



# **Dynamic Online Profiling of Users' Content Consumption Preferences**

**Rani Nelken**

**Director of Research, Outbrain**

---

# What Is Outbrain?

## INTERNAL DISCOVERY

### OF THE Class

Schools - thought

### udent success

SEARCH

POWERED BY Google

SHARE THIS



Print  
Email  
More sharing

Recommend 210



PHOTO ILLUSTRATION/THINKSTOCK

University of Phoenix

# Imagine

an education  
built toward  
the right career.  
Yours.

Learn more >

ADVERTISEMENT

### Promoted Stories

- End of an Era: The Car in America Is Track
- This app gives 401(k) PC Maga
- How Absolute New York Times
- Bad Neighbor Warning Signs
- Misperceptions Work Environm
- 13 Amazing Uses

### More from CNN

- 'Miracle material' graphene one step closer to commercial use
- Astronomers find 'diamond engagement ring' in space
- Oregon governor performs CPR on woman
- 'Very confident' black boxes detected
- Hyperloop vs. world's fastest trains
- 'Kill switch' may be standard on U.S. phones in 2015

# Distribution Partners

560

MILLION

MONTHLY  
UNIQUE VISITORS  
GLOBALLY



UNIVISION®  
COMMUNICATIONS INC



Over

25

BILLION

PAGE VIEWS

PER MONTH

---

Over

**200 BILLION**

**RECOMMENDATIONS**

**SERVED PER MONTH**



# The Lighthouse

Help people discover content they can trust to be interesting, relevant, and timely for them



# Personalization

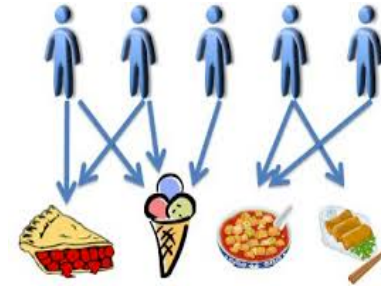
We use 2 approaches:

Content Based



This talk

Collaborative Filtering





# User profiles lifecycle

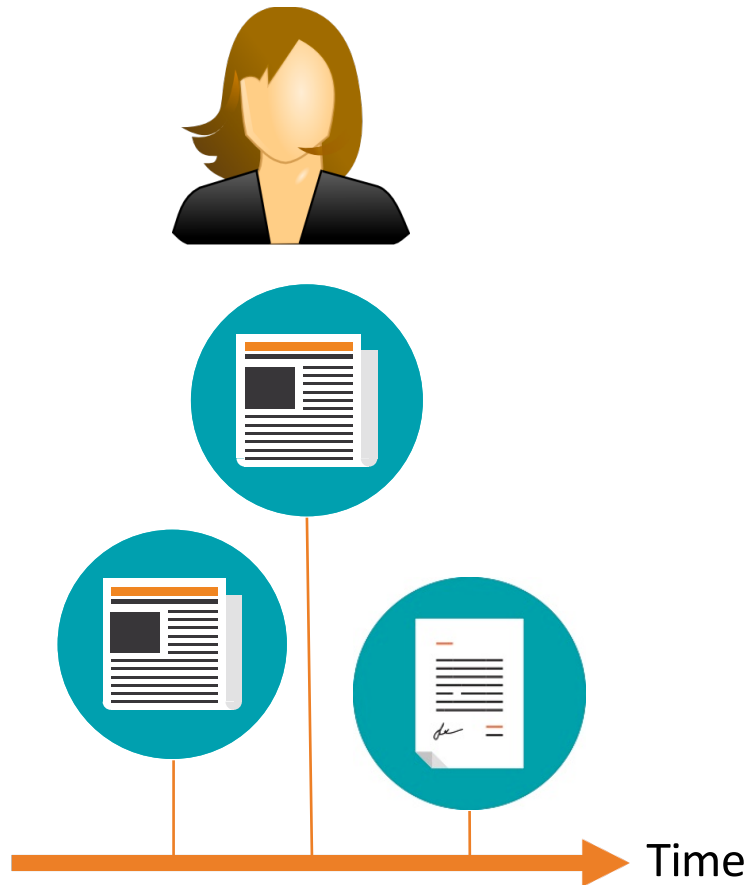
- Extract predictive features from documents
- Construct user profiles incrementally
- Match content with user profile

# Constructing the profile

Interesting algorithmic questions due to:

- Content understanding challenges
- Online updates
- Dynamic interests
- Scale

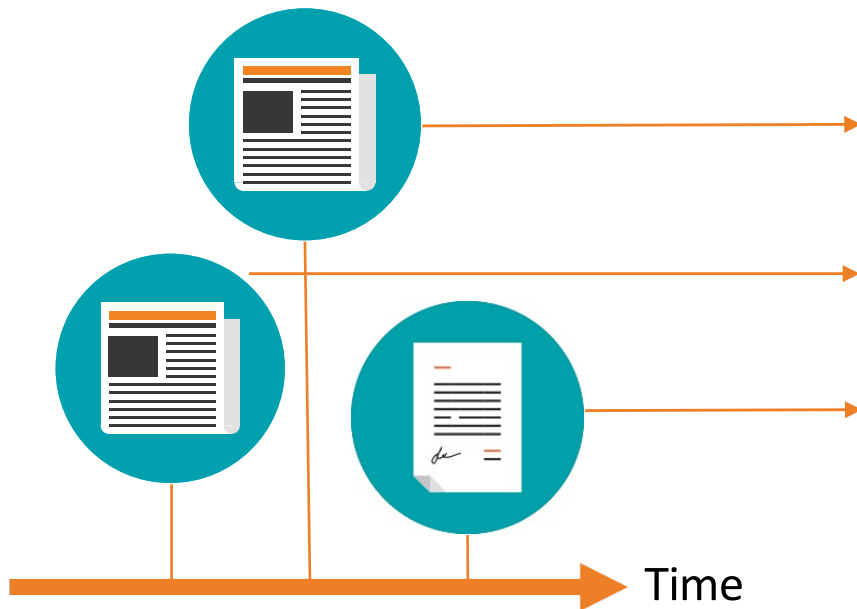
# User interacts with documents



Interaction types:

Page Views, Clicks, Shares, Likes, Video play/pause, ...

# Extract rich features from documents



- Sites
- NLP features: categories, topics, entities
- Cartesian products of features

# NLP: What is a document **about**?

- Classic “Bag of words” approach is too high-dimensional
- We want to encode document meaning in a more compact space

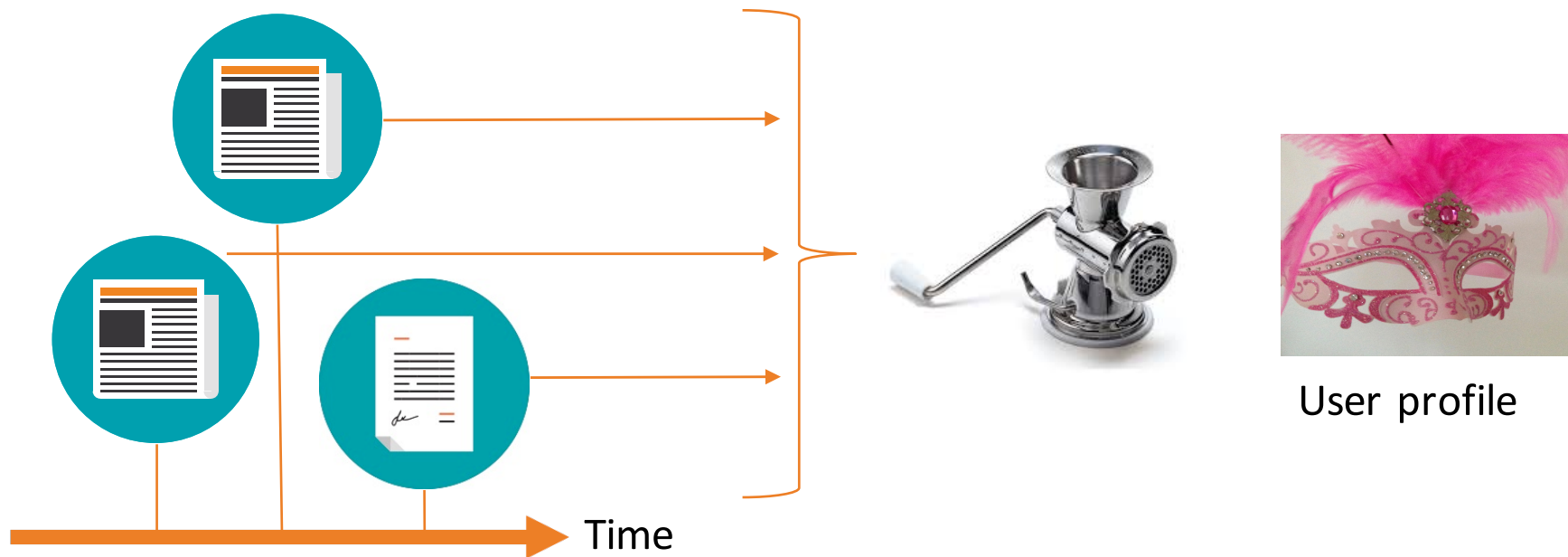
# Reduced dimensionality

- Categories
  - *Supervised* classification relative to a fixed set of categories
- Topics
  - *Unsupervised* probabilistic model
  - topic = distribution over words
  - doc = distribution over topics
- Most *relevant* Named entities  
(person, organization, location)





# Aggregate features into a profile



---

# Learning to Aggregate: requirements

- Incremental
- Dynamic
- Low space-complexity

# Learning to Aggregate: solutions

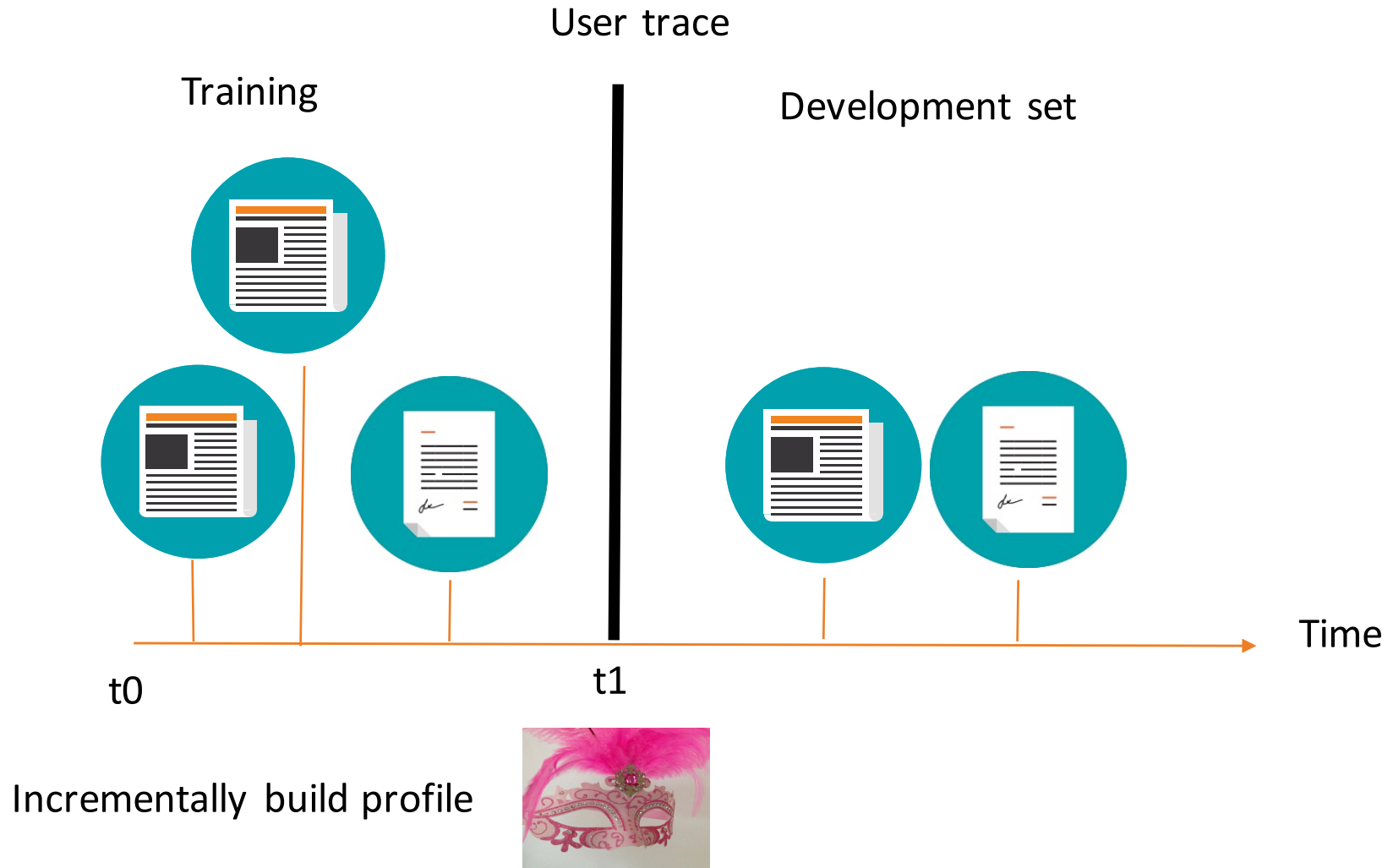
- Streaming algorithms:

“Lossy” counting over streams: count items but *selectively* forget/decay

- Sketching algorithms:

Compact hashing schemes with partial overlap and redundancy

# Offline Evaluation



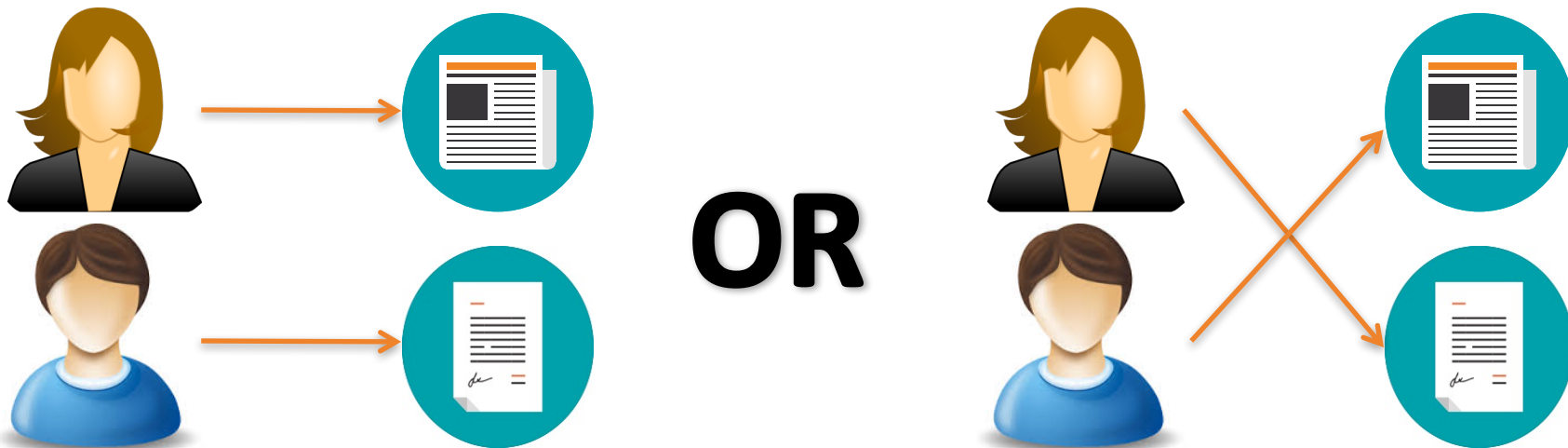
# Offline evaluation: difficulties

- Positive interactions are observed but negative examples are not
- Heavy bias due to publishers' editorial staff



# Offline prediction: Solution

- Pick a pair of users and a pair of docs from their development set
- Does the model prefer actual pairing or counterfactual one?



# Conclusion

- Content based profiling is extremely powerful
- Interesting algorithmic challenges in building dynamic online profiles
- Interesting software architecture challenges for a scalable implementation



Thank you

rani at outbrain dot com,

 @RaniNelken

We're hiring

[www.outbrain.com/about/careers](http://www.outbrain.com/about/careers)