How Good Is Your Record Linkage System, Really?

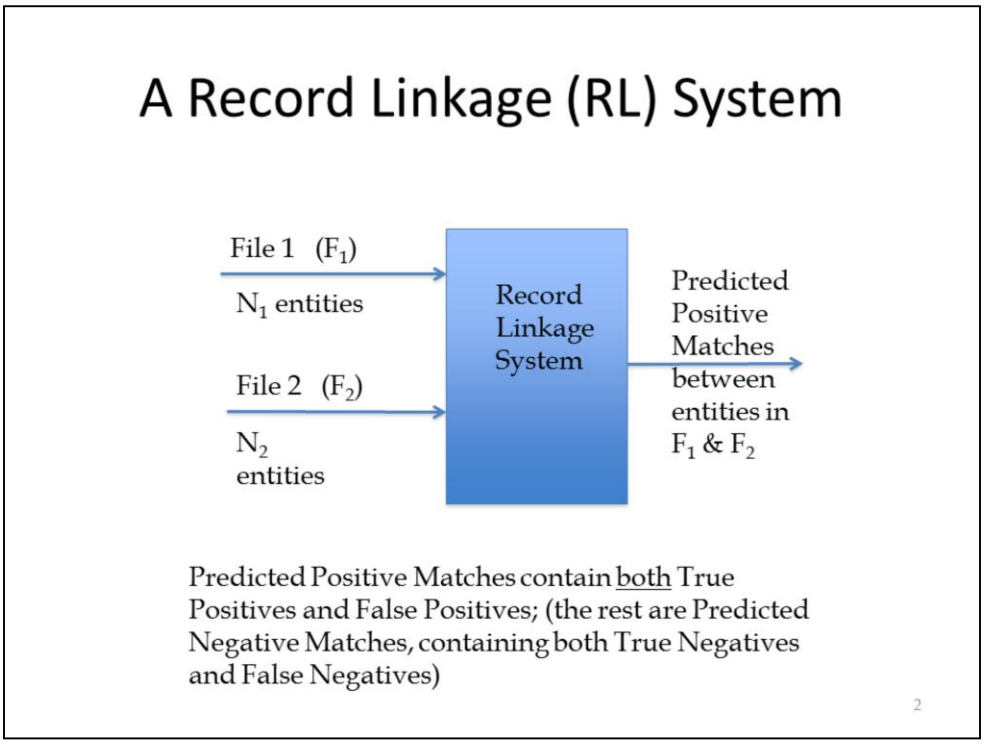
A presentation to FedCASIC
Administrative and Linked Records
Thursday, 20 Mar 2014

K. Bradley Paxton, Ph. D.
ADI, LLC
brad.paxton@adillc.net

1

The use of Administrative Records is increasing in many agencies, and is being studied at Census for use in 2020.

It can save a lot of money, but it's real value ("goodness") depends on matching quality, and further, how it is tuned for optimal use – both requiring testing.



Think of F_1 as Census data, and F_2 as tax data, with duplicates removed…

Typically, N_1 and N_2 are comparable, but they don't have to be the same.

Think of "entities" as persons or households, typically.

If a Census record is correctly matched with a tax record, say, then improved Census data can result; however, if the linkage is incorrect, it could be made worse.

Confusion Matrix

	System Positive Predictions	System Negative Predictions	Row Sums
Data Positive Match Truth	True Positives	False Negatives (Type II Error)	True Positive Matches
Data Negative Match Truth	False Positives (Type I Error)	True Negatives	True Negative Matches
Column Sums	Predicted Positive Matches	Predicted Negative Matches	

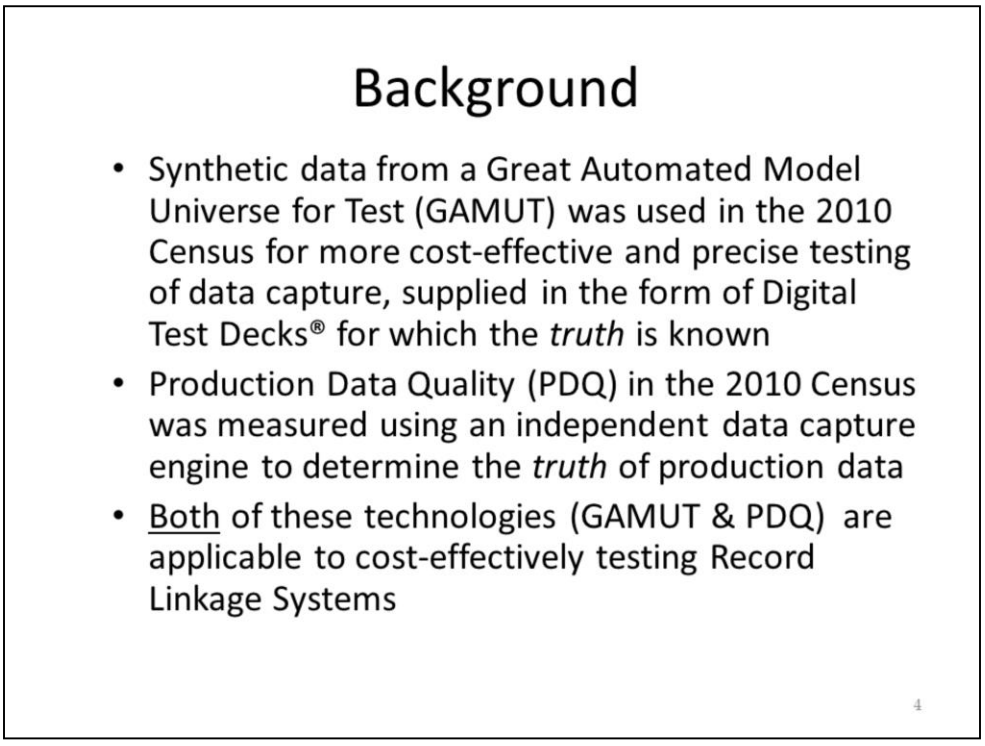
3

If you run a test and can't estimate <u>all</u> the numbers in the black box, you haven't run a good enough test and can't optimize your RL system!

It is not sufficient to just say one has more "matches", as that does not tell you how many of your Predicted Positive Matches are False Positives.

Digging deeper, you need to know how many matches "escaped" your system, (the False Negatives).

If you are able to test your RL system in a given state and determine all the elements in the matrix, than you can optimize RL system performance and maximize your return on investment.

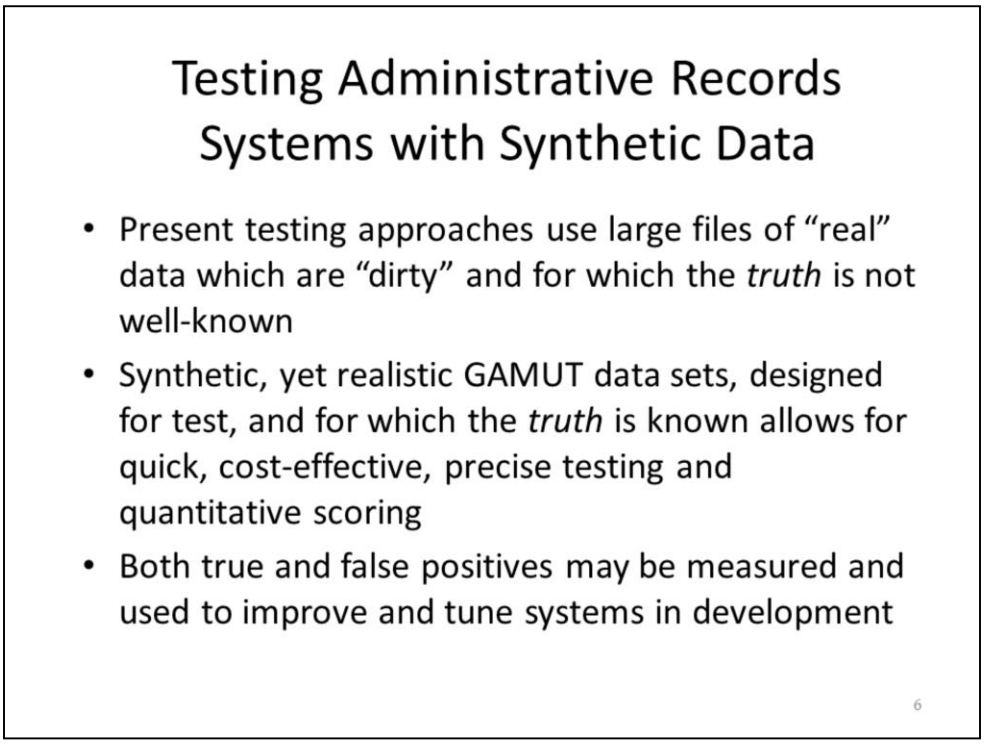


Both of these approaches are very cost-effective because they replace a lot of human effort with automation.

Basic Idea for Today

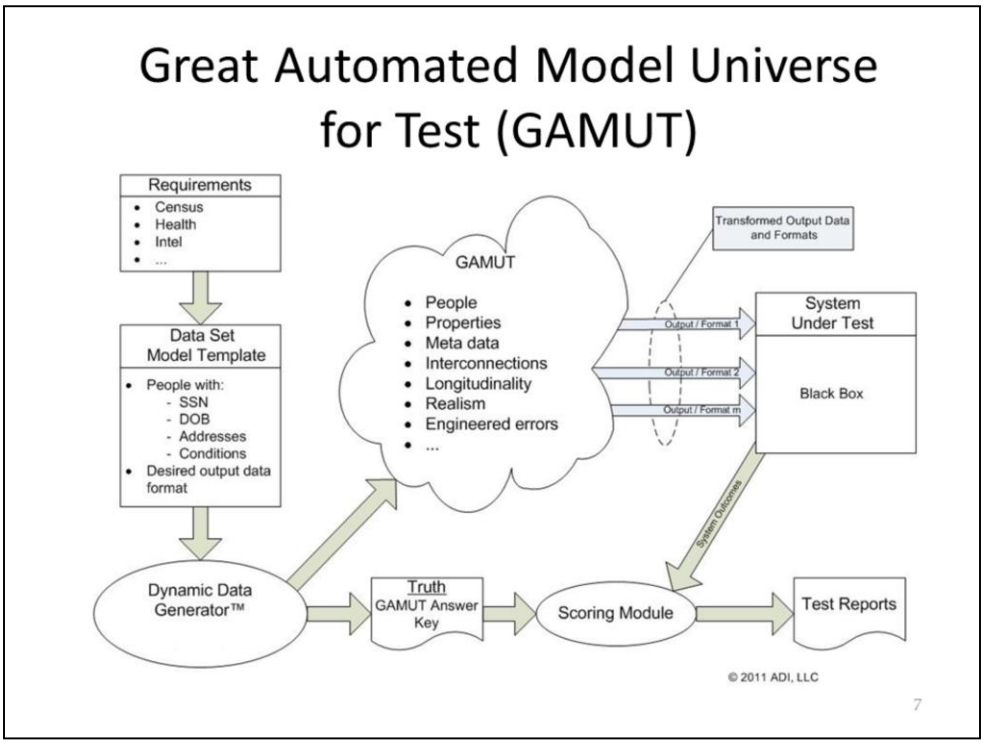
- The GAMUT technology is best for system development testing, when good data, designed for test, is generally private or unavailable or both (Ref.1)
- The PDQ technology is best for production testing with real data to determine how good your matching really is and learn how to improve it (Ref.2)
- They can overlap to provide enriched testing and optimization benefits as you phase out of development and get into production

5



Actually, synthetic data can be **better** for development testing than real data!!!

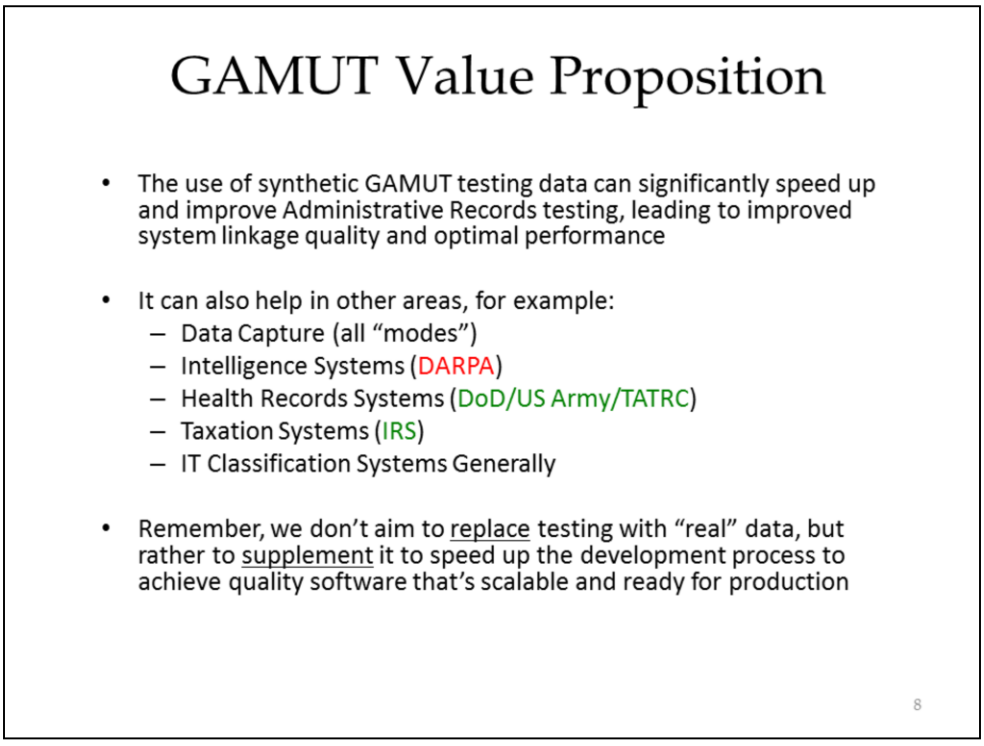
This is because you know the truth and can introduce engineered errors or variations that need to be tested.



I prefer to think about the SUT first, and then consider what kind of data is needed for a particular test plan.

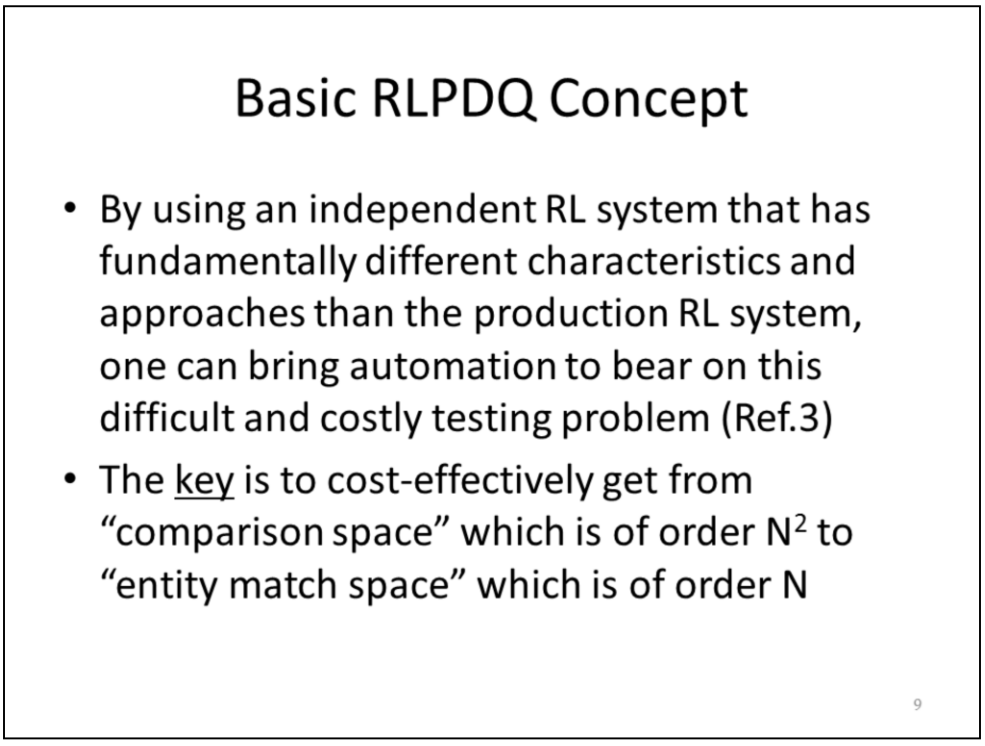
The job of the SUT is to ingest various data streams from the GAMUT and infer some facts about the GAMUT that are not apparent, like does a person in one data stream match a person in another data stream.

Usually, looking at these data streams gives you only a little “peek” at what’s really in the GAMUT.



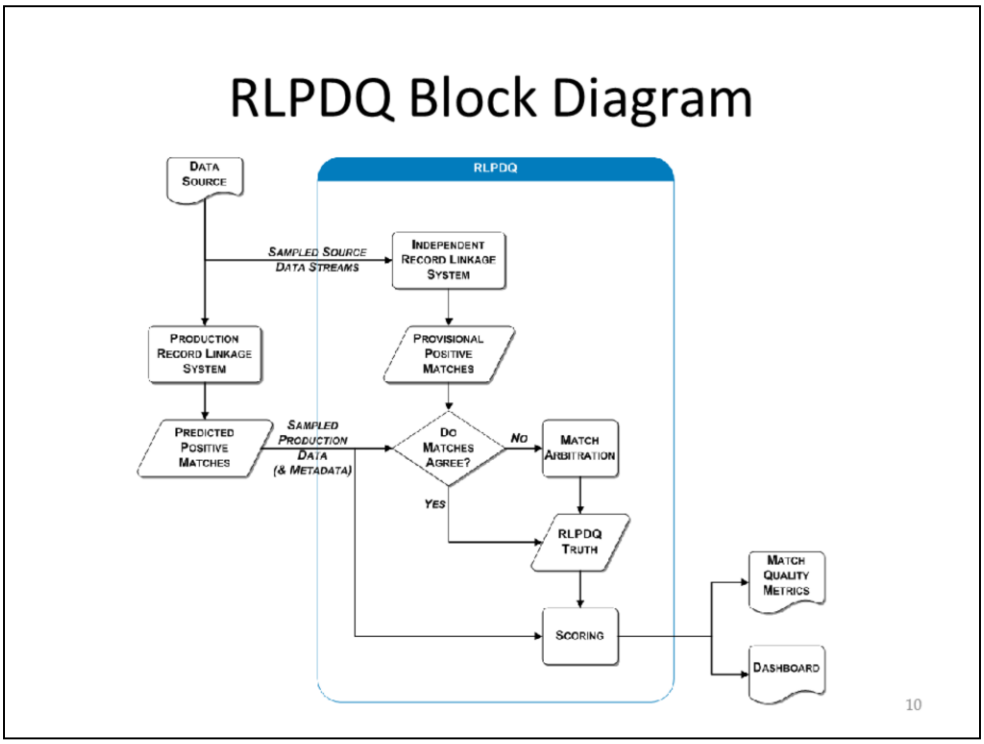
Four modes of data capture are: paper (self-administered questionnaire), internet, telephone (CATI), and enumerator (CAPI).

Actually, because synthetic data is DESIGNED FOR TEST, it is actually better for testing than real data, especially in the development stages.



The independent RL system can, say, use different technology to estimate matches, weight data fields differently, and perhaps use different blocking techniques.

Using automation greatly reduces the testing workload and cost.

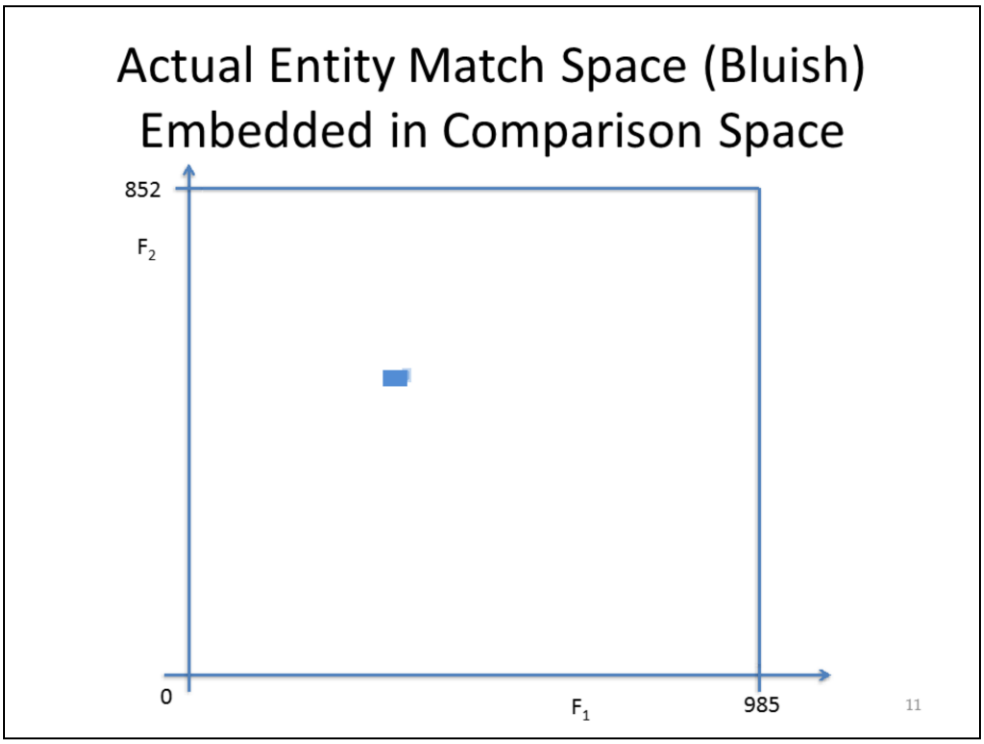


The data source supplies the two files – census and tax data, say.

Matches both systems agree upon are highly likely to be correct, and this is the bulk of the answers you seek.

Even if the two engines don't agree, most of the time ONE got it right!

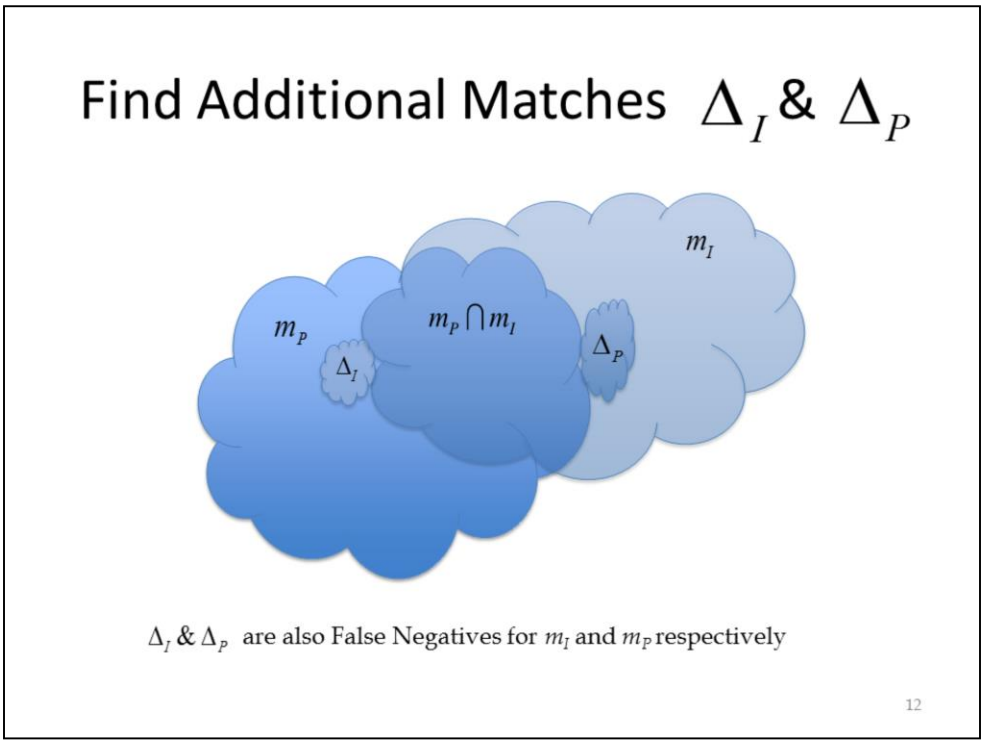
Arbitration on what's left involves humans looking at entity pairs, using automation to reduce effort.



So, say both files are roughly only a thousand or so records (N); even then, the number of record pairs that must be examined is about a million (N^2).

The primary PDQ job is to quickly and efficiently get you focused on "Entity Match Space" (order N), rather than "Comparison Space" (order N^2).

This is an actual example result "to scale" - It's a little hard to see the small blue blob, so…



Here is a close-up of our little blue blob…

This is "entity matching space" detailed in Reference 3.

NOT to scale – usually the overlap region is most of it.

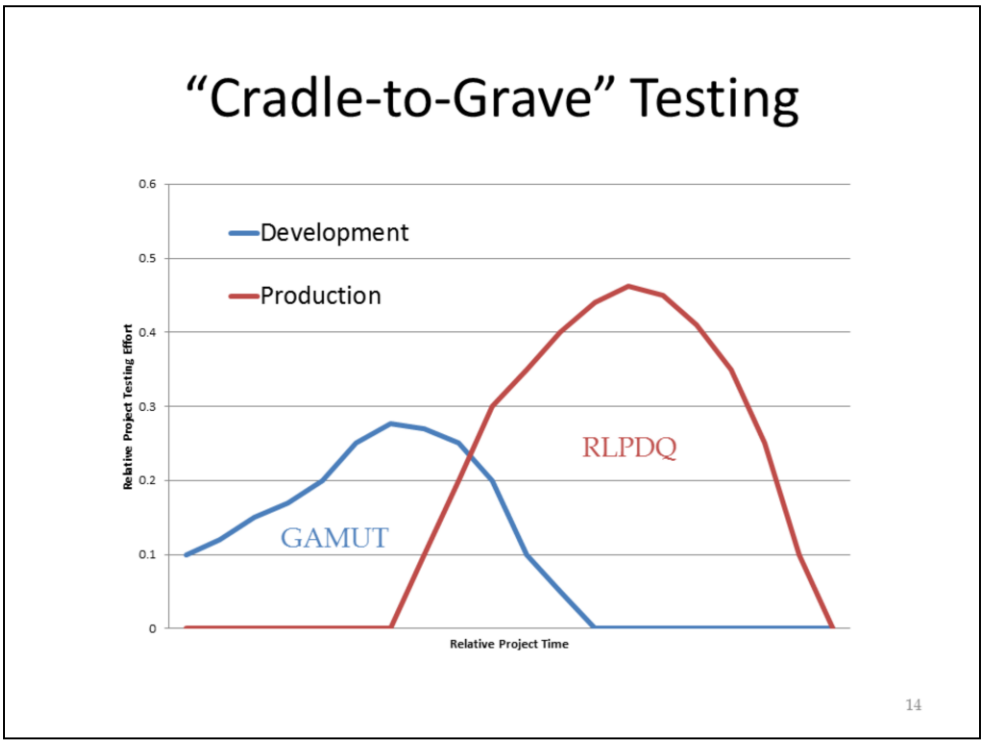
Getting a handle on False Negatives can help tune your system for maximum value.

RLPDQ Value Proposition

- An RLPDQ system can be a cost-effective tool to quantitatively test RL systems operating on real data streams
- It uses automation to rapidly get from the burdensome comparison space to the more readily analyzed entity match space
- Using this approach speeds up RL systems testing, facilitates tuning, and can improve production output quality.

13

So, now suppose you use both of these test technologies -



Ideally, you can use both techniques as the project progresses from development through production.

In the overlap region, as production ramps up, you can learn more about your RL System by comparing both sets of test outputs, as they tend to discover different types of errors.

In particular, RLPDQ is effective at uncovering "escapes" (False Negatives) in your real production data.

Conclusions

- GAMUT synthetic data is useful for development testing
- An RLPDQ system is useful for production testing
- Use of both techniques during production ramp-up leads to increased learning opportunities for better quality and greater efficiency

15

Since the two methods get at False Negatives in different ways, that increases the chances that you uncover these “hard to find” errors.

Thank you!

References

1. Paxton, K. Bradley, and Hager, Thomas, *Use of Synthetic Data in Testing Administrative Records Systems*, Proceedings, Federal Committee on Statistical Methodology (FCSM), Washington, DC, 2012

1. Paxton, K. Bradley, Spiwak, Steven P., Huang, Douglass, and McGarity, James K., *Testing Production Data Capture Quality*, Proceedings, Federal Committee on Statistical Methodology (FCSM), Washington, DC, 2012

1. Paxton, K. Bradley, *Testing Record Linkage Production Data Quality*, . In JSM Proceedings, Government Statistics Section. Montreal, Canada: American Statistical Association. Pgs. 1157-1171, 2013.

Contact:
K. Bradley Paxton, ADI, LLC
brad.paxton@adillc.net

13 Mar 2014

16

Shoot me an e-mail or give me your business card, and I'll send you these references.