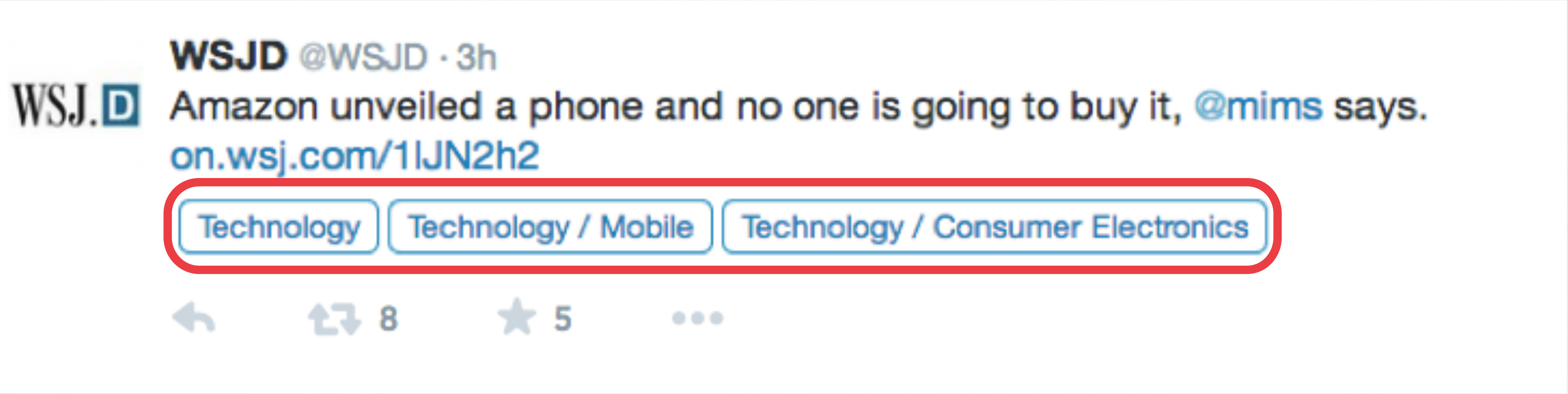
Large Scale High-Precision Topic Modeling on Twitter

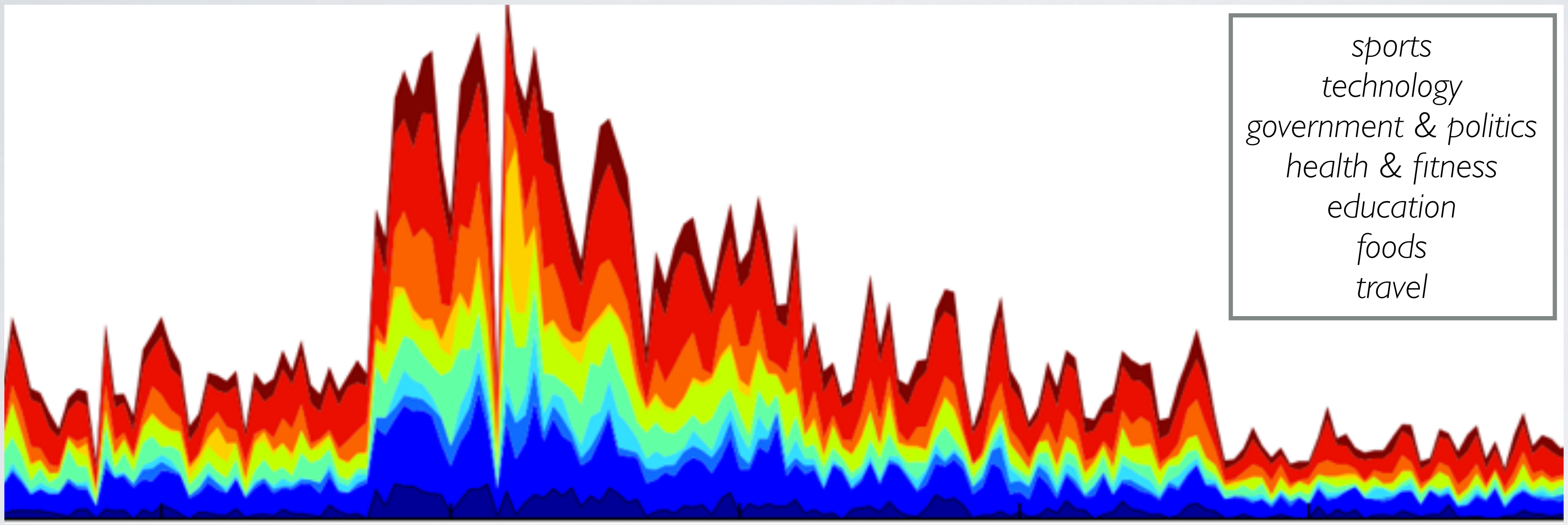
Shuang Yang, Alek Kolcz
Andy Schlaikjer, Pankaj Gupta

Topic modeling of Tweets



Many Use Cases

Business intelligence & analytics

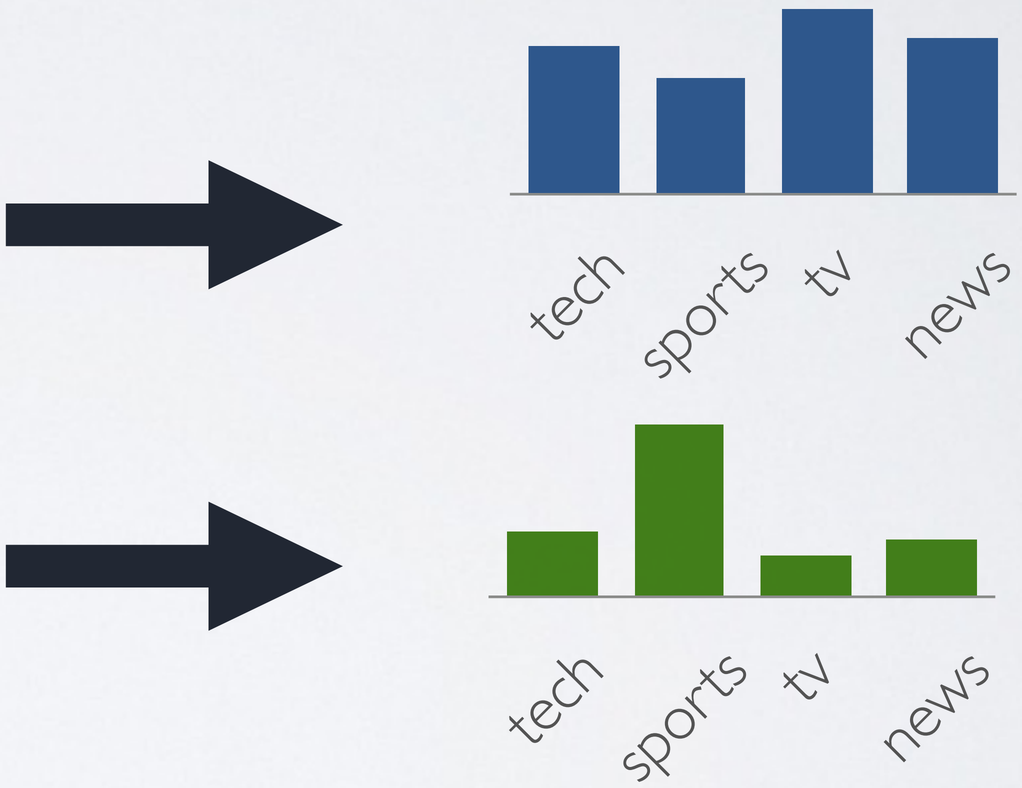


Many Use Cases

Business intelligence & analytics

User interest modeling

	Tweets consumed	Tweets produced
@userX		
@userY		

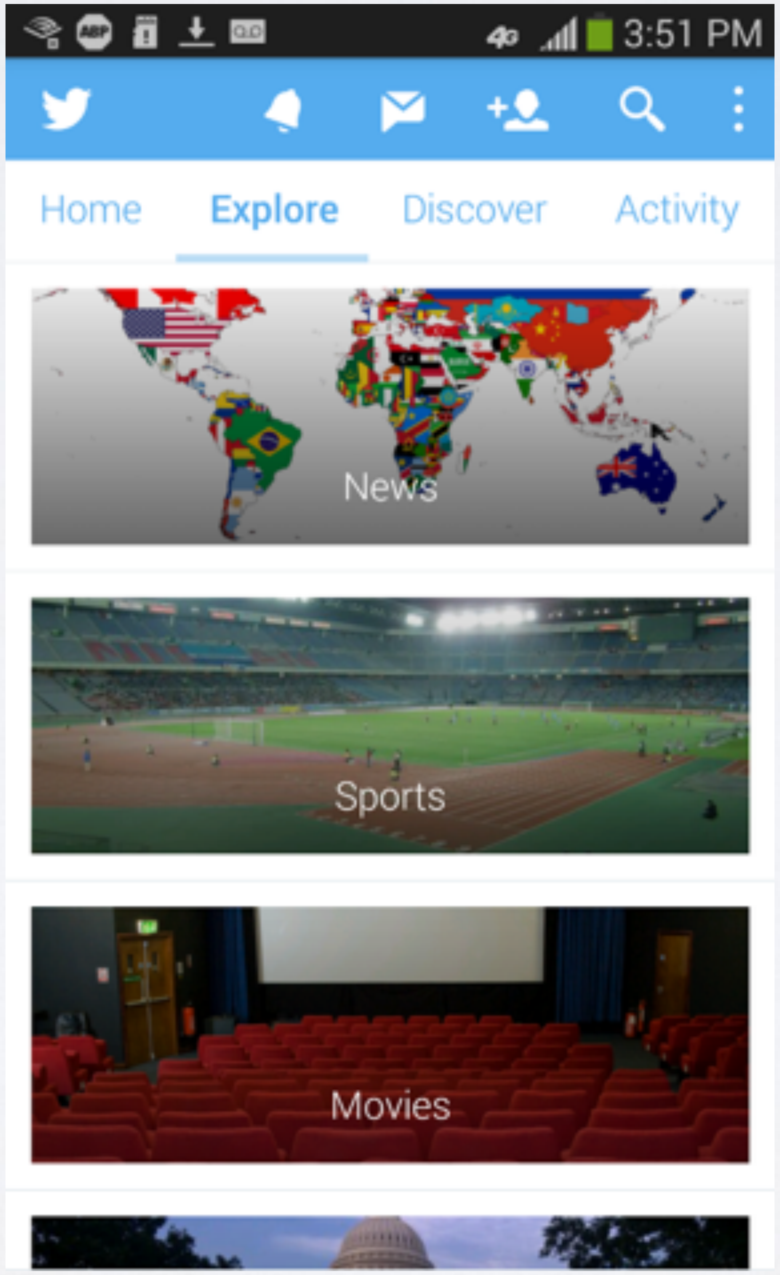


Many Use Cases

Business intelligence & analytics

User interest modeling

Topic channels



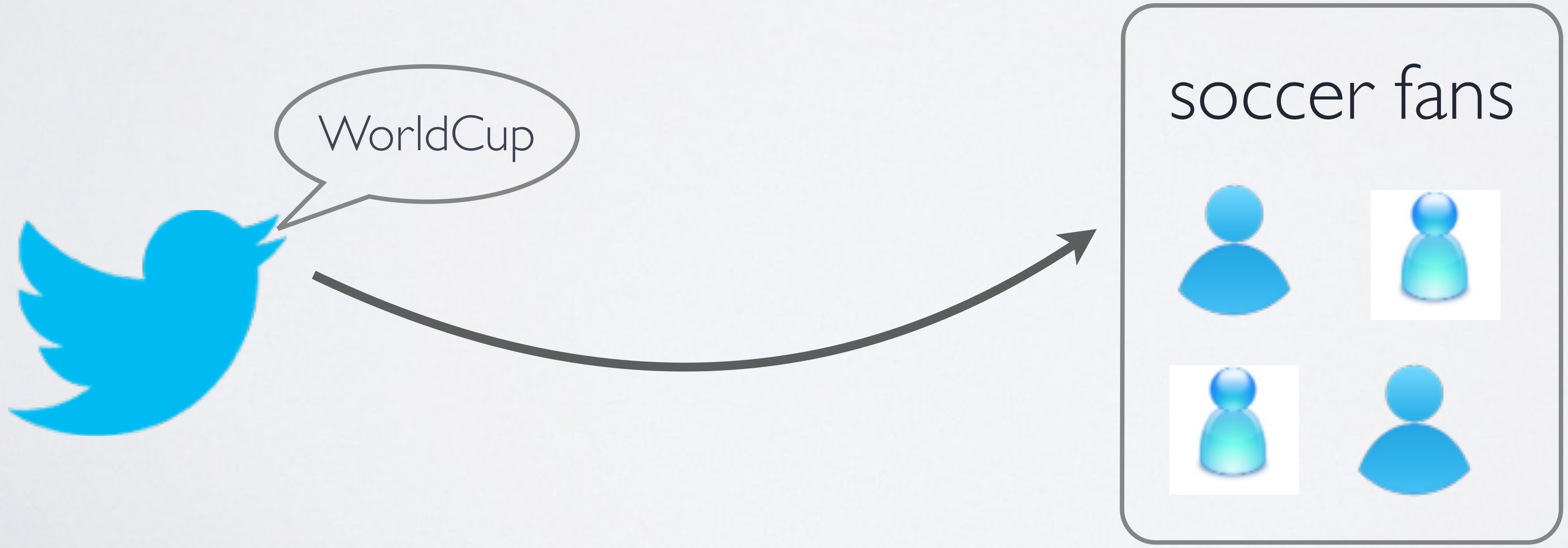
Many Use Cases

Business intelligence & analytics

User interest modeling

Topic channels

Personalization & recommendation



Many Use Cases

Business intelligence & analytics

User interest modeling

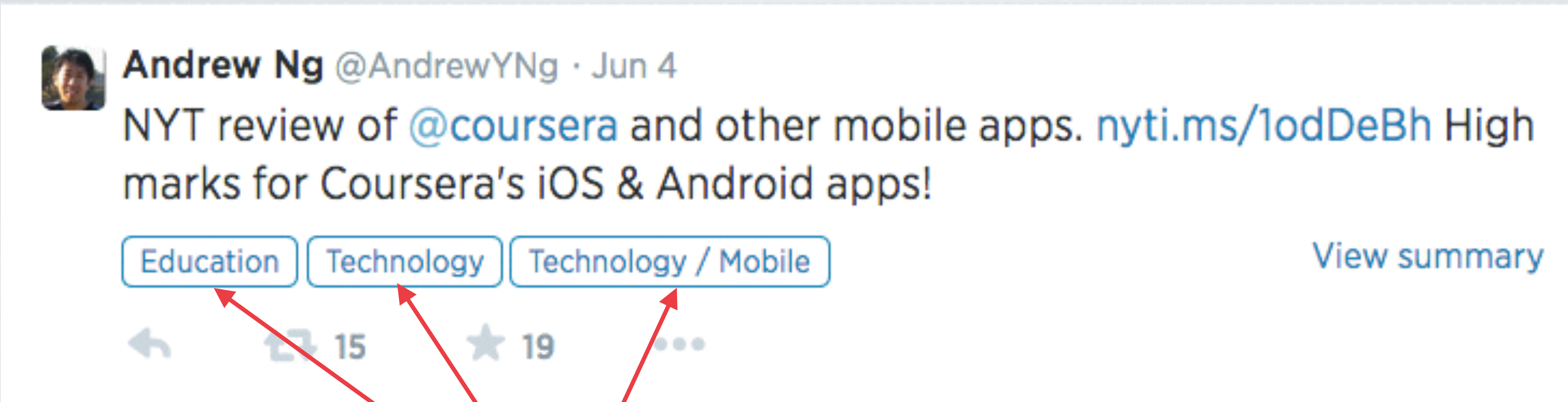
Topic channels

Personalization & recommendation

Many more:
- Ads targeting
- News feed ranking
- Ads CTR prediction
- ROI optimization
- Search index
- Event summarization
- ...

Quality requirement

90+% precision



A topic tag is correct with at least 90% confidence.

Quality requirement

90+% precision

Extremely challenging:

- **sparse**: <140 chars, ~7 unigram terms
- **noisy**: can contain any strings, e.g, “w00t”, “gr8”, “5sos”
- **ambiguous**: 0 (conversational) or multiple topics
- **dynamic:** trending topics change rapidly with new terms, events and entities emerging every second

Existing approaches?

Don’t Work!

Existing approaches?

Don’t Work!

LDA (latent Dirichlet allocation) & variants

- ~40% precision, 90% precision = mission impossible
- Not easy to align the results to a predefined taxonomy
- Topics will reshuffle if retrained over a different data set and / or the granularity of topics are refined

Existing approaches?

Don't Work!

LDA (latent Dirichlet allocation) & variants

- ~40% precision, 90% precision = mission impossible
- Not easy to align the results to a predefined taxonomy
- Topics will reshuffle if retrained over a different data set and / or the granularity of topics are refined

Topic filtering / language models:

- Works well only for relatively focused topics (e.g, NBA)
- Hard to scale to a large number of general topics

Meet Jubjub

Twitter topic modeling system

On full-scale Twitter data
- 270M+ MAUs, 500M+ tweets/day, 400B+ total

In real-time
- 150K+ requests per second, sub-ms latency

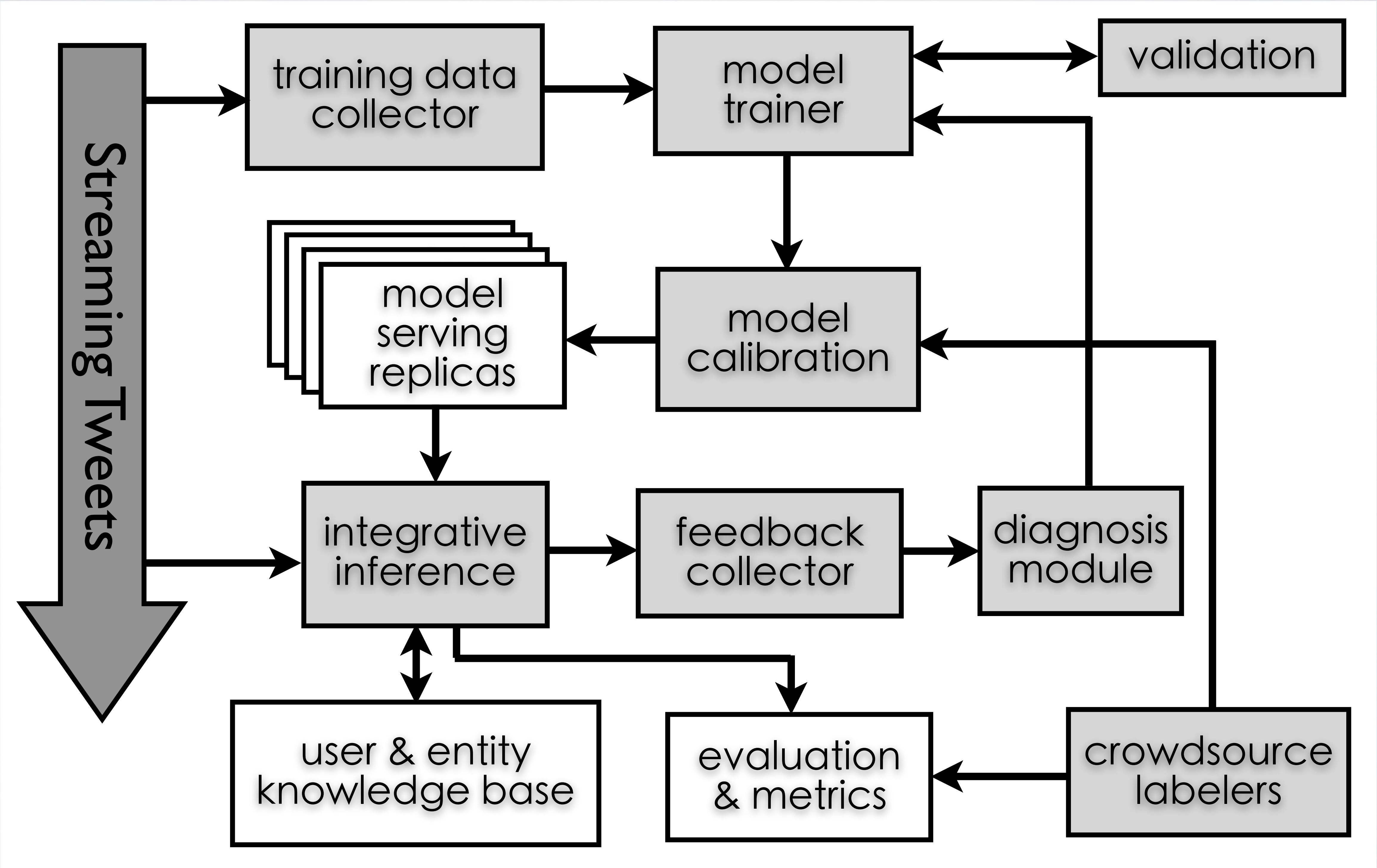
Over a structured taxonomy
- 350+ atomic topics, ~7-depth DAG/tree-taxonomy

At ~93% precision
- ~40% coverage (on English-language tweets)

How?

Technical solutions

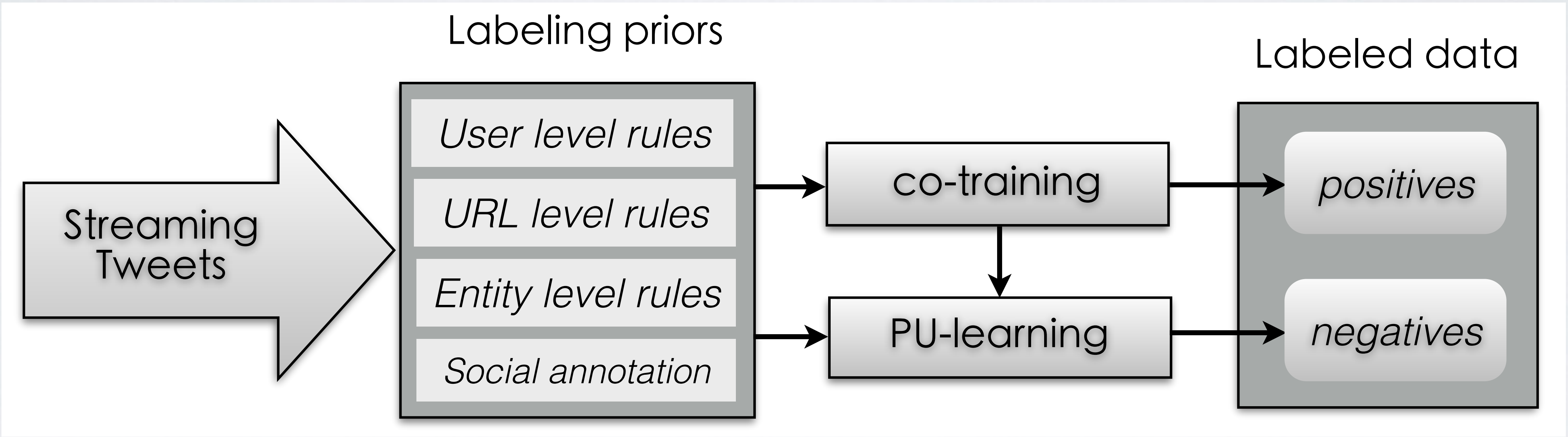
Architecture overview



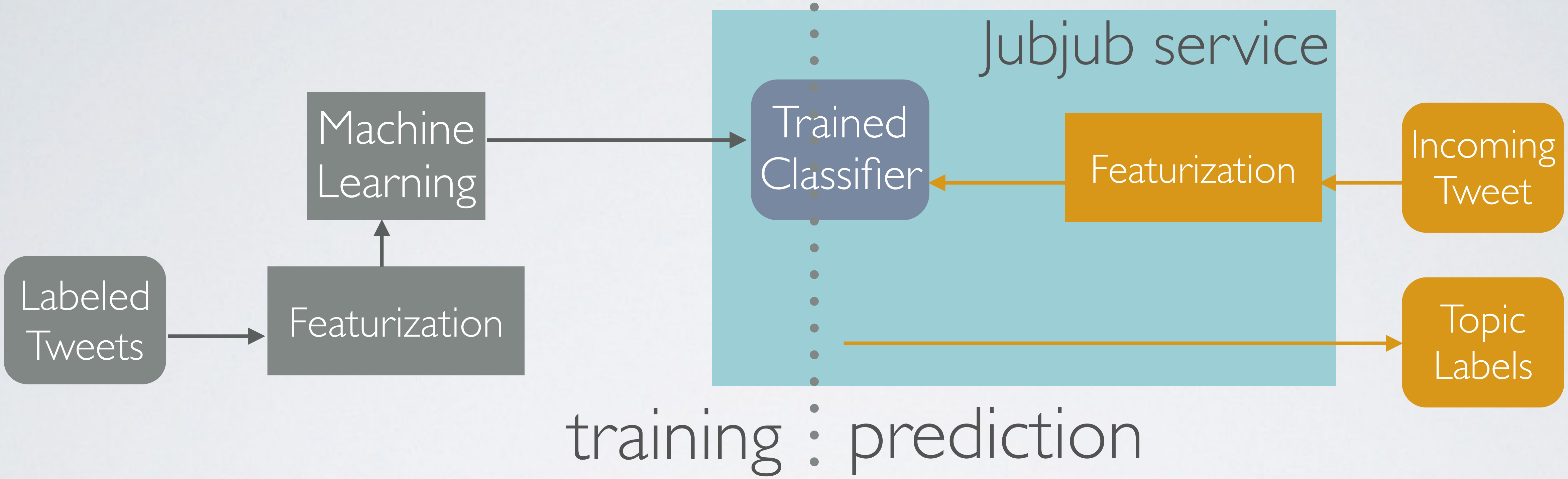
Labeled data acquisition

Human annotation is prohibitive for the scale we consider

Automatically collect high-quality labeled data from Twitter stream



Tweet text classification



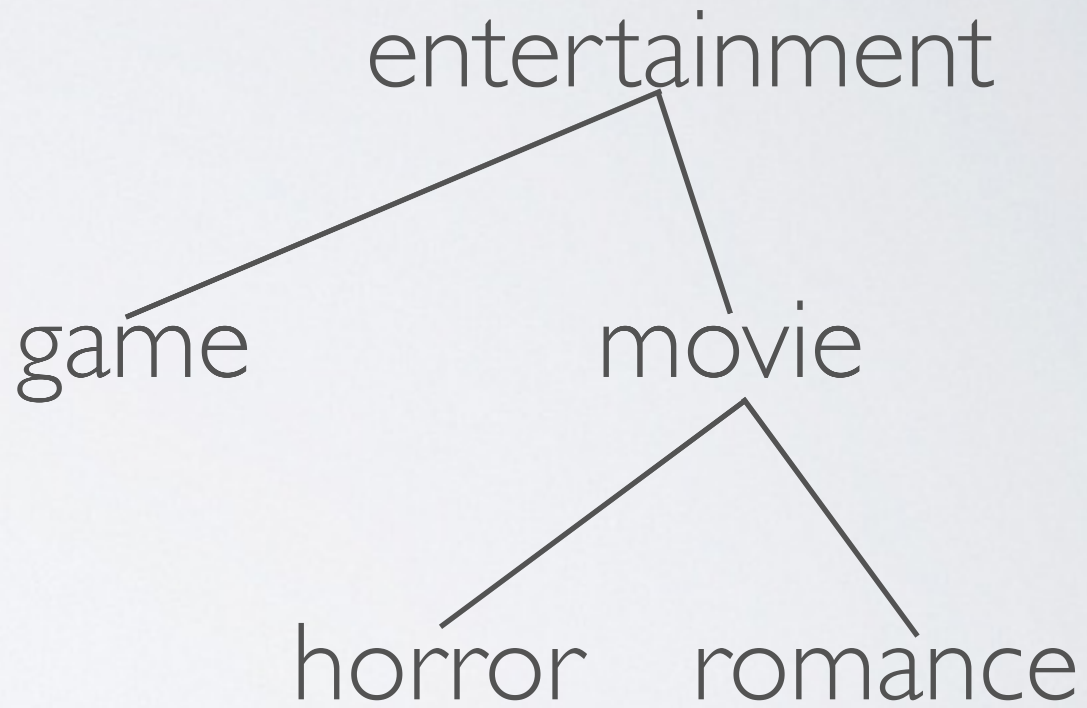
- Extract feature efficiently on the fly, robust to misspelling and abbreviation, without expensive preprocessing (stemming, pre-pruning)

- Training high-quality classifiers at scale

Label correlation

Relational classification

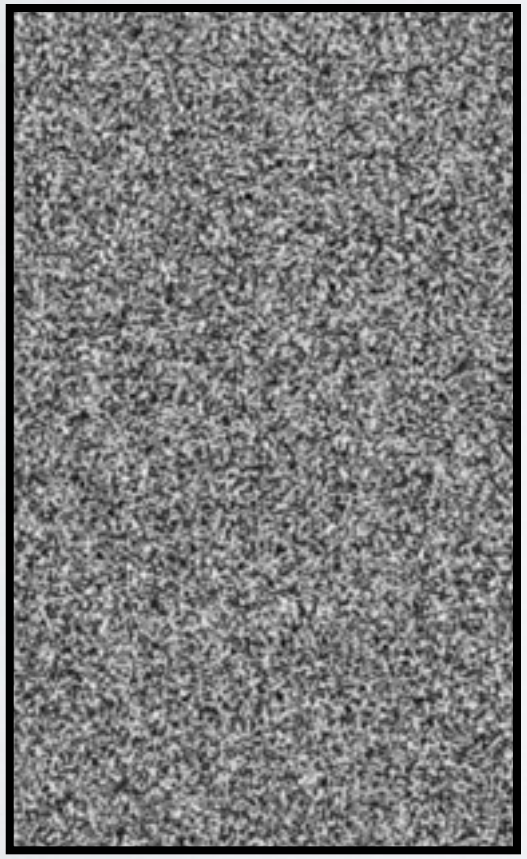
- Data sharing via label propagation
- Parameter sharing via hierarchical regularization
- Cost sensitive training



Two stage learning

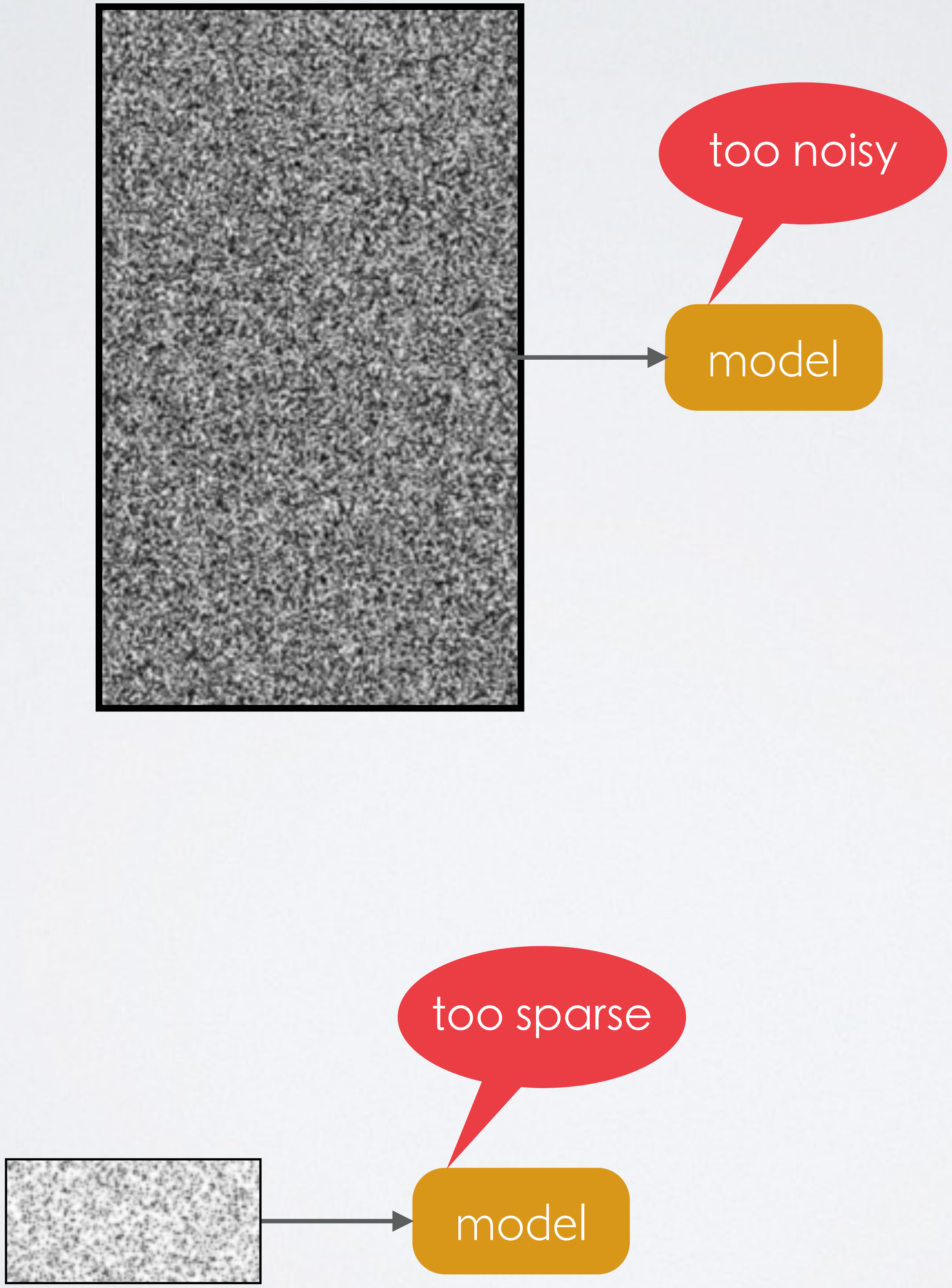
Two types of data with different (volume, signal-noise ratio):

- Large amount (virtually unlimited) of noisy data,
- Small amount of good (human-labeled) data



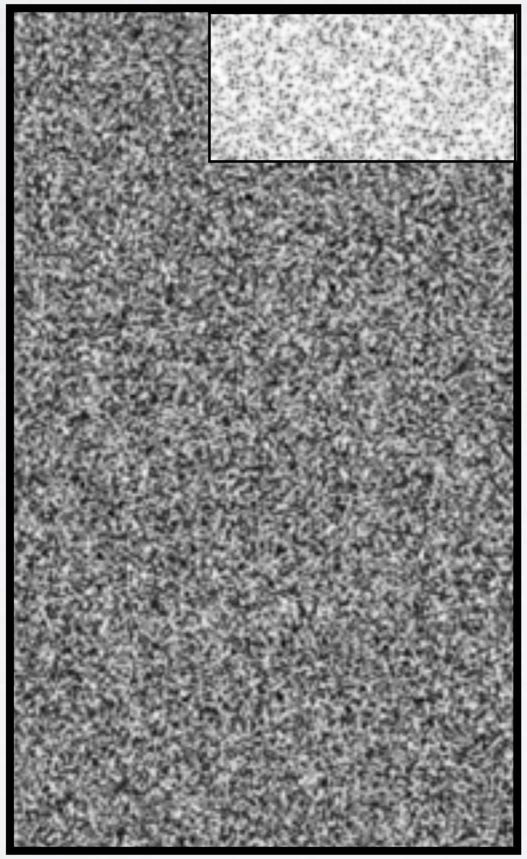
Two stage learning

Models trained on either data set is poor



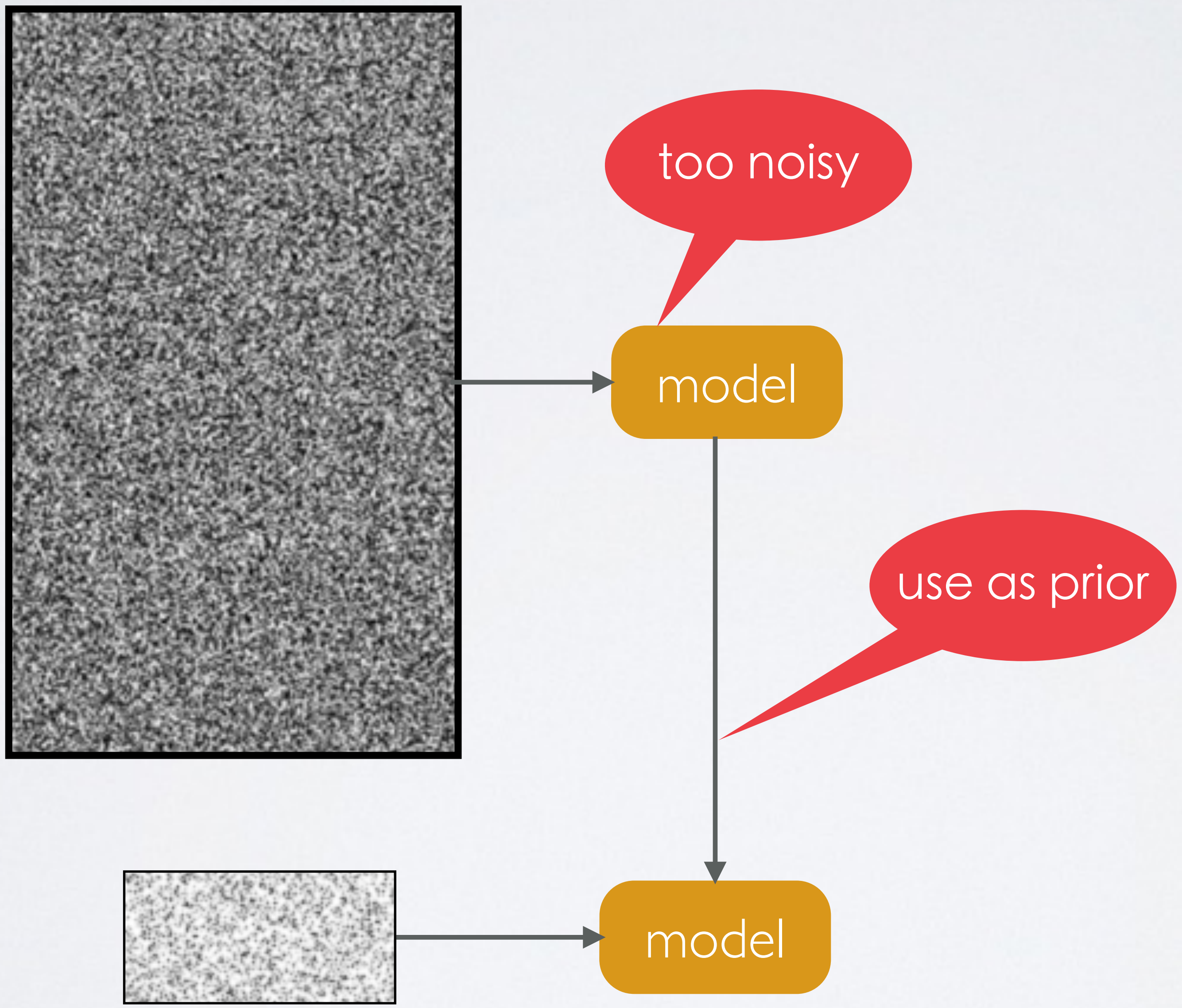
Two stage learning

Training on combined data won't either (noisy data dominate in amount)



Two stage learning

Two stage training: fine tune noisy model on good data



Diagnosis & corrective learning

Detect model mistakes via wisdom of crowd

Diagnose systematic model mistakes and correct it on the fly

Decision Rejection

Chatter tweets:

- A large proportion of tweets are non-topical conversations
- Detect chatter tweets and reject scoring them to save latency

Topical Topics
* potter, harry, tumblr, fandom, weirdly, gifs, gif, hogwarts, fictional, rewatch, hp, rowlingpuns, obvs, deathly * rep., ballot, officials, mayor, gov, votes, investigation, ballots, mayor's, committee, capitol, district, courthouse, senate * donation, charity, donated, awareness, donations, donate, autism, raise, help, funds, fundraiser, fundraising, donating, support * investors, financial, banks, debt, markets, goldman, finance, banking, stocks, economic, earnings, ipo, bank, equity, investment
Chatter Topics
* foh, lmaooo, lmaooooo, lml, lmaoo, deadass, henny, lmaooooo, niggas, djzeeti, smfh, nah, nigga, tho * coworker, washer, dangit, dryer, yeah, i've, beeping, oh, 6ish, i'll, 10ish, nope, 4am, kinda, probably * thanks, enjoyed, congrats, big, next, week, soon, coming, everyone, well, incredible, wow, meeting, achievement, weekend * someone, because, hate, sometimes, anymore, she, person, tell, her, aren't, saying, sleep, enough, without, ask, real, money

Low confidence tweets:

- Inference on tweets less than 10chars are error prone.

Model calibration:

- Topics whose models cannot meet a target precision are tuned off (th = 1.0).
- Threshold estimation in the context of data drift

Quality evaluation

Quality labeled data via crowdsource annotation:

- Binary question: Is this tweet about the assigned topic?: (tweet, topic)
- easier task, lower cost, less likely to make mistakes
- Biased, cannot be used for recall evaluation

Quality control strategies:

- Topic probes.
- Worker level quality monitor and admission
- Confidence estimation
- …

Beyond text

Tweet = Envelope

- tweet text
- embedded url
- author
- engagers
- entities
  * #hashtag
  * @mentions
  * named-entities ...
- contexts:
  * time stamp
  * geo
  * social
  * media
  * taxonomy …

Derive topics from other signals

URL:
- Webpage (crawled) text classification
- Cache classification results for seen URLs

Author:
- Known-For topics derived from content production and social annotation

Engager:
- Interested-In topics derived from content consumption and interest graph

Entities (e.g., #hashtags, @mentions, named-entities)
- A trained high-precision retrieval model

Integrative inference

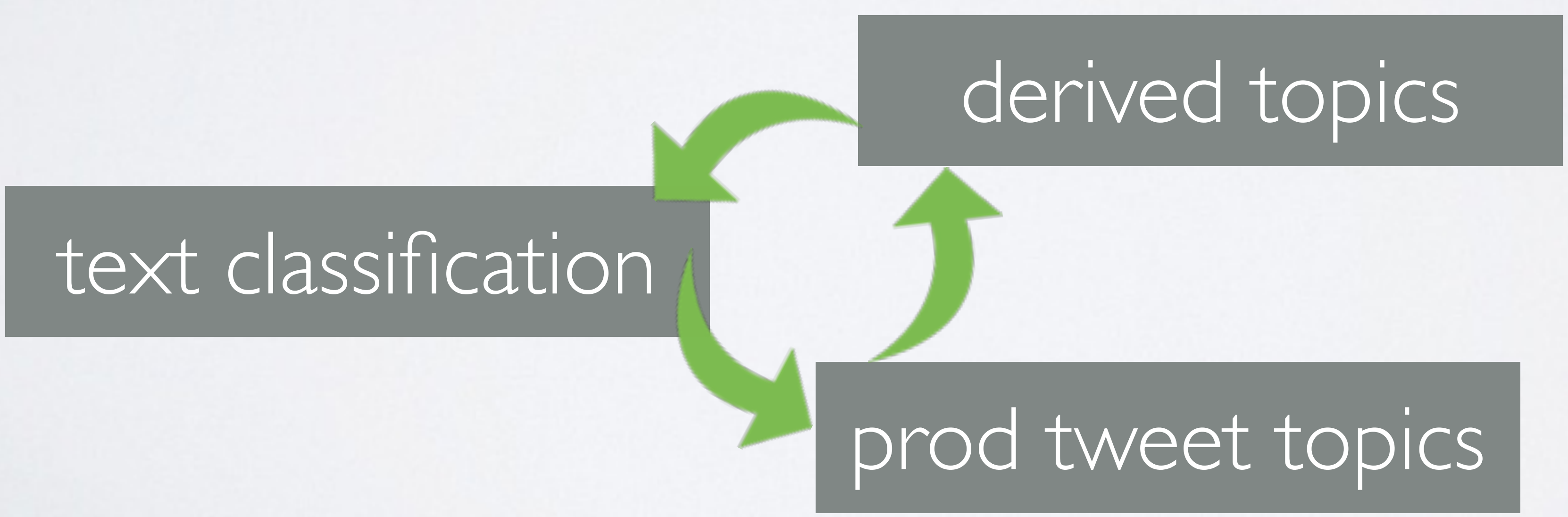
Topical signals as multiple noisy experts:

- Multi-modality: each expert is only good at its area of expertise

Integrative inference: beat the best expert in hindsight

	alan	bod	alice	dave	eric	ensemble
instance 1	☹	☺	☹	☺	☺	☺
instance 2	☺	☹	☹	☺	☹	☺
instance 3	☺	☹	☺	☹	☹	☺
instance 4	☹	☺	☹	☺	☺	☺
instance 5	☹	☺	☺	☹	☹	☺
instance 6	☺	☹	☹	☺	☹	☺

Close the inference loop



Summary

Jubjub Twitter topic modeling system

- Infer topics for tweets which are noisy, sparse and ambiguous in nature
- At full Twitter scale
- In Real-time
- Over a structured taxonomy of 300+ topics
- At 93% precision with ~40% coverage

A full stack of topic modeling techniques

- Auto. labeled data acquisition at no cost
- Efficient and effective featurization
- Multi-class multi-label classification at scale
- Relational regularization
- Diagnosis and corrective learning
- Two stage training
- Closed-loop integrative inference
- Human computation for quality evaluation

Thank you!
Want to know more?

- Come to our Poster: (@)

- Read our paper: (Yang, Kolcz, Schlaikjer, Gupta: Large scale high-precision topic modeling on Twitter, KDD' 14.)

- Keep in touch: (Follow @syang on Twitter)