Solving Information-Integration Problems Using Category Theory

David I. Spivak, Ryan Wisnesky, Joshua Tan, Jee Chung

October 21, 2015

Part I

Project Summary

Project Summary

Overview: For large organizations, simple logistical questions can be surprisingly hard to answer. A question like "how many employees are tax-exempt?" may require querying hundreds of databases using multiple data models and possibly inconsistent definitions: for example, is a "contractor" also an "employee"? Over the past five years, our team has developed a new technology for performing information-integration tasks—such as querying, combining, and evolving databases—based on category theory, a new branch of mathematics which originated in 1945 [3]. Category theory gives us theoretical guidance missing from the widely used relational model of data [5], and we have used it to build a prototype software tool, FQL (`categoricaldata.net/fql.html`), for integrating databases more quickly and accurately than existing relational tools. The goal of this project is to determine the commercial viability of our categorical approach to information integration.

Intellectual Merit: Information integration is a perennial topic in computer science, being first identified as a problem of significant interest in the 1970s [1]. There are tools, such as extract-transform-load (ETL) tools, that perform certain information integration tasks, but these tools are invariably ad-hoc [1]. As a natural mathematical framework for describing models and the relationships between them, category theory provides *principled* techniques for solving the information integration problems that plague scientists, engineers, and commercial enterprises. Although category theory has revolutionized several areas of computer science – for example, functional programming [2] – category theory is only beginning to be applied to problems in information integration.

Broader Impacts: According to the market intelligence firm IDC, 40% of all enterprise IT budgets are dedicated to solving information-integration problems, and total annual spending of information-integration software alone is over $5 billion [4]. Information-integration problems are currently solved using a combination of manual data manipulation and custom software, all based on the relational model of data. The categorical model we have developed has been formally proven to provide more accurate solutions to certain information-integration problems than the relational model and to solve certain information-integration problems that the relational model cannot [6]. And because relational databases are a special case of categorical databases, our approach can easily interoperate with existing technologies. Hence, practical software based on our categorical model has the potential to significantly ease the burden of challenging real-world IT problems. In addition, the success of an information-integration tool based on category theory may inspire other efforts to leverage category theory for data management tasks.

Contents

Part II

Project Description

1 I-Corps team

Our team recently co-founded Categorical Informatics, Inc. with the goal of providing software and services to solve information-integration problems using the mathematics of category theory. This very early stage venture is attempting to commercialize research that has been done over the past five years at MIT. We recently (Oct 1, 2015) began a small ($100k, 6-month) Phase I SBIR from the National Institute of Standards and Technology (NIST) to complete some almost-finished research questions and to harden (i.e., document, bug-fix, etc) the prototype/proof-of-concept codebase we have developed. We have been been engaged in initial market exploration and pilots (e.g., with NIST) over the past year, and while progress is being made, the venture has not achieved any significant revenue or achieved product-market fit. We believe that NSF I-Corp represents a critical opportunity. Our team has become increasingly convinced that our research has significant economic value and is cautiously optimistic that it has the potential to be successfully transitioned and commercialized. Unfortunately, we have also come to recognize that success will require a more aggressive and disciplined approach to customer discovery and validation. We need to better understand the market, prioritize the features we develop, understand how to package them into compelling products, and discover what problems people are most willing to pay to solve.

Our team consists of 1) David Spivak (dspivak@math.mit.edu), a research scientist in the MIT department of mathematics who studies the use of category-theory to model information-bearing structures of all kinds, from databases to social networks; 2) Ryan Wisnesky (ryan@catinf.com), previously a postdoc at MIT and now president of Categorical Informatics, Inc, who developed the initial proof-of-concept software that demonstrates our categorical approach to information integration; 3) Josh Tan (josh@catinf.com), a mathematician from NYU with experience in both category theory and several start-ups; and 4) Jee Chung (jeechung@alum.mit.edu), Director of Enterprise Systems at GMO (a financial institution) with 25 years of experience solving information-integration problems in large enterprises. David Spivak and Ryan Wisnesky are Co-PIs, Joshua Tan is the EL, and Jee Chung is the IM.

2 Lineage of proposed innovation

The team was selected for, and participated in the MIT VMS I-Corps Site Program in early 2015, receiving a microgrant (NSF award #1347267), which afforded us an introduction to the customer discovery process.

David Spivak has been working on the theory of categorical databases since 2008. His main interest has been in the translation of data between models, via data migration functors. This work has taken on many forms, from simplicial databases, which use a mix between combinatorial and topological methods, to the simplified model in which database schemas

are simply categories and instances are set-valued functors. He was encouraged to discover that the pullback and left and right Kan extensions of presheaves along a schema mapping mimicked constructions found in databases, namely project, join, and union, but that some Kan extensions corresponded to genuinely new data migration possibilities.

In 2010, David Spivak met Ryan Wisnesky (then a computer science Ph.D. student at Harvard), who had been studying database query languages from a category-theoretic perspective. Ryan Wisnesky had also been collaborating with the information-integration department of IBM research on the traditional (i.e., relational) approach to information-integration, and he quickly recognized both the potential of David Spivak's theoretical work and the need to create working software so as to try out the theory in practice. Ryan Wisnesky and David Spivak then collaborated to flesh out the computer science of functorial data migration, which Ryan Wisnesky implemented as the prototype FQL software tool (categoricaldata.net/fql.html).

In the summer of 2015, the I-Corp Site customer-discovery activities led Ryan to GMO, where he was introduced to Jee Chung, who, having spent 25 years doing information-integration in large enterprises using current ad-hoc tools, immediately recognized the potential for functorial data migration to (eventually) make his life easier by providing a more principled way to perform information integration.

2.1 Broader impacts of the proposed work

The ability to store and access information, e.g., in a database or a set of tables, is crucial for any organization, from science lab to commercial enterprise to government. There are many scenarios in which one database has to "talk to" another, i.e., to share information without compromising its integrity. This can occur on a small scale, such as when two scientific labs want to share data, or on a large scale, such as during acquisitions and mergers. Successfully translating data between disparate systems typically presents a major problem, as it is difficult to coordinate the models. This can be aided with category theory, a branch of abstract mathematics designed to build bridges between disparate models. Our categorical database model has been constructed for this purpose. Our aim is to leverage the high-level mathematical insights to mitigate the enormous complexity and reduce the cost of data integration tasks.

The development of an industrial-strength tool will also facilitate the dissemination of category theory, because it provides a concrete solution to a difficult problem, while staying very close to the mathematical foundation. Bringing mathematics down to earth with such a tool serves the pedagogical purpose of enabling users to play with very abstract mathematics

in the form of migrating data between database models.

3 Description of the potential commercial impact

According to the market intelligence firm IDC, 40% of all enterprise IT budgets are dedicated to solving information-integration problems, and total annual spending of information-integration software alone is over $5 billion [4]. Information-integration problems are currently solved using a combination of manual data manipulation and custom software, all based on the relational model of data. The categorical model we have developed has been formally proven to provide more accurate solutions to certain information-integration problems than the relational data model and to solve certain information-integration problems that the relational model cannot [6]. And because relational databases are a special case of categorical databases, our approach can easily interoperate with existing technologies. Hence, practical software based on our categorical model has the potential to significantly ease the burden of challenging real-world IT problems.

4 Brief description of project plan

Categorical Informatics, Inc was recently spun out of MIT and is being initially funded by a small ($100k) 6 month Phase I SBIR grant from the National Institute of Standards and Technology (NIST). This funding is dedicated to completing almost-finished research questions and to begin hardening the prototype/proof-of-concept codebase (creating documentation, fixing bugs, etc). This transition period is projected to run from October 1, 2015 until March 31, 2016. However, even a hardened version of the current codebase is not a product: it is a technology demonstration, and we have not yet achieved product-market fit. Hence, during this transition period it is critical that Categorical Informatics determine the exact product that must be developed in order to secure a phase II SBIR and to attract venture capital funding. In addition to determining the form of the product – e.g., a programming language, a Java library, a cloud service, etc – it is also critical to determine the market that this product fits. Based on preliminary pilot studies, we believe a process-ŋoriented industry such as manufacturing would be a good initial target market. Product engineers in manufacturing are particularly interested in tools that can answer questions about designs as they pass through the entire product lifecycle; better feedback leads directly to cost ŋsavings downstream in commissioning (supply chains) and redesign. We need to develop domain-expertise to determine if this is indeed a good initial market, and because the underlying technology is domain agnostic, we believe we should explore other possible markets (e.g., finance, healthcare) as

well.

5 Results from prior NSF support

The only prior NSF support we have had was through the I-Corps Site program, a $1,500 micro-grant distributed by MIT's Venture Mentoring Service. Through this grant, a new company called *Categorical Informatics Inc*, recently spun out of MIT, was introduced to the customer-discovery process. This provided a bit of the background needed for the company to successfully commercialize the technological results of prior categorical informatics research at MIT.

Intellectual Merit: Information integration is a perennial topic in computer science, being first identified as a problem of significant interest in the 1970s [**?**]. There are tools, such as extract-transform-load (ETL) tools, that perform certain information integration tasks, but these tools are invariably ad-hoc [**?**]. As a natural mathematical framework for describing models and the relationships between them, category theory provides *principled* techniques for solving the information integration problems that plague scientists, engineers, and commercial enterprises. Although category theory has revolutionized several areas of computer science—for example, functional programming [2]—category theory is only beginning to be applied to problems in information integration. It is this application that we addressed using the I-Corps Site grant.

Broader Impacts: According to the market intelligence firm IDC, 40% of all enterprise IT budgets are dedicated to solving information-integration problems, and total annual spending of information-integration software alone is over $5 billion [**?**]. Information-integration problems are currently solved using a combination of manual data manipulation and custom software, all based on the relational model of data. The categorical model we have developed has been formally proven to provide more accurate solutions to certain information-integration problems than the relational model and to solve certain information-integration problems that the relational model cannot [6]. And because relational databases are a special case of categorical databases, our approach can easily interoperate with existing technologies. Hence, practical software based on our categorical model has the potential to significantly ease the burden of challenging real-world IT problems. In addition, the success of an information-integration tool based on category theory may inspire other efforts to leverage category theory for data management tasks. The specific results of our work on the I-Corps Site program was that we began our exploration of the market need for our information integration tool.

Part III

References Cited

References cited

[1] Doan, A.; Halevy, Z. Ives. (2012) *Principles of Data Integration*. Morgan Kaufmann Publishing.

[2] Moggi, E. (1991) *Notions of computation and monads*. Information and Computation Journal Volume 93 Issue 1.

[3] Eilenberg, S.; MacLane, S. (1945) *General theory of natural equivalences.* Trans. Amer. Math. Soc. 58, pp. 231 – 294.

[4] Bernstein, P.; Haas, L. (2008) *Information Integration in the Enterprise* Communications of the ACM Volume 51 Number 9, pp. 72 – 79.

[5] Abitebould, S.; Hull, R.; Vianu, V.. (1995) *Foundations of Databases*. Addison-Wesley Longman Publishing.

[6] Spivak, D.I.; Wisnesky, R. (2015) *Relational Foundations for Functorial Data Migration*. Proceedings of the International Symposium on Database Programming Languages (DBPL), ACM.

Part IV

Biographical Sketches

David I. Spivak

(Co-Principal Investigator)

Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Building E17 Room 417
Cambridge, MA 02143
dspivak@mit.edu

Professional Preparation

University of Maryland, College Park, BS 1996
University of California, Berkeley, PhD 2007

Appointments

2007 – 2010: Visiting Assistant Professor, University of Oregon
2010 – 2013: Postdoctoral Associate, Massachusetts Institute of Technology
Since 2013: Research Scientist, Massachusetts Institute of Technology

Related Publications

- Spivak, D.I. (2012) "Functorial Data Migration". *Information and Communication*. Vol 217, pp. 31 – 51.
- Spivak, D.I. Wisnesky, R. (2015) "Relational Foundations for Functorial Data Migration". *Proceedings of the International Symposium on Database Programming Languages (DBPL)*, ACM.
- Spivak, D.I. (2014) *Category Theory for the Sciences*. Cambridge: MIT Press. 486 pages.
- Spivak, D.I. (2014) "Database queries and constraints via lifting problems." *Mathematical structures in computer science*.
- Wisnesky, R.; Spivak, D.I.; Schultz, P.; Subrahmanian, E. (2015) "Functorial data migration: from theory to practice". *NIST Interagency/Internal Report (NISTIR)*.

Other Significant Publications

- Giesa, T.; Spivak, D.I.; Buehler, M.J. (2012) "Category theory based solution for the building block replacement problem in materials design". *Advanced Engineering Materials*. DOI: 10.1002/adem.201200109
- Spivak, D.I. (2014) "Categories as mathematical models". To appear in *Categories for the Working Philosopher*.
- Spivak, D.I.; Kent, R.E. (2012) "Ologs: a categorical framework for knowledge representation". *PLoS ONE* 7(1): e24274. doi:10.1371/journal.pone.0024274.

- Forssell, H.; Gylterud, H.K.; Spivak, D.I. (2015) "Type theoretical databases". To appear in: *Logical Foundations of Computer Science*.
- Gross, J.; Chlipala, A.; Spivak, D.I. (2014) "Experience Implementing a Performant Category-Theory Library in Coq". *5th conference on interactive theorem proving (ITP'14)*.

Synergistic Activities

- Taught "Category theory for scientists" in Spring 2013 at MIT, a first-of-its-kind course on applied category theory. The 18 students were from math, materials science, computer science, neuroscience, and other fields. The course textbook *Category Theory for the Sciences* has led to interest from people of a wide variety of backgrounds, both geographically and in terms of mathematical sophistication.
- Hired and worked with Ryan Wisnesky to create a software tool, FQL, which can be used to teach category theory, including left and right Kan extensions as "data migration functors". A company, *Categorical Informatics Inc.* is now being spun out of MIT, to commercialize this product.
- Hired researchers from a variety of background as postdocs. Out of four postdocs hired to date, one is a woman, one is latino, and another is half latino.
- Mentored fourteen undergraduate students of all backgrounds on nineteen research projects in applied category theory since 2010.
- Co-organized the first Computational Category Theory conference at NIST in 2015 and the "Workshop on topology and abstract algebra for biomedicine" at the Pacific Symposium on Biocomputing (PSB) 2016.

PhD Thesis Advisor: Peter Teichner (University of California, Berkeley).

Postgraduate-scholar sponsors: Daniel Dugger, Haynes Miller.

Collaborators: Spencer Breiner (NIST), Markus Buehler (MIT), Adam Chlipala (MIT), Subrahmanian Eswaran (CMU), Henrik Forssell (University of Oslo), Tristan Giesa (MIT), Jason Gross (MIT), Hakon R. Gylterud (Stockholm University), Al Jones (NIST), Eugene Lerman (UIUC), Marco Pérez (University of Mexico), Dylan Rupel (Notre Dame), Patrick Schultz (MIT), Dmitry Vagner (Duke), Ryan Wisnesky (Categorical Informatics Inc).

Former PhD Students: None.

Ryan Wisnesky
(Co-Principal Investigator)

Categorical Informatics, Inc
250 Main St No 426035
Cambridge, MA 02142
ryan@catinf.com

Professional Preparation
Stanford University, BS, Mathematics and Computer Science, 2006
Stanford University, MS, Computer Science, 2006
Harvard University, Ph.D, Computer Science, 2014

Appointments
2014 – 2015: Postdoctoral Associate, Massachusetts Institute of Technology, Math Dept
2015 – present: President, Categorical Informatics, Inc.

Related Publications

- *David I. Spivak, Ryan Wisnesky.* **Relational Foundations for Functorial Data Migration**. Proceedings of the 15th International Symposium on Database Programming Languages (DBPL 2015).
- *Patrick Schultz, David I. Spivak, Ryan Wisnesky.* **Functorial Data Migration: From Theory to Practice**. NIST Interagency/Internal Report (NISTIR) (Publication ID: 919457, 2015)
- *Koutrika, G; Wisnesky, R; Hernandez, M; Krishnamurthy, R; Popa, L* **HIL: A High-Level Scripting Language for Entity Integration**. Proceedings of the 16th International Conference on Extending Database Technology (EDBT 2013)
- *Bogdan Alexe, Douglas Burdick, Mauricio A. Hernandez, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ioana R. Stanoi, Ryan Wisnesky*. **High-Level Rules for Integration and Analysis of Data: New Challenges**. Festschrift celebrating Peter Buneman (PBF 2013).
- *Dessloch, S; Hernandez, M; Wisnesky, R; Radwan, A; Zhou, J* **Orchid: Integrating Schema Mapping and ETL.** Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE 2008).

Five Other Significant Publications

- *Gregory Malecha, Ryan Wisnesky.* **Using Dependent Types and Tactics to Enable Semantic Optimization of Language-Integrated Queries**. Proceedings of the 15th International Symposium on Database Programming Languages (DBPL 2015)

- *Gregory Malecha, Greg Morrisett, and Ryan Wisnesky.* **Trace-based Verification of Imperative Programs with I/O.** Journal of Symbolic Computation Special Issue on the Automated Specification and Verification of Web Systems (JSC-WWV 2010).
- *Adam Chlipala, Gregory Malecha, Greg Morrisett, Avraham Shinnar, and Ryan Wisnesky.* **Effective Interactive Proofs for Higher-order Imperative Programs.** Proceedings of the 14th ACM SIGPLAN International Conference on Functional Programming (ICFP 2009).
- *Gregory Malecha, Greg Morrisett, Avraham Shinnar, and Ryan Wisnesky.* **Toward a Verified Relational Database Management System.** Proceedings of The 37th ACM SIGPLAN - SIGACT Symposium on Principles of Programming Languages (POPL 2010).
- *Ryan Wisnesky, Mauricio Hernandez, and Lucian Popa.* **Mapping Polymorphism**. Proceedings of the 13th International Conference on Database Theory (ICDT 2010).

Five Synergistic Activities

- Maintains an active collaboration with the information-integration department of IBM research, pursuing traditional (non-categorical) approaches to information integration.
- Serves on program committees and reviews papers in the database programming languages space.
- Coordinates open-source software development efforts related to category theory and information management besides the FQL tool described in this proposal (see `categoricaldata.net`).
- Works on automated reasoning projects that, while applicable to the categorical approach to information integration, are also applicable to formal verification (see DBPL, POPL, and ICDT papers above)
- Does outreach to the functional programming community (e.g., by giving talks about categorical databases at Boston's Haskell group)

PhD Thesis Advisor: Greg Morisett (now at Cornell)

Postgraduate-scholar sponsors: David Spivak

Collaborators: Adam Chlipala (MIT), Subrahmanian Eswaran (CMU), Mauricio Hernandez (IBM), Gregory Malecha (UCSD), Patrick Schultz (MIT), Avi Shinnar (IBM) David Spivak (MIT), Lucian Popa (IBM)

Former PhD Students: None.

Joshua Tan

(Entrepreneurial Lead)

Categorical Informatics, Inc
250 Main St No 426035
Cambridge, MA 02142
josh@catinf.com

Professional Preparation
New York University, MS in Mathematics, 2015
Yale University, BA in Ethics, Politics & Economics and Humanities, 2010

Appointments
Business Developer, Categorical Informatics Inc, October 2015-present
Programmer, Ufora, 2014
Co-Founder, MQF, 2010
Founder, Yale College Consulting Club, 2006-2008

MS Thesis Advisor: Misha Gromov and Sylvain Cappell

Jee Chung

(Industry Mentor)

Grantham, Mayo, Van Otterloo & Co. LLC (GMO)
40 Rowes Wharf
Boston, MA 02110
jeechung@alum.mit.edu

Professional Preparation
MIT, BS Physics 1989

Appointments
Head, Enterprise Systems at GMO, 2011-present
VP, Software architecture and application development, State Street Corporation, 2004-2011
Senior Consultant, Ab Initio Software Corporation, 2003-2004
Senior Technical Director, Context Integration, 1997-2003
Data Architect, Fidelity Investments, 1996-1997
Senior Consultant, Context integration, 1994-1996
Database Administrator, Morgan Stanley, 1989-1994