Data Quality and Customer Behavioral Modeling

Daniel Krasner

Chief Data Scientist, Sailthru

Co-founder, KFit Solutions

Wolfram Data Summit 2012

Washington, DC

Agenda

- A bit of background
- What is Data Science
- Why is Data Quality Important
- A few examples
- Contact Info

A Bit of background

- Education: pure mathematics
  - PhD Columbia University
- Academics posts: MSRI, UCLA
- In search of other opportunities
- Startups, data science, NYC tech
- Steep learning curve
- Data Science at Sailthru
- KFit Solutions (with Matt DeLand)

What Is Data Science

- Data Science strives to answer two questions:
  - What has happened (and why)?
    - Reporting
    - Data visualization
    - Statistical inference
    - Causality
  - What will happen?
    - Predictive Behavioral Modeling
    - Trend/Anomaly/Pattern Detection
    - Recommendation Systems

Data Quality

- Data dictates the model
  - Good data beats good modeling
  - Data science is inherently experimental
- You can log a lot but not everything
  - Volume and quality are not synonymous
- “Chicken and the egg” of data collection
- Granularity
- What are you trying to build
- It’s not just the data

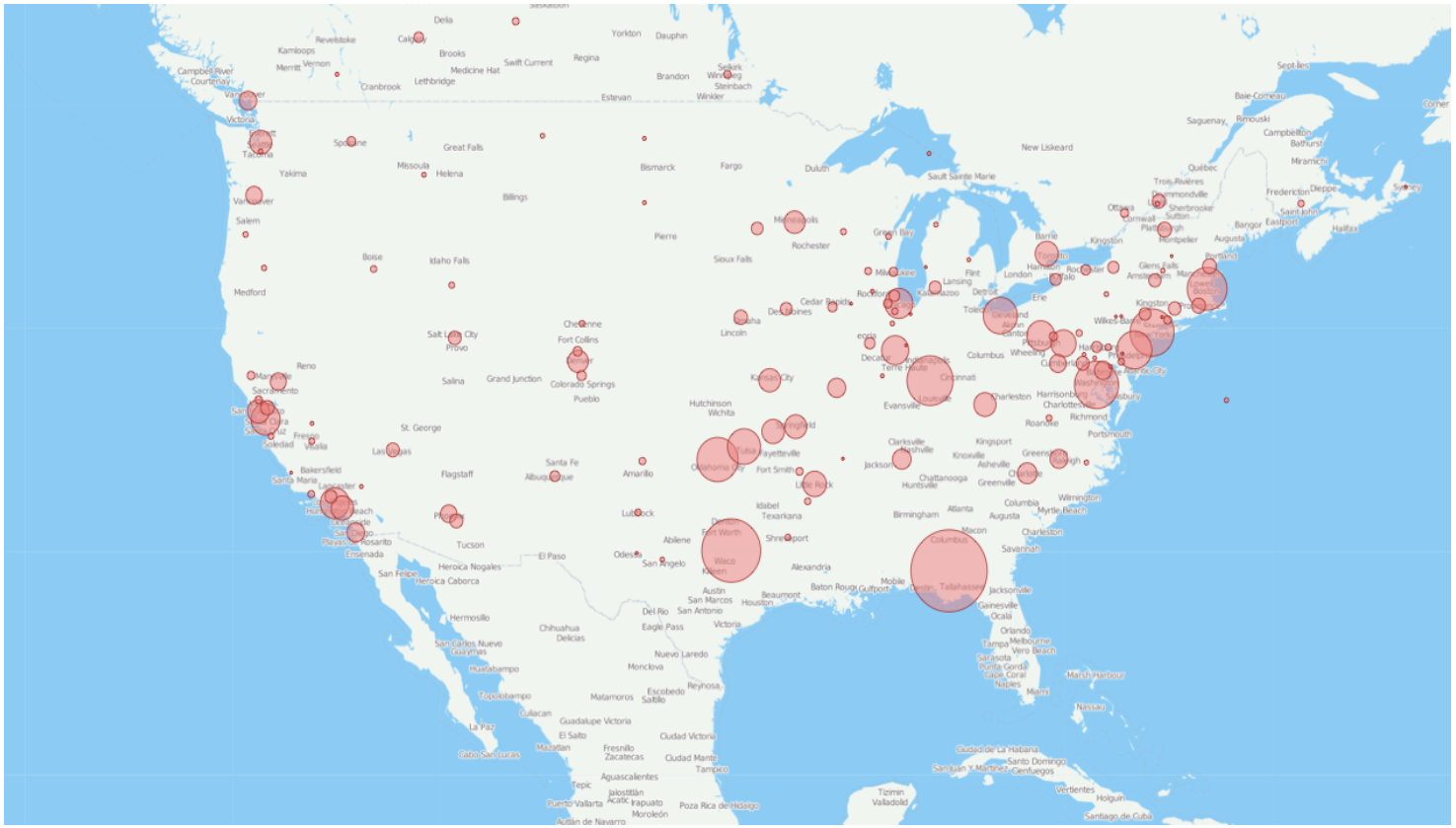
A few examples

- Sailthru
  - Behavioral email and ad platform (SaaS)
  - About 200 e-commerce/e-publishing clients
  - Collects huge amount of user action data
    - Every user has a profile
    - Opens, clicks, purchases, page-views, geo, devices, on site action
    - Most everything is time-stamped
    - NoSQl backend (one of the first Mongo DB in prod.)
  - Key interests:
    - Provide actionable tools to clients

A Few Examples

– Insight into user behavior, consumption

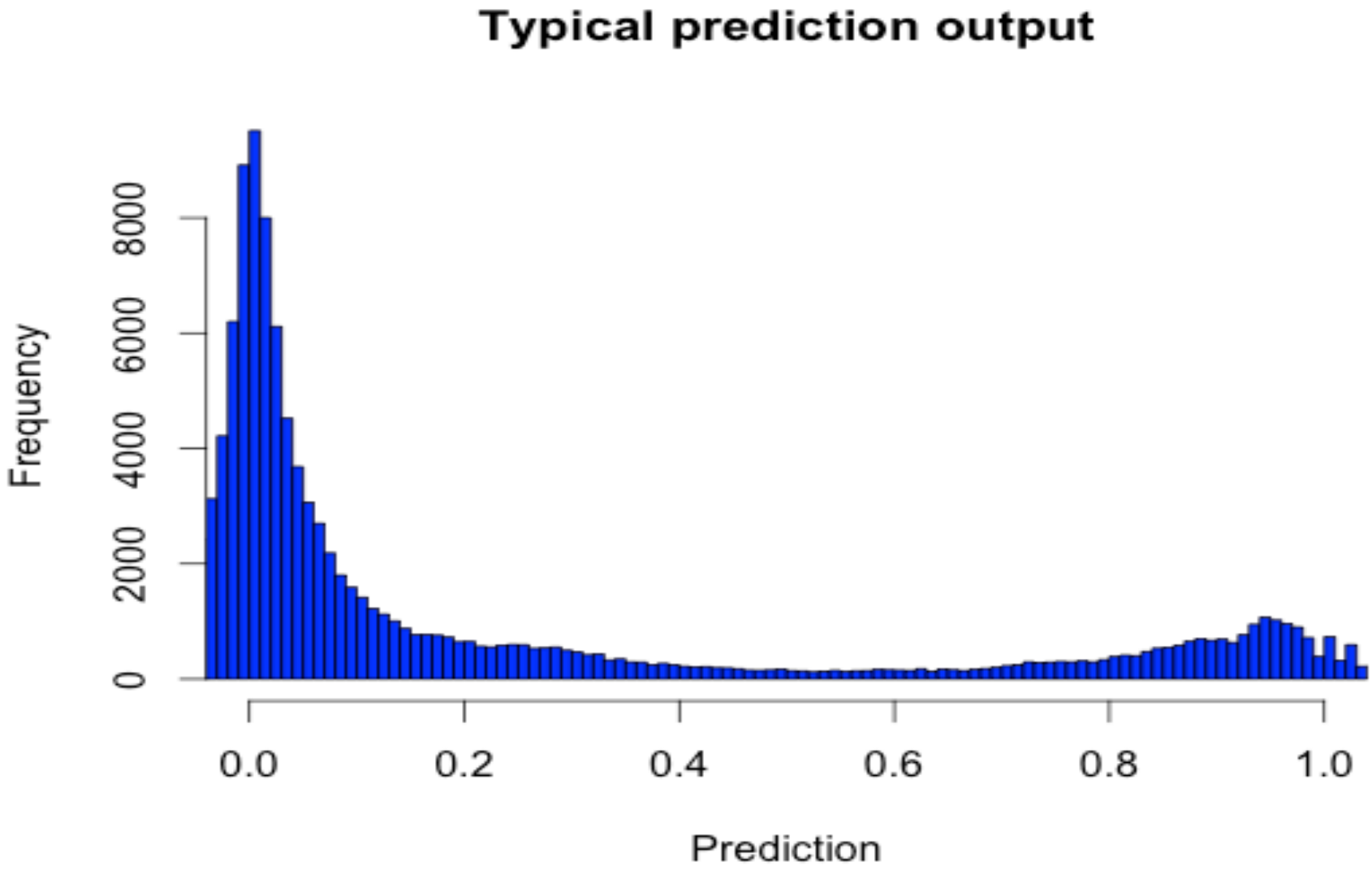
- Strives to make every bit of reporting actionable



A Few Examples

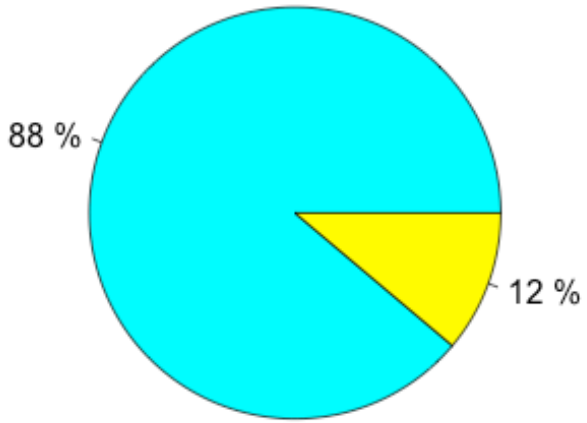
- Predicting behavior
  - Engagement, purchases, page views, CLV
- Engagement Model
  - Probability that a user will open in the next week
    - Every client's model is retrained on a weekly basis
    - Not "real-time" prediction
    - Monitoring system in place
    - Results piped into the dashboard
    - Combination of python, R, bash
    - Quiet accurate

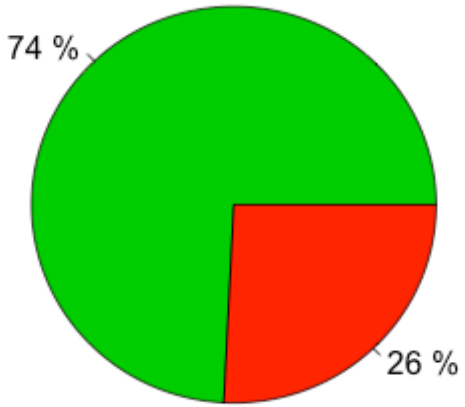
Engagement Model



Engagement Model

Typical Engaged-Disengaged Prediction Accuracy





A number of clients have > 85% accuracy on both true-positives and true-negatives

News Media

- Recent News client
  - Very little individual user data
  - Small percentage of users cookie id'ed
  - Hard to create an ad recommendation system
  - Large amount of news dissemination data
    - Articles, updates, impressions, geo, dozens of sources and domains
  - Big potential in modeling and understanding news trend, topics consumption patterns

Contact

- Sailthru:
  - https://www.sailthru.com/
  - daniel@sailthru.com
- Kfit Solutions:
  - http://kfitsolutions.com/
  - daniel@kfitsolutions.com

Thank you!