Information Integration

(Still) An Achilles Heel of Computing

Joachim Hammer

Dept. of Computer & Information Science

University of Florida

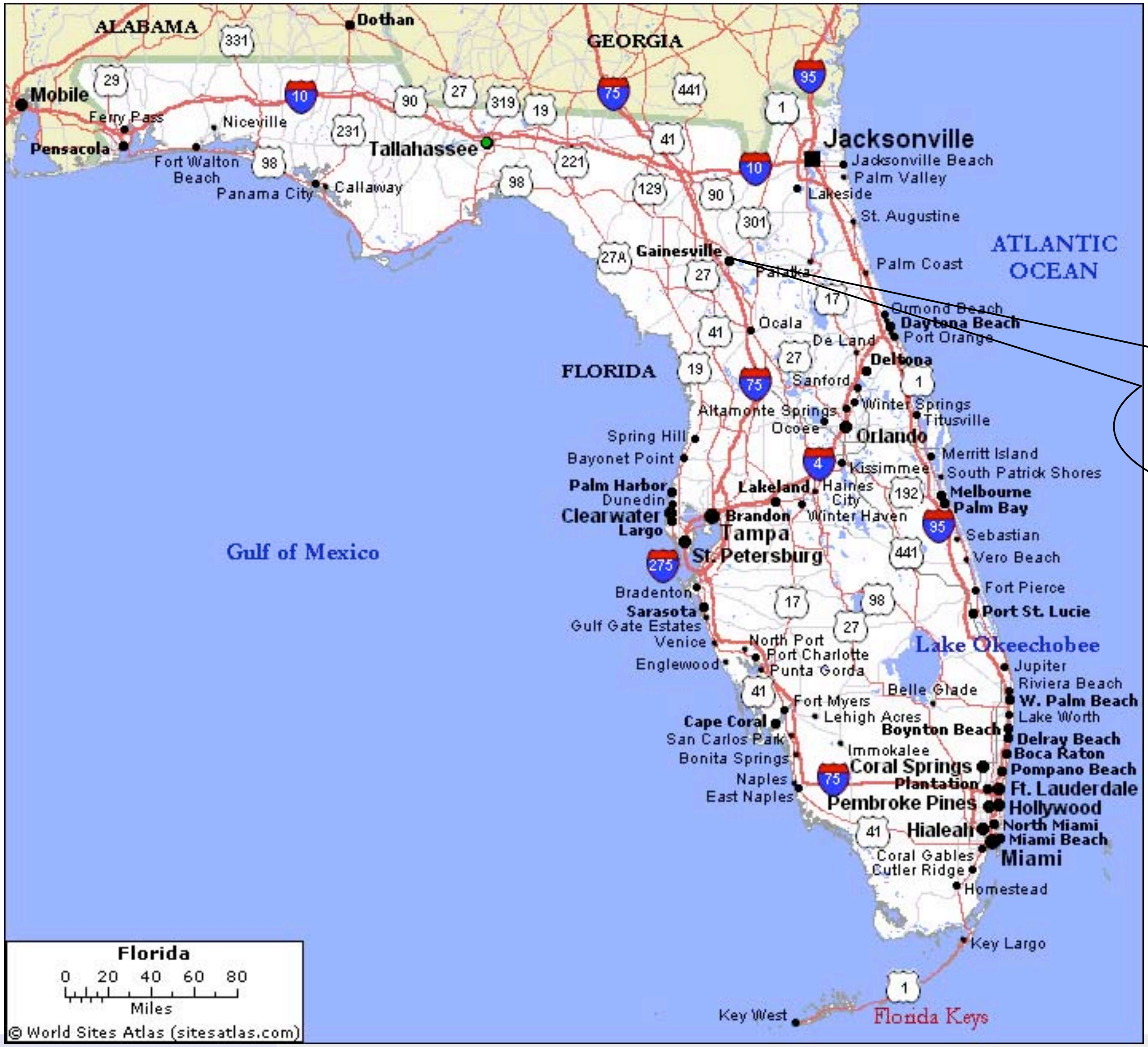
The University of Hong Kong
ECOM/ICOM Programme
August 3, 2006

Career Highlights

- Associate Professor (since 2003)
  - Dept. of CISE, University of Florida (since 1997)
  - Interim Director, Database Research & Dev. Center (since July 2005)
- Research Scientist (1994 - 1997)
  - Database Group, CS Dept., Stanford University
- Ph.D. & M.S., Computer Science & Applied Math (1994, 1990)
  - CS Dept., University of Southern California
  - Thesis: Resolving Semantic Heterogeneity in a Federation of Autonomous, Heterogeneous Database Systems
  - Advisor: Prof. Dr. Dennis McLeod

- Visiting Professor, Center for Computing Technologies (TZI) at the University of Bremen (Sept. 2004 - June 2005)
- Visiting Professor, CS Dept., The University of Hong Kong (since Sept. 2001)

Where is Gainesville, FL?



Gainesville, FL
pop.: 240,000

University of Florida

- Founded in 1860 as the state’s land-grant *Florida Agricultural College*
  - Became University in 1906
- Went from 102 students in 1860 to 46,000 students in 2005
  - Among the five largest universities in the US
  - Oldest, largest and most comprehensive University in FL
- 900+ buildings (including 170 with classrooms and laboratories) occupying 2,000-acre campus
  - Including residence halls for 7,000 undergraduate and 2,200 graduate students with their families

Dept. of Computer & Information Science & Engineering

- 48 faculty members (all ranks)
- ~ 1,100 undergraduate and graduate students seeking Bachelor's, Master's, and Ph.D. degrees
- Areas of strengths
  - Computer Graphics
  - Database and Information Systems
  - High-Performance Computing
  - Computer Vision

Database Research & Development Center

- University center affiliated with the Department of CISE
  - Funded entirely through research grants brought in by participating faculty
  - State-of-the art,1000+ sq. ft lab housing workstations and multi-processor file/compute servers for specialized system development and prototyping work
  - 100 Mbps fiber optics local area network with fiber connections to a campus-wide backbone network
- Members
  - **Faculty:** Alin Dobra, Joachim Hammer (interim director), Chris Jermaine, Tamer Kahveci, Markus Schneider, Stanley Su (Prof. emeritus)
  - **Students:** ~ 30 Ph.D. and M.S., 5-10 undergraduates (senior projects)

Rest of Talk

- Information Integration
  - Motivation - Problem Description - Challenges
- State-of-the-Art
  - Core technologies (success stories)
  - Areas of continuing research
- Morpheus Data Transformation Project at UF
  - Goals
  - Current state
- Summary and Future Directions

Motivation

- Two enterprises have agreed to merge…

Employee Database for Company A, headquartered in Canada

EmpID	Name	Total Compensation	Shares
99999	Last, First	999k	9999999

# of shares owned

Canadian $$
after tax
lunch allowance

Employee Database for Company B, headquartered in US

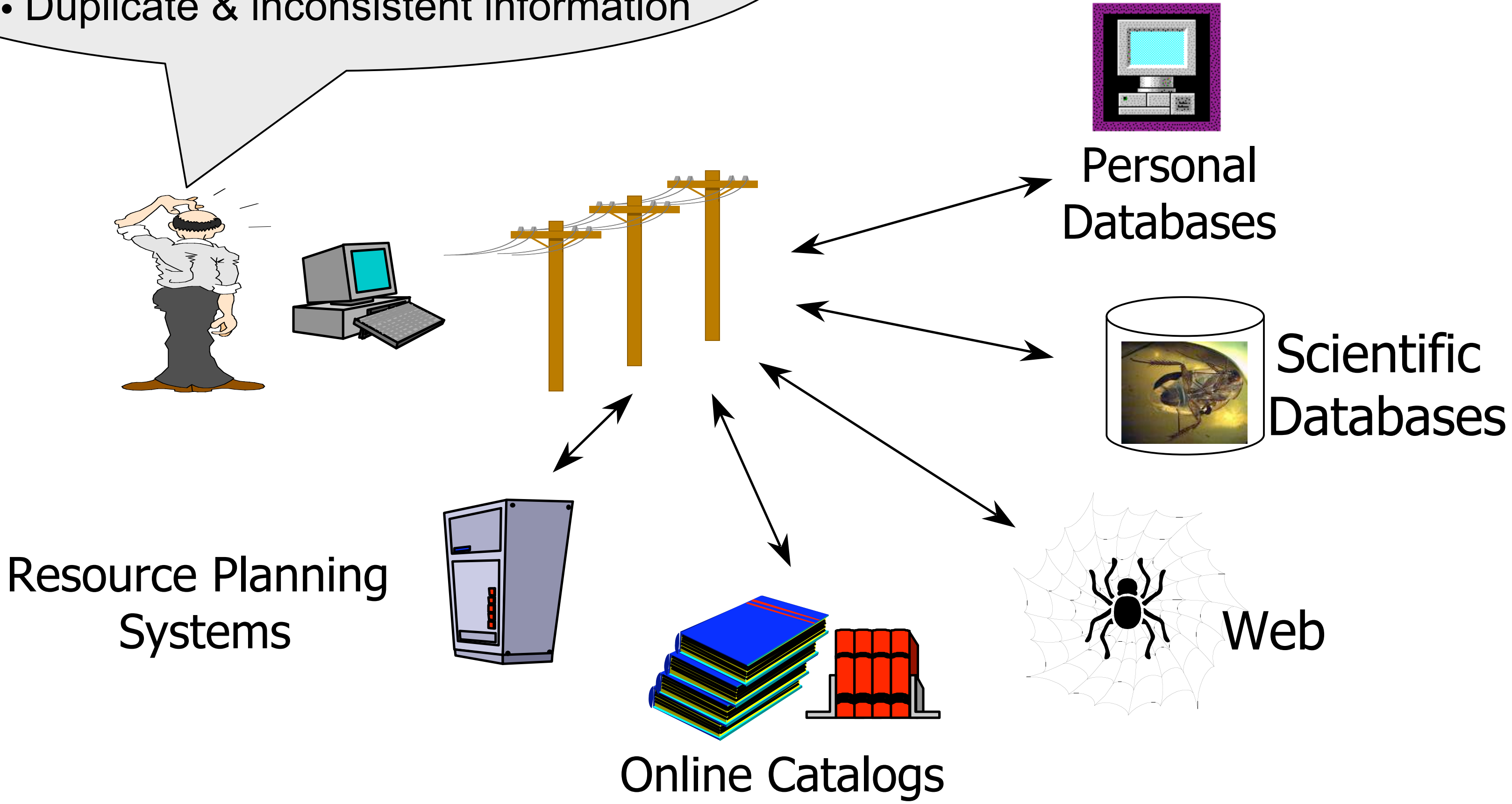
SS#	First	MI	Last	Wages	Stock Options
99999	First	MI	Last	99k	9999999

total # of shares
including options

US $$
before tax

Integration Problems are Everywhere…

- Different interfaces
- Different data representations
- Duplicate & inconsistent information



Application Areas that are Driving Research (Funding)

- Emergency Management
  - Emergency response planning
  - Damage assessment
- Homeland Defense
  - Threat prediction and detection
  - Coalition forming
- Extended Enterprise/Supply Network
  - Decision/negotiation support to improve performance and customization
  - Support for autonomous, cooperating logistics processes
- Customer Modeling/Validation/RM
  - What are the characteristics of people who go elsewhere?

Some more Observations

- Increasingly difficult for users to leverage information resources despite the supposed availability of data
  - Data globally dispersed, often over wide area
    - Within and across organizations
  - Little will or ability to share the data
    - Privacy/security concerns
    - People are grumpy/lazy/petty
  - Difficult to make sense of data that is being retrieved
    - No semantics
    - Human query processor
  - Lots of new and interesting data sources
    - E.g., sensor networks, RFIDs, …
- **How do we get knowledge to decision makers in a unified view?**

Information Integration

- Wikipedia:
  - "… Field of study of techniques attempting to merge information from disparate sources despite differing conceptual, contextual and typographical representations."
- Context of this talk:
  - Problem of making information from multiple, heterogeneous data sources available to users in a unified manner
    - Materialized (data warehouse) or virtual

Information Integration is a very Hard Problem…

- Listed on all listed on all four self assessments* of the DBMS community as a “grand challenge”
  - Problem that cannot be solved easily, and is intended as a “call to action” for a given field
- Conclusion in 2003: Scaleable solutions to information integration remain as elusive as they were decade ago

*

1. The Lowell Database Research Self Assessment, *The Computing Research Repository (CoRR)*, vol. cs.DB/0310006, 2003
2. The Asilomar Report on Database Research, *SIGMOD Record*, vol. 27, pp. 74-80, 1998
3. Database Systems: Achievements and opportunities, *Communications of the ACM*, vol. 34, pp. 110-120, 1991
4. Future Directions in DBMS research - the Laguna Beach Participants, *SIGMOD Record (ACM Special Interest Group on Management of Data)*, vol. 18, pp. 17-26, 1989

Challenges

- Understanding of source schema and data
  - Lack of semantics
    - Most data models in use do not capture adequate metadata
    - No standard way to represent semantics
  - Lots of ways to model information
    - E.g., Bill Kent: “The many forms of a single fact.” In Thirty-Fourth IEEE Computer Society International Conference (COMPCON Spring '89), San Francisco 1989
- Data cleaning and reconciliation
  - How do deal with missing data?
  - Identification of duplicate records without global ID space
- Complex data transformations
  - E.g., salary in Canadian $$ (net after taxes with a lunch allowance) to my wages (US dollars, gross)

Examples

- The many ways to represent calendar dates
  - `October 2, 2003`
  - `10/02/03`
  - `02/10/2003`
  - `Oct. 2 2003`
  - `02-10-2003`
- Two attributes do not semantically have to mean the same thing, even if they have a common representation
- **`“2 days”`** could mean…
  - Two calendar days
  - Two business days (excluding weekends and holidays)
  - Two Federal Express days (which excludes Sundays)
  - Two Wall Street trading days (which excludes weekends and certain other days)
  - Two London trading days (which excludes weekends and another collection of days)

More Challenges

- One person’s data is another person’s metadata
  - Relation “Sells-to” with attributes `salesperson` and `customer` vs relations “Sells-to-MacDonalds”, “Sells-toWendys”, “Sells-to-BK”, etc. with attribute `salesperson`
- Real-time processing (“on-the-fly” data integration)
- New types of data and sources (e.g., streams from sensors)
  - Both sources and the data they provide are very dynamic
  - Rethink the traditional “store-and-query” approach
- Lack of realistic test data makes it hard to experiment, validate and compare existing approaches

State-of-the-Art

- Gazillions of research papers – mostly since late 1980’s
- Mostly focused on “schema matching” problem
  - Your “wages” is my “salary”
- Success stories
  - Flexible architectures for data integration
    - Federations, data warehousing, mediators
  - Multi-source query processing techniques including methods for optimizing queries across multiple data sources
  - Tools for rapid wrapping of data sources
- But, current integration approaches rely on manual coding of mediation and connection software - ***Not scalable***

Multi-faceted Problem

- Data management community
  - Data warehousing
  - Data mining
- A.I.
  - Knowledge representation & ontologies
  - Machine learning
- Web community
  - Web service
  - Service oriented architecture
- Standards committees

Data Base Context

- Lots of early work on federated databases
  - Sharing architectures and languages, view integration
- Wrapper & mediator toolkits
  - Stanford TSIMMIS project (Garcia-Molina et al.)
  - IBM’s Garlic project (Schwarz et al.) → IBMS’s Information Integrator
  - Maryland Wrapper project (Rashid et al.)
- Schema matching
  - Survey by Bernstein & Rahm

Data Base Context Cont’d

- Complete integration systems
  - Orchestra (Penn)
  - Clio (IBM Almaden)
  - Integration Wizard (Univ. Florida)
- Source exploration and knowledge extraction
  - Extraction of schema from databases analysis, e.g., SEEK project (Hammer et al.)
- Data mining
  - Cleaning of data (e.g., exploratory data analysis)
  - Approximate query processing
- High initial investment on configuring systems, very domain/problem specific, technologies only useful when sources are relatively static

Data Warehouse Context

- Major purpose of ETL (Extract-Transform-Load) tools
  - Informatica
  - Data Stage (Ascential now IBM)
  - Altova’s Mapforce, Itemfield’s Contentmaster, …
- All provide some high level scripting language for data access, extraction, and transformation
- One CIO of a major e-commerce warehouse said:
  - Warehouse schema changes once a week
  - “High pole in the tent” is writing/converting transforms
  - Transforms are only partly in Informatica; rest in data base procedures, custom code, …
- I.e., problem is converting your salary (Canadian $$; net after taxes with a lunch allowance) to my wages (USA dollars, gross)
  - By a human programmer

A.I. Context – Knowledge Representation

- Gazillion papers -- since the beginning of time
  - KRL
  - Loom
  - Classic …
- Semantic web continues this thrust
  - RDF
  - RDF schema
  - OWL (lite, full, DL, …)
- Reasoning about data meaning – by a program; not a human
  - Sophisticated meta data systems

A.I. Context – Ontologies

- Ontology editors and knowledge representation frameworks
  - E.g., Protégé - Stanford University
- Tools for designing, merging, and sharing ontologies
  - E.g., Ontology algebra (Wiederhold et al.)
- Ontology libraries
  - Core ontologies containing elements that are as generic and method-independent as possible
  - E.g., http://www.daml.org/ontologies/
- Lots of tools but designing useful ontology remains difficult

Web Context

- Significant efforts to integrating applications and data on the Web
- Lots of technologies and recommendations
- Web services
  - Standard based, loosely coupled (composability, agility), platform independent (interoperability), etc.
  - SOAP, WSDL, UDDI, …
- Service Oriented Architecture (SOA)
  - Merger of Web services and enterprise computing architectures
- Overall, viable approaches to overcome information integration challenges in Web context
  - Verdict is out on how efficient/scaleable this will be

Standards

- Get rid of the integration problem
- Big issue
  - No shortage of proposals
  - Difficult to get elephants to co-operate
- RosettaNet
  - Very slow adoption rate
- Will work where…
  - There is an 800 pound gorilla (WalMart, Dell, etc.)
  - Where there are a small number of actors incented to cooperate (airlines)

Data Transformations: Morpheus Project

- A transform construction tool (TCT)
  - High-level scripting tool in which to write transforms
  - But “open” to other tool environments
- A repository in which to store transforms and data types
  - With sophisticated browsing tools
- Based on POSTGRES
  - Leverages POSTGRES ADT system.
  - Provides a DBMS-based transformation system
- Joined effort with Mike Stonebraker (MIT)
- Funded by Microsoft
- Work in progress
  - First prototype demonstration at SIGMOD 2006, Chicago, IL in July 2006

UNIVERSITY OF FLORIDA

UF
Pete Dobbins
Christan Grant
Joachim Hammer
Dev Oliver
Umut Sargut
Rebecca Wells

MIT
Tiffany Dohzen
Mujde Pamuk
Mike Stonebraker

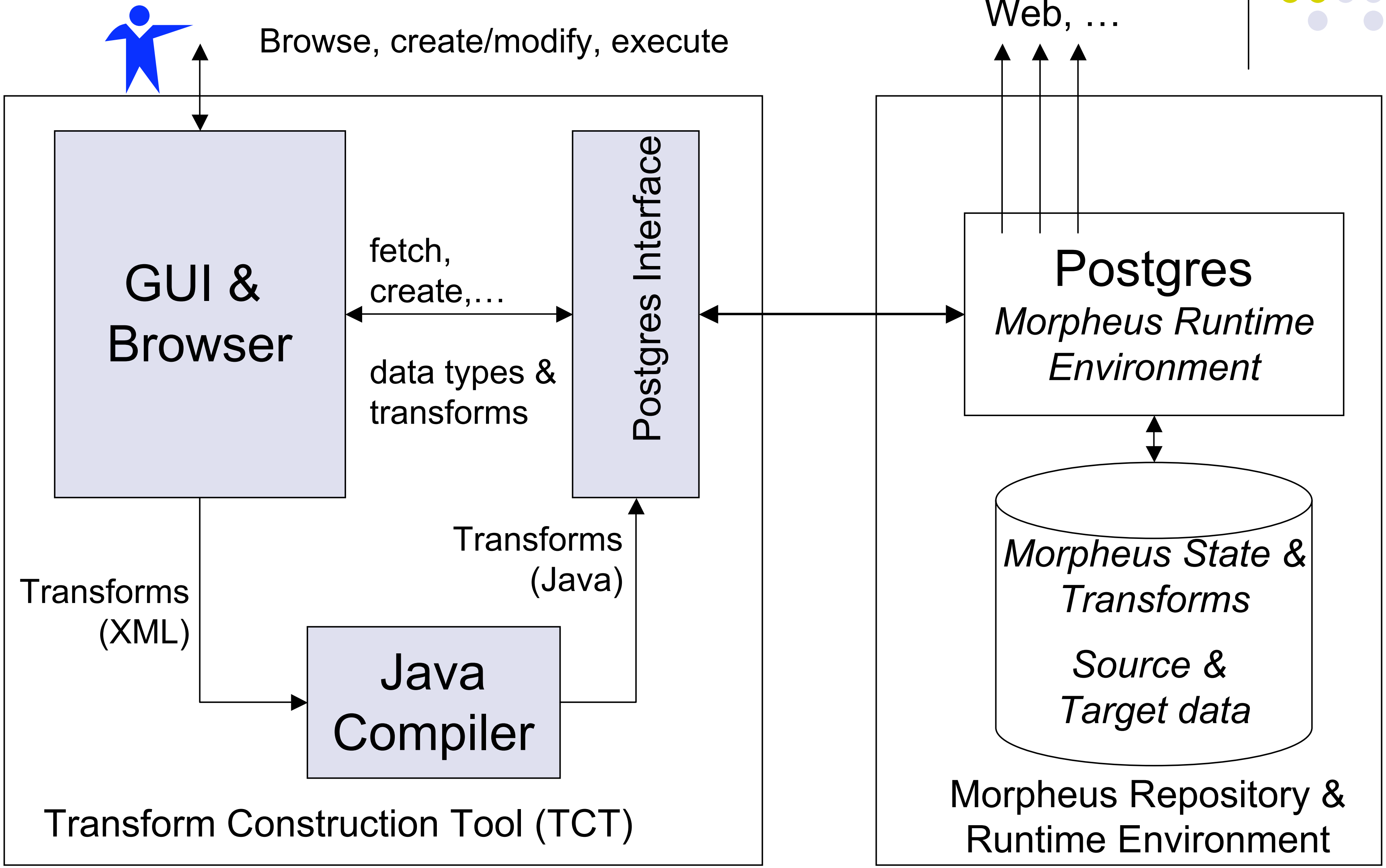
Goals

- Current research on schema matching does not represent a solution to the major information integration challenge that we see
  - Independently constructed schemas **never** have identical data elements (see previous example)
- Although automated integration tools may some day be available, need a solution that is
  - Efficient, scalable, powerful enough to address integration needs TODAY
- Goal is to develop tool that makes it easy for humans to build data transformations and share them with others

Context

	match	transform
Data	DBMS Schema matching	**Morpheus**
Text	Most A.I. efforts, taxonomies, etc.	Language translation, some A.I. efforts

Morpheus Architecture


Call-outs to external data, Web, …
Browse, create/modify, execute
GUI & Browser
fetch, create,…
data types & transforms
Postgres Interface
Postgres
Morpheus Runtime Environment
Morpheus State & Transforms
Source & Target data
Transforms (XML)
Transforms (Java)
Java Compiler
Transform Construction Tool (TCT)
Morpheus Repository & Runtime Environment

Transform Construction Tool

- Workflow-based
- High level primitives
  - E.g., rearranger, table lookup, misc. computation primitives, macro (superbox)
  - Postgres user-defined functions

Morpheus Repository

- Get (say) 50,000 popular transforms in MR
  - Have to take on faith that there is a critical mass of transforms..
- Basic design cycle
  - Find closest transform to what you need
  - Alter it or compose it with new stuff to get required transform
  - New transform is automatically added to MR
- Transformation of individual or bulk data handled by POSTGRES

Morpheus Repository

- Search by keyword in text descriptions
- Search by classification hierarchy of input or output data types
- Search by classification hierarchy of transforms
- Basic browsing paradigm
  - Use search to find something of interest
  - Browse in any “direction” to find nearby transforms

Example:
Convert MIT Student Record

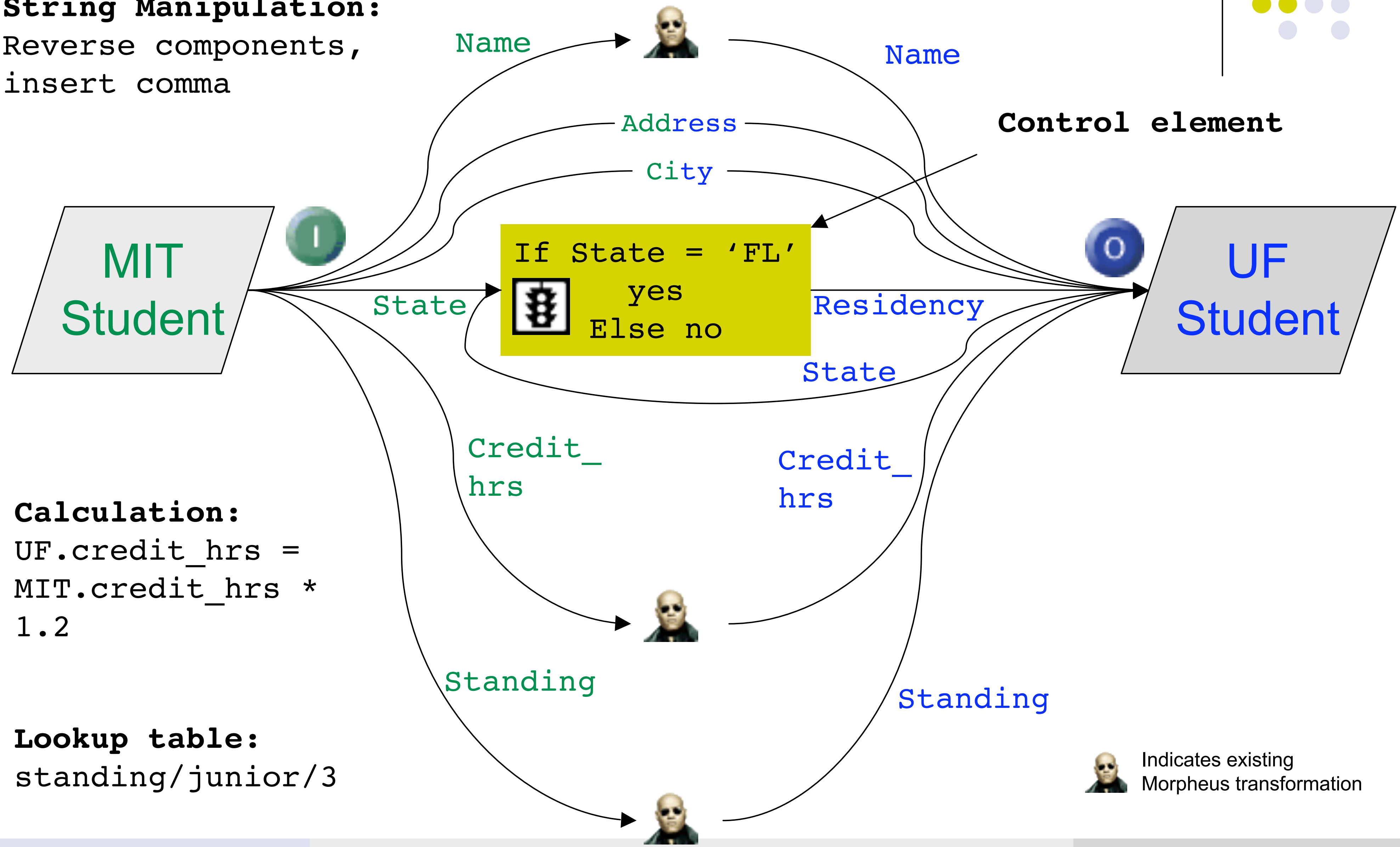
Name:	Mujde Pamuk
Address:	44 Foobar Str
City:	Atlanta
State:	GA
Credit hrs:	70
Standing:	3

… to UF Student Record

Name:	Pamuk, Mujde
Address:	44 Foobar Str
City:	Atlanta
State:	GA
Credit hrs:	84
Residency:	no
Standing:	junior

Transformation: StudentConversion

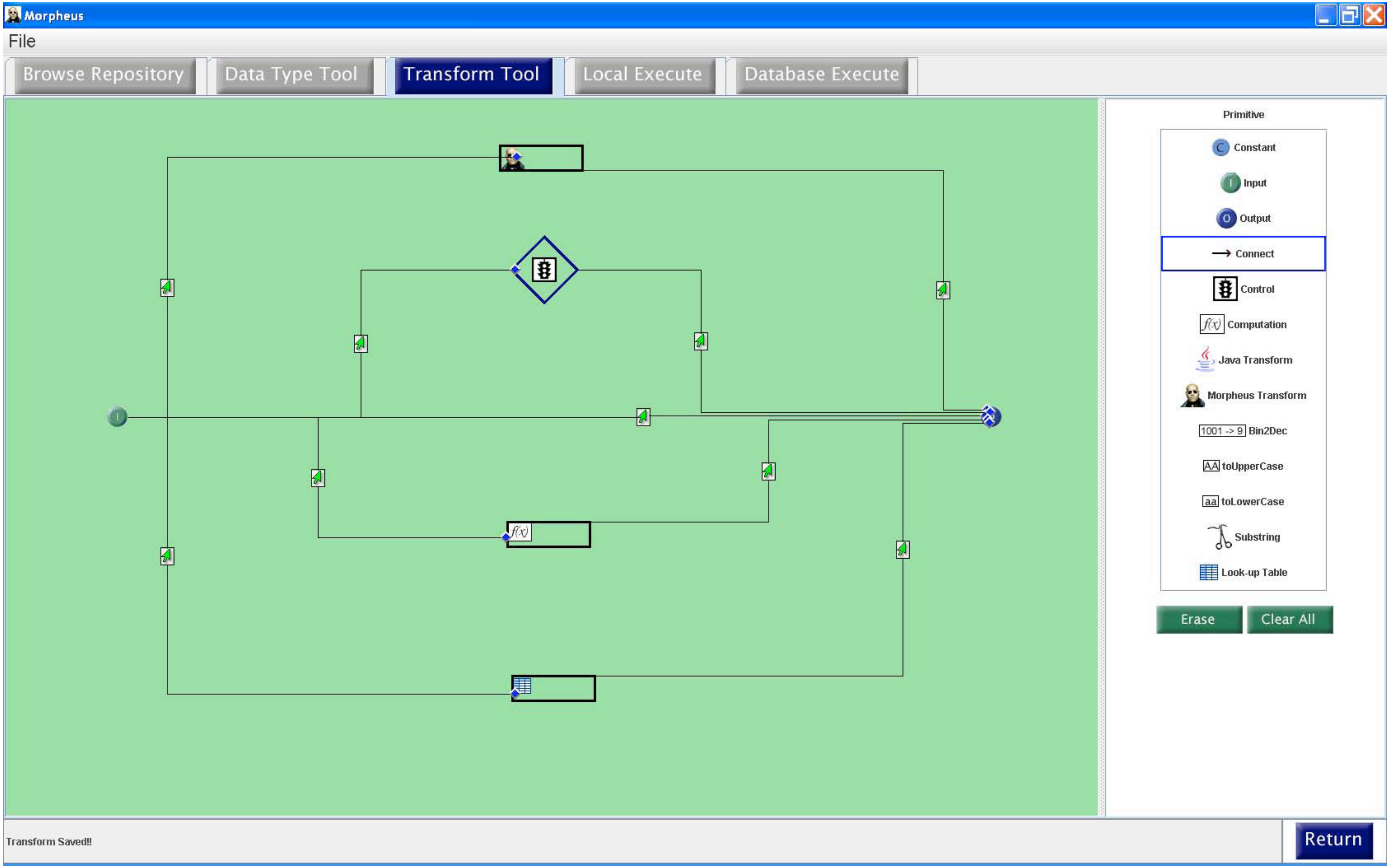
String Manipulation:
Reverse components, insert comma



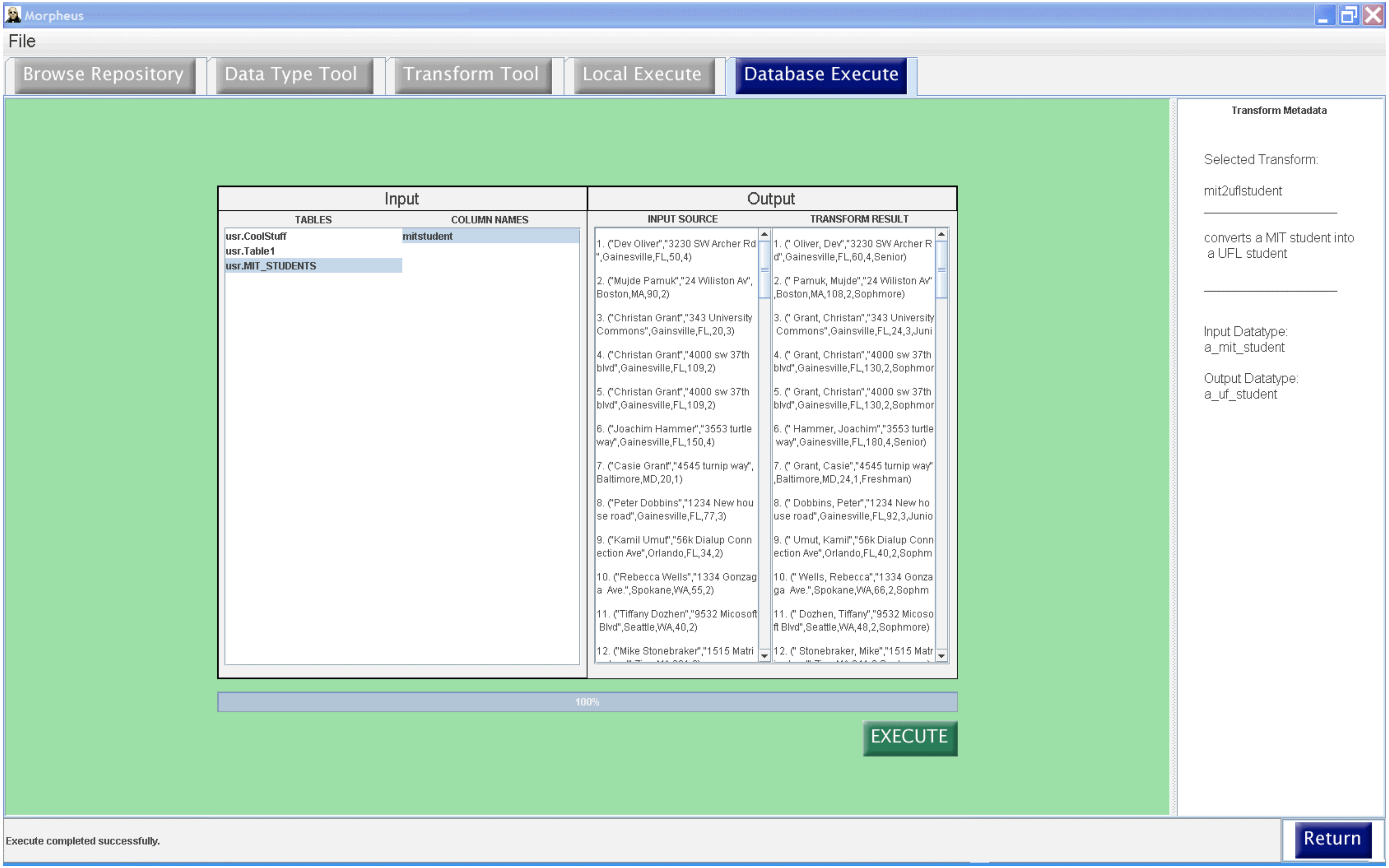
Calculation:
UF.credit_hrs = MIT.credit_hrs * 1.2

Lookup table:
standing/junior/3

Indicates existing Morpheus transformation



Transform Saved!!



Input

TABLES	COLUMN NAMES
usr.CoolStuff	mitstudent
usr.Table1	
usr.MIT_STUDENTS	

Output

INPUT SOURCE	TRANSFORM RESULT
1. ("Dev Oliver","3230 SW Archer Rd",Gainesville,FL,50,4)	1. (" Oliver, Dev","3230 SW Archer Rd",Gainesville,FL,60,4,Senior)
2. ("Mujde Pamuk","24 Wiliston Av",Boston,MA,90,2)	2. (" Pamuk, Mujde","24 Wiliston Av",Boston,MA,108,2,Sophmore)
3. ("Christan Grant","343 University Commons",Gainsville,FL,20,3)	3. (" Grant, Christan","343 University Commons",Gainsville,FL,24,3,Juni
4. ("Christan Grant","4000 sw 37th blvd",Gainesville,FL,109,2)	4. (" Grant, Christan","4000 sw 37th blvd",Gainesville,FL,130,2,Sophmor
5. ("Christan Grant","4000 sw 37th blvd",Gainesville,FL,109,2)	5. (" Grant, Christan","4000 sw 37th blvd",Gainesville,FL,130,2,Sophmor
6. ("Joachim Hammer","3553 turtle way",Gainesville,FL,150,4)	6. (" Hammer, Joachim","3553 turtle way",Gainesville,FL,180,4,Senior)
7. ("Casie Grant","4545 turnip way",Baltimore,MD,20,1)	7. (" Grant, Casie","4545 turnip way",Baltimore,MD,24,1,Freshman)
8. ("Peter Dobbins","1234 New house road",Gainesville,FL,77,3)	8. (" Dobbins, Peter","1234 New house road",Gainesville,FL,92,3,Junio
9. ("Kamil Umut","56k Dialup Connection Ave",Orlando,FL,34,2)	9. (" Umut, Kamil","56k Dialup Connection Ave",Orlando,FL,40,2,Sophm
10. ("Rebecca Wells","1334 Gonzaga Ave.",Spokane,WA,55,2)	10. (" Wells, Rebecca","1334 Gonzaga Ave.",Spokane,WA,66,2,Sophm
11. ("Tiffany Dozhen","9532 Micosoft Blvd",Seattle,WA,40,2)	11. (" Dozhen, Tiffany","9532 Micosoft Blvd",Seattle,WA,48,2,Sophmore)
12. ("Mike Stonebraker","1515 Matri	12. (" Stonebraker, Mike","1515 Matr

Transform Metadata

Selected Transform:

mit2uflstudent

converts a MIT student into a UFL student

Input Datatype:
a_mit_student

Output Datatype:
a_uf_student

Execute completed successfully.

Future of Morpheus

- Add missing functionality
  - Protection system
  - More elaborate error checking
- Ability to include Web services in transformation
- Improved visualization system
- Integrate with existing transformation tools, e.g., Microsoft's SSIS
- Conduct a thorough performance study to evaluate whether Morpheus idea can support 100's of users executing 1000's of transformations on large data sets
- If Morpheus is successful, will be further proof that software sharing and reuse can work!

More on Morpheus

`www.cise.ufl.edu/~jhammer/morpheus`

- Technical Reports and Unpublished Work
  - P. Dobbins, T. Dohzen, C. Grant, J. Hammer, D. Oliver, M. Pamuk, U. Sargut, and R. Wells, "The Morpheus Data Integration System," submitted to *Conference on Innovative Database Research (CIDR)*. Asilomar, CA, 2007
- Conference Papers
  - T. Dohzen, M. Pamuk, S.-W. Seong, J. Hammer, and M. Stonebraker, "Data integration through transform reuse in the Morpheus project," in *ACM SIGMOD International Conference on Management of Data (Demo Track).* Chicago, IL: ACM, 2006, pp. 736-738
  - J. Hammer, O. Topsakal, and M. Stonebraker, "THALIA: Test Harness for the Assessment of Legacy Information Integration Approaches," in *21st International Conference on Data Engineering (ICDE)*. Tokyo, Japan: IEEE, 2005, pp. 485-486

Summary

- Information integration is hard problem
  - Pain can be felt throughout the business world
- Has attracted much attention from research community
  - Numerous success stories
  - Lots of start-ups offering viable solutions to help ease the pain
- Personal focus is on basic problems of making data accessible by humans and decision support algorithms
  - Data repositories are rapidly growing in size (many new formats and data types)
  - New applications / new requirements
- Contributions to specific areas of research
  - Knowledge extraction to support access to legacy data sources
    - Generation of software to connect heterogeneous, distributed processes
  - Data transformations: Morpheus project

Future Directions

- Need to continue efforts to make integration approaches more “user-friendly”
  - Domain expert merely as guidance rather than human integrator
- Shift in research focus to supporting “on-the-fly” integration to support more dynamic data sources
  - Perhaps millions of data sources
  - Degree of integration is lower
  - May need to accept approximations of results
- Other important areas:
  - Deep Web integration
- Will information integration ever be completely automated?
  - Probably not in my lifetime
  - However, with increase in collaboration among the different areas more progress is possible