The University Of Sheffield.

New version of slides at
http://www.dcs.shef.ac.uk/~fabio/SSMS09.pdf

Information Extraction from Text

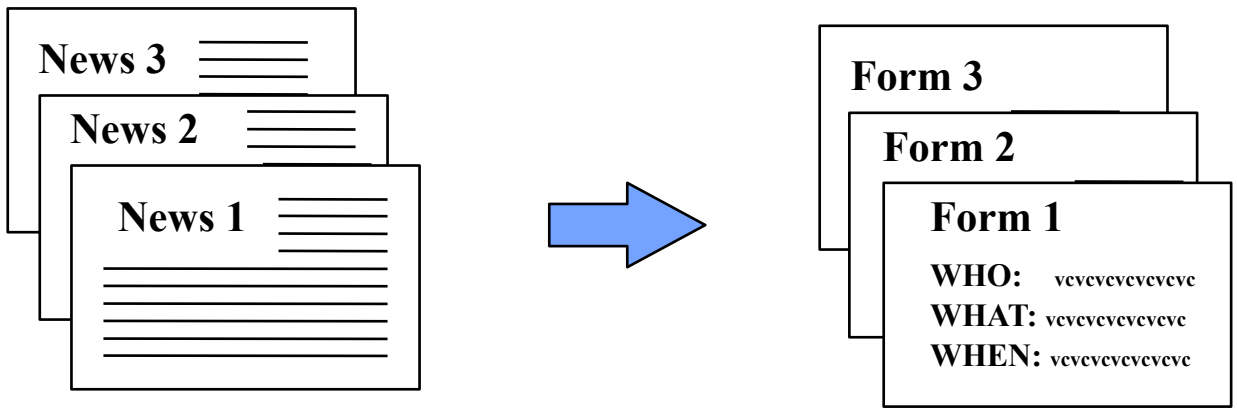
Prof. Fabio Ciravegna
Professor of Language and Knowledge Technologies
Head of Organisations Information and Knowledge Group
Department of Computer Science
The University of Sheffield

fabio@dcs.shef.ac.uk

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Information Extraction



- automatically extracting pre-specified information from natural language texts
  - salient facts about pre-specified types of events, entities or relationships.
- populating a structured information source from a semi-structured, unstructured, or free text, information source.

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Why Texts and IE?

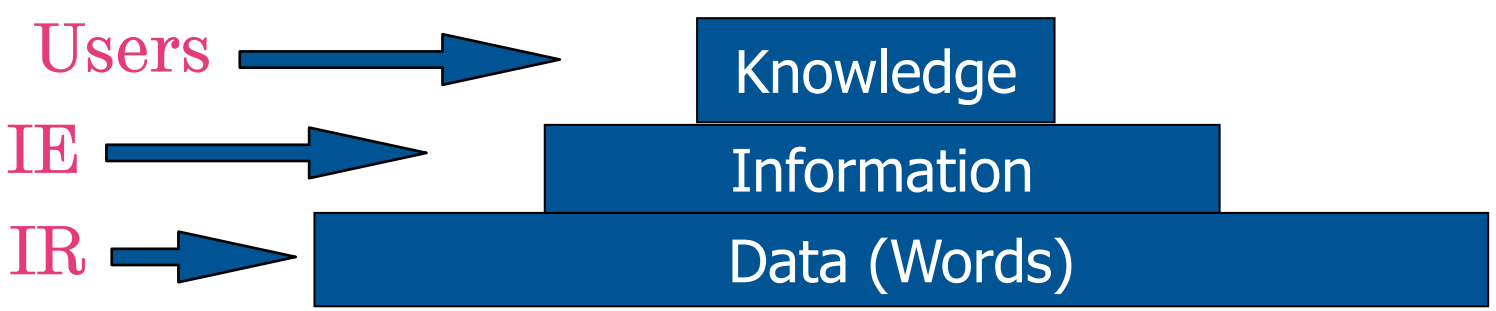
- Textual documents are pervasive (e.g. Web)
  - Contained knowledge cannot be queried
    - Q: How many cases of swine flu have been identified in the UK in the last three months that involve children under 5 years old
    - therefore knowledge cannot be
      - Used by automatic systems
      - Easily managed by humans
- IE can identify information in documents
  - e.g. to populate a database/ontology
  - e.g. to annotate documents
- Method: some forms of language analysis

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

IE Vs Information Retrieval

	IR	IE
Task	Data Indexing	Information Extraction
Returns upon User Query	Relevant Documents	Relevant Information
Query Generality	Full	Limited to target information



IE tasks

WASHINGTON, D.C. (October 5, 1999) - nQuest Inc. today announced that Paul Jacobs. former Vice-President of E-Commerce at SRA International has joined the company's executive management team as president.

Named Entities — Relation Extraction — Event Extraction

IE tasks

WASHINGTON, D.C. (October 5, 1999) - nQuest Inc. today announced that Paul Jacobs. Vice-President of E-Commerce at SRA Internati has joined the company's executive management as president.

Company: nQuest Inc.
Date: today
InPerson: Paul Jacobs
InRole: president

Company: SRA International
OutPerson: Paul Jacobs
OutRole: Vice-President of E-Commerce,

Named Entities → Relation Extraction → Event Extraction

The University Of Sheffield.

The Basics of IE

Most Information for this section derives from
Sunita Sarawagi: Information Extraction,
Foundations and Trends in Databases, Vol. 1, No. 3 (2007) 261–377

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Taxonomical Classification

- The type of input documents
- The type of input resources available for seeding extraction
- The method used for extraction
- The output of extraction
- The IE tasks

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

What Input?

- Template-based documents
  - e.g. forms, database-backed Web pages, etc.
- Semi-Structured documents
  - that follow some pre-defined styles
    - Resumés
    - Seminar announcements
    - Newspaper articles
- Open ended sources
  - No pre-defined style identifiable
    - e.g. Web pages

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

- Semi-structured documents present strong regularity in
  - Formatting
  - Language
  - Lexicon
- Requirement for IE is to be able to model Regularity in a seamless way

Path: itc.it!itc.it!not-for-mail
From: Carola Dori <dori@itc.it>
Newsgroups: itc.seminari
Subject: seminario V. RASKIN, 24 novembre 1997
Date: 12 Nov 1997 14:17:27 +0100
Organization: Istituto Trentino di Cultura - IRST
Lines: 51
Sender: news@itc.it
Distribution: local
Message-ID: <64ca97$ebf@wonder.itc.it>
NNTP-Posting-Host: wonder.itc.it

AVVISO DI SEMINARIO
="="="============================""

Lunedi' 24 novembre 1997
ore 14:30

Sala seminari edificio Est
ITC-IRST, Povo

="="="============================""

"SOME CHOICES IN COMPUTATIONAL LEXICAL SEMANTICS"

Victor Raskin
Professor and Chair, Linguistics
Purdue University
W. Lafayette, IN 47907-1356 USA

Abstract:

The modern computational lexical semantics reached a point in its development when it has become necessary to define the premises and

Example:
<p> Capitol Hill- *1* br twnhme. D/W W/D. Pkg incl $*675*. *3* BR upper flr no gar. $*995*. (206)999-9999

The University Of Sheffield.

What resources?

- Input to an IE system:
  - Documents + Task (e.g. ontology)
- Additional sources of information:
  - Labelled documents (most usual case, useful also for testing)
  - Gazetteers (i.e. list of terms - e.g. proper names, etc.)
  - Structured databases of potential answers
    - Presenting **labelled** structured information equivalent to that that should be extracted from text
      - but not the **same** data
  - Unlabelled structures
    - e.g. tables from the Web that are likely to contain relevant information but the information is not labelled

© Fabio Ciravegna, University of Sheffield

Rest of the Rest

- Dimensions for IE Methods
  - Hand coded Vs Learning
  - Rule-based or Statistics
- Types of output
  - Entities
    - All mentions in document Vs just one mention per document
  - Relations
    - Within sentence or across sentences

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

IE Tasks

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Entity Extraction

- Entities are typically noun phrases and comprise of one to a few tokens in the unstructured text
- Covers
  - Entity Spotting,
  - Classification
  - Unique Identification

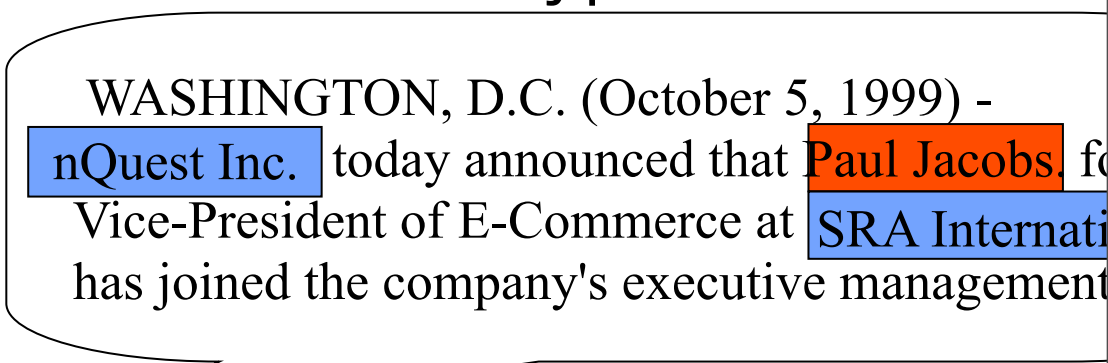
- Spotting and Classification a generally conflated into one single task

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Classification

- Can include up to some hundreds of types
  - e.g. ACE competition
- Examples:
  - Named Entity Recognition:
    - Classic tasks (e.g. MUC conferences)
      - Includes Named Entities, Time Expressions and Numerical Expression
  - Terminology recognition
    - Recognition of technical terminology in specialistic documents
      - E.g. names of genes, parts of an aircraft, etc.



© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Unique Identification

- Requires matching the identified and classified name against a previous list and decide if
  - it is an individual already known
  - it is a new individual
- In Semantic Web this is equivalent to generate a URI for the name

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Rule-Based Methods for Entity Extraction

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Why Rules?

- Many real-life extraction tasks can be conveniently handled through a collection of rules, which are either hand-coded or learnt from examples
- A typical rule-based system consists of:
  - a collection of rules
  - a set of policies to control the firings of multiple rules

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Basic rules

- Rules tend to have the form
  - *Contextual Pattern -> Action*
- E.g. Finite State Transducer Rules

```
Rule: Company1                                          from gate.ac.uk
    ( ( {Token.orthography == upperInitial} )+
        {Lookup.kind == companyDesignator}
    ):match
 -->
    :match.NamedEntity = { kind=company, rule="Company1" }
```

© Fabio Ciravegna, University of Sheffield

Token Features

- The String
- Orthography type
- Part of Speech
- Gazetteer information
- Any other information provided by any type of preprocessing

Word	Lemma	PoS	case	Gaze
the	the	Art	low	
seminar	Seminar	Noun	low	
at	at	Prep	low	
4	4	Digit	low	
pm	pm	Other	low	timeid
will	will	Verb	low	

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

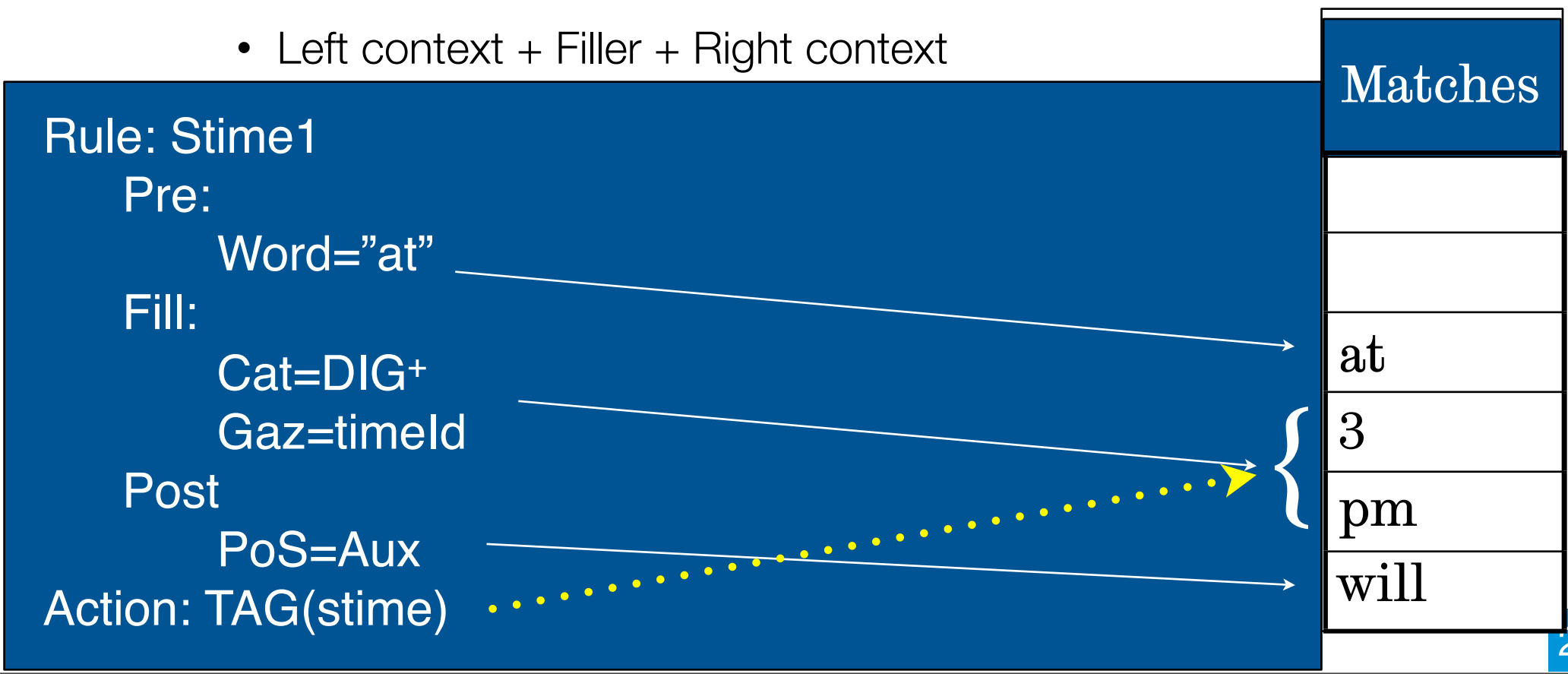
Types of Entity Rules

- Identifying an entity requires recognition of a portion of the document and to insert an XML tag
  - SGML tags in the old days
- Three approaches tried in literature
  - Whole entity recognition
    - E.g. Annie (Cunnigham 2001), Rapier (Califf 1999), etc.
  - Boundary recognition
    - E.g. $(LP)^2$ (Ciravegna 2001), BWI (Kushmerick 2001)
  - Multiple entity recognition
    - E.g. Whisk (Soderland 1999)

© Fabio Ciravegna, University of Sheffield

Rules to Identify Entities

- The classic approach uses rules that model a whole entity
  - No dependency among entities
  - Rule models
    - Left context + Filler + Right context



© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Rules to identify boundaries

- Rules model
  - Left context + Right context of each tag
- Different rules to identify <entity> and </entity>
  - <entity> recognised independently from </entity>

```
Rule: Stime1
      Pre:
            Word=”seminar”:
            Word=”at”:
      Post:
            Cat=DIG+
            Gaz=timeId

Action: TAG(<stime>)
```

Matches
The
seminar
at
3
pm

The University Of Sheffield.

Multiple Entities Rules

- Identify more than entity
  - Model the dependency that sometimes exist between entities
    - especially order in very structured pages

Example:

<p> Capitol Hill- ***1*** br twnhme. D/W W/D. Pkg incl $***675***. ***3*** BR upper flr no gar. $***995***. (206)999-9999

Rule:

ID:7
Pattern: * ('Capitol Hill') * (*Digit*) * '$' (*Number*)
Output: Rental {Neighborhood $1} {Bedrooms $2} {Price $3}

Rule from: STEPHEN SODERLAND:
Learning Information Extraction Rules for Semi-structured and Free Text,
Machine Learning 1, 44()

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Discussion

- Multi-entity rules are useful only if text very structured
- Single entity rules are more natural for manual writing
  - People are very good at generalising
- Boundary rules are better for learning
  - Require less examples to generalise

```
("at"|"starting from")+ <TIME> DIGIT +("PM"|"AM") </TIME>
```

To learn Single Entity Rules 4 examples are required:

- "at"+ <TIME> DIGIT + PM </TIME>"
- "at"+ <TIME> DIGIT + AM </TIME>"
- "starting from"+ <TIME> DIGIT + AM </TIME>"
- "starting from"+ <TIME> DIGIT + PM </TIME>"

To learn Boundary rules 2 examples are needed:

- "at"+ <TIME> DIGIT +PM </TIME>"
- "starting from"+ <TIME> DIGIT +AM </TIME>"

  - Worst case scenario is obviously multiple entity recognition

Fabio Ciravegna: Adaptive Information Extraction from Text by Rule Induction and Generalisation in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001) , Seattle, August 2001.

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Organising Rule Collections

- When rules are fired
  - More than one can apply for a specific span of text
  - Which rule is to be applied?
- Strategies
  - Unordered rules with ad-hoc strategies
    - E.g. Prefer rules marking larger span of text (longer entities)
      - E.g. <ORG> IBM Corp. </ORG> preferred to <ORG> IBM </ORG>
  - Ordered set of rules
    - E.g. rules are sorted by precision on the training corpus

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Rule Learning Algorithms

- Given an annotated corpus
  - Derive a minimal set of rules that cover all (and only) the annotated examples
    - Or at least to maximise recall and precision
  - As determining the optimal rule set is intractable
    - Existing algorithms follow a greedy hill climbing strategy
      - Learn one rule at a time i.e.:

(1) Rset = set of rules, initially empty.
(2) While there exists an entity $\mathbf{x} \in D$ not covered by any rule in Rset
    (a) Form new rules around $\mathbf{x}$.
    (b) Add new rules to Rset.
(3) Post process rules to prune away redundant rules.

Sunita Sarawagi: Information Extraction, Foundations and Trends in Databases, Vol. 1, No. 3 (2007) 261–377

© Fabio Ciravegna, University of Sheffield

Bottom-Up Rule Formation

- For each annotated example
  - Create 1 rule by selecting a window of words to the left and right of entity/tag
    - Completely overfitting the example
    - Likely 100% precision, very low recall
      - Will cover just the current example (plus all the repetitions)
  - Drop constraints on words in window
  - Identify best rule (set) covering example
  - Remove all other instances covered by rules
    - Covering algorithm

© Fabio Ciravegna, University of Sheffield

Example

the seminar at <time> 4 pm will

Condition	Additional Knowledge				Action
Word	Lemma	LexCat	case	SemCat	Tag
the	the	Art	low		
seminar	Seminar	Noun	low		
at	at	Prep	low		stime
4	4	Digit	low		
pm	pm	Other	low	timeid	
will	will	Verb	low		

Fabio Ciravegna:
Adaptive Information Extraction from Text by Rule Induction and Generalisation
in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001) , Seattle, August 2001.

Example

the seminar at <time> 4 pm will

Condition	Additional Knowledge				Action
Word	Lemma	LexCat	case	SemCat	Tag
at	at	Prep	low		stime
4	4	Digit	low		
pm	pm	Other	low	timeid	

Fabio Ciravegna:
Adaptive Information Extraction from Text by Rule Induction and Generalisation
in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001) , Seattle, August 2001.

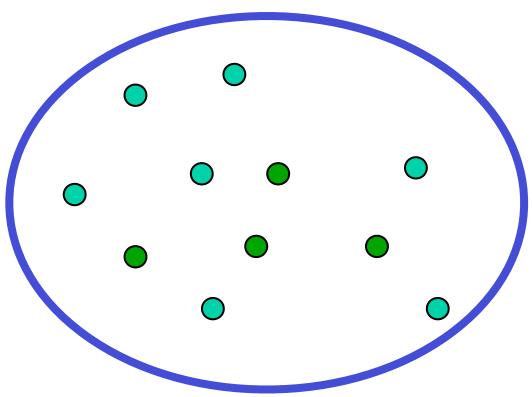
Example

the seminar at <time> 4 pm will

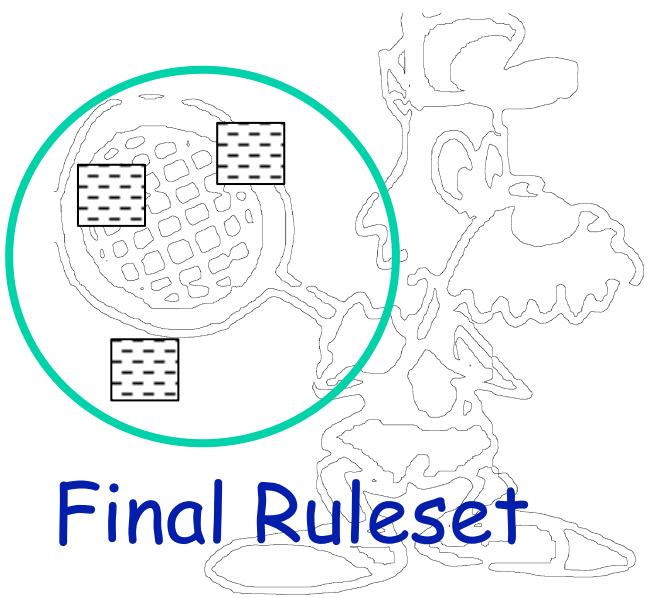
Condition	Additional Knowledge				Action
Word	Lemma	LexCat	case	SemCat	Tag
	at				stime
		Digit			
				timeid	

Fabio Ciravegna:
Adaptive Information Extraction from Text by Rule Induction and Generalisation
in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001) , Seattle, August 2001.

Rule Induction & Generalization



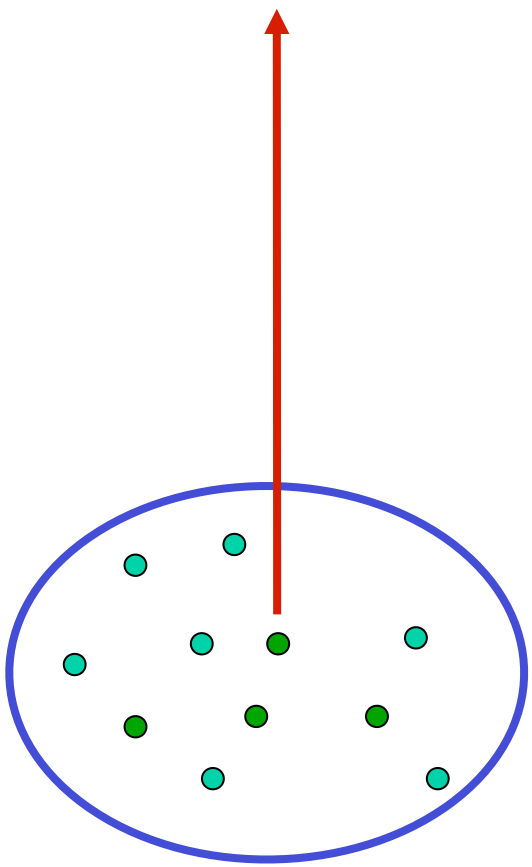
Positive Examples



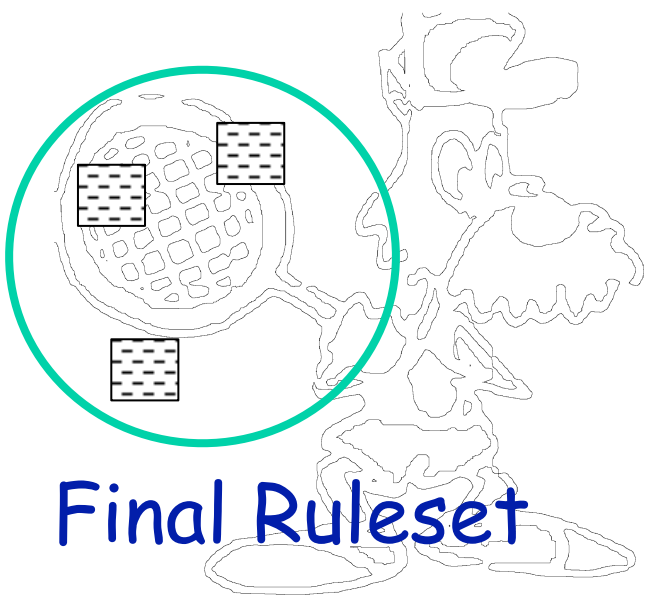
Final Ruleset

Rule Induction & Generalization

Rule 



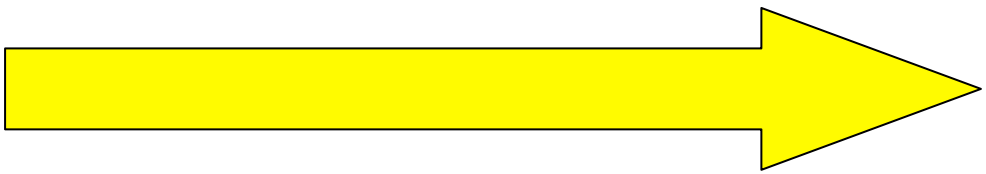
Positive Examples

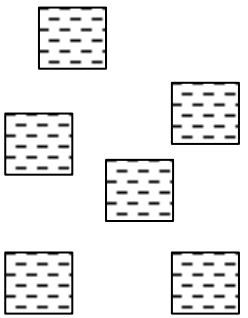


Final Ruleset

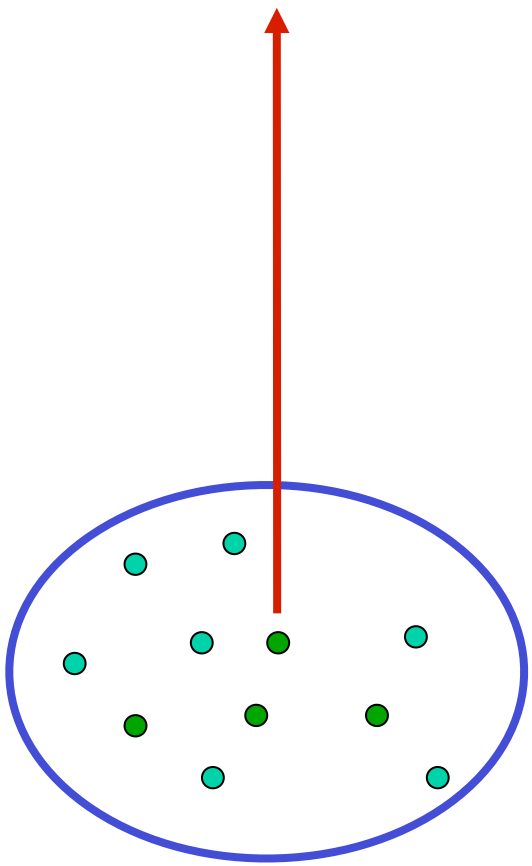
Rule Induction & Generalization

Rule

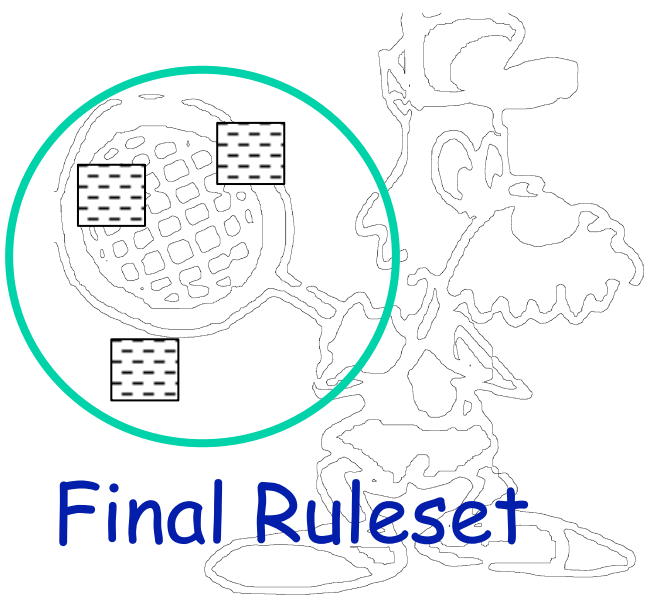




Generalizations



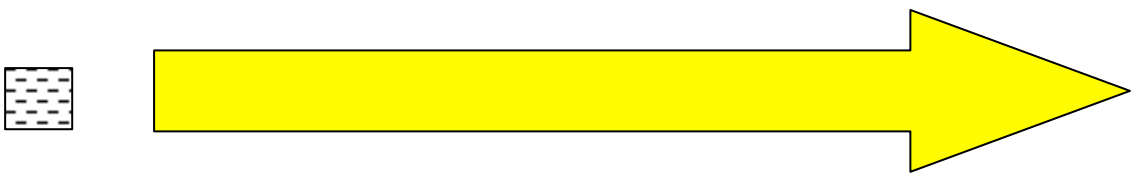
Positive Examples

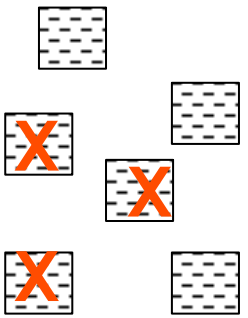


Final Ruleset

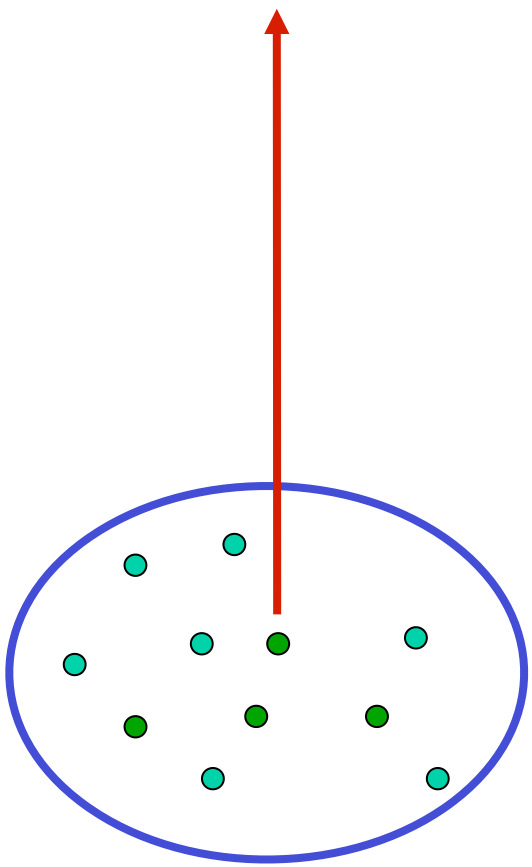
Rule Induction & Generalization

Rule

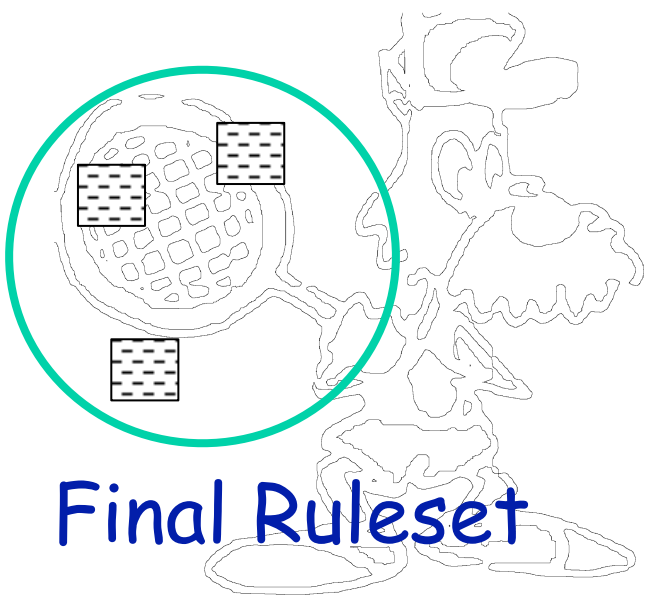




Generalizations



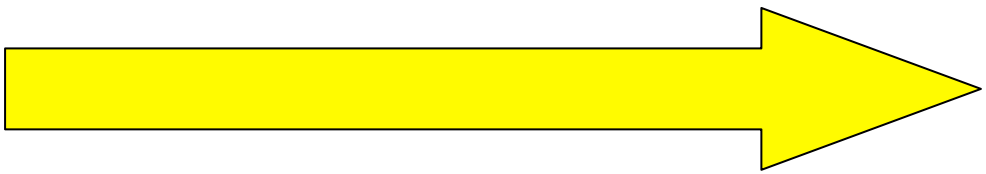
Positive Examples

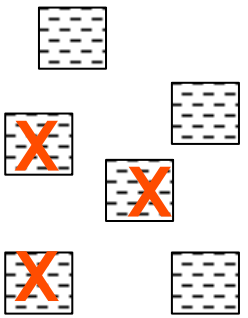


Final Ruleset

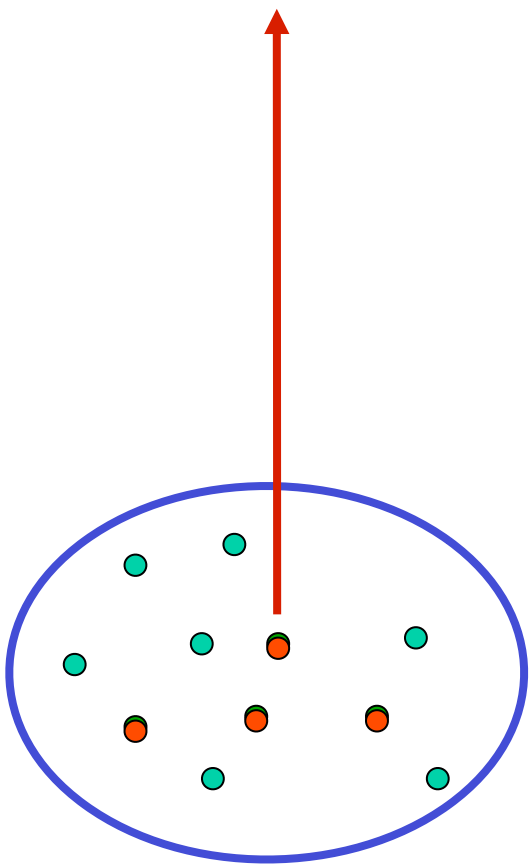
Rule Induction & Generalization

Rule

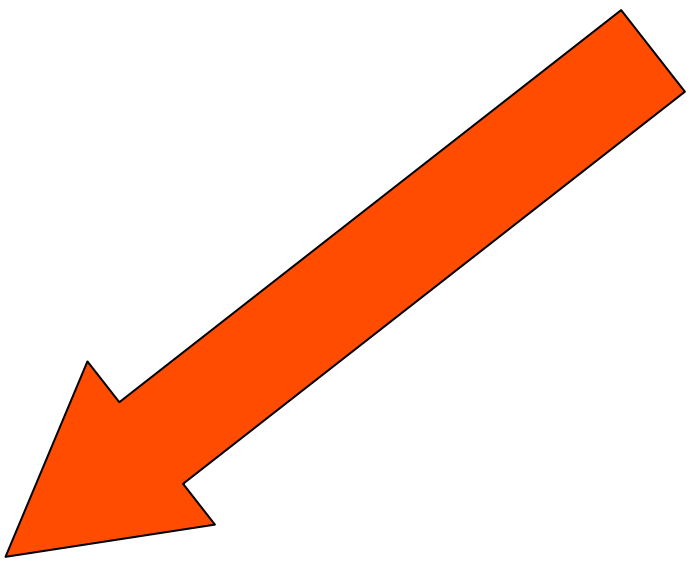


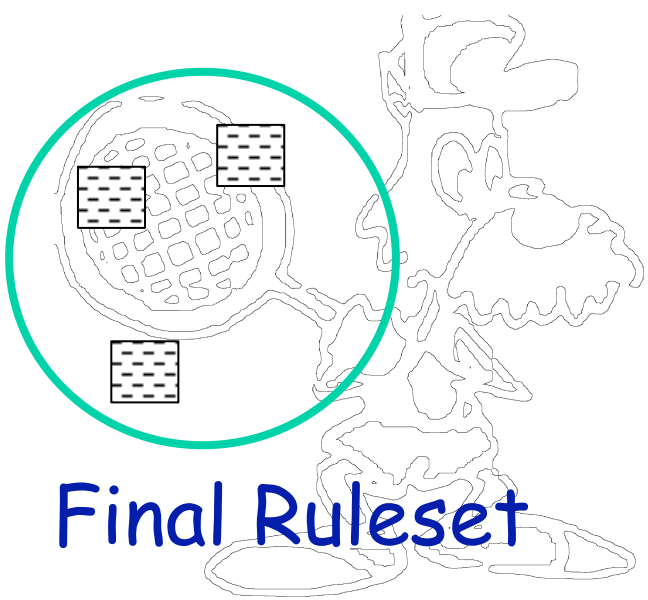


Generalizations



Positive Examples

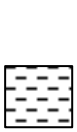


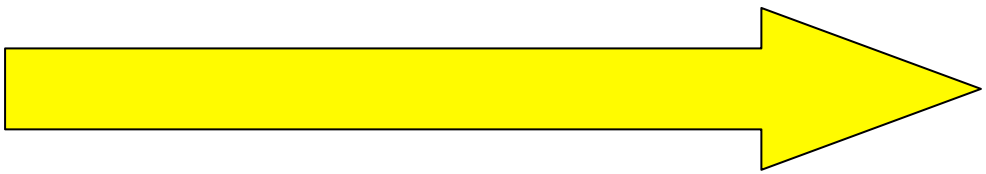


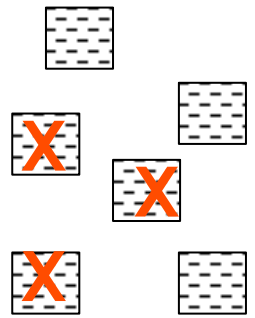
Final Ruleset

Rule Induction & Generalization

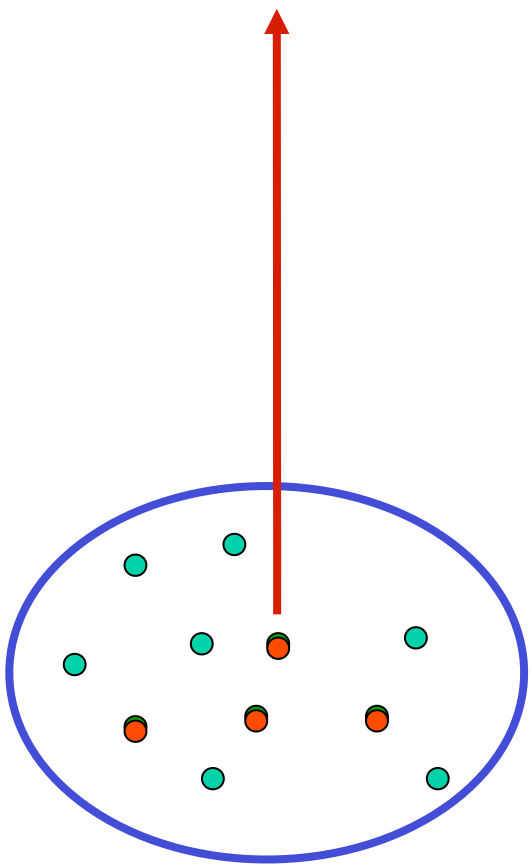
Rule



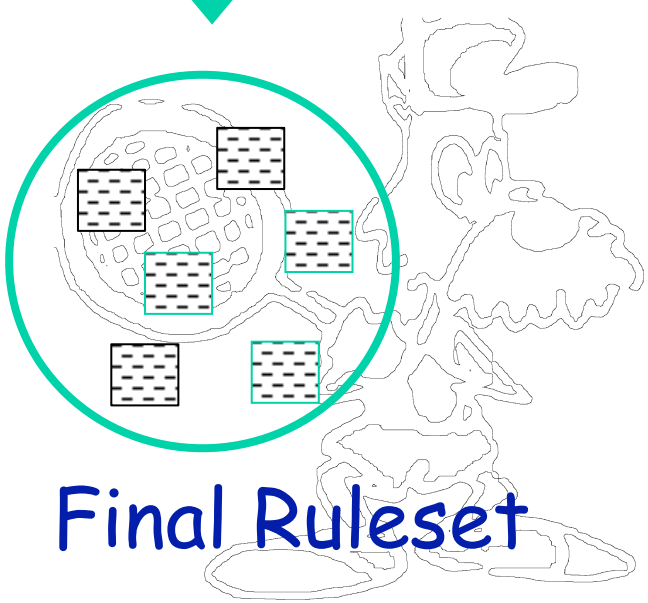




Generalizations

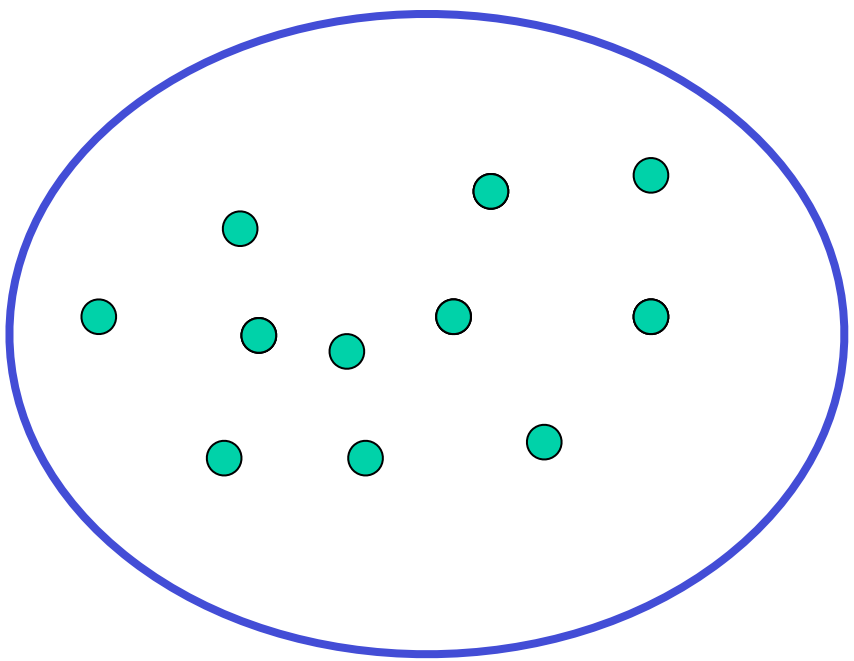


Positive Examples

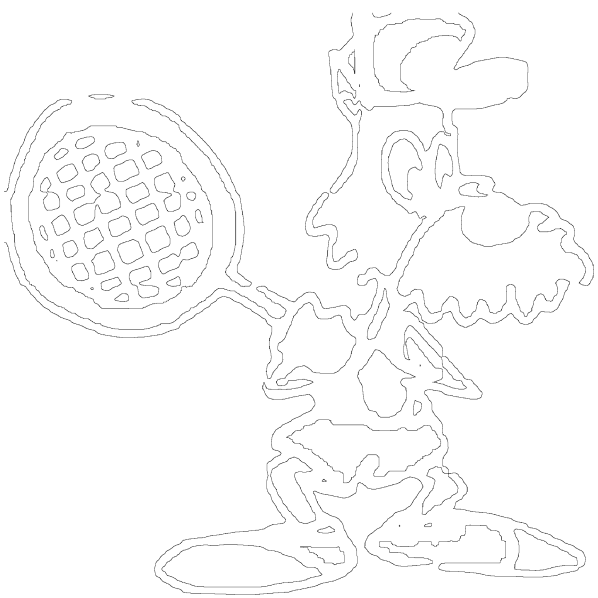


Final Ruleset

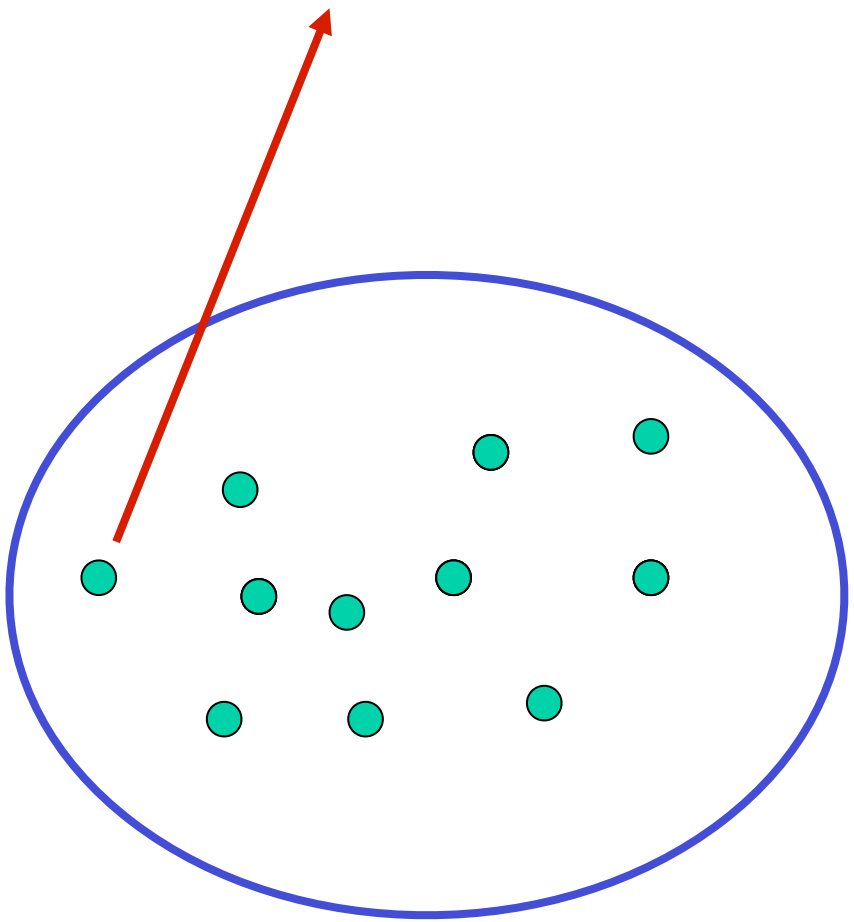
Covering Algorithm



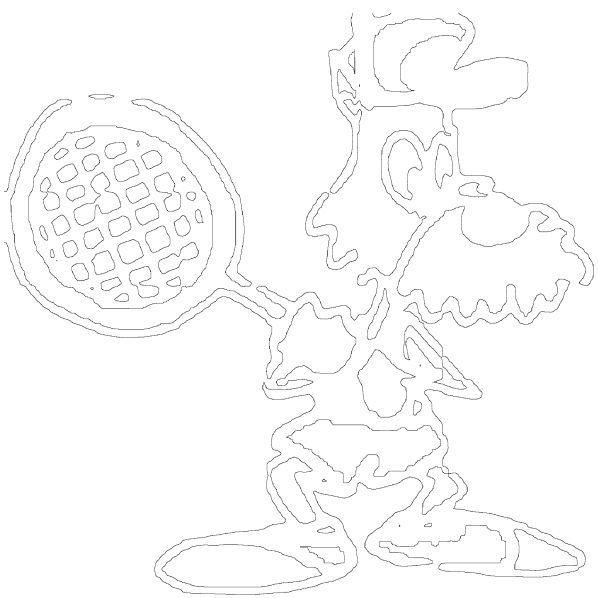
Positive Examples



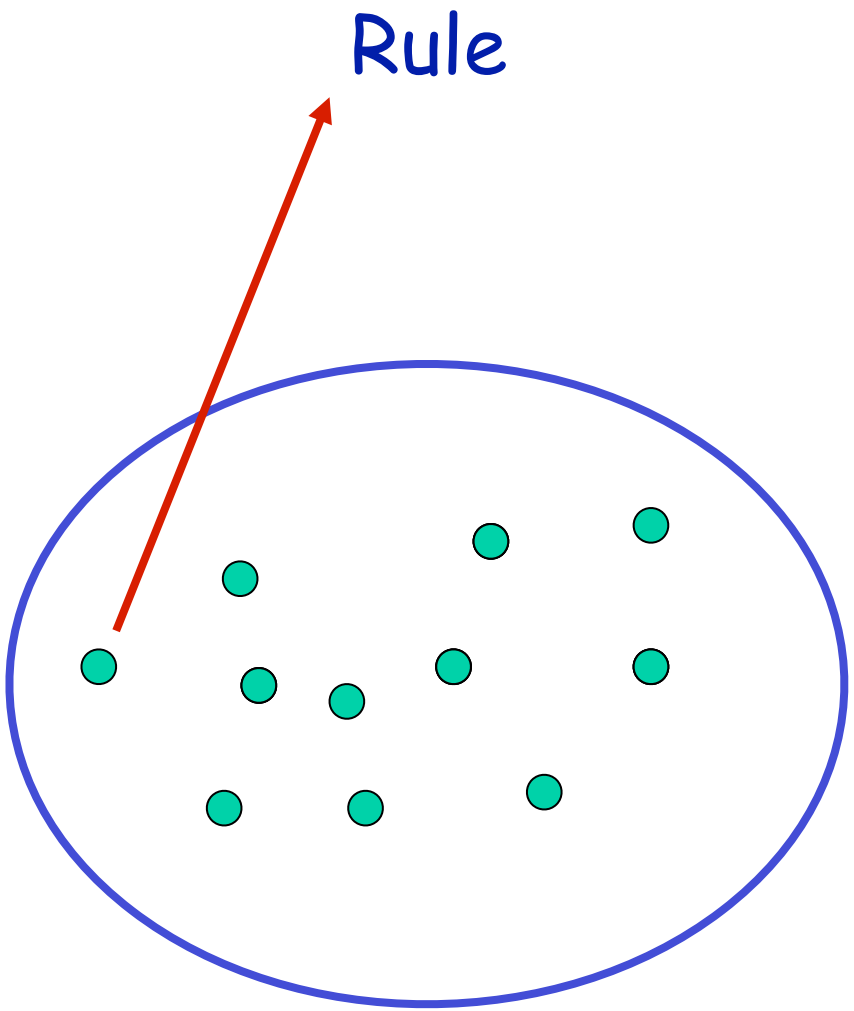
Covering Algorithm

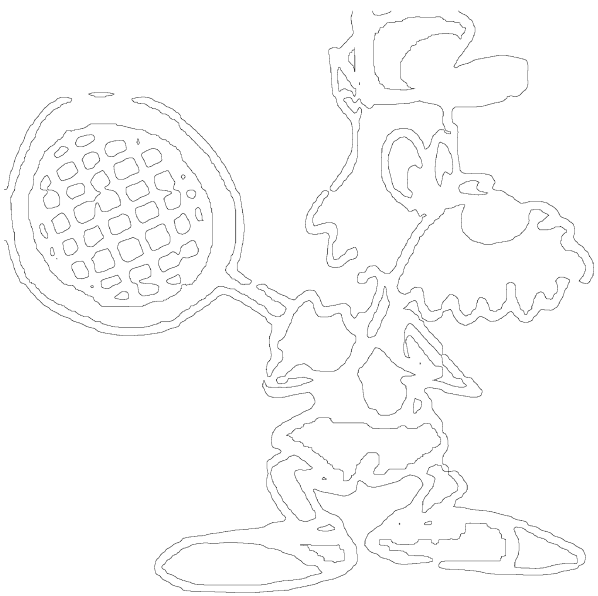


Positive Examples

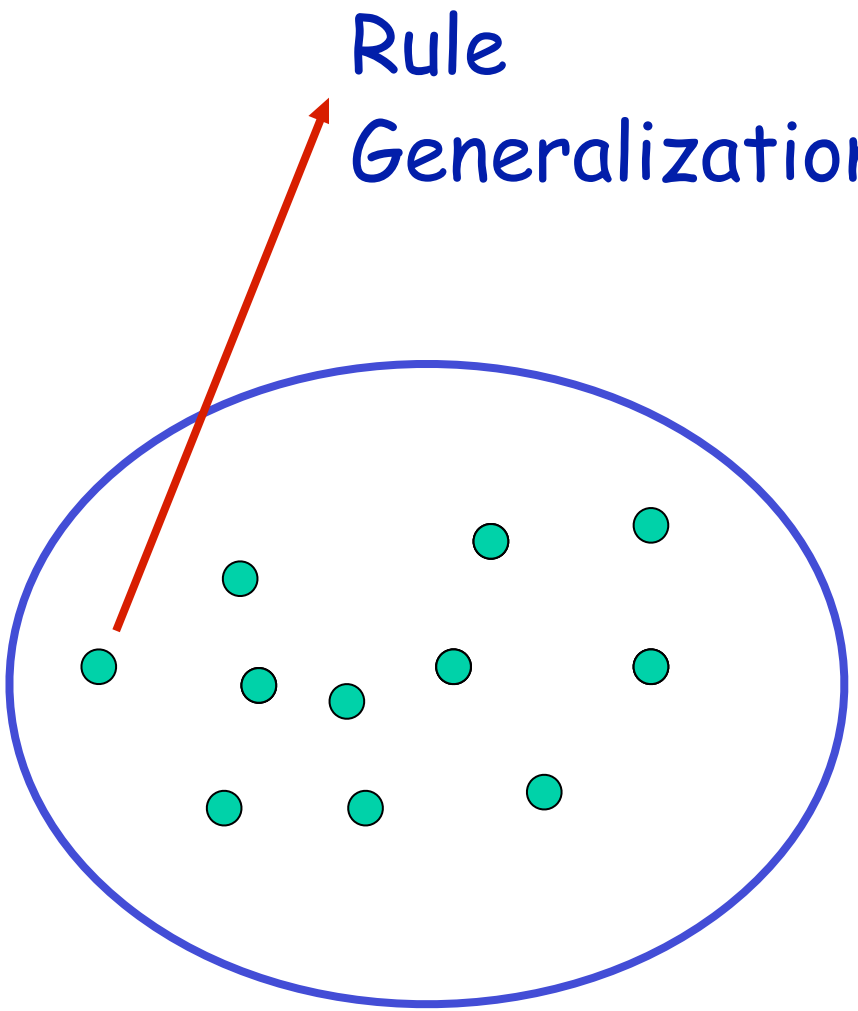


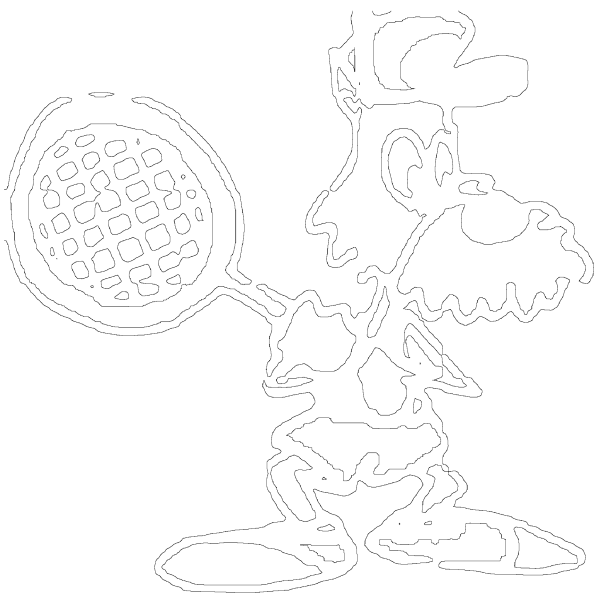
Covering Algorithm



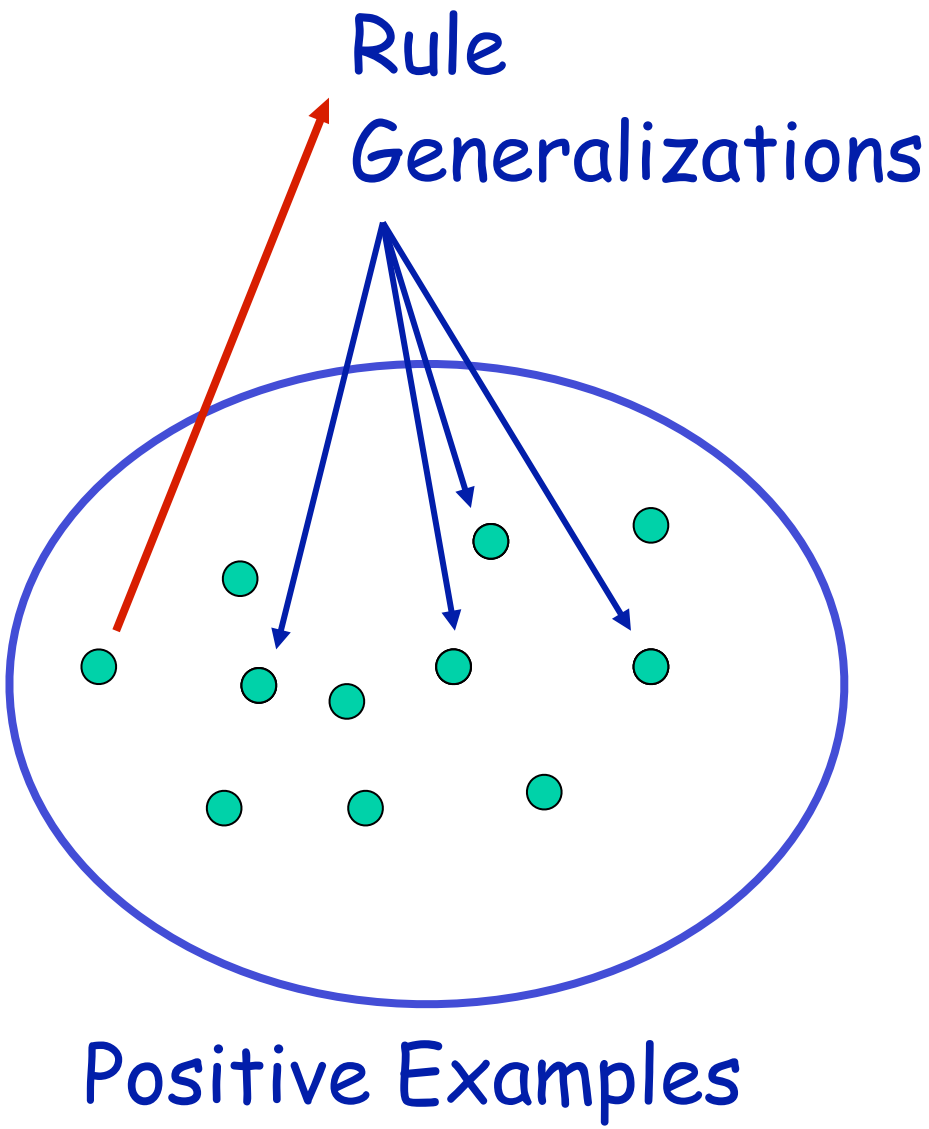


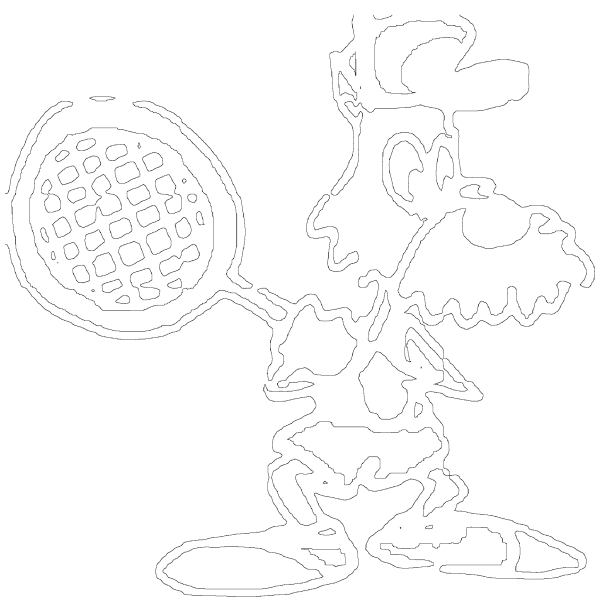
Covering Algorithm



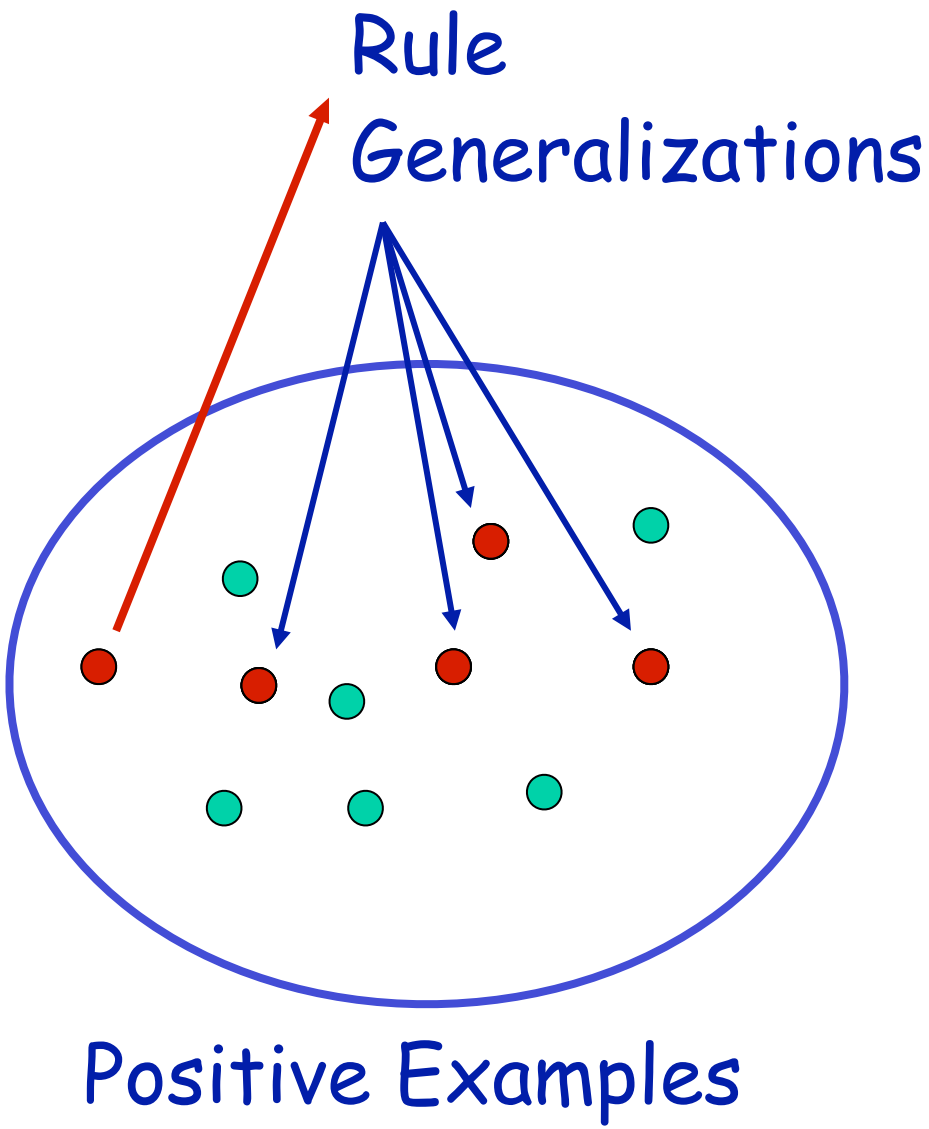


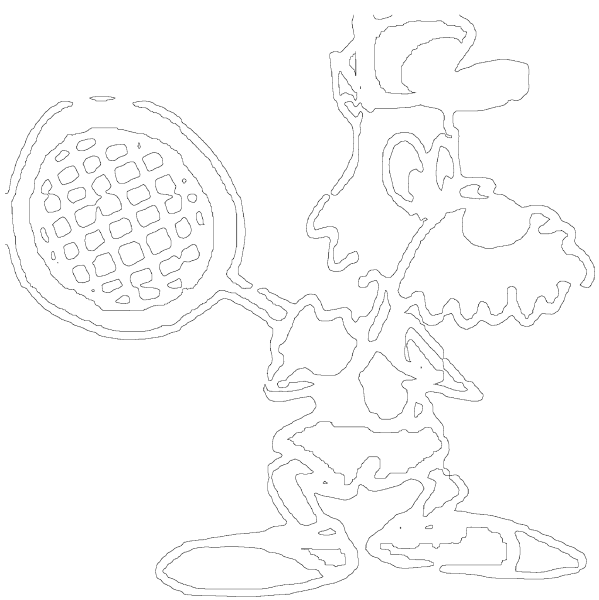
Covering Algorithm





Covering Algorithm





The University Of Sheffield.

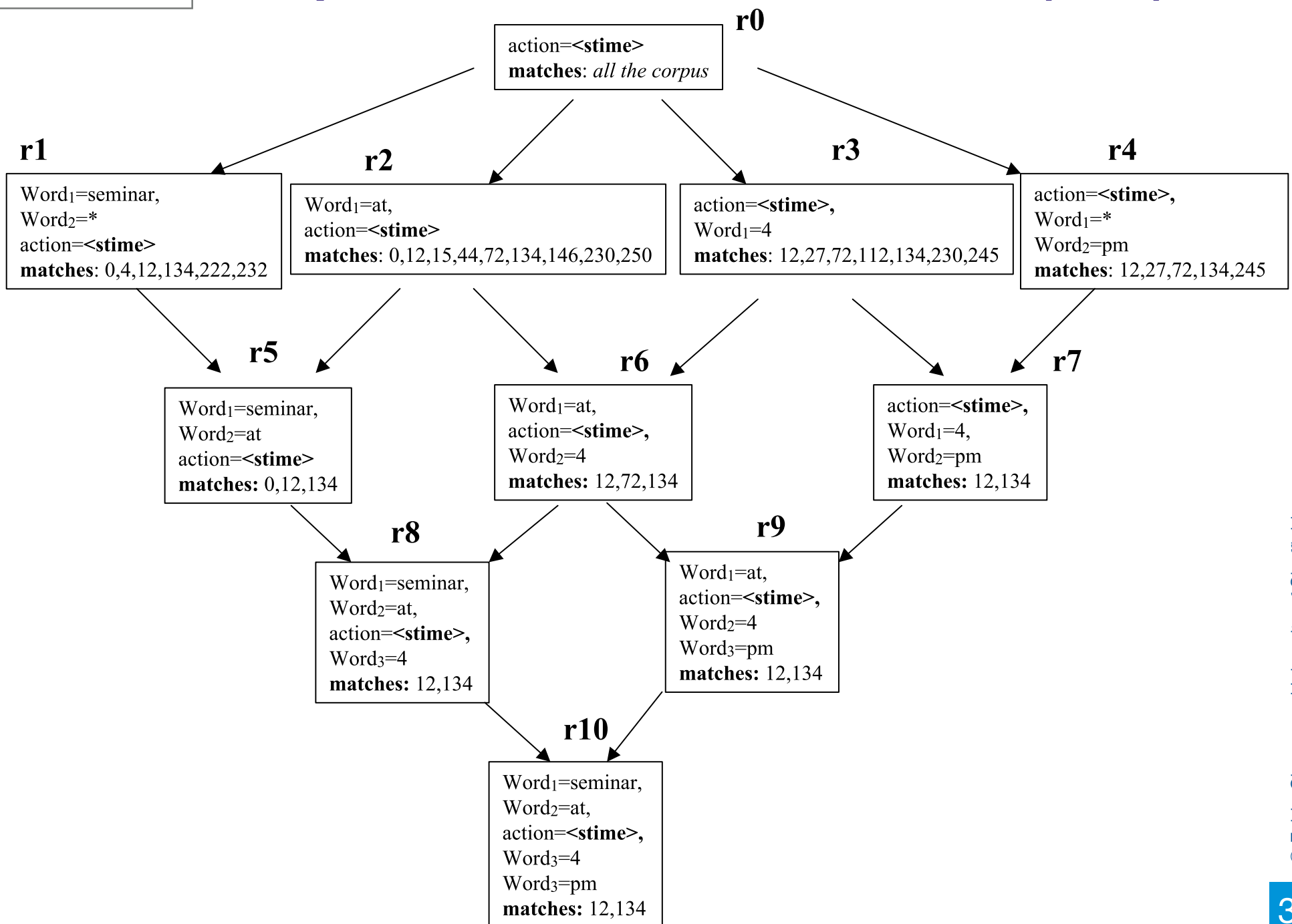
Top Down Algorithm

- Starts from an empty rule
  - will match the whole corpus
  - 100% recall, low precision
- Progressively insert constraints on words to raise precision while keeping recall
  - Stop when rules are overfitting examples

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Top-down version of $(LP)^2$



© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Statistical Methods for Entity Recognition

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Token Level Models

- Text is treated as a sequence of tokens
  - extraction problem is to assign an entity label to each token

The	seminar	will	start	at	3	.	30	pm	this	afternoon
null	null	null	null	null	stime	stime	stime	stime	null	null

OR

The	seminar	will	start	at	3	.	30	pm	this	afternoon
null	null	null	null	null	B start	C stime	C stime	E stime	null	null

B=Begins C=Continues E=Ends

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Algorithms for Token Models

- Token labels depends on
  - Token Features
  - Features of previous tokens
- Most popular algorithms:
  - Conditional Random Fields
    - Considers: features of current token and the token immediately preceding token (window=1)

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Entity Level

- Consider the input as a sequence of entities
  - Will try and tag group of words as entities

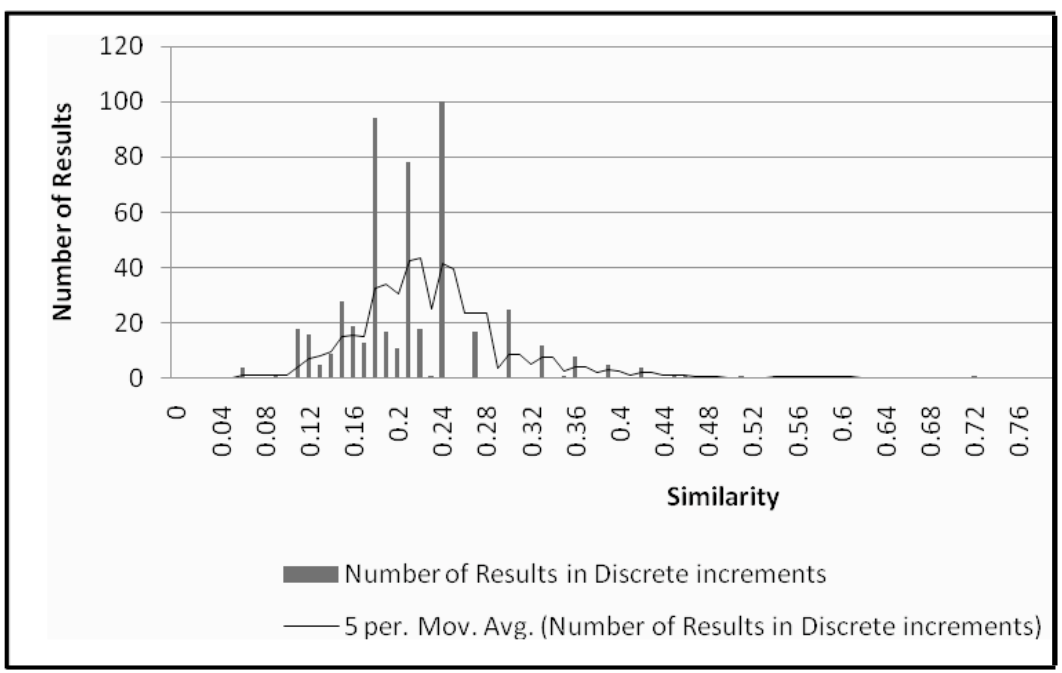
The	seminar	will	start	at	3	.	30	pm	this	afternoon
e1=null					e2=stime				e3=null	

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Similarity-based Algorithms

- Measure distance of group of words to a gazetteer list
  - It is possible to show how (especially long) terms can be discriminated from the noise
  - Very popular in terminology recognition



Jonathan Butters and Fabio Ciravegna: Using String Distance Metrics For Terminology Recognition, in Proceedings of the sixth international conference on Language Resources and Evaluation, LREC 2008, Marrakech, May 2008

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Algorithms

- The models output a y for which the score

  $$s(y)=w\ f(x,y)$$

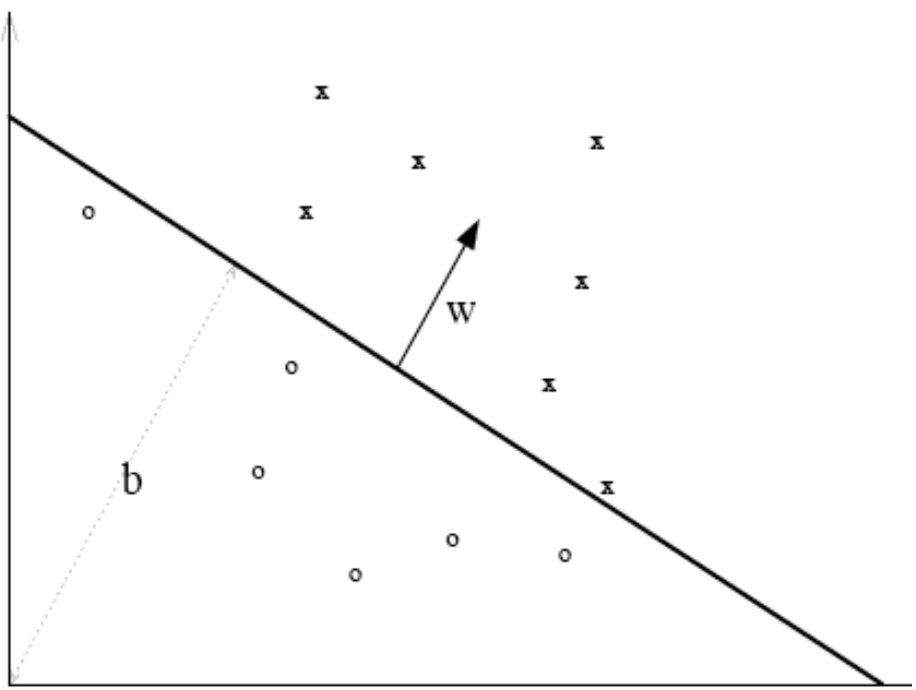
  is maximum
  - f(x,y) is a feature vector defined over input *x* and output *y*
  - For token models ***y*** is a sequence of labels for x
  - For segment models, y is a segmentation of the input x

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Perceptrons

- Decision function is a hyperplane in input space
  - The Perceptron Algorithm (Rosenblatt, 57)



A slide from Nello Cristianini: http://www.support-vector.net/icml-tutorial.pdf

© Fabio Ciravegna, University of Sheffield

Support Vector Machines

- An SVM is a way to train a standard perceptron
- They
  - Selects special points (support vectors) to determine the hyperplane
    - they learn how to weight the features by solving a big optimization problem

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

SWM-based classification

- Given some training data, a set of points of the form

$$\mathcal{D} = \{(\mathbf{x}_i, c_i) | \mathbf{x}_i \in \mathbb{R}^p, c_i \in \{-1, 1\}\}_{i=1}^n$$

  - x_i is a p-dimensional vector of real numbers

- Where c_i is either +1 or -1 (i.e. it is the class it belongs to)
- We want to find the maximum-margin hyperplane which divides the points having $c_i = 1$ from those having $c_i = -1$

$$\mathbf{w} \cdot \mathbf{x} - b = 0$$

- The vector w is a normal vector: it is perpendicular to the hyperplane

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

SVN (ctd)

- We want to choose the w and b to maximize the margin, or distance between the parallel hyperplanes that are as far apart as possible while still separating the data.
- These hyperplanes can be described by the equations

$$\mathbf{w} \cdot \mathbf{x} - b = 1$$

and

$$\mathbf{w} \cdot \mathbf{x} - b = -1.$$

- To cut a long story short, the classification problem can be rewritten as

Minimize (in $\mathbf{w}, b$)

$$\frac{1}{2}\|\mathbf{w}\|^2$$

subject to (for any $i = 1, \ldots, n$)

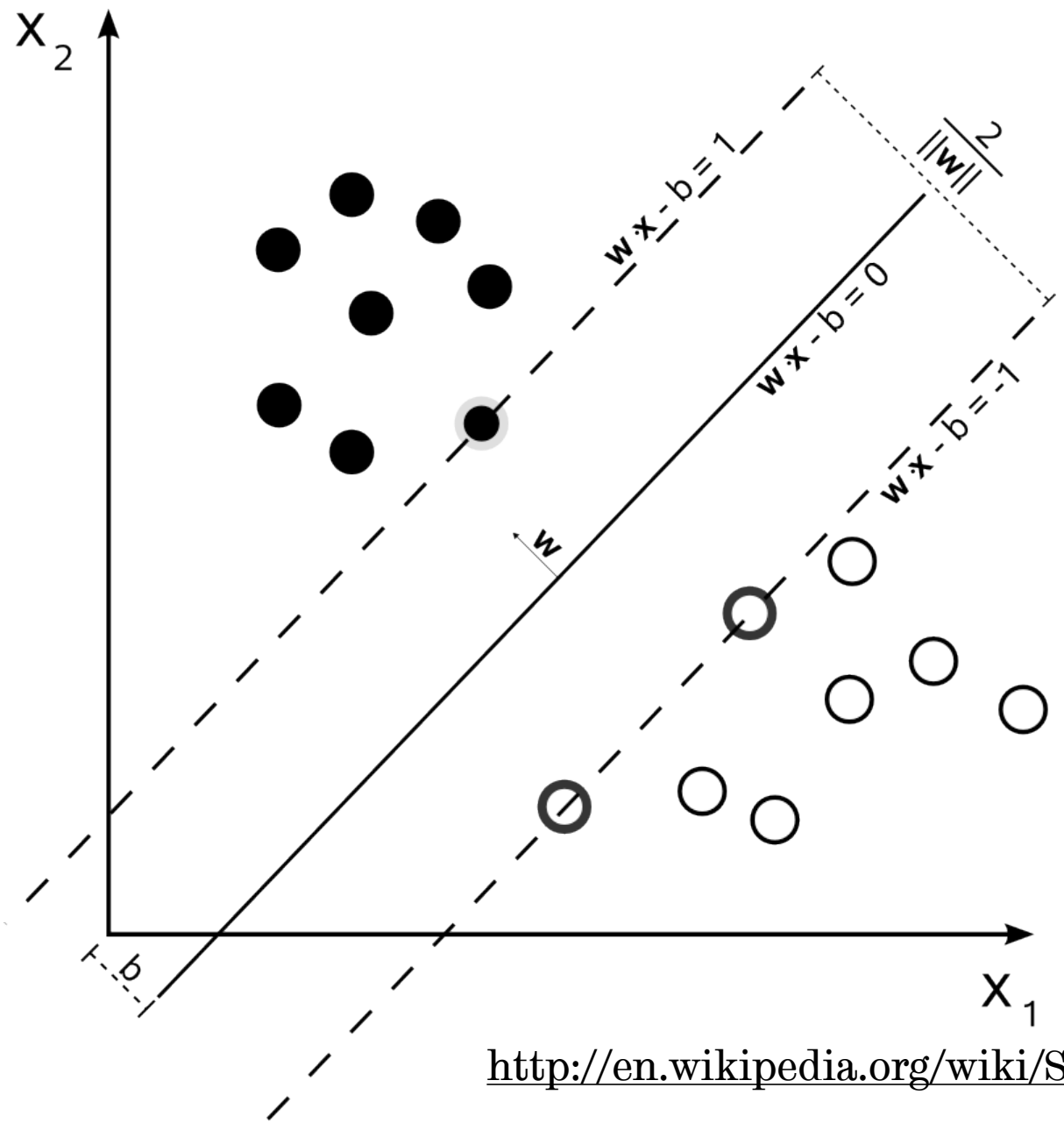
$$c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1.$$

- Where ci={-1,1}

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

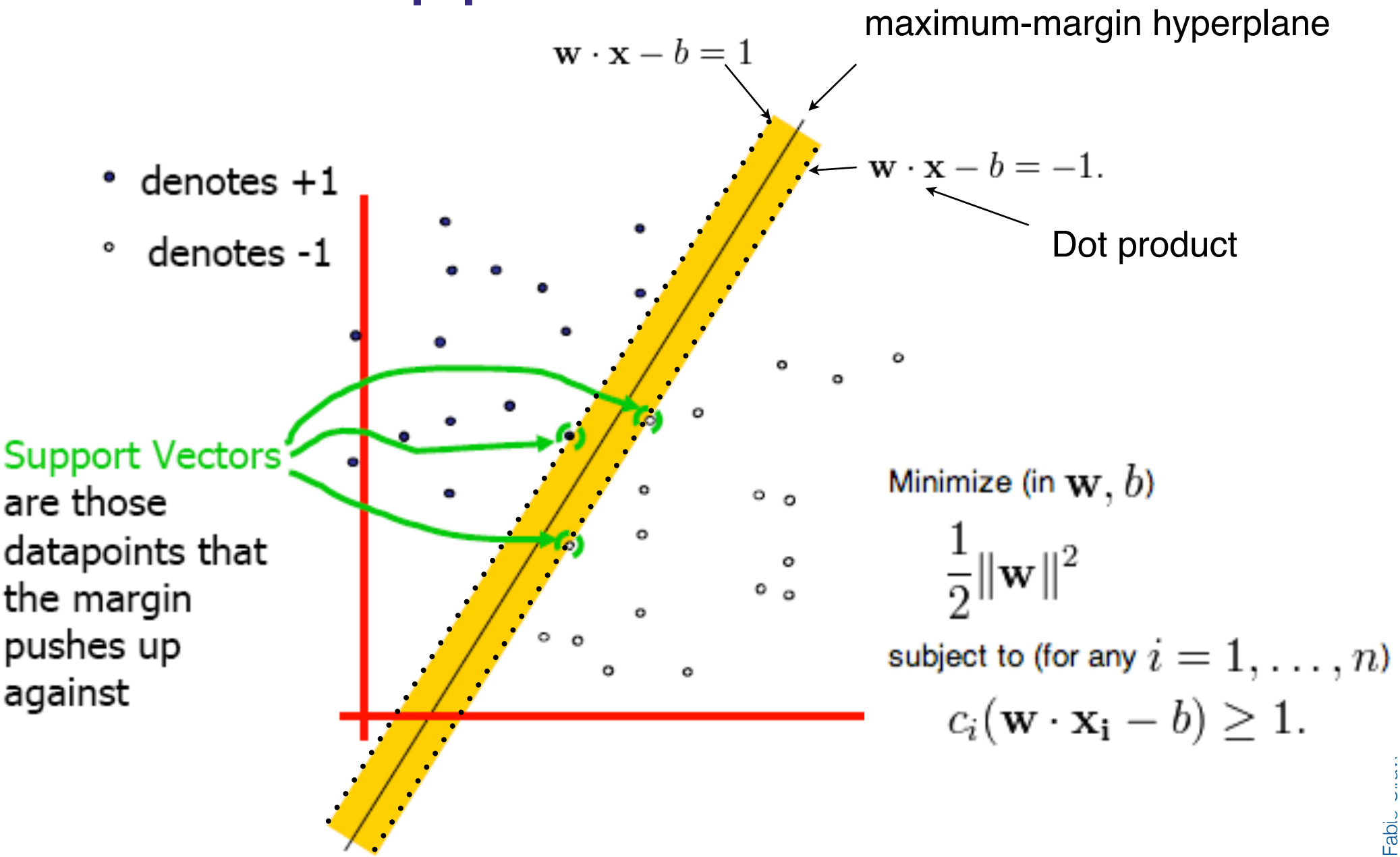
Graphically



http://en.wikipedia.org/wiki/Support_vector_machines

© Fabio Ciravegna, University of Sheffield

Support Vectors



Copyright © 2001, 2003, Andrew W. Moore

http://www.cs.cmu.edu/afs/cs/Web/People/awm/tutorials/svm15.pdf

© Fabi...

The University Of Sheffield.

Max-Margin Algorithms

- Support Vector Machines extended for training on structured models
  - Goal is to find a *w* such that the score $s(y_i) = w\, f(x_i, y_i)$ has a margin of at least $err(y, y_i)$ more than $s(y_i) = w\, f(x_i, y)$
    - i.e. the score of labelling correctly a vector x is at least more than the score of labelling the vector with a wrong label
      - the difference is core is determined by the user function err(y, yi)

Algorithm 3.2. Train$(D = \{(\mathbf{x}_\ell, \mathbf{y}_\ell)\}_{\ell=1}^N, \mathbf{f} : f_1 \cdots f_K), C$

1. **Output:** $\mathbf{w} = \text{argmin}\frac{1}{2}||\mathbf{w}||^2 + C\sum_{\ell=1}^N \max_{\mathbf{y}}(\text{err}(\mathbf{y}, \mathbf{y}_\ell) + \mathbf{w}\cdot\mathbf{f}(\mathbf{x}_\ell, \mathbf{y}) - \mathbf{w}\cdot\mathbf{f}(\mathbf{x}_\ell, \mathbf{y}_\ell))$
2. **Initialize** $\mathbf{w}^0 = \mathbf{0}$, Active constraints = Empty.
3. **for** $t = 1 \cdots \text{maxIters}$ **do**
4. **for** $\ell = 1 \cdots N$ **do**
5. $\hat{\mathbf{y}} = \text{argmax}_{\mathbf{y}}(\text{err}(\mathbf{y}, \mathbf{y}_\ell) + \mathbf{w} \cdot \mathbf{f}(\mathbf{x}_\ell, \mathbf{y}))$
6. **if** $\mathbf{w}\cdot\mathbf{f}(\mathbf{x}_\ell, \mathbf{y}_\ell) < \text{err}(\hat{\mathbf{y}}, \mathbf{y}_\ell) + \mathbf{w}\cdot\mathbf{f}(\mathbf{x}_\ell, \hat{\mathbf{y}}) - \xi_\ell - \epsilon$ **then**
7. Add $(\mathbf{x}_\ell, \hat{\mathbf{y}})$ to the set of constraints.
8. $\mathbf{w}, \xi$ = solve QP with active constraints.
9. **Exit** if no new constraint added.

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Relation Extraction

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

The Task

- Recognising relationships among entities
- Two aproaches
  - Known entities, discover relations
  - Known entity and relation, discover the target of the relation (entity)
  - Known relation, discover entities
    - Entities are unmarked
      - Unsupervised methods

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Relationship identification

- Given a text snippet *x* and two marked entities E1 and E2 in x, identify if there is any of the relationships Y between E1 and E2 .
  - Possible relations: all domain relations plus the relation null
- Problem is simpler than entity extraction
  - because only a scalar prediction is required
    - instead of a vector of predictions
- However, relationship extraction is considered a harder problem
  - because it requires a combination of
    - local and nonlocal noisy clues from diverse syntactic and semantic structures in the text

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Tokens

- The tokens around and in-between the two entities hold strong clues for relationship extraction.
  - ⟨Company⟩ Kosmix ⟨/Company⟩ is located in the ⟨Location⟩ Bay area ⟨/Location⟩.
- Tokens can often be
  - Stemmed
    - Located== locat
  - Morphological analised
    - Located==locate
  - PoS tagged (e.g. Noun, verb, etc.)

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Chunk Parsing

- Major syntactic groups are identified, but no syntactic relation is established among them
  - Easy to implement
  - Virtually error proof
  - Works also for documents with linguistic idiosyncrasies and or formatted text (e.g. Tables)

⟨NG⟩ Haifa ⟨/NG⟩, <VG> located </VG> <NG> 53 miles </NG> from <NG> Tel Aviv ⟨/NG⟩, <VG> will host </VG> <NG> ICML </NG> in <NG> 2010 </NG>.

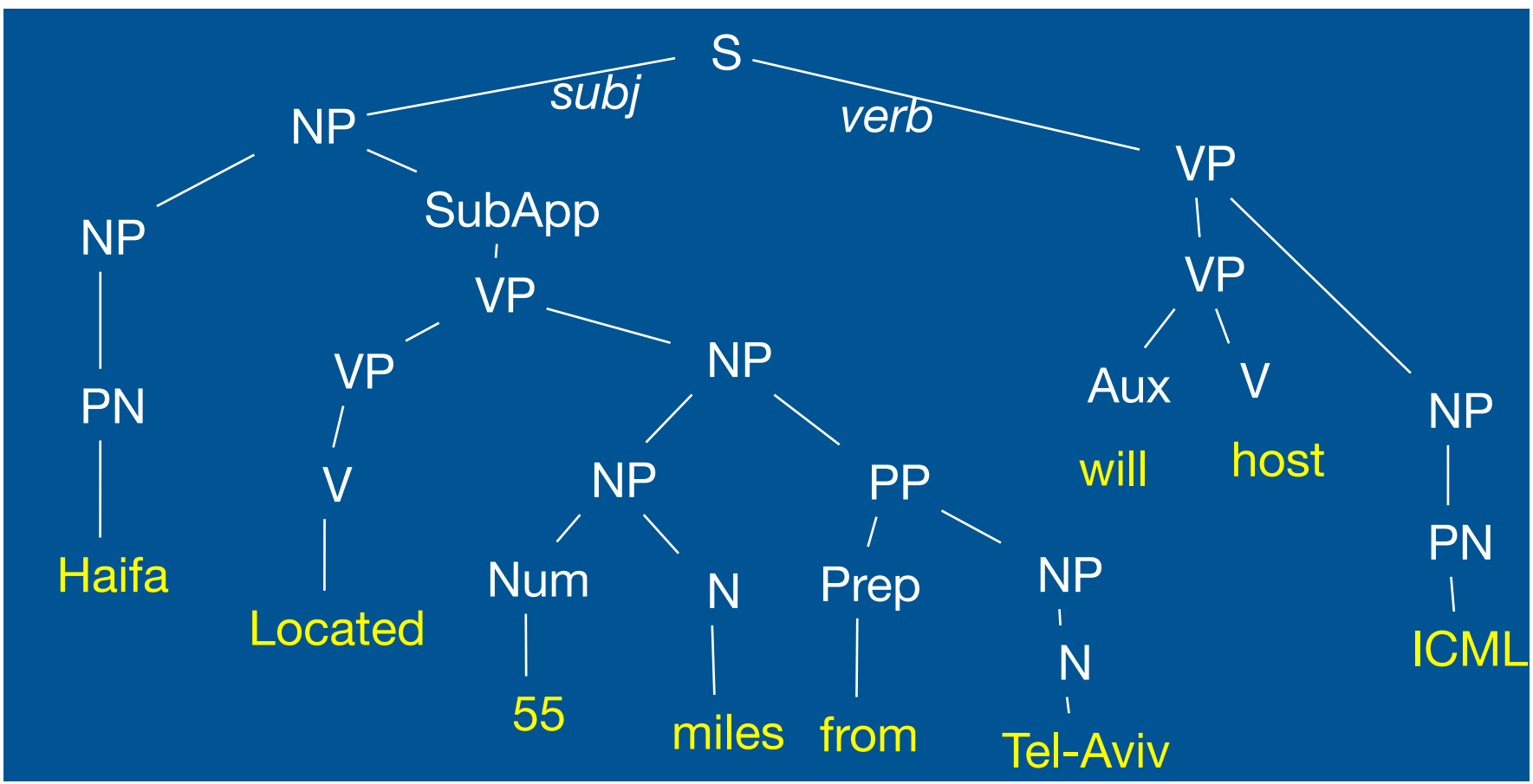
Steven P. Abney, Parsing By Chunks, in Principle-Based Parsing, 257-278, Kluwer Academic Publishers

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Parse trees

- Establishes syntactic relations among words and group of words according to a NL grammar
  - Easy to recognise that Haifa will host ICML, not Tel-Aviv
  - Difficult to model language perfectly, hence several broken trees



© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Dependency graph

- It links each word to the words that depend on it
  - Relations may or may not be typed
- Full parse trees are expensive to create.
  - A dependency graph is often as adequate as a parse tree
    - Some complexities may be lost, hence more robust than parse tree

ersity of Sheffield

The University Of Sheffield.

Issues in Features

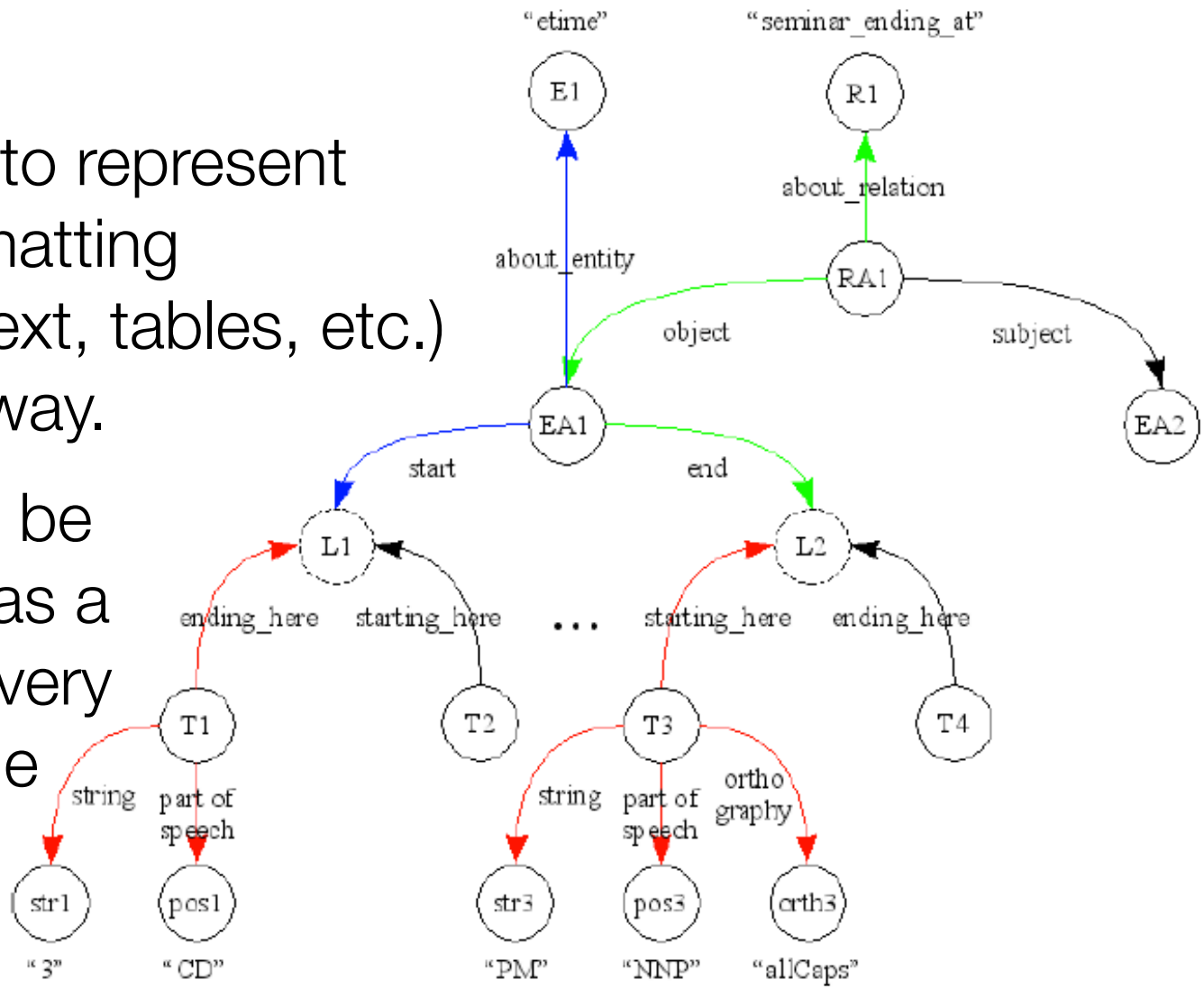
- While in entity recognition features are always related to a token
  - Now we have different structural forms
    - Token features, trees, graphs
- This implies more difficulty in learning
  - Methods:
    - Feature-based methods that extract a flat set of features from the input and then invoke an off-the-shelf classifier (e.g. SVM)
    - Representing feature space as a graph
      - Application for Kernel Methods for graph learning
        - Kernel-based methods that design special kernels to capture the similarity between structures
    - Rule-based methods
      - Similar to those for entity extraction

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

F-Space as Graph

- Tokens, features and relations represented as graphs
  - It is possible to represent also text formatting (e.g. HTML text, tables, etc.) in the same way.
  - Learning can be represented as a task of discovery of walks in the graph



Jose' Iria and Fabio Ciravegna: A Methodology and Tool for Representing Language Resources for Information Extraction 5th International Conference on Language Resources and Evaluation (LREC 2007), Genoa, 24-25-26 May, 2006

© Fabio Ciravegna, University of Sheffield

Kernel Trick

- A method for using a linear classifier algorithm to solve a non-linear problem
  - by mapping the original non-linear observations into a higher-dimensional space,
  - where the linear classifier is subsequently used;
- this makes a linear classification in the new space equivalent to non-linear classification in the original space.

Source Wikipedia

© Fabio Ciravegna, University of Sheffield

Kernel Functions

- Kernel functions enable Kernel Methods to operate in the feature space
  - without ever computing the coordinates of the data in that space,
  - but rather by simply computing the inner products between the images of all pairs of data in the feature space

Source Wikipedia

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Kernel Example

- Let T and T ′ represent the dependency trees of two different training instances
  - X = (x, E1 , E2) and X ′ = (x′ , E′1 , E′2)
- The kernel function K(X, X ′) is defined in a dependency tree
  - P= Shortest path connecting (E1,E2)
  - P'= Shortest path connecting (E1',E2')
  - P1...Pk set of properties along the path
- Two nodes are considered similar if the value of many of these k properties are common

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Example (ctd)

- The node similarities are used to define the kernel function as follows:

$$K(P,P') = \begin{cases} 0 & \text{if } P, P' \text{ have different lengths} \\ \lambda \prod_{k=1}^{|P|} \text{CommonProperties}\,(P_k, P'_k) & \text{otherwise,} \end{cases}$$

- where CommonProperties (Pk , P′k) measures the number of properties common between the kth node along the paths P and P' , respectively.
- Kernel value is high when the length of the shortest path between the two entities is the same in both sentences and the nodes along the path share many common properties.

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Further Details

- For further details and a thorough analysis of the state of the art, see:

Sunita Sarawagi:
Information Extraction,
Foundations and Trends in Databases, Vol. 1, No. 3 (2007) 261–377

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Evaluation in IE

© Fabio Ciravegna, University of Sheffield

Importance of Evaluation in IE

- IE was born from a series of competitive evaluations organised by DARPA in the US
  - MUC Conferences, 1989-1998
    - IE as a departure from IR but using the same types of measures of accuracy
    - The idea was to understand what worked and what not in text analysis
      - Finding a way to compare IE systems and approaches in a controlled way
- Evaluation is in IE's DNA
  - Publishing IE papers without evaluation is not considered acceptable

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Organising Evaluation

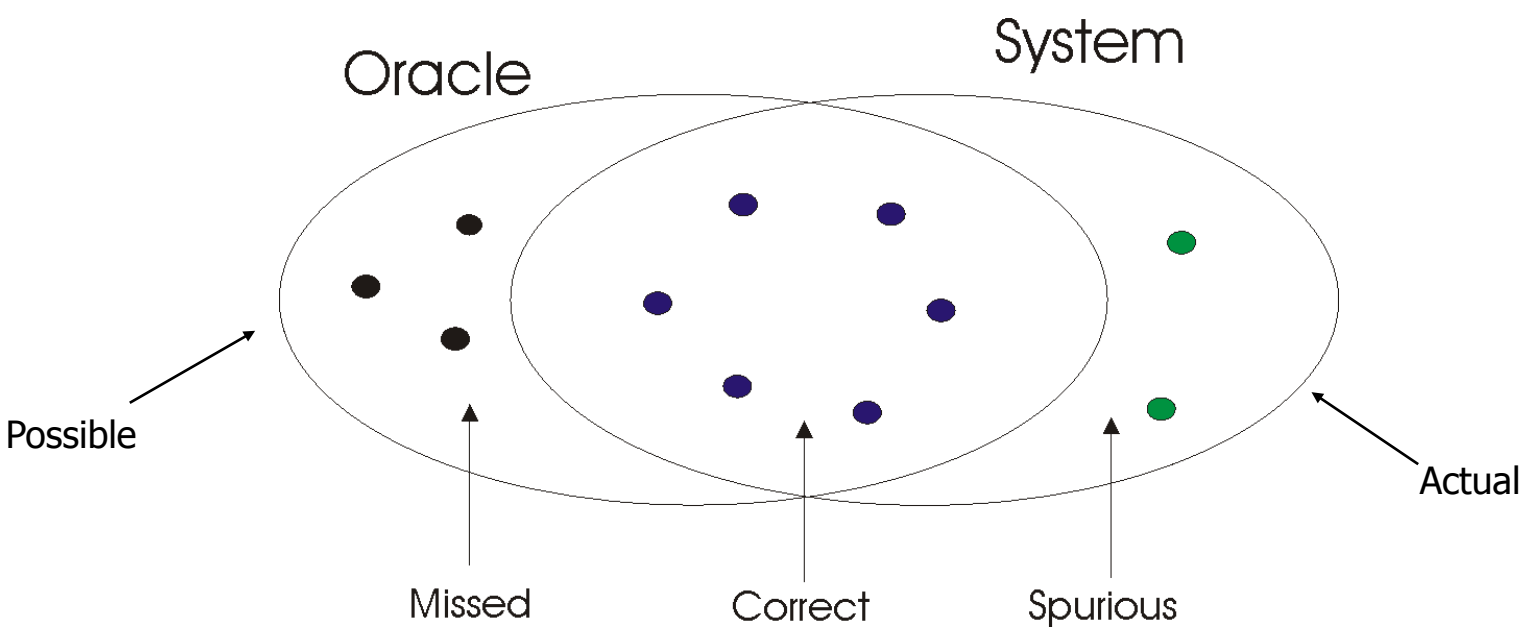
- You will need:
  - An annotated training corpus
    - That you will use to develop rules or to train a machine learning algorithm
  - A result scorer
    - A tool that automatically computes accuracy of the system against an annotated corpus
    - E.g. The MUC Scorer
  - An annotated test corpus
    - To be used blindly to test results
      - Please note that run on test corpus should be a one off test
        - Test corpus is not be used to fine tuning accuracy in any way
        - E.g. By looking at the results and changing your rules or by tuning the learning parameters

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

The Rationale Behind

- **Precision**: how correct is the average answer provided by the system
- **Recall**: how many (correct) pieces of information are retrieved by the system
- **F-measure**: allows comparative evaluations



© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Evaluation Measures

$$\text{Recall} = \frac{\text{CORRECT} + (\text{PARTIAL} * 0.5)}{\text{POSSIBLE}}$$

$$\text{Precision} = \frac{\text{CORRECT} + (\text{PARTIAL} * 0.5)}{\text{ACTUAL}}$$

$$F(\beta) = \frac{(\beta^2 + 1) * \text{PREC} * \text{REC}}{\beta^2 * \text{PREC} + \text{REC}}$$

F-Measure is to be used to compare systems

In all evaluations all the three measures must be published

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Issues in Evaluation

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Issues Affecting Evaluation

- The Algorithm
- The feature set used
- The leniency in assessing results
  - the availability of standard annotated corpora do not guarantee that the experiments performed with different approaches and algorithms proposed in the literature can be reliably compared
    - Data problems
    - Problems of experimental design
    - Problems of presentation

Alberto Lavelli, Mary E Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson:
Evaluation of Machine Learning-based Information Extraction Algorithms: Criticisms and Recommendations, Language Resources and Evaluation, Volume 42, Issue 4 (December 2008).

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Leniency in Evaluation

- Data Problems
  - Errors in data, branching corpora, templates Vs markup
- Experimental design
  - Training/Test Set selection
    - e.g. 50/50 Vs 80/20
  - Tokenization
  - How to count matches (see below)

Alberto Lavelli, Mary E Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson:
Evaluation of Machine Learning-based Information Extraction Algorithms: Criticisms and Recommendations,
Language Resources and Evaluation, Volume 42, Issue 4 (December 2008).

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Issues in Evaluation

- Fragment evaluation:
  - How leniently should inexact identification of filler boundaries be assessed?
- Counting multiple matches:
  - When a learner predicts multiple fillers for an entity, how should they be counted?
- Filler variation:
  - When text fragments having distinct surface forms refer to the same underlying entity, how should they be counted?

Alberto Lavelli, Mary E Califf, Fabio Ciravegna, Dayne Freitag, Claudio Giuliano, Nicholas Kushmerick, Lorenza Romano, and Neil Ireson:
Evaluation of Machine Learning-based Information Extraction Algorithms: Criticisms and Recommendations,
Language Resources and Evaluation, Volume 42, Issue 4 (December 2008).

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Why use NL Features?

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

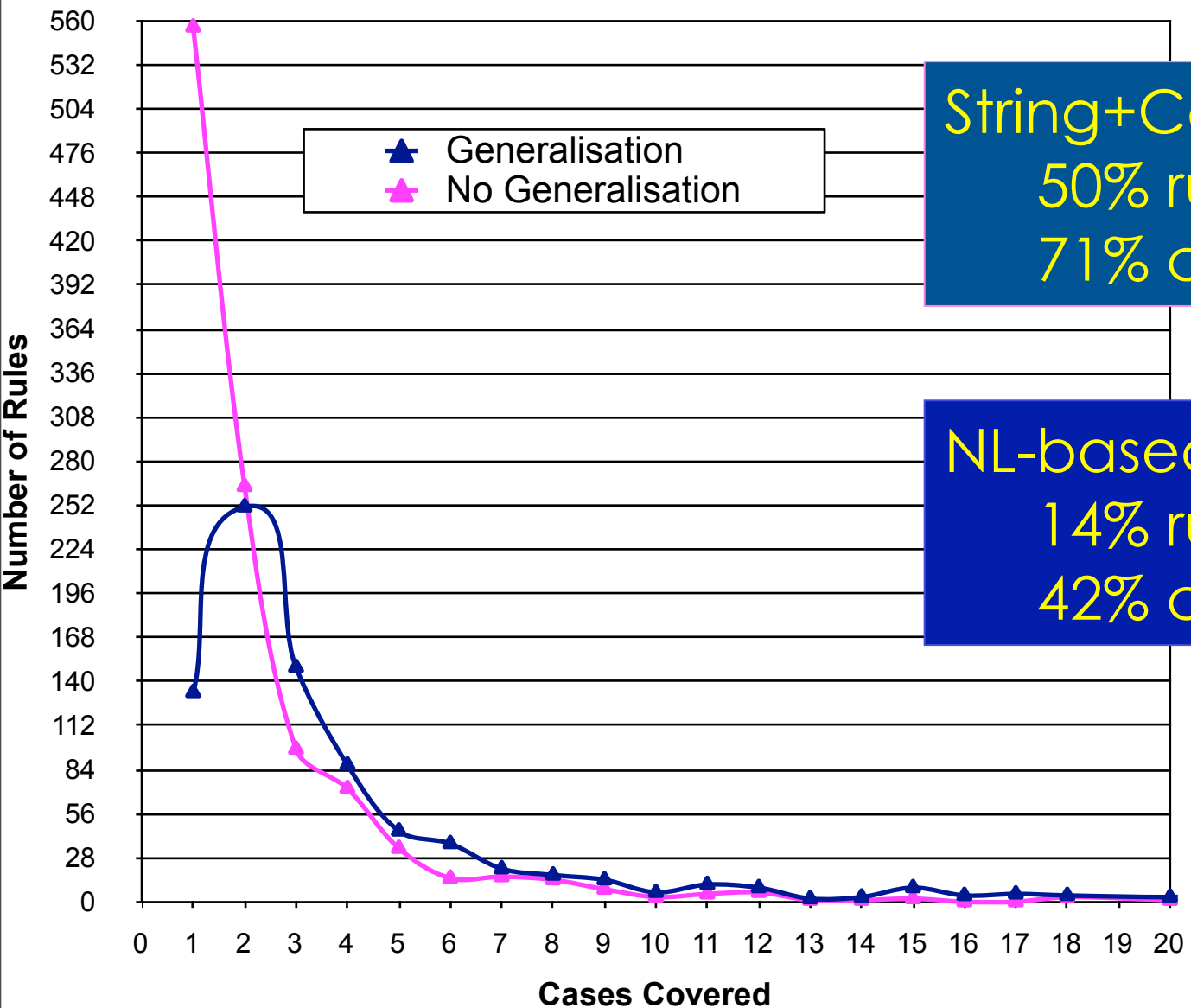
Why use NL Features

- Words are not independent pieces of data
  - There is a clear relationship among them
    - Syntax and semantics of the language influence their composition
      - Some grouping are required and their presence can be indicative of specific regularities
  - Some elements are non enumerable
    - E.g. Numbers, time expressions, etc.
      - Rules that try to learn from numbers are extremely inefficient
        - Most of the times try to make discrete what is continuous
- Linguistic features allow to generalise over the flat data structure
  - As humans we would never learn a rule like “John Smith’s talk” ==> speaker: John Smith
    - But rather:
      <person name>’s talk ==> speaker: <person name>

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Effect of NL Features(2)
Reduction in Data Sparsity



String+Case
- 50% rules cover 1 case
- 71% cover up to 2 cases

NL-based features
- 14% rules cover 1 case
- 42% cover up to 2 cases

Corpus: CMU seminar announcements

Algorithm: $(LP)^2$

Fabio Ciravegna:
Adaptive Information Extraction from Text by Rule Induction and Generalisation
in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001) , Seattle, August 2001

© Fabio Ciravegna, University of Shef

The University Of Sheffield.

Effect of features

- Effectiveness of using
  - NL features (G)
  - Just string and case (NG)

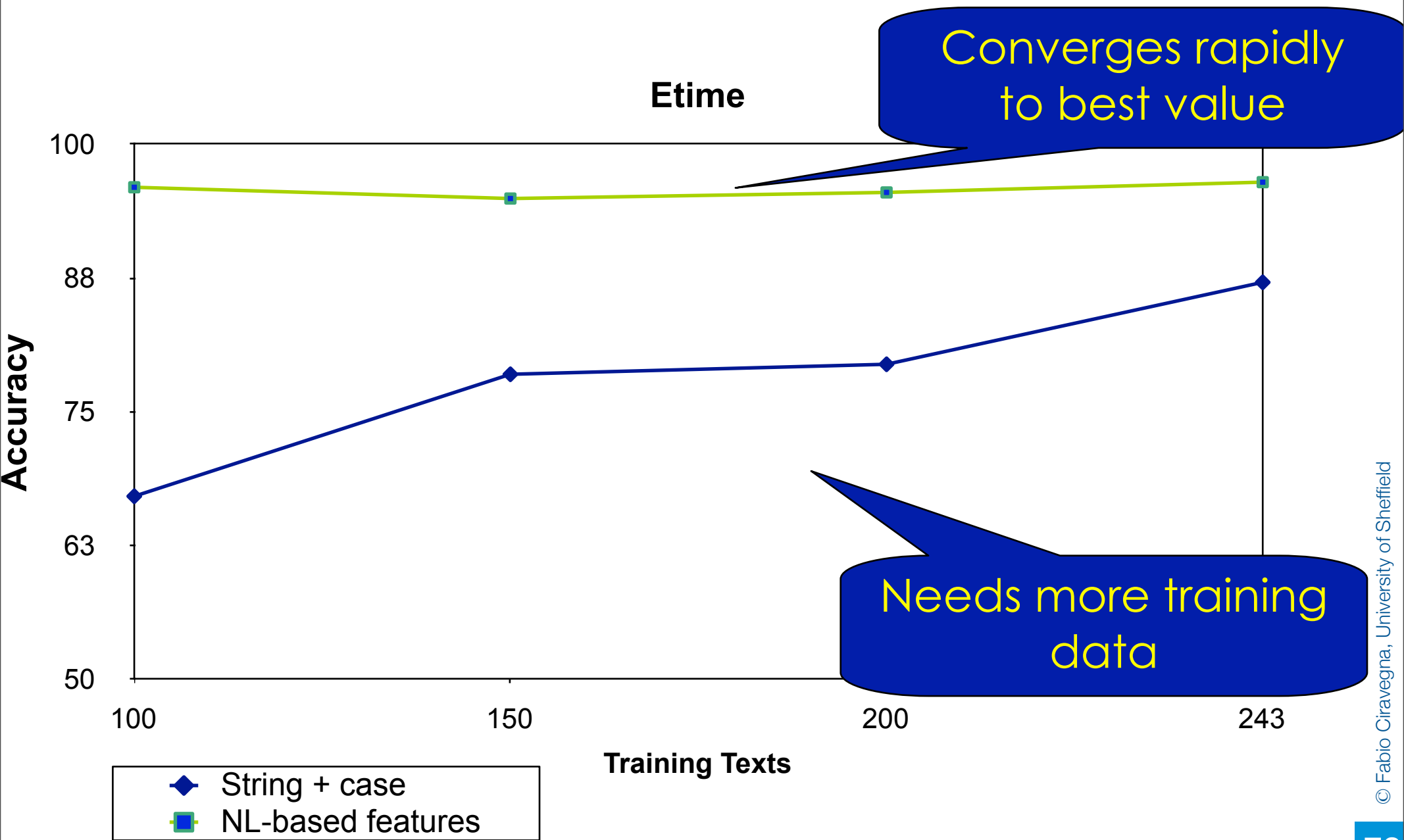
Slot	$(LP)^2_G$	$(LP)^2_{NG}$
speaker	72.1	14.5
location	74.1	58.2
stime	100	97.4
etime	96.4	87.1
All slots	**89.7**	**78.2**

Fabio Ciravegna:
Adaptive Information Extraction from Text by Rule Induction and Generalisation
in Proceedings of 17th International Joint Conference on Artificial Intelligence (IJCAI 2001) , Seattle, August 2001

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Effect of Features(4)



© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

A comparative Evaluation

The Pascal Challenge on ML-based Information Extraction

Neil Ireson, Fabio Ciravegna, Marie Elaine Califf, Dayne Freitag, Nicholas Kushmerick, Alberto Lavelli:
Evaluating Machine Learning for Information Extraction,
22nd International Conference on Machine Learning (ICML 2005), Bonn, Germany, 7-11 August, 2005

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Pascal Challenge on ML-based IE

- Fair comparison of ML algorithms for IE through a controlled application of the methodology
- Summary assessment of the general benefit of state-of-the-art ML to the problem of IE.
- Identification of any challenges not adequately addressed by ML approaches.
- Publication of an extensive testbed to enable comprehensive, comparable research beyond the lifetime of the challenge

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Pascal Challenge Corpus

- A corpus of 1,100 documents
  - 850 Workshop Call for Papers (CFPs)
  - 250 Conference CFPs
- Training Corpus (400 Workshop CFPs)
  - Randomly divided into 4 sets of 100 documents.
    - Each of these sets is further randomly divided into 10 subsets of 10 documents.
- Test Corpus (200 Workshop CFPs)

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Pascal Challenge: Tasks

- Given all the available training documents learn the textual patterns necessary to extract the annotated information
- Learning Curve: Examine the effect of limited training resources on the learning process by incrementally adding the provided subsets to the training
- Active Learning: Examine the effect of selecting which documents to add to the training data

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Pascal Challenge: Result Sample

	Rule Learning			SVM			CRF		
Slot	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
ws name	65.6	24.1	35.2	62.9	53.9	58.0	61.8	*57.6*	59.6
ws acronym	*88.7*	*84.4*	*86.5*	73.8	52.3	61.2	80.6	35.8	49.6
ws date	76.9	63.2	69.4	81.0	66.6	73.1	82.2	69.3	*75.2*
ws home	*86.4*	61.9	72.1	65.6	87.0	*74.8*	67.8	66.5	67.1
ws location	62.1	40.2	48.8	61.1	*67.4*	64.1	73.7	57.6	64.7
ws submission	*87.6*	*85.1*	*86.4*	71.9	76.37	74.0	74.7	68.0	71.2
ws notification	88.9	*88.9*	*88.9*	86.7	82.1	84.3	87.0	77.4	81.9
ws camerarady	87.6	*86.5*	*87.0*	76.4	73.6	75.0	77.7	79.1	78.4
conf name	79.2	42.2	*55.1*	64.9	41.1	50.3	64.3	40.0	49.3
conf acronym	92.2	*88.8*	*90.5*	61.9	34.8	44.5	57.6	42.8	49.1
conf home	65.6	28.0	*39.3*	36.8	9.3	14.9	38.9	9.3	15.1

© Fabio Ciravegna, University

The University Of Sheffield.

Annotating Documents to IE Train Systems

Can we really ask people to annotate documents?

Most slides are from Ziqi Zhang, University of Sheffield

© Fabio Ciravegna, University of Sheffield

Do People Like Annotating?

- No, they hate it
  - They will try not to do it or do it quickly
- It is time and energy consuming
  - It is not their job
    - Unless they are professional annotators
  - They are not rewarded for it
- It is tiring
- It is error prone
- But most of all: is it possible to annotate documents with sufficient accuracy to train an IE system?

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

The archaeotools Experience

- A project funded by AHRC/EPSRC/JISC in the UK. In collaboration with the University of York (Archaeology Department)
- Goal:
  - Building an e-archaeology application to allow archaeologists to discover, share, and analyse datasets and legacy publications
- Role of IE: To identify in several collections of documents:
  - Pacenames: around 2,000 in corpus
    - Yorkshire, Cambridge, The London Tower, Baker Street, St. Paul, Church road.
  - Subjects: around 10,000
    - Roman pottery, spearhead, animal remains, church, courtyard, plates, vessel
  - Temporals: around 4,000
    - Roman, Saxon, AD1078, 300BC, 43 - 801AD, circa 1771, Victorian era, Bronze Age

http://nlp.shef.ac.uk/wig/research/ArchaeoTools.html

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

IE in Aracheotools

- Based on SVN
  - The TRex tool http://t-rex.sourceforge.net/
- Training based on corpora annotated by 5 expert archaeologists
  - training documents 42, length: up to several hundreds of pages
  - total documents to tag by machine learning: 967
  - total documents to tag by rules: 3991
- Annotation process was geared at high quality
  - Annotation instructions were clarified through several iterations
    - Our archaeologists colleagues, they clearly explained the task to annotators, went through examples with them
    - The IE experts went through several confusing examples with archaeologists to clarify their doubts
  - One senior researcher was appointed to make final decision in case of doubts from any annotators
  - Annotators were very motivated and the task was part of their job!!!

© Fabio Ciravegna, University of Sheffield

IE challenges – annotation quality

IAA F-measure – Inter-Annotator-Agreement F-measure, Hripcsak and Rothschild (2005).

		Annotator A	
		Positive	Negative
Annotator B	Positive	a	b
	Negative	c	d

- ✓ Treating A's annotations as gold standard, and B's as reference
- ✓ Precision of B = a/(a+b), Recall of B = a/(a+c)
- ✓ F-measure of B = 2a/(2a+b+c)
- ✓ Equivalent to the standard P, R, F metrics used for evaluating IE systems

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Annotation quality (ctd.)

- IAA F-measure – Inter-Annotator-Agreement F-measure
  - ✓ Figures obtained from a shared corpus annotated by three different annotators

	Place name	Subject	Temporal
Lowest IAA between any two annotators	66.2	49	67.2
Highest IAA between any two annotators	*80*	*63*	*83.3*

© Fabio Ciravegna, University of Sheffield

Annotation Quality (ctd.)

- Individual annotator v.s. all annotators
  - ✓ Create a corpus that is annotated by different annotators, train a learning system on n% of the corpus and evaluate on the other 1-n% (combined corpus)
  - ✓ For each annotator, select the documents annotated only by himself/herself, train a learning system on n% of the corpus and evaluate on the other 1-n% (individual corpus)
  - ✓ The assumption is each annotator's behaviour is consistent w.r.t. him/herself, but may be noisy to others
  - ✓ … thus the second approach (individual corpus) should produce better learning systems than the first (combined corpus) even if number of training examples decreases

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Annotation Quality (ctd.)

- Individual annotator v.s. all annotators

	Combined corpus		Average figure from Individual corpus (5 annotators)		Annotations found in the corresponding Individual corpus as % of the Combined corpus
	P	R	P	R	
Place name	53.3	64.7	**59.3**	**74**	45%
Subject	56.5	50.1	**74.4**	**70.1**	30%
Temporal	62.4	59.4	**74.9**	**76**	35%

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Annotation Quality (ctd.)

- What about training a system on one annotator's corpus (A) and test on all the other annotators' corpora?

Training corpus	Testing corpora	Temporal			Place name			Subject		
Annotator	Annotator	P	R	F	P	R	F	P	R	F
All	All	*62.4*	*59.4*	*61*	*53.3*	*64.7*	*58*	*56.5*	*50.1*	*53*
A	B,C,D,E	**71.2**	**72.9**	**72**	57.5	49.3	**53**	**59.8**	**73.9**	**66**
B	A,C,D,E	**70.7**	**75.5**	**73**	54.5	43.6	**48**	**81.5**	**53.6**	**65**
C	A,B,D,E	**68**	**79.1**	73	47.1	70.6	57	44.4	85.5	58
D	A,B,C,E	**69.1**	**73.4**	**71**	52.7	57.2	**55**	**68.5**	**63.2**	**66**
E	A,B,C,D	**68.9**	**60**	**64**	45.4	64.2	**53**	**75.3**	**64.2**	**69**
			AVG	71		AVG	53		AVG	65

- Training on individuals reports higher results for ⅔ fields!
- Training on C: always reports lower precision higher recall!
- Training on B: always reports excellent P, but R for Place and Subject name is the lowest

© Fabio Ciravegna

The University Of Sheffield.

In Summary

- Temporal:
  - P&R: higher than in the combined version in all cases
    - P:[+6%,+9%], R: [+1%,+20%], Average F=+10%
- Place names:
  - It reports worse results than the combined version
    - One dramatic case of fall in R: -21%!!!
  - P: [-8%,+3&], R:[-21%,+6%], Avg F=-5%
- Subject
  - It reports better average F-measure but one case of severe loss in precision
  - P: [-12%,+26&], R:[+3%,+35%], Avg F=+12%

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Annotation Quality - Conclusions

- In general archaeology is a difficult domain, with many uncertainty and ambiguity even for humans
- Inconsistency between annotators generated noise that influences learning system
- Very careful evaluation of the quality of annotation must always be implemented
  - Aka possibility/ability for the annotators to perform good quality annotation
- Never ever suppose that humans are 100% correct
  - For complex tasks they may perform at 80% accuracy!!!!
    - Always ask users to annotate (at least partially) overlapping sets of documents
      - So to be able to check their agreement

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Available tools for IE

Where to go if you need not to reinvent the wheel

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Available Tools

- Gate: general architecture for NLP
  - Mainly specialised for IE after 2001
  - Provides a long list of open source libraries
  - It contains Annie, a NER architecture
  - www.gate.shef.ac.uk
- UIMA
  - Largely inspired by Gate
  - IBM developed, Industrially supported
- Open Calais
  - Web Service to automatically create rich semantic metadata for submitted content (NER, Events, etc.)
  - http://www.opencalais.com/

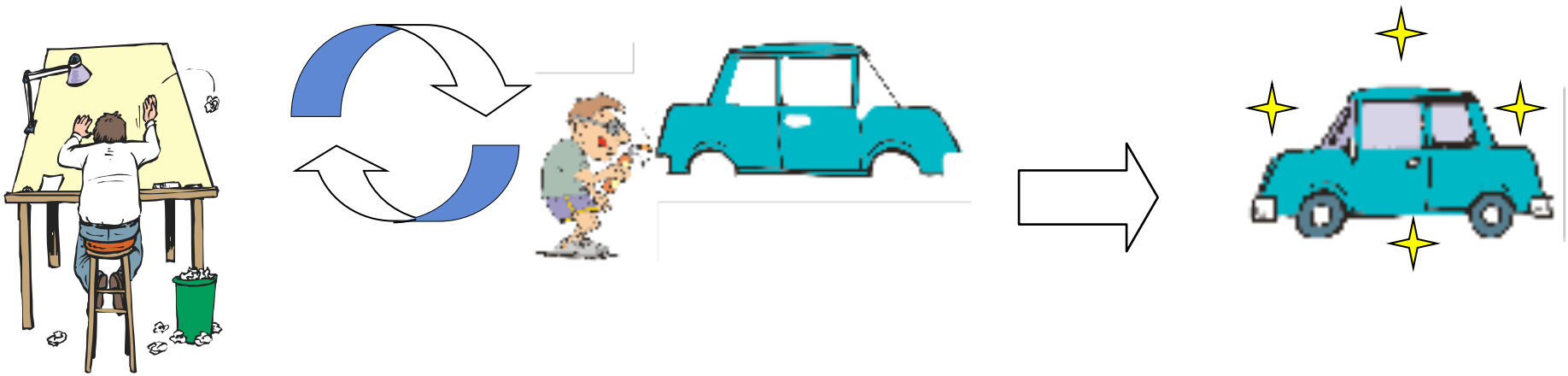
© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

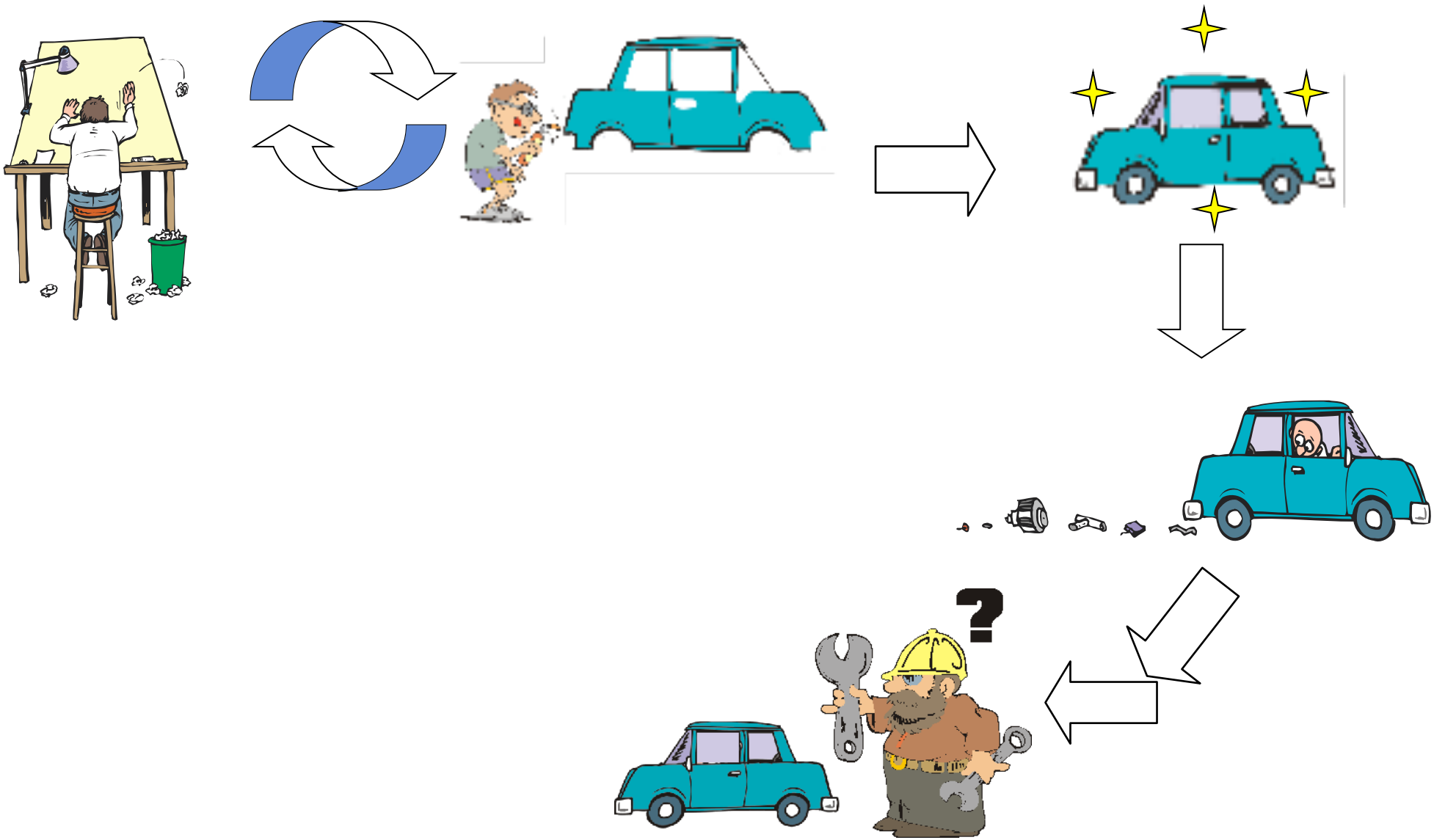
Using IE for Knowledge Management

© Fabio Ciravegna, University of Sheffield

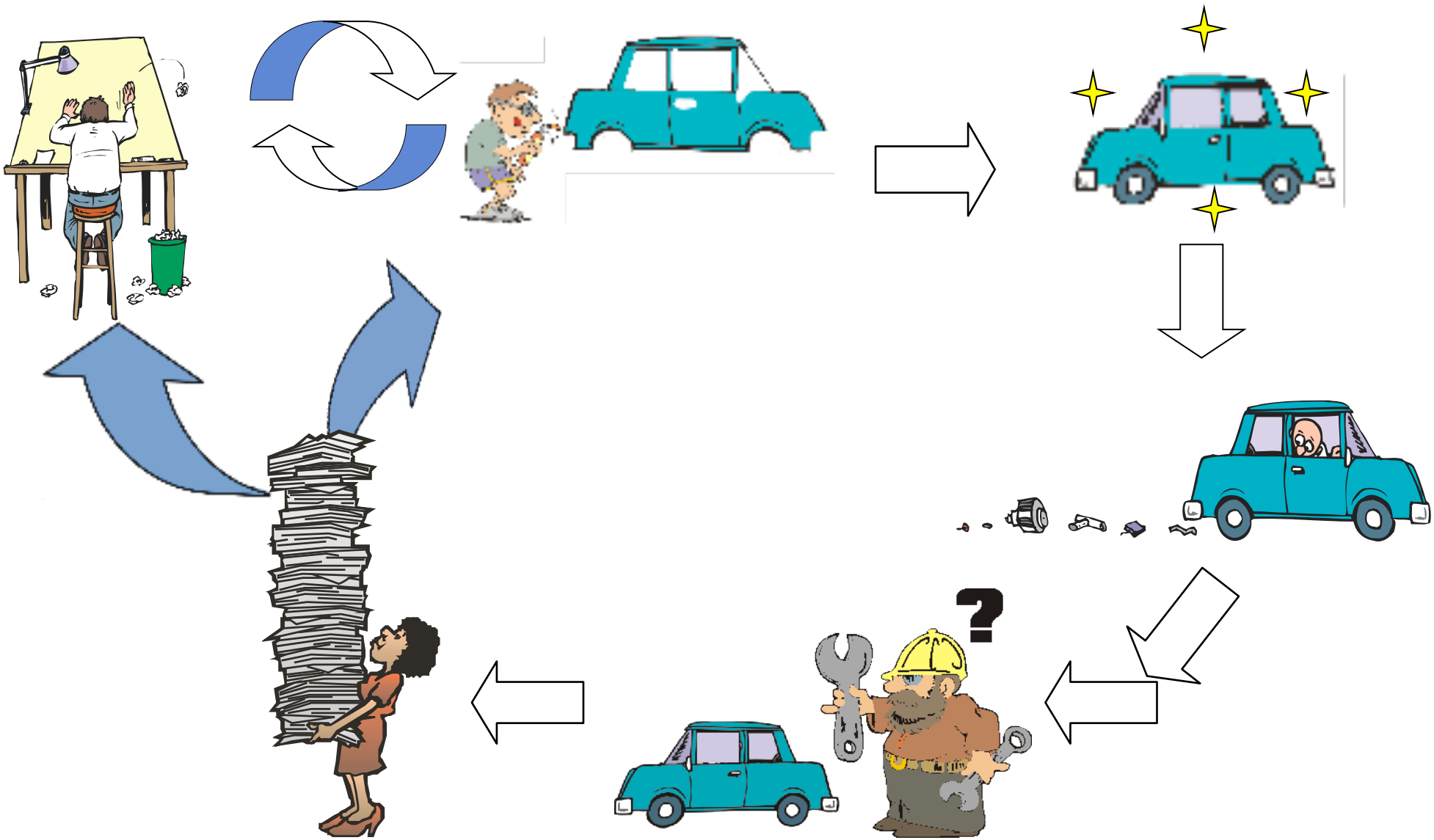
Traditional Knowledge Management



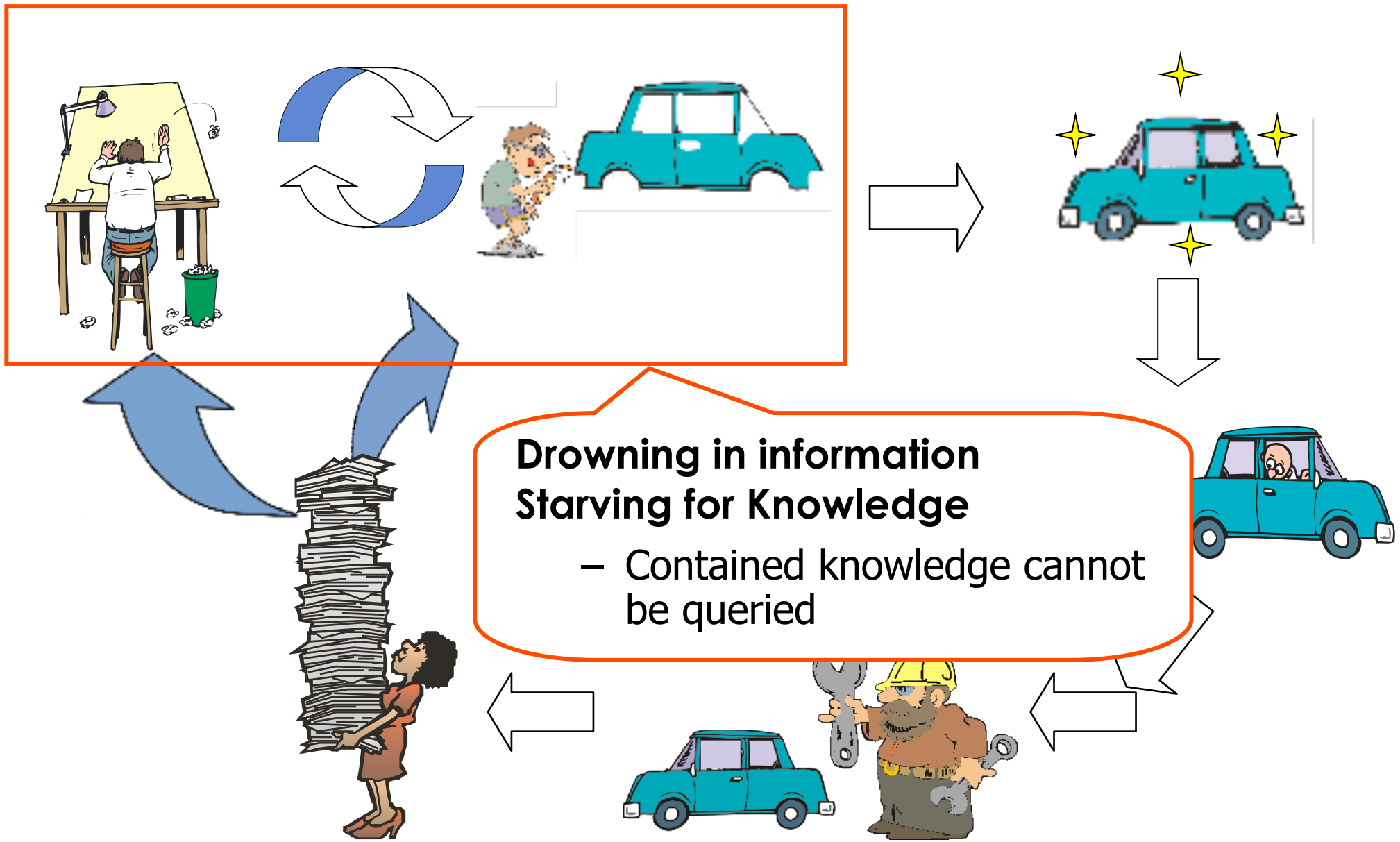
Traditional Knowledge Management



Traditional Knowledge Management



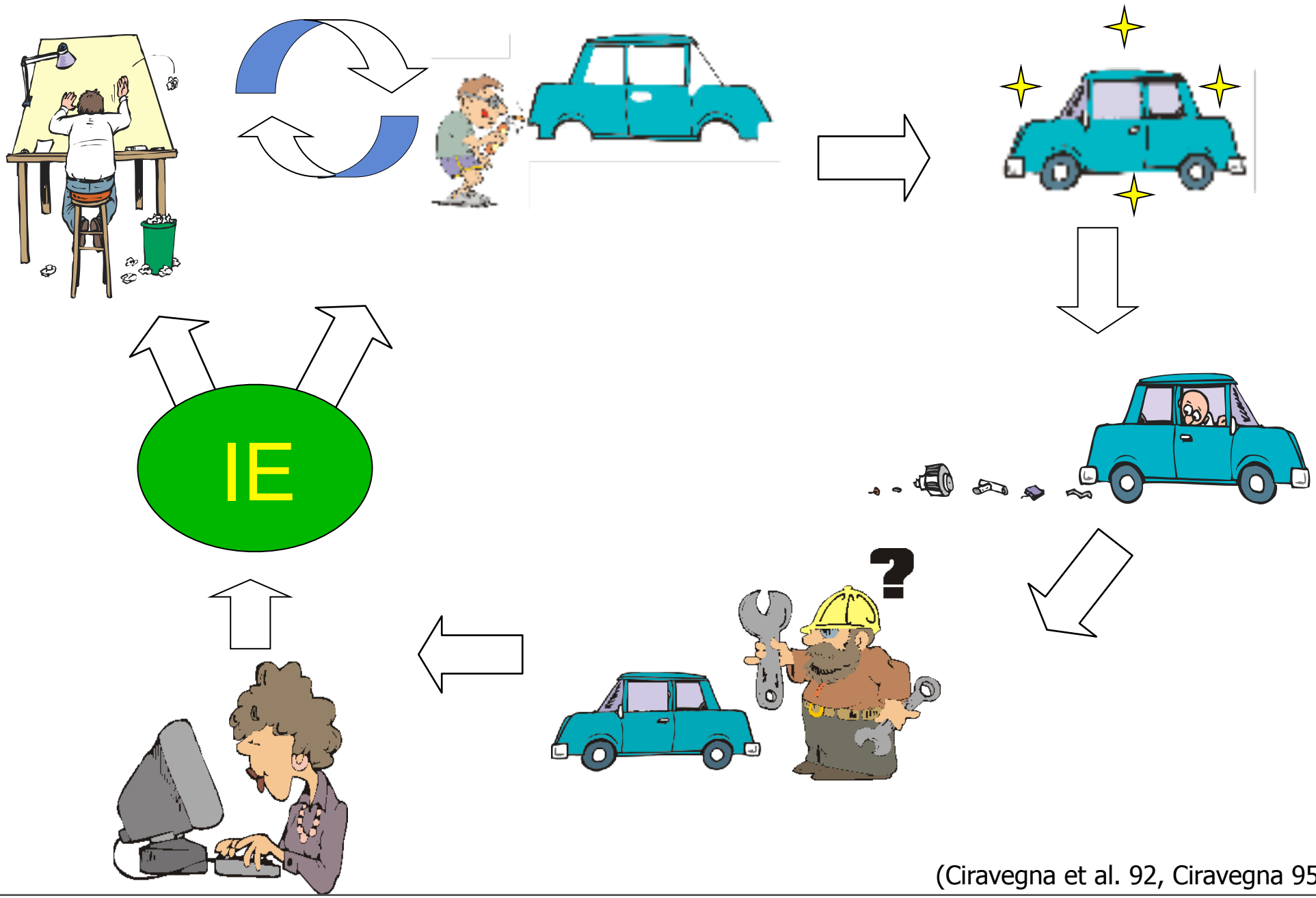
Traditional Knowledge Management



Drowning in information
Starving for Knowledge

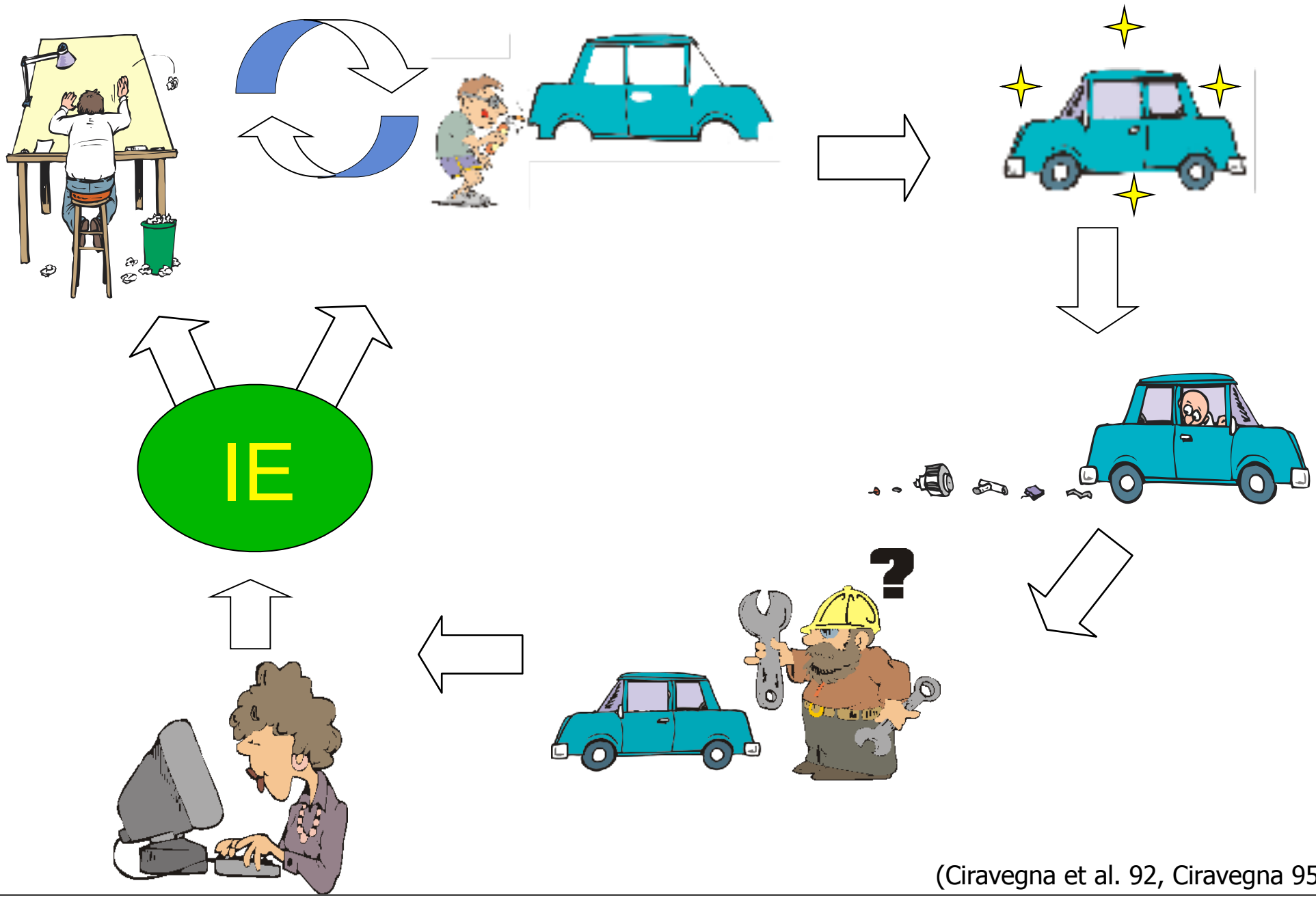
- Contained knowledge cannot be queried

Knowledge Management using IE



(Ciravegna et al. 92, Ciravegna 95)

Knowledge Management using IE



(Ciravegna et al. 92, Ciravegna 95)

Knowledge Management using IE

REF.: 00140/89

STRUCTURED DATA: <licence plate number, model, km, ...>

TOPIC: Mancato funzionamento motorino avviamento.

TEXT: Sulle auto per presentazione a stampa specializzata si verifica il mancato funzionamento del motorino avviamento durante prova pergola (motorino EY8 0, 8/72).

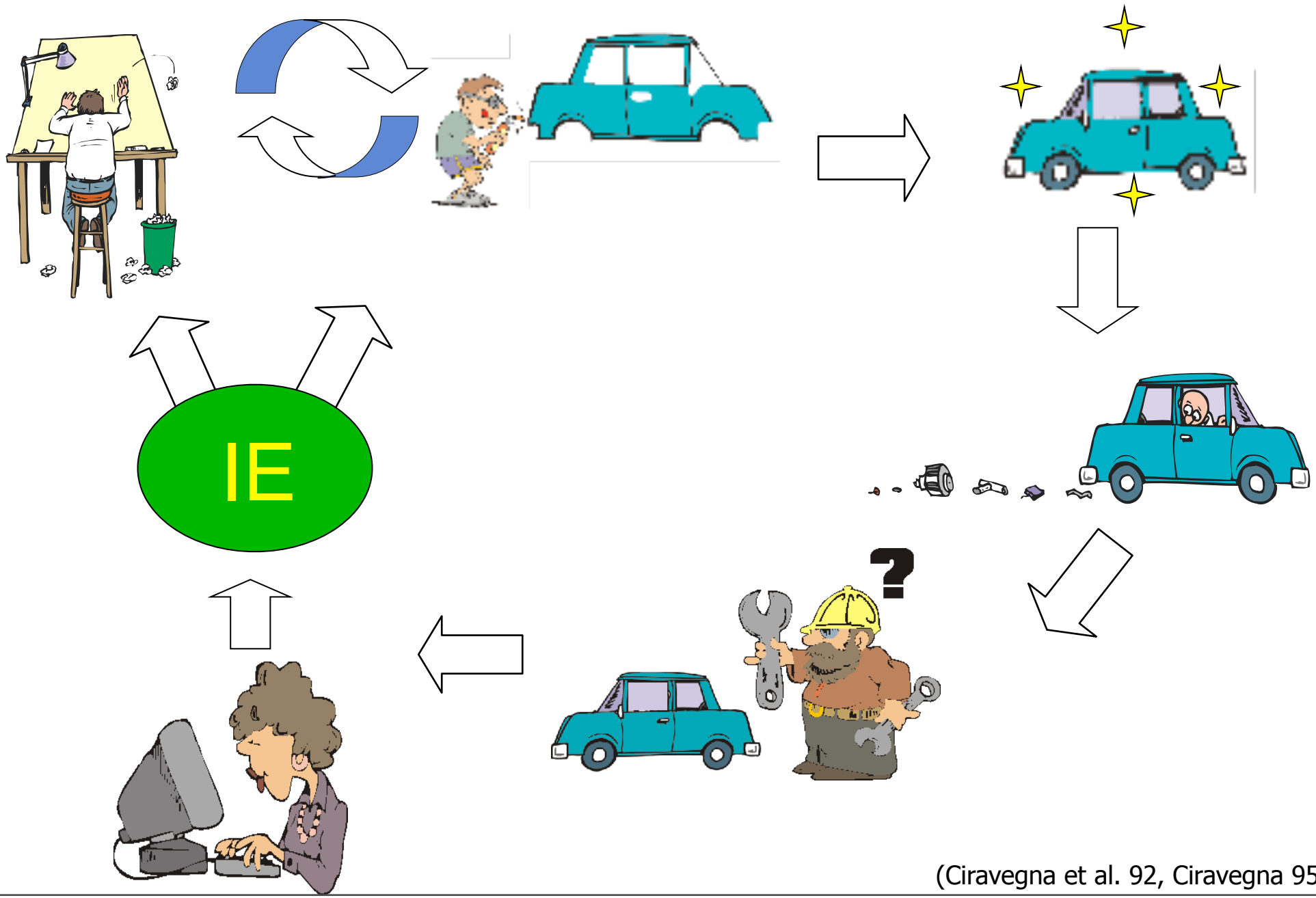
FIRST DIAGNOSIS: Antonioli 24/06/89: vedere scheda 0014/89.

DIAGNOSIS: Bianchi 25/06/89: Anomalia causata da ossidazione con conseguente bloccaggio innesto alberino scorrimento, e mancata chiusura contatti elettromagnete. Il particolare è stato inviato ai laboratori per ulteriori controlli.

Giorgioni 28/06/89 l'ossidazione e' stata causata dall'utilizzo di materiale non idoneo alle prescrizioni.

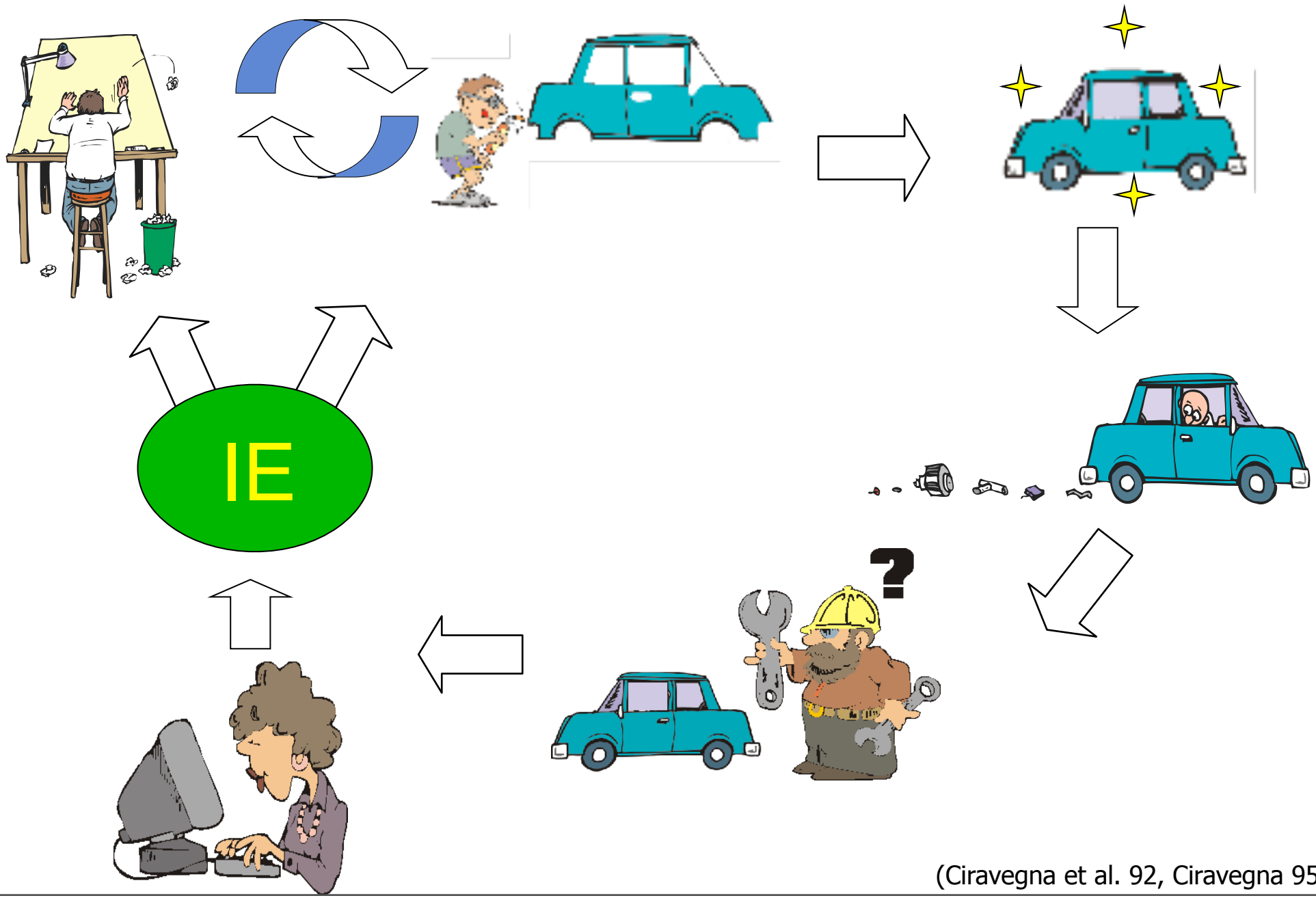
(Ciravegna et al. 92, Ciravegna 95)

Knowledge Management using IE



(Ciravegna et al. 92, Ciravegna 95)

Knowledge Management using IE



(Ciravegna et al. 92, Ciravegna 95)

Knowledge Management using IE

MAIN FAULT: NON-FUNCTIONAL (COD. A124)

Part: starter motor (cod: 0129AIX2)

CAUSED BY: FAILURE TO CLOSE (COD. A156)

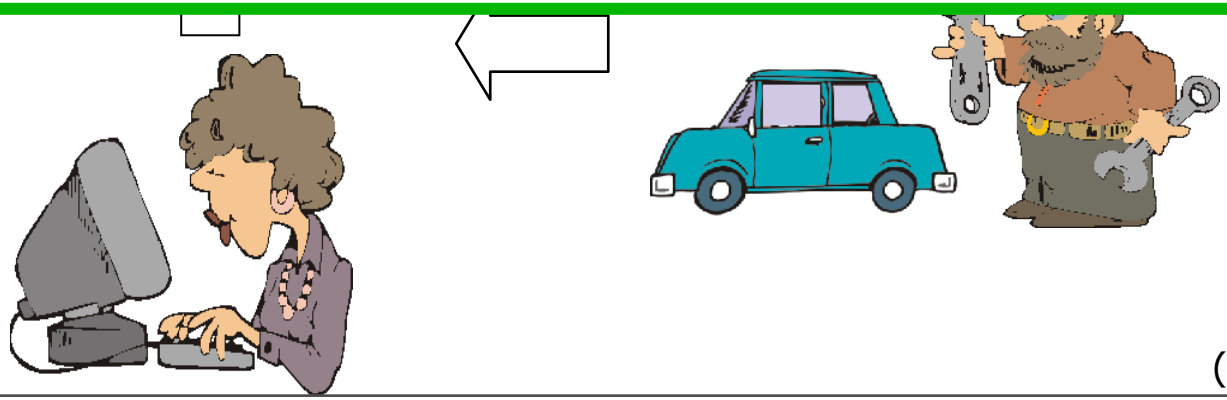
Part: electromagnetic contacts starter motor (cod 0129OOT9)

CAUSED BY: BLOCKAGE (COD A345)

Part: starter drive pinion (cod. 0129OOT9)

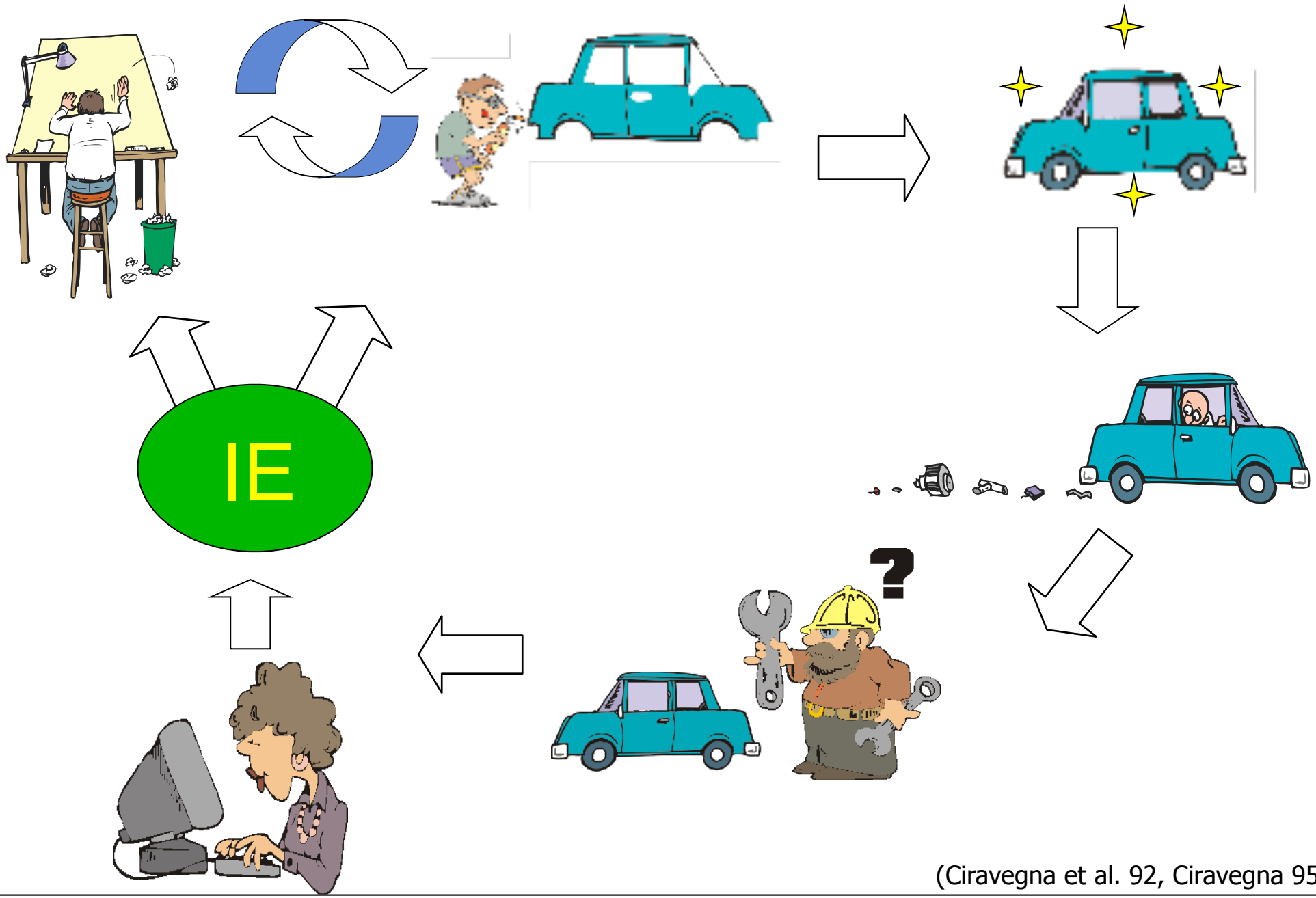
CAUSED BY: OXIDATION (COD A567)

Part: starter drive pinion (cod. 0129OOT9)



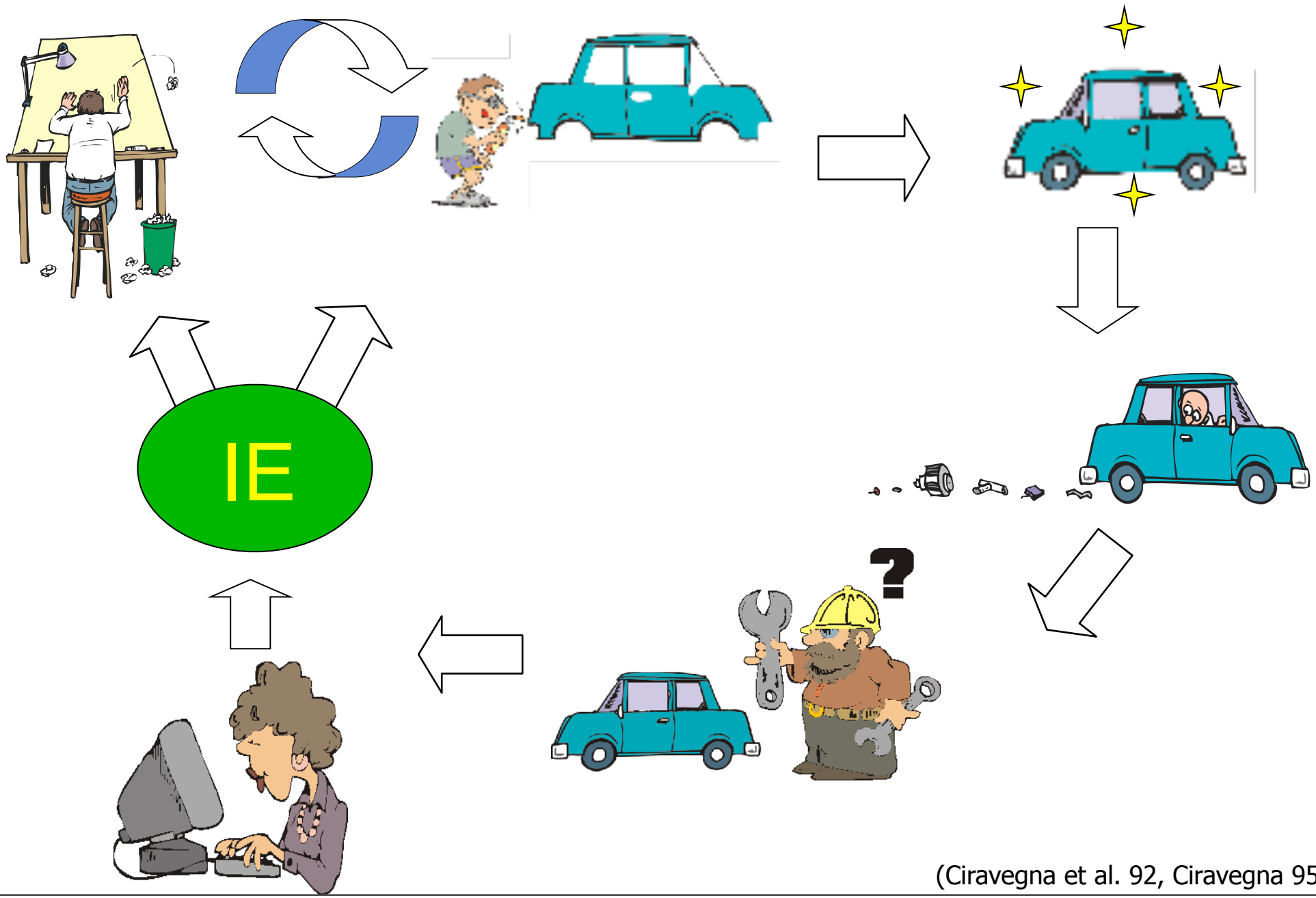
(Ciravegna et al. 92, Ciravegna 95)

Knowledge Management using IE



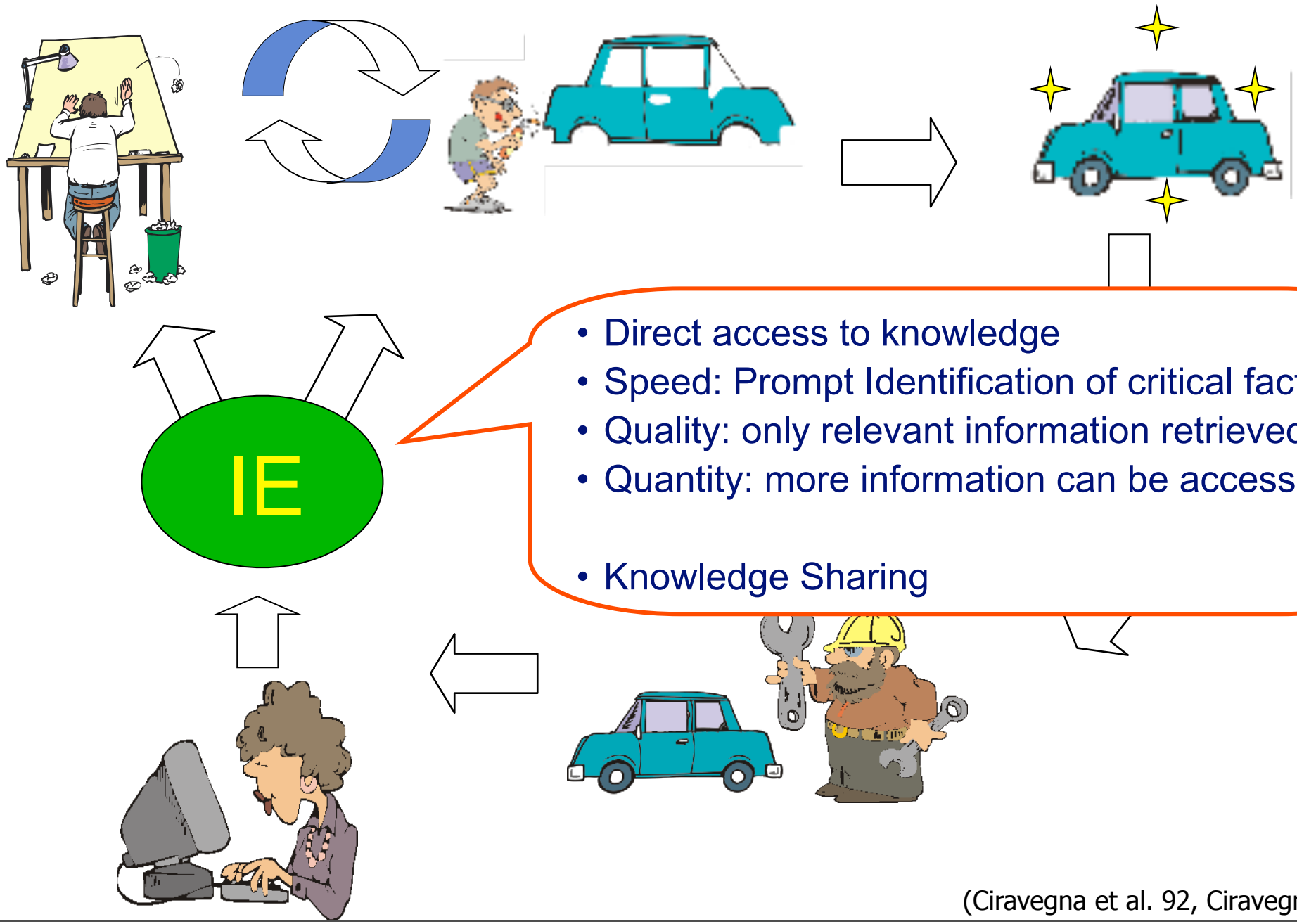
(Ciravegna et al. 92, Ciravegna 95)

Knowledge Management using IE



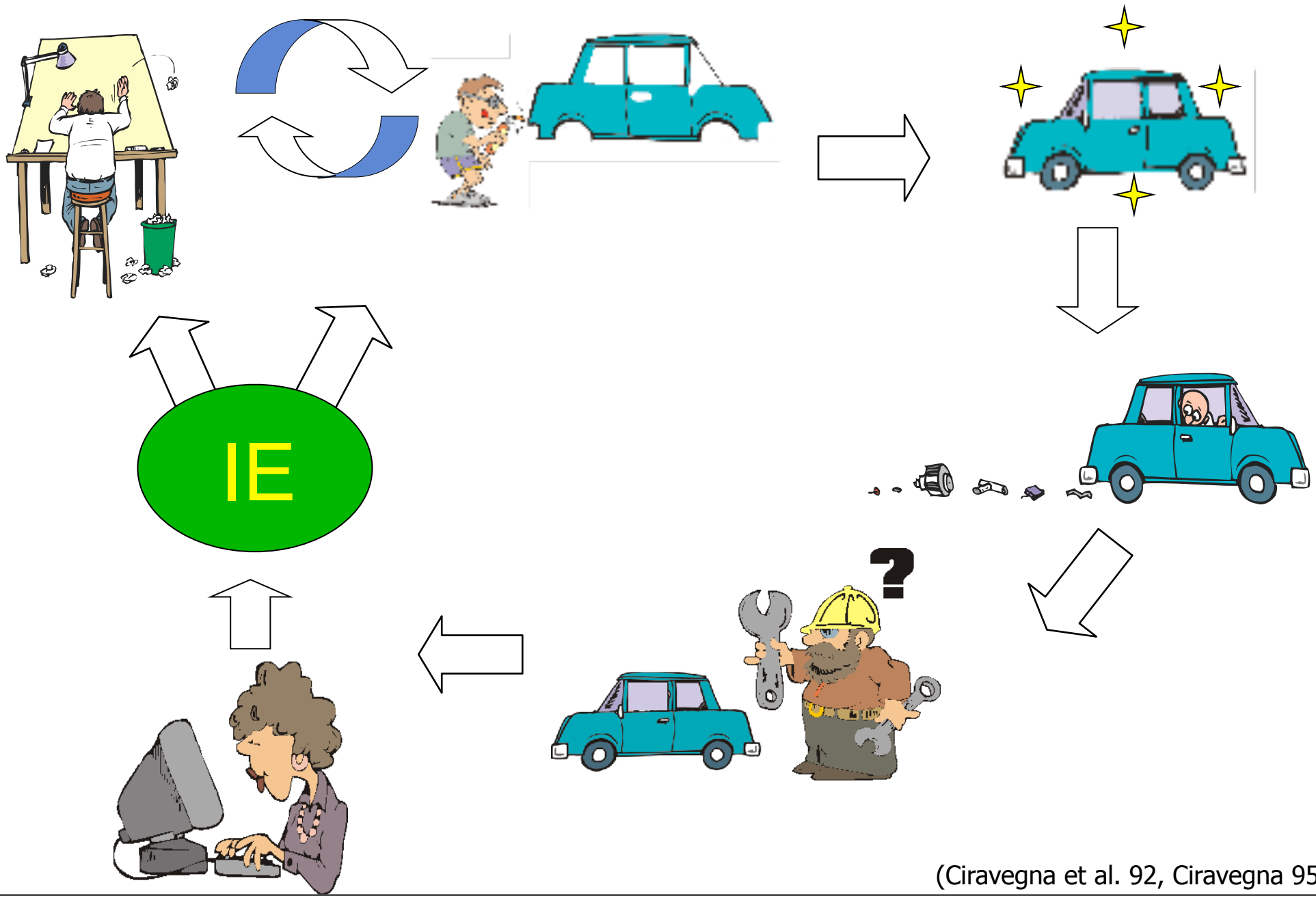
(Ciravegna et al. 92, Ciravegna 95)

Knowledge Management using IE



(Ciravegna et al. 92, Ciravegna 95)

Knowledge Management using IE



(Ciravegna et al. 92, Ciravegna 95)

Jet engine example

- a jet engine can produce ~1Gbyte of vibration data per hour of flight;
  - if irregularities are found, part of the data can be stored
  - reports can be written (event reports)
  - pictures can be taken

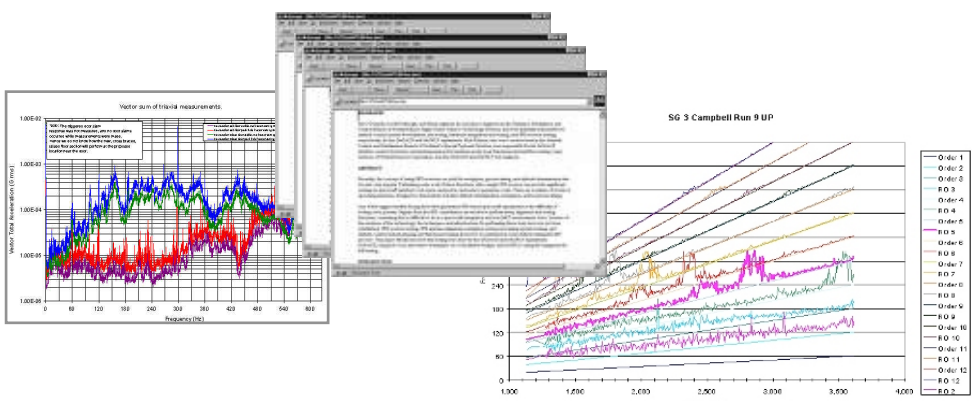


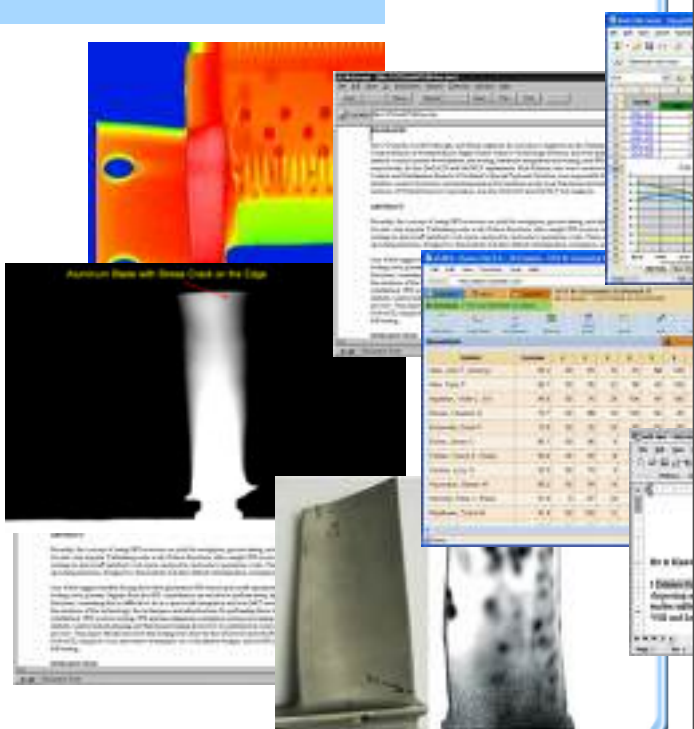
image © www.rolls-royce.com

© Fabio Ciravegna, University of Sheffield

Jet engine example (3)

When engine is serviced (e.g. overhaul)

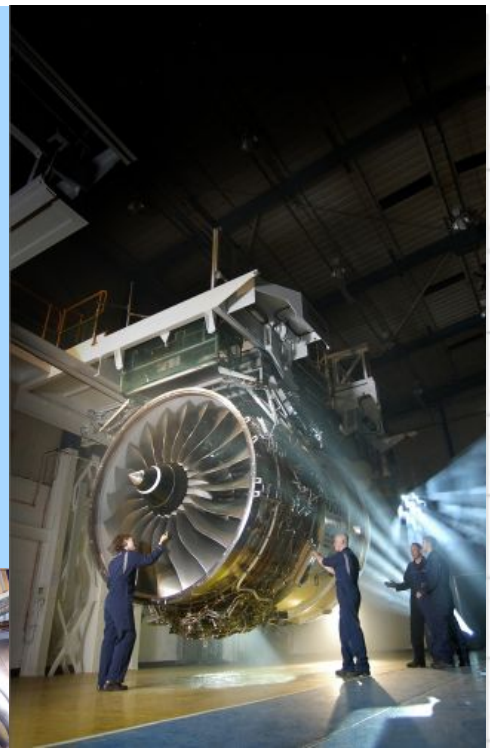
- financial information is produced.
- if issues are found,
  - pictures are taken
  - reports are written
  - engine is tested

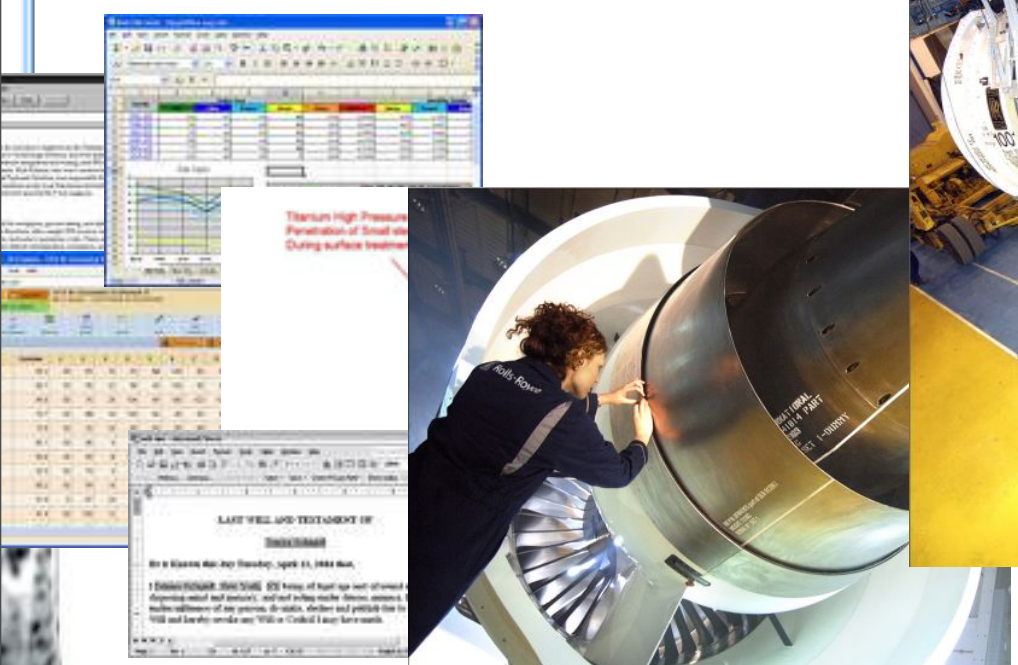


© Fabio Ciravegna, University of Sheffield

Jet engine example (4)

- If problem is recurring (or suspected so)
  - a problem resolution group is established
    - existing evidence is retrieved
    - further evidence is collected
    - a learned lesson is generated
    - same problem is investigated across models





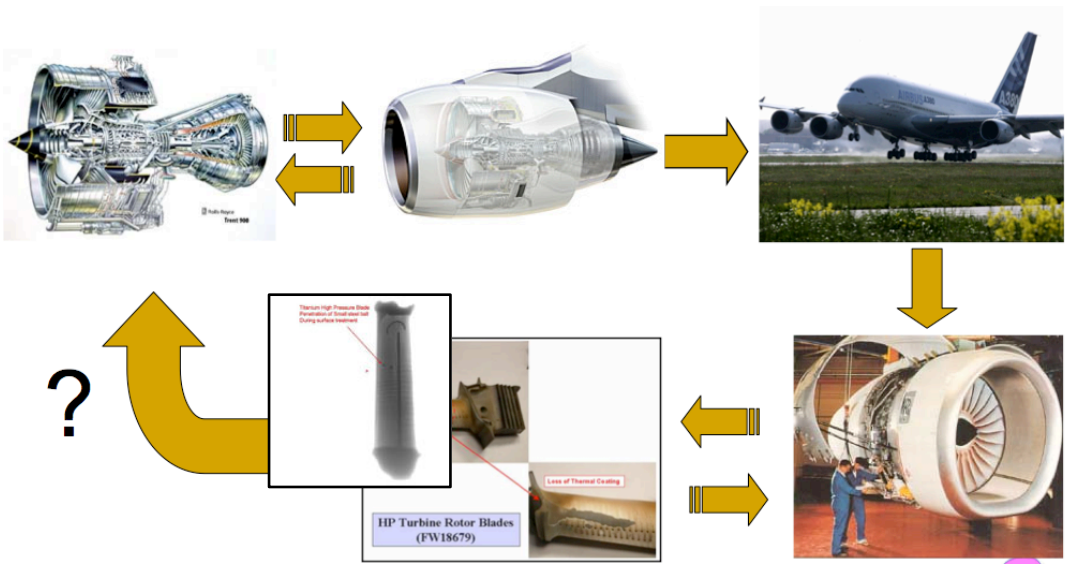
images © www.rolls-royce.com

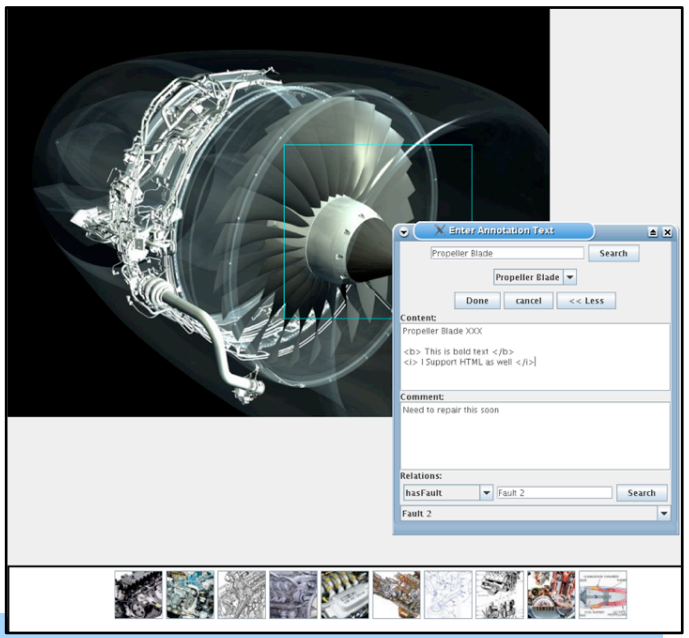
Document Type
AROC proforma
AROC results
Development
EHM data
Emails
ONWING emails
Images
Lab findings
Monitoring Requir
Presentations
Procedures
RCP
Risk Assessment
Solution Reports
Technical Reports
TS&O Reports

© Fabio Ciravegna, University of Sheffield

Jet engine example (4)

- Lifecycle "folder" will easily sum up to several Terabytes
- Folder will contain highly interrelated information stored in different media





- Goal for Knowledge Management:
- Making information available independently from
  - Data format (structured/unstructured)
  - The archive
- Making it available for automatic processing
- Making it easily accessible and manageable despite its size

© Fabio Ciravegna, University of Sheffield

Knowledge Management 2010+

- Organisation archives are following a Web-like trend
  - Massive shift towards multi- and cross-media
    - Text, Images, Data
    - Videos

An image is worth 1,000 words
- But it is very difficult to index

  - Large scale

in 2007: 4 billion cameras

    - Dramatic reduction of memory and storage cost
    - Increase in speed and capacity
  - Number and size of distributed archives of information
    - Large organisations' Intranets as mini webs
      - Thousands of computers
      - Hundreds of repositories
      - Hundreds of millions of documents

New technical information doubles every 2 years:
•Expected to double every 72 hours by 2010

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Issues for IE

- Ontologies can
  - Be large
  - Change frequently
  - Be distributed
    - Multiple ontologies for same domain
      - To cover several point of views
- Corpora can be large in terms of
  - Size
    - Several hundreds of thousands of documents
  - Number
    - Several dozens or hundreds of documents

Here goes the discuss on large scale

Changes in ontology

Changes in corpora

Etc.

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Requirements for IE in KM

- Ontologies can
  - Be large
    - Easy porting of IE system to several hundreds of classes should be possible
      - Who is going to annotate the documents for training?
  - Change frequently
    - Modifications to IE system should be possible every time there is a modification
      - Who is going to re-annotate the documents?
  - Be distributed
    - Multiple ontologies for same domain
      - To cover several points of view
    - Who is going to annotate the documents with the different points of view?

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Requirements for IE (ctd)

- Corpora can be large in terms of
  - Size
    - Several hundreds of thousands of documents
    - IE systems must be efficient
  - Number
    - Several dozens or hundreds of documents
    - Who is going to annotate the different corpora for each of the several ontologies?

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Originality of Information

- The data worldwide is
  - 25% original; 75% replicated
  - 25% from the workplace; 75% not
  - 95% unstructured and growing

This presentation is made with 75% of recycled material

Source: Michael Brodie, Verizon Inc. Sept 2007

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Some Considerations

- Most companies work in one specific domain
  - Documents produced by a jet engine company will be about jet engines
- Different points of view (ontologies)
  - Will still concern the same domain
    - Just slightly differently interpreted
      - E.g. Design view Vs manufacturing view
- Most resources similar or compatible
- Not all documents are equally important
  - A document is important if
    - It contains relevant information AND/OR
    - If people re-use it

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

A Partial Solution

- Porting to a new corpus can be seen as a process of adaptation of resources already trained in the same domain
  - Rather than a new training exercise where we start from scratch
  - i.e. We start from a system trained on a similar corpus in the same domain as a starting point of the new training exercise
- Process:
  - Annotate documents with trained system
  - Ask users to correct annotations
  - Retrain system

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

What Advantages?

- Correcting annotations is an easier task from a cognitive point of view
  - [Ciravegna et al. 02] showed that correcting annotations can reduce annotation time by 80% even in a short exercise
- Preliminary imprecise annotation can even be useful to users in desperate need of information
  - Annotation can be corrected by users as part of the information seeking process
    - With a suitable interface that does not get in the way of their job

Fabio Ciravegna, Alexiei Dingli, Daniela Petrelli and Yorick Wilks:
User-System Cooperation in Document Annotation based on Information Extraction
in Asuncion Gomez-Perez, V. Richard Benjamins (eds.): Knowledge Engineering and Knowledge Management (Ontologies and the Semantic Web), Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), 1-4 October 2002 - Siguenza (Spain), Lecture Notes in Artificial Intelligence 2473, Springer Verlag

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Issues

- (Confirmed) user annotations must be treated differently from unconfirmed system annotations
  - Training must be able to take into consideration uncertainty in annotations
    - 100% correct in user
    - Less so in case of system
- Training system must be modified to cope with this issue

© Fabio Ciravegna, University of Sheffield

Pervasive annotation

- Pervasive annotation of relevant documents
  - the more relevant documents are, the more they will be accessed reused,
    - the more they will be annotated as a side effect.
- The more they are annotated,
  - the more they become retrievable
    - (and hence considered for reuse).
  - As different users performing different tasks can reuse the same documents
    - relevant documents will be annotated with multiple ontologies relevant to different tasks.
- This satisfies our requirements for
  - high quality annotation for the most relevant documents
  - pervasive annotation using different ontologies

Pervasive (ctd)

- effortless and pervasive process, in which the quality of annotations depend directly on the usefulness of the document itself,
  - most popular documents receive most careful annotations.
- To sum up, the annotation as a side effect of reuse addresses the bootstrapping of annotation and the subsequent learning phase.

Pervasive (ctd)

- pervasive annotation of documents with multiple ontologies also has an important side effect on knowledge acquisition.
  - When the same snippet of text is annotated with two unrelated parts of two separate ontologies, this is a sign that the two ontologies are potentially related. If the situation happens regularly, a knowledge gardener (light admin role) supervising ontology development, can be given suggestions to review and add same-as relations between ontology parts, increasing the interconnections across ontologies.

Risks

- This cannot provide mass penetration into the very large scale of the documents containing the long tail of knowledge.
- Following the well known rule of 80/20, we can expect that only large majority of annotations
  - advancement in terms of knowledge management,
  - the risk is that a sort of local maximum is reached
    - Uknown unknown are missed

Addressing the long tail

- annotations provided on popular docs used to train and extract from the long tail
  - "with enough eyeballs, all bugs are shallow"
- annotation of other documents may be sub-optimal.
  - But at least it will enable discovery of documents so far ignored.
  - If they are relevant,reuse will naturally start and therefore the annotation will improve.
- annotation will naturally improve with the annotation of more documents,
  - hence meeting our requirement that annotation improves with scale:
    - the more an ontology is popular the more documents are annotated, the more
      - the system will enable annotation of legacy documents not yet annotated.

The University Of Sheffield.

Towards Web Scale

EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

IEEE Intelligent Systems, vol. 24, no. 2, pp. 8-12, March/April, 2009

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Moving to Web Scale

- Completely manual approaches are unsuitable to large-scale corpora,
  - error prone and inter-annotator disagreement can cause sensible differences in output quality
- Automated methods work on small or medium-scale corpora (up to hundreds of thousands of documents)
  - generally require precise and consistent identification of text snippets to help train an underlying learner.
  - a (set of) specific uniform corpus(-ora) is (are) expected to be addressed
- But how about the Web?
  - None of the limitations above enable web scale analysis

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Lack of Annotations

- The first lesson of Web-scale learning is to use available large-scale data rather than hoping for annotated data that isn't available
  - Useful semantic relationships can be automatically learned without needing any manually annotated data from
    - the statistics of search queries and the corresponding results or from the accumulated evidence of Web-based text patterns
    - formatted tables

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Memorization

- Memorization is a good policy if you have a lot of training data.
  - Build a huge database of probabilities of short sequences of consecutive words (n-grams) from a corpus of billions or trillions of words
  - Memorizing specific phrases over large scale is more effective than general patterns
  - Use general rules only when they improve translation over memorization
    - E.g dates and numbers

© Fabio Ciravegna, University of Sheffield

Some Results

- It is possible to show how a relational logic and a 100-million-page corpus can answer questions such as
  - "what vegetables help prevent osteoporosis?"
  - by isolating and combining the relational assertions that
    - "kale is high in calcium" and
    - "calcium helps prevent osteoporosis."

S. Schoenmackers, O. Etzioni, and D.S. Weld, "Scaling Textual Inference to the Web," *Proc. 2008 Conf. Empirical Methods in Natural Language Processing* (EMNLP 08), Assoc. for Computational Linguistics, 2008, pp. 79–88.

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Using Tables

- The Web contains hundreds of millions of independently created tables and possibly a similar number of lists that can be transformed into tables
- Researchers reliably extracted 2.5 million distinct schemata from a collection of 150 million tables
  - They
    - represent how different people organize data
    - provide a rich collection of column values, and values that they decided

M.J. Cafarella et al., "WebTables: Exploring the Power of Tables on the Web," *Proc. Very Large Data Base Endowment* (VLDB 08), ACM Press, 2008, pp. 538–549.

© Fabio Ciravegna, University of Sheffield

Query Logs

- It is possible to combine data from multiple tables with other sources, e.g. unstructured Web pages or Web search queries
- From them it is possible to identify classes such as Company and then find hundreds of names of companies
  - Then it is possible to identify their attributes (e.g. CEO, stock prices, etc.)
- It works for thousands of classes and tens of thousands of attributes
  - 90 percent precision over the top 10 attributes per class

M. Paşca, "Organizing and Searching the World Wide Web of Facts. Step Two: Harnessing the Wisdom of the Crowds," *Proc. 16th Int'l World Wide Web Conf.*, ACM Press, 2007, pp. 101–110

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Conclusions

© Fabio Ciravegna, University of Sheffield

Conclusions

- We have seen tasks for IE and some models to perform it
- We have mainly focussed on limited scale and precise domains
- The main requirement for IE is to reduce training requirements in order to enable porting over large scale
  - In number of ontologies
  - In number of corpora
  - In number of documents
- Accuracy is not necessarily a major requirement
  - E.g. Search engines

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

Conclusions (ctd)

- Web Scale learning can provide
  - The big breakthrough for IE use (i.e. To integrate search engine results)
  - An excellent starting point for building resources to be adapted for specific applications in a semi-automated way
    - As mentioned IE can be seen as an exercise to port a system to a similar corpus/domain

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

© Fabio Ciravegna, University of Sheffield

The University Of Sheffield.

100 Years Of Excellence.

© Fabio Ciravegna, University of Sheffield