CANADIAN DATA SCIENCE WORKSHOP 2018

# Data Science Needs and Stress Points

**BRIAN KENG | @bjlkeng**

**Chief Data Scientist**
Rubikloud Technologies
brian.keng@rubikloud.com

**Adjunct Professor, Data Science**
Rotman School of Management
University of Toronto

rubikloud®

## THESIS

Rubikloud uses AI to deliver **intelligent decision automation** to enterprise retailers through its cloud-native platform.

### CLOUD NATIVE

### ARTIFICIAL INTELLIGENCE

### RETAIL SPECIFIC

## KEY FACTS

**2013**

Founded

**86**

Employees (and growing)

Including 25 in
Data Science fields

**$45M**

Total investment from:
Horizons Ventures, Access Industries, MaRS IAF & Private Angels

**Multinational**

Clients in North America,
Europe & East Asia.

Offices in Hong Kong,
London & Netherlands

**5**

Patents: 3 Filed

2 Provisional

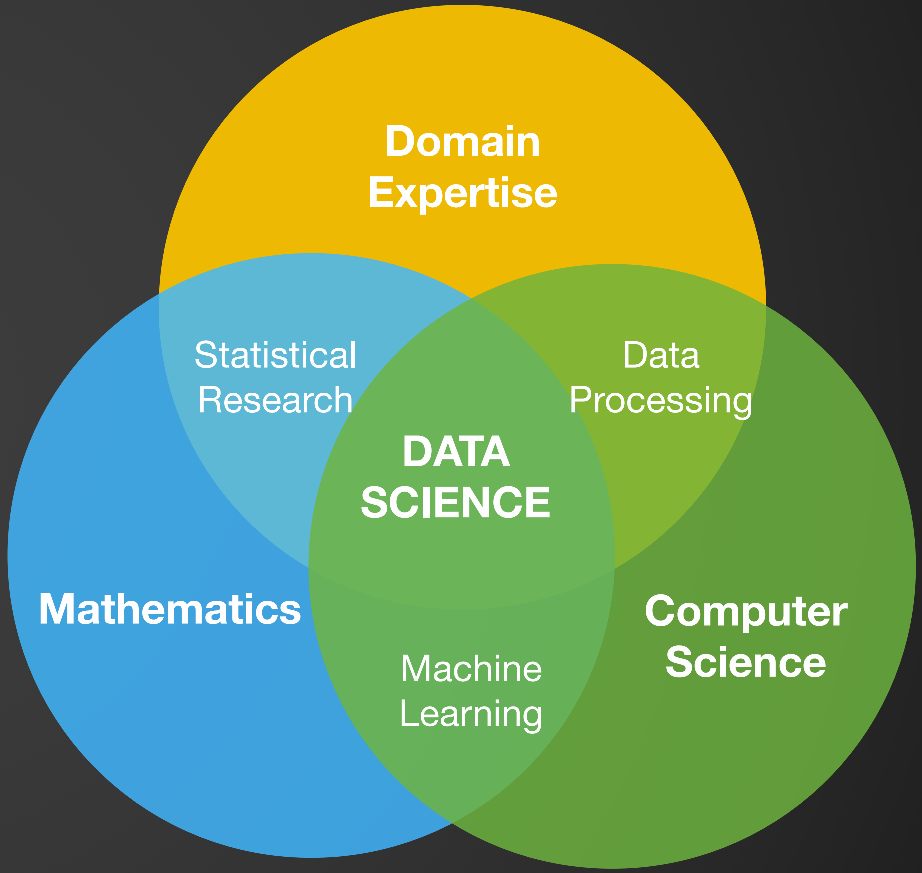
# What is Data Science, Machine Learning, Data Mining, and Statistics?

**Machine Learning**
is a set of methods that can:

- Automatically detect patterns in data
- Use uncovered patterns to predict future data
- Perform other kinds of decision making under uncertainty

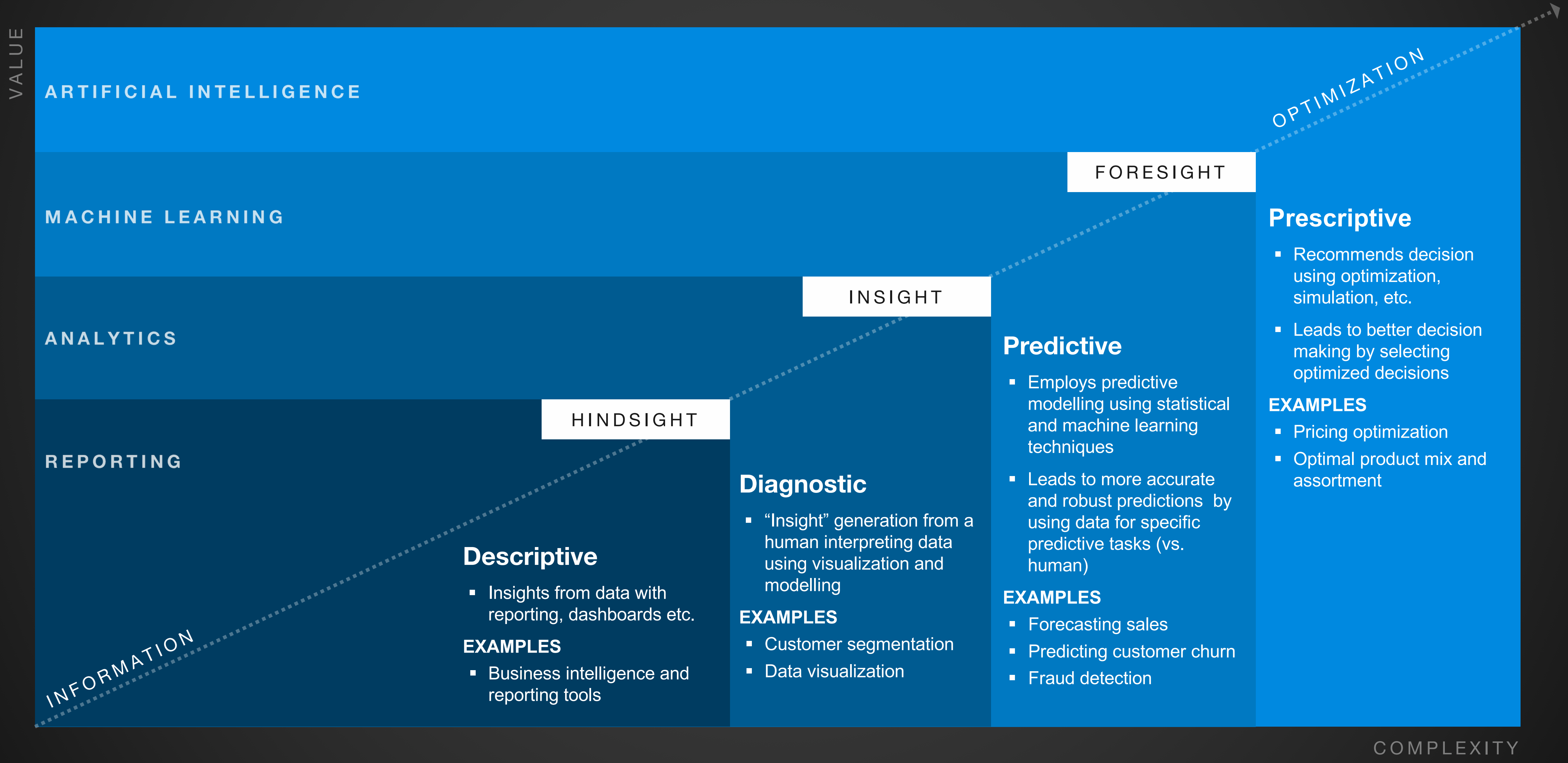
**Data Mining**
is the process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

**Statistics**
is the study of the collection, analysis, interpretation, presentation, and organization of data.



*Source: Palmer, Shelly, Data Science for the C-Suite. New York: Digital Living Press, 2015. Print*

# Artificial Intelligence as *Intelligent Decision Automation*



### Descriptive

- Insights from data with reporting, dashboards etc.

**EXAMPLES**

- Business intelligence and reporting tools

### Diagnostic

- “Insight” generation from a human interpreting data using visualization and modelling

**EXAMPLES**

- Customer segmentation
- Data visualization

### Predictive

- Employs predictive modelling using statistical and machine learning techniques
- Leads to more accurate and robust predictions by using data for specific predictive tasks (vs. human)

**EXAMPLES**

- Forecasting sales
- Predicting customer churn
- Fraud detection

### Prescriptive

- Recommends decision using optimization, simulation, etc.
- Leads to better decision making by selecting optimized decisions

**EXAMPLES**

- Pricing optimization
- Optimal product mix and assortment

# Data Science Related Roles

## DATA SCIENTISTS

**300+ applicants in 2018**
**50+ interviews, 1 hire**

- Applied Modelling (Machine Learning, Statistics)
- Software Development (Python)
- Data Technology (Spark, Hadoop, AWS/Azure/GCP, SQL)

## MACHINE LEARNING ENGINEER

**400+ applicants in 2018**
**40+ interviews, 1 hire**

- Software Development (Python)
- Data Technology (Spark, Hadoop, AWS/Azure/GCP, SQL)
- Applied Modelling (Machine Learning, Statistics)

## DATA ANALYST

**1000+ applicants in 2018**
**30+ interviews, 0 hires**

- Domain Knowledge (retail, CRM/loyalty, promotion, digital)
- Data Analysis, Presentation & Visualization
- Data Technology (SQL, Excel, Powerpoint, Python)

## RESEARCH SCIENTIST

**100+ applicants in 2018**
**7 interviews, 0 hires**

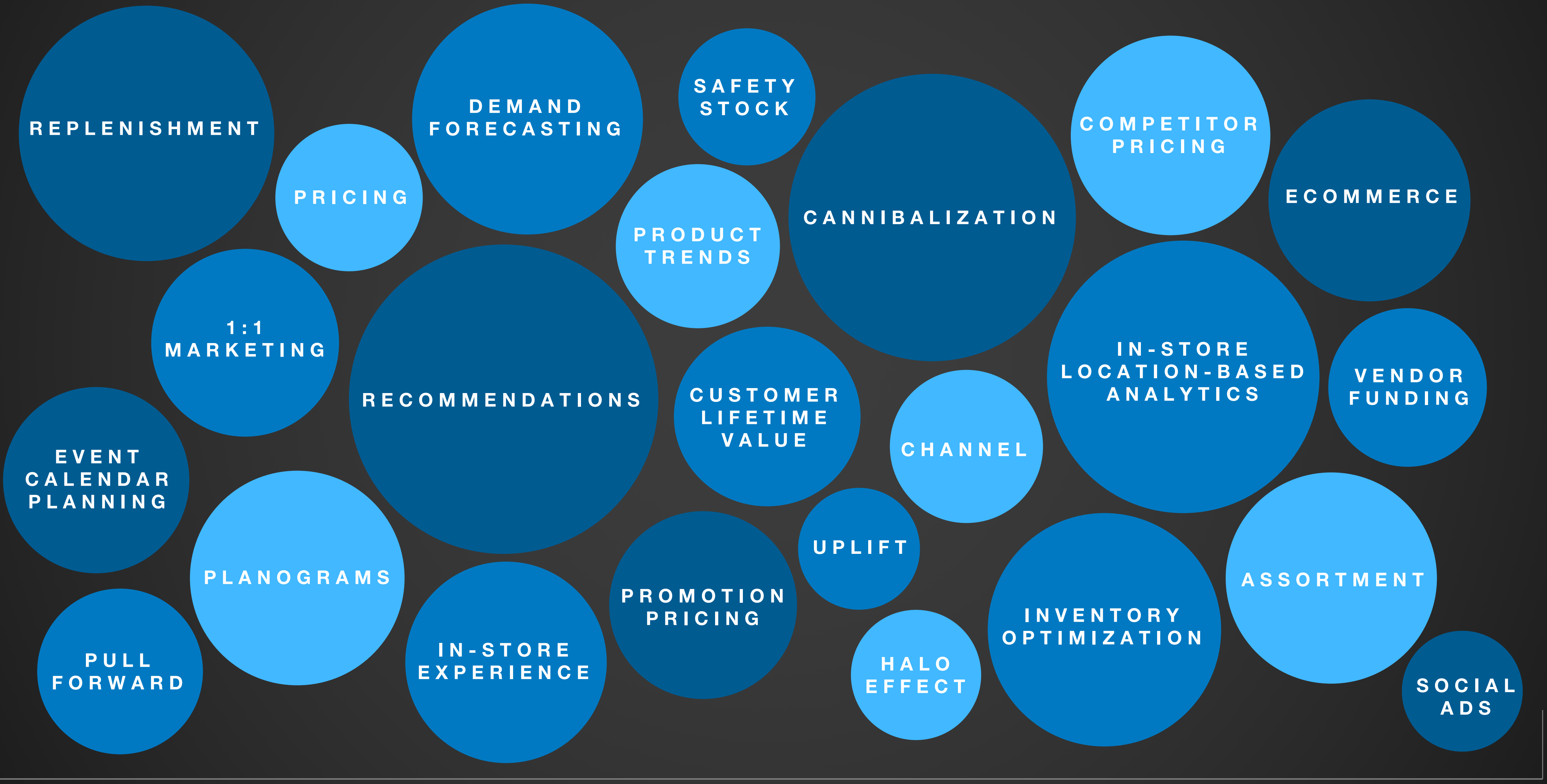
- Research (Machine Learning, Statistics, Operations Research)

# Retail: Simple right?

CUSTOMER

PRODUCTS

# Retail: A big hairy mess



REPLENISHMENT

PRICING

DEMAND FORECASTING

SAFETY STOCK

PRODUCT TRENDS

CANNIBALIZATION

COMPETITOR PRICING

ECOMMERCE

1:1 MARKETING

RECOMMENDATIONS

CUSTOMER LIFETIME VALUE

CHANNEL

IN-STORE LOCATION-BASED ANALYTICS

VENDOR FUNDING

EVENT CALENDAR PLANNING

PLANOGRAMS

PROMOTION PRICING

UPLIFT

ASSORTMENT

IN-STORE EXPERIENCE

HALO EFFECT

INVENTORY OPTIMIZATION

PULL FORWARD

SOCIAL ADS

RETAIL

# Data Science Retail Research Problems

## Customer Representations

- Multifaceted
- Time-series nature of data
- Complexity of purchased items (brand, price, category etc.
- Randomness in customer behavior
- Limited observability

## Data Sparsity

- Lots of data (100M Customers, 100K products, 10B Transactions, 1K stores)
- Predictions occur at granular level (customer+product + store)
- New customers, products, promotions are constantly being introduced

## Product Ontologies

- Products data is messy, incomplete, silo’ed
- Need for semantic representations (e.g. Google Knowledge Graph)
- Combine multiple incomplete data sources (text, images, proprietary databases)

## Novel Data Sources

- Social data (e.g. influencers, trends)
- IoT sensors (e.g. shelves, store traffic, warehouses etc.)

## Reinforcement Learning

- Customer Relationship Management (CRM) Loyalty Programs
- “Blackbox” optimization techniques for many complex feedback systems

# Academic Partnerships

### DATA SCIENCE RESEARCH INTERNSHIP

- Master's/PhD students who do academic research in strategic areas for the company
- Collaboration with University of Toronto and McGill University (3 students)
- Jointly-funded via Mitacs/NSERC Engage Grants
- Students work 3 days a week at company with access to data, compute, mentorship, etc.

### ACADEMIC-INDUSTRY PARTNERSHIP WITH UNIVERSITY OF GUELPH (PROFESSOR GRAHAM TAYLOR)[1]

- Rubikloud provided platform (RubiOne), data, retail problem for final project of Intro. to ML
- Students competed in a "Kaggle-like" competition on anonymized retail dataset
- Used as an example for academic-industry partnership in the Vector Institute 1000 AIMs program (Appendix D)

### DATA SCIENCE AND MACHINE LEARNING INTERNSHIPS

- Interns from UW Co-op program (4-month, undergraduate), UofT PEY (16 month, undergraduate), UofT MScAC (8 month, Master's)
- Learn fundamental skills in software development, data manipulation/frameworks, and real issues surrounding deployment of models in the real world

### WORKSHOP ON ICDM BIG DATA AND DATA SCIENCE IN RETAIL (ICDM 2017) [2]

- Organized a workshop on big data and data science
- Papers focused on retail specific problems

[1] https://rubikloud.com/lab/lstm-rfm-lmfao-making-sense-data-science-acronyms-deep-dive/

[2] https://rubikloud.com/Retail-Science-Workshop/

# Data Availability for Research

## Available Data Sources

- **Retail Clients**: Transactions, Customers, Products, Inventory, Promotions, Store, Margin etc.
- **1st Party**: Rubikloud Tool data (retailer input-ed data)
- **3rd Party**: Industry surveys, market data, competitive pricing, demographics, crawled etc.
- Varying levels of cleanliness: Missing fields, semantic consistency, joinable fields etc.

## Security

- No PII data but sensitive
- Must be on Rubikloud infrastructure
- NDA for external parties even for anonymized data

## Accessibility

- On-Premise
- RubiOne (data science IDE: Jupyter/cloud based)
- Cost can be a factor depending on compute (cloud infrastructure)

Brian Keng
brian.keng@rubikloud.com
@bjlkeng

# THANK YOU