Sistemi Intelligenti 2
2009 / 2010

Umberto Straccia

ISTI-CNR, Pisa, Italy
http://www.straccia.info
straccia@isti.cnr.it

Uncertainty, Vagueness, and the Semantic Web

Sources of Uncertainty and Vagueness on the Web

- (Multimedia) Information Retrieval:
  - To which degree is a Web site, a Web page, a text passage, an image region, a video segment, . . . relevant to my information need?
- Matchmaking
  - To which degree does an object match my requirements?
    - if I'm looking for a car and my budget is *about* 20.000 €, to which degree does a car's price of 20.500 € match my budget?

- Semantic annotation / classification
  - To which degree does e.g., an image object represent or is about a dog?


“White Dog Cafe”

- Information extraction
  - To which degree am I’m sure that e.g., SW is an acronym of “Semantic Web”?

- Ontology alignment (schema mapping)
  - To which degree do two concepts of two ontologies represent the same, or are disjoint, or are overlapping?
    - For instance, to which degree are are SUVs and Sports Cars overlapping?

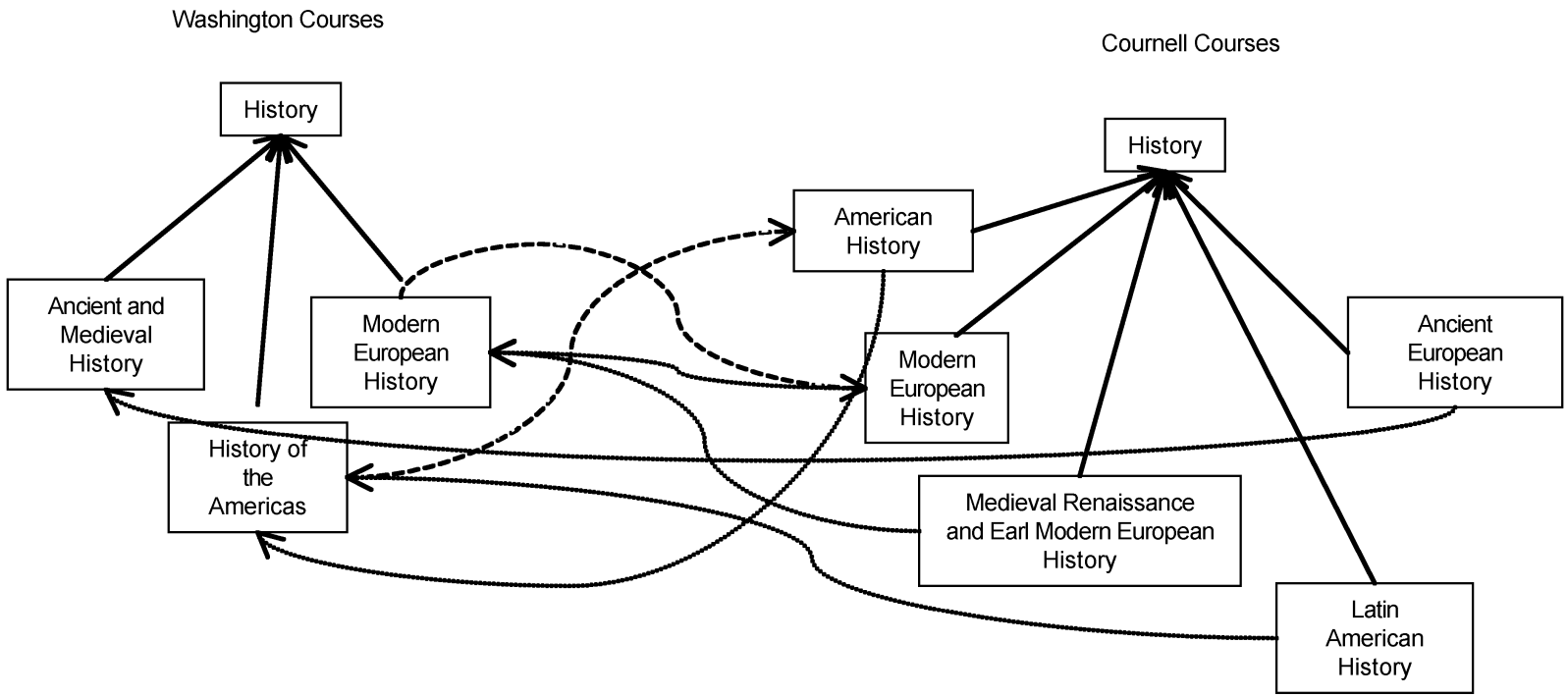
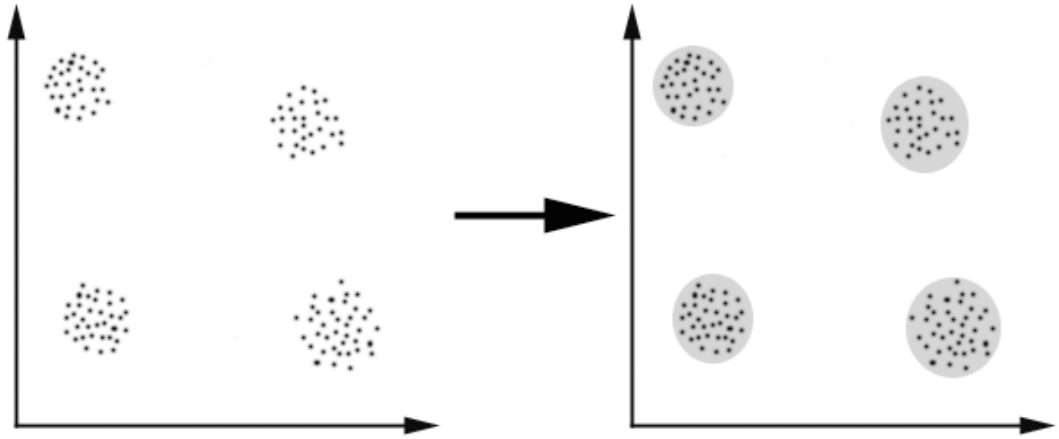


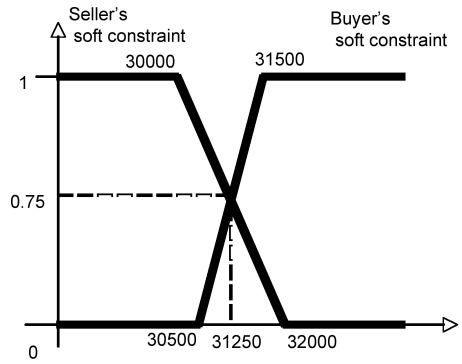
Figure: The excerpt of two ontologies and category matchings

- Similarity: To which degree are two objects similar?
- Clustering: Do a set of objects from a group (cluster) of similar objects ?



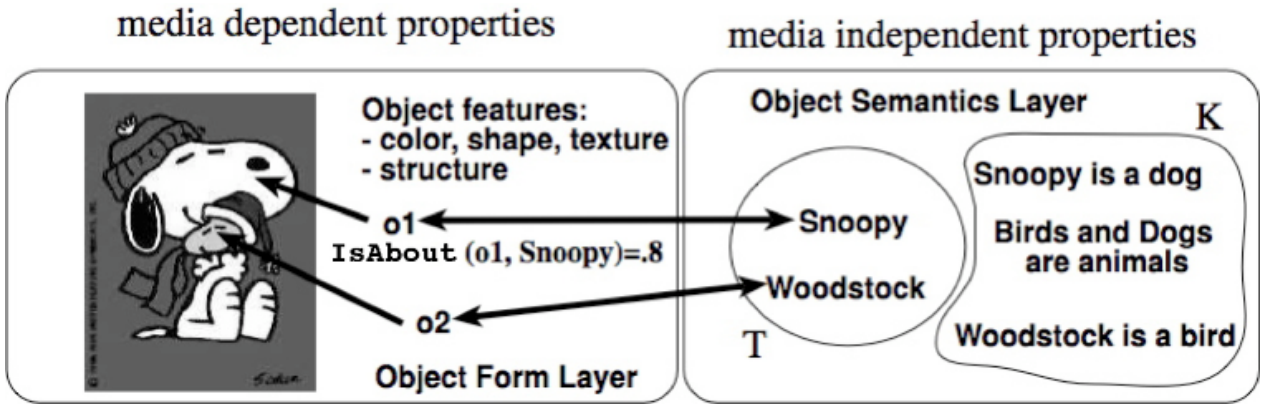
- Representation of background knowledge
  - To some degree birds fly.
  - To some degree Jim is a blond and young.

Example (Matchmaking)



- A car seller sells an Audi TT for 31500 €, as from the catalog price.
- A buyer is looking for a sports-car, but wants to to pay not more than around 30000 €
- Classical DLs: the problem relies on the crisp conditions on price.
- More fine grained approach: to consider prices as vague constraints (fuzzy sets) (as usual in negotiation)
  - Seller would sell above 31500 €, but can go down to 30500 €
  - The buyer prefers to spend less than 30000 €, but can go up to 32000 €
  - Highest degree of matching is 0.75 . The car may be sold at 31250 €.

Example (Multimedia information retrieval)

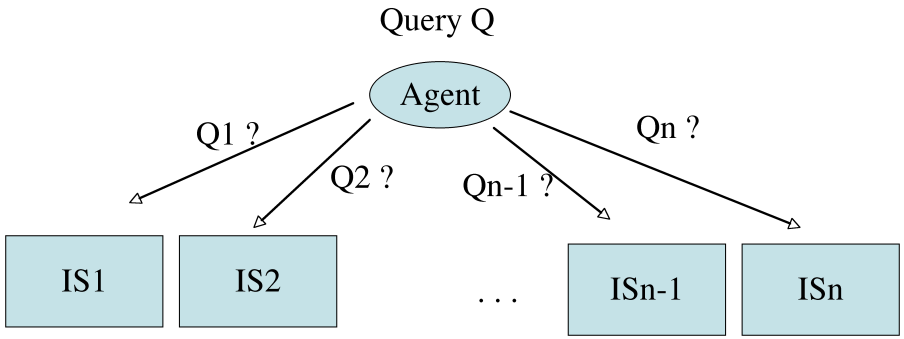


IsAbout		
ImageRegion	*Object ID*	*degree*
*o*1	*snoopy*	0.8
*o*2	*woodstock*	0.7
⋮	⋮	

"Find top-k image regions about animals"

$$Query(x) \leftarrow ImageRegion(x) \wedge isAbout(x, y) \wedge Animal(y)$$

Example (Distributed Information Retrieval)



Then the agent has to perform automatically the following steps:

1. The agent has to select a subset of relevant resources $\mathscr{S}' \subseteq \mathscr{S}$, as it is not reasonable to assume to access to and query all resources (resource selection/resource discovery);
2. For every selected source $\mathcal{S}_i \in \mathscr{S}'$ the agent has to reformulate its information need Q_A into the query language $\mathcal{L}_i$ provided by the resource (schema mapping/ontology alignment);
3. The results from the selected resources have to be merged together (data fusion/rank aggregation)

Example (Database query)

HotelID	*hasLoc*
*h*1	*hl*1
*h*2	*hl*2
⋮	⋮

ConferenceID	*hasLoc*
*c*1	*cl*1
*c*2	*cl*2
⋮	⋮

hasLoc	*hasLoc*	*distance*
*hl*1	*cl*1	300
*hl*1	*cl*2	500
*hl*2	*cl*1	750
*hl*2	*cl*2	800
⋮	⋮	

hasLoc	*hasLoc*	*close*	*cheap*
*hl*1	*cl*1	0.7	0.3
*hl*1	*cl*2	0.5	0.5
*hl*2	*cl*1	0.25	0.8
*hl*2	*cl*2	0.2	0.9
⋮	⋮	⋮	⋮

"Find top-k cheapest hotels close to the train station"

$$q(h) \leftarrow \mathit{hasLocation}(h, hl) \wedge \mathit{hasLocation}(\texttt{train}, cl) \wedge \mathit{close}(hl, cl) \wedge \mathit{cheap}(h)$$

Example Decision Making

Electrical power dispatching system in the case of shortage of electrical power

- There are four regions of a city
- We have to decide to which to give electricity in the case of shortage of electrical power
- The criteria we are considering is based on the electricity demand of
  - Residential area
  - Shopping centers
  - Clubs and recreation centers
  - Educational centers
  - Medical urgent care centers

Example Decision Making (cont.)

Shall I go hiking this weekend?

- It typically snows about 5% of the days during the winter
- The Weather Channel (TWC) says there is a 70% chance of snow on this weekend
- Question: What is the chance that it will snow this weekend?

Example (Health-care: diagnosis of pneumonia)

ICSI

INSTITUTE FOR CLINICAL SYSTEMS IMPROVEMENT

Health Care Guideline:

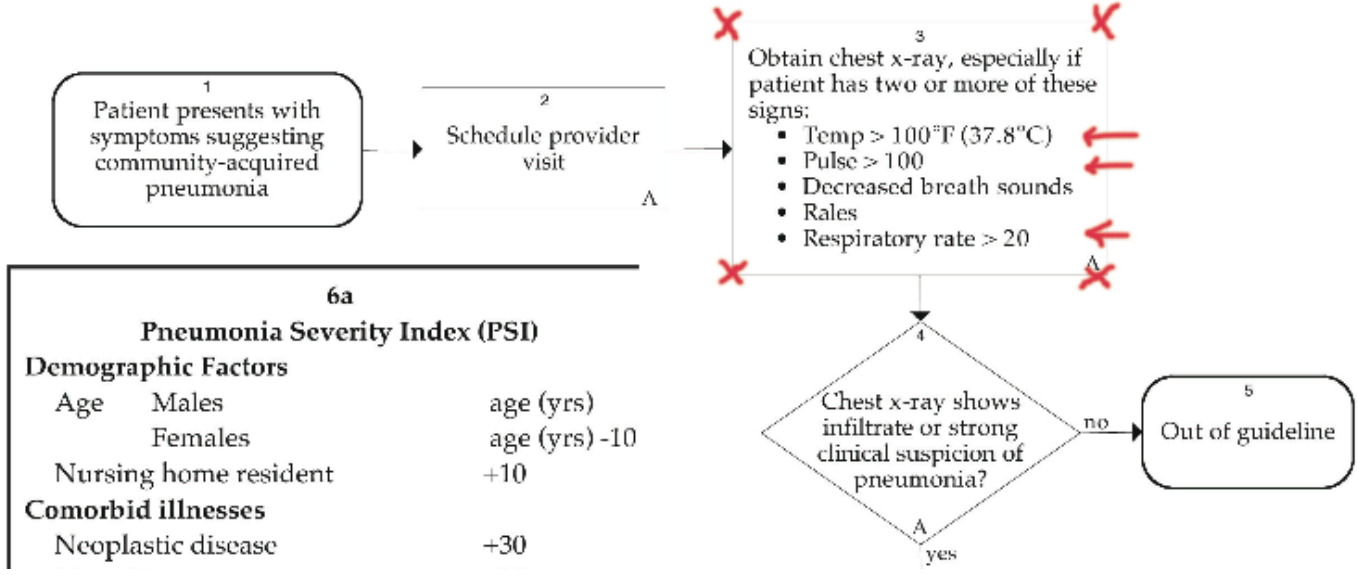
Community-Acquired Pneumonia in Adults

Seventh Edition
May 2006

Work Group Leader
John Degelau, MD
Internal Medicine,
HealthPartners Medical Group

Work Group Members

Family Medicine
Garrett Trobec, MD

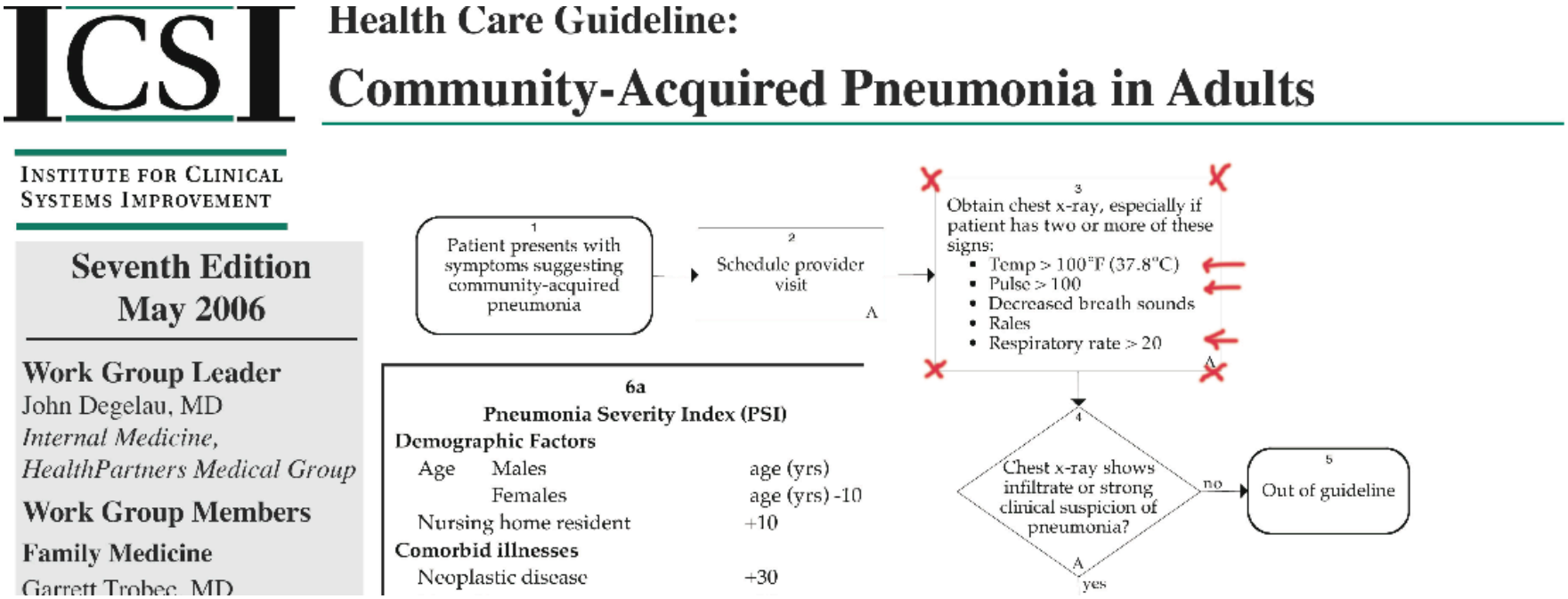


6a

Pneumonia Severity Index (PSI)

Demographic Factors		
Age	Males	age (yrs)
	Females	age (yrs) -10
Nursing home resident		+10
Comorbid illnesses		
Neoplastic disease		+30

Example (Health-care: diagnosis of pneumonia)



ICSI

INSTITUTE FOR CLINICAL SYSTEMS IMPROVEMENT

Health Care Guideline:

Community-Acquired Pneumonia in Adults

Seventh Edition
May 2006

Work Group Leader
John Degelau, MD
Internal Medicine,
HealthPartners Medical Group

Work Group Members

Family Medicine
Garrett Trobec, MD

1
Patient presents with symptoms suggesting community-acquired pneumonia

2
Schedule provider visit
A

3
Obtain chest x-ray, especially if patient has two or more of these signs:
- Temp > 100°F (37.8°C)
- Pulse > 100
- Decreased breath sounds
- Rales
- Respiratory rate > 20

A

4
Chest x-ray shows infiltrate or strong clinical suspicion of pneumonia?
A
yes
no

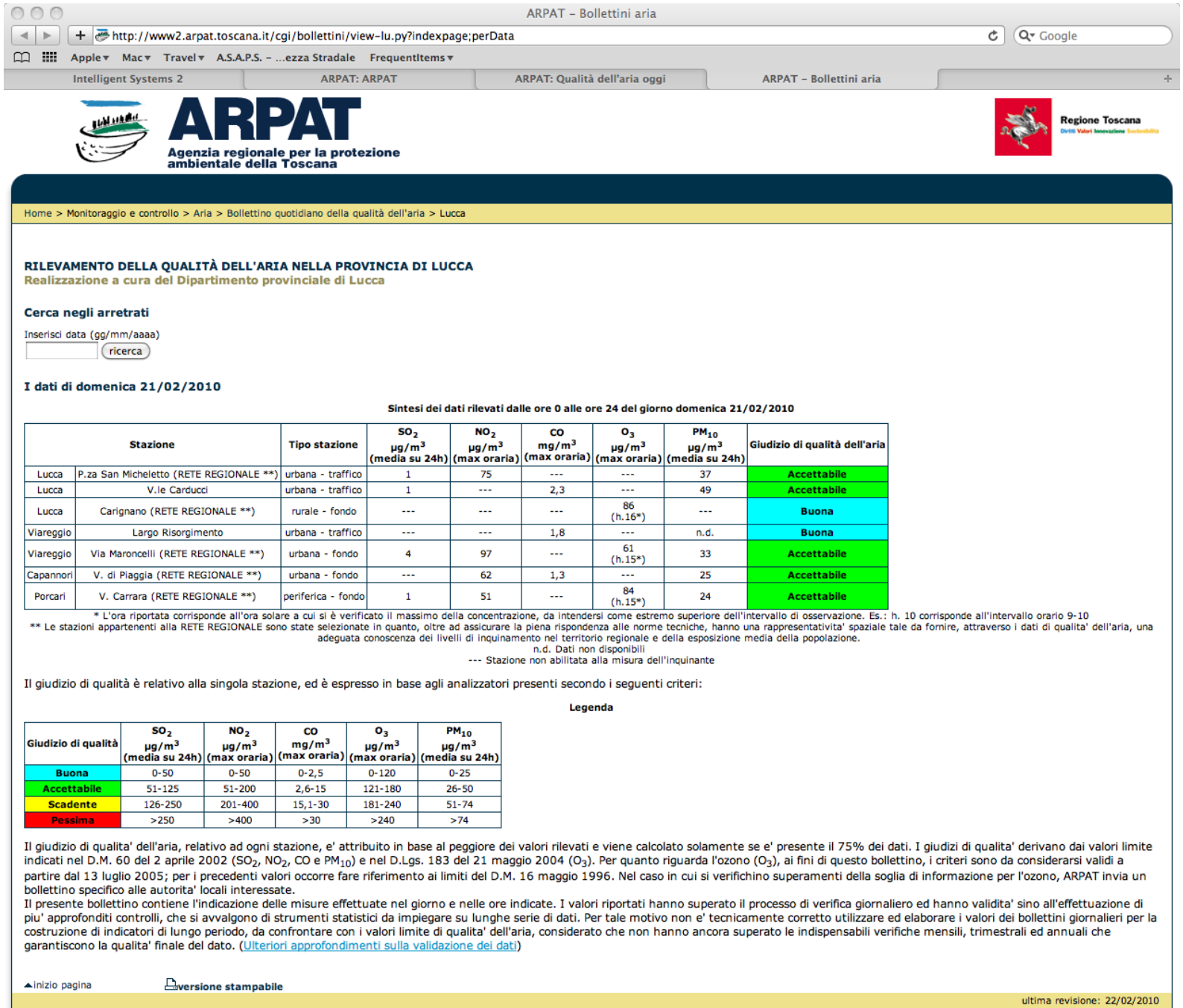
5
Out of guideline

6a
Pneumonia Severity Index (PSI)

Demographic Factors		
Age	Males	age (yrs)
	Females	age (yrs) -10
Nursing home resident		+10
Comorbid illnesses		
Neoplastic disease		+30

- E.g., $\mathit{Temp} = 37.5$, $\mathit{Pulse} = 98$, $\mathit{RespiratoryRate} = 18$ are in the “danger zone” already
- Temperature, Pulse and Respiratory rate, . . . : these constraints are rather imprecise than crisp

ARPAT: Air quality in the province of Lucca



ARPAT – Bollettini aria

http://www2.arpat.toscana.it/cgi/bollettini/view-lu.py?indexpage;perData

Apple Mac Travel A.S.A.P.S. – ...ezza Stradale FrequentItems

Intelligent Systems 2 | ARPAT: ARPAT | ARPAT: Qualità dell'aria oggi | ARPAT – Bollettini aria

ARPAT
Agenzia regionale per la protezione ambientale della Toscana

Regione Toscana

Home > Monitoraggio e controllo > Aria > Bollettino quotidiano della qualità dell'aria > Lucca

RILEVAMENTO DELLA QUALITÀ DELL'ARIA NELLA PROVINCIA DI LUCCA
Realizzazione a cura del Dipartimento provinciale di Lucca

Cerca negli arretrati

Inserisci data (gg/mm/aaaa)
ricerca

I dati di domenica 21/02/2010

Sintesi dei dati rilevati dalle ore 0 alle ore 24 del giorno domenica 21/02/2010

Stazione		Tipo stazione	SO_2 µg/m³ (media su 24h)	NO_2 µg/m³ (max oraria)	CO mg/m³ (max oraria)	O_3 µg/m³ (max oraria)	PM_{10} µg/m³ (media su 24h)	Giudizio di qualità dell'aria
Lucca	P.za San Micheletto (RETE REGIONALE **)	urbana - traffico	1	75	---	---	37	Accettabile
Lucca	V.le Carducci	urbana - traffico	1	---	2,3	---	49	Accettabile
Lucca	Carignano (RETE REGIONALE **)	rurale - fondo	---	---	---	86 (h.16*)	---	Buona
Viareggio	Largo Risorgimento	urbana - traffico	---	---	1,8	---	n.d.	Buona
Viareggio	Via Maroncelli (RETE REGIONALE **)	urbana - fondo	4	97	---	61 (h.15*)	33	Accettabile
Capannori	V. di Piaggia (RETE REGIONALE **)	urbana - fondo	---	62	1,3	---	25	Accettabile
Porcari	V. Carrara (RETE REGIONALE **)	periferica - fondo	1	51	---	84 (h.15*)	24	Accettabile

* L'ora riportata corrisponde all'ora solare a cui si è verificato il massimo della concentrazione, da intendersi come estremo superiore dell'intervallo di osservazione. Es.: h. 10 corrisponde all'intervallo orario 9-10
** Le stazioni appartenenti alla RETE REGIONALE sono state selezionate in quanto, oltre ad assicurare la piena rispondenza alle norme tecniche, hanno una rappresentativita' spaziale tale da fornire, attraverso i dati di qualita' dell'aria, una adeguata conoscenza dei livelli di inquinamento nel territorio regionale e della esposizione media della popolazione.
n.d. Dati non disponibili
--- Stazione non abilitata alla misura dell'inquinante

Il giudizio di qualità è relativo alla singola stazione, ed è espresso in base agli analizzatori presenti secondo i seguenti criteri:

Legenda

Giudizio di qualità	SO_2 µg/m³ (media su 24h)	NO_2 µg/m³ (max oraria)	CO mg/m³ (max oraria)	O_3 µg/m³ (max oraria)	PM_{10} µg/m³ (media su 24h)
Buona	0-50	0-50	0-2,5	0-120	0-25
Accettabile	51-125	51-200	2,6-15	121-180	26-50
Scadente	126-250	201-400	15,1-30	181-240	51-74
Pessima	>250	>400	>30	>240	>74

Il giudizio di qualita' dell'aria, relativo ad ogni stazione, e' attribuito in base al peggiore dei valori rilevati e viene calcolato solamente se e' presente il 75% dei dati. I giudizi di qualita' derivano dai valori limite indicati nel D.M. 60 del 2 aprile 2002 (SO_2, NO_2, CO e PM_{10}) e nel D.Lgs. 183 del 21 maggio 2004 (O_3). Per quanto riguarda l'ozono (O_3), ai fini di questo bollettino, i criteri sono da considerarsi validi a partire dal 13 luglio 2005; per i precedenti valori occorre fare riferimento ai limiti del D.M. 16 maggio 1996. Nel caso in cui si verifichino superamenti della soglia di informazione per l'ozono, ARPAT invia un bollettino specifico alle autorita' locali interessate.
Il presente bollettino contiene l'indicazione delle misure effettuate nel giorno e nelle ore indicate. I valori riportati hanno superato il processo di verifica giornaliero ed hanno validita' sino all'effettuazione di piu' approfonditi controlli, che si avvalgono di strumenti statistici da impiegare su lunghe serie di dati. Per tale motivo non e' tecnicamente corretto utilizzare ed elaborare i valori dei bollettini giornalieri per la costruzione di indicatori di lungo periodo, da confrontare con i valori limite di qualita' dell'aria, considerato che non hanno ancora superato le indispensabili verifiche mensili, trimestrali ed annuali che garantiscono la qualita' finale del dato. (Ulteriori approfondimenti sulla validazione dei dati)

▲inizio pagina versione stampabile

ultima revisione: 22/02/2010

Sintesi dei dati rilevati dalle ore 0 alle ore 24 del giorno domenica 14/02/2010

Stazione		Tipo stazione	SO_2 µg/m³ (media su 24h)	NO_2 µg/m³ (max oraria)	CO mg/m³ (max oraria)	O_3 µg/m³ (max oraria)	PM_{10} µg/m³ (media su 24h)	Giudizio di qualità dell'aria
Lucca	P.za San Micheletto (RETE REGIONALE **)	urbana - traffico	1	75	---	---	56	Scadente
Lucca	V.le Carducci	urbana - traffico	2	---	2	---	75	Pessima
Lucca	Carignano (RETE REGIONALE **)	rurale - fondo	---	---	---	87 (h.18*)	---	Buona
Viareggio	Largo Risorgimento	urbana - traffico	---	---	1,7	---	n.d.	Buona
Viareggio	Via Maroncelli (RETE REGIONALE **)	urbana - fondo	1	121	---	60 (h.17*)	45	Accettabile
Capannori	V. di Piaggia (RETE REGIONALE **)	urbana - fondo	---	79	2	---	59	Scadente
Porcari	V. Carrara (RETE REGIONALE **)	periferica - fondo	2	72	---	82 (h.16*)	63	Scadente

Giudizio di qualità	SO_2 µg/m³ (media su 24h)	NO_2 µg/m³ (max oraria)	CO mg/m³ (max oraria)	O_3 µg/m³ (max oraria)	PM_{10} µg/m³ (media su 24h)
Buona	0-50	0-50	0-2,5	0-120	0-25
Accettabile	51-125	51-200	2,6-15	121-180	26-50
Scadente	126-250	201-400	15,1-30	181-240	51-74
Pessima	>250	>400	>30	>240	>74

"Il giudizio di qualita' dell'aria, relativo ad ogni stazione, e' attribuito in base al peggiore dei valori rilevati e viene calcolato solamente se e' presente il 75% dei dati. I giudizi di qualita' derivano dai valori limite indicati nel D.M. 60 del 2 aprile 2002 (SO2, NO2, CO e PM10) e nel D.Lgs. 183 del 21 maggio 2004 (O3)."

http://www.comune.capannori.lu.it/node/6008

Il sindaco: "L'ordinanza che vieta l'uso dei caminetti nelle case è un dovere amministrativo per tutelare la salute dei cittadini. Non c'erano alternative"

L'ordinanza che vieta l'accensione dei caminetti e delle stufe a legna nelle case che possiedono un impianto di riscaldamento a metano, gpl o gasolio è una decisione che tocca il "cuore" di Capannori, tradizionalmente comune agricolo, che vede nell'immagine del caminetto l'espressione più vera e più bella del focolaio domestico.

Sono consapevole che il caminetto, nell'immaginario collettivo, è il simbolo della casa.

Non è stato semplice, quindi, scegliere di adottare il provvedimento in questione.

Chiedere ai capannoresi di non utilizzare i caminetti se possono scaldare casa con un impianto di riscaldamento alternativo è stato - ci tengo a puntualizzarlo - un dovere a cui, come sindaco, non potevo sottrarmi.

Un dovere nei confronti del mio territorio e dell'intera comunità.

La relazione degli esperti, infatti, non lasciava margine a altre soluzioni.

Cito testualmente quanto rilevato dall'Arpat: "Nel caso della stazione di Capannori è emersa un'elevata presenza di particolato originato dalla combustione di biomasse. Considerate le caratteristiche della Piana Lucchese si deve ritenere che tale frazione del Pm10 sia dovuta essenzialmente a un significativo uso di legna in stufe e caminetti tradizionali, a basse efficienza energetica, che non garantiscono una completa combustione e sono quindi rilevanti sorgenti emissive di varie tipologie di inquinanti, fra cui il Pm10. Un altro possibile contributo di cui è difficile stimare la rilevanza è costituito dalla diffusa abitudine a bruciare nei campi e nei giardini gli scarti vegetali. Il contributo dalla combustione di biomasse alla concentrazione di Pm10 diventa addirittura del 47 per cento nei giorni in cui nella stazione è superato il limite di 50 microgrammi al metro cubo come media giornaliera".

Firmare l'ordinanza che vieta l'uso dei caminetti e delle stufe a legna fino al 31 marzo prossimo significa aver avuto il coraggio di una scelta che, sebbene faccia discutere, mi permette di tutelare la salute dei cittadini.

Ritengo che questo sia il primo compito di un sindaco.

Ancora una volta, quindi, confido nelle buone pratiche dei capannoresi e nell'attenzione con cui guardano alla vita e all'ambiente.

Assieme abbiamo già superato prove difficili, che poi si sono rivelate fonti di grandi soddisfazioni per Capannori.

Anche questa volta, chiedo ai cittadini di fare un gesto - quello di spegnere i caminetti e le stufe a legna - per difendere la qualità dell'aria del nostro territorio.

In questo modo, assieme, difenderemo anche il diritto alla salute di ciascuno di noi.

Il sindaco,
Giorgio Del Ghingaro

Pagina creata il: 3/2/2010

"..."

Uncertainty vs. Vagueness: a clarification

- What does the value (usually in $[0, 1]$) of the degree mean?
- There is often a misunderstanding between interpreting a degree as a measure of uncertainty or as a measure of vagueness !
- The value 0.83 has a different interpretation in “Birds fly to degree 0.83” from that in “Hotel Verdi is close to the train station to degree 0.83”

Uncertainty

- Uncertainty: statements are true or false
  - But, due to lack of knowledge we can only estimate to which probability/possibility/necessity degree they are true or false
- For instance, a bird flies or does not fly
  - we assume that we can clearly define the property “can fly”
- The probability/possibility/necessity degree that it flies is 0.83
- E.g., under probability theory this may mean that 83% of the birds do fly, while 17% of the birds do not fly
  - Note: e.g., a chicken has to be classified as either flying or non-flying thing

Example

- Sport Car:

$$\forall x, hp, sp, ac\ SportCar(x) \iff HP(x, hp) \wedge Speed(x, sp) \wedge Acceleration(x, ac) \wedge hp \geq 210 \wedge sp \geq 220 \wedge ac \leq 7.0$$

		
audi_tt	*mg*	*ferrari_enzo*

- Ferrari Enzo is a Sport Car: $HP = 651, Speed \geq 350, Acc. = 3.14$
- MG is not a Sport Car: $HP = 59, Speed = 170, Acc. = 14.3$
- Is Audi TT 2.0 a Sport Car ? $HP = unknown, Speed = 243, Acc. = 6.9$
- We can estimate from a training set (Naive Bayes Classification)

$$\begin{aligned} Pr(SportCar|AudiTT) &= Pr(AudiTT|SportCar) \cdot Pr(SportCar) \cdot (1/Pr(AudiTT)) \\ &\approx \frac{Pr(speed \leq 243|SportCar) \cdot Pr(accel \geq 6.9|SportCar) \cdot Pr(SportCar)}{Pr(speed \leq 243) \cdot Pr(accel \geq 6.9)} \end{aligned}$$

- Sport Car:

$$\forall x, hp, sp, ac\ SportCar(x) \iff HP(x, hp) \wedge Speed(x, sp) \wedge Acceleration(x, ac) \wedge hp \geq 210 \wedge sp \geq 220 \wedge ac \leq 7.0$$

		
audi_tt	*mg*	*ferrari_enzo*

- Note: Audi TT 2.0 is not a Sport Car: $HP = 200, Speed = 243, Acc. = 6.9$
- Explicit definition of Sport Car is too sharp
- We can estimate from a training set (Naive Bayes Classification)

$$Pr(SportCar|MyCar) = Pr(MyCar|SportCar) \cdot Pr(SportCar) \cdot (1/Pr(MyCar))$$

$$\approx \frac{Pr(MyCar.hp_{\leq}|SportCar) \cdot Pr(MyCar.speed_{\leq}|SportCar) \cdot Pr(MyCar.accel_{\geq}|SportCar) \cdot Pr(SportCar)}{Pr(MyCar.hp_{\leq}) \cdot Pr(MyCar.speed_{\leq}) \cdot Pr(MyCar.accel_{\geq})}$$

Vagueness

- Vagueness: statements involve concepts for which there is no exact definition, such as
  - tall, small, close, far, cheap, expensive, “is about”, “similar to”.
- A statements is true to some degree, which is taken from a truth space (usually $[0, 1]$).
- E.g., “Hotel Verdi is close to the train station to degree 0.83”
  - the degree depends on the distance
- E.g., “The image is about a sun set to degree 0.75”
  - the degree depends on the extracted features and the semantic annotations

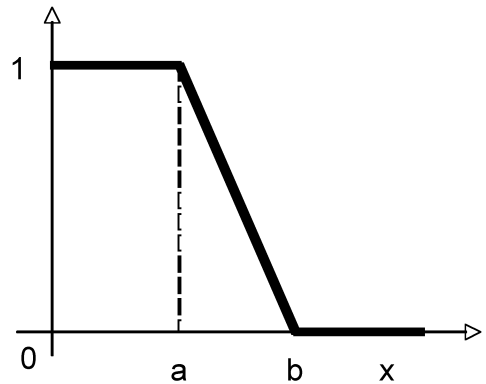
Example

- Sport Car:

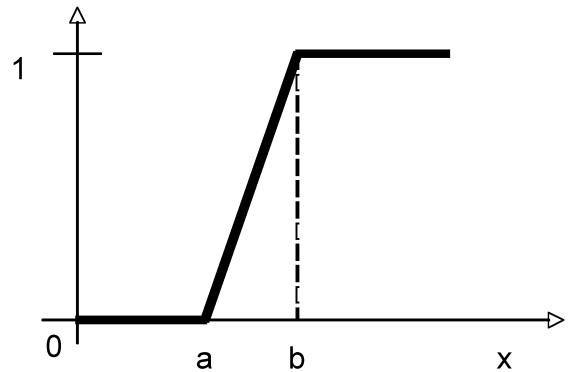
$$\forall x, hp, sp, ac\ SportCar(x) \iff 0.3HP(x, hp) + 0.2Speed(x, sp) + 0.5Accel(x, ac)$$

- Each feature, gives a degree of truth depending on the value and the membership function

$$\begin{aligned} HP(x, hp) &= rs(180, 250)(hp) \\ Speed(x, sp) &= rs(180, 240)(sp) \\ Accel(x, ac) &= ls(6.0, 8.0)(ac) \end{aligned}$$



ls(a,b)

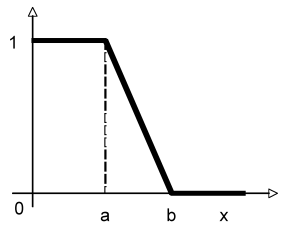


rs(a,b)

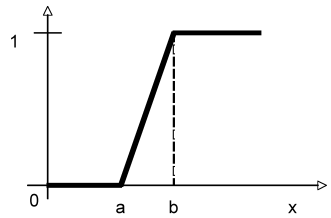
- Degree of truth of $SportCar(AudiTT)$: $0.3 \cdot 0.28 + 0.3 \cdot 1.0 + 0.5 \cdot 0.55 = 0.447$

- The fuzzy membership functions can be learned from a training set (large literature)

$$\begin{aligned} HP(x, hp) &= rs(192, 242)(hp) \\ Speed(x, sp) &= rs(193, 234)(sp) \\ Accel(x, ac) &= ls(6.5, 7.5)(ac) \end{aligned}$$



ls(a,b)



rs(a,b)

- Learned Training Sport Class:

$$\forall x, hp, sp, ac\ TrainingSportCar(x) \iff 0.3HP(x, hp) + 0.2Speed(x, sp) + 0.5Accel(x, ac)$$

- Now, a classification method can be applied: e.g. kNN classifier

$$\forall x, hp, sp, ac\ SportCar(x) \iff \sum_{y \in Top_k(x)} Similar(x, y) \cdot TrainingSportCar(y)$$

$$\begin{aligned} \forall x, hp, sp, ac\ Similar(x, y) \iff\ & 0.3 \cdot HP(x, hpx) \cdot HP(y, hpy) + 0.2 \cdot Speed(x, spx) \cdot Speed(y, spy) + \\ & + 0.5 \cdot Accel(x, acx) \cdot Accel(y, acy) \end{aligned}$$

where $Top_k(x)$ is the set of top-k ranked most similar cars to car x

Imperfect Information

- Mixing uncertainty and vagueness:
  - "Probably it will be hot tomorrow"
    - Crisp quantifier ("probably") over vague statement
  - "In most cases, a bird does fly"
    - Vague quantifier ("most") over crisp statement
- The notion of imperfect information covers concepts such as

uncertainty	"Nancy is likely John's girlfriend"
vagueness	"John's girlfriend is blond"
incompleteness	"John's girlfriend is Nancy or Mary"
imprecision	"The hight of John's girlfriend is in between 165cm and 170cm"
contradiction	"John's girlfriend, Nancy, lives in Rome. Nancy is living in Florence."

Uncertainty vs. Vagueness

- The distinction between uncertainty and vagueness is not always clear: depends on the assumptions
- (Multimedia) Information Retrieval:

Query: "I'm looking for a house"

System Answer:  score/degree 0.83

- What's behind the computational model?

- Probabilistic model
  - Assumption: a multimedia object is either relevant or not relevant to a query q
  - Score: The probability of being a multimedia object o relevant (Rel) to q

$$score := Pr(Rel \mid q, o)$$

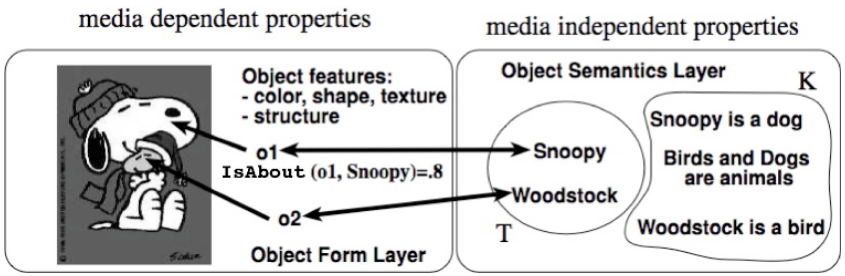
- Vague/Fuzzy model
  - Assumption: a multimedia object o is about a *semantic index term* ($t \in \mathbb{T}$) to some degree in $[0, 1]$
  - The mapping of objects $o \in \mathbb{O}$ to semantic entities $t \in \mathbb{T}$ is called *semantic annotation*

$$F : \mathbb{O} \times \mathbb{T} \rightarrow [0, 1]$$

$F(o, t)$ indicates to which degree the multimedia object o is about the semantic index term t

  - Score: The evaluation of how much the multimedia object o is about the the information need q

$$score := F(o, q)$$



- In other cases there may be both approaches as well
- For instance, in Ontology Alignment, what about the degree n of the mapping

$$\langle SUV, Van, \cap, n \rangle \ ?$$

- Probabilistic model: a car is a SUV (Van) or is not a SUV (Van)
- Then, e.g. from a training set, compute

$$n = Pr(SUV \cap Van)$$

- Fuzzy model: a car is to some degree a SUV and to some other degree a Van
- Then, e.g. from a training set, compute

$$n = kNNSUV(x) \cdot kNNVan(x)$$