Motivation: Forecasting the Present via Social Media

Real world phenomena like unemployment and disease behavior have been estimated using social media – a process known as **nowcasting** [1].

These estimates are **cheaper** and **faster** than traditional data collection methods like phone surveys.

Faster and cheaper means **saved money** and **better policy**.

**Problem:** Nowcasting Requires Ground Truth Data

Yet, economists [3] still want estimates for targets lacking ground truth data:

- 10-day auto sales (once used, but no longer available [2])
- Physical movement (e.g., out of parents house, for job)
- etc.

**Solution:** Substitute Domain Knowledge for Ground Truth

Our target users have some **semantic** and **signal** domain knowledge about their target phenomena:

- **User’s Domain Knowledge**
  - Semantic Knowledge (keywords, etc.)
  - Signal Knowledge (seasonal trends, etc.)

**Example User Query (Target: Unemployment Behavior)**

Unemployment-related keywords

Seasonal trends of unemployment (drawn by user)

**User Interaction Loop**

Like a search engine, **users iteratively submit queries**:

1. **“Ground Truth Data”**
   - 10/17: I need a job 483
   - 12/26: I love you 5092
   - 1/28: Justin Bieber 940,291

2. **Topics Used**:
   - Relevant Topic 1
   - Relevant Topic 2

3. **Resulting Target Estimate**:
   - 2012: Correlation with query signal: 0.897
   - 2013: Resulting Target Estimate: 0.897
   - 2014: Resulting Target Estimate: 0.897
   - 2015: Resulting Target Estimate: 0.897

**Users Given Trade-Off Between Query Components**

Users can **explore a set of query results in real-time**, produced by varying the relative influence of the semantic and signal query components:

**Architectural Trade-off**

- **User Query**
  - Scoring and Ranking
  - Aggregation

**Scalability with the Cloud**

- **Current Prototype System**:
  - 40 billion tweets (collected 2011-2015)
  - 150 million topic-signal pairs (after threshold filtering)
  - Query processing runtime: ~20s (on 10 EC2 c3.8xlarge servers)
  - Topic-signal pairs support daily updating

**Related Work**


**Contact**

dol@umich.edu  •  http://web.eecs.umich.edu/~dol