Connect

beyond learning – reasoning; why

…… logic

……….. and its limits

……………… fundamental, uncertainty

……………………reasoning under uncertainty

…………………………. back to learning - from text

connecting the dots: motivation

"who is the leader of USA?"

facts … [X is prime-minister of C] … [X is president of C]

no such fact [X is leader of USA] … now what?

X is president of C => X is leader of C – *rules* (*knowledge*)

✓ Obama is president of USA => Obama is leader of USA

example of *reasoning* ..

reasoning can be tricky:

Manmohan Singh is prime-minister of India

Pranab Mukherjee is president of India

"who is the leader of India"

… *much* more *knowledge* is needed

reasoning and web-intelligence

"book me an American flight to NY ASAP"

"this New Yorker who fought at the battle of Gettysburg was once considered the inventor of baseball"

Alexander Cartwright or Abner Doubleday – *Watson got it right*

"who is the Dhoni of USA?"

– *analogical reasoning* - X is to USA what Cricket is to India (?)

\+ *abductive reasoning* – there *is no* US baseball team … so ?

find *best possible answer*ˆ

\+ *reasoning under uncertainty* … who is the "most" popular ?

Semantic Web:

- web of linked *data*, *inference* rules and engines, query
  - pre-requisite: extracting *facts* from text, as well as *rules*

logic: propositions

A, B – ‘propositions’ (either True or False)

A <u>and</u> B is True: A=True and B=True ($A \wedge B$)

A <u>or</u> B is True: either A=True or B=True ($A \vee B$)

if A then B (same as if A=True then B=True)

is the same as saying A=False <u>or</u> B=True

also written as:

A=> B is equivalent to **$\sim A \vee B$**

check: A=T, ~A=F, so ($\sim A \vee B$) =T only when B=T

Important:

if A=F, ~A=T, so ($\sim A \vee B$) is true regardless of B being T or F

logic: predicates

Obama is president of USA:

isPresidentOf (Obama, USA) - *predicates, variables*

X is president of C => X is leader of C

R: isPresidentOf (X, C) => isLeaderOf (X, C)

plus – the above is stating a rule for *all* X,C - *quantification*

"Obama is president of USA": *fact*

F: isPresidentOf (Obama, USA)

using rule R and fact F,

isLeaderOf (Obama, USA) is *entailed*

(*unification*: X *bound* to Obama; C *bound* to USA)

Q: isLeaderOf (X, USA) – *query*

reasoning = answering queries or deriving new facts

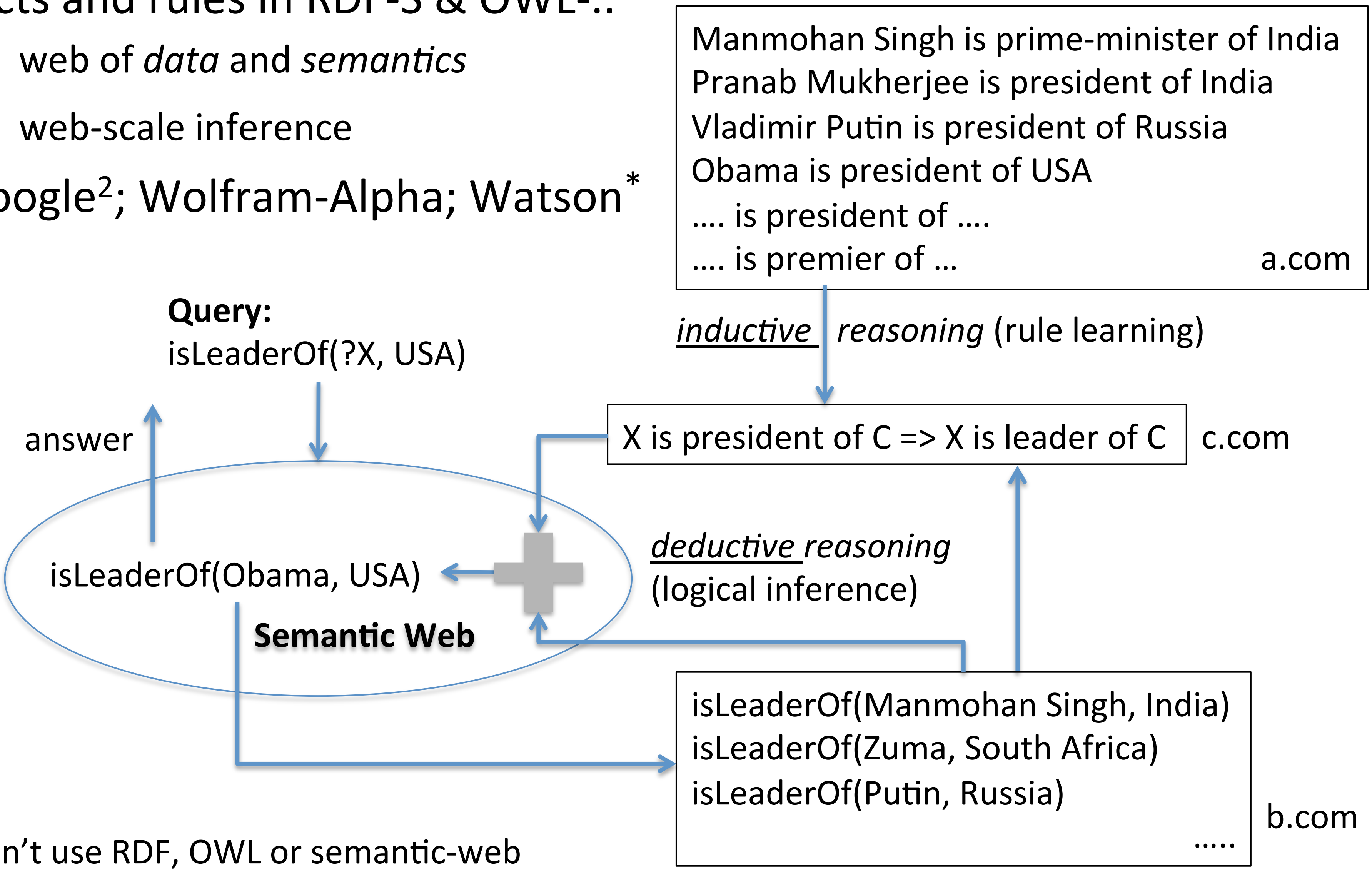
using *unification* + *inference* = *resolution*

semantic web vision

facts and rules in RDF-S & OWL-..

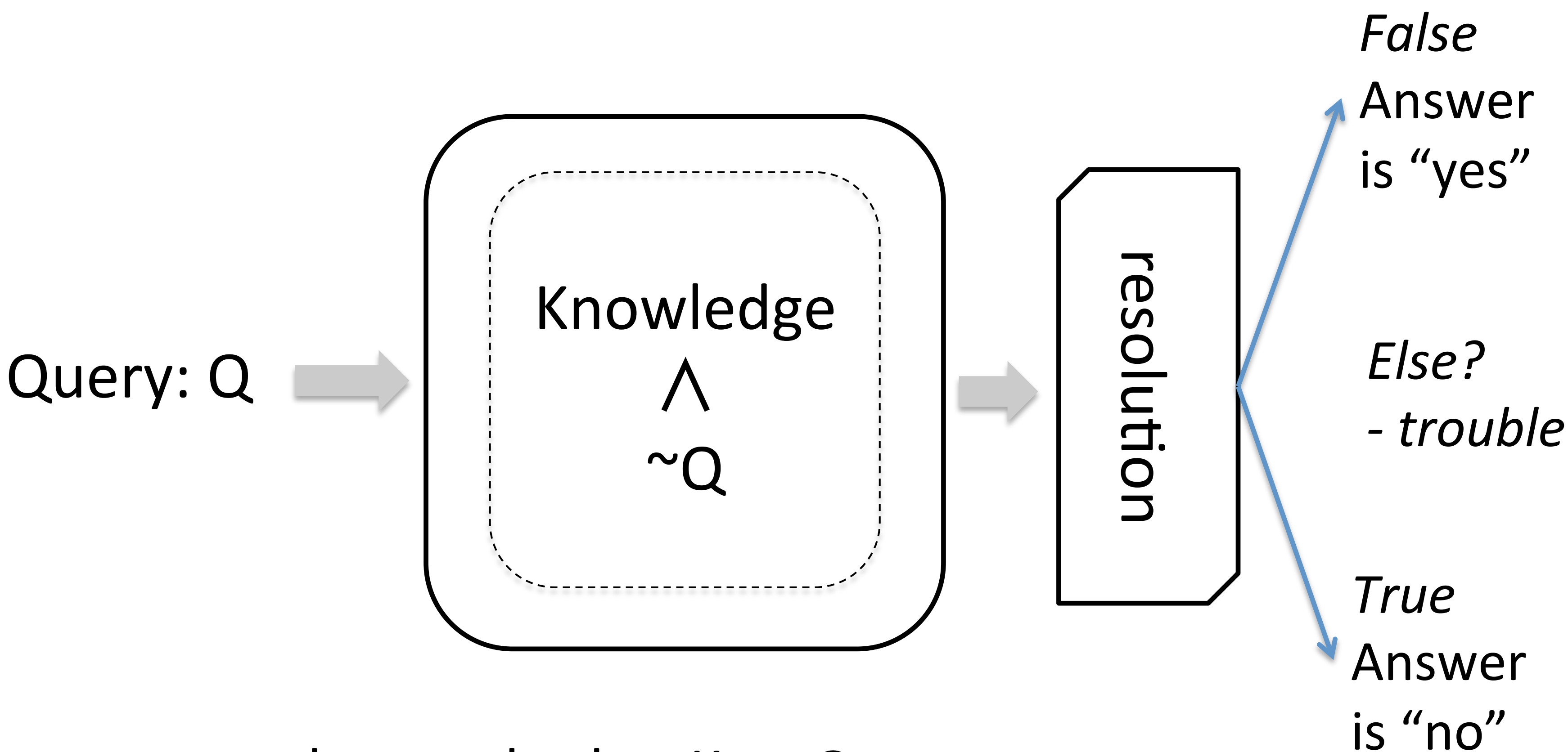
- web of *data* and *semantics*
- web-scale inference

Google[2]; Wolfram-Alpha; Watson*



*don't use RDF, OWL or semantic-web *technology* though they have similar intent, spirit …

logical inference: resolution



we want to know whether K => Q

i.e. ~K ∨ Q is True

i.e. K ∧ ~Q is False !

in other words K augmented with ~Q *entails* falsehood, for sure

logic: fundamental limits

resolution may never end; *never* (whatever algorithm!)

- ➢ undecidability

  predicate logic undecidable (Godel, Turing, Church …)

- ➢ intractability

  propositional logic is decidable, but intractable (SAT and NP ..)

? whither automated reasoning, semantic-web..?

fortunately:

OWL-DL,OWL-lite (description logic: leader $\sqsubset$ person …)

decidable; still intractable in worst case

Horn logic (rules, i.e., person $\wedge$ bornIn(C) => citizen(C) …)

undecidable (except with caveats); but tractable

logic and uncertainty

predicates *A, B, C*

1. For all x, $A(x) => B(x)$.
2. For all x, $B(x) => C(x)$

1 and 2 *entail* For all x, $A(x) => C(x)$ <u>fundamental</u>

however, consider the *uncertain* statements:

1': For **most** x, $A(x) => B(x)$. *"most firemen are men"*

2'. For **most** x, $B(x) => C(x)$. *"most men have safe jobs"*

it *does **not*** follow that "For **most** x, $A(x) => C(x)$" !

logic and causality

- if the sprinkler was on then the grass is wet

  S => W

- if the grass is wet then it had rained

  W => R

therefore it follows, i.e. *S => R* is *entailed*

which states “the sprinkler is on, so it had rained”

- ➢ *problem is that causality was treated differently in each statement => absurdity*

causality and classification

if S then W (W is an observable *feature* of S)

$$S \longrightarrow W$$

if R then W (W is an observable *feature* of R)

$$\begin{tikzcd} R \arrow[r] & W \arrow[l, bend left, dashed] \end{tikzcd}$$

if *W* is observed then *R* happened *(abduction)*

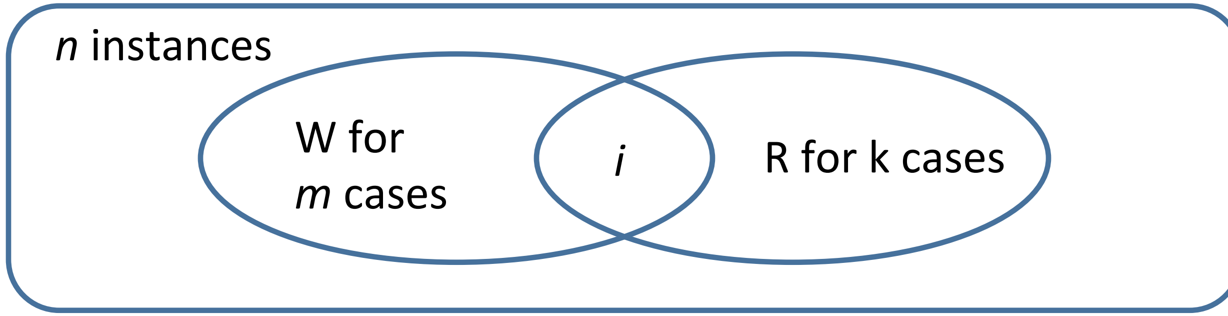
concluding which *class of event* observed S or R

abductive reasoning

= from effects to likely causes

probability tables and ‘marginalization’

#	W	R
1	y	n
2	y	y
3	n	n



consider $p(\mathrm{R,W})$

to get $p(\mathrm{R})$ we can ‘sum out’ W: $p(\mathrm{R}) = \sum_{\mathrm{W}} p(\mathrm{R,S})$

this is called *marginalization* of *W*

notice that marginalization is equivalent to *aggregation* on column P:

$\sum_{\mathrm{W}} p(\mathrm{R,W}) = {}_{\mathrm{R}}G_{\mathrm{SUM(P)}}\ T^{\mathrm{R,W}}$

or, in SQL:
SELECT R, SUM(P) from $T^{\mathrm{R,W}}$
GROUP BY R

R	P
y	k/n
n	(n-k)/n

$= \sum_{\mathrm{w}}$

R	W	P
y	y	i/n
n	y	(m-i)/n
y	n	(k-i)/n
n	n	(n-m-k+i)/n

$P(\mathrm{R,W}) = T^{\mathrm{R,W}}$

probability tables and Bayes rule ...

R	W	P
y	y	i/n
n	y	(m-i)/n
y	n	(k-i)/n
n	n	(n-m-k+i)/n

p(R,W)

=

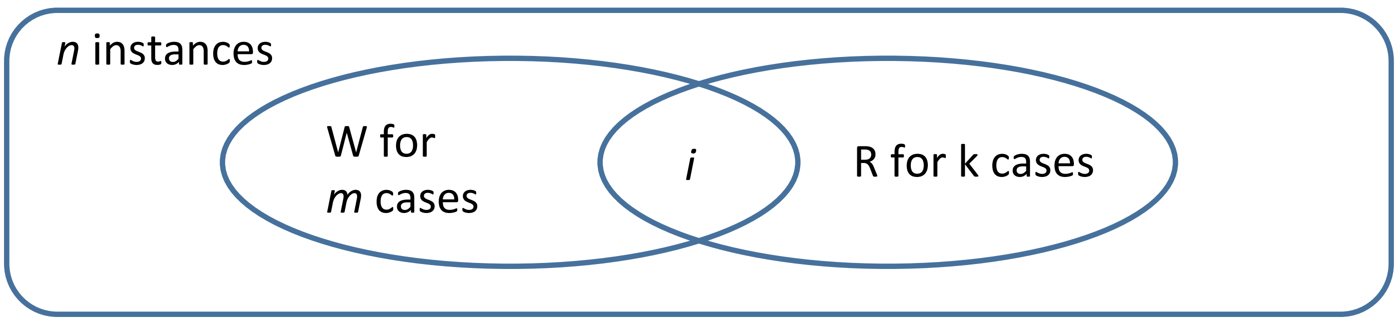
R	W	P
y	y	i/m
n	y	(m-i)/m
y	n	k-i/(n-m)
n	n	(n-m-k+i)/(n-m)

p(R|W)

*

W	P
y	m/n
n	(n-m)/n

p(W)



probability tables and Bayes rule …

R	W	P
y	y	i/n
n	y	(m-i)/n
y	n	(k-i)/n
n	n	(n-m-k+i)/n

p(R,W)

=

R	W	P
y	y	i/m
n	y	(m-i)/m
y	n	k-i/(n-m)
n	n	(n-m-k+i)/(n-m)

p(R|W)

*

W	P
y	m/n
n	(n-m)/n

p(W)

$T_0^{R,W}$ $T_1^{R,W}$ T_2^{W}

notice that the product $p(R|W)\ p(W) = T_1^{R,W} \bowtie_B T_2^{W}$

i.e., the *join* of the two tables T_1 and T_2 on the common attribute W!

so, probability tables (also called *potentials*) can be multiplied in SQL!

SELECT R, SUM(P1*P2) from $T_1^{R,W}$, T_2^{W} WHERE W1=W2 GROUP BY R

probability tables and *evidence*

R	W	P
y	y	i/n
n	y	(m-i)/n
y	n	(k-i)/n
n	n	(n-m-k+i)/n

$\mathbf{e}^{(B=y)}$ =

R	W	P
y	y	i/n
n	y	(m-i)/n

=

R	W	P
y	y	i/m
n	y	(m-i)/m

$* \; m/n$

$P(R,W)$ $\quad$ $P(R,W)\,\mathbf{e}^{(W=y)}$ $\quad$ $P(R|W=y)$ $\quad$ $*\; p(W=y)$

$= T^{R,W}$

SELECT R,W,P from $T^{R,W}$ WHERE W=y

if we restrict $p(R,W)$ to entries where *evidence* W=y holds:

$$p(R,W)\,\mathbf{e}^{(W=y)} = p(R|W=y) * p(\mathbf{e}^{(W=y)})$$

applying evidence is equivalent to the *select* operator on $T^{R,W}$

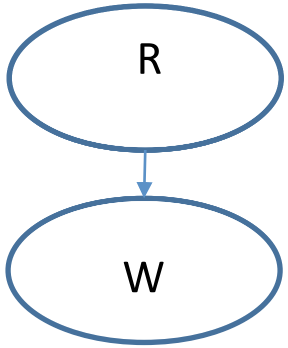
$$P(R,W)\,\mathbf{e}^{(W=y)} = \sigma_{W=y}\, T^{R,W}$$

so the ***a posteriori*** probability of R given evidence **e** is just:

$$P(R|\mathbf{e}^{(W=y)}) = p(R,W)\,\mathbf{e}^{(W=y)} \,/\, p(\mathbf{e}^{(W=y)})$$

A	P
y	i/m
n	(m-i)/m

naïve Bayes classifier



p(W|R)

W	R	P
y	y	.9
n	y	.1
y	n	.2
n	n	.8

p(R)

R	P
y	.2
n	.8

$p(R,W) = p(R|W)\ p(W) = p(W|R)\ p(R)$ - Bayes Rule again

given evidence $\mathbf{e}^{W=y}$:

$p(R,W)\mathbf{e}^{W=y} = p(R|W=y)p(W=y) = p(W=y|R)p(R)$

or, $\boldsymbol{p(R|W=y)} = p(W=y|R)\ p(R)/p(W=y) = \boldsymbol{\sigma\ p(W=y|R)\ p(R)}$

in SQL:

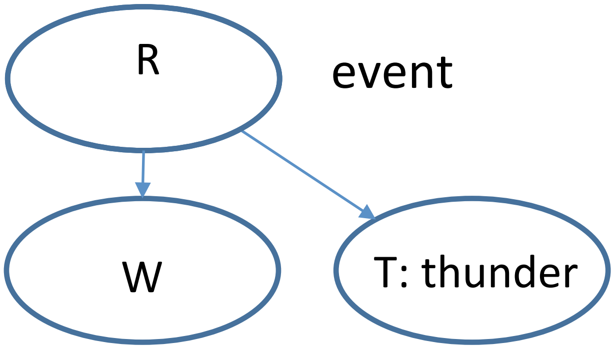
SELECT R, SUM(P*P) FROM $T_1^{R,W}$, T_2^{R} WHERE W=y

AND R=R GROUP BY R

normalizing so that $\Sigma_R=1$: $p(R=y) = .18 / (.18+.16) = 53\%$

R	P
y	.9*.2 = .18
n	.2*.8 = .16

naïve Bayes classifier: multiple features



p(W|R)

W	R	P
y	y	.9
n	y	.1
y	n	.2
n	n	.8

p(R)

R	P
y	.2
n	.8

$p(R,T|W{=}y) = \sigma\, p(W{=}y|R)\, p(T|R)\, p(R)$ – but we need to "sum out" T

SELECT R, SUM(P*P*P) FROM $T_1^{W,R}\ T_2^{T,R}\ T_3^{R}$

WHERE W=y AND $R_1{=}R_2,\ R_2{=}R_3$ GROUP by R

same result!

obviously,

since there

was *no* new **evidence!**

R	P
y	.9*.2 = .18
n	.2*.8 = .16

=

R	P
y	.9*.2 = .18
n	.2*.8 = .16

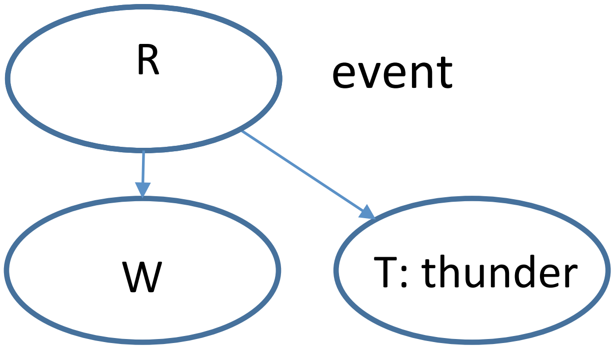
$\bowtie_R$

p(T|R)

T	R	P
y	y	.8
n	y	.2
y	n	.1
n	n	.9

NOTE: same as Bayesian classifier with *partial* evidence …

naïve Bayes classifier: new evidence



event

$p(W|R)$

W	R	P
y	y	.9
n	y	.1
y	n	.2
n	n	.8

$p(R)$

R	P
y	.2
n	.8

$p(R|W{=}y,T{=}y) = \sigma\, p(W{=}y|R)\, \boldsymbol{p(T{=}y|R)}\, p(R)$

SELECT R, SUM(P*P*P) FROM $T_1^{W,R}\ T_2^{T,R}\ T_3^{R}$

WHERE W=y, T=y AND $R_1=R_2,\ R_2=R_3$ GROUP by R

now,

new evidence

gives:

R	P
y	.9*.2*.8 = .144
n	.2*.8*.1 = .016

=

R	P
y	.9*.2 = .18
n	.2*.8 = .16

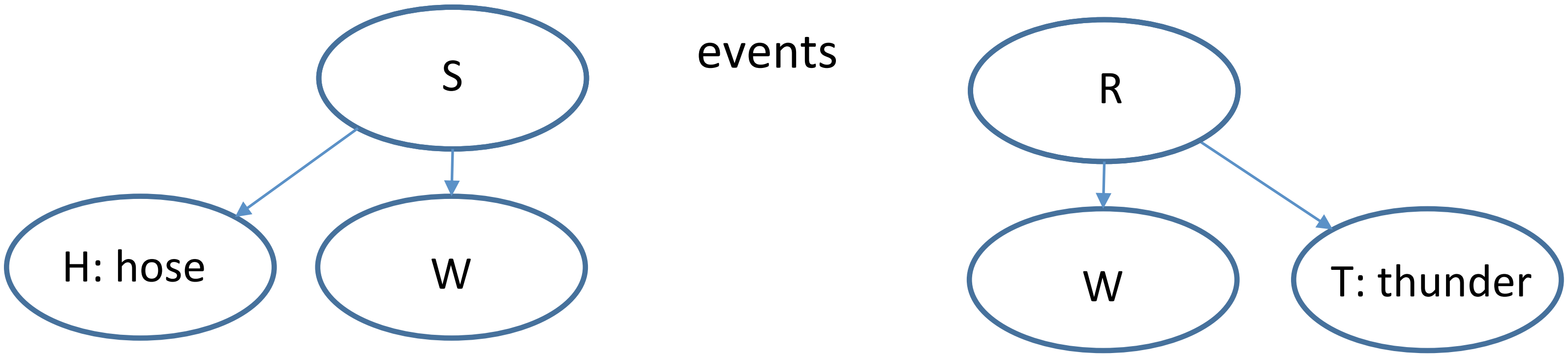
$\bowtie_R$

$p(T|R)$

T	R	P
y	y	.8
n	y	.2
y	n	.1
n	n	.9

$p(R{=}y|\mathbf{e}) = .144/(.144+.016) = .9$ or 90%! .... **belief revision**

multiple naïve Bayes classifiers

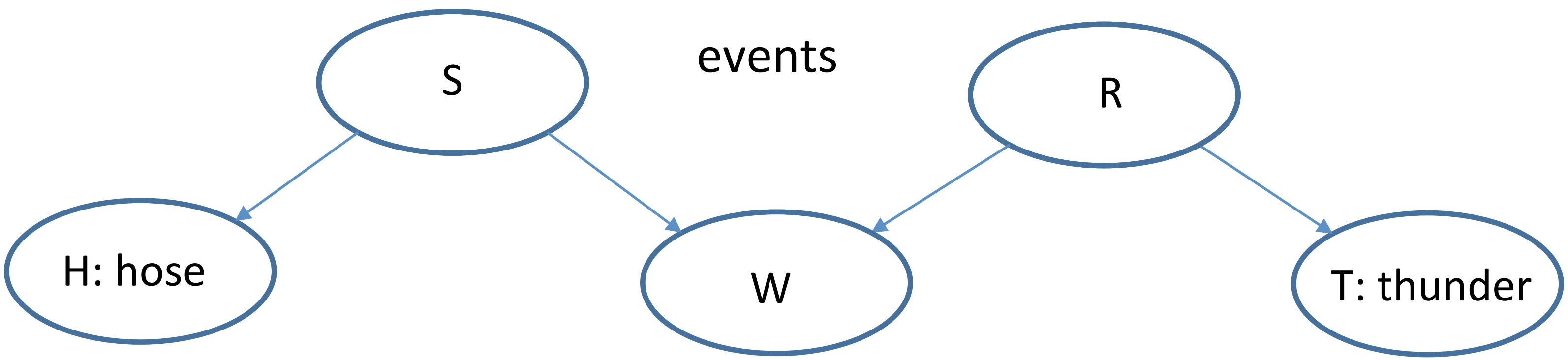


$p(\mathrm{R}\,|\,\mathrm{W,T}) = \sigma_1\, p(\mathrm{W}\,|\,\mathrm{R})\; p(\mathrm{T}\,|\,\mathrm{R})\; p(\mathrm{R})$

$p(\mathrm{S}\,|\,\mathrm{H,W}) = \sigma_2\, p(\mathrm{H}\,|\,\mathrm{S})\; p(\mathrm{W}\,|\,\mathrm{S})\; p(\mathrm{S})$

but … W is the same observation …

Bayesian network



$p(H,W,T,S,R) = p(H,W,T,S|R)\ p(R) = p(H,W,T|S,R)\ p(S|R)\ p(R)$

independence of S,R; also features H,T,W | S,R independent:

$p(H,W,T,S,R) = p(H|S,R)\ p(W|S,R)\ p(T|S,R)\ p(S)\ p(R)$

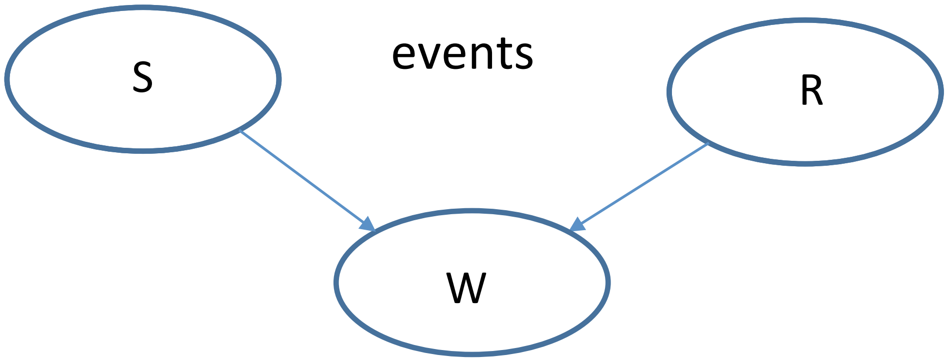
and… this is tricky … H,R and T,S also independent

$p(H,W,T,S,R) = p(H|S)\ p(W|S,R)\ p(T|R)\ p(S)\ p(R)$ □

general rule - only conditionals based on parents needed

use SQL as before: apply evidence, join and aggregate

inference in Bayesian networks



CPT
$p(W|S,R)$
not joint!

W	S	R	P
y	y	y	.9
y	y	n	.7
y	n	y	.8
y	n	n	.1
n	n	n	.9
n	n	y	.2
n	y	n	.3
n	y	y	.1

$p(R)$

R	P
y	.2
n	.8

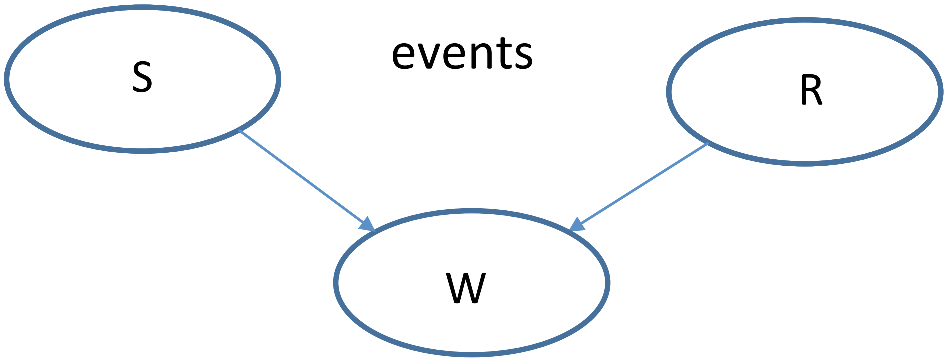
$P(S)$

S	P
y	.3
n	.7

$p(R,S,W) = p(R,S|W)\ p(W)$ and
$p(R,S,W) = p(W|S,R)\ p(S)\ p(R)$
evidence$_1$: "grass is wet", W=y
$\Sigma_S\, p(R,S|W{=}y) = \sigma\, \Sigma_S\, p(W|R,S)\ p(S)\ p(R)\ \mathbf{e}^{W=y}$ in SQL:
SELECT R, SUM(P*) FROM T.. WHERE W=y, R=R,S=S
GROUP BY R
normalizing so that sum is 1:
$p(R{=}y|W{=}y) = .166/(.166+.224) = .42$
i.e. 42%

W	R	P
y	y	(.9*.3+.8*.7)*.2=.166
y	n	(.7*.3+.1*.7)*.8=.224

inference continued: "explaining away" effect



p(R)

R	P
y	.2
n	.8

P(S)

S	P
y	.3
n	.7

W	S	R	P
y	y	y	.9
y	y	n	.7
y	n	y	.8
y	n	n	.1
n	n	n	.9
n	n	y	.2
n	y	n	.3
n	y	y	.1

evidence$_1$: "grass is wet", W=y

AND evidence$_2$: "sprinkler on", S=y

$p(R|W,S) = \sigma\, p(W|R,S)\, p(S)\, p(R)\, \mathbf{e}^{W=y,S=y}$ in SQL:

SELECT R, SUM(P*) FROM T.. WHERE W=y, S=y, R=R

GROUP BY R

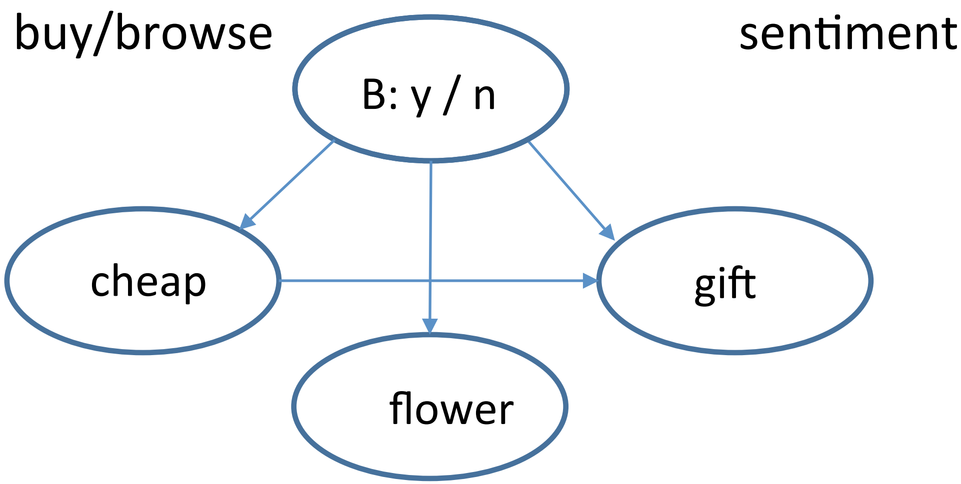
normalizing so that sum is 1:

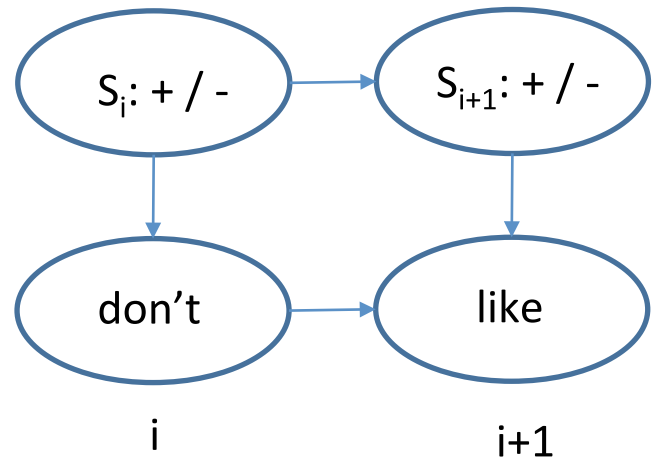
$p(R=y|W=y,S=Y) = .054/(.054+.168) = .24$

W	R	P
y	y	(.9*.3)*.2=.054
y	n	(.7*.3)*.8=.168

i.e. 24%: *less than the earlier* 42% - belief propagation

Bayes nets: beyond independent features





if 'cheap' and 'gift' are *not* independent, $P(G|C,B) \neq P(G|B)$

(or use $P(C|G,B)$, depending on the order in which we *expand* $P(G,C,B)$)

"I don't like the course" and "I like the course; don't complain!"

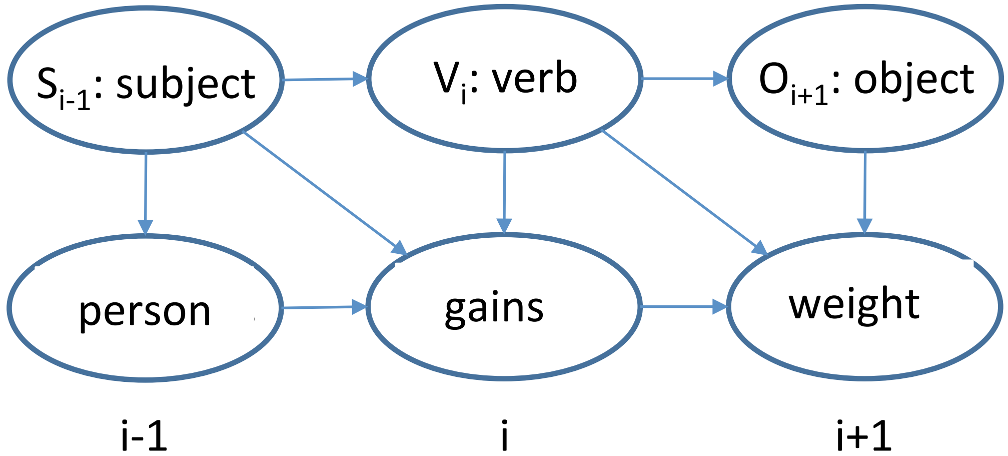
first, we might include "don't" in our list of features (also "not" …)

still – might not be able to disambiguate: need *positional order*

$P(x_{i+1}|x_i, S)$ for each position *i*: hidden markov model (HMM)

we may also need to accommodate 'holes', e.g. $P(x_{i+k}|x_i, S)$

where do facts come from? learning from text



suppose we want to learn *facts* of the form <subject, verb, object> from text

single class variable is not enough; (i.e. we have many y_j in data [Y,X])

further, positional order is important, so we can use a (different) HMM ..

e.g. we need to know $P(x_i | x_{i-1}, S_{i-1}, V_i)$

whether 'kill' following 'antibiotics' is a verb will depend on whether 'antibiotics' is a subject

more apparent for the case <person, gains, weight>, since 'gains' can be a verb or a noun

problem reduces to estimating *all* the a-posterior probabilities $P(S_{i-1}, V_i, O_{i+1})$

for every *i* , and also allowing 'holes' (i.e., $P(S_{i-k}, V_i, O_{i+p})$) and find the *best* facts from a collection of text? …. many solutions; apart from HMMs - CRFs

after finding all facts from lots of text, we cull using support, confidence, etc.

open information extraction

Cyc (older, semi-automated): 2 billion facts

Yago – largest to date: 6 billion facts, linked i.e., a graph

e.g. <Albert Einstein, wasBornIn, Ulm>

Watson – uses facts culled from the web internally

REVERB – recent, lightweight: 15 million S,V,O triples

e.g. <potatoes, are also rich in, vitamin C>

1. part-of-speech tagging using NLP classifiers (trained on labeled corpora)
2. focus on verb-phrases; identify nearby noun-phrases
3. prefer proper nouns, especially if they occur often in other facts
4. extract more than one fact if possible:

“Mozart was born in Salzburg, but moved to Vienna in 1781” yields

<Mozart, moved to, Vienna>, in addition to <Mozart, was born in, Salzburg>

belief networks: learning, logic, big-data & AI

- network *structure* can be learned from data
- applications in [genomic] medicine
  - medical diagnosis
  - gene-expression networks
  - how do phenotype traits arise from genes
- logic and uncertainty
  - belief networks bridging the gap:
  - (Pearl Turing award; Markov logic n/w …)
- big-data
  - inference can be done using SQL – map-reduce works!
- hidden-agenda:
  - deep belief networks
  - linked to connectionist models of brain