

A decorative graphic on the left side of the slide, consisting of a black crosshair intersecting a blue square, a red square, and a yellow square.

A Bootstrapping Approach for Stakeholder Identification in Public Comment Corpora

Jaime Arguello & Jamie Callan

International Conference in Digital Government Research

May 22, 2007



Motivation

- Authors of public comments differ from one another in meaningful ways:
 - Level of impact
 - “I am a grandparent of children with environmentally aggravated illness.”
 - “As a grandmother of 12 wonderful children, I urge you ...”
 - Expertise
 - “As a chemist that works in the area of fish analysis for mercury, I am concerned ...”
 - “Even as a lay person, I know that mercury is one of the most toxic substances known.”
 - Geographical location
 - “I am a resident of California and someone who is aware ...”
 - “As a resident of a city with high mercury levels I am concerned about ...”
 - Level of representation
 - In EPA’s Mercury Corpus, > 10% parents, of which 80% mothers
 - In DOI’s Polar Bear Corpus, only 0.4% parents, but 77% wildlife advocates



Motivation

- Government agencies may want to know who is represented in a corpus of public comments and may want to focus their attention on the subsets of comments from particular communities
- With only a search interface, this is difficult to do in large corpora
 - EPA's Mercury Corpus: 500,000 docs (12,249 stakeholders)
- Our goal is to provide:
 - a “bird's eye view” of the stakeholder landscape
 - a navigational aid

[a former chemist \(2\)](#)
[a former chemist for cleveland air pollution control \(1\)](#)
[practicing chemist \(3\)](#)
[a practicing chemist \(2\)](#)
[a practicing chemist for over 25 years \(1\)](#)
[research chemist \(3\)](#)
[a research chemist \(2\)](#)
[a research chemist in industry \(1\)](#)
[retired research chemist \(3\)](#)
[a retired research chemist \(2\)](#)
[a retired research chemist with years of experience in](#)

economist (5)

[an economist \(3\)](#)
[a former economist with the epa \(1\)](#)
[a professional economist with extensive business exper](#)

environmental scientist (24)

[an environmental scientist \(18\)](#)
[an environmental scientist who had completed air permits](#)
[an environmental scientist formerly from nj \(2\)](#)
[an environmental scientist with 30 years of experience in t](#)
[an environmental scientist at cornell university \(1\)](#)

epidemiologist (12)

[a public health nurse epidemiologist \(2\)](#)
[an epidemiologist from new york medical college \(1\)](#)
[a chronic disease epidemiologist \(1\)](#)
[an epidemiologist \(1\)](#)
[a physician and epidemiologist \(1\)](#)
[pediatrician and epidemiologist \(2\)](#)
[a pediatrician and epidemiologist with long standing co](#)
[a pediatrician and epidemiologist \(1\)](#)
[phd epidemiologist \(4\)](#)
[a phd epidemiologist \(2\)](#)
[a phd epidemiologist who teaches and does research o](#)

geologist (4)

[a geologist \(3\)](#)
[a retired geologist \(1\)](#)

physicist (7)

[a physicist \(2\)](#)
[a research physicist \(1\)](#)
[a professional physicist \(1\)](#)
[a physicist with long experience of the dangers of merc](#)
[astronomer \(2\)](#)
[an amateur astronomer \(2\)](#)

psychologist (62)

[a psychologist \(16\)](#)
[a school psychologist \(12\)](#)
[a child psychologist \(8\)](#)
[a psychologist in michigan and work with special educ](#)
[a licensed psychologist \(2\)](#)
[a psychologist who works with developmentally disable](#)
[a health psychologist \(1\)](#)
[a psychologist working with infants and toddlers \(1\)](#)
[developmental/cognitive psychologist \(1\)](#)
[a licensed child psychologist in fl \(1\)](#)
[a clinical developmental psychologist \(1\)](#)
[developmental psychologist \(15\)](#)
[a developmental psychologist \(12\)](#)
[a developmental psychologist who has worked with ma](#)
[a developmental psychologist who has been watching t](#)

research scientist (8)

[a research scientist \(4\)](#)
[a research scientist in the pharmaceutical industry \(2\)](#)
[a research scientist who studied the effects of mercury con](#)
[a research scientist at the university of delaware \(1\)](#)

research worker (22)

[a professional health researcher \(2\)](#)
[a scientist and professional researcher of pharmacolog](#)
[the american a former field investigator for the epa in r](#)
[a neuroscience researcher at the university of californ](#)
[biomedical researcher \(2\)](#)

environmental scientist (24)

[an environmental scientist \(18\)](#)

[an environmental scientist who had completed air permits for major power plants \(2\)](#)

[an environmental scientist formerly from nj \(2\)](#)

[an environmental scientist with 30 years of experience in the environmental field \(1\)](#)

[an environmental scientist at cornell university \(1\)](#)

epidemiologist (12)

[a public health nurse epidemiologist \(2\)](#)

[an epidemiologist from new york medical college \(1\)](#)

[a chronic disease epidemiologist \(1\)](#)

[an epidemiologist \(1\)](#)

[a physician and epidemiologist \(1\)](#)

pediatrician and epidemiologist (2)

[a pediatrician and epidemiologist with long standing concerns about environmental pollution \(1\)](#)

[a pediatrician and epidemiologist \(1\)](#)

phd epidemiologist (4)

[a phd epidemiologist \(2\)](#)

[a phd epidemiologist who teaches and does research on children s envirnomenal health \(2\)](#)

geologist (4)

[a geologist \(3\)](#)

[a retired geologist \(1\)](#)

physicist (7)

[a physicist \(2\)](#)

[a research physicist \(1\)](#)

[a professional physicist \(1\)](#)

[a physicist with long experience of the dangers of mercury \(1\)](#)

[astronomer \(2\)](#)



Stakeholder definition

- Stakeholder = a group or community of which the author is a member (mentioned explicitly)
 - As a former employee of a power company, I know ...
 - As a woman of child-bearing age, I now have ...
 - I am an avid fisherman and I do not agree ...
- Sounds subjective?
 - Human agreement (f-measure): **overlap** = 0.70, **exact** = 0.53
 - Open question: what's the extent of the stakeholder mention?
 - “As a person that spent 2 two years recovering from mercury and arsenic poisoning, I am more than appalled that you ...”
 - “As a person that spent 2 two years recovering from mercury and arsenic poisoning, I am more than appalled that you ...”
- Stakeholder types **not** addressed
 - Impacted group (not necessarily the author) (e.g., children)
 - Entities being regulated (e.g., coal-fired power plants)



Bootstrapping Approaches for Information Extraction: The General Idea

- **Input:** seeds
 - “a mother”
- Find all contexts where seeds occur
 - “As a mother, I worry about mercury pollution.”
- Extract patterns from contexts
 - “As _____, I worry about mercury pollution.”
- **Score and keep only the best!!!**
- Find other noun-phrases (NPs) that co-occur with patterns
 - “As a neurologist, I worry about mercury pollution.”
- **Score and keep only best!!!**
- Add NPs to seeds
 - “a mother”, “a neurologist”
- Repeat

Bootstrapping Approaches for Information Extraction: Ingredients

- The bootstrapping framework

- A scoring function for

- Patterns
- Entities



- A representation of context

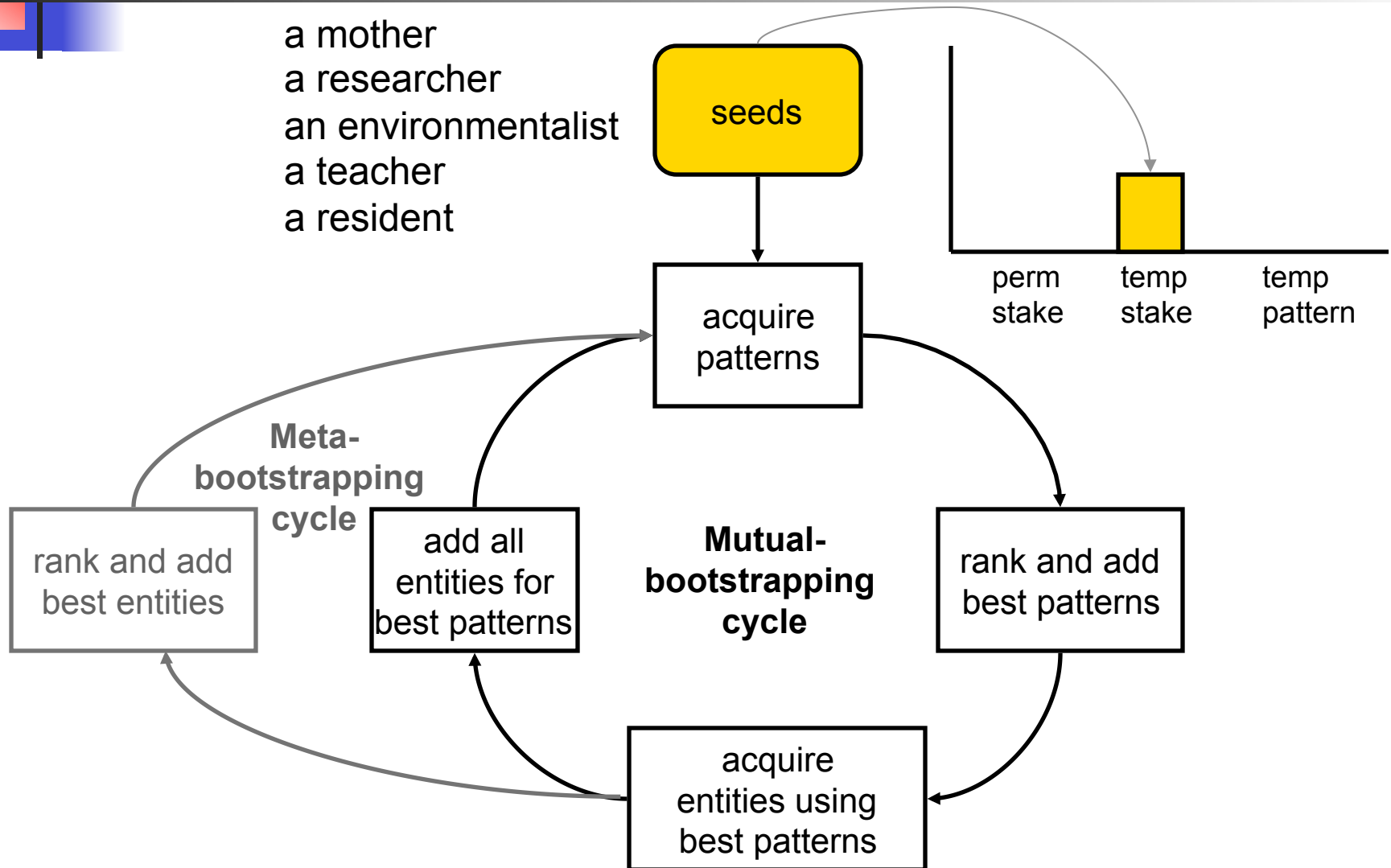
- “As _____, I worry about mercury pollution.”

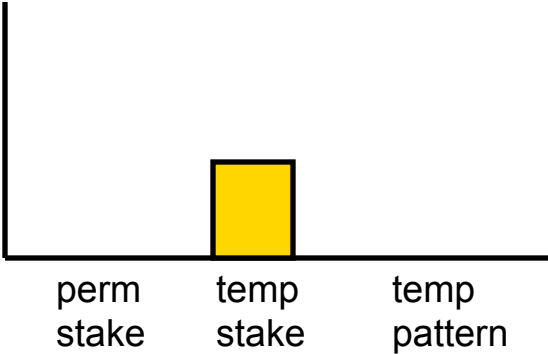


- Definition: extraction pattern template (**EPT**)
- What features of the context are the most informative for discriminating b/w target (**stakeholder**) and non-target (**non-stakeholder**) NPs?

Meta (2-cycle) Bootstrapping Approach

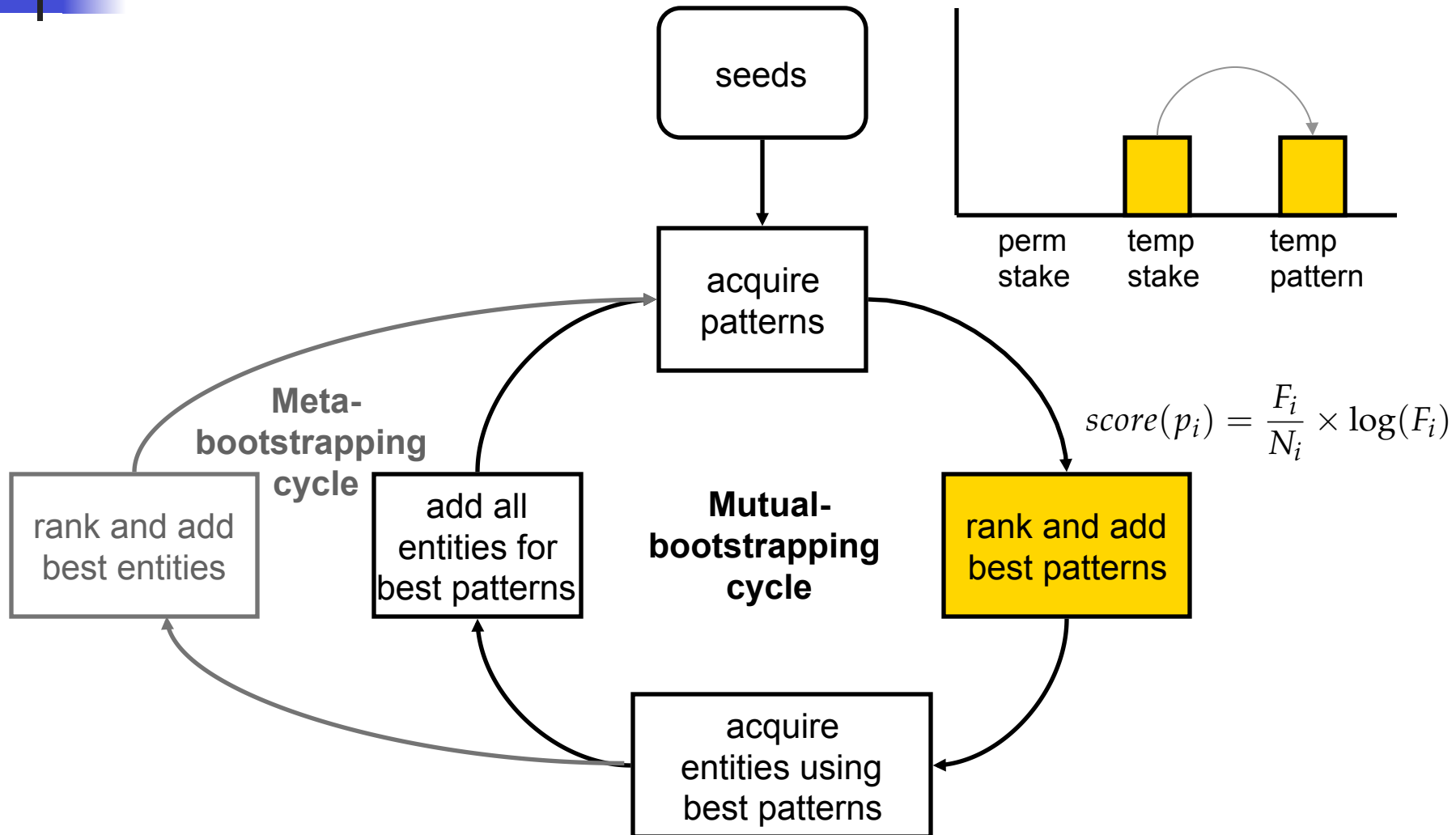
a mother
a researcher
an environmentalist
a teacher
a resident



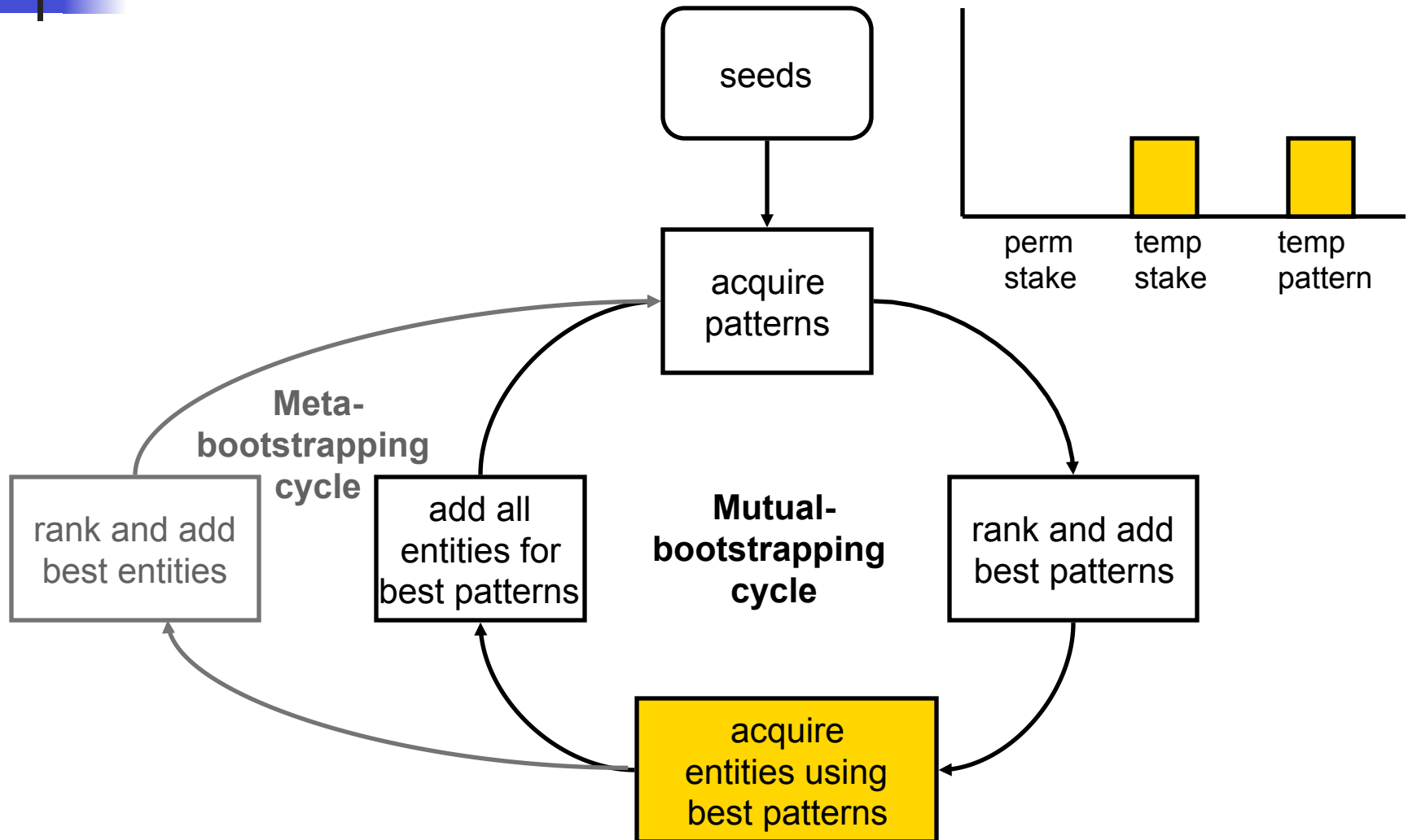


(Riloff & Jones, 99)

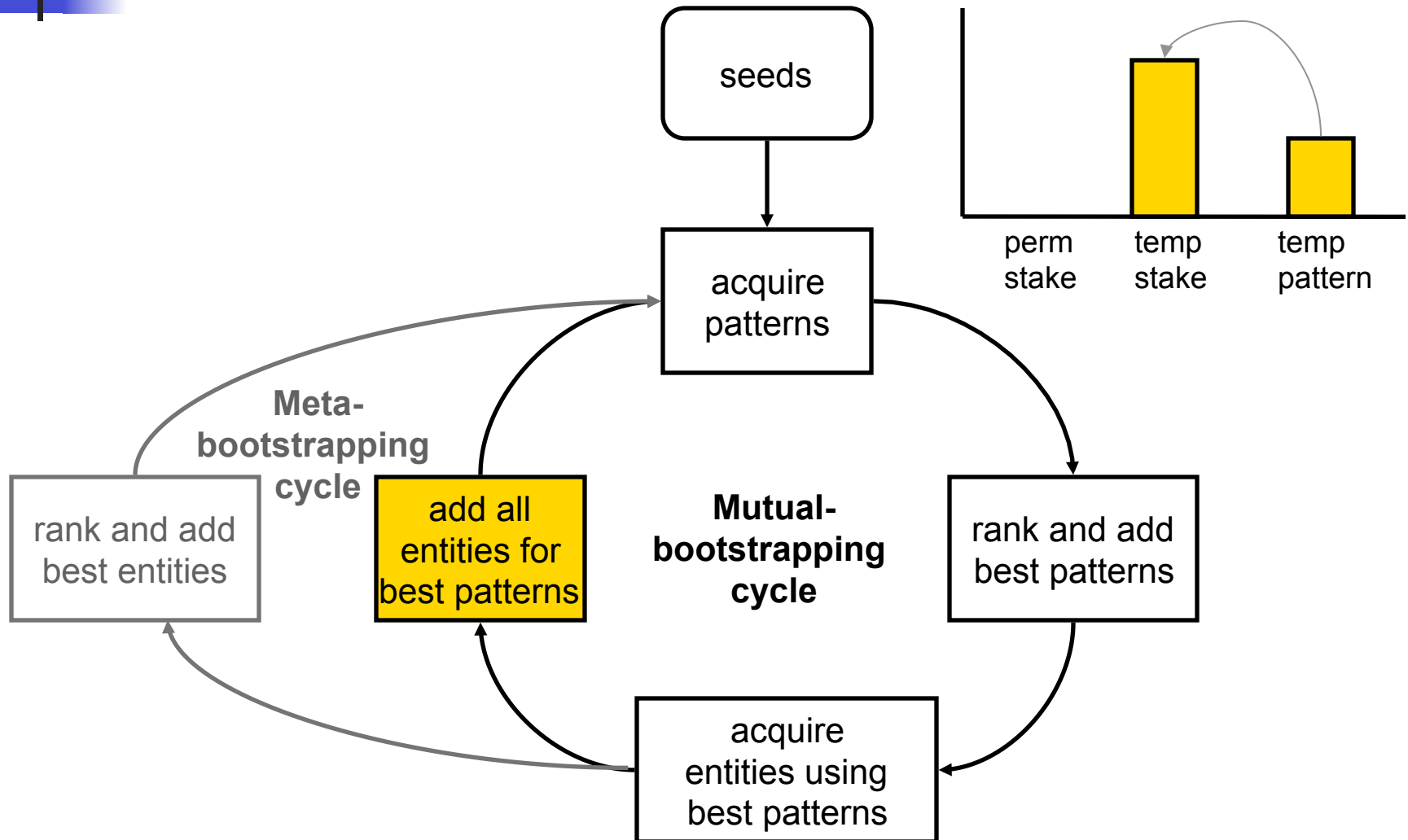
Meta (2-cycle) Bootstrapping Approach



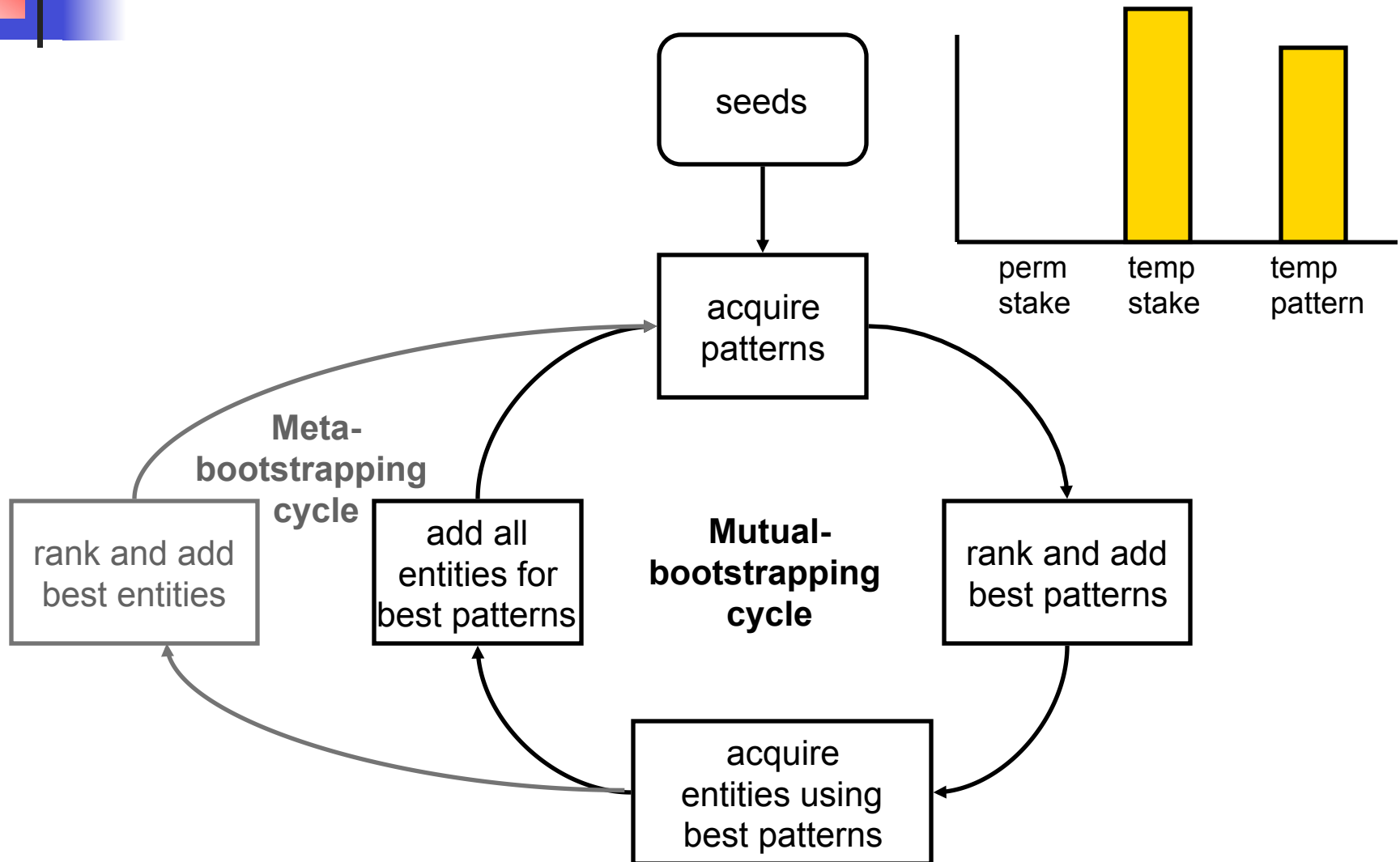
Meta (2-cycle) Bootstrapping Approach



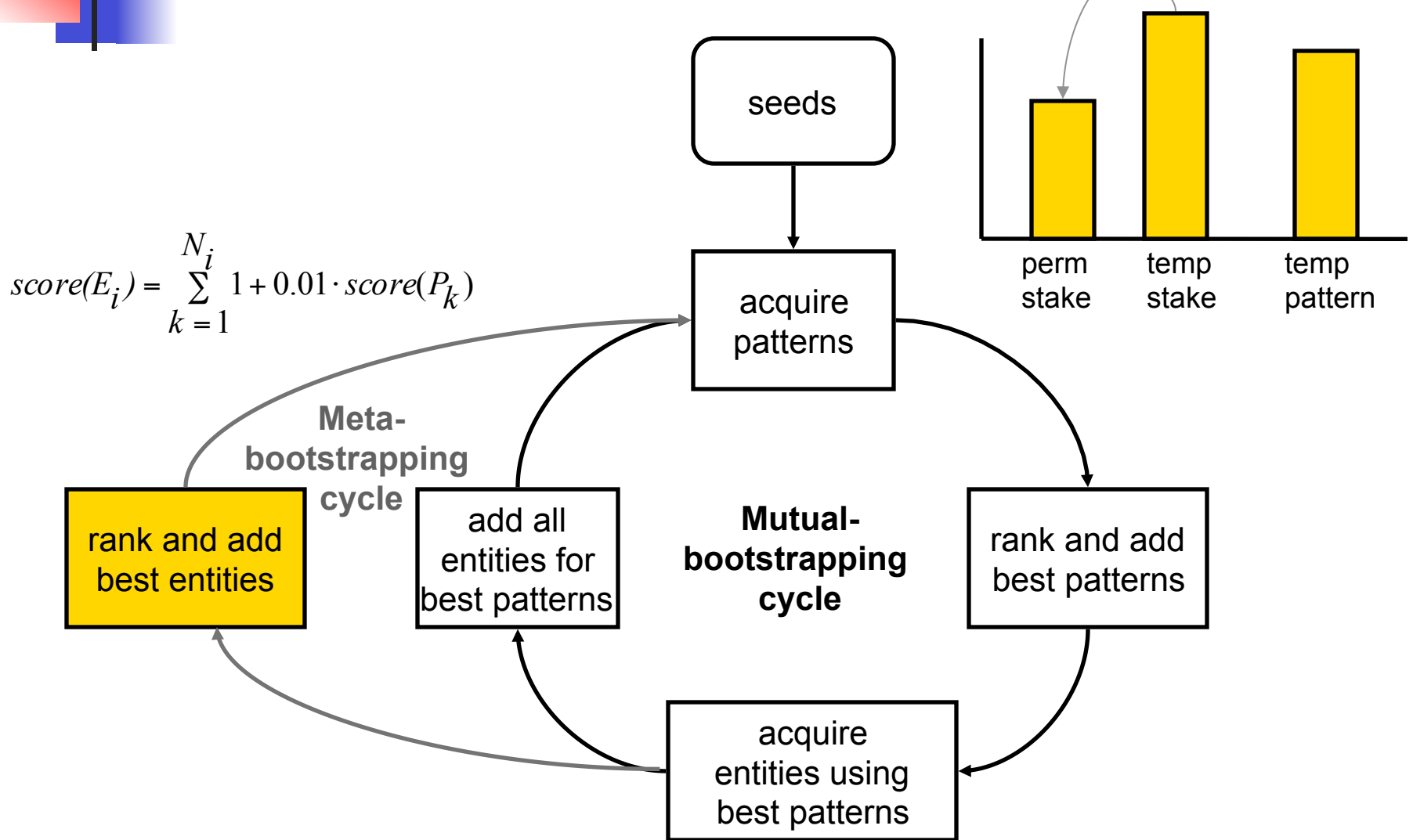
Meta (2-cycle) Bootstrapping Approach



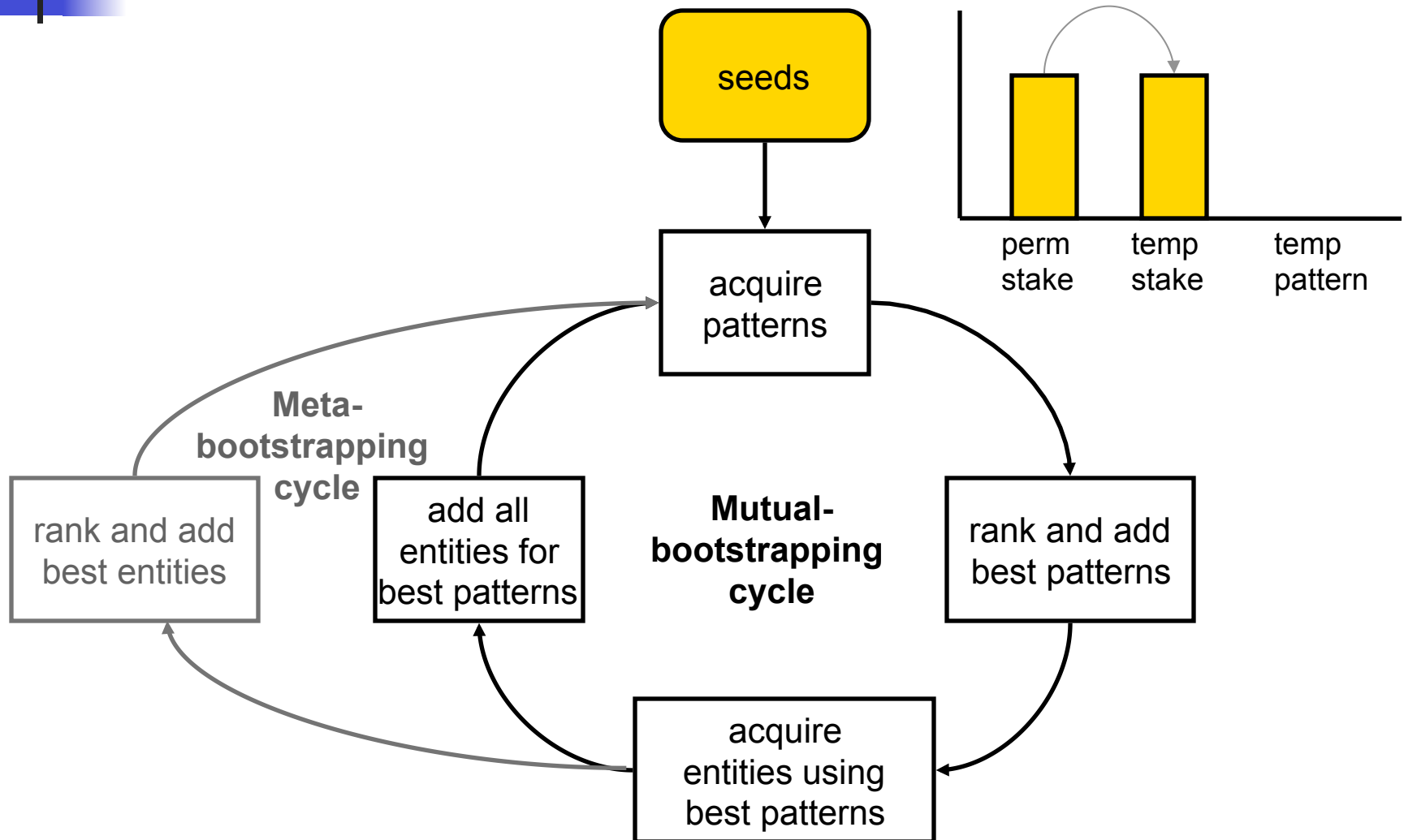
Meta (2-cycle) Bootstrapping Approach



Meta (2-cycle) Bootstrapping Approach



Meta (2-cycle) Bootstrapping Approach





3 Extraction Patterns Templates

1) Surface-based patterns

- $W_{L1} W_{L2} \text{ ______ } W_{R1} W_{R2}$
- $W_{XY} = \{\mathbf{word}, \mathbf{pos}\}$

2) WordNet- (WN) based patterns

- $W_{L1} W_{L2} \text{ ______ } W_{R1} W_{R2}$
- $W_{XY} = \{\mathbf{pos}\}$, *except if...*
 - (1) $\text{part-of-speech}(W_{XY}) = \text{pronoun}$ - to avoid conflating 1st & 2nd person pronouns
 - “You are the agency responsible for protecting the environment, and I ...”
 - (2) $\text{part-of-speech}(W_{LY}) = \text{verb}$, to avoid conflating verbs that refer to the author and verbs that refer to a 3rd entity
 - “I have a child that suffers from mercury contamination, and I ...”
- Stakeholder subsumed by “person” in WordNet



3 Extraction Pattern Templates (cont' d)

3) Sundance-based patterns

- Sundance information extraction (IE) engine (chunking, shallow parsing, pronoun resolution, etc.) Used in prior work to extract: person, building name, victim, weapon, location, company name, professional title, ...
- **17** templates (**15** verb-centric, **2** noun-centric), e.g.:

<subject> <aux. verb> <direct object> → <stakeholder> have friends

<subject> <aux. verb> <adjective phrase> → <stakeholder> am very concerned

<subject> <active infinitive verb> → <stakeholder> needs to protect

<subject> <active verb> → <stakeholder> wishes

<infinitive verb> <preposition> <np> → worry as <stakeholder>

- Main assumption: target class NPs occur as arguments of some verbs more than others (?)
 - For, “person”, perhaps (e.g., subject of verbs like “*feel*”, “*think*”, “*say*”)
 - But, for “stakeholders”?



Heuristics (Surface- and WN-based patterns)

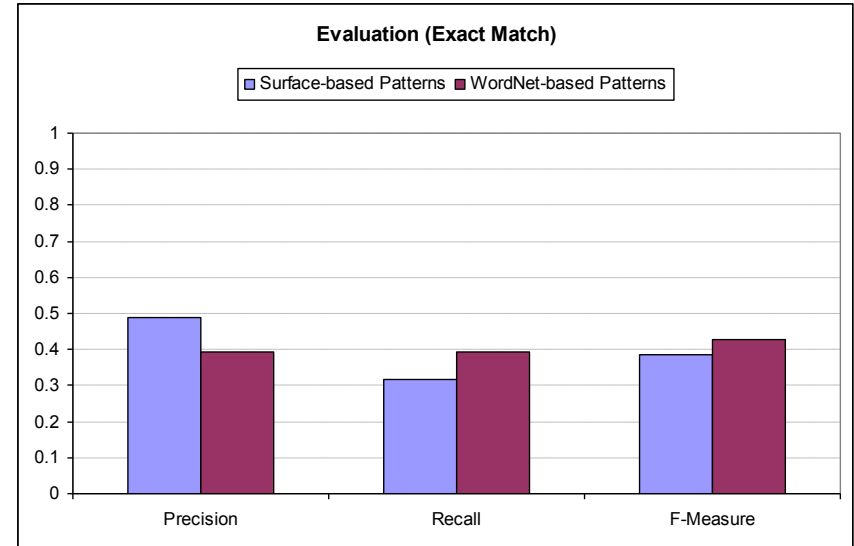
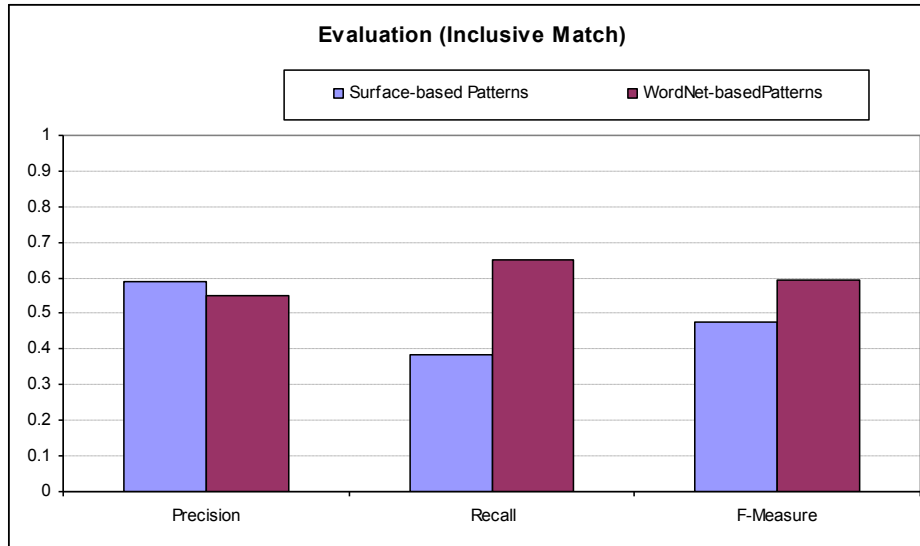
- NP-expansion (greedy NP chunker)
 - “I **am** a mother **of two** boys and I think that”
 - “I **am** a mother of two boys **and I** think”
- Head-NP querying
 - Long stakeholder NPs are rare
 - “a person who has children living by the infamous 4 corners power plant in New Mexico”
- List handling
 - “I am a husband, a father, a teacher, and a concerned North American.”
- Adjective/Adverb padding
 - Allow optional adverb/adjective within learned patterns
 - “I am *practically* an advocate for the environment.”



Evaluation Methodology

- EPA's Mercury Corpus: > 500,000 documents
- Unsupervised Training
 - 120,000 documents, after duplicate detection
 - 5 seeds: **a biologist, an environmentalist, a resident, a citizen, an American**
 - Best 80 learned extraction patterns applied to test set
- Test set (annotated)
 - 1,020 documents
 - 60 stakeholders detected (about 1 in 20 documents)
- Measures: precision (P), recall (R), f-measure (F1):
 - Exact Match: reference NP = predicted NP
 - Inclusive Match: (reference NP within predicted NP) **OR**
(predicted NP within reference NP)

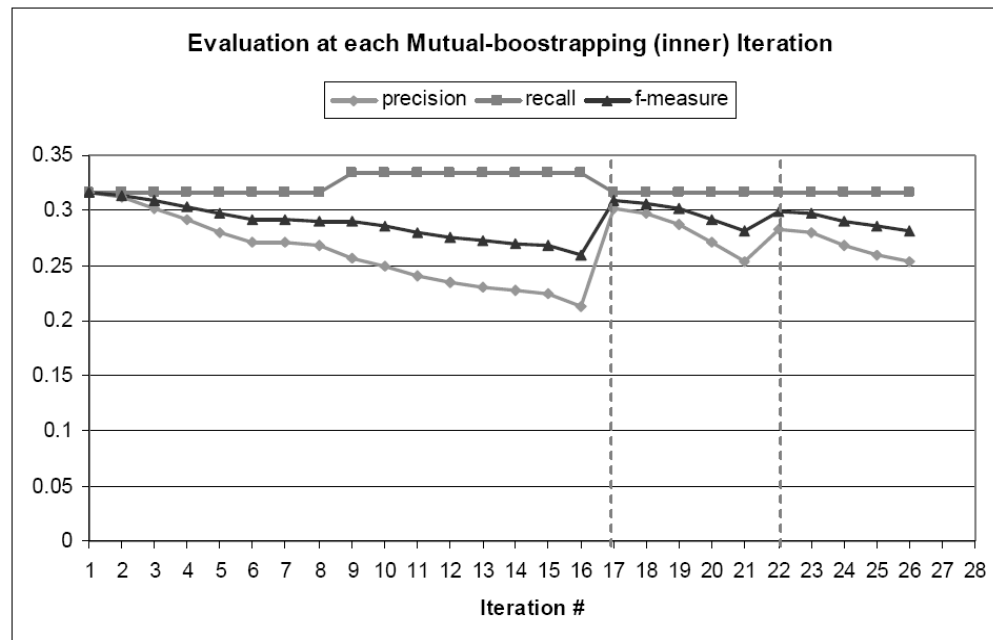
Results (Surface- & WN-based patterns)



- Surface-based patterns suffer from low generalization, but have slightly higher precision
- WordNet-based patterns
 - Achieve higher recall by generalizing from the local context
 - Avoid loss in precision by imposing semantic constraint on extracted NP (NP is a “person”)

Results (Sundance-based patterns)

- None of the learned patterns occur in the test set
- Alternate evaluation
 - The system with seeded with all entities extracted by Sundance-based pattern “I am <dobj>”
 - best pattern on test set
 - Maximum advantage to perform well on this test set





WN-based error analysis

- Effective extraction patterns follow heavy-tailed distribution
 - A few patterns extract many stakeholders and many patterns extract only a few stakeholders
- Bootstrapping algorithm rejects rare patterns due to low recall
- This is representation challenge
 - Local context should be overlooked in the presence of more meaningful long-distance evidence
 - “As a mother, **the effect of mercury on children** concerns me.”
- Ambiguous contexts
 - “As a policy maker, I urge you to take immediate action”



Sundance-based patterns error analysis

- Pronoun resolution
 - “As a [[former employee]_{NP} of [the power industry]_{NP}]_{NP}, **I know** there
...”
 - “Being a [[child development]_{NP} and [healthcare specialist]_{NP}]_{NP}, **I am**
...”
 - <subj> ActVp → I (NP antecedent) **know/am**
- Verb-centric assumption doesn't hold for stakeholders
 - Stakeholders do not occur exclusively as the subject/object of a verb set
 - “As a mother, I care about ...”
 - “This government cares only about ...”



Conclusions

- Stakeholder mentions can be identified with about .60 precision/recall in a bootstrapping framework.
- About 40% of the stakeholders are being missed
 - Rare contexts
 - “As a mother, the effects of mercury on children concerns me.”
 - Ambiguous contexts
 - “As the agency responsible for the environment, I think ...”
 - Implicit mentions
 - “I live 5 miles away from a coal-fired power plant.”
- What now?
 - Not committing to a single context representation
 - Organizing stakeholders (> 12,000 stakeholders in Mercury corpus)
 - Component technology for other text mining apps (e.g., sentiment analysis, identifying constructive comments)