It’s not just about Models and Data: Context is a Catalyst for Insight

Keeping People “In the Loop” with Data and Analytics

Mary Roth, IBM Research, Distinguished Engineer

July 31, 2017

I-SIM 2017
INFORMS SIMULATION SOCIETY 2017 Research Workshop
Data analytics is an iterative, nested, complex process with many people and tools in the loop.
Today, working together is often a series of disconnected interactions. How did we reach this conclusion? How have other teams approached this? What did they DO? How did we reach this conclusion?

Capturing and persisting this ‘in-between’ knowledge is key!
...What if we **connected** these interactions to the data and tools?
We have done 4 user studies, with two user types

**Business Analyst**

Broad study of analysts in multiple roles in a Fortune 100 company

Focused study based on potential users of an enterprise data lake for 900 marketing professionals

**Data Scientist**

In depth study of scientists analyzing genomic data to improve food safety

Limited study of corporate practitioners who work with clients to create insights from data

*Here is what we learned…*
Challenge 1. Getting started is hard…

“As an advanced analytics professional, I spend **80% of my time identifying best sources**, tables …before even creating an analytic model…the **quality of sources and linkages really do matter**”

“When I first started it was like, ‘you need to go to this source, this person, and spin around the room couple of times, and then you will get the answer you need.’”

…and analytics is an iterative, freeform journey

“They continuously get new data and update the references. **Whenever they have a new reference we have to repeat all this mapping again.**”

What’s needed? **Contextual search** to quickly find relevant information and experts
Challenge 2. Teams are diverse…

**Scientific Consortium**

One team, many degrees!

*Engineering, Business, Marketing, Statistics, Biology, Physics, Epidemiology, Anthropology, Computer Science, Food Science*

BA, BS, PhD, MEng

and sharing is *ad hoc*…

“we share code via email”

…but vital:

What’s needed? 

---

**Fortune 100 company**

900+ users

80% understand the data but do not program

“Email is not good. *I need to invest a lot of effort in reading the email and trying to understand sort of what the context of this plot is, …* Often I say, I guess I will look at it later.”

---

*Seamless conversation* as the common denominator *across tools*
Challenge 3. Context is critical!

“Tribal knowledge -- Who is the person that knows this? -> that ends up being an email trail today. If I have a question, I would like the community to provide the answer. In context. With evidence.”

“A big part of it is also knowledge transfer, training...”

“What’s needed? Provenance information captured automatically and transparently

“[People] only understand a slice...”

149 workflows invoked 42,767 times
producing 271,010 files and 116,568 graphs
How can we support the end-to-end collaboration needed? Social networks provide one model of how to approach this

**Context. Conversation. Provenance.**

People with many different backgrounds can collaborate... They see what their friends are doing

They have group discussions. In context.

They ask for recommendations

Search takes their context into account
Our solution is a social network in which people, data and analytics can all participate…

**A collaboration environment for data and analytics**

- Capture **context** and provenance around data and analytics
- Provide a **social, collaborative** user experience
- **Recommend** people, data, applications, visualizations
- Capture provenance information

**A data ecosystem to acquire, catalogue, govern, find, and use data in context**

- **Curate** and **contribute** data
- Search and browse **data catalog**
- **Provision** data to analytic tools
- **Govern** data access
...Using familiar tools to contribute their expertise

**Domain Experts** contribute semantics and business meaning

**IT Staff** govern, prepare and contribute integrated data

**Data Scientists** contribute data processing and analytics expertise

**Executives** contribute business insights through conversation and comments

**Analysts** contribute domain expertise, analysis and results
Example: Using next generation sequencing data for Food Safety

- **Technicians**: Collect samples and record metadata about samples in community notebooks.

- **Biologists**: Perform DNA/RNA extraction and sequencing, record metadata, upload results to community notebook.

- **Bioinformatics Experts**: Launch apps, upload results, perform analytics on sequencing output and upload results and visualizations to community notebook.

- **Quality Assurance Engineers**: Review results and make recommendations.

- **Metagenomics Community**: Read/write notes. Samples Notebook, DNA/RNA Notebook, BLAST Results, Annotation App, BLAST App.

- **Metagenomics Analytics Software**: Review results, launch apps, upload results.

149 workflows invoked 42,767 times producing 271,010 files and 116,568 graphs.
Example: Using next generation sequencing data for Food Safety

Workflows are run, and data and results are visualized and interpreted in the context of a group discussion.

David runs the workflows and posts analysis results showing correct eukaryote identification of matrix (fish)

Niina confirms correct identification of species

Bobby asks a question about length distribution of RNA fragments.

And then provides the answer to his question
Example: Adding social context and self-service to an Enterprise Data Lake

900+ users
80% understand the data but do not program

Comments enable users to share tribal knowledge, enabling contextual search.

Recommendations accelerate productivity

Ranking and use statistics provide informal measure of quality and utility.
Example: Adding social context and self-service to Watson Analytics

Activity streams show what colleagues are working on in related tools

Contextual search includes colleagues, queries and data related to the search terms. And ranked results provide statistics showing use.
At the core is a graph that captures **context** about how people work.

**Executives** contribute business insights through conversation and comments.

**Data Scientists** contribute analysis and results.

**IT Staff** govern, prepare and contribute integrated data.

**Analysts** contribute semantics and business meaning.
What context do we capture in the graph?
The graph is both populated and consumed by 3rd party tools.
The graph is indexed to enable contextual search and recommendations

All nodes are indexed, including comments and tags

Relationships in the graph are used for ranking and to establish trust and quality in the results

The graph contains the user’s context, which is used to return the most relevant results
Provenance information is captured via embeddable widgets

Embeddable widgets link data to code, capture comments, suggest and record provenance based on code analysis.
...contributing provenance information to the larger project

The graph captures execution history, as well as input and output

Enabling graphical exploration of provenance data across multiple applications
PowerBlock, a clean energy start up, is working to develop a model to predict the energy efficiency of thousands of older real estate properties. This model will be used to identify potential buildings for investors seeking to finance green energy retrofits.

The team consists of:
- Harsha, the business lead
- Kevin, the analyst
- Jorge, the data scientist
- Horst and Ignacio, who did a similar project last year
Harsha kicks off a group discussion in context

Harsha, the business lead, provides a sample data file and asks the team to create a model to predict which buildings in NYC might be energy inefficient.

Jorge and Kevin get to work. They notice the sample is quite small, and decide to look for other projects that have done something similar.
Kevin quickly finds related projects and experts.

Kevin does a search on ‘energy’, and the search result returns data, people and previous projects that involved energy.

Kevin finds Ignacio and Horst, whom he was unaware were working on energy projects.
“Jorge, our field engineers in New York City are visiting too many unqualified properties, and it is costing us too much time and money. Can we build a model to predict which commercial properties have inefficient energy usage pattern?”

How does contextual search work?
Kevin also finds all of the context for the energy project Horst and Ignacio worked on.

This context shows they did a similar project in Boston. It not only contains the data, but also a link to their Jupyter notebook and their insights as to which machine learning algorithm worked the best.
Kevin shares what he learned with the team.

Kevin attaches the context and the data files to a note and adds a comment to share the insight into which machine learning algorithm to use.

Jorge provisions the data file directly from the conversation to a Zeppelin notebook.
Jorge builds a predictive model and adds his results to the conversation.

Jorge uses the Boston and NYC data and Machine Learning algorithm from the Boston project and shares his results from Zeppelin directly into the conversation.
Harsha validates Jorge’s results using Watson Analytics

Harsha is able to provision Jorge’s data directly to Watson Analytics to validate his results. And, he can capture his work in Watson Analytics and add it directly to the conversation.
Conclusion

To accelerate data analytics and help the humans in the loop, we must
– Help them get started
– Help diverse (and often distributed) teams work together effectively
– Help them track their progress and understand context and provenance

A conversational paradigm, inspired by social networking, is one powerful approach

The magic ingredients include
– A rich metadata graph with schematic, semantic, collaborative and usage metadata
– Powerful services driven by that metadata and invoked as part of a conversation
– Tracking app invocations, dialog and results as part of the conversational record

This allows humans to stay firmly in the center of the analytic process while progressing faster through their investigations, with better results.