Swimming Lessons for the Data Lake

Becoming a Marketing Analyst for the Next Ten Years!
Today
Agenda

• The rising waters of change
• Swimming lessons for the Lake
• The Horizon Beyond the Lake
• Three “New” Values to keep afloat
  ❑ Ethics
  ❑ Agility
  ❑ Diversity
The Rising Waters of Change

Analytic Change is Here
The Rising Waters of Change

The Future is here

- BIG Unstructured data
- Data Lakes
- Open Source
- Machine Learning
- Artificial Intelligence
- IoT
- Blockchain
- Bitcoin
The Swimming Pool where we learned to swim
AKA The Relational Database

The relational database has:
• A defined structure of tables
• Optimization to ensure efficient queries
• Clean data
• Governance processes
• Established ETL
• Proven tools to access the data
• A variety ways analysts of differing skill levels can access information
A New Place to Try to Swim
AKA The Data Lake

The data lake can:
• Contain significantly more data
• Be unstructured – both the table structure and the data itself tend to be more free form
• Contain unclean data
• Have few established governance processes
• Be accessed through many different tools
• Be difficult to navigate
• Be real-time and responsive to change
So, why might we want to swim here?
✓ Adventure
✓ Discovery
✓ Challenge
✓ Closer connection to nature
What does this mean for the Analyst?
What this means for the analyst

More Data: More Findings

Big data can mean data that is:

✓ Closer to real time
✓ More comprehensive
✓ More detailed

➔ Better answers
What this means for the analyst
Cleaning Your Own Water

Source data  ETL, Warehouse  Analytic Interface
What this means for the analyst

Cleaning Your Own Water

Data governance in the Data Lake is “on the way out” rather than “on the way in”

Analysts need to be data savvy; responsible for cleaning their data and making the right connections between data sets.

The Analyst cannot rely on the curation of others.

This may require some “heavy lifting” which you must document so it can be repeated.
What this means for the analyst

Longer running queries

- The Data Lake is rarely optimized for analytics; it is intended as a repository to efficiently store vast amounts of data, not to process it.
- It can be difficult to join tables efficiently.
- Our lazy coding ways may be exposed!
- Queries of a few rows can take as long as millions of rows.
What this means for the analyst

More Programming Skills

• To access the lake, analysts require extensive programming skills

• Technical know-how on how to more efficiently generate queries

• Extensive coding to aggregate and transform data
What this means for the analyst

Data Wrangler

“... because of the diversity of data, you spend a lot of your time being a data janitor, before you can get to the cool, sexy things that got you into the field in the first place.”

Matt Mohebbi
Iodine co-founder and data scientist

“We really need better tools so we can spend less time on data wrangling and get to the sexy stuff.”

Michael Cavaretta
Ford Motor Company data scientist
What this means for the analyst

New Tools

• New Species: From Pigs to Pythons, Elephants to Bee Hives..... Mother nature is putting us to the test!

• Adaptation: It is requiring adaptation of existing tools to respond to the change.
Its Back to School for the Analyst
Again and Again
New is not always better
Be Wary of Every Bright Shiny Thing
The Horizon Beyond the Lake
Adapting to Change
Beyond the Lake

Floods, hurricanes and droughts require humans to adapt our behaviour in the wake of Global Warming

→ The trends in data and analytics require analysts to do the same
Three Pillars of “Data Science”
Different Analytic Approaches

Understand

• Develop an understanding of the data and the business
• Reporting, query

Improve

• Enhance the business through prediction and algorithms
• Predictive models, recommendation engines

Create

• Build entirely new products or features using the data.
The Pyramid of “Data Science”

Distribution of Analysts

Create

Improve

Understand

Fewer people

Many people
What does the future hold?

Two Tsunamis at work

DEMAND AND DATA GROWS

• Massive growth in data availability

• Predicted shortage in data science skills

• IBM Predicts Demand For Data Scientists Will Soar 28% By 2020

AUTOMATION AND MACHINE LEARNING GROWS

• Concept is not new; but data is what makes it powerful.

• With enough cases to learn from - inference is great.

• Gartner predicts 40% of all data science tasks will be automated by 2020.
Introducing the Citizen Data Scientist
Filling the Skills Gap through Automation

Gartner defines a Citizen Data Scientist as a “person who creates or generates models that use advanced diagnostic analytics or predictive and prescriptive capabilities, but whose primary job function is outside the field of statistics and analytics.”

Gartner also predicts that citizen data scientists will surpass data scientists in the amount of advanced analysis produced by 2019.
You’re not that special anymore!

Automation ‘can’ replace you

• Predictive Analytics experience has always been scarce and, therefore, valuable.

• Experience is still important but it is no longer so imperative

• Automation will allow you test multiple options, streamline the process and rely less on past experience and gut.

➔ Will you really be replaced?
You’re not that special anymore!

But you can still be critical

Experience may not be *as* valuable but it is still important

AND other skills are more critical than ever:

- Domain expertise and business understanding
- The ability to define a problem
- Understanding the data – where it is, what it is, how it is collected, how it is structured.
- Interpretation and storytelling
- Implementation of results
Learn to Tell Stories
Create Business Value NOT Just Models

Interpretation, Communication, Storytelling Skills, Business Knowledge

Advanced Programming, Development
ByCatch

Ethical Implications
A Breach of Trust
The Ethics of Data Use

A privacy breach..... Or an ethical one?
Agility is key
The End of the Waterfall Methodology

Agile analytics
• Iterative and Incremental
• Minimum viable product
• Deliver in stages (sprints)
• Regular and early buy-in from stakeholders
Toronto, the most ethnically diverse city in the world.

1,162,635 people reported themselves as being part of a visible minority.

Source: Census 2006, Canada.
FOR TORONTO.....

DIVERSITY
OUR
STRENGTH

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.
And Diversity is key to Adapting to Change

Although change may initially push us to be more singular

“It is now widely recognized that climate change and biodiversity are interconnected. Biodiversity is affected by climate change, with negative consequences for human well-being, but biodiversity, through the ecosystem services it supports, also makes an important contribution to both climate-change mitigation and adaptation.”

UN Convention on Biological Diversity
You cannot watch from the dock
You are going to need to get wet!
Presenter
Emma Warrillow, President, Data Insight Group Inc. (DiG)

A marketing strategist with a talent for numbers, Emma Warrillow uses analytics as the foundation for customer-centric marketing strategies. Emma has worked in both the corporate and consulting side of marketing analytics for more than two decades.

Through her company, DiG (Data Insight Group Inc.), Emma and her team of senior database and analytics professionals help companies understand their customer data and make sense of what it is saying about their customers.

With a Masters degree in Management Sciences (University of Waterloo) and an undergraduate degree in Mathematics and Statistics (Queen’s University), Emma is uniquely qualified to understand the analytics and business of customer relationship strategies.

She is a recognized expert in her field and is a highly sought-after speaker, judge, panelist, and frequent contributor to industry publications.

Emma makes data make sense.

https://twitter.com/DataInsightGrp
Your feedback counts!

Don't forget to complete the session survey in your conference mobile app.

1. Go to the Agenda icon in the conference app.
2. Find this session title and select it.
3. On the sessions page, scroll down to Surveys and select the name of the survey.
4. Complete the survey and click Finish.