Datameer®

Easier, Faster, Smarter

Data Science without the Scientist

Matt Schumpert
10.30.13

Agenda

- Background
- First principles
- Mind-blowing fun fact
- Current state & challenges
- Suggestions for making life easier
- Demo!

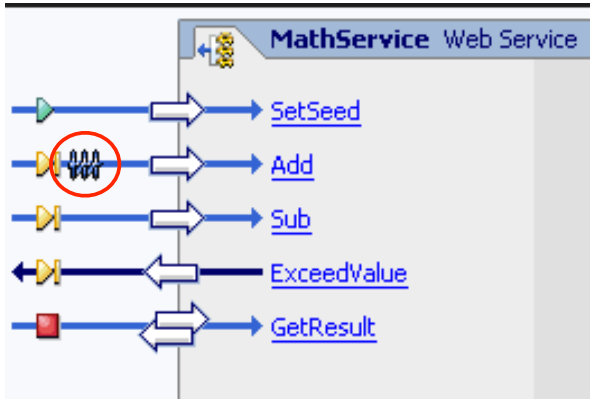
Me

- Enterprise infrastructure software guy
- Focused on abstraction and customers
- Likes simplicity

A favorite example...

Buffered Web Services:

"When a buffered operation is invoked by a client, the method operation goes on a JMS queue and WebLogic Server deals with it asynchronously by transparently creating a Message Driven Bean to consume the message. As with Web Service reliable messaging, if WebLogic Server goes down while the method invocation is still in the queue, it will be dealt with as soon as WebLogic Server is restarted. When a client invokes the buffered Web Service, the client does not wait for a response from the invoke, and the execution of the client can continue"



1. First Principles

First Principles from an Expert

- Instrument everything
- Invest in infrastructure
- Put all your data in one place
- Data first, questions later
- Keep raw data forever
- Let everyone party on the data
- Produce tools to support the whole lifecycle

- Jeff Hammerbacher

2. Mind-boggling fun fact

190,000 unfilled data scientist jobs by 2018

-McKinsey

Signal-to-Noise Ratio is Dropping!

3. Current state + challenges

Hallmarks of Traditional Analytics

- Esoteric skills
- Long cycle times
- Low transparency
- Data & application silos
- Mired in data prep
- Sampling (guesstimation)
- Expensive!
- Extremely valuable work products

Current Recipe:

- Pull historical data
- Sample
- Cleanse / Pre-process
- Design / implement model
- Train
- Hand-code / Integrate
- Deploy
- Fine-Tune, rinse and repeat

Science != Everyday Decisions

There **must** be a better way!

Apply traditional tools to big data?

SAS	R	Mahout
Expensive Not Scalable Silo'ed	Requires Coding Retraining Clunky Architecture	Coding Required Immature Limited Support

And what about the rest of the (big data) story?

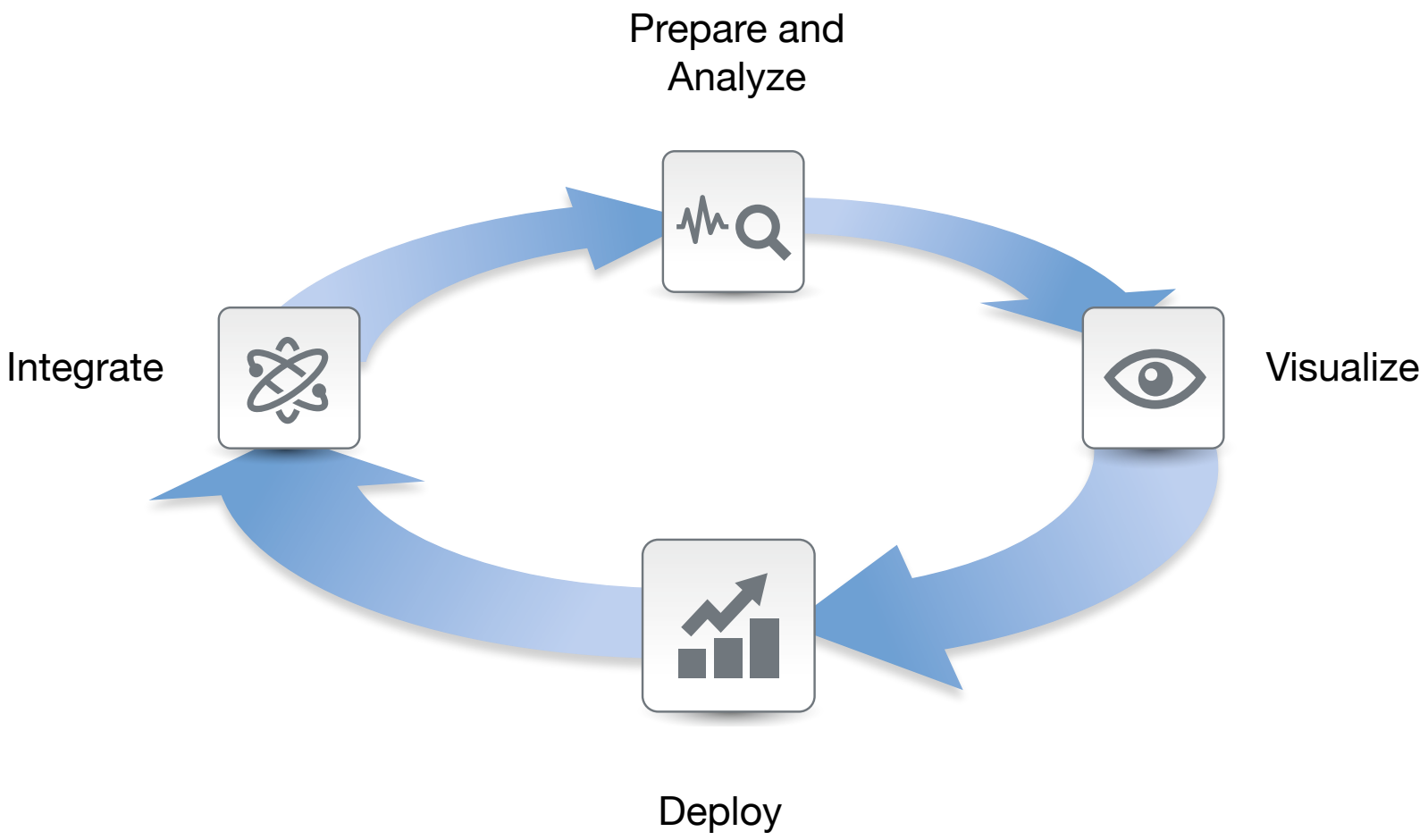
Big Data Analytics is NOT (just):

- A sexy new visualization tool
- Machine learning / Predictive analytics
- Data science
- Hadoop
- The data warehousing movie replayed

Big Data Analytics IS:

- A granular, complete and current understanding of your operations and customers
- Answering questions at the speed of business
- Relevancy in all customer interactions
- Closed-loop decisioning that's data-driven
- *Managing* data through a ***lifecycle***

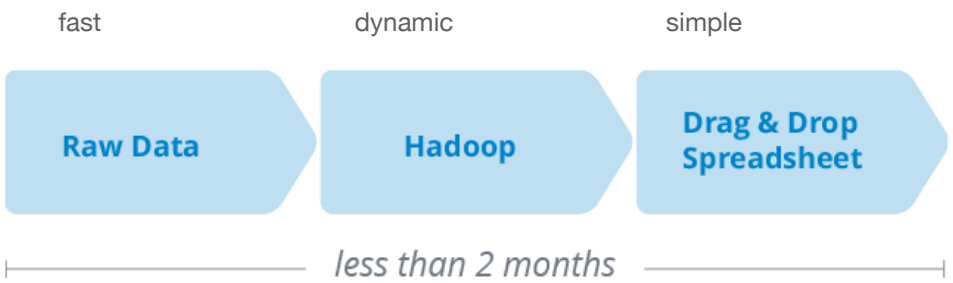
The Big Data Analytics Lifecycle



A lesson from data warehousing / BI

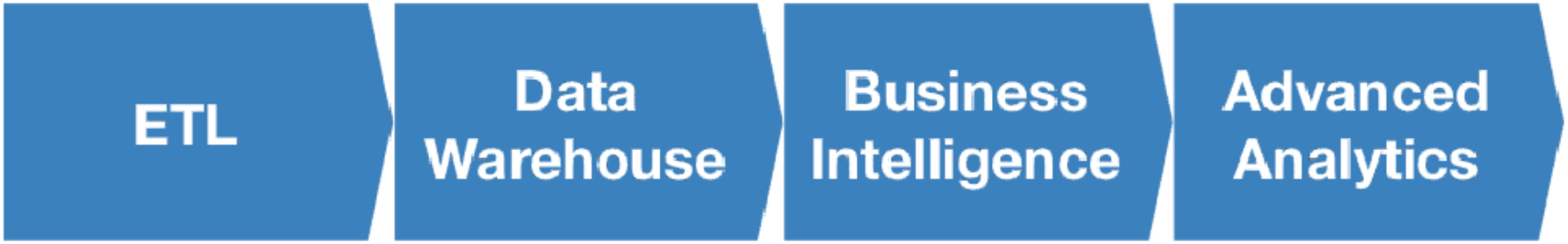
traditional / schema-on-write:

agile / schema-on-read:



Source: TDWI

Don’t rebuild Rome... again!!



ETL	Data Warehouse	Business Intelligence	Advanced Analytics
$250K	$300K	$100K	$300K
4 People	3 People	8 People	5 People
2 months	3 months	3 months	2 months

There **must** be a better way!

4. Making life easier

How (without army):

- Speak the language of the business
- Generate (don’t write) code
- Simplify data integration and preparation
- Move the computation (analytics) to the data

Esoteric Language == Obscurity

K-Means

CART

Mutual Information

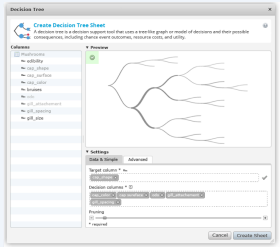
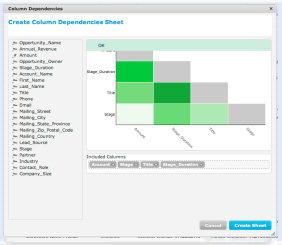
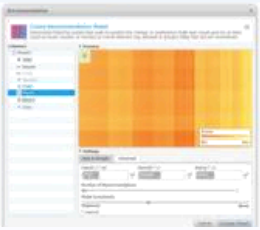
Matrix Factorization

Random Forest?

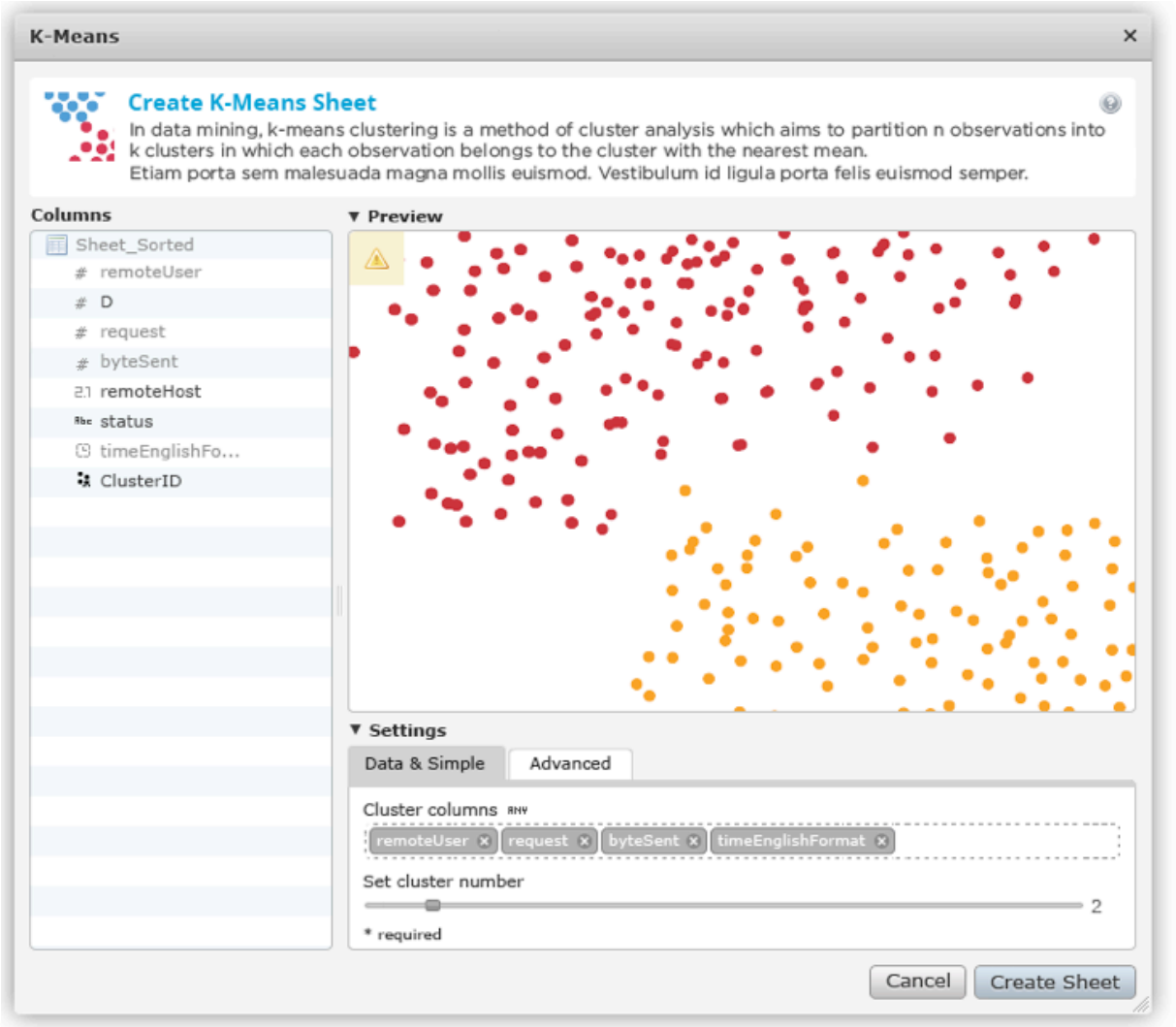
Logistical Regression

Support Vector Machine??

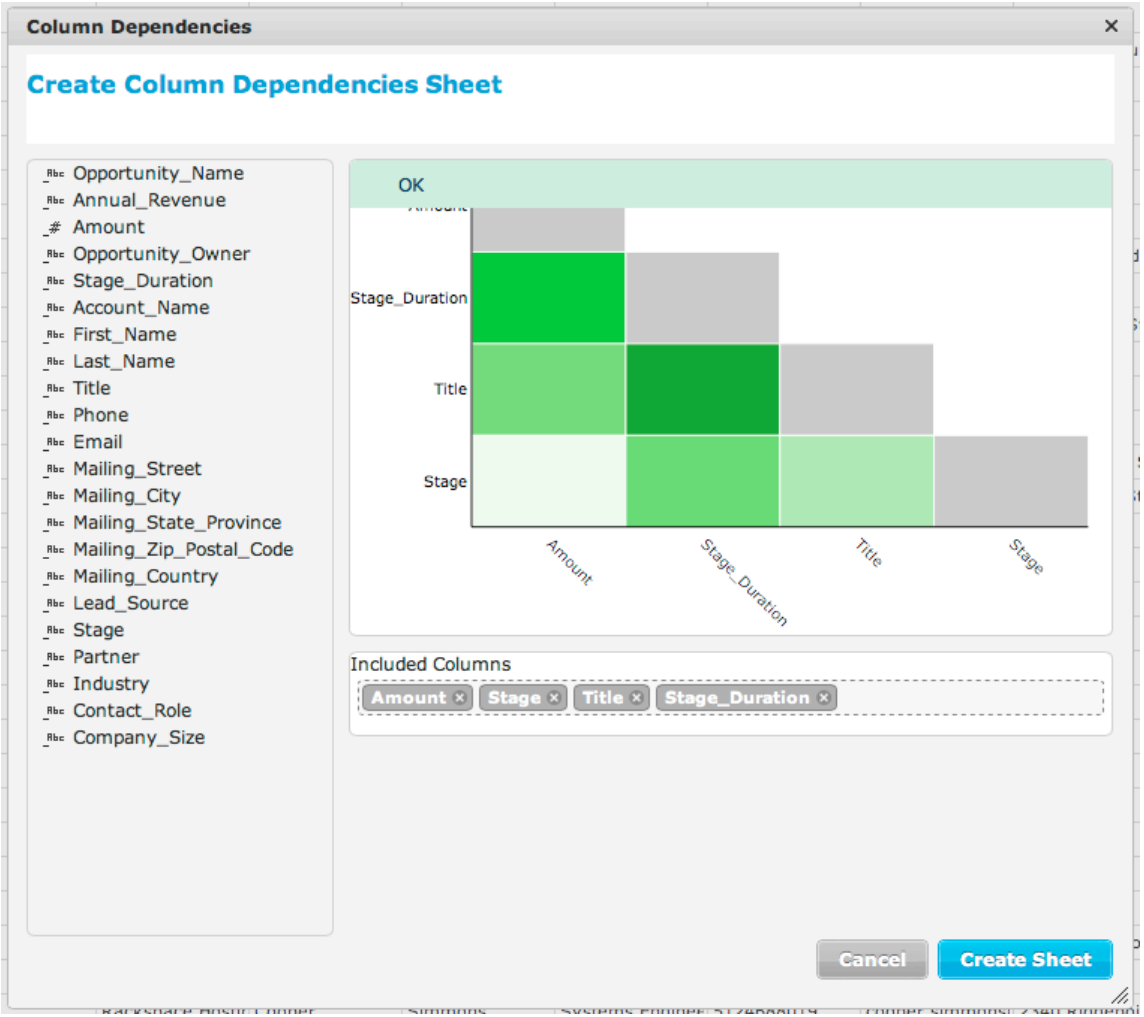
Algorithms can be straightforward!

Algorithm		Description
Clustering		Automatically finds patterns to group data
Decision Tree		Automatically identifies the attributes and the likelihood they lead to a result
Column Dependencies		Automatically quantifies how much an attribute influences another attribute
Recommendations		Automically predict interests of a person based on historical observations from many people

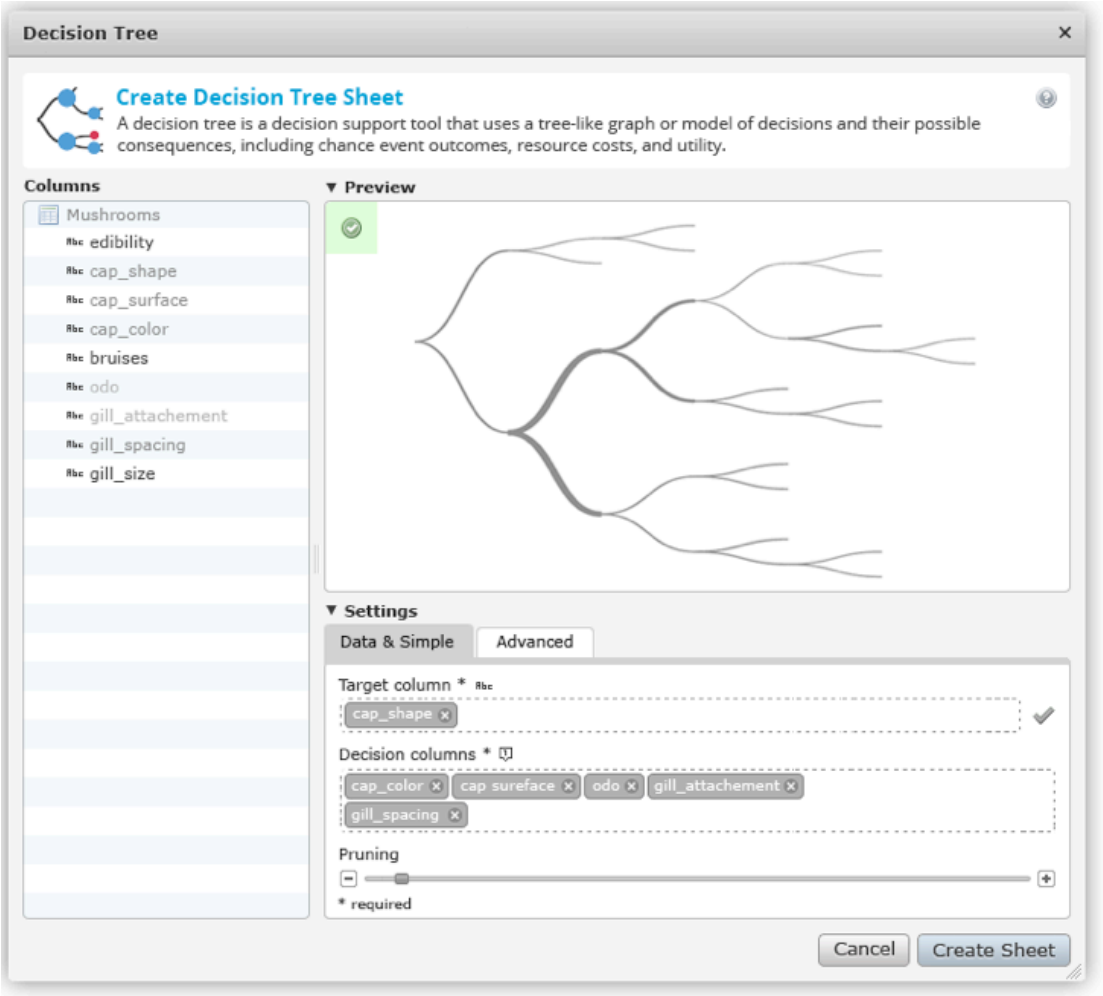
Clustering



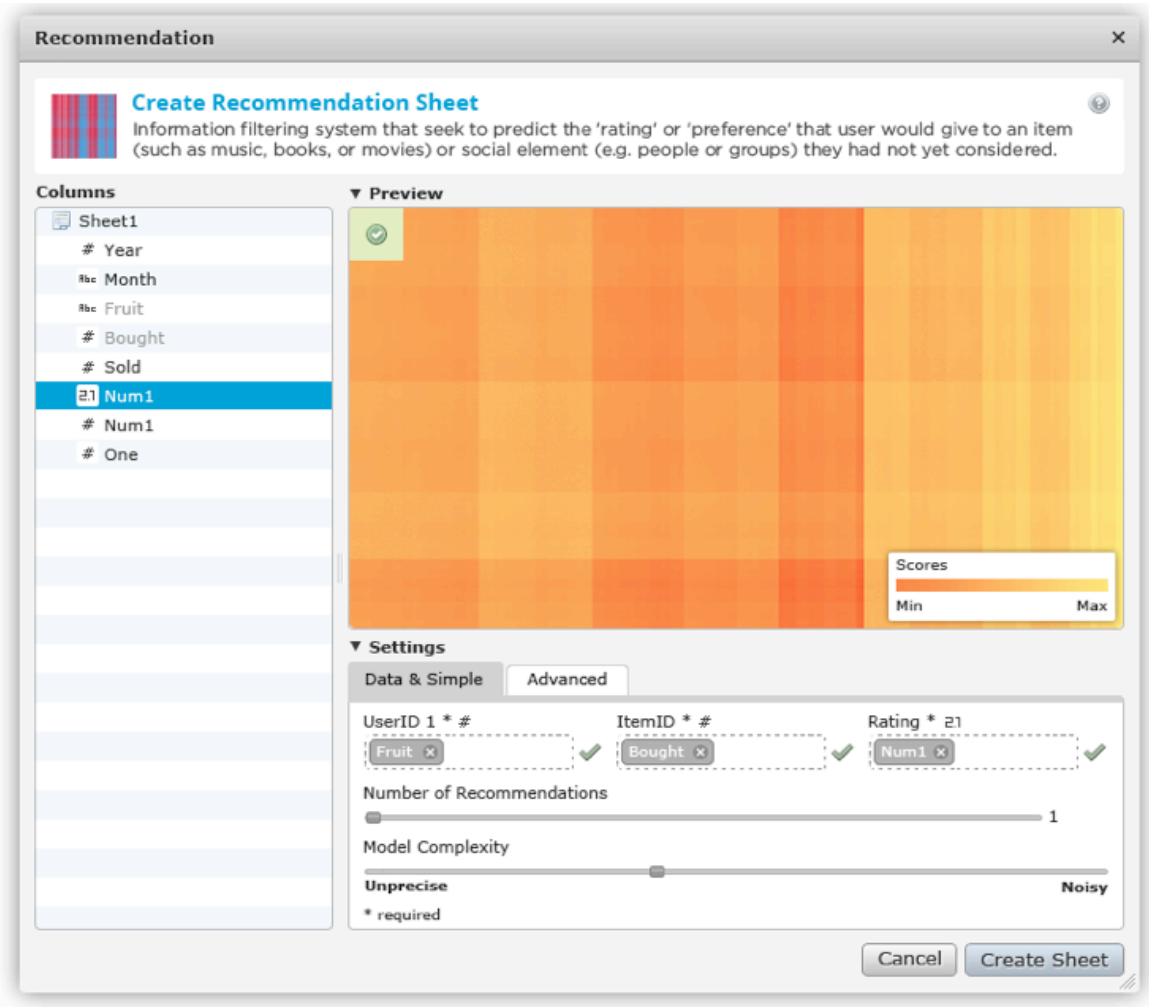
Column Dependencies



Decision Trees



Recommendations



Example:
Fraud Investigation
Sales Conversion

DEMO

Data Wrangling

DEMO

Datameer Smart Analytics for Hadoop Now Available

New self-service data mining functionality lets business users find patterns and relationships in their data without a data scientist

SAN MATEO, Calif., Sept. 30, 2013 -- Empowering business users to find insights in their data even faster, Datameer today announced the public availability of Datameer Smart Analytics, an optional data mining add-on for Datameer 3.0. Extending the self-service functionality of its data integration, analytics and visualization application for Hadoop, Smart Analytics let non-technical users apply popular data mining algorithms to find patterns and explore relationships in their data. A free trial of Datameer 3.0 with Smart Analytics is available at http://www.datameer.com/Datameer-trial.html.

Data Mining for the Masses

@Datameer