Information Quality through Semantic Models

Enterprise Data Forum
Pittsburgh, November 2002

Joshua Fox
joshua@unicorn.com
http://www.unicorn.com

Agenda

- The Challenge
- The Goal
- The Role of the Semantic Model in Enterprise IQ
- IQ Imperatives
- The Development Process

The Challenge

Lack of Information Quality in the Enterprise

The Data Problem

- Enterprises have limited knowledge of
  - Where data is
  - What data means
  - Who is using it and how
  - Impact of a change

The Result of Lack of Knowledge

- Low quality business information
- Lack of aggregated information
- Reluctance to make critical changes – new processes or apps.
- Needless effort in manual integration and cleansing of data

Problems with Metadata

Missing, inaccurate, duplicate, and otherwise low-quality metadata makes repeated manual re-analysis necessary.

Syntax-Only Metadata

- No discernable semantics:

  CCPQ: NUMBER
- Barely discernable semantics:

  INVCID: CHAR(50)
- Semantics discernable to human reader only:

  EMPLOYEE_NAME: VARCHAR(50)

Metadata with Poorly-Defined Semantics

DB Table

EMPLOYEES

A human reader understand that this refers to "Employees," but does "Employees" include contractors, part-timers, new hires who haven't started yet, etc.?

Contradictory Metadata

- In Customer Database 1

  ID: VARCHAR(50)
- In Customer Database 2

  Identifier: NUMBER(0)

How do we identify customers? Are these different ID's or the same ID, where the VARCHAR is always parseable as a number?

Missing Metadata

Cobol Record

```
000014Jones0000
0Katherine000000
Y37344XX74CQXXXX
```

Business-Rule-Constrained Metadata

Customer DB

Name: VARCHAR	`Acme Widgets Inc.`
Is_Platinum: BIT	`TRUE`
Last_Years_Sales: INTEGER	`228,000`

The application sets *Is_Platinum* to TRUE if sales are >200,000, but that fact is not recorded in the metadata.

IQ Goals

Information Quality in the Enterprise

Information Quality

- Pioneered by Larry English
  - http://www.infoimpact.com
  - *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*
- Other Information Quality terms:
  - Data Quality
  - Total Data Quality Management (TdQM)

IQ Goals

- Agreed business data semantics/meaning
- Automatically generate accurate data transforms
- Make overlapping databases consistent
- Automatically generate data cleansing scripts
- Identify overlapping sources
- Maintain accurate transformations/queries

The Solution

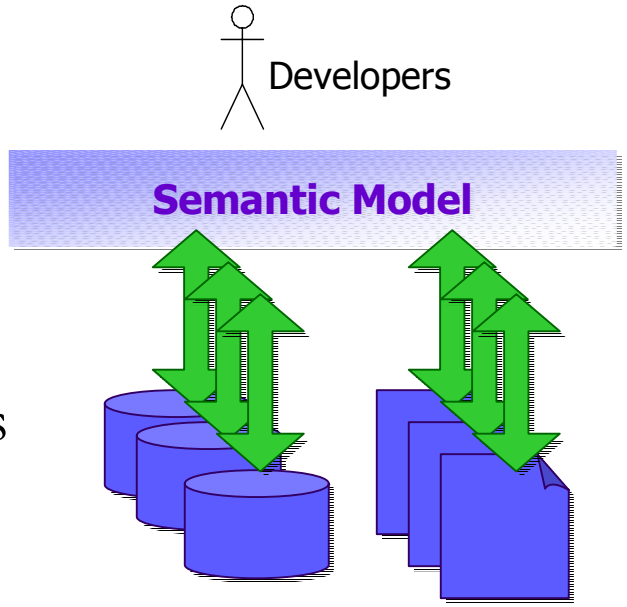
A Central Semantic Model

Semantic Model

A Semantic Model serves as a *single view* which unifies enterprise databases and message formats into a single asset by expressing *shared business semantics*.

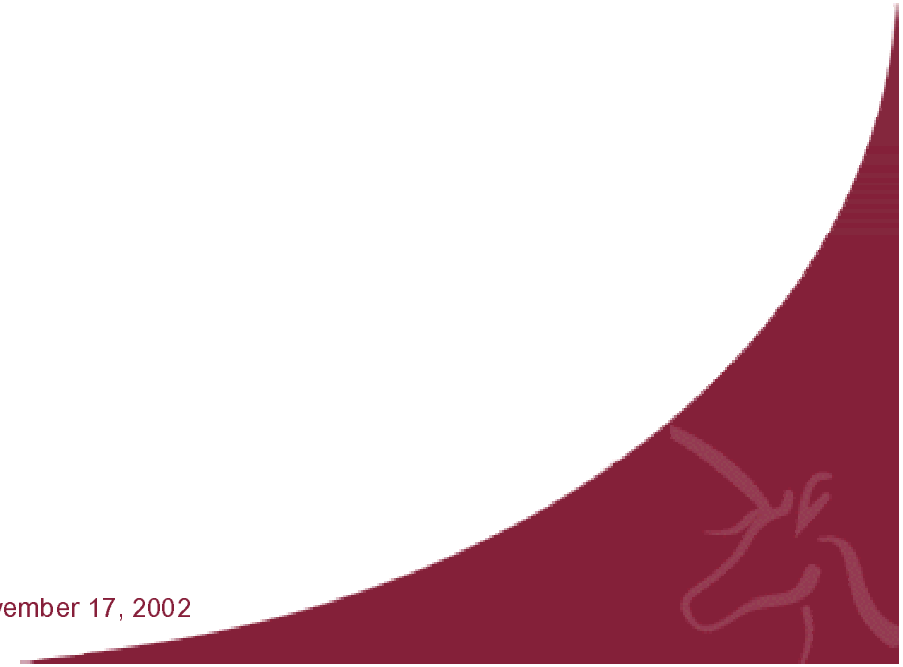
Understand Your Data - What Does It Mean?

- Capture business vocabulary in an Information Model
  - Hierarchy of Packages
  - Entities & properties
  - Business rules
  - Leverage off-the-shelf industry models
  - Consolidate distributed logical models

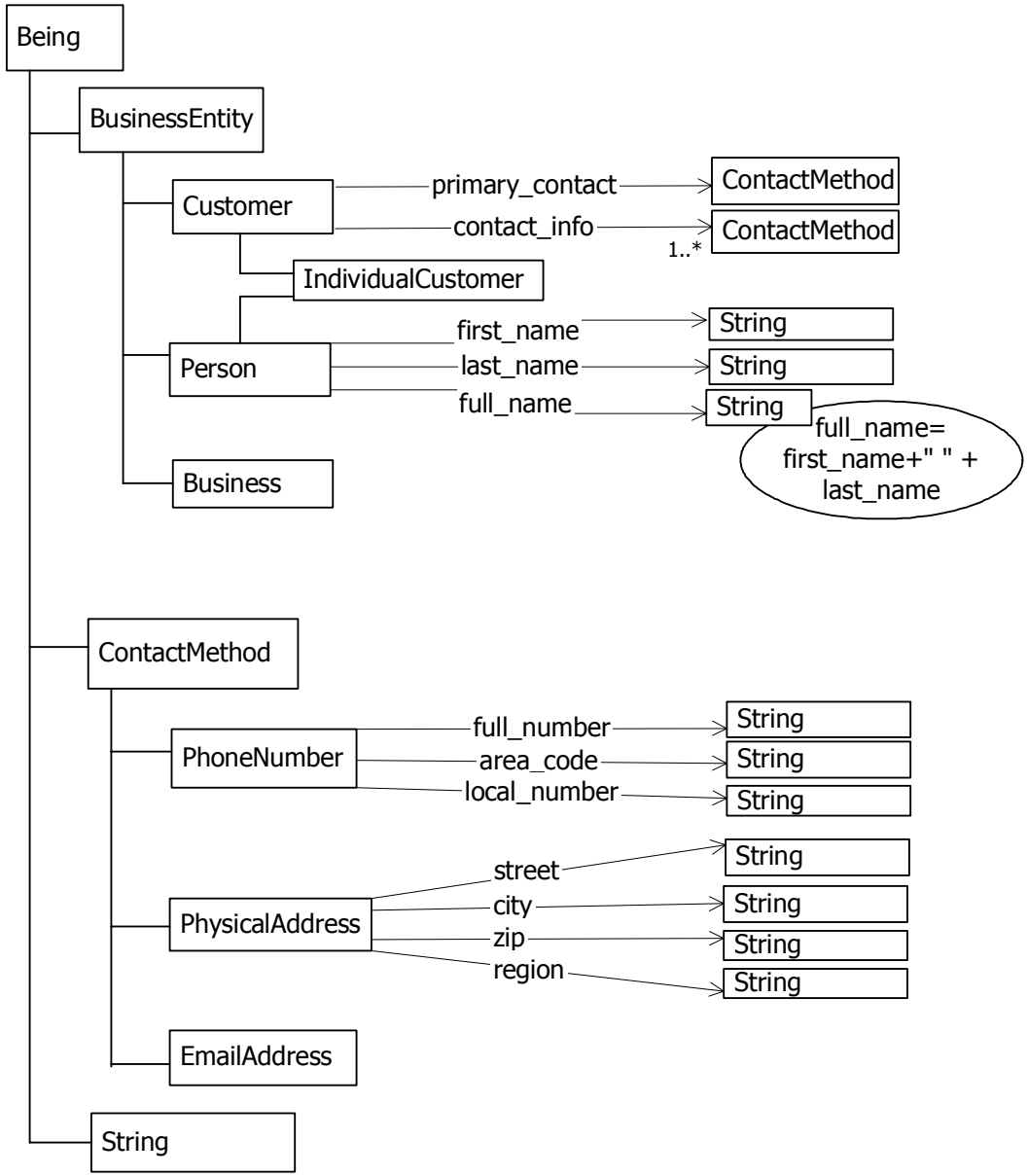


The Semantic Model

- Rich
- Central
- Active



Example Semantic Model

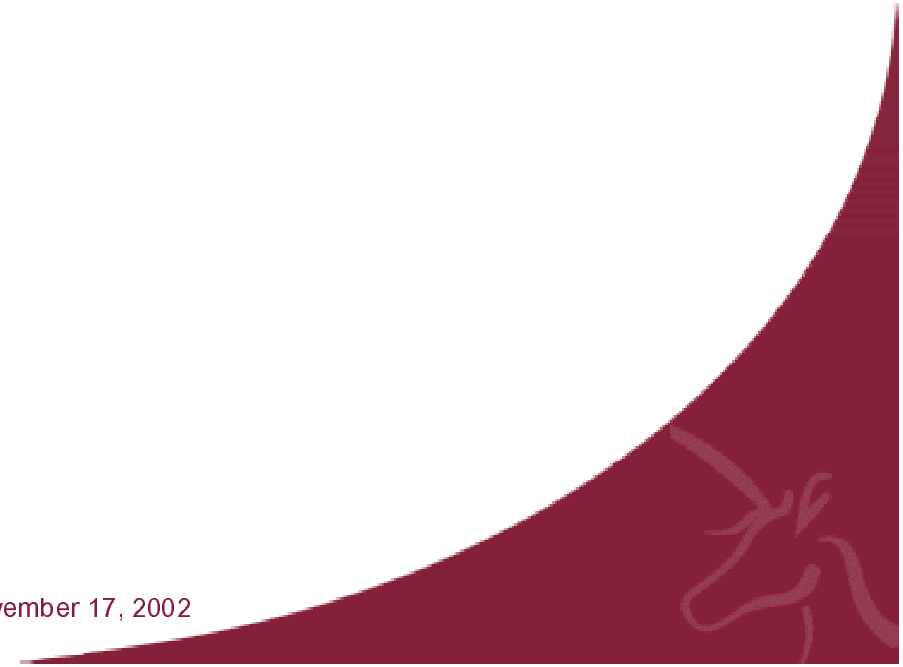


Rich Model

- Based on Ontology
  - The formalization of semantics of real-world entities
  - A science developed in academia for decades
  - Gained recent popularity in Tim Berners-Lee's Semantic Web

Classes

- Sets of business-domain real-life instances.
- Compare to OO classes or ER entities
- Inheritance



Properties

- Relationships between classes
- Relate one instance of class A to one or more instances of class B
- Can be defined as “Unique” (each instance of class B has at most one instance of A related to it)

Business Rules

- Constraining the value of a property, relating the value of properties

- E.g., enumeration: `State.abbreviation` must be one of `"AK", "AL", "AZ"…`

- E.g., look-up table:

State name	abbreviation
Alaska	AK
Alabama	AL
Arizona	AZ

Business Rules

- E.g.,

```
Person.full_name =
    last_name + " " + first_name
```

- E.g.,

```
miles = km * 1.6
```

Making the Central Model into a Semantic Hub

Map metadata (schemas) to ontological concepts

- RDB schemas,
- XML Schemas (DTD, XSD)
- COBOL Copybooks
- ERwin models

Mapping Schema Concepts to Semantic Concepts

- Map Simple Types (Columns, Attributes) to properties in the model
- Map Complex Types (Tables, Entities, Groups) to classes in the model
- Map constraints and other logic as expressed in the schemas into Business Rules

Example Schemas

Schema: CRM System
DB Type: Oracle 8i
Hostname: Athena

Table 1: Individual Customer

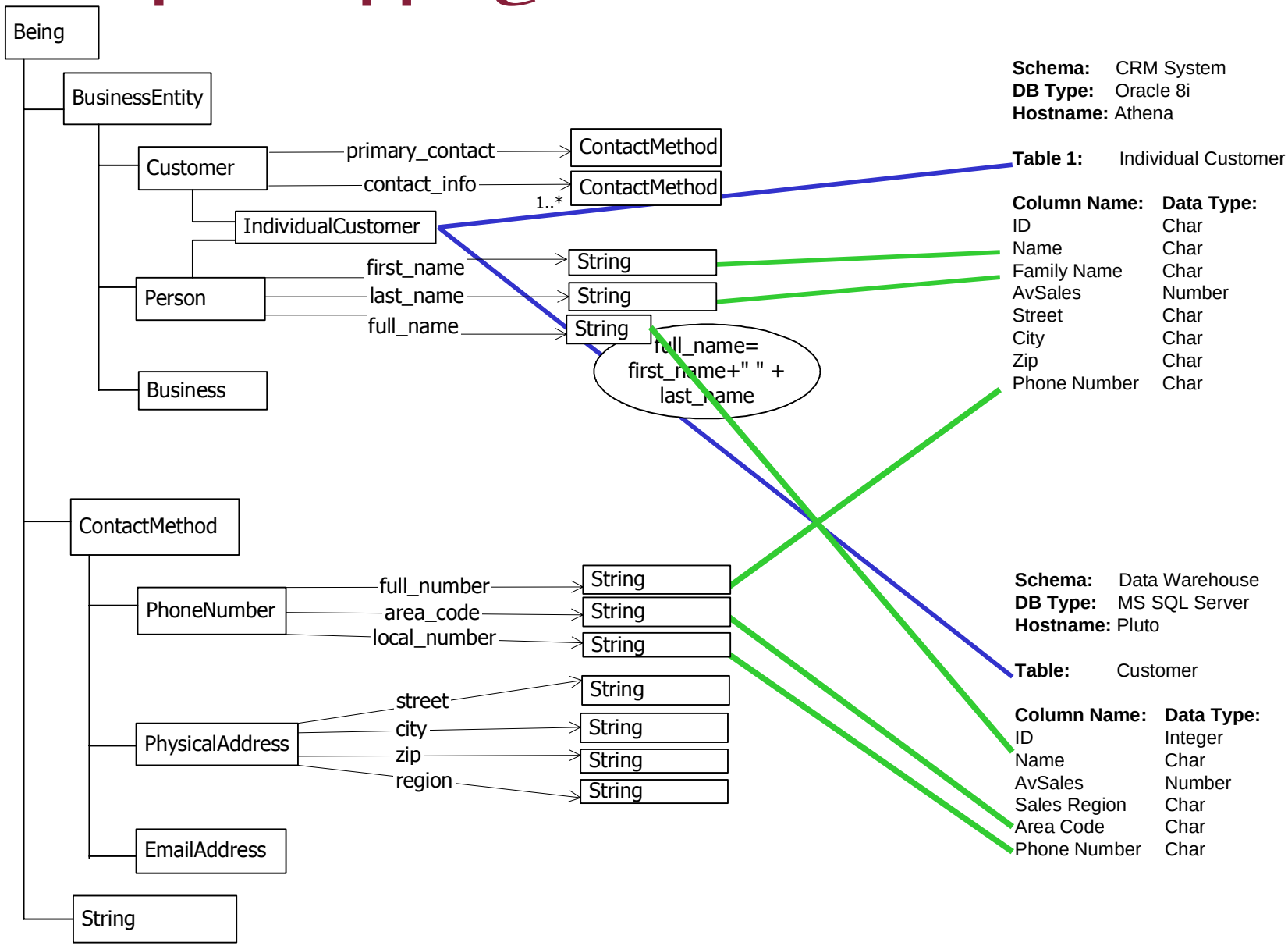
Column Name:	Data Type:	
ID	Char	Always a number, though in a string
Name	Char	Actually just first - name
Family Name	Char	
AvSales	Number	Actually meant as yearly sales
Street	Char	
City	Char	
Zip	Char	
Phone Number	Char	

Schema: Data Warehouse
DB Type: MS SQL Server
Hostname: Pluto

Table: Customer

Column Name:	Data Type:	
ID	Integer	Always a number, though in a string
Name	Char	Note double use of “name” ; full name calculated from first & last names
AvSales	Number	Actually monthly sales, YearlySales /12
Sales Region	Char	(one of {NW, NE, S, SW, W} , based on zip)
Area Code	Char	(first 3 digits of full number)
Phone Number	Char	(actually only the local part, remaining digits of full number)

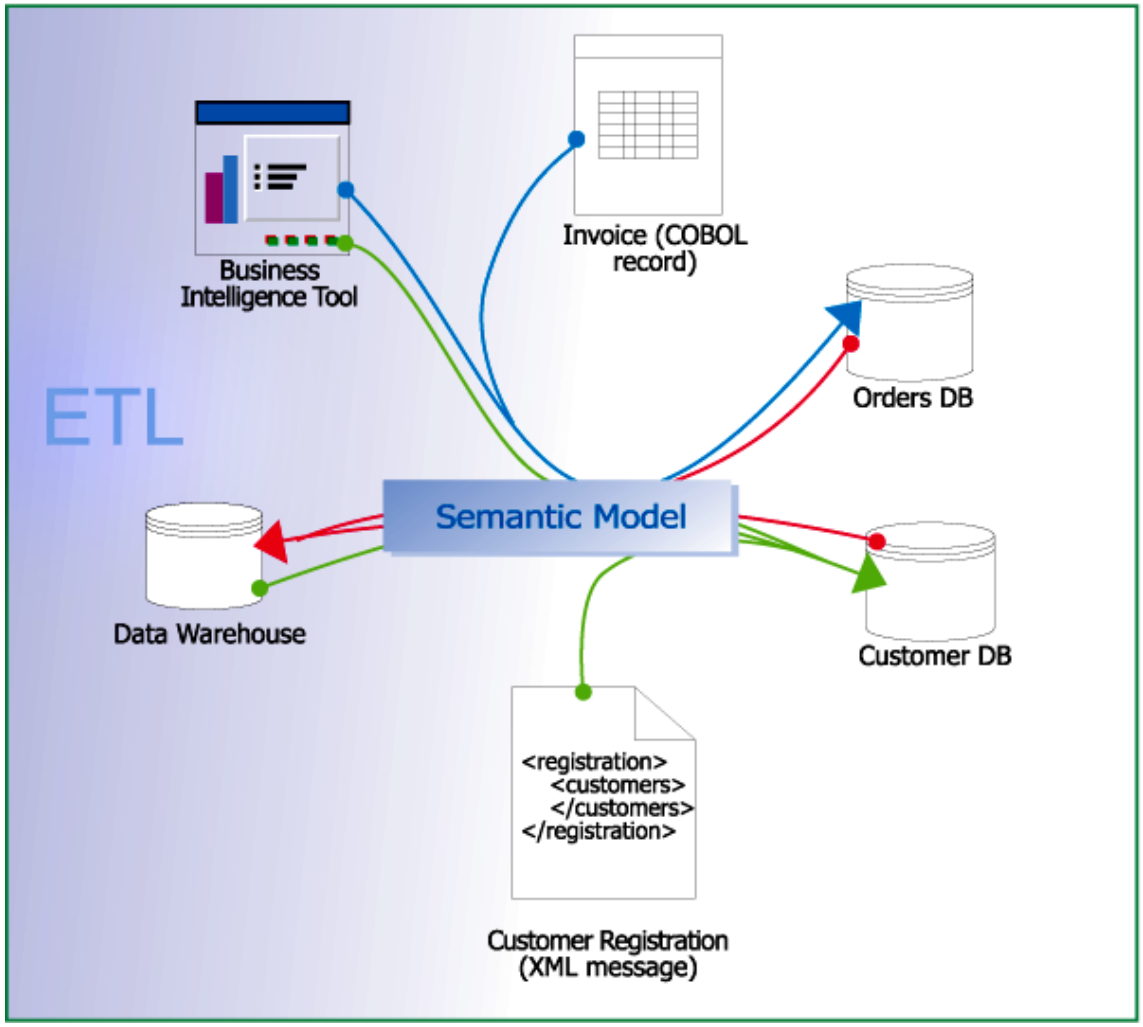
Example Mapping



An Active Model

- Active model: Not just a presentation of semantics, the active model *does* things for us.
- Once the model has been created and mapped to metadata, we have a formal encoding of everything needed to automatically generate code (SQL, XSLT, etc.) to
  - Query information on a semantic concept, wherever it resides in the enterprise
  - Transform data from one schema to another

Applications of Active Model

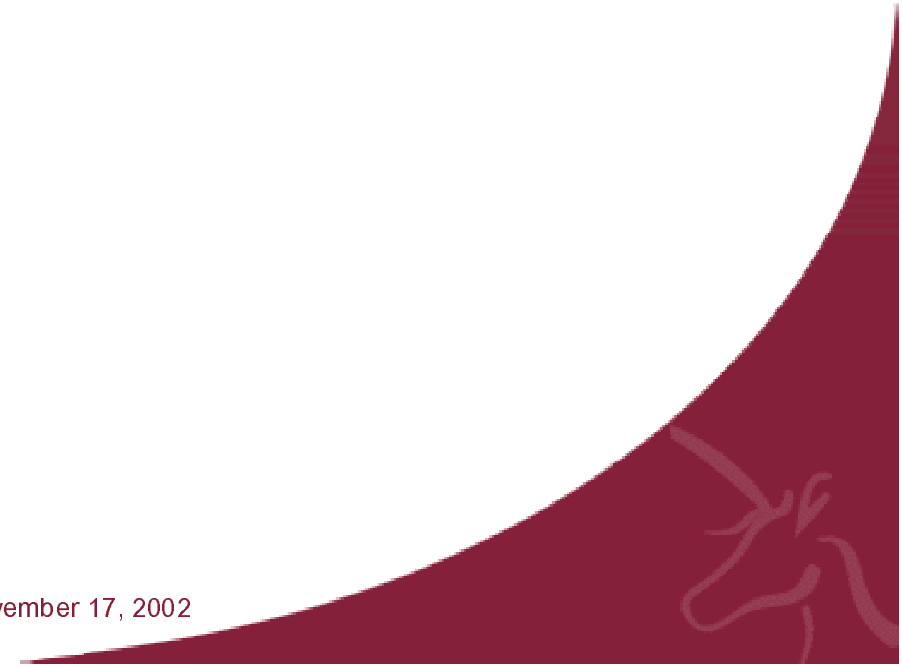


The Process

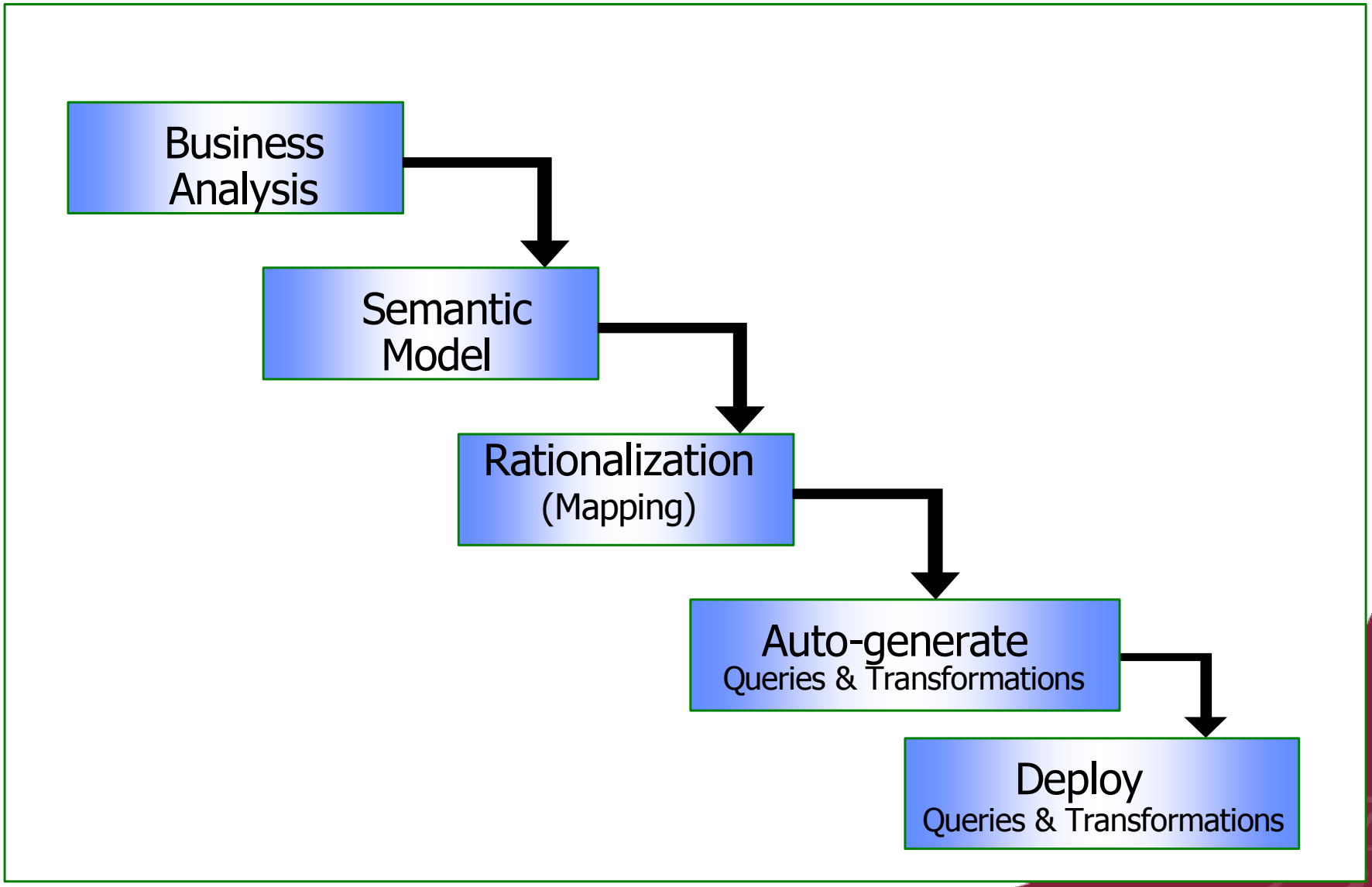
Integrating IQ into the Enterprise

Development Process

- Analysis
- Modeling
- Mapping
- Deployment



Process Overview



Conclusion

- The rich semantic model helps implement business IQ coherently across the enterprise
- Disjointed *data* is transformed into meaningful *information*.

Feedback

Joshua Fox

joshua@unicorn.com

http://www.unicorn.com

tel. 1-866-2-UNICORN x115