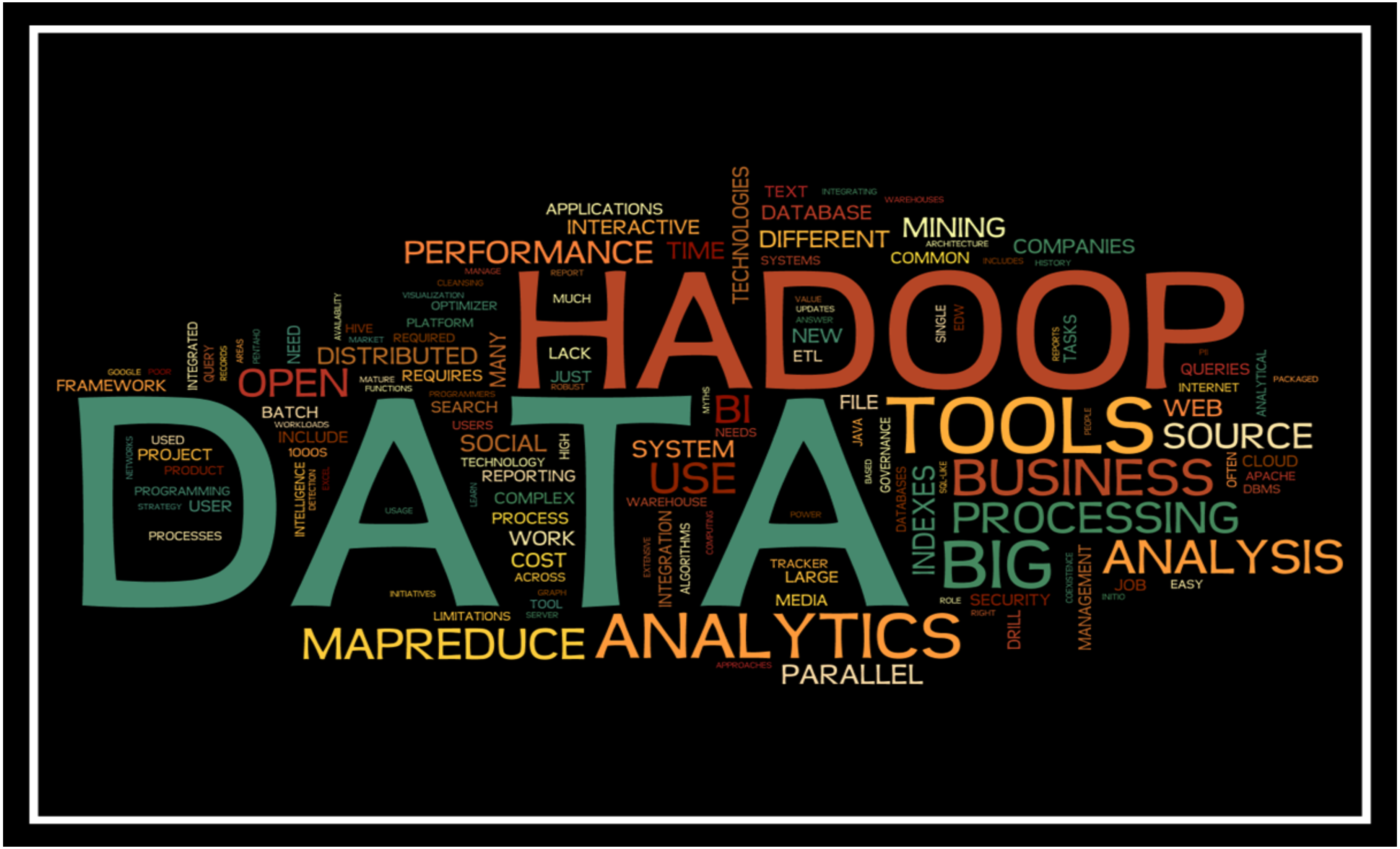
Big Data Overview



About Us?

- Jeff Wilts
  - AVP Information Management @ Canadian Tire
  - Jeff.Wilts@Cantire.Com
- Feel free to connect on LinkedIn

Big Data – Current Scope

Develop an Enterprise Big Data Data Hub, which will:

Ingest data from a variety of sources, regardless of:

- Data types
- Data format
- Volume

Aggregate and/or join the data across the sources

Provide the data to one or more external analytic tools for use in decision making processes

Allowing us to more effectively and efficiently answer business questions regarding trends, forecasts, etc.

Data Driven Decision Making

Big Data helps define and answer questions, and ultimately should change the way we do things here at Canadian Tire.

How do we optimize product assortment?

How do we understand our customers and communities better?

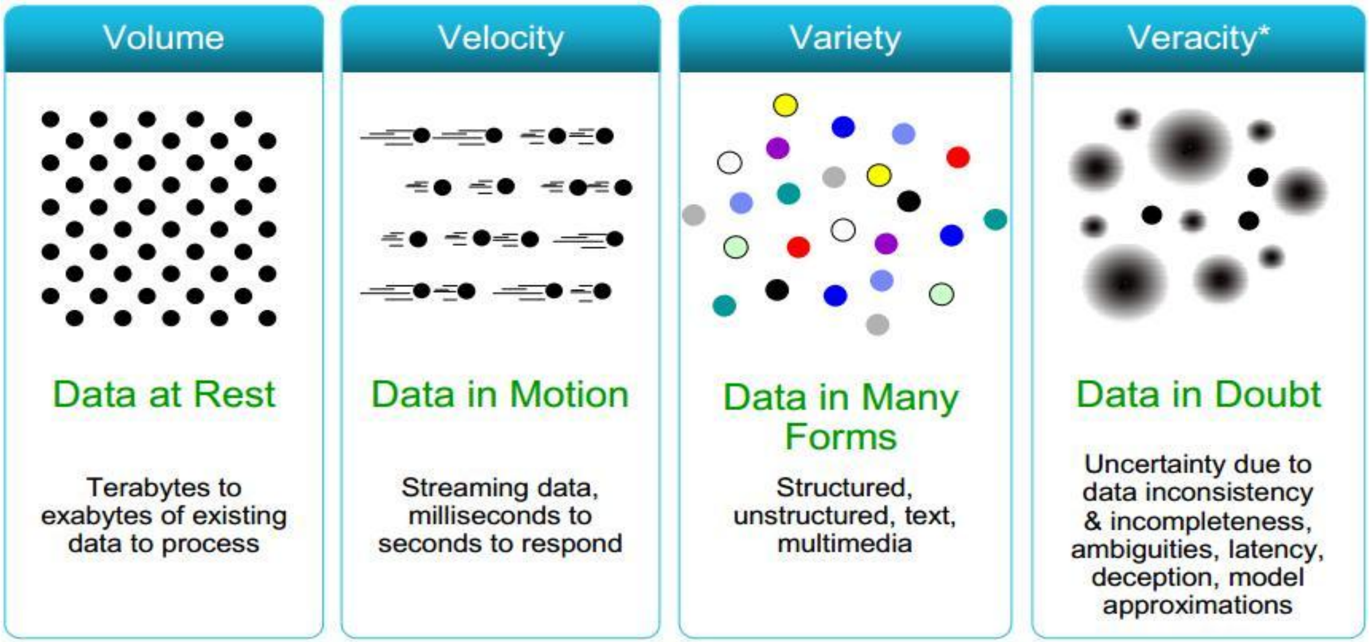
How can we lower lead times, improve in stock position, but also lower waste?

What is the impact of eCommerce promotions and activity beyond digital?

How much influence is competitor, price, weather, demographics, customer service, product quality, special events having on store sales?

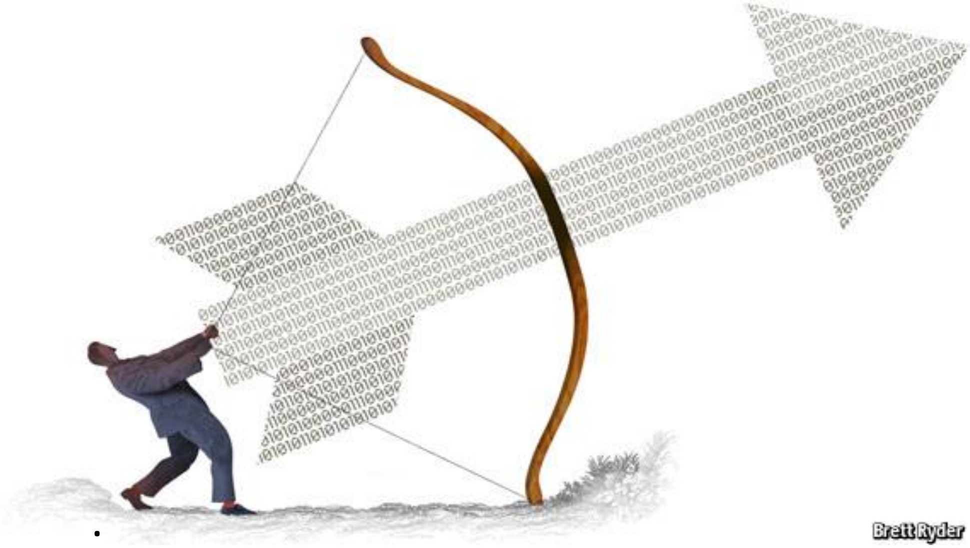
Big Data

- New emerging "class" of analytics
- Requires new tools, processes, and thinking
- Has ethical ramifications



"Big Data" requires at least one of these 4 dimensions exist

Big Data Benefits Are Realized From Incremental Improvements



In the past, cost of data collection forced off line businesses to look for big wins

But Big Data enables new equation
Small changes
X huge number of instances
X long time
= Significant Benefit

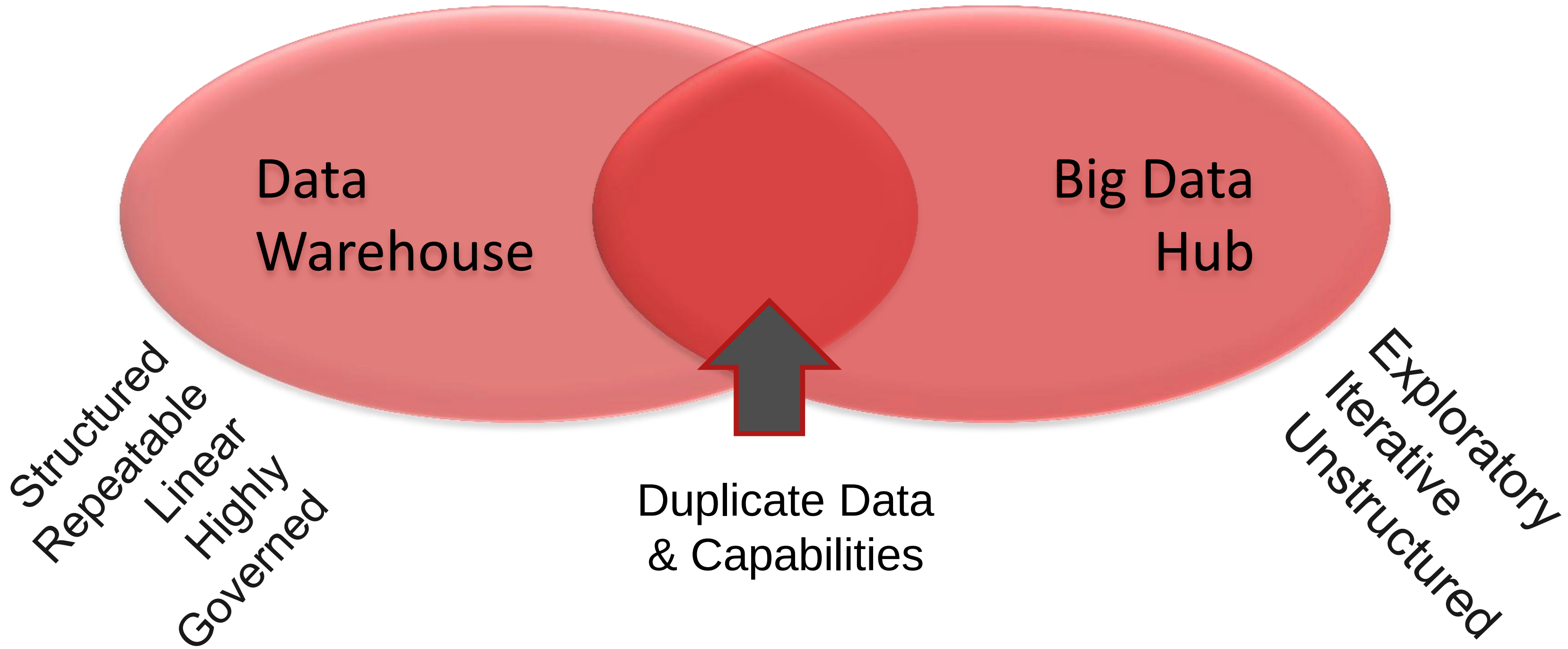
.

UPS's 60,000 delivery vans, cutting each route by just one mile saves $50m in fuel and other costs a year.

Individual improvements of 0.5% to 1% in productivity add up to a 22% rise in the teams' overall productivity.

Endless trial-and-error testing of small things can be worth it.

But We already Have a Data Warehouse?



DW	DW	BD	BD
Auditable	Linear	Fast Fail	Creative
Governed	Model for Quality	Cheap & Fast	Volume Hides Noise
Model what is important	SQL	Land Everything	Java / Map-Reduce

Data Warehouse vs. Big Data - Focus

Area	EDW	Big Data
Focus	Reports, KPIs, trends	Patterns, correlations, models
Process	Static, comparative	Exploratory, experimentation, visual
Data sources	Pre-planned, added slowly	Chosen on the fly, on-demand
Transformation	Up front, carefully planned	ELT, on-demand, in-database, enrichment
Data quality	Single version of truth	Tolerant of “good enough”; probabilities
Data model	Logical / relational / formal	Conceptual / semantic / informal
Results	Report what happened	Predict what will happen
Analysis	Hindsight	Forecast, Insight

How Did We Get Started?

- To deploy Big Data required a number of decisions.

Care & Foster

- Who Does:
  - Ongoing builds
  - Admin & Support
  - Training & Governance

Use Case

- What to do first?

SI

- Who will help
  - Tech Stand Up
  - Use Case Build

Technology

- Which tools
- How (Cloud, Internal etc)

Key Tools Used For Our Business Cases

- Apache Hadoop – the most advanced Big Data platform at the moment. It allows processing of large data sets across clusters of computers using simple programming models
- Apache Hive – data warehouse infrastructure built on top of Hadoop
- R – software environment for statistical computing and graphics
- Java and Python – most popular general purpose programming languages at the moment
- Pig – High level programming language for creating MapReduce programs in Hadoop

CTC EIM Technology Stack

Development
- Pig
- Access
- python
- Hive
- PL/I
- PL/SQL
- Kornshell
- DataStage
- Java
- The Perl Programming Language

Connectivity
- Flume
- WebSphere MQ
- Ipswitch File Transfer
- Connect:Direct
- Sqoop

Productivity
- PVCS
- maven
- Excel
- DbVis Software
- Quest Software

Database and Storage
- IBM MDM
- hadoop
- IBM DB2
- Microsoft SQL Server 2012
- Oracle
- Netezza
- cloudera – Ask Bigger Questions

Data Quality, Governance and Security
- Vormetric
- Data Advantage Group
- ataccama

Operations and Monitoring
- Apache Tomcat
- APPFLUENT
- Tivoli software

Analytics & Reporting
- IBM Information Analyzer
- Cognos
- sas
- MicroStrategy
- Impala
- R
- Google Developers – Google Charts

CTC Applications
- InTireNet
- Business Information Glossary
- Automotive Infrastructure

Platform Tools
- Microsoft Windows Server System
- z/OS
- Linux
- AIX
- vmware
- WebSphere

How We Evaluated Tools

Categories
Vendor responsiveness
Open Source Commitment
Solution Characteristics
SQL Capabilities
Administration
High Availability
Security and Access Logging
DR Solution
Costs
Licensing and Support Costs

Tool Selection Summary Lessons

- Lots of options for tools
  - Vendors will give you conflicting information
- Accurate sizing is problematic
  - You will get it wrong initially
- Virtual vs. Physical is important
- Install vanilla
  - Products are still immature

Our Partner(s) - SI

- We didn't have skills in house.
- Takes lots of roles to build / manage / exploit

S.I. considerations

Lots of companies can provide SI expertise:

- **Direct by Tools Vendor**
- **Big Data Specialist Firm**
- **General SI / Consulting Firm**

SI Evaluation Criteria

Strategic Considerations
- Company Viability
- Skills Maturity
- Ability to Execute
- Company Direction
- Ease of doing Business

Experience
- Technology Stack
- Use Case
- Industry
- Bench Size / Depth

Technical Fulfillment
- Technical Innovation
- Platform Skills
- Self Service Skills
- Communications
- Scalability
- Solution Flexibility

TCO
- Initial Rates
- Start up costs
- On-going support costs
- Onshore / Offshore

Support
- Unique skills required
- In-house support
- Vendor relationships
- Vendor Accreditations
- Support risk

Full Outsource ⟷ Full In House

Care and Foster

Current State:

- Very little in-house technical knowledge of Hadoop technology or administration requirements.

Actions:

- Determine how environment is to be administered and maintained
- Train development and administration staff in Hadoop technology
- Create governance model around Hadoop infrastructure
- Communicate and integrate new CTC-focussed Hadoop standards and best-practices internally and to our development partners

Questions:

- Who provides Training and input to Governance/Standards?
- How do we ensure fast-fail doesn’t become ungoverned mess?
- How do we manage 2 types of uses? (Systematic & ad hoc)

Care and Foster Selection Criteria

- Experience in building and assisting customers with Hadoop Technologies
- Technology Focus (Open Source vs. Proprietary)
- Monitoring and Administration
- High Availability Tools
- Security Infrastructure
- Skills Required to Support/Maintain infrastructure similar to CTC

The Business Case

- NIMBY
- Big Data needed to be tied to enhancing a corporate operational function.
- We have positioned as an enhancement to other things.
  - Improve Quality
  - Faster turn around
  - More complete insights / visibility
  - Lower long term costs

Use Cases We Considered First

What We Settled On First

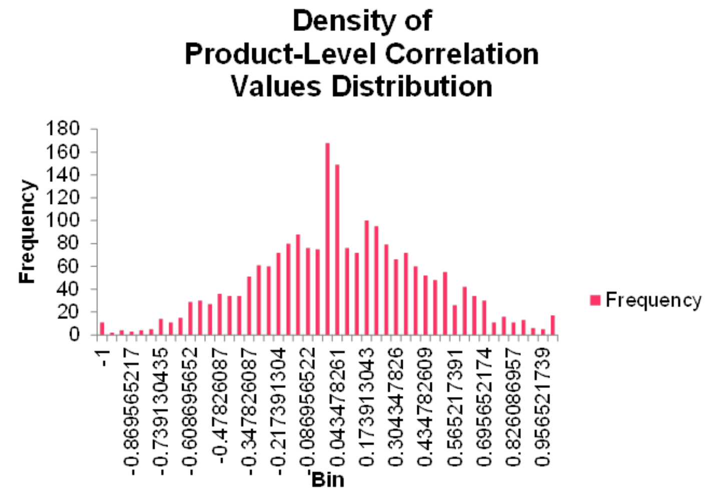
- Learn all we can about what it takes to run a Big Data Environment

- In depth POS analysis and what drives sales?

What we have Learned - Technically

1. Integration – How to retrieve and load data, what tools and patterns
2. Analytics – How to organize data and use tools to enable analytics
3. Architecture and Operations – Learning curve on scaling up for performance, security, backup and recovery etc

- Extract and Load using Sqoop
- Query and manipulate data using Hive/Pig
- Correlate Sales to Product Reviews using R Hadoop
- Export to Excel for visualization



Big Data – Technical Insights

- Even a small environment (15 nodes) with minimal support enables a lot of capability.
- Environment scales and fault tolerance works.
- Built templates that significantly reduces time it takes to land data (weeks to days)
- Different skills required to use load, manipulate and analyze data
- Pig, Hive, R – External skilled resources were essential
- Hadoop Administrator was required – Have now hired

More strenuous test

Build a foundation for explaining changes in store sales

- **What was impact of price, weather, competitor, etc**
  1. 3 Years POS Sales – All products all stores
  2. 3 Years of Weather Data – All Weather Stations, every day
     - Temperature, snowfall with lags, rainfall, first snow event
  3. Product Rating from eCommerce customers
  4. Competitor – type and presence
  5. Store retail area size
- Multiple Linear Regression was used as mathematical modeling tool
- Determine sensitivity of every product sold over 3 year period to price, weather competitor, and product review

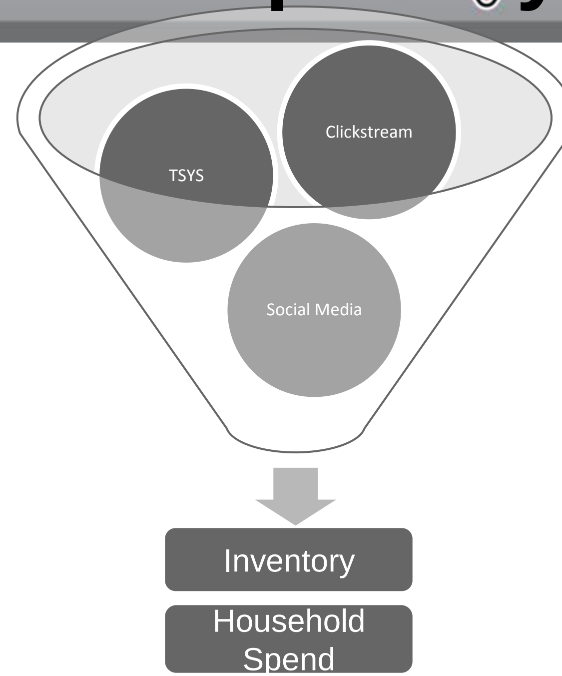
- Apply those factors to individual stores to explain year over year change in sales

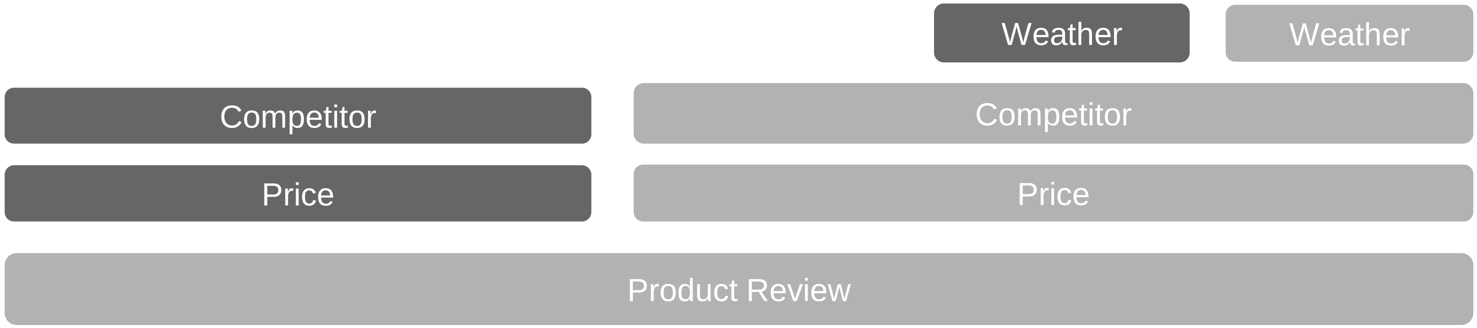
Case 1 – Sales Impacts Analysis

Sales Impact Use Case Proved Capability

Build model that provides the foundation for analyzing and understanding the factors that influence year over year change in store performance

- 2.6 Billion POS transactions 2011-2013 aggregated to monthly level
- Linear model allowed approx. R^2 = 30% of monthly sales revenue variability to be explained by price, store size, product review ratings and presence of competitor.





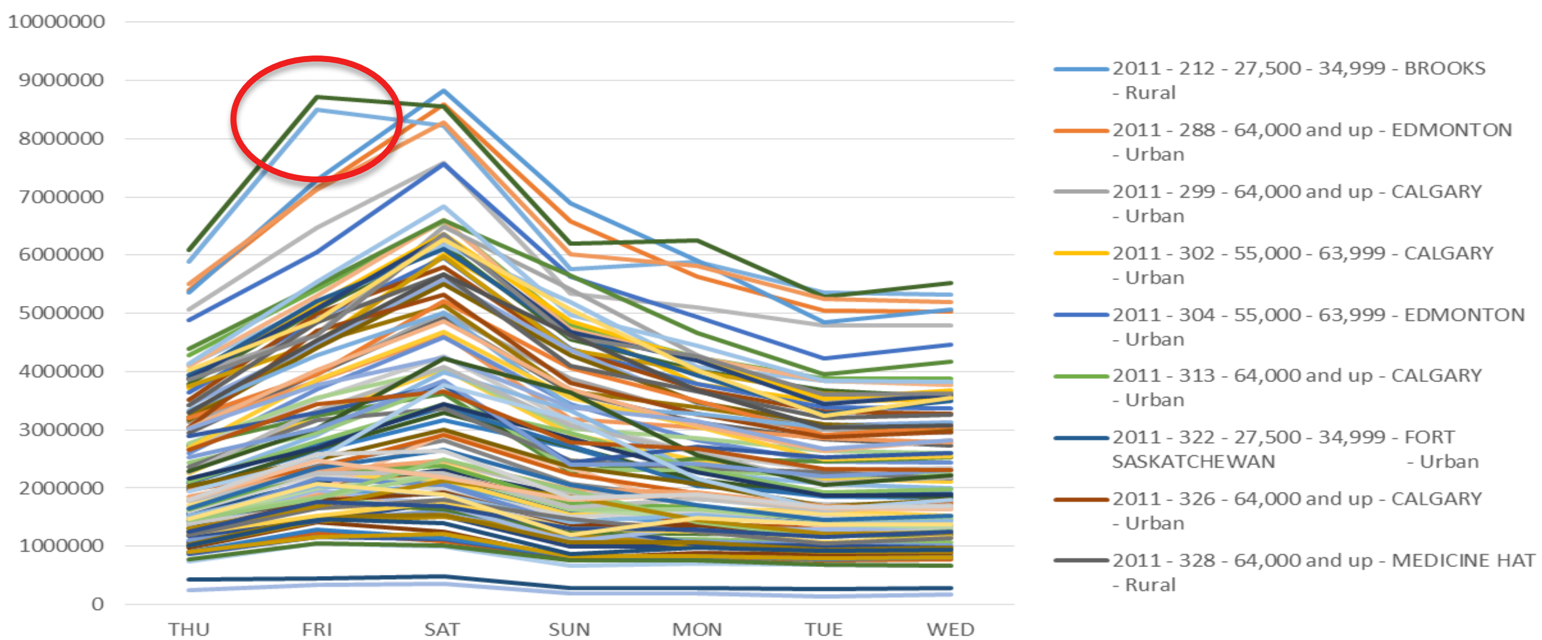
Incrementally add factors and mature model

Example: Top Products That Responded to Weather Model

Product	Fine Line	Sub Category	Factors
CABLE,ROOF 80'LENGTH	ROOF HEATING CABLE	ROOF REPAIR & MAINTENANCE	Snowfall, Rainfall
TWINKLER PINS ASST	HALLOWEEN COSTUMES AND ACCESSORIES	HALLOWEEN & HARVEST	Snowfall, Rainfall
JUN WICHITA BL 3G	UPRIGHT EVERGREENS	NURSERY	Temperature, Snowfall (next week)
FAKE BLOOD	HALLOWEEN COSTUMES AND ACCESSORIES	HALLOWEEN & HARVEST	Rainfall, Temperature
LILC FR HYBRD STD 7G	TOPIARIES AND STANDARDS	NURSERY	Snowfall
205/55R16 94V AL PA3	SOP MICHELIN WINTER TIRES	Special Order Winter Tires	First Snow, Snowfall (week before)
PEL STV 2000 SQ FEET	WOODSTOVES	WOODSTOVES & ACCESSORIES	Snowfall
WP9035 NEW WTR PMP	Water Pumps, New	ENGINE COOLING	Temperature, Rainfall
FG0279 FUEL PUMP	Fuel Pumps, Electric	FUEL SYSTEMS	Price, Snowfall

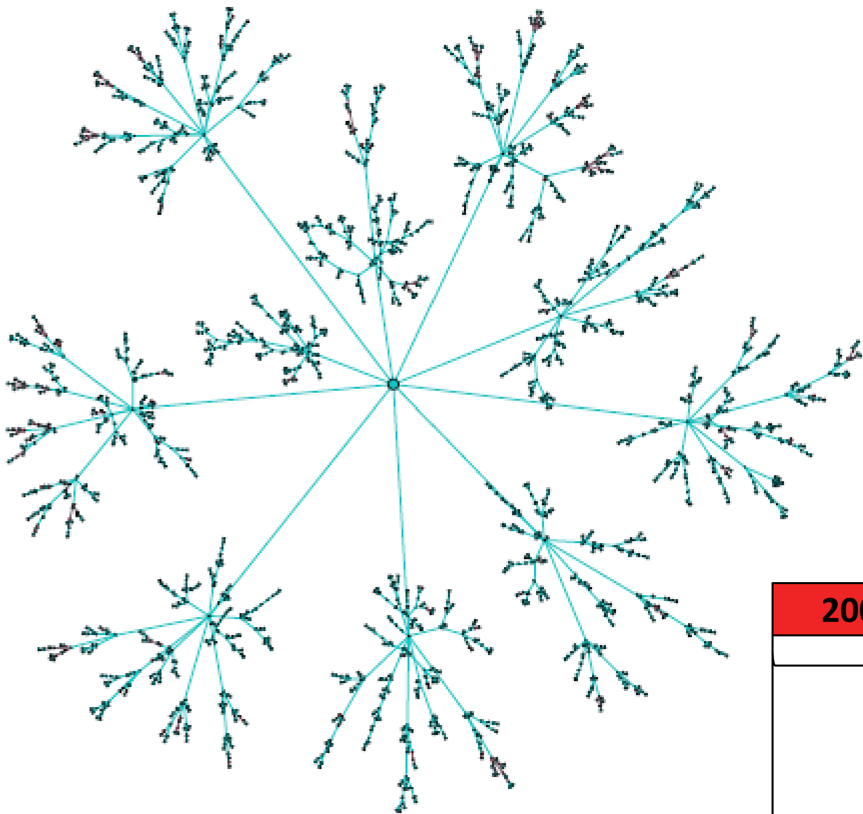
Exploring Weekly Sales Patterns

These Rural Stores have a different pattern, why? Are they particularly good at Friday's or are they poor at weekend's? How can we exploit and learn from this?



Case 2 – Swimming Pools

Predictive Modeling Overview



Decision tree learning uses a decision tree as a predictive model which maps observations about an item (store/SKU in our case) to conclusions about the item's target value (expected sales). For the pools modeling challenge, team has executed several decision tree based algorithms:

1. **C5 Decision Tree**
2. **CHAID Decision Tree**
3. **Random Forrest**

The pools model uses ~175 variables and there are many paths that the Store/SKU combinations follow – thus, a detailed review and validation of the logic embedded in the tree is difficult. So how can we judge accuracy? We test the model against unseen data.

2009	2010	2011	2012	2013	2014

1. Build model on historical data

2. Test on unseen data

3. Generate 2014 recos

<u>Benefits of predictive modeling</u> are that it is:

1. **Fact based** (i.e. business intuition and insight are used to make hypotheses; science is used to test them)
2. **Measurable** (i.e. we can compare models/approaches for accuracy in advance)
3. **Repeatable** (i.e. results will reflect new data inputs)
4. **Improvable** with new data or new insight

Predictive Modeling for Pools – Summary of data used

Team explored a large dataset (5 years of history where available) for predictive value in explaining the year over year variability in pool sales

- 490 CT Stores & Dealers with unique behaviors
- 49 pool SKUs in 3 finelilnes
- 1,600+ FSAs w/ 830,000 + postal codes
- Promotions
- 9.5M Family Households
- 5M+ Sales Transactions
- Pool opening and closing kits sales trends
- 156K + OMC Transactions
- 2,600+ Weather stations
- 9500+ Stockouts
- Vehicle counts in store trade areas
- 14.3M dwellings

Output:
Forecast for every Store/SKU combination that can be rolled up to any regional level as required

Case 3 – Automotive Assortment

What is the Big Deal?

The automotive parts business is unlike any other business at Canadian Tire

- It has a unique complexity, that if we are to be successful, requires a different approach
- Our customers needs are best met when we are knowledgeable about their cars, about the repair cycles for the various parts their vehicles require and our customers propensity to wait (or not) for the parts they need

Our customers needs vary:

- With the age, price, make, model and engine size of their vehicle
- With the amount of miles the vehicle is driven, the condition of the roads that are driven on
- How the vehicle is used and the time of year (season)

- They also vary based on where the customer lives as the climate across the country impacts wear of different components

- There are many different factors that influence need and the purchasing decision

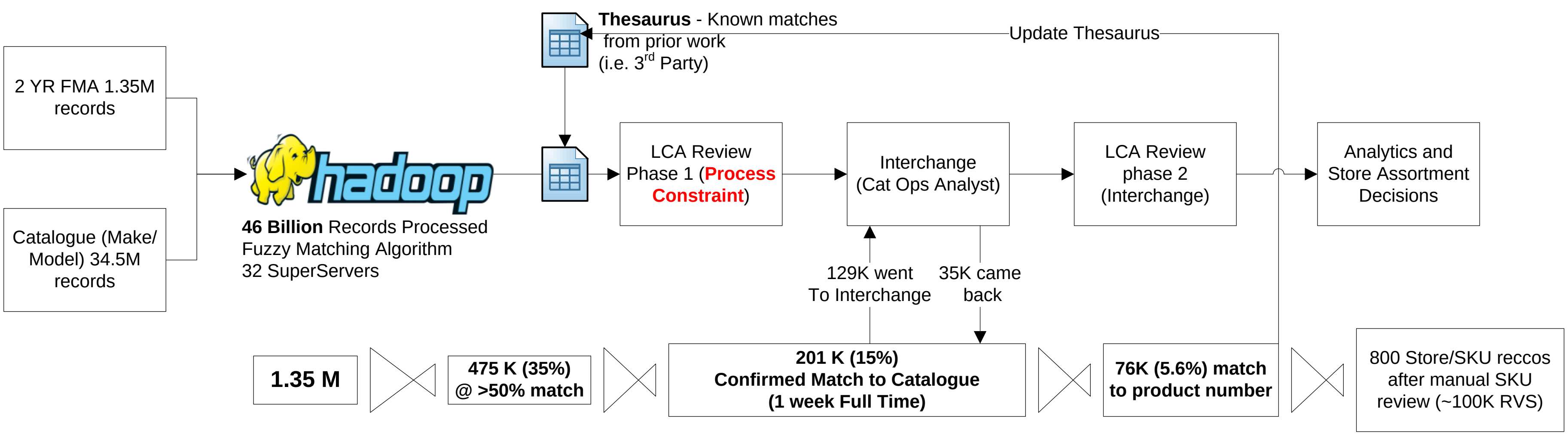
- The result of all of this is that no two stores markets are the same, and therefore no two stores assortment needs are the same

The complexity of the Auto Parts business requires a unique assortment planning approach

Hadoop cluster enabled us to decipher a far greater proportion of free-form text FMA transactions

17525	1	70%	07K145215A	07K109345		38%	SEAL	CAMSHAFT SEAL	100%	VOLKSWAGEN	VOLKSWAGEN	100%	BEETLE	BEETLE
128836	1	70%	CA707B	20707		38%	Control Arm	R. CONTROL ARM	100%	FORD	FORD	100%	MUSTANG	MUSTANG
106706	1	70%	22-382	02820		38%	rack pinion napa	PINION RACE	92%	DODGE-RAM	DODGE-RAM TRUCK	100%	RAM 1500 PICKUP	RAM 1500 PICKUP
161498	1	70%	260-5240	CAK540		38%	control arm	UPPER CONTROL ARM	100%	CHEVROLET TRUCK	CHEVROLET TRUCK	100%	SILVERADO 1500 PU	SILVERADO 1500 PU

23497	1	93%	T-48	T48		86%	COOLANT BOTTLE	COOLANT REC TANK CAP	100%	FORD	FORD	100%	FOCUS	FOCUS
53425	1	93%	NU1735	MU1735		86%	FUEL PUMP ASSEM	FP MODULE ASSEMBLY	92%	CHEVROLET	CHEVROLET TRUCK	100%	BLAZER (S10)	BLAZER (S10)
157680	1	93%	7L1Z-1A189-/	9L3Z 1A189-A		86%	TPMS-4	TPMS SENSOR	100%	FORD	FORD	100%	FOCUS	FOCUS
2098	1	93%	PT2658	2658		86%	FRONT A/T SEAL	A/T FRONT SEAL	88%	FORD	FORD TRUCK	100%	RANGER PICKUP	RANGER PICKUP



Insights

Use Cases

- **MOST IMPORTANT DECISION**
- Start with a real business problem.
  - If you can't explain the problem then don't start.
  - Learning about the technology is *not* a use case.
- Make sure use case ties to corporate strategy.

- Think iteratively (Start Small and Grow)

Big Data – Business Insights

On the path to testing capability, we picked up a few insights

Detection of Foreign Merchandise in Automotive

- Fuzzy logic works faster and can handle more complexity

Parts Reorder Decisions in Automotive

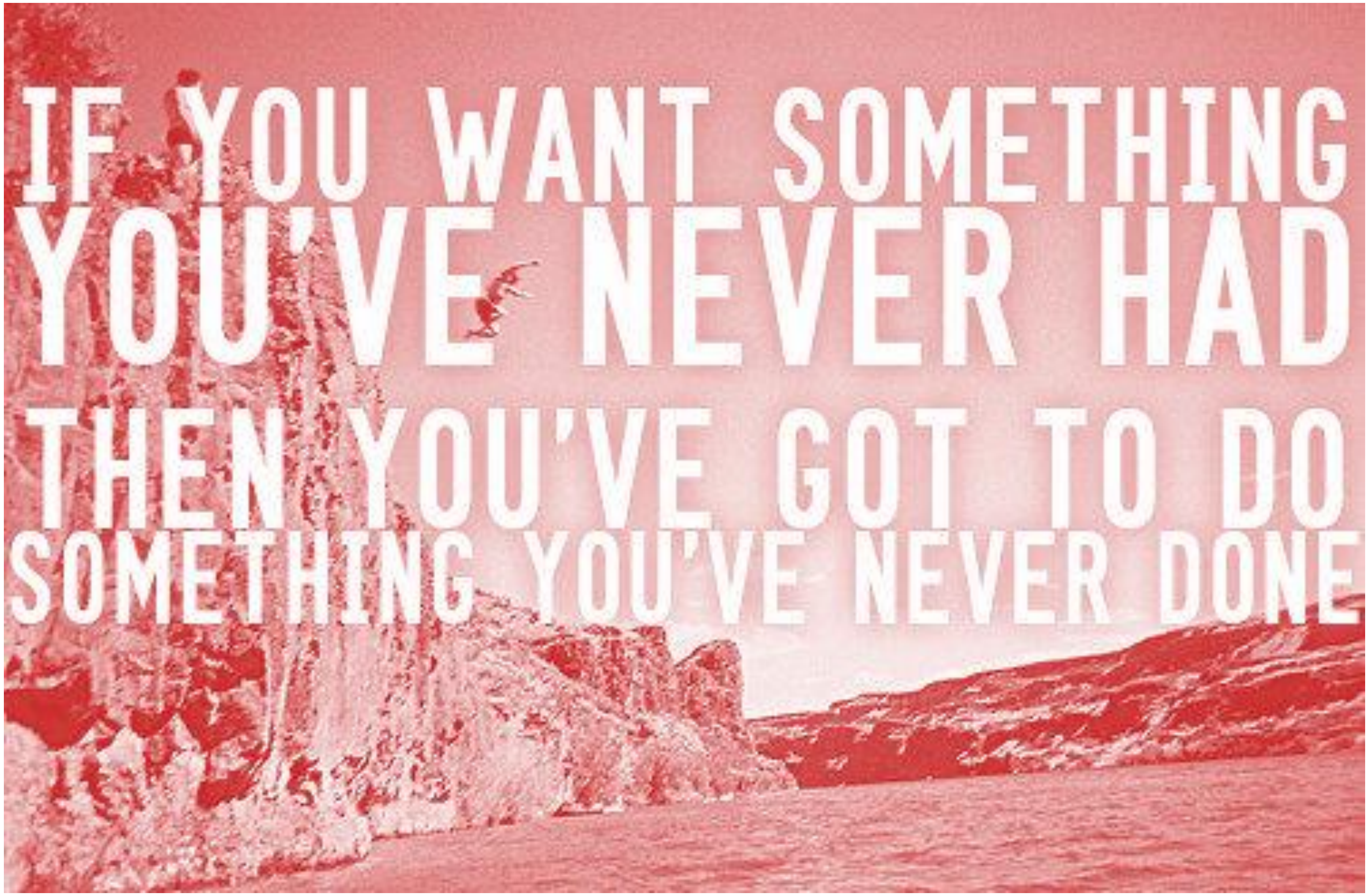
- Machine learning (random forest) provides better faster decision trees

Store Performance Drivers – Multiple Regression

- Quantifying the combined impact of weather, price, competitor is possible and this type of analysis should be continued to increase confidence.
  - Then include out of stock, customer service, community events, flyer, ecommerce etc.

- Market basket clustering and visualization is easily done

Where to Next?



Retail Understanding

Merchandising		Store Operations	Marketing	Supply Chain	Finance
Customer Loyalty Programs	Customer Behavior Analysis	Optimized In-Store Experience	Market Share Analysis	Delivery Precision	Enterprise Financial Value Model
Individually Tailored Offers	Market Basket Analysis	Store format, profile, location & space	Positioning & Competitor Analysis	Optimized in-Stock Management	Profit & Loss Analysis
Customer Surveys	Localized Assortment mix & Cannibalization Analysis	Assortment space, Placement & Visualization	Macro-economics, Demographics & External Analysis	Operational Performance	Customer & Product Profitability
Customer Communicaiton	Price-elasticity & Price Strategy	New Store Effects Analysis	Market Trends	Inventory Availability	Working Capital Analysis
Customer Segmentation	Pricing Simulation	Work Force Effectiveness	Category / Product Trends	Logistics Network Visbility	Cash Flow Analysis
Customer Management	Ad Effectiveness	Cash Movement Analysis		Supplier Performance	In-/tangible Asset, Liability & Qualitative Analysis
	New Products & Innovations	Shrinkage Analysis		Sharing Informaiton	Actual vs. Budget vs. Forecast Analysis
				Collaborative Activities & Services	Budgeting

ADVICE

- Embrace – Fast Fail, Early Fail
- Think “Iterative and Continuous”
- Get creative with sourcing data
- Think Relative and Trending – Not Absolutes
- You will need to “cobble” and “tinker”
- Simple and efficient beats out complex and complete
- Collaboration and exec sponsorship is essential
- Don’t chase vanity, chase actionable.