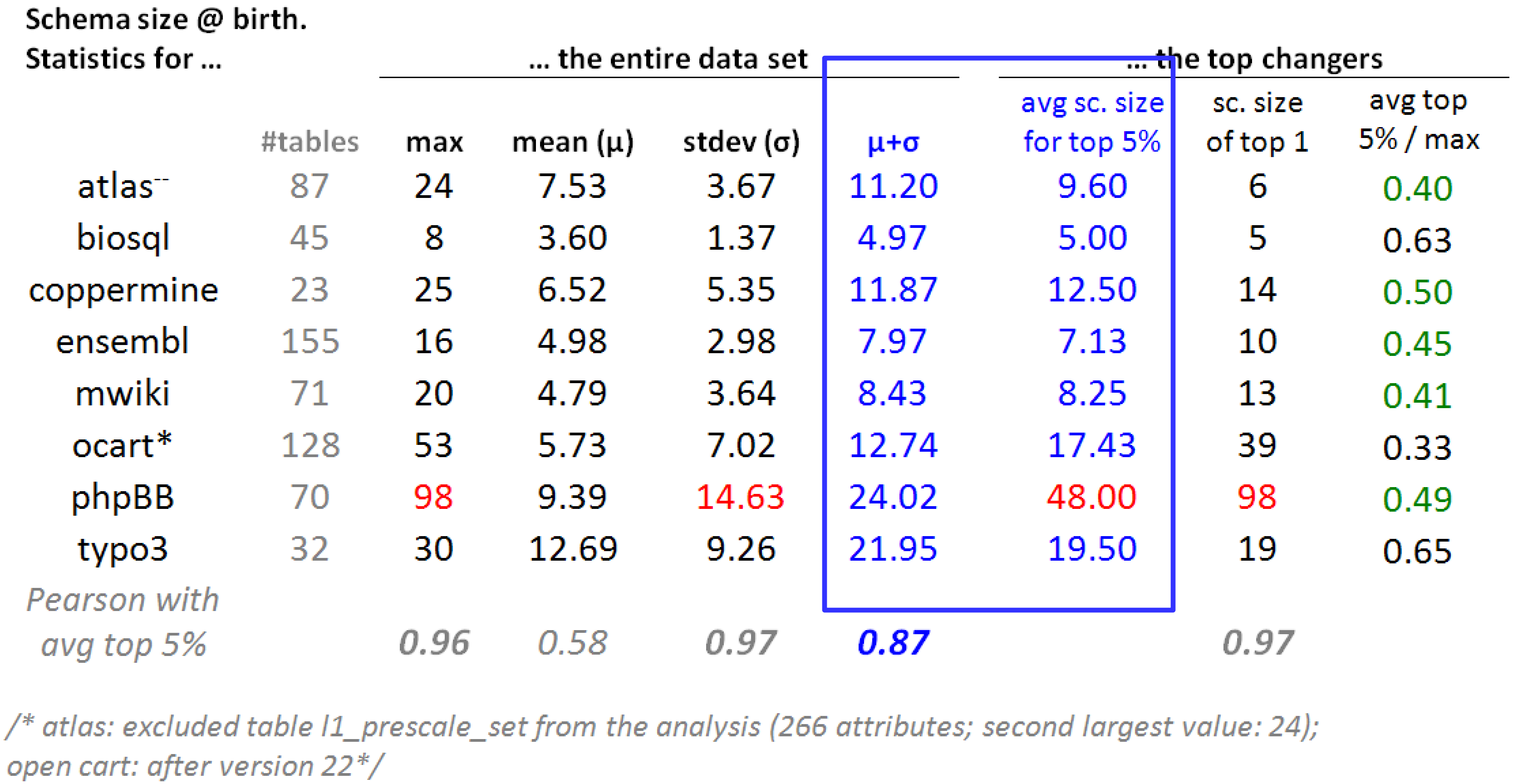


Schema size @ birth. Statistics for ...		... the entire data set				... the top changers		
	#tables	max	mean (μ)	stdev (σ)	μ+σ	avg sc. size for top 5%	sc. size of top 1	avg top 5% / max
atlas^{--}	87	24	7.53	3.67	11.20	9.60	6	0.40
biosql	45	8	3.60	1.37	4.97	5.00	5	0.63
coppermine	23	25	6.52	5.35	11.87	12.50	14	0.50
ensembl	155	16	4.98	2.98	7.97	7.13	10	0.45
mwiki	71	20	4.79	3.64	8.43	8.25	13	0.41
ocart*	128	53	5.73	7.02	12.74	17.43	39	0.33
phpBB	70	98	9.39	14.63	24.02	48.00	98	0.49
typo3	32	30	12.69	9.26	21.95	19.50	19	0.65
Pearson with avg top 5%		***0.96***	***0.58***	***0.97***	***0.87***		***0.97***	

/ atlas: excluded table l1_prescale_set from the analysis (266 attributes; second largest value: 24); open cart: after version 22*/*

For every dataset: we selected the top 5% of tables in terms of this sum of updates and we averaged the schema size at birth of these top 5% tables.

Top changers tend to have medium schema sizes



Schema size @ birth. Statistics for ...		... the entire data set				... the top changers		
	#tables	max	mean (μ)	stdev (σ)	μ+σ	avg sc. size for top 5%	sc. size of top 1	avg top 5% / max
atlas^{--}	87	24	7.53	3.67	11.20	9.60	6	0.40
biosql	45	8	3.60	1.37	4.97	5.00	5	0.63
coppermine	23	25	6.52	5.35	11.87	12.50	14	0.50
ensembl	155	16	4.98	2.98	7.97	7.13	10	0.45
mwiki	71	20	4.79	3.64	8.43	8.25	13	0.41
ocart*	128	53	5.73	7.02	12.74	17.43	39	0.33
phpBB	70	98	9.39	14.63	24.02	48.00	98	0.49
typo3	32	30	12.69	9.26	21.95	19.50	19	0.65
Pearson with avg top 5%		***0.96***	***0.58***	***0.97***	***0.87***		***0.97***	

/ atlas: excluded table l1_prescale_set from the analysis (266 attributes; second largest value: 24); open cart: after version 22*/*

The average schema size for the top 5% of tables in terms of their update behavior is close to one standard deviation up from the average value of the schema size at birth(i.e., very close to $mu$+$sigma$). *//except phpBB*

Top changers tend to have medium schema sizes

Schema size @ birth. **Statistics for ...**		**... the entire data set**				**... the top changers**		
	#tables	max	mean (μ)	stdev (σ)	μ+σ	avg sc. size for top 5%	sc. size of top 1	avg top 5% / max
atlas^{--}	87	24	7.53	3.67	11.20	9.60	6	0.40
biosql	45	8	3.60	1.37	4.97	5.00	5	0.63
coppermine	23	25	6.52	5.35	11.87	12.50	14	0.50
ensembl	155	16	4.98	2.98	7.97	7.13	10	0.45
mwiki	71	20	4.79	3.64	8.43	8.25	13	0.41
ocart*	128	53	5.73	7.02	12.74	17.43	39	0.33
phpBB	70	98	9.39	14.63	24.02	48.00	98	0.49
typo3	32	30	12.69	9.26	21.95	19.50	19	0.65
Pearson with avg top 5%		***0.96***	***0.58***	***0.97***	***0.87***		***0.97***	

/ atlas: excluded table l1_prescale_set from the analysis (266 attributes; second largest value: 24); open cart: after version 22*/*

- In 5 out of 8 cases, the average schema size of top-changers within 0.4 and 0.5 of the maximum value (practically the middle of the domain) and never above 0.65 of it.

- Pearson: the maximum value, the standard deviation of the entire data set and the average of the top changers are very strongly correlated.

Wide tables have a medium number of updates

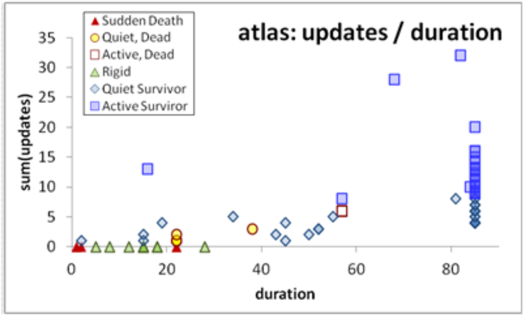
Total amt. of updates. Statistics for …		… the entire data set					… the top 5% with respect to schema size at birth (top wide)			
	#tables	max	mean (μ)	stdev (σ)	μ+σ	max/2	avg upd. of top 5%	upd. of top 1	avg of top 5% / max	Top up. in wide?
atlas	88	32	5.86	11.81	11.81	16.0	**12.60**	20	0.39	N
biosql	45	22	5.38	11.91	11.91	11.0	**8.00**	0	0.36	N
coppermine	23	18	3.30	7.98	7.98	9.0	**13.50**	9	0.75	Y
ensembl	155	87	10.38	27.05	27.05	43.5	**28.22**	0	0.32	N
mwiki	71	43	6.92	16.03	16.03	21.5	**17.75**	19	0.41	Y
ocart*	128	42	2.56	8.56	8.561	21.0	**14.55**	2	0.35	Y
phpBB	70	97	6.33	22.17	22.17	48.5	**43.00**	97	0.44	Y!
typo3	32	61	7.53	20.89	20.89	30.5	**2.00**	1	0.03	N
Pearson with avg top 5%			*0.27*	*0.59*	*0.50*	*0.74*		*0.79*		

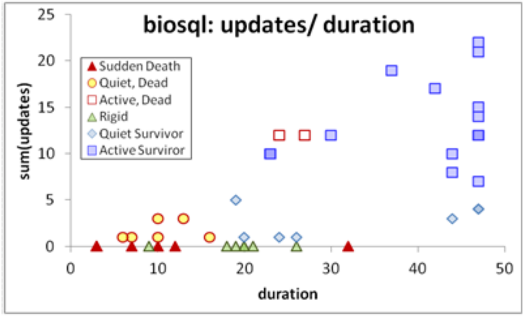
For each data set, we took the top 5% in terms of schema size at birth (**top wide**) and contrasted their update behavior wrt the update behavior of the entire data set.

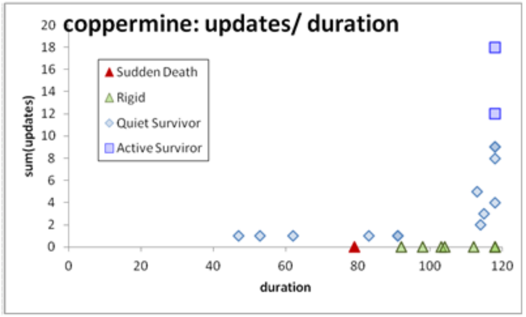
Typically, the avg. number of updates of the top wide tables is close to the 50% of the domain of values for the sum of updates (i.e., the middle of the y-axis of the comet figure, measuring the sum of updates for each table).

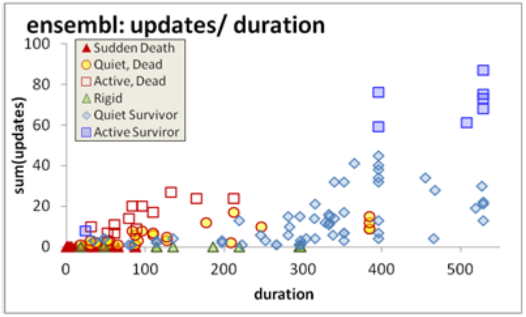
This is mainly due to the (very) large standard deviation (twice the mean), rather than the -- typically low -- mean value (due to the large part of the population living quiet lives).

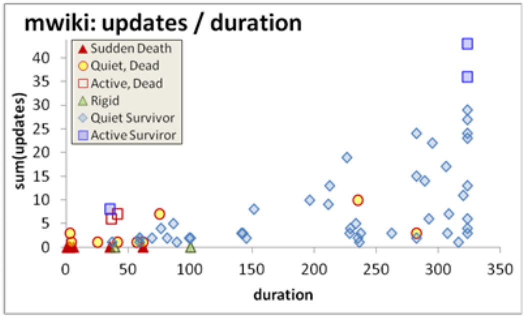
INVERSE GAMMA

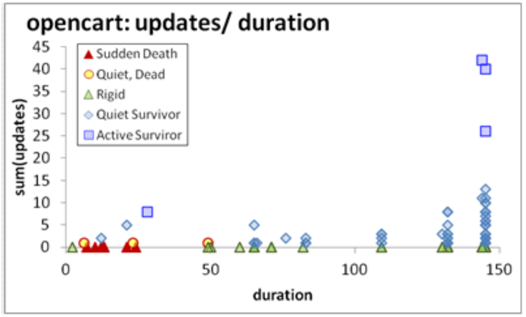

atlas: updates / duration

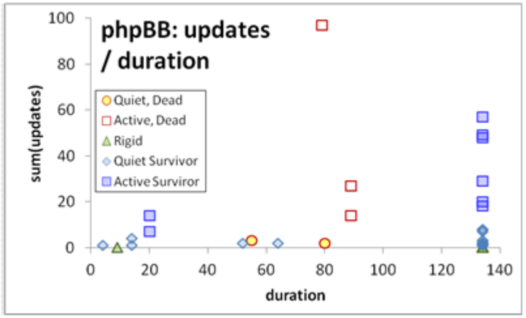

biosql: updates/ duration

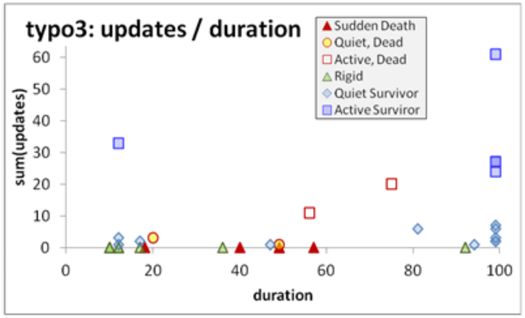

coppermine: updates/ duration


ensembl: updates/ duration

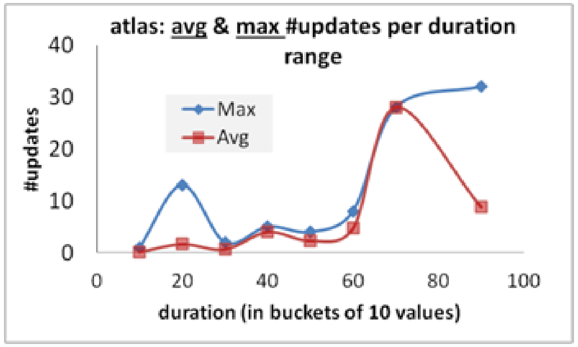

mwiki: updates / duration

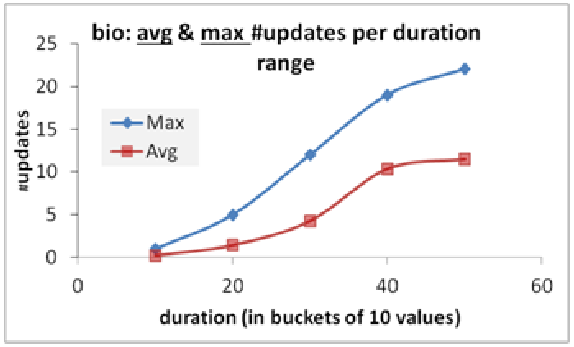

opencart: updates/ duration

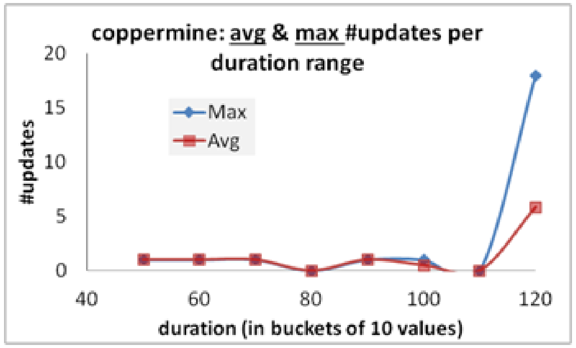

phpBB: updates / duration

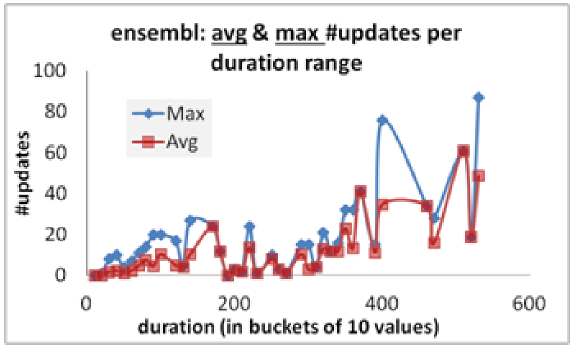

typo3: updates / duration

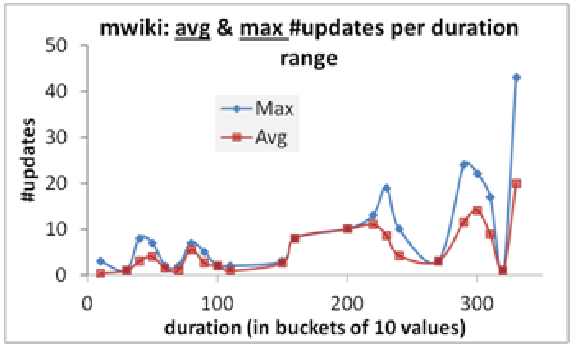
Skyline & Avg for Inverse Gamma

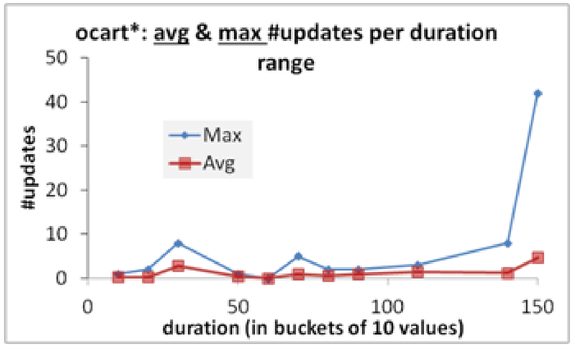


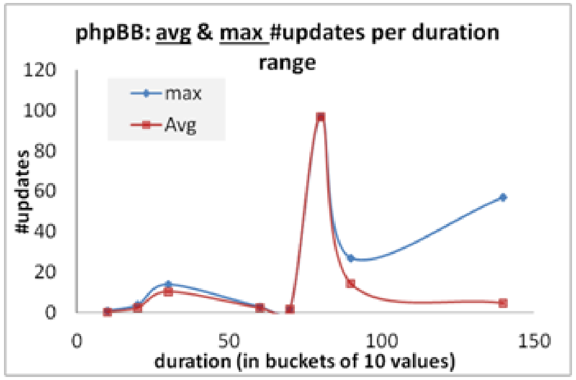


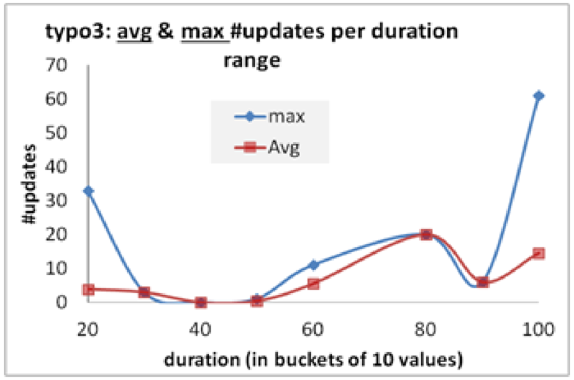




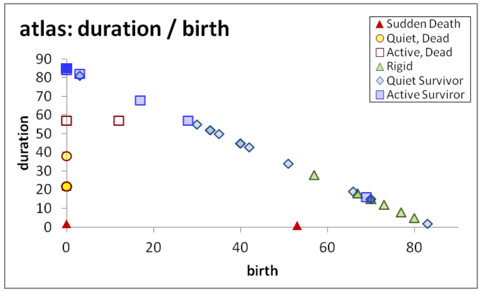


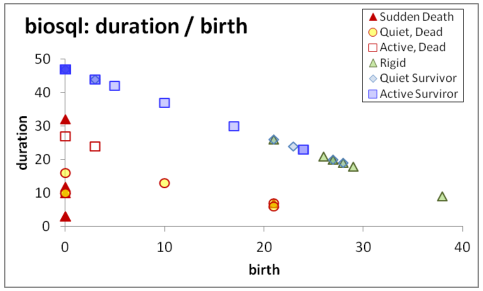


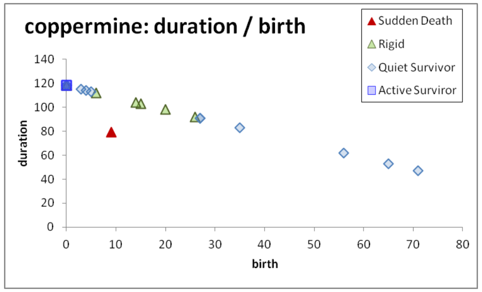


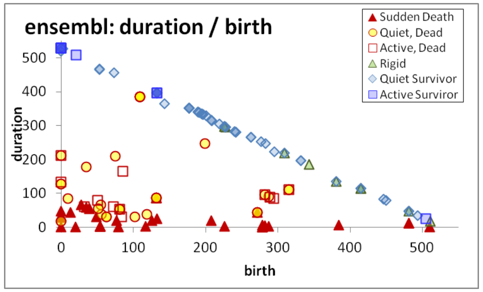


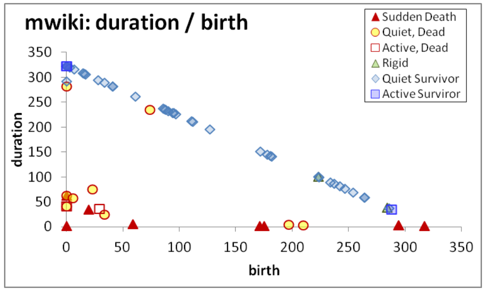
THE EMPTY TRIANGLE PATTERN

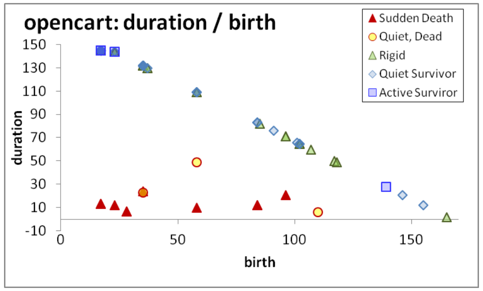

atlas: duration / birth

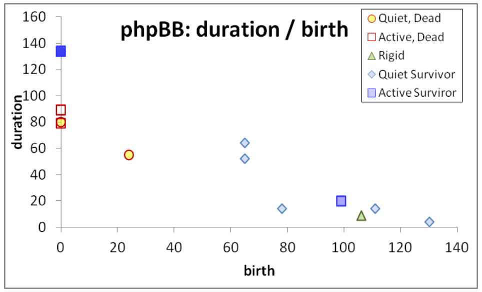

biosql: duration / birth

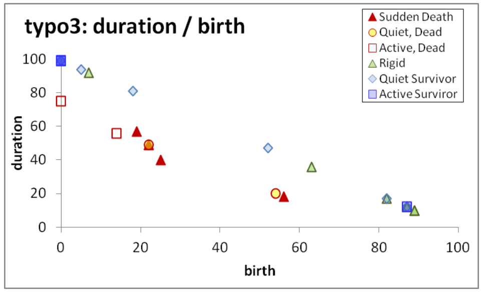

coppermine: duration / birth

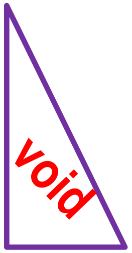

ensembl: duration / birth


mwiki: duration / birth


opencart: duration / birth


phpBB: duration / birth


typo3: duration / birth



Top changers: early born, survivors, often with long durations, and often all the above

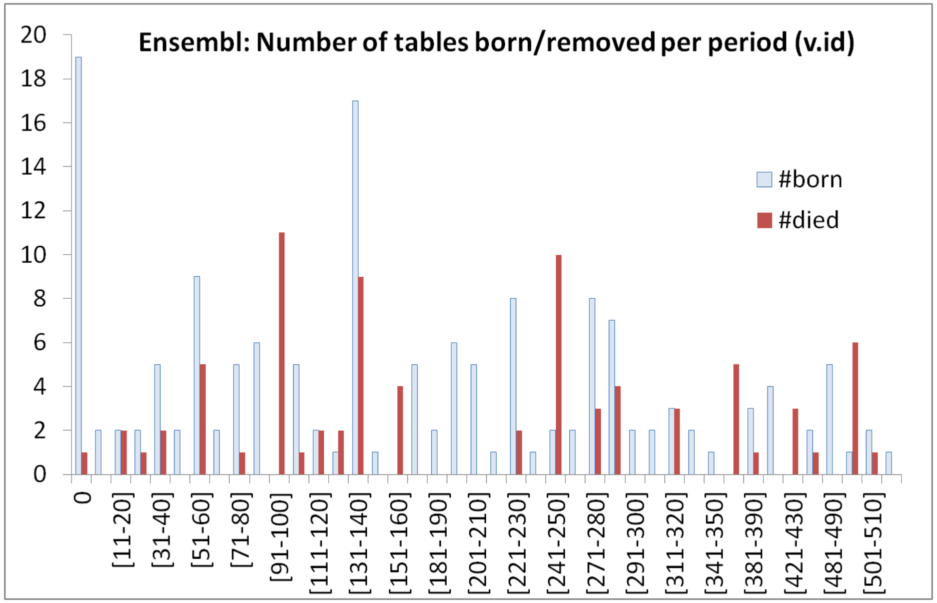
	atlas	biosql	coppermine	ensembl	mwiki	ocart*	phpBB	typo3
Tables	88	45	23	155	71	128	70	32
Active	27	16	2	23	5	4	11	7
active tables(%)	31%	36%	9%	15%	7%	3%	16%	22%
As percentages over active								
Born early	96%	81%	100%	78%	80%	75%	82%	86%
Survivors	93%	88%	100%	48%	60%	100%	73%	71%
Long duration	85%	69%	100%	22%	40%	75%	55%	57%
Born early, survive, live long	85%	69%	100%	22%	40%	75%	55%	57%

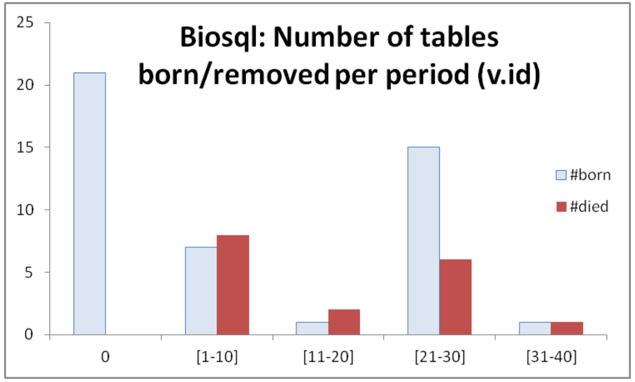
- In all data sets, active tables are born early with percentages that exceed 75%
- With the exceptions of two data sets, they survive with percentage higher than 70%.
- The probability of having a long duration is higher than 50% in 6 out of 8 data sets.
- Interestingly, **the two last lines are exactly the same sets of tables in all data sets!**
    - An active table with long duration has been born early and survived with prob. 100%
    - An active, survivor table that has a long duration has been born early with prob. 100%

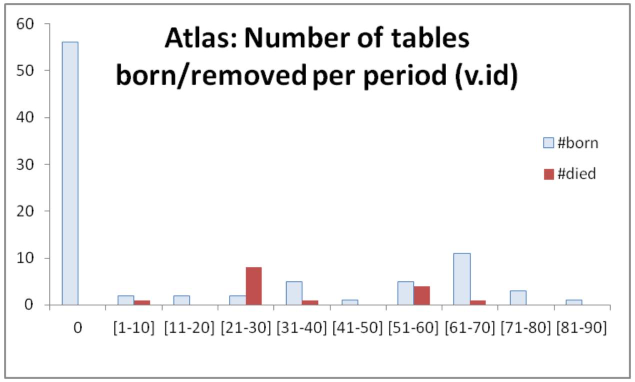
Dead are: quiet, early born, short lived, and quite often all three of them

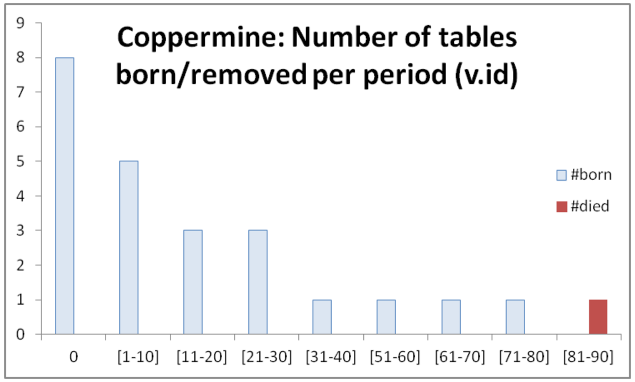
	atlas	biosql	coppermine	ensembl	mwiki	ocart*	phpBB	typo3
tables	88	45	23	155	71	128	70	32
dead	15	17	1	80	21	14	5	9
dead tables(%)	17%	38%	4%	52%	30%	11%	7%	28%
As percentages over # dead								
Few updates	87%	88%	100%	85%	90%	100%	40%	78%
Early born	80%	82%	100%	70%	62%	71%	100%	78%
Short-lived	80%	76%	0%	89%	90%	100%	0%	22%
Few upd's, early born, short duration	60%	59%	0%	51%	43%	71%	0%	0%
Do tables die of old age?								
long durations	48	14	18	13	23	86	57	12
long durations, dead	0	0	0	0	1	0	0	0
Dead among long-lived (%)	0%	0%	0%	0%	4%	0%	0%	0%

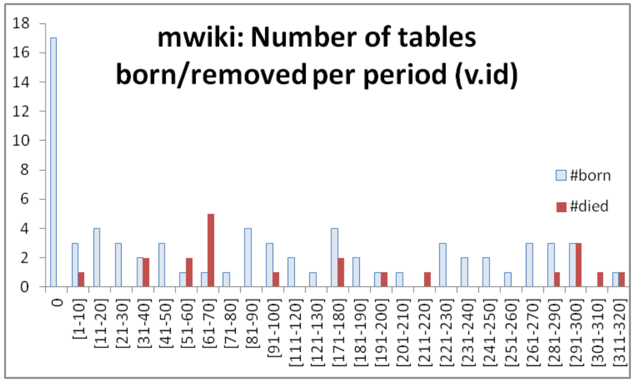
Most births &deaths occur early (usually)

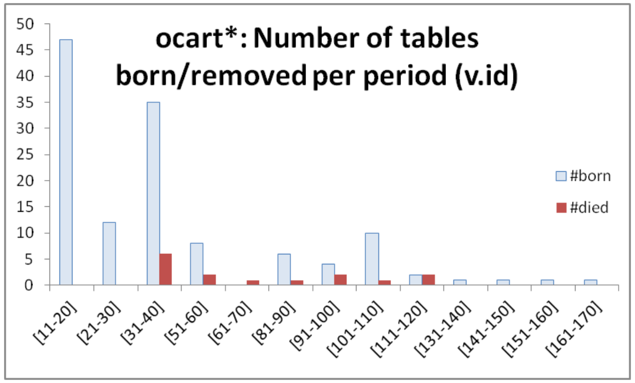


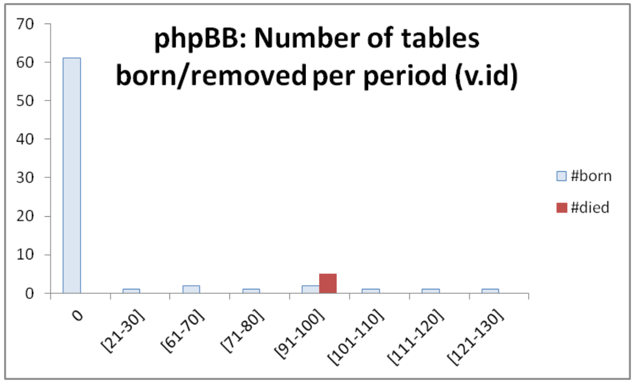


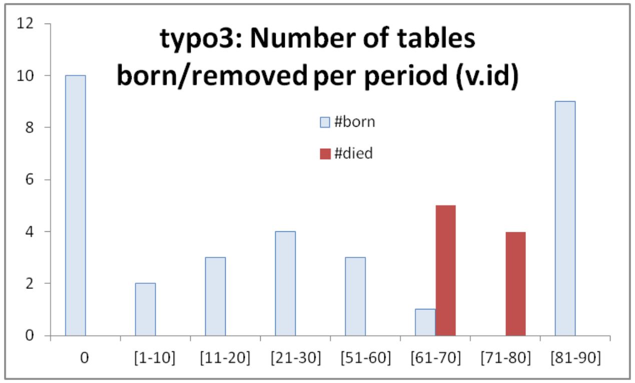


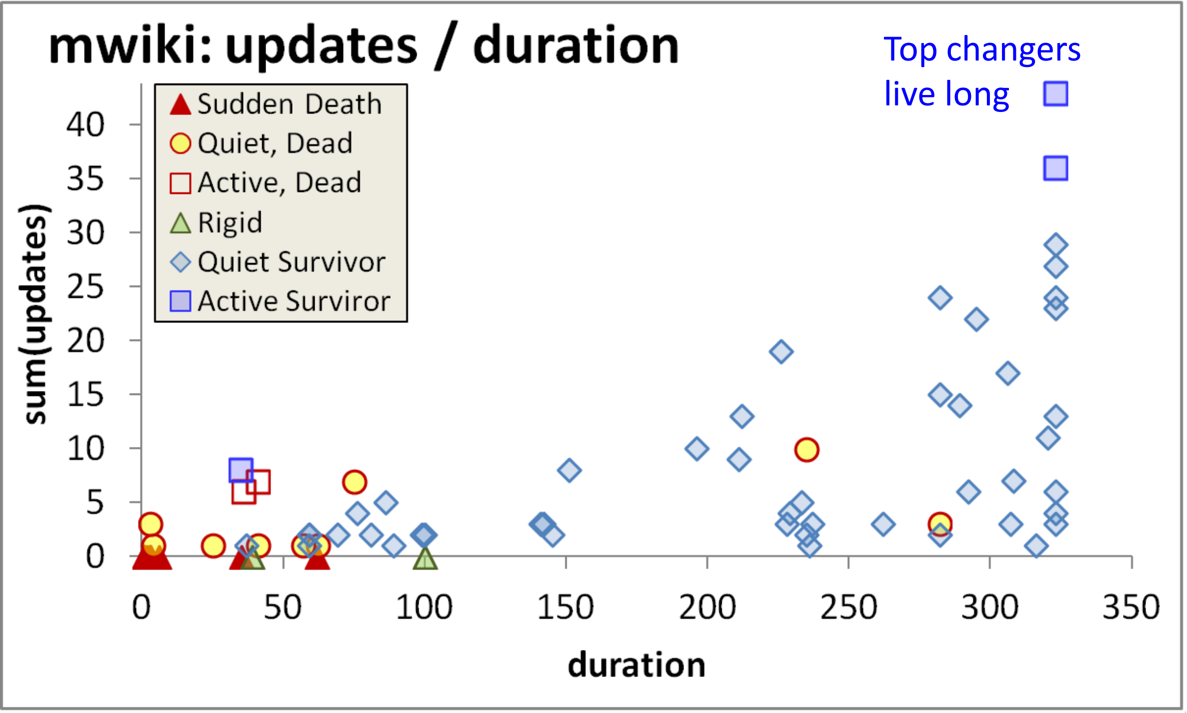




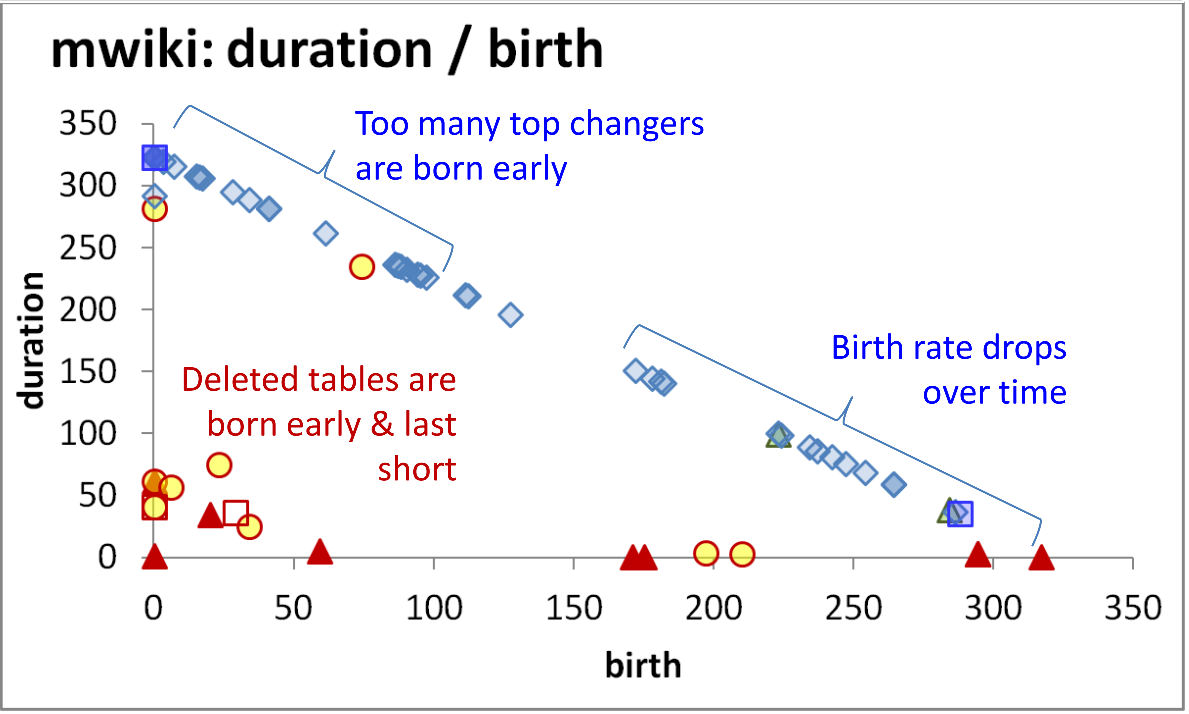







mwiki: updates / duration

Top changers live long


mwiki: duration / birth

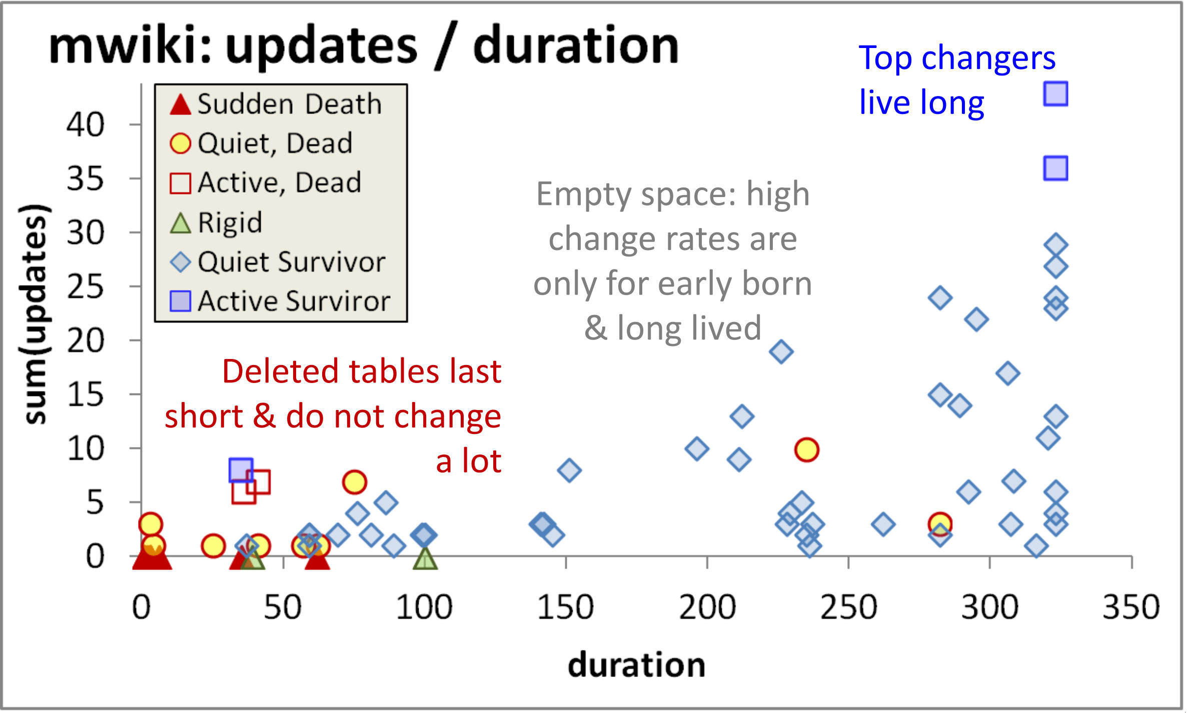
Too many top changers are born early

Deleted tables are born early & last short

Birth rate drops over time

Longevity and update activity correlate !!

- Remember: top changers are defined as such wrt ATU (AvgTrxnUpdate), not wrt sum(changes)
- Still, they dominate the sum(updates) too! (see top of inverse Γ)
- See also upper right blue part of diagonal: too many of them are born early and survive => live long!

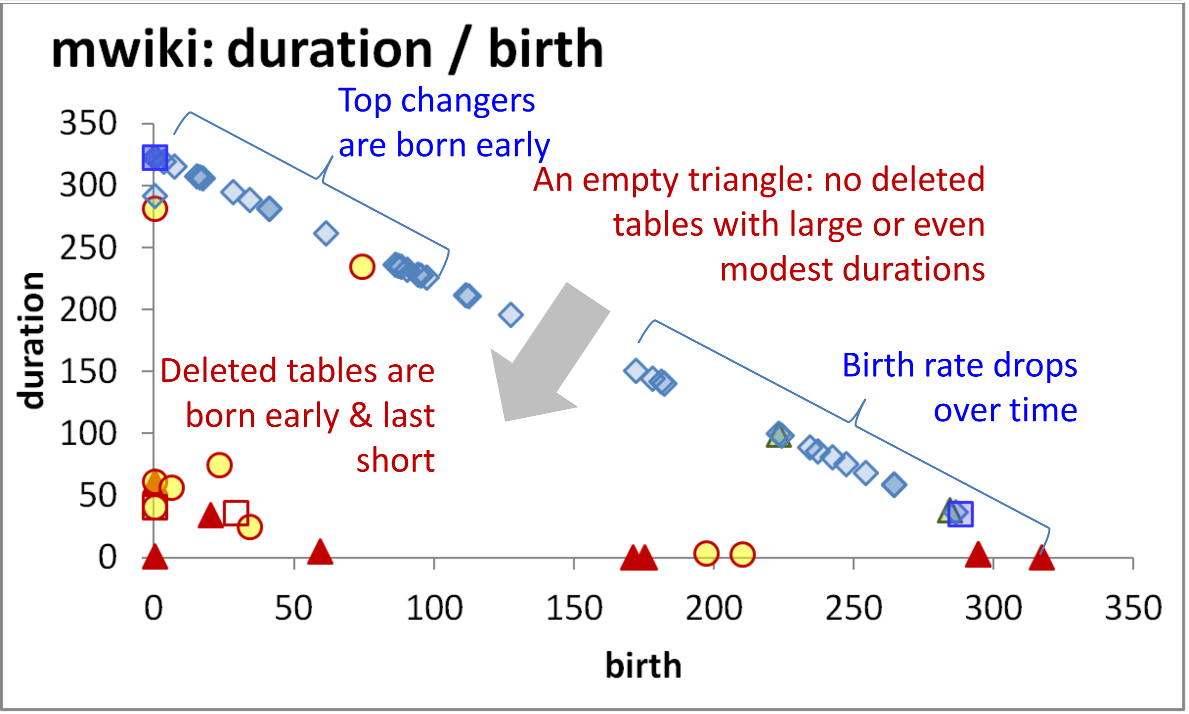


mwiki: updates / duration

Top changers live long

Empty space: high change rates are only for early born & long lived

Deleted tables last short & do not change a lot



mwiki: duration / birth

Top changers are born early

An empty triangle: no deleted tables with large or even modest durations

Deleted tables are born early & last short

Birth rate drops over time

All in one

- Early stages of the database life are more "active" in terms of births, deaths and updates, and have higher chances of producing deleted tables.
- After the first major restructuring, the database continues to grow; however, we see much less removals, and maintenance activity becomes more concentrated and focused.