Two New Quality Metrics for Measuring Big Data Record Linkage Systems

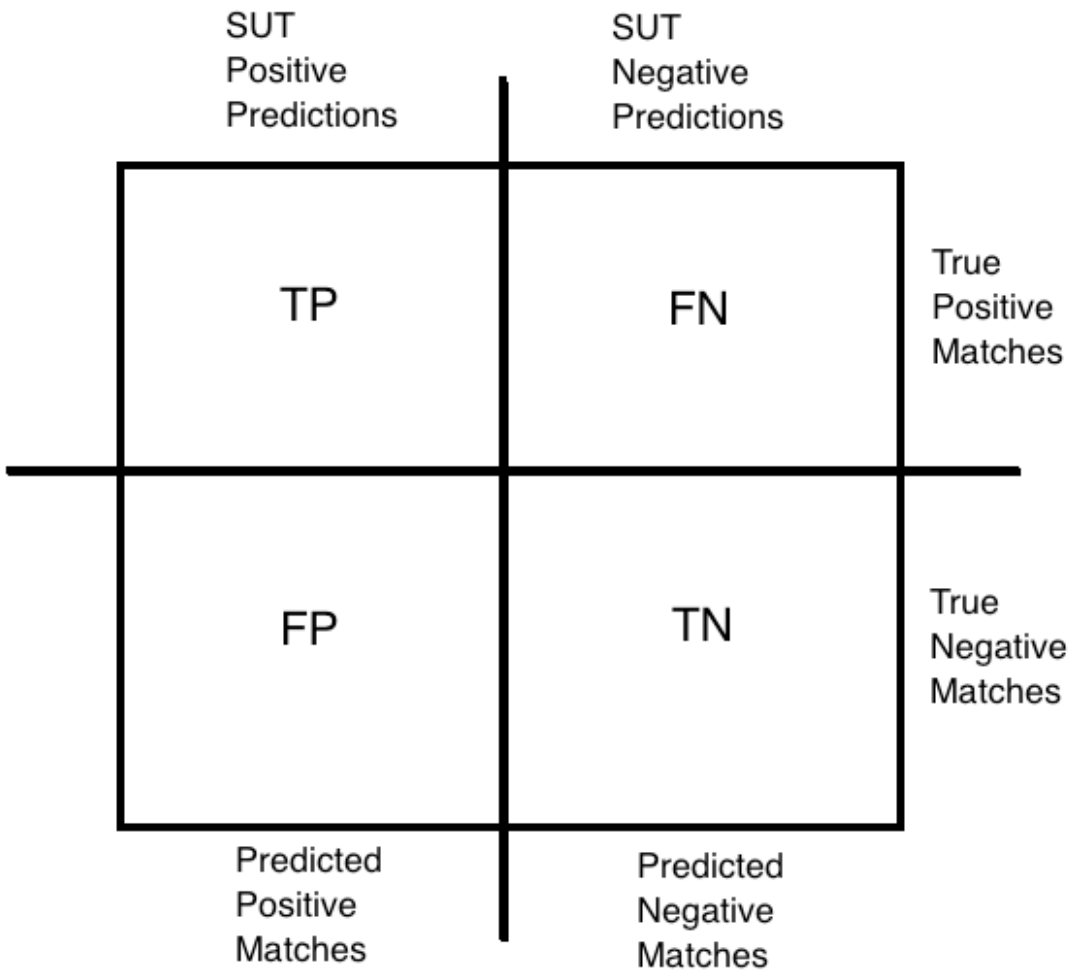
FedCASIC
Suitland, MD
May 3, 2016

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

Background

- Record Linkage (RL) systems are being increasingly used to create better “big data”
- At JSM2014, we explained two different methods for efficiently testing RL systems (Ref.1)
- To analyze RL test data, and determine when an actual improvement has been made, one needs to understand the trade-off between precision and recall
- Using this trade-off, we present two new quality metrics that can help improve RL systems through a continuous cycle of testing and tuning

Typical Confusion Matrix



Precision & Recall

- Precision: $c = TP/(TP + FP)$

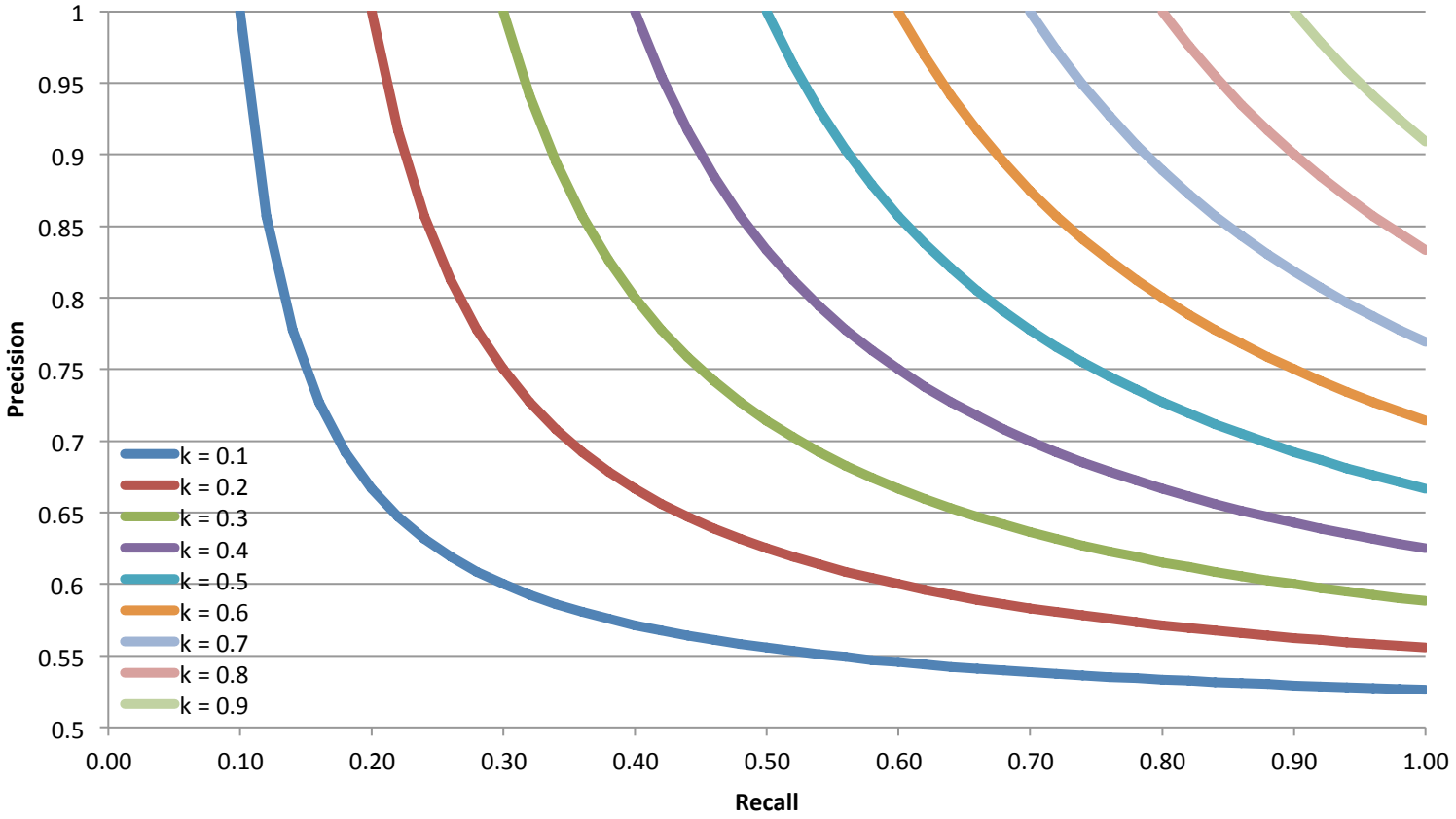
(the fraction of predicted matches that are correct)

- Recall: $r = TP/(TP + FN)$

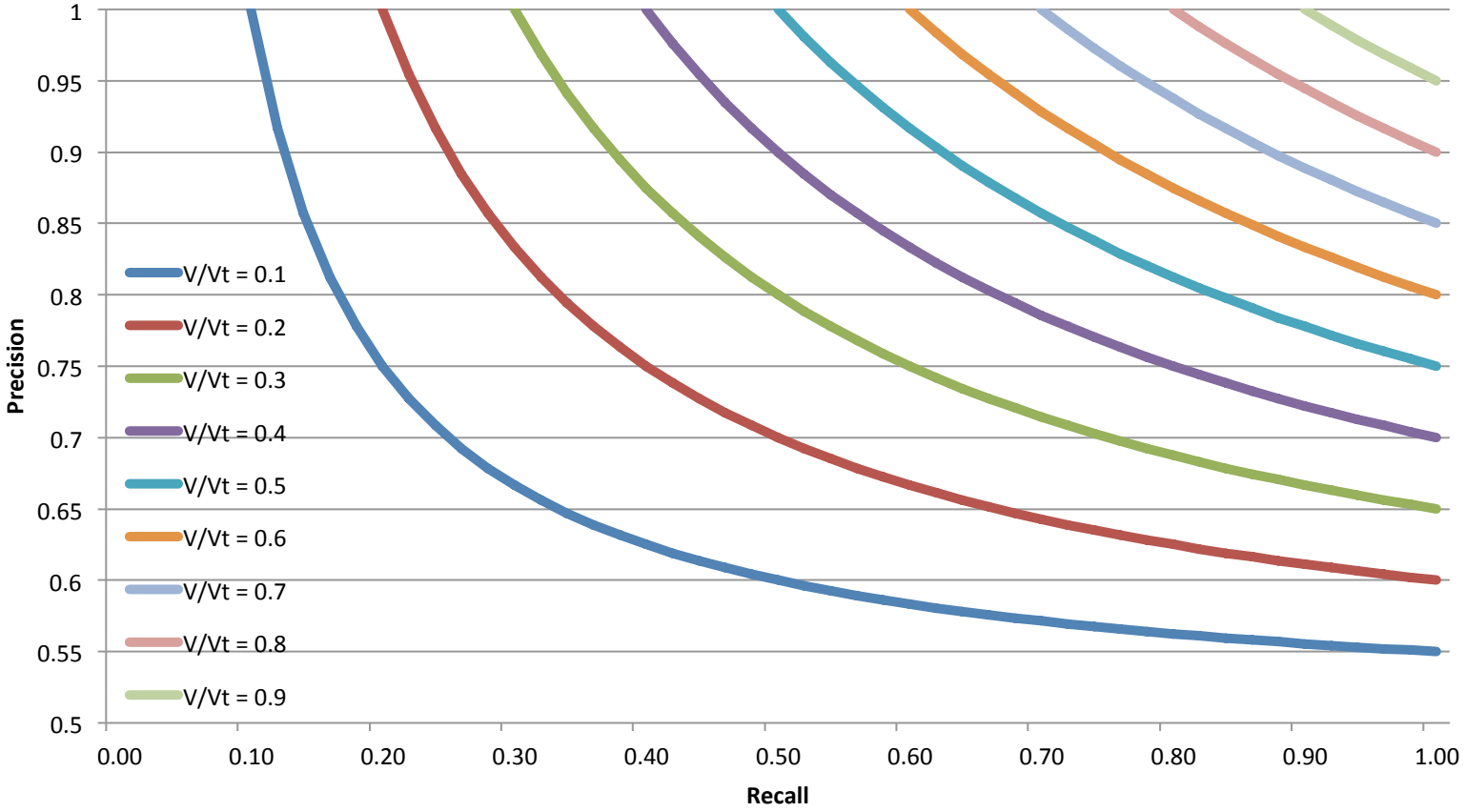
(the fraction of correct matches that are predicted)

- There is a fundamental trade-off between precision & recall!

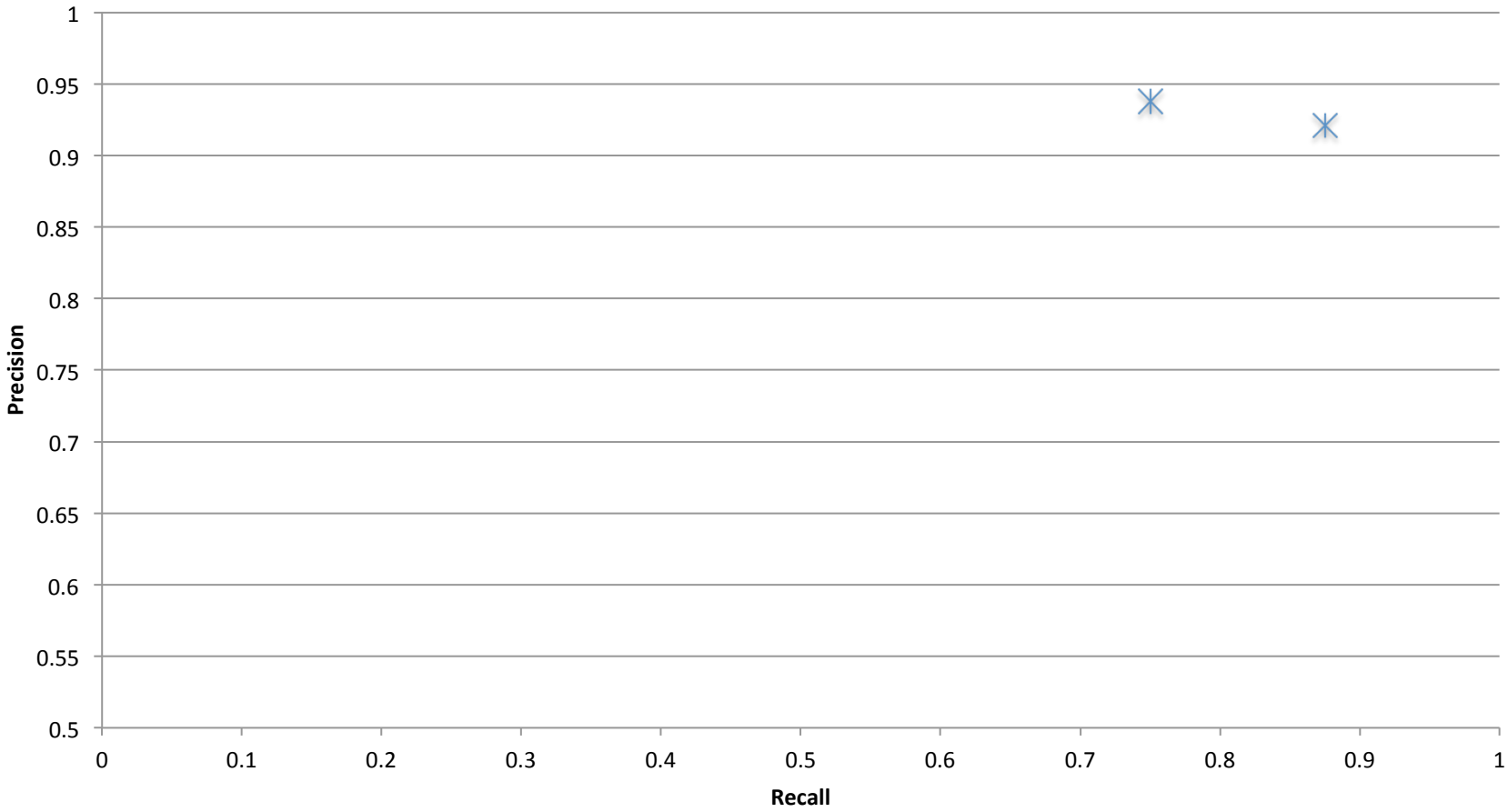
Precision vs. Recall
(for Constant Quality Metric k)



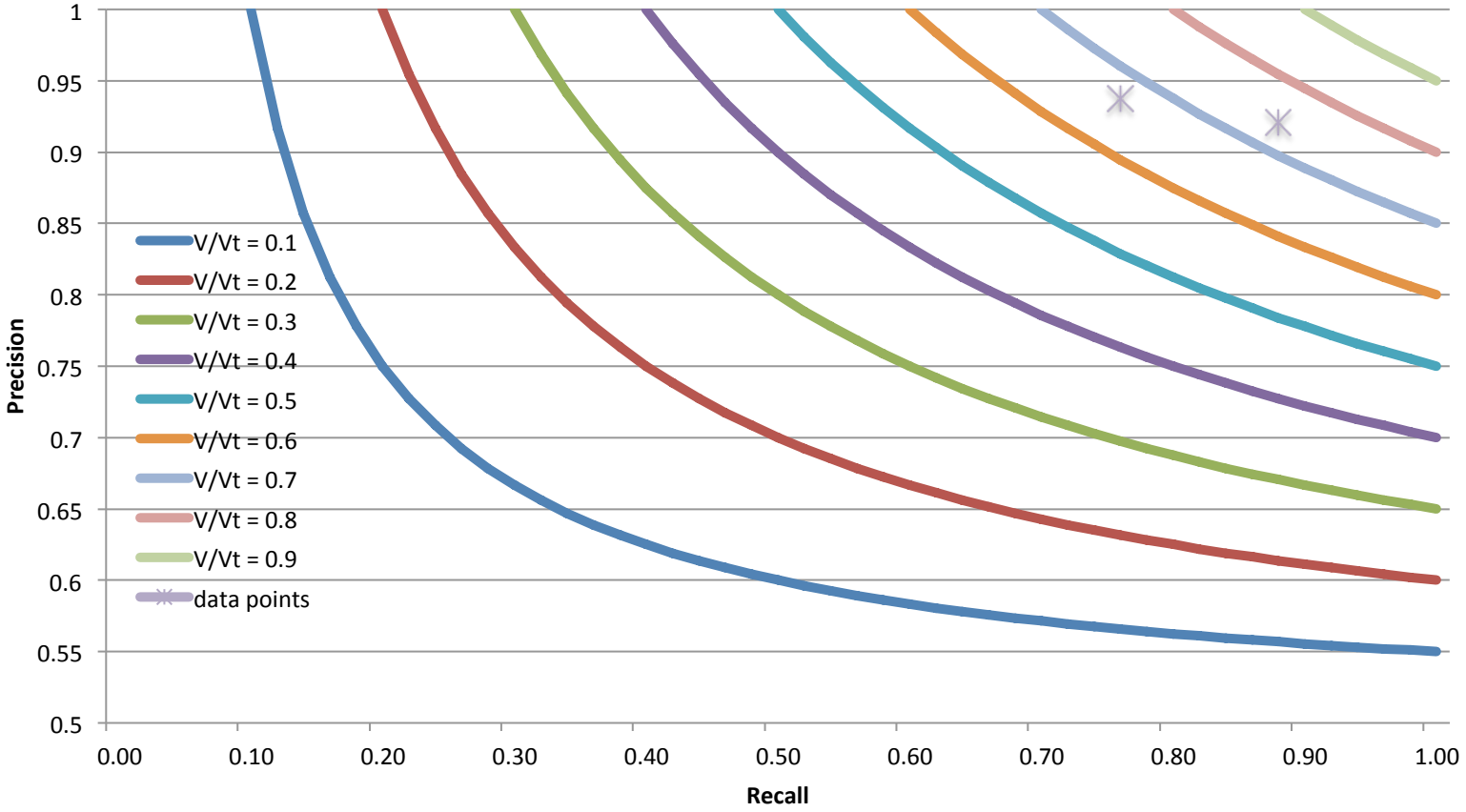
Precision vs. Recall (for Constant Value Metric V/V_t)



Which Test Result is Better?



Aha! This helps!



Conclusions

- A new quality metric (k) helps determine the best RL system in terms of accuracy
- A new (Value) metric estimates the relative savings due to RL system improvements
- A continuous cycle of testing and tuning the Census RL system could save billions of dollars
- I predict that essentially all the 2030 Census will be done with Record Linkage!

References

1. Paxton, K. Bradley, *Using Record Linkage to Create Big Data? How Good Is It?* In JSM Proceedings, Government Statistics Section. Washington, DC: American Statistical Association, 2014, pgs. 742-754.

2. Christen, Peter, *Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection*, Springer, 2012, p.167.

3. Alvarez, Sergio A., *An Exact Analytical Relation among Recall, Precision, and Classification Accuracy in Information Retrieval*, Technical Report BC-CS-2002-01, Computer Science Department, Boston College, June 2002.