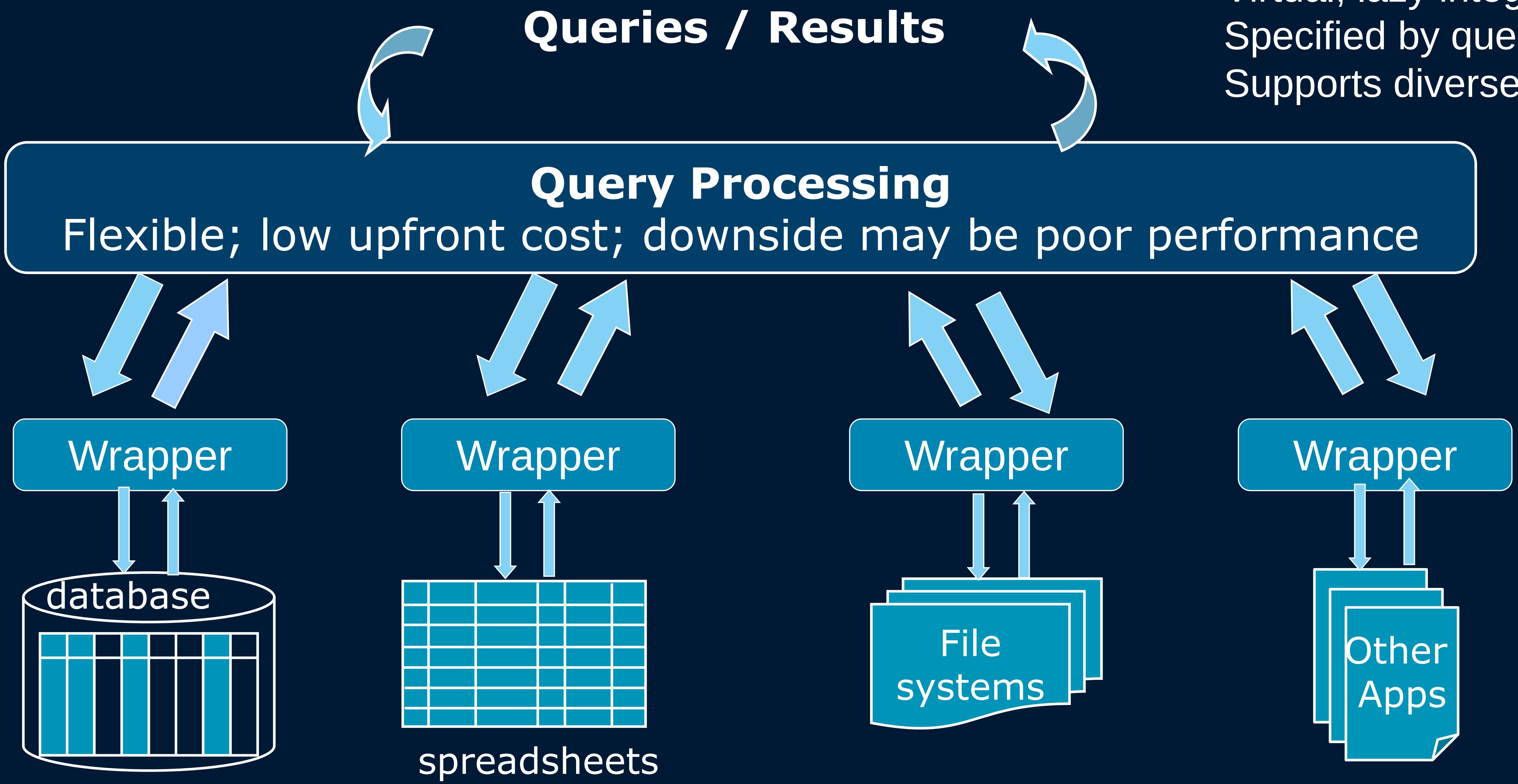
Theory and Practice: Collaborating with Ron Fagin on Two Projects in Information Integration

Laura Haas
IBM Fellow and Director, IBM Research Accelerated Discovery Lab

The Adventure Begins…

Heterogeneous Federation: Garlic

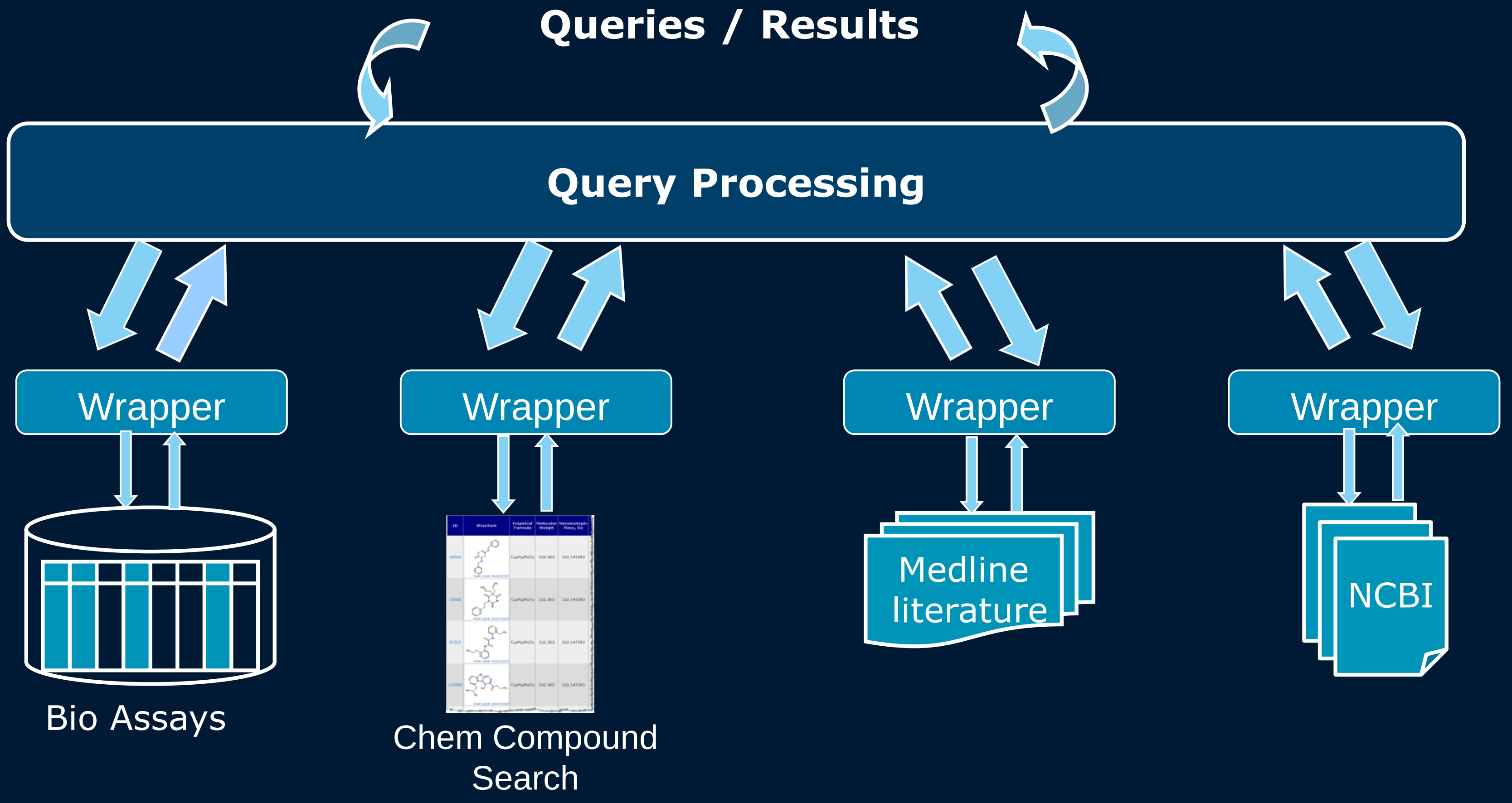
Virtual, lazy integration
Specified by queries
Supports diverse data

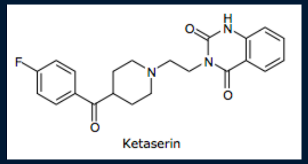


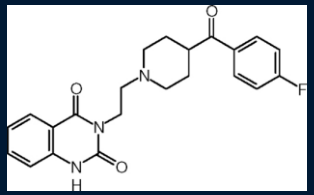
Queries / Results

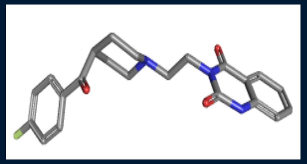
Query Processing
Flexible; low upfront cost; downside may be poor performance

Our first real applications were in life sciences (pharma)









Ketanserin, 74050-98-9, Ketanserina, Ketanserine, Ketanserinum, Ketanserin tartrate, Perketal, Serefrex, Sufrexal, Taseron, C22H22FN3O3, CHEMBL51, R-41468, CHEBI:6123, R-41,468, Tocris-0908, 3-(2-(4-(4-Fluorobenzoyl)piperidin-1-yl)ethyl)quinazoline-2,4(1H,3H)-dione, AC1L1GSK, Spectrum2_001713, EINECS 277-680-2, Biomol-NT_000096, UNII-97F9DE4CT4

The prototypical query for drug discovery:

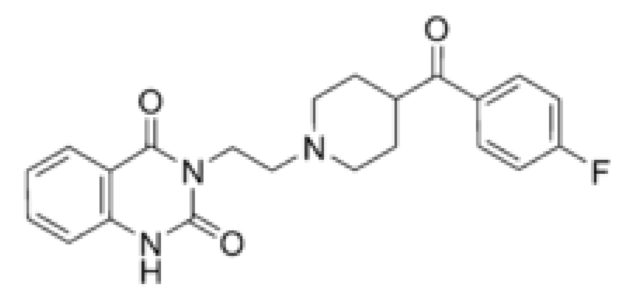
- *“Find a compound with a structure like this one and assay results in this range”*
- Example:
  - Show me all the compounds similar to ketanserin that have been tested against members of the serotonin family and have an ic50 < 1E-8 with molecular weight between 375 and 450, and a logP value between 4 and 6

With Heterogeneous Sources, You Get Heterogeneous Semantics

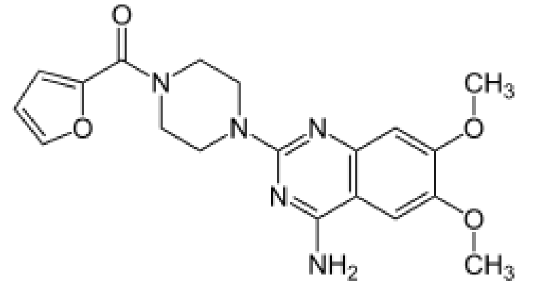
- Searching a database for compounds where “375 < Molwt < 450” yields a set

  Ambrisentin (378), Prazosin (383), Trimetaphen cansilate (365), Ketanserin (395)

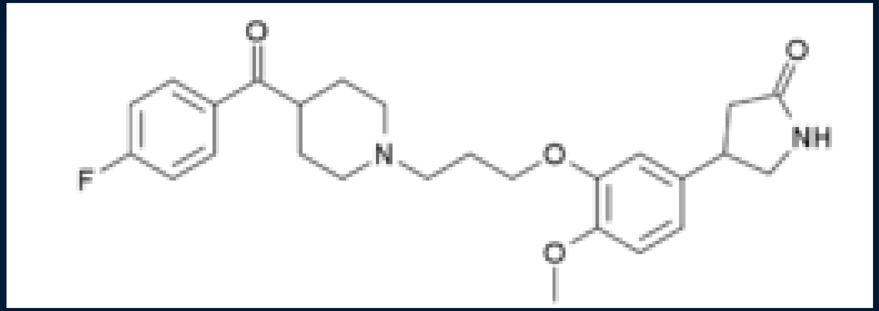
- Using a compound search engine to look for “Structure like Ketanserin” yields a sorted list



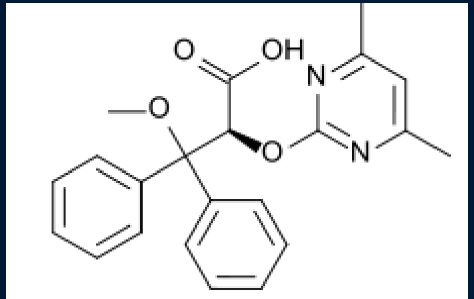
Ketanserin, .99



Prazosin, .85



Lidanserin, .63

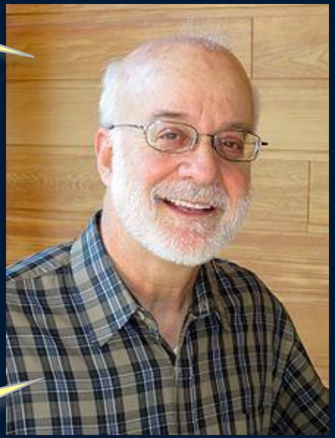


Ambrisentin, .12

- How do we make sense of a query like (375 < Molwt < 450) ∧ (Structure like Ketanserin) ?
- What about (375 < Molwt < 450) ∨ (Structure like Ketanserin) ?
- And what about (Structure like Ketanserin) ∧ (Usage like ‘reduce hypertension’) ?

Who You Gonna Call?

Simple questions can be surprisingly hard to answer!



Ron, help!

What do you mean? Be precise!

I get it! Use fuzzy logic

Ok, but how do I do that - fast?

What do you mean? Be precise!

I have an algorithm that finds the top k with only √n database accesses

Is that the best you can do?

I proved that you can't do better than √n – it's optimal

The Rest is History

- R Fagin: Combining Fuzzy Information from Multiple Systems in PODS 1996 has been cited over 860 times
- We eventually implemented it in Garlic
  - It wasn’t easy, It required a series of unnatural acts to ensure it was used correctly.
  - E Wimmers, L Haas, M Roth, C Braendli: Using Fagin's Algorithm for Merging Ranked Results in Multimedia Middleware. CoopIS 1999 was cited 43 times
- Influenced other IBM products, including
  - Watson Bundled Search system
  - InfoSphere Federation Server
  - WebSphere Commerce
- Ron and friends (Lotem and Naor) eventually came up with a better algorithm
  - R Fagin, A Lotem, M Naor: Optimal Aggregation Algorithms for Middleware. PODS 2001 has been cited more than 1800 times
  - Won the Best Paper Award in PODS 2001
  - PODS Test of Time Award in 2011
  - IEEE Technical Achievement Award in 2011
  - Gödel Prize in 2014
  - Gems of PODS talk, 2016
- Laura never understood how this algorithm could be “more optimal” than the original*

* Well, ok, I get it, but there’s a lesson in here about different communities’ idea of precision! ☺

Meanwhile, Garlic Had Its Own Successes

- Made heterogeneous federation mainstream and commercially available
  - Leveraged a commercial query processing engine and handled all SQL queries
  - Relatively few, simple, object-relational extensions to accommodate diverse sources
  - Multiple IBM products and ultimately the basis for a new line of business for IBM and the industry’s Information Integration market

- Made it (more) practical
  - Cost-based optimization, where wrappers provide the input on capabilities and costs
  - Extensible wrapper architecture, optimizer-controlled caching

- Applied it to a compelling problem – the killer app for life sciences

And We Learned a Lot

- Federation is fabulous for rapid proof-of-concepts and iterative development
  - Still being sold today
  - Enabling technology for BigSQL, BLU, and other “hybrid” data systems

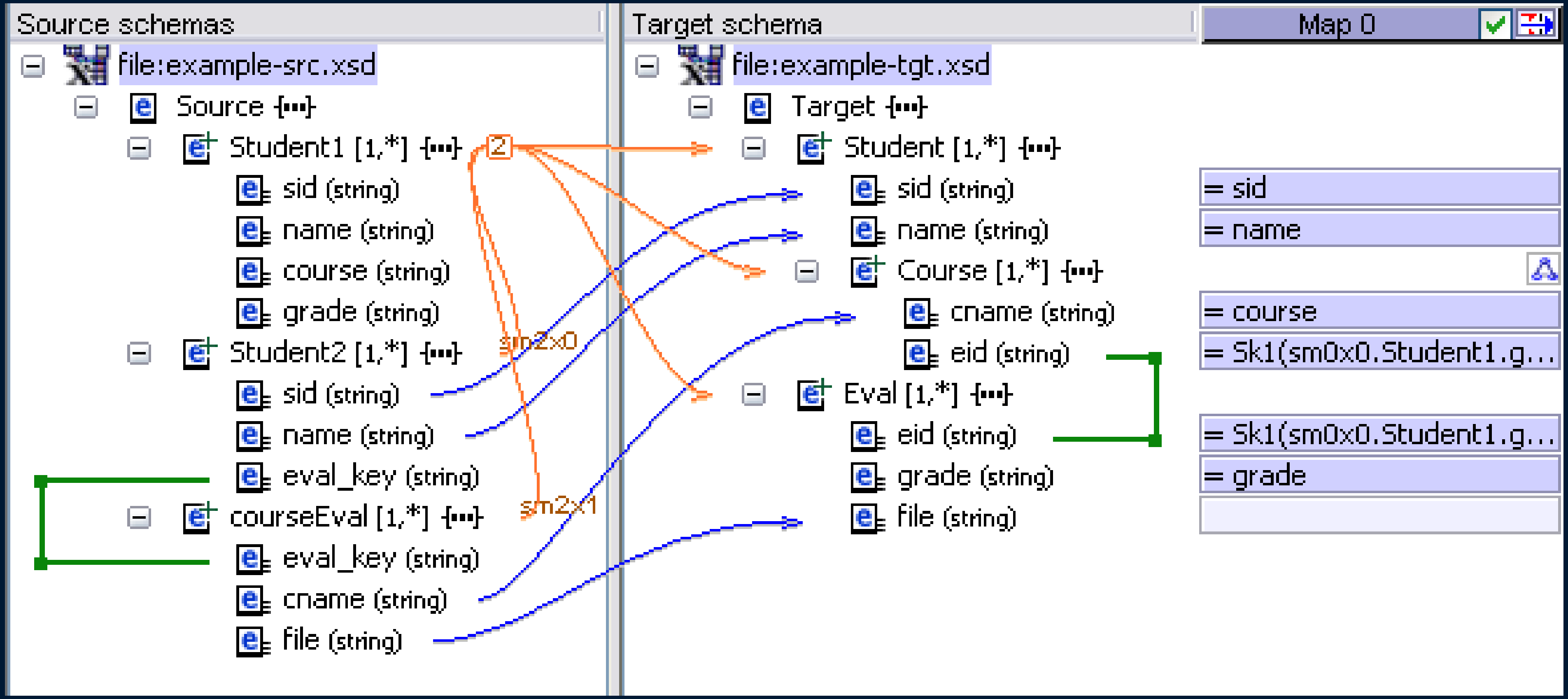
- If you give your clients some rope, they’ll hang themselves
  - All the power of SQL is a lot of power
  - Simplicity is nice but misleading
  - Some queries cannot be done efficiently if the data is distributed

- Configuring the system (setting up access to remote data) could be easier
  - Nicknames had to be defined and linked to (simple) queries
  - Should be able to generate the DDL easily
  - Really just a matter of mapping attributes…

WAIT! That could be interesting!

Clio: Schema Mapping Creation

Clio



Source Schema ⟶ Target Schema

Key ideas: Use correspondences, preserve data semantics

The mapping is a high-level specification we can compile into a transformation script

Clio could generate SQL, XSLT, Java, …

Miller, Haas, Hernández. *Schema Mapping as Query Discovery*. VLDB 2000

Haas, Hernández, Ho, Popa, Roth. *Clio grows up: from research prototype to industrial tool.* SIGMOD Conference 2005

But It’s Not So Simple!

Source

EMP	MGR
Fagin	Haas
Clarkson	Haas
Haas	Welser

Target

EMP	DEPT

DEPT	MGR

Three Possible Solutions – Which One Is Best?

Source

EMP	MGR
Fagin	Haas
Clarkson	Haas
Haas	Welser

Target

EMP	DEPT
Fagin	Haas
Clarkson	Haas
Haas	Welser

DEPT	MGR
Haas	Haas
Welser	Welser

EMP	DEPT
Fagin	d_1
Clarkson	d_1
Haas	d_2

DEPT	MGR
d_1	Haas
d_2	Welser

EMP	DEPT
Fagin	d_1
Clarkson	d_2
Haas	d_3

DEPT	MGR
d_1	Haas
d_2	Haas
d_3	Welser

Who You Gonna Call?

This Time, Things Went Much More Smoothly! Why?

1. Having "discovered" the problem, I left
2. Ron had playmates who could speak his language

Phokion Kolaitis

Renee Miller

Lucian Popa

And later,

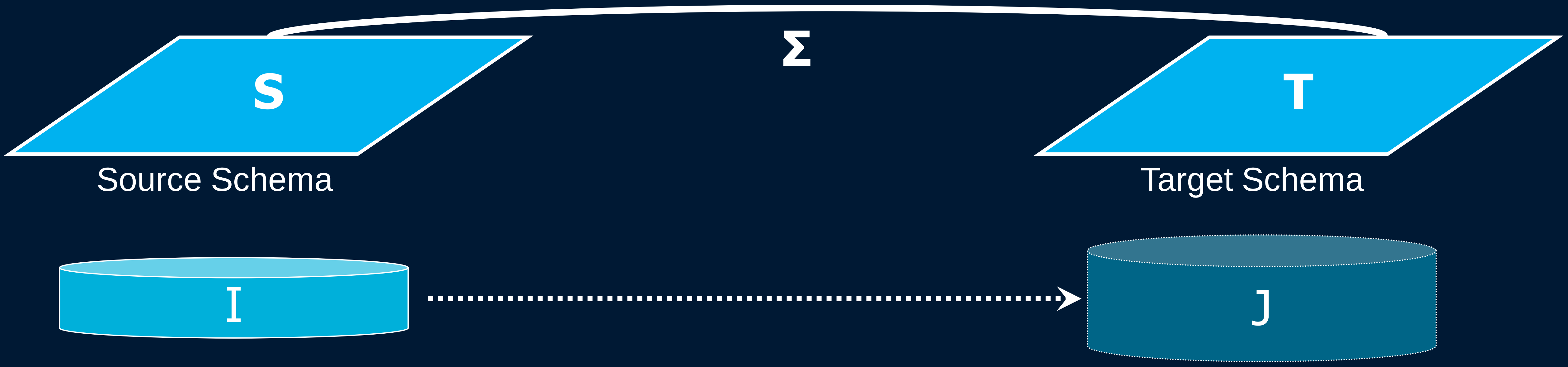
Wang-Chiew Tan

Let's start from scratch and lay the foundations for data exchange!

What is Data Exchange?

Data exchange is an old, but continual, database problem

- Phil Bernstein—2003: “*Data exchange is the oldest database problem*”
- **EXPRESS**: IBM San Jose Research Lab—1977
  - Transforms data between hierarchical databases
- Data exchange underlies:
  - Data warehousing, ETL (Extract-Transform-Load), …



So What Did They Do?

- Answered the question: which solution should we produce?
  - Defined a “universal” solution to be one as general as possible
  - Third solution is universal if there are no target constraints

EMP	DEPT
Fagin	d_1
Clarkson	d_2
Haas	d_3

DEPT	MGR
d_1	Haas
d_2	Haas
d_3	Welser

- Figured out how to deal with target constraints specified by equality-generating dependencies (*egds*)
  - For example, DM(d,m) ∧ DM(d',m)) → ($d = d'$)
  - If this egd is a target constraint, then second solution is universal

EMP	DEPT
Fagin	d_1
Clarkson	d_1
Haas	d_2

DEPT	MGR
d_1	Haas
d_2	Welser

- Figured out how to find the universal solution
  - Use the “chase” (a tool from database design) to generate the target from the source efficiently
  - The egds tell when to equate labeled nulls
- Explored and solved many further problems
  - Mapping composition
  - Mapping inversion

This Work Also Had a Huge Impact

- Technology used in many products and research systems
  - In Federation to configure schemas and generate views
  - In Content management systems to transform between XML representations
  - In DB design tools to convert between different information models
  - In application development tools to map between relational data and object-oriented programming models

- Created a rigorous foundation for the study of integration semantics

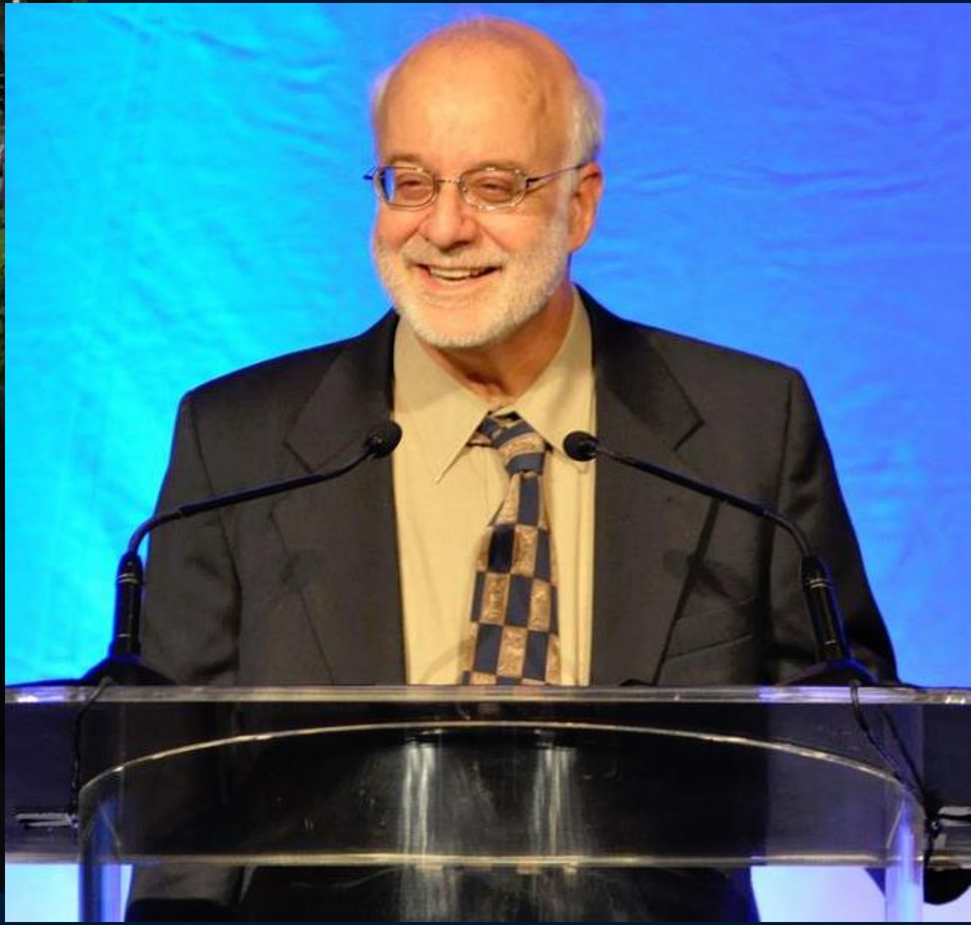
- Spawned a subfield for the systematic investigation of the semantics and uses of schema mappings
  - For data integration and data exchange
  - For schema evolution and metadata management

- Highly influential
  - 1st paper won the International Conference on Database Theory Test of Time Award in 2013

    Over 1000 citations; 2nd most highly cited paper of the decade in the journal TCS
  - Follow-up paper on composition won the PODS Test of Time Award in 2014
  - Led to many PhD dissertations

Ron and I Have Been Through a Lot Together

- Bridging our differences was not easy, but we both were rewarded



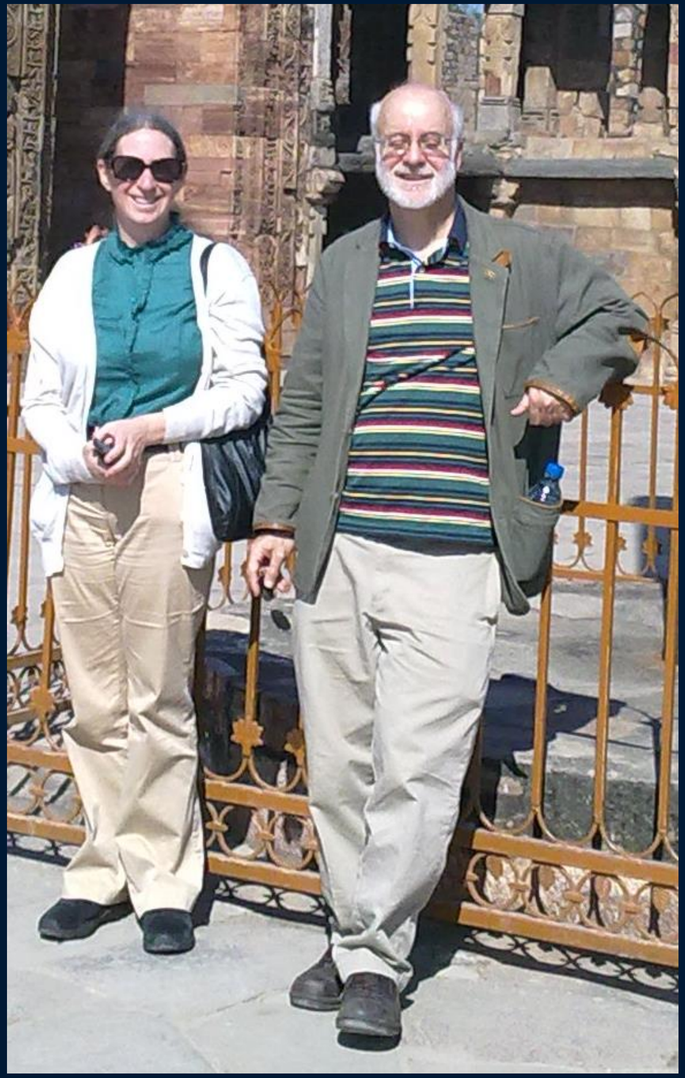
Ron and I Have Done a Lot Besides Science Together

- We eat a lot

- And we are both VERY competitive!

We won the Almaden Olympics – twice!

We’ve been around the world



I am honored to have him as my friend and colleague