Survival in schema evolution: putting the lives of **survivor** and **dead tables** in counterpoint

Panos Vassiliadis, Apostolos Zarras

Department of Computer Science and Engineering
University of Ioannina, Hellas

http://www.cs.uoi.gr/~pvassil/projects/schemaBiographies/

Why is schema evolution so important?

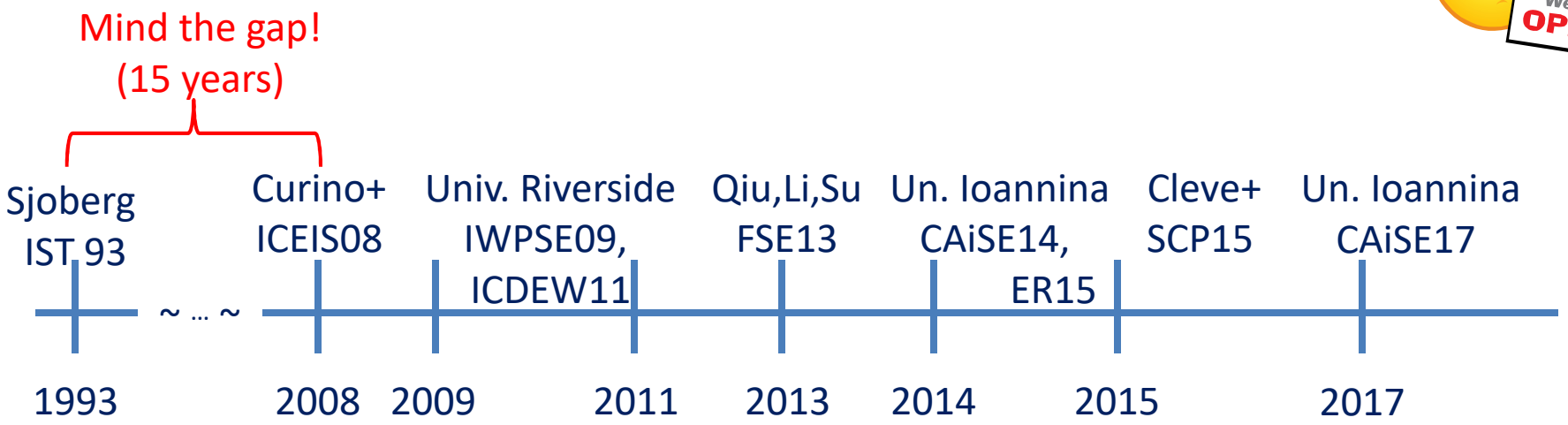
- Software and DB **maintenance** makes up for **at least 50% of all resources spent in a project**.
- **Databases are rarely stand-alone: typically, an entire ecosystem of applications is structured around them =>**
- **Changes in the schema can impact a large** (typically, not traced) **number of surrounding app's**, without explicit identification of the impact.

Is it possible to **"design for evolution"** and **minimize the impact of evolution** to the surrounding applications?

… But first, we need to know the "patterns of evolution" of relational schemata! …

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 &
- We are now presented with the opportunity to study the version histories of such "open source databases"



Our take on the problem

- To address the quest for finding patterns in the evolution of relational schemata, we have …
  - Collected version histories for the schemata of 8 open-source projects
    - CMS's: MediaWiki, TYPO3, Coppermine, phpBB, OpenCart
    - Physics: ATLAS Trigger
    - Biomed: 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)
  - Performed exploratory research where we statistically study / mine these measures, to **extract patterns & regularities for the lives of tables**

- **Available at:**

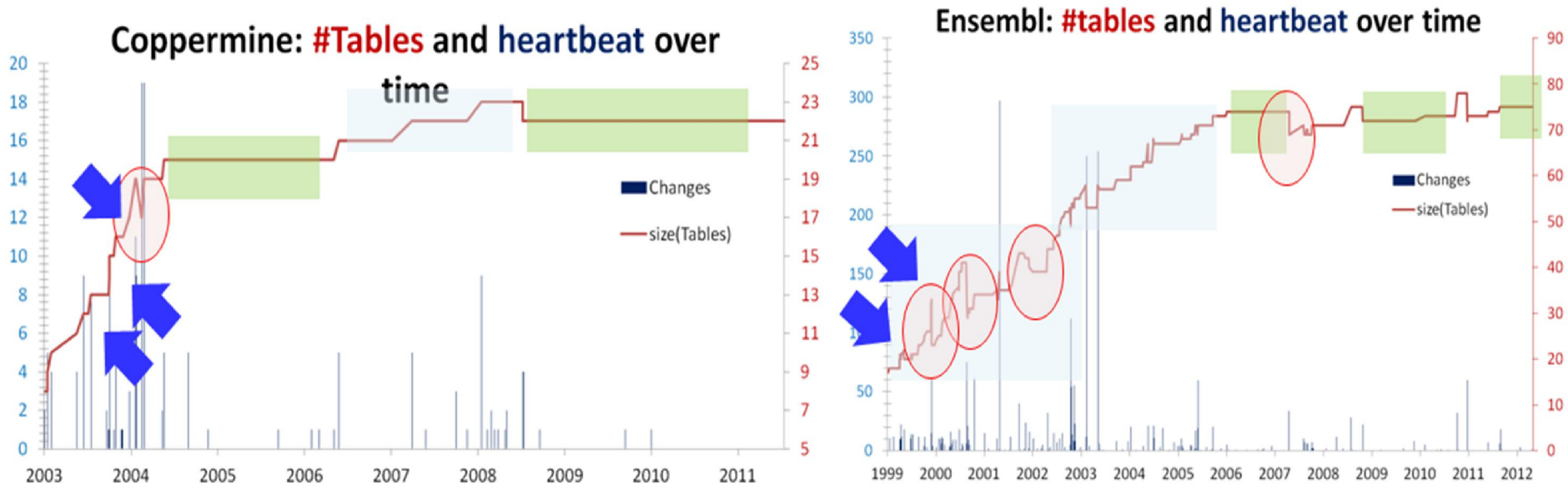
  https://github.com/DAINTINESS-Group/EvolutionDatasets

Scope of our studies

- **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!

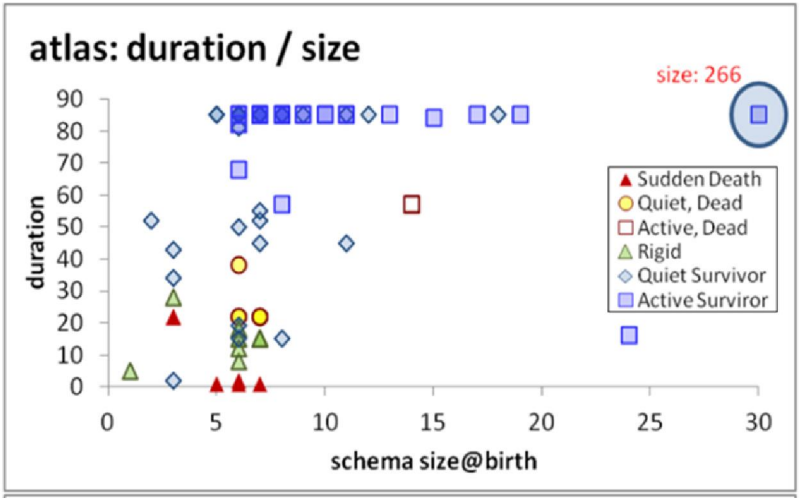
FoSS Dataset	Versions	Lifetime	Tables @ Start	Tables @ End
ATLAS Trigger [A]	84	2 Y, 7 M, 2 D	56	73
BioSQL [B]	46	10 Y, 6 M, 19 D	21	28
Coppermine [C]	117	8 Y, 6 M, 2 D	8	22
Ensembl [B]	528	13 Y, 3 M, 15 D	17	75
MediaWiki [C]	322	8 Y, 10 M, 6 D	17	50
OpenCart [C]	164	4 Y, 4 M, 3 D	46	114
phpBB [C]	133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

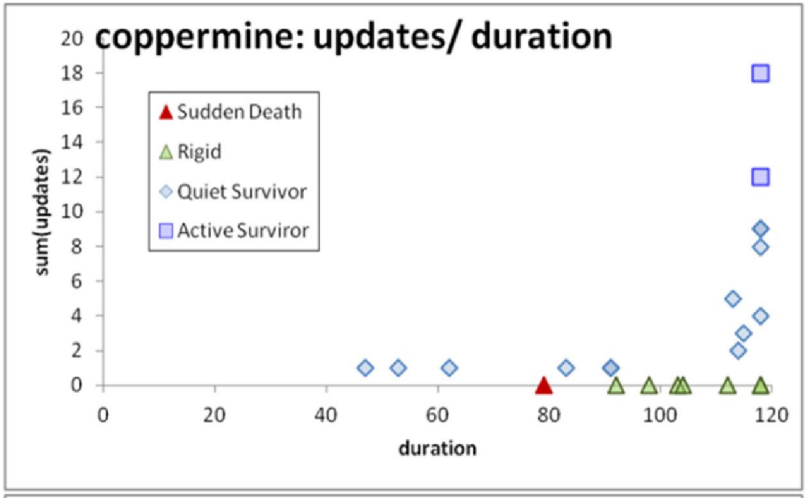
What we have found for schema evolution [CAiSE 14, IS 15]

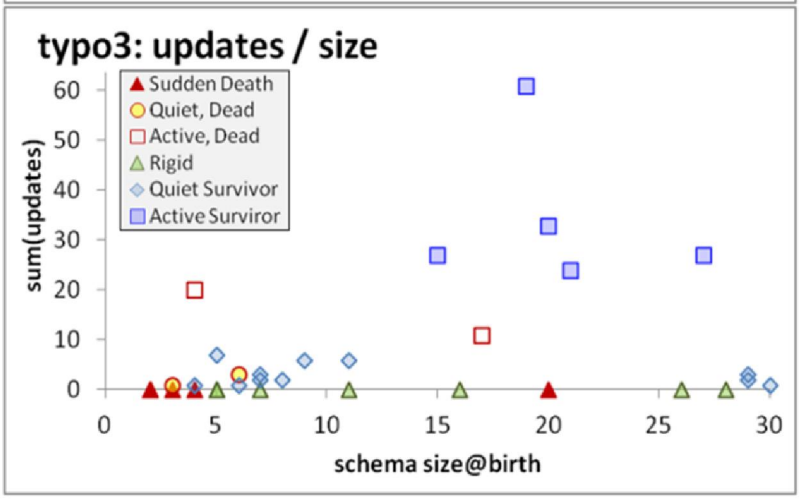


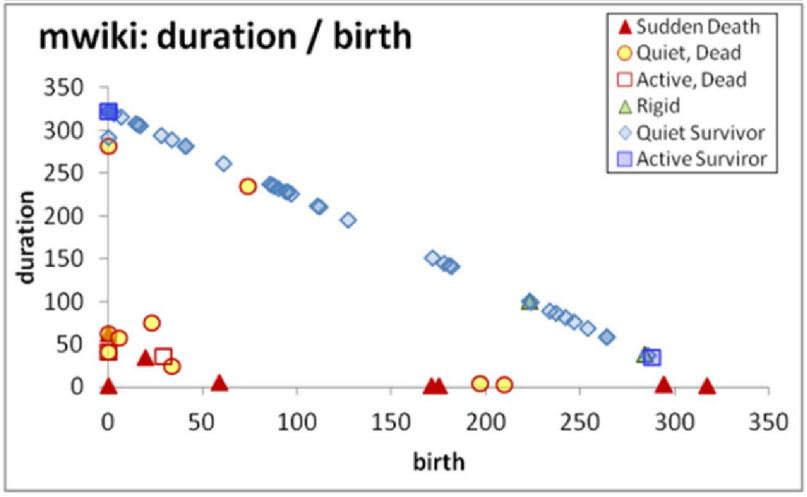
Schema growth over time (red continuous line) along with the heartbeat of changes (spikes) for two of our datasets. Overlayed darker green rectangles highlight the **calmness** versions, and lighter blue rectangles highlight **smooth expansions**. Arrows point at periods of **abrupt expansion** and circles highlight **drops in size**. [IS15]

What we know so far for table evolution [ER 15, IS 17]









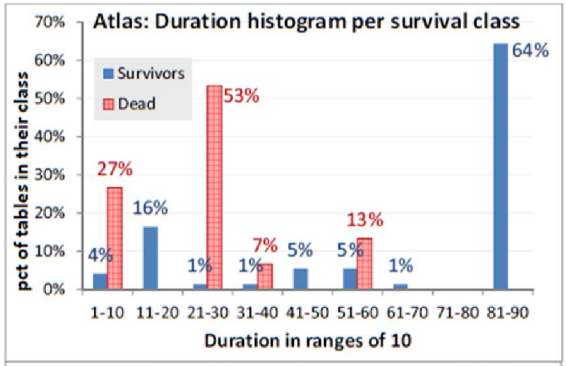
What we don't know yet…

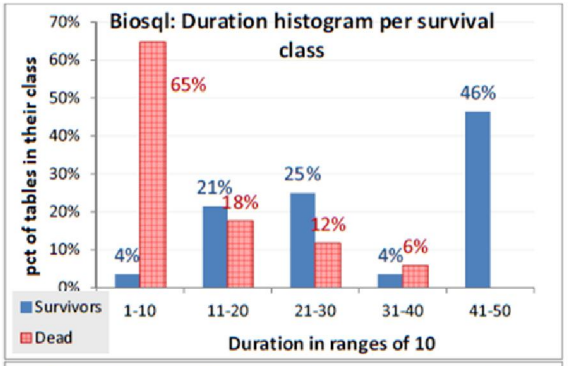
- Although we have fused the discrimination of survivor vs. dead tables in the graphical representations, the 4 patterns do not tell us …
- **… how do survivors differ from dead tables with respect to the combination of duration and activity profile**

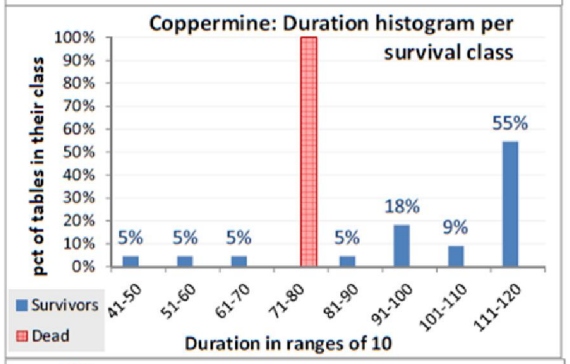
- Also studied [not part of the paper]: year of birth, schema size, schema resizing

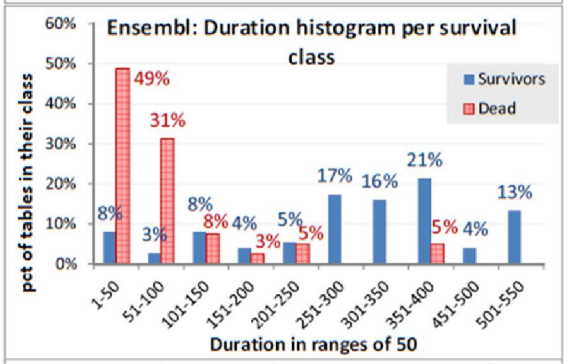
Compute the histograms of durations for both dead and survivors, and you get …

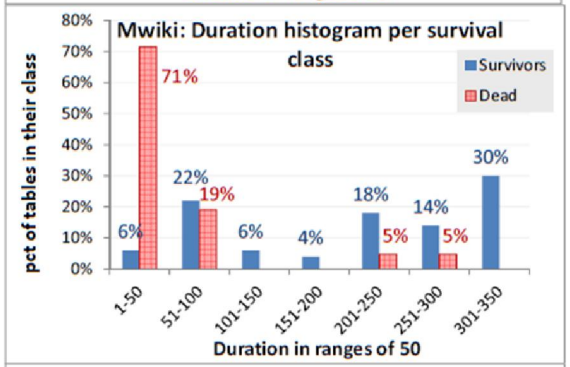
OPPOSITE SKEWED DURATIONS

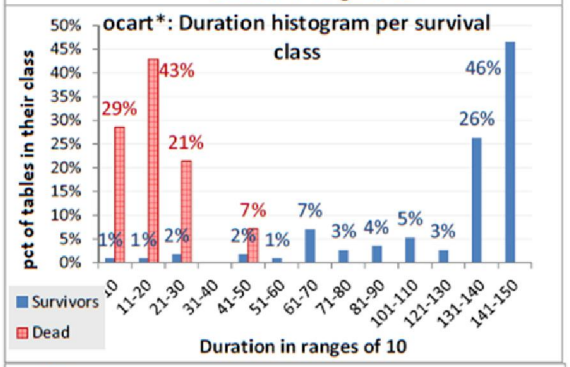


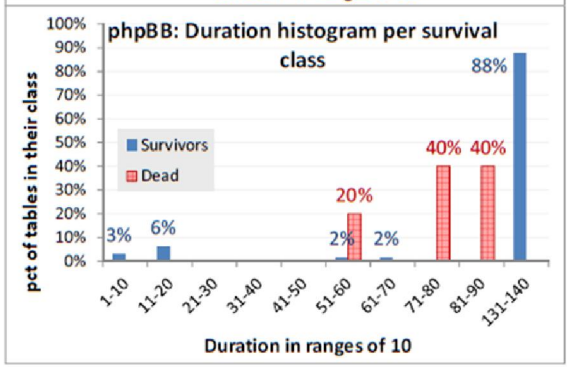


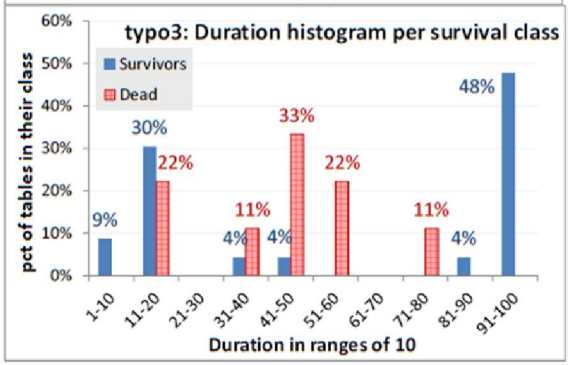












The oppositely skewed durations pattern

Histograms for the durations of dead vs. survivor tables

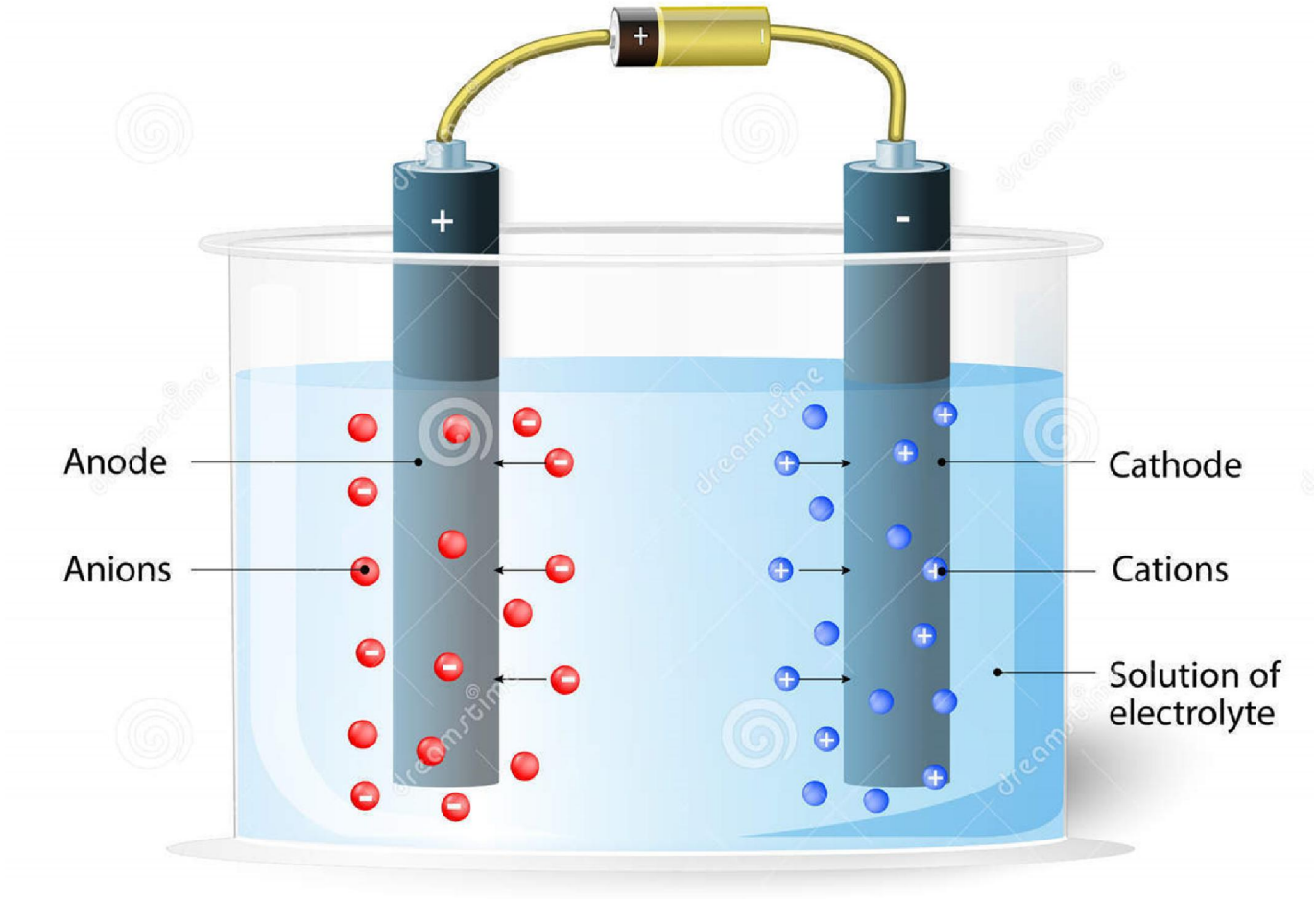
- **The dead tables are strongly biased towards short durations** (positively skewed),
- often with **very large percentages of them being removed very shortly after birth**.

- **Survivor tables are mostly heavy-tailed at the other end of the spectrum (negatively skewed), i.e., at high (frequently: max) durations.**

Not only are the durations of dead vs survivors "opposite", but also the activity profile is inverse, resulting in …

ELECTROLYSIS PATTERN FOR TABLE ACTIVITIES

Electrolysis in chemistry



https://www.dreamstime.com/stock-illustration-electrolysis-experimental-set-up-electrolysis-process-passing-electric-current-cations-move-towards-cathode-get-image68978953 (C) Designua | Dreamstime.com

Duration is related to the Life & Death Class of the tables!

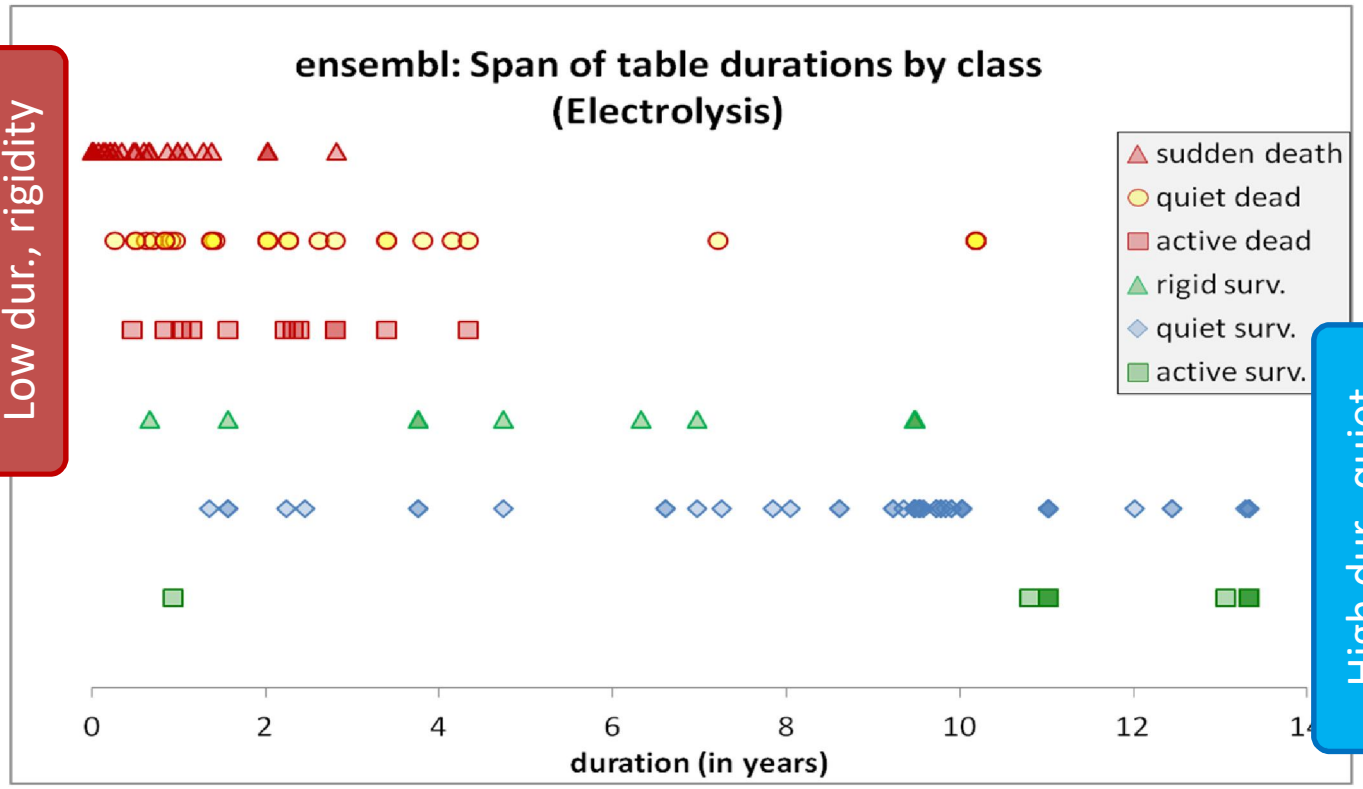
DEAD

Rigid

Quiet

Active

Low dur., rigidity



High dur., quiet

SURVIVORS

Rigid

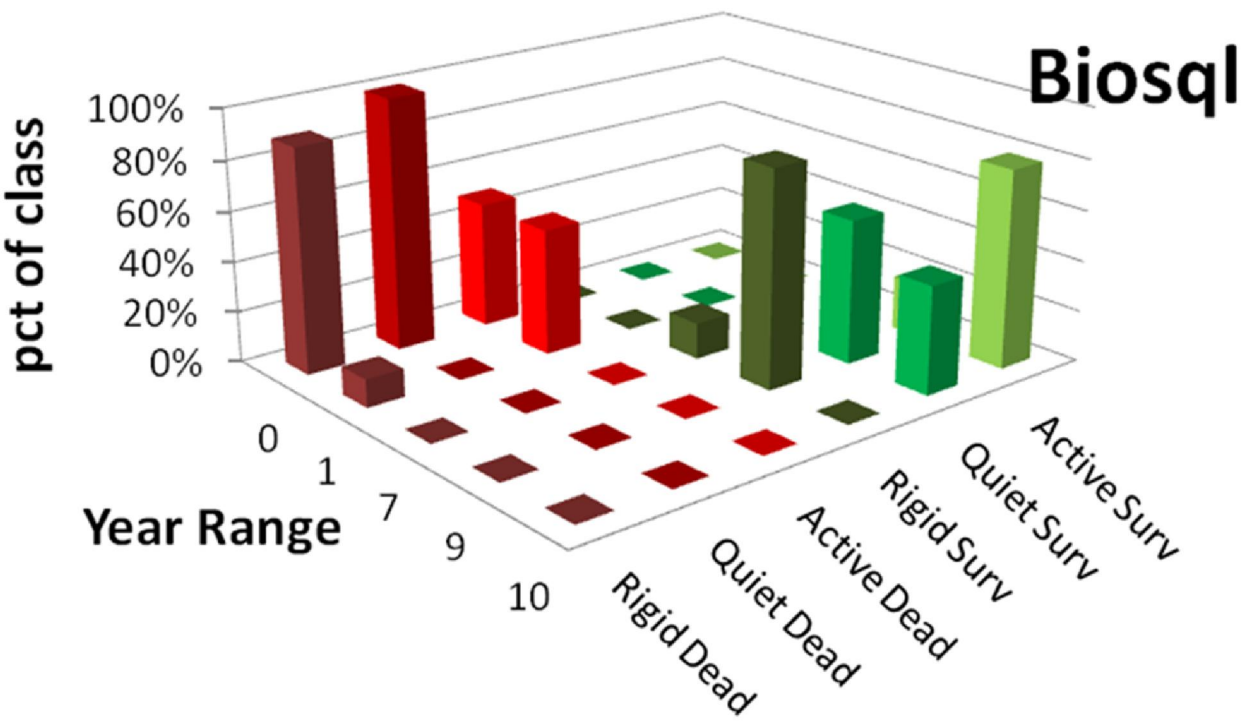
Quiet

Active

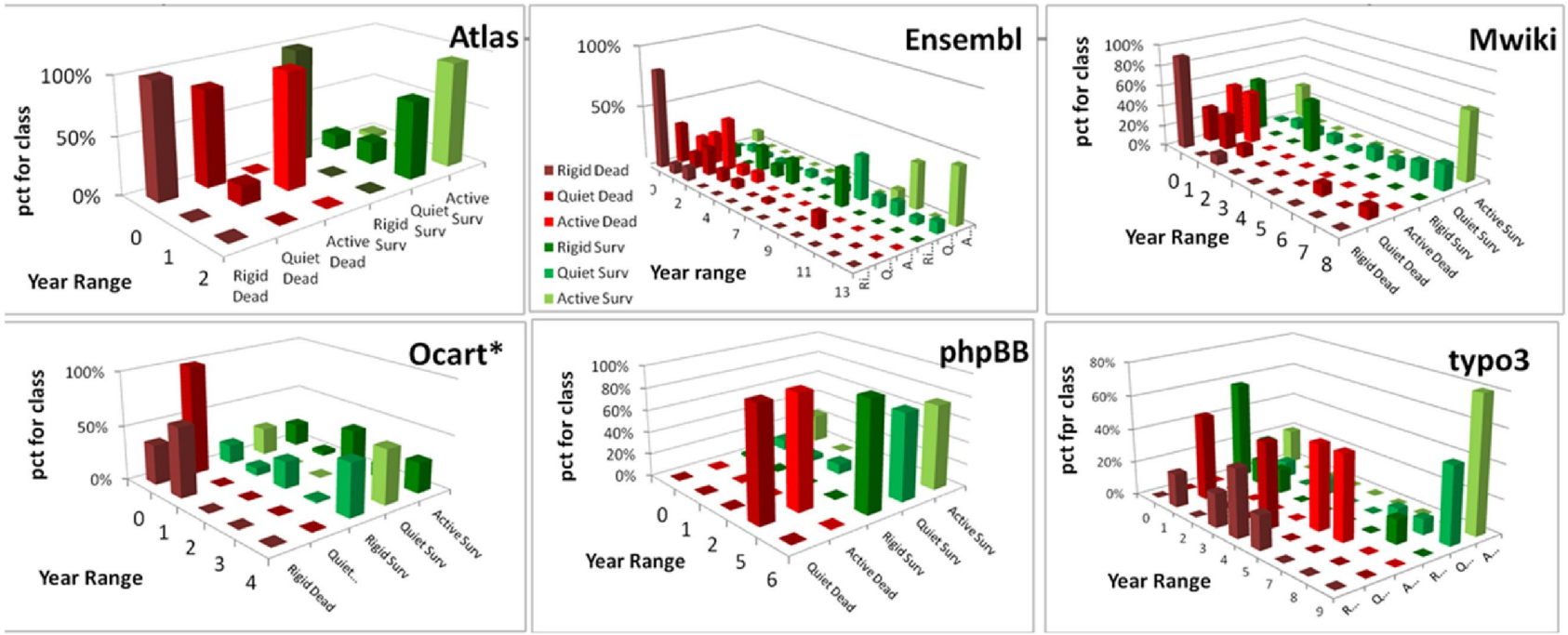
(a) **Survival**: DEAD vs SURVIVORS

(b) **Activity**: Rigid (no change) vs Active (change rate > 10%) vs Quiet (all in between)

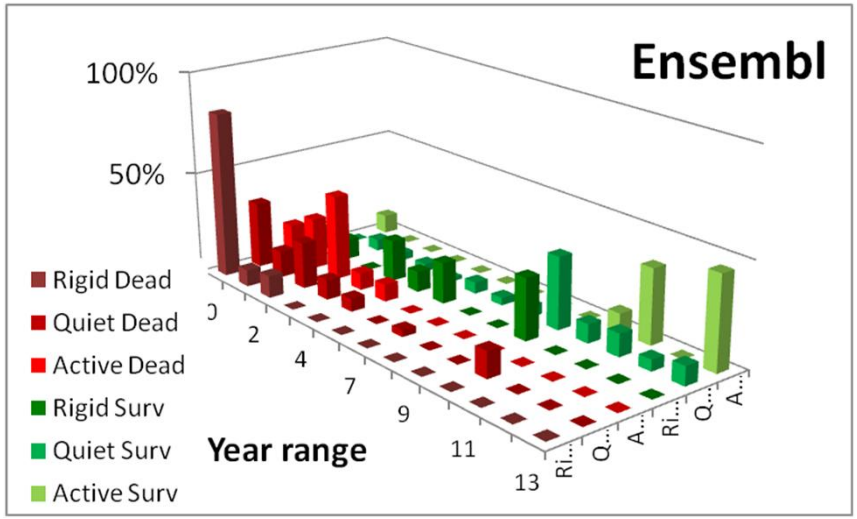
(c) **Life And Death (LAD)** class: Survival x Activity



Attn: all pct's are *per class*



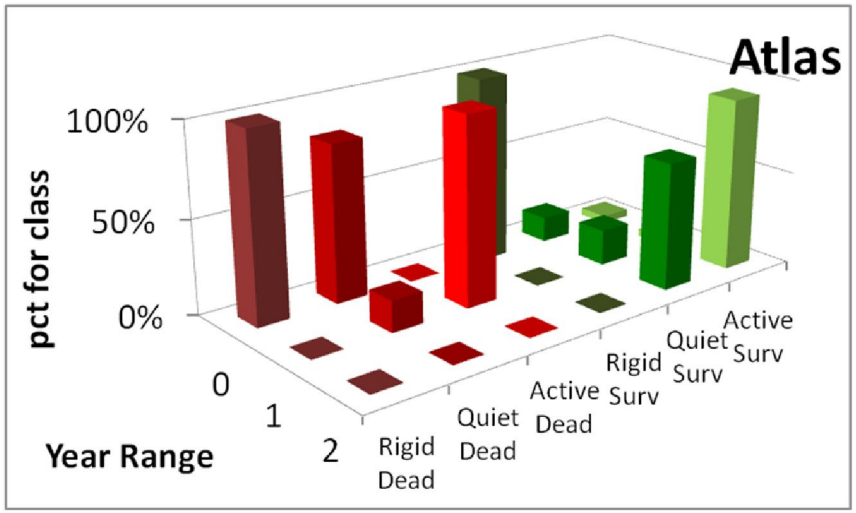
The electrolysis pattern

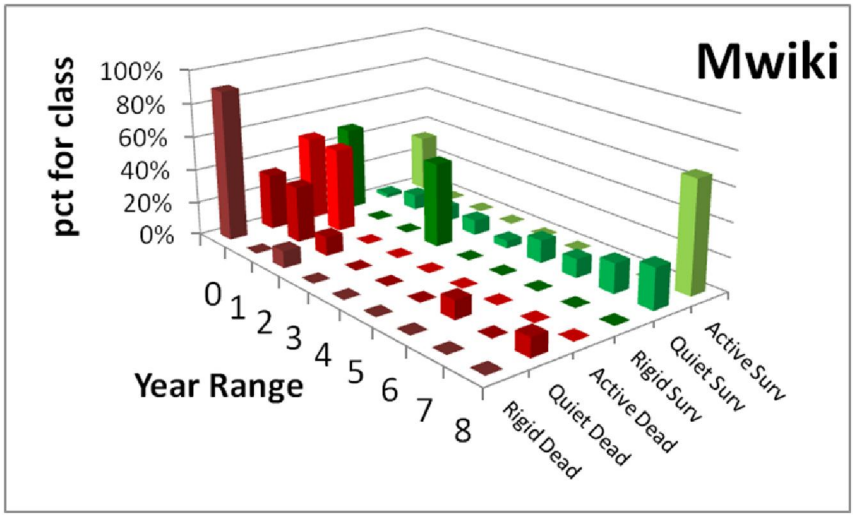


- **Dead tables demonstrate much shorter lifetimes** than survivor ones,
- can be located at short or medium durations, and **practically never at high durations**.
- With few exceptions, the **less active dead tables are, the higher the chance to reach shorter durations**.

- **Survivors** expose the inverse behavior, i.e., **mostly located at medium or high durations**.
- The **more active survivors are, the stronger they are attracted towards high durations**, with a significant such inclination for the few active ones that cluster in very high durations.

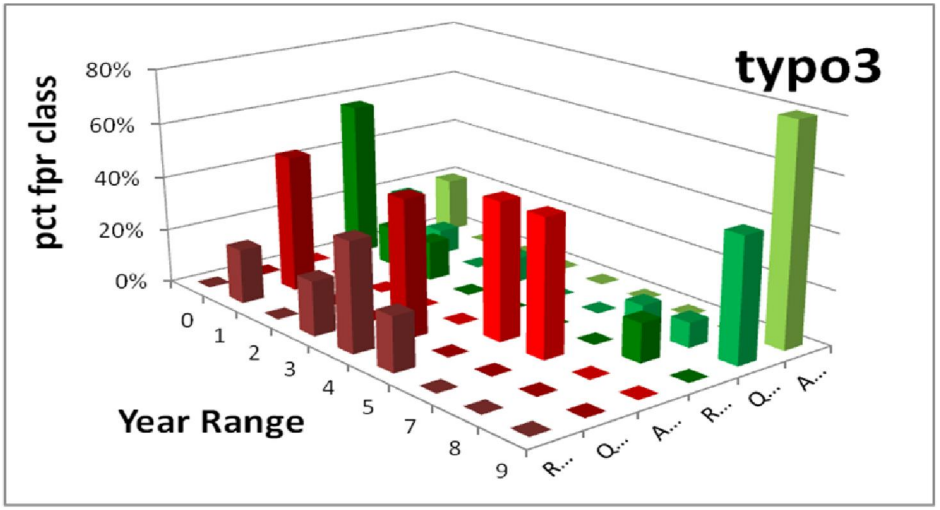
The electrolysis pattern: survivors

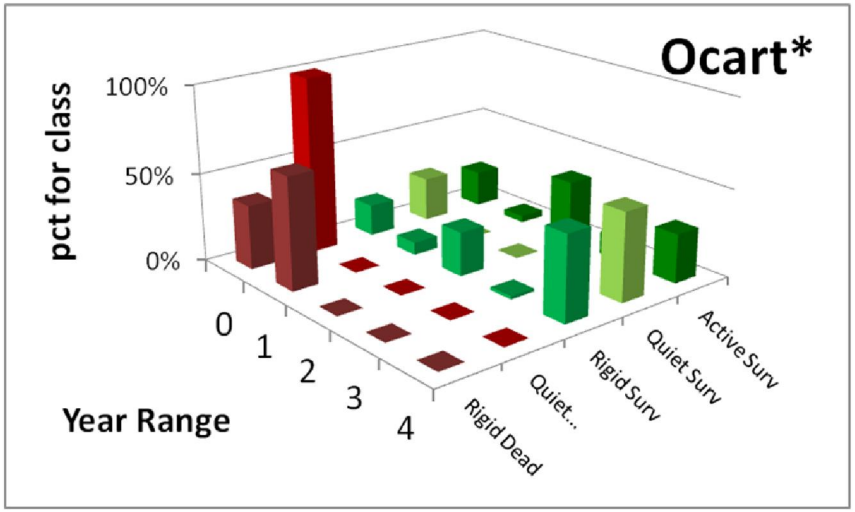




- The extreme clustering of active survivors to high durations
- The wider spread of (quite numerous) quiet survivors to a large span of durations with long trails of points
- The clustering of rigid survivors, albeit not just to one, but to all kinds of durations (frequently, not as high as quiet and active survivors)

The electrolysis pattern: dead





- The total absence of dead tables from high durations
- The clustering of rigid dead at low durations,
- the spread of quiet dead tables to low or medium durations, and
- the occasional presence of the few active dead, that are found also at low or medium durations, but in a clustered way

For each data set, for each LifeAndDeath class, **percentage** of tables per duration range **over the total of the data set** (for each data set, the sum of all cells adds up to 100%)

Atlas	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0% -20%)	5%	0%	0%	1%	1%	0%	7%
[20%-80%)	3%	7%	2%	11%	13%	3%	40%
[80%-100%]	0%	0%	0%	0%	28%	25%	53%
	8%	7%	2%	13%	42%	28%	100%

Biosql	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	20%	13%	4%	0%	0%	0%	38%
[20%-80)	0%	0%	0%	2%	0%	0%	2%
[80%-100%]	0%	0%	0%	13%	16%	31%	60%
	20%	13%	4%	16%	16%	31%	100%

Copperm.	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	0%	0%	0%	0%	0%	0%	0%
[20%-80)	4%	0%	0%	0%	13%	0%	17%
[80%-100%]	0%	0%	0%	30%	43%	9%	83%
	4%	0%	0%	30%	57%	9%	100%

Ensembl	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	23%	13%	5%	1%	3%	1%	37%
[20%-80)	1%	7%	3%	5%	23%	0%	54%
[80%-100%]	0%	0%	0%	0%	9%	6%	8%
	24%	20%	8%	6%	35%	7%	100%

Mwiki	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	13%	7%	3%	1%	6%	1%	23%
[20%-80)	1%	4%	0%	1%	31%	0%	58%
[80%-100%]	0%	1%	0%	0%	27%	3%	20%
	14%	13%	3%	3%	63%	4%	100%

Ocart*	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	3%	2%	0%	8%	5%	1%	23%
[20%-80)	5%	0%	0%	17%	16%	0%	43%
[80%-100%]	0%	0%	0%	17%	22%	2%	34%
	9%	2%	0%	42%	44%	3%	100%

phpBB	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	0%	0%	0%	1%	3%	3%	7%
[20%-80)	0%	1%	0%	0%	4%	0%	11%
[80%-100%]	0%	1%	4%	49%	24%	9%	81%
	0%	3%	4%	50%	31%	11%	100%

typo3	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	3%	3%	0%	16%	9%	3%	34%
[20%-80)	13%	3%	3%	3%	6%	0%	31%
[80%-100%]	0%	0%	3%	3%	19%	13%	34%
	16%	6%	6%	22%	34%	16%	100%

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

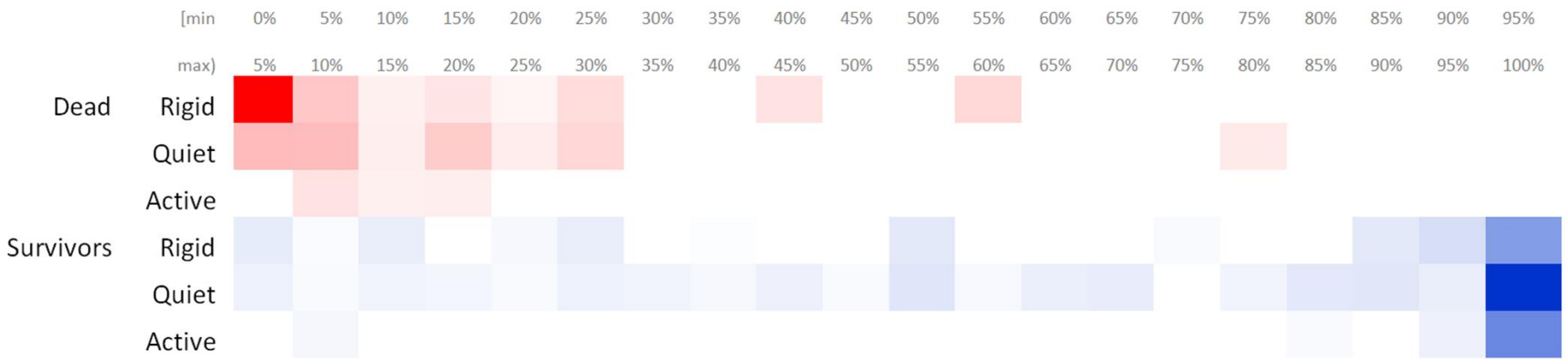
http://www.cs.uoi.gr/~pvassil/projects/schemaBiographies

Indicative, average values over all datasets: for each LifeAndDeath class, **percentage** of tables per duration range **over the total of the data set**

	Rigid Dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	8%	5%	2%	4%	3%	1%	*23%*
[20%-80%)	3%	3%	1%	5%	13%	0%	*26%*
[80%-100%]	0%	0%	1%	14%	24%	12%	*51%*
	12%	*8%*	*3%*	*23%*	*40%*	*14%*	*100%*

An acute reader might express the concern whether it would be better to gather all the tables in one single set and average over them. We disagree: each data set comes with its own requirements, development style, and idiosyncrasy and putting all tables in a single data set, not only scandalously favors large data sets, but integrates different things. We average the behavior of schemata, not tables here.

… electrolysis as a heatmap …



- For each *LifeAndDeath* value, and for each duration range of 5% of the database lifetime, we computed the percentage of tables (over the total of the data set) whose duration falls within this range.
- We removed cells that corresponded to only one data set

The resulting heatmap shows the polarization in colors: brighter color signifies higher percentage of the population

- Background
- Durations' study
- Electrolysis
- Discussion

Main Findings

Open Issues

DISCUSSION & OPEN ISSUES

Gravitation to Rigidity

- Although the majority of survivor tables are in the quiet class, we can quite emphatically say that **it is the absence of evolution that dominates**!
  - Survivors vastly outnumber removed tables.
  - Similarly, rigid tables outnumber the active ones, both in the survival and, in particular, in the dead class.
  - Schema size is rarely resized, and only in survivors (not in the paper).
  - Active tables are few and do not seem to be born in other but early phases of the database lifetime.
- Evidently, not only survival is also stronger than removal, but **rigidity is also stronger a force than variability** and the combination of the two forces further lowers the amount of change in the life of a database schema.

Gravitation to rigidity: death

- **Why dead tables have short durations and die mostly rigid?**
  - We believe its due to the cost that deletions have for the maintenance of the software that surrounds the database.
  - The earlier a table is removed, the smaller the cost of maintaining the surrounding code is. If the table starts being used by queries spread in the code, the cost to to locate, maintain and test the application code that uses it is high.

Gravitation to rigidity: life

- Who survives? Why do survivors last long?
  - Due to the reluctance for removals, it appears that **after a certain period, practically within 10%-20% of the databases' lifetime, tables begin to be "safe"…**
  - … add to this that the starting versions of the database already include a large percentage of the overall population of tables …
  - … and you get a right-heavy, left-tailed, negatively skewed distribution of survivor tables (for 6 out of 8 data sets, **survivor durations reaching the final bucket of the respective histogram exceed 45%**).

Gravitation to rigidity: life

- **Tables with high durations that survive spend their lives mostly quietly** (with the few occasional maintenance changes) – again minimizing the impact to the surrounding code.
- **The high concentration of the few active tables to very high durations and survival is related to the gravitation to rigidity:**
  - … the early phases of the database lifetime typically include more table births
  - … after the development of a substantial amount of code, too high rate of updates becomes harder; this results in very low numbers of active tables being born later.
  - So, **the pattern should not be read so much as "active tables are born early", but rather as "we do not see so many active tables being born in late phases of the database life".**

Activity & Duration

- **Rigid tables find it hard to attain high durations** (unless found in an environment of low change activity).
  - Shortly after there are born, they are in the high-risk group of being removed.
  - Rigid tables have the highest migration probability (a single upd => quiet).
- Long duration and high activity are also correlated
  - Long duration is practically a pre-requisite of high activity (very rare exceptions)
  - Lack of late born active tables explains the long duration of the few active ones

Risks for developers

- Young rigid tables are the high risk group for being removed
  - Tables mostly survive; when they don't, tables typically die shortly after their birth and quite often, rigid
- If a table surpasses infant mortality it will likely survive to live a rigid or, more commonly, a quiet live.
- There is a small group of active tables, going through significant updates. Look for them in the early born, survivors.

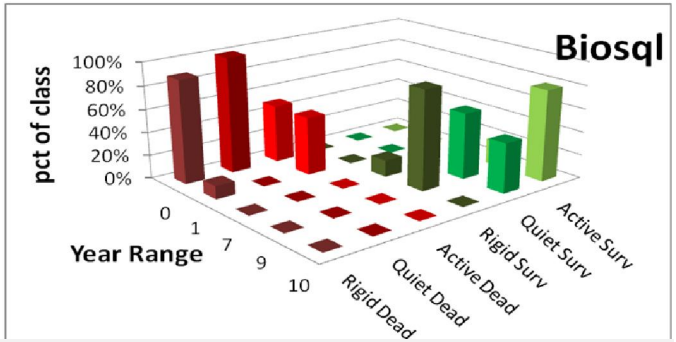
- Soon after a table is born, the development of code that depends on it should be kept as restrained as possible
- After the period of infant mortality, it is fairly safe to say that (unless the table shows signs of significant update activity), gravitation to rigidity enters the stage and the table's evolution will be low.

Future work

- Related literature suggests that database **evolution cools down after the first versions**. Is it true?
- Collect posted comments and expressed user requirements at the public repositories and try to figure out **why** change is happening the way it does.
  - Automating this effort is a very ambitious goal in this context.
- Finally, the **validation of existing research** results with **more studies from other groups**, different software tools, hopefully extending the set of studied data sets, is imperative to allow us progressively to move towards `laws' rather than `patterns‘ of change in the field of understanding schema evolution.

Danke schön!
Thank you!

- Yes, we can indeed find **patterns in the lives of tables,** during schema evolution!
- **Survivors, mostly long-lived (esp. active ones) and quietly active are radically different than dead tables, being mostly short-lived and rigid!**
- **Gravitation to rigidity rules:** we see more absence than presence of schema evolution!



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

http://www.cs.uoi.gr/~pvassil/projects/schemaBiographies

AUXILIARY SLIDES

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

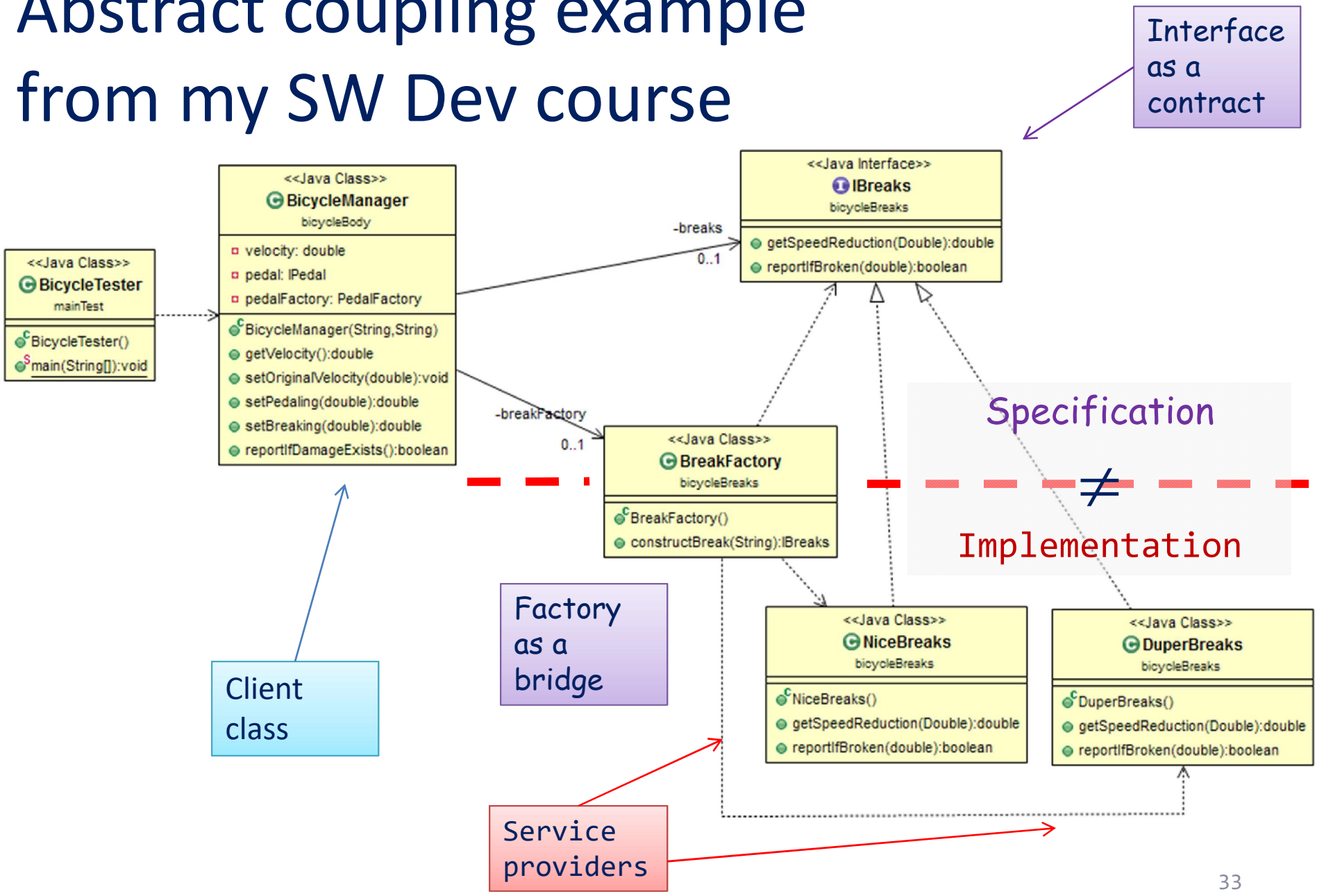
- How do databases change?
- In particular, how does the schema of a database evolve over time?

- 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?

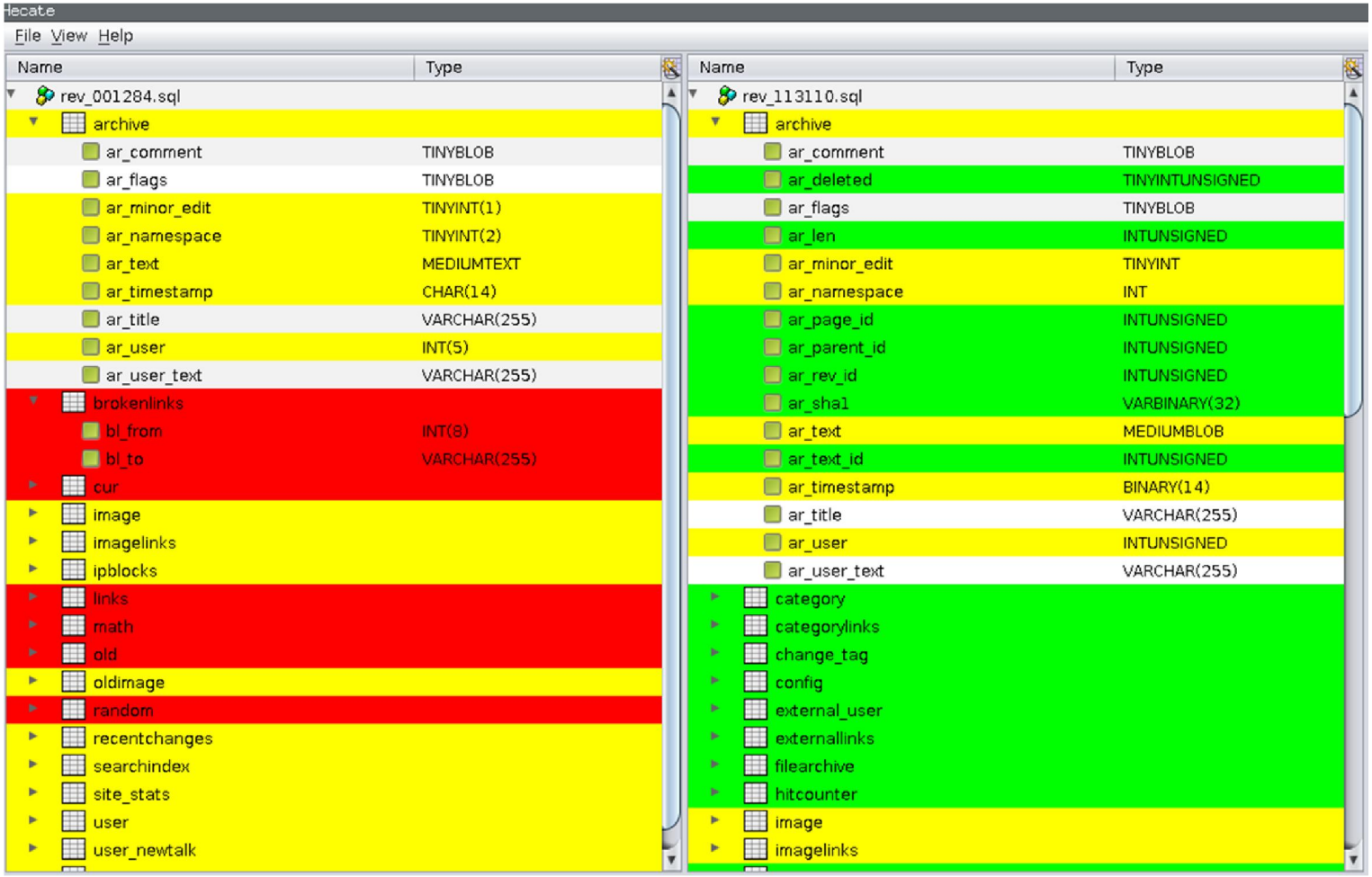
Why care for the “laws”/patterns of schema evolution?

- 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



Hecate: SQL schema diff extractor



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

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

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

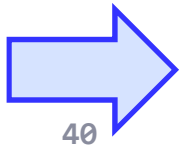
Measures and Terminology

- *SurvivalClass*: this measure classifies a tables as (a) a *survivor* table (with the value of 20 in our data) if the table has survived (i.e., was present at the last known version of the database schema) or (b)a *dead* table (with the value of 10 in our data), if its last known version is prior to the last known version of the schema history.
- *ATU*: Average Transitional amount of Updates is the ratio *SumUpd / Duration*
- *ActivityClass*: characterization of how "active" a table is. Takes the value 0 for *rigid* tables that go through zero updates in their life, 2 for *active* tables, having ATU larger than 0.1 and sumUpd larger than 5 (see [ER 2015]), and 1 for the rest of the tables, characterized as *quiet* tables.
- *LifeAndDeath Class*: the Cartesian product of the measures *SurvivorClass* and *ActivityClass*. The *LifeAndDeath Class* characterizes a table both with respect to its survival and to its update profile during its lifetime. The measure's domain includes six values produced by the combination of {dead, survivor} x {rigid, quiet, active} (and conveniently computed as the sum *SurvivorClass + ActivityClass* in our data).

Can we generalize out findings broadly?

External validity

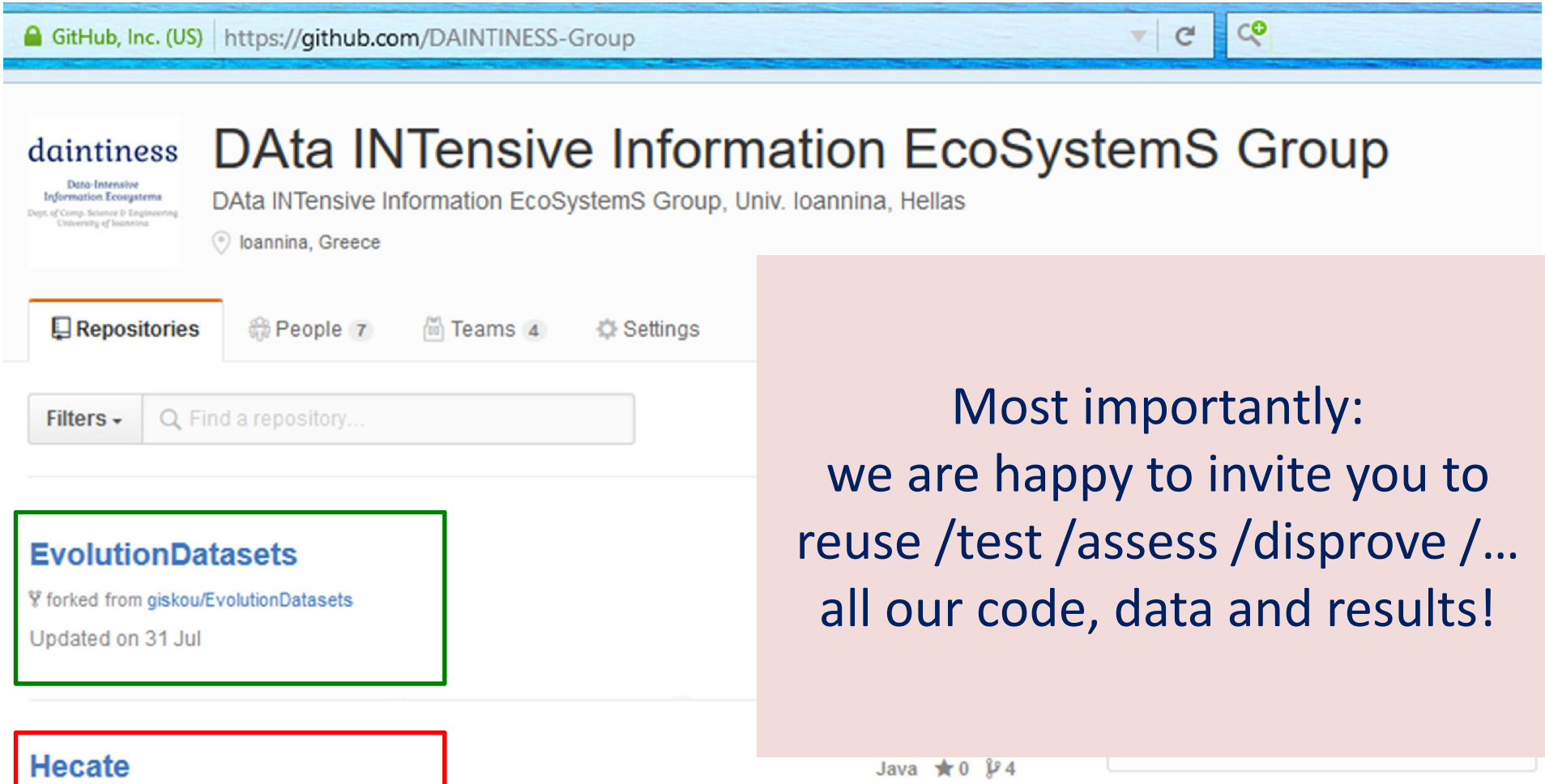
- We perform an **exploratory study to observe frequently occurring phenomena** within the scope of the aforementioned population
- **Are our data sets representative enough?** Is it possible that the observed behaviors are caused by sui-generis characteristics of the studied data sets?
  - Yes: we believe 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



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/projects/schemaBiographies

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

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

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.

Timeline of empirical studies

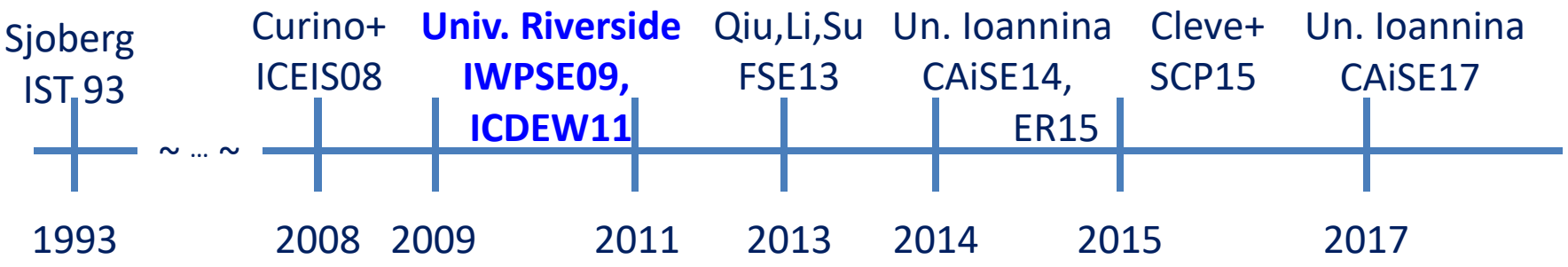
Curino+ @ ICEIS08: Mediawiki for 4.5 years
100% increase in the number of tables
142% in the number of attributes.

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

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

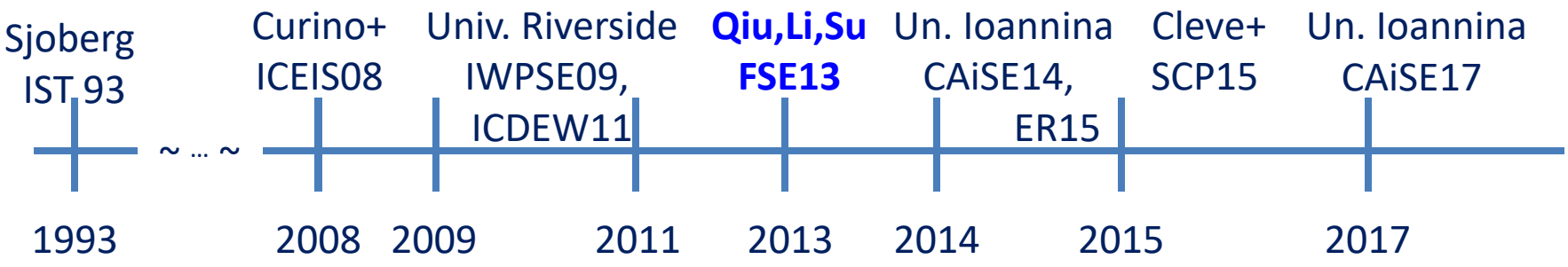


Timeline of empirical studies

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.

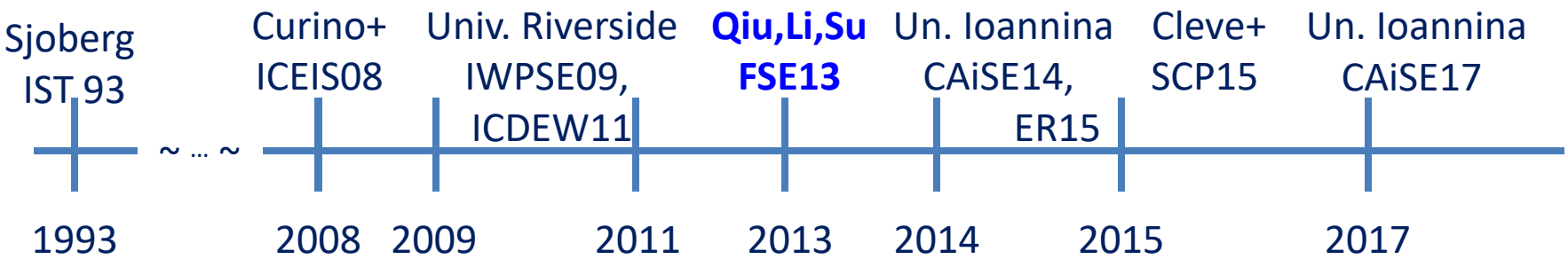


Timeline of empirical studies

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

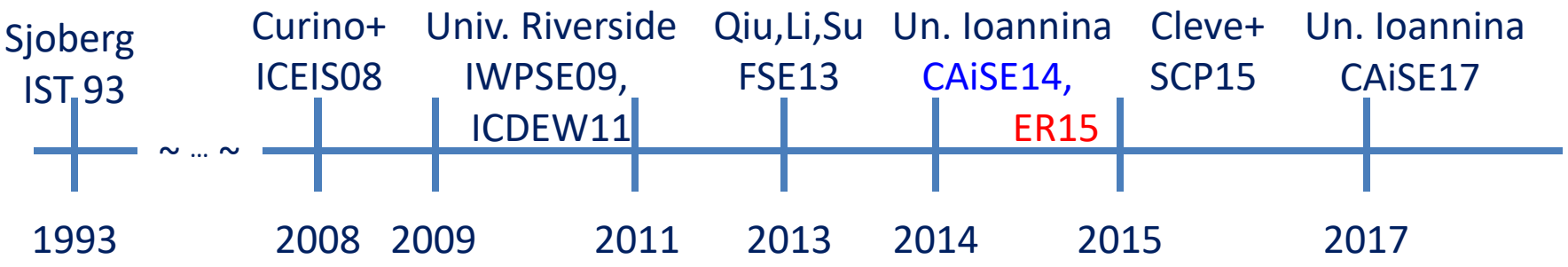
- Code and db changed in the same revision: 50.67% occasions
- Code change was in a previous/subsequent version than the one where the database schema change: 16.22% of occasions
- database changes not followed by code adaptation: 21.62% of occasions
- 11.49% of code changes were unrelated to the database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



Timeline of empirical studies

CAiSE14: DB level
ER'15: Table level



Timeline of empirical studies

Cleve+ Science Comp. Progr. 2015: Oscar, an open source electronic medical record system

- schema grows over time
- deletions are rare
- change is infrequent: most tables have less than 4 changes

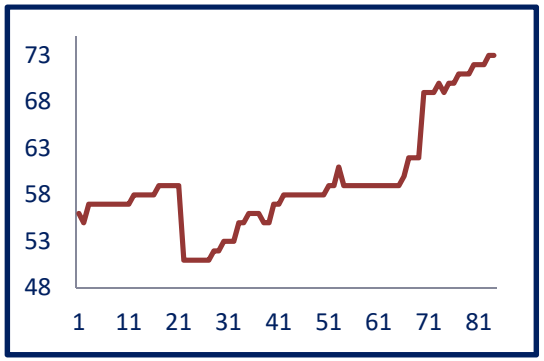
CAISE 14 / INF. SYSTEMS 15

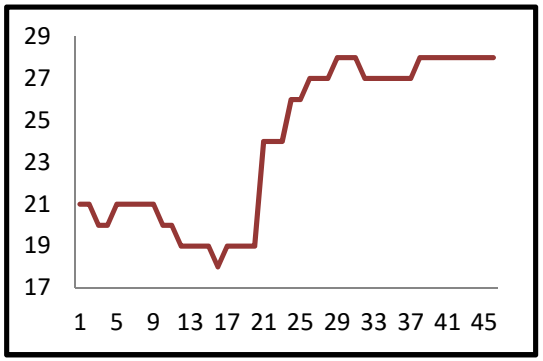
Datasets

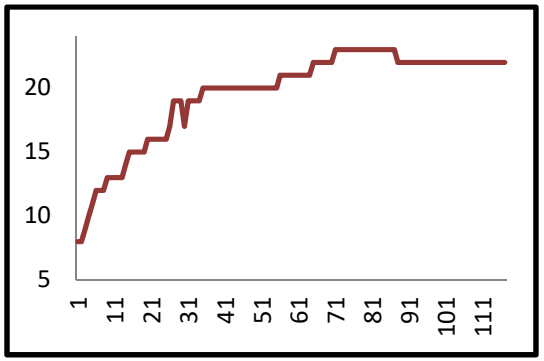
https://github.com/DAINTINESS-Group/EvolutionDatasets

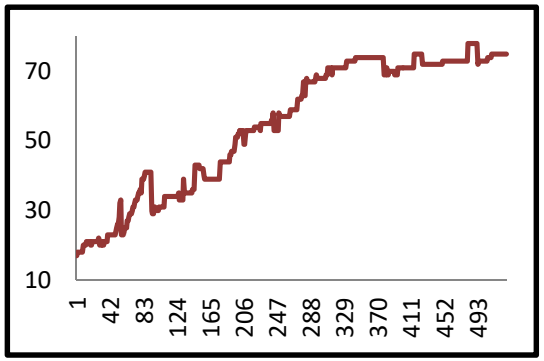
- Content management Systems
  - MediaWiki, TYPO3, Coppermine, phpBB, OpenCart
- Medical Databases
  - Ensemble, BioSQL
- Scientific
  - ATLAS Trigger

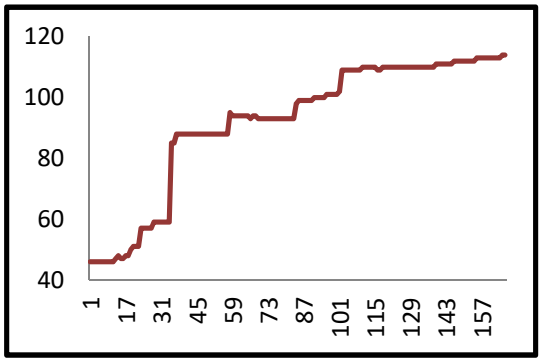
Schema Size (relations)

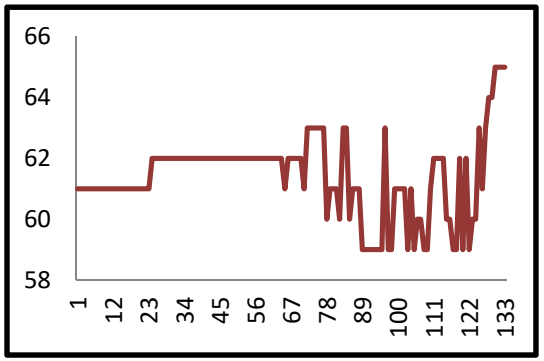


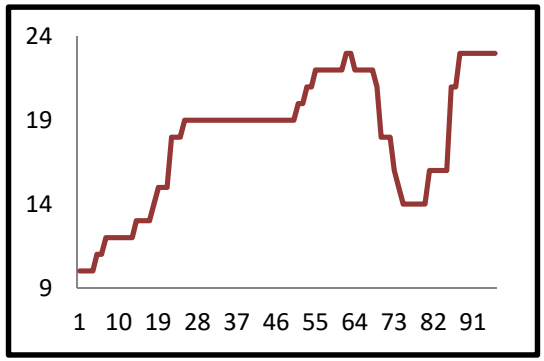


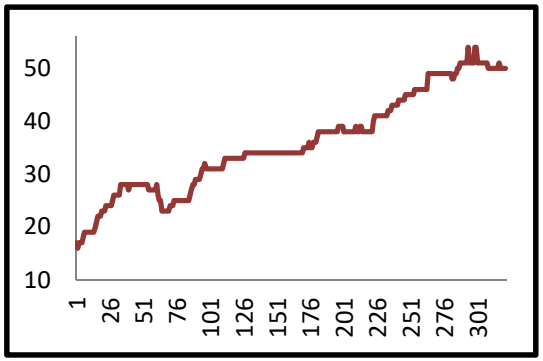












CaiSE'14: Main results

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

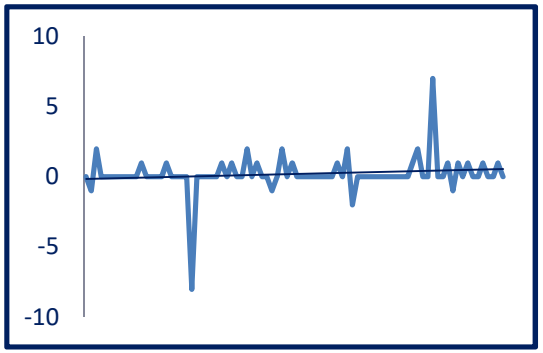
Patterns of change: no consistently constant behavior

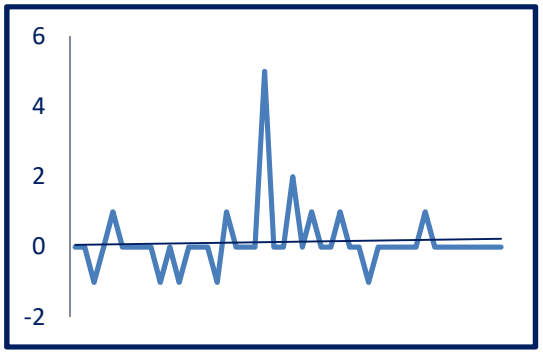
- Changes reduce in density as databases age
- Change follows three patterns: **Stillness**, **Abrupt change** (up or down), **Smooth growth upwards**
- Change frequently follows **spike** patterns
- **Complexity** does **not** increase with age

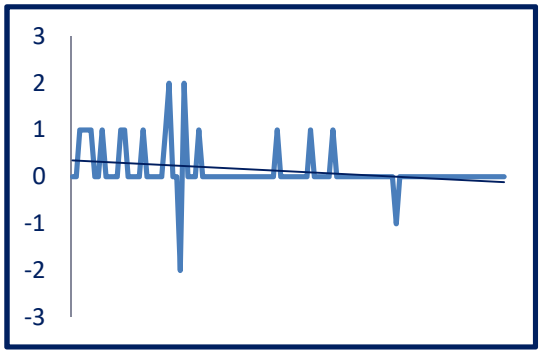
What we have found for schema evolution [CAiSE 14, IS 15]

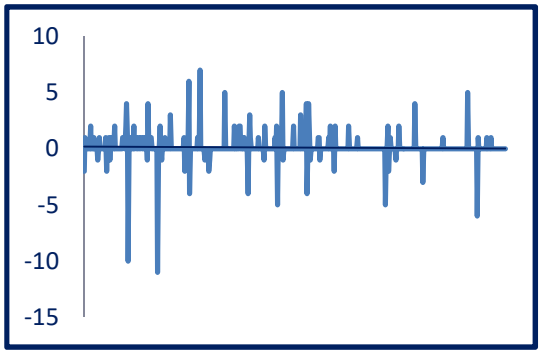
- Schemata grow over time in order to satisfy new requirements, albeit not in a continuous or linear fashion, but rather, with bursts of concentrated effort interrupting longer periods of calmness.
- Growth is small, with average growth being close to zero.
- Growth comes with drops in schema size that signify the existence of perfective maintenance.

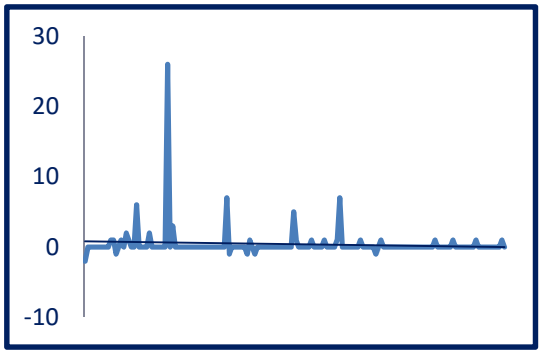
Schema Growth (diff in #tables)

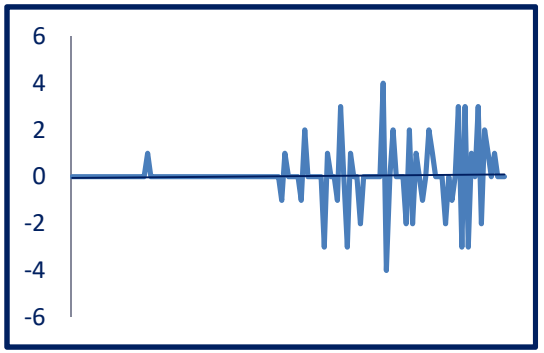


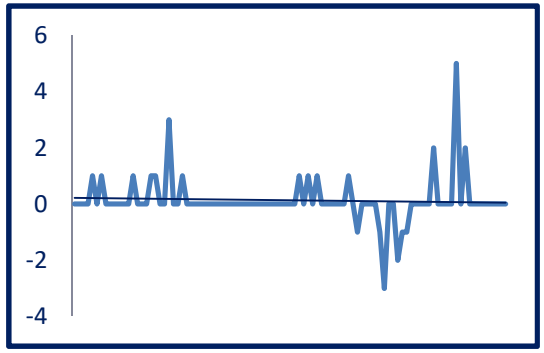


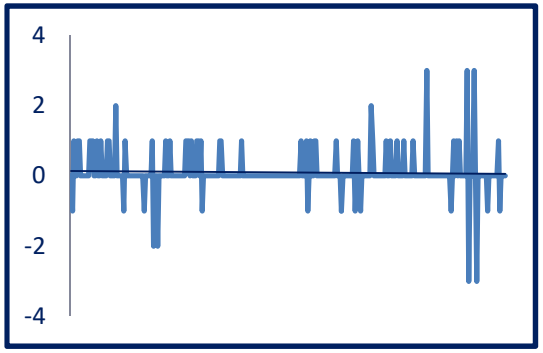




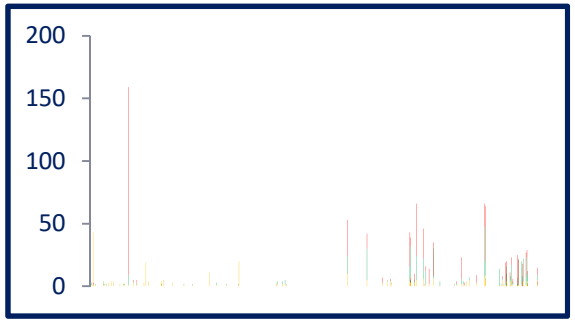


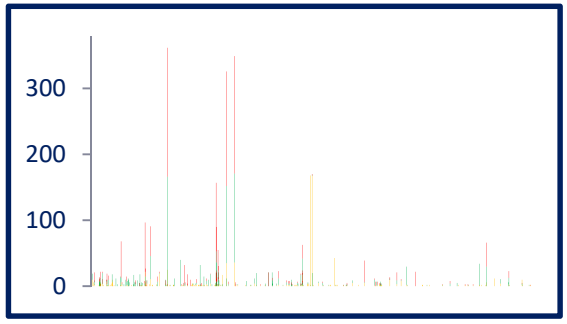


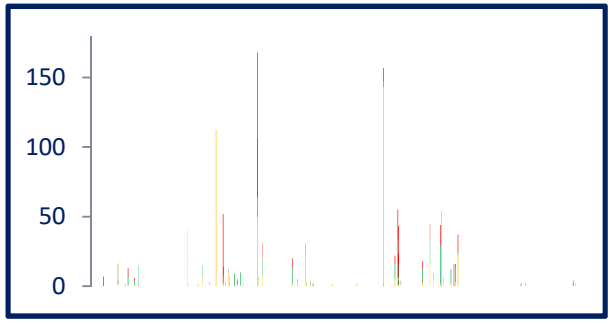


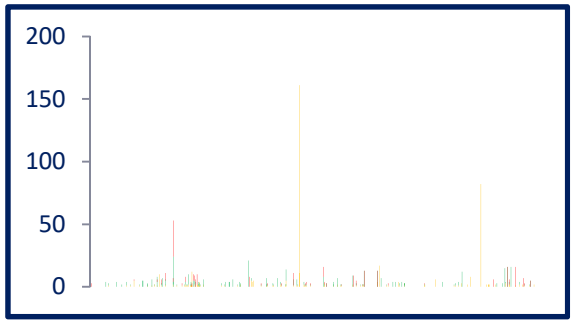


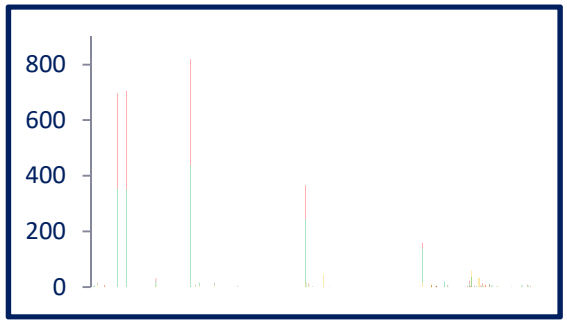
Change over time

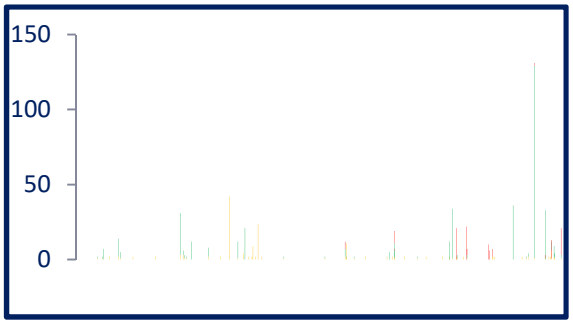


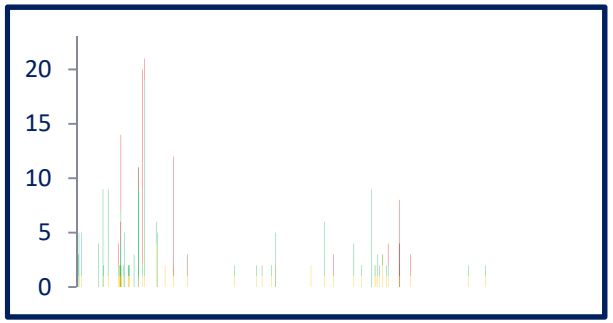




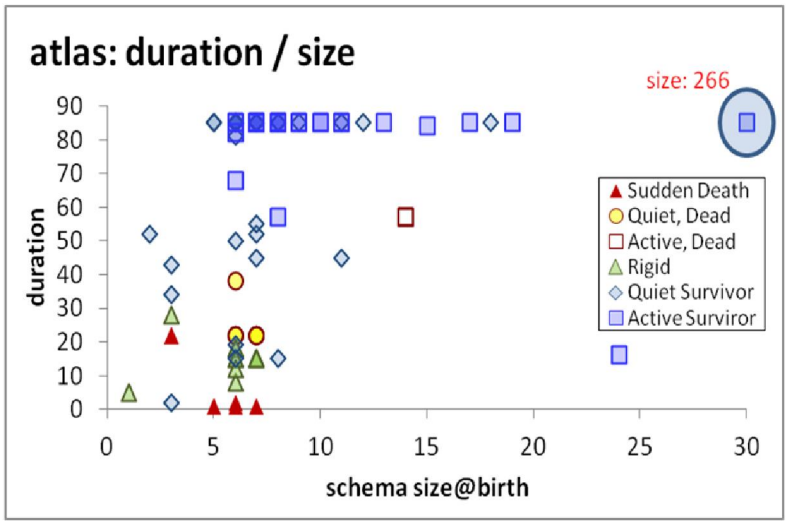


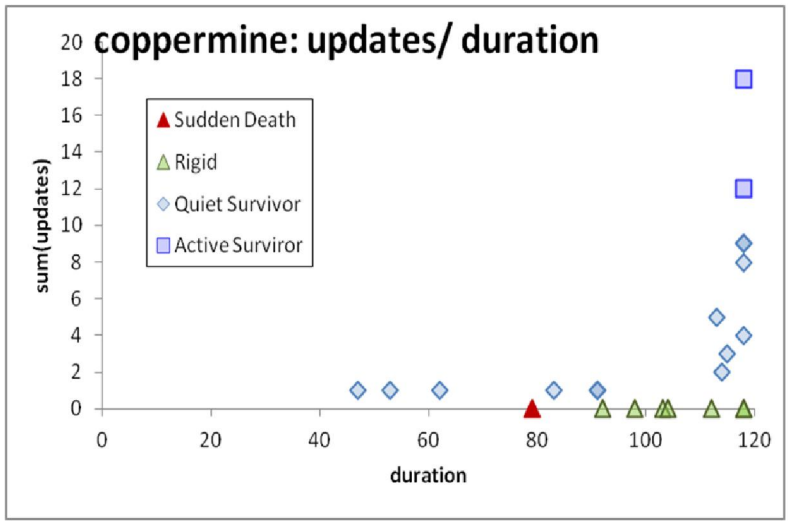


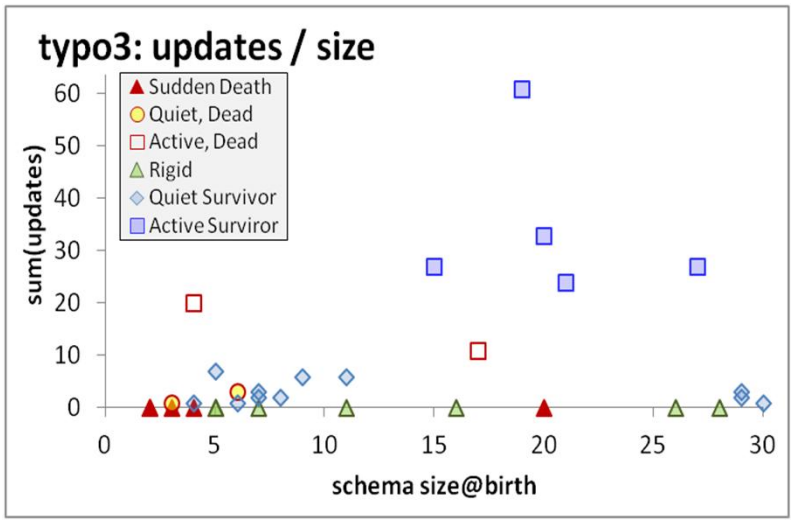


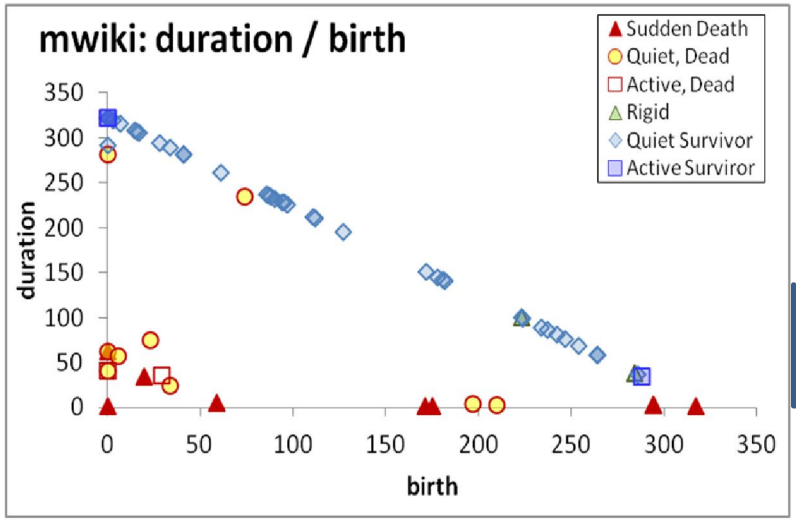


ER 2015 / IS 2017









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

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

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 …

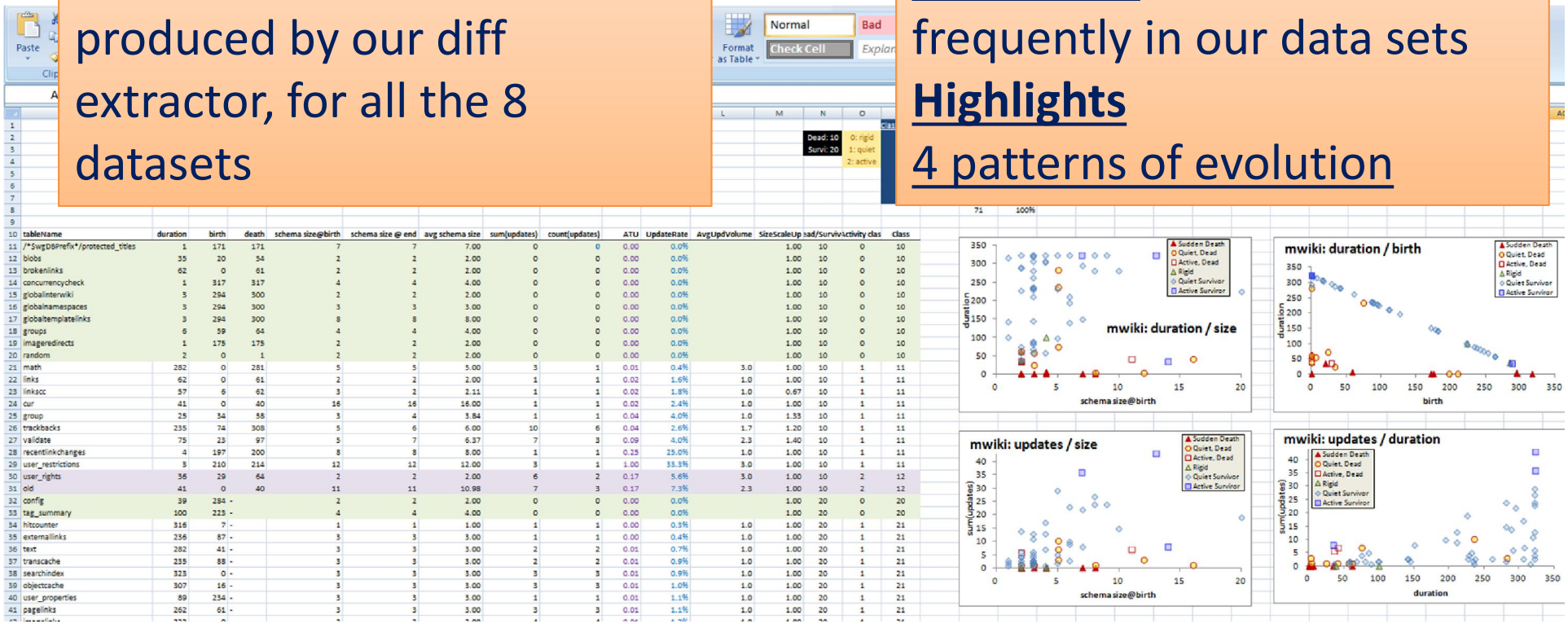
	<5	5-10	>10
atlas	10,23%	68,18%	21,59%
biosql	75,56%	24,44%	0,00%
coppermine	52,17%	30,43%	17,39%
ensembl	54,84%	38,06%	7,10%
mediawiki	61,97%	19,72%	18,31%
phpbb	40,00%	44,29%	15,71%
typo3	21,88%	31,25%	46,88%
opencart	57,20%	33,05%	9,75%
Average	46,73%	36,18%	17,09%

THE FOUR PATTERNS

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



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

What we know so far for table evolution [ER 15, IS 17]

- The *Γ pattern* indicates that tables with large schemata tend to have long durations and avoid removal;
- The *Comet pattern* indicates that the tables with most updates are frequently the ones with medium schema size;
- The *Inverse Γ pattern* indicates that tables with medium or small durations produce amounts of updates lower than expected, whereas tables with long duration expose all sorts of update behavior.
- The *Empty Triangle* pattern indicates a significant absence of tables of medium or long durations that were removed – thus, an empty triangle – signifying mainly short lives for deleted tables and low probability of deletion for old timers.

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

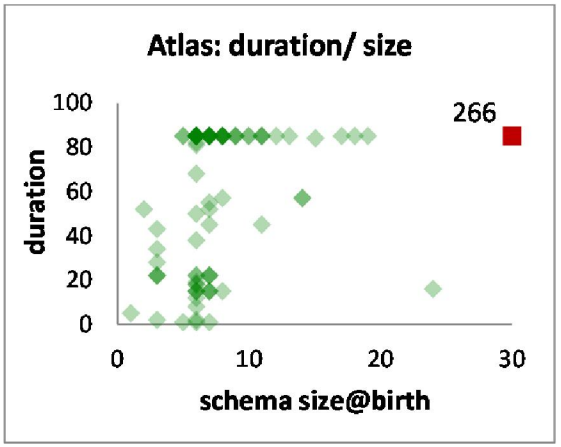
- How are these measures interrelated?

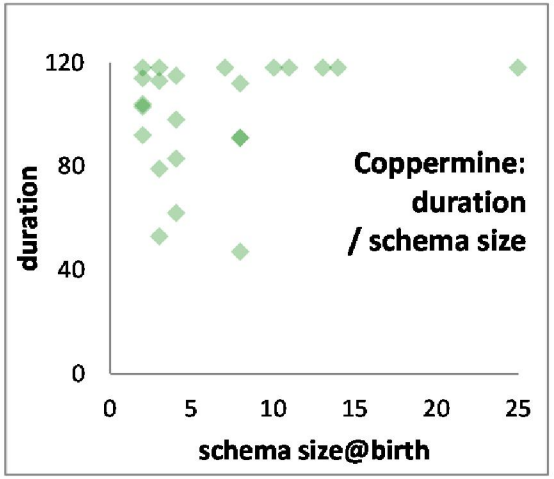
SCHEMA SIZE, CHANGE AND DURATION

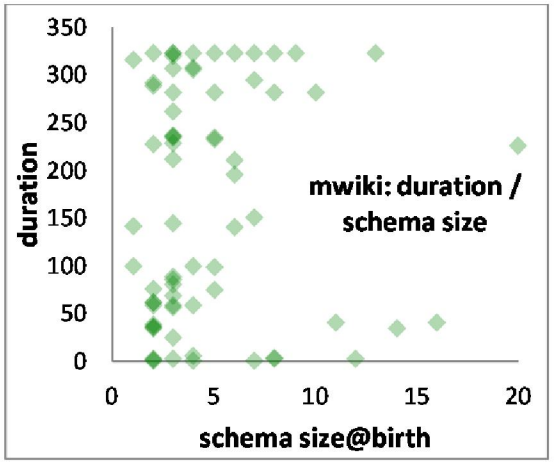
The Gamma Γ Pattern: "if you 're wide, you survive"

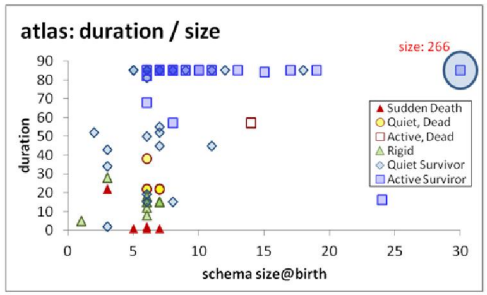
- The Gamma phenomenon:
  - tables with small schema sizes can have arbitrary durations, //small size does not determine duration
  - larger size tables last long

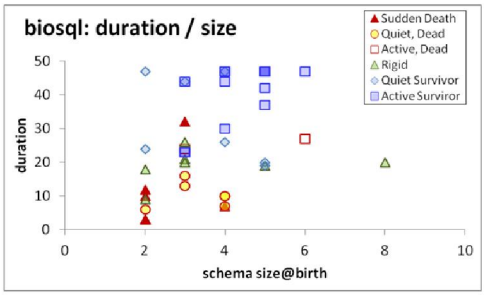
- Observations:
  - whenever a table exceeds the critical value of 10 attributes in its schema, its chances of surviving are high.
  - in most cases, the large tables are created early on and are not deleted afterwards.

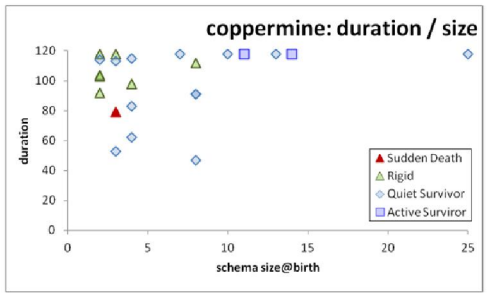


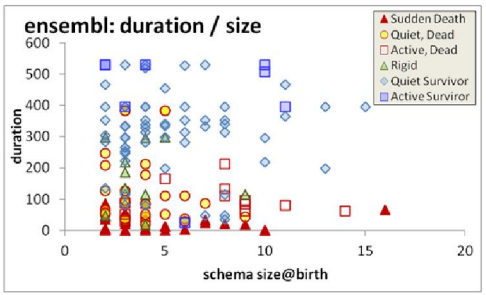


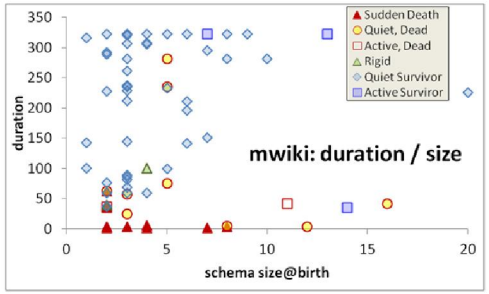


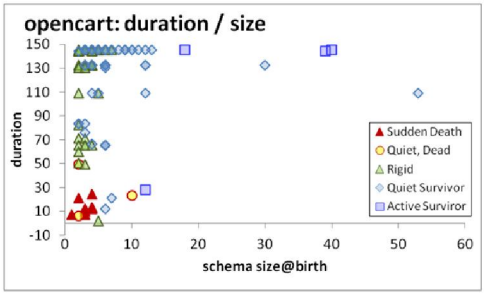


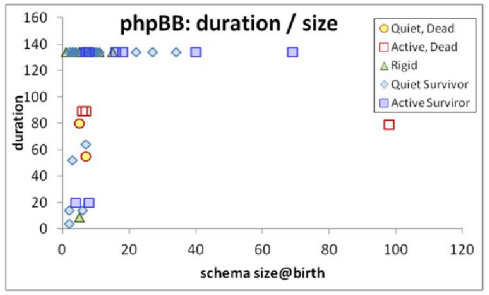


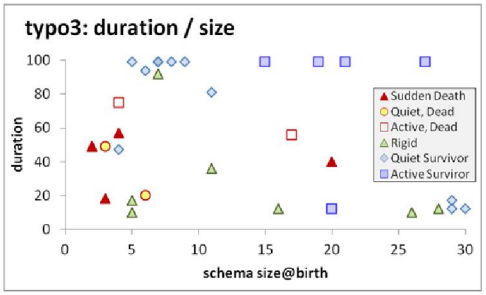












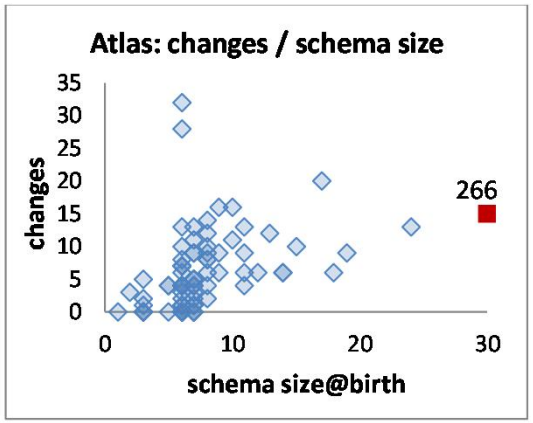
Exceptions

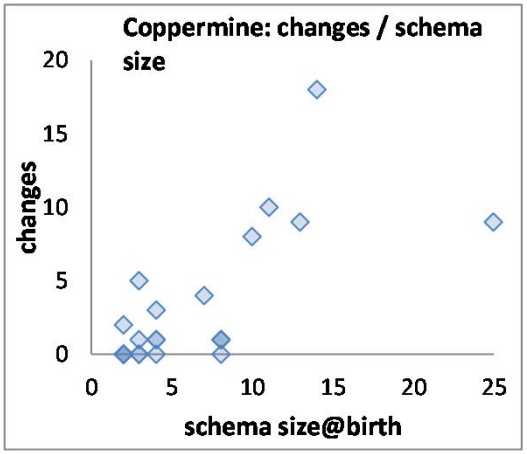
- Biosql: nobody exceeds 10 attributes
- Ensembl, mwiki: very few exceed 10 attributes, 3 of them died
- typo: has many late born survivors

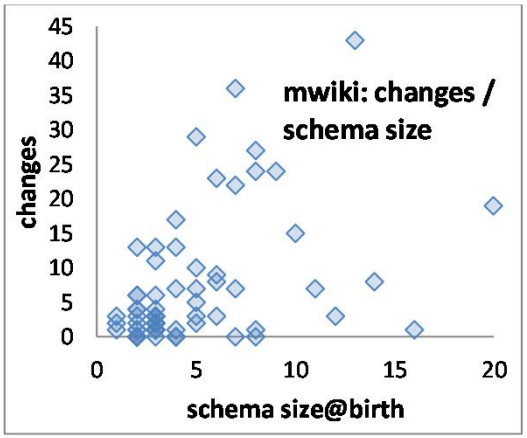
The Comet Pattern

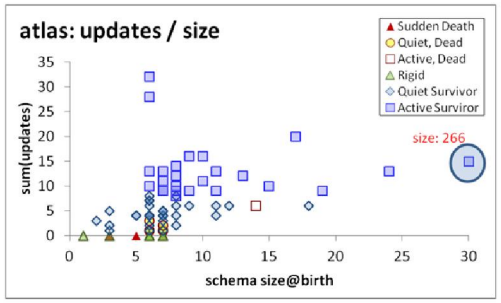
"Comet " for change over schema size with:

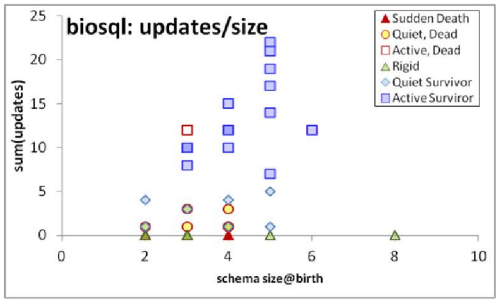
- a large, dense, nucleus cluster close to the beginning of the axes, denoting small size and small amount of change,
- medium schema size tables typically demonstrating medium to large change
  - The tables with the largest amount of change are typically tables whose schema is on average one standard 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.

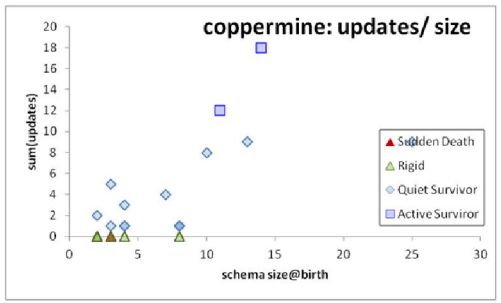


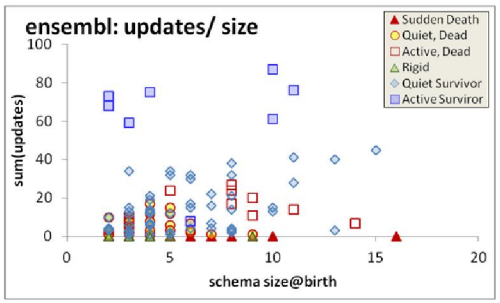


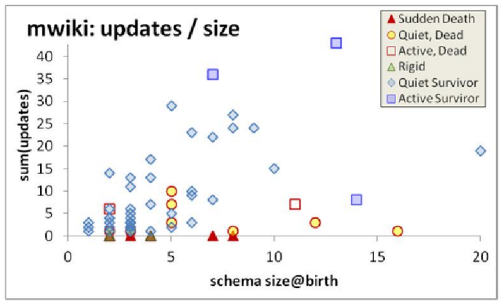


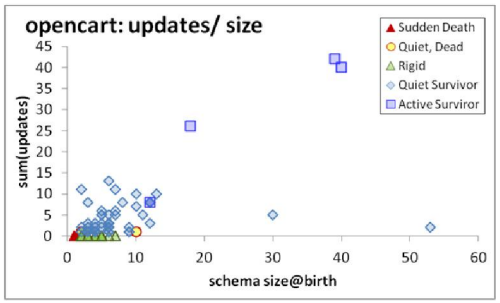


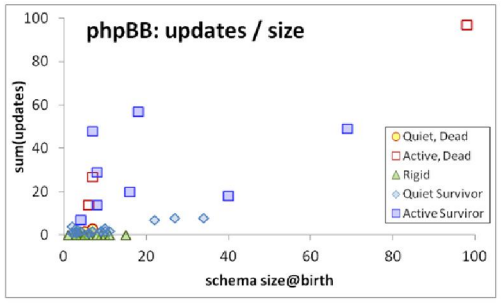


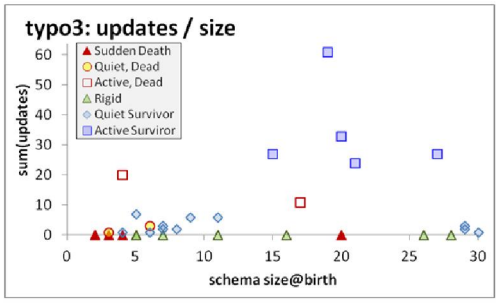




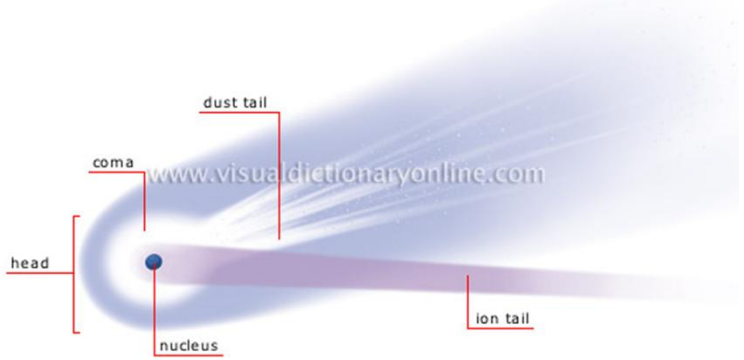


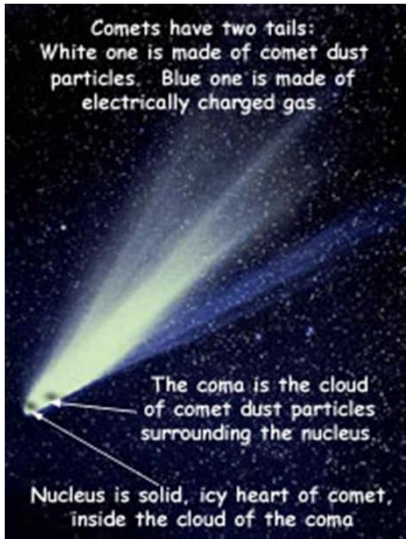






http://visual.merriam-webster.com/astronomy/celestial-bodies/comet.php

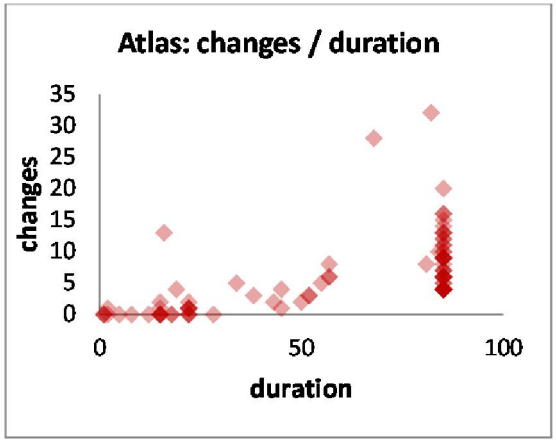


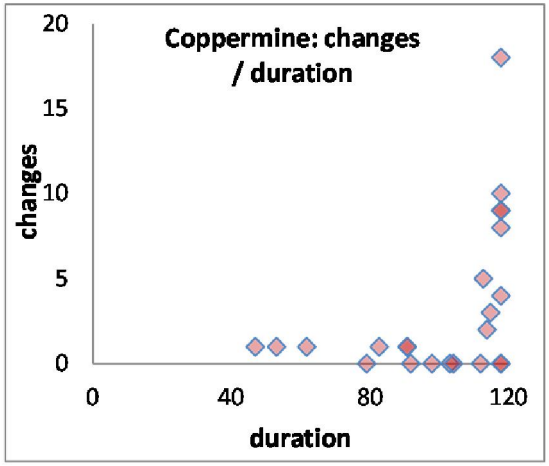


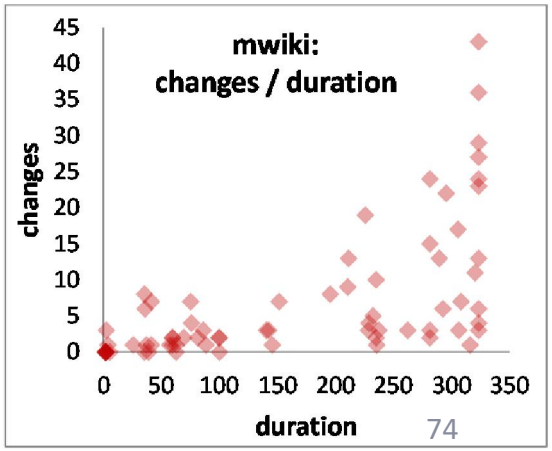
http://spaceplace.nasa.gov/comet-nucleus/en/

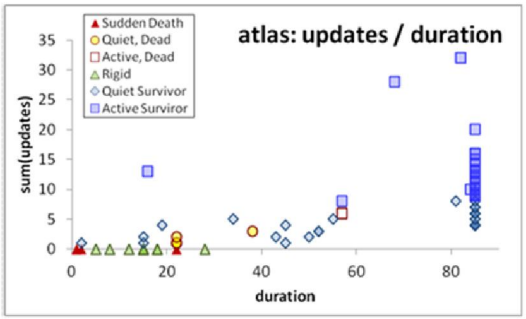
The inverse Gamma pattern

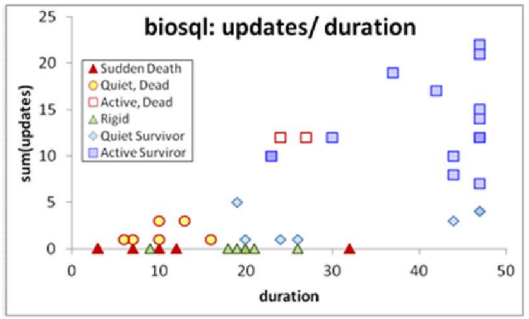
- The correlation of change and duration is as follows:
  - small durations come necessarily with small change,
  - large durations come with all kinds of change activity and
  - medium sized durations come mostly with small change activity (Inverse Gamma).

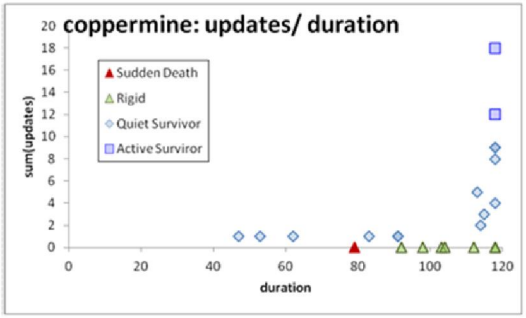


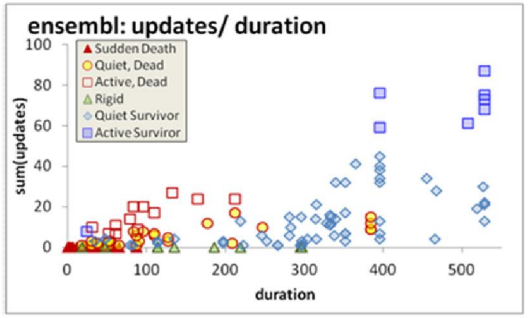


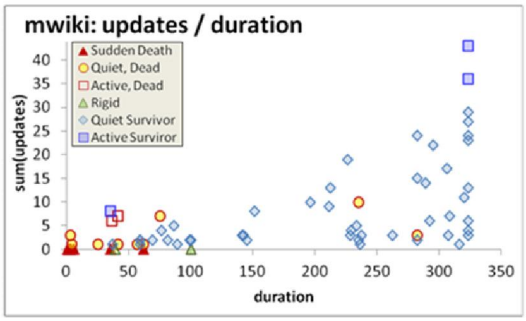


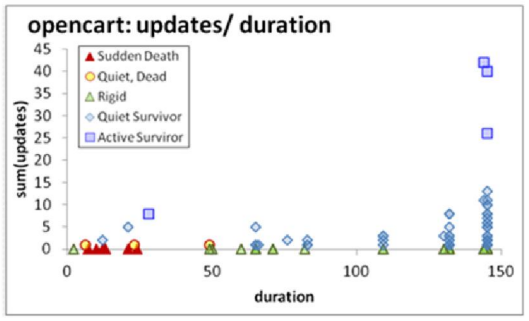

atlas: updates / duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

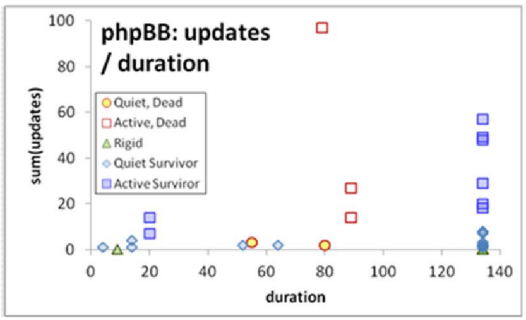

biosql: updates/ duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

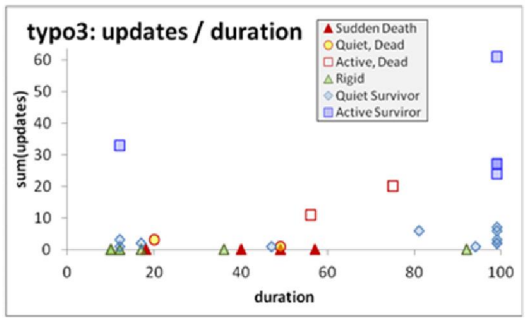

coppermine: updates/ duration
Sudden Death
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration


ensembl: updates/ duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration


mwiki: updates / duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration


opencart: updates/ duration
Sudden Death
Quiet, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration


phpBB: updates
/ duration
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration


typo3: updates / duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
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%**

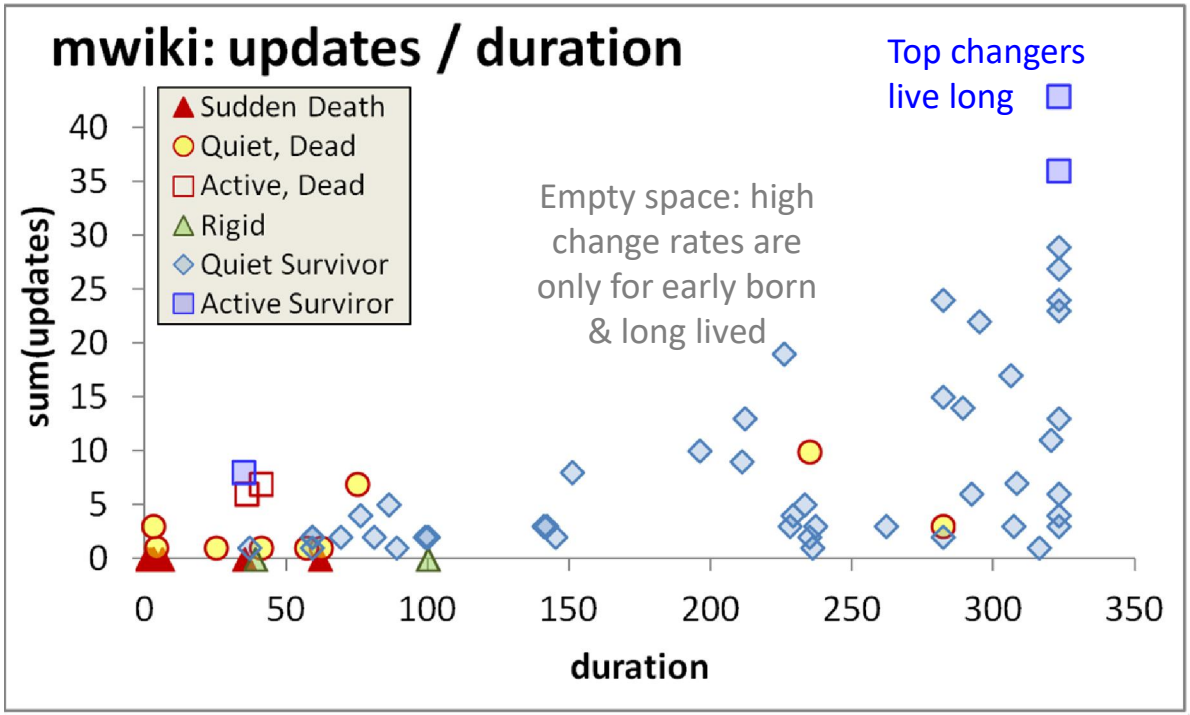
Non-survivors

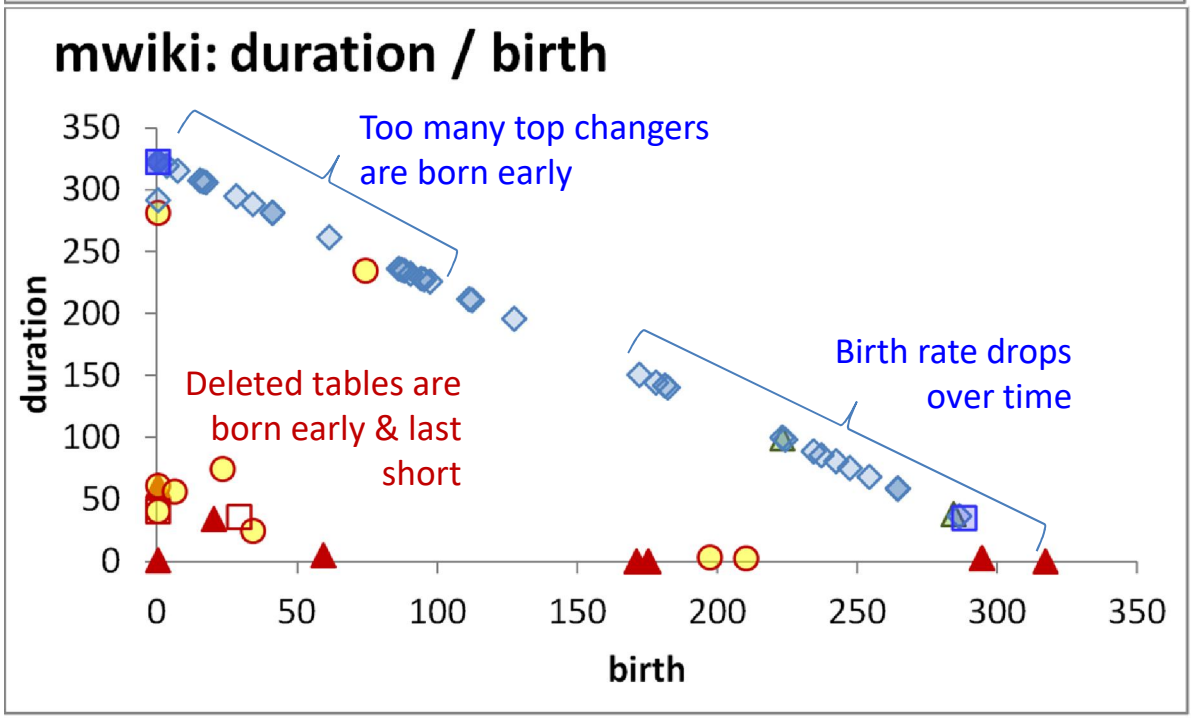
- Sudden deaths mostly
- Quiet come ~ close
- Too few active

Survivors

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

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

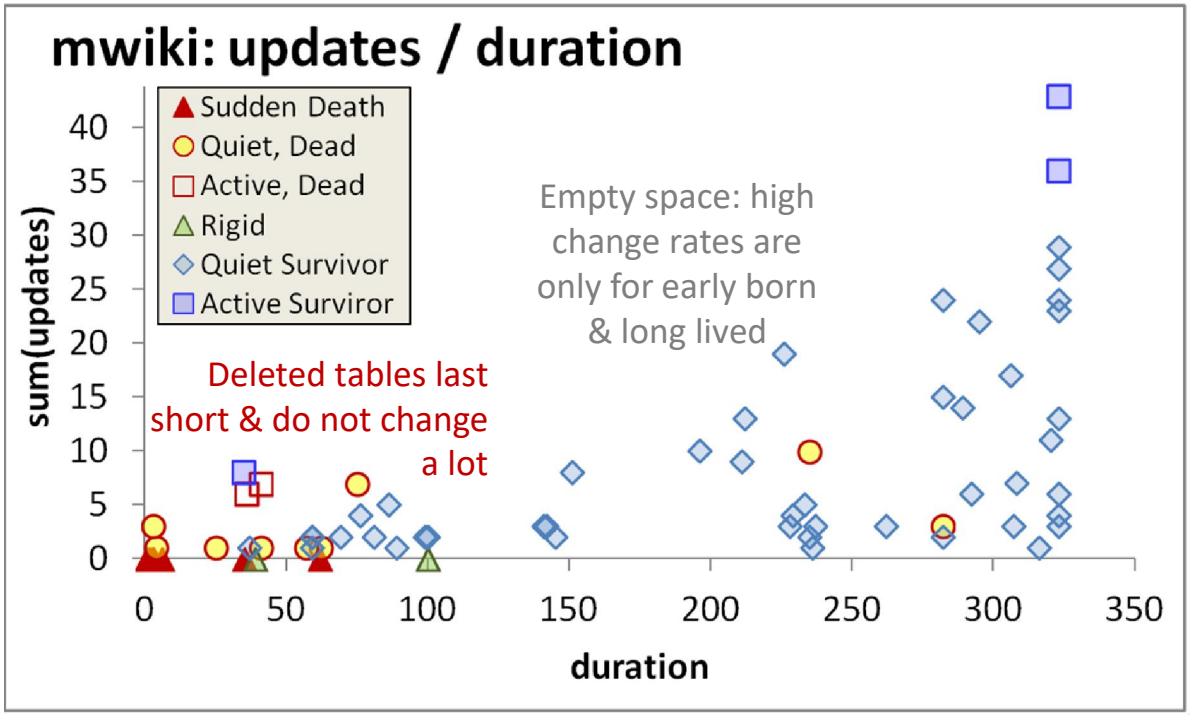


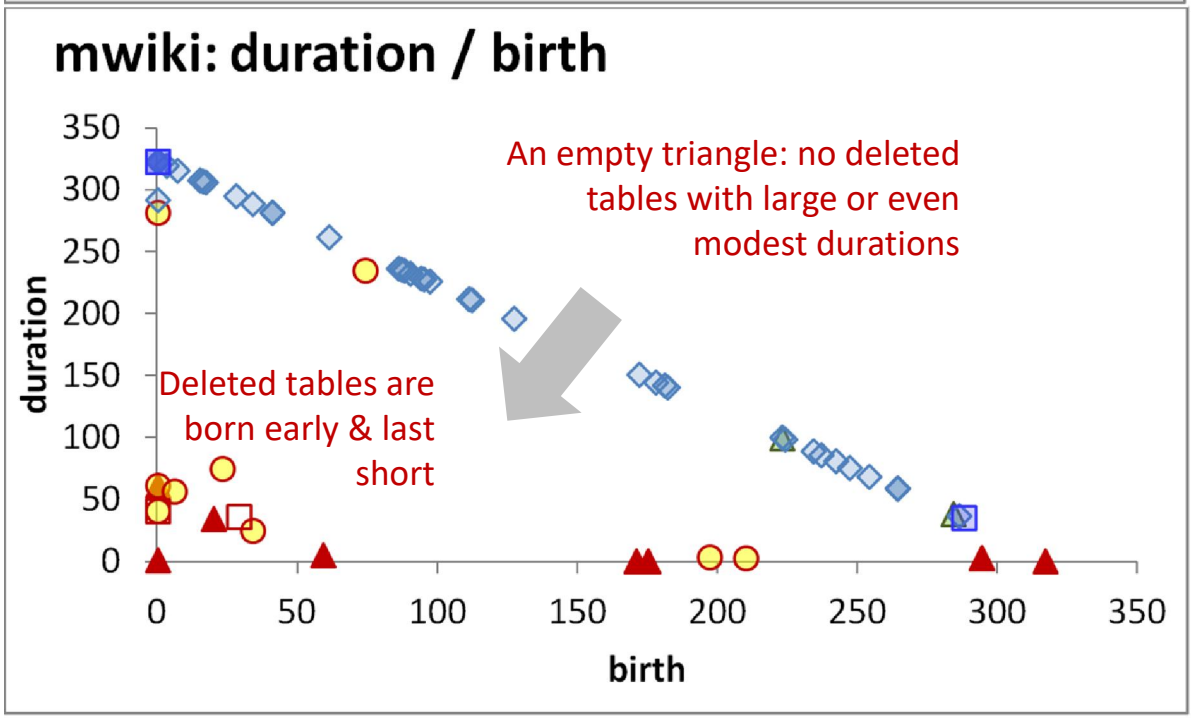


Longevity and update activity correlate !!

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

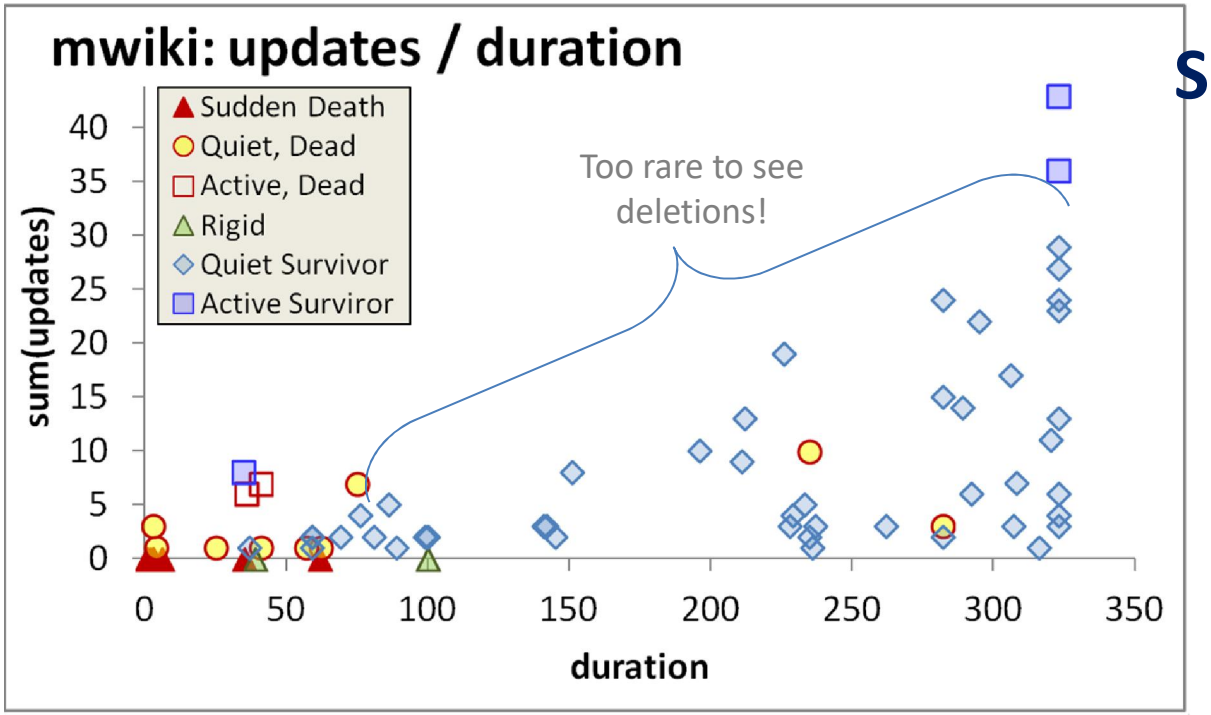
- are long lived,
- typically come from the early versions of the database
- due to the combination of high ATU and duration => they have high total amount of updates, and,
- frequently survive!

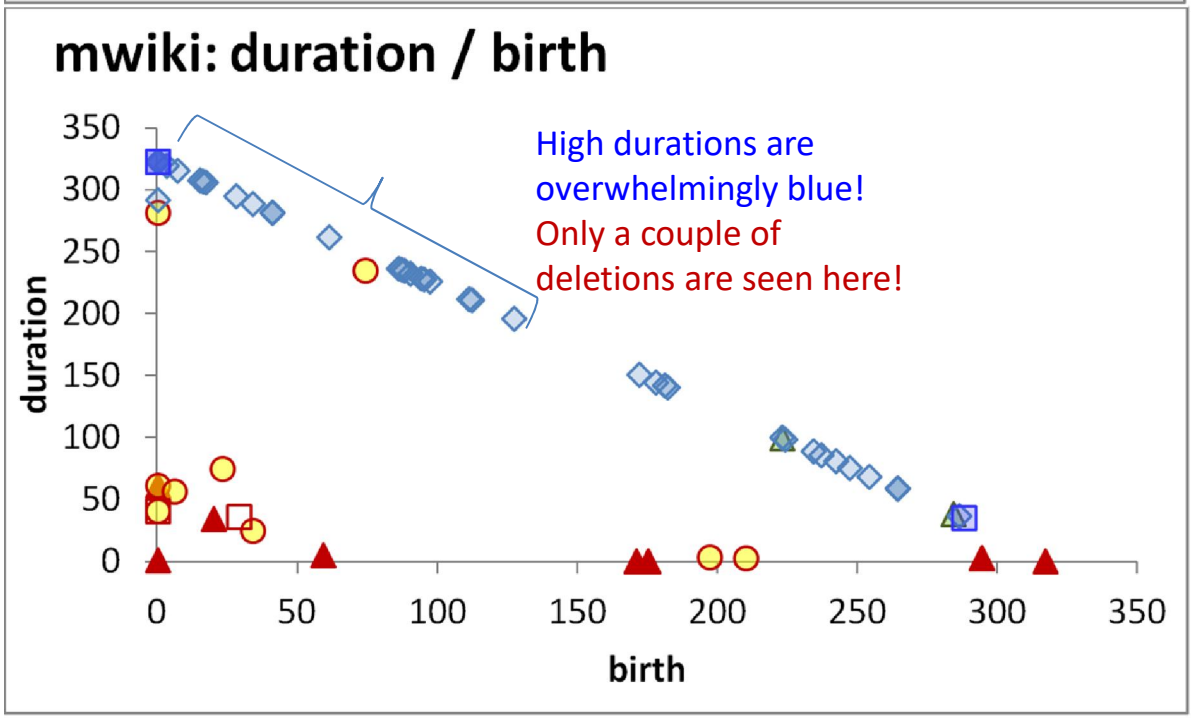




Die young and suddenly

- 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).

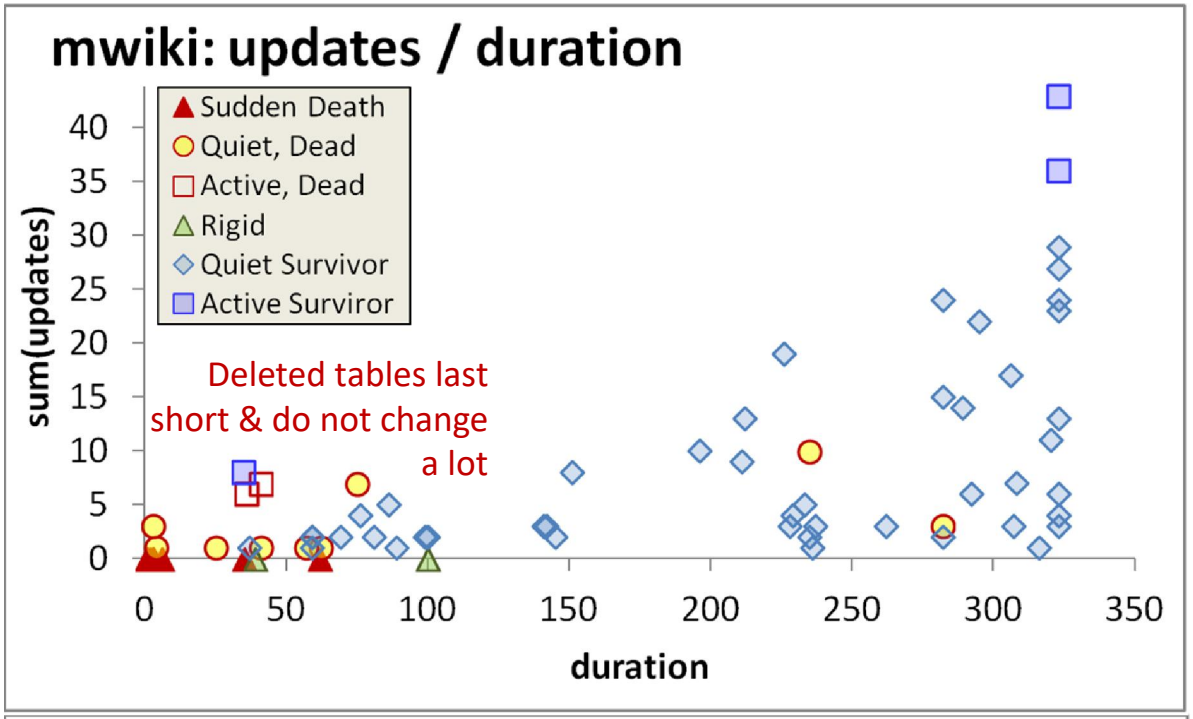


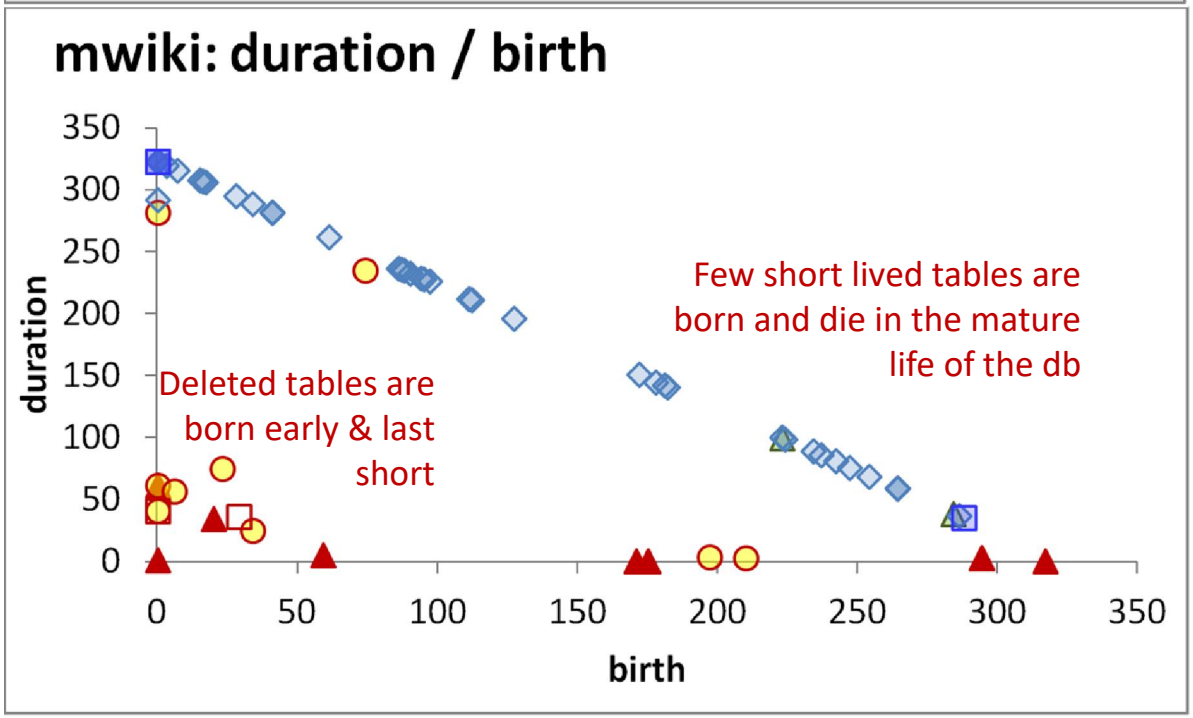


Survive long enough & you 're probably safe

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)!

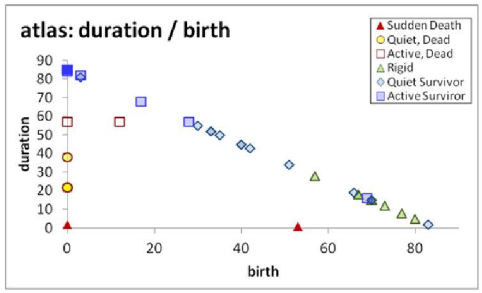


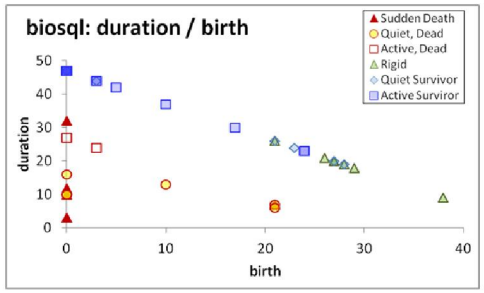


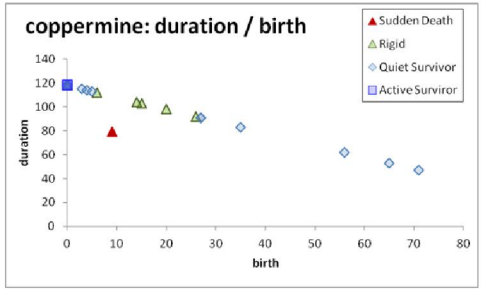
Die young and suddenly

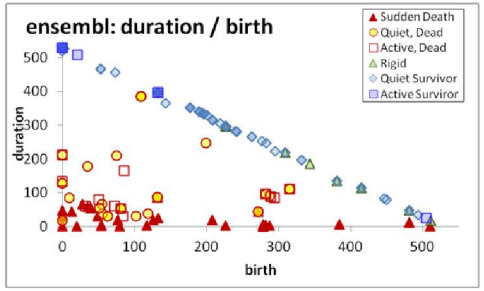
[Early life of the db] There is a very large concentration of the deleted tables in a small range of newly born, quickly removed, with few or no updates, resulting in very low numbers of removed tables with medium or long durations.

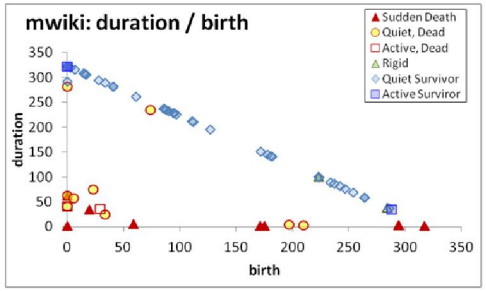
[**Mature db**] After the early stages of the databases, we see the birth of tables who eventually get deleted, but they mostly come with very small durations and sudden deaths.

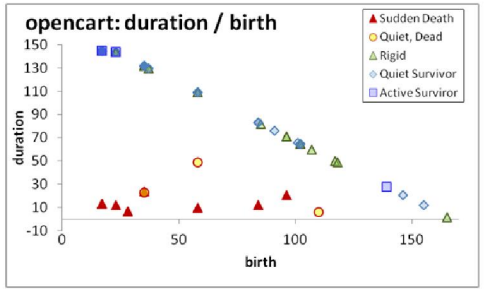

atlas: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth

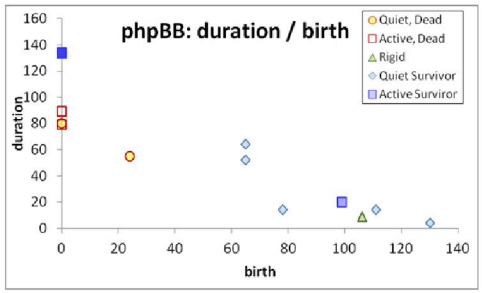

biosql: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth

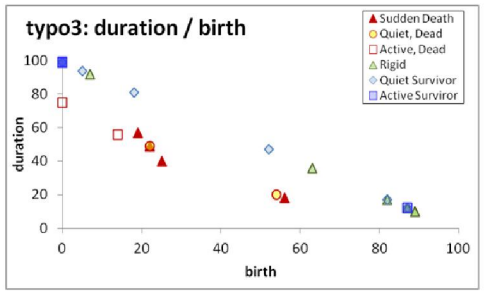

coppermine: duration / birth
Sudden Death
Rigid
Quiet Survivor
Active Survivor
duration
birth

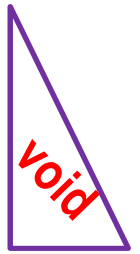

ensembl: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth


mwiki: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth


opencart: duration / birth
Sudden Death
Quiet, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth


phpBB: duration / birth
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth

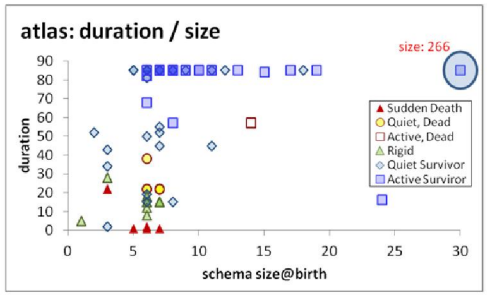

typo3: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth

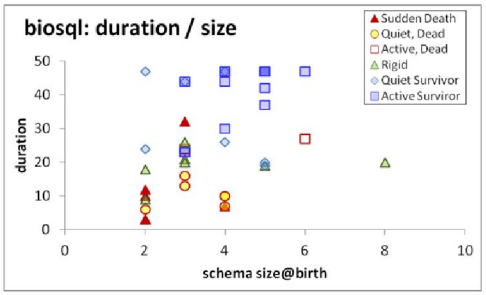

void

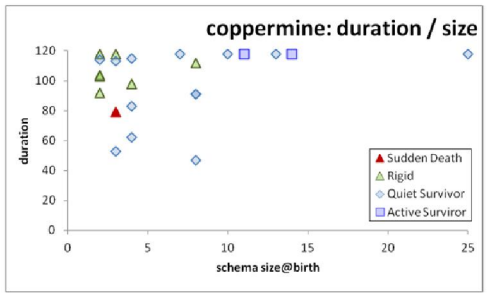
Schema size @ birth / duration

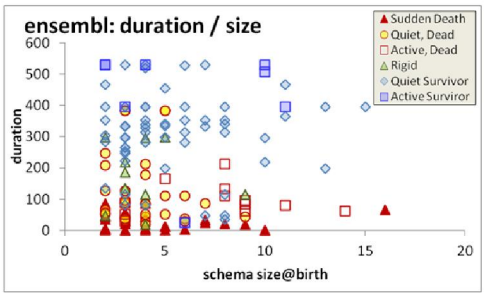
Only the thin die young, all the wide ones seem to live forever

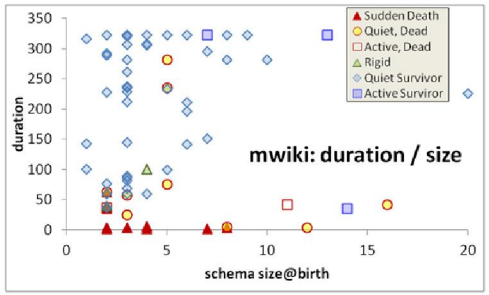
THE GAMMA PATTERN

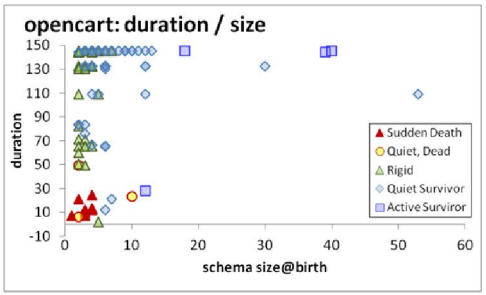


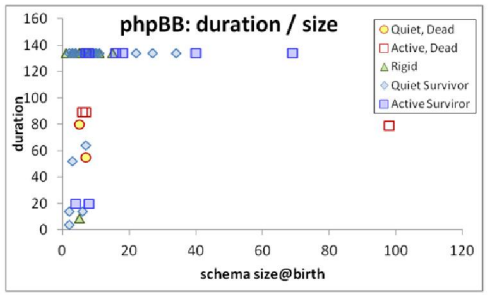


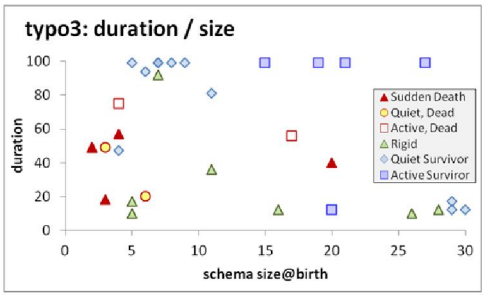












Exceptions

- 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 …*		
	# 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

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

	# Tables	# Wide tables	*As pct over #Tables…* …Wide	…Wide of long duration	*As pct over the set of Wide Tables …* … 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

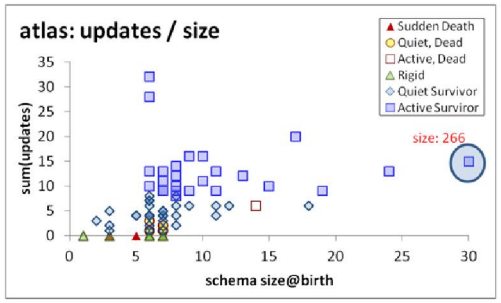
Subset relationship

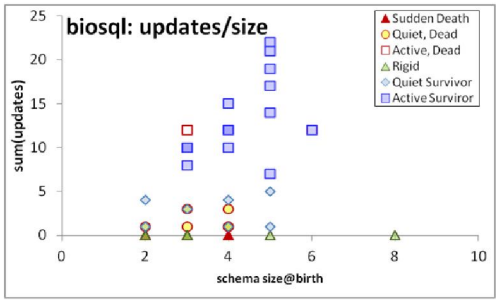
In all data sets, if a wide table has a long duration within the upper part of the Gamma, this deterministically (100% of all data sets) signifies that the table was also early born and survivor.

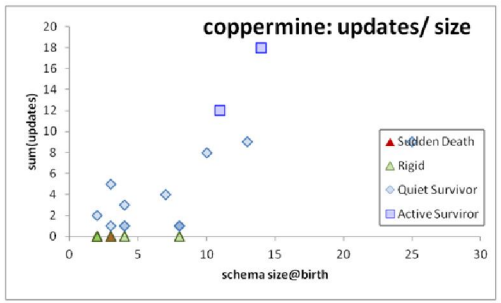
If a wide table is in the top of the Gamma line, it is deterministically an early born survivor.

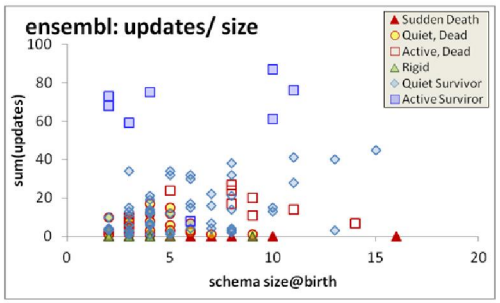
Schema size and updates

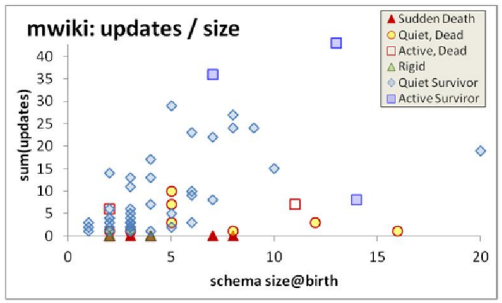
THE COMET PATTERN

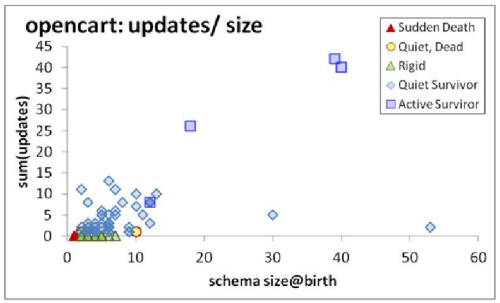


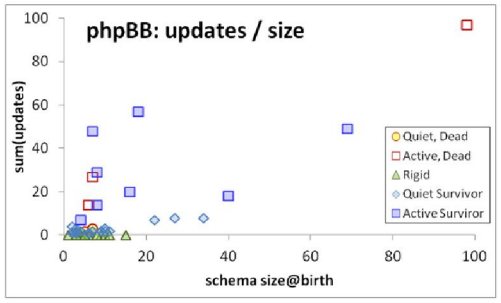


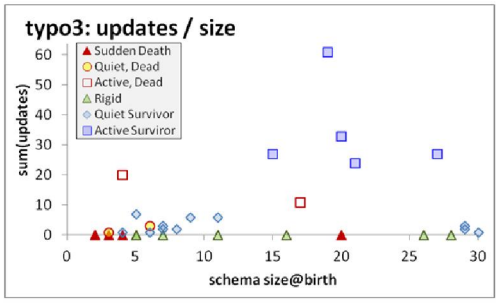




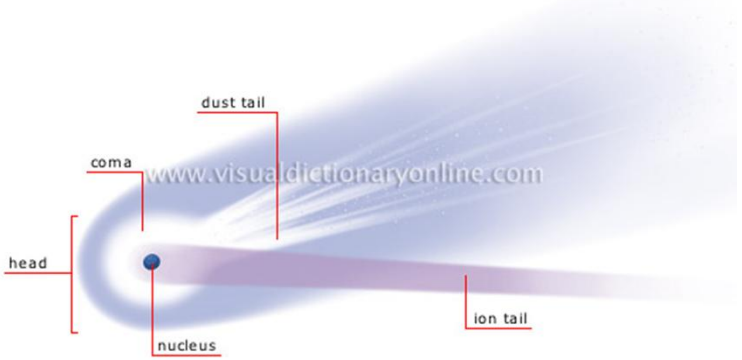


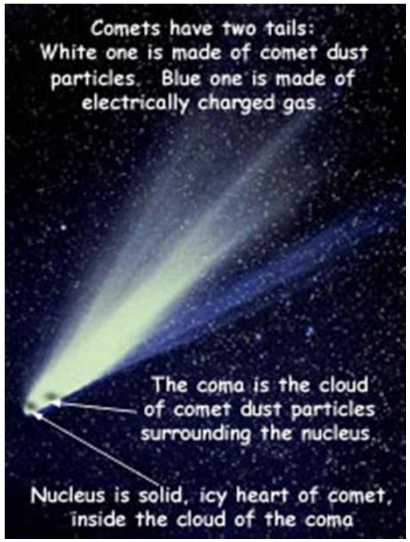






http://visual.merriam-webster.com/astronomy/celestial-bodies/comet.php





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*/*

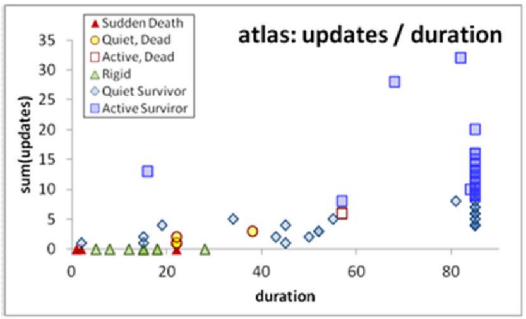
- In 5 out of 8 cases, the average schema size of top-changers within 0.4 and 0.5 of the maximum value (practically the middle of the domain) and never above 0.65 of it.
- Pearson: the maximum value, the standard deviation of the entire data set and the average of the top changers are very strongly correlated.

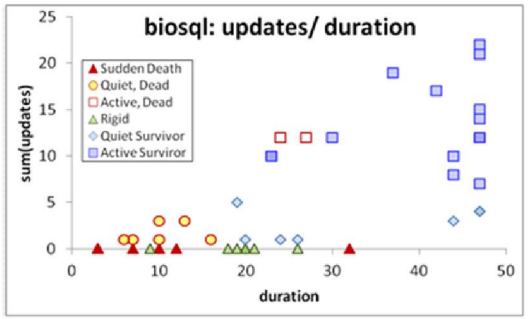
Wide tables have a medium number of updates

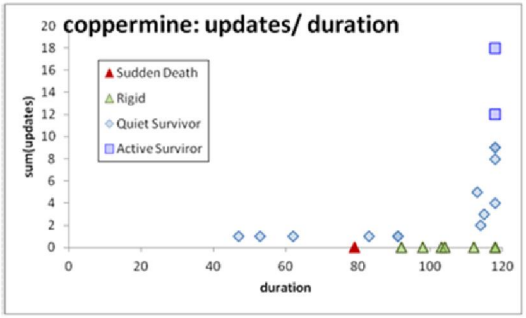
Total amt. of updates. Statistics for ...		... the entire data set					... the top 5% with respect to schema size at birth (top wide)			
	#tables	max	mean (μ)	stdev (σ)	μ+σ	max/2	avg upd. of top 5%	upd. of top 1	avg of top 5% / max	Top up. in wide?
atlas	88	32	5.86	11.81	11.81	16.0	**12.60**	20	0.39	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*		

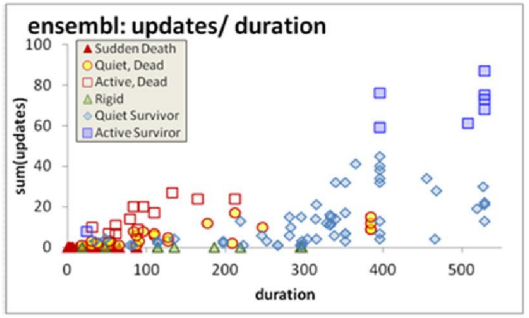
For each data set, we took the top 5% in terms of schema size at birth (**top wide**) and contrasted their update behavior wrt the update behavior of the entire data set.
Typically, the avg. number of updates of the top wide tables is close to the 50% of the domain of values for the sum of updates (i.e., 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).

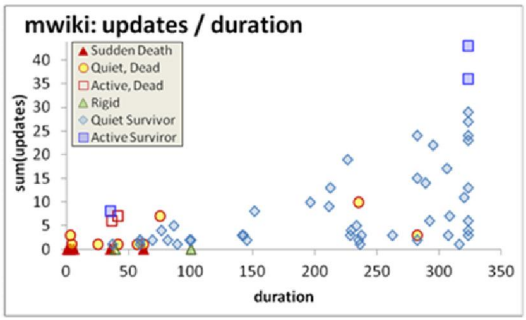
INVERSE GAMMA

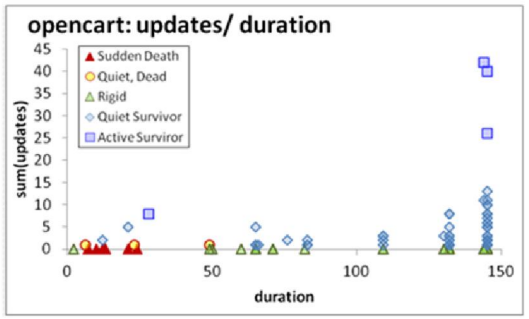

atlas: updates / duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

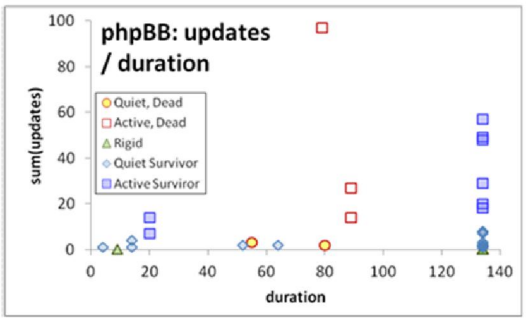

biosql: updates/ duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

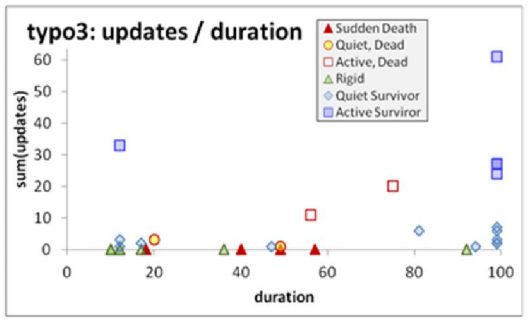

coppermine: updates/ duration
Sudden Death
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

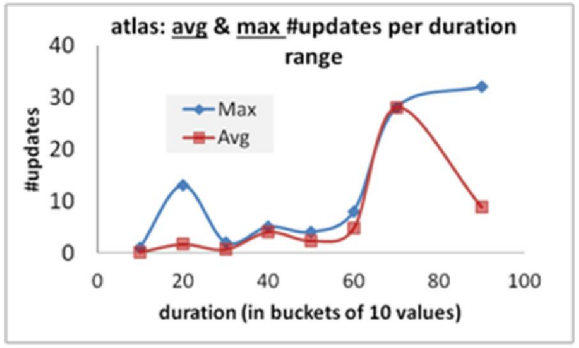

ensembl: updates/ duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

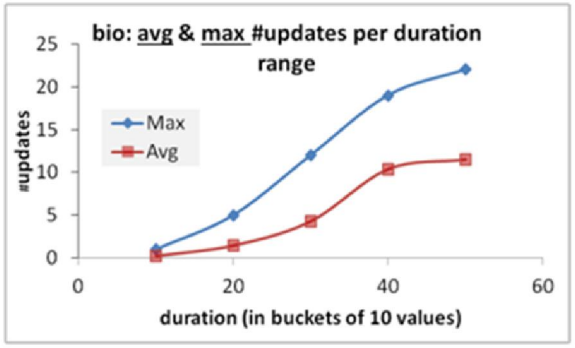

mwiki: updates / duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

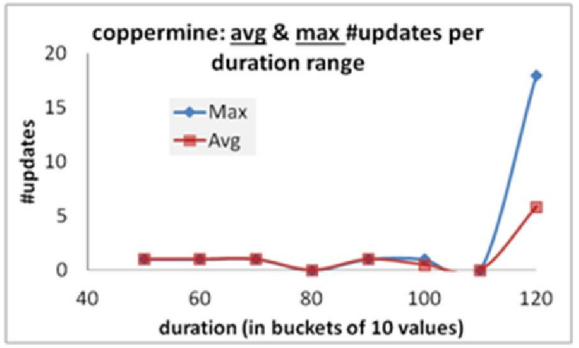

opencart: updates/ duration
Sudden Death
Quiet, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

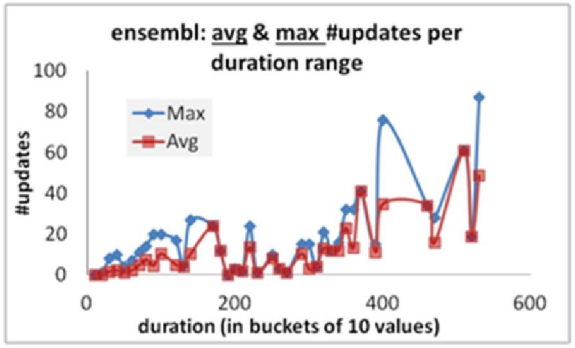

phpBB: updates
/ duration
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

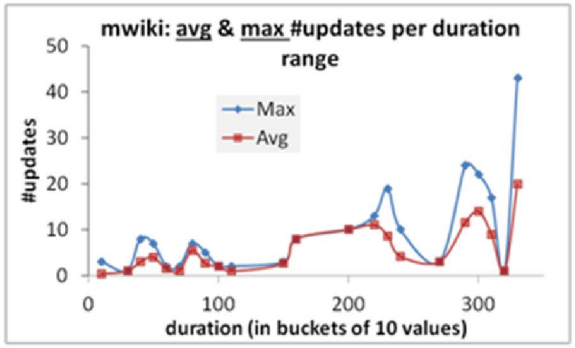

typo3: updates / duration
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Surviror
sum(updates)
duration

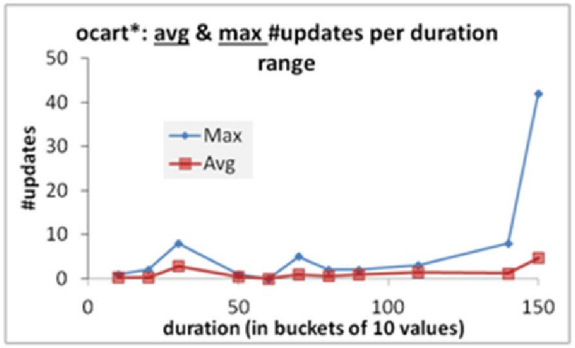


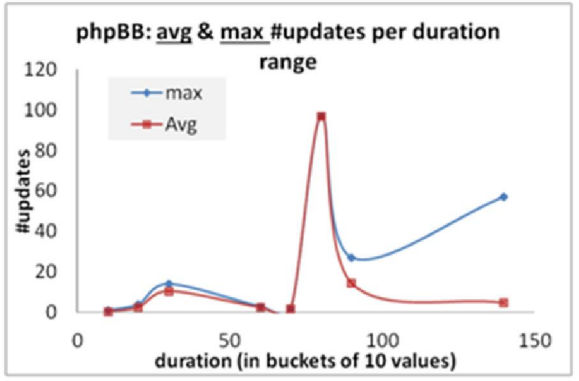


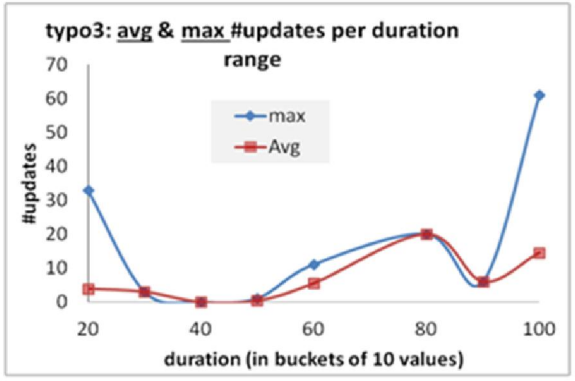






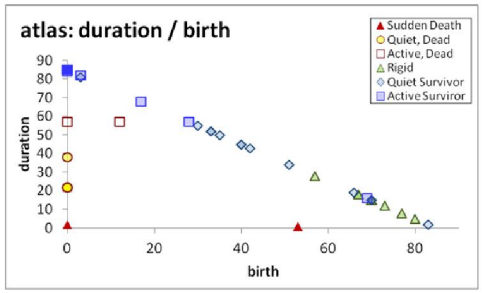


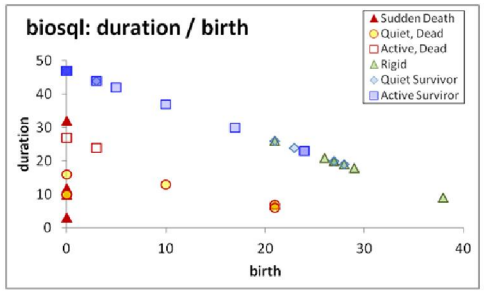


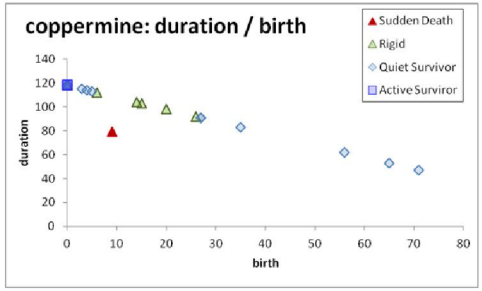


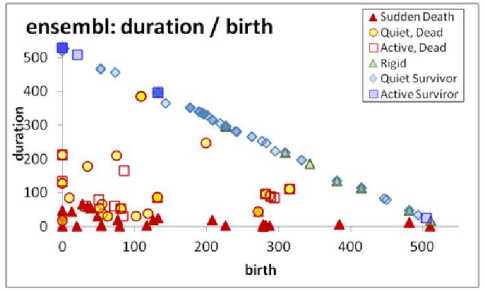
Skyline & Avg for Inverse Gamma

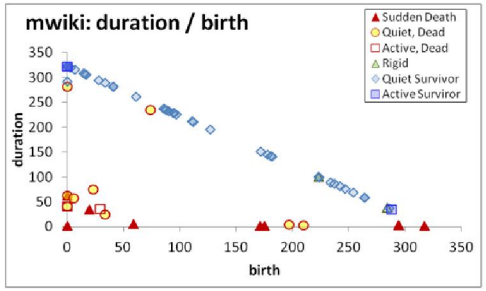
THE EMPTY TRIANGLE PATTERN

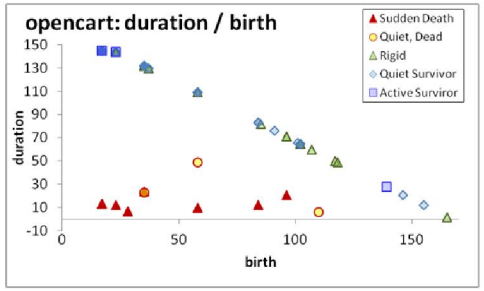

atlas: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth

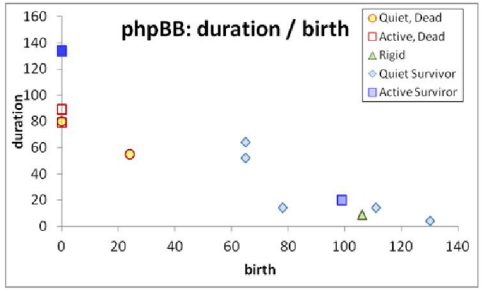

biosql: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth

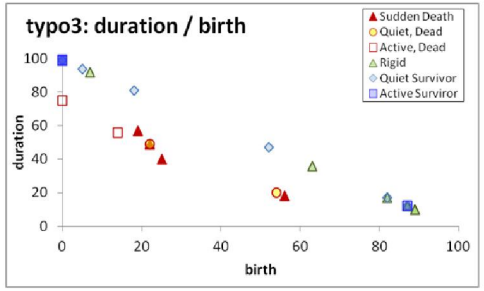

coppermine: duration / birth
Sudden Death
Rigid
Quiet Survivor
Active Survivor
duration
birth

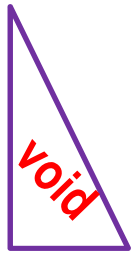

ensembl: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth


mwiki: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth


opencart: duration / birth
Sudden Death
Quiet, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth


phpBB: duration / birth
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth


typo3: duration / birth
Sudden Death
Quiet, Dead
Active, Dead
Rigid
Quiet Survivor
Active Survivor
duration
birth


void

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

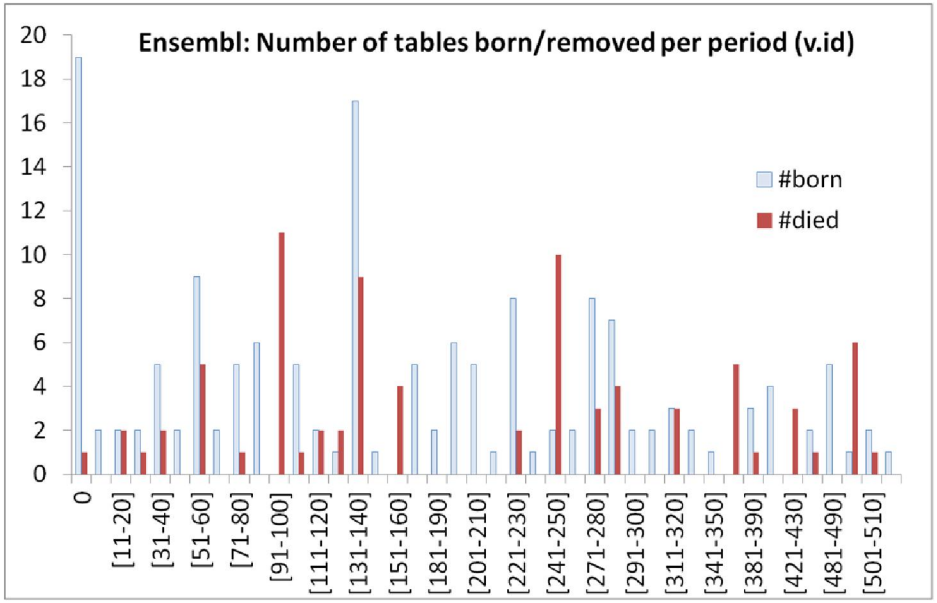
	atlas	biosql	coppermine	ensembl	mwiki	ocart*	phpBB	typo3
Tables	88	45	23	155	71	128	70	32
Active	27	16	2	23	5	4	11	7
active tables(%)	31%	36%	9%	15%	7%	3%	16%	22%
As percentages over active								
Born early	96%	81%	100%	78%	80%	75%	82%	86%
Survivors	93%	88%	100%	48%	60%	100%	73%	71%
Long duration	85%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

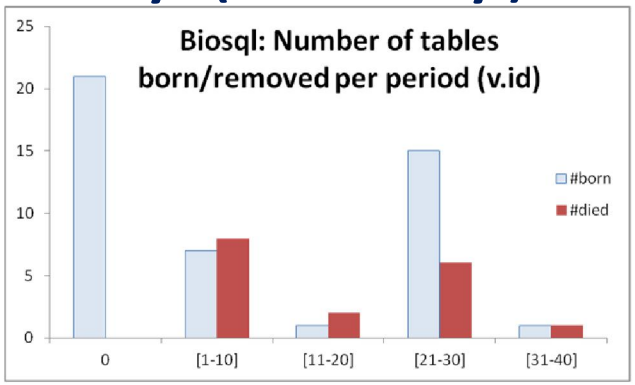
- In all data sets, active tables are born early with percentages that exceed 75%
- With the exceptions of two data sets, they survive with percentage higher than 70%.
- The probability of having a long duration is higher than 50% in 6 out of 8 data sets.
- Interestingly, **the two last lines are exactly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

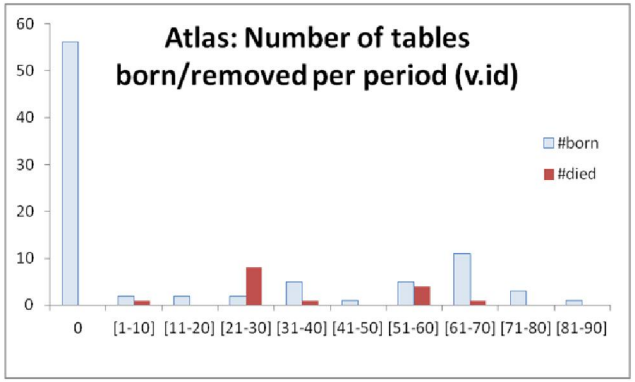
Dead are: quiet, early born, short lived, and quite often all three of them

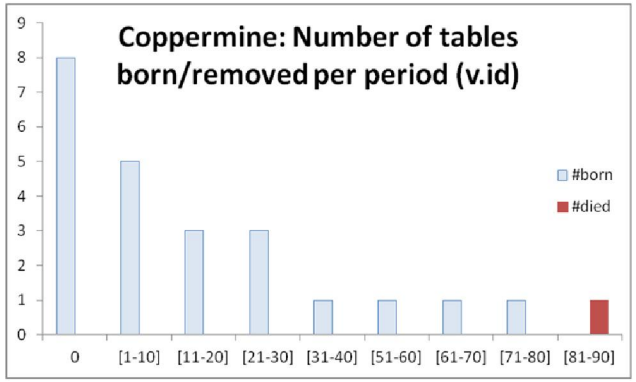
	atlas	biosql	coppermine	ensembl	mwiki	ocart*	phpBB	typo3
tables	88	45	23	155	71	128	70	32
dead	15	17	1	80	21	14	5	9
dead tables(%)	17%	38%	4%	52%	30%	11%	7%	28%
As percentages over # dead								
Few updates	87%	88%	100%	85%	90%	100%	40%	78%
Early born	80%	82%	100%	70%	62%	71%	100%	78%
Short-lived	80%	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%

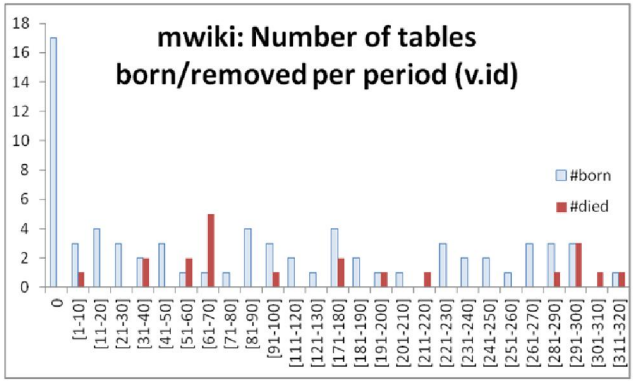
Most births &deaths occur early (usually)

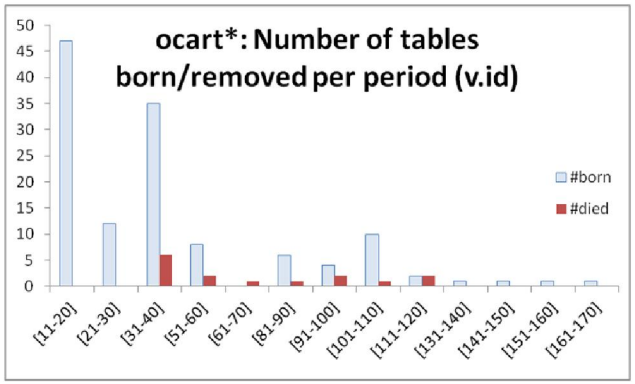


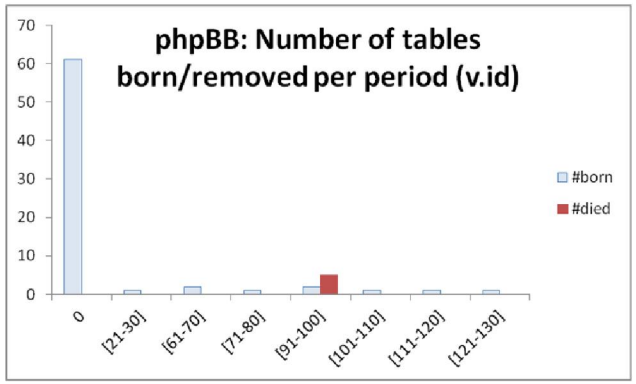


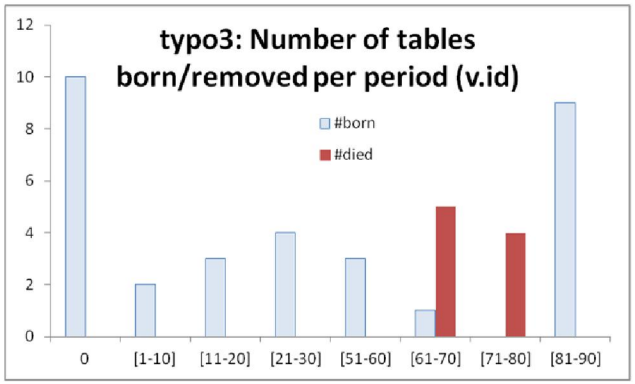


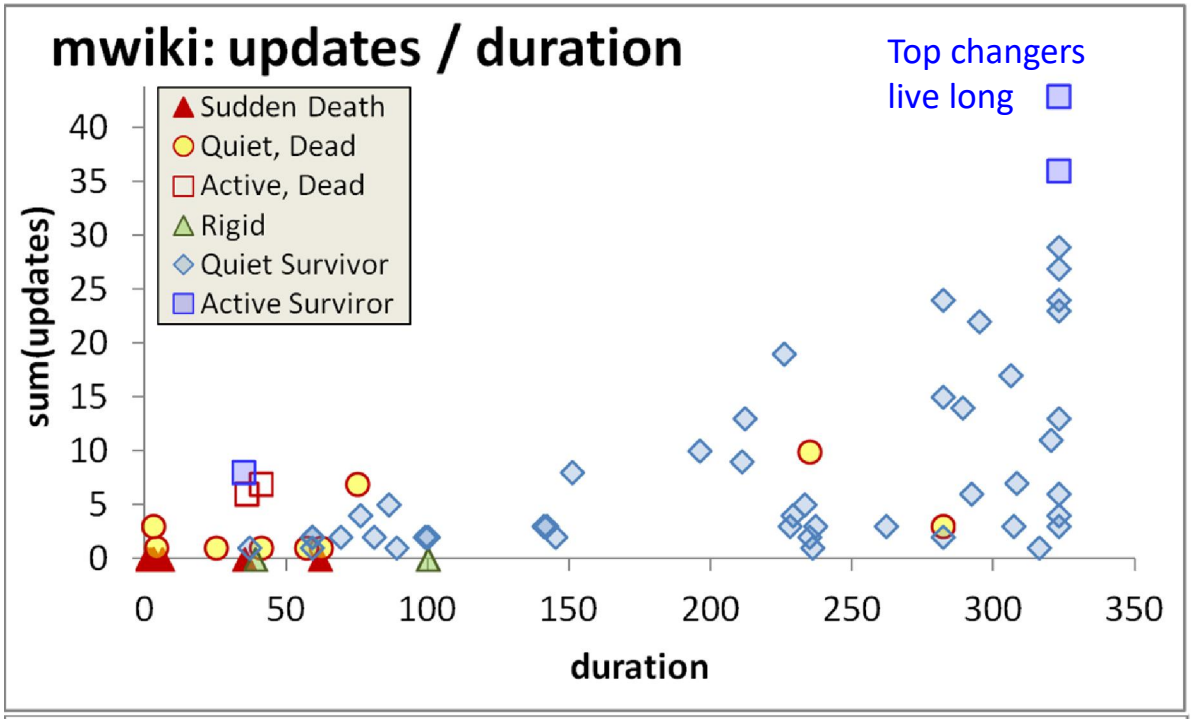


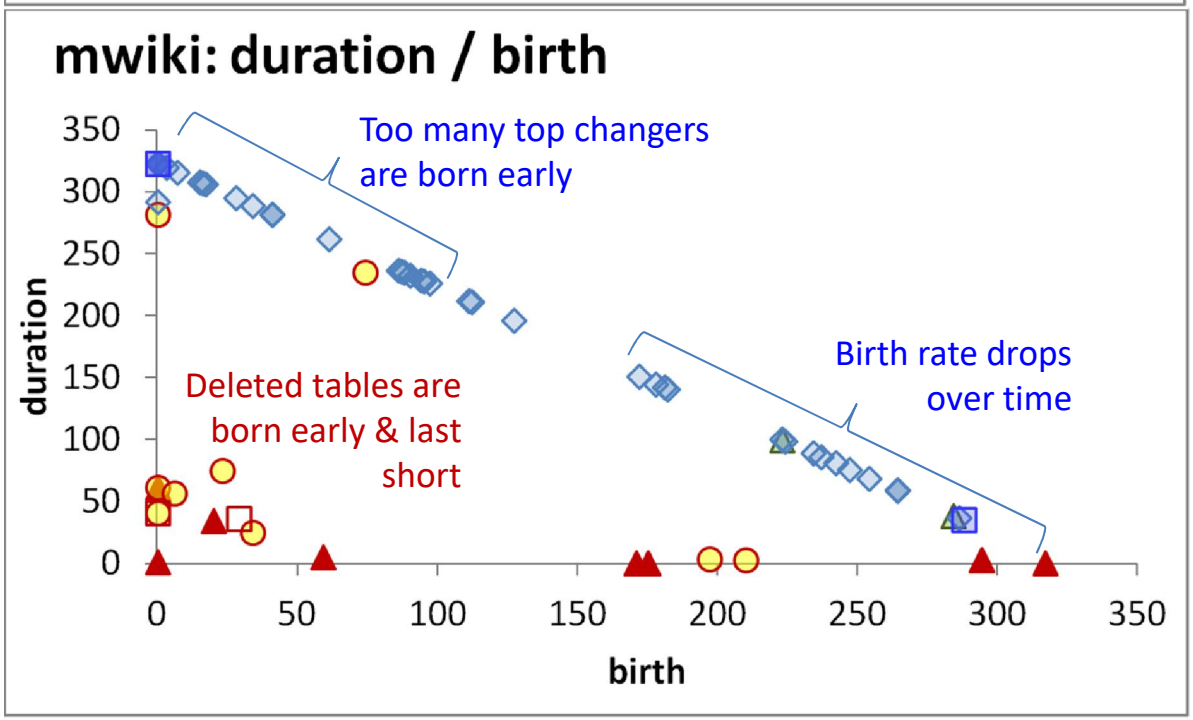






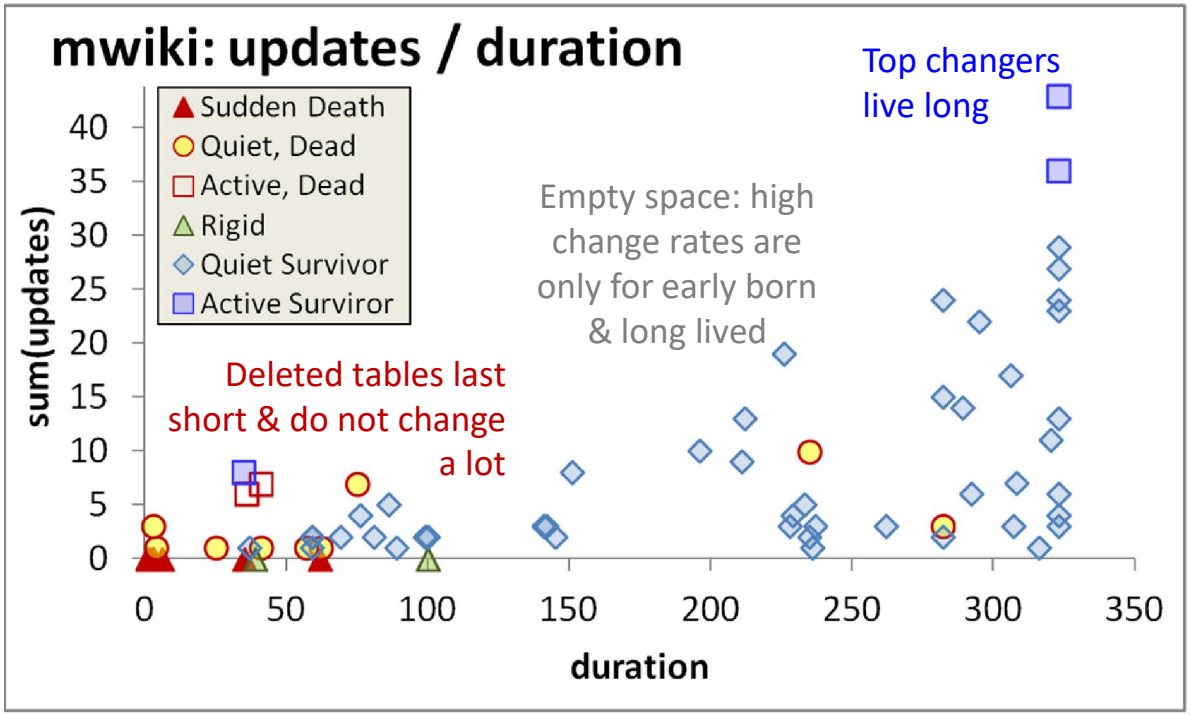


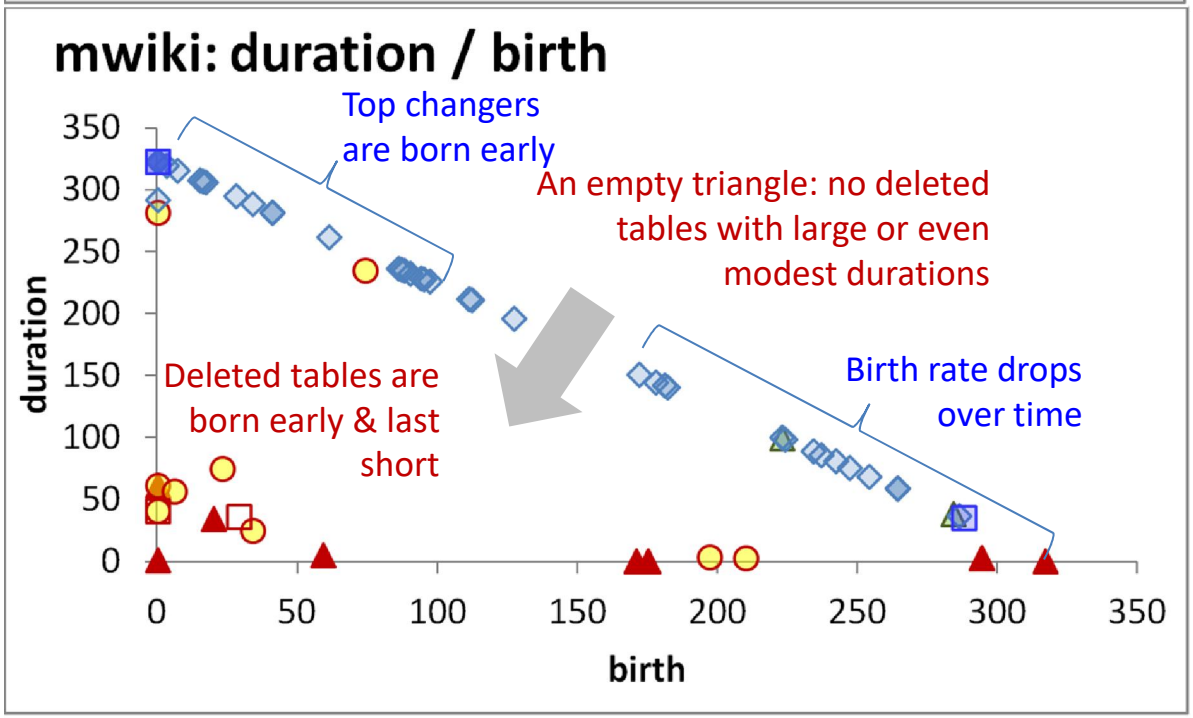




Longevity and update activity correlate !!

- 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!





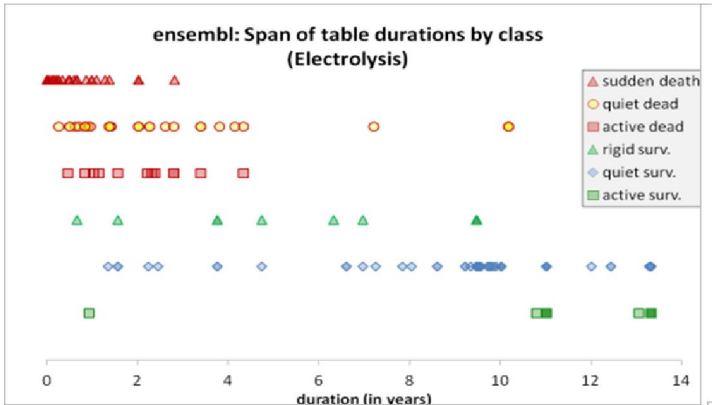
All in one

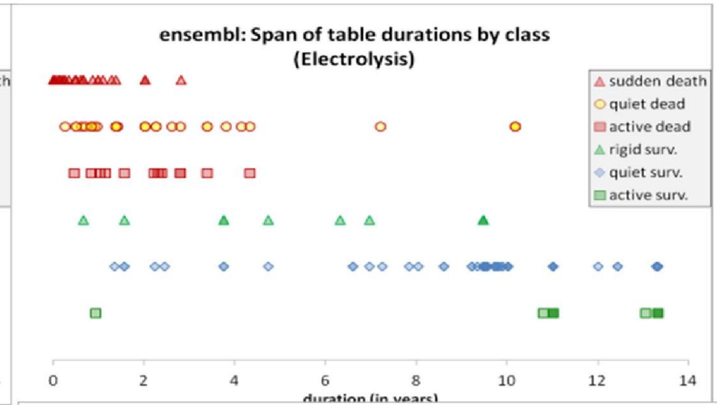
- Early stages of the database life are more "active" in terms of births, deaths and updates, and have higher chances of producing deleted tables.
- After the first major restructuring, the database continues to grow; however, we see much less removals, and maintenance activity becomes more concentrated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

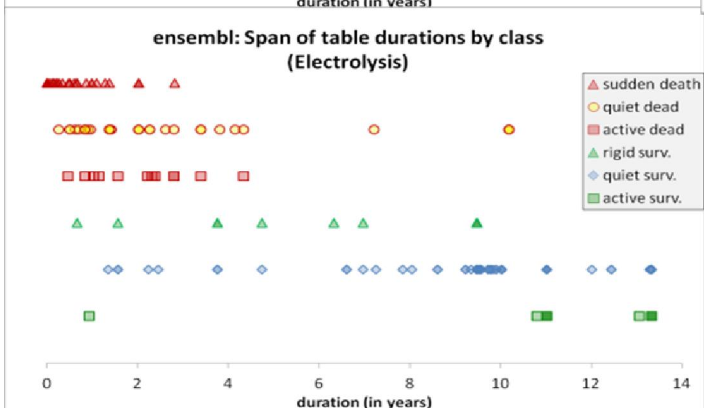
ELECTROLYSIS @ CAISE 2017

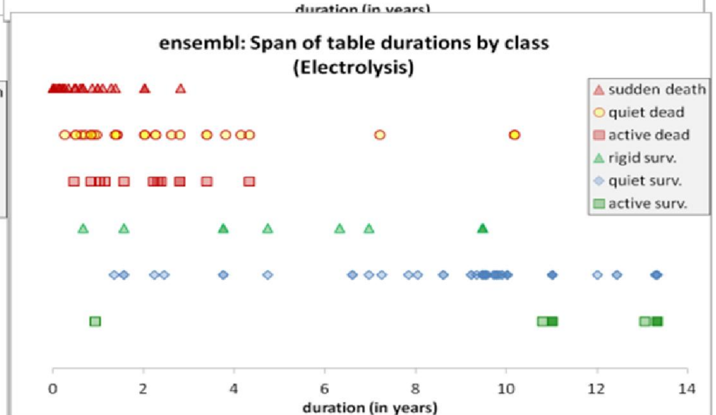
Why do we see what we see

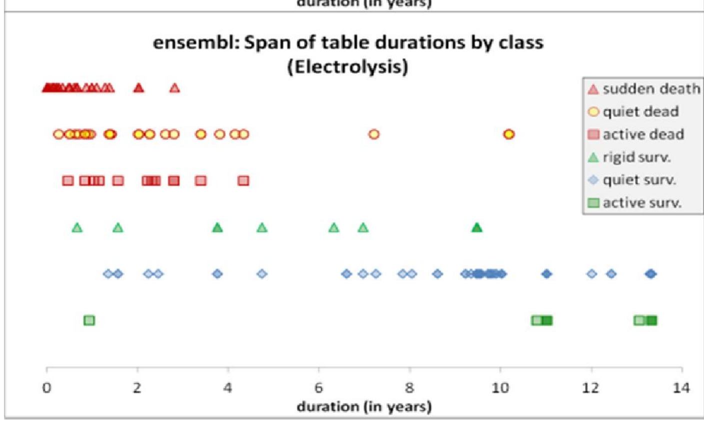
- We believe that this study strengthens our theory that schema evolution antagonizes a powerful **gravitation to rigidity**.
- DB's = "dependency magnets"
  - all the application code relies on them but not vice versa, =>
  - avoiding schema evolution reduces the need for adaptation and maintenance of application code

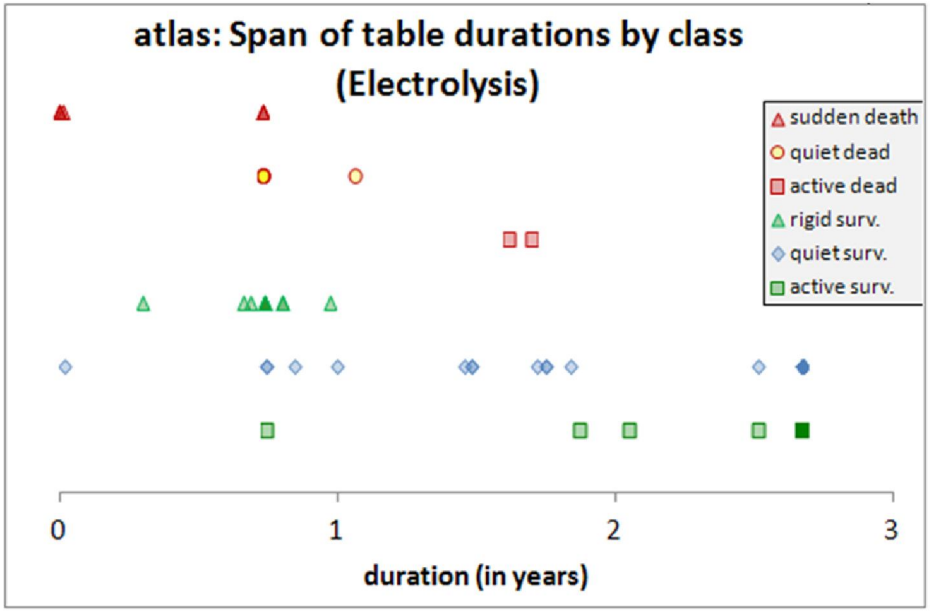

ensembl: Span of table durations by class
(Electrolysis)
sudden death
quiet dead
active dead
rigid surv.
quiet surv.
active surv.
0
2
4
6
8
10
12
14
duration (in years)

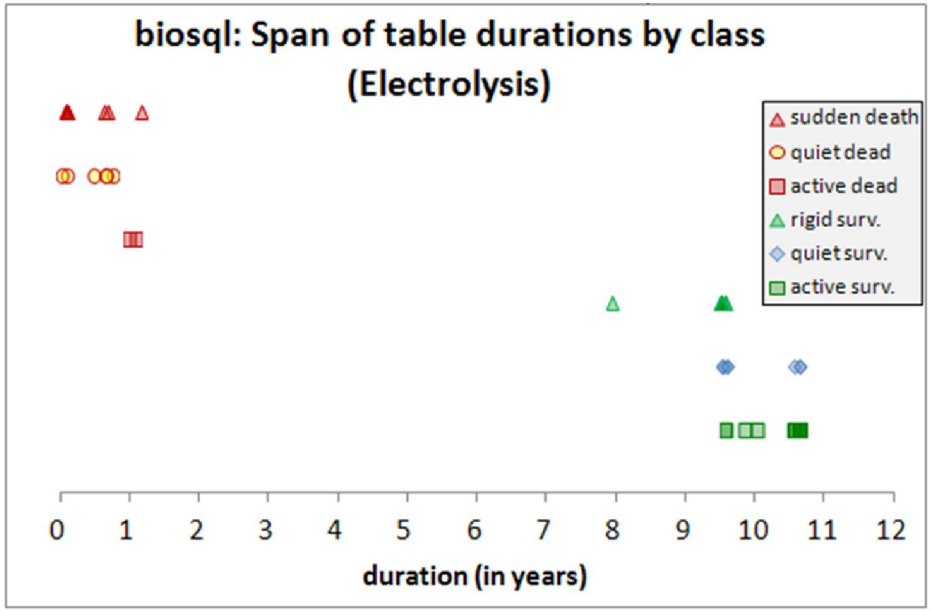

ensembl: Span of table durations by class
(Electrolysis)
sudden death
quiet dead
active dead
rigid surv.
quiet surv.
active surv.
0
2
4
6
8
10
12
14

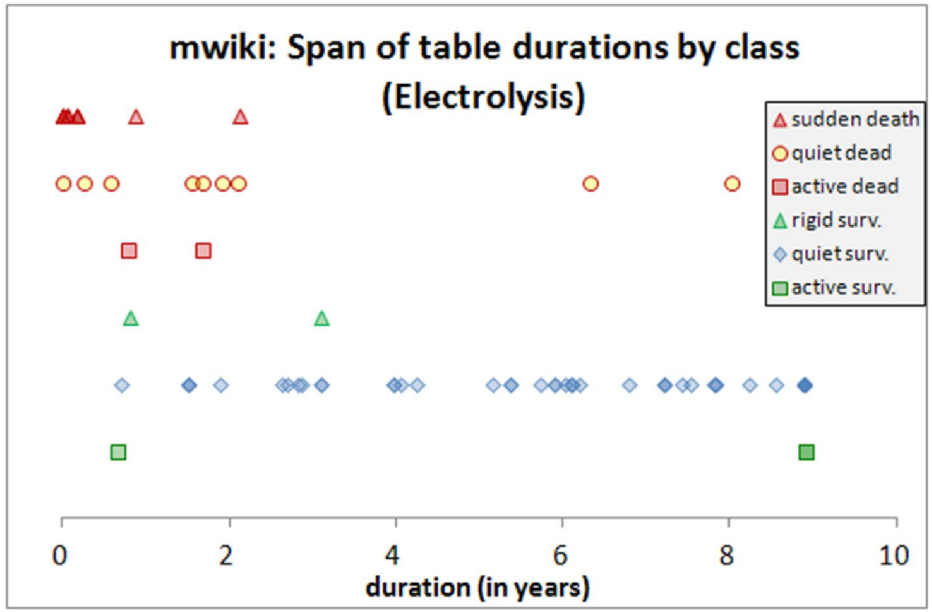

ensembl: Span of table durations by class
(Electrolysis)
sudden death
quiet dead
active dead
rigid surv.
quiet surv.
active surv.
0
2
4
6
8
10
12
14
duration (in years)

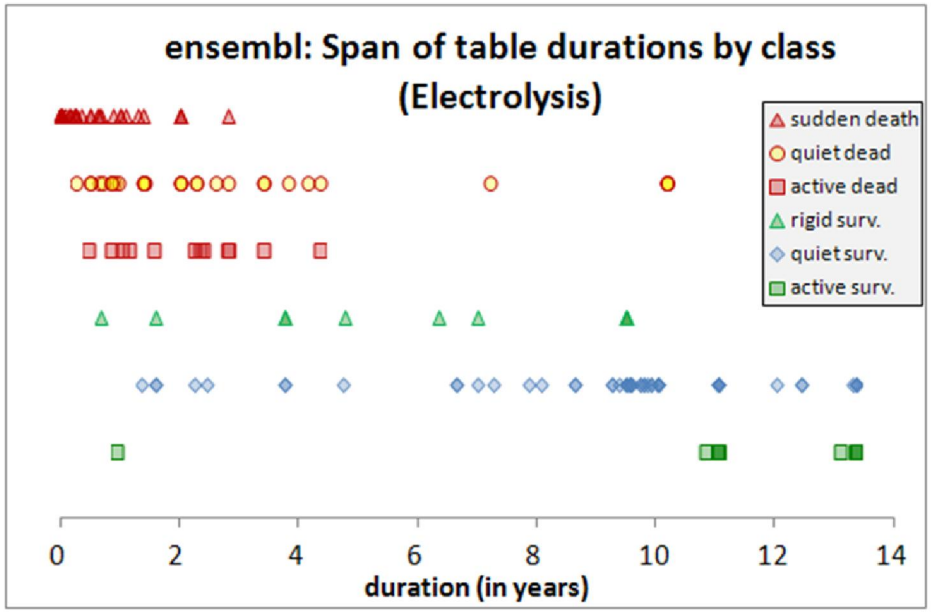

ensembl: Span of table durations by class
(Electrolysis)
sudden death
quiet dead
active dead
rigid surv.
quiet surv.
active surv.
0
2
4
6
8
10
12
14
duration (in years)


ensembl: Span of table durations by class
(Electrolysis)
sudden death
quiet dead
active dead
rigid surv.
quiet surv.
active surv.
0
2
4
6
8
10
12
14
duration (in years)


atlas: Span of table durations by class (Electrolysis)
sudden death
quiet dead
active dead
rigid surv.
quiet surv.
active surv.
0
1
2
3
duration (in years)


biosql: Span of table durations by class (Electrolysis)
sudden death
quiet dead
active dead
rigid surv.
quiet surv.
active surv.
0
1
2
3
4
5
6
7
8
9
10
11
12
duration (in years)


mwiki: Span of table durations by class (Electrolysis)
sudden death
quiet dead
active dead
rigid surv.
quiet surv.
active surv.
0
2
4
6
8
10
duration (in years)


ensembl: Span of table durations by class (Electrolysis)
sudden death
quiet dead
active dead
rigid surv.
quiet surv.
active surv.
0
2
4
6
8
10
12
14
duration (in years)

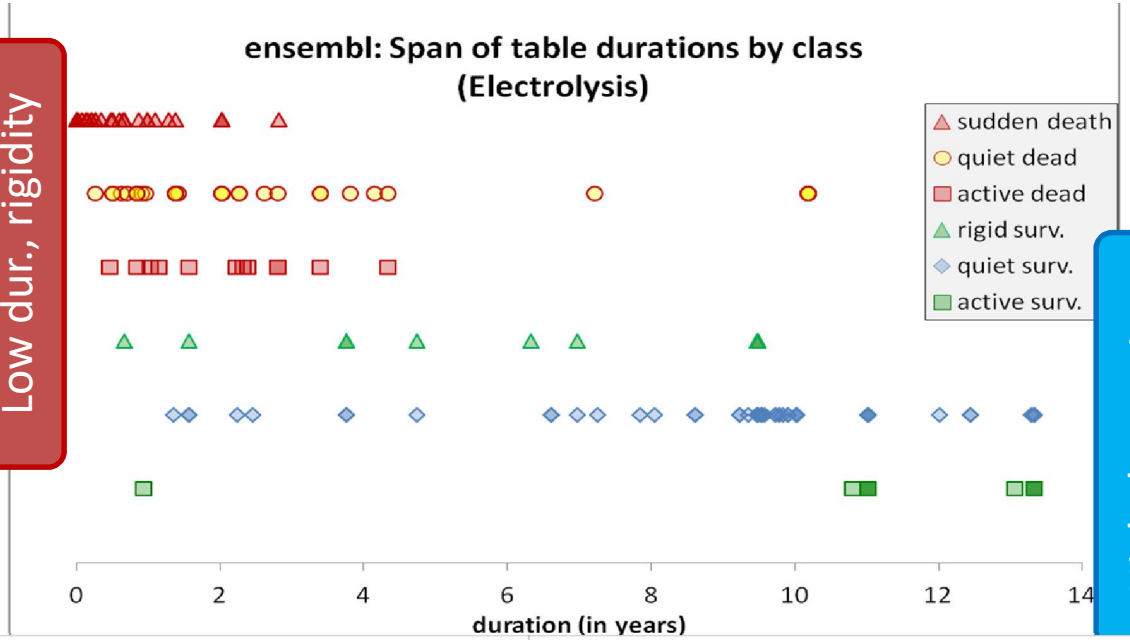
DEAD

Rigid

Quiet

Active

Low dur., rigidity



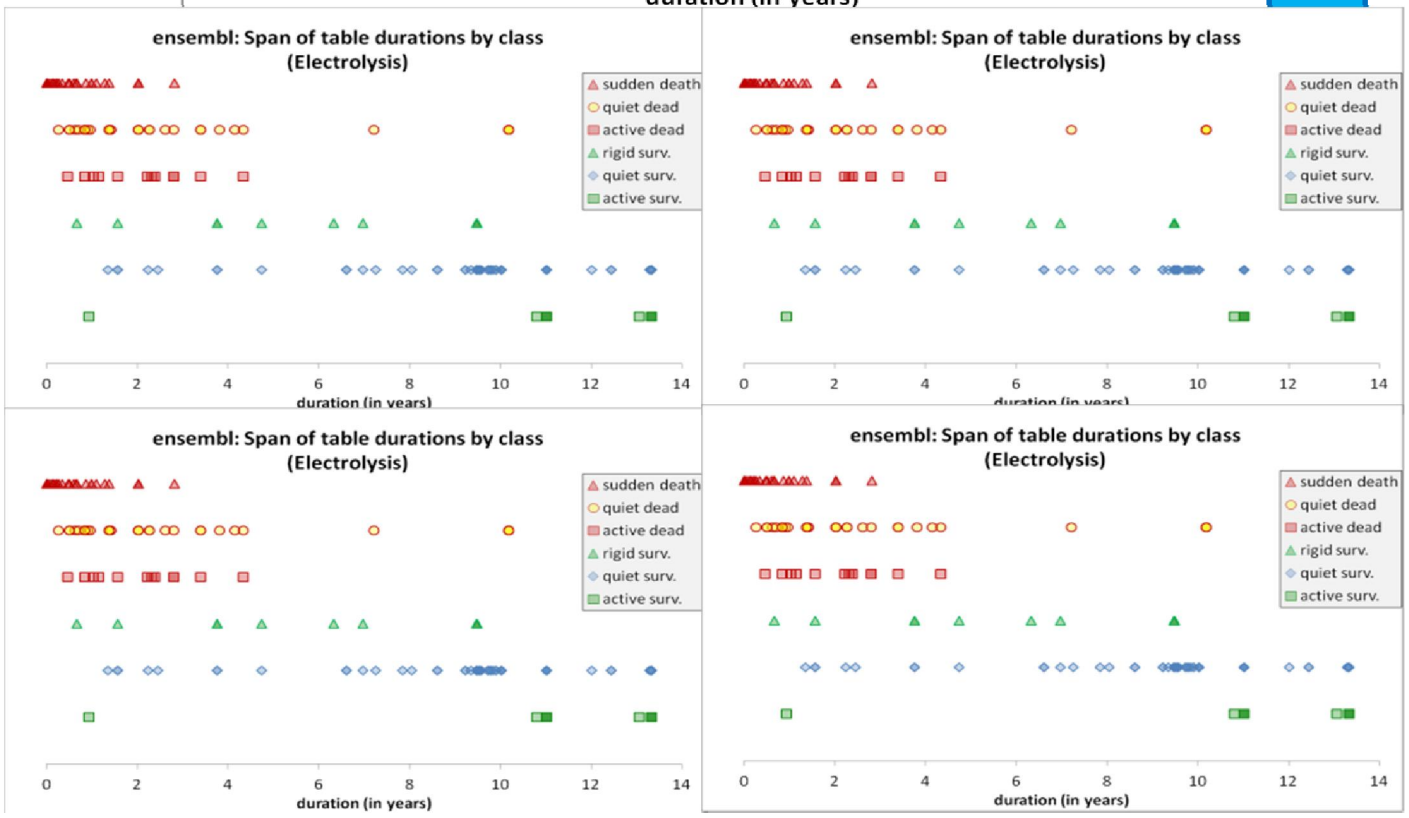
High dur., quiet

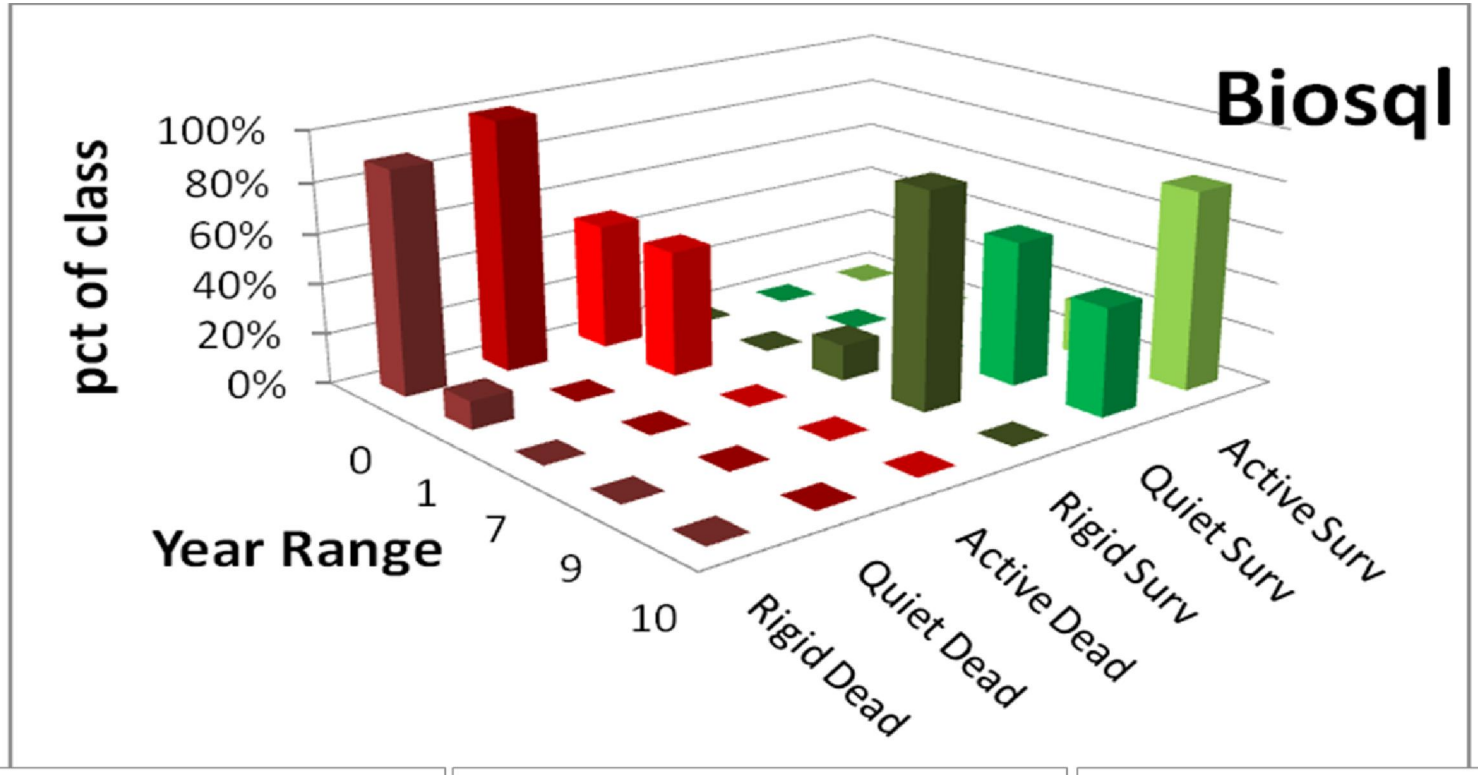
SURVIVORS

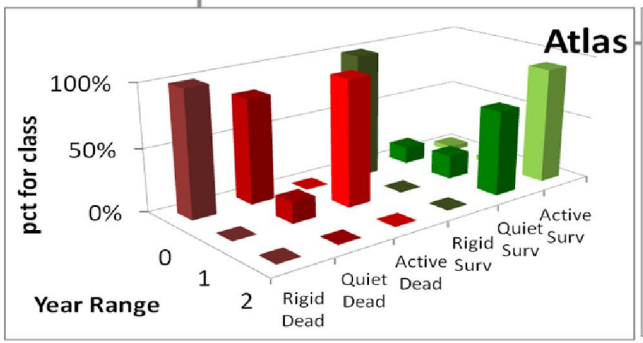
Rigid

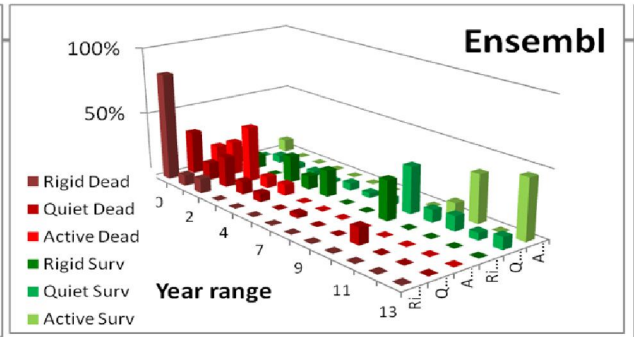
Quiet

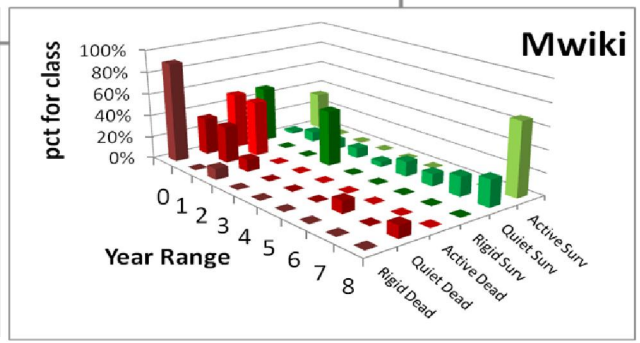
Active

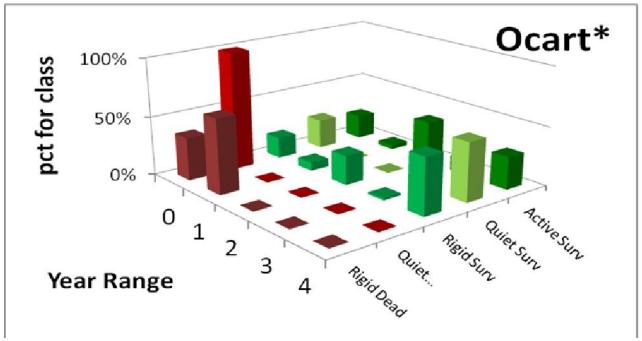


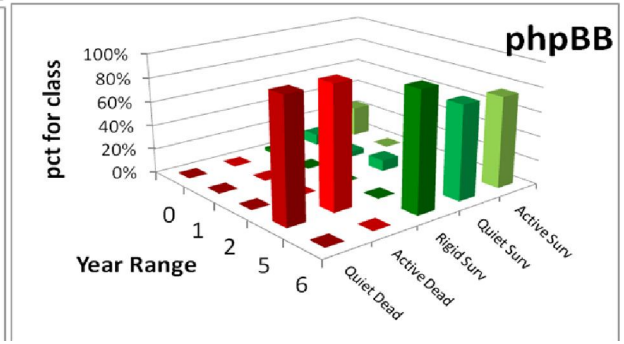

Biosql
pct of class
100%
80%
60%
40%
20%
0%
Year Range
0
1
7
9
10
Rigid Dead
Quiet Dead
Active Dead
Rigid Surv
Quiet Surv
Active Surv

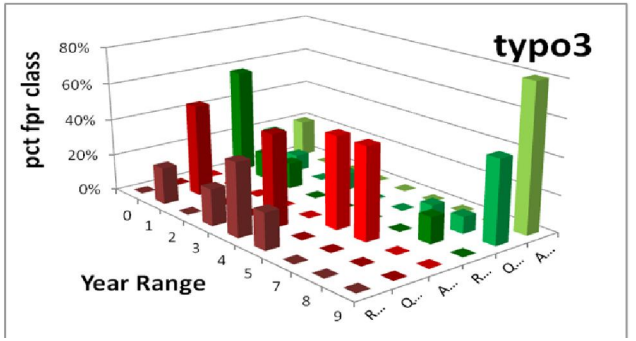

Atlas
pct for class
100%
50%
0%
Year Range
0
1
2
Rigid Dead
Quiet Dead
Active Dead
Rigid Surv
Quiet Surv
Active Surv


Ensembl
100%
50%
Rigid Dead
Quiet Dead
Active Dead
Rigid Surv
Quiet Surv
Active Surv
Year range


Mwiki
pct for class
100%
80%
60%
40%
20%
0%
Year Range
0
1
2
3
4
5
6
7
8
Rigid Dead
Quiet Dead
Active Dead
Rigid Surv
Quiet Surv
Active Surv


Ocart*
pct for class
100%
50%
0%
Year Range
0
1
2
3
4
Rigid Dead
Quiet...
Rigid Surv
Quiet Surv
Active Surv

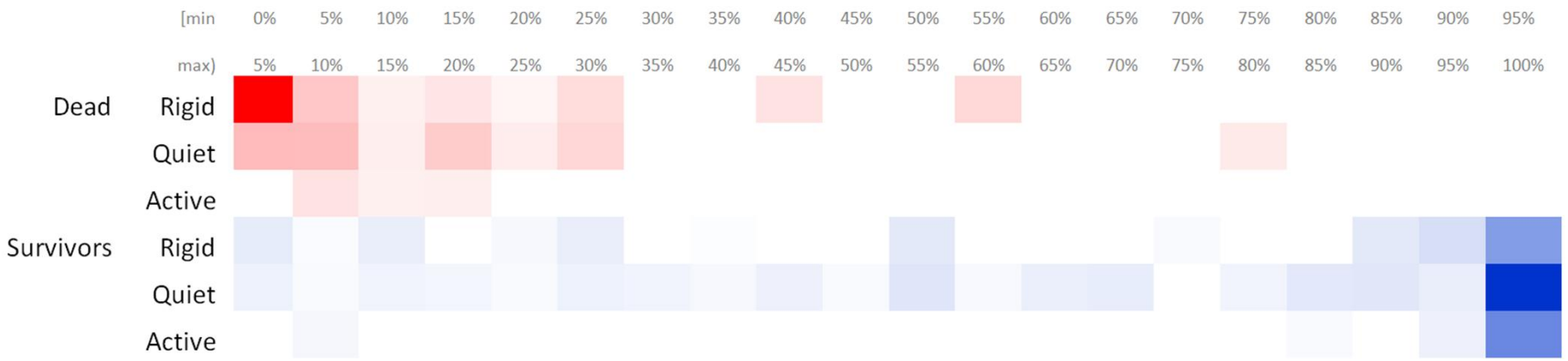

phpBB
pct for class
100%
80%
60%
40%
20%
0%
Year Range
0
1
2
5
6
Quiet Dead
Active Dead
Rigid Surv
Quiet Surv
Active Surv


typo3
pct fpr class
80%
60%
40%
20%
0%
Year Range
0
1
2
3
4
5
7
8
9
R...
Q...
A...
R...
Q...
A...

The data to support the pattern

- We have performed an in-depth study of how tables are distributed in different durations. To group durations, we have split the duration of each schema lifetime into periods of 5%. Then, for each *LifeAndDeath* value, and for each duration range of 5% of the database lifetime, we computed the percentage of tables whose duration falls within this range.

… electrolysis as a heatmap …



- For each *LifeAndDeath* value, and for each duration range of 5% of the database lifetime, we computed the percentage of tables (over the total of the data set) whose duration falls within this range.
- We removed cells that corresponded to only one data set

The resulting heatmap shows the polarization in colors: brighter color signifies higher percentage of the population

For each data set, for each LifeAndDeath class, **percentage** of tables per duration range **over the total of the data set** (for each data set, the sum of all cells adds up to 100%)

Atlas	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0% -20%)	5%	0%	0%	1%	1%	0%	7%
[20%-80%)	3%	7%	2%	11%	13%	3%	40%
[80%-100%]	0%	0%	0%	0%	28%	25%	53%
	8%	7%	2%	13%	42%	28%	100%

Biosql	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	20%	13%	4%	0%	0%	0%	38%
[20%-80)	0%	0%	0%	2%	0%	0%	2%
[80%-100%]	0%	0%	0%	13%	16%	31%	60%
	20%	13%	4%	16%	16%	31%	100%

Copperm.	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	0%	0%	0%	0%	0%	0%	0%
[20%-80)	4%	0%	0%	0%	13%	0%	17%
[80%-100%]	0%	0%	0%	30%	43%	9%	83%
	4%	0%	0%	30%	57%	9%	100%

Ensembl	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	23%	13%	5%	1%	3%	1%	37%
[20%-80)	1%	7%	3%	5%	23%	0%	54%
[80%-100%]	0%	0%	0%	0%	9%	6%	8%
	24%	20%	8%	6%	35%	7%	100%

Mwiki	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	13%	7%	3%	1%	6%	1%	23%
[20%-80)	1%	4%	0%	1%	31%	0%	58%
[80%-100%]	0%	1%	0%	0%	27%	3%	20%
	14%	13%	3%	3%	63%	4%	100%

Ocart*	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	3%	2%	0%	8%	5%	1%	23%
[20%-80)	5%	0%	0%	17%	16%	0%	43%
[80%-100%]	0%	0%	0%	17%	22%	2%	34%
	9%	2%	0%	42%	44%	3%	100%

phpBB	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	0%	0%	0%	1%	3%	3%	7%
[20%-80)	0%	1%	0%	0%	4%	0%	11%
[80%-100%]	0%	1%	4%	49%	24%	9%	81%
	0%	3%	4%	50%	31%	11%	100%

typo3	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	3%	3%	0%	16%	9%	3%	34%
[20%-80)	13%	3%	3%	3%	6%	0%	31%
[80%-100%]	0%	0%	3%	3%	19%	13%	34%
	16%	6%	6%	22%	34%	16%	100%

Indicative, average values over all datasets: for each LifeAndDeath class, **percentage** of tables per duration range **over the total of the data set**

	Rigid Dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv	
[0-20%)	8%	5%	2%	4%	3%	1%	*23%*
[20%-80%)	3%	3%	1%	5%	13%	0%	*26%*
[80%-100%]	0%	0%	1%	14%	24%	12%	*51%*
	12%	*8%*	*3%*	*23%*	*40%*	*14%*	*100%*

An acute reader might express the concern whether it would be better to gather all the tables in one single set and average over them. We disagree: each data set comes with its own requirements, development style, and idiosyncrasy and putting all tables in a single data set, not only scandalously favors large data sets, but integrates different things. We average the behavior of schemata, not tables here.

Do certain LifeAndDeath classes have high concentrations in particular data ranges?

The following tables are important. Many findings for **survivor** tables refer to it.

For each data set, for each LifeAndDeath class, **percentage** of tables per duration range **over the total of their Life&Death class** (for each data set, for each column, percentages add up to 100%)

Atlas	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv
[0% -20%)	57%	0%	0%	9%	3%	0%
[20%-80%)	43%	100%	100%	91%	30%	12%
[80%-100%]	0%	0%	0%	0%	68%	88%
	100%	*100%*	*100%*	*100%*	*100%*	*100%*

Copperm.	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv
[0-20%)	0%			0%	0%	0%
[20%-80%)	100%			0%	23%	0%
[80%-100%]	0%			100%	77%	100%
	100%			*100%*	*100%*	*100%*

Mwiki	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv
[0-20%)	90%	56%	100%	50%	9%	33%
[20%-80%)	10%	33%	0%	50%	49%	0%
[80%-100%]	0%	11%	0%	0%	42%	67%
	100%	*100%*	*100%*	*100%*	*100%*	*100%*

phpBB	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv
[0-20%)		0%	0%	3%	9%	25%
[20%-80%)		50%	0%	0%	14%	0%
[80%-100%]		50%	100%	97%	77%	75%
		100%	*100%*	*100%*	*100%*	*100%*

Biosql	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv
[0-20%)	100%	100%	100%	0%	0%	0%
[20%-80%)	0%	0%	0%	14%	0%	0%
[80%-100%]	0%	0%	0%	86%	100%	100%
	100%	*100%*	*100%*	*100%*	*100%*	*100%*

Ensembl	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv
[0-20%)	97%	65%	67%	20%	9%	9%
[20%-80%)	3%	35%	33%	80%	65%	0%
[80%-100%]	0%	0%	0%	0%	26%	91%
	100%	*100%*	*100%*	*100%*	*100%*	*100%*

Ocart*	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv
[0-20%)	36%	100%		19%	13%	25%
[20%-80%)	64%	0%		41%	38%	0%
[80%-100%]	0%	0%		41%	50%	75%
	100%	*100%*		*100%*	*100%*	*100%*

typo3	Rigid dead	Quiet Dead	Active Dead	Rigid Surv	Quiet Surv	Active Surv
[0-20%)	20%	50%	0%	71%	27%	20%
[20%-80%)	80%	50%	50%	14%	18%	0%
[80%-100%]	0%	0%	50%	14%	55%	80%
	100%	*100%*	*100%*	*100%*	*100%*	*100%*

Average values over all datasets: for each LifeAndDeath class, **percentage** of tables per duration range **over the total of their LifeAndDeath class** (for each data set, for each column, percentages add up to 100%)

	Rigid dead	Quiet Dead	Active Deac	Rigid Surv	Quiet Surv	Active Surv
[0-20%)	57%	53%	44%	21%	9%	14%
[20%-80)	43%	38%	31%	36%	29%	2%
[80%-100%]	0%	9%	25%	42%	62%	84%
	100%	*100%*	*100%*	*100%*	*100%*	*100%*

An acute reader might express the concern whether it would be better to gather all the tables in one single set and average over them. We disagree: each data set comes with its own requirements, development style, and idiosyncrasy and putting all tables in a single data set, not only scandalously favors large data sets, but integrates different things. We average the behavior of schemata, not tables here.

What is the distribution of tables per activity class for the dead tables that have durations less than the 20% of the database's life?

What is the distribution of tables per activity class for the survivor tables that have durations longer than the 80% of the database's life?

The following table is important. Many findings for **dead** tables refer to it.

Zoom into low 20% of durations for the dead and upper 20% for the survivors

	Pct of durations shorter than 20% of db life for Dead tables over the …				*Pct of durations longer than 80% of db life for Survivor tables over the …*			
	… Dead	… Rigid	… Quiet	…Active	… Surv	… Rigid	… Quiet	…Active
atlas	27%	57%	0%	0%	64%	0%	68%	88%
biosql	100%	100%	100%	100%	96%	86%	100%	100%
coppermine	0%	0%	-	-	86%	100%	77%	100%
ensembl	80%	97%	65%	67%	32%	0%	26%	91%
mediawiki	76%	90%	56%	100%	42%	0%	42%	67%
opencart*	50%	36%	100%	-	46%	41%	50%	75%
phpBB	0%	-	0%	0%	88%	97%	77%	75%
typo3	22%	20%	50%	0%	48%	14%	55%	80%

We count the number of tables, per LifeAndDeath class, for the respective critical duration range, and we compute the fraction of this value over the total number of tables pertaining to this LifeAndDeath class (columns Rigid, Quiet, Active). For the Dead and Surv columns, we divide the total number of dead/survivor tables belonging to the respective critical duration over the total number of dead/survivor tables overall.

In more than half of the cells of the table, the percentage reaches or exceeds 50%

	Pct of durations shorter than 20% of db life for Dead tables over the …				*Pct of durations longer than 80% of db life for Survivor tables over the …*			
	… Dead	… Rigid	… Quiet	…Active	… Surv	… Rigid	… Quiet	…Active
atlas	27%	57%	0%	0%	64%	0%	68%	88%
biosql	100%	100%	100%	100%	96%	86%	100%	100%
coppermine	0%	0%	-	-	86%	100%	77%	100%
ensembl	80%	97%	65%	67%	32%	0%	26%	91%
mediawiki	76%	90%	56%	100%	42%	0%	42%	67%
opencart*	50%	36%	100%	-	46%	41%	50%	75%
phpBB	0%	-	0%	0%	88%	97%	77%	75%
typo3	22%	20%	50%	0%	48%	14%	55%	80%

We count the number of tables, per LifeAndDeath class, for the respective critical duration range, and we compute the fraction of this value over the total number of tables pertaining to this LifeAndDeath class (columns Rigid, Quiet, Active). For the Dead and Surv columns, we divide the total number of dead/survivor tables belonging to the respective critical duration over the total number of dead/survivor tables overall.

Dead Tables

- ***All kinds of dead tables are strongly inclined (a) to rigidity, and (b) to small durations.***
  - **The less active tables are the more they are attracted to short durations**.
  - The attraction of dead tables, especially rigid ones, to (primarily) low or, (secondarily) medium durations is significant and only few tables in the class of dead tables escape this rule.
    - Interestingly, in all our datasets, the only dead tables that escape the barrier of low and medium durations are a single table in mediawiki, another one in typo3, and the 4 of the 5 tables that are simultaneously deleted in phpBB.

Dead tables

1. ***Rigid dead tables, which is the most populated category of dead tables, strongly cluster in the area of low durations (lower than the 20% of the database lifetime) …***
   - …with percentages of 90% – 100% in 3 of the 6 data sets
2. ***Quiet dead tables, which is a category including few tables, are mostly oriented towards low durations***.
   - Specifically, there are 5 data sets with a high concentration of tables in the area of low durations; for the rest, the majority of quiet dead tables lie elsewhere
3. The **very few active dead**, have mixed behaviors.

Survivor tables

- It is extremely surprising that, the vast majority of **active survivors exceed 80% of the database lifetime** in all datasets.
  - With the exception of three data sets in the range of 67%-75%, ***the percentage of active survivors that exceeds 80% of the db lifetime exceeds 80%, and even attains totality in 2 cases.***
  - ***Active survivor tables are not too many; however, it is amazing how long they live***. If one looks into the detailed data and in synch with the empty triangle pattern of [IS16], ***the top changers are very often of maximum duration, i.e., early born and survivors***.
    - This should be read as: no top-changer tables are born later!

Survivor tables

- ***Rigid survivors demonstrate a large variety of behaviors***.
- ***Quiet survivors, being the (sometimes vast) majority of survivor tables,*** are ***mostly gravitated towards large durations, and secondarily to medium ones.***
  - In 6 out of 8 data sets, the percentage of quiet survivors that exceeds 80% of db lifetime surpasses 50%.
  - In the two exceptions, medium durations is the largest subgroup of quiet survivors.
  - Still, quiet survivors also demonstrate short durations too, so overall, their span of possible durations is large.
  - Notably, in all data sets, there are quiet survivors reaching maximum duration.

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/projects/schemaBiographies

Threats To Validity

- With respect to the *measurement validity* of our work, we have tested (i) our automatic extraction tool, Hecate, for the accuracy of its automatic extraction of delta's and measures, and (ii) our human-made calculations.
- The *external validity* of our study is supported by several strong statements: we have chosen data sets with
  - fairly long histories of versions,
  - a variety of domains (CMS's and scientific systems),
  - a variety in the number of their commits (from 46 to 528), and,
  - a variety of schema sizes (from 23 to 114 at the end of the study);
- We have also been steadily attentive to work only with phenomena that are common to all the data sets.
- **Do not to interpret our findings as laws** (that would need confirmation of our results by other research groups), **but rather as patterns.**