# From Stored Data to Data Stories: Building Data Narratives with Open Source Tools

Paul J. Kowalczyk

2019-10-29

**From Stored Data to Data Stories:**
**Building Data Narratives with Open Source Tools**

ODSC West 2019
9 AM Wednesday, October 30

Paul J Kowalczyk, PhD
Senior Data Scientist
Solvay / Bristol Research & Innovation Centre
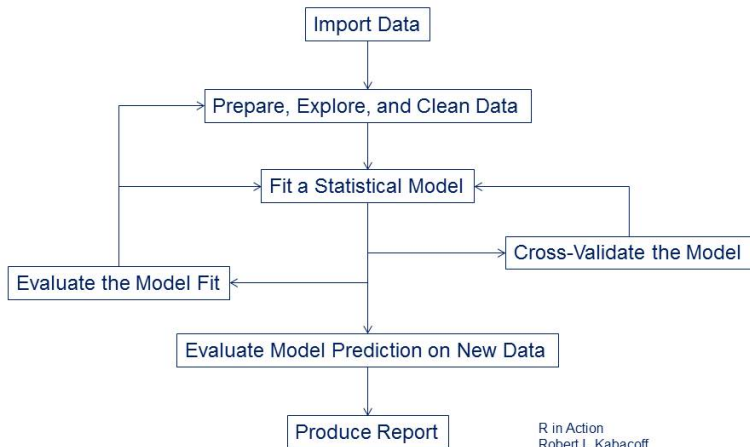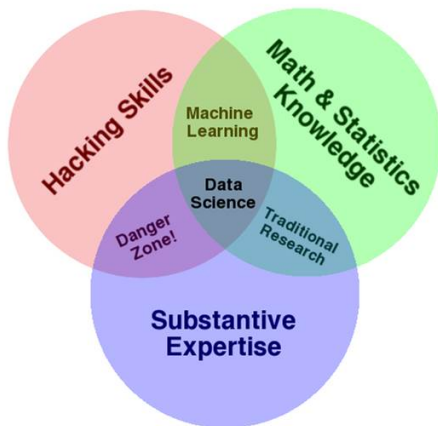paul.kowalczyk@solvay.com

**Code & Data**:
https://github.com/pjkowalczyk/StoredDataStories

# Outline

| Time | Activity |
| --- | --- |
| 9:00-9:15 | Introduction |
| 9:15-9:45 | Ground Truth |
| 9:45-10:00 | Practicum I |
| 10:00-10:30 | Notebooks |
| 10:30-10-45 | Practicum II |
| 10:45-11:15 | Presentations |
| 11:15-11:30 | Parcticum III |
| 11:30-12:00 | Documents |
| 12:00-12:15 | Practicum IV |
| 12:15-12:30 | Wrap-Up |

# Workflow

Import Data

Prepare, Explore, and Clean Data

Fit a Statistical Model

Cross-Validate the Model

Evaluate the Model Fit

Evaluate Model Prediction on New Data

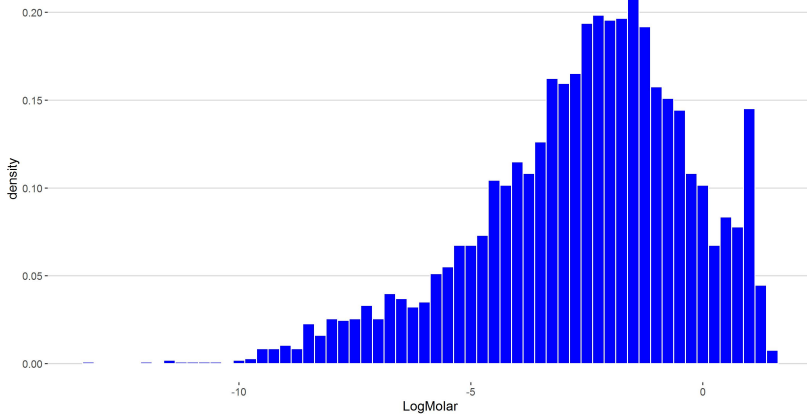Produce Report

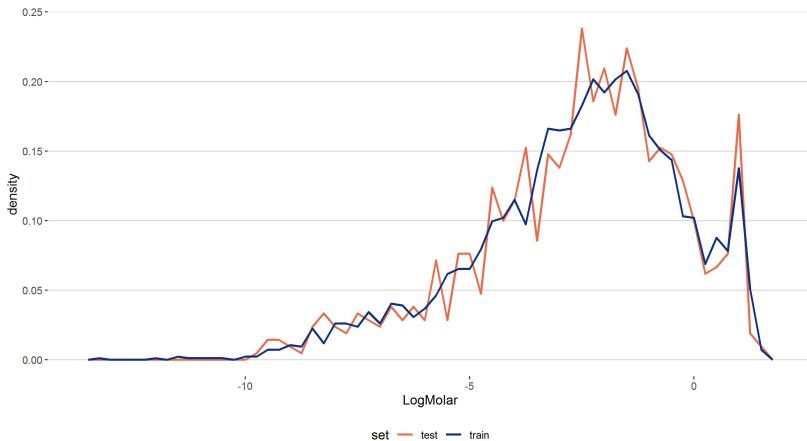# Drew Conway Data Science Venn Diagram

# The Data Science Venn Diagram for this Workshop
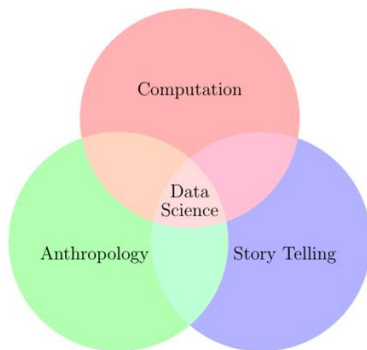
# Water Solubility

# Train / Test Split

# Full Correlations
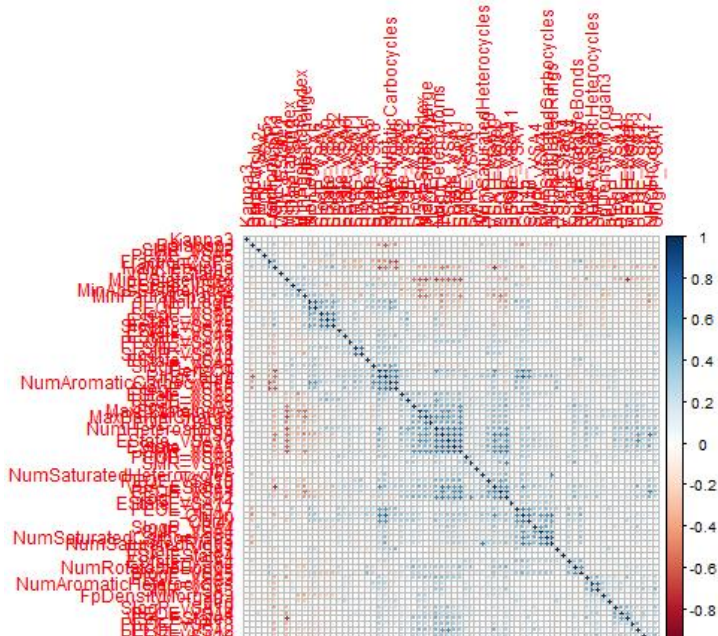
# Reduced Correlations

# Case Study

Build predictive models for

- **water solubility** (regression)
    - OECD Test No. 105
- **biodegradability** (classification)
    - OECD Test No. 301