Biographic Entity Resolution

Challenges of managing and sharing

John N. Dvorak
Senior Level IT Architect
September 16, 2014
Global Identity Summit

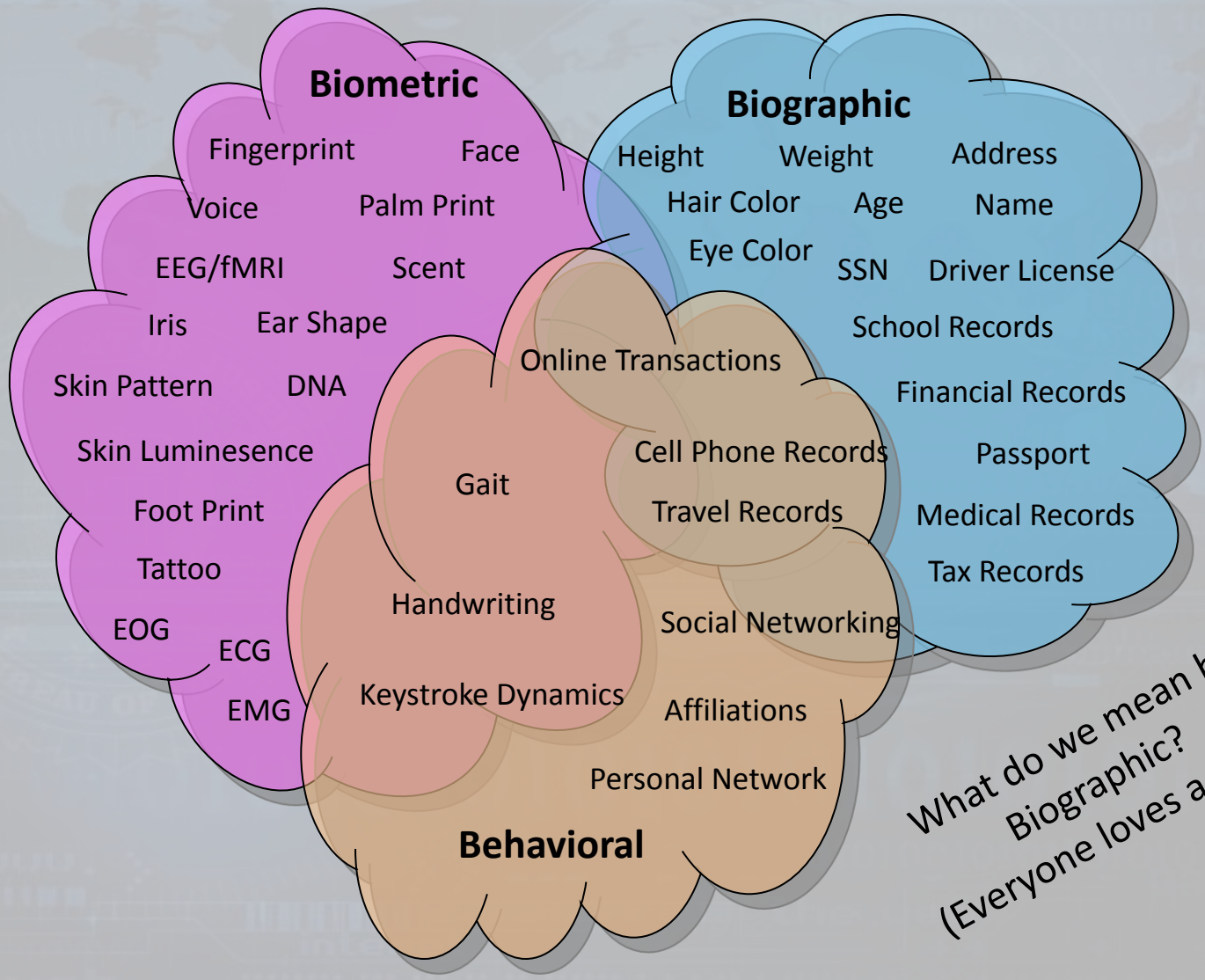
Discussion Topics

- Challenges of data
- Resolving entities
- Managing and sharing resolved entities
- Future (*a.k.a. the world John would like to see*)

Brief (very!) Introduction

- **Entity Resolution:** The process of determining whether two or more references to real-world objects such as people (individuals), places, or things are referring to the same object or to different objects. This concept is sometimes referred to as Entity Correlation, Entity Disambiguation, or Record Linkage, and includes related concepts such as Identity Resolution. (from draft DARA)
- **Entity Map:** Complete enriched entity data that includes the linkage of relationships between people, places, things, and characteristics of data resulting from an entity resolution process.
- For a deeper discussion about biographic identity challenges and technologies, attend Dr. Keith Miller's Biographic Identity Technologies session, Wednesday, 9 AM (Track C)

Relationships in Identification



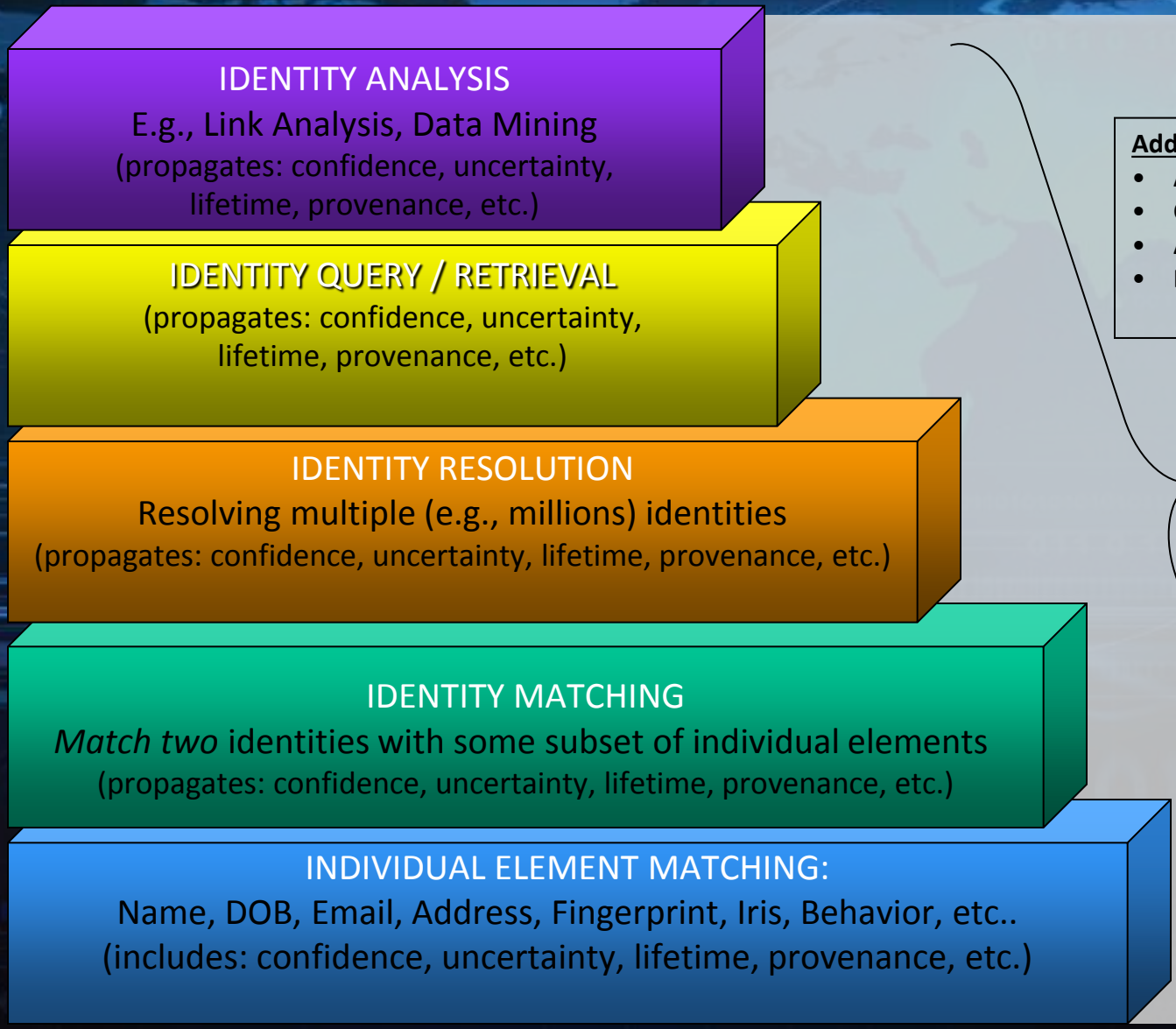
What do we mean by Biographic?
(Everyone loves a Venn)

The Challenges of Data

- Of course, volume, velocity, variety, veracity
- Data comes from a variety of **sources** (government, commercial)
- Data comes in a variety of **formats** (structured, unstructured, semi-structured, questionably structured)
- Data sources do not contain a single, reliable, **unique** identifier to link records
- There are many errors & ambiguities! (i.e. data is dirty)
  - Transliteral
  - Typographic
  - Cultural variation
  - Mis-fielding
  - Mis-coding

Last	First	DOB
Doe	John	19501203
Doe	Jack	19500312
Smith	Jayne	19820703
Doe	J	195003
Jane	Smith	19820703
Wright	Albert	2025558362

Layers of Use of Biographic Identity



Additionally – apply at all levels

- **Authentication**
- **Credentialing**
- **Access Control**
- **Privacy**
  - **(includes Anonymization)**

Source: MITRE Corporation 2014

The Judgment Problem

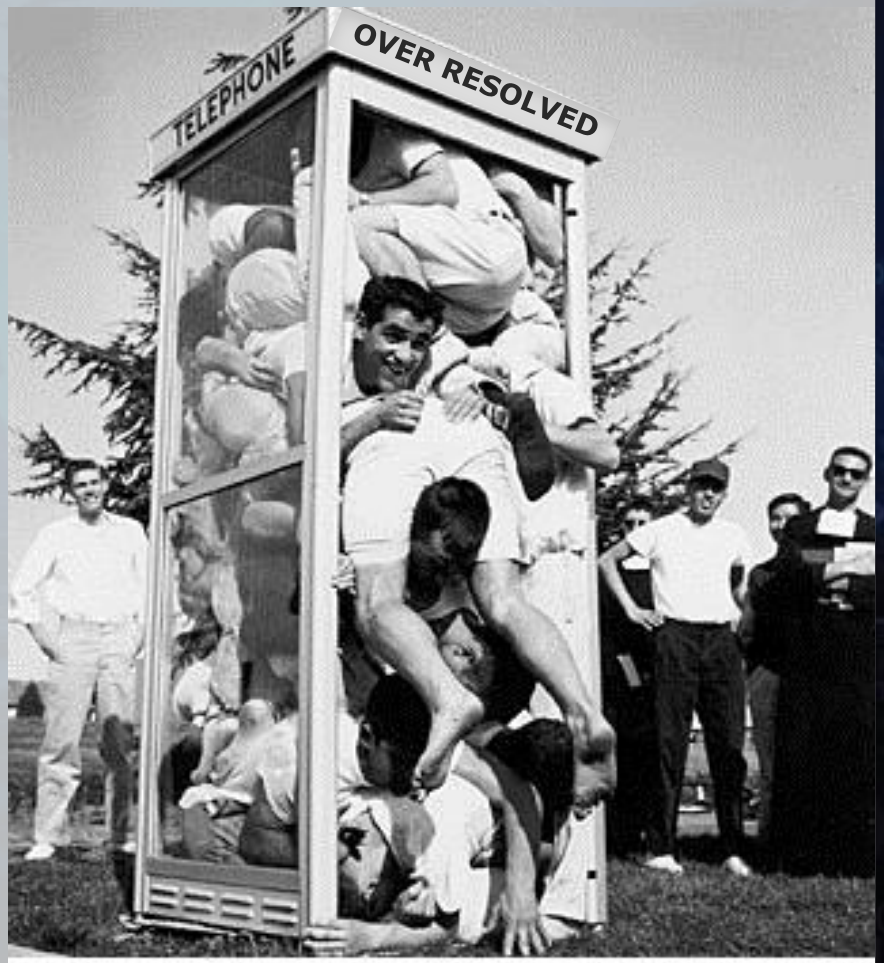
(or We can't automate everything)

- Automation provides **candidate** matches, based on carefully constructed rules
- **Judging** which records should be linked is an **analytical** task
- **Manual review** is usually necessary and manual "override" needed to correct errors

Match criteria must be balanced

Beware **Over Resolving!**

Making match criteria more tolerant to small errors gets better results, but can lead to over-resolving data



Photograph by Joe Munroe/Ohio Historical Society Collections

Entity Resolution…

- Automates manual tasks, improving time to analysis
- Broadens search attributes and richness of results

But…

- Can mislead analysis if we don't understand how business rules apply to results
- Poses another world of challenges to sharing

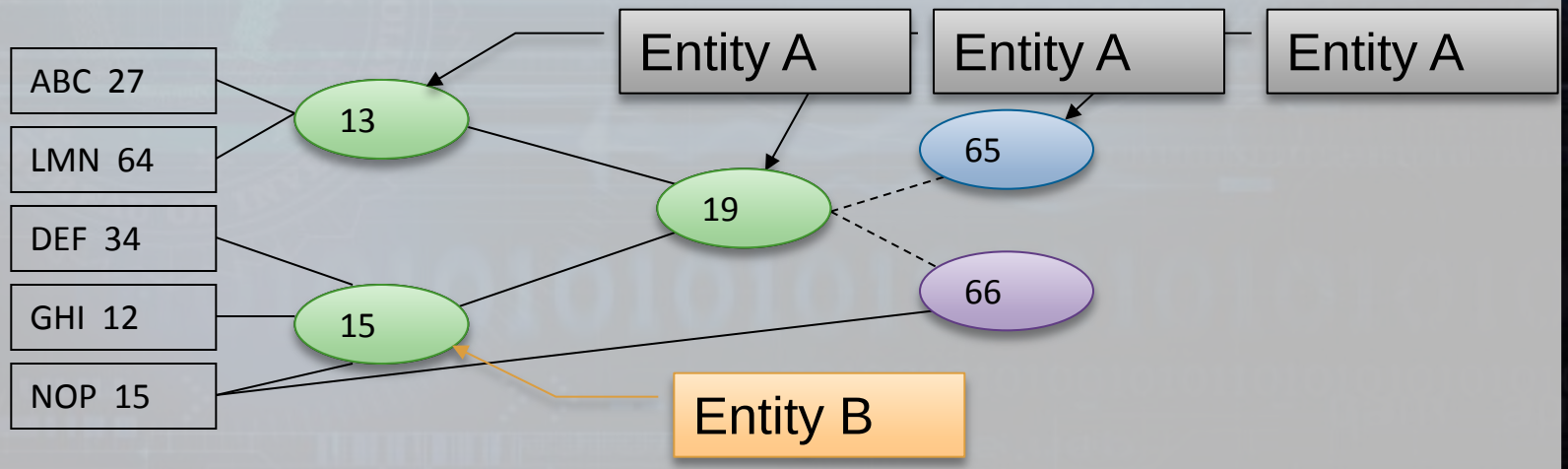
Entity Management

Challenges of Entity Management

- How do we **address change**? (change process, training)
- How do we meaningfully warn an entity owner of new information? (alerting)
- How do we keep up with the **volume** of change when human intervention is needed? When and how are we compelled to care?
- How do we maintain **provenance**? (origin(s) and *who*, *what*, *whens* of change)

Tracking pedigree: An example

Entities are merged and branched over time, all of which needs to be tracked and made visible (provenance/pedigree):



And now…

Assuming you have tackled the small problems of entity extraction, resolution and local management, and also assuming you don't live on an isolated island of analytic wonderment…

You need to share what you know

And process what others know

(restricted by laws & policies, not technology)

Sharing Resolved Entities

Sharing resolved entities – the next frontier

The Challenge: How do we share correlated entity maps across agencies with disparate architectures, technologies, policies, etc?

(Hint: answer cannot be “everyone use same proprietary software solution”)

Related questions such as:

- How do shared entities maps get inserted in a way that is **useful** for the operator?
- How do I treat the data? An external dataset? Trusted? Untrusted? Partially trusted?
- On what fields do we **match**?
- What matching algorithm did the originator use? Same? Different? Do we trust it? (Do we care and when?)
- Do I have access to the **source records**? Provenance?
- How do we **react to candidate data** received? How and when do I share my enriched data with the originator?

We need to be careful not to institutionalize bad data!

More complex data, more complex challenges

Today:

- Technical: proprietary matching algorithms, proprietary scoring mechanisms, non-standard export formats, isolated innovation
- Business: policy variation, inconsistent processes, often separately managed modalities

A hopeful future:

- Open standards for sharing correlated data/entities, compatible (or at least “understandable”) policies, sharable algorithms/analytics, broad innovation within and between many modalities

A (shameless) plug for the DARA

SA IPC's Data Aggregation Working Group (DAWG) under PM-ISE is developing the **Data Aggregation Reference Architecture (DARA).**

- Response to **NSISS Priority Objective 10**: "Develop a reference architecture to support a consistent approach to data discovery and entity resolution and data correlation across disparate datasets"
- Among its objectives: "Define a reference architecture that enables entity resolution and data correlation, and disambiguation across multiple data aggregation investments."

Data Aggregation Reference Architecture public statement:

http://www.ise.gov/blog/ise-bloggers/building-data-aggregation-reference-architecture-%E2%80%93-industry%E2%80%99s-help

Associated RFI:

https://www.fbo.gov/?s=opportunity&mode=form&id=e5ce39a84c1e09afab844a9487866a68&tab=core&_cview=0

SA IPC: Information Sharing and Access Interagency Policy Committee
NSSIS: National Strategy for Information Sharing and Safeguarding

Other Places of Interest

NIEM
https://www.niem.gov

Global Federated Identity and Privilege Management (GFIPM)
http://www.gfipm.net

DOJ's Global Reference Architecture (GRA)
https://it.ojp.gov/GRA

Open source:

SEARCH's (search.org) Entity Resolution Toolkit (ERS), sponsored by DOJ's National Institute of Justice
https://github.com/entityresolution

We advance and evolve

U.S. Department of Justice
Federal Bureau of Investigation
Washington, D.C. 20535

John N. Dvorak
Senior Level IT Architect
John.dvorak@ic.fbi.gov