Configurable Analytic Flows at Scale: A New Challenge for the BPM Community

Richard Hull, IBM T.J. Watson Research Center

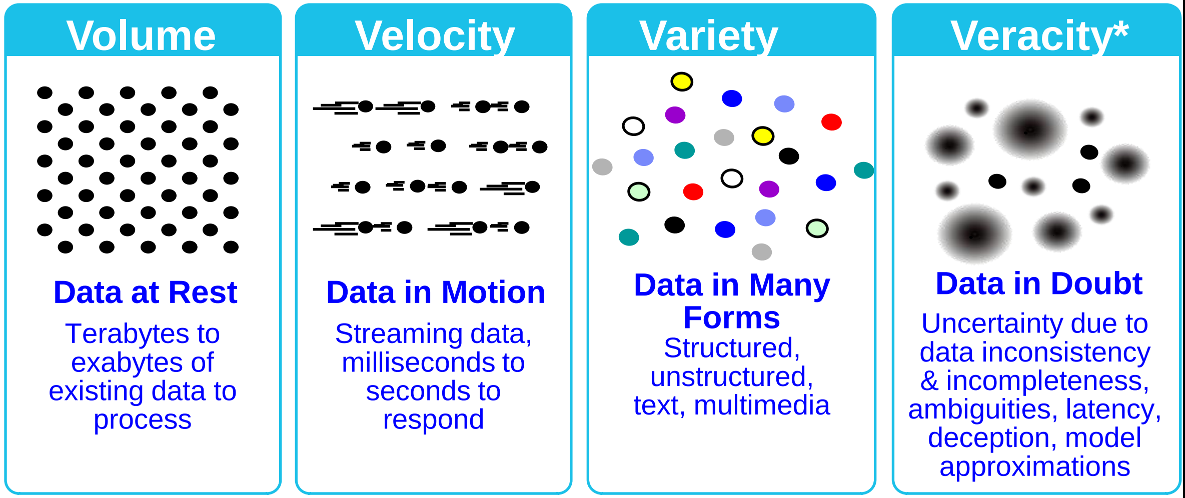
8 September 2014

Presented at the 3rd Intl. Workshop on Data- and Artifact-Centric BPM (DAB),
as part of the 12th Intl. Conf. on Business Process Management (BPM),
held in Eindhoven, The Netherlands

IBM Copyright © 2014

Big Data Analytics

- ***“Big Data”***
  - Volume
  - Variation
  - Velocity
  - Veracity
- ***“Analytics”***
  - Descriptive
  - Predictive
  - Prescriptive
- ***Big Data Analytics is bringing value across all industries***
  - Logistics
  - Retail
  - Healthcare
  - Energy
  - Education
  - Born-on-the-web companies
  - …



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

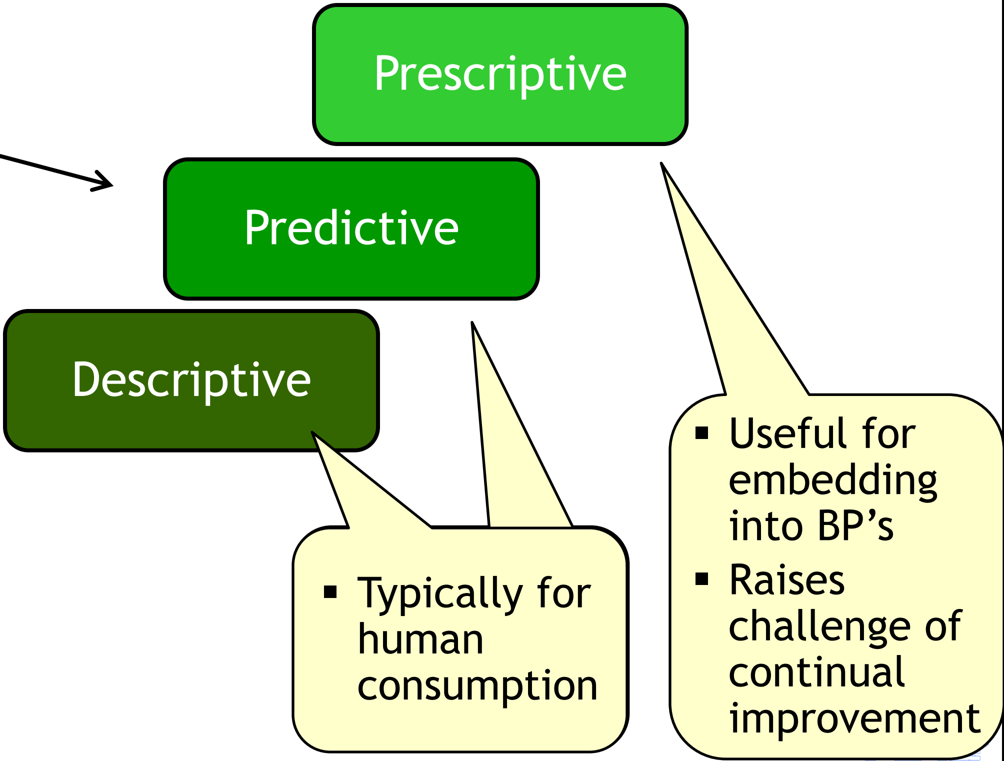
Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations



Prescriptive

Predictive

Descriptive

- Typically for human consumption

- Useful for embedding into BP’s
- Raises challenge of continual improvement

IBM Copyright © 2014

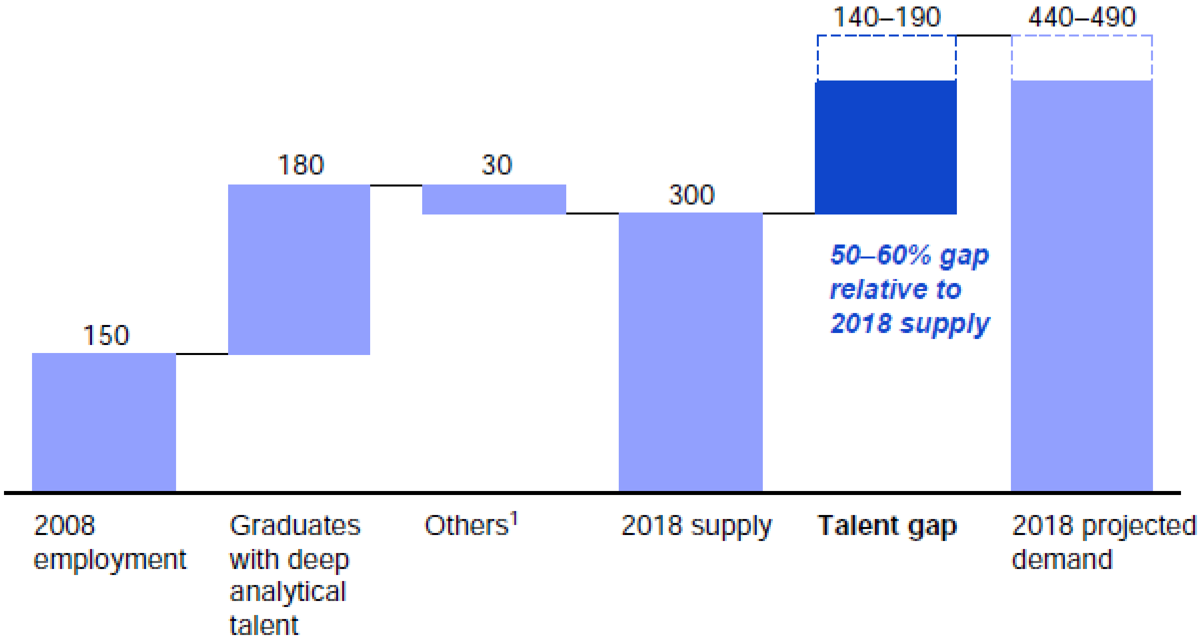
"Big Data Analytics": A major force in early 21st century

- McKinsey
  - Big Data will become the basis for competition
  - Big Data will underpin new waves of productivity growth
  - *140,000 to 190,000 more deep analytical talent positions in US*
  - *1.5 Million more data-savvy managers needed in US*
- Key sectors include healthcare, retail, manufacturing, also education

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

Thousand people



[1] Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

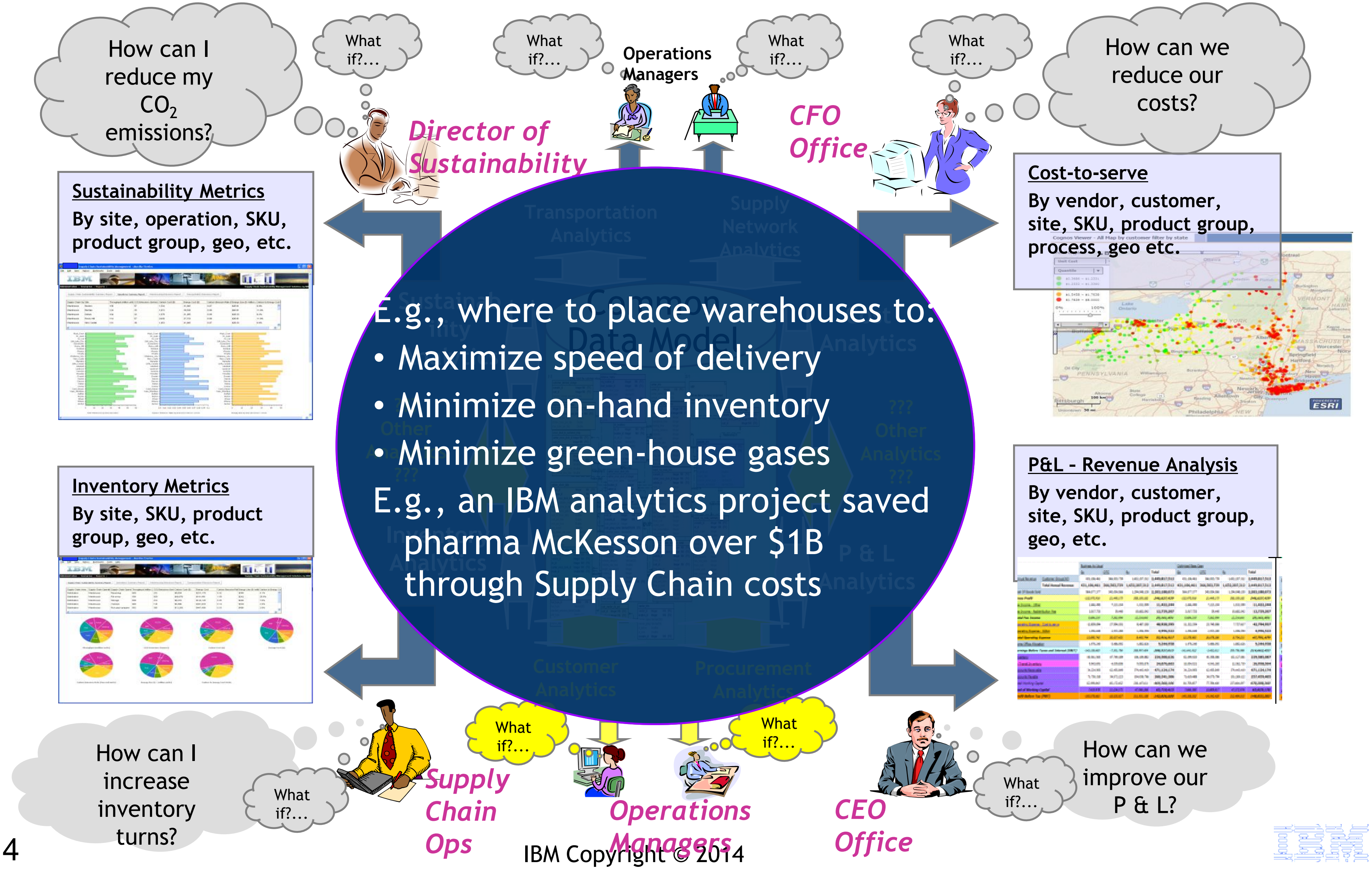
SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

From McKinsey 2011:
Big Data: The next frontier of innovation, competition and productivity

IBM Copyright © 2014

Example: Supply Chain Management

Big Data Analytics increasingly relevant to Business Operations



How can I reduce my CO_2 emissions?

Director of Sustainability

Sustainability Metrics

By site, operation, SKU, product group, geo, etc.

Operations Managers

CFO Office

How can we reduce our costs?

Cost-to-serve

By vendor, customer, site, SKU, product group, process, geo etc.

Inventory Metrics

By site, SKU, product group, geo, etc.

P&L - Revenue Analysis

By vendor, customer, site, SKU, product group, geo, etc.

E.g., where to place warehouses to:

- Maximize speed of delivery
- Minimize on-hand inventory
- Minimize green-house gases

E.g., an IBM analytics project saved pharma McKesson over $1B through Supply Chain costs

How can I increase inventory turns?

Supply Chain Ops

Operations Managers

CEO Office

How can we improve our P & L?

IBM Copyright © 2014

Example: Social Media & Text Analytics

- RetailerXX wants to sell to the “Millennials” - ages 16-25
- Who are the Millennials, anyway, and how do they shop ??

- ***IBM analyzed over 3 BILLION tweets***
- ***Created 7 “clusters” of Millennial shoppers***

For example…

Fashion on a Dime Persona

**Illustrative*

- ***Loves going to the Mall***, whether it is to purchase at a department store or at Forever 21
- Young Millennial who has a positive sentiment towards RetailerXX’s but is ***not brand loyal***
- Prefers discounts and is highly incentivized by personalized offers
- ***Shares everything with their friends***, from their latest purchase to their dream vacation
- ***Follows latest fashion news*** and gossip, dreams of going to Fashion Week, and feels like they be
st

Fashion on a Dime*

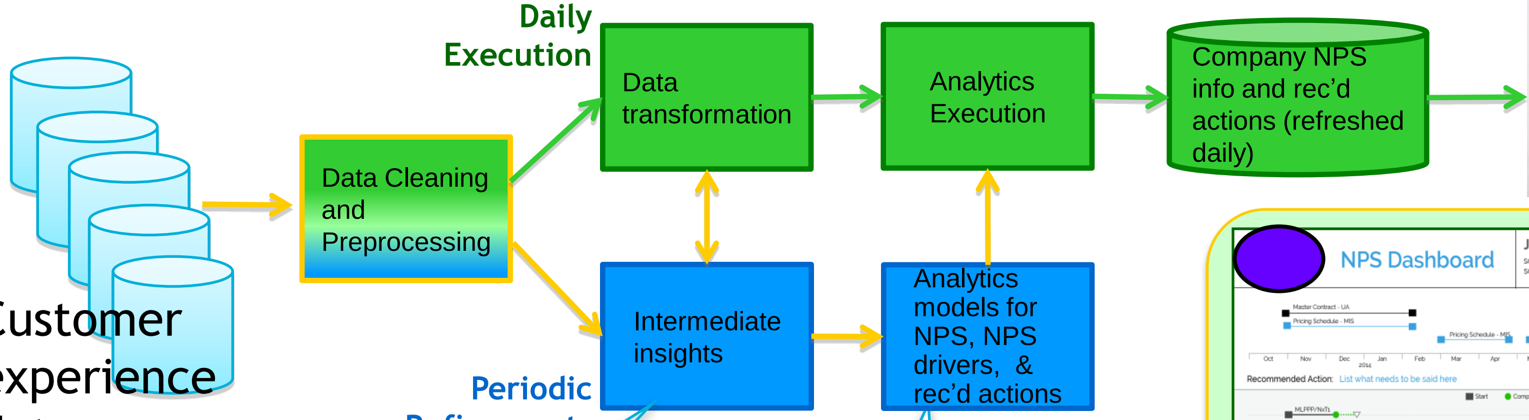
Info Extraction / Text Analytics increasingly present in Big Data applications

IBM Copyright © 2014

“Actionable Customer Satisfaction” for B2B sales

IBM Major Telco

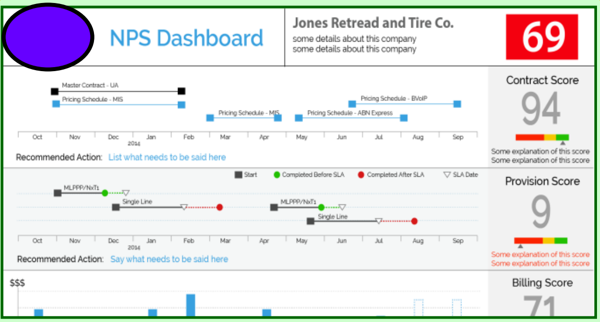
Analytics infer customer sat, key drivers, mediating actions


Daily Execution
Data Cleaning and Preprocessing
Data transformation
Analytics Execution
Company NPS info and rec’d actions (refreshed daily)
Intermediate insights
Analytics models for NPS, NPS drivers, & rec’d actions
Periodic Refinements

Customer experience data

(e.g., Purchasing Provisioning, Tickets, …)

B2B Seller NPS Dashboard



[redacted] NPS Dashboard

Jones Retread and Tire Co.
some details about this company
some details about this company

69

Recommended Action: List what needs to be said here

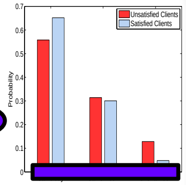
Contract Score
94
Some explanation of this score
Some explanation of this score

Recommended Action: Say what needs to be said here

Provision Score
9
Some explanation of this score
Some explanation of this score

Billing Score
71

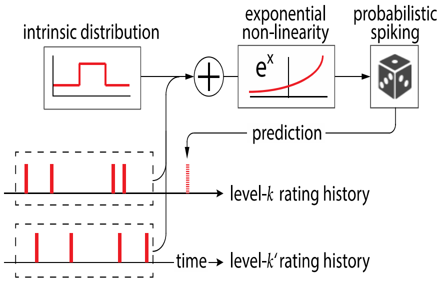
Representative Insight & Mitigating Actions



[redacted]

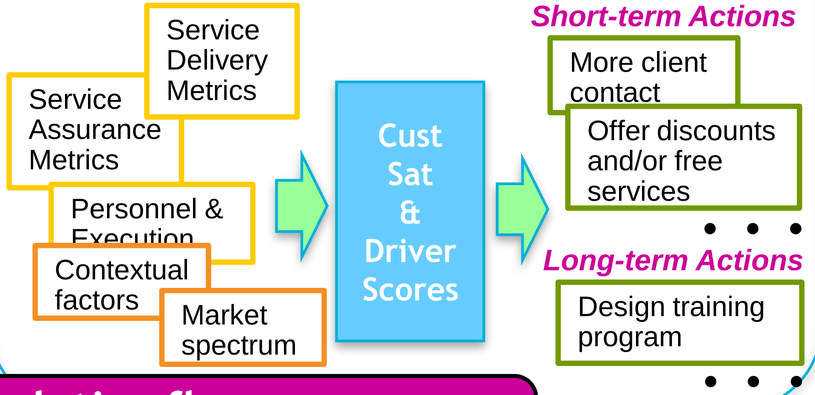
- **Service Delay Threshold:** Customers averaging > [redacted] day ticket response show lower customer satisfaction
- Potential Actions:
  - Auto escalate tickets after 4 days
  - Pre-communicate to customer when repair estimate > [redacted] days

Predicting drivers through Mechanistic Models



Based on recently

Cust Sat Explained from Drivers to Recommended Actions



Service Delivery Metrics

Service Assurance Metrics

Personnel & Execution

Contextual factors

Market spectrum

Cust Sat & Driver Scores

Short-term Actions

More client contact

Offer discounts and/or free services

• • •

Long-term Actions

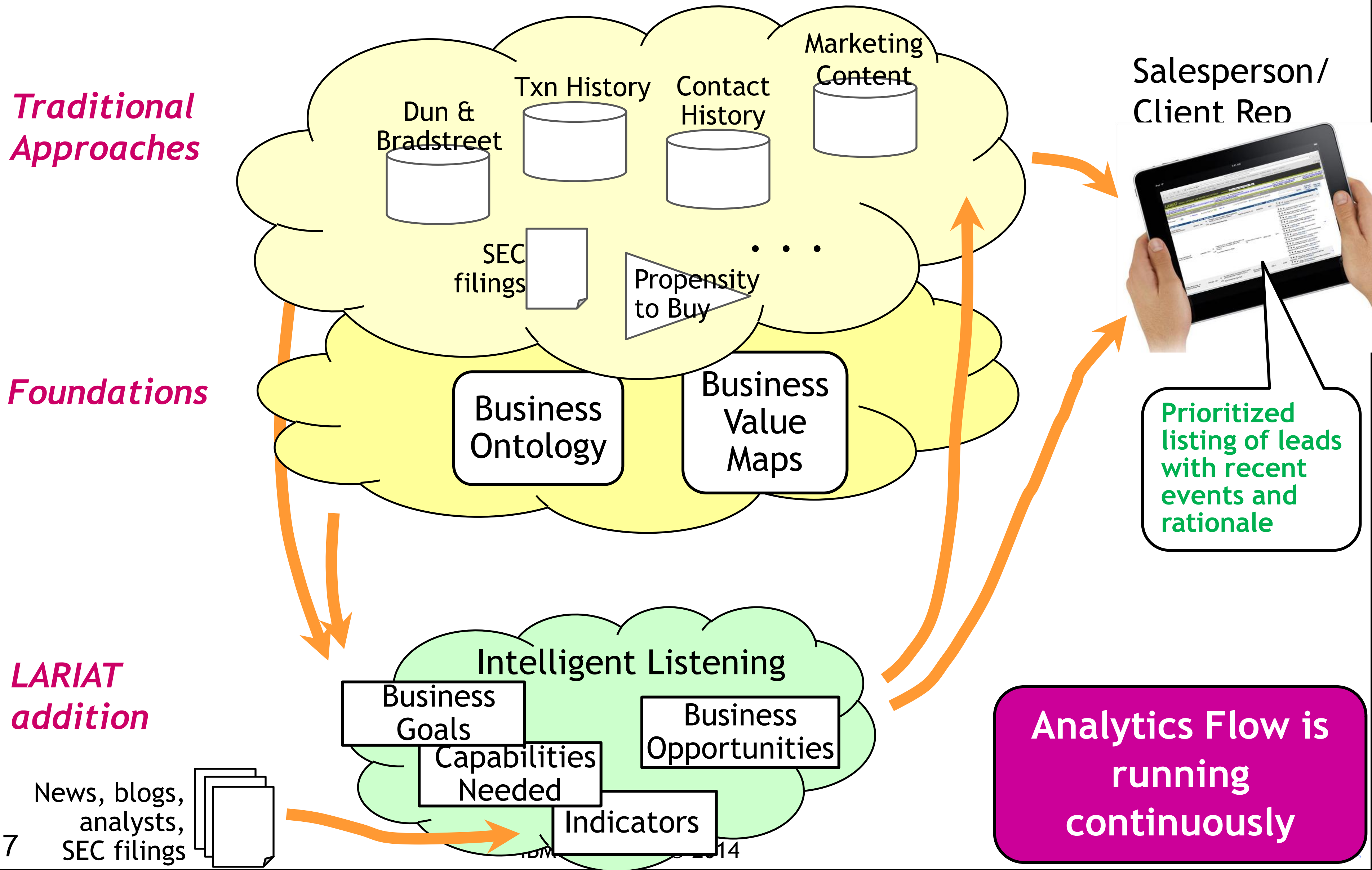
Design training program

• • •

Repeated execution of analytics flow embedded into on-going Business Process

IBM Copyright © 2014

LARIAT adds timely listening to traditional approaches to B2B Lead-to-Revenue management



Analytics Flow is running continuously

LARIAT output: Data about Companies (detail from Smarter Process sales team view)

Company

DUNS number

Priority level

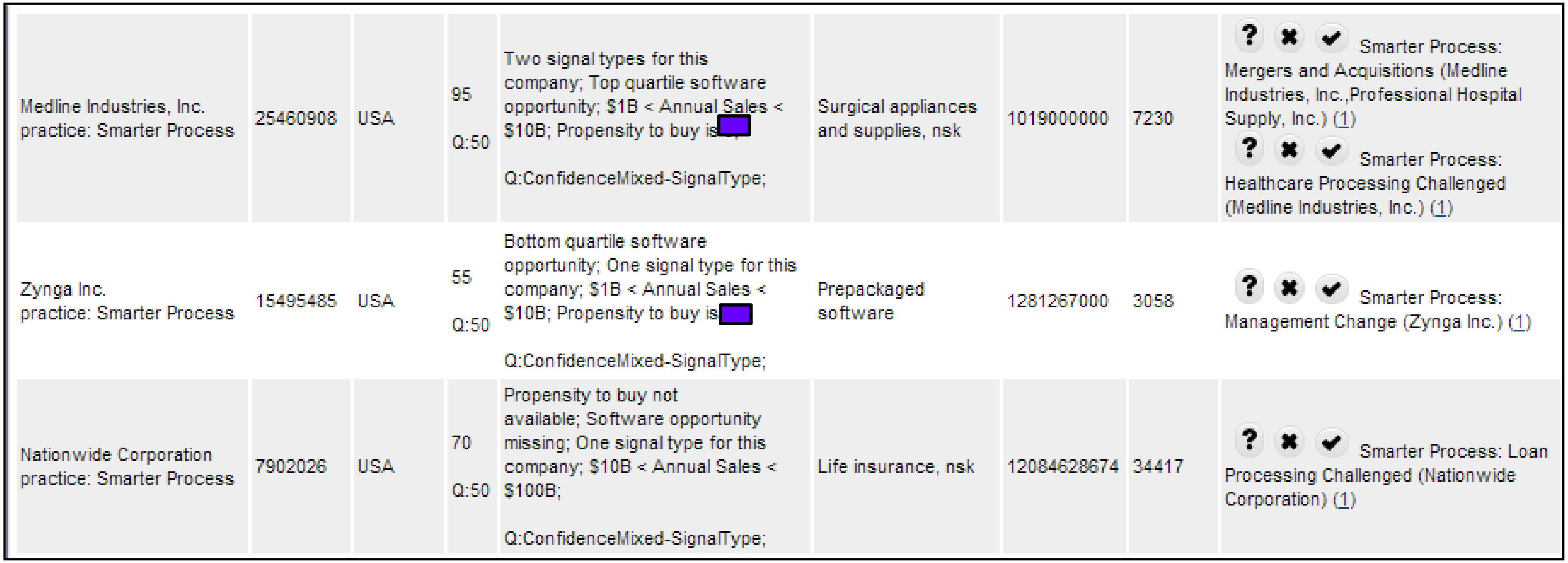
Rationale for priority

Industry

Revenue

Employee count

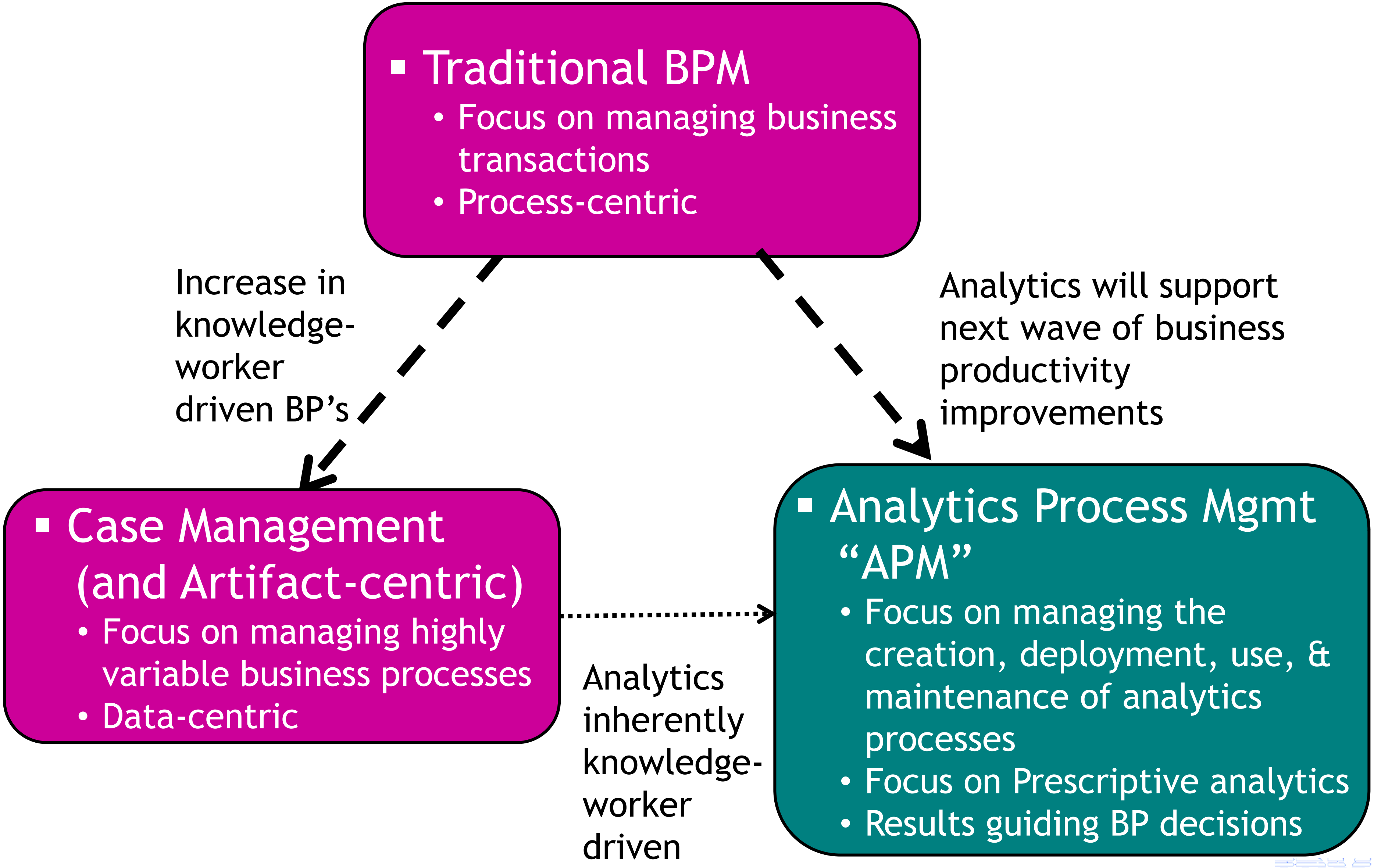
Relevant indicators



Medline Industries, Inc. practice: Smarter Process	25460908	USA	95 Q:50	Two signal types for this company; Top quartile software opportunity; $1B < Annual Sales < $10B; Propensity to buy is 🟪; Q:ConfidenceMixed-SignalType;	Surgical appliances and supplies, nsk	1019000000	7230	? ✖ ✔ Smarter Process: Mergers and Acquisitions (Medline Industries, Inc.,Professional Hospital Supply, Inc.) (1) ? ✖ ✔ Smarter Process: Healthcare Processing Challenged (Medline Industries, Inc.) (1)
Zynga Inc. practice: Smarter Process	15495485	USA	55 Q:50	Bottom quartile software opportunity; One signal type for this company; $1B < Annual Sales < $10B; Propensity to buy is 🟪 Q:ConfidenceMixed-SignalType;	Prepackaged software	1281267000	3058	? ✖ ✔ Smarter Process: Management Change (Zynga Inc.) (1)
Nationwide Corporation practice: Smarter Process	7902026	USA	70 Q:50	Propensity to buy not available; Software opportunity missing; One signal type for this company; $10B < Annual Sales < $100B; Q:ConfidenceMixed-SignalType;	Life insurance, nsk	12084628674	34417	? ✖ ✔ Smarter Process: Loan Processing Challenged (Nationwide Corporation) (1)

IBM Copyright © 2014

A new kind of BPM



- Traditional BPM
  - Focus on managing business transactions
  - Process-centric

Increase in knowledge-worker driven BP’s

Analytics will support next wave of business productivity improvements

- Case Management (and Artifact-centric)
  - Focus on managing highly variable business processes
  - Data-centric

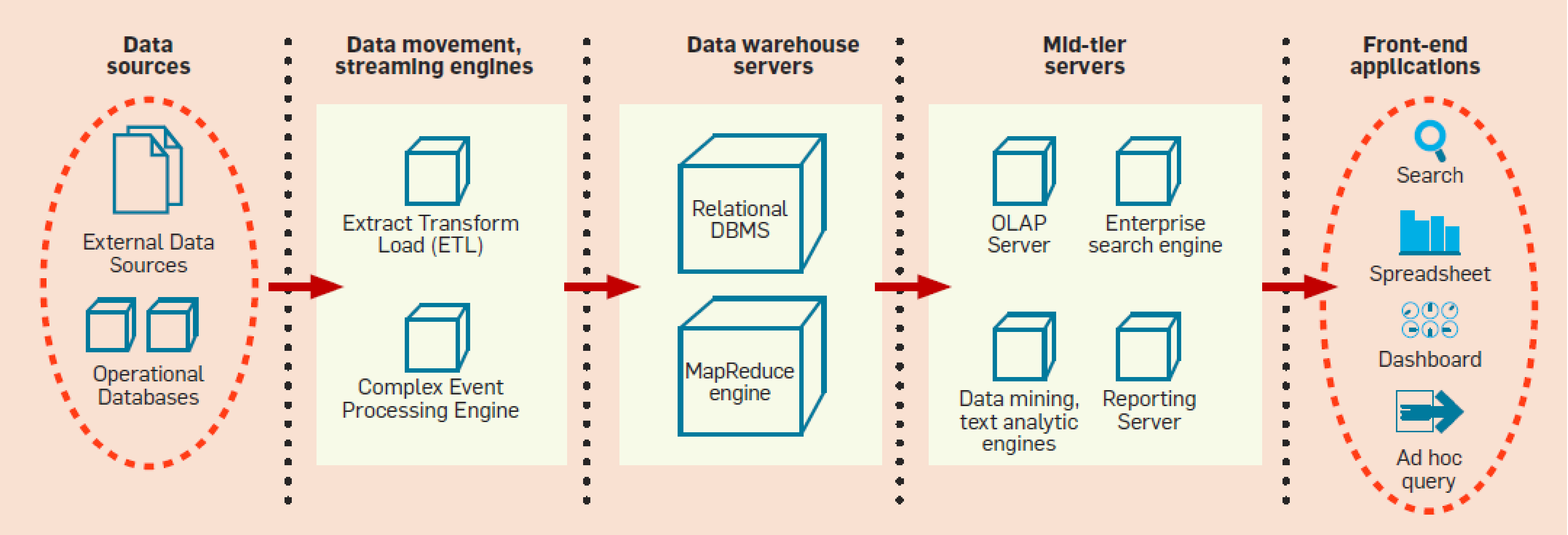
Analytics inherently knowledge-worker driven

- Analytics Process Mgmt “APM”
  - Focus on managing the creation, deployment, use, & maintenance of analytics processes
  - Focus on Prescriptive analytics
  - Results guiding BP decisions

IBM Copyright © 2014

Research community has not been thinking about repeating analytics flows used by BPs

- CACM Survey of Business Intelligence [Chaudhuri, Dayal, Narasayya 2011]
  - The “product” of analytics is for human consumption, not BPs



- CACM Technical Challenges in Big Data [Jagadish et al 2014]
  - Again, the “product” of analytics is for human consumption, not BPs

- [Troung and Dustar 2012]: “Research on how to manage analysis algorithms and how to provide an open platform for third parties to develop, search and share algorithms is quite open.”

IBM Copyright © 2014

Agenda

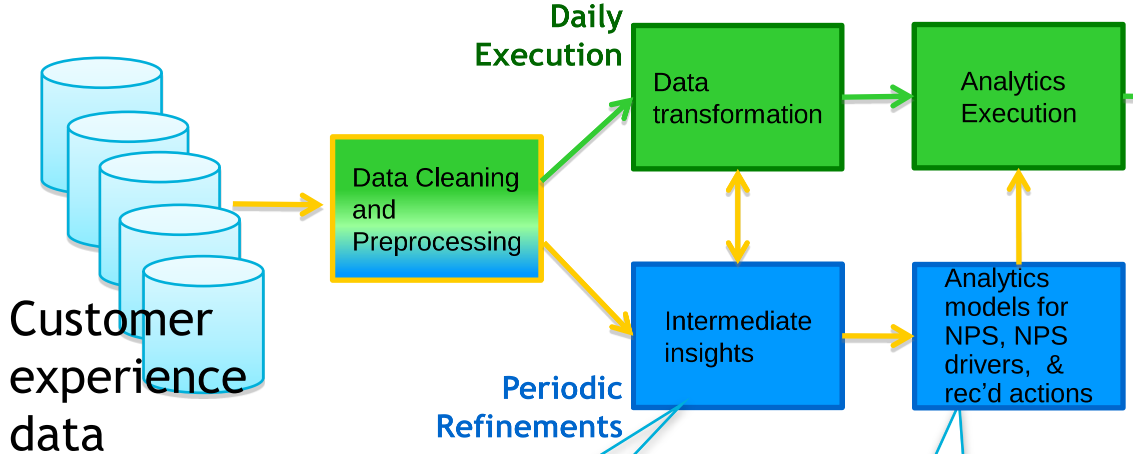
- Drill-down on representative Analytics Processes
- What makes APM different/hard?
- The ProkoFieV framework
  - Functionality
  - Variation
  - Provenance
- Relevant techniques/tools
- Some foundational research questions

IBM Copyright © 2014

“Actionable Customer Satisfaction” for B2B sales

IBM Major Telco

Analytics infer customer sat, key drivers, mediating actions



Customer experience data

(e.g., Purchasing Provisioning, Tickets, …)

Data Cleaning and Preprocessing

Daily Execution

Data transformation

Analytics Execution

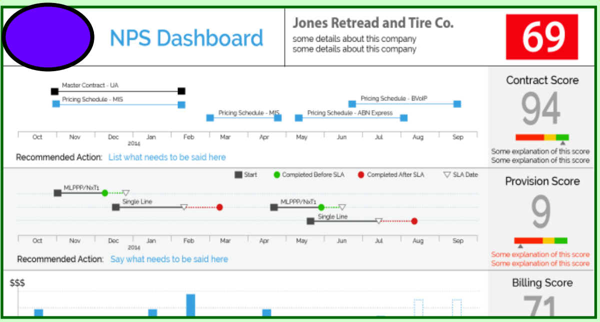
Company NPS info and rec’d actions (refreshed daily)

Periodic Refinements

Intermediate insights

Analytics models for NPS, NPS drivers, & rec’d actions

B2B Seller NPS Dashboard



[redacted] NPS Dashboard

Jones Retread and Tire Co.

some details about this company
some details about this company

69

Recommended Action: List what needs to be said here

Contract Score

94

Some explanation of this score
Some explanation of this score

Recommended Action: Say what needs to be said here

Provision Score

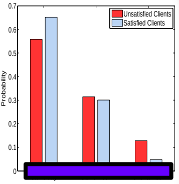
9

Some explanation of this score
Some explanation of this score

Billing Score

71

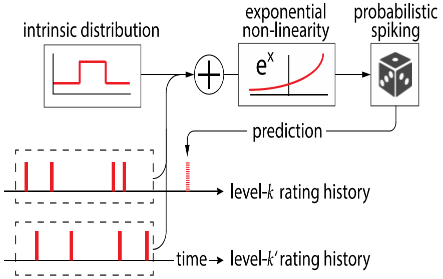
Representative Insight & Mitigating Actions



[redacted]

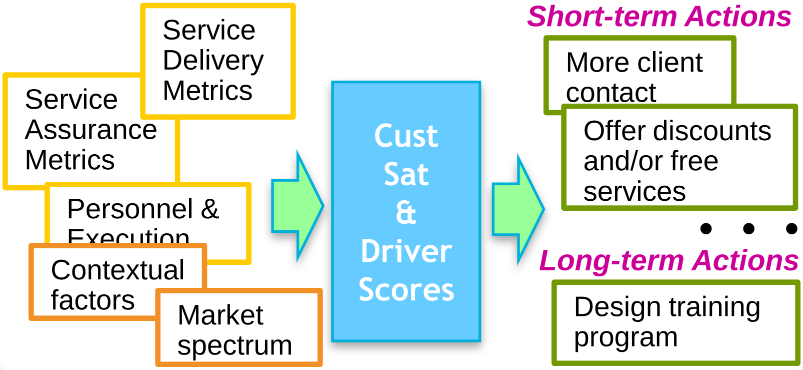
- **Service Delay Threshold:** Customers averaging >[redacted] day ticket response show lower customer satisfaction
- Potential Actions:
  - Auto escalate tickets after 4 days
  - Pre-communicate to customer when repair estimate > [redacted] days

Predicting drivers through Mechanistic Models



Based on recently

Cust Sat Explained from Drivers to Recommended Actions



Service Delivery Metrics

Service Assurance Metrics

Personnel & Execution

Contextual factors

Market spectrum

Cust Sat & Driver Scores

Short-term Actions

More client contact

Offer discounts and/or free services

• • •

Long-term Actions

Design training program

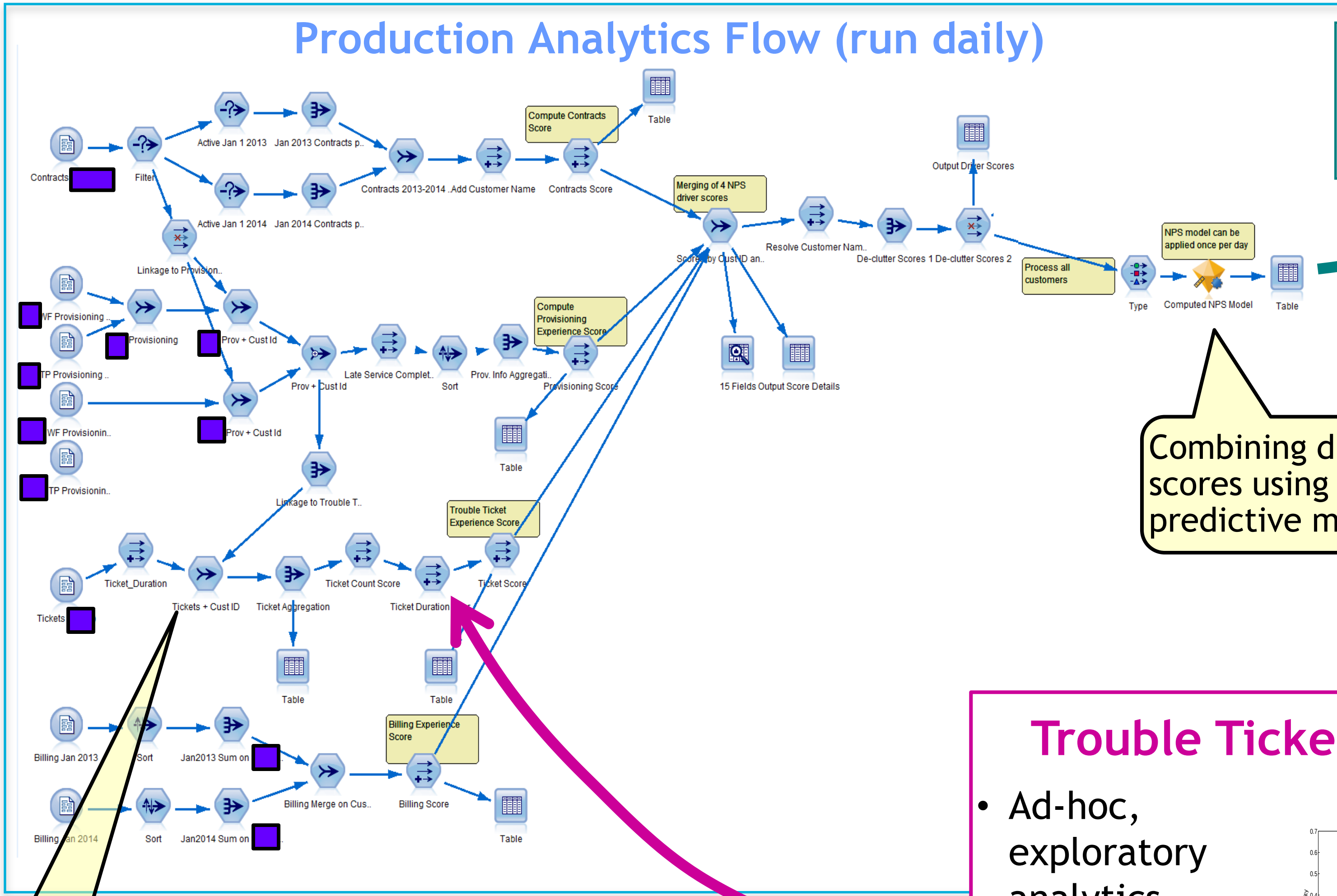
• • •

Repeated execution of analytics flow embedded into on-going Business Process

IBM Copyright © 2014

Actionable Customer Satisfaction: Production Flow and “Feeder Analytics”

Production Analytics Flow (run daily)



B2B Seller NPS Dashboard

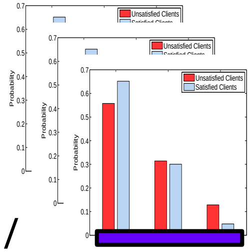
Combining driver scores using predictive model

Computing Trouble Tickets Driver

Statistical insights incorporated into Production Flow

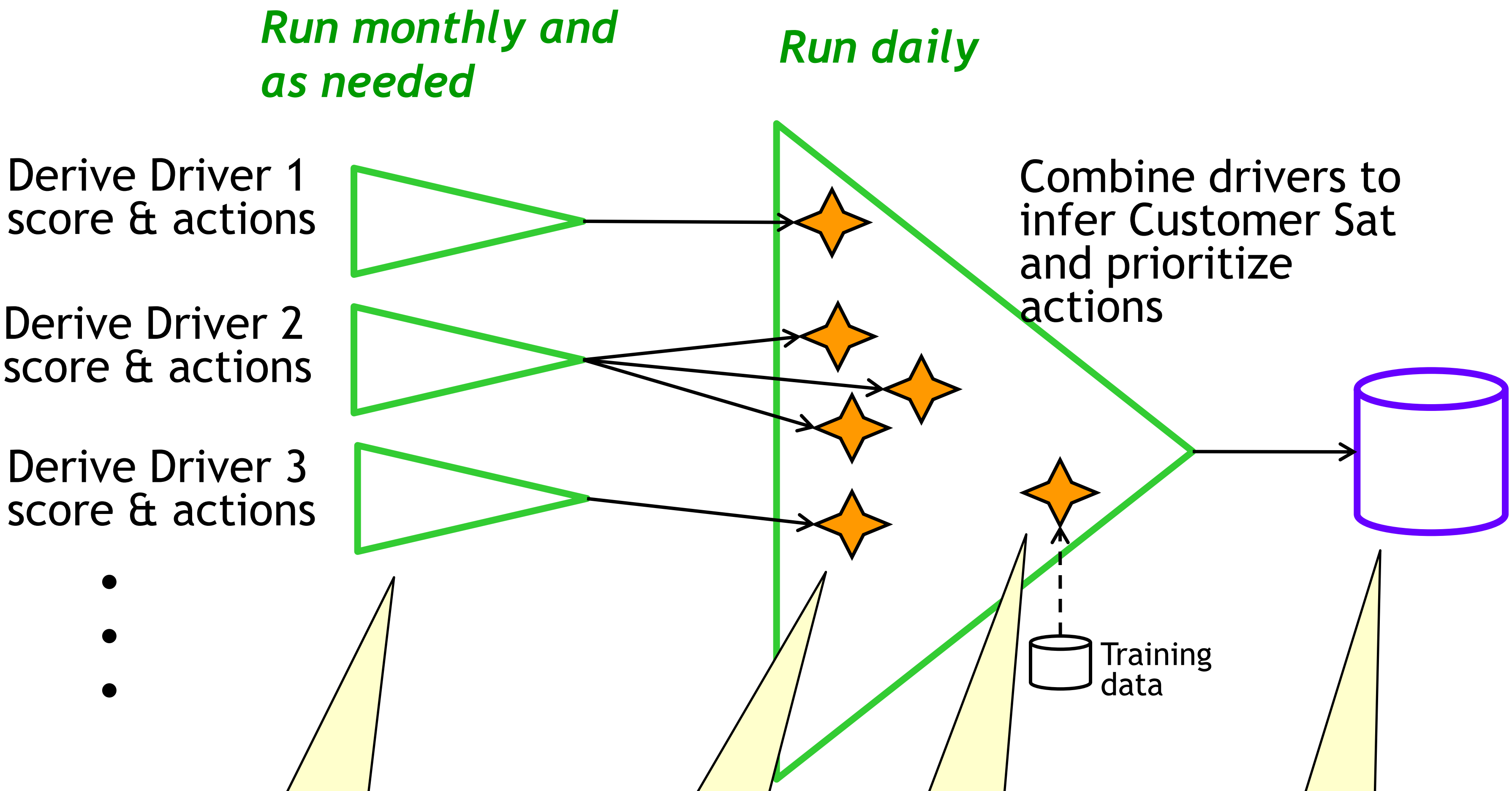
Trouble Tickets Driver

- Ad-hoc, exploratory analytics
- Found ‘[redacted]-Day Tipping Point”
- Need to validate/ & possibly refine monthly



IBM Copyright © 2014

Actionable Customer Sat Analytics Flow (abstracted)



Run monthly and as needed
Run daily
Derive Driver 1 score & actions
Derive Driver 2 score & actions
Derive Driver 3 score & actions
Combine drivers to infer Customer Sat and prioritize actions
Training data

These sub-flows produce statistical models (policies, algorithms) for inferring driver scores

The driver models are embedded into the top-level customer sat flow

Training data may be used to create statistical model for the top-level flow

Daily output includes customer sat, driver scores, prioritized actions

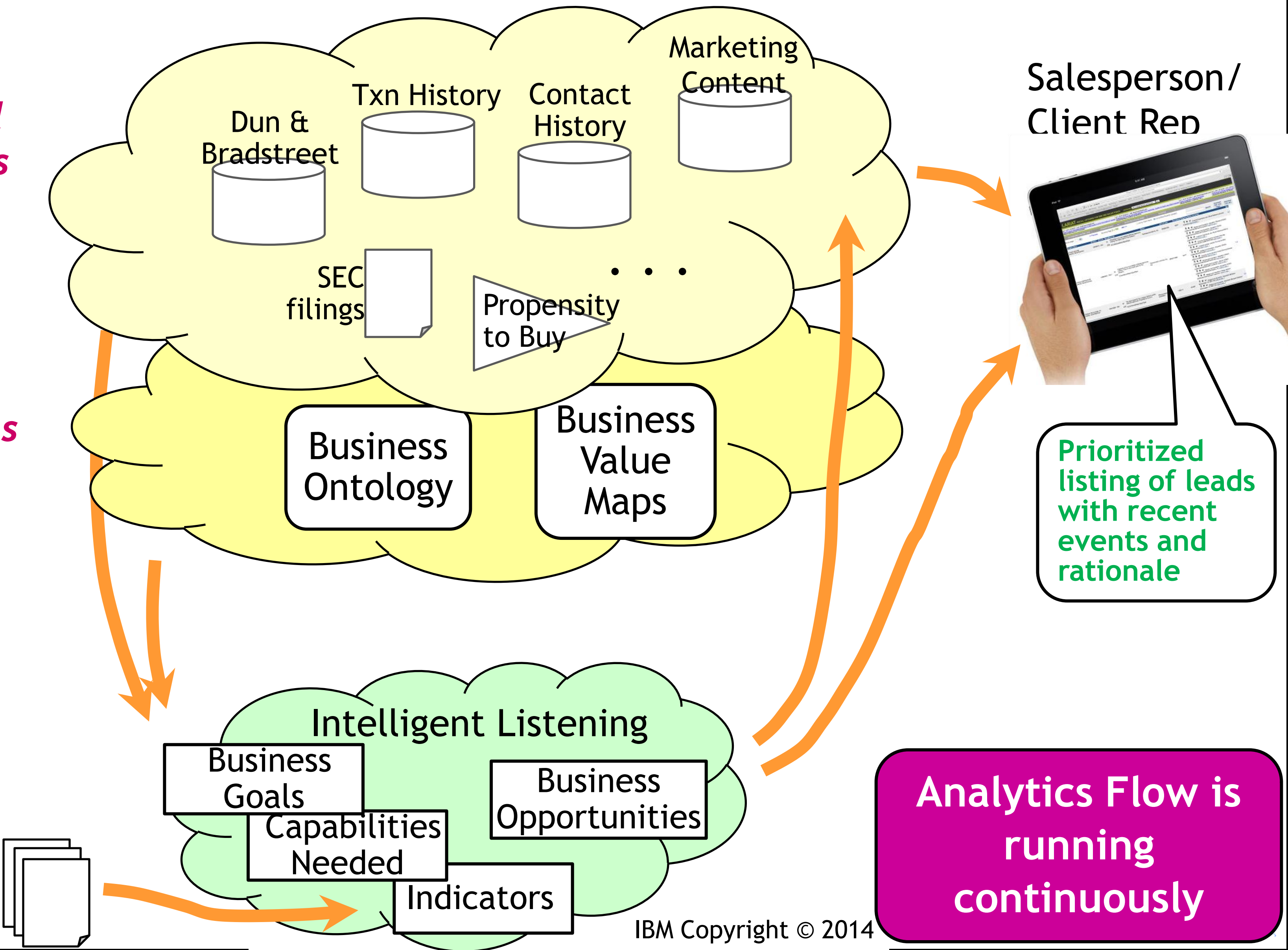
IBM Copyright © 2014

LARIAT adds timely listening to traditional approaches to B2B Lead-to-Revenue management

Traditional Approaches

Foundations

LARIAT addition



Dun & Bradstreet

Txn History

Contact History

Marketing Content

SEC filings

Propensity to Buy

• • •

Business Ontology

Business Value Maps

Intelligent Listening

Business Goals

Capabilities Needed

Business Opportunities

Indicators

News, blogs, analysts, SEC filings

Salesperson/ Client Rep

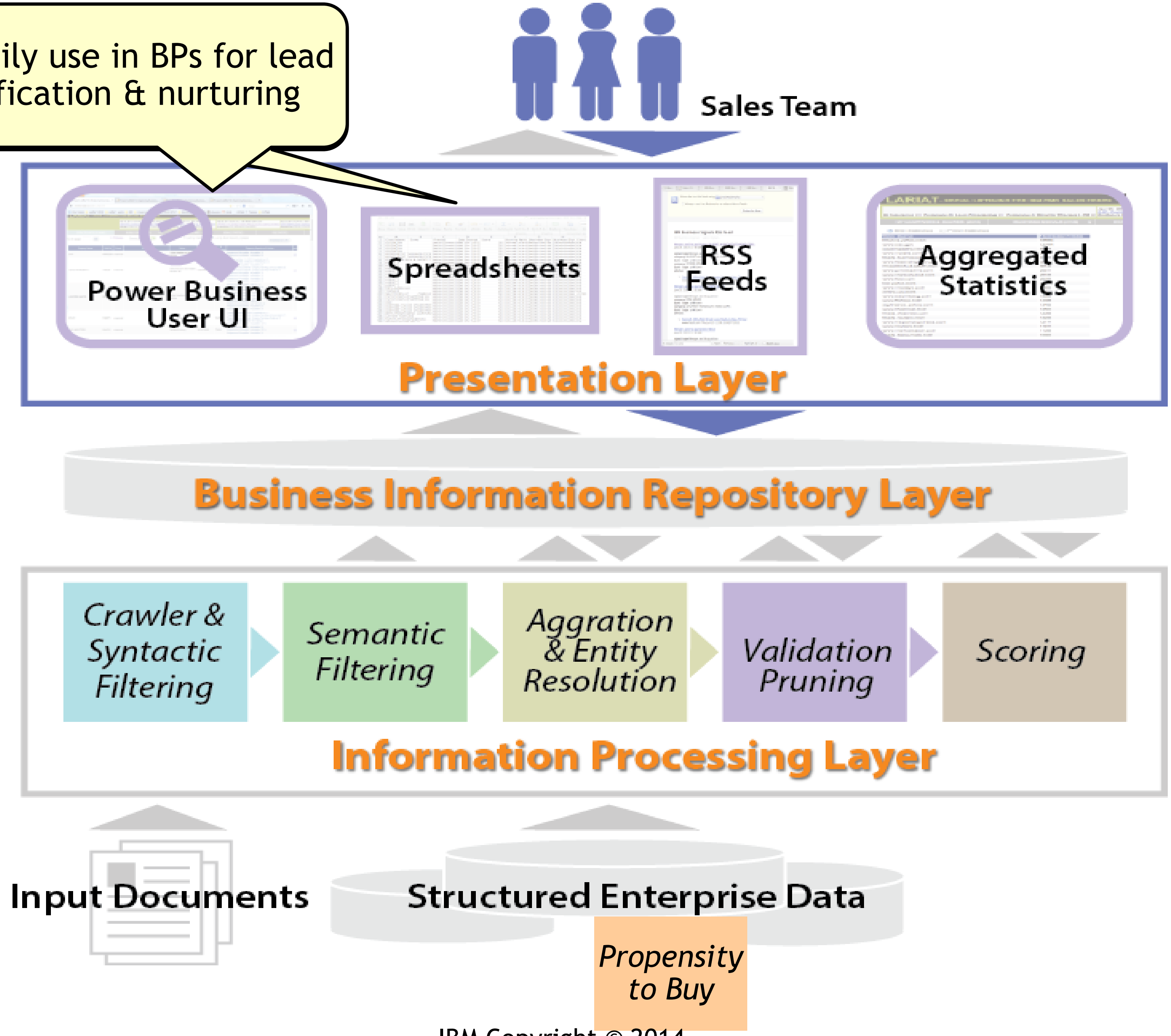
Prioritized listing of leads with recent events and rationale

Analytics Flow is running continuously

IBM Copyright © 2014

LARIAT Functional Components and Processing Flow overview

For daily use in BPs for lead identification & nurturing



IBM Copyright © 2014

Stakeholders around an Analytics Flow Solution (example)

Subject Matter Experts

Identify business patterns

Ontologies

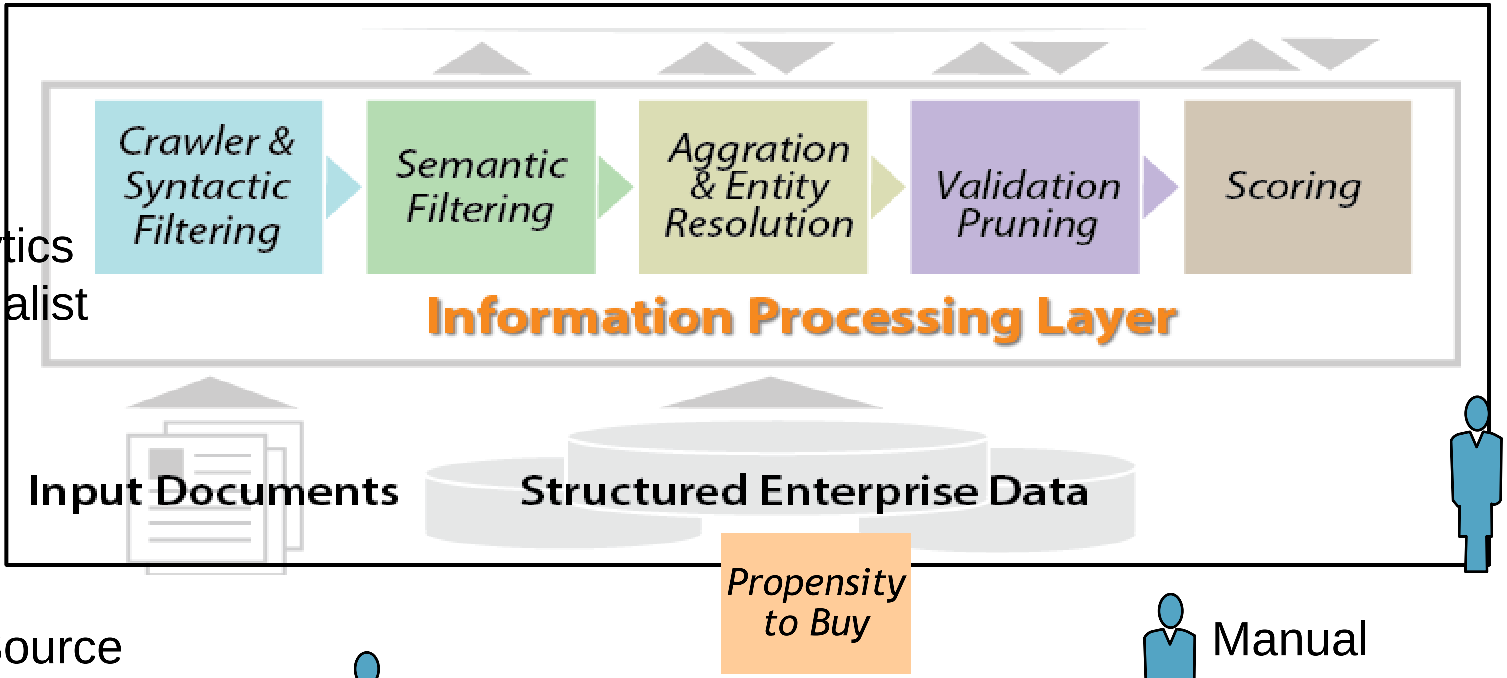
Business causality relationships

P/L Owner

- Manages ROI
- Very concerned about metrics
- Wants explanations

Visualization

Sales Mgmt



Crawler & Syntactic Filtering

Semantic Filtering

Aggration & Entity Resolution

Validation Pruning

Scoring

Information Processing Layer

Input Documents

Structured Enterprise Data

Propensity to Buy

Text Analytics Specialist

Create text extractors based on business patterns

Sellers

Measure & refine

Source selection

ETL Specialist

Statistical Analytics Team

Maintain/tune Propensity-to-buy

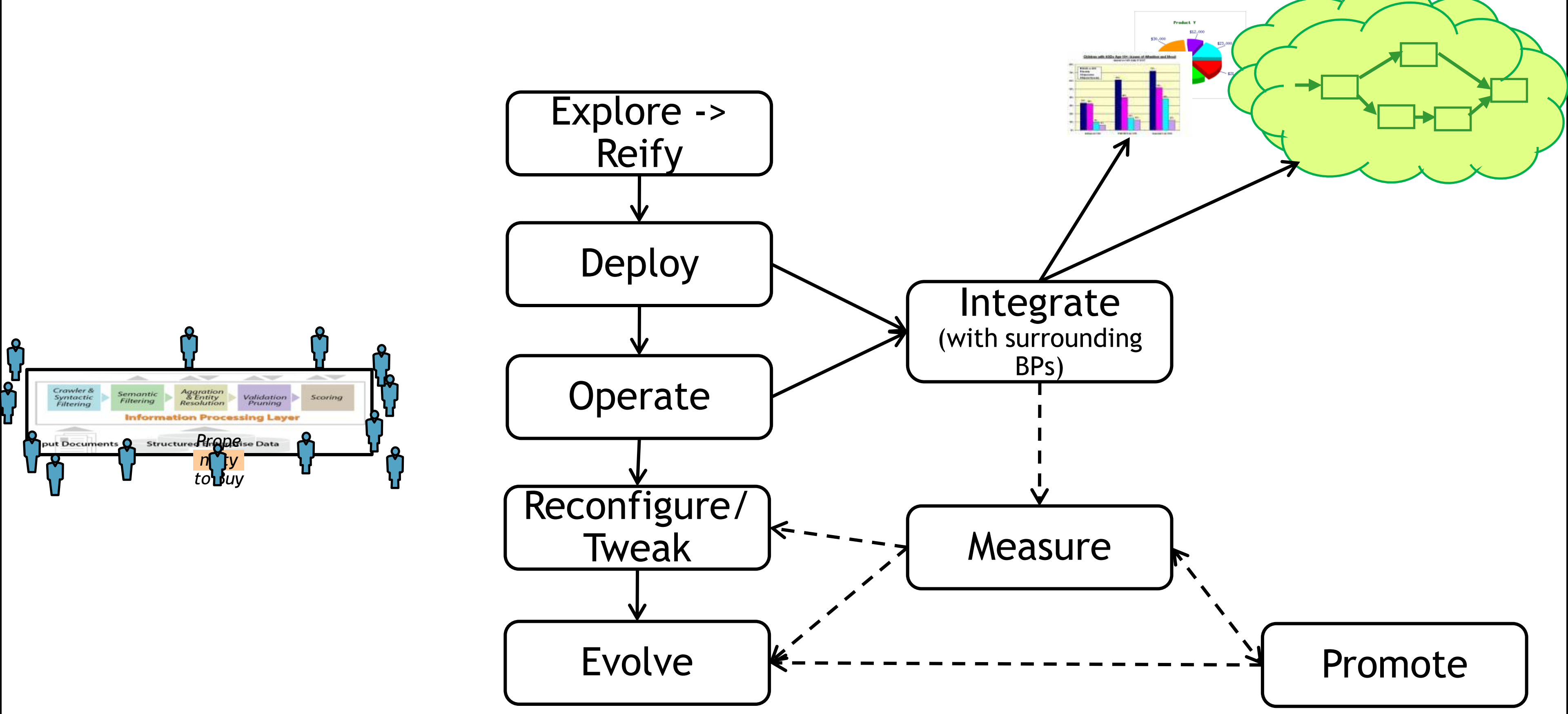
Manual validation team

Source acquisition

IT Support

IBM Copyright © 2014

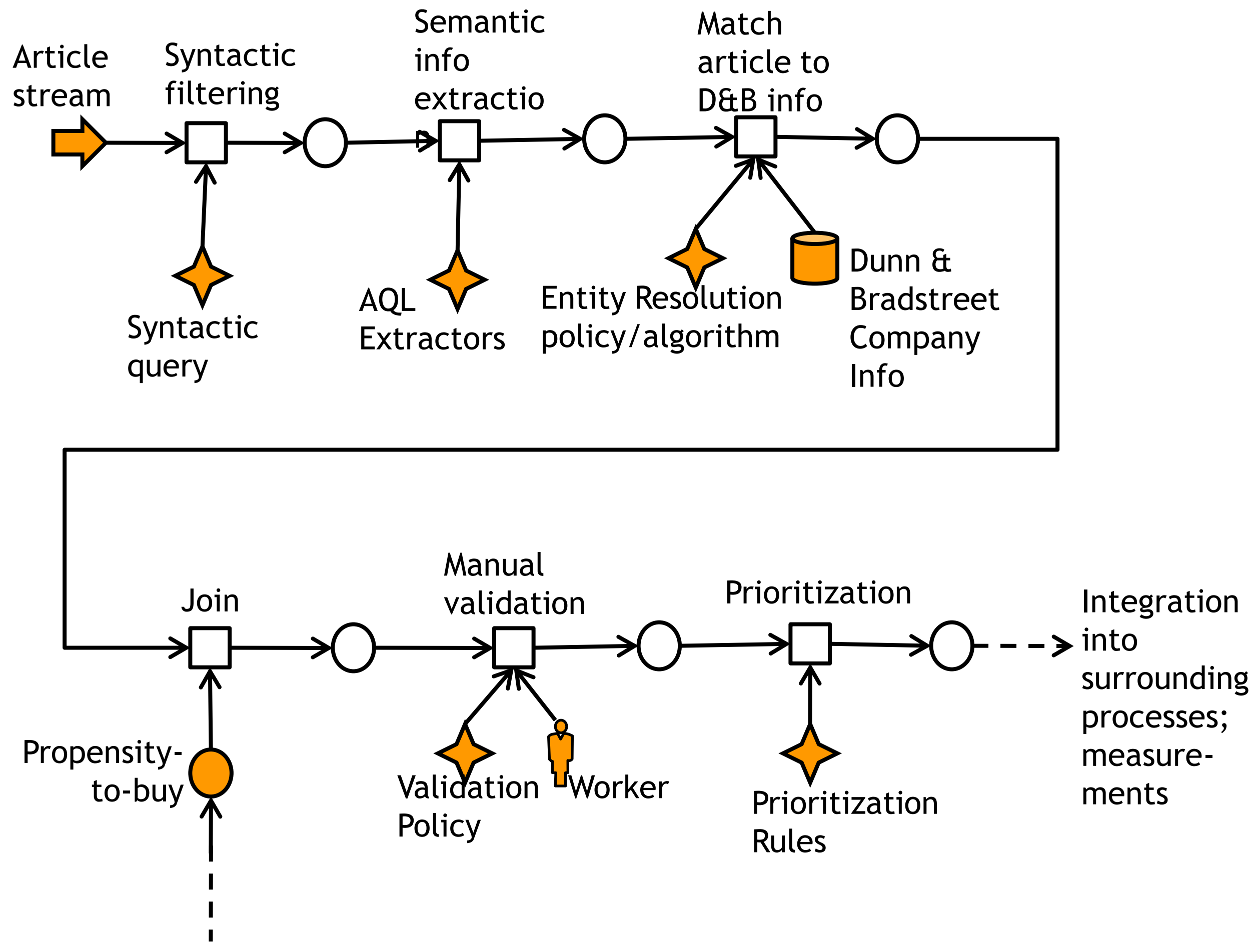
Top-level BPs for a repeating analytics flows*


Crawler & Syntactic Filtering
Semantic Filtering
Aggration & Entity Resolution
Validation Pruning
Scoring
Information Processing Layer
put Documents
Structured Enterprise Data
Propensity to Buy
Explore -> Reify
Deploy
Operate
Reconfigure/ Tweak
Evolve
Integrate (with surrounding BPs)
Measure
Promote

- Each of these top-level BPs is knowledge-worker intensive
- Case Management/Biz Artifacts is natural approach to support these
- This will enable strong measurement & governance of the effectiveness of both the analytics flows and the personnel that are working on/with them

*Related to, and expanding on, CRISP-DM

The core entity type: Configurable Analytics Flow



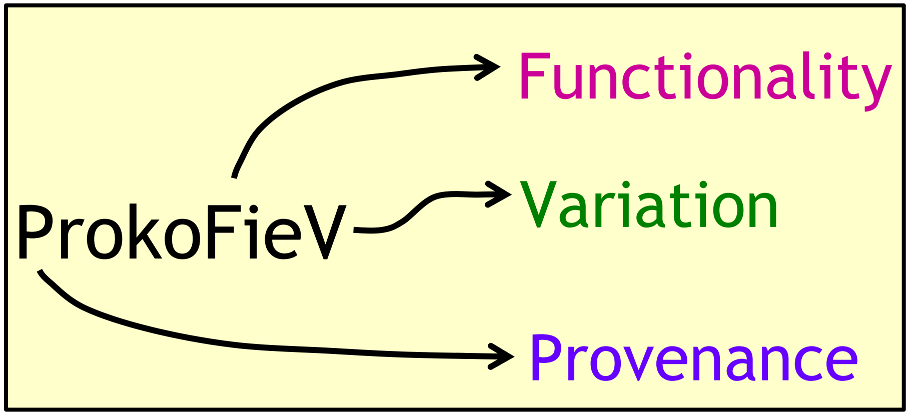
- Full flow might execute, or a subgraph
- Multiple points of configurability
  - Mainly based on changing data or logic
- A vehicle for retaining ***provenance*** of computed data
  - Prospective: flow design
  - Retrospective: info about instance
- Provides anchor for measurements and identifying attributions

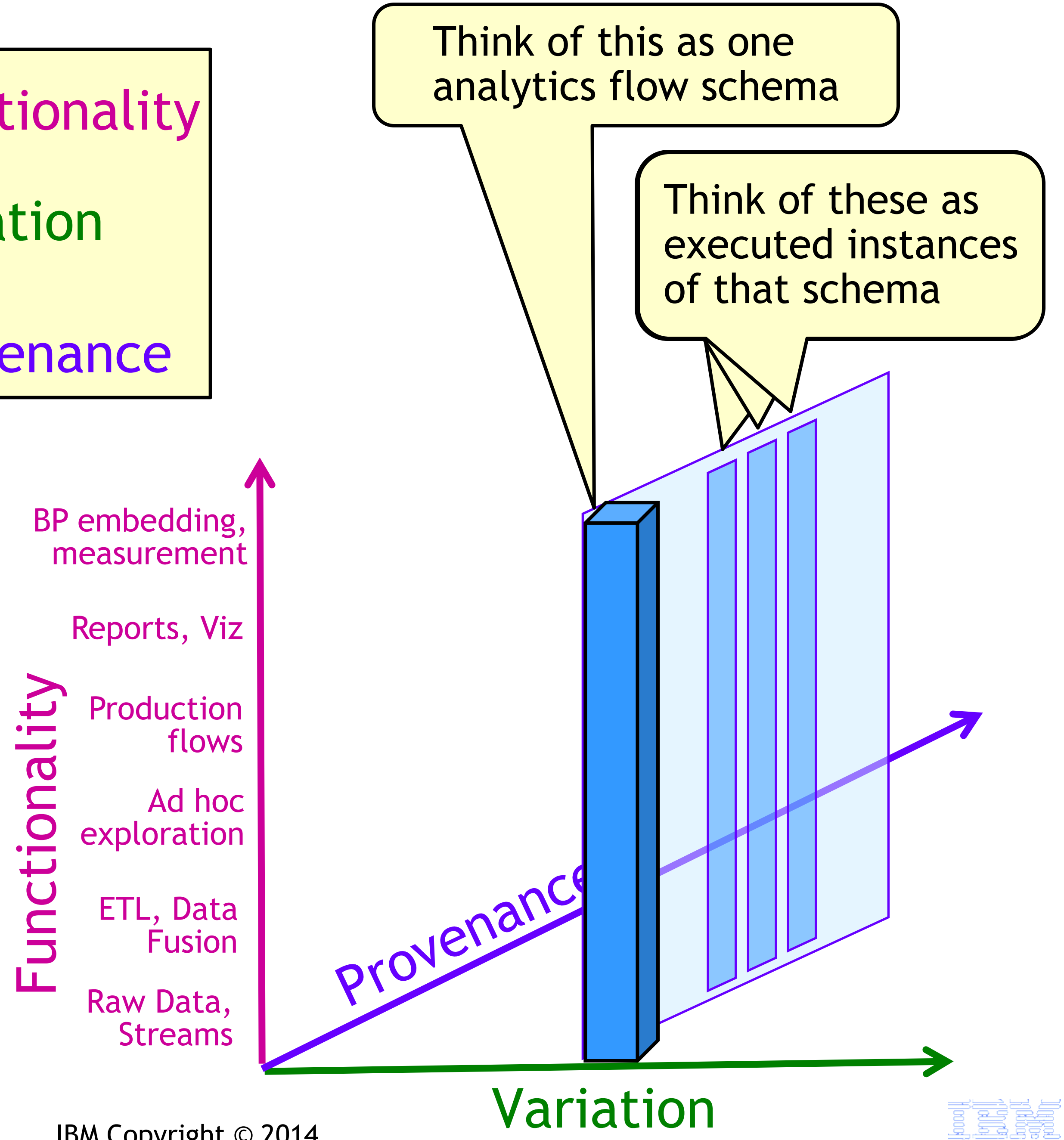
Notes:

- Flows are Directed Acyclic Graphs (DAGs)
- ***Evolution/Variation*** can be accomplished with simple manipulations, e.g., add node, delete node, etc.

IBM Copyright © 2014

Broader Perspective: A 3-dimensional view of this space




Think of this as one analytics flow schema

Think of these as executed instances of that schema

IBM Copyright © 2014

Relevant techniques/tools

- ETL
  (Extract-Transform-Load)
  - Broad array of techniques for gathering/ curating data for use in analytics/data mining
  - No higher-level tools to help workers manage/record/govern their ETL work

- CRISP-DM
  (CRoss-Industry Standard Process for Data Mining)
  - Framework for Data Mining (including refinements)
  - Primarily a methodology; comprehensive mgmt platforms not available
  - Focus on finding one-off insights

- Case Management
  - Good fit: The top-level BPs for APM are very knowledge-worker driven
  - We should identify some template schemas

- Scientific Workflow
  - The analytics flows themselves are quite similar to scientific workflows
  - However, analytics flows emphasize measurement, attribution, refinement

- BPM Adapatability
  - Frameworks/tools to manage variation of BPs, at instance level and schema level
  - Analytics flows are DAGs (simpler); but provenance and queries against collections of flows are important

- IT Governance
  - Standardized practices for ensuring that IT processes are effectively serving business objectives
  - Analytics flows are a blend of biz and IT

IBM Copyright © 2014

Some key challenges (overview)

- A precise model of Configurable Analytics Flows
  - ***Capabilities***: Provenance, Support for Measurement, Variation/Evolution
  - ***Abstraction*** over the heterogeneity of underlying components/tools

- From exploratory flow(s) to a reified flow
  - The challenge of being ***light-weight***

- Extract-Transform-Load (ETL)
  - Some tools are fairly mature but the work is ***still very time-consuming***

- Enabling rich collaboration in Analytics Flow eco-system
  - Vision for ***factorization of logical components*** to enable broad-scale, cloud-hosted crowd sourcing across all areas of APM

Additional challenges

- Case Mgmt templates for the top-level BPs for APM
- How should Case Mgmt be extended to work better on Analytics Flows?

IBM Copyright © 2014

Power of a Good Model << animated slide >>

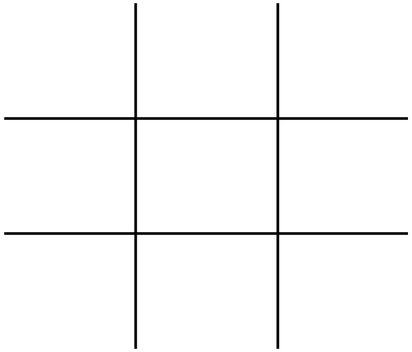
Good models go beyond description - they support action

- Selecting the right model for the job matters

Example: “Game of 15”
Winner: First one to reach exactly 15 with any 3 chips

First model – A is 🟥 and B is 🟨 – what is B’s move?

Second model –

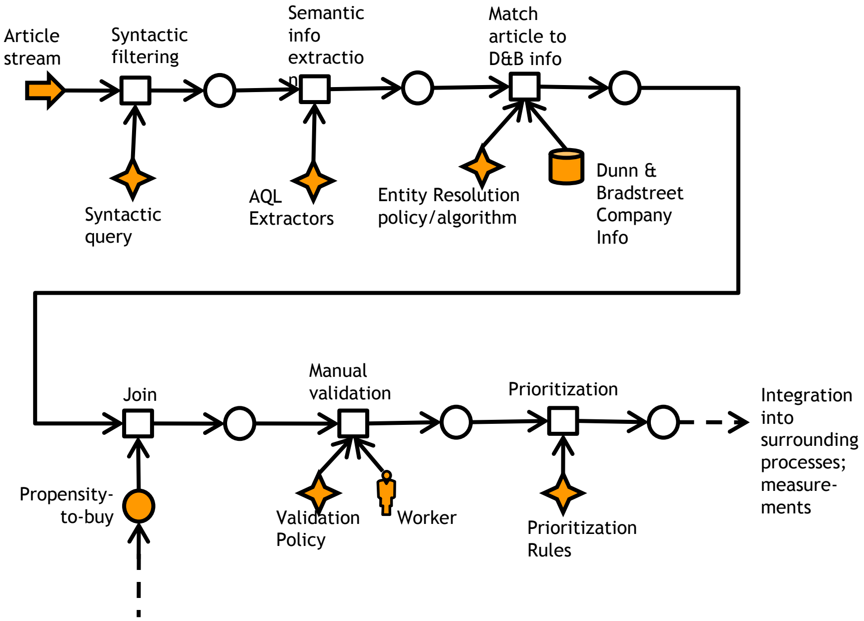


– B’s move is 6!

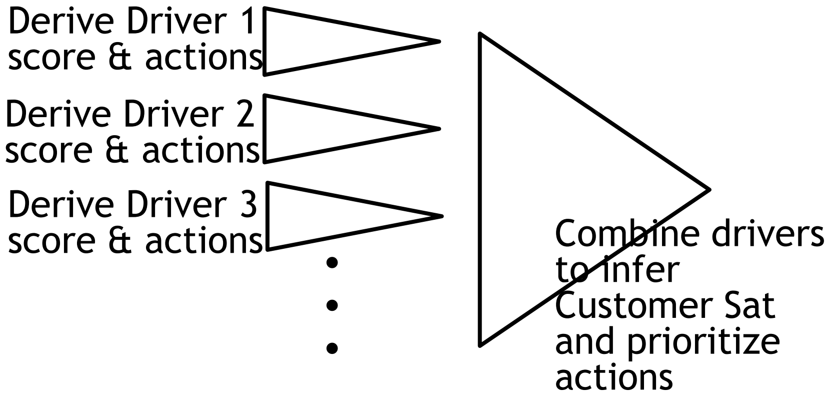
<< special thanks to David Cohn >>

Configurable Analytics Flows as a key abstraction layer

Flow from LARIAT


Article stream
Syntactic filtering
Semantic info extractio n
Match article to D&B info
Syntactic query
AQL Extractors
Entity Resolution policy/algorithm
Dunn & Bradstreet Company Info
Join
Manual validation
Prioritization
Integration into surrounding processes; measure-ments
Propensity-to-buy
Validation Policy
Worker
Prioritization Rules

Flow from Actionable Customer Sat. (abstracted)


Derive Driver 1 score & actions
Derive Driver 2 score & actions
Derive Driver 3 score & actions
Combine drivers to infer Customer Sat and prioritize actions

Is this the useful model? Is there a more useful one?

IBM Copyright © 2014

Configurable Analytics Flows: Requirements and Approaches

Requirements

- Intuitive, conceptually transparent
- Numerous ways to work with the flows
  - Ad hoc exploration
  - Re-use, including re-use of sub-flows
  - Rich query ability over large sets of flows, including visualize answers
- Enables easy comparison between flows based on measurements
  - Crucial for achieving ultimate biz goal of the analytics
- Provenance of flow outputs is intuitive, conceptually transparent
  - Important for measurement, compliance, governance

Starting points from Scientific Workflow

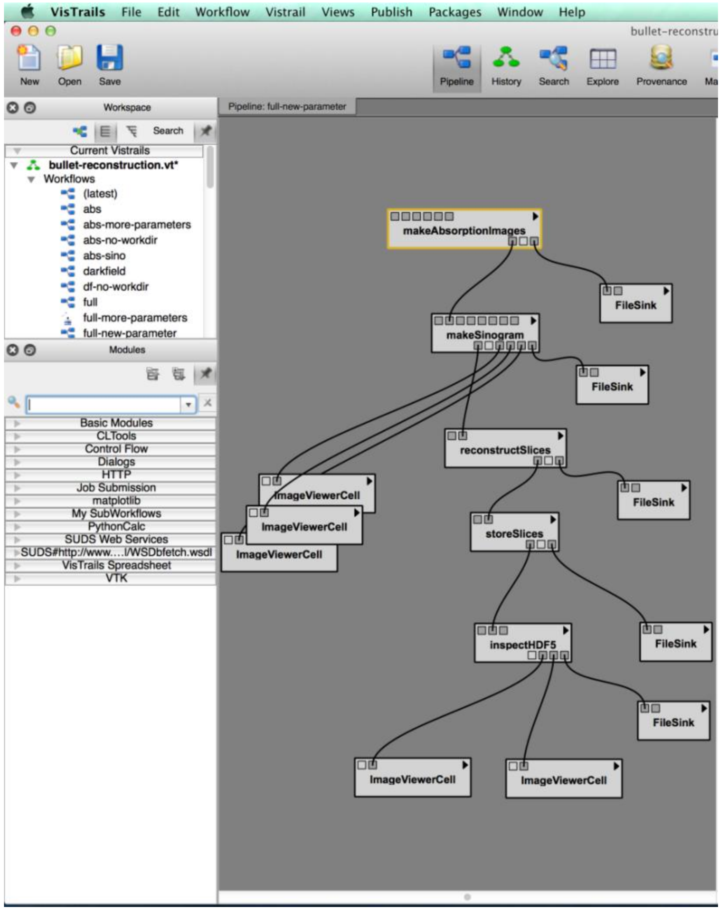
- Considerable work on provenance, executability, optimization, tools

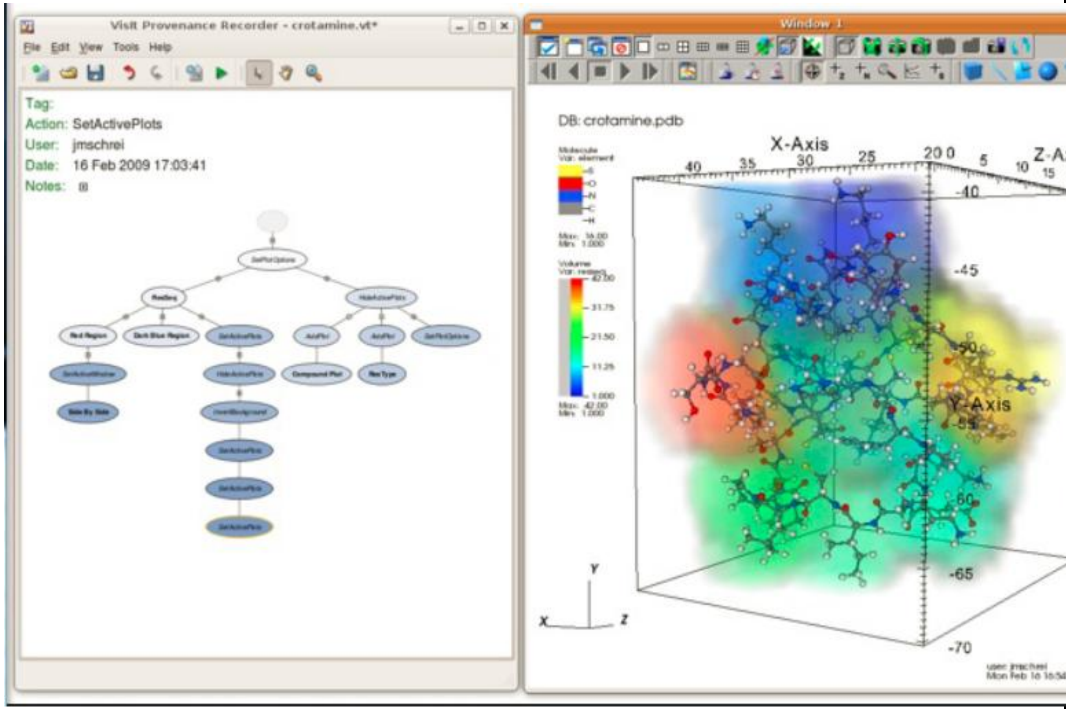
Additional research needed

- Adapt query/visualization to better support measurement
- Develop a theory of sub-flows, sub-flow composition, queries on sub-flows
- Find simple/intuitive ways to describe flows, to enable "executive level" explanations for flow outputs and differences between flows

IBM Copyright © 2014

VisTrails: Flows and Flow Provenance Tree




VisIt Provenance Recorder - crotamine.vt*
File Edit View Tools Help
Window 1
DB: crotamine.pdb
X-Axis
Y-Axis
Z-A
X
Y
Z

Tag:
Action: SetActivePlots
User: jmschrei
Date: 16 Feb 2009 17:03:41
Notes:

IBM Copyright © 2014

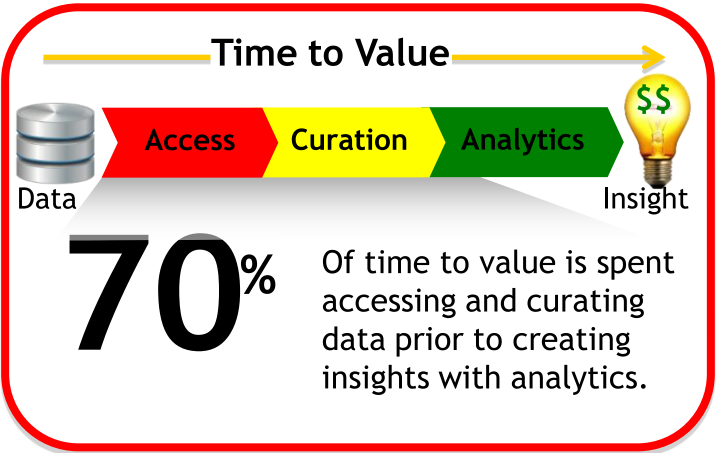
From ad hoc flows to reified flow

- Context:
  - Data scientists often explore a variety of perspectives and analytics models before identifying insights that can bring deep value
    - Heterogeneous data/tools may be used
  - Several flows might be used for testing/measurement
  - Finally, some flow(s) will be reified and put into production use
    - Perhaps in a tool different from the ad hoc exploration tool(s)
- Challenge: Data Scientists typically can't keep track of their flows
  - Capture of flows
  - Access to flows (and sub-flows): Queries over flow collections
  - Mapping from highly flexible ad hoc tools to production tool
- Starting points from Scientific Workflow
  - E.g., Kepler, Taverna, SWIFT, VisTrails use flow models, with query support
  - Approaches to "capture" of flows
    - Use operating system logs (e.g., PASS)
    - Logically centralized workflow tool – record all (e.g., Kepler, VisTrails) or delegate prov capture to components (e.g., Provenance-Aware SOA project/standard)
  - VisTrails designed to support ad hoc, exploratory flow creation
    - Focus on outputs used by humans, not embedded into BPs
    - ***Representation of sets of flows, and query access, needs strengthening***
    - ***Can we create something even more light-weight, unobtrusive (cf REST, JSON)***

IBM Copyright © 2014

The ETL Challenge

- [NY Times 8/17/2014] -- *50% to 80% of Analytics work is "data wrangling" or "data munging" or "data janitor work"*
  - Timothy Weaver, CIO of Del Monte Foods: data wrangling big data's "iceberg" issue



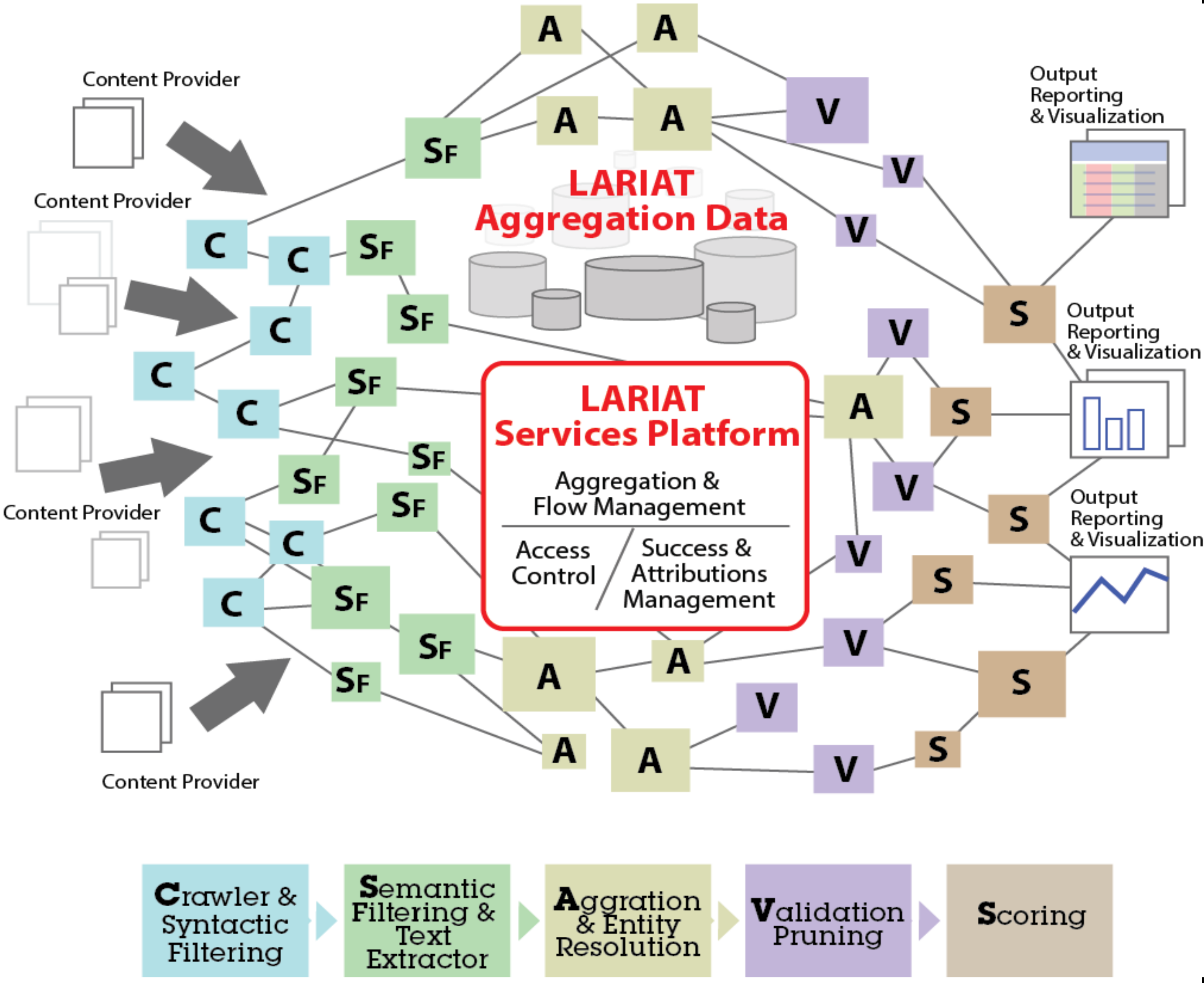
70% Of time to value is spent accessing and curating data prior to creating insights with analytics.

- *State of the art in ETL (e.g., [Chaudhuri, Dayal, Narasayya 2011]):*
  - Gather data and place into a warehouse
  - Variety of tools are now mature
    - Consistency mgmt, e.g. "…, California, Canada"
    - String manipulations, entity resolution, e.g., "California" -> "CA"
    - Extracting structure from strings, e.g., parse "Coby MP3 512MB MP-C756 – Blue."
    - Instance-level key/foreign-key idenification
    - Data load and refresh (e.g., by triggers, by log scraping)
- *The data-centric BPM community can provide help !*
  - We know: process, data mgmt, variation, knowledge work, collaboration
  - Starting point may be to apply ideas from Configurable Analytics Flows to ETL
    - Extend warehouse focus to include process-centric data capture
    - Enable capture of ad hoc ETL explorations
    - Simplify reification of "good" ETL flows
    - Enable better re-use through use of ontologies, semantic web

IBM Copyright © 2014

Vision for Factoring Analytics Flow (Illustration)

- An environment where multiple parties can contribute to different portions of the LARIAT flow?
- Data-centricity & basic analytics flow provide backbone
  - Cf. variation in traditional BPM
- Multi-tenancy:
  - Different end-users given access to subsets of flow & output
- Compensation based on Attribution
  - Challenge: how to determine attribution



[Callery et al, SCC, 2014]

IBM Copyright © 2014

Conclusions / Call to Action

- Analytics Process Management (APM) is the next big research challenge in BPM
- Data-centric BPM community is best positioned group to attack this
- Case Management is well-suited for the top-level BPs of APM
- Configurable Analytics Flows are a good abstraction layer for modeling the fundamental activity of APM
- While Scientific WF provides a starting point, there are many challenges in adapting to the BP context
  - Stemming from repeating flows, heterogeneity of stakeholders, measurement & feedback loop, explanation to executives, …
- [Troung and Dustar 2012]: "Research on how to manage analysis algorithms and how to provide an open platform for third parties to develop, search and share algorithms is quite open."

IBM Copyright © 2014

Acknowledgements

This thinking is based on projects/collaborations with several people, including

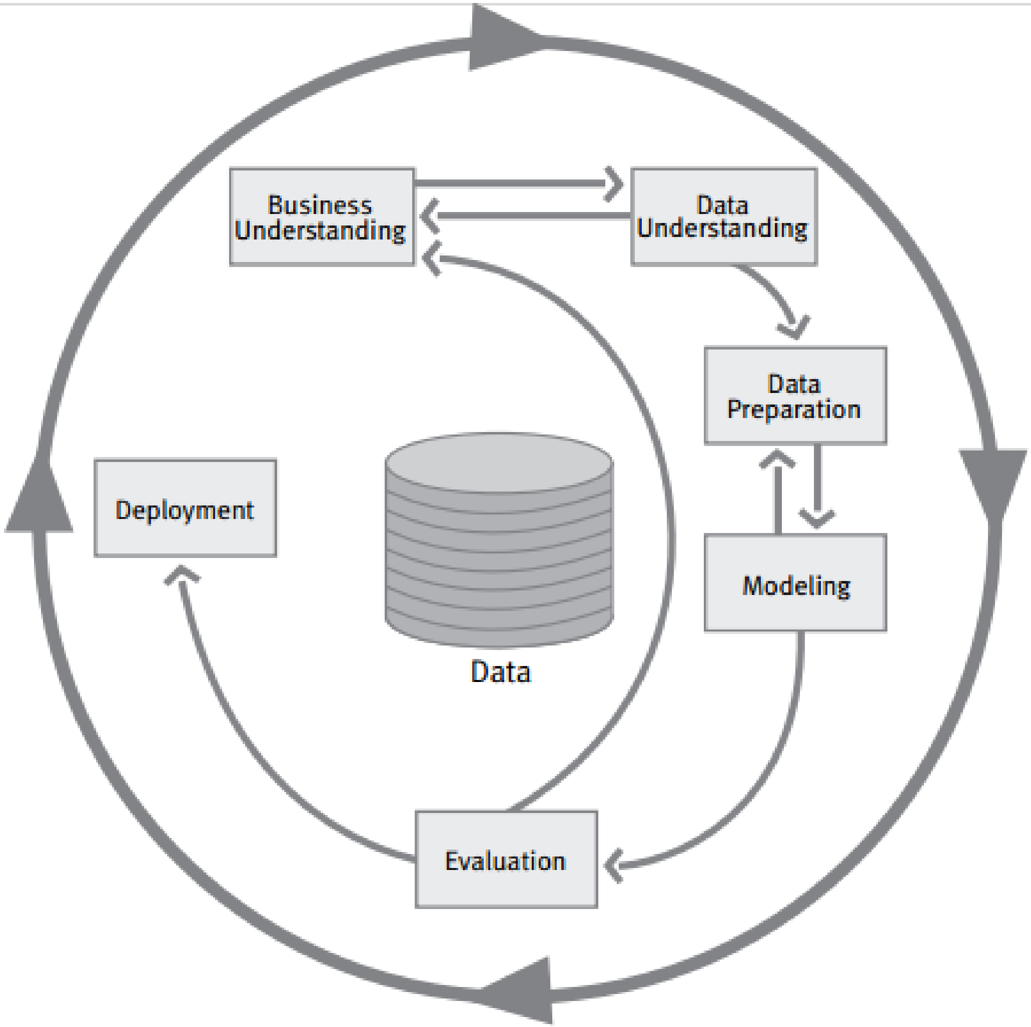
- LARIAT: Matt Callery, Terry Heath, Danny Oppenheim, Noi Sukaviriya, Roman Vaculin
- Actionable Customer Satisfaction: Krishna Ratakondra, Jeff Robinson, Anshul Sheopuri, Dashun Wang
- BOLO: Elham Kabheri, Yang (Daniel) Li, Matt Reiman, Roman Vaculin

IBM Copyright © 2014

Backup slides

IBM Copyright © 2014

CRISP-DM: Standardized method for performing iterative Data Mining



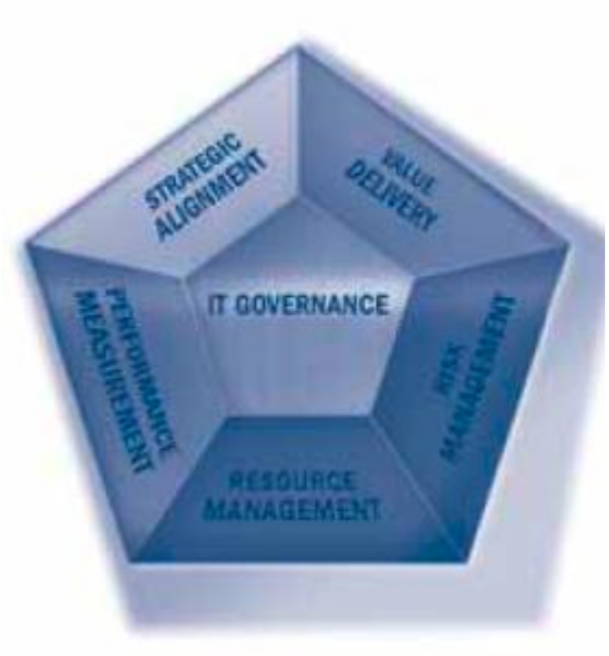
Cross-Industry Standard Process for Data Mining

- Identify business challenges & questions
- Understand the available data
- Prepare data
  - Cleansing
  - Transformation
  - Integration
- Create analytical model(s)
  - Myriad of alternatives to fit broad variety of applications
- Evaluate & refine models
- Deploy

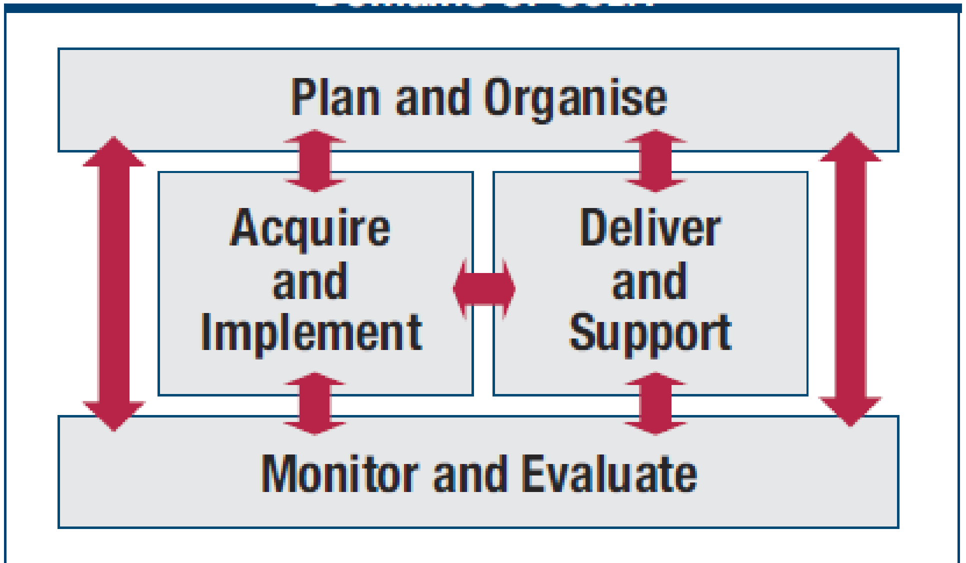
- Iterate
  - 1 or 2 month cycle
  - Each iteration builds value, infrastructure and experience

IT Governance (COBIT)

The 5 focus areas in COBIT



The 4 interrelated Domains of COBIT



Plan and Organise

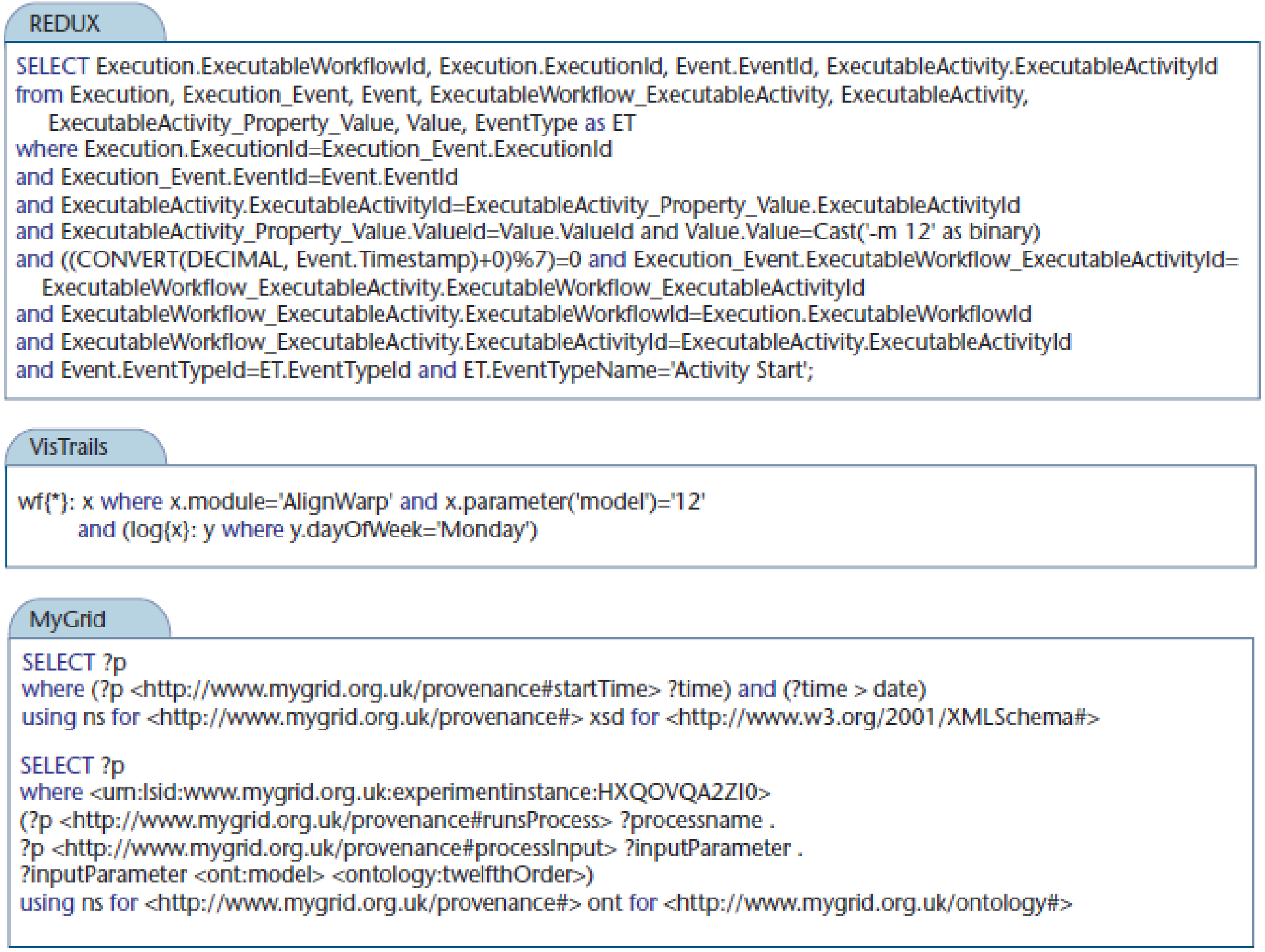
Acquire and Implement

Deliver and Support

Monitor and Evaluate

- COBIT assumes a fairly rigid separation between IT and Biz roles
- With Analytics Flows, some roles lie at interface of IT and Biz, e.g.,Data Scientist, UI Designer/implementer
- Approaches to manage and measure these roles requires an extension of COBIT

Querying sets of flows in Scientific WF Systems



REDUX

```
SELECT Execution.ExecutableWorkflowId, Execution.ExecutionId, Event.EventId, ExecutableActivity.ExecutableActivityId
from Execution, Execution_Event, Event, ExecutableWorkflow_ExecutableActivity, ExecutableActivity,
    ExecutableActivity_Property_Value, Value, EventType as ET
where Execution.ExecutionId=Execution_Event.ExecutionId
and Execution_Event.EventId=Event.EventId
and ExecutableActivity.ExecutableActivityId=ExecutableActivity_Property_Value.ExecutableActivityId
and ExecutableActivity_Property_Value.ValueId=Value.ValueId and Value.Value=Cast('-m 12' as binary)
and ((CONVERT(DECIMAL, Event.Timestamp)+0)%7)=0 and Execution_Event.ExecutableWorkflow_ExecutableActivityId=
   ExecutableWorkflow_ExecutableActivity.ExecutableWorkflow_ExecutableActivityId
and ExecutableWorkflow_ExecutableActivity.ExecutableWorkflowId=Execution.ExecutableWorkflowId
and ExecutableWorkflow_ExecutableActivity.ExecutableActivityId=ExecutableActivity.ExecutableActivityId
and Event.EventTypeId=ET.EventTypeId and ET.EventTypeName='Activity Start';
```

VisTrails

```
wf{*}: x where x.module='AlignWarp' and x.parameter('model')='12'
       and (log{x}: y where y.dayOfWeek='Monday')
```

MyGrid

```
SELECT ?p
where (?p <http://www.mygrid.org.uk/provenance#startTime> ?time) and (?time > date)
using ns for <http://www.mygrid.org.uk/provenance#> xsd for <http://www.w3.org/2001/XMLSchema#>

SELECT ?p
where <urn:lsid:www.mygrid.org.uk:experimentinstance:HXQOVQA2ZI0>
(?p <http://www.mygrid.org.uk/provenance#runsProcess> ?processname .
?p <http://www.mygrid.org.uk/provenance#processInput> ?inputParameter .
?inputParameter <ont:model> <ontology:twelfthOrder>)
using ns for <http://www.mygrid.org.uk/provenance#> ont for <http://www.mygrid.org.uk/ontology#>
```

Figure 5. Provenance query implemented by three different systems. REDUX uses SQL, VisTrails uses a language specialized for querying workflows and their provenance, and myGrid uses SPARQL.

- REDUX:
  SQL against underlying relational store
- VisTrails:
  domain-specific query language
- MyGrid:
  SPARQL against RDF store

[Freire et al, CISE, 2008]