

# Scenario Description

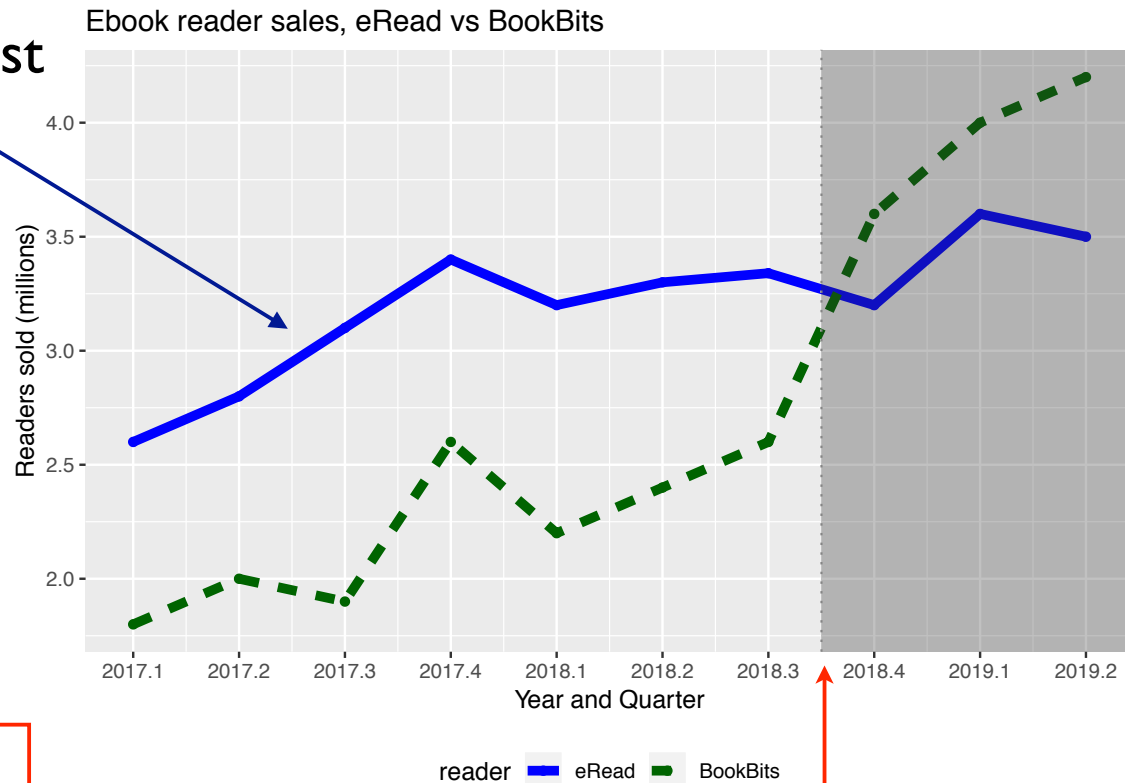
- WVCorp: the company you (the data scientist) work for
  - WVCorp has user forums and discussion boards for each of their products, where customers can discuss issues and features.
  - “Buzz”: when a topic on the user forum has a very high activity level -- considered an indication of user interest in that topic.
- eRead: WVCorp’s ebook reader product
- TimeWrangler: WVCorp’s time-management app
- BookBits: A competitor’s ebook reader product
- GCal: a third-party cloud-based calendar infrastructure that TimeWrangler can integrate with

# Predicting Buzz

WVCorp Data Science Team  
Notional Project Sponsor Presentation

# A Lost Opportunity

**eRead:** Formerly best selling indie ebook reader



Estimated \$25M  
lost revenue on  
product sales

**BookBits v.2  
with shared bookshelves  
introduced**

# Could we Have Caught This?

- eRead Forum discussions:
  - Sharing a booklist with a friend, to grab from as they pleased
  - Sharing a book with a group of friends (first-come-first-serve)
- Whenever these questions arose, the discussion was lively
  - Suggestions, work-arounds, kludges, “me too”s
  - A shared bookshelf (like BookBits) would have met these recurring needs
- There was **Buzz** around this issue! But we ignored it. Or didn't find it.
  - Labor intensive to continually keep up with forum activity

# Goal: Catch it Early

- Predict which topics on our product forums will have persistent buzz
  - Features customers want
  - Existing features users have trouble with
- Persistent buzz: real, ongoing customer need
  - not ephemeral or trendy issues

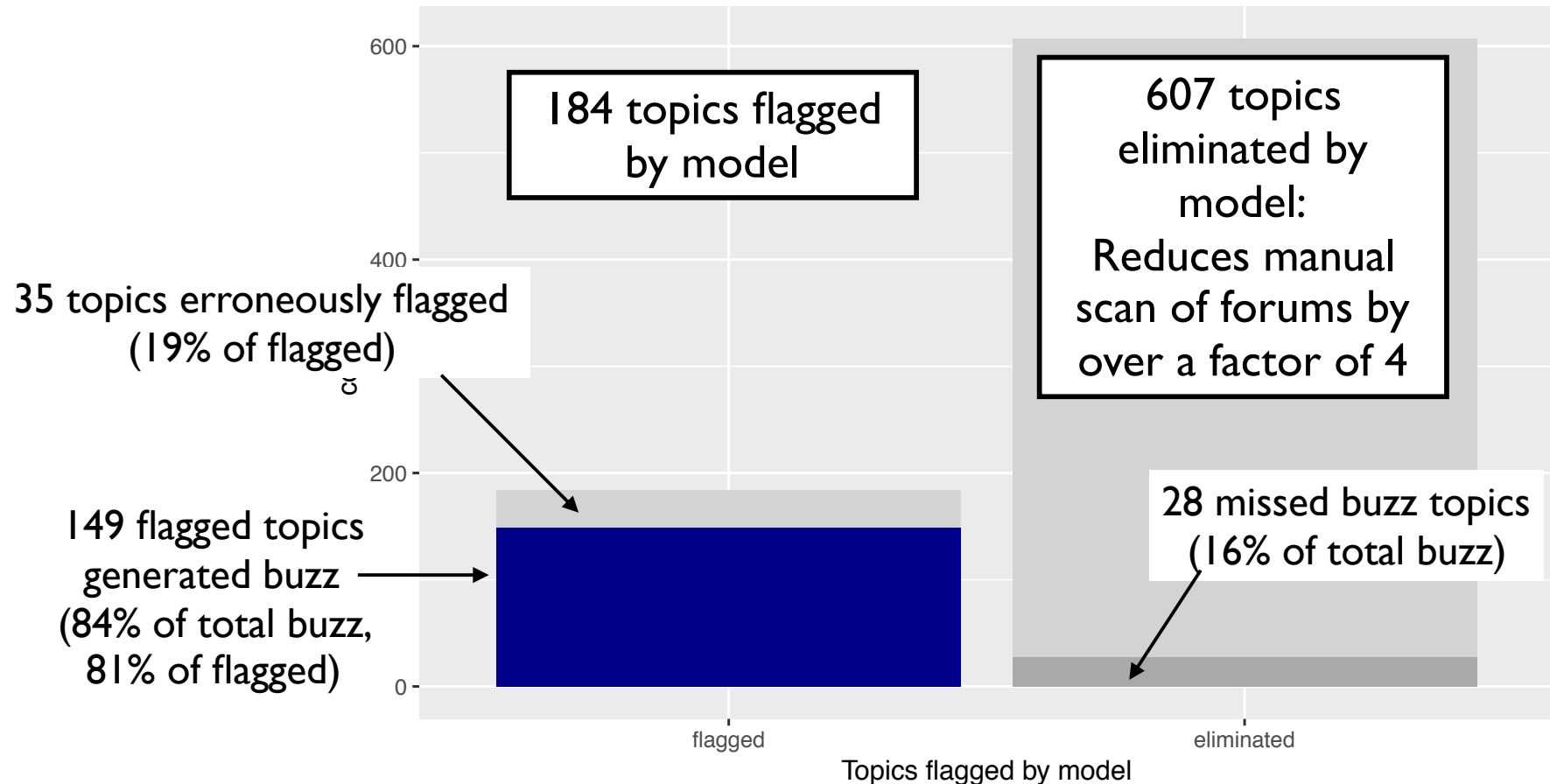
# Pilot Study

- Collected three weeks of data from forum
- Trained model on Week 1 to identify which topics will buzz in Weeks 2/3
  - **Buzz = Sustained increase of 500+ active discussions in topic/day, relative to Week 1, Day 1**
- Compared predicted results to topics that actually buzzed
- Feedback from team of five product managers -- how useful were the results?

# Pilot results promising

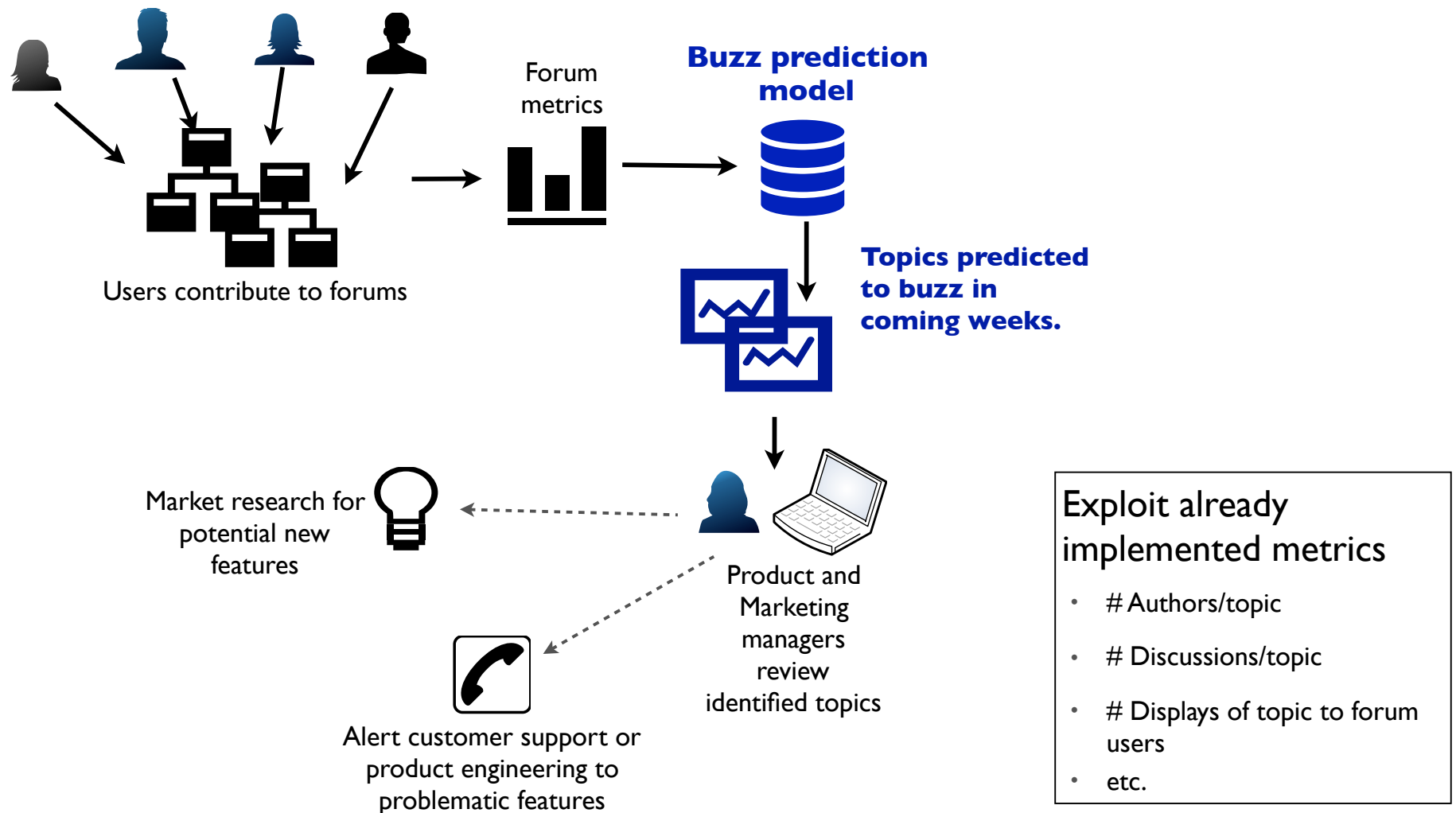
Buzz identification: results

791 topics total; 177 generated buzz



PMs: 75% of identified topics produced “valuable insight”

# How it Works





# Buzz Model

- Random Forest Model
  - Many “experts” voting
  - Runs efficiently on large data
  - Handles a large number of input variables
  - Few prior assumptions about how variables interact, or which are most relevant
  - Very accurate

# Example:

## Catching An Issue Early

- Topic: TimeWrangler → GCal Integration
  - # discussions up since GCal v. 7 release
    - GCal events not consistently showing up; mislabeled.
    - TimeWrangler tasks going to wrong GCalendar
  - **Hot on forums before hot in customer support logs**
    - Forum activity triggered the model two days after GCal update
    - Customer support didn't notice for a week

# Next Steps

- Further reduce PM workload, give them better customer intelligence.
  - New metrics for better prediction
    - Record if discussion activity is growing/shrinking, and how fast
    - Why do new forum users join? What question did they come to ask?
  - Goal: Find 98% of impending buzz, 10% false positive rate
- Efficiently route buzz info to relevant Product Managers, Marketing, and Customer Support groups

**Thank You**