Mayank Kejriwal
University of Southern California

Wrappers - II

Active Learning & Wrappers

- Active Learning
  - **Idea:** system selects most informative examples to label
  - **Advantage:** fewer examples to reach same accuracy
- Wrappers
  - One wrapper may use hundreds of extraction rules
    - Small reduction of *examples per rule* => big impact on user
  - Need more than 95% accuracy!
    - That would be 5% incorrect data
    - Select most informative examples to get to 100% accuracy

Which example should be labeled next?

SkipTo(**Phone:**)

Training Examples

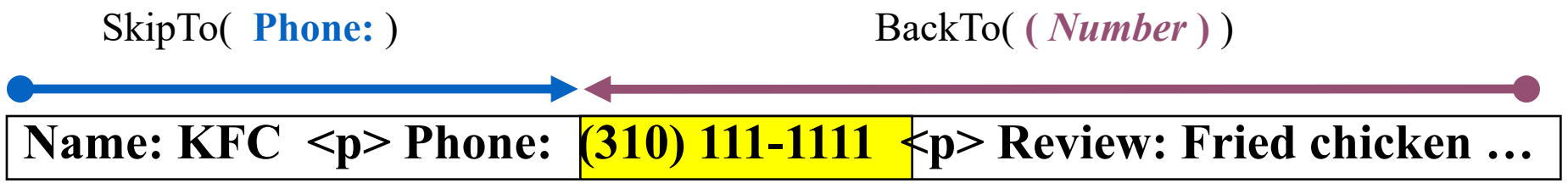
Name: Joel's <p> Phone: (310) 777-1111 <p>Review: The chef…
Name: Kim's <p> Phone: (213) 757-1111 <p>Review: Korean …

Unlabeled Examples

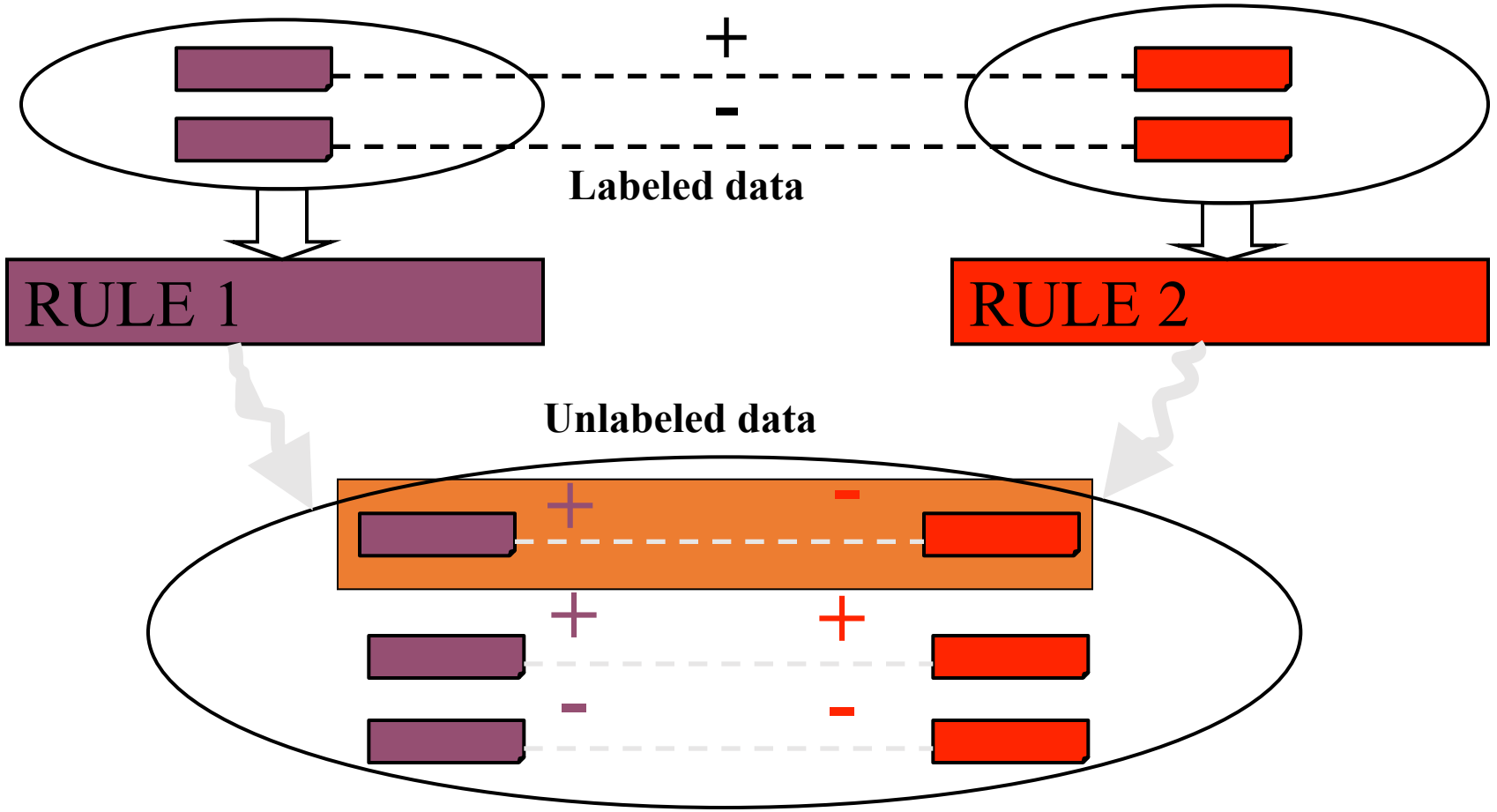
Name: Chez Jean <p> Phone: (310) 666-1111 <p> Review: …
Name: Burger King <p> Phone:(818) 789-1211 <p> Review: ...
Name: Café del Rey <p> Phone: (310) 111-1111 <p> Review: ...
Name: KFC <p> Phone: (800) 111-7171 <p> Review:...

Multi-view Learning

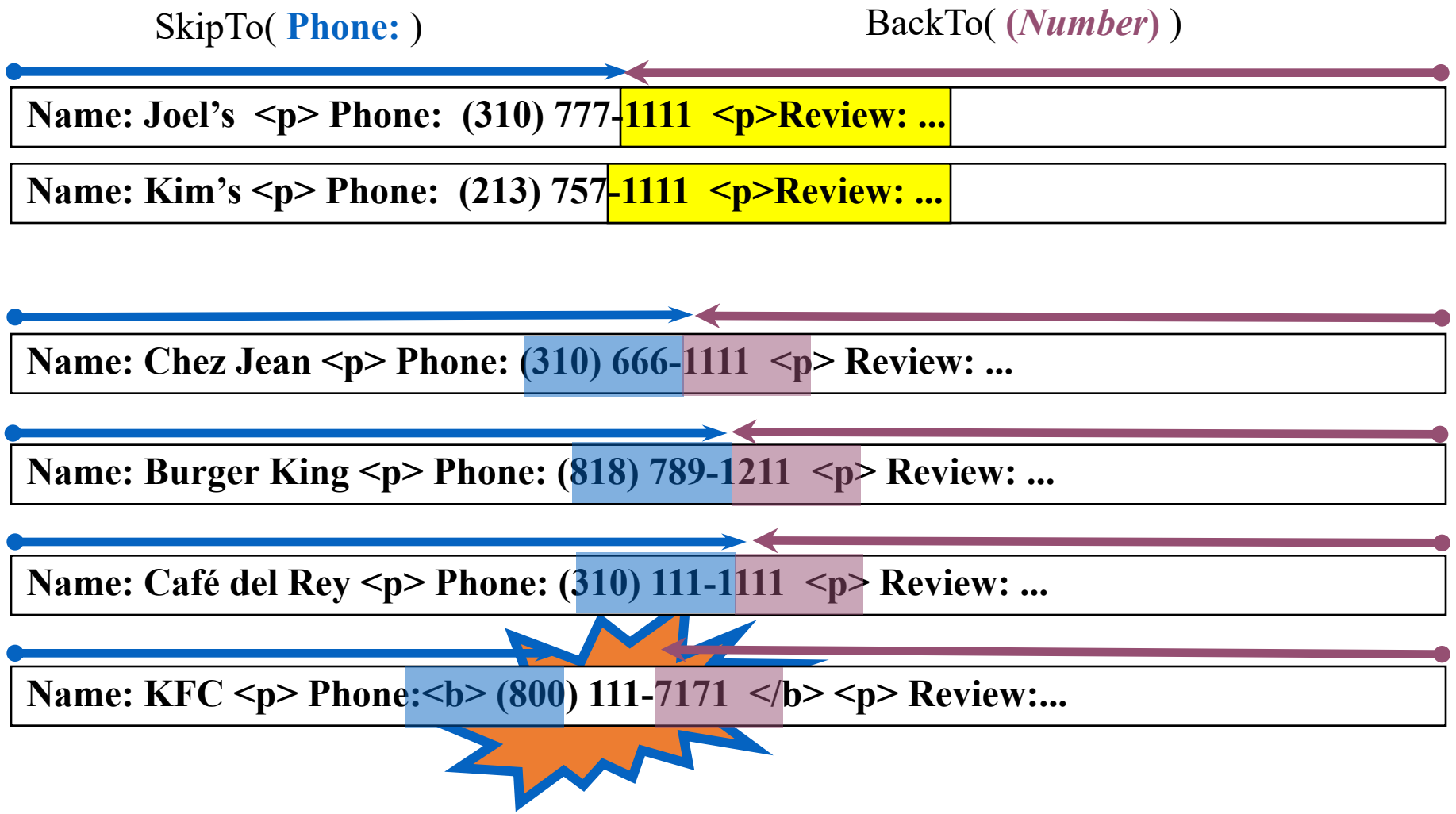
Two ways to find start of the phone number:



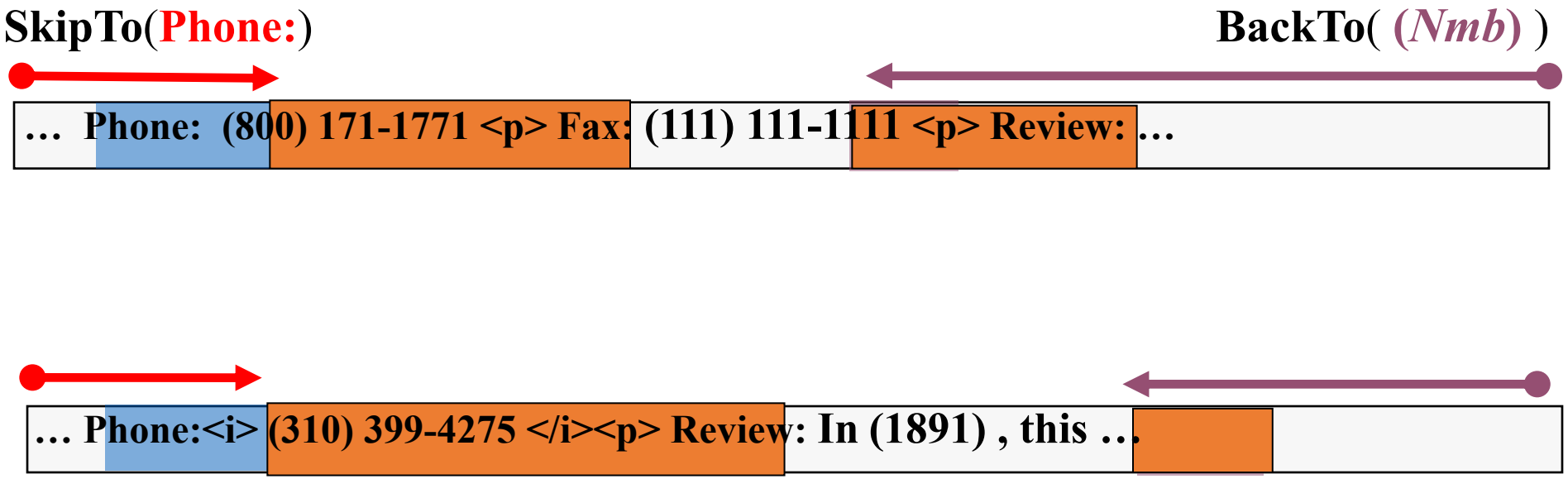
Multi-view Learning: Co-Testing



Co-Testing for Wrapper Induction



Not all queries are equally informative



Weak Views

- Learn “content description” for item to be extracted
  - Too general for extraction
    - ***(Nmb) Nmb – Nmb*** can’t tell a ***phone number*** from a ***fax number***
  - Useful at ***discriminating*** among ***query candidates***
  - Learned field description
    - Starts with: ***(Nmb)***
    - Ends with: ***Nmb – Nmb***
    - Contains: ***Nmb Punct***
    - Length: **[6,6]**

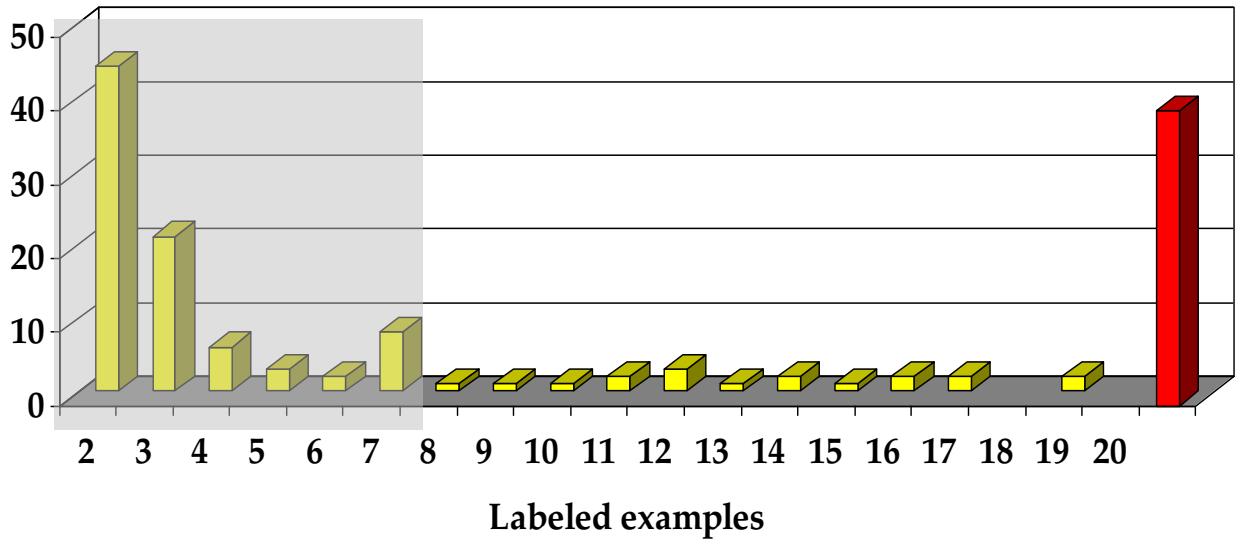
Naïve & Aggressive Co-Testing

- Naïve Co-Testing:
  - Query: randomly chosen contention point
  - Output: rule with fewest mistakes on queries

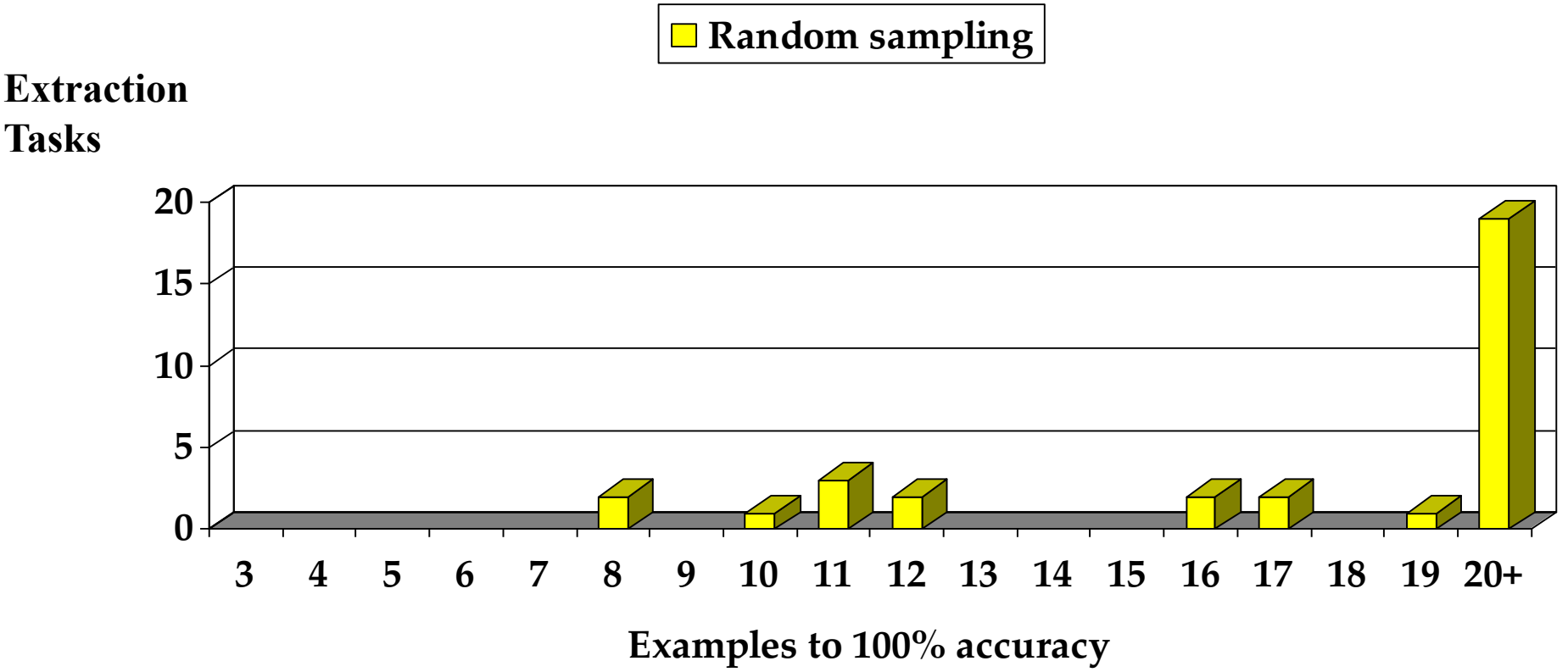
- Aggressive Co-Testing:
  - Query: contention point that most violates weak view
  - Output: committee vote (2 rules + weak view)

Empirical Results: 33 Difficult Tasks

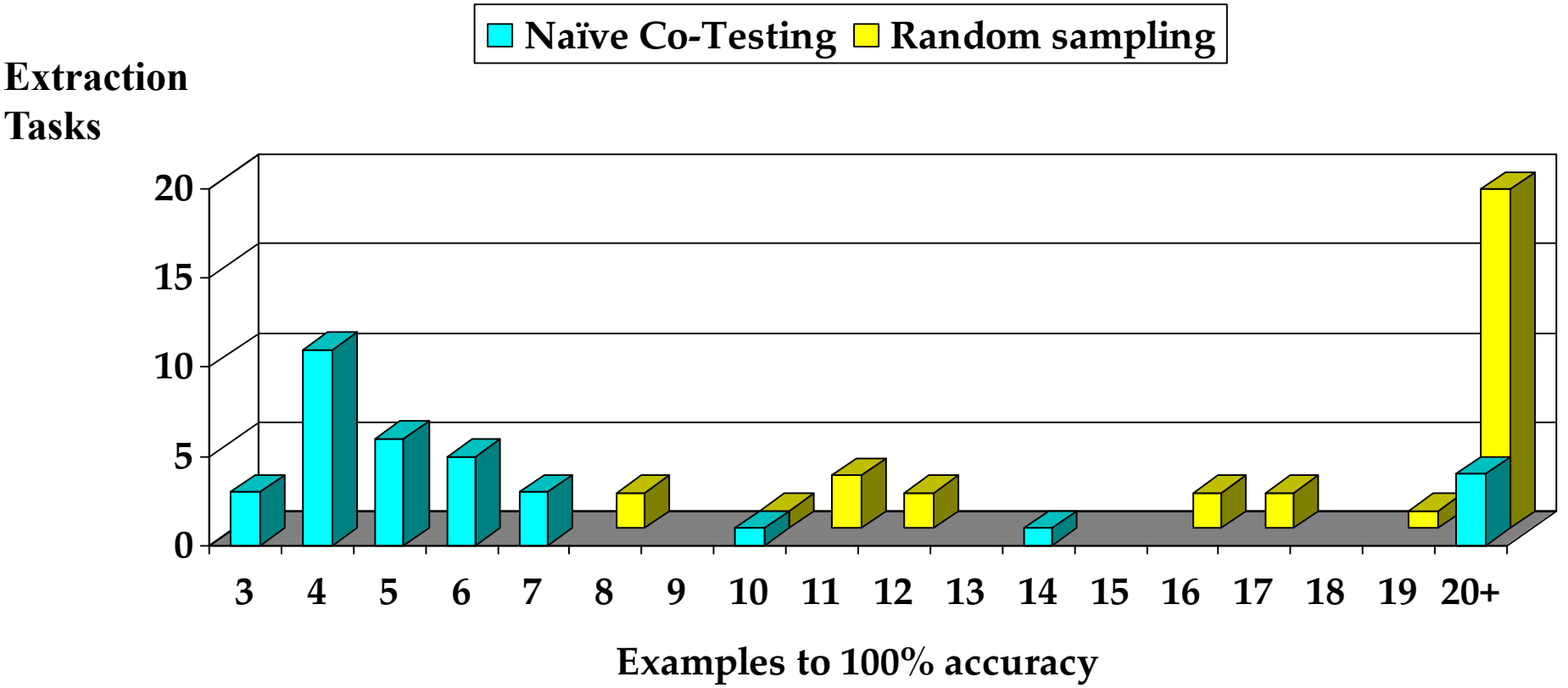
- 33 ***most difficult*** of the 140 extraction tasks
  - Each view: > 7 labeled examples for best accuracy
  - At least 100 examples for task



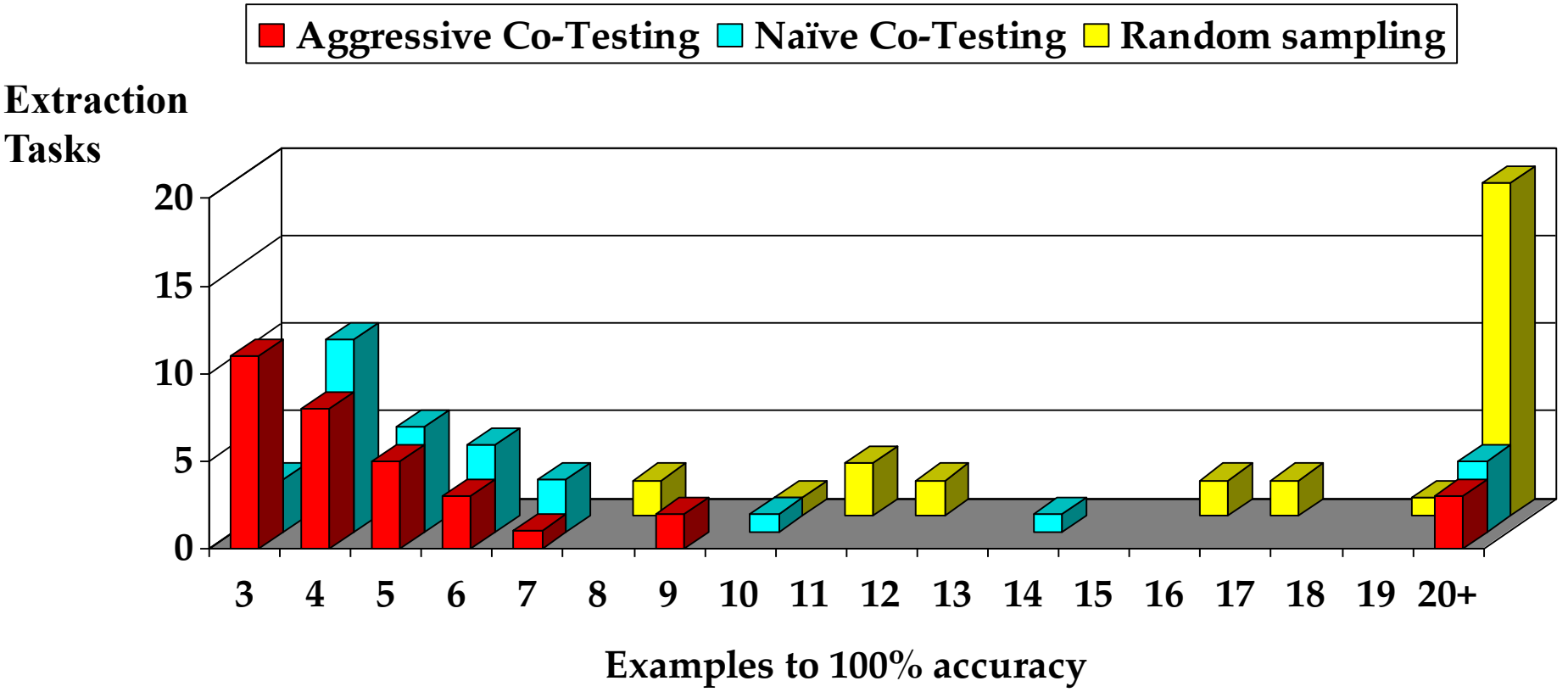
Results in 33 Difficult Domains



Results in 33 Difficult Domains



Results in 33 Difficult Domains



Summary

- Advantages:
  - Powerful extraction language (eg, embedded list)
  - One hard-to-extract item does not affect others
- Disadvantage:
  - Does not exploit item order (sometimes may help)