

# Opinion Integration Through Semi-supervised Topic Modeling

**Yue Lu and Chengxiang Zhai**

University of Illinois at Urbana-Champaign



**ILLINOIS**  
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



# Why Opinion Integration?

- Web 2.0 → huge amount of opinions
- What have been said about Hillary Clinton?

**190,451 posts**

**How to digest all?**

**4,773,658 results**

**Facebook**



**Google Blog Search BETA**

**WIKIPEDIA The Free Encyclopedia**

# Two Kinds of Opinions

How to benefit from both?

Expert opinions
<ul style="list-style-type: none"><li>• CNET editor's review</li><li>• Wikipedia article</li></ul>
 <ul style="list-style-type: none"><li>• Well-structured</li><li>• Easy to access</li></ul>
 <ul style="list-style-type: none"><li>• Maybe biased</li><li>• Outdated soon</li></ul>

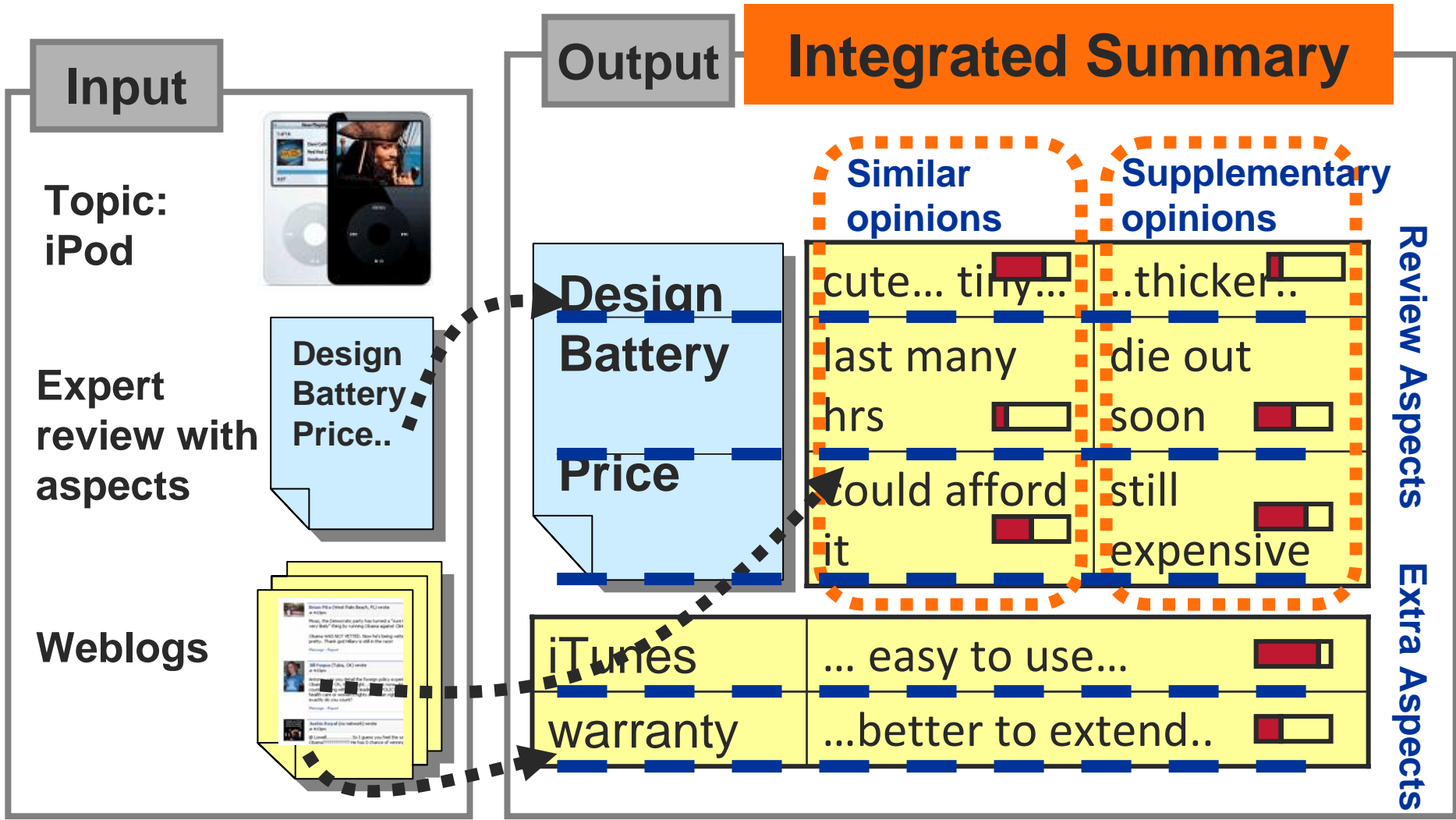
Ordinary opinions
<ul style="list-style-type: none"><li>• Forum discussions</li><li>• Blog articles</li></ul>
 <ul style="list-style-type: none"><li>• fragmental</li><li>• Hard to access</li></ul>
 <ul style="list-style-type: none"><li>• Represent the majority</li><li>• Up to date</li></ul>



# Research Questions



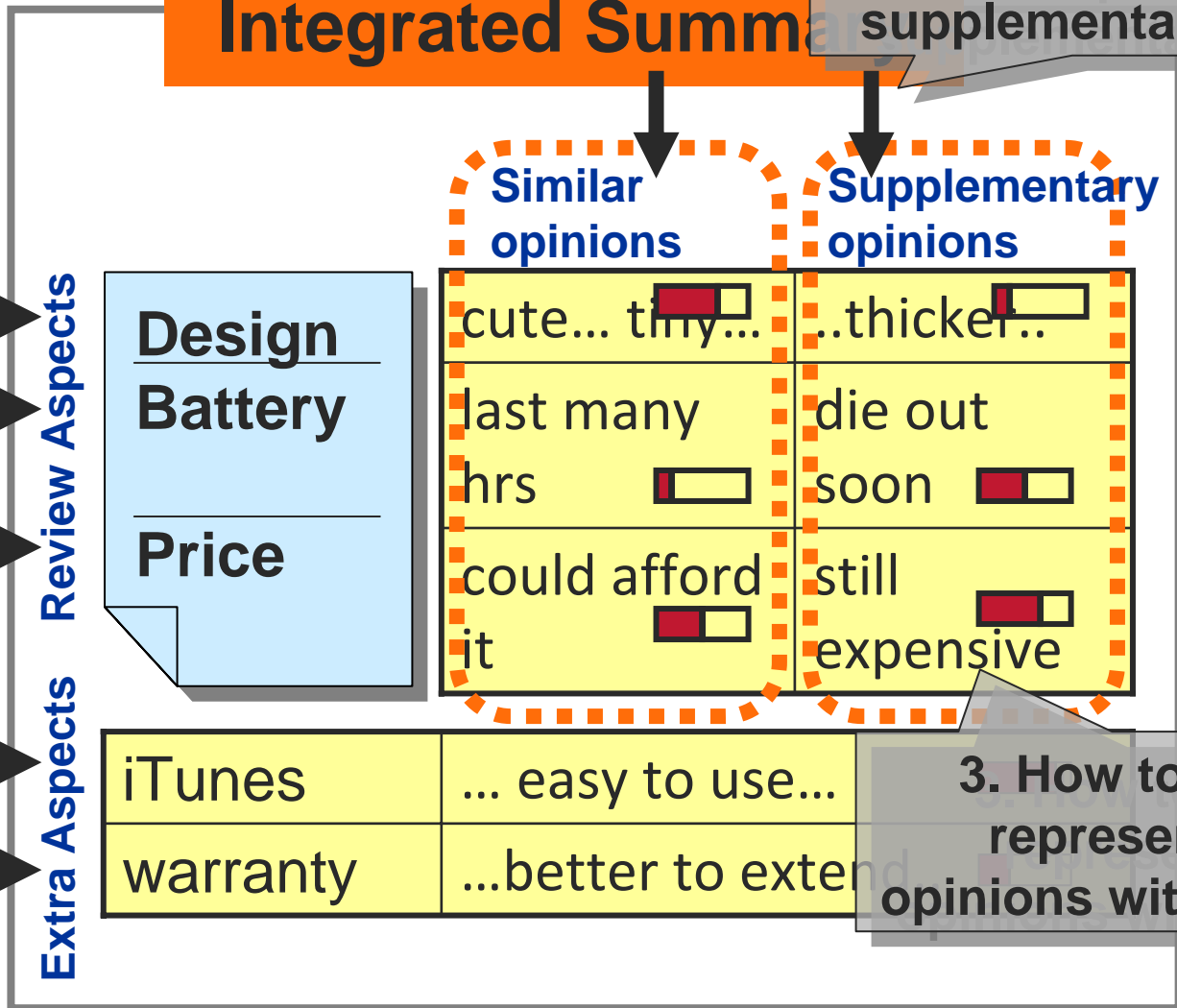
- How do we formalize the problem of opinion integration?
- How do we solve the problem in a general way?
- How do we evaluate it?



1. How to align opinions to expert aspects or extra aspects?

## Integrated Summary

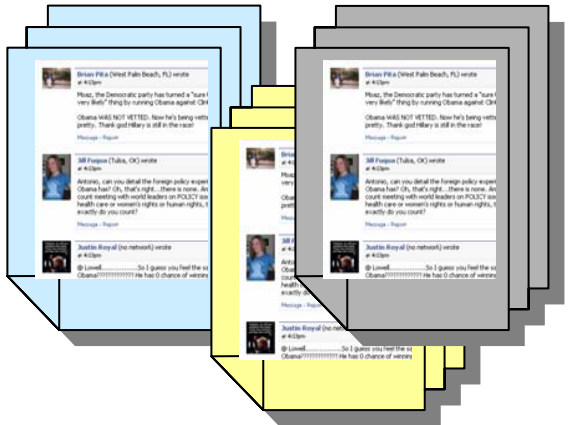
2. How to distinguish similar opinions with supplementary ones?



3. How to extract representative opinions with support?

- Step 1: opinion sentences retrieval

General Weblogs



Weblogs on iPod



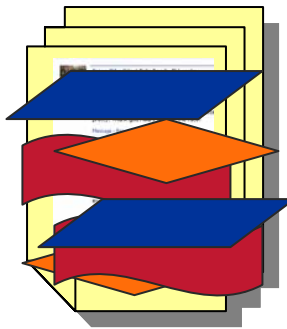
Query = "iPod"

- Step 2: opinion integration using probabilistic topic models (3 subtasks)

# Subtask 1: Aspect Alignment

Align opinion sentences to aspects

Weblogs



Review Aspects

**Design**

**Battery**

**Price**

cute... tiny... ..thicker..

last many hrs die out soon

could afford it still expensive

Extra Aspects

iTunes

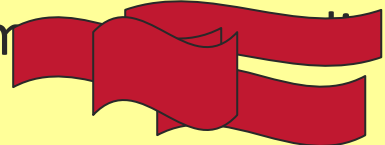


... easy to use...

warranty

...better to extend..

Separate sim opinions from supp ones



Review Aspects  
Extra Aspects

<b>Design</b>	cute... tiny... ..thicker..
	last m  out soon
	could afford it  still expensive
<b>Battery</b>	
<b>Price</b>	
<b>iTunes</b>	 e...
<b>warranty</b>	...better to extend..

# Subtask 3: Opinion Summary

Summarize each block with representative sentences and support


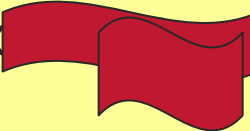


**Representative Opinion (RO)**


- Representative sentence: 
- Support = 3 

Review Aspects

Extra Aspects

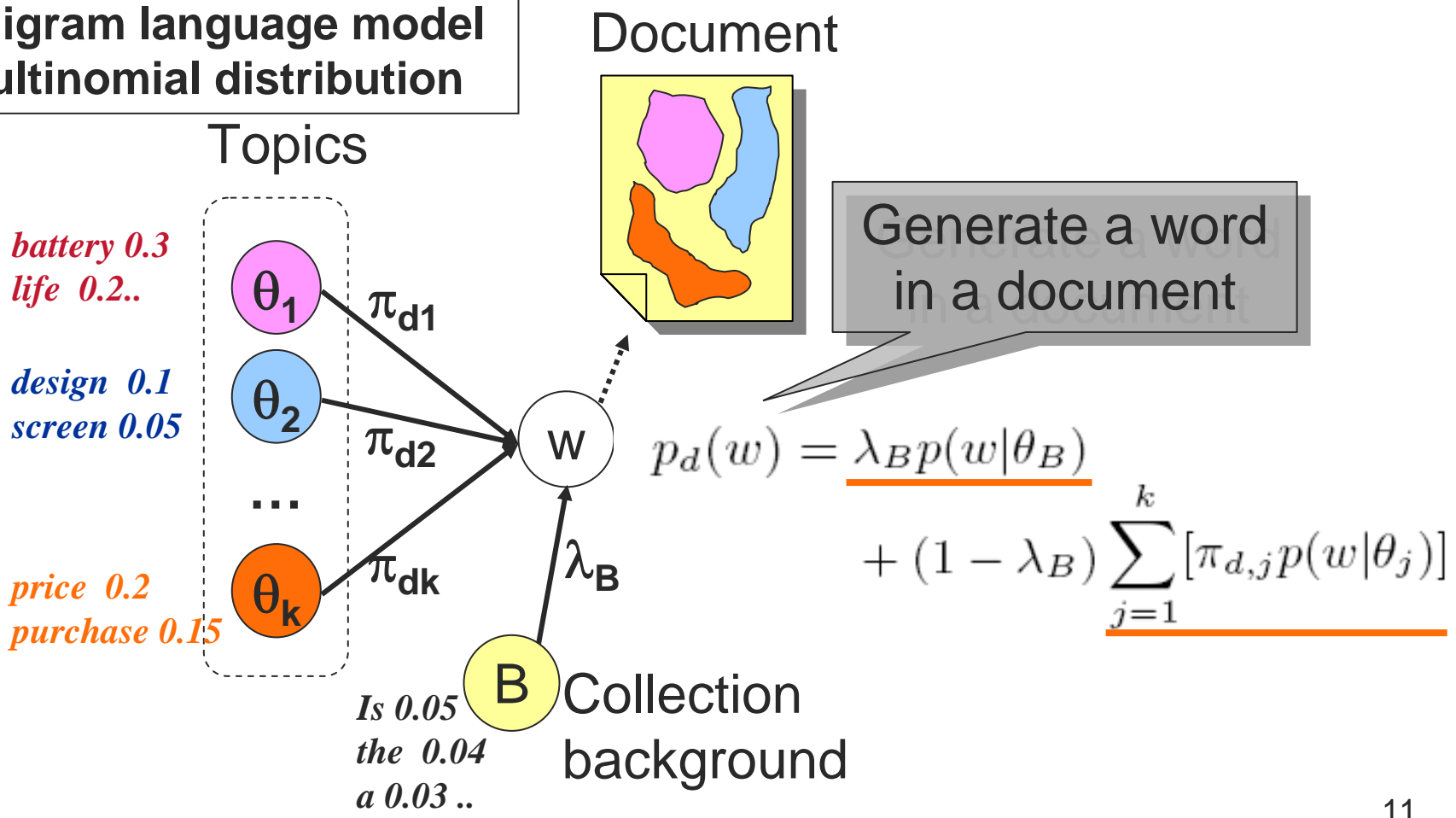
<b>Design</b>
<b>Battery</b>
<b>Price</b>

cute... tiny...	..thicker..
I  y hrs	c  pn
could afford it 	still expensive 

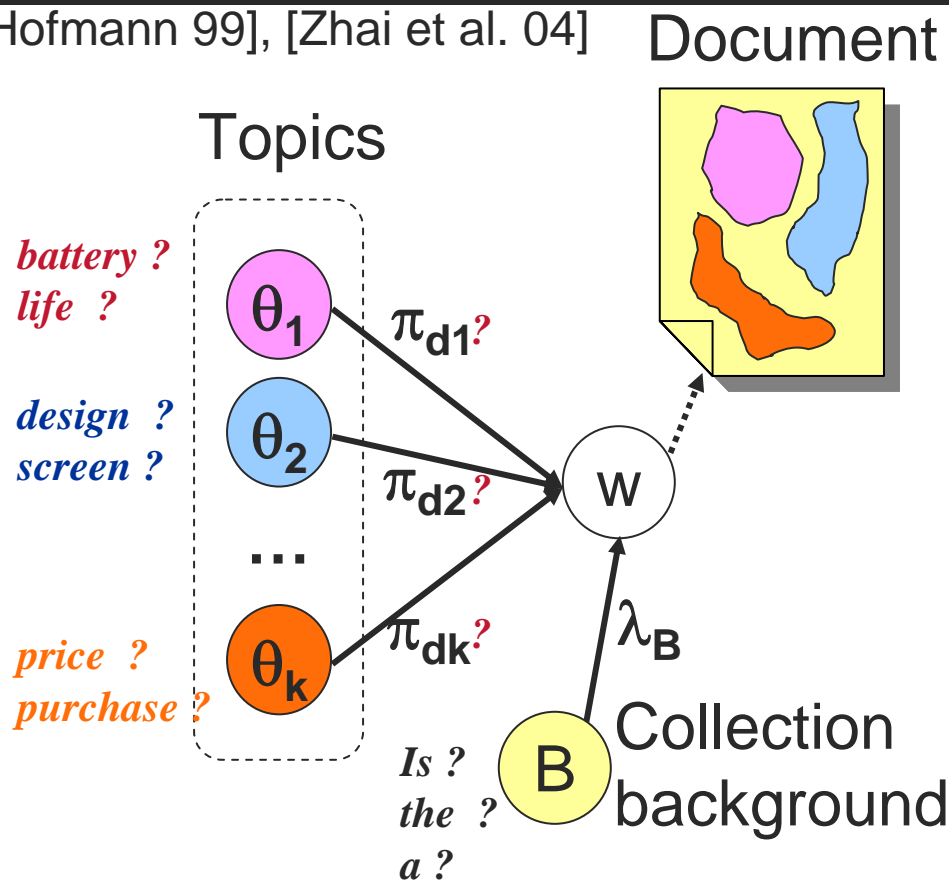
iTunes	... easy 
warranty	...better to extend..

[Hofmann 99], [Zhai et al. 04]

**Topic model**  
= unigram language model  
= multinomial distribution



[Hofmann 99], [Zhai et al. 04]



Generate a word in a document

$$p_d(w) = \lambda_B p(w|\theta_B) + (1 - \lambda_B) \sum_{j=1}^k [\pi_{d,j} p(w|\theta_j)]$$

Log-likelihood of the collection

$$\log p(\mathcal{C}_O | \Lambda) = \sum_{d \in \mathcal{C}_O} \sum_{w \in V} -\{c(w, d) \times \log p_d(w)\}$$

Estimated with Maximum Likelihood Estimator (MLE) through an EM algorithm

$$\hat{\Lambda} = \arg \max_{\Lambda} \log p(\mathcal{C}_O | \Lambda)$$

# Basic PLSA: Problem?

Expert review with aspects

Design..  
Battery..  
Price..  
...

Extracted topics may not align with expert review aspects

**Solution:**  
conjugate priors  
Semi-supervised PLSA

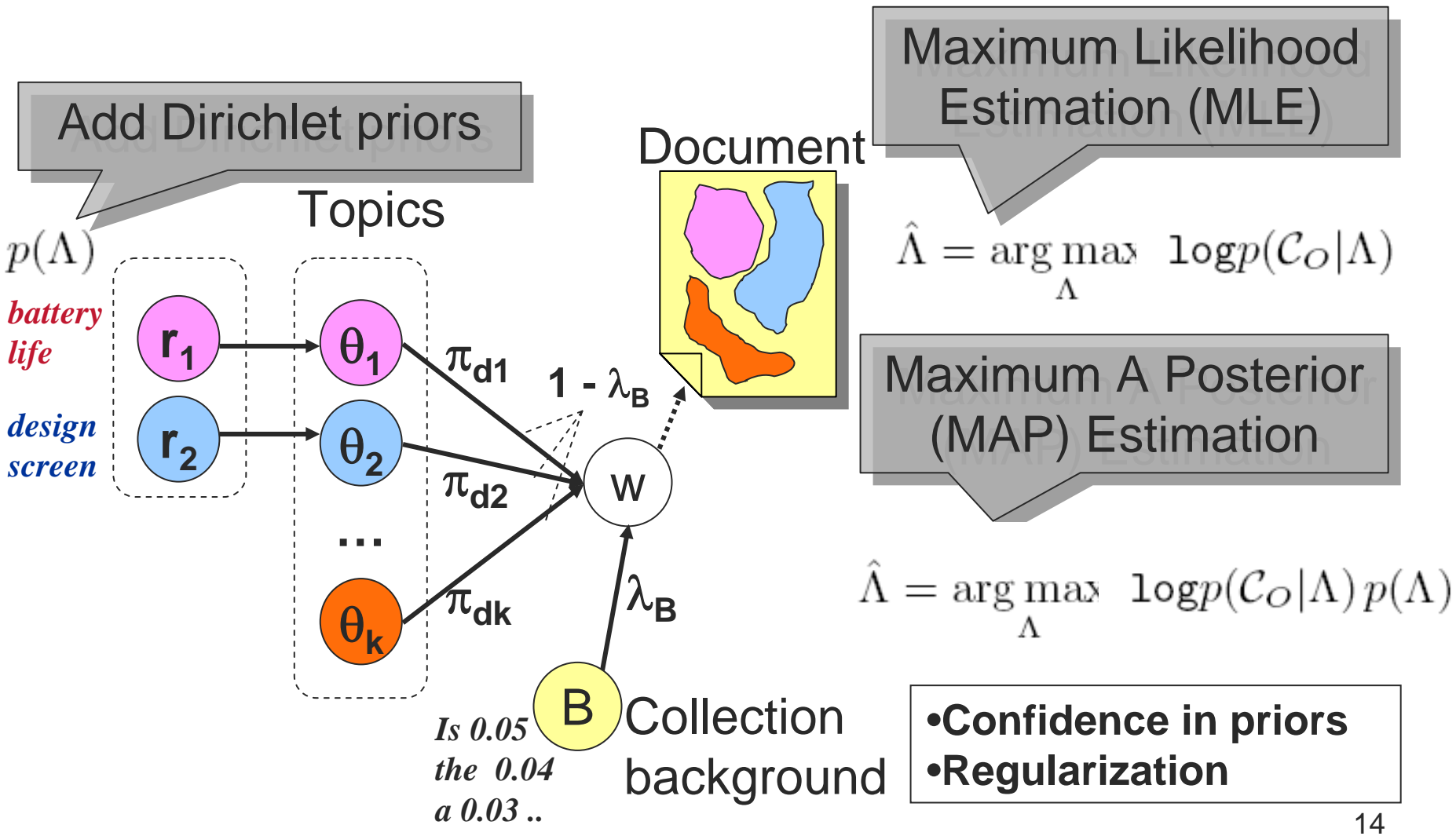
Weblogs on iPod



**Basic PLSA**

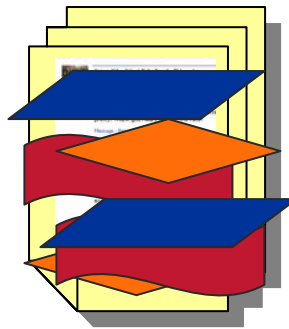


iPod nano  
iPod shuffle  
iPod touch  
...



# Subtask 1: Aspect Alignment

Weblogs



Review Aspects

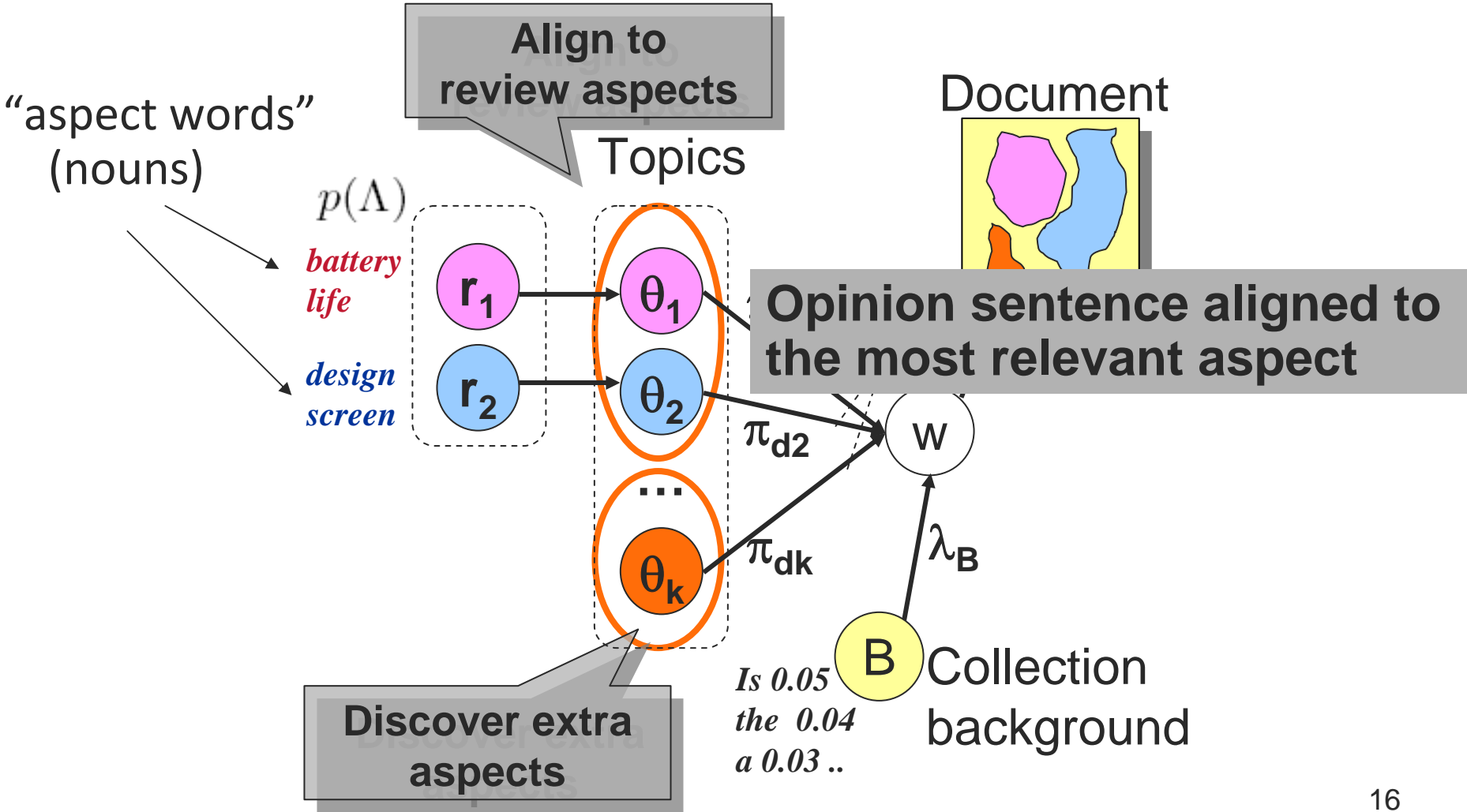
<b>Design</b>
<b>Battery</b>
<b>Price</b>

cute... tiny...	..thicker..
last many hrs	die out soon
could afford it	still expensive

Extra Aspects

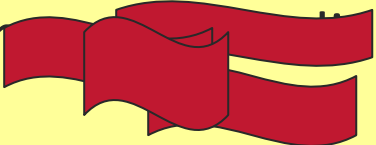


iTunes	... easy to use...
warranty	...better to extend..

# Subtask 1: Aspect Alignment

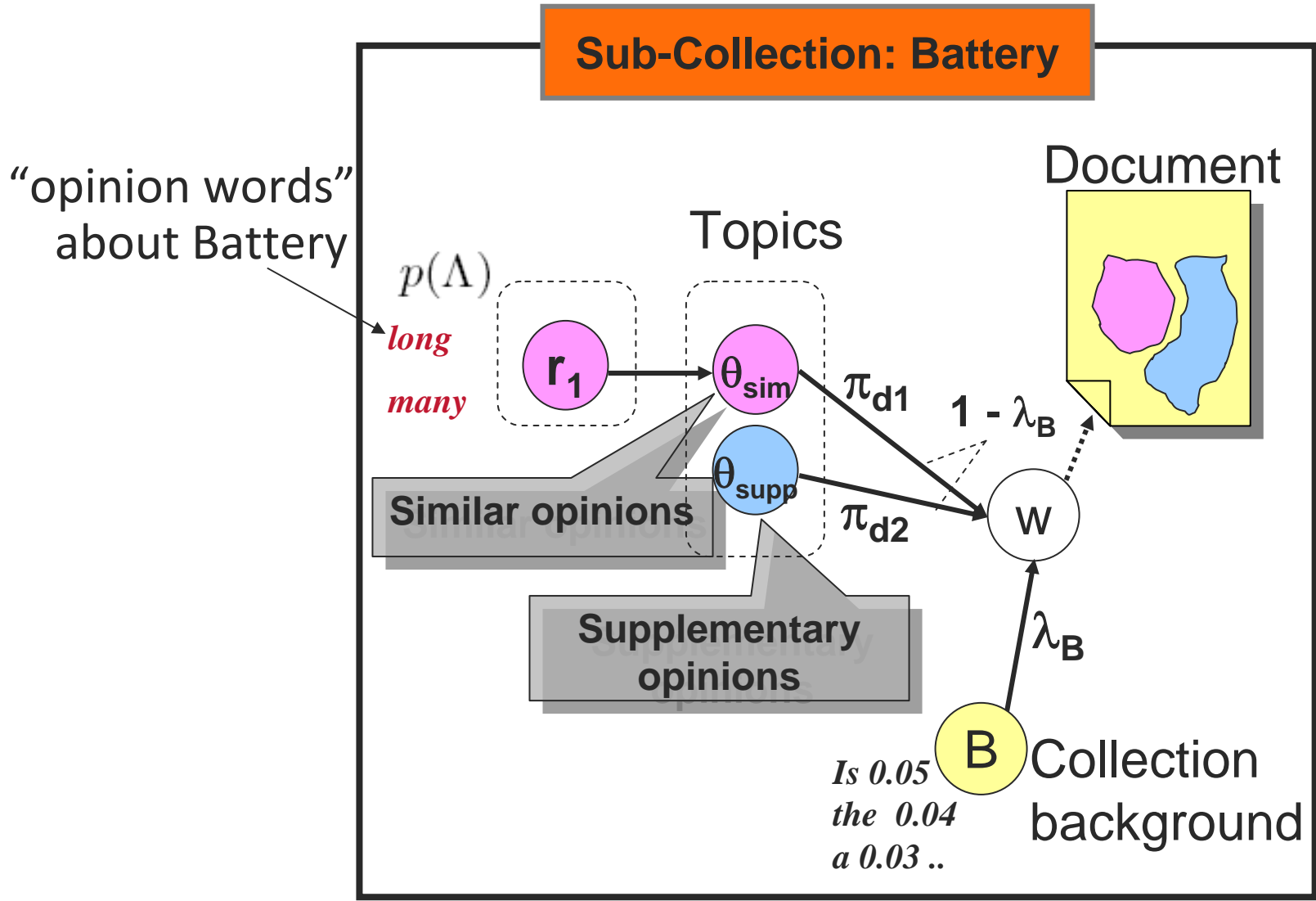


# Subtask 2: Opinions Separation

Review Aspects  
Extra Aspects



<b>Design</b>	cute... tiny... ..thicker..
	last m  out soon
	could afford it  still expensive
<b>Battery</b>	
<b>Price</b>	
<b>iTunes</b>	 e...
<b>warranty</b>	...better to extend..

# Subtask 2: Opinions Separation



# Subtask 3: Opinion Summary


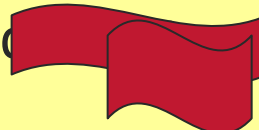


**Representative Opinion (RO)**


- Representative sentence: 
- Support = 3 

Review Aspects

Extra Aspects

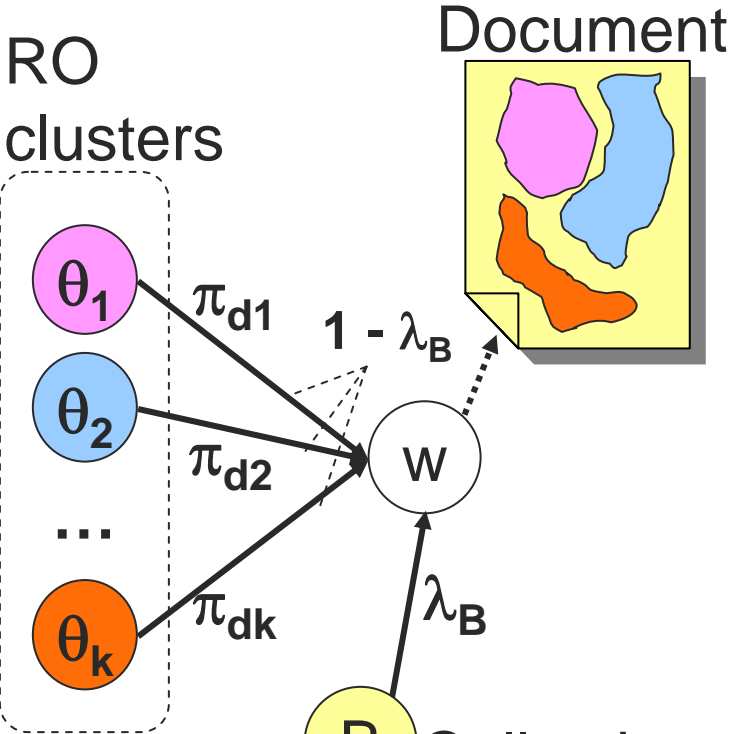
<b>Design</b>
<b>Battery</b>
<b>Price</b>

cute... tiny...	..thicker..
I  y hrs	c  pn
could afford it 	still expensive 

iTunes	... easy 
warranty	...better to extend..

# Subtask 3: Opinion Summary

## Sub-Collection: A Block



Centroid Sentence 1 ←

Centroid Sentence 2 ←

...

Support = cluster size

*Is 0.05  
the 0.04  
a 0.03 ..*

- Expert review data:

Topic	Source	# words	# aspects
iPhone	CNET	4434	19
Barack Obama	Wikipedia	312	14

- Ordinary opinion data:

Topic	Query Terms	# articles	# sentences
iPhone	iPhone	552	3000
Barack Obama	Barack+Obama	639	1000

- Opinion Integration with review aspects

Review article	Similar opinions	Supplementary opinions
<p>You can make emergency calls, but you can't use any other functions...</p>	N/A	<p>... methods for <b>unlocking</b> the iPhone have emerged on the past few weeks, involve tinkering with the iPhone hardware...</p>
<p>rated battery life <b>hours talk time, 24 hours of music</b> playback, <b>7 hours of video</b> playback, and 6 hours on Internet use.</p>	<p>Up to <b>8 Hours of Talk Time</b>, 6 Hours of Internet Use, <b>7 Hours of Video</b> Playback or <b>24 Hours of Audio</b> Playback</p>	<p>Playing relatively high bitrate VGA H.264 videos, our iPhone lasted almost exactly <b>9 freaking hours</b> of continuous playback with cell and WiFi on (but Bluetooth...</p>

Activation

Confirm the opinions from the review

Unlock/hack iPhone

Battery

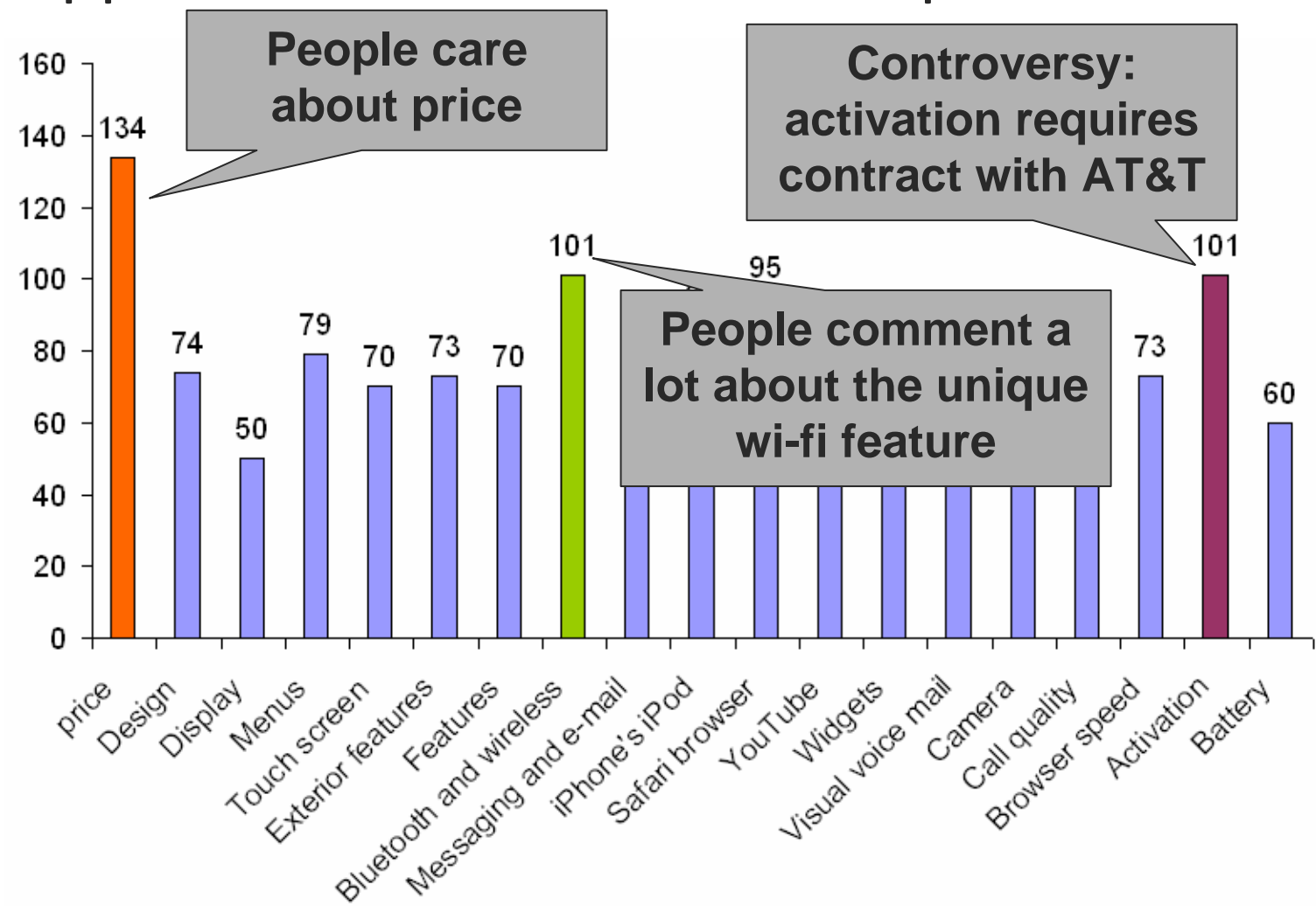
Additional info under real usage

- Opinions on extra aspects

support	Supplementary opinions on extra aspects
15	<p>You may have heard of <b>iASign</b> — an iPhone app that allows you to <b>activate</b> your phone with iTunes rigamarole.</p> <p><b>Another way to activate iPhone</b></p>
13	<p><b>Cisco</b> has owned the <b>trademark</b> on the name "<b>iPhone</b>" since 2000, when it acquired InfoG... which originally registered the name</p> <p><b>iPhone trademark originally owned by Cisco</b></p>
13	<p>With the imminent availability of the iPhone, a look at 10 things current smartphones like the <b>Nokia N95</b> have been able to do that the <b>iPhone can't currently match...</b></p> <p><b>A better choice for smart phones?</b></p>

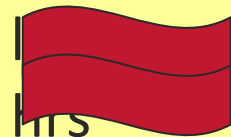
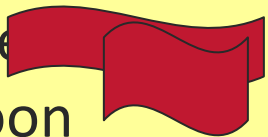
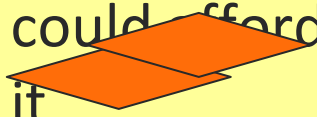

# Results: Product (iPhone)


- Support statistics for review aspects



- Goal
  - Evaluate human agreement (**how hard is opinion integration?**)
  - Evaluate how our approach could reproduce human choice (**how well is our method doing?**)
- Method
  - Ask 3 users to perform 3 tasks
  - Tasks designed from the Obama example

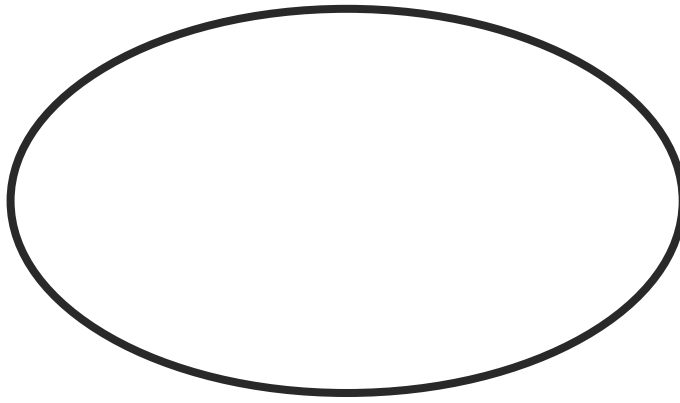
# Task 1: Distinguish Extra Aspects

<b>Design</b>	cute... tiny...	..thicker..
	I  y hrs	die  soon
	could afford it 	still  expens

iTunes	... easy 
warranty	better to extend..

7 extra aspects

34 opinions



- Result
  - Low human agreement (1/7)
  - Our method recovers 52.4% of user choices on avg

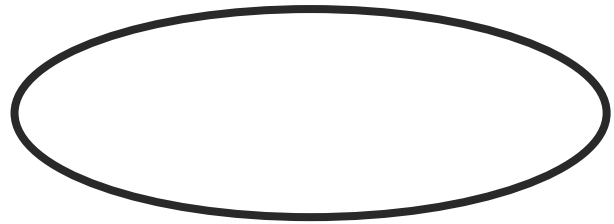
# Task 2: Aspect Alignment

- Mix 27 opinions
- Label each with one of 14 aspects

**Review Aspects**

<b>Design</b>	cute... tiny...	..thicker..
<b>Battery</b>	last many hrs	die out soon
<b>Price</b>	could afford it	still expensive

## Results:



- Users agree on 13/27 = 48% sentences
- Our method recovers 10.67/27 = 40% sentences on avg.

Review Aspects

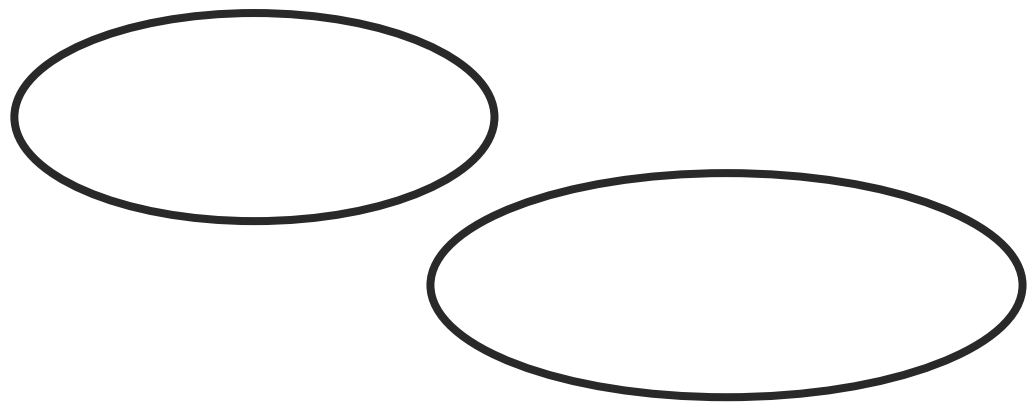
**Design**  
**Battery**

---

**Price**

cute... tiny...	..thicker..
I y hrs	die soon
could afford it	still expens

- Mix one sim opinion with many supp opinions
- Select one opinion most similar to the review opinion
- Result: recovers 60% of human choice



- **Novel problem:** opinion integration
- **Unified approach:** semi-supervised probabilistic topic modeling
- Many potential **interesting applications**
- Future Work
  - More rigorous evaluation
  - More general setup: many expert reviews instead of one

# Thank you!



ILLINOIS  
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

