ADI, LLC, USA

Use of Synthetic Data in Testing Administrative Records Systems

A presentation to the FCSM

Tuesday, 10 Jan 2012

ADI ADVANCED DOCUMENT IMAGING

gunnison Consulting Group Inc.

Some Background on ADI, LLC

- Synthetic data from ADI was used in the 2010 Census for more cost-effective and precise testing of data capture
- This data was supplied in Digital Test Decks®, corresponding image files, and scripts for testing data capture modes other than paper
- Independently designed and developed a generic and powerful "Dynamic Data Generator™" (DDG) for creating synthetic test data
- Also doing medical (IBM) and intelligence (DARPA) synthetic data sets

Security Aspects

- Program security around real data precludes engaging industry for scientific study, market research, and for consistent evaluation of multiple vendors
  - In Medical records, there are *HIPAA* laws
  - In Census records, there is *Title 13*
  - In IRS records, there is *Title 26*
  - In SSA records, there is the *Privacy Act of 1974 (5 U.S.C. § 552a)*
  - …
- Our synthetic data is realistic, but not real!

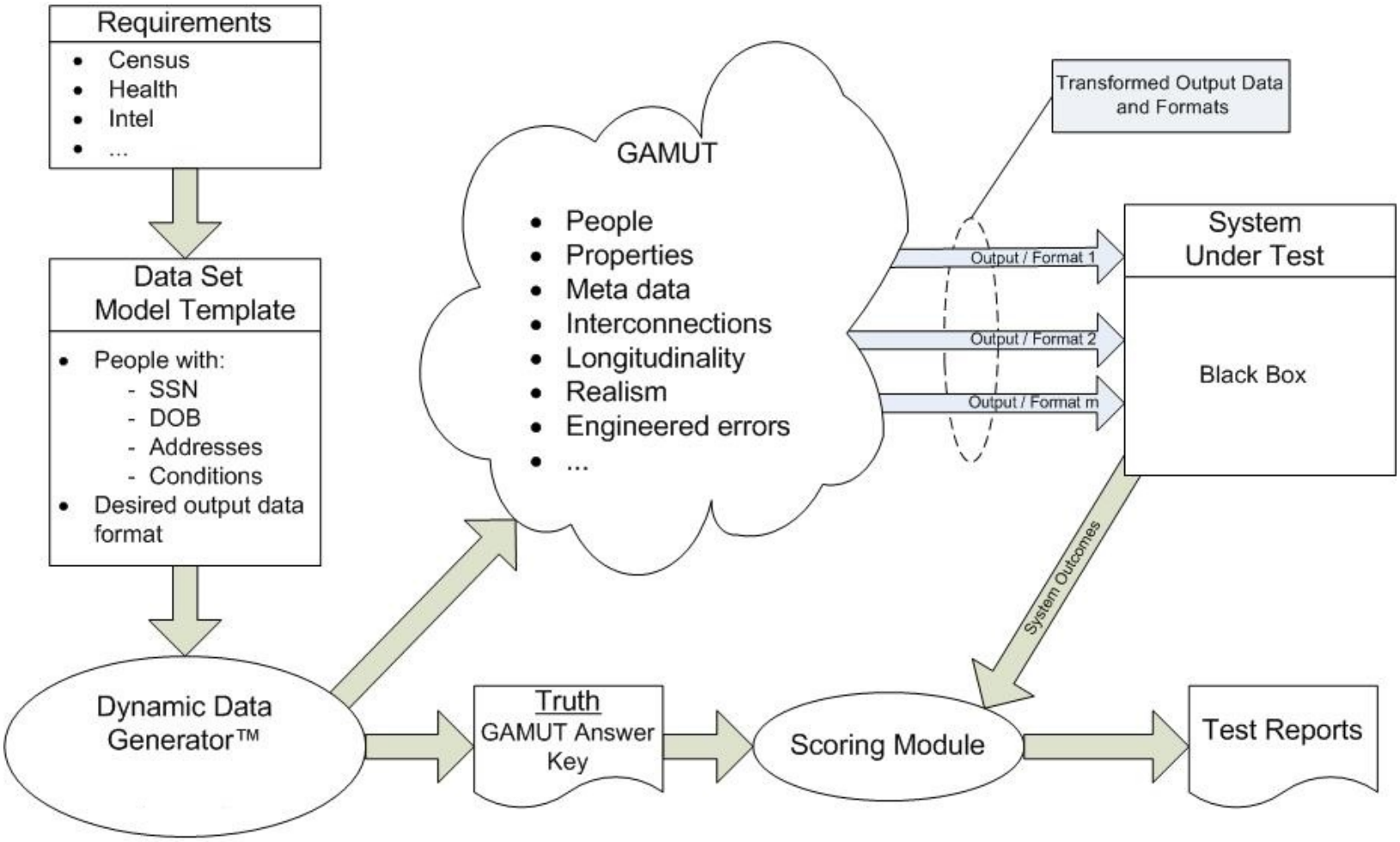
Testing Administrative Records Systems with Synthetic Data

- Administrative Records will be very useful to Census, but testing the systems that are being developed to use them is extremely difficult
- Present testing approaches use large files of "real" data for which the "truth" is not known
- Synthetic, yet realistic data sets, designed for test, and for which the truth is known allows for quick, cost-effective and precise testing and quantitative scoring
- Both true and false positives may be measured and used to improve systems in development

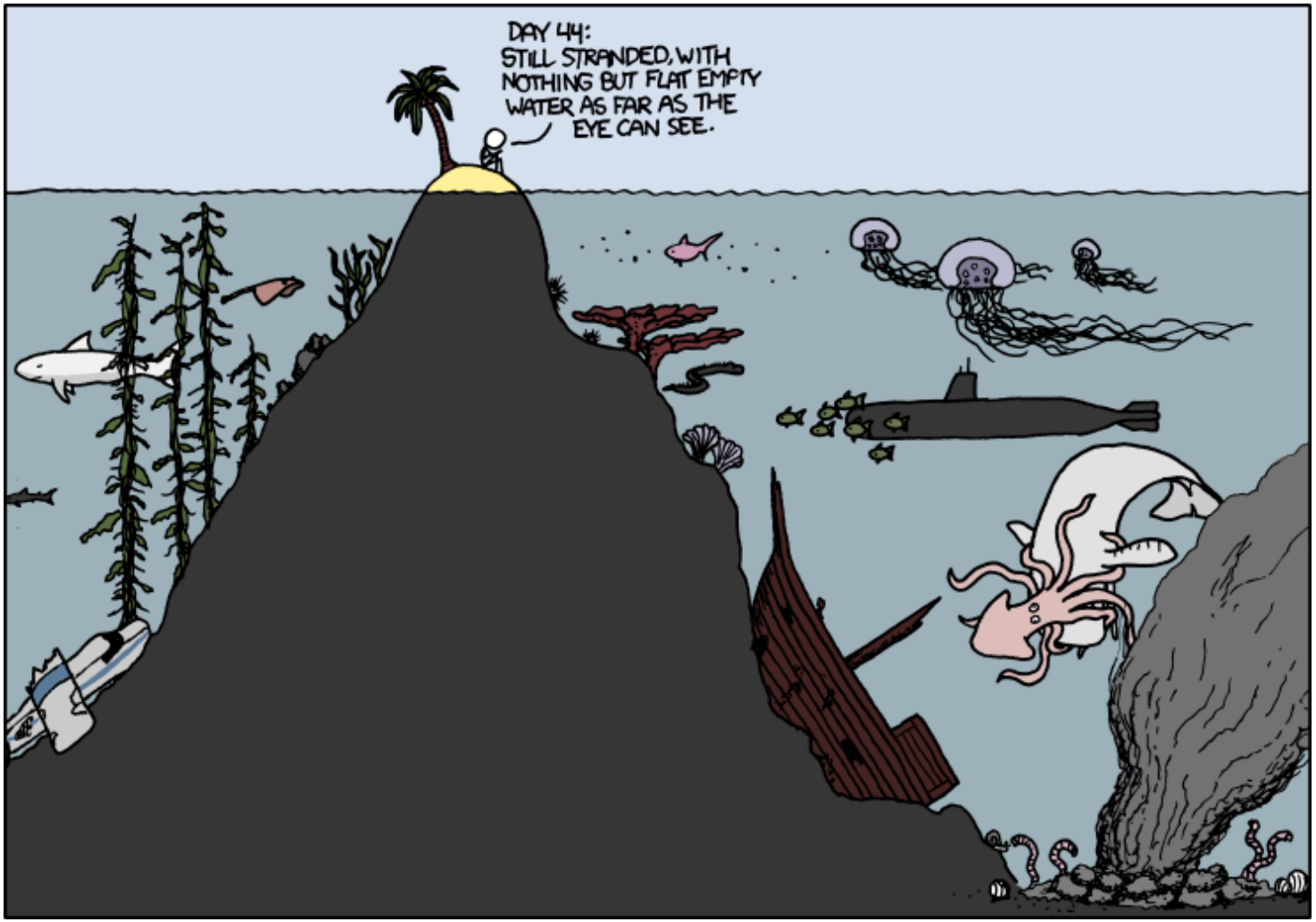
ADI ADVANCED DOCUMENT IMAGING

gunnison Consulting Group Inc.

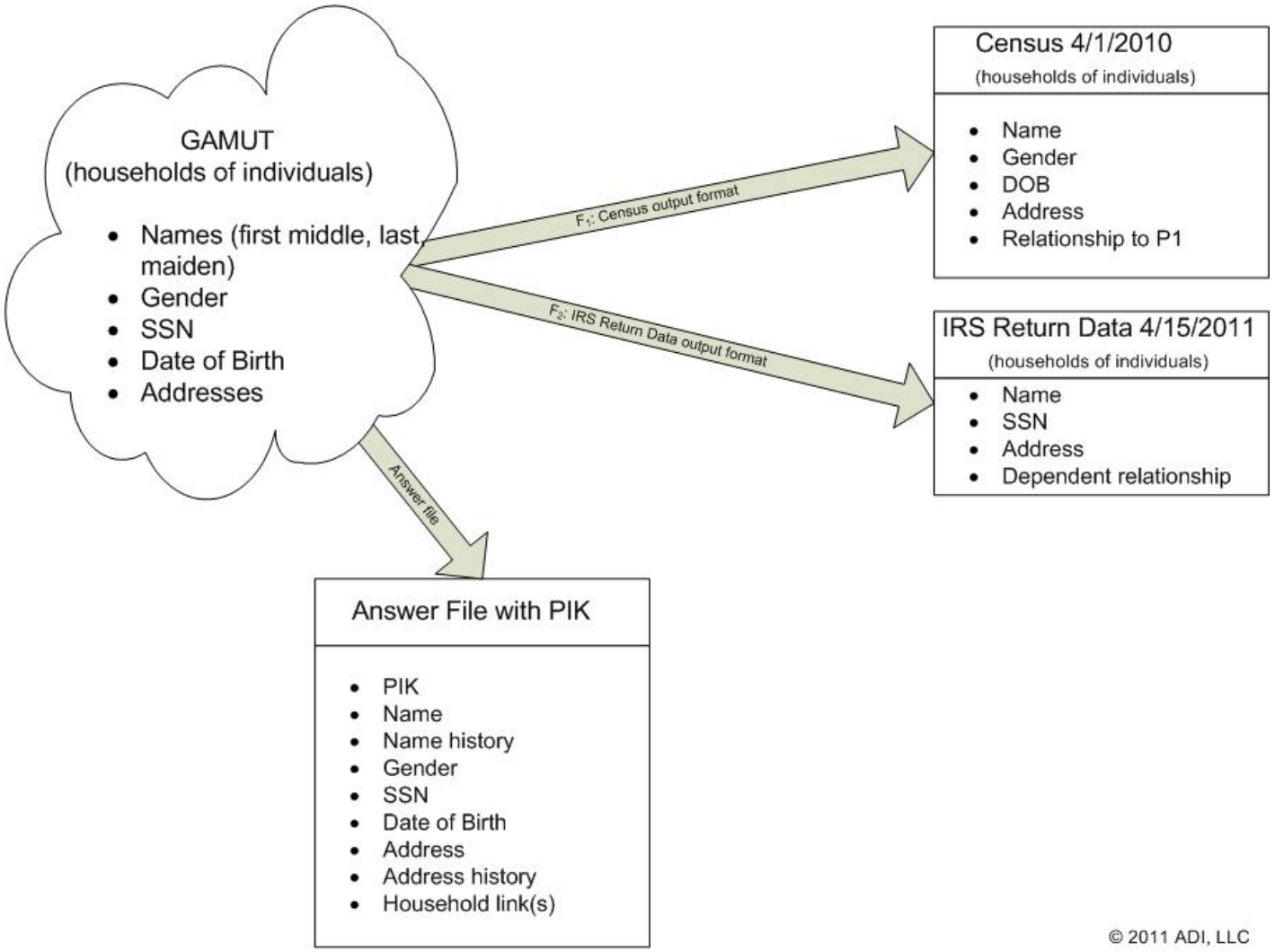
Great Automated Model Universe for Test (GAMUT)



A "Peek" at the GAMUT?



Today's GAMUT Example



Demo GAMUT Characteristics

- (Only) about 1000 synthetic households generated for this demo GAMUT
- Two data feeds were made: Census and Tax (IRS)
- Geographic scope:
  - DC, New Mexico, West Virginia

Data Feed Characteristics

- Census Data Feed:
  - Snapshot on 1 Apr 2010
  - Names, DOB, Gender, Relationships
  - Addresses
  - PIK Numbers
- IRS Return Data Feed:
  - Snapshot on 15 Apr 2011
  - SSNs
  - Names, Addresses
  - Dependent Relationships

Some GAMUT Demo "Features"

- Census
  - Dupes 2%
  - Person 1 DOB missing or morphed (1-2%)
  - Name morphing 2%
  - Coverage 99%
- Tax
  - Filer SSN can be both husband and wife
  - Filer name can be concatenation of both
  - Moves 10%
  - Coverage 85%

Test Example: Person Matching

- Using this data, we explain how testing can be done using GAMUT and how to analyze the results with a classic Receiver Operating Characteristic (ROC) technique
- For this example, we are just looking at testing a hypothetical RL system that does matching of Census feed Person 1 to Tax Filers in Tax feed

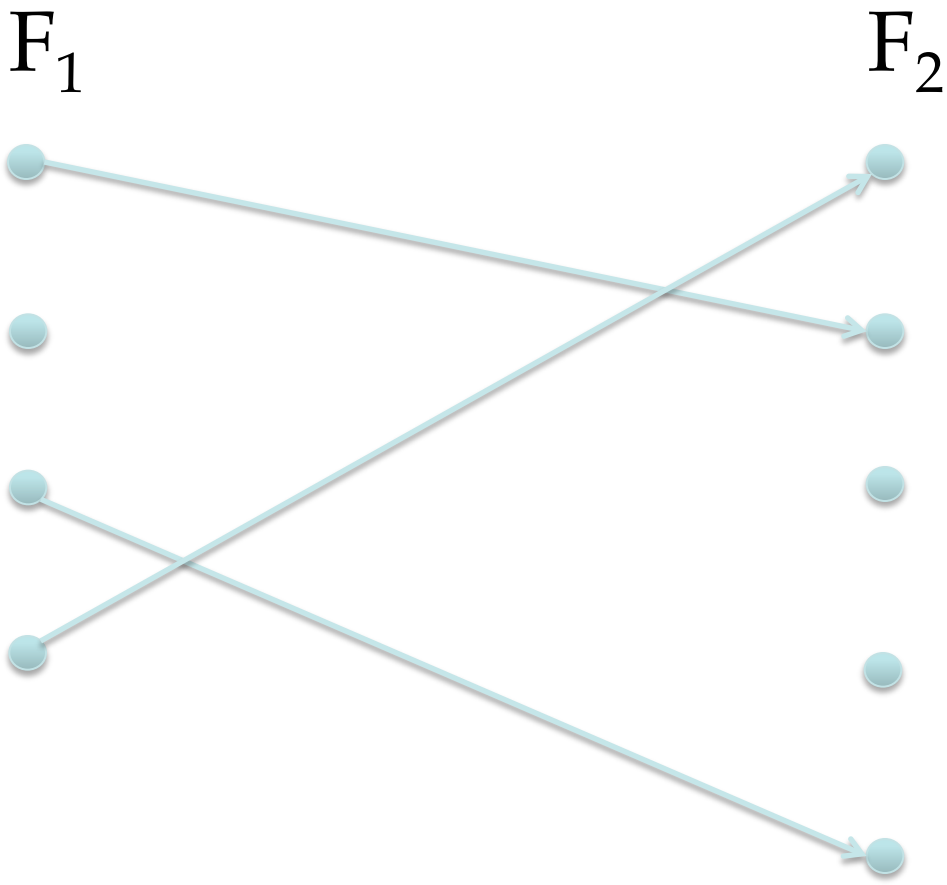
Test Plan: Person Matching

- Output/Format 1 is F_1 = Census Data
- Output/Format 2 is F_2 = IRS Tax Data
- Say for each unique person in F_1, the System Under Test (SUT) is to predict the best person match(s) in F_2, if any
- Say there are N matches in the Truth, adding up both positive and negative matches
- The GAMUT Truth is M positive matches
  - Therefore $M \leq N$

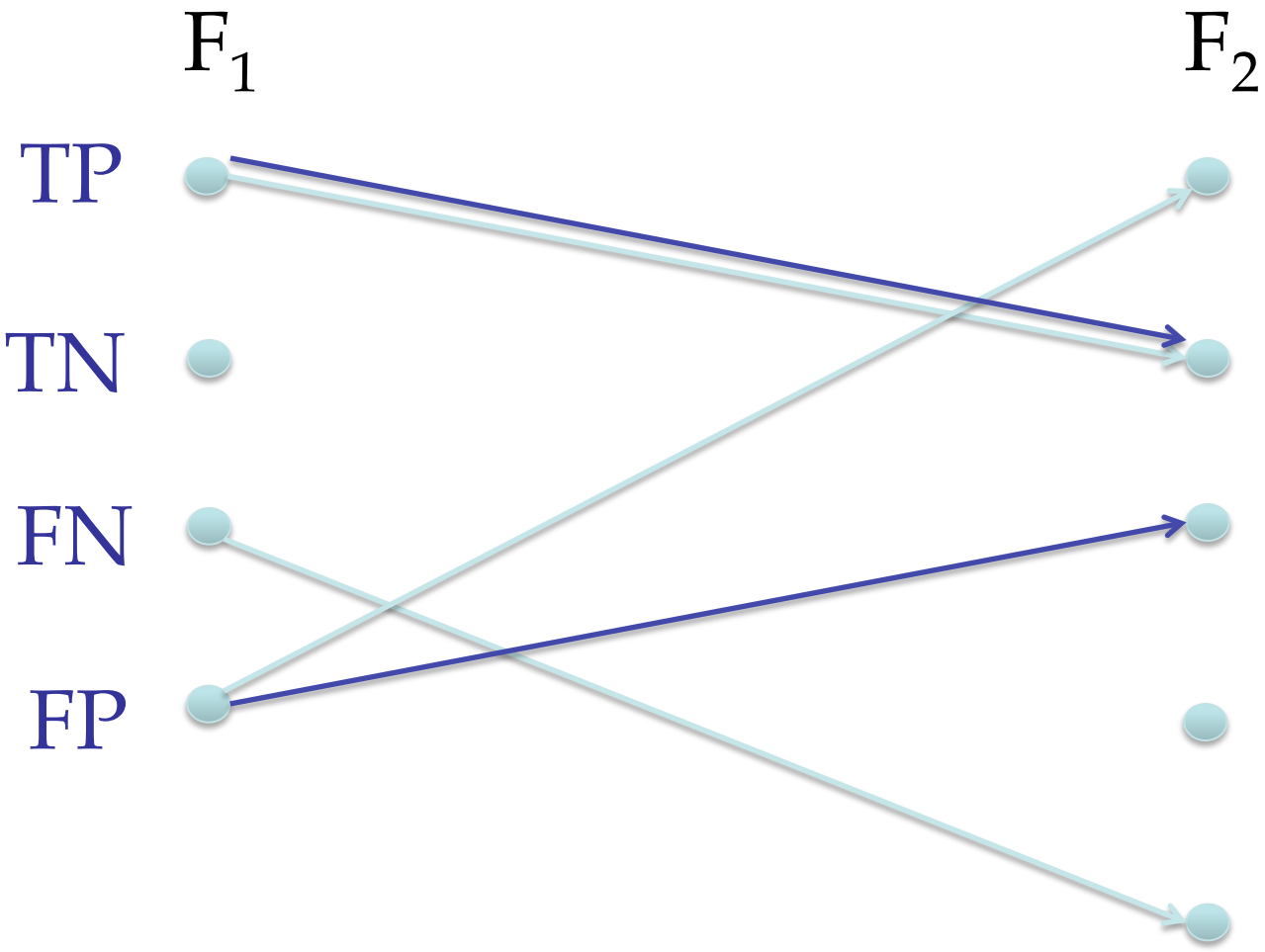
Test Plan: Cont.

- The SUT predicts m matches ($0 \leq m \leq N$)
- Of the m matches, GAMUT Truth says cm of them are correct ($0 \leq c \leq 1$): "True Positives"
- Therefore m - cm = m(1 - c) are "False Positives" (Type I errors)
- Also, one can compute:
  - "False Negatives" = M - cm (Type II errors)
  - "True Negatives" = N – M - m(1 - c)

Example of Test Truth



Example of Test Truth with Classification System Results



Confusion Matrix

		SUT Prediction	SUT Prediction	Row Sums
		Positive Match	Negative Match	
Data Truth	Positive Match	**TP** cm	**FN** M - cm	M
Data Truth	Negative Match	**FP** m(1 - c)	**TN** N – M - m(1 - c)	N - M
Column Sums		m	N – m	N

FP are Type I errors; FN are Type II

Case A

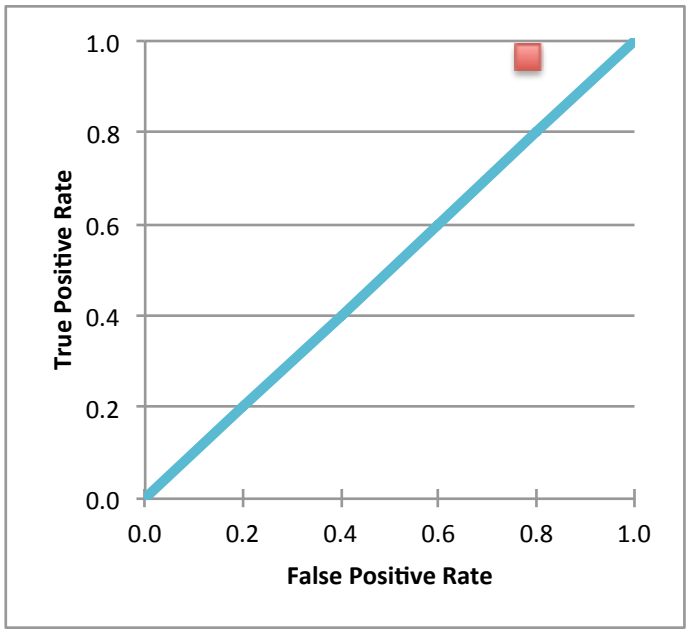
Example Test – Case A

Generic ROC Plot and Confusion Matrix (Case A)

N	M	m	c
985	848	925	0.8843

	Prediction of S.U.T.		
	Pos	Neg	
Pos	818	30	848
Neg	107	30	137
	925	60	985

TPR	FPR	A	f
0.965	0.781	0.861	0.923



ADI ADVANCED DOCUMENT IMAGING

gunnison Consulting Group Inc.

Case B

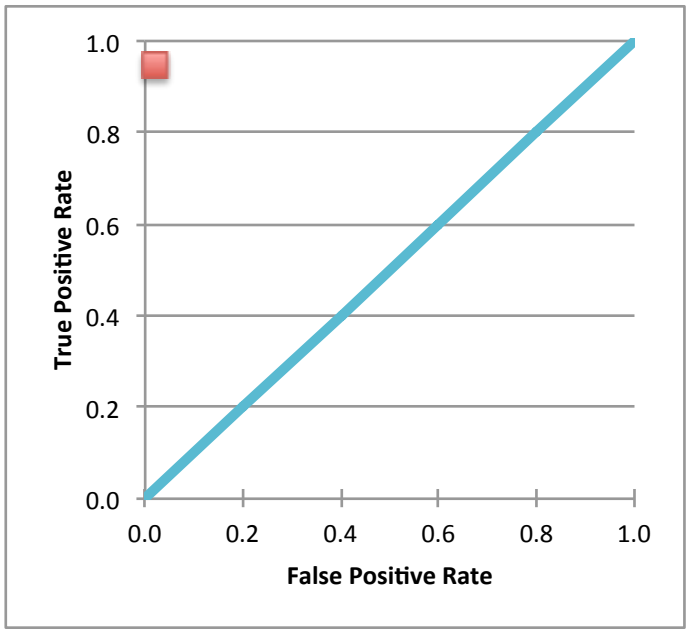
Example Test – Case B

Generic ROC Plot and Confusion Matrix (Case B)

N	M	m	c
985	848	808	0.9963

	Prediction of S.U.T.		
	Pos	Neg	
Pos	805	43	848
Neg	3	134	137
	808	177	985

TPR	FPR	A	f
0.949	0.022	0.953	0.972



ADI ADVANCED DOCUMENT IMAGING

gunnison Consulting Group Inc.

Conclusions

- The use of synthetic GAMUT testing data can significantly speed up and improve Administrative Records testing at Census, leading to improved system performance

- It can also help in other areas, for example:
  - Record Linkage Generally
  - Data Capture (all "modes")
  - Health Records Systems
  - Intelligence Systems
  - Census 2020 Research and Evaluations

- Remember, we don't aim to replace testing with "real" data, but rather to supplement it to speed up the development process to achieve quality software that's scalable and ready for production

Questions or Comments?

- Contact:
  - Brad Paxton brad.paxton@adillc.net
  - Steve Spiwak steve.spiwak@adillc.net
  - Tom Hager tom.hager@adillc.net
- ADI Website:
  - www.adillc.net
- Sample data available on request