Relation Extraction

Bill MacCartney

CS224U

23 April 2018

[with slides adapted from many people, including Dan Jurafsky, Rion Snow, Jim Martin, Chris Manning, William Cohen, Michele Banko, Mike Mintz, Steven Bills, and others]

Goal: ”machine reading”

Reading the Web: A Breakthrough Goal for AI

I believe AI has an opportunity to achieve a true breakthrough over the coming decade by at last solving the problem of reading natural language text to extract its factual content. In fact, I hereby offer to bet anyone a lobster dinner that by 2015 we will have a computer program capable of automatically reading at least 80% of the factual content [on the] web, and placing those facts in a structured knowledge base. The significance of this AI achievement would be tremendous: it would immediately increase by many orders of magnitude the volume, breadth, and depth of ground facts and general knowledge accessible to knowledge based AI programs. In essence, computers would be harvesting in structured form the huge volume of knowledge that millions of humans are entering daily on the web in the form of unstructured text.

— Tom Mitchell, 2004

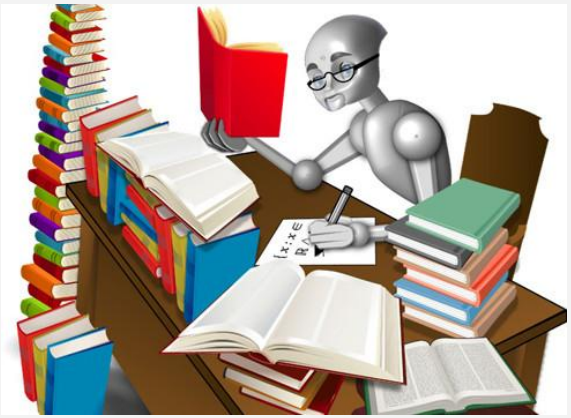


illustration from DARPA

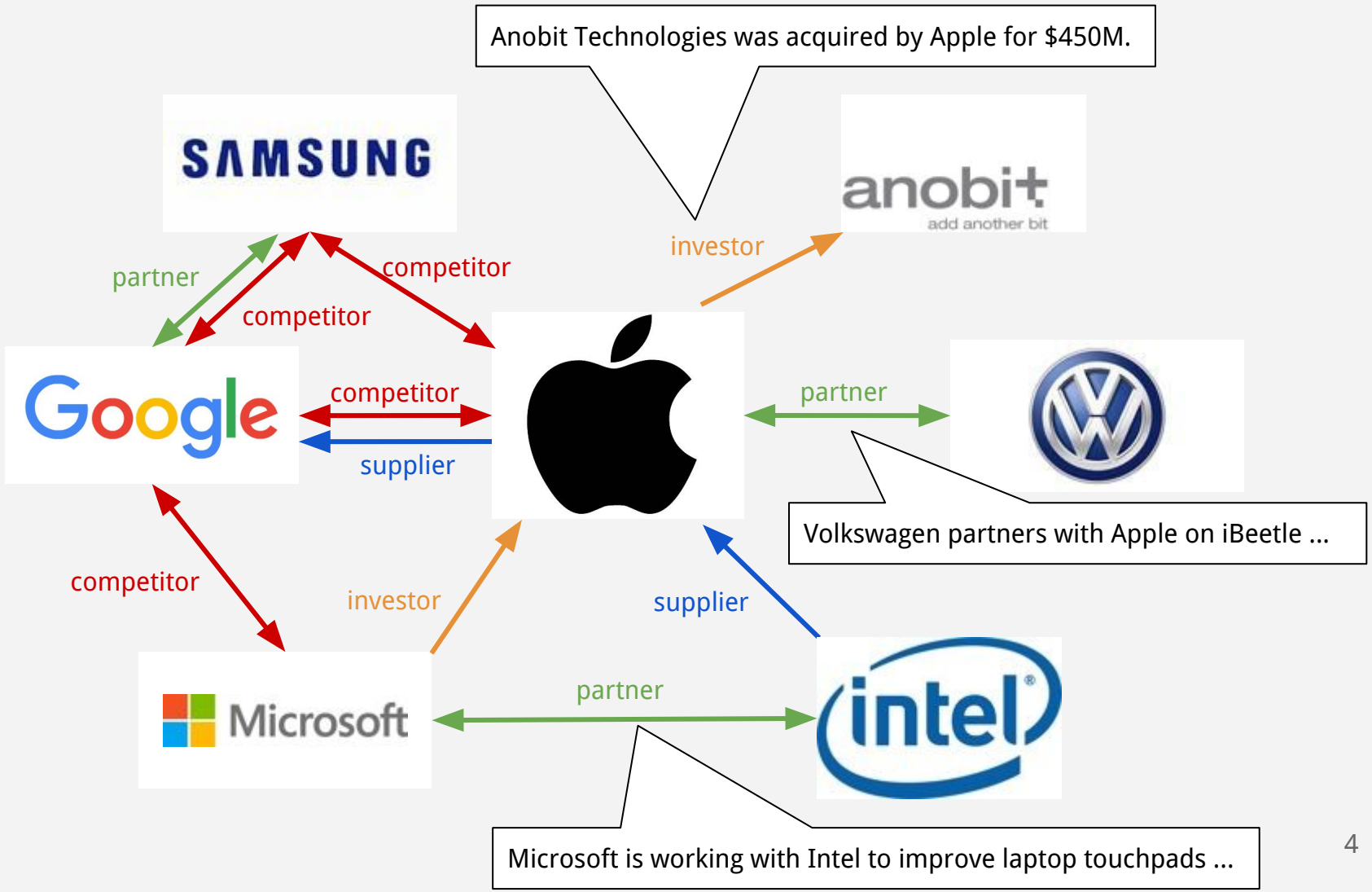
Relation extraction example

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Subject	Relation	Object
American Airlines	subsidiary	AMR
Tim Wagner	employee	American Airlines
United Airlines	subsidiary	UAL

example from Jim Martin

Example: company relationships



Example: gene regulation


Involvement of Tumor Necrosis Factor Receptor-associated Protein 1 (TRAP1) in Apoptosis Induced by β-Hydroxyisovalerylshikonin*

Subject	Relation	Object
p53	**is_a**	protein
Bax	**is_a**	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...	...	...

textual abstract:
summary for human

structured knowledge extraction:
summary for machine

Lexical semantic relations

Many NLP applications require understanding relations between word senses: synonymy, antonymy, hyponymy, meronymy.

WordNet is a machine-readable database of relations between word senses, and an indispensable resource in many NLP tasks.

http://wordnetweb.princeton.edu/perl/webwn

```
vehicle
    craft
        aircraft
            airplane
            dirigible
            helicopter
        spacecraft
        watercraft
            boat
            ship
            yacht
    rocket
        missile
        multistage rocket
    wheeled vehicle
        automobile
        bicycle
        locomotive
        wagon
```

WordNet is incomplete

But WordNet is manually constructed, and has many gaps!

In WordNet 3.1	Not in WordNet 3.1
insulin progesterone	leptin pregnenolone
combustibility navigability	affordability reusability
HTML	XML
Google, Yahoo	Microsoft, IBM

Esp. for specific domains: restaurants, auto parts, finance

Esp. neologisms: iPad, selfie, bitcoin, twerking, Hadoop, dubstep

Example: extending WordNet

Mirror ran a headline questioning whether the killer's actions were a result of playing Call of Duty, a first-person shooter game ...

Melee, in video game terms, is a style of elbow-drop hand-to-hand combat popular in first-person shooters and other shooters.

Tower defense is a kind of real-time strategy game in which the goal is to protect an area/place/locality and prevent enemies from reaching ...

```
video game
  action game
    ball and paddle game
      Breakout
    platform game
      Donkey Kong
    shooter
      arcade shooter
        Space Invaders
      first-person shooter
        Call of Duty
      third-person shooter
        Tomb Raider
  adventure game
    text adventure
    graphic adventure
  strategy game
    4X game
      Civilization
    tower defense
      Plants vs. Zombies
```

Example: extending Freebase

Freebase: 20K relations, 40M entities, 70B facts

Curation is an ongoing challenge — things change!

Relies heavily on relation extraction from the web

/film/film/starring

Wonder Woman	Gal Gadot
Dunkirk	Tom Hardy
Tomb Raider	Alicia Vikander

/organization/organization/parent

tbh	Facebook
Kaggle	Google
LinkedIn	Microsoft

/music/artist/track

Frank Ocean	Chanel
Cardi B	Bodak Yellow
Selena Gomez	Bad Liar

/people/person/date_of_death

Barbara Bush	2018-04-17
Milos Forman	2018-04-14
Winnie Mandela	2018-04-11

Approaches to relation extraction

1. Hand-built patterns
2. Bootstrapping methods
3. Supervised methods
4. Distant supervision
5. Other related work

Approaches to relation extraction

1. Hand-built patterns
2. Bootstrapping methods
3. Supervised methods
4. Distant supervision
5. Other related work

Patterns for learning hyponyms

- Intuition from Hearst (1992)

  Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

- What does *Gelidium* mean?
- How do you know?

Patterns for learning hyponyms

- Intuition from Hearst (1992)

  Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

- What does *Gelidium* mean?
- How do you know?

Hearst's lexico-syntactic patterns

Ys such as X ((, X)* (, and/or) X)

such Ys as X…

X… or other Ys

X… and other Ys

Ys including X…

Ys, especially X…

Hearst, 1992. Automatic Acquisition of Hyponyms.

Examples: “Ys, especially X”

The best part of the night was seeing all of the tweets of the performers, especially Miley Cyrus and Drake. ✓

Those child stars, especially Miley Cyrus, I feel like you have to put the fault on the media. ✓

Kelly wasn’t shy about sharing her feelings about some of the musical acts, especially Miley Cyrus. ✓

Rihanna was bored with everything at the MTV VMAs, especially Miley Cyrus. ✗

The celebrities enjoyed themselves while sipping on delicious cocktails, especially Miley Cyrus who landed the coveted #1 spot. ✗

None of these girls are good idols or role models, especially Miley Cyrus. ✗

Examples: “X was founded by Y”

NeXT was founded by Steve Jobs in 1985, after he was ousted from Apple Computers by John Sculley. ✓

Since 2002, when Blue Origin rival SpaceX was founded by Elon Musk, venture investment in the sector has increased markedly. ✓

Microsoft was founded by Paul Allen and Bill Gates on April 4, 1975, to develop and sell BASIC interpreters for the Altair 8800. ✓

The first successful commercial winery in New York was founded by Jean Jacques in 1839, at Washingtonville, on the west bank of the Hudson. ✗

One of the most obscure and fascinating companies implicated in the Panama Papers was founded by Jürgen Mossack in 1977. ✗

The largest annual space event on Earth was founded by the United Nations General Assembly and has been running every year since 1999. ✗

Examples: founder patterns

Elon Musk, the creator and founder of SpaceX, poked fun at the chaos his rocket launch caused for Californians on social media Friday night.

The co-founder of PayPal, Elon Musk, established SpaceX in 2002 with the goal of increasing space travel by reducing the cost of space launches.

Elon Musk co-founded PayPal and Tesla Motors, and created the space corporation SpaceX, which is credited with sending the first ...

SpaceX was founded in 2002 by entrepreneur Elon Musk with the goal of reducing space transportation costs.

Elon Musk is the founder, CEO and lead designer at Space Exploration Technologies (SpaceX), where he oversees ...

When Elon Musk first founded his rocket-ship company SpaceX, he had no idea how it would make a profit.

Problems with hand-built patterns

- Recall is not that great
  - Any finite set of patterns will fail to match many potential extractions
- Precision is not great either!
  - Many pattern-driven extractions are just wrong
  - Hearst: 66% accuracy on hyponym extraction
- Requires hand-building patterns for each relation!
  - And for every language!
  - Hard to write; hard to maintain

Approaches to relation extraction

1. Hand-built patterns
2. Bootstrapping methods
3. Supervised methods
4. Distant supervision
5. Other related work

Bootstrapping approaches

- If you have:
  - some seed instances of the relation
  - (or some patterns that work pretty well)
  - and lots & lots of unannotated text (e.g., the web)
- … can you use those seeds to do something useful?
- Bootstrapping can be considered *semi-supervised*

Bootstrapping example

- Target relation: *burial place*
- Seed tuple: [*Mark Twain*, *Elmira*]
- Grep/Google for "Mark Twain" and "Elmira"

  "Mark Twain is buried in Elmira, NY."

  → X is buried in Y

  "The grave of Mark Twain is in Elmira"

  → The grave of X is in Y

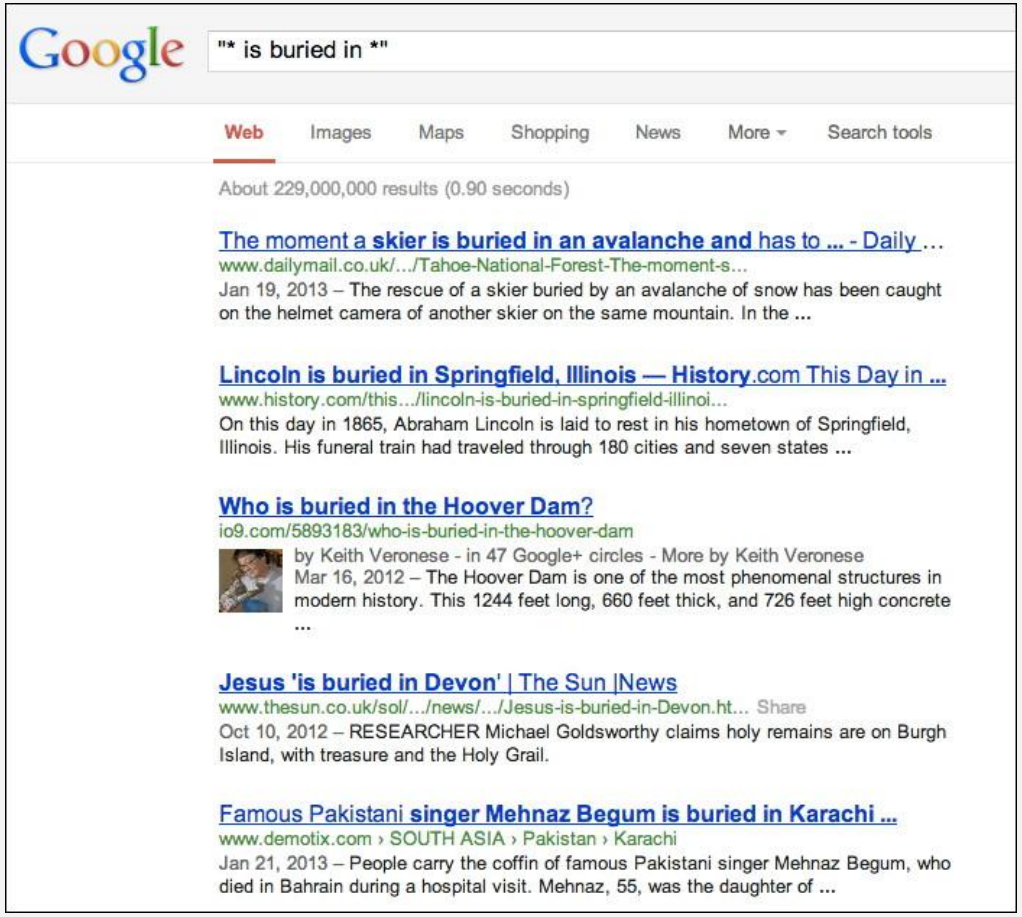
  "Elmira is Mark Twain's final resting place"

  → Y is X's final resting place

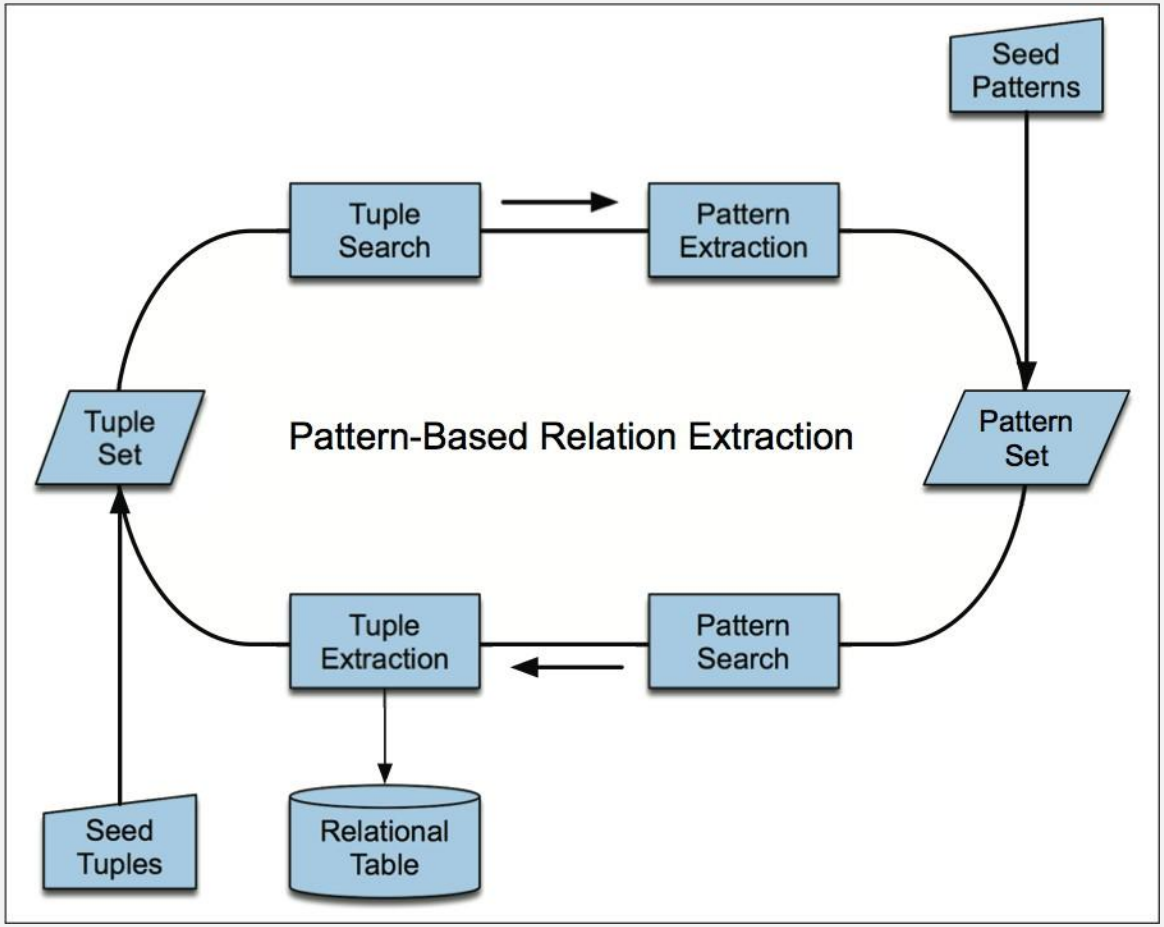
- Use those patterns to search for new tuples

slide adapted from Jim Martin

Bootstrapping example



The bootstrapping loop



slide adapted from Jim Martin

DIPRE (Brin 1998)

Extract (author, book) pairs

Start with these 5 seeds:

Author	Book
Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

Learn these patterns:

URL Prefix	Text Pattern
`www.sff.net/locus/c.*`	<LI><B>*title*</B> by *author* (
`dns.city-net.com/~lmann/awards/hugos/1984.html`	*title* by *author* (
`dolphin.upenn.edu/~dcummins/texts/sf-award.htm`	*author* || *title* || (

Iterate: use these patterns to get more instances & patterns…

Bootstrapping problems

- Requires that we have seeds for each relation
  - Sensitive to original set of seeds
- Big problem of semantic drift at each iteration
- Precision tends to be not that high
- Generally have lots of parameters to be tuned
- No probabilistic interpretation
  - Hard to know how confident to be in each result

Approaches to relation extraction

1. Hand-built patterns
2. Bootstrapping methods
3. Supervised methods
4. Distant supervision
5. Other related work

Supervised relation extraction

For each pair of entities in a sentence, predict the *relation type* (if any) that holds between them.

The supervised approach requires:

- Defining an inventory of relation types
- Collecting labeled training data (the hard part!)
- Designing a feature representation
- Choosing a classifier: Naïve Bayes, MaxEnt, SVM, ...
- Evaluating the results

An inventory of relation types

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (General affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	*None*
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-to-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

Relation types used in the ACE 2008 evaluation

Labeled training data

Source	Training epoch	Approximate size
English Resources		
Broadcast News	3/03 – 6/03	55,000 words
Broadcast Conversations	3/03 – 6/03	40,000 words
Newswire	3/03 – 6/03	50,000 words
Weblog	11/04 – 2/05	40,000 words
Usenet	11/04 – 2/05	40,000 words
Conversational Telephone Speech	11/04-12/04 (differentiated by topic vs. eval)	40,000 words
Arabic Resources		
Broadcast News	10/00 – 12/00	30,000+ words
Newswire	10/00 – 12/00	55,000+ words
Weblog	11/04 – 2/05	20,000+ words

Datasets used in the ACE 2008 evaluation

Feature representations

- Lightweight features — require little pre-processing
  - Bags of words & bigrams between, before, and after the entities
  - Stemmed versions of the same
  - The types of the entities
  - The distance (number of words) between the entities
- Medium-weight features — require base phrase chunking
  - Base-phrase chunk paths
  - Bags of chunk heads
- Heavyweight features — require full syntactic parsing
  - Dependency-tree paths between the entities
  - Constituent-tree paths between the entities
  - Tree distance between the entities
  - Presence of particular constructions in a constituent structure

Classifiers

Now use any (multiclass) classifier you like:

- multiclass SVM
- MaxEnt (aka multiclass logistic regression)
- Naïve Bayes
- etc.

Zhou et al. 2005 results

Type	Subtype	#Testing Instances	#Correct	#Error	P	R	F
AT		**392**	**224**	**105**	**68.1**	**57.1**	**62.1**
	Based-In	85	39	10	79.6	45.9	58.2
	Located	241	132	120	52.4	54.8	53.5
	Residence	66	19	9	67.9	28.8	40.4
NEAR		**35**	**8**	**1**	**88.9**	**22.9**	**36.4**
	Relative-Location	35	8	1	88.9	22.9	36.4
PART		**164**	**106**	**39**	**73.1**	**64.6**	**68.6**
	Part-Of	136	76	32	70.4	55.9	62.3
	Subsidiary	27	14	23	37.8	51.9	43.8
ROLE		**699**	**443**	**82**	**84.4**	**63.4**	**72.4**
	Citizen-Of	36	25	8	75.8	69.4	72.6
	General-Staff	201	108	46	71.1	53.7	62.3
	Management	165	106	72	59.6	64.2	61.8
	Member	224	104	36	74.3	46.4	57.1
SOCIAL		**95**	**60**	**21**	**74.1**	**63.2**	**68.5**
	Other-Professional	29	16	32	33.3	55.2	41.6
	Parent	25	17	0	100	68.0	81.0

Table 4: Performance of different relation types and major subtypes in the test data

Supervised RE: summary

- Supervised approach can achieve high accuracy
  - At least, for *some* relations
  - If we have lots of hand-labeled training data
- But has significant limitations!
  - Labeling 5,000 relations (+ named entities) is expensive
  - Doesn't generalize to different relations, languages
- Next: beyond supervised relation extraction
  - Distantly supervised relation extraction
  - Unsupervised relation extraction

Approaches to relation extraction

1. Hand-built patterns
2. Bootstrapping methods
3. Supervised methods
4. Distant supervision
5. Other related work

Distant supervision paradigm

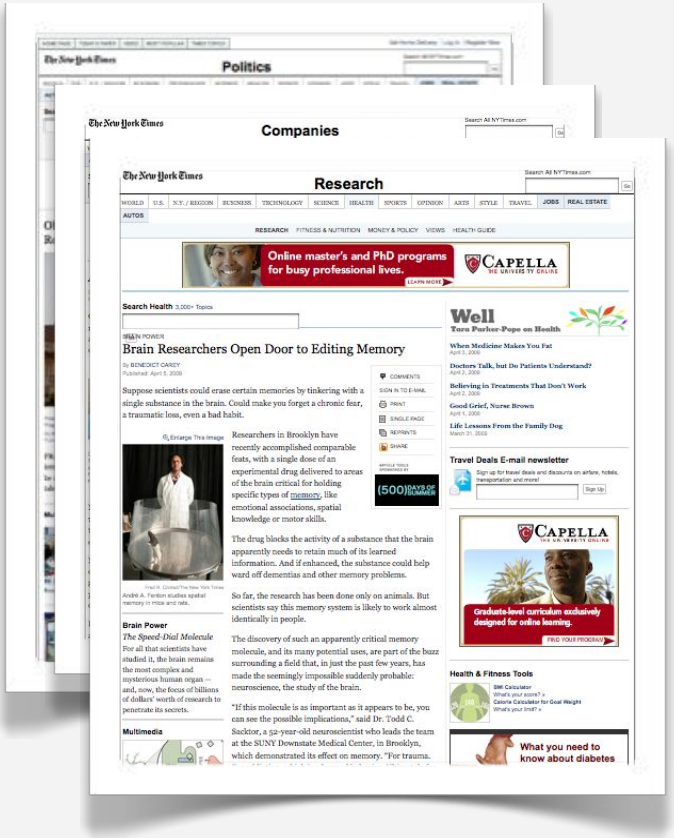
Snow, Jurafsky, Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. NIPS 17

Mintz, Bills, Snow, Jurafsky. 2009. Distant supervision for relation extraction without labeled data. ACL-2009.

- Hypothesis: If two entities belong to a certain relation, any sentence containing those two entities is likely to express that relation
- Key idea: use a *database* of relations to get lots of training examples
  - instead of hand-creating a few seed tuples (bootstrapping)
  - instead of using hand-labeled corpus (supervised)

Hypernyms via distant supervision

We construct a noisy training set consisting of occurrences from our corpus that contain a hyponym-hypernym pair from WordNet.



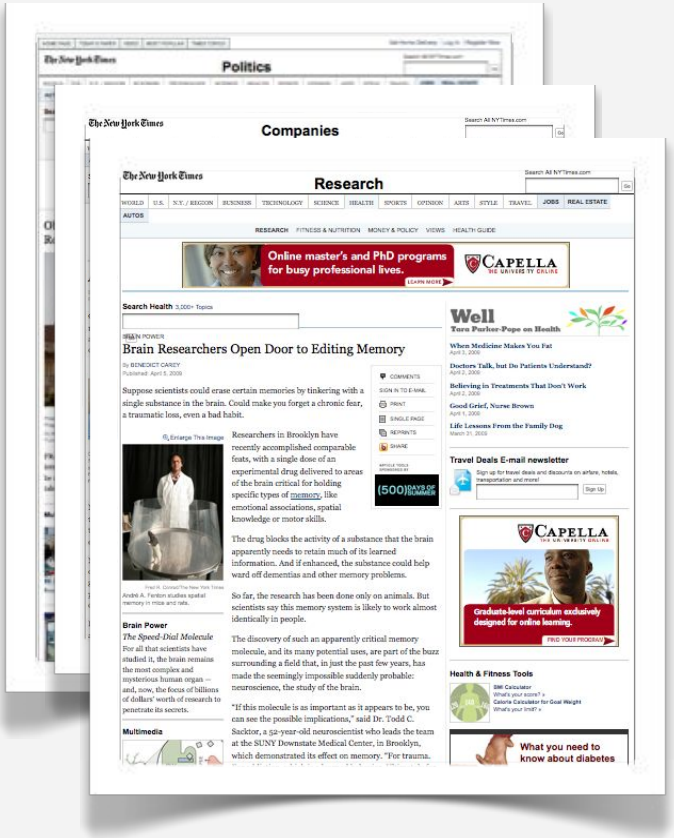
This yields high-signal examples like:

"...consider authors like Shakespeare..."
"Some authors (including Shakespeare)..."
"Shakespeare was the author of several..."
"Shakespeare, author of *The Tempest*..."

slide adapted from Rion Snow

Hypernyms via distant supervision

We construct a noisy training set consisting of occurrences from our corpus that contain a hyponym-hypernym pair from WordNet.



This yields high-signal examples like:

"...consider authors like Shakespeare..."
"Some authors (including Shakespeare)..."
"Shakespeare was the author of several..."
"Shakespeare, author of *The Tempest*..."

But also noisy examples like:

"The author of *Shakespeare in Love*..."
"...authors at the Shakespeare Festival..."

slide adapted from Rion Snow

Learning hypernym patterns

1. Take 6M newswire sentences

   *... doubly heavy hydrogen **atom called deuterium** ...*

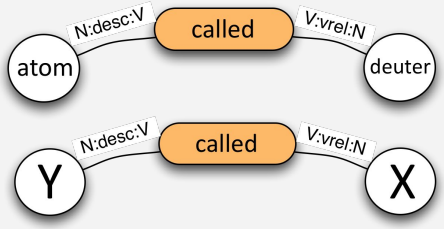
2. Collect noun pairs

   e.g. (atom, deuterium)
   752,311 pairs from 6M sentences of newswire

3. Is pair a hypernym in WordNet?

   14,387 yes; 737,924 no

4. Parse the sentences



5. Extract patterns

   69,592 dependency paths with >5 pairs

6. Train classifier on patterns

   logistic regression with 70K features
   (converted to 974,288 bucketed binary features)

slide adapted from Rion Snow

One of 70,000 patterns

Pattern: <superordinate> called <subordinate>
or: <Y> called <X>

Learned from cases such as:

(sarcoma, cancer)	…an uncommon bone cancer called osteogenic sarcoma and to…
(deuterium, atom)	…heavy water rich in the doubly heavy hydrogen atom called deuterium.

New pairs discovered:

(efflorescence, condition)	…and a condition called efflorescence are other reasons for…
(O'neal_inc, company)	…The company, now called O'Neal Inc., was sole distributor of…
(hat_creek_outfit, ranch)	…run a small ranch called the Hat Creek Outfit.
(hiv-1, aids_virus)	…infected by the AIDS virus, called HIV-1.
(bateau_mouche, attraction)	…local sightseeing attraction called the Bateau Mouche…

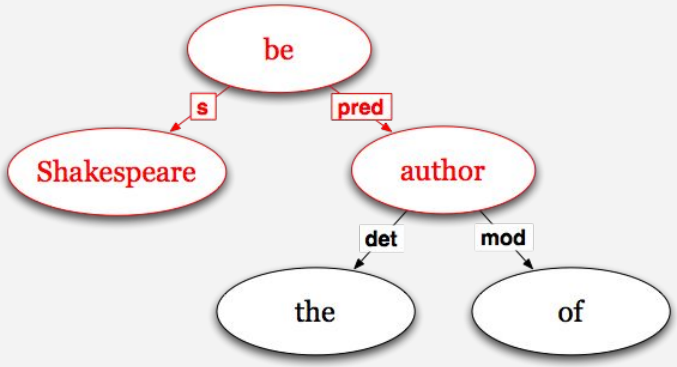
slide adapted from Rion Snow

Syntactic dependency paths

Patterns are based on paths through dependency parses generated by MINIPAR (Lin, 1998)

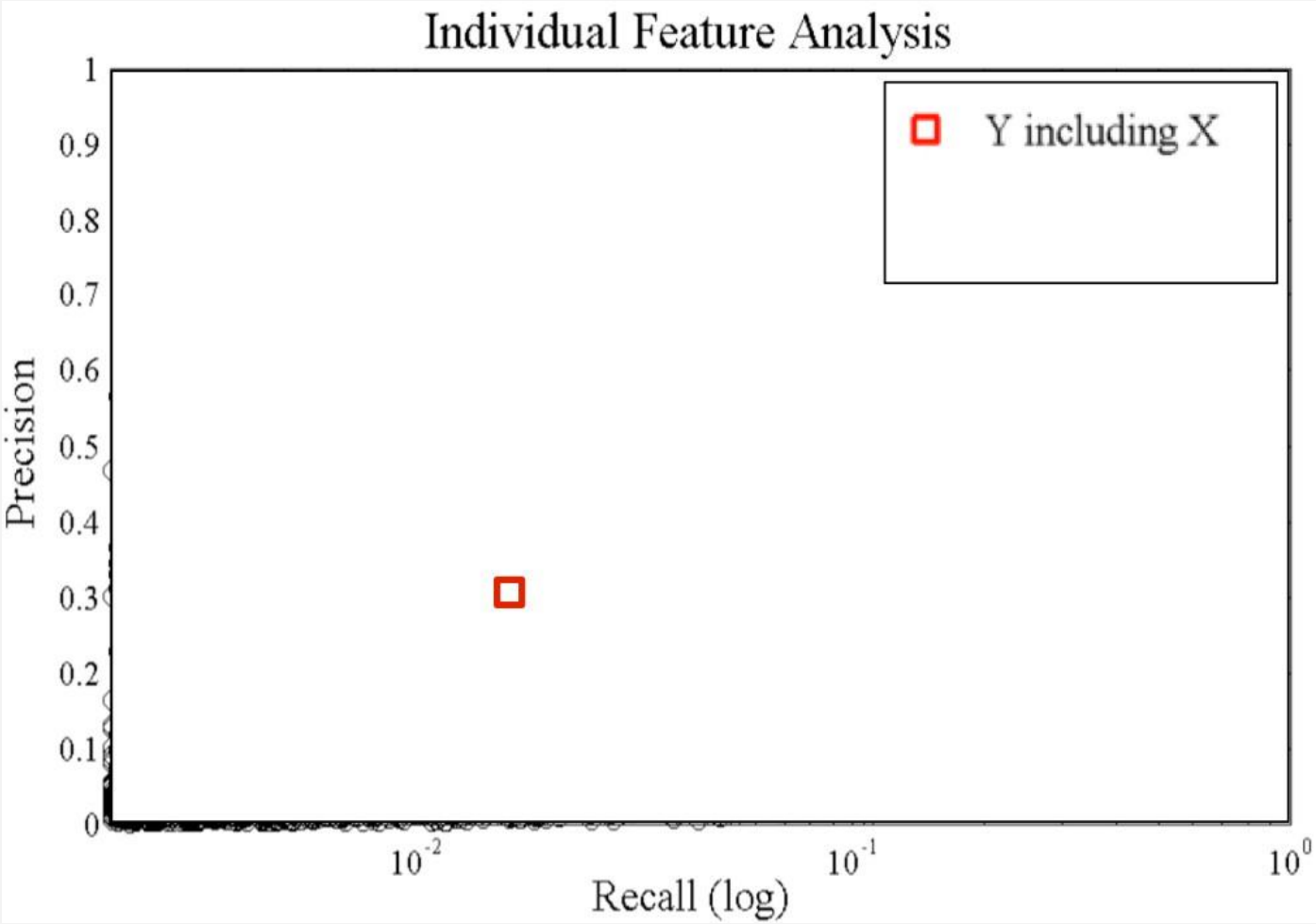
Example word pair: (Shakespeare, author)
Example sentence: "Shakespeare was the author of several plays…"

Minipar parse:



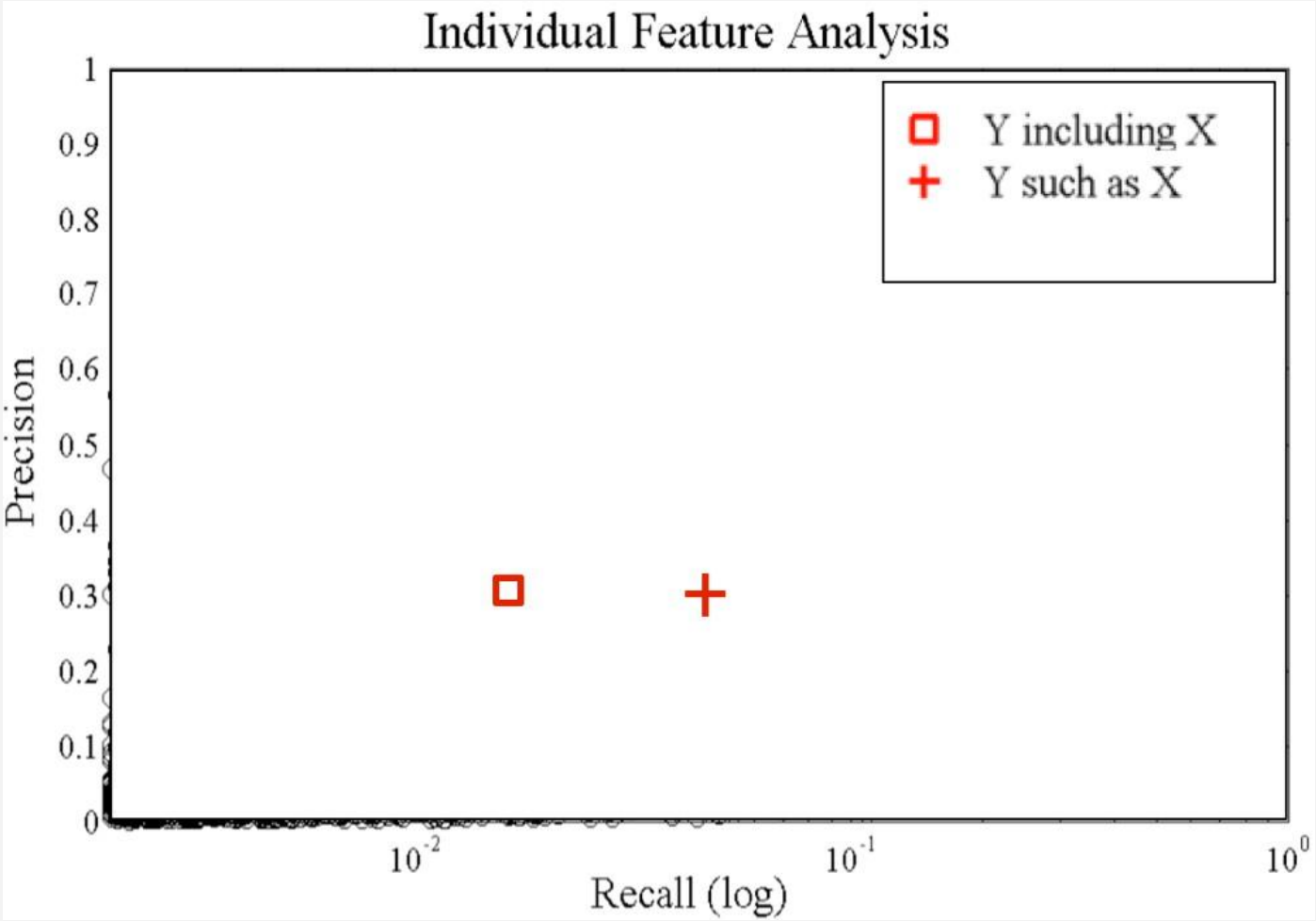
Extract shortest path:
-N:s:VBE, be, VBE:pred:N

P/R of hypernym extraction patterns



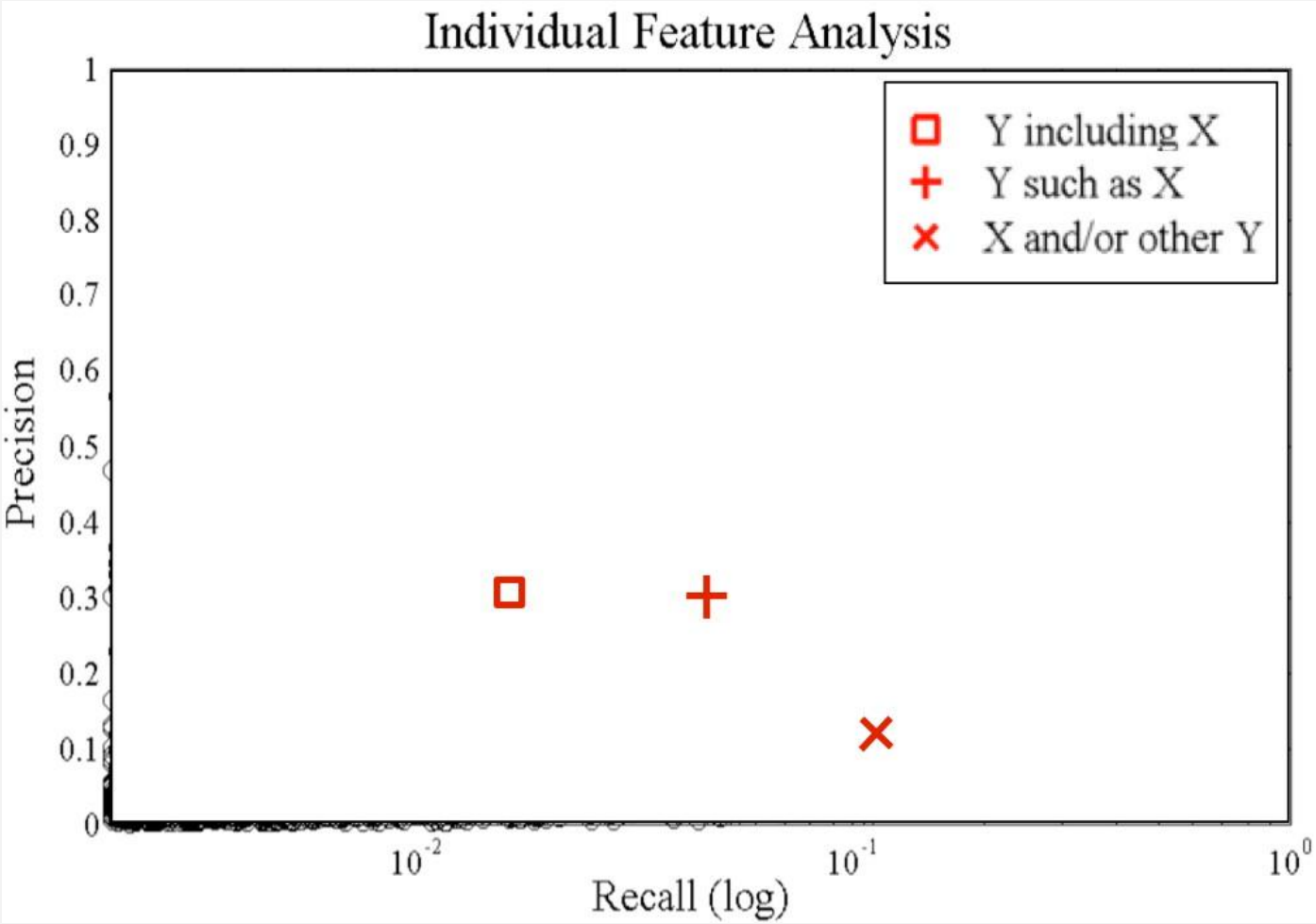
slide adapted from Rion Snow

P/R of hypernym extraction patterns



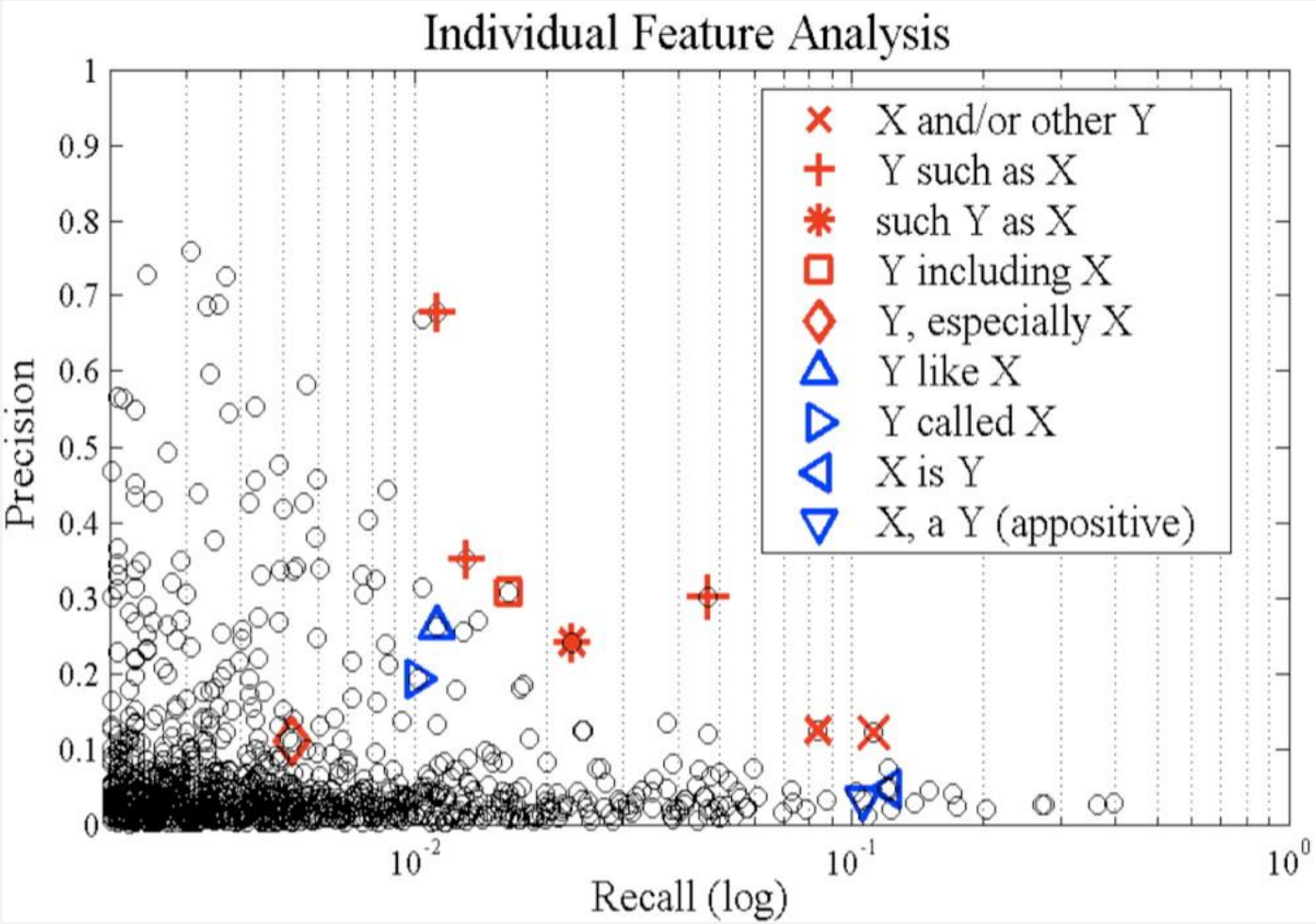
slide adapted from Rion Snow

P/R of hypernym extraction patterns



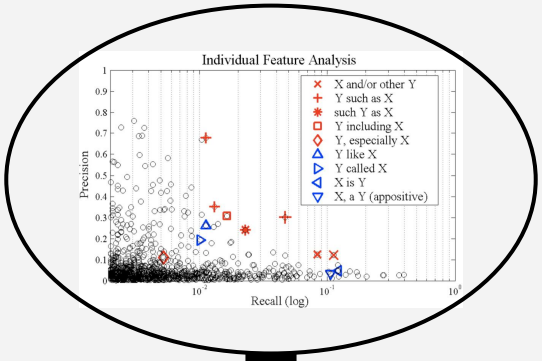
slide adapted from Rion Snow

P/R of hypernym extraction patterns



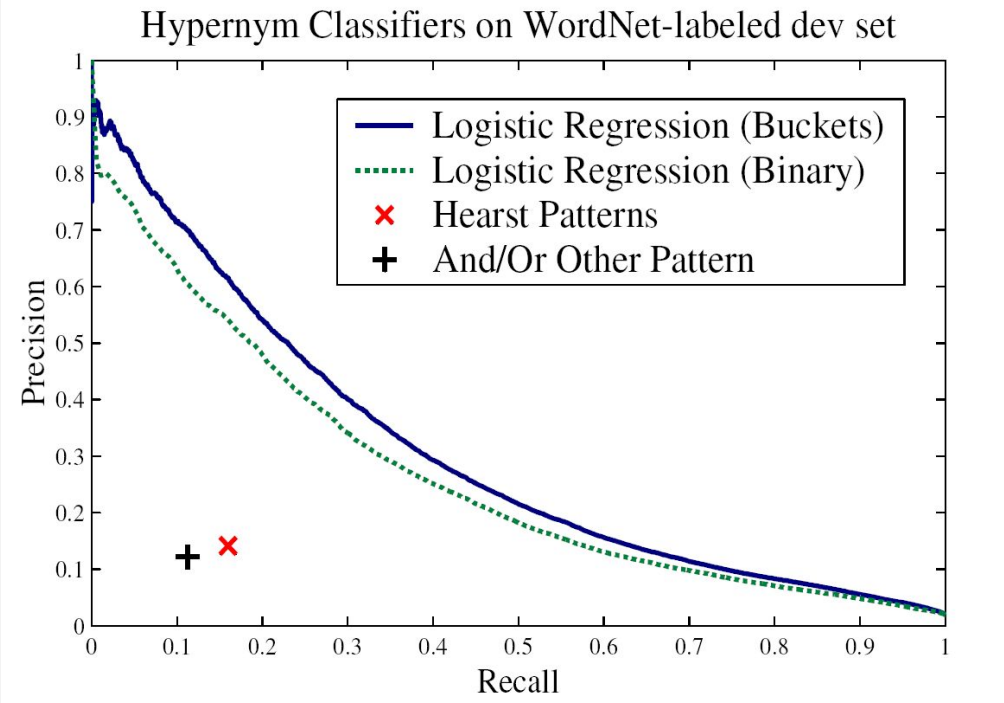
slide adapted from Rion Snow

P/R of hypernym classifier



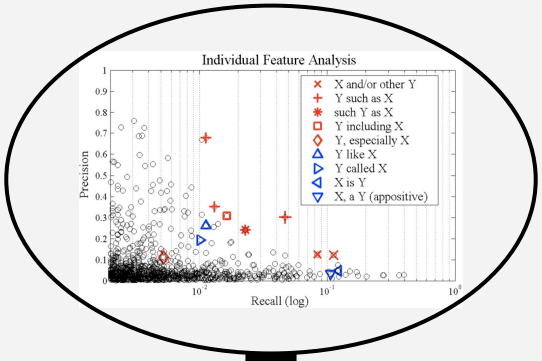
logistic regression

$$P(R|E) = \frac{1}{1 + e^{-\sum w_i x_i}}$$



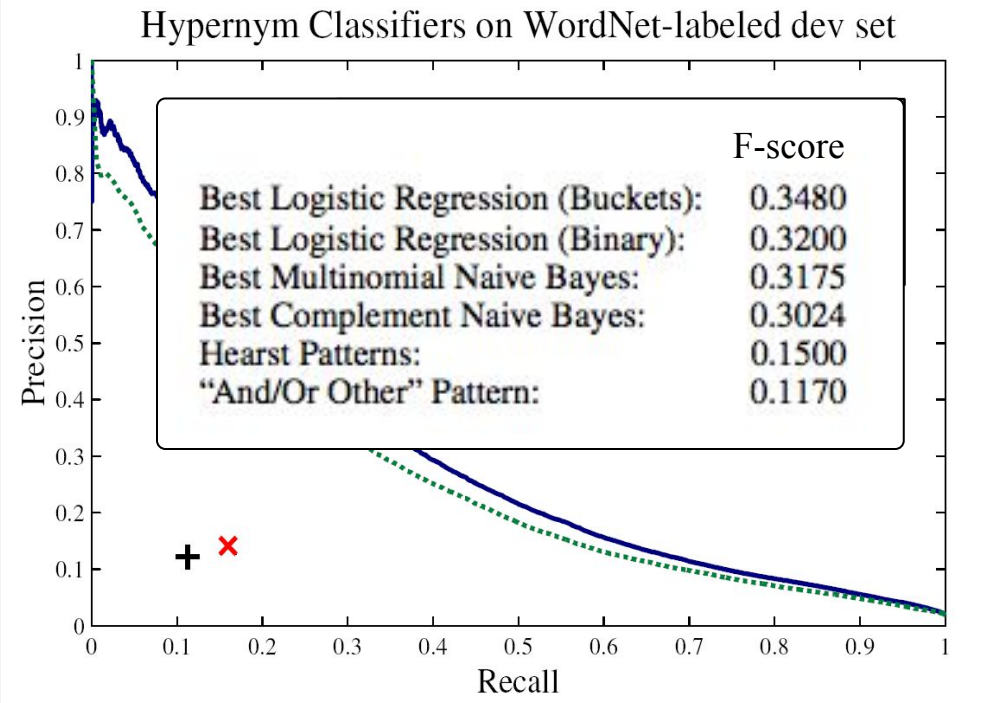
10-fold Cross Validation on
14,000 WordNet-Labeled Pairs

P/R of hypernym classifier



logistic regression

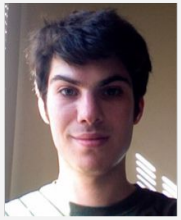
$$P(R|E) = \frac{1}{1 + e^{-\sum w_i x_i}}$$

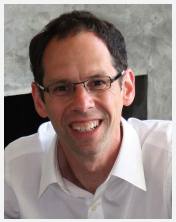


10-fold Cross Validation on
14,000 WordNet-Labeled Pairs

What about other relations?

Mintz, Bills, Snow, Jurafsky (2009).
Distant supervision for relation extraction without labeled data.



Training set

Freebase

102 relations
940,000 entities
1.8 million instances

Corpus

WIKIPEDIA

1.8 million articles
25.7 million sentences

Frequent Freebase relations

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Collecting training data

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Training data

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Collecting training data

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Collecting training data

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Collecting training data

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

Collecting training data

Corpus text

Bill Gates founded Microsoft in 1975.
Bill Gates, founder of Microsoft, …
Bill Gates attended Harvard from…
Google was founded by Larry Page …

Freebase

Founder: (Bill Gates, Microsoft)
Founder: (Larry Page, Google)
CollegeAttended: (Bill Gates, Harvard)

Training data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

(Larry Page, Google)
Label: Founder
Feature: Y was founded by X

Negative training data

Can't train a classifier with only positive data! Need negative training data too!

Solution?
Sample 1% of unrelated pairs of entities.

Result: roughly balanced data.

Corpus text

Larry Page took a swipe at Microsoft...
...after Harvard invited Larry Page to...
Google is Bill Gates' worst fear ...

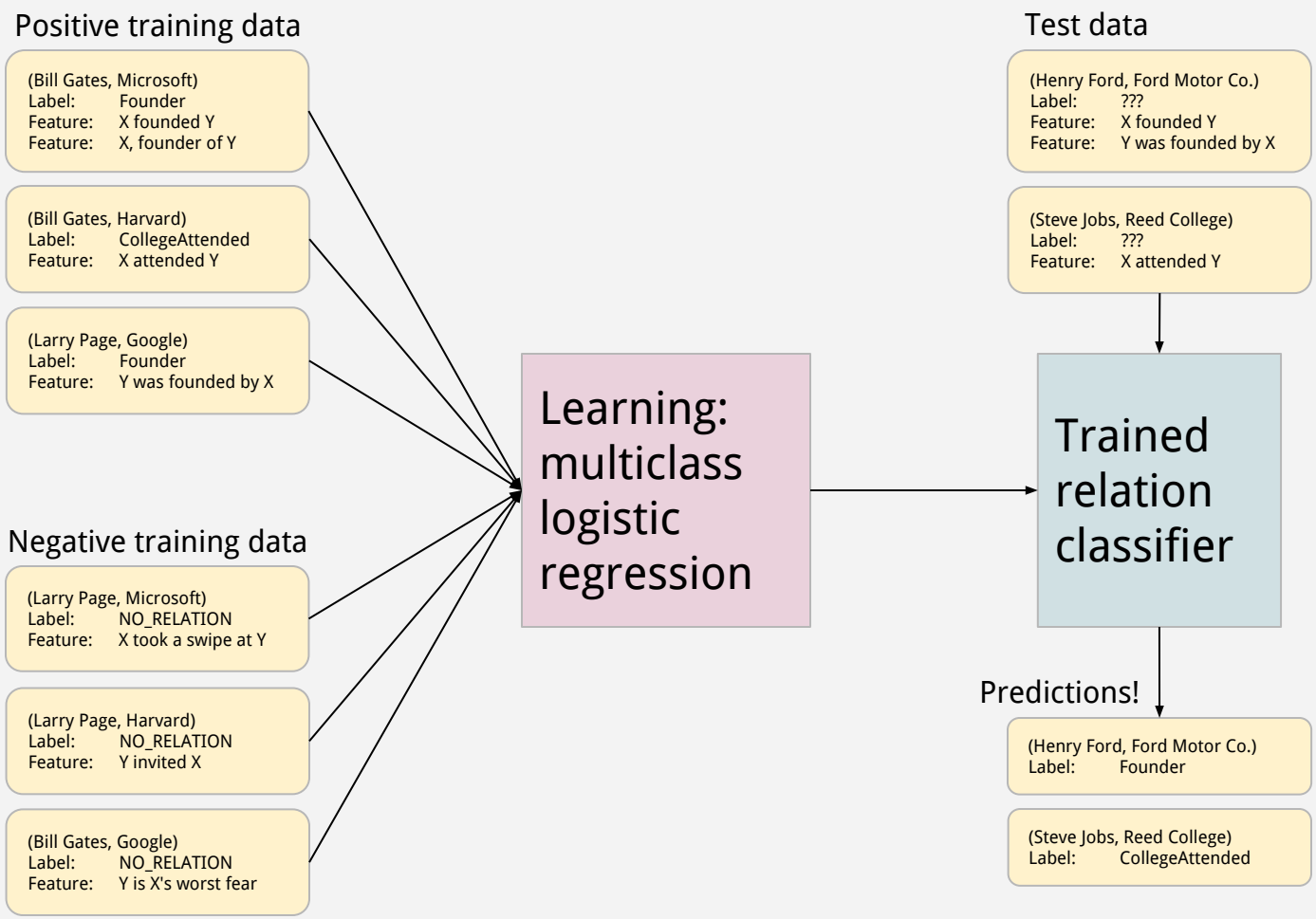
Training data

(Larry Page, Microsoft)
Label: NO_RELATION
Feature: X took a swipe at Y

(Larry Page, Harvard)
Label: NO_RELATION
Feature: Y invited X

(Bill Gates, Google)
Label: NO_RELATION
Feature: Y is X's worst fear

The experiment



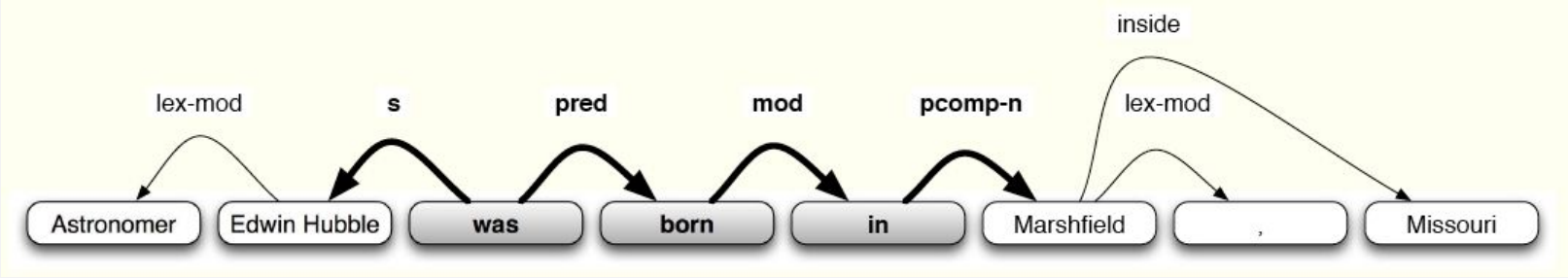
Benefits of distant supervision

- Has advantages of supervised approach
  - leverage rich, reliable hand-created knowledge
  - relations have canonical names
  - can use rich features (e.g. syntactic features)

- Has advantages of unsupervised approach
  - leverage unlimited amounts of text data
  - allows for very large number of weak features
  - not sensitive to training corpus: genre-independent

Lexical and syntactic features

Astronomer Edwin Hubble was born in Marshfield, Missouri.



Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]

High-weight features

Relation	Feature type	Left window	NE1	Middle	NE2	Right window
/architecture/structure/architect	LEX$\curvearrowleft$		ORG	, the designer of the	PER	
	SYN	designed $\Uparrow_s$	ORG	$\Uparrow_s$ designed $\Downarrow_{by-subj}$ by $\Downarrow_{pcn}$	PER	$\Uparrow_s$ designed
/book/author/works_written	LEX		PER	s novel	ORG	
	SYN		PER	$\Uparrow_{pcn}$ by $\Uparrow_{mod}$ story $\Uparrow_{pred}$ is $\Downarrow_s$	ORG	
/book/book_edition/author_editor	LEX$\curvearrowleft$		ORG	s novel	PER	
	SYN		PER	$\Uparrow_{nn}$ series $\Downarrow_{gen}$	PER	
/business/company/founders	LEX		ORG	co - founder	PER	
	SYN		ORG	$\Uparrow_{nn}$ owner $\Downarrow_{person}$	PER	
/business/company/place_founded	LEX$\curvearrowleft$		ORG	- based	LOC	
	SYN		ORG	$\Uparrow_s$ founded $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	
/film/film/country	LEX		PER	, released in	LOC	
	SYN	opened $\Uparrow_s$	ORG	$\Uparrow_s$ opened $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	$\Uparrow_s$ opened
/geography/river/mouth	LEX		LOC	, which flows into the	LOC	
	SYN	the $\Downarrow_{det}$	LOC	$\Uparrow_s$ is $\Downarrow_{pred}$ tributary $\Downarrow_{mod}$ of $\Downarrow_{pcn}$	LOC	$\Downarrow_{det}$ the
/government/political_party/country	LEX$\curvearrowleft$		ORG	politician of the	LOC	
	SYN	candidate $\Uparrow_{nn}$	ORG	$\Uparrow_{nn}$ candidate $\Downarrow_{mod}$ for $\Downarrow_{pcn}$	LOC	$\Uparrow_{nn}$ candidate
/influence/influence_node/influenced	LEX$\curvearrowleft$		PER	, a student of	PER	
	SYN	of $\Uparrow_{pcn}$	PER	$\Uparrow_{pcn}$ of $\Uparrow_{mod}$ student $\Uparrow_{appo}$	PER	$\Uparrow_{pcn}$ of
/language/human_language/region	LEX		LOC	- speaking areas of	LOC	
	SYN		LOC	$\Uparrow_{lex-mod}$ speaking areas $\Downarrow_{mod}$ of $\Downarrow_{pcn}$	LOC	
/music/artist/origin	LEX$\curvearrowleft$		ORG	based band	LOC	
	SYN	is $\Uparrow_s$	ORG	$\Uparrow_s$ is $\Downarrow_{pred}$ band $\Downarrow_{mod}$ from $\Downarrow_{pcn}$	LOC	$\Uparrow_s$ is
/people/deceased_person/place_of_death	LEX		PER	died in	LOC	
	SYN	hanged $\Uparrow_s$	PER	$\Uparrow_s$ hanged $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	$\Uparrow_s$ hanged
/people/person/nationality	LEX		PER	is a citizen of	LOC	
	SYN		PER	$\Downarrow_{mod}$ from $\Downarrow_{pcn}$	LOC	
/people/person/parents	LEX		PER	, son of	PER	
	SYN	father $\Uparrow_{gen}$	PER	$\Uparrow_{gen}$ father $\Downarrow_{person}$	PER	$\Uparrow_{gen}$ father
/people/person/place_of_birth	LEX$\curvearrowleft$		PER	is the birthplace of	PER	
	SYN		PER	$\Uparrow_s$ born $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC	
/people/person/religion	LEX		PER	embraced	LOC	
	SYN	convert $\Downarrow_{appo}$	PER	$\Downarrow_{appo}$ convert $\Downarrow_{mod}$ to $\Downarrow_{pcn}$	LOC	$\Downarrow_{appo}$ convert

Experimental set-up

- 1.8 million relation instances used for training
  - Compared to 17,000 relation instances in ACE
- 800,000 Wikipedia articles used for training, 400,000 different articles used for testing
- Only extract relation instances not already in Freebase

It works!

Ten relation instances extracted by the system that weren't in Freebase

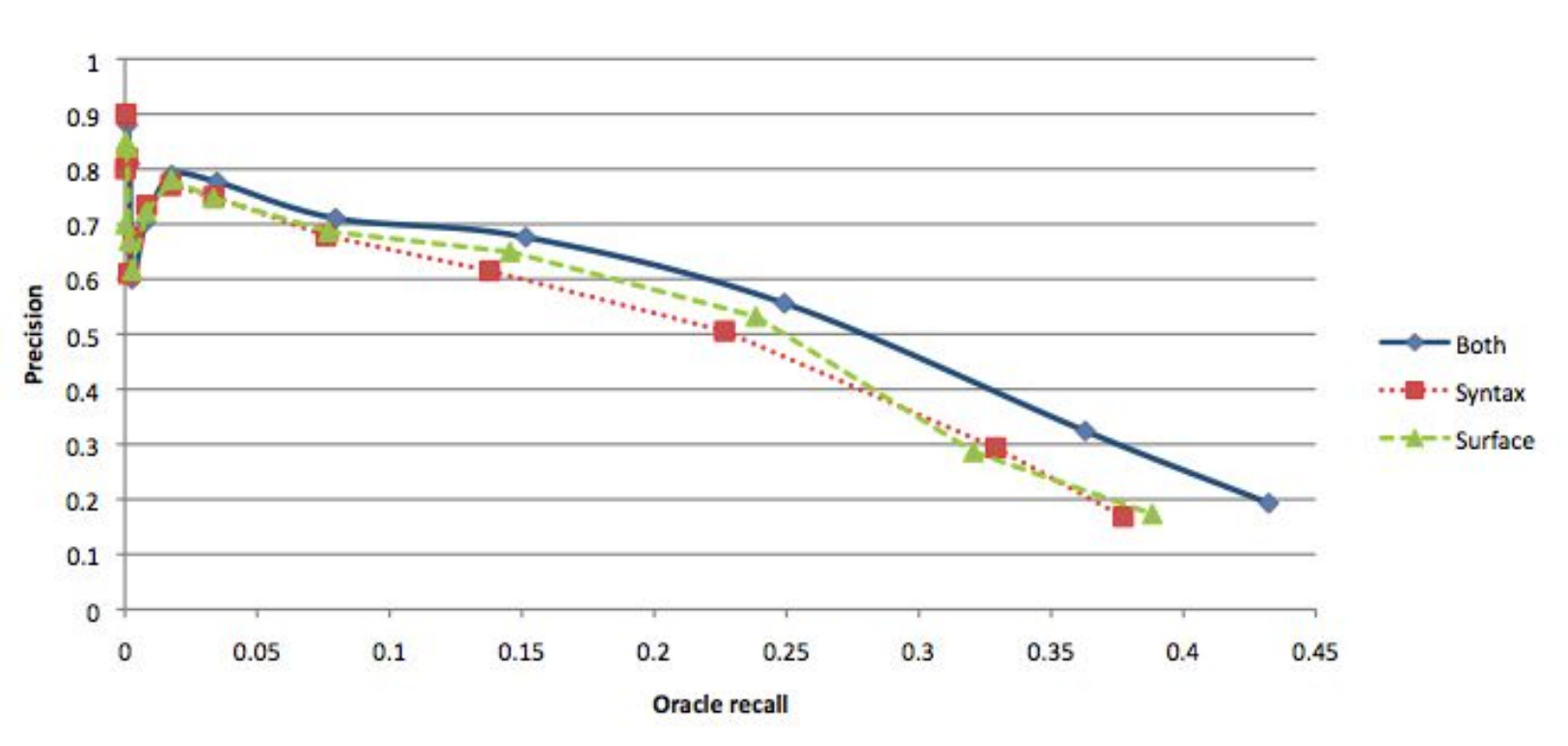
Relation name	New instance
/location/location/contains	Paris, Montmartre
/location/location/contains	Ontario, Fort Erie
/music/artist/origin	Mighty Wagon, Cincinnati
/people/deceased_person/place_of_death	Fyodor Kamensky, Clearwater
/people/person/nationality	Marianne Yvonne Heemskerk, Netherlands
/people/person/place_of_birth	Wavell Wayne Hinds, Kingston
/book/author/works_written	Upton Sinclair, Lanny Budd
/business/company/founders	WWE, Vince McMahon
/people/person/profession	Thomas Mellon, judge

Evaluation

- Held-out evaluation
  - Train on 50% of gold-standard Freebase relation instances, test on other 50%
  - Used to tune parameters quickly without having to wait for human evaluation

- Human evaluation
  - Performed by evaluators on Amazon Mechanical Turk
  - Calculated precision at 100 and 1000 recall levels for the ten most common relations

Held-out evaluation



Automatic evaluation on 900K instances of 102 Freebase relations.
Precision for three different feature sets is reported at various recall levels.

Distant supervision: takeaways

- The distant supervision approach uses a database of known relation instances as a source of supervision
- We're classifying pairs of entities, not pairs of entity mentions
- The features for a pair of entities describe the patterns in which the two entities have co-occurred across many sentences in a large corpus
- Can make use of 100x or even 1000x more data than in the supervised paradigm

Approaches to relation extraction

1. Hand-built patterns
2. Bootstrapping methods
3. Supervised methods
4. Distant supervision
5. Other related work

What else is out there?

- *Open information extraction* (OpenIE) aims to extract *all* relations from text, without supervision or any fixed set of relations.

```
(Google, is based in, Mountain View)
(Mountain View, is home to, Google)
```

- *Knowledge base completion* (KBC) aims to use information in a KB to fill in missing entries.

```
(AB, country_of_birth, Iceland)
=> (AB, speaks_language, Icelandic)
```

OpenIE demo

http://openie.allenai.org/

For next time

- Read Mintz et al. 2009
- Start working through the relation extraction codebook, rel_ext.ipynb