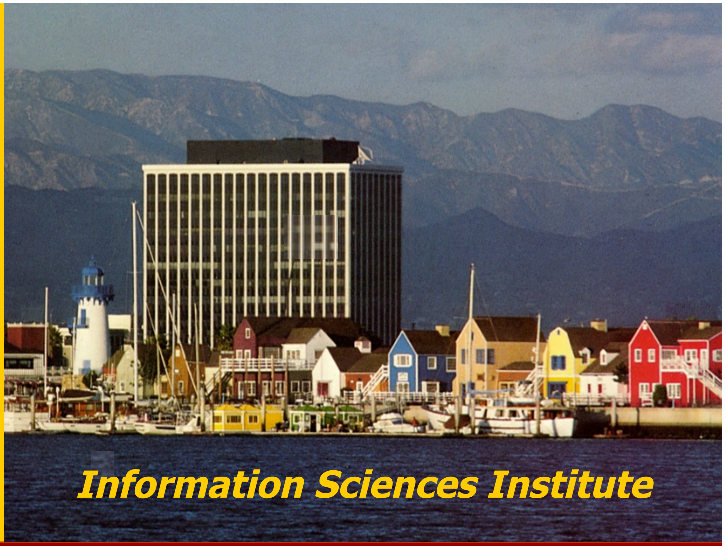
USC Viterbi
School of Engineering



Exploiting Structure within Data for Accurate Labeling using Conditional Random Fields

Aman Goel, Craig A. Knoblock, Kristina Lerman
Department of Computer Science & Information Sciences Institute
University of Southern California

USC

Outline

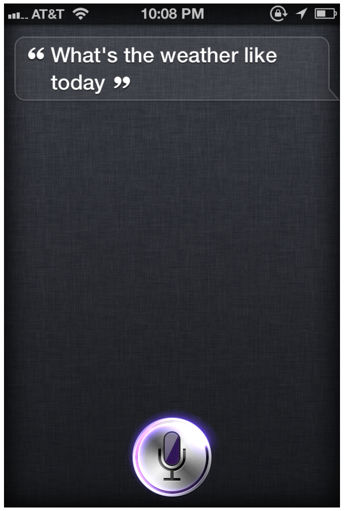
- Problem
- Existing approaches
- Our approach
- Experiments
- Real word application
- Conclusion
- Q&A (5 mins)

15 mins

Problem definition

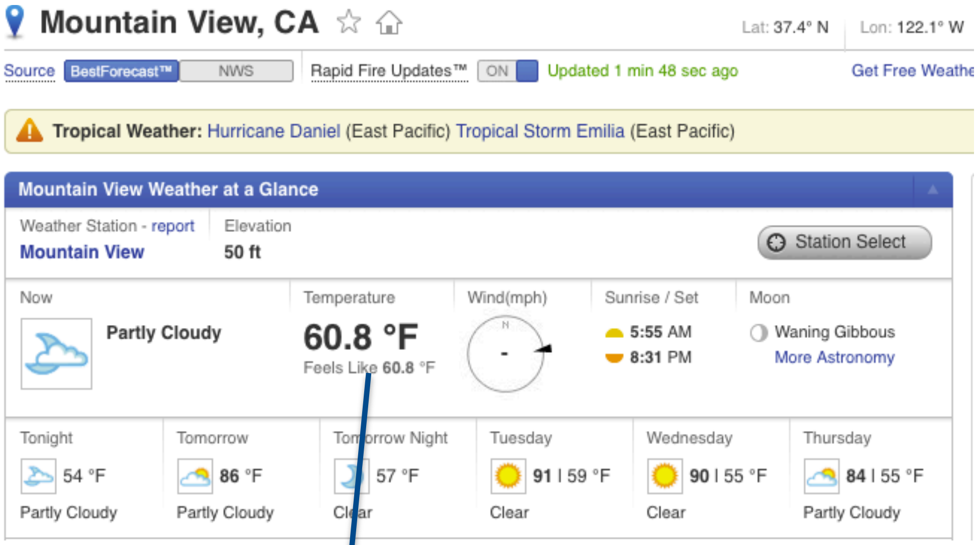
Aman Goel	=>	Name
2667 Ellendale Place, Los Angeles, CA 90007	=>	Address
(323) 246-7180	=>	Telephone Number

Motivation


AT&T 10:08 PM

“ What's the weather like today ”

Visit a website


Mountain View, CA ☆ ⌂ Lat: 37.4° N Lon: 122.1° W

Source BestForecast™ NWS Rapid Fire Updates™ ON Updated 1 min 48 sec ago Get Free Weathe

⚠ **Tropical Weather:** Hurricane Daniel (East Pacific) Tropical Storm Emilia (East Pacific)

Mountain View Weather at a Glance

Weather Station - report **Mountain View**

Elevation **50 ft**

Station Select

Now	Temperature	Wind(mph)	Sunrise / Set	Moon
Partly Cloudy	**60.8 °F** Feels Like **60.8** °F	N	**5:55** AM **8:31** PM	Waning Gibbous More Astronomy

Tonight	Tomorrow	Tomorrow Night	Tuesday	Wednesday	Thursday
54 °F	**86** °F	57 °F	**91** | 59 °F	**90** | 55 °F	**84** | 55 °F
Partly Cloudy	Partly Cloudy	Clear	Clear	Clear	Partly Cloudy

Can’t use !!!

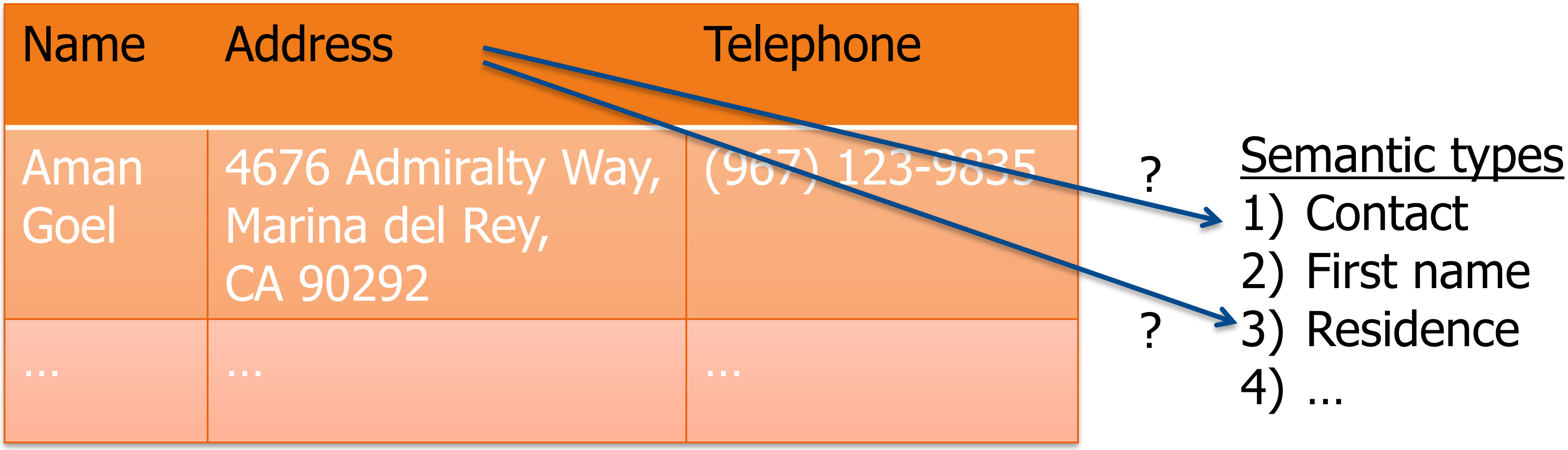
Extract strings

Mountain View, Partly Cloudy, 60.8 ºF, 5:56 AM

Existing approaches

- Vertical approach (schema matching)
- Horizontal approach (hidden markov model)

Vertical approach (schema matching)



Name	Address	Telephone
Aman Goel	4676 Admiralty Way, Marina del Rey, CA 90292	(967) 123-9835
…	…	…

Semantic types

1) Contact
2) First name
3) Residence
4) …

Approach

- Using dictionaries and string edit distances

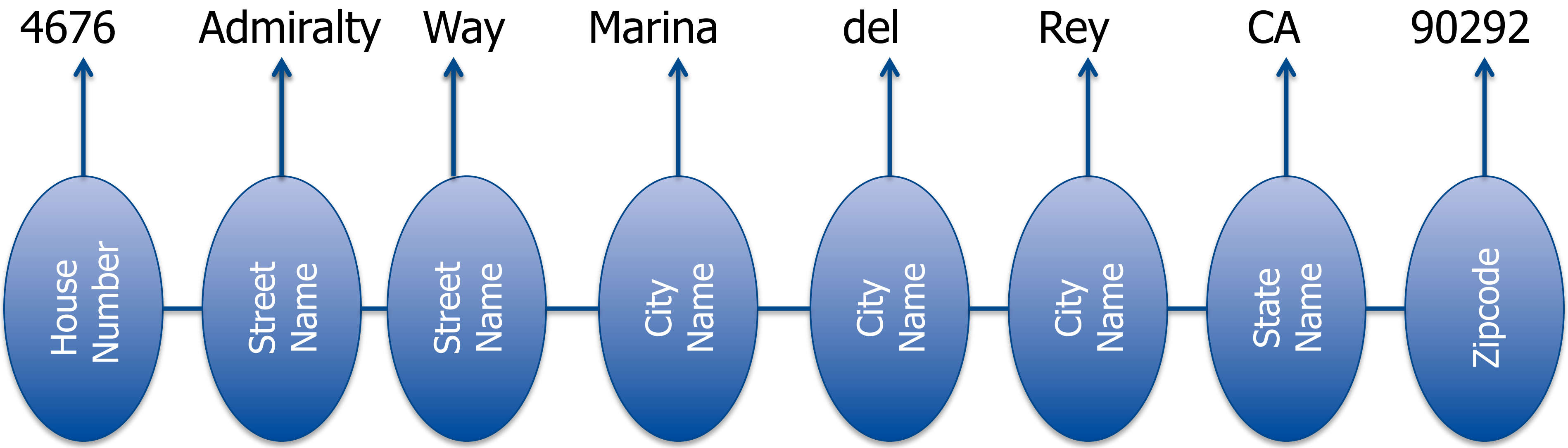
Problems

- Assume relational data
- Missing column names

Horizontal approach (HMMs)

Automatic segmentation of text into structured records:

V. Borkar, K. Deshmukh and S. Sarawagi, SIGMOD, 2001



4676 Admiralty Way Marina del Rey CA 90292

<u>Problems</u>

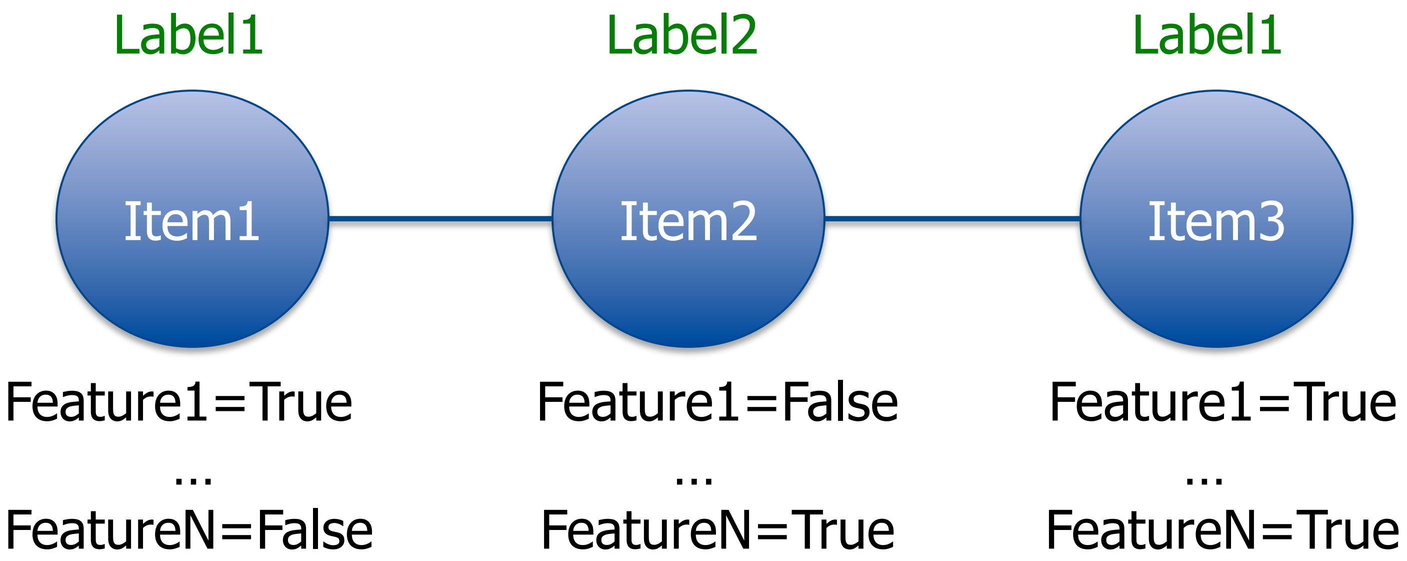
- Complex semantic type is already assumed

Our approach

Observations

- Ordering relationship between complex types
  - Name before Address
- Ordering relationship between the tokens
  - House number before Street name
- Relationship between tokens in different fields
  - (mph, inches) vs (kmph, mm)

CRF



Label1 Label2 Label1

Item1 Item2 Item3

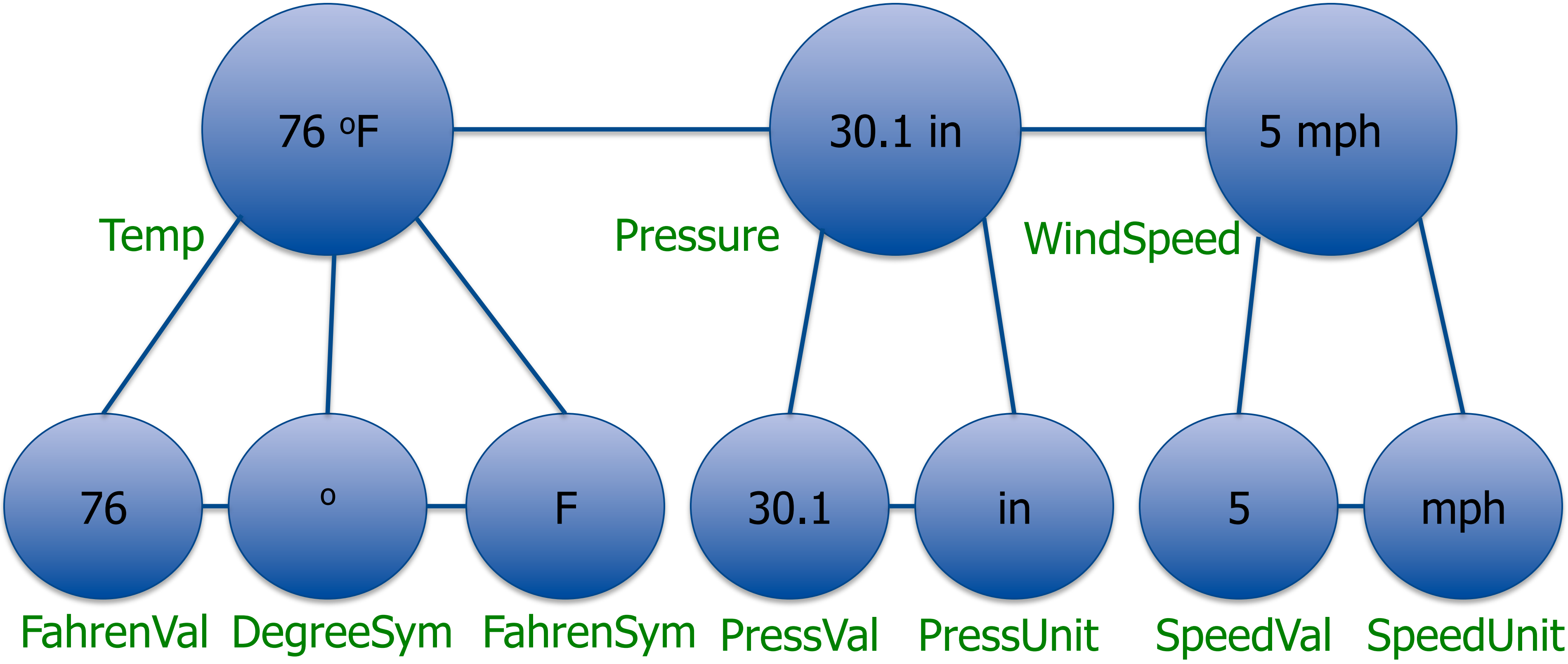
Feature1=True
…
FeatureN=False

Feature1=False
…
FeatureN=True

Feature1=True
…
FeatureN=True

Feature functions		Weights
• Label1 is followed by Label2	- - - - - - - - - - -	0.7
• Label1 has Feature1=True	- - - - - - - - - - -	1.5

CRF model from data



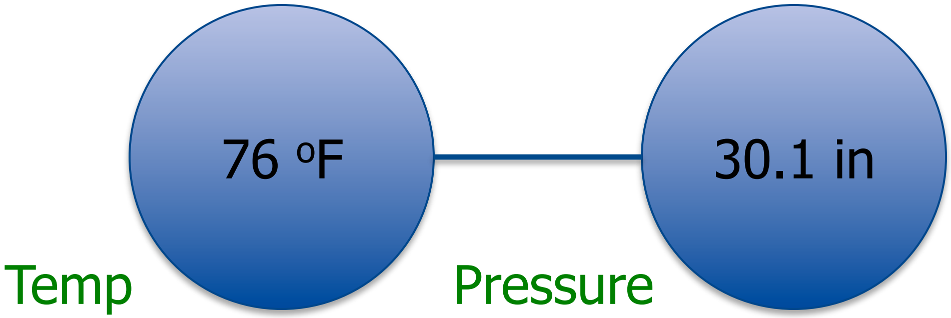
Features

Token features:

- Is_capitalized
- 7_characters_in_token
- Starts_with_B
- Number_is_in_100s
- Has_2_precision_digits
- Is_dollar_sign
- Is_negative
- Number_starts_with_9
- Is_all_caps
- Is_percent_sign
- …

Feature functions

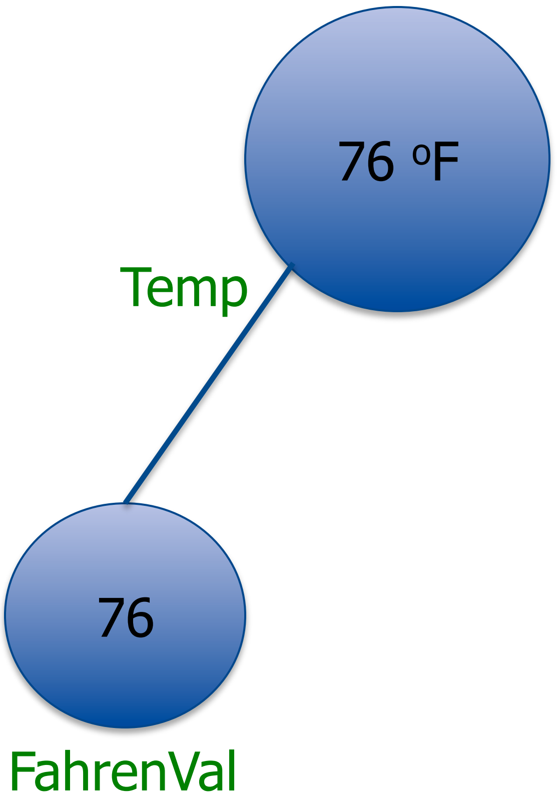
- Relationship between field nodes



Temp_field_is_followed_by_Pressure_field

Feature functions

- Relationship between field and token node



76 °F

Temp

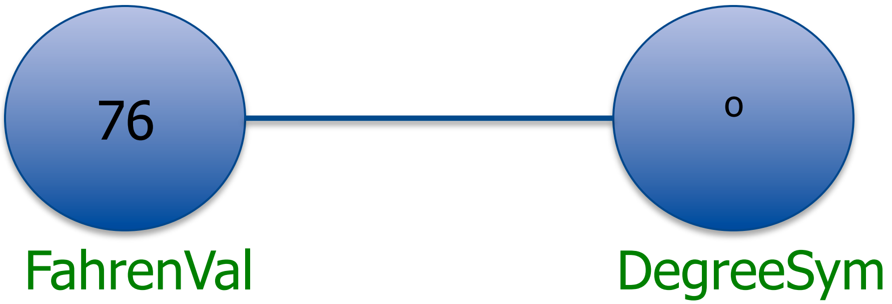
76

FahrenVal

Temp_field_has_token_FahrenVal

Feature functions

- Relationship between neighboring token nodes



76

o

FahrenVal

DegreeSym

FahrenVal_token_is_followed_by_DegreeSym_token

Feature functions

- Relationship between token label and its feature

FahrenVal

FahrenVal_token_is_in_10s

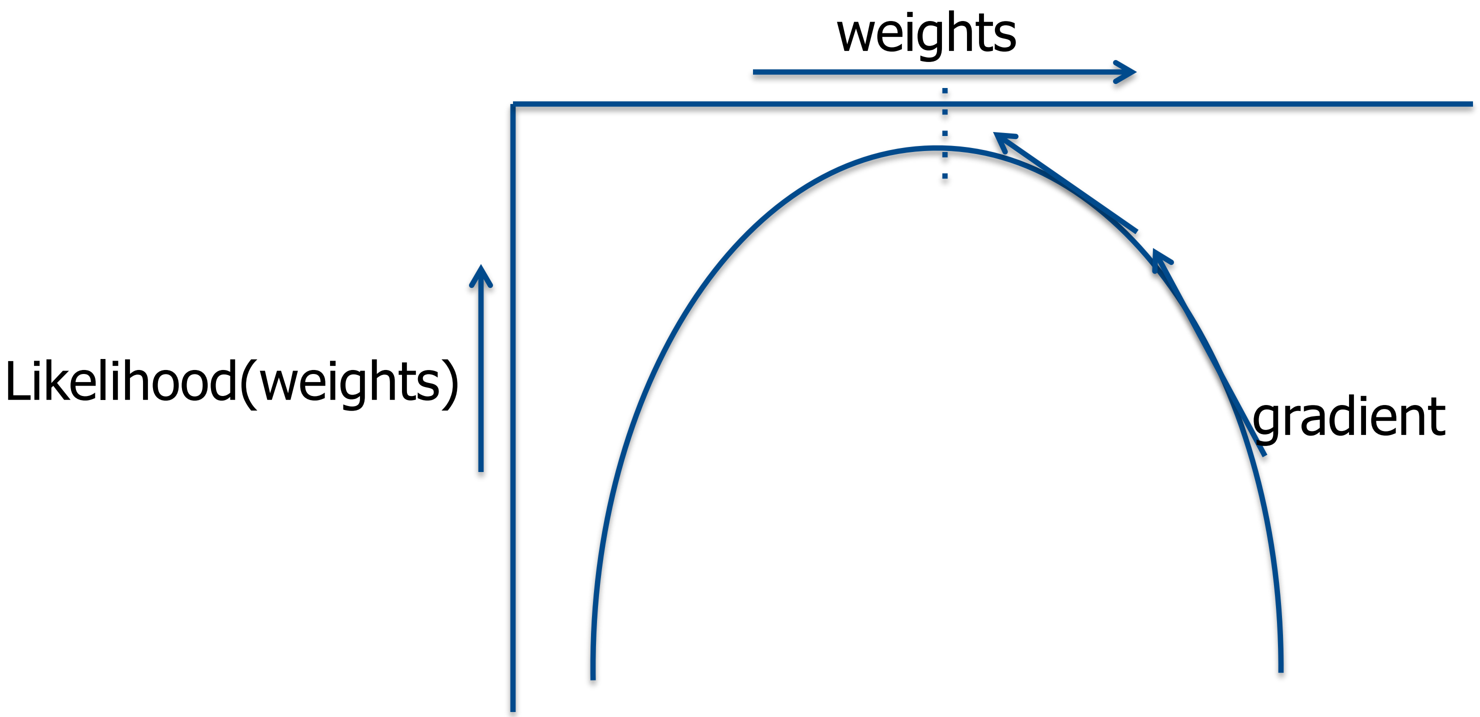
FahrenVal_token_has_0_precision_digits

Log-likelihood

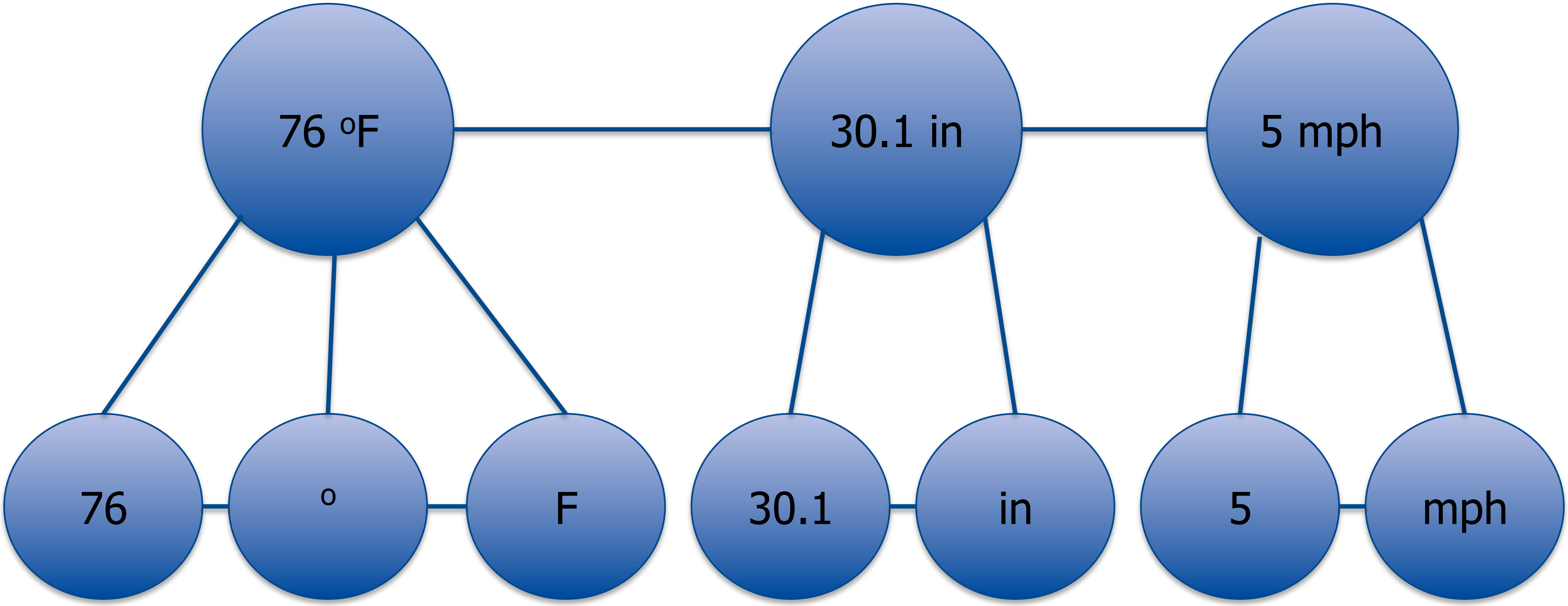
$$1/Z(x) * exp(\sum_{c} (\sum_{k} \underset{\text{weights}}{w_k} \underset{\text{feature functions}}{f_k}(\overset{\text{labels}}{y_c}, \overset{\text{features}}{x})))$$

Learning weights

- Convex optimization problem



Complexity (e^N)



Largest fully connected sub-graph size = 3

Experiment setup

- Collected 400 tuples from 4 websites
- Trained on 300 tuples from 3 sites
- Tested on 100 labeled examples from held-out site

	Weather	Flight	Geocoding
# field types	15	8	5
# token types	37	17	12

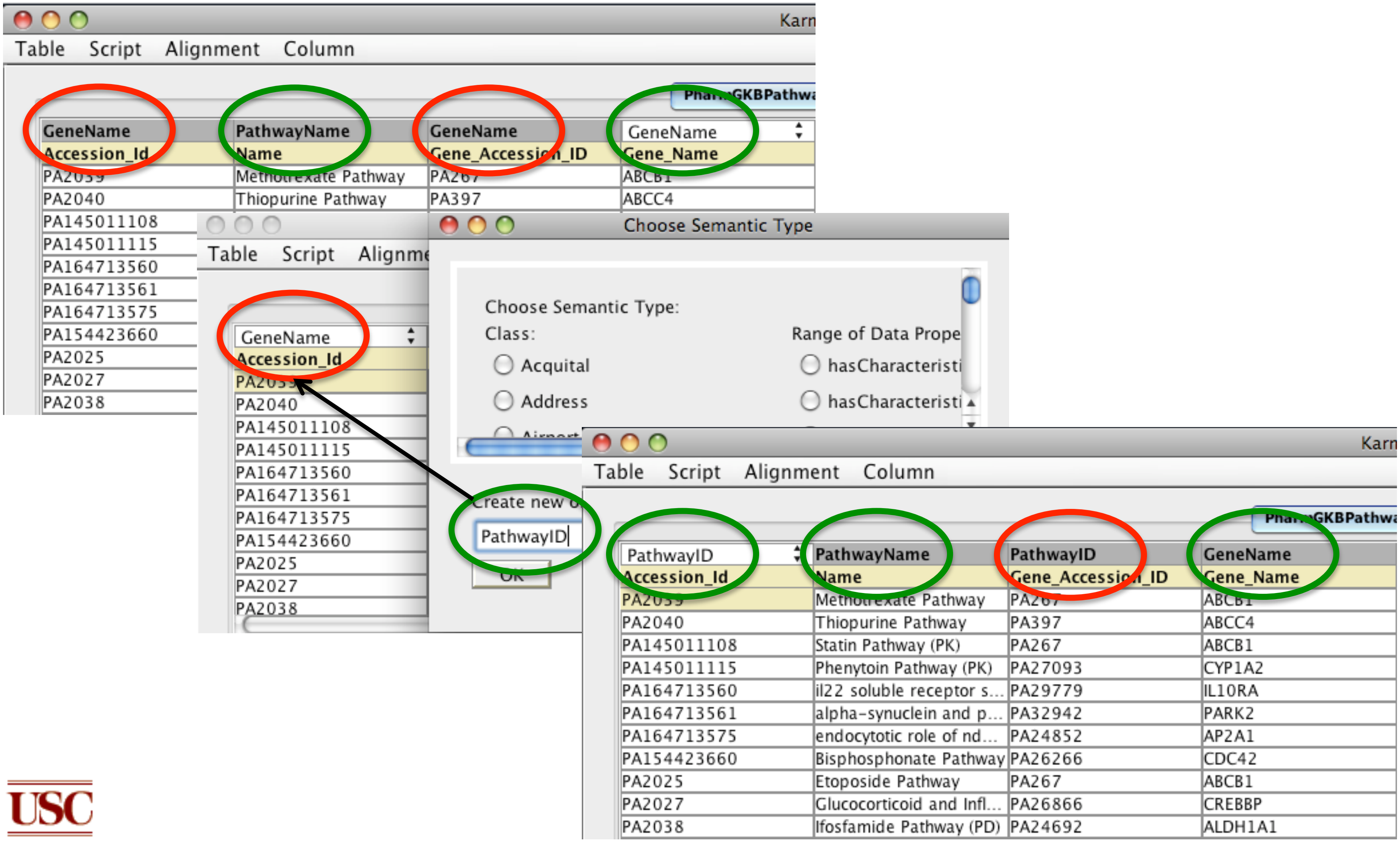
Results

	Weather	Flights	Geocoding
Field labeling accuracy	0.89	0.97	0.98
Token labeling accuracy	0.86	0.87	0.90
Labeling accuracy using regular expressions*	0.65	0.42	0.36

* Semantic labeling of online information sources:

K. Lerman, A. Plangrasopchok, and C. A. Knoblock, IJSWIS, 2006

Karma uses this approach



GeneName Accession_Id	PathwayName Name	GeneName Gene_Accession_ID	GeneName Gene_Name
PA2039	Methotrexate Pathway	PA267	ABCB1
PA2040	Thiopurine Pathway	PA397	ABCC4
PA145011108			
PA145011115			
PA164713560			
PA164713561			
PA164713575			
PA154423660			
PA2025			
PA2027			
PA2038			

GeneName Accession_Id
PA2039
PA2040
PA145011108
PA145011115
PA164713560
PA164713561
PA164713575
PA154423660
PA2025
PA2027
PA2038

PathwayID Accession_Id	PathwayName Name	PathwayID Gene_Accession_ID	GeneName Gene_Name
PA2039	Methotrexate Pathway	PA267	ABCB1
PA2040	Thiopurine Pathway	PA397	ABCC4
PA145011108	Statin Pathway (PK)	PA267	ABCB1
PA145011115	Phenytoin Pathway (PK)	PA27093	CYP1A2
PA164713560	il22 soluble receptor s...	PA29779	IL10RA
PA164713561	alpha-synuclein and p...	PA32942	PARK2
PA164713575	endocytotic role of nd...	PA24852	AP2A1
PA154423660	Bisphosphonate Pathway	PA26266	CDC42
PA2025	Etoposide Pathway	PA267	ABCB1
PA2027	Glucocorticoid and Infl...	PA26866	CREBBP
PA2038	Ifosfamide Pathway (PD)	PA24692	ALDH1A1

Conclusion

Contributions

- Accurately identifying complex semantic types
- Also identify token types
- Fast