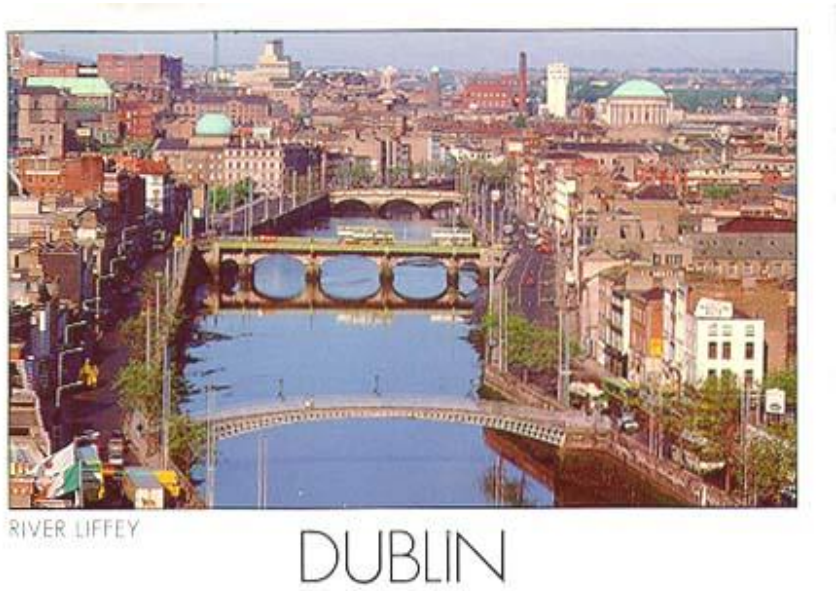
Distributional Footprints of Deceptive Product Reviews

Song Feng, Longfei Xing, Anupam Gogar and Yejin Choi

Stony Brook University, Stony Brook, NY

Motivation

Motivation

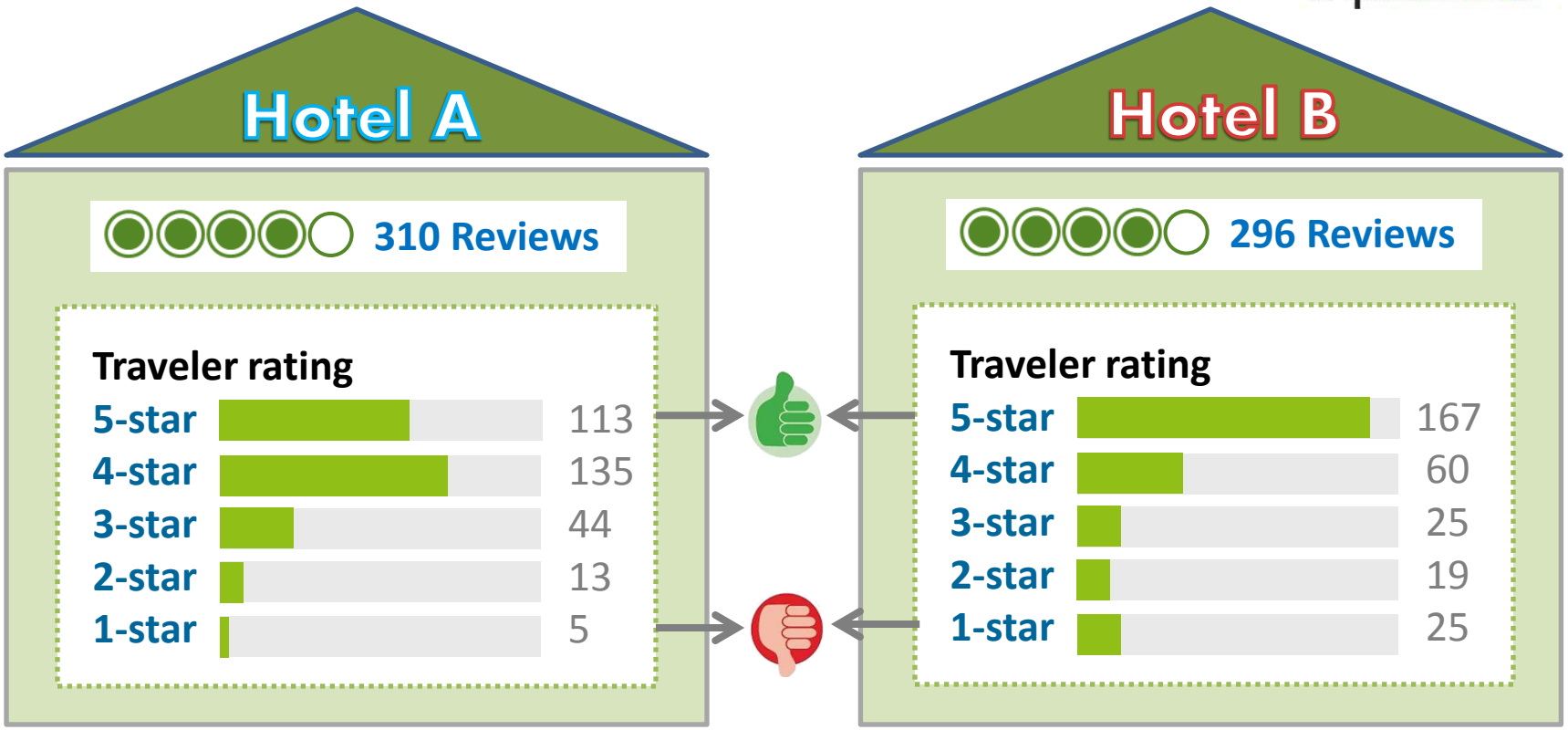


Motivation


Hotel A
310 Reviews
Traveler rating
5-star 113
4-star 135
3-star 44
2-star 13
1-star 5


Hotel B
296 Reviews
Traveler rating
5-star 167
4-star 60
3-star 25
2-star 19
1-star 25

Motivation


Hotel A
310 Reviews
Traveler rating
5-star 113
4-star 135
3-star 44
2-star 13
1-star 5
Hotel B
296 Reviews
Traveler rating
5-star 167
4-star 60
3-star 25
2-star 19
1-star 25

Motivation


Hotel A
310 Reviews
Traveler rating
5-star 113
4-star 135
3-star 44
2-star 13
1-star 5


Hotel B
296 Reviews
Traveler rating
5-star 167
4-star 60
3-star 25
2-star 19
1-star 25

Motivation


Hotel A
310 Reviews
Traveler rating
5-star 113
4-star 135
3-star 44
2-star 13
1-star 5


Hotel B
296 Reviews
Traveler rating
5-star 167
4-star 60
3-star 25
2-star 19
1-star 25

Motivation


Hotel A
310 Reviews
Traveler rating
5-star 113
4-star 135
3-star 44
2-star 13
1-star 5
4-star > # 5-star


Hotel B
296 Reviews
Traveler rating
5-star 167
4-star 60
3-star 25
2-star 19
1-star 25

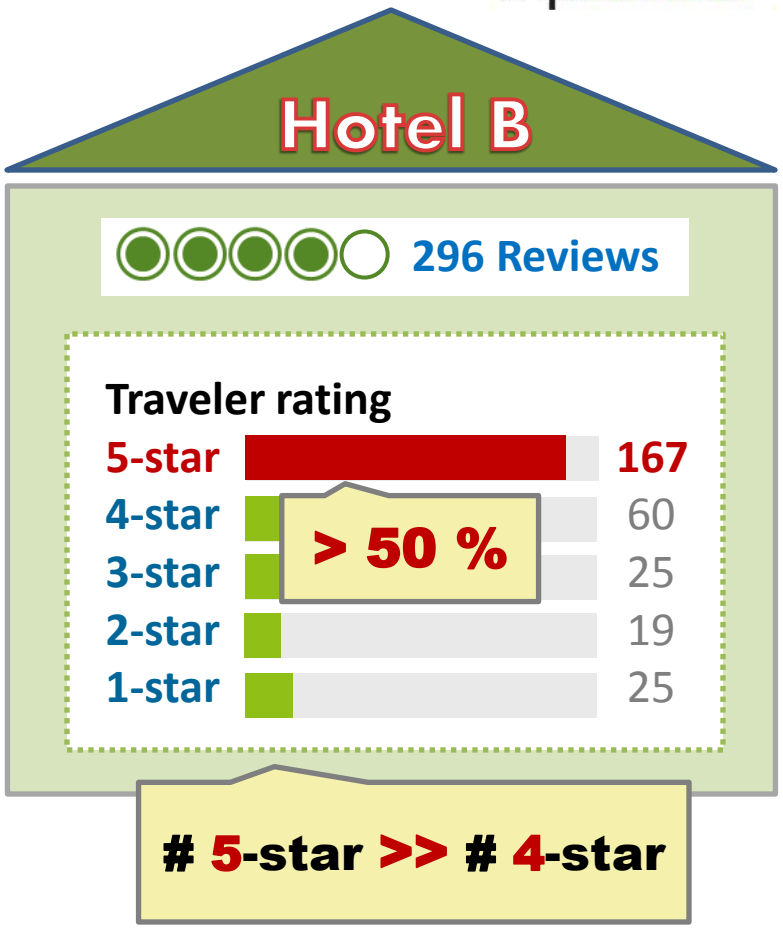
Motivation


Hotel A
310 Reviews
Traveler rating
5-star 113
4-star 135
3-star 44
2-star 13
1-star 5
4-star > # 5-star


Hotel B
296 Reviews
Traveler rating
5-star 167
4-star 60
3-star 25
2-star 19
1-star 25
5-star >> # 4-star

Motivation


Hotel A
310 Reviews
Traveler rating
5-star 113
4-star 135
3-star 44
2-star 13
1-star 5
4-star > # 5-star


Hotel B
296 Reviews
Traveler rating
5-star 167
4-star 60
3-star 25
2-star 19
1-star 25
> 50 %
5-star >> # 4-star

Motivation


Hotel A
310 Reviews
Traveler rating
5-star 113
4-star 135
3-star 44
2-star 13
1-star 5


Hotel B
296 Reviews
Traveler rating
5-star 167
4-star 60
3-star 25
2-star 19
1-star 25

Motivation

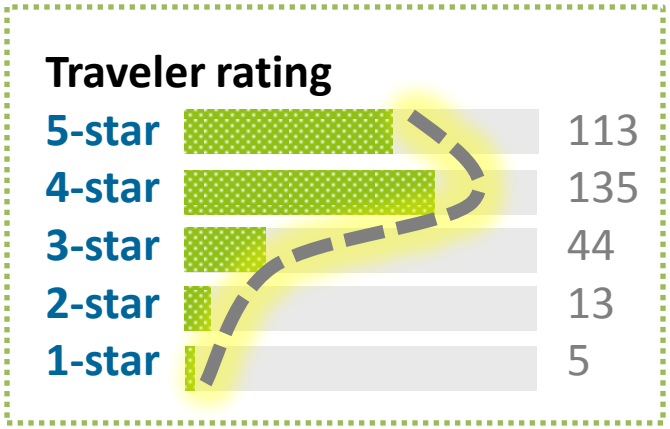

Hotel A
310 Reviews
Traveler rating
5-star 113
4-star 135
3-star 44
2-star 13
1-star 5


Hotel B
296 Reviews
Traveler rating
5-star 167
4-star 60
3-star 25
2-star 19
1-star 25

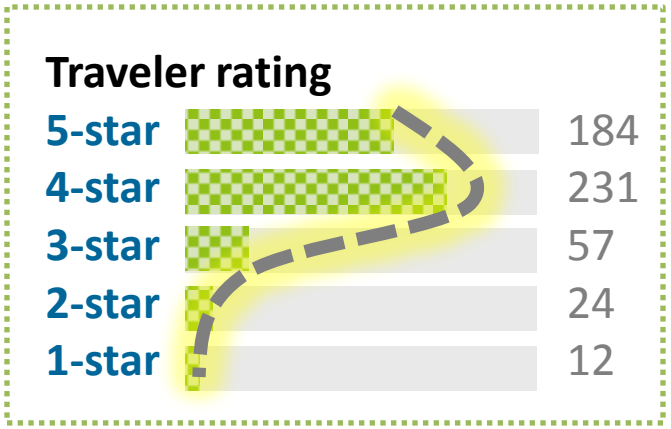
Motivation

tripadvisor®

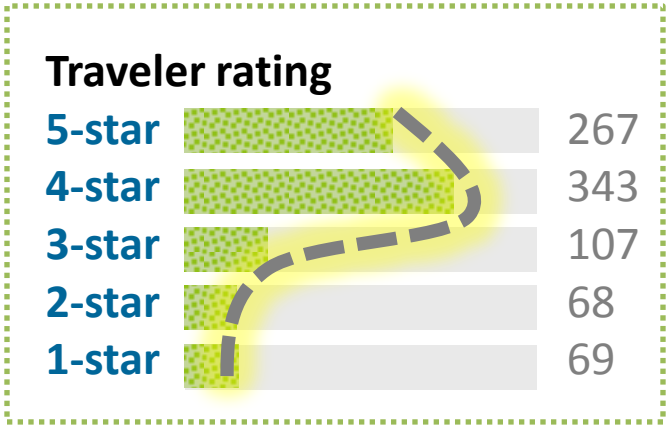
(1)



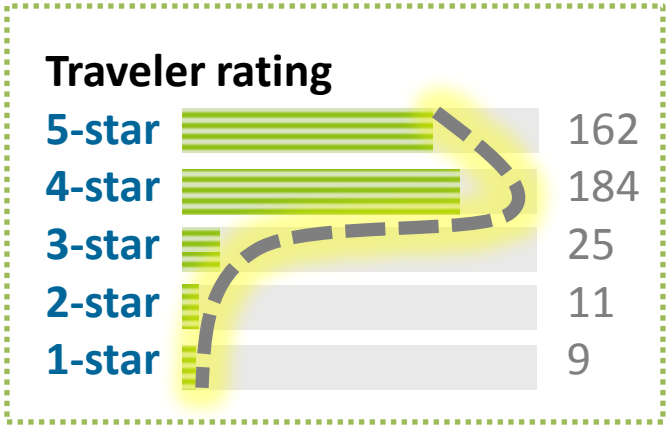
(2)



(3)

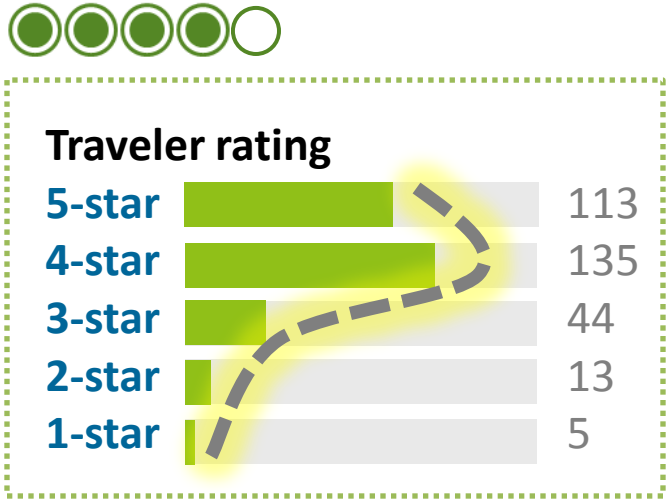


(4)

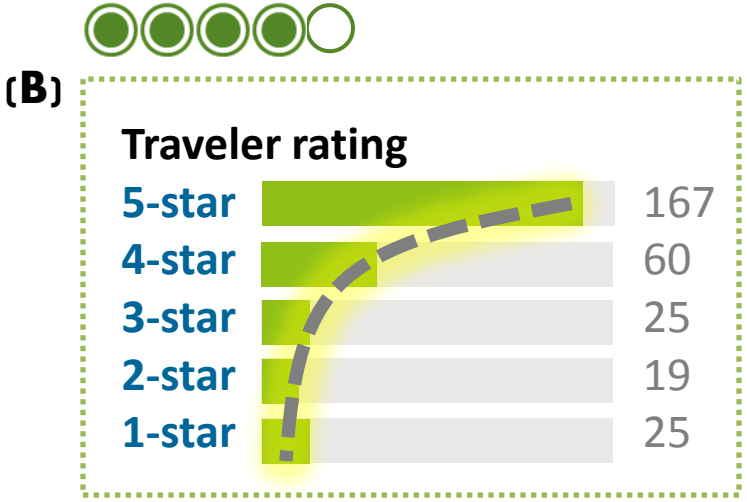


Motivation

(A)

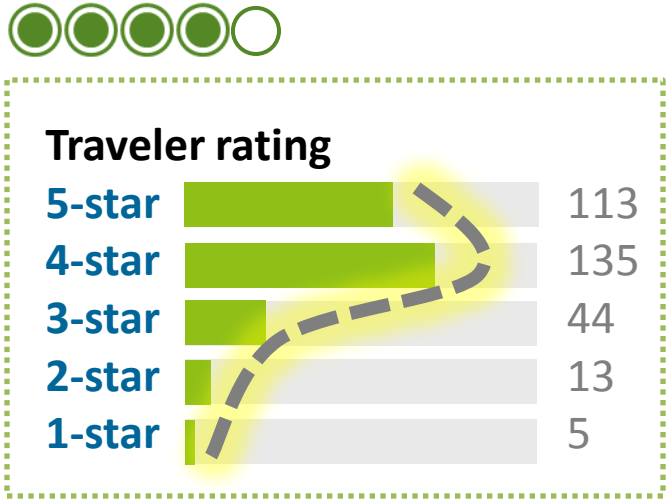


(B)

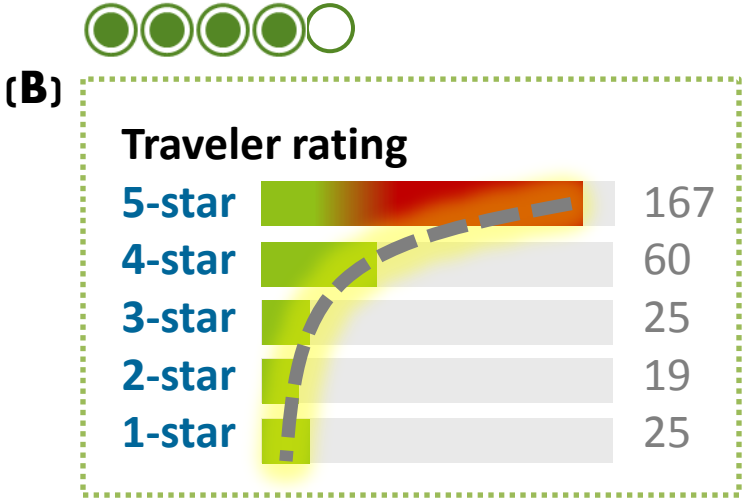


Motivation

(A)

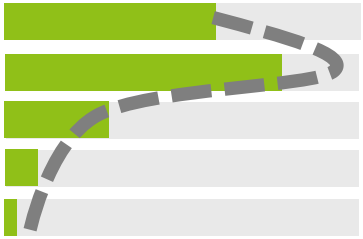


(B)

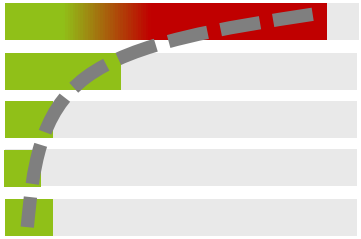


Motivation

- **Natural**



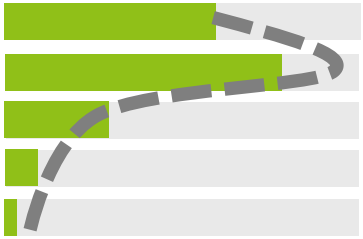
- **Distorted**



Motivation

Rating distributions Deceptive reviews

- **Natural**



- **Distorted**

Data

Product reviews – amazon.com

- Meta data (by *Jindal and Liu*, 2008).

Hotel reviews – tripadvisor®

- Meta data
- Review Text

Data

Hotels: 4000 hotels,

21 cities,

English-speaking countries.

Period: 2007-2011.

Reviews: 840,000 in total.

Data

Reviewers: **Single-time** Reviewers

Multi-time Reviewers

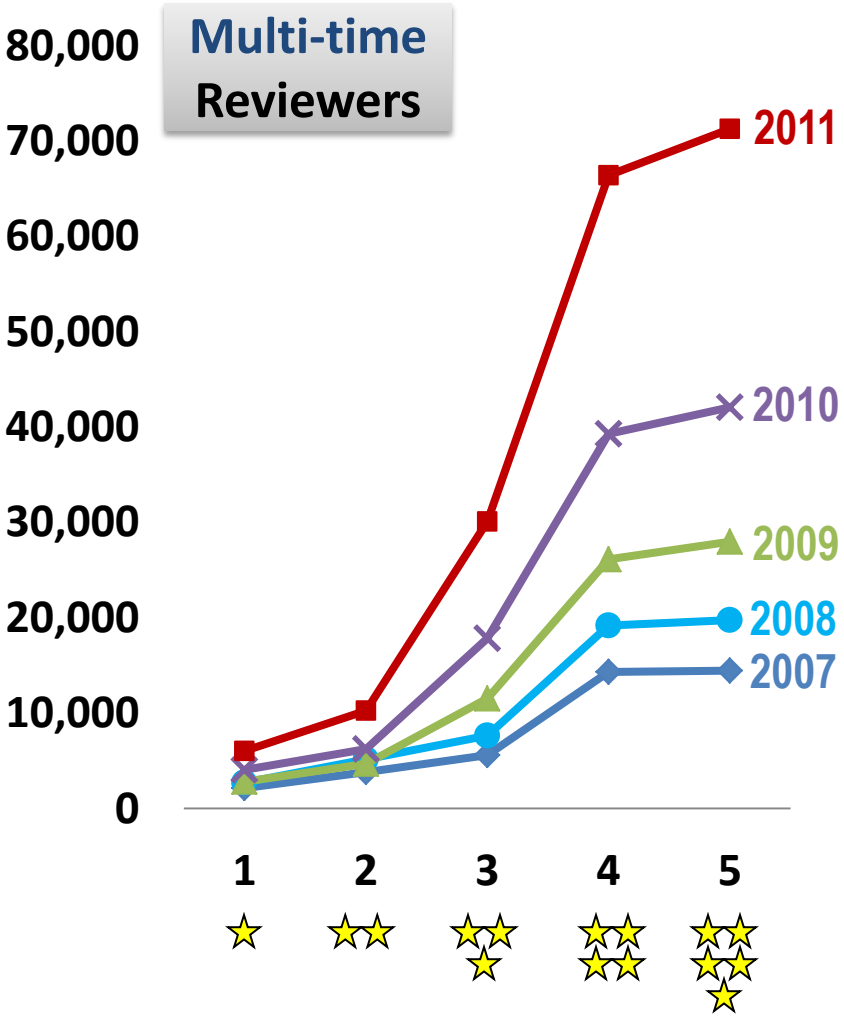
Data

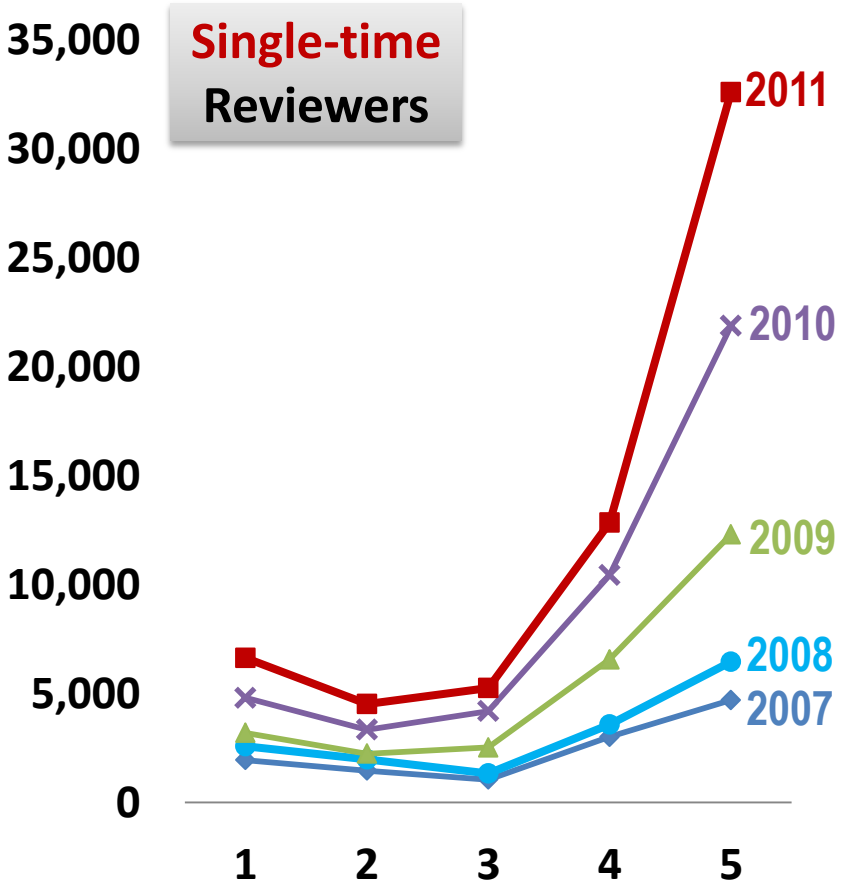
Reviewers: **Single-time** Reviewers

Multi-time Reviewers

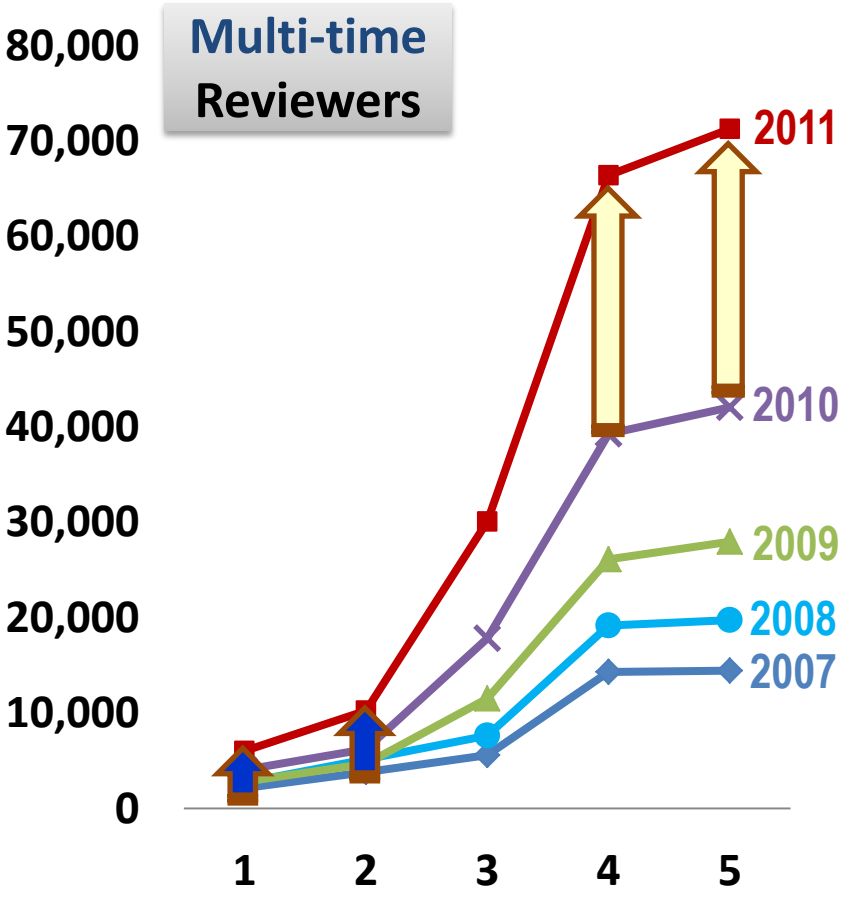
Ott et al., WWW 2012

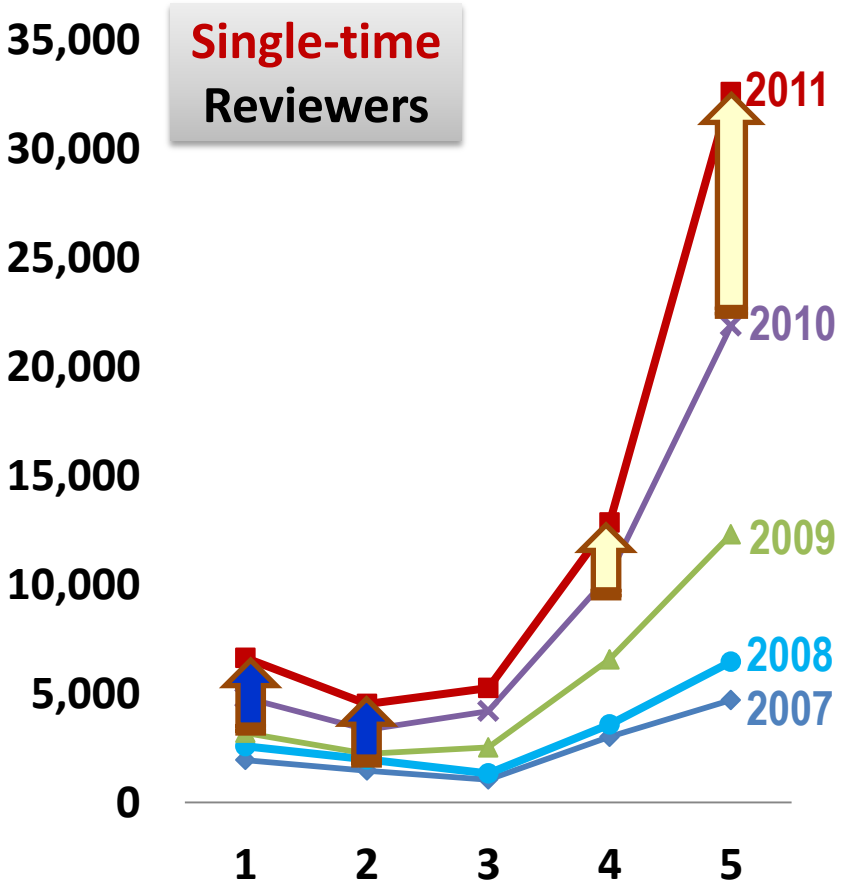
Rating Distribution (Yearly)



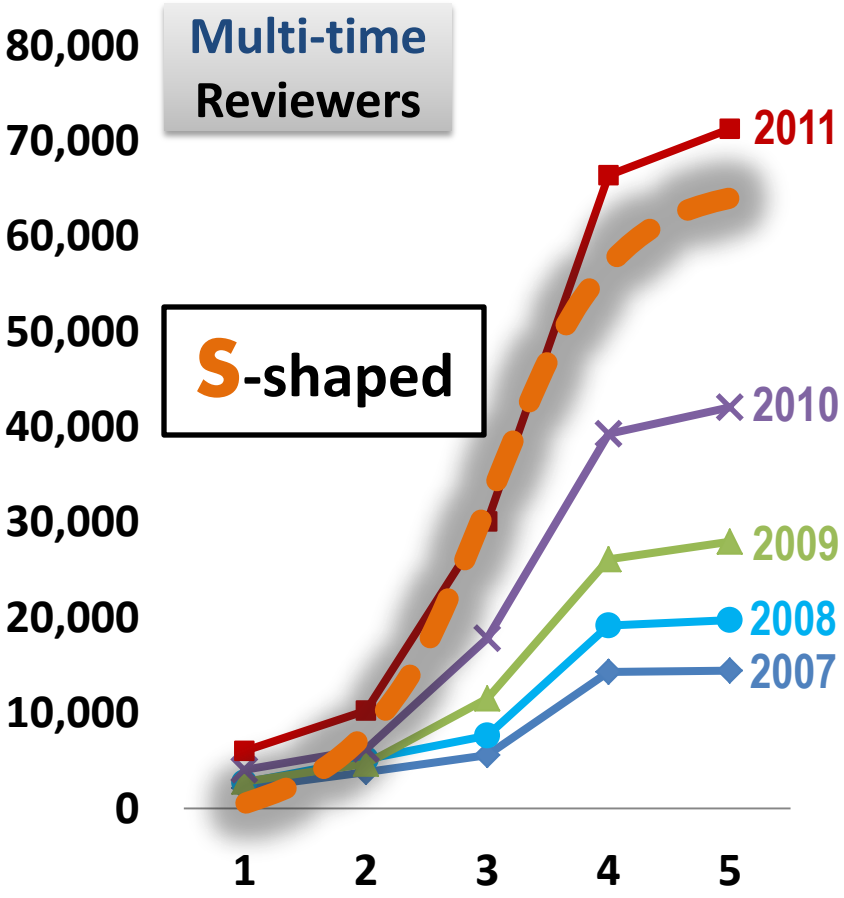


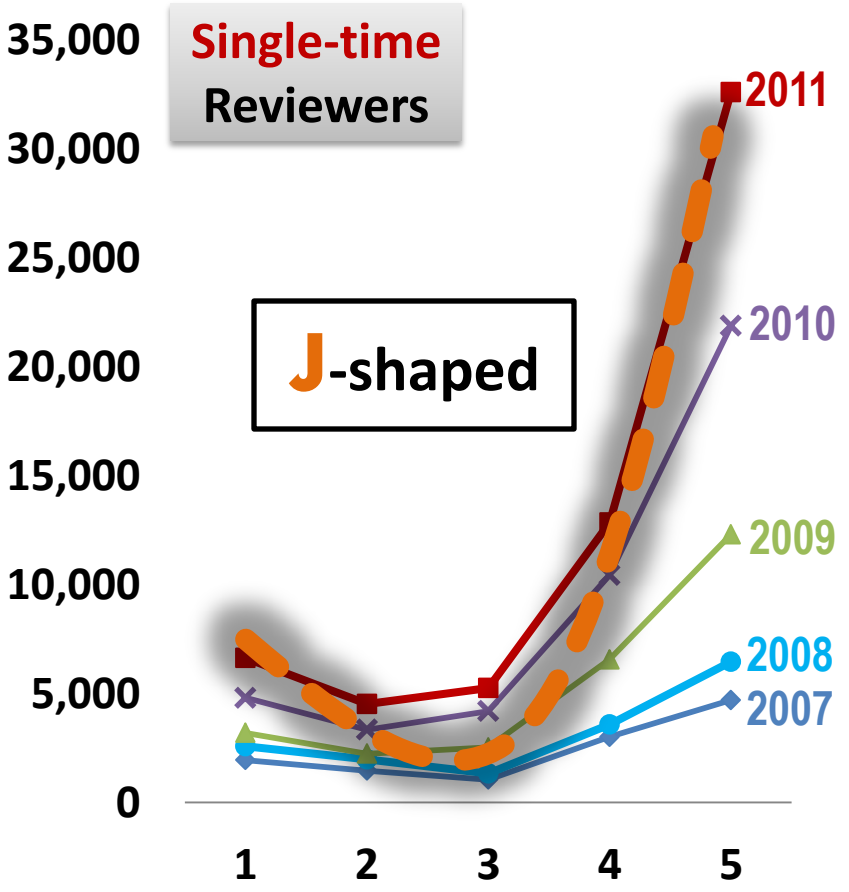
Rating Distribution (Yearly)





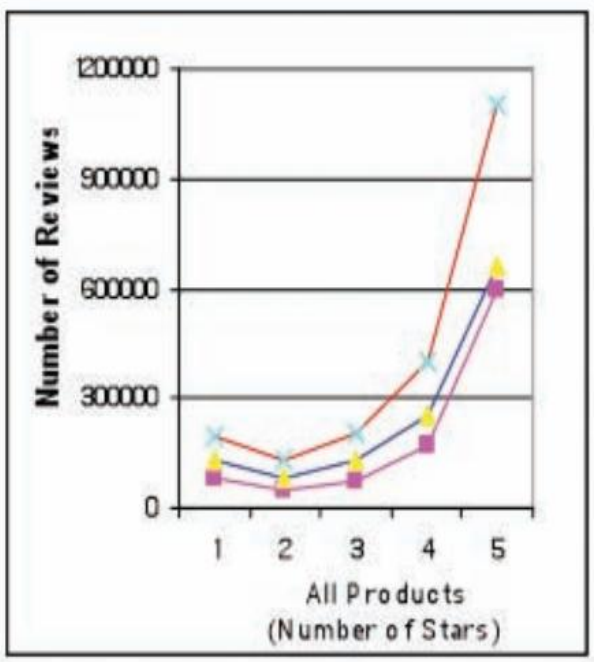
Rating Distribution (Yearly)

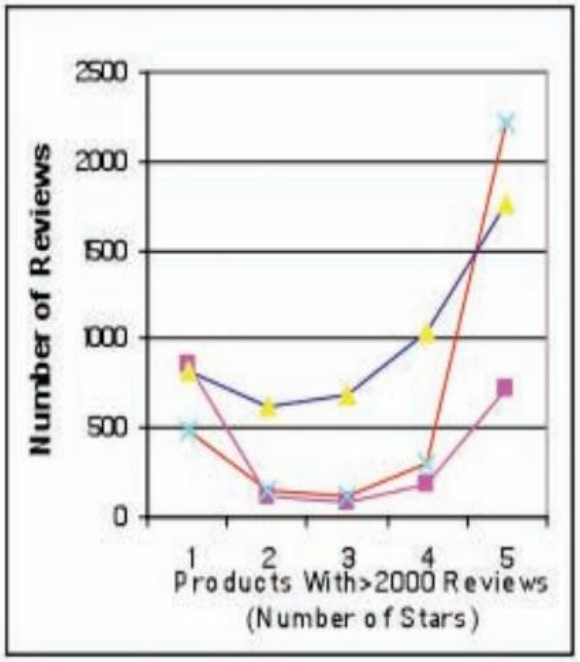


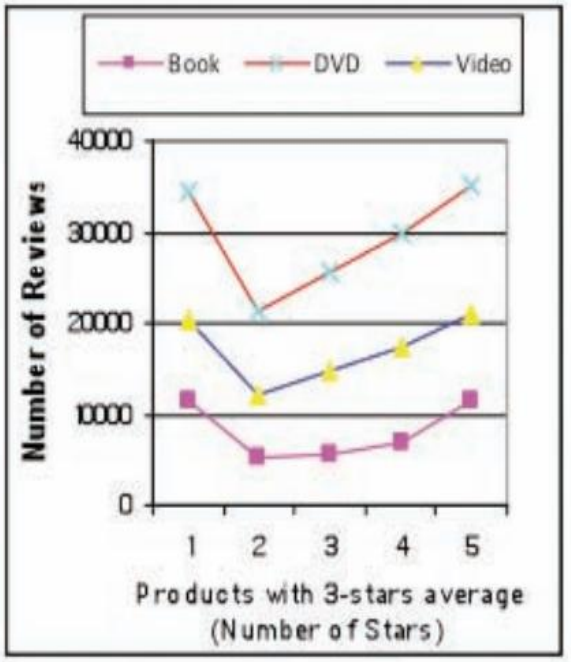


Related Work I

- "**J-shape**" distribution of product reviews
  Hu et al. (2009)







Related Work II

- Spam review detection
  -- **rather than fake** reviews
  -- spam reviews: obvious advertisement,
  often completely irrelevant information

Jindal and Liu (2007, 2008)
Liu et al (2007)
Jindal et al. (2010)
Lim et al. (2010)

Related Work III

- Fake review detection
  -- validation based on human labeling
  ➔ **susceptible to human errors in telling apart real reviews and fake reviews!**
  (Ott et al. report < 62% accuracy of human judges if judging based only on the content of the review)

  G. Wu et al. (2010)
  Jindal et al. (2010)
  Lim et al. (2010)
  Mukherjee et al. (2011, 2012)

Related Work IV

***Ott et al.*, 2011 @ ACL**

- Created reliable "gold standard data" for the first time
  - 400 ***manufactured*** fake reviews using Amazon Mechanical Turk
  - 400 collected (downloaded) truthful reviews
- Nearly **90**% accuracy via *supervised learning*
- Based on the linguistic content of the reviews.

Three Contributions

1. Characterization of rating distributions
   ➔ **Natural** vs. **distorted** rating distributions
2. Detection strategies to identify deceptive business entities & reviews
3. Novel evaluation methodologies.
   - ❖ Avoid human judges
     (because they are not good at catching fakes)
     -- Ott et al. 2011 report human accuracy ~ 60%
   - ❖ Avoid manufacturing fake reviews
     (because they are costly)

Three Contributions

1. **Characterization of rating distributions**
   - ➔ **Natural vs. distorted rating distributions**
2. Detection strategies to identify deceptive business entities & reviews
3. Novel evaluation methodologies.
   - ❖ Avoid human judges
     (because they are not good at catching fakes)
     -- Ott et al. 2011 report human accuracy ~ 60%
   - ❖ Avoid manufacturing fake reviews
     (because they are costly)

Rating Distribution

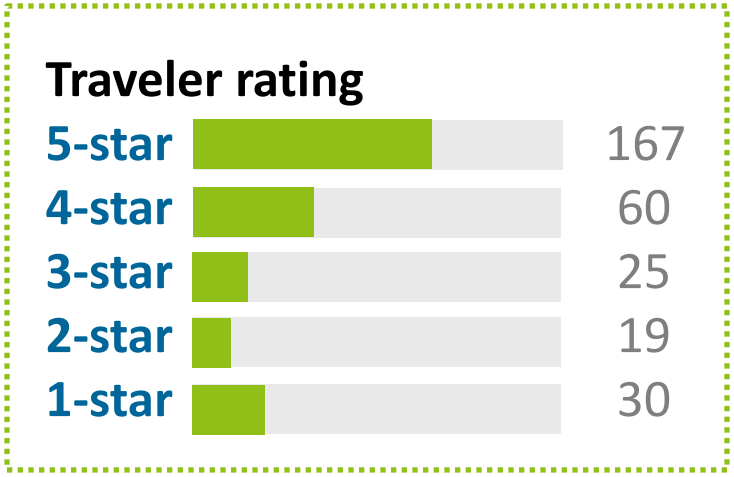
Shape of distribution

$\widehat{D}$

Rating Distribution

Shape of distribution

$\widehat{D}$

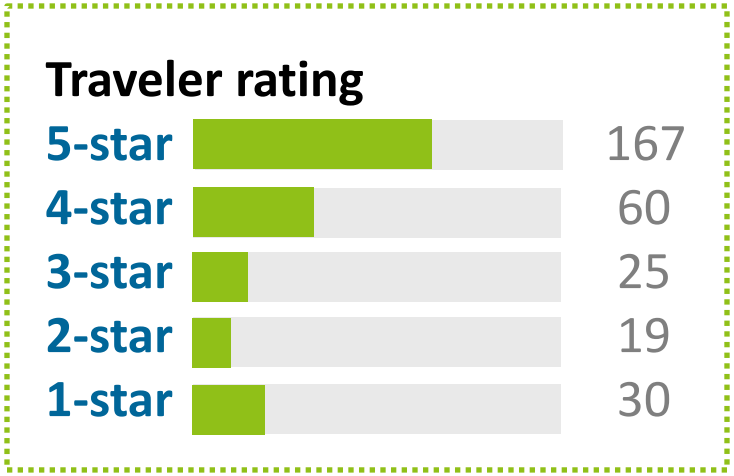


Rating Distribution

Shape of distribution

$\widehat{D}$:= rating scores sorted (in descending order)

by # of corresponding counts

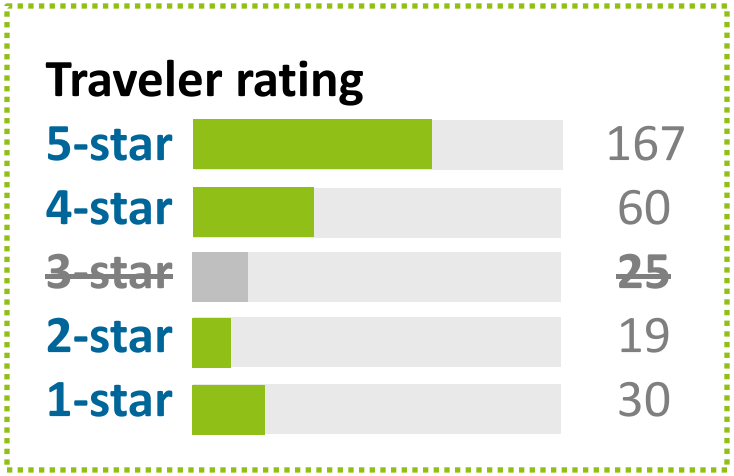


Rating Distribution

Shape of distribution

$\widehat{D}$:= rating scores sorted (in descending order)

by # of corresponding counts

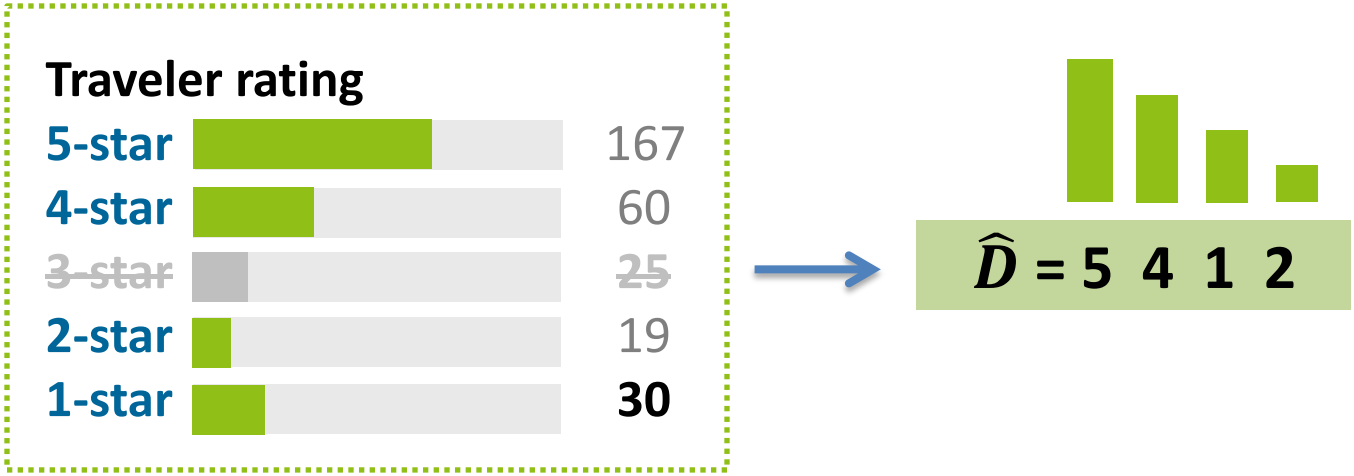


Rating Distribution

Shape of distribution

$\widehat{D}$:= rating scores sorted (in descending order)

by # of corresponding counts

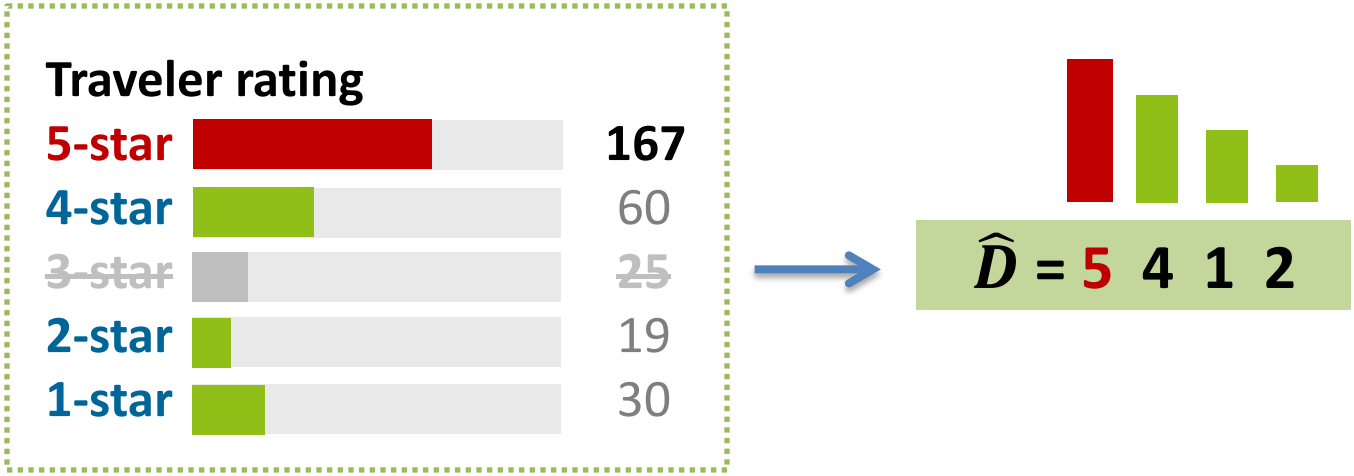


Rating Distribution

Shape of distribution

$\widehat{D}$:= rating scores sorted (in descending order)

by # of corresponding counts

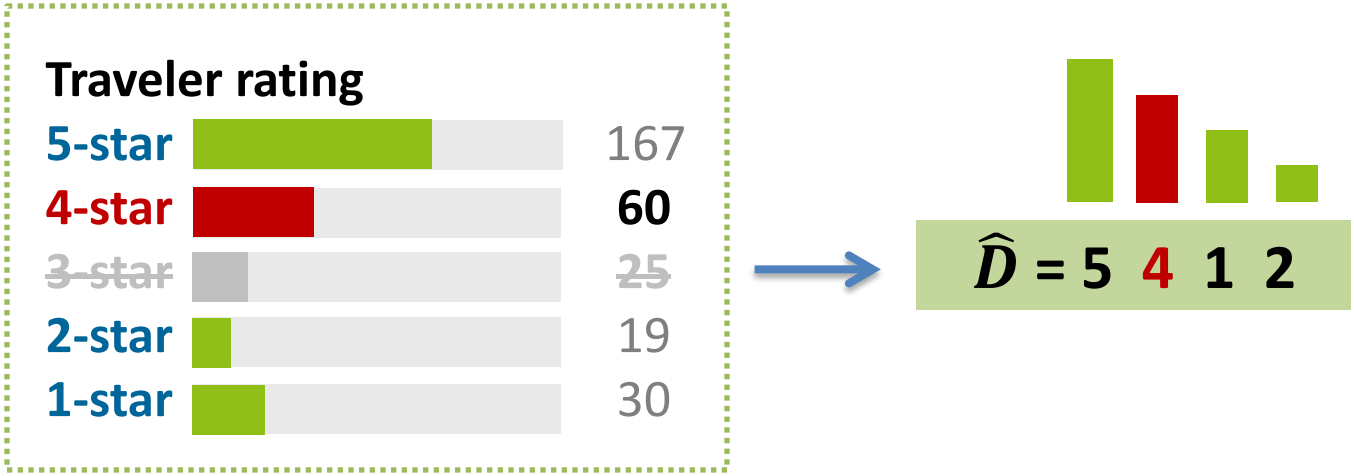


Rating Distribution

Shape of distribution

$\widehat{D}$:= rating scores sorted (in descending order)

by # of corresponding counts

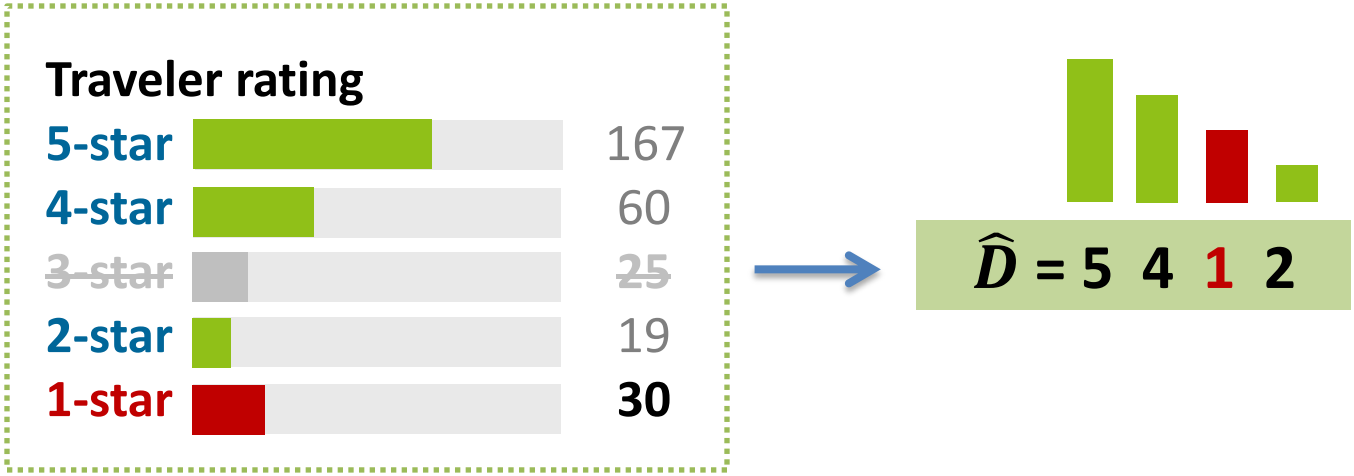


Rating Distribution

Shape of distribution

$\widehat{D}$:= rating scores sorted (in descending order)

by # of corresponding counts

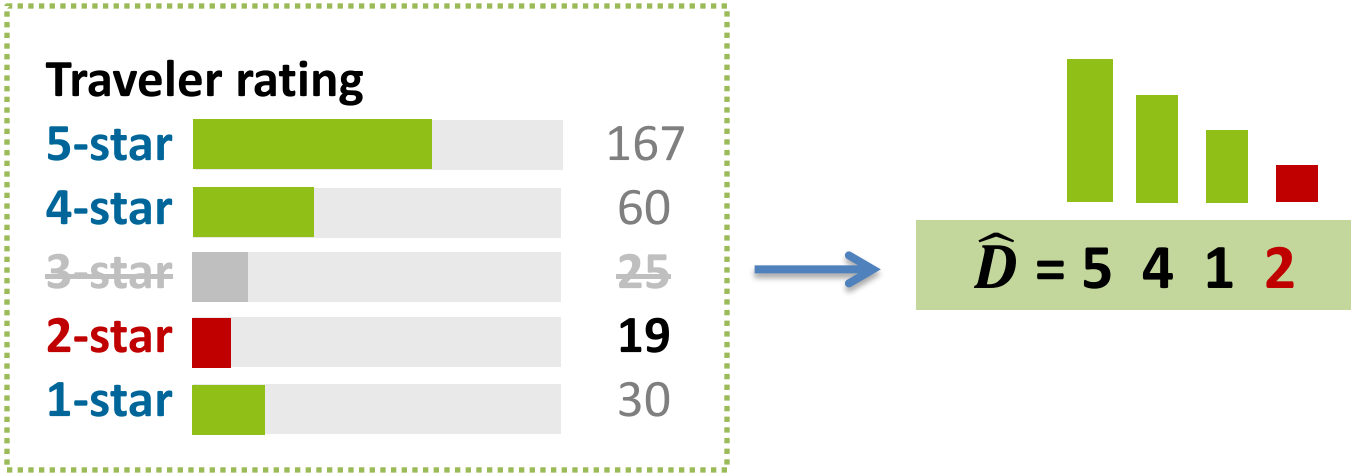


Rating Distribution

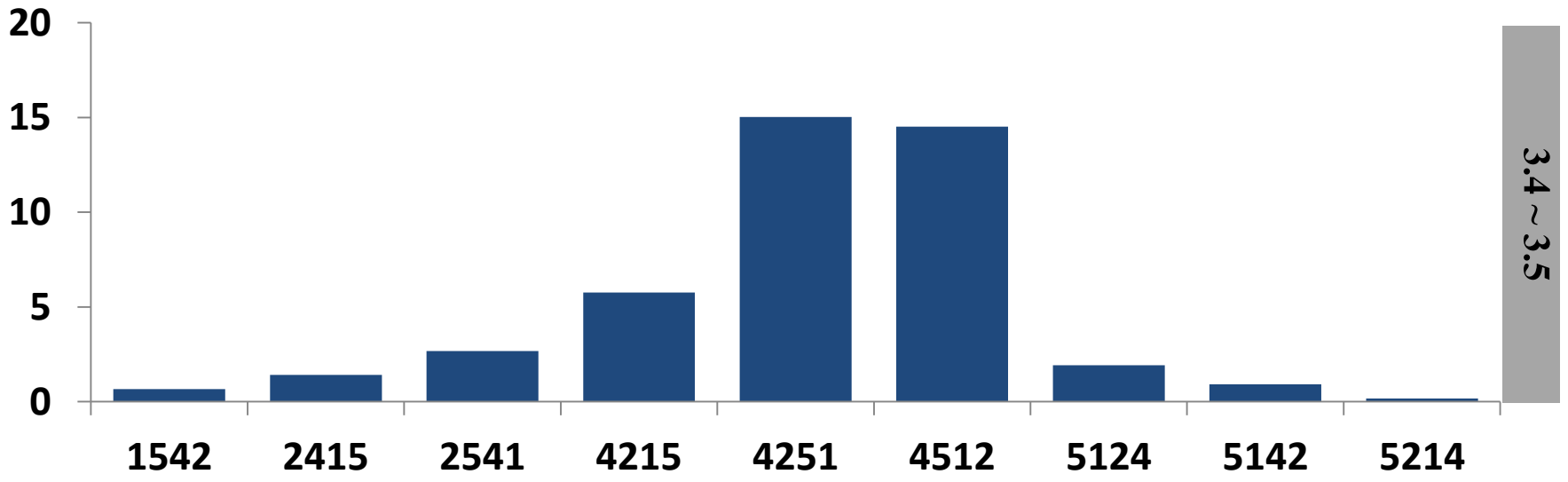
Shape of distribution

$\widehat{D}$:= rating scores sorted (in descending order)

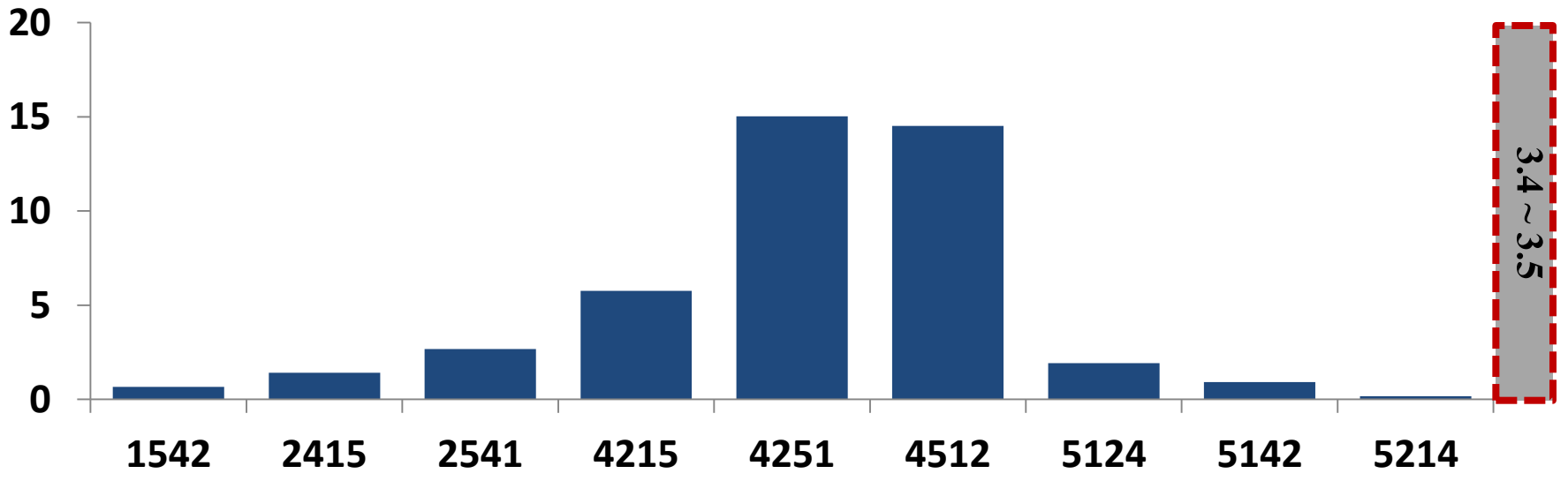
by # of corresponding counts



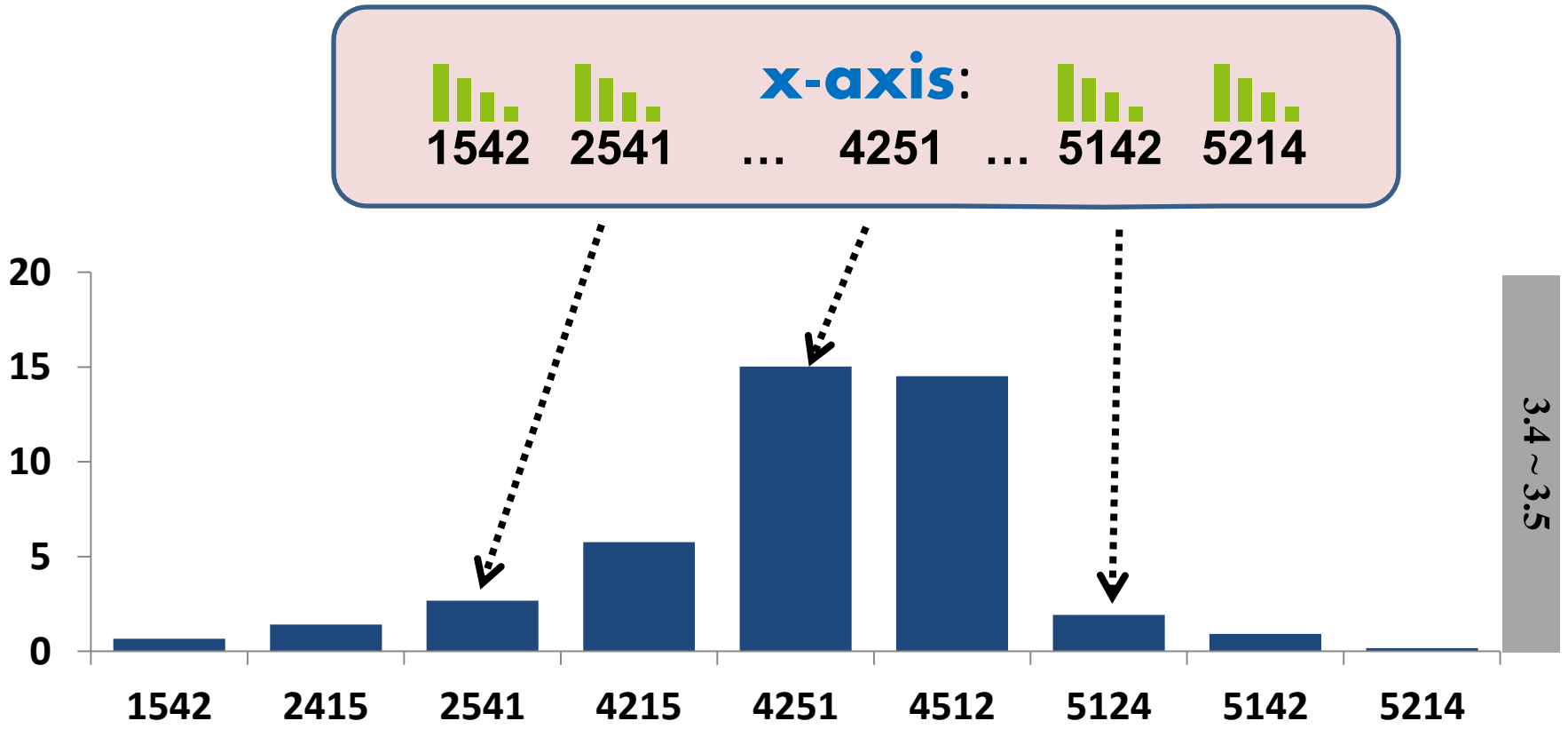
Distribution of $\widehat{D}$



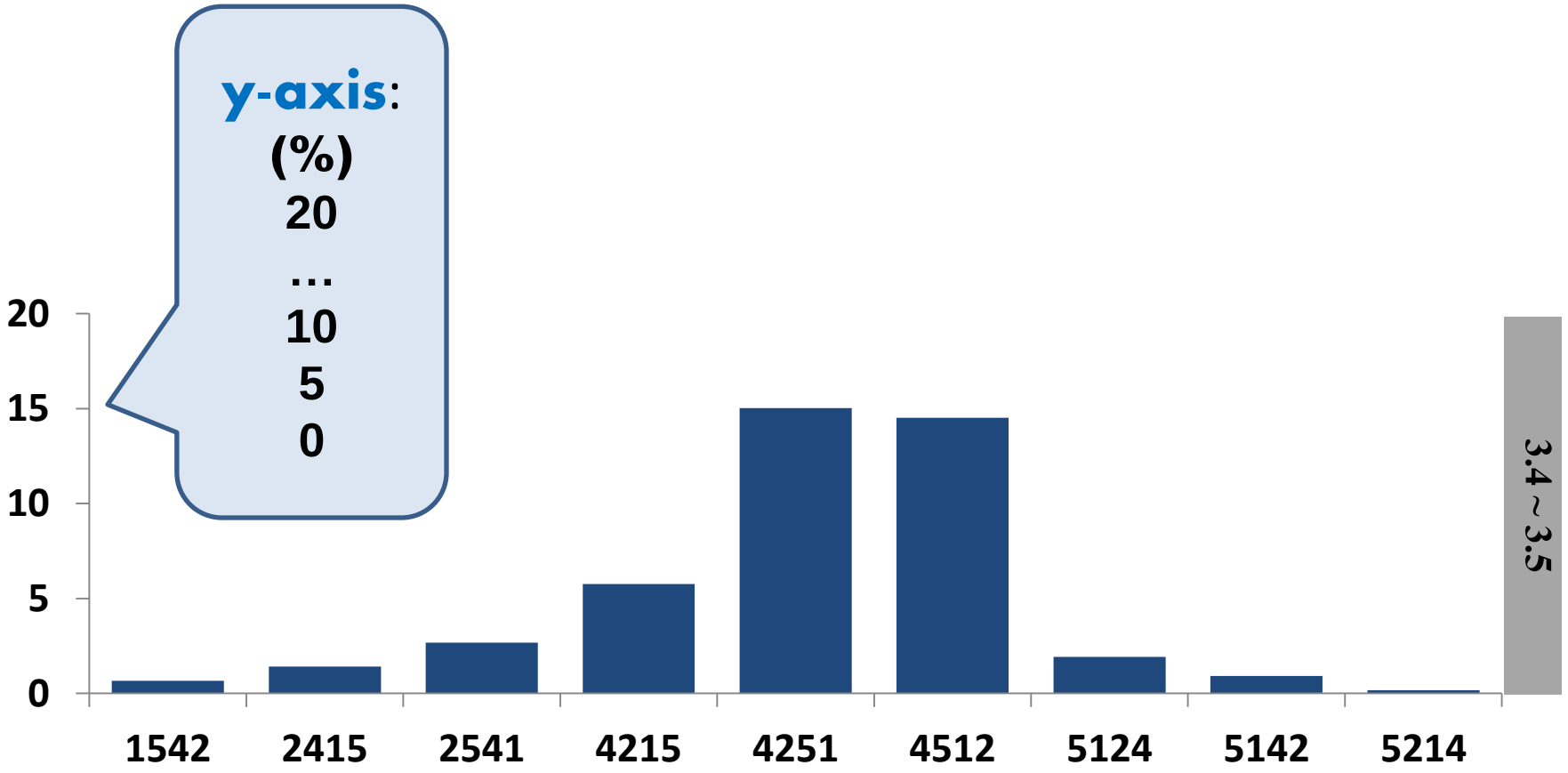
Distribution of $\widehat{D}$



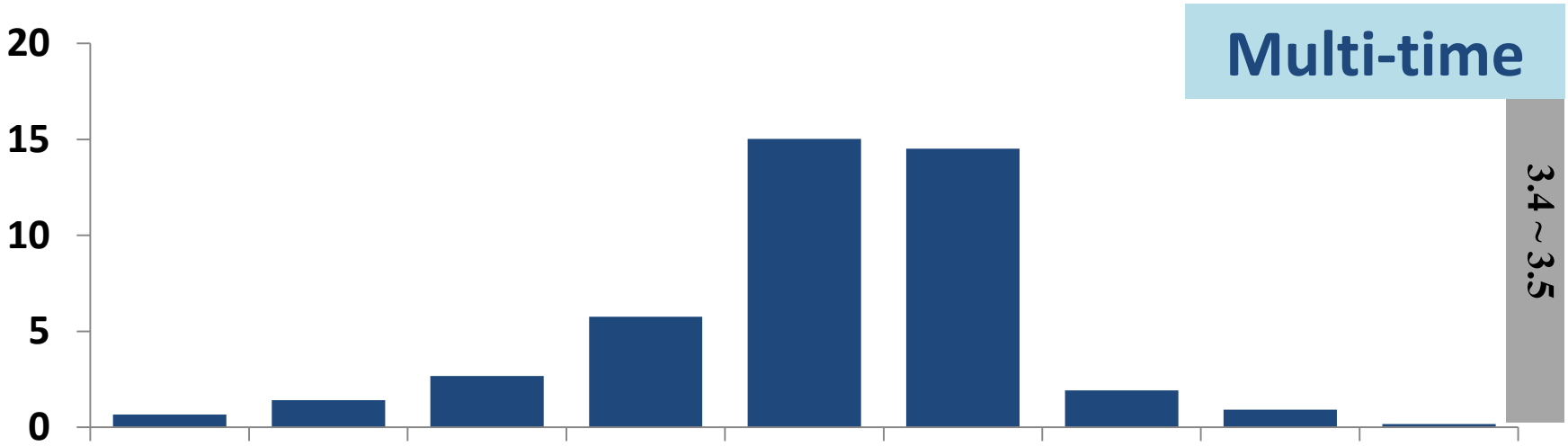
Distribution of $\widehat{D}$



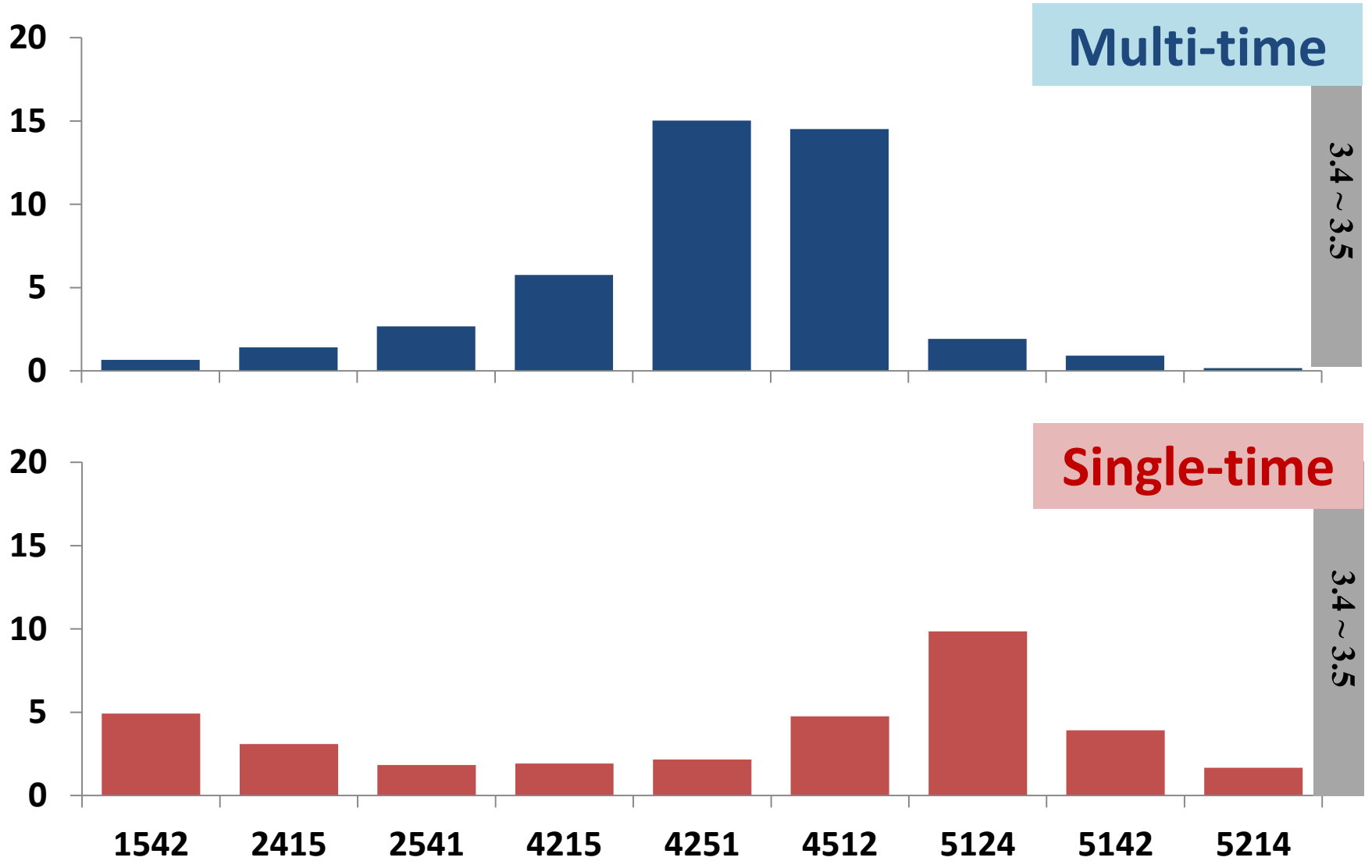
Distribution of $\widehat{D}$



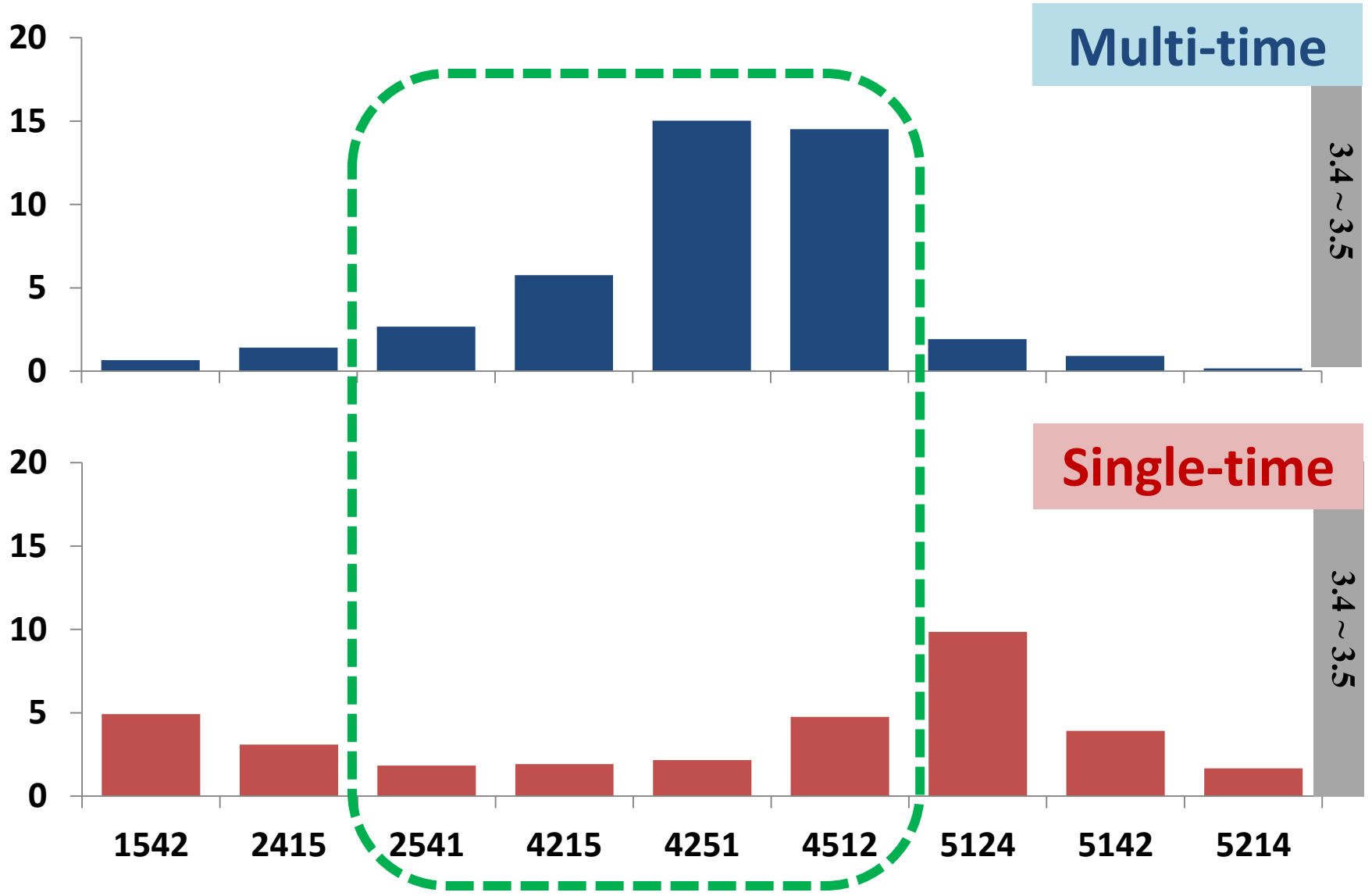
Distribution of $\widehat{D}$



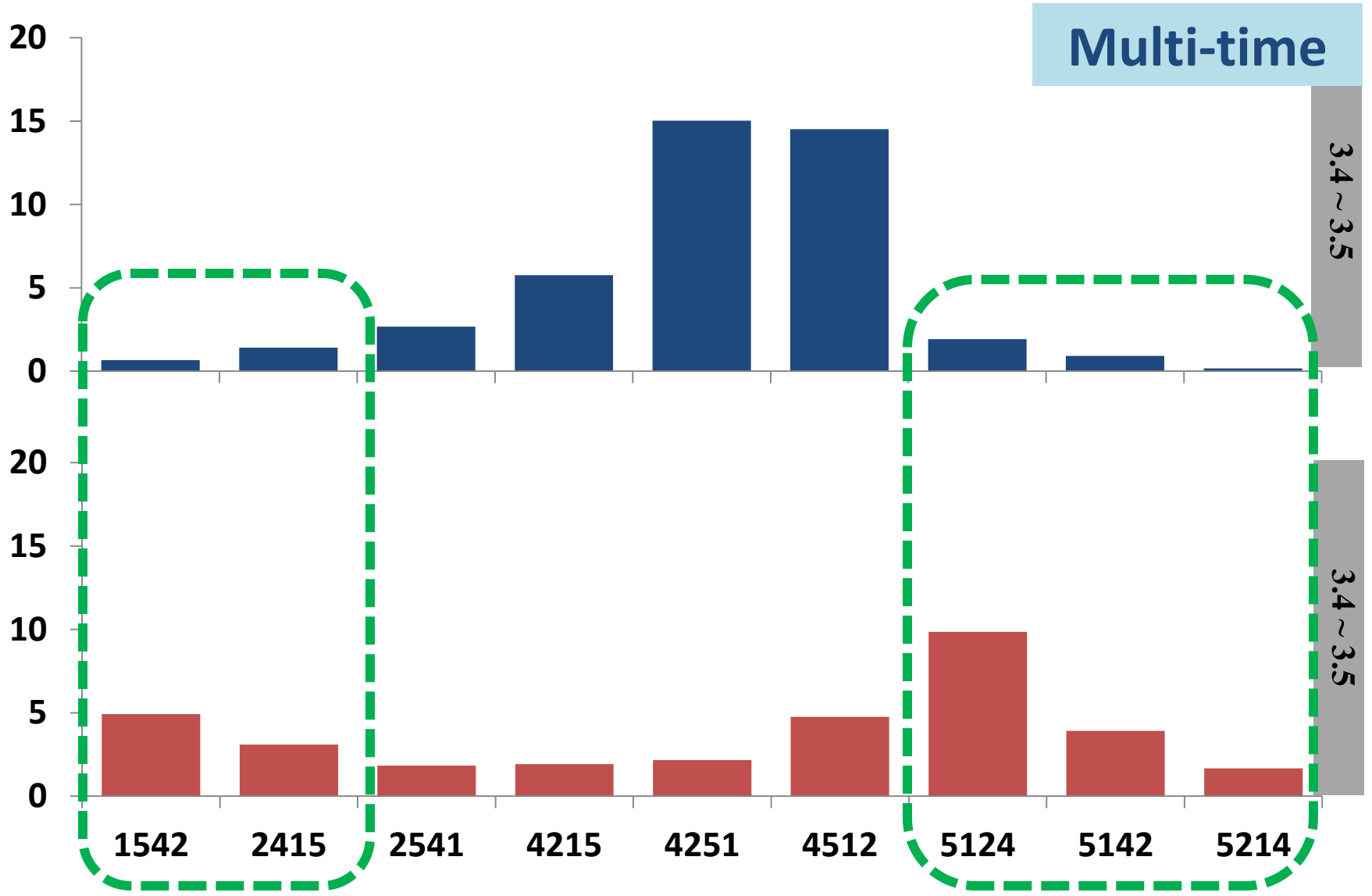
Distribution of $\widehat{D}$



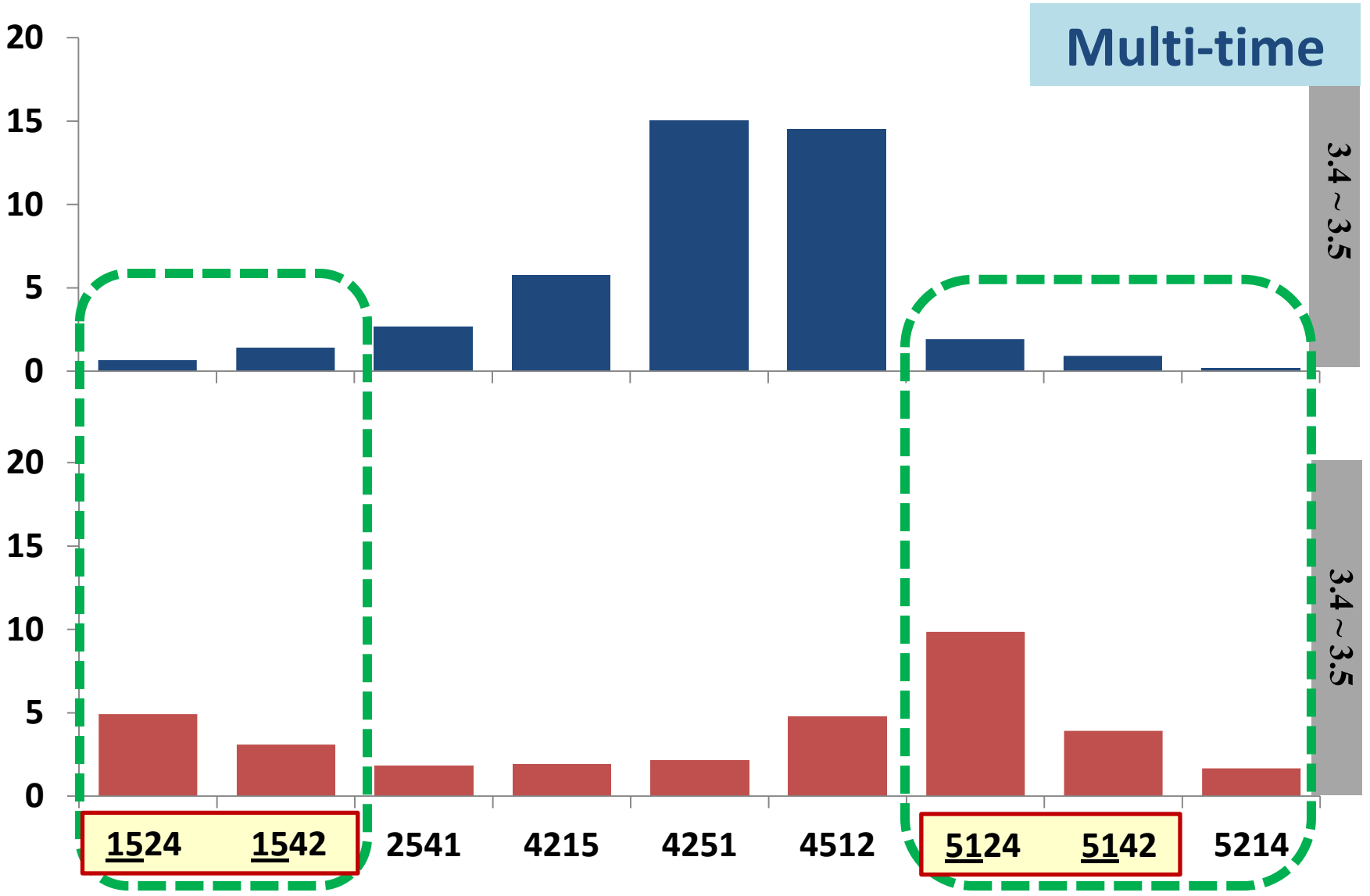
Distribution of $\widehat{D}$



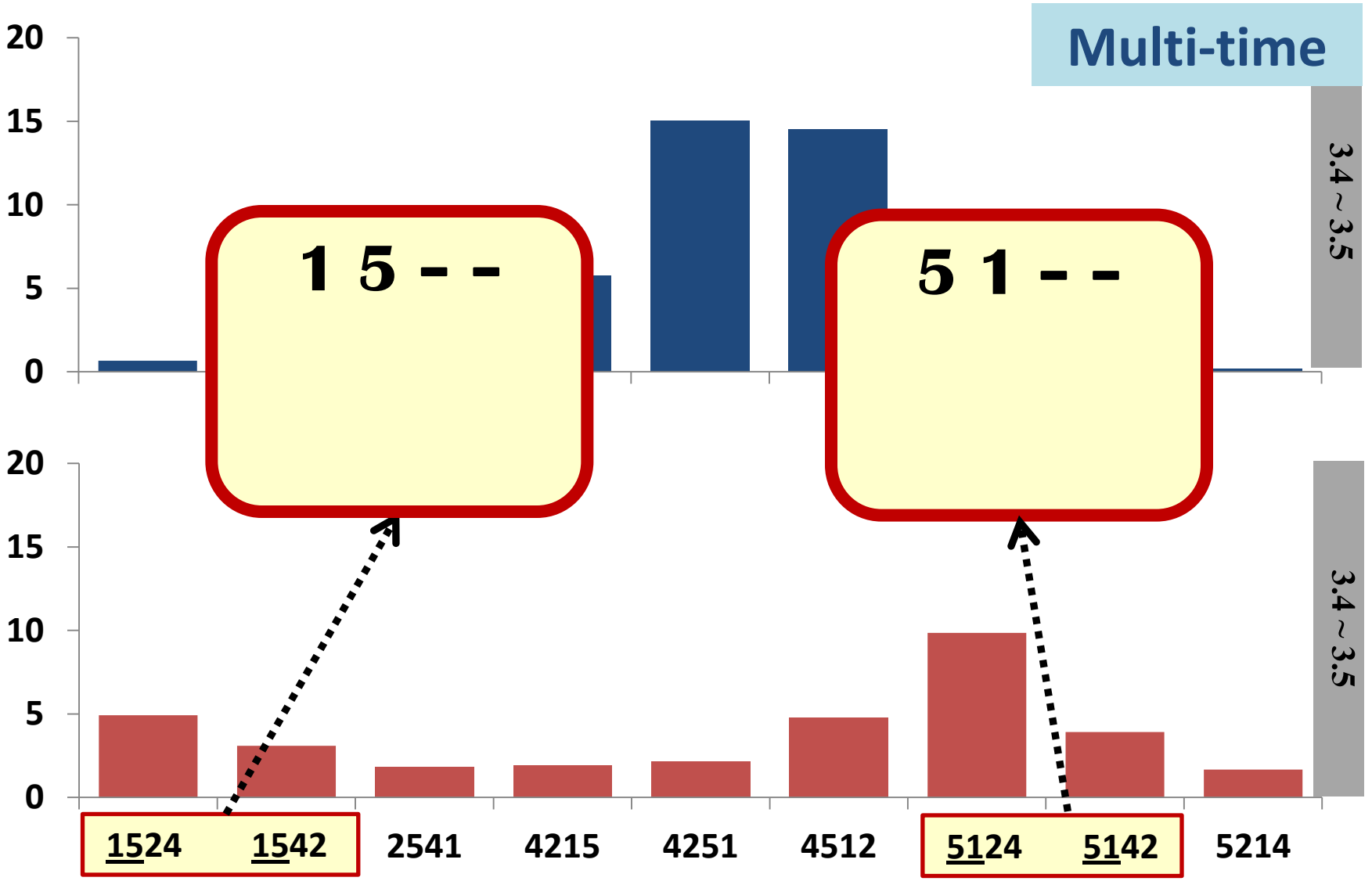
Distribution of $\widehat{D}$



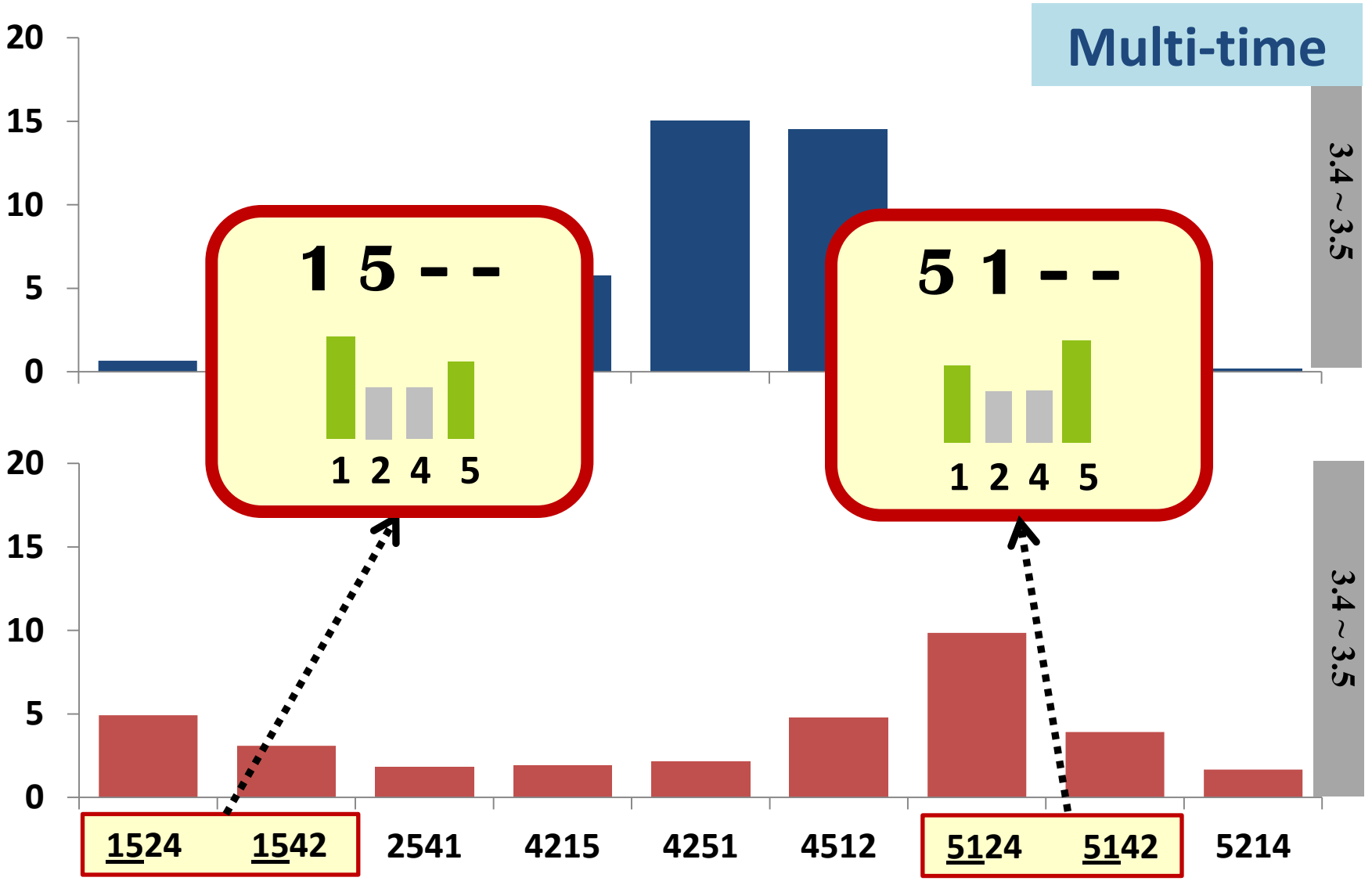
Distribution of $\widehat{D}$



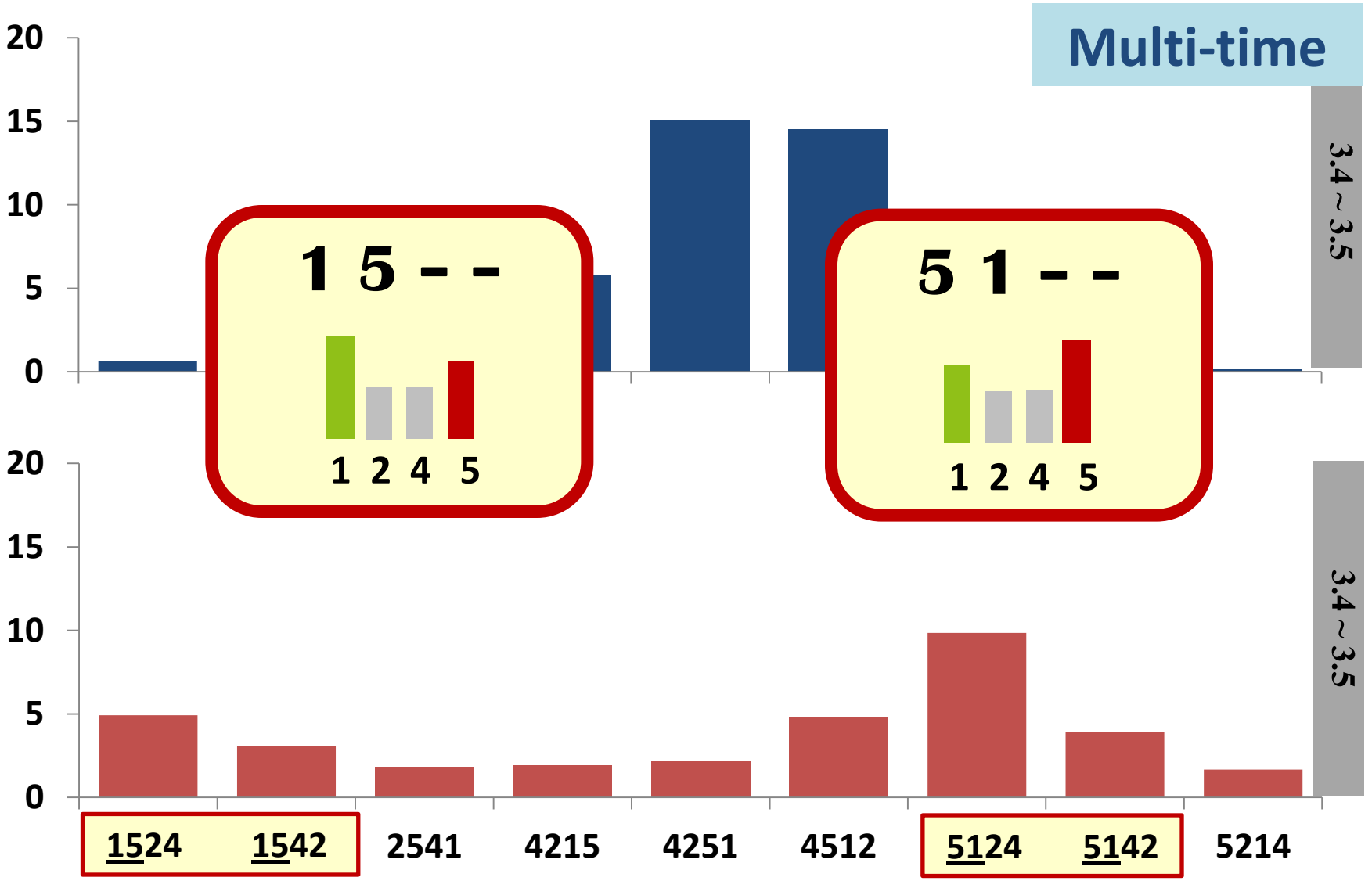
Distribution of $\widehat{D}$



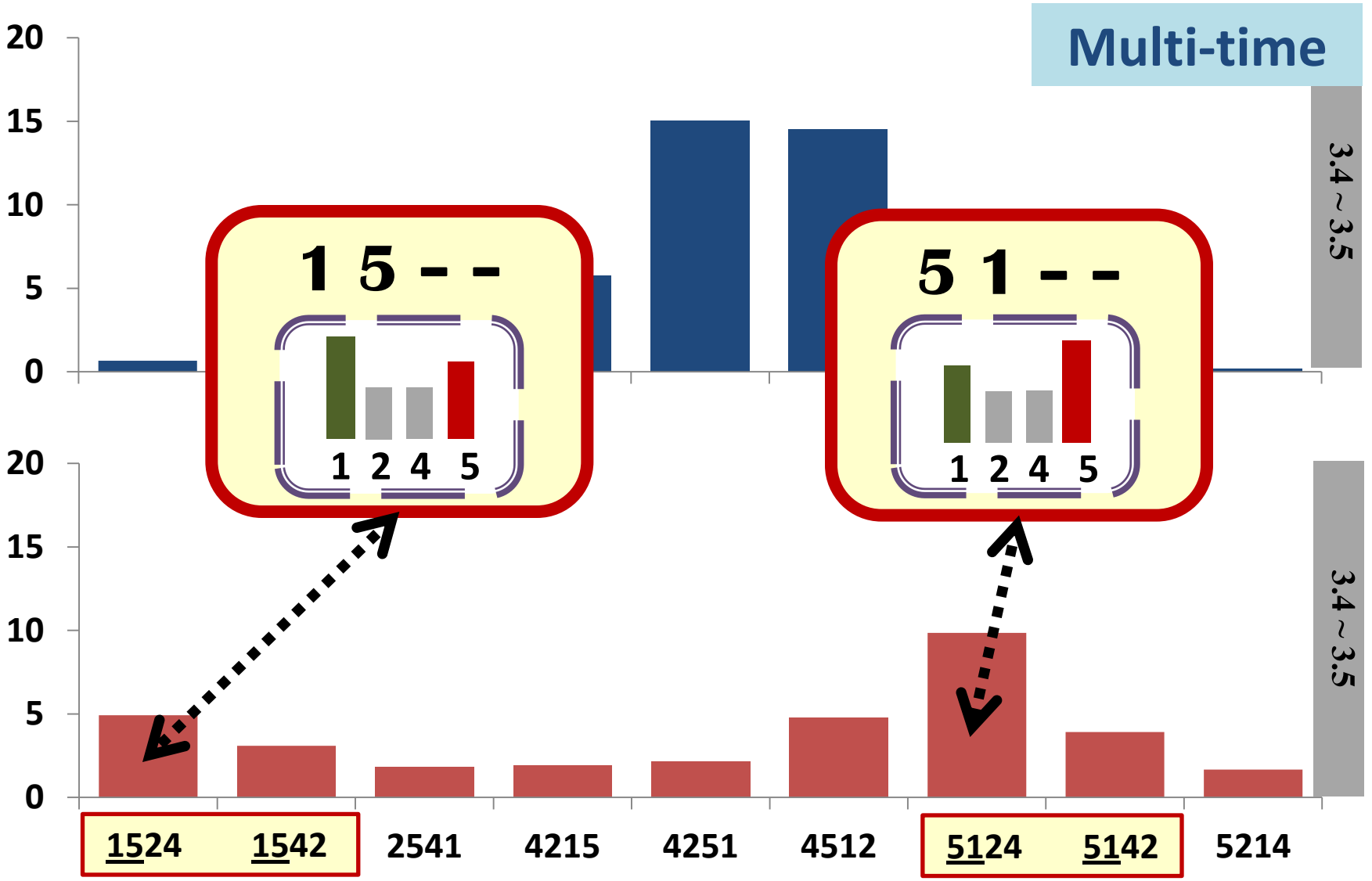
Distribution of $\widehat{D}$



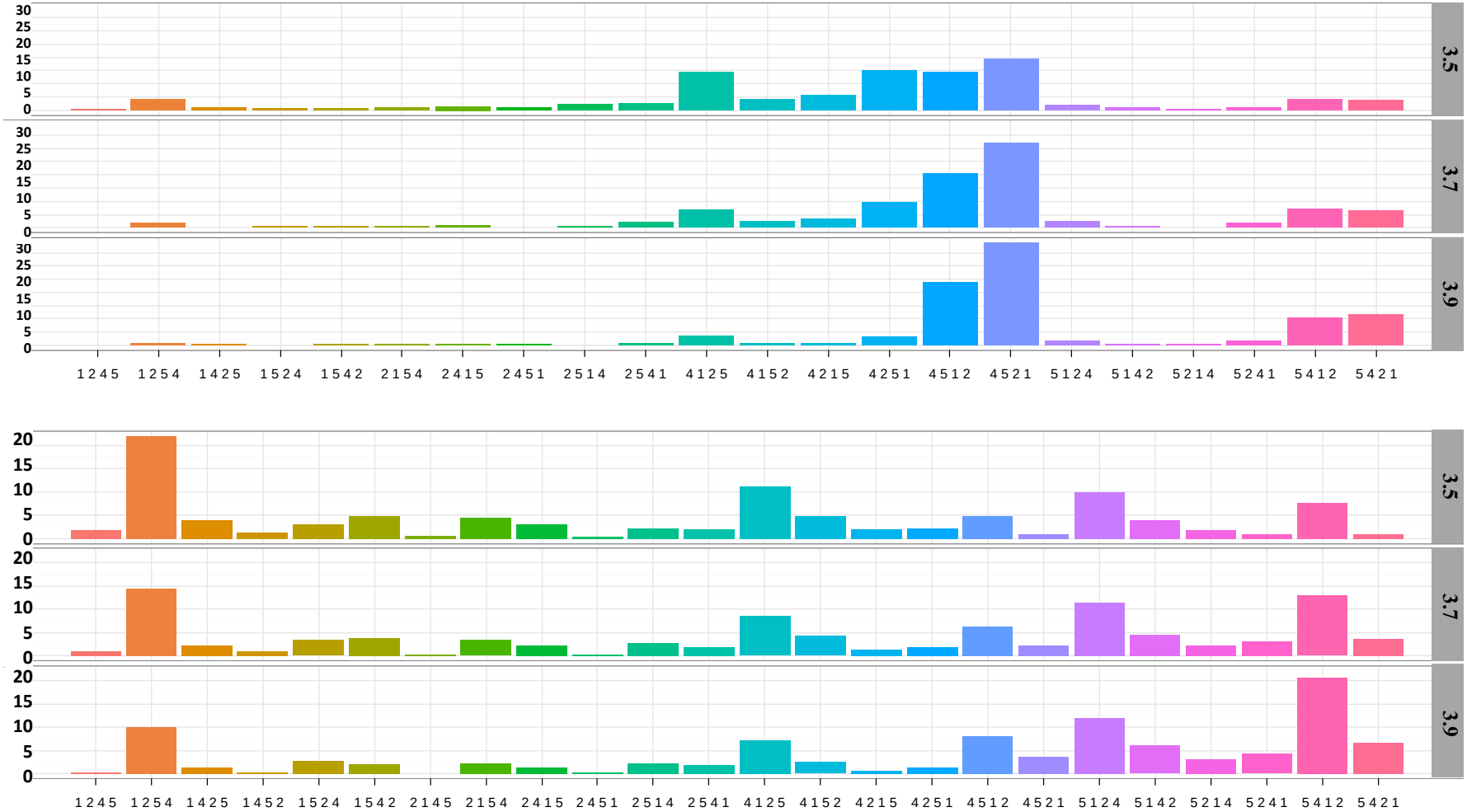
Distribution of $\widehat{D}$



Distribution of $\widehat{D}$



More (colorful) details in the paper !



Three Contributions

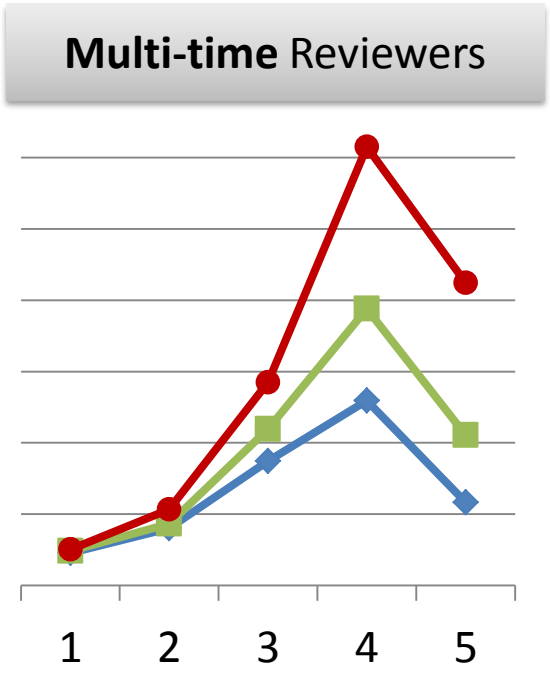
1. Characterization of rating distributions
   - ➔ **Natural** vs. **distorted** rating distributions
2. **Detection strategies to identify deceptive business entities & reviews**
3. Novel evaluation methodologies.
   - ❖ Avoid human judges
     (because they are not good at catching fakes)
     -- Ott et al. 2011 report human accuracy ~ 60%
   - ❖ Avoid manufacturing fake reviews
     (because they are costly)

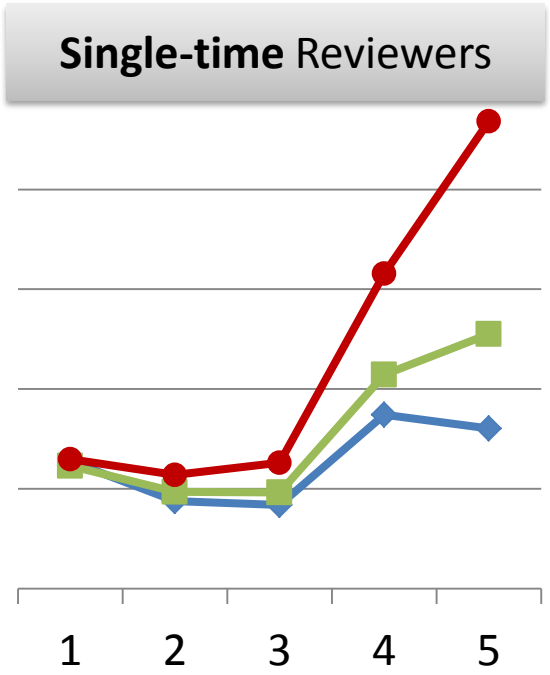
Strategies: Suspicious Hotels

- **Average rating discrepancy**
  - $\delta(h) = \bar{r}_S(h) - \bar{r}_M(h)$
  - Sort $\delta(h)$ in descending order, select hotels with top-ranked $\delta(h)$.
- **Rating distribution: *highly positive/negative***
  - $\tau(h) = \dfrac{\frac{|5star|_S}{|1star|_S}}{\frac{|5star|_M}{|1star|_M}}$
  - Sort $\tau(h)$ in descending order, select hotels with top-ranked $\tau(h)$.
- **Temporal boost in rating**
  - Monthly rating is much greater than that of months before and after. (*Jindal et al.* 2010)

Strategies: Suspicious Hotels

- **Average rating discrepancy**
  - $\delta(h) = \bar{r}_S(h) - \bar{r}_M(h)$
  - Sort $\delta(h)$ in descending order, select hotels with top-ranked $\delta(h)$.

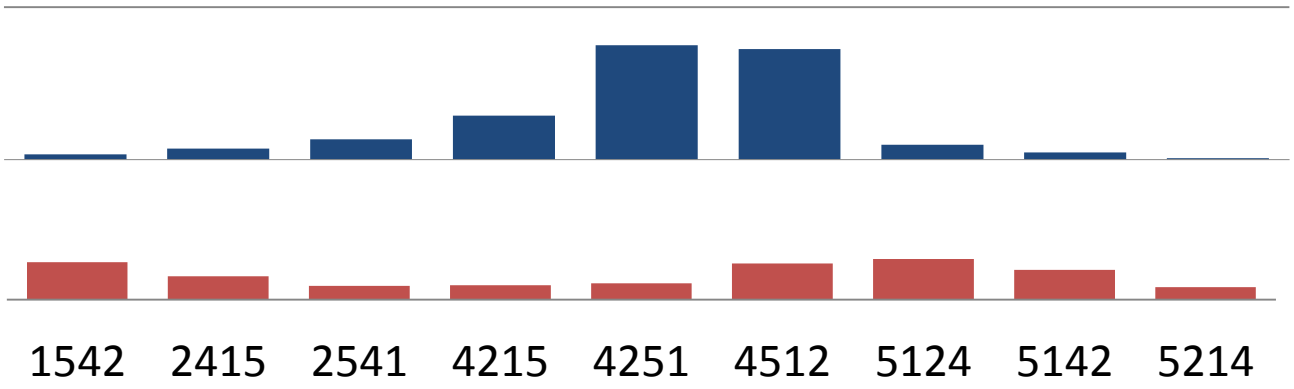




Strategies: Suspicious Hotels

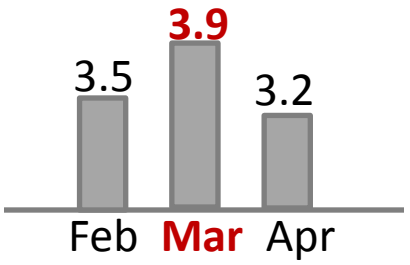
- **Average rating discrepancy**
  - $\delta(h) = \bar{r}_S(h) - \bar{r}_M(h)$
  - Sort $\delta(h)$ in descending order, select hotels with top-ranked $\delta(h)$.

- **Rating distribution: *positive/negative***
  - $\tau(h) = \dfrac{\frac{|5star|_S}{|1star|_S}}{\frac{|5star|_M}{|1star|_M}}$
  - Sort $\tau(h)$ in descending order, select hotels with top-ranked



Strategies: Suspicious Hotels

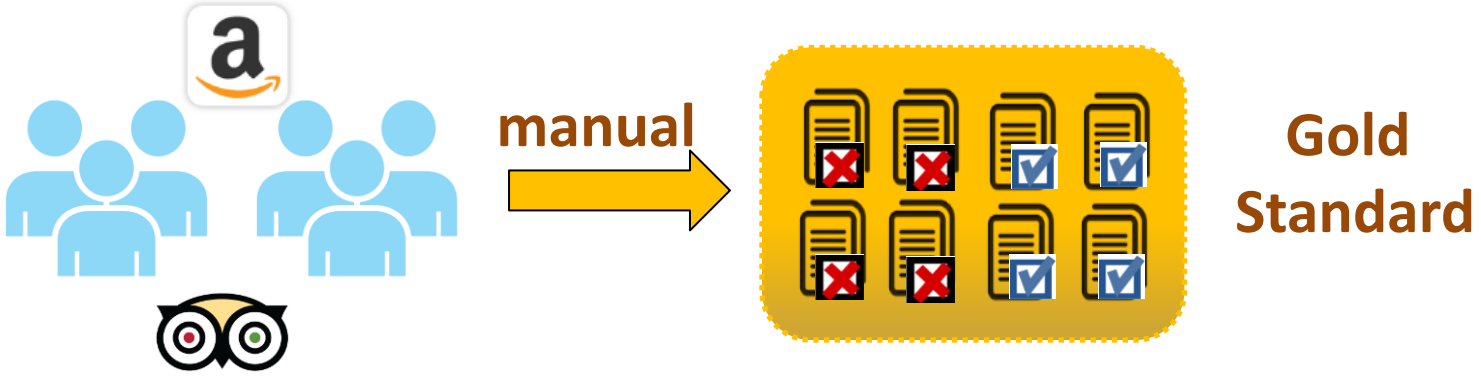
- **Average rating discrepancy**
  - $\delta(h) = \bar{r}_S(h) - \bar{r}_M(h)$
  - Sort $\delta(h)$ in descending order, select hotels with top-ranked $\delta(h)$.
- **Rating distribution: *highly positive/negative***
  - $\tau(h) = \dfrac{\frac{|5star|_S}{|1star|_S}}{\frac{|5star|_M}{|1star|_M}}$
  - Sort $\tau(h)$ in descending order, select hotels with top-ranked $\tau(h)$.
- **Temporal boost in rating**
  - Monthly rating is much greater than that of months before and after. (*Jindal, et al.* 2010)



Three Contributions

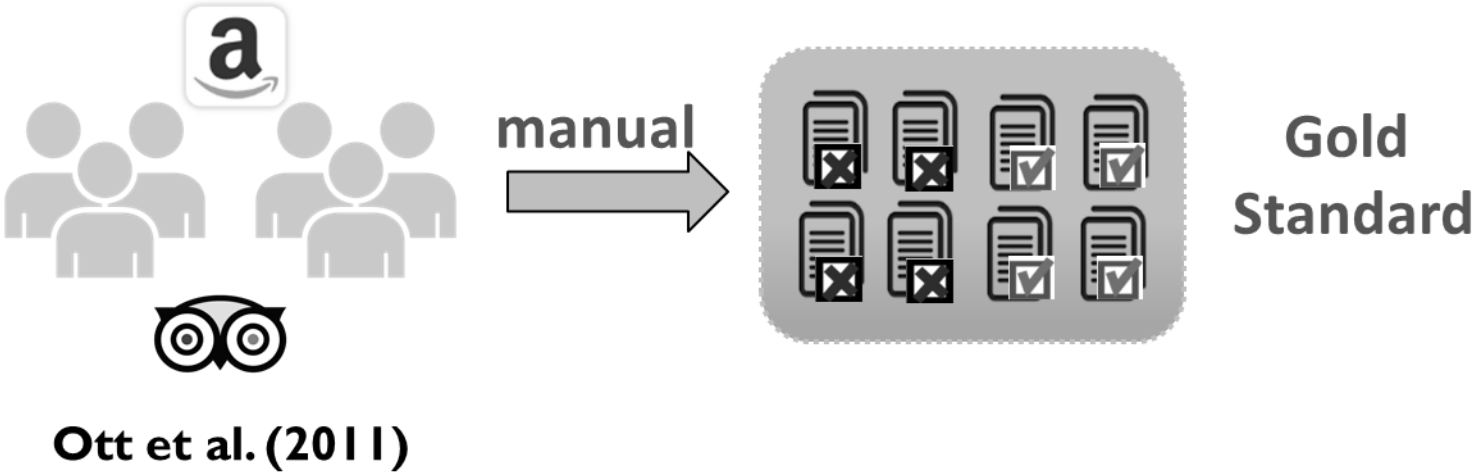
1. Characterization of rating distributions
   - ➔ **Natural** vs. **distorted** rating distributions
2. Detection strategies to identify deceptive business entities & reviews
3. **Novel evaluation methodologies.**
   - ❖ Avoid human judges
     (because they are ***not good*** at catching fakes)
     -- Ott et al. 2011 report human accuracy < 62%
   - ❖ Avoid manufacturing fake reviews
     (because they are ***costly***)

Pseudo-Gold Standard Data



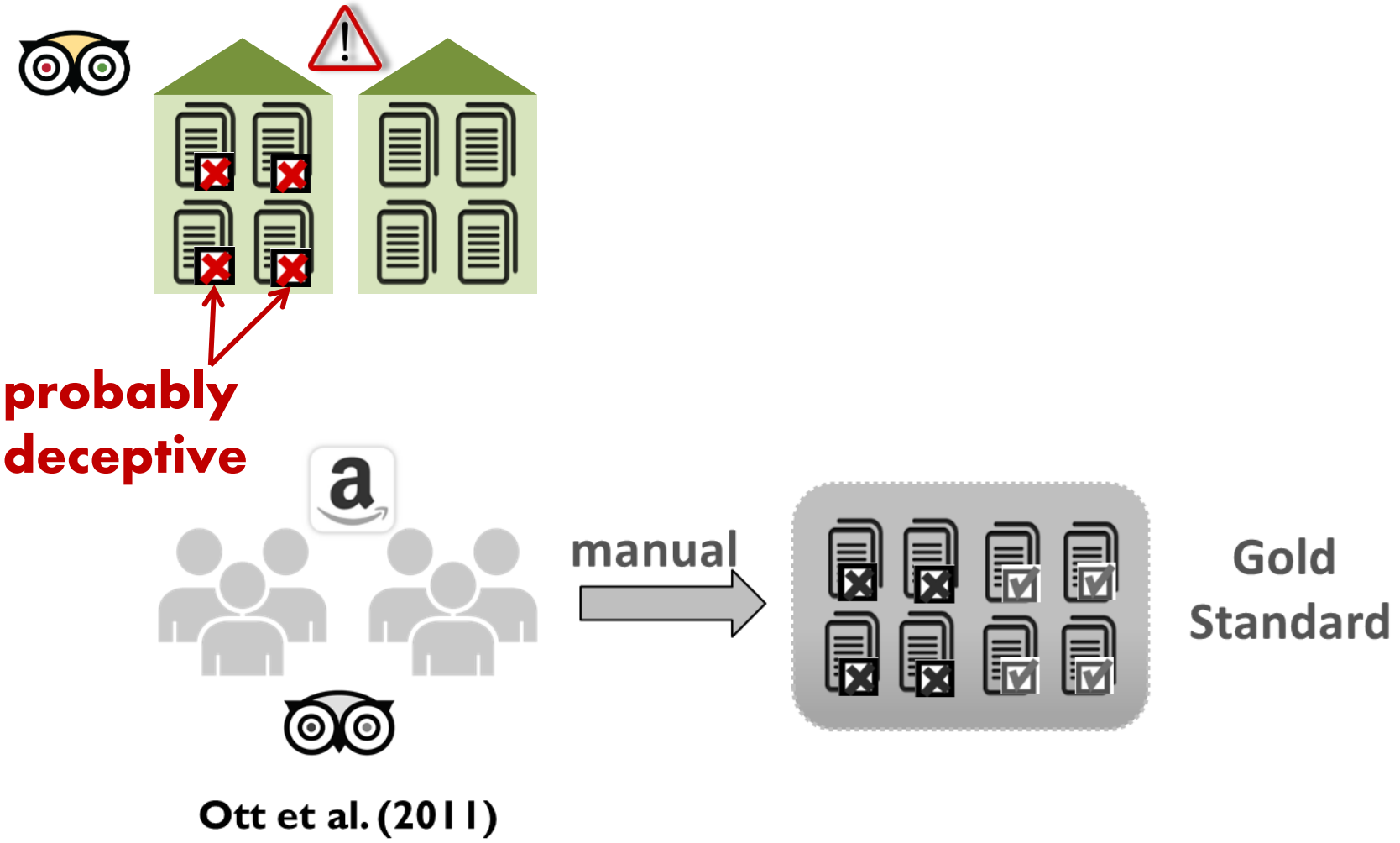
Ott et al. (2011)

Pseudo-Gold Standard Data

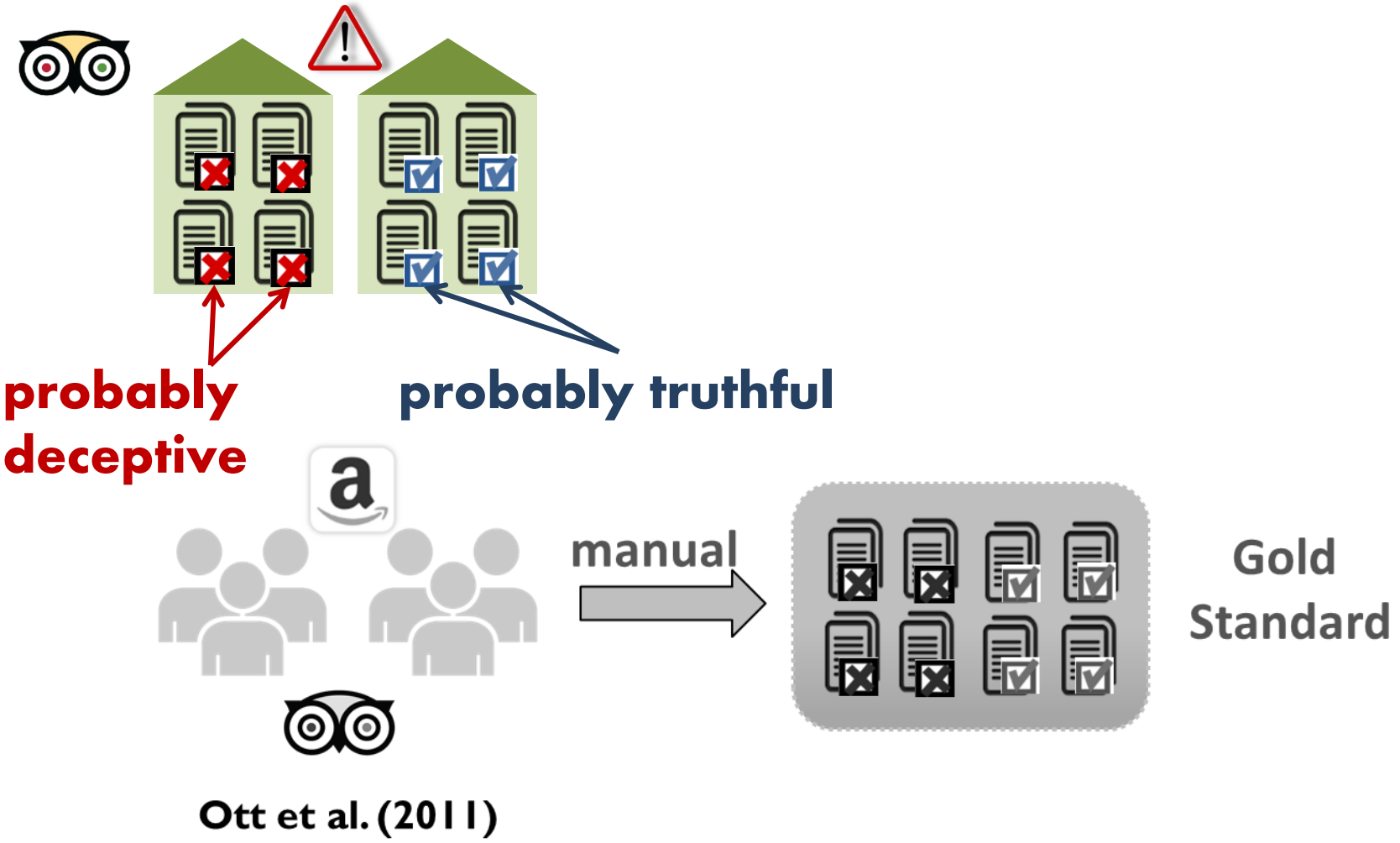


Ott et al. (2011)

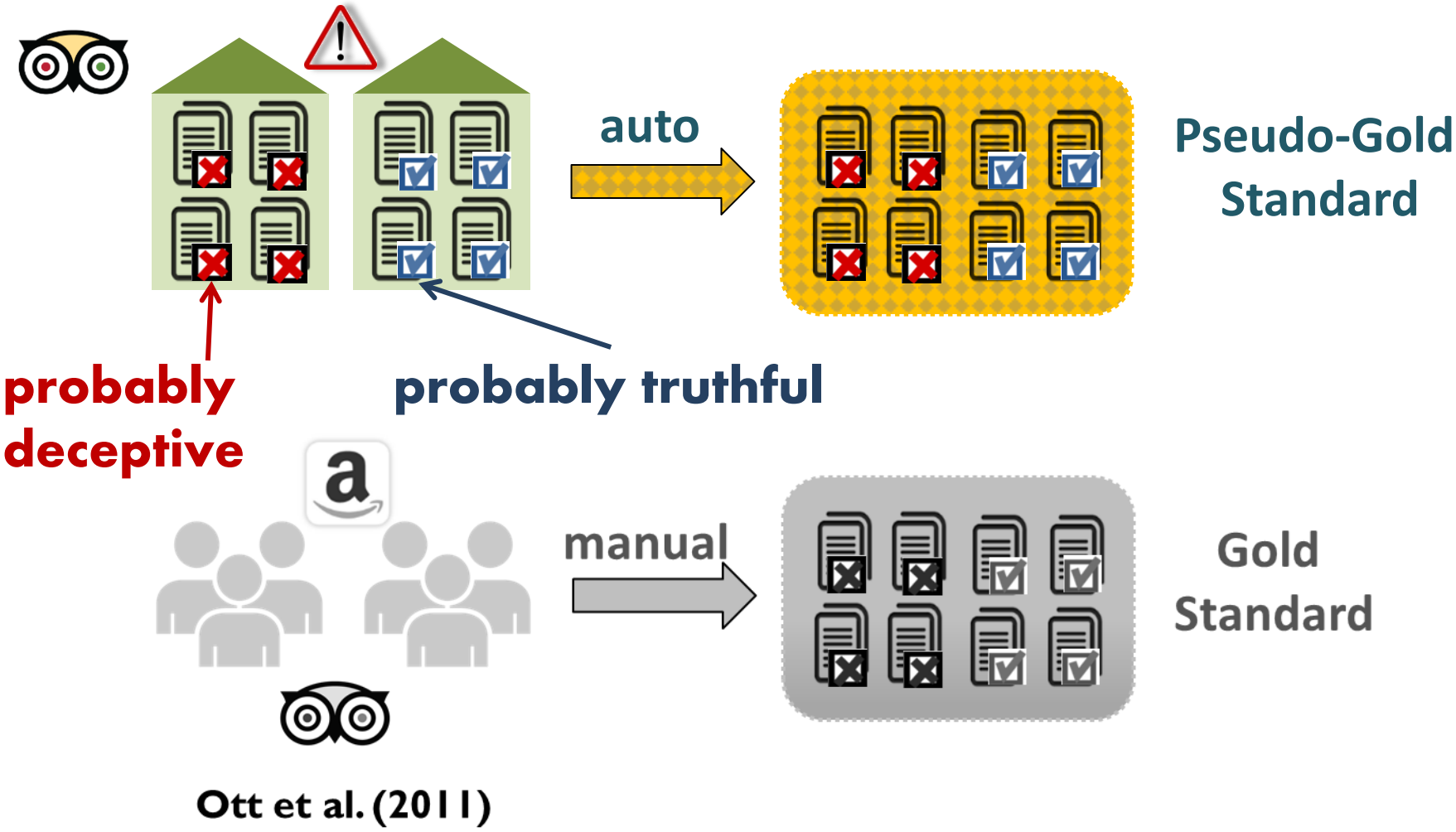
Pseudo-Gold Standard Data



Pseudo-Gold Standard Data

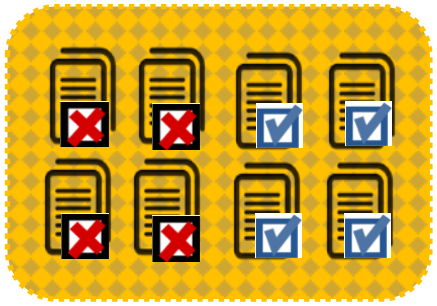


Pseudo-Gold Standard Data



Pseudo-Gold Standard Data

auto



Pseudo-Gold Standard

Ott et al. (2011)

manual

Gold Standard

Evaluation via Machine Learning

Training Data	Testing Data
Pseudo-gold standard	Gold Standard

- **Gold standard** datasets (*Ott et al. 2011*)
  - 400 truthful reviews from Tripadvisor.
  - 400 deceptive positive reviews by AMT.
- **Pseudo-gold standard** datasets
  - Three strategies
  - Positive: 5-star

Evaluation via Machine Learning

Training Data	Testing Data
Gold standard	Pseudo-gold standard
Pseudo-gold standard	Gold standard
Pseudo-gold standard	Pseudo-gold standard

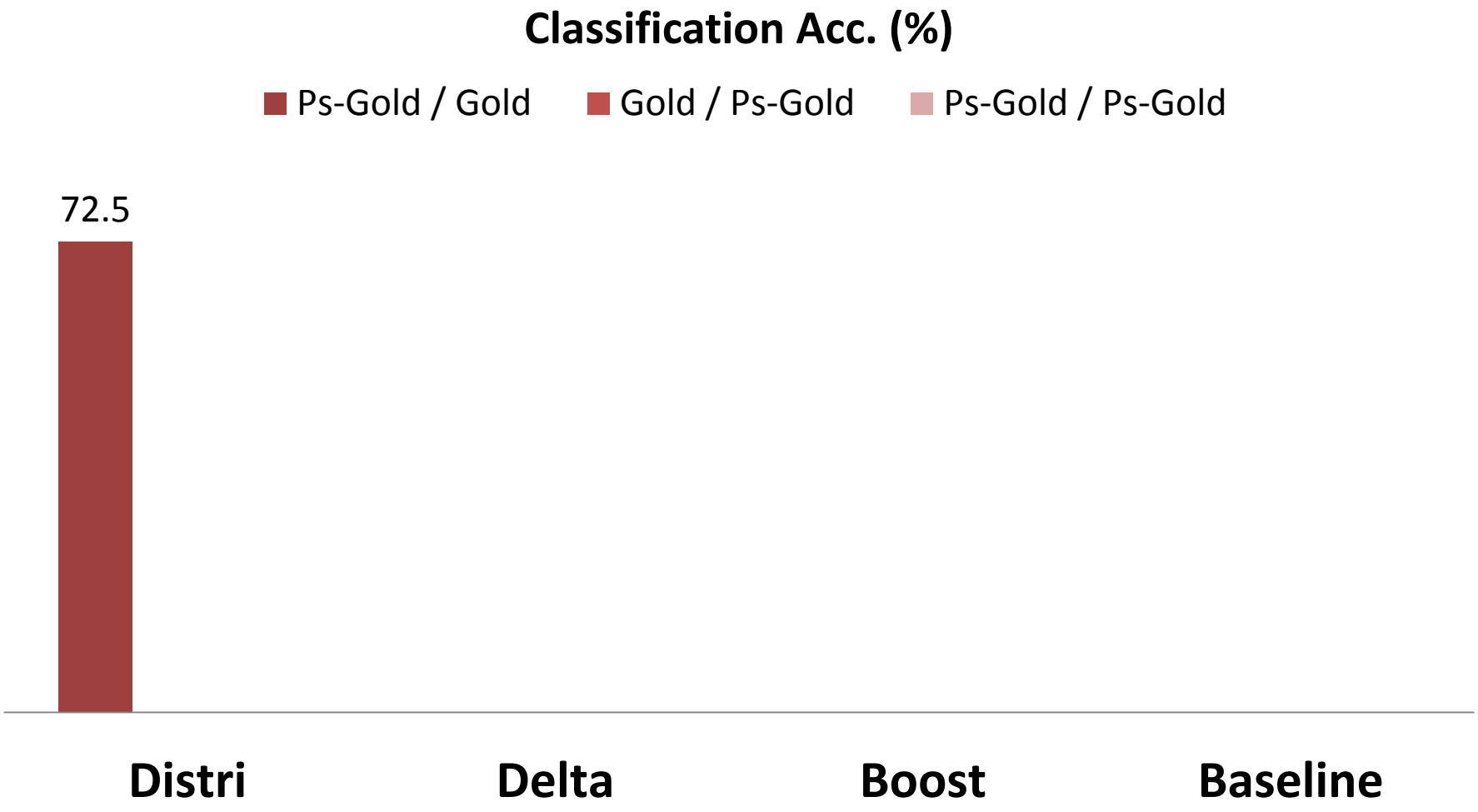
- **Gold standard** datasets (*Ott et al. 2011*)
  - 400 truthful reviews from Tripadvisor.
  - 400 deceptive positive reviews by AMT.
- **Pseudo-gold standard** datasets
  - Three strategies
  - Positive: 5-star

Evaluation via Machine Learning

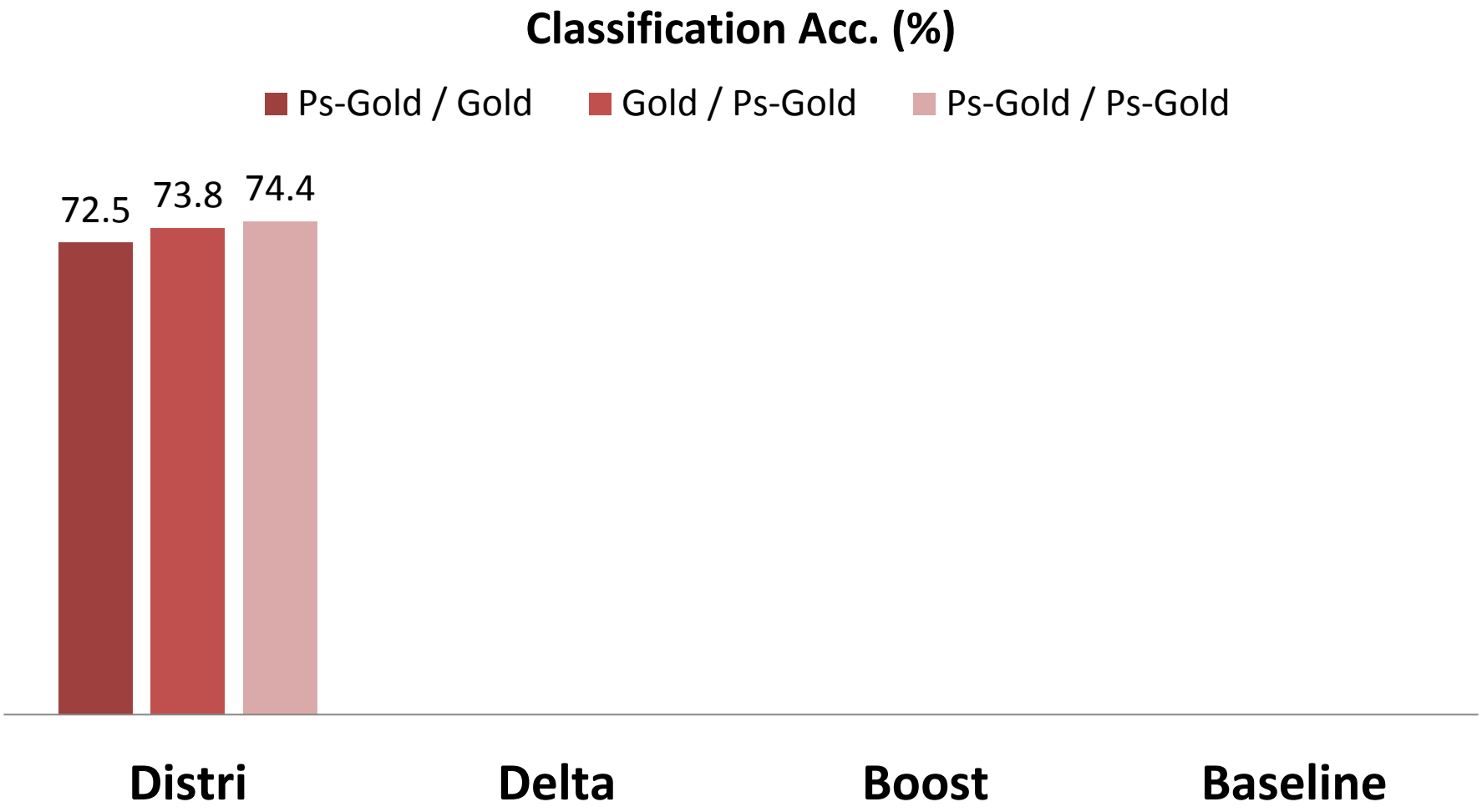
Experiment settings

- SVM classifier: LIBSVM (*Chang and Lin*, 2012)
- 80% training, 20% testing
- 5-fold cross validation
- Term frequency of unigrams

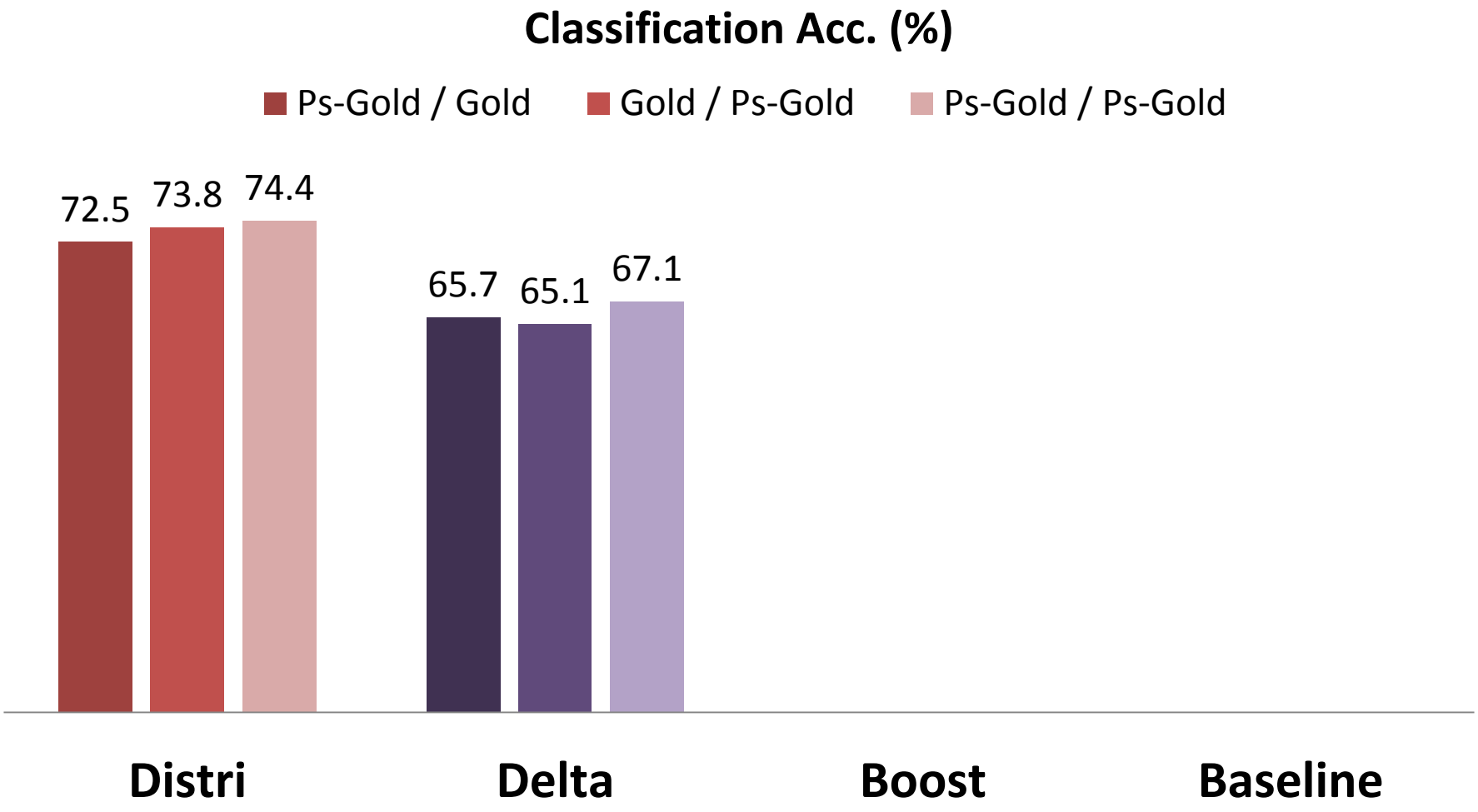
Evaluation: Three Strategies



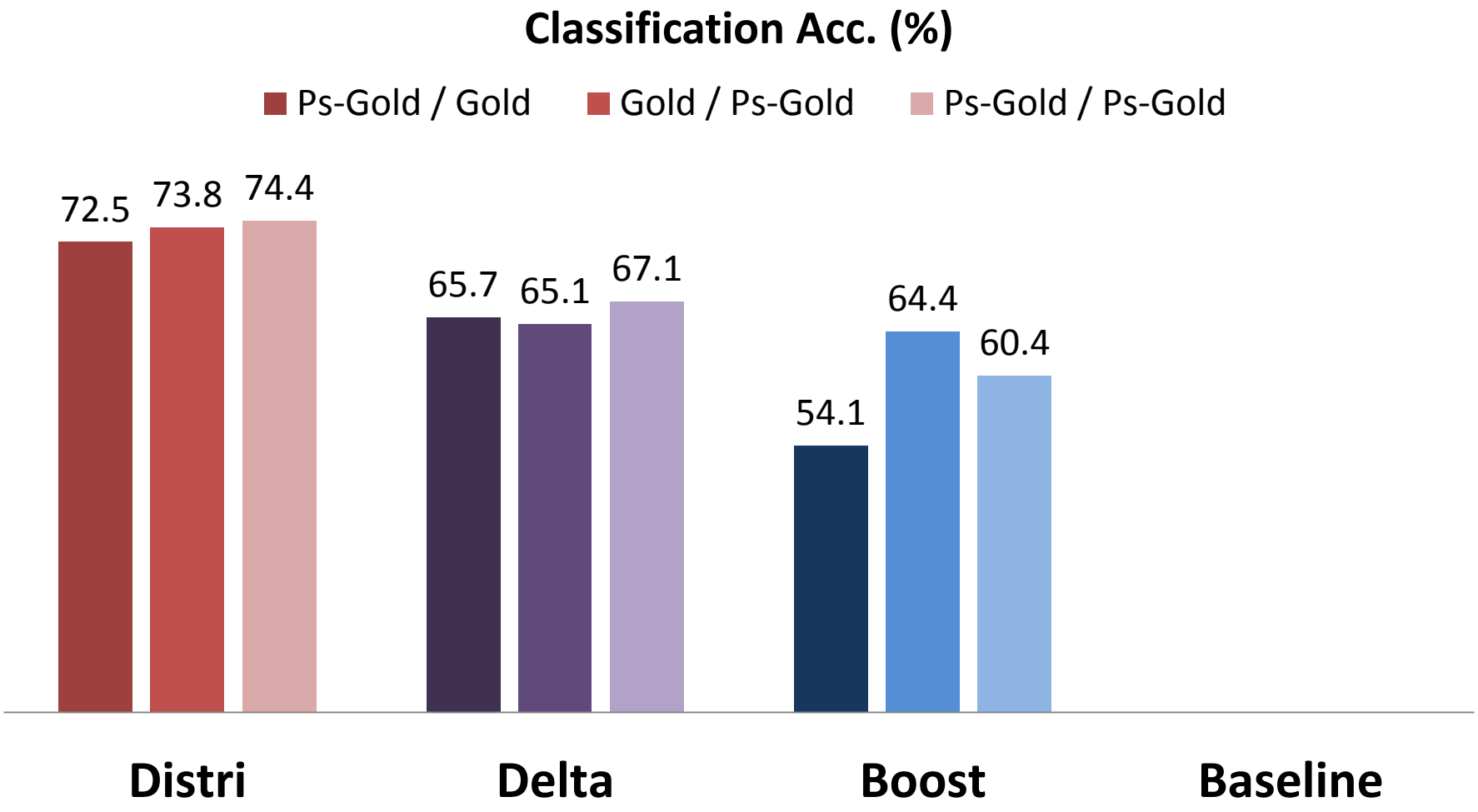
Evaluation: Three Strategies



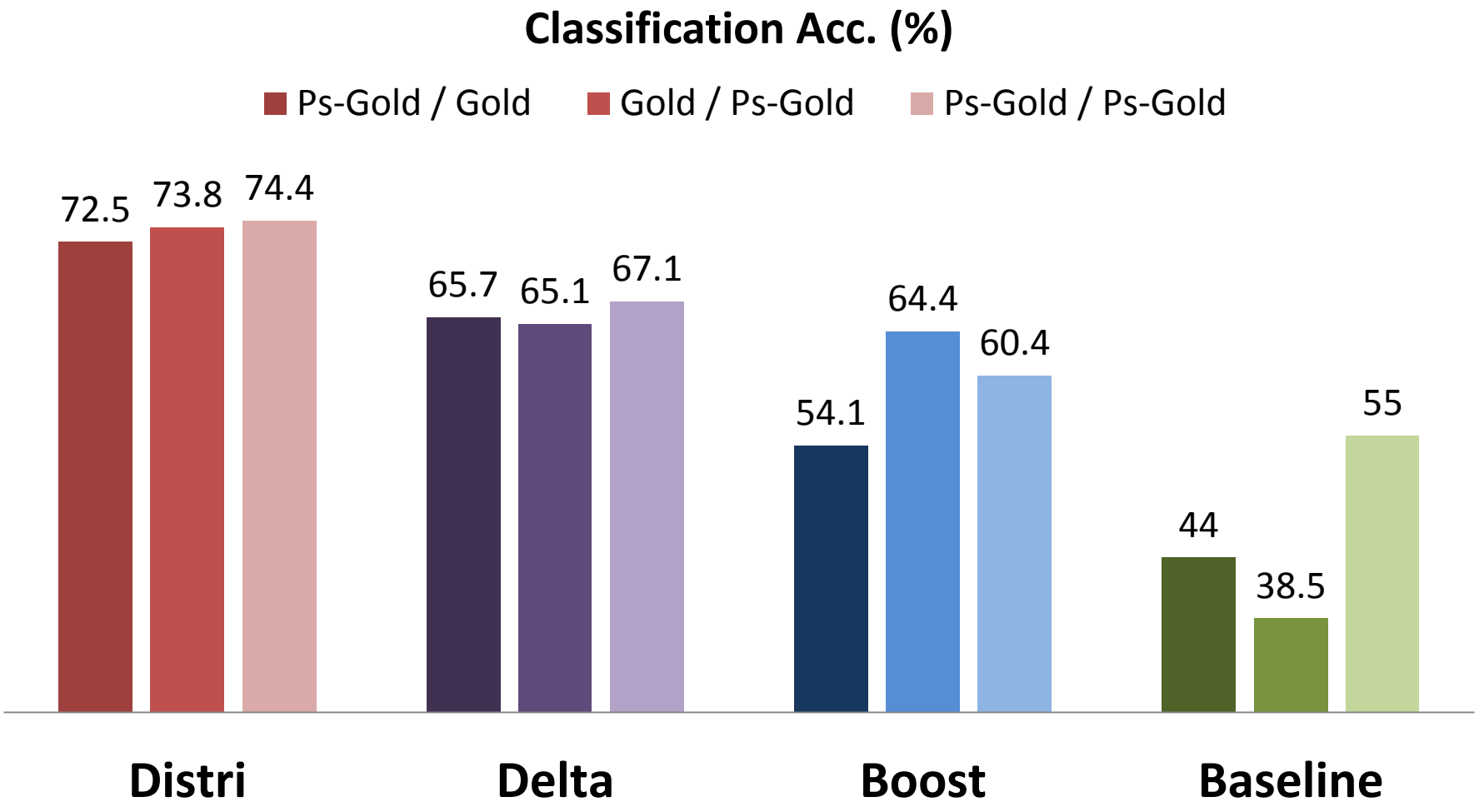
Evaluation: Three Strategies



Evaluation: Three Strategies



Evaluation: Three Strategies



Conclusion

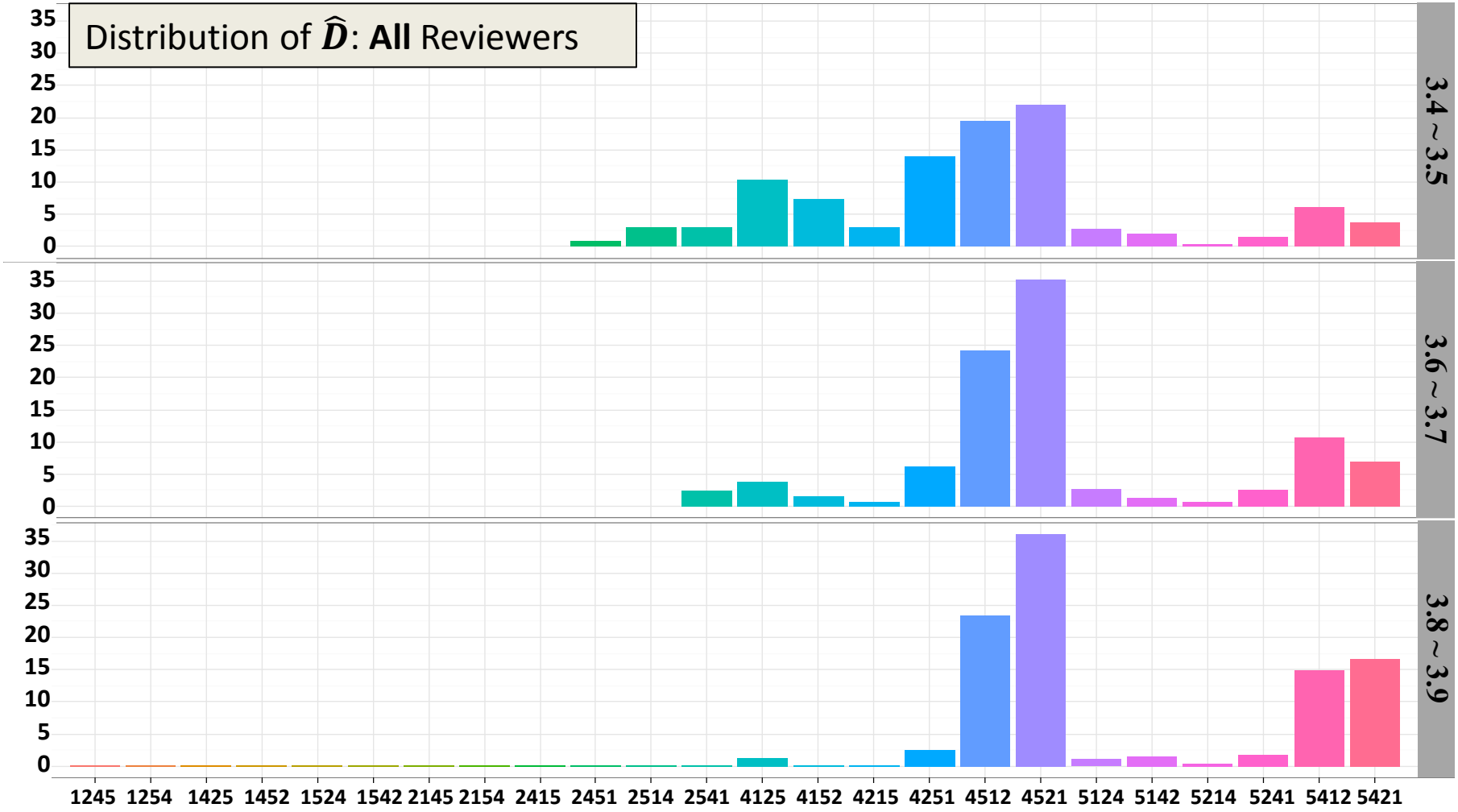
- Introduced "***natural distribution of opinions***"
  - Quantitative analyses on TripAdvisor & Amazon
- Strategies for detecting deception
  - based only on the distributional footprint (metadata)
  - without relying on textual content ➔ not susceptible to newly trained fake reviewers or domain change!
- Novel evaluation techniques
  - not dependent on human judges (unreliable)
  - not dependent on human labor (costly)

➔ Pseudo-gold standard data!

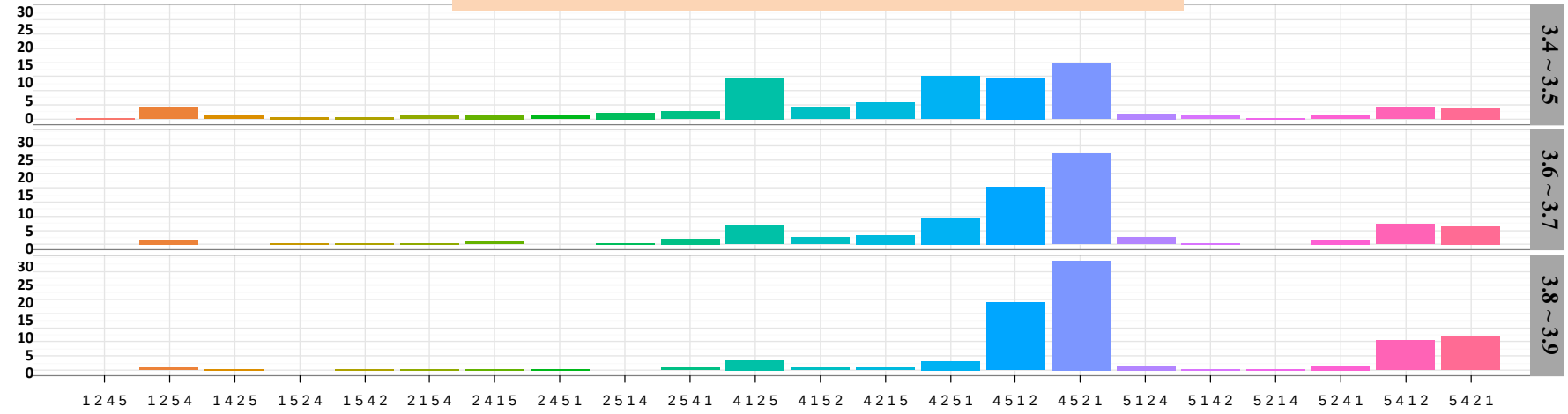
(noisy, but of reasonable quality (~74%), and cheap!)

Questions? Thank you!!!

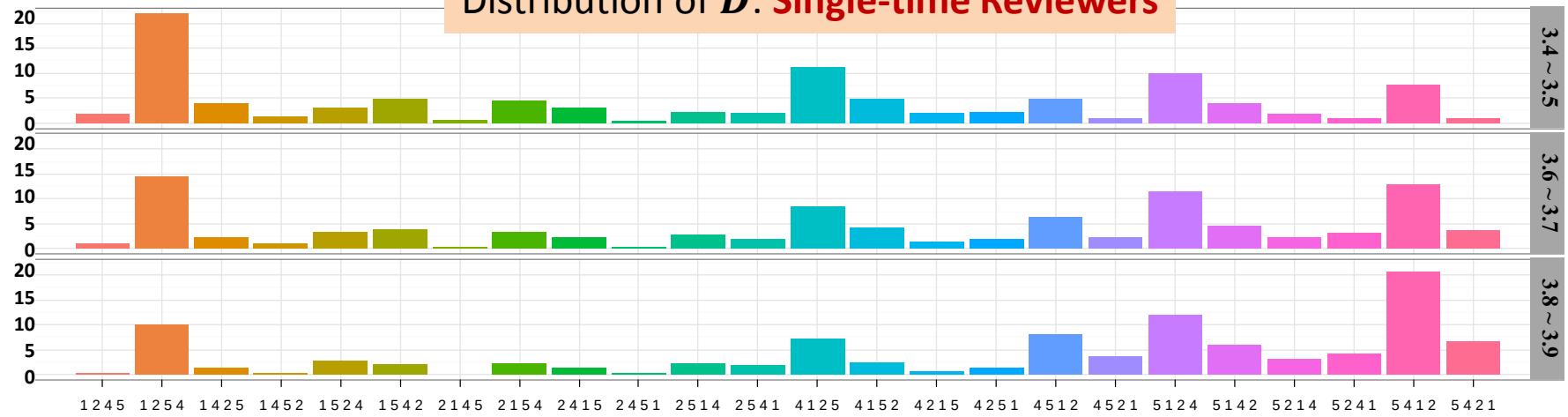
Distribution of Distribution



Distribution of $\widehat{D}$: Multi-time Reviewers



Distribution of $\widehat{D}$: Single-time Reviewers



Strategies: Pseudo-truthful reviewers

- ⊕ Number of reviews
  - #Historical reviews (= 10).
- ⊖ Review post dates (*Lim et al*. 2010)
  - ✧ Multiple posts in a very short period (within 2 days).
- ⊖ Rating discrepancy
  - ✧ Ratings are always greatly deviated from average rating ().

Strategies: Suspicious Reviews

Authorship-based (content independent)

✧ **Truthful reviews**

 Multi-time

or

 Pseudo-truthful

✧ **Deceptive reviews**

 Suspicious

and

 Single-time