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

Keeping People “In the Loop” with Data and Analytics

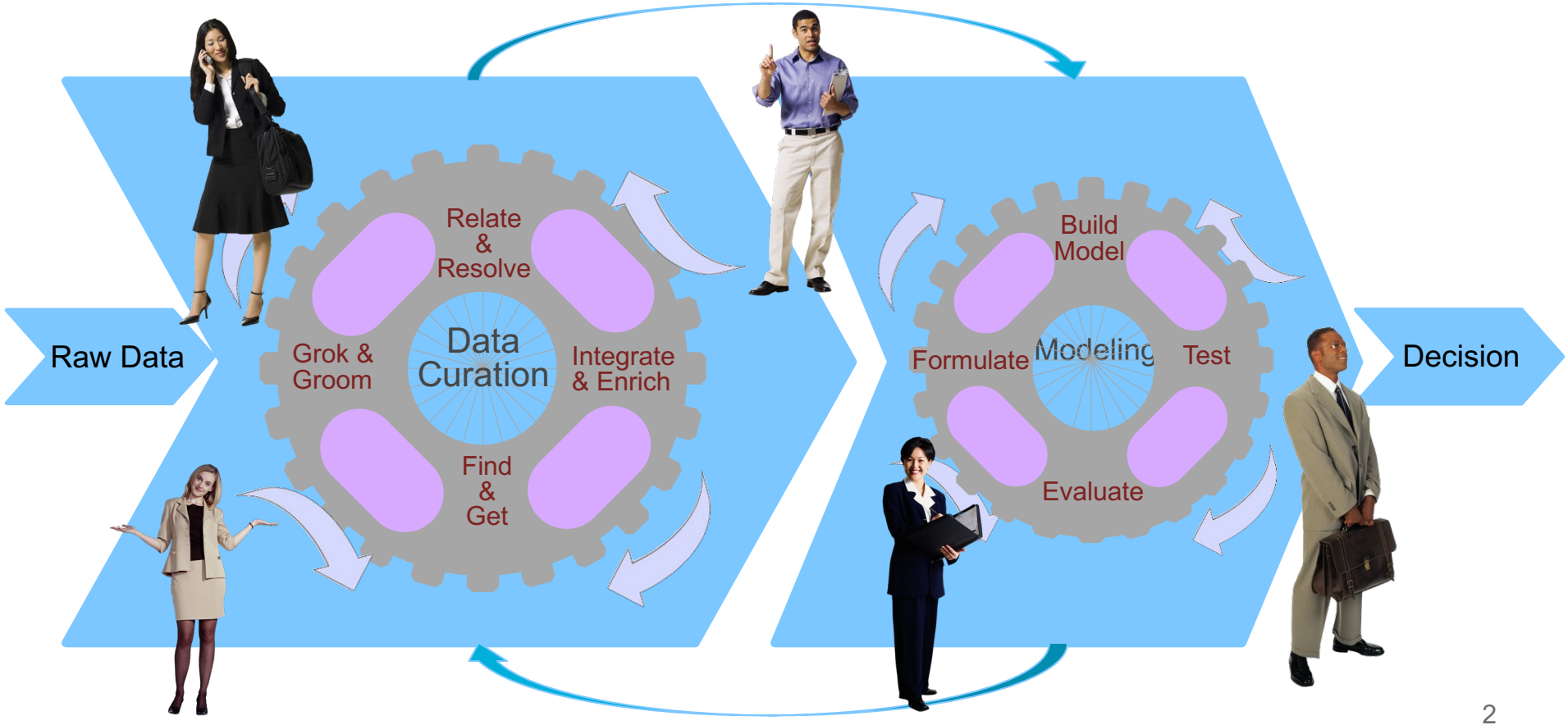
Mary Roth, IBM Research, Distinguished Engineer

July 31, 2017

I-SIM 2017
INFORMS SIMULATION SOCIETY 2017
Research Workshop

IBM

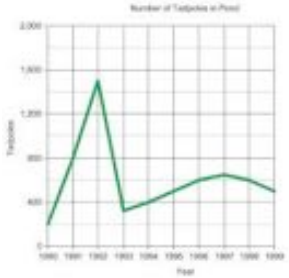
Data analytics is an iterative, nested, complex process with many people and tools in the loop



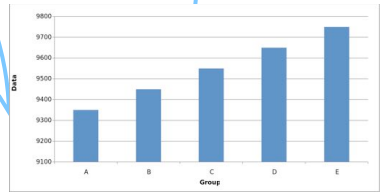
Today, working together is often a series of disconnected interactions

How have other teams approached this?

What did they DO?

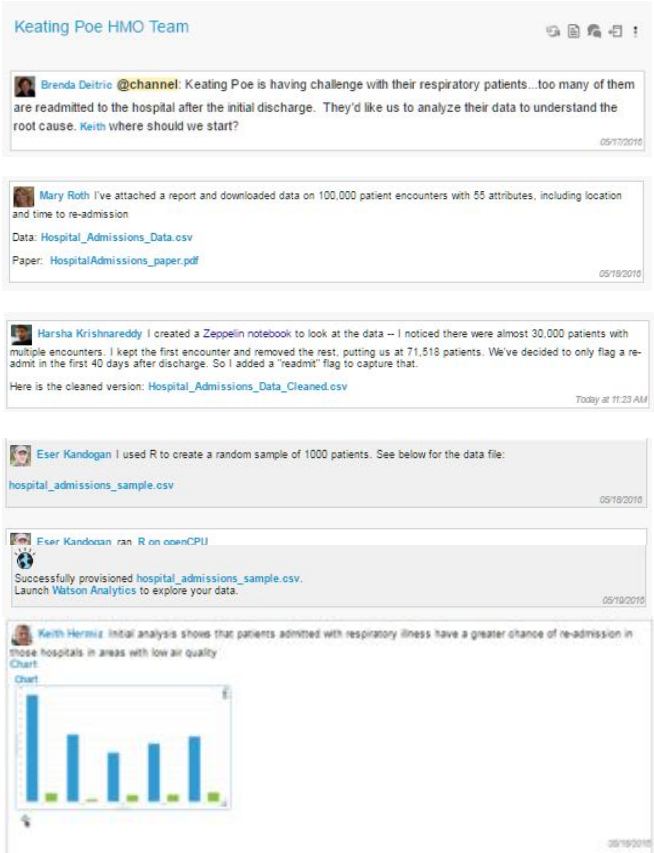


How did we reach this conclusion?



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:

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."

What's needed? → ***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...**"

"**[People] only understand a slice...**"

149 workflows invoked 42,767 times producing 271,010 files and 116,568 graphs

What's needed? →

***Provenance information* captured automatically and transparently**

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.

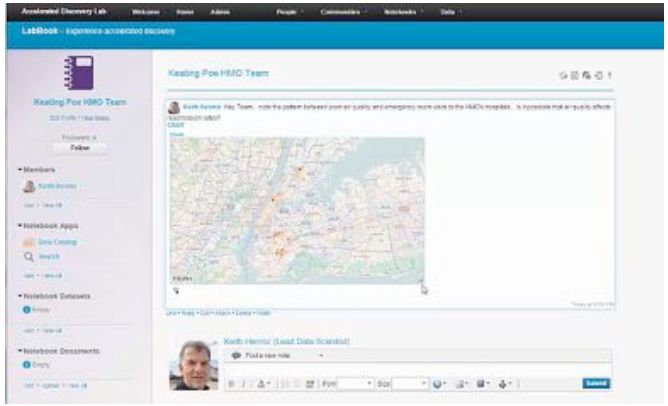
They ask for recommendations

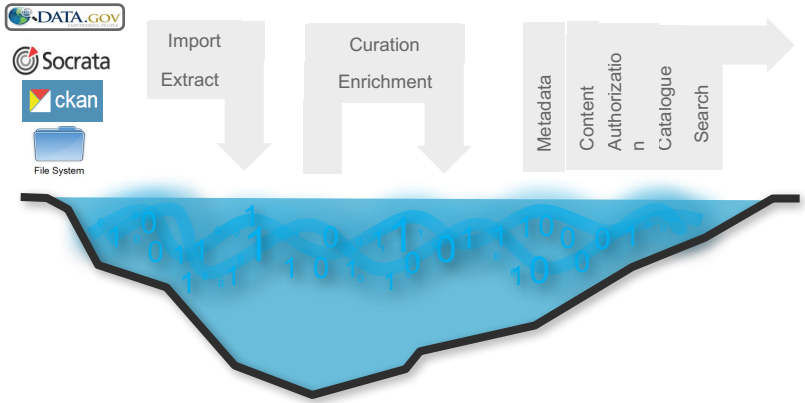
Search takes their context into account

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

slack

Executives contribute business insights through conversation and comments

Keating Poe HMO Team

Brenda Deitric @channel: Keating Poe is having challenge with their respiratory patients...too many of them are readmitted to the hospital after the initial discharge. They'd like us to analyze their data to understand the root cause. Keith where should we start?

Mary Roth I've attached a report and downloaded data on 100,000 patient encounters with 55 attributes, including location and time to re-admission

Data: Hospital_Admissions_Data.csv

Paper: HospitalAdmissions_paper.pdf

05/18/2016

Harsha Krishnareddy I created a Zeppelin notebook to look at the data -- I noticed there were almost 30,000 patients with multiple encounters. I kept the first encounter and removed the rest, putting us at 71,518 patients. We've decided to only flag a re-admit in the first 40 days after discharge. So I added a "readmit" flag to capture that.

Here is the cleaned version: Hospital_Admissions_Data_Cleaned.csv

Today at 11:23 AM

Eser Kandogan I used R to create a random sample of 1000 patients. See below for the data file:

hospital_admissions_sample.csv

05/18/2016

Successfully provisioned hospital_admissions_sample.csv.
Launch Watson Analytics to explore your data.

Keith Hermiz Initial analysis shows that patients admitted with respiratory illness have a greater chance of re-admission in those hospitals in areas with low air quality

Chart

Analysts contribute domain expertise analysis and results

IBM Watson Analytics

AnyLogic

Arena Simulation Software

slack

Domain Experts contribute semantics and business meaning

DataWorks IBM

IT Staff govern, prepare and contribute integrated data

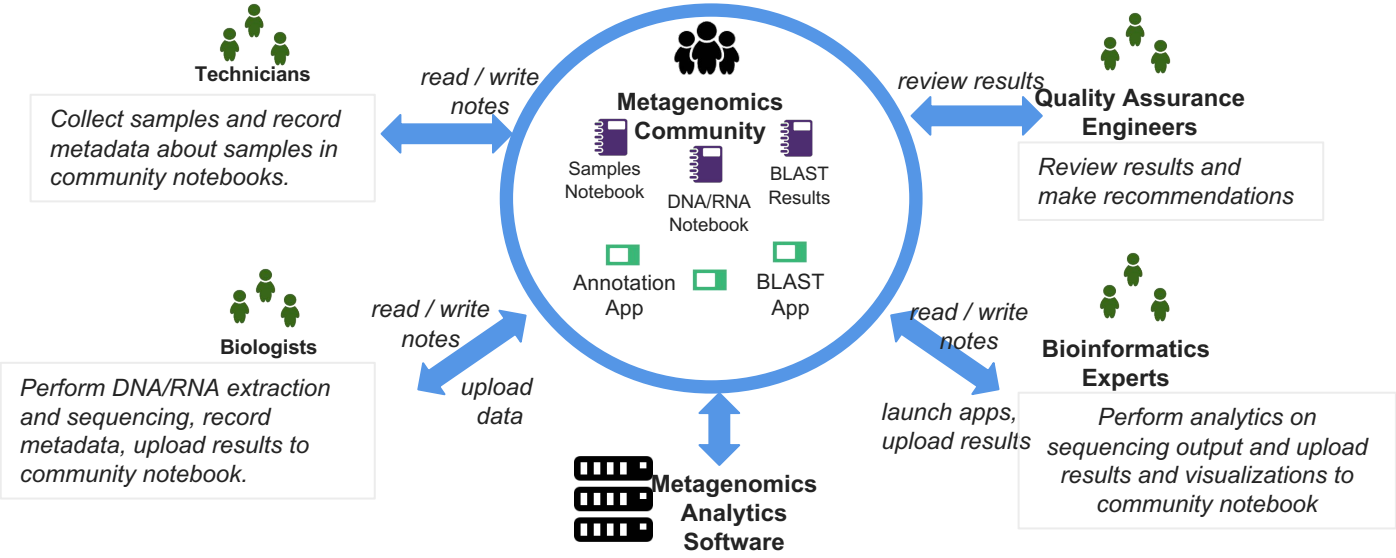
Zeppelin

jupyter

Data Scientists contribute data processing and analytics expertise

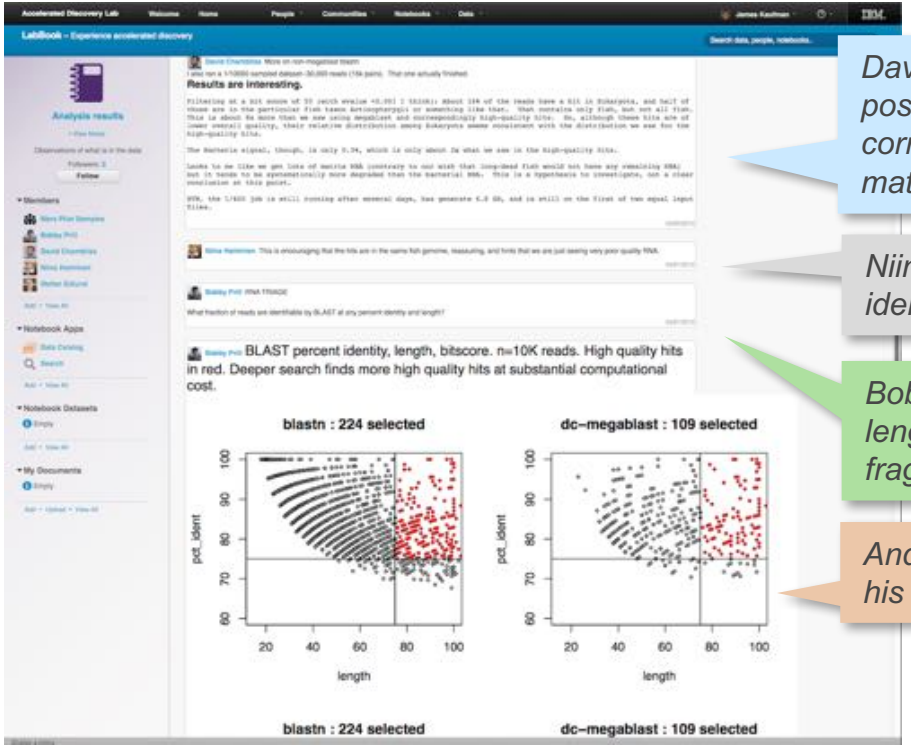
Example: Using next generation sequencing data for Food Safety

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

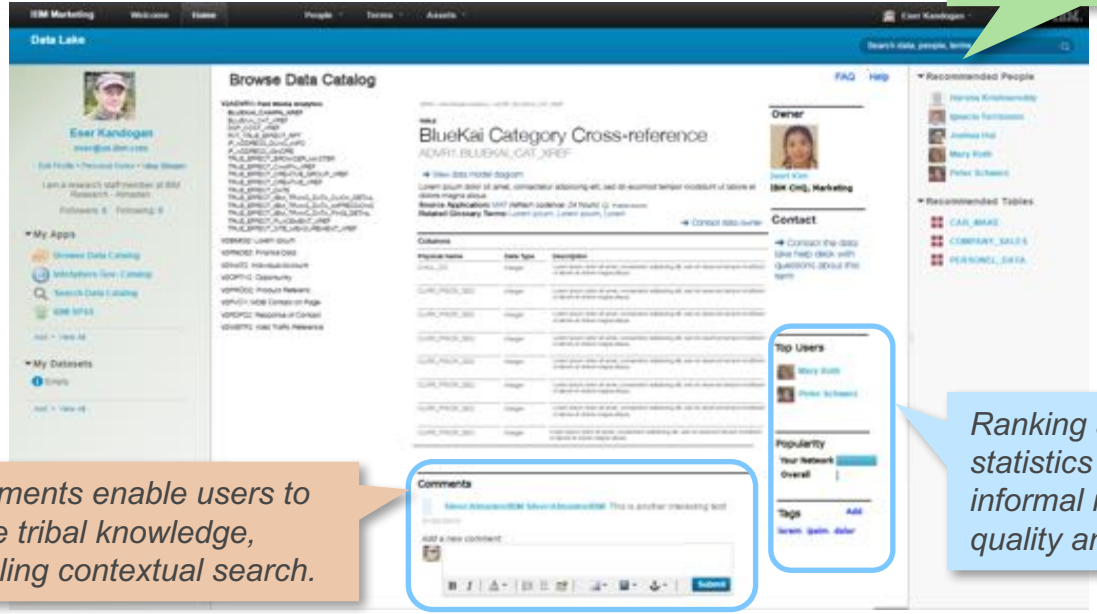
Bobby asks a question about length distribution of RNA fragments.

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

Recommendations accelerate productivity



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

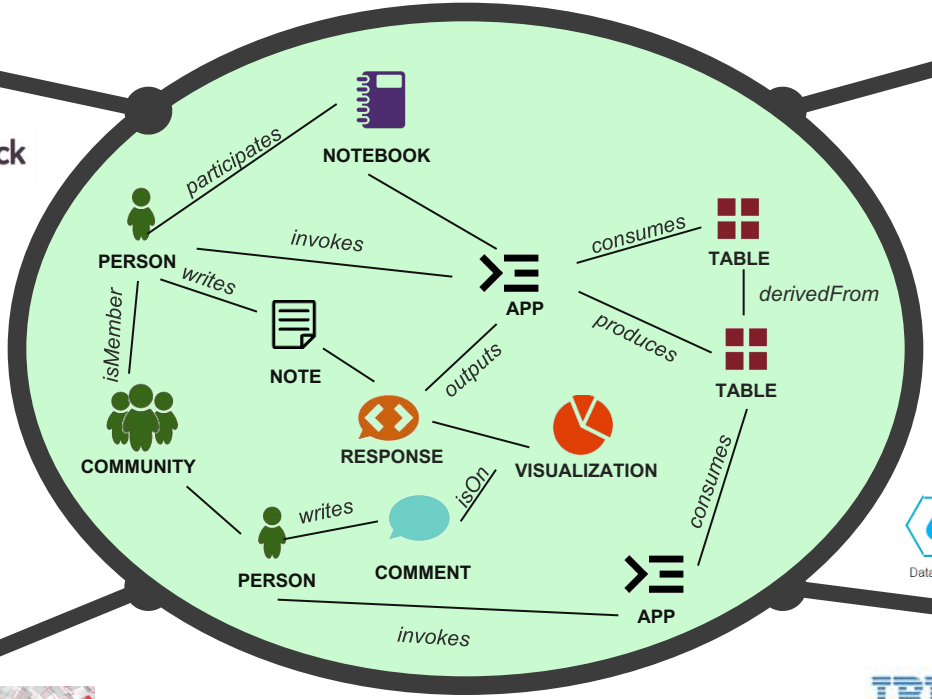
Ranking and use statistics provide informal measure of quality and utility.

Example: Adding social context and self-service to Watson Analytics

At the core is a graph that captures context about how people work

Executives
contribute business insights through conversation and comments

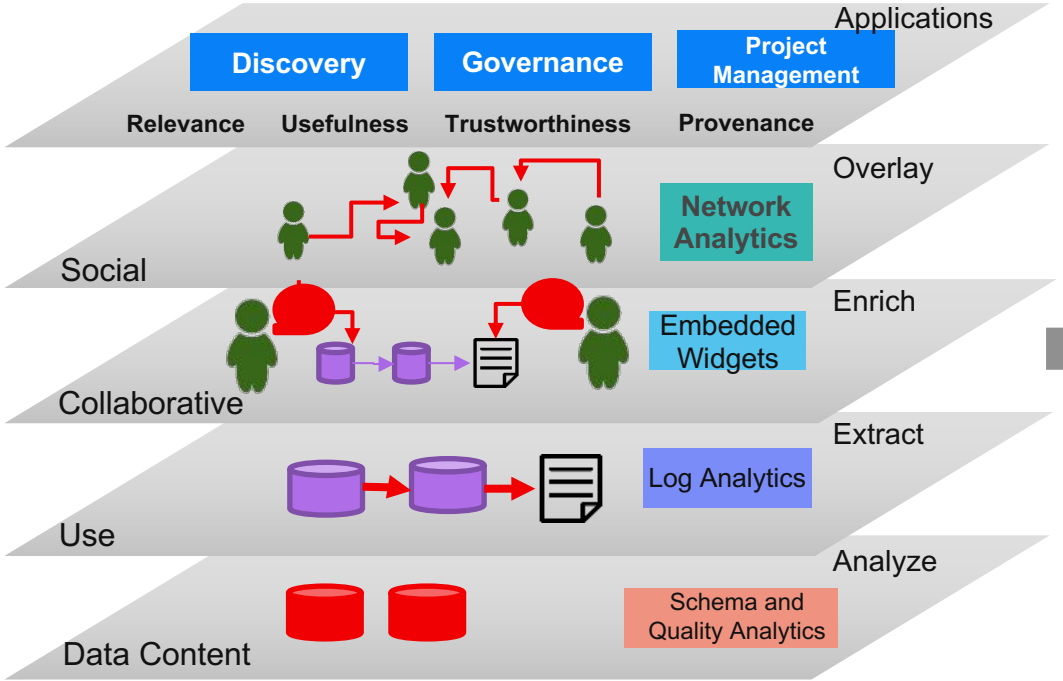
Analysts
contribute semantics and business meaning

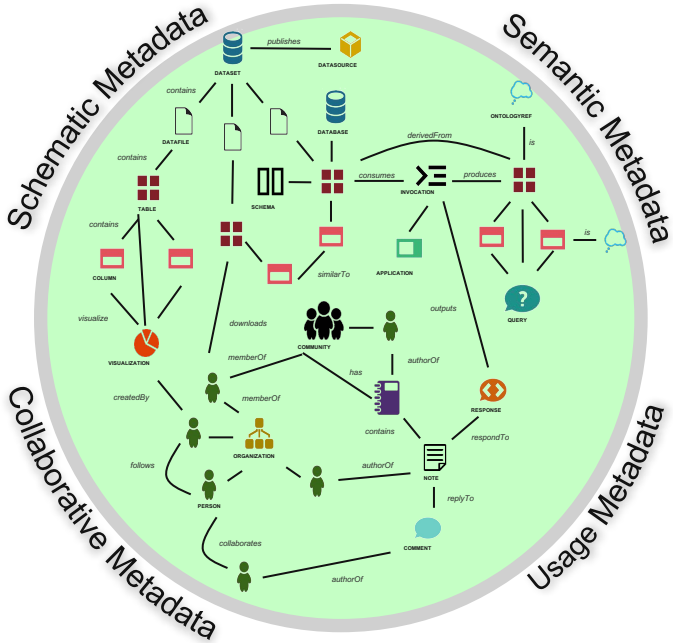


Data Scientists
contribute analysis and results

IT Staff
govern, prepare and contribute integrated data

What context do we capture in the graph?





The graph is both populated and consumed by 3rd party tools

Source systems

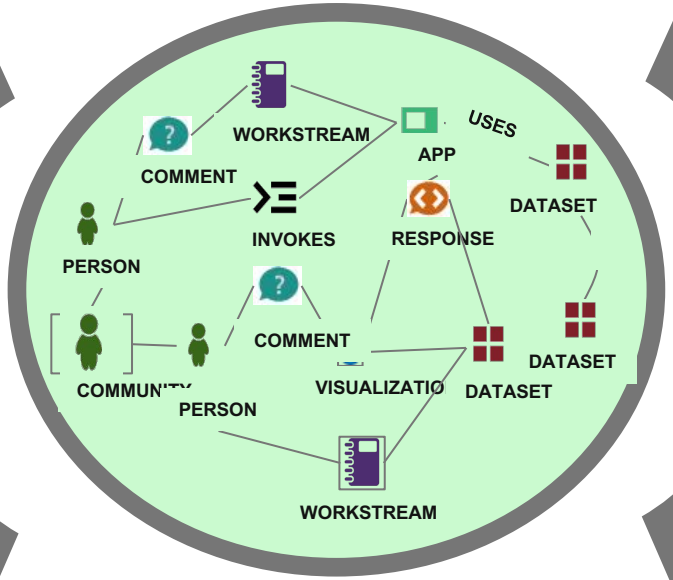
Social Networking Tools

Analyst Tools

Data Integration Tools

REST API + LOG EXTRACTION

Contextual Usage Graph



REST API

Embeddable Widgets

Social Widgets

Contextual Search

Activity Streams

Contextual Graph Browser

Recommendations

User Interfaces

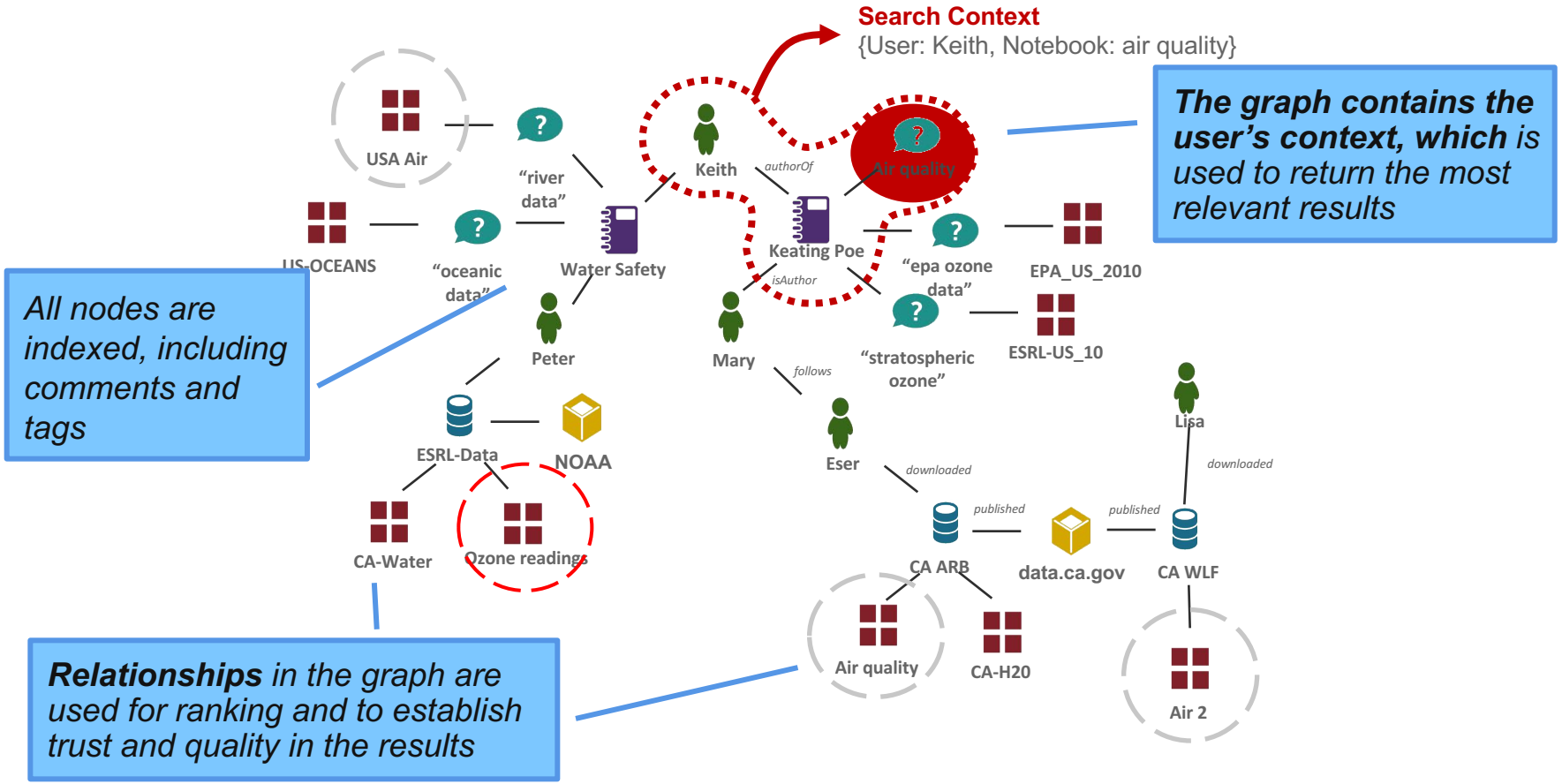
Business users

Analysts

Data Scientists

IT staff

The graph is indexed to enable contextual search and recommendations



Provenance information is captured via embeddable widgets

Zeppelin Notebook Interpreter Datasets Projects Connected

Bank Customer Analysis default

Load Data Into Table FINISHED

```
import org.apache.commons.io.IOUtils
import java.net.URL
import java.nio.charset.Charset

// Zeppelin creates and injects sc (SparkContext) and sqlContext (HiveContext or SqlContext)
// So you don't need create them manually

// load bank data
// @step { "input": [{ "alias": "abc", "location": "/tmp/abc.txt", "file_type": "datafile" }, {
"alias": "abc2", "location": "/tmp/abc2.txt",  "file_type": "datafile" }], "output": [{ "alias": "abc
-output", "location": "/tmp/abc-output2.txt",  "file_type": "datafile" }], "comments": "this is trial"
, "additional_params": { "key1": "value1", "key2": "value2" } }

val bankText = sc.parallelize(
    IOUtils.toString(
        new URL("https://s3.amazonaws.com/apache-zeppelin/tutorial/bank/bank.csv"),
        Charset.forName("utf8")).split("\n"))

case class Bank(age: Integer, job: String, marital: String, education: String, balance: Integer)

val bank = bankText.map(s => s.split(";")).filter(s => s(0) != "\"age\"").map(
    s => Bank(s(0).toInt,
            s(1).replaceAll("\"", ""),
            s(2).replaceAll("\"", ""),
            s(3).replaceAll("\"", ""),
            s(5).replaceAll("\"", "").toInt
        )
```

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

```
import java.nio.charset.Charset
bankText: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[521] at parallelize at <console>:124
defined class Bank
bank: org.apache.spark.sql.DataFrame = [age: int, job: string, marital: string, education: string, balance: int]
```

Comments

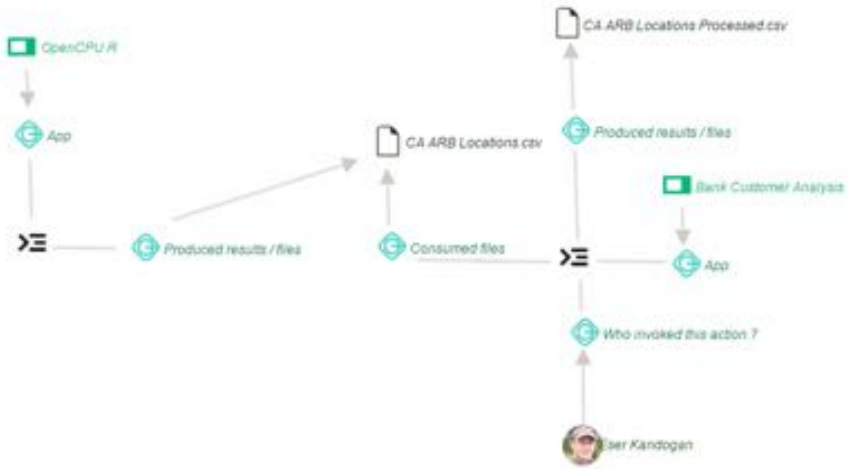
Eser Kandogan This script will load bank.csv from AWS

Eser Kandogan The last time it failed for ...

Execution History

Eser Kandogan executed Yesterday at 11:51 AM and took 3 seconds [Explore Provenance]
Consumed:
Produced: ActiveLabbookUsers.csv, CA ARB Locations.csv

Eser Kandogan executed Today at 9:08 AM and took 2 seconds [Explore Provenance]
Consumed:
Produced: ActiveLabbookUsers.csv, CA ARB Locations.csv



Enabling graphical exploration of provenance data across multiple applications

Demo: Building a model to predict energy efficiency for light commercial properties

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:

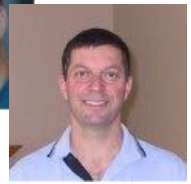
- Harsha, the business lead
- Kevin, the analyst

slack

jupyter

Zeppelin

- Jorge, the data scientist



- Horst and Ignacio, who did a similar project last year

Harsha kicks off a group discussion in context

PowerBlock

Harsha S. Kapre JORGE and team: Our field engineers in New York City are visiting too many unqualified properties, and it is costing us too much time and money. We have so much data -- can we build a model to predict which commercial properties have inefficient energy usage pattern? Here is a sample of buildings they have already identified in New York City. Can you build a model from this?

NYCBuildings.csv

JORGE A. CASTANON Hi Harsha. Yes, we can work on this. Kevin, this data
can leverage?

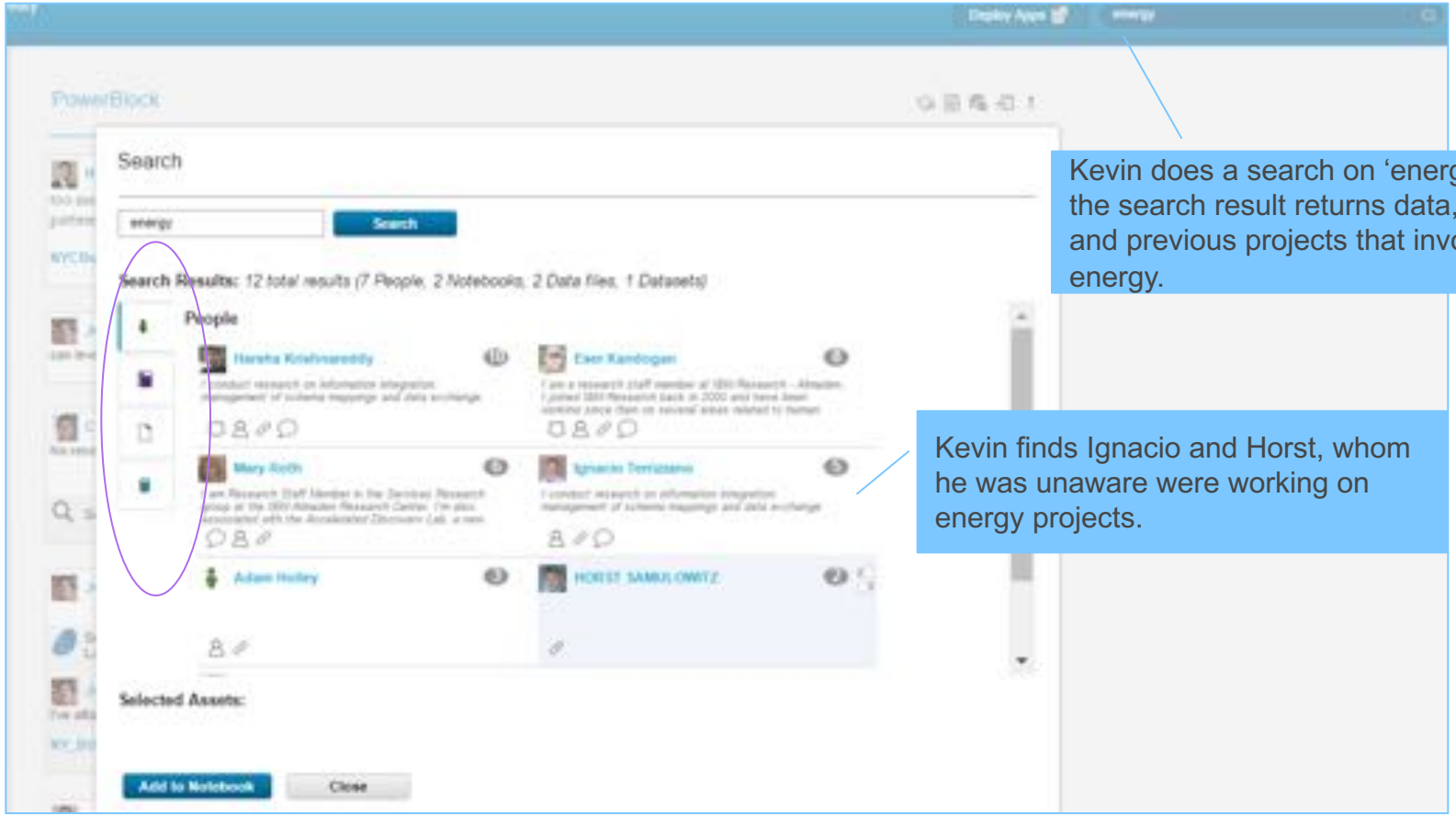
Charles K. Shank ran Search Jorge, Horst worked on a similar project last year. I think we should be able to combine his data with this new file, and use his recommendations for the right ML algorithm in order to build a predictive model. I'm attaching the dataset with all of the information here.

Selected: Boston Restaurants Energy Usage, Boston Restaurants Energy Usage

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..

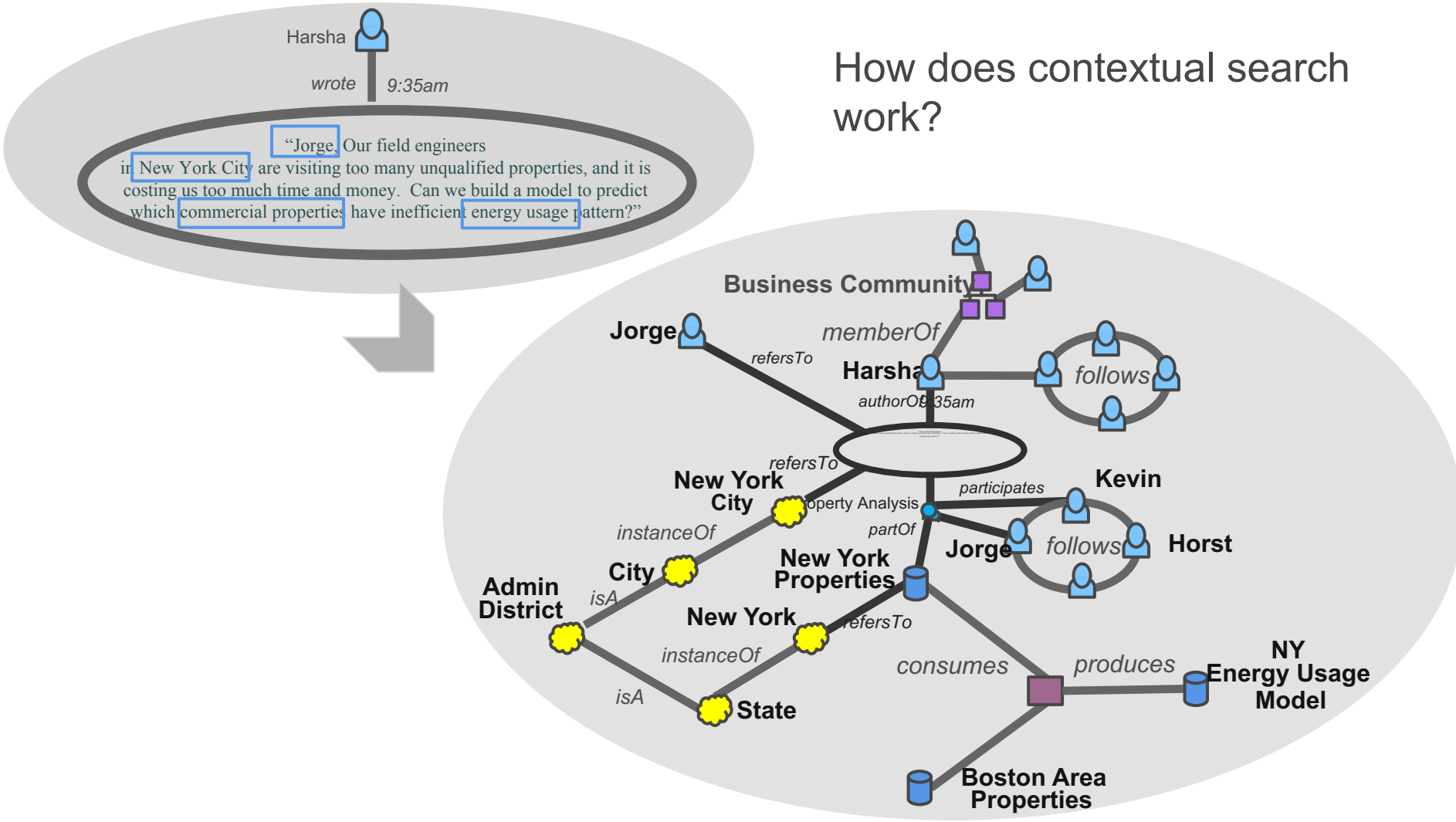
Kevin quickly finds related projects and experts



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.

How does contextual search work?



Kevin quickly finds related projects and experts

Kevin also finds all of the context for the energy project Horst and Ignacio worked on.

PowerBlock

Search

energy

Search

Search Results: *12 total results (7 People, 2 Notebooks, 2 Data files, 1 Datasets)*

Notebooks

PowerBlock

Boston Restaurants Energy Usage

Boston Restaurants Energy Usage

Ignacio Terrizzano Horst, I am working with the city of Boston to monitor energy usage by restaurants. I've attached some data on over 2000 restaurants. Can you detect any patterns in the data?

BOSTON_RESTAURANTS_ENERGY_USAGE_DATA.csv

HORST SAMULOWITZ

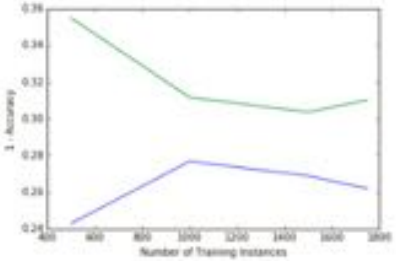
Yes, I have created a Jupyter notebook to run CADS. I will then provide the best ML algorithm trajectory for the data.

Boston Restaurants Energy usage analysis

HORST SAMULOWITZ

After looking at the CADS analysis, I am confident that using R SVM Classification is a suitable approach. There is a nice downward tendency on the validation set's performance error as the number of instances increases, as depicted in the chart.

Performance Error Graph for currently 2nd-ranked pipeline
[TH Finished] Data -> FS: Weka Chi Squared -> Sampler -> CL: R SVM
Training (blue) VS Validation (green)



This context shows they did a similar project in Boston. It not only contains the data, but also a link to their Juypter 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.

Charles K. Shank ran Search Jorge, Horst worked on a similar project last year. I think we should be able to combine his data with this new file, and u
his recommendations for the right ML algorithm in order to build a predictive model. I'm attaching the dataset with all of the information here.

Selected: Boston Restaurants Energy Usage, Boston Restaurants Energy Usage

JORGE A. CASTANON ran Zeppelin Data Provision

Successfully provisioned NYCBuildings.csv and BOSTON_RESTAURANTS_ENERGY_USAGE_DATA.csv.
Launch Zeppelin to explore your data.

JORGE A. CASTANON Harsha – Kevin and I were able to build a model based on data from field engineers in Boston as well as your New York City data.
I've attached a file with the predictions for New York here. How do the results compare?
NY_BUILDING_ENERGY_EFFICIENCY_PREDICTION.csv

Harsha S. Kapre ran Watson Analytics Data Provision

Successfully provisioned clusterEnergyLocation.csv.
Launch Watson Analytics to explore your data.

Harsha S. Kapre Wow! The results are promising & your model has identified several of
really save us time!
domestic_gas 01

Zeppelin Notebook Job Datasets Projects

Exploring datafile: NYCBuildings.csv

```
%spark
```

```
%sh
cd /home/hvkrishn/zepp/notebook_datastore
mkdir 2BYWYTDWK
```

```
import com.ibm.labbook.client.Labbook;

var labbook = new Labbook("https", "pond.almaden.ibm.com", 9420, "hvkrishn@us.ibm.com", "<labbook-password>")
    labbook.getFile("DATAFILE_de0c81ca-21a6-4749-a3ff-5fe8415336d3", "NYCBuildings.csv", "/home/hvkrishn/zepp/notebook_datastore/2BYWYTDWK");

//Data file downloaded to:  /home/hvkrishn/zepp/notebook_datastore/2BYWYTDWK/NYCBuildings.csv
```

Jorge builds a predictive model and adds his results to the conversation

Zeppelin Notebook Job Datasets Projects

PowerBlock - Buildings Energy Efficiency Pred...

```
%spark.pyspark

//Adding the result file back to Labbook
labbook.postFile("DATASET_a64b9576-2dfa-44d0-94cf-2aa73d2ed10b", "/home/hvkrishn/zepp/notebook_datastore/2BXU3MHK4/NY_BUILDING_ENERGY_EFFICIENCY_PREDICTION.csv");
```

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.

Charles K. Shank ran ... ink we should be able to combine his data with this new file, and use his recommendations for the ... ing the dataset with all of the information here.

Selected: Boston ... rgy Usage

JORGE A. CASTANON ran Zeppelin Data Provision

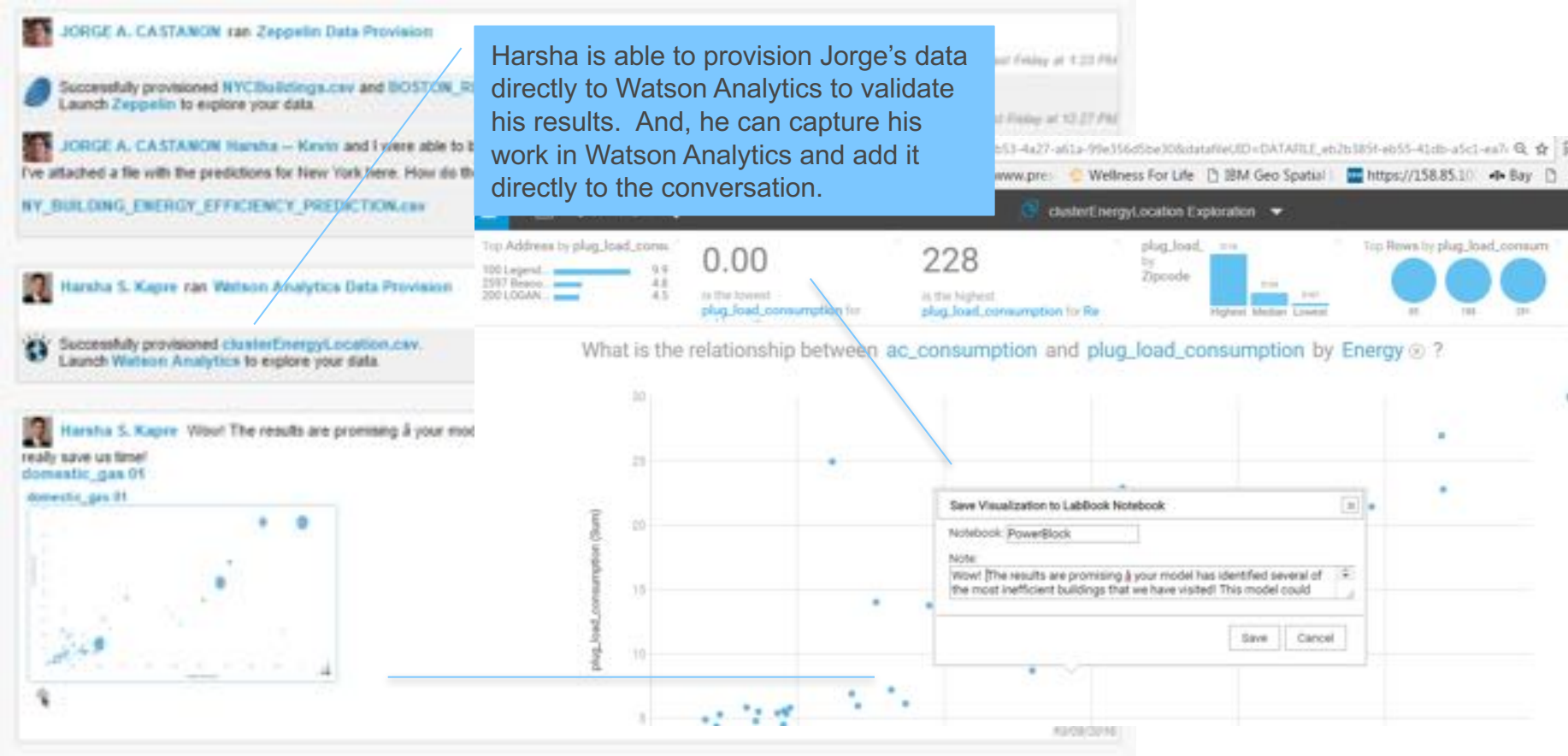
Successfully provisioned NYCBuildings.csv and BOSTON_RESTAURANTS_ENERGY_USAGE_DATA.csv.
Launch Zeppelin to explore your data.

JORGE A. CASTANON Harsha -- Kevin and I were able to build a model based on data from field engineers in Boston as well as your New York City data. I've attached a file with the predictions for New York here. How do the results compare?

NY_BUILDING_ENERGY_EFFICIENCY_PREDICTION.csv

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 loopS, 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.