

Using Ontologies for Data and Semantic Integration

Monica Crubézy

Stanford Medical Informatics, Stanford University

~ ~

November 4, 2003

Ontologies

- Conceptualize a domain of discourse, an area of expertise
 - Concepts (drug, patient, gene, clinical trial)
 - Properties, or attributes (dosage, age, location)
 - Relationships (contra-indications, body parts)
- Adhere to a modeling formalism, such as:
 - Frame-based representation
 - Description logics

Protégé

- A general-purpose environment for **ontology-editing** and **knowledge-base construction**
 - Open-source, freely available (protégé.stanford.edu)
 - Interoperable with standards for knowledge representation (OKBC, RDF/S and more recently OWL)
 - Extensible in many ways (GUI, plugins, storage)
- Main frame-based modeling constructs
 - **Classes** represent concepts, organized in hierarchy
 - **Slots** represent properties of classes, with restriction **facets** on their values (e.g., type, cardinality, range)
 - **Instances** represent individual members of a class, with particular values for slots
 - **Instance-valued slots** hold relationships with other concepts

GLIF: Ontology for Clinical Guidelines

The screenshot shows the Protegé 2.0 beta interface. On the left is a class hierarchy tree. The main window displays the details for the `Action_Step` class, which is a `STANDARD-CLASS`. The details include a name field, a documentation field, a role dropdown set to `Concrete`, and a table of template slots.

Name	Type	Cardinality	Other Facets
iteration_info	Instance	single	classes={Iteration_Specification}
exceptions	Instance	multiple	classes={Guideline_Exception}
strength_of_recommendation	Instance	single	classes={Strength_Of_Evidence_Or_Recommendation}
triggering_events	Instance	multiple	classes={Triggering_Event}
tasks	Instance	multiple	classes={Action_Specification}
next_step	Instance	single	classes={Guideline_Step}
duration_constraint	Instance	single	classes={Duration_Interval}
didactics	Instance	multiple	classes={Supplemental_Material_List}
strength_of_evidence	Instance	single	classes={Strength_Of_Evidence_Or_Recommendation}
name	String	single	

Class hierarchy

List of slots for class Action_Step

GLIF: An instance of Action_Step

Automatically-generated
instance-knowledge
entry form

Specific values
fill slots

X-Ray (type=Action_Step, name=CoughStudy_00091)	
Name	Triggering Events
X-Ray	
Tasks	V C + -
◆ Chest X-Ray	
Next Step	V C + - Exceptions
◆ Treatment of cough	
Iteration Info	V C + -
Didactics	V C + - Strength Of Evidence
◆ X-Ray should be ordered in nearly all patients with chron	◆ Grade II-2

Ontologies for Data Integration

1. Hold reference/standard models and data repositories (e.g., the GLIF ontology)
 - Existing examples speak for themselves
2. Integrate data, metadata, and semantics of multiple data sources
 - A template ontology approach
3. Enable reconciliation and translation of data between different models
 - An ontology-mapping approach

1. (Standard) Ontologies in Biomedicine

- Pervasive
 - From controlled terminologies to full-blown ontologies
 - Across the entire scope from biology to medicine
- Many exemplars
 - Unified Medical Language System (UMLS)
 - Medical terminology and concept description (GALEN/OpenGALEN)
 - Foundational Model of Anatomy
 - Guideline models (GLIF, SAGE)
 - Gene Ontology (GO)
 - Pharmacogenomics ontology (PharmGKB)
 - ...

2. Integrating Data and Semantics

Syntactic differences

<sales="Robitussin">25</sales>
<sales="Pepto-Bismol">100</sales>

versus

Item	Sold
Robit.	25
PeptoB.	100

Differences are usually explicit,
but may be hard to reconcile.

Semantic differences

"Sales" means cases sold per week.
"Robitussin" means all Robitussin-
branded medication.

versus

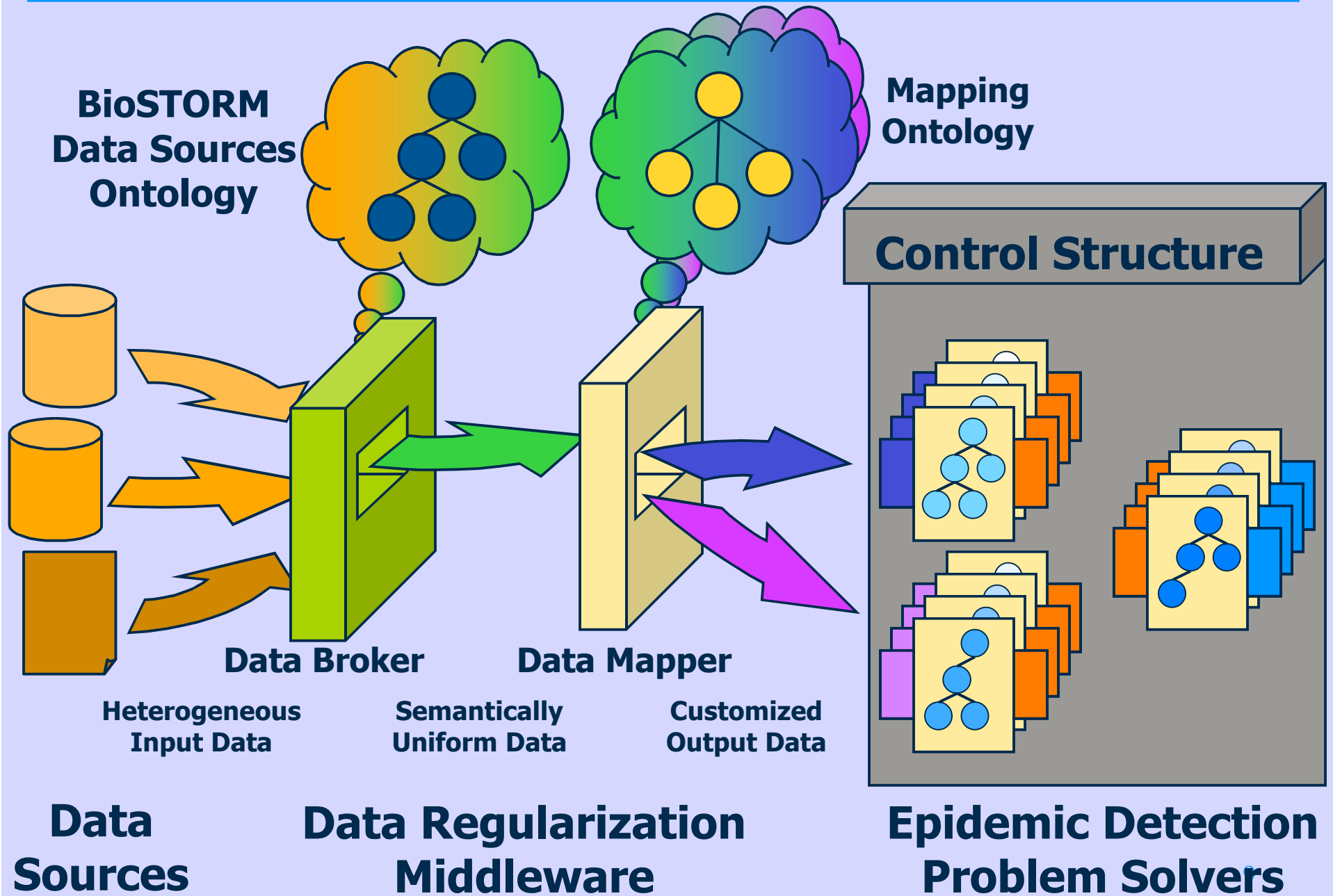
"Sales" is average number of
bottles sold per hour.
"Robitussin" only refers to
Robitussin DM.

Differences can be subtle and
implicit.

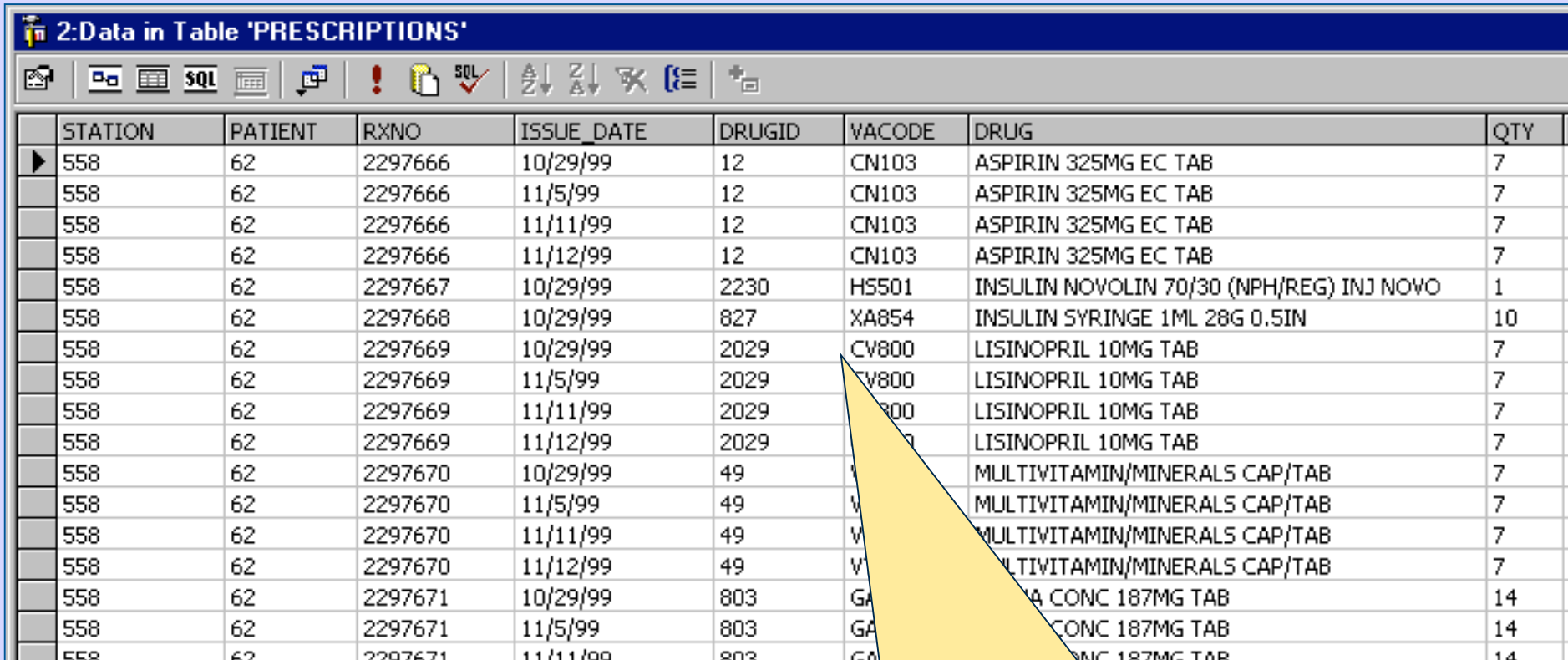
Integrating Data for Epidemic Detection

- The **BioSTORM** Project:
 - Biological Spatio-TempORal Module
 - Within DARPA-funded **BioALIRT** program for epidemics surveillance based on non-traditional, pre-diagnostic data
- Purpose:
 - To federate diverse non-traditional data sources (e.g., ER visits, 911 calls, absenteeism reports, pharmacy sales)
 - To enable space/time analysis of data by various computational methods, for early epidemics detection

Integrating Data for Epidemic Detection



Veterans Affair Data



STATION	PATIENT	RXNO	ISSUE_DATE	DRUGID	VACODE	DRUG	QTY
558	62	2297666	10/29/99	12	CN103	ASPIRIN 325MG EC TAB	7
558	62	2297666	11/5/99	12	CN103	ASPIRIN 325MG EC TAB	7
558	62	2297666	11/11/99	12	CN103	ASPIRIN 325MG EC TAB	7
558	62	2297666	11/12/99	12	CN103	ASPIRIN 325MG EC TAB	7
558	62	2297667	10/29/99	2230	H5501	INSULIN NOVOLIN 70/30 (NPH/REG) INJ NOVO	1
558	62	2297668	10/29/99	827	XA854	INSULIN SYRINGE 1ML 28G 0.5IN	10
558	62	2297669	10/29/99	2029	CV800	LISINOPRIL 10MG TAB	7
558	62	2297669	11/5/99	2029	CV800	LISINOPRIL 10MG TAB	7
558	62	2297669	11/11/99	2029	CV800	LISINOPRIL 10MG TAB	7
558	62	2297669	11/12/99	2029	CV800	LISINOPRIL 10MG TAB	7
558	62	2297670	10/29/99	49	WV800	MULTIVITAMIN/MINERALS CAP/TAB	7
558	62	2297670	11/5/99	49	WV800	MULTIVITAMIN/MINERALS CAP/TAB	7
558	62	2297670	11/11/99	49	WV800	MULTIVITAMIN/MINERALS CAP/TAB	7
558	62	2297670	11/12/99	49	WV800	MULTIVITAMIN/MINERALS CAP/TAB	7
558	62	2297671	10/29/99	803	GA800	GA CONC 187MG TAB	14
558	62	2297671	11/5/99	803	GA800	GA CONC 187MG TAB	14
558	62	2297671	11/11/99	803	GA800	GA CONC 187MG TAB	14

Several relational tables
Large space of data values
Semantics known to database creators

911 Emergency Call Data

2:Data in Table 'Calls1999'

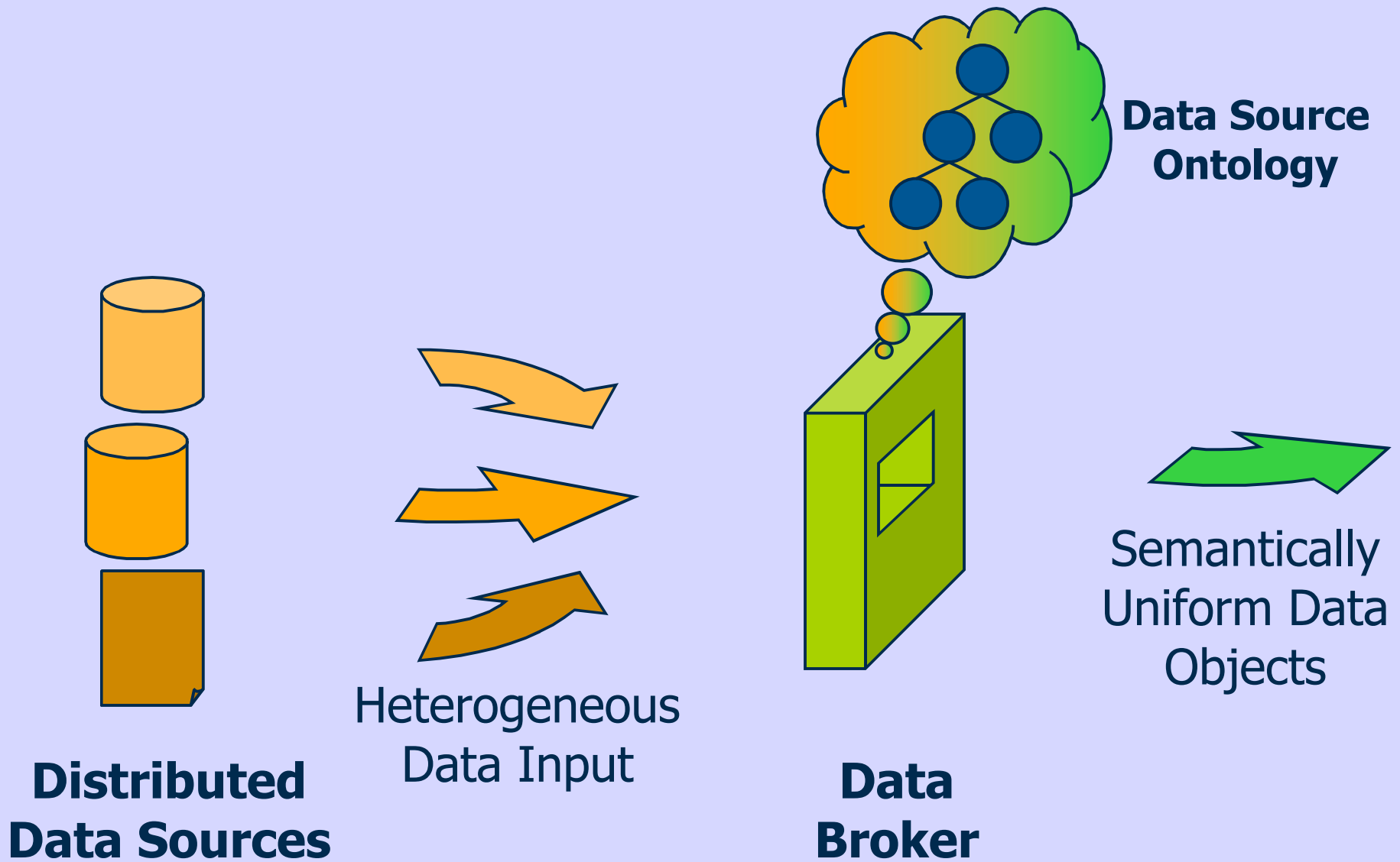
	MATCH_STATUS	MATCH_SCORE	INCIDENT_NUM	DATE	CALL_URGENCY	CALL_TYPE_MAIN	CALL_TYPE_MODIFIER	DISPOSITION	ZIP
	M	95	F99000001	1/1/99	M	20	10	CAN	94102
	M	95	F99000006	1/1/99	M	21	7	CAN	94102
	U	0	F99000008	1/1/99	M	20	3	PDT	<NULL>
▶	M	100	F99000009	1/1/99	R	20	5	PDT	94109
	M	100	F99000010	1/1/99	M	19	3	DMC	94114
	T	84	F99000011	1/1/99	M	7	4	CPP	94103
	M	100	F99000013	1/1/99	M	20	3	CAN	94115
	M	100	F99000015	1/1/99	R	23	1	PDT	94118
	M	100	F99000016	1/1/99	M	20	8	SFG	94110
	T	90	F99000017	1/1/99	M	14	3	SPD	94110
	M	100	F99000021	1/1/99	M	20	5	MTZ	94111
	M	100	F99000023	1/1/99	M	20	3	PDT	94102
	M	100	F99000024	1/1/99	R	5	1	PDT	94102
	M	100	F99000025	1/1/99	R	17	1	CAN	94102
	T	90	F99000026	1/1/99	Y	10	2	CAN	94114
	M	100	F99000028	1/1/99	M	19	3	STF	94102
	M	100	F99000030	1/1/99	M	24	4	SFG	94112
	M	100	F99000032	1/1/99	M	23	3	SFG	94118

**One table in a relational database
Constrained space of data values
Arbitrary and unclear semantics**

Data Integration Approaches

- **Integration of explicit local models of each source**
 - Database schema matching and query distribution
 - Ontology merging, alignment & integration
- **Description of data sources using a single global model of entire domain of knowledge**
 - SIMS (ISI): tie multiple DBs with rich semantics & construct complex queries
 - TAMBIS (U. Man.,UK): represent, access & query multiple molecular biology DBs
 - caBIO (NCI): model cancer biology & provide methods to query remote DBs transparently

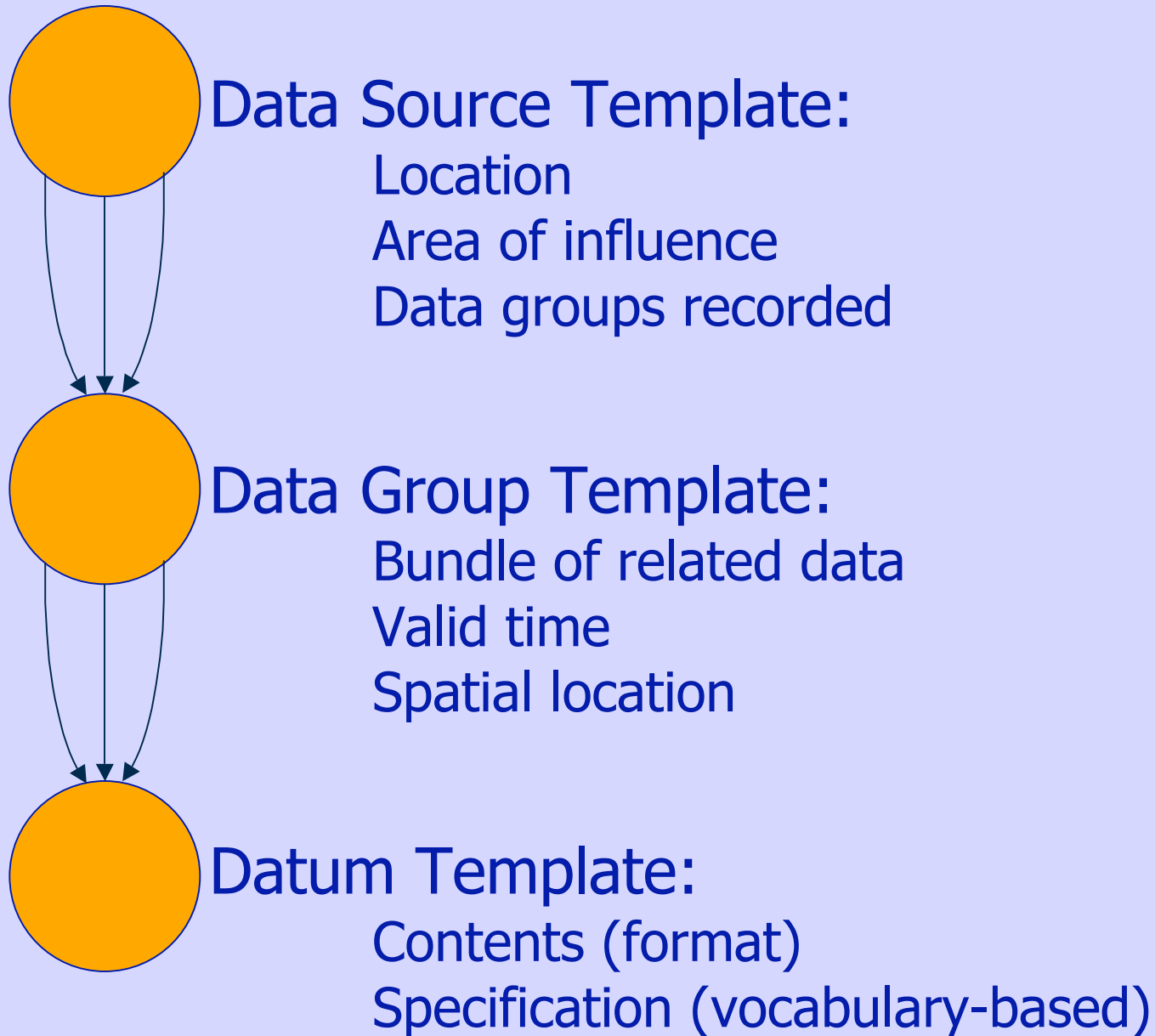
A Template Data Source Ontology



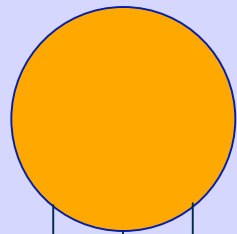
A Template Data Sources Ontology

- A template ontology for contextualizing diverse data sources
 - Hybrid of local and global approaches
 - Extensible & customizable framework for describing data and their context in a way they can be compared and operated on homogeneously
- Rationale
 - Require minimal ontological commitment of data sources
 - Preserve richness of data sources & flexibility in data use
 - Introduce no bias to data integration (left to analytical methods)
 - Ensure semantic uniformity of heterogeneous data

Template-based Approach



SF 911 Data Source Ontology

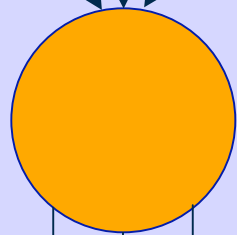


SF 911 Dispatch Center

Located at Hunter's Point

Receives Data from Greater SF

Receives "911 Call" Data

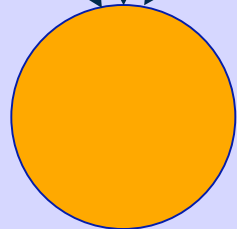


911 Call

Contains: "Call Urgency", "Call Type,"
"Call Disposition," etc.

Valid on a specified date

Call occurred at a specified location



Call Type

Contents: string

Specification: Semantics of the string

The Template Data Sources Ontology

The screenshot shows the Protégé-2000 interface for the 'Biostorm SAIL 3-0' ontology. The left pane displays a class hierarchy starting with ':THING' and ':SYSTEM-CLASS', with 'Data_Provider' selected. A yellow arrow points from 'Data_Provider' in the hierarchy to the right pane. The right pane shows the 'Data_Provider' class details, including its name, documentation, and a table of template slots.

Relationship Superclass

- :THING^A
- :SYSTEM-CLASS^A
- Data_Provider^A**
 - School
 - Employer
 - Pharmacy
 - Emergency_Room
 - Individual_Patient
 - School_District
 - Computer_Simulation
 - Insurance_Claims
 - Research_Group
 - Emergency-911_Call_Center
 - Hospital
 - Environmental_Measurement
- Data_Group_Specification
- Datum_Specification
- Data_Group
- Datum^A
- Datum_Specification_Component_Classes
- Utility_Classes^A

Name

Data_Provider

Documentation

A Data Provider is an entity that produces data. Specifically, a "Data Provider" is expected to produce a stream of "Measurement" objects. The slots below provide a location to enter appropriate metadata necessary to contextualize a data provider and provide a list of the measurements provided. Click the "+" button to add a new piece of metadata to the description of the data source. Double-click on a slot to see more information.

Template Slots

Name	Type	Cardinality
Specifications_of_Data_Groups_Produced ^o	Instance	multiple
Measurement_Stream ⁱ	Instance	multiple
Data_Provider_Contacts	Instance	required multiple
Data_Provider_Unique_Identifier	String	required single
Data_Provider_Name	String	required single
Parent_Data_Provider ⁱ	Instance	single

**Classes of
Data Sources**

An Instance of Data Source

The screenshot shows a software window titled "Menlo Park VA (VA_Hospital)". It contains several sections:

- Data Provider Name:** Menlo Park VA
- Data Provider Unique Identifier:** Stanford-002
- Specifications of Measurements Produced:** A list of five data groups: (Stanford) VA Conditions Data Group, (Stanford) VA Demographics Data Group, (Stanford) VA Encounters Data Group, (Stanford) VA Prescriptions Data Group, and (Stanford) VA Vitals Data Group. A yellow arrow points from this list to a callout box.
- Data Collection Location (1 values):** A section with a "Location Name" field containing "Address of Menlo Park VA". Below it are fields for "Street Address" (245678), "Street Name" (Foothill Blvd), "City Name" (Palo Alto), and "Zip Code" (94305).
- Parent Data Provider:** Veterans' Affairs Hospital System

Associated set of Measurements (“data groups”)

An Instance of a Data Group

(Stanford) VA Prescriptions Data Group (Measurement_Specification)

Measurement Specification Name
(Stanford) VA Prescriptions Data Group

LOINC Term(s)

Name	Property_Measured	Kind-of-Property_Measu...	Time_Aspect_of_Measu...	Scale_of_Measurements
Rx_Number	Rx_Number	Real Number	Point_in_Time	Measured in Integers
Dispensed on Date	Dispense_Date	Date	Point_in_Time	Narrative Measurement
Drug ID	Drug_ID	Categorical Measure	Point_in_Time	Measured in Integers
Ingredient	Ingredient	Text	Point_in_Time	Narrative Measurement
Display Name	Display_Name	Text	Point_in_Time	Narrative Measurement
Quantity of Drug	Drug_Quantity	Count of Integers	Point_in_Time	Measured in Integers
Special Instructions Given	Special_Instructions_Give	Text	Point_in_Time	Narrative Measurement
Division Issuing Prescript	Division	Text	Point_in_Time	Narrative Measurement
Rx Expires on on Date	Expire_Date	Date	Point_in_Time	Narrative Measurement
Severity	Severity	Text	Point_in_Time	Narrative Measurement
Affected Medications	Affected_Med		Point_in_Time	Narrative Measurement
Last Refill Date	Last_Fill_Date		Point_in_Time	Narrative Measurement
Days Supplied	Days_Supply		Point_in_Time	Measured in Integers
Daily Dose	Daily_Dose	Re	Point_in_Time	Measured in Real Numbe
Station Number	Station	Row	Point_in_Time	Measured in Integers
Fill Type	Fill_Type			
Provider Name	Provider_Name			Measurement

Associated LOINC-based vocabulary and specification of properties

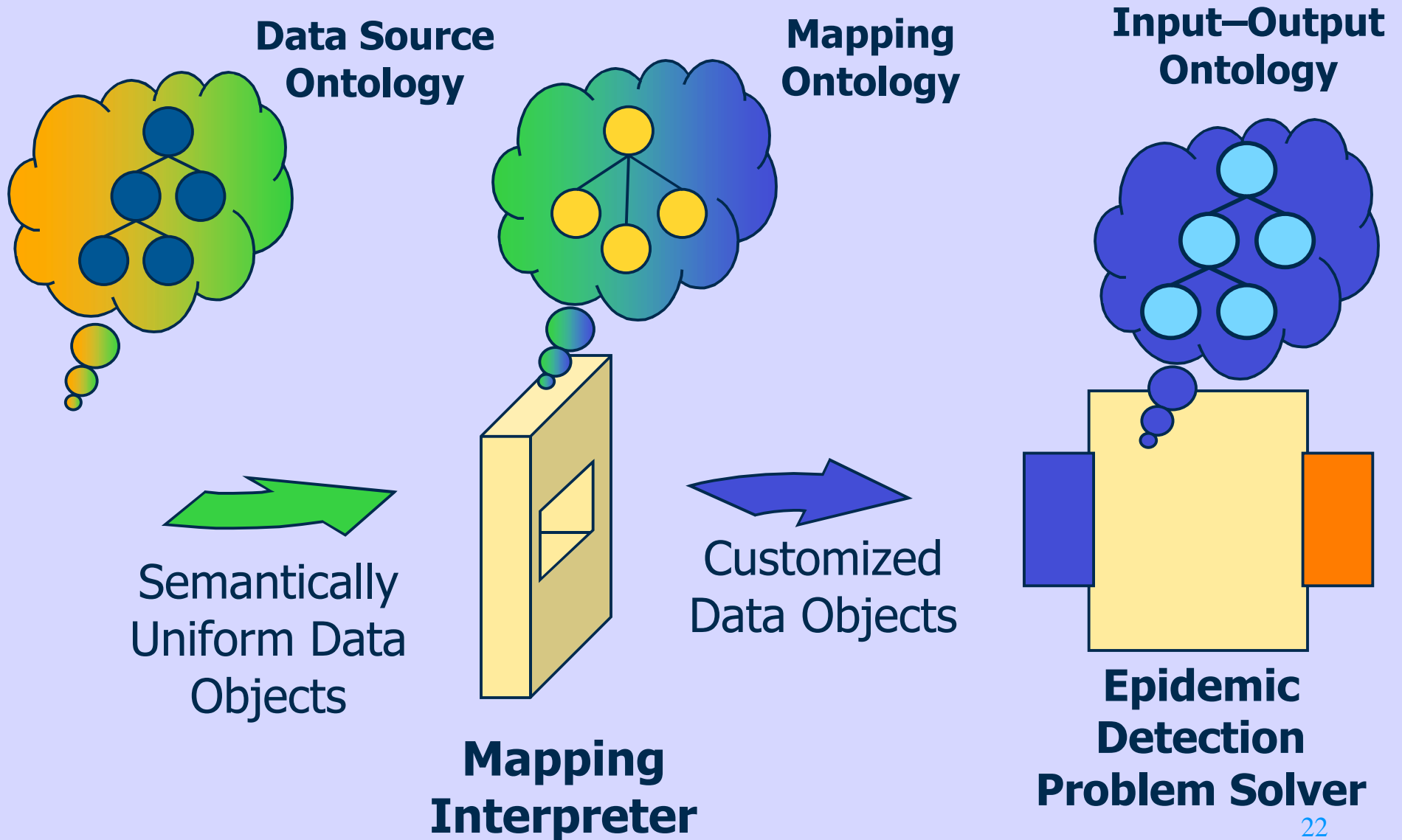
Providing Uniform Context to Data

- **Semantics**
 - Common language for describing and comparing surveillance data sources, for which no standards currently exist
 - Extensible framework for incorporation of new data sources
- **Metadata**
 - Shared repository for enumerating available data sources in machine-processable form
 - Explicit and extensible vocabulary consistent with LOINC standard for describing attributes of data and sources
- **Data**
 - Storage as instances of the ontology, OR
 - Definition of how data can be accessed from data sources

3. Reconciling Diverse Ontologies

- Many ontologies in biomedicine are federated models that fully or partially resemble standardization efforts
- But:
 - It is hard to agree on reference ontologies
 - We cannot expect people to adopt them (in the course of defining the standard, and even after)
 - Various reference and proprietary models need to interact in component-based architectures
- So, tools are needed to align different models and translate data represented in a given model to and from another model

Operating on Data in Multiple Ways



Conceptual and Syntactic Mismatch

Notion of a "Data Group"

S	recordedTemporalData	Instance	classes={TemporalDatum}
S	dataGroupSpecification	Instance	classes={DataGroupSpecification}
S	uid	String	
S	recordedSpatialData	Instance	classes={SpatialDatum}
S	originatingDataProvider	Instance	classes={DataProvider}
S	recordedLOINCData	Instance	classes={LOINCDatum}

Notion of an "Individual Event"

S	validEvent	Boolean	
S	date	Instance	classes={TimePoint}
S	dataSource	Class	parents={DataSourceType}
S	illnessCategory	Class	parents={IllnessCategory}
S	location	Instance	classes={GeoReference}

Conceptual and Syntactic Mapping

Notion of a "Data Group"

S	recordedTemporalData	Instance	classes={TemporalDatum}
S	dataGroupSpecification	Instance	classes={DataGroupSpecification}
S	uid	String	
S	recordedSpatialData	Instance	classes={}
S	originatingDataProvider	Instance	classes={}
S	recordedLOINCData	Instance	classes={}

- filter out invalid events
- extract & reformat source, date, location
- abstract illness category
- drop uid

Notion of an "Individual Event"

S	validEvent	Boolean	
S	date	Instance	classes={TimePoint}
S	dataSource	Class	parents={DataSourceType}
S	illnessCategory	Class	parents={IllnessCategory}
S	location	Instance	classes={GeoReference}

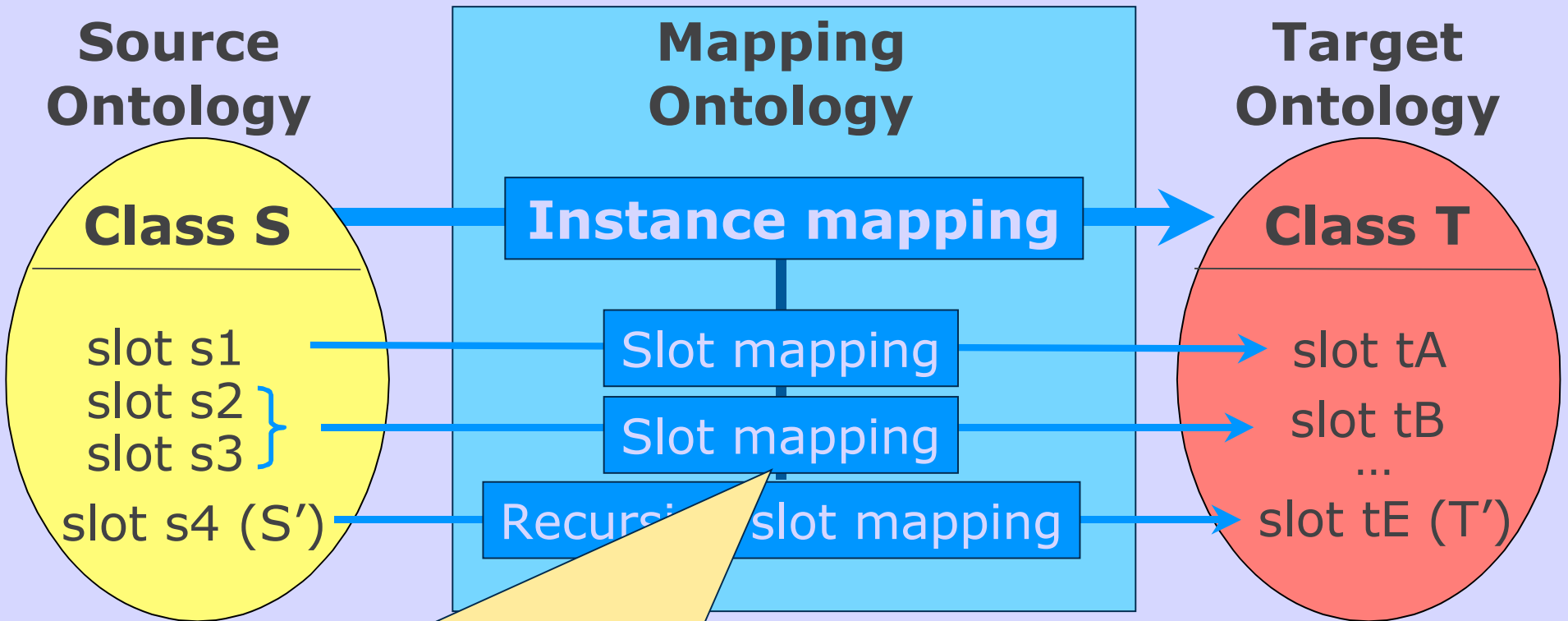
Ontology Mapping for Data Exchange

- Conceptual alignment
 - change in domain of discourse
 - difference in the level of knowledge granularity
 - split and join of concepts & attributes
- Value transformation
 - abstraction, reduction
 - aggregation or dispatch
 - format change (unit change)
 - custom computation (functional transformation)

Explicit Mapping Relations

- Isolate connections between ontologies
 - Each component ontology remains unchanged
 - Mapping relations express concept-level and attribute-level correspondences
 - Components focus and operate on their own view, format of knowledge & data
- Define mediation of data between ontology-based components
 - Mapping relations include the specification of rules of transformation of values
 - Components do not have to handle knowledge transformation internally

An Ontology of Mapping Relations



- renaming: $\text{value}(tA) = \text{value}(s1)$
- constant: $\text{value}(tD) = \text{constant}$
- lexical: $\text{value}(tB) = \text{"*}<s2>* / 20*<s3>*"}$
- functional: $\text{value}(tE) = \text{function}()$
- recursive: $\text{value}(tC) = \text{instance (auxiliary instance-mapping)}$

Mapping Data Groups to Individual Events

**instance
mapping**

**constant slot
mapping**

Source-class V C + -

◆ DataGroup

Apply mapping to subclasses of source-class?

Target-class V C + -

◆ IndividualEvent

Slot-maps V C + -

◆ to-validEvent
◆ SF911_recordedLOINCDData-to-illnessCategory
◆ SF911_recordedSpatialData-to-location
◆ SF911_recordedTemporalData-to-date
◆ SF911_to-dataSource

Condition

```
<LANG:Python>slotValueIsOfClass("originatingDataProvider", "Emergency911CallCenter") and  
"*<originatingDataProvider.dataProviderName>*" == "San Francisco 911 EMS Dispatch" and  
isValidSF911Record()
```

Reverse-mapping

On-demand

Slot-map-name

SF911_to-dataSource

Const-val

Dispatch911

Target-slot V C + -

◆ dataSource

Mapping Data Groups to Individual Events

recursive slot mapping

Slot-map-name
SF911_recordedSpatialData-to-location

Source-slot V C + -
recordedSpatialData

Target-slot V C + -
location

Mappings V C + -
SF911-SpatialDatum-to-ZIPCode

Subclasses-accept...

on-demand instance mapping

Source-class V C + -
SpatialDatum

Apply mapping to subclasses of source class

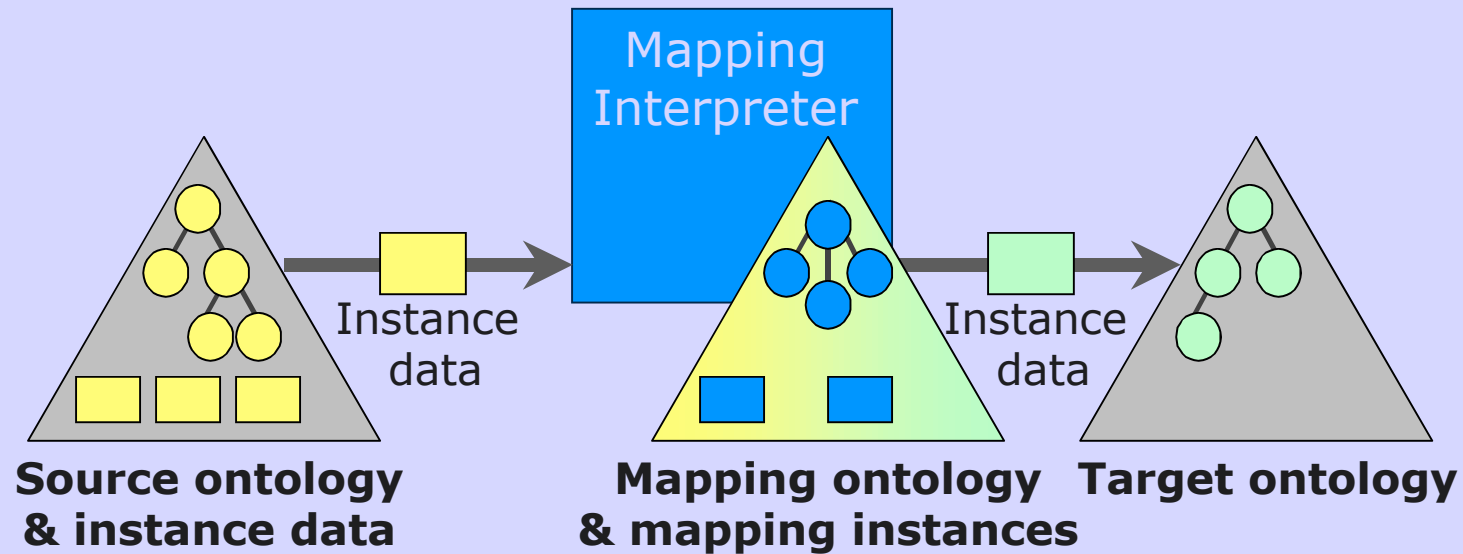
Target-class V C + -
ZIPCode

Slot-maps V C + -
zipCodes-to-identifier

Condition
<LANG:Python>isOfClass("ZIPCodeArea") and
"*<datumSpecification.name>*" ==
"911 Call Location (ZIP)"

Reverse-map... On-demand

Mapping Interpreter



- Processes the mapping relations between one or more source ontologies and a target ontology
- Produces a set of instances of the target ontology from the existing instances of the source ontology

Results of Mapping Interpretation

Source "Data Group" instance → Resulting target "Individual Event" instance

Unique ID: SMI-DATA-001

Originating Data Provider: San Francisco 911 EMS Dispatch

Data Group Specification: 911 Call Record

Recorded LOINC Data

Type	Specification	Contents
StringDatum	Match Status	T
IntegerDatum	Match Score	90
StringDatum	Incident Number	F99000002
StringDatum	Call Urgency	M
IntegerDatum	Main Call Type	20
IntegerDatum	Call Type Modifier	10
StringDatum	Call Disposition	HEA

Recorded Spatial Data

Type	Specification
ZIPCodeArea	911 Call Location (ZIP)
CensusBlockGroups	911 Call Location (Block Group)

Recorded Temporal Data

Type	Specification
Datetime	Date of 911 Call

ValidEvent

DataSource: Dispatch911

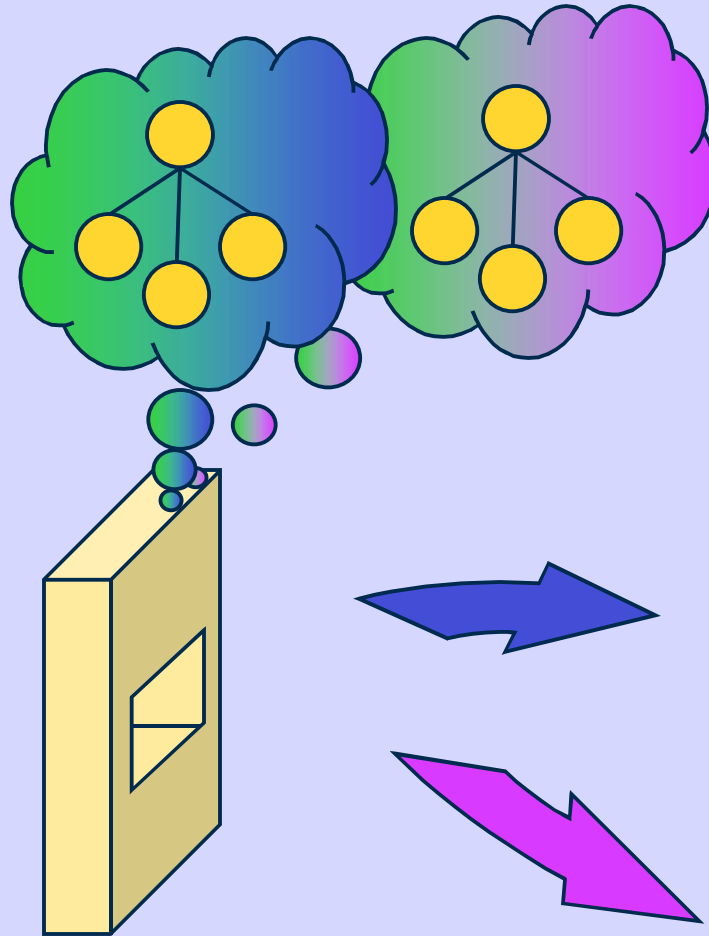
Date: 1999

Location: 94012

IllnessCategory: trauma

Varying Problem Solvers

**Mapping
Ontologies**



**Input–Output
Ontologies**

**Mapping
Interpreter**

**Problem
Solvers**

Benefits of Ontology-based Data Integration

1. Modeling data with ontologies
 - Provides rich, machine-processable semantics to data
 - Facilitates knowledge communication and sharing
2. Integrating data with a template ontology
 - Enables software components to operate on data in a uniform way
 - Facilitates access to existing data sources for any new customer component
 - Eliminates the need for customer components to be reprogrammed when a new data source is added

Benefits of Ontology-based Data Integration

- Integrating data models by ontology mapping
 - Isolates ontological connections and data-level transformations for instance migration
 - Enables flexible, interconnected, component-based architectures
 - Each component relies on its own ontology
 - Components remain independent
 - Component coupling is explicit and maintainable

Perspectives

- Data integration will always be needed!
 - Before standards are agreed upon and used
 - When information systems need to integrate and analyze multiple data sources
 - When system components need to access or rely on different ontologies
- Adaptations to be made for richer ways of modeling ontologies (DLs in particular)
- Combination with other data-integration approaches: matching, merging, alignment

Aknowledgements

- At Stanford Medical Informatics
 - Zachary Pincus
 - Samson Tu, Mor Peleg
 - Natasha Noy
 - Prof. Mark Musen
- Funding agencies
 - National Library of Medicine
 - National Institute for Standards and Technology
 - National Cancer Institute
 - Defence Advanced Research Project Agency

- Stanford Medical Informatics

<http://smi.stanford.edu>

- The Protégé project

<http://protege.stanford.edu>

- Monica Crubézy

<http://smi.stanford.edu/people/crubezy>
crubezy@smi.stanford.edu