

Utilizing Human Data Validation for KPI Analysis *and* Machine Learning

Dan Morris
Radius Intelligence



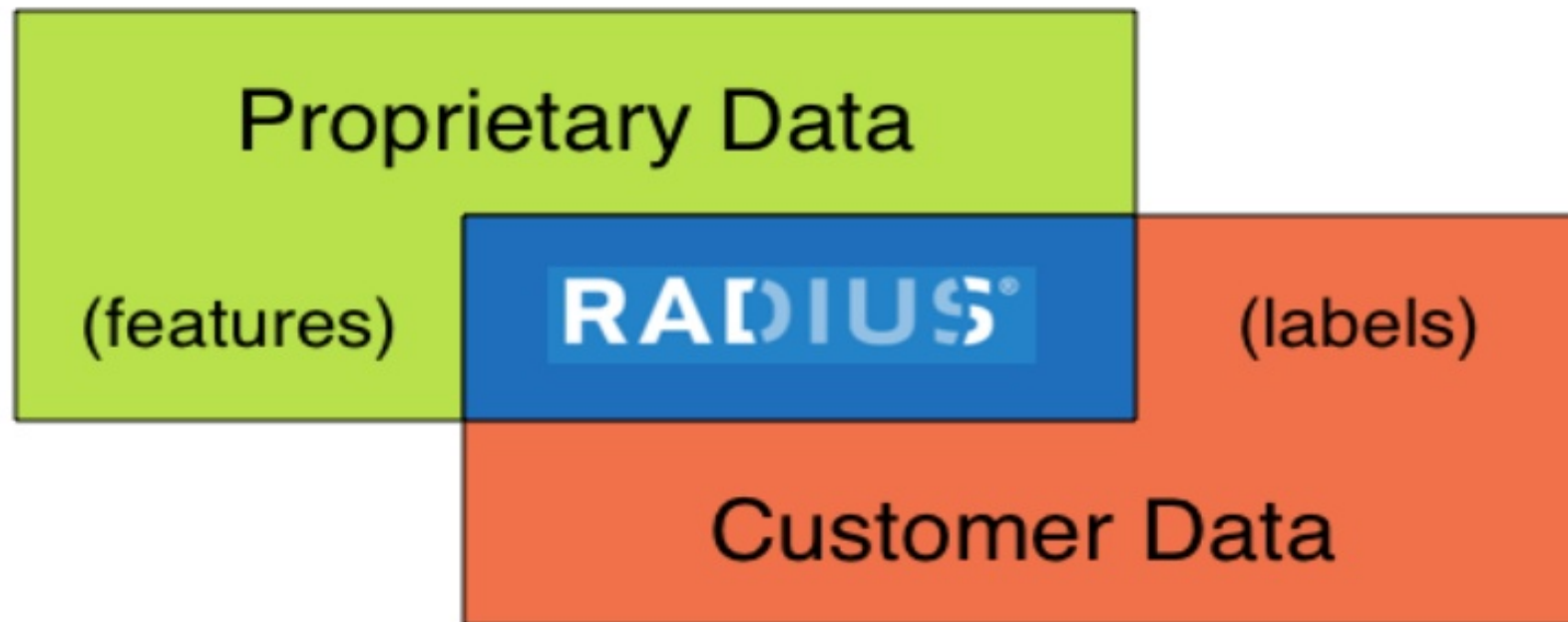
SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO

Overview and Key Takeaways

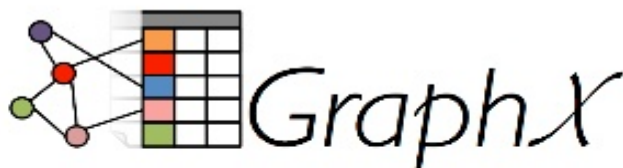
- Data science problems @ Radius
- Human validation: costs and benefits
- Sampling and experimentation for multiple consumers
- Positive feedback cycles in production



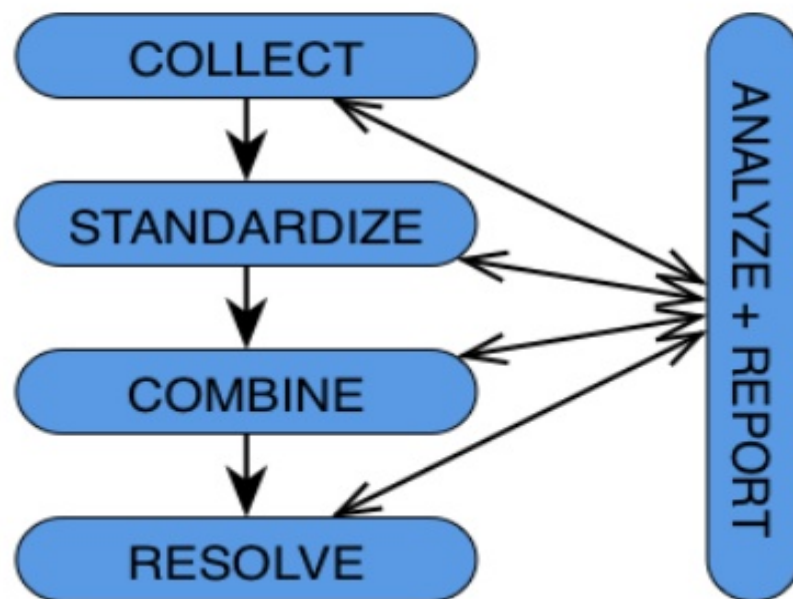
Radius - B2B Predictive Marketing



Radius - Data Engineering



MLlib



SPARK SUMMIT 2016

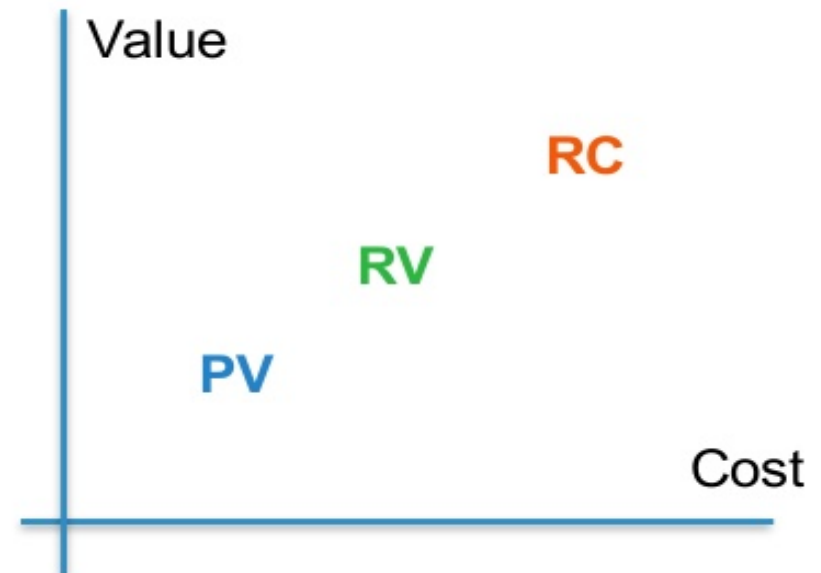
Why Human Validation?

- Business firmographic data is a difficult data problem
- Our sources face the same challenges that we do
- Each source must be considered a “proposal”
- Independent Human Validation is *(the closest thing to)* ground truth



Degrees of Human Validation

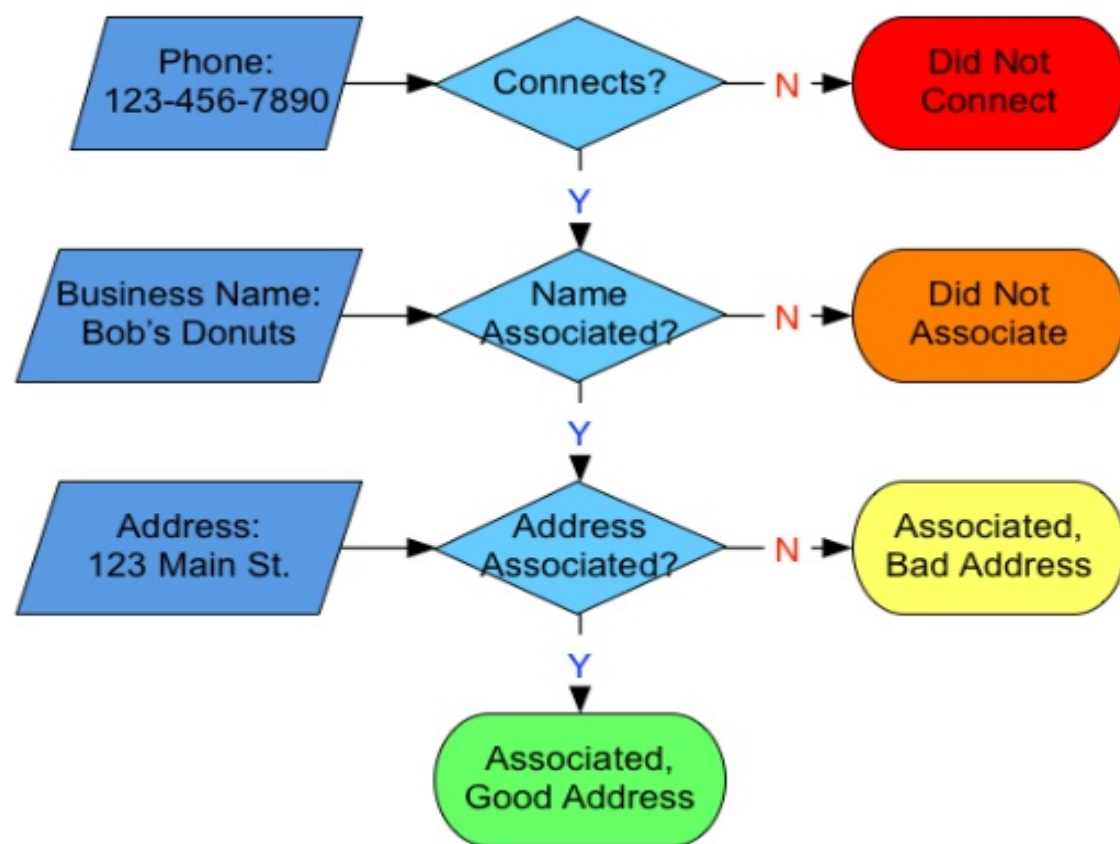
- Prompted Validation
- Research Validation
- Research Curation



Prompted Validation

Example Assignment:

Verify the Phone Number and
Address of a Business



Research Validation

Example Assignment:

**Determine if a Business
Belongs to a Chain / Franchise**

Business Name: **Bob's Donuts**

Address: **123 Main St.**

Website: www.bobsdonuts.biz

Industry: **Limited Service Restaurants**

Is Chain: **(Y / N / U)**

Chain Type: **(Local / Regional / National)**



Research Curation

Example Assignment:

**Where is the Headquarters of
this Company Located?**

Company Name: **Bob's Donuts Inc.**

Website: www.bobsdonuts.biz

Has many locations: **(Y / N / U)**

HQ Location: **???**

Source of information: **???**



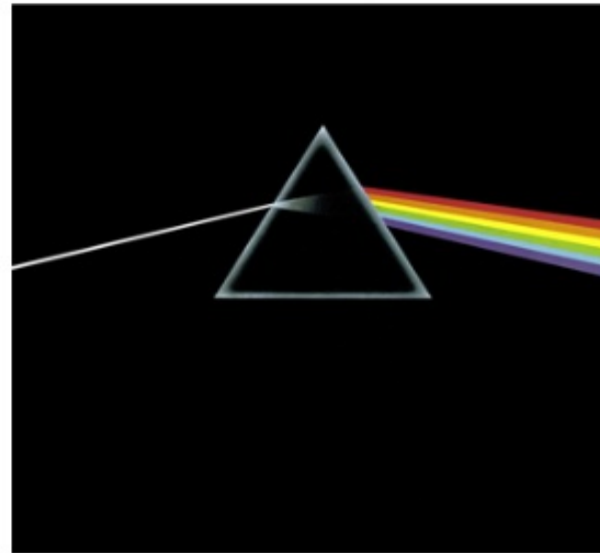
Human Validation: Benefits

- Ground Truth
 - Supervised ML
 - Internal Metrics
 - Competitive Analysis
- Our customers are humans, too!



Human Validation: Costs

- Money
- Time
- Us and Them



Cost: Money

- *Validated* data costs more than *aggregated* data



Validation + Data Science

>



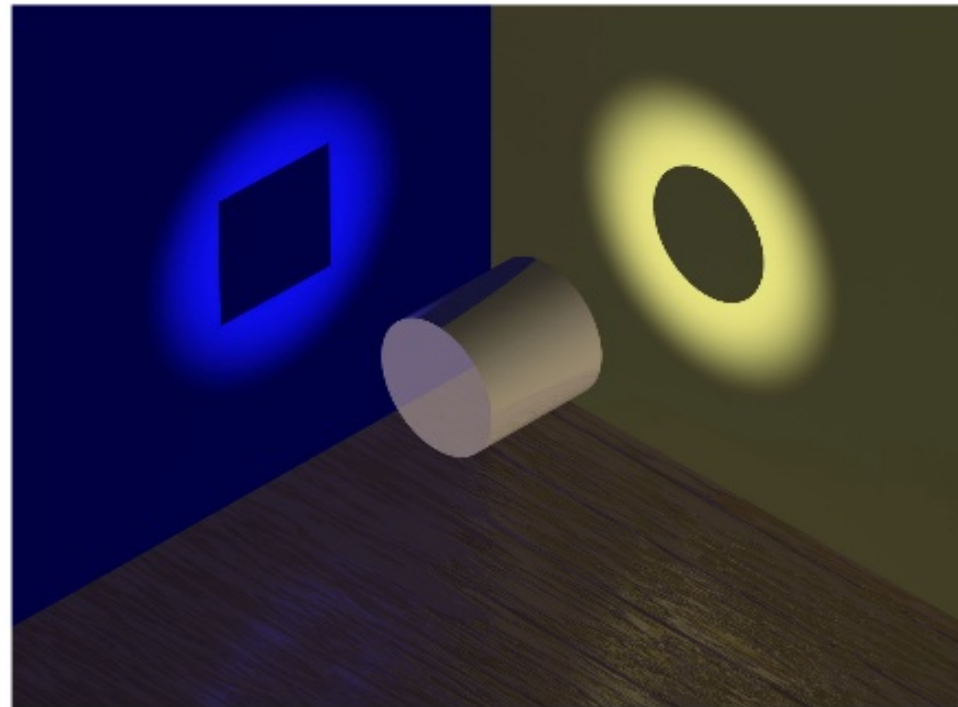
Pure Validation

Cost: Time

- Automated experimental framework
 - Shift bottleneck to validation teams
- Parallelized validation improves turnaround time
 - Be mindful of differences in teams / validators
- Decay / Obsolescence of validations



Cost: Us and Them



Clearly communicate expectations and interpretations



SPARK SUMMIT 2016

Uses for Validated Data

- KPI Analysis
- ML Training Sets
- Spot Hypothesis Validation



Challenge: minimize number of validations while meeting all downstream needs



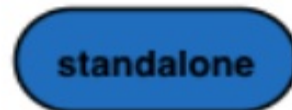
SPARK SUMMIT 2016

Multiple-Consumer Sampling

Standalone vs. Chain Experiment
1 value per 1 location == Easy Sampling!



Business Locations



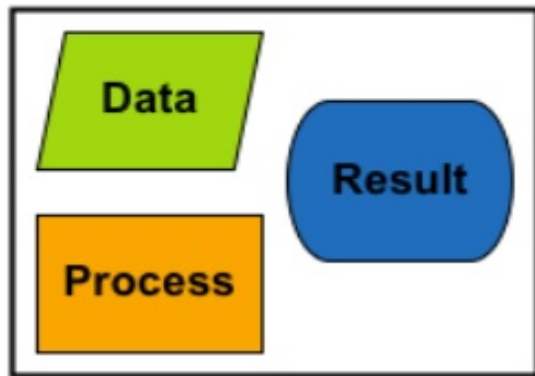
Multiple-Consumer Sampling

Phone Accuracy Experiment

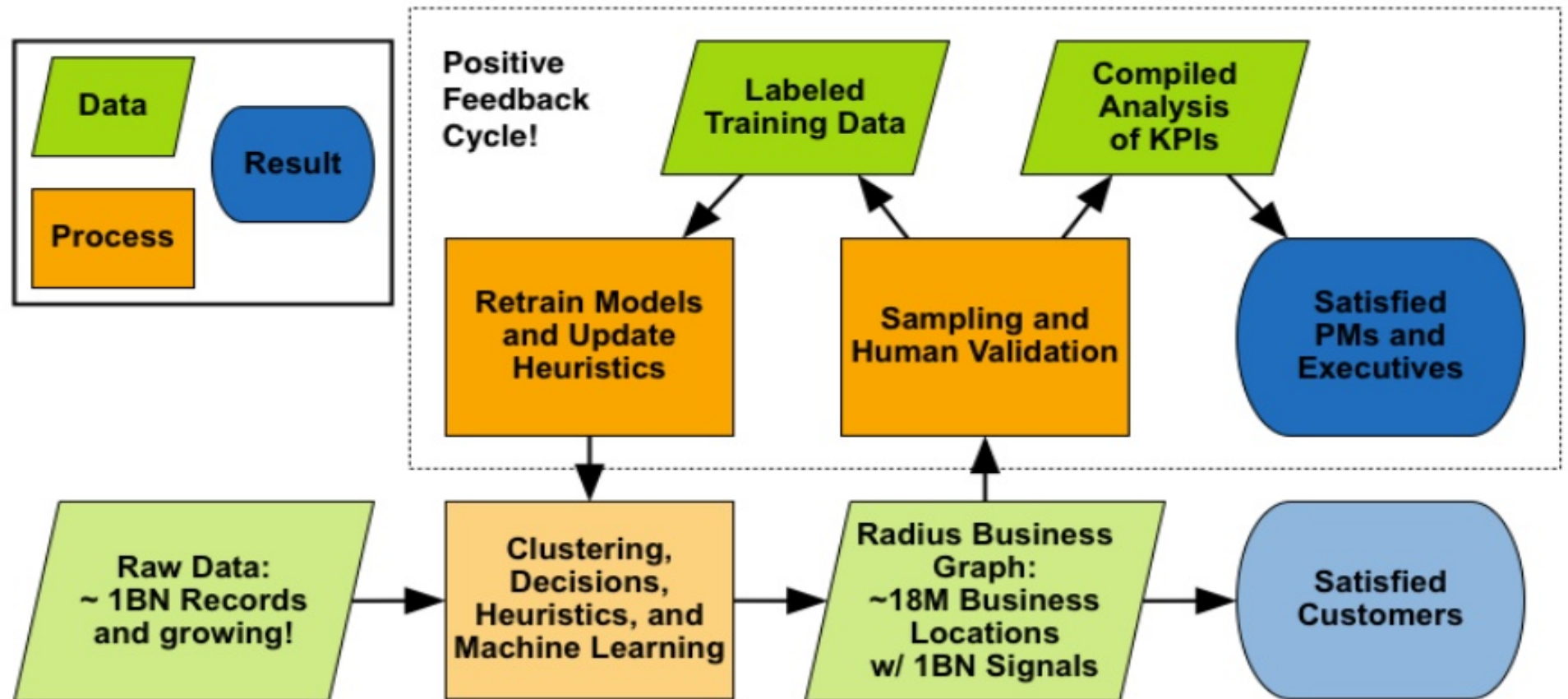
(0, 1, 2, 3, ...) values to 1 location == Difficult Sampling.



Basic Production Pipeline



Positive Feedback Production Cycle



THANKS!

email me: dan.morris@radius.com

stalk me: [@djsensei](#)

connect me: [linkedin.com/in/danielepmorris](https://www.linkedin.com/in/danielepmorris)

work with me: radius.com/jobs



SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO