

On the Authoritative Data Sources: One Data Element at a Time

Applications

*DAMA-National Capital Region
Chapter Meeting
March 9, 2010, Washington DC*



Kelvin K. T. Huang, Ph.D.
CEO, Taskco Corporation
Adjunct Professor, Fordham University
Research Affiliate, MIT Information Quality Program
huangkt@mit.edu
1-914-3747468

Taskco

ADS Applications

- Health Care
- Intelligence Community
- Finance
- Data Source Quality and Operators

The Beach Program – Univ. of Sydney

Bettering the Evaluation and Care of Health

- The BEACH© program continuously collects information about the clinical activities in general practice in Australia including
 - Characteristics of the GPs
 - Patients seen
 - Reasons people seek medical care
 - Problems managed, and for each problem managed (direct link)
 - medications prescribed, advised, provided, clinical treatments and procedures provided
 - referrals to specialists and allied health services
 - test orders including pathology and imaging
- The BEACH database currently includes about 1,100,000 GP-patient encounter records (July 2009).
- The **weighted average monthly treatment cost (WAMTC)** is calculated for specific classes of medicines as part of the Pharmaceutical Benefits Pricing Authority's (PBPA) annual price reviews of PBS medicines. The WAMTC is calculated from dosage regimen prescribed by GPs.

Designated Data Source

- PBPA Selected BEACH as the **designated data source** for calculating WAMTC for the following therapeutic drug classes:
 - ACE inhibitors (Acels)
 - Calcium Channel Blockers (CCBs)
 - Proton Pump Inhibitors (PPIs)
 - HMG CoA Reductase Inhibitors (Statins) - Regular potency
 - HMG CoA Reductase Inhibitors (Statins) - High potency
- Other dosage data may be useful for those companies who want to make a comparison between BEACH and the designated data sources.
 - Angiotensin II Receptor Antagonists (ATRAAs)
 - H-2 Receptor Antagonists (H2RAAs)
- Pharmaceutical companies with products subject to WAMTC price adjustments may be interested in purchasing these dosage data for the next price reviews for Acels, CCBs, PPIs, Statins-Regular, Statins-High.
- Dosage data for interim periods may also be useful for tracking WAMTC trends before the start of the next pricing round.

We Failed to Connect the Dots

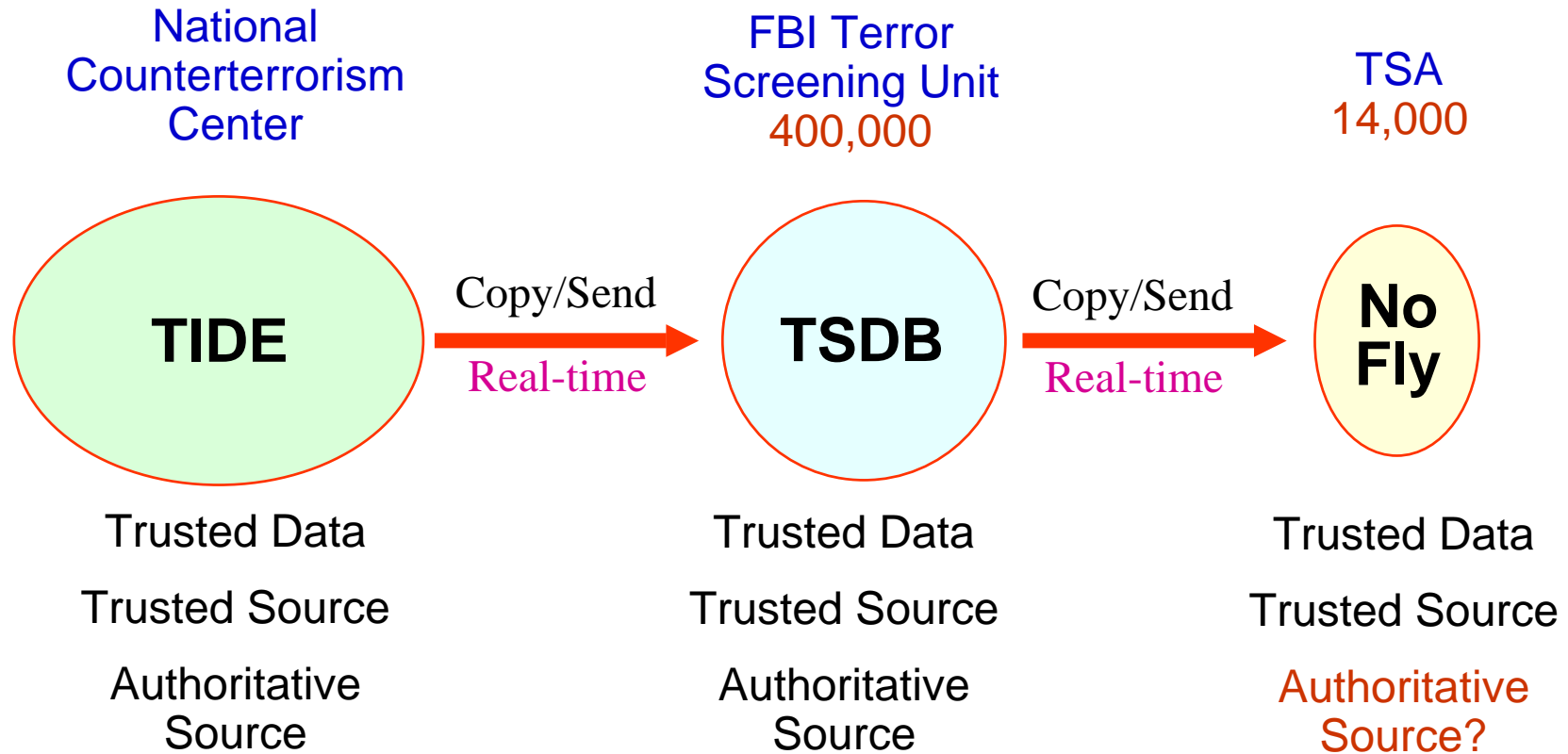
The 23-year-old Nigerian attempted to destroy an Flight 253 on its way from Yemen to Detroit On 2009 Christmas Day

- Umar Farouk Abdulmutallab, was in TIDE, the **Terrorist Identities Datamart Environment** owned by the **National Counterterrorism Center** in McLean, VA.
- The **FBI Terror Screening Unit** goes through the TIDE database and identifies those who are the real threat and places them in their **Terrorism Screening Data Base**, which contains about 400,000 individuals.
- Then the FBI passes these names to **TSA**, who maintains its own **"No-fly" list** of about 14,000 people.
- FBI did not deem Abdulmutallab a real threat, he never made it to TSA's "no-fly" list.

What happen?

- FBI failed in their analysis by not putting Abdulmutallab on the list?
- Disconnected Data – how the data was shared or, in this case, not shared?
 - This common practice of copying data and sending it to others and calling it 'sharing' is really problematic
- President Obama said; "There was a mix of human and systemic failures that contributed to this potential catastrophic breach of security. We had the intelligence information, but **failed to connect-the-dots.**"

Disconnected Data



Official government assessment:

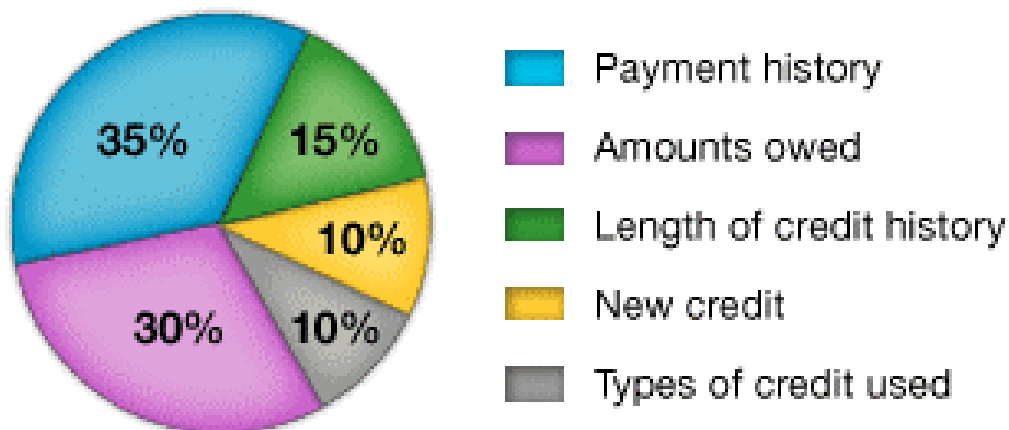
"A failure of intelligence analysis, whereby the CT community failed before December 25 to identify, correlate and fuse into a coherent story all the discrete pieces of intelligence held by the US Government related to an emerging terrorist plot against the US..."

What can go wrong?

- Notion of trusted data coming from a 'trusted source'
 - When you copy data you have trusted data, but no authority other than yourself.
 - After data has been copied, it puts the burden on the recipient to ensure it has not been tampered with.
- Make sense of context
 - An individual who purchased a ticket in cash and has no luggage (TSA)
 - His father reported him to the US Embassy in Nigeria (TIDE).
- Connect the dots
 - Allow TSA analyst with the proper credentials gain access to the full or limited set of information.
 - When data is everywhere and the fusing and correlating must be done by an analyst with tools
 - Support real-time "connect-the-dots" bi-directional analysis – connects directly to the same source, but constrains, or limits the data based on the user's credentials and policies.

Credit Report – FICO Score

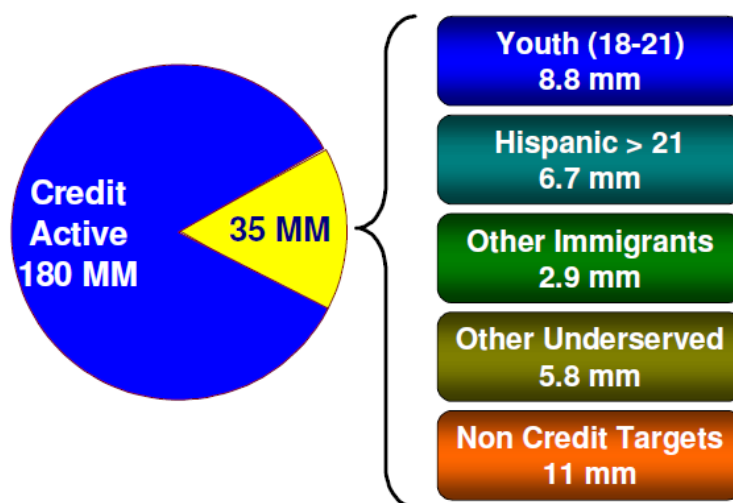
- Identification – name, address, SS#
- Employment – current and previous
- Credit – accounts, limits, payments
- Inquiries – requests for info.
- Public Record – foreclosure, tax liens, bankruptcy.
(Negative information remains for 7 years. Bankruptcy remains 10.)



Non-Traditional Data Sources

■ The thin-file problem – lack of sufficient credit information

- According to Fair Isaac's data, 15 percent of the country's adult population falls into the thin-file category (that would earn no FICO score), while 10 percent would garner no "hits" whatsoever (again, no FICO score).
- Nearly 18 million Americans have files too thin to produce a credit score, and another 17 million have no files at all



Source: Experian Data Reporting Prepaid Executive Roundtable April 28, 2005

- Interested financial institutions – mortgage lenders, auto lenders, insurers and credit card companies, as well as the major government sponsored entities in the mortgage market.
- More data sources enable better algorithms – By using the random-tradeline approach, it can identify consumers who may have been delinquent on a bank card but always paid the mortgage

Alternative Authoritative Data Sources

- Alternative Data Sources
 - Incorporate nontraditional data such as consistent rental, telecommunication and/or utility bill payments as an alternative means to measure creditworthiness.
 - Add demographic information
 - Reaching out to the telecom and utility companies to show them the evidence that supplying data to credit bureaus helps their bottom line as well.
- Key Points from Center for Financial Services Innovation (CFSI)
 - More work needs to be done, including identifying the most useful data points.
 - Scale will not be achieved until alternative data are included in automated underwriting processes and standards are developed
 - Data providers and lenders must target the right market segments
 - One data aggregator collects information about consumer debit accounts—
 - if a consumer opened a new checking account or bounced a check. Other data providers provide similar kinds of positive and negative information about consumer's behavior in regard to phone bills or payment plan performance--i.e., whether the consumer always pays his magazine subscriptions or rent-to-own furniture bills

Data Sources & Quality Proposition

- We want to use data and feature attributes from a **variety of reliable sources**
- All data is good, but it's **not created equal**. Some needs to be precise; much doesn't
- We must capture and use information on **data accuracy and precision** (or “quality”) in order to effectively leverage the data.
- **Accuracy** is the degree of veracity (closeness to the actual value) or “bulls eye” while **Precision** is the degree of reproducibility, or “grouping”.



High **accuracy**, low **precision**



High **accuracy**, low **precision**

Taskco

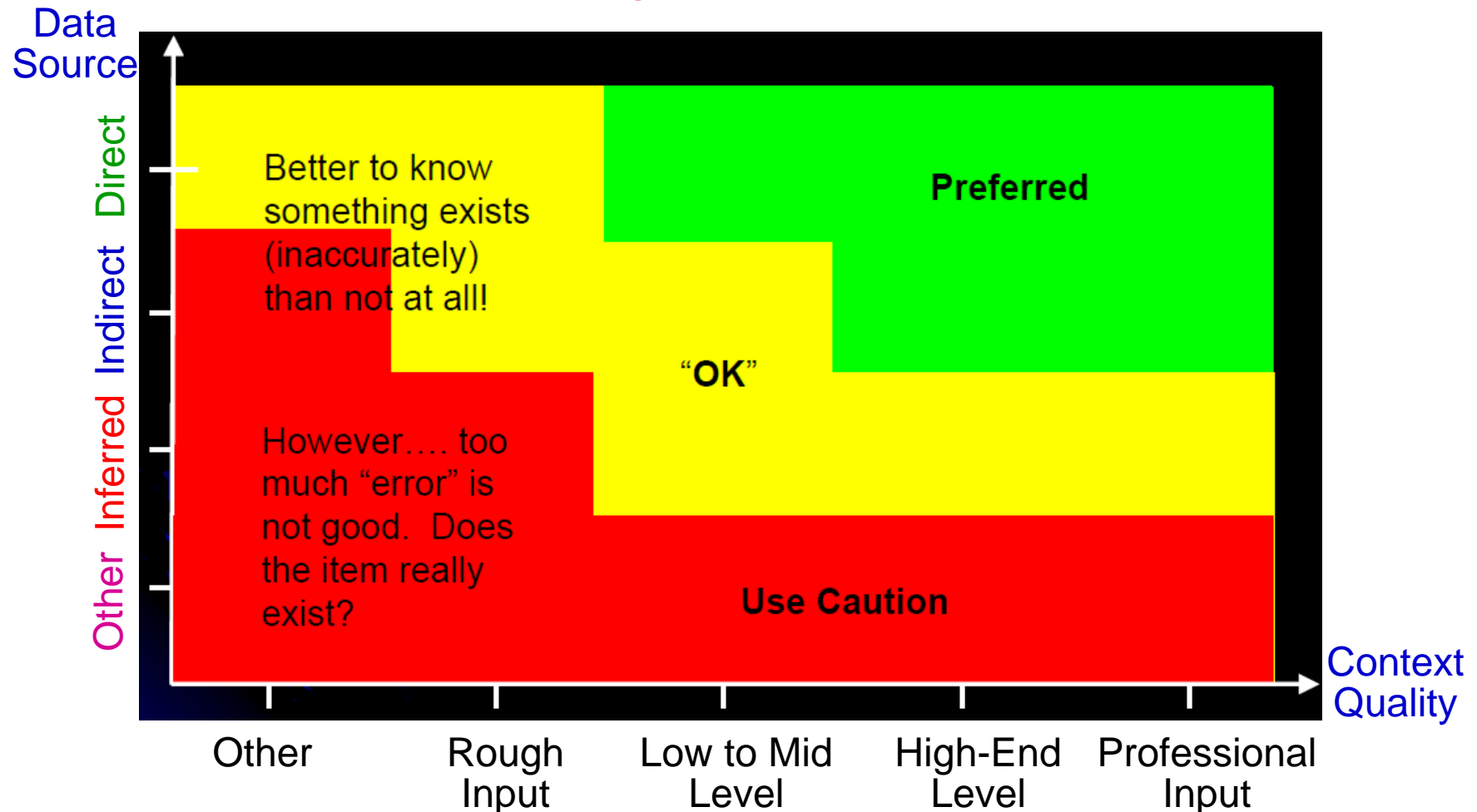
Data Quality Solutions for ADS

- Develop metrics to quantify “quality”
 - Context quality (How accurately do we know the data elements?)
 - Data Quality (How representative is the data we are locating?)
- Provide guidance on the accuracy required
 - What leveled is needed (e.g. edit or addition)?
- Develop a quality matrix, with recommendations
 - Provide quality combinations for data collection
- Store quality metrics for each data point collected
- Provide editing and analytical capabilities
 - Sort, report, edit, replace, etc. by any metric

Reliability – Data Source Quality Matrix

Defining the “Accuracy Required” is the third axis for the matrix

Confidence Ranking (Discrete or Probability)



Current & Future Works

■ ADS Theoretical Foundation

- Logical Operators
- Arithmetical Operators
- Temporal Operators
- Quality Operators

■ Algebraic Structure

■ Reliability Model

■ Security

Questions?



Kelvin K.T. Huang
kthuang@taskco.us
Taskco Corp
1-240-4037403
03/09/2010

The difference is in having Quality Data

■ Contact us

- Kelvin K.T. Huang (kthuang@taskco.us)

■ Resources:

- Medical Doctor, Health Informatics
- CFA & Accountant, Financial Informatics
- Military Retiree, Military Informatics

■ What we do

- A consultancy and service provider
- A solution partner in all aspects of data management and analytics