

# DataServices WORLD

## Workshop: Data Quality, Data Access and Data Services

---

Thursday November 20, 2008 – 3pm

# **The Root Cause: Data is Never Perfect**

---

**In the real world, database information is  
never 100 percent perfect,  
never 100 percent consistent,  
never 100 percent complete,  
and it never can be.**

# A Human-to-Computer Gap

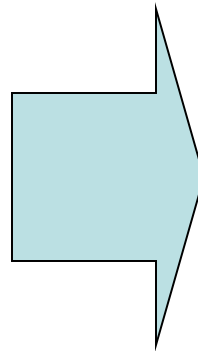
---

- **Humans usually perceive information approximately and easily tolerate data errors and variations**
  - *Humans and computers introduce inconsistencies*
  - *Other inconsistencies are intentionally introduced (e.g. by customers trying to avoid billing)*
- **Computer software is usually exact and unforgiving**
  - *Determining equality or inequality is easy*
    - *“Damianakis” = “Damianakis”*
    - *“Damianakis” ≠ “Smith”*
  - *Determining similarity is difficult*
    - *“Damianakis” ≈ “Damamakis”*

# Imperfect Data: Key Strategic Initiatives are at Risk

- Inconsistencies which wouldn't trouble a human, make data useless to a computer

- *can't retrieve it*
- *can't compare it*
- *can't match it*
- *can't link it*



- MDM / CDI
- Data Integration
- BI
- M&A
- Compliance
- Claims matching
- Data Warehouse
- CRM
- ...

# Solving the Problem

---

**Try to make the data perfect...  
(Sisyphus anyone?)**

*OR*

**Enable Services and Applications to  
handle the inconsistencies  
(which humans do naturally...)**

# What Makes this Problem Difficult?

---

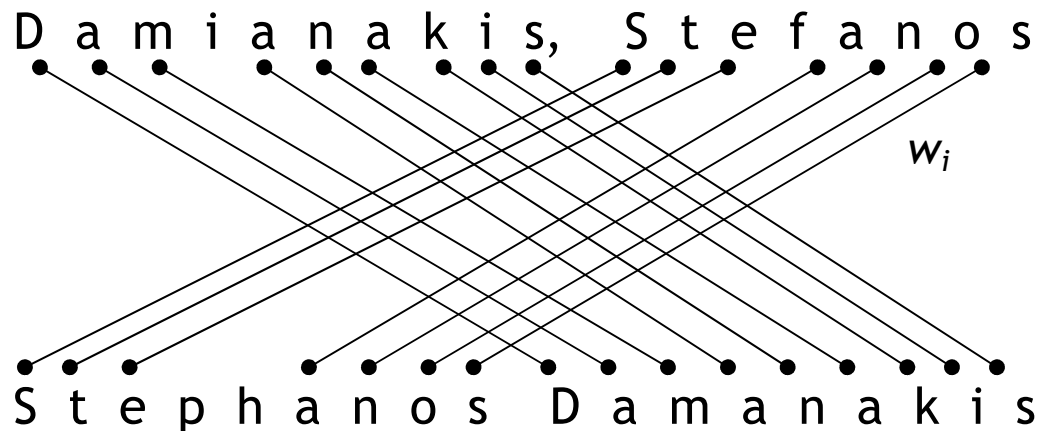
- **Data is never perfect – it will never be perfect**
  - *data changes over time in ways that change*
  - *CPUs and DBMSes are based on an “exact” world*
- **Human expertise is impossible to extract and program explicitly**

# Innovation Powered by Mathematics

Problem	Conventional Solutions	Mathematical Innovations	Advantages of Mathematics
Data Matching (compare data elements)	Soundex, NYSIIS, Edit Distance, etc	Mathematically Model Human Similarity (bipartite graphs)	<ul style="list-style-type: none"><li>• Superior Accuracy</li><li>• Symmetric error-tolerance</li><li>• No Guessing (of rules and parameters)</li><li>• Computational Efficiency &amp; Scalability</li><li>• Data Independence<ul style="list-style-type: none"><li>• people, assets, products, companies, claims, transactions, etc.</li></ul></li></ul>
Record Matching (compare/link database records)	Custom, Manual Matching Rule Sets with optional statistical parameters (probabilistic)	Mathematically Model Human Decisions (machine learning)	<ul style="list-style-type: none"><li>• Engineering Efficiency<ul style="list-style-type: none"><li>• easy to maintain and refine</li></ul></li><li>• Multi-lingual</li><li>• Real-time</li><li>• Sparse data support built-in</li><li>• Embeddable</li><li>• Quick and easy deployment</li><li>• DBMS independent</li></ul>

# Advanced Bipartite Graph Matching

- Uses Bipartite Graphs to compute similarity metrics
- Mathematically models a human notion of similar
- Captures a truer, richer notion of similarity than conventional methods



# Advantages of Mathematical Modeling

---

- **No need to clean data**
  - *Software excels with imperfect data*
- **No matching rules to build and maintain**
  - *No need to:*
    - *Check for different errors in the various fields*
    - *Guess and enumerate all possible record matching rules*
- **Pre-deployment**
  - *Less time and effort to deploy*
- **Post-deployment**
  - *Virtually no ongoing management to maintain matching accuracy*

# The Bottom Line

---

- **All Enterprise applications must deal with the imperfections inherent in “real-world” data otherwise...**
  - *applications will not work correctly*
  - *silos will never be truly connected*
  - *true value will not be realized by the enterprise*