International Conference on Digital Disease Detection

#DigDisDet

Michael Paul

Johns Hopkins University and Microsoft Research (internship)

Joint work with Eric Horvitz and Ryen White (Microsoft Research)

Understanding Cancer Patients through Search Engine Query Logs

- What are the information needs of a person diagnosed with cancer?
- How do these needs change over time?

Understanding Cancer Patients through Search Engine Query Logs

- What can we learn about a cancer patient from their search queries?

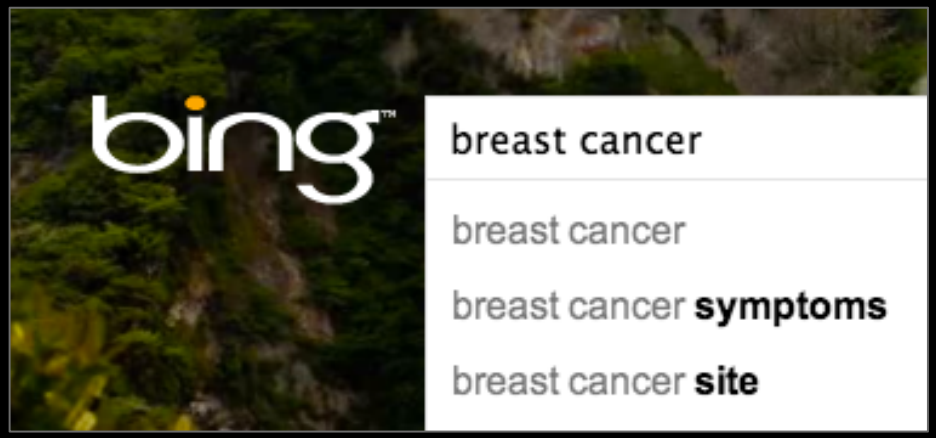
- Insights into what a person is thinking and planning over the course of an illness

Search Logs

- 12 TB of search query logs
  - from Bing + Internet Explorer & toolbar
  - 18 months of history
- Logs are anonymized; users consent to share

Search Logs

- 140,000 users searched “breast cancer” at least 3 times



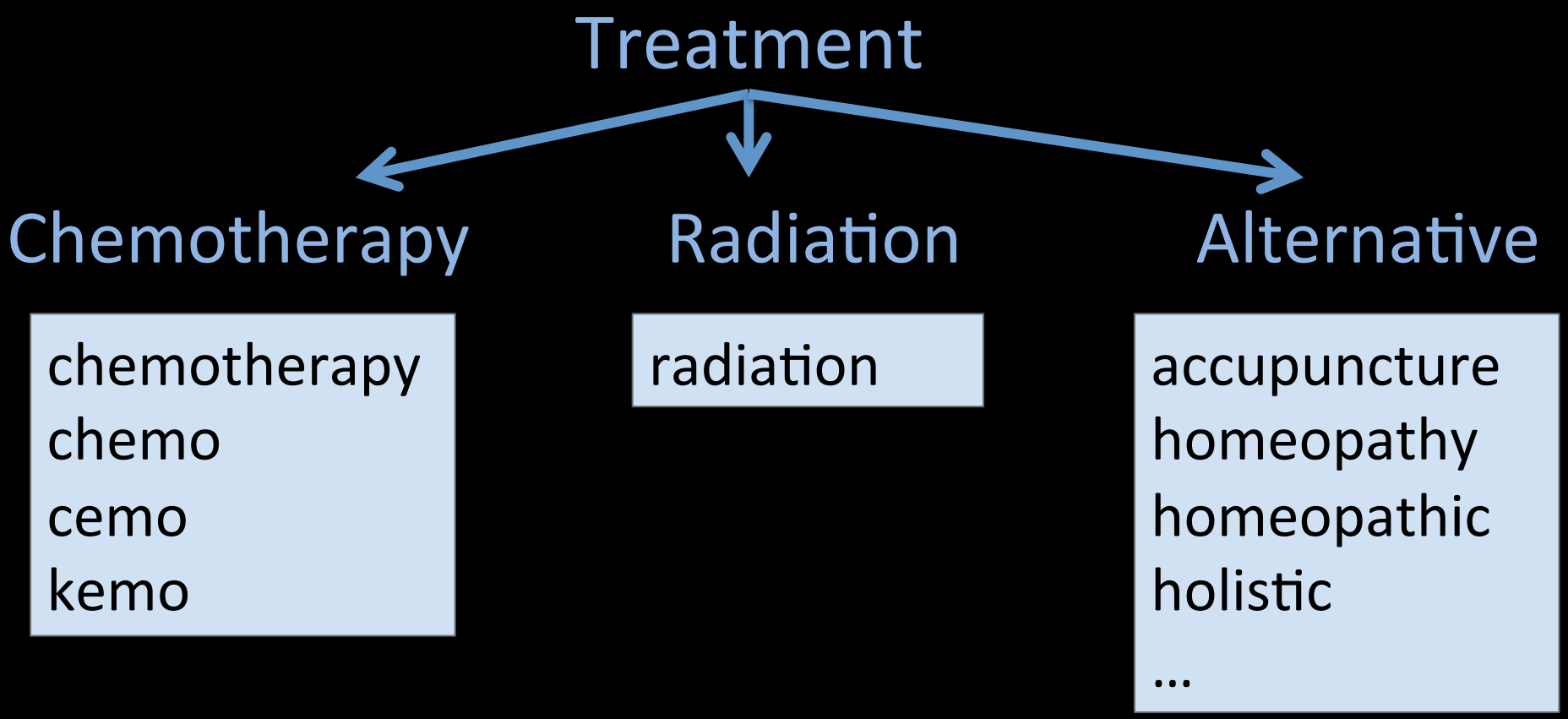
- Collected queries containing relevant terms

Ontology of Search Terms

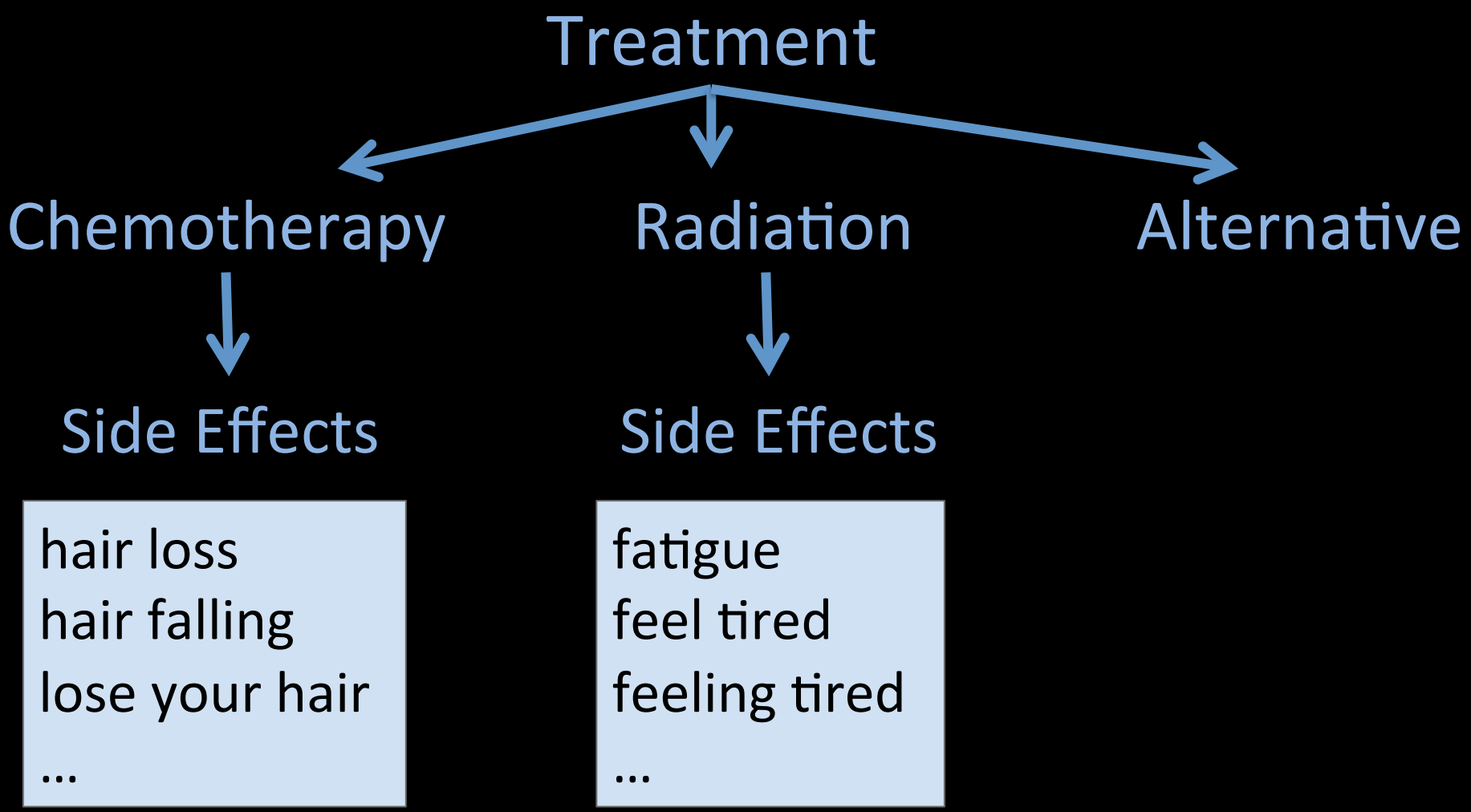
Treatment

treatment
treatments
medication
medications
…

Ontology of Search Terms

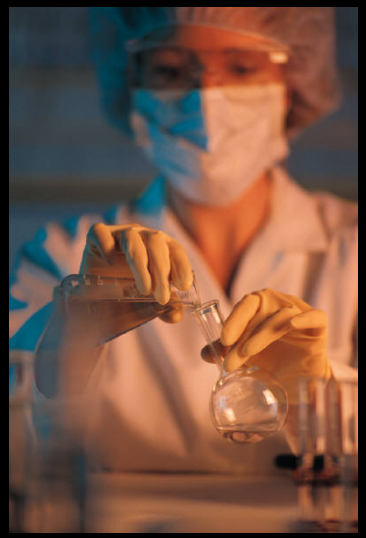


Ontology of Search Terms



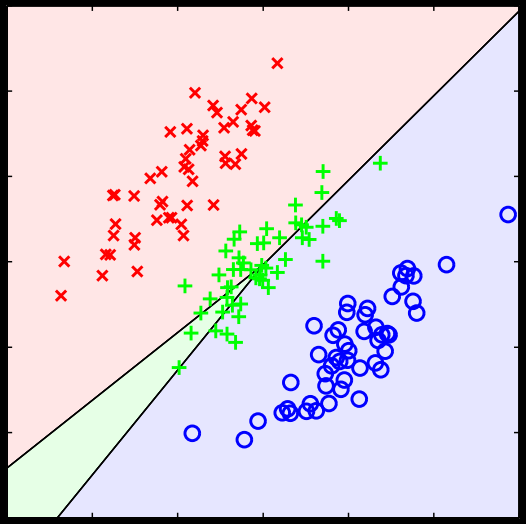
Identifying Likely Cancer Patients

- Looked at partial search histories of 480 users
- Identified 107 users who appeared to have been recently diagnosed



Automatic Classification

- Supervised machine learning
- Trained models using 480 annotated users



Automatic Classification

- Cross-validation performance:
  - 94% precision
  - 29% recall
- Over 2,000 users identified by high-precision classifier as likely to have cancer

Identifying Day of Diagnosis

- Annotated the timelines of the 107 users identified as likely recent cancer patients

- Identified the plausible day of diagnosis

Automatic Classification

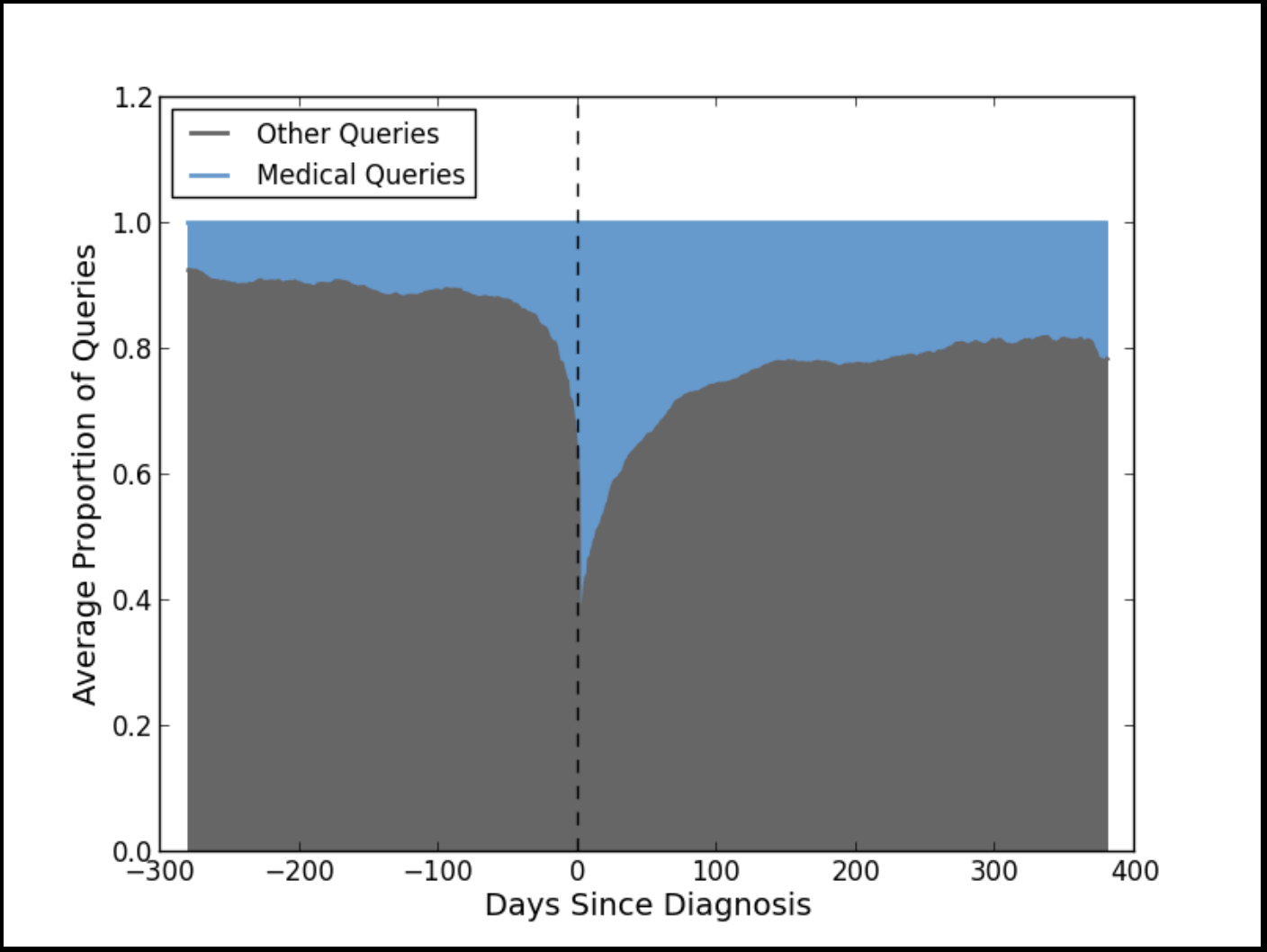
- Cross-validation performance:
  - 41% exactly correct
  - 75% within 7 days

Search Volume Over Time

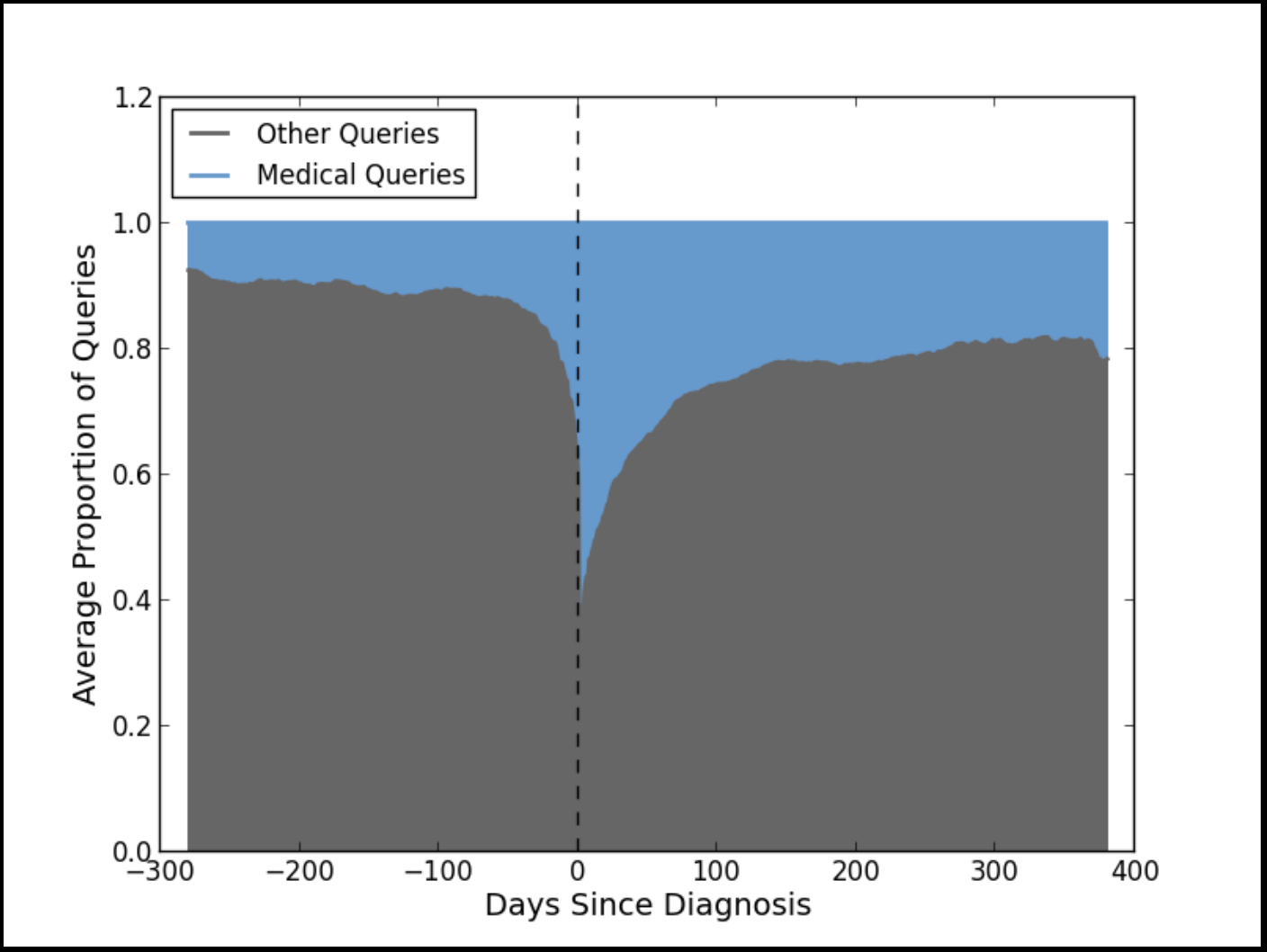
- Can align all user histories around the day of diagnosis

- Analyzed the information users search for before and after diagnosis

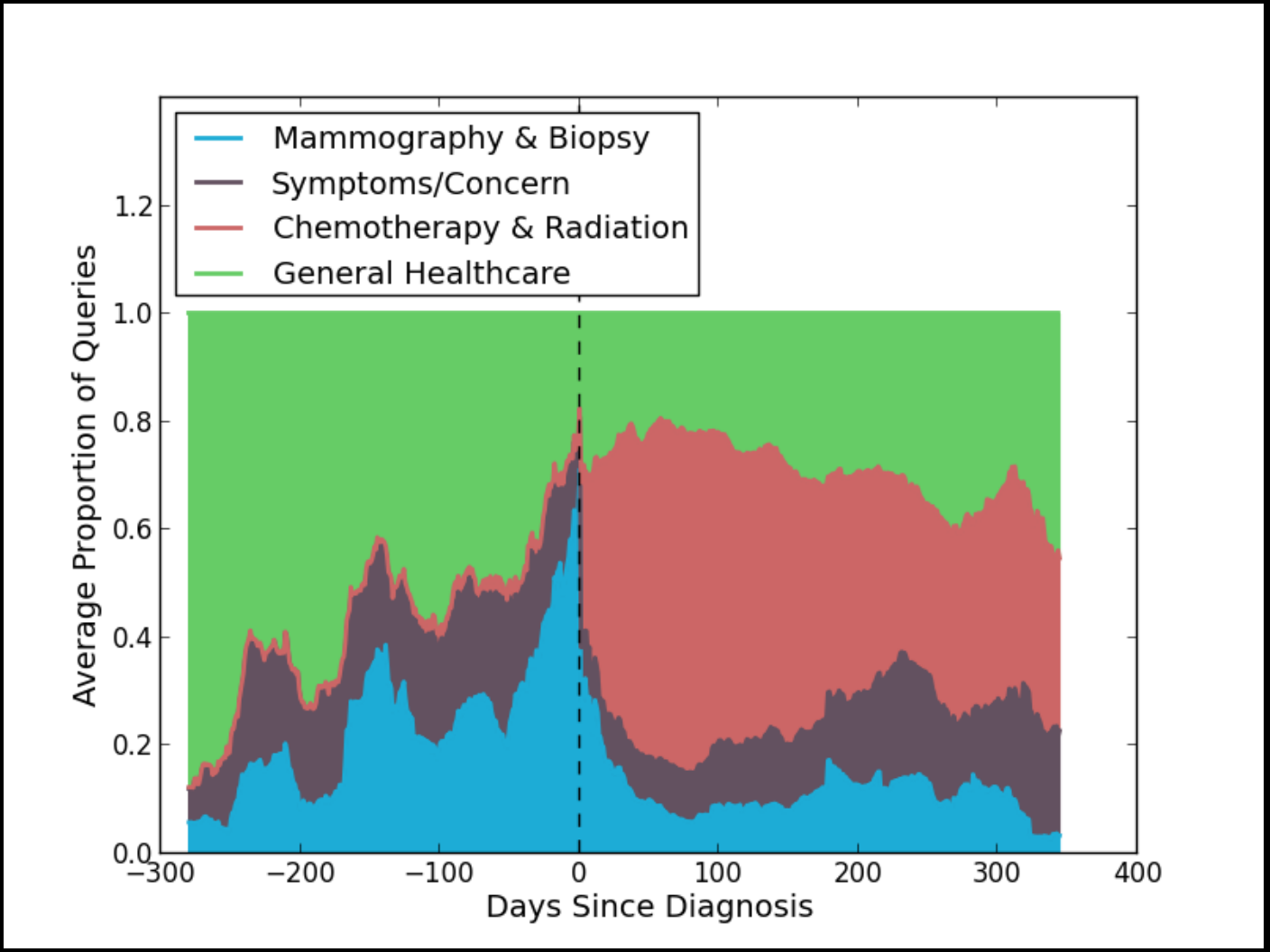
Search Volume Over Time



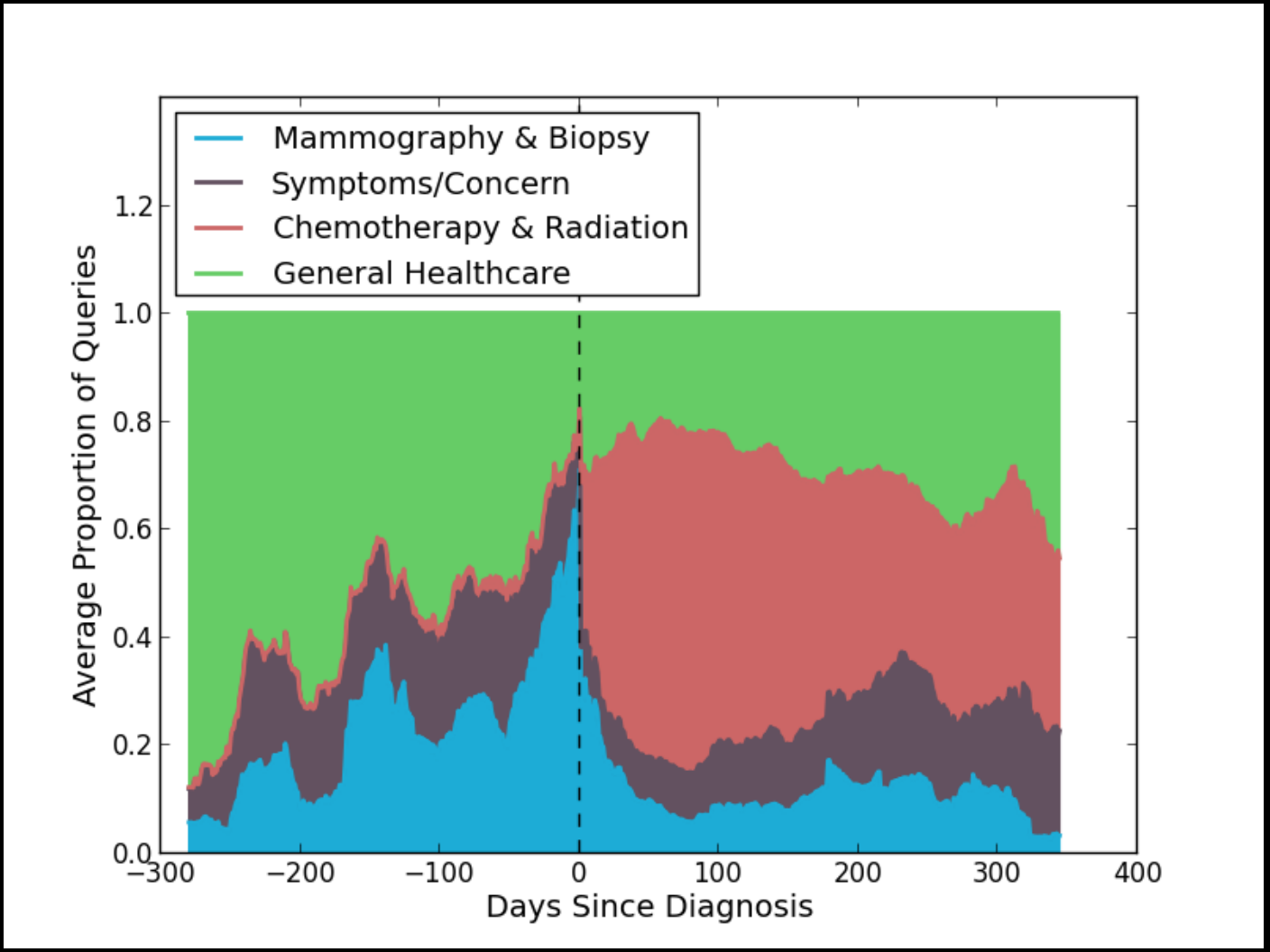
Search Volume Over Time



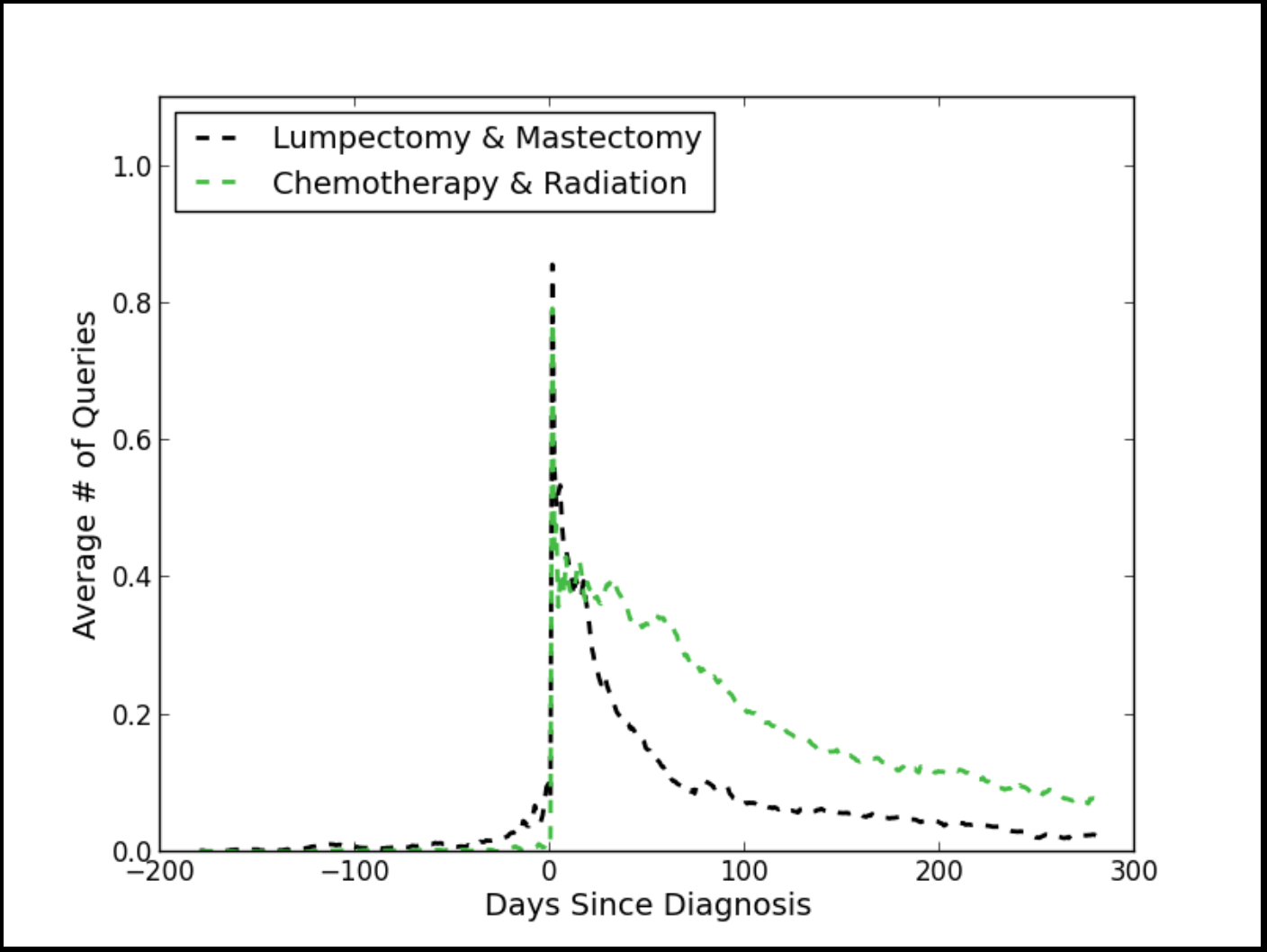
Search Volume Over Time



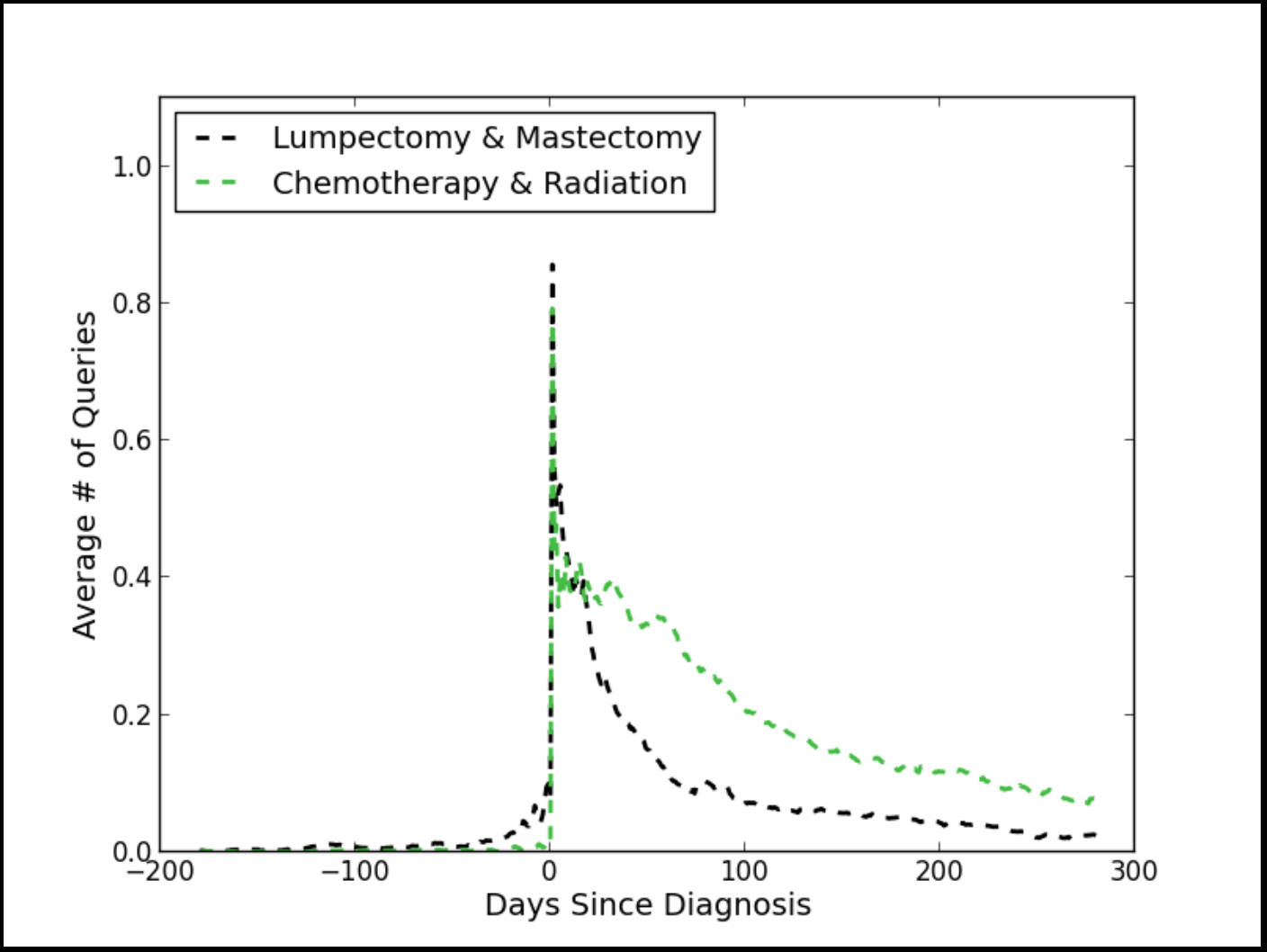
Search Volume Over Time



Search Volume Over Time



Search Volume Over Time



Future Directions

- Deeper analysis of search queries during key points in history

- Break down search histories based on age of user, stage of cancer, or other attributes