ΕΜΠ ΔΠΜΣ

Εφαρμοσμένες Μαθηματικές Επιστήμες

Αλγόριθμοι Εξόρυξης Πληροφορίας

Διάλεξη01**Εισαγωγή**

Η πληροφορία είναι ζωτική

Τεχνητή Γονιμοποίηση
- Συλλογή ωαρίων
- Γονιμοποίηση με σπέρμα συντρόφου ή δότη
- Παράγονται αρκετά έμβρυα, κάποιο από αυτά μεταφέρεται στη μήτρα

Το πρόβλημα: επιλογή 'βέλτιστου' (με μεγαλύτερη πιθανότητα επιβίωσης) εμβρίου

Η επιλογή βασίζεται σε 60 καταγεγραμμένα χαρακτηριστικά του εμβρύου (μορφολογία, ωάριο, σπέρμα, λεμφικό θυλάκιο κ.ά.). Ένας εμβρυολόγος αδυνατεί να λάβει υπ' όψιν το σύνολο των χαρακτηριστικών.

Λύση: Αλγόριθμος Εξόρυξης Πληροφορίας

Ερευνητική ομάδα στη Μ. Βρετανία αναζητεί μεθόδους αυτοματοποίησης της διαδικασίας επιλογής, βασισμένη σε εκτενές αρχείο ιστορικών δεδομένων.

Η πληροφορία είναι ζωτική

Κτηνοτροφία
- Παραγωγή γάλακτος και κρέατος
- Τυπικά, το 1/5 των αγελάδων ενός κοπαδιού θανατώνεται κάθε χρόνο

Το πρόβλημα: επιλογή χείριστου συνόλου προς σφαγή

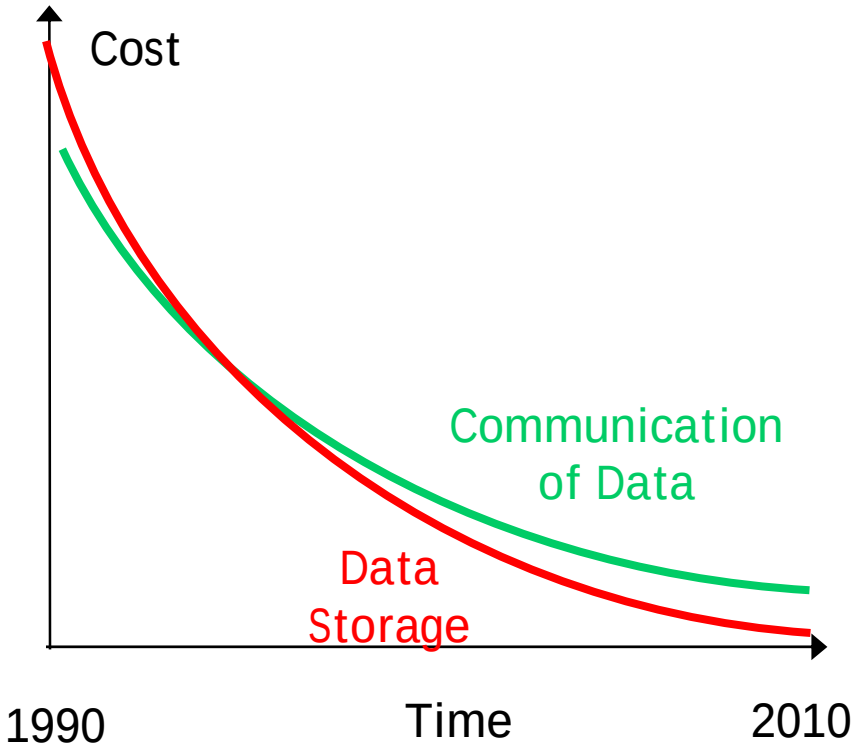
Η απόφαση στηρίζεται σε χαρακτηριστικά παραγωγικότητας, φυλής, γονιμότητας, υγείας, συμπεριφοράς κ.ά.

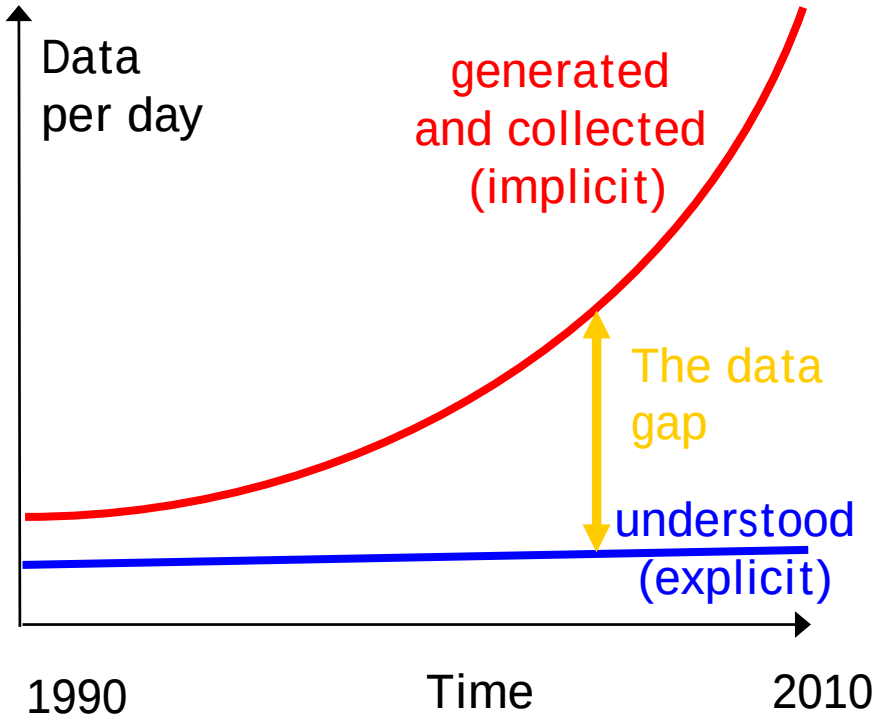
Περίπου 700 χαρακτηριστικά για κάθε 'μονάδα παραγωγής γάλακτος' συλλέγονται από εταιρεία στη Νέα Ζηλανδία, η οποία διατηρεί αρχείο για εκατομμύρια 'μονάδες'.

Λύση: Αλγόριθμος Εξόρυξης Πληροφορίας

Επιχειρείται η εξακρίβωση των κανόνων εκείνων που χρησιμοποιούνται από επιτυχημένους κτηνοτρόφους, ώστε να διαδοθεί η γνώση και εμπειρία τους.

Τα δεδομένα αφθονούν





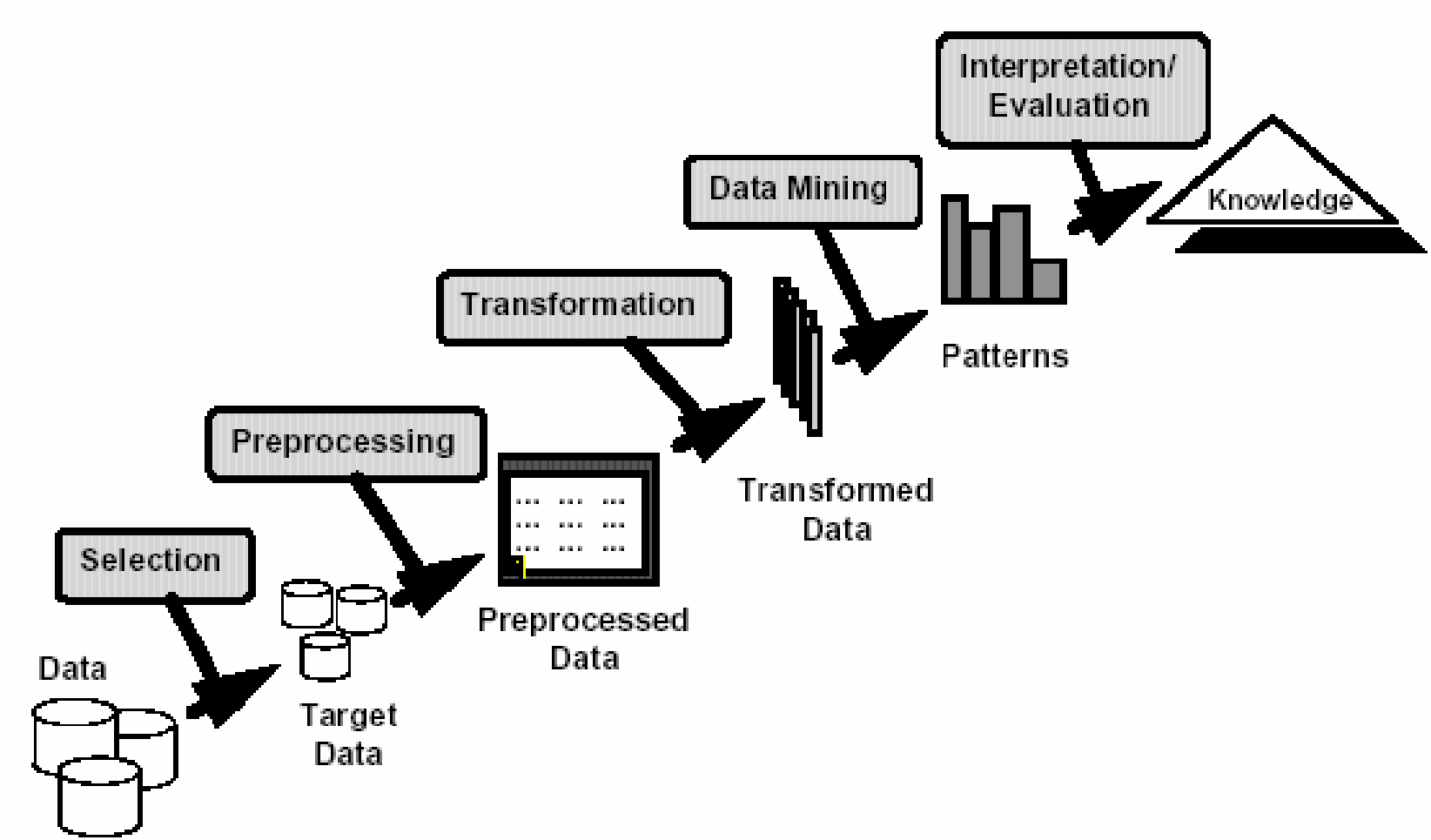
Malthus Law of Information:

- Το νέο πληροφοριακό περιεχόμενο διπλασιάζεται κάθε χρόνο
- Ο χρόνος που δαπανάται για την κατανάλωση πληροφοριών παραμένει σταθερός

Τα δεδομένα αφθονούν

- Μόνο ένα μικρό ποσοστό (5-10%) των συλλεγόμενων δεδομένων τυγχάνει ανάλυσης
- Μία τυπική επιχειρησιακή βάση δεδομένων σήμερα περιέχει συχνά μεγάλο αριθμό εγγραφών (10^8-10^{12}) δεδομένων πολλών διαστάσεων (10-10^4 μεταβλητές)
- Τελικά: "*We are drowning in data, but starving for knowledge!*"
- Πώς μπορούν να εξερευνηθούν εκατομμύρια εγγραφών εκατοντάδων μεταβλητών, ώστε να ανακαλυφθούν πρότυπα (patterns)?

Από τα δεδομένα στην πληροφορία και τη γνώση

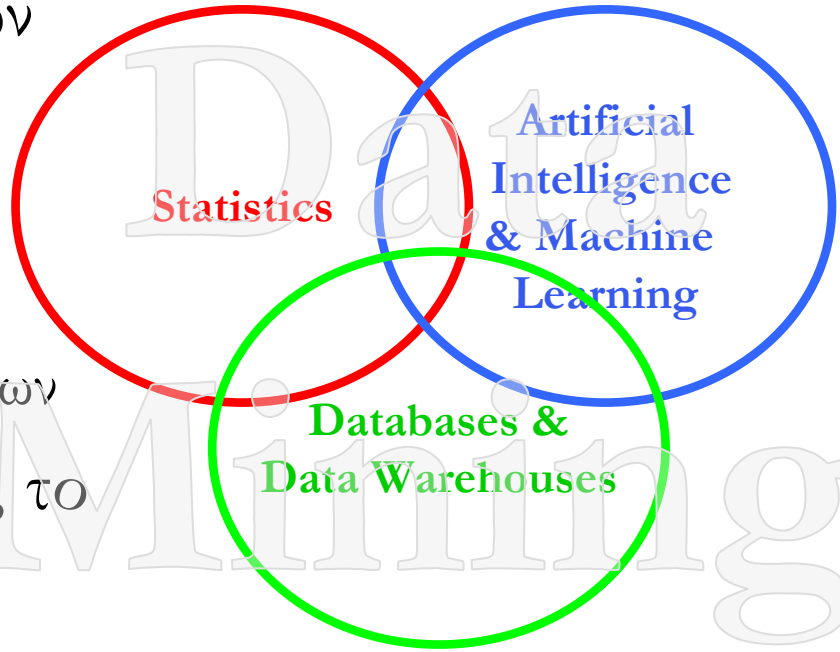


Εξόρυξη πληροφορίας και γνώσης από δεδομένα

- *Data mining*: όρος που χρησιμοποιείται – λανθασμένα– για να περιγράψει το σύνολο της διαδικασίας εξόρυξης γνώσης από βάσεις δεδομένων (*Knowledge Discovery in Databases*)
- Ορισμός: *The nontrivial extraction of implicit, previously unknown, and potentially useful information from data.*
- Εναλλακτικά: *Statistics at scale, speed and simplicity.*
- Στο εξής: *Εξόρυξη Πληροφορίας / Γνώσης από Δεδομένα ≡ Data Mining*

Αφετηρία

- Το ερευνητικό πεδίο αποτελεί τομή μεθόδων και εργαλείων που πηγάζουν από
  - Στατιστική
  - Μηχανική Μάθηση
  - Βάσεις & αποθήκες δεδομένων
- Αποτελεί σύγχρονη εξέλιξη, το πρώτο σχετικό συνέδριο πραγματοποιήθηκε το 1995.
- Πειραματική επιστήμη!



Περιεχόμενα μαθήματος

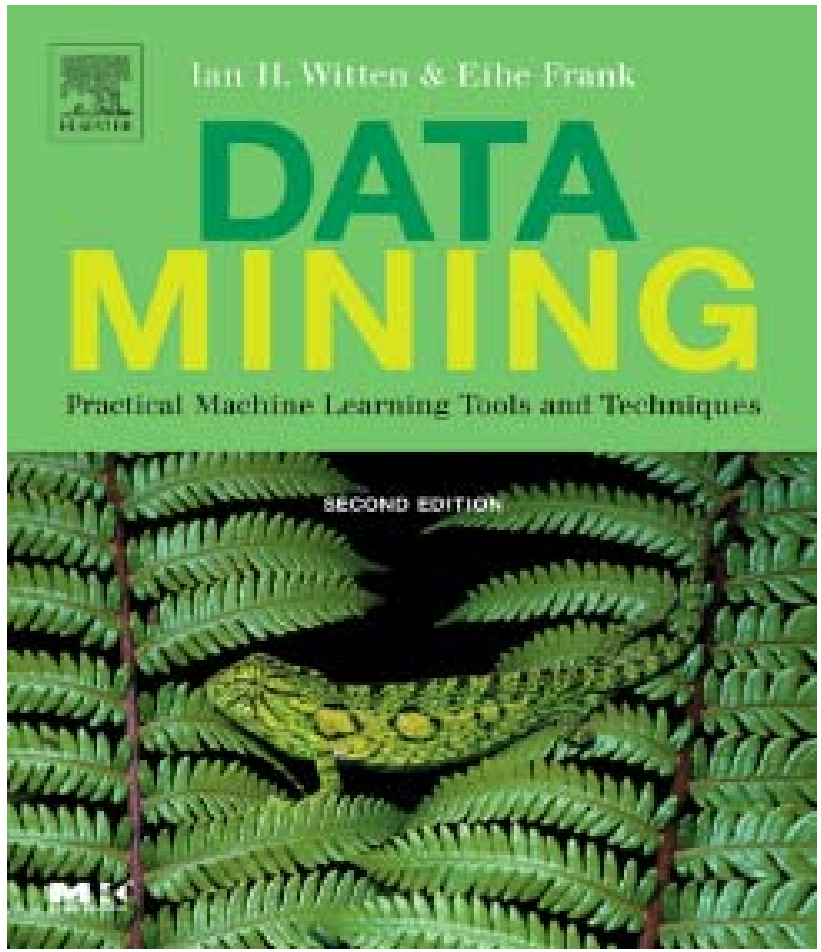
- Διάλεξη01: Εισαγωγή
- Διάλεξη02: Συνιστώσες δεδομένων, οπτικοποίηση & εξερεύνηση
- Διάλεξη03: Προεπεξεργασία & επιλογή δεδομένων
- Διάλεξη04: Απεικόνιση γνώσης, αξιοπιστία & αποτίμηση
- Διάλεξη05: Αλγόριθμοι εκμάθησης (κανόνες ταξινόμησης, δένδρα αποφάσεων)
- Διάλεξη06: Αλγόριθμοι εκμάθησης (αλγόριθμοι ομαδοποίησης, κανόνες συσχέτισης)

Περιεχόμενα μαθήματος

- Διάλεξη07: Αλγόριθμοι εκμάθησης (κανόνες Bayes, νευρωνικά δίκτυα)
- Διάλεξη08: Αλγόριθμοι εκμάθησης (μετα-μαθησιακοί αλγόριθμοι)
- Διάλεξη09: Παρουσίαση Εργασιών
- Διάλεξη10: Εισαγωγή στο Σημασιολογικό Ιστό (Σ.Πόνης)
- Διάλεξη11: Εισαγωγή στις Οντολογίες (Σ.Πόνης)
- Διάλεξη12: Λογισμικό Επιχειρηματικής Ευφυΐας SAP (Σ.Πόνης)

Βιβλίο

Οι διαλέξεις στηρίζονται στο "*Data Mining, Practical Machine Learning Tools and Techniques*", Witten & Frank, Morgan Kaufmann, Ιούνιος 2005.



Λογισμικό

WEKA: Waikato Environment for Knowledge Analysis

http://www.cs.waikato.ac.nz/ml/weka/

blog

http://datamining coursentua simor.blogspot.com/

Πρόβλημα: Ασθενής καταναλωτική πίστη

- Ισχυρά ανταγωνιστική αγορά
- Βάση δεδομένων πελατών με χαρακτηριστικά επιλογών και προφίλ τους
- Ανάλυση προτύπων συμπεριφοράς παλαιών πελατών
- Εντοπισμός κρίσιμων διακριτών χαρακτηριστικών πιστών ή πρώην πελατών
- Ανάδειξη πελατών υψηλής πιθανότητας διακοπής συνεργασίας
- Ειδικός χειρισμός συγκεκριμένων ομάδων πελατών, υπερβολικά κοστοβόρος για εφαρμογή του στο σύνολο των πελατών
- Άλλες εφαρμογές…
- *"In today's highly competitive, customer-centered, service-oriented economy, data is the raw material that fuels business growth—if only it can be mined."*

Πρότυπα

- Επομένως αναζητούνται αλγόριθμοι εντοπισμού προτύπων (patterns) και κανονικοτήτων σε δεδομένα
- Το πρόβλημα είναι κάθε άλλο παρά καινούργιο, ωστόσο σήμερα
  - Τα δεδομένα είναι αποθηκευμένα σε ηλεκτρονική μορφή
  - Ο όγκος δεδομένων και επομένως ο αριθμός πιθανός προτύπων είναι τεράστιος
  - Η αναζήτηση είναι (ημι)αυτοματοποιημένη
- Ισχυρά πρότυπα ⟶ αξιόπιστες προβλέψεις
  - Πρόβλημα 1: τα περισσότερα πρότυπα είναι χαμηλού βαθμού ενδιαφέροντος
  - Πρόβλημα 2: τα πρότυπα είναι πιθανόν ανακριβή ή πλαστά
  - Πρόβλημα 3: τα δεδομένα είναι διαστρεβλωμένα ή ελλιπή

Περιγραφή προτύπων

- Περιγραφή προτύπων
  - Μαύρο κουτί: μη κατανοητοί μηχανισμοί
  - Διαφανές κουτί: αποκαλύπτει τη δομή του προτύπου → δομική περιγραφή
- Οι δομικές (structural) περιγραφές αναπαριστούν τα πρότυπα με σαφώς ορισμένο (ρητό, explicit) τρόπο, με σκοπό την
  - Πρόβλεψη
  - Κατανόηση και επεξήγηση πρόβλεψης
- Θα πραγματευτούμε την εύρεση και περιγραφή δομικών προτύπων σε δεδομένα με τεχνικές που ανήκουν στο πεδίο της *Μηχανικής Μάθησης* (*Machine Learning*)

Περιγραφή δομικών προτύπων

- Παράδειγμα: κανόνες *if-then* για τη σύσταση περί φακών επαφής

Age	Spectacle prescription	Astigmatism	Tear production rate	**Recommended lenses**
Young	Myope	No	Reduced	**None**
Young	Hypermetrope	No	Normal	**Soft**
Pre-presbyopic	Hypermetrope	No	Reduced	**None**
Presbyopic	Myope	Yes	Normal	**Hard**
…	…	…	…	**…**

```
If tear production rate = reduced
   then recommendation = none
Otherwise, if age = young and astigmatic = no
   then recommendation = soft
```

Παραδείγματα

- Οι διαλέξεις αφορούν τη (μηχανική) μάθηση μέσω παραδειγμάτων, επομένως δε θα μπορούσαν παρά να περιέχουν πλήθος παραδειγμάτων.
- Χρησιμοποιούνται διάφορα σύνολα δεδομένων (datasets) από ποικιλία πεδίων αναφοράς
- Η έκταση των παραδειγμάτων είναι μη ρεαλιστική
- Ωστόσο είναι ικανή για τη λεπτομερειακή μελέτη και κατανόηση των αλγορίθμων

Πρόβλημα Καιρού

- Συνθήκες διεξαγωγής ενός ορισμένου παιχνιδιού

Outlook	Temperature	Humidity	Windy	**Play**
Sunny	Hot	High	False	**No**
Sunny	Hot	High	True	**No**
Overcast	Hot	High	False	**Yes**
Rainy	Mild	Normal	False	**Yes**
…	…	…	…	**…**

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes
```

Κανόνες Ταξινόμησης & Συσχέτισης

- Κανόνες Ταξινόμησης (Classification rules)
  - **`If outlook = sunny and humidity = high then play = no`**
  - Προβλέπουν την κατηγορία στην οποία ανήκει το συγκεκριμένο **παράδειγμα** (**instance**, **example**), πχ **`play/no play`**.
- Κανόνες Συσχέτισης (Association rules)
  - **`If temperature = cool then humidity = normal`**
  - **`If windy = false and play = no then outlook = sunny and humidity = high`**
  - Συσχετίζουν τις τιμές των διάφορων **χαρακτηριστικών** (**attributes**) μεταξύ τους

Πρόβλημα Καιρού με μικτά χαρακτηριστικά

- Τα χαρακτηριστικά λαμβάνουν πιθανόν αριθμητικές τιμές

Outlook	Temperature	Humidity	Windy	**Play**
Sunny	85	85	False	**No**
Sunny	80	90	True	**No**
Overcast	83	86	False	**Yes**
Rainy	75	80	False	**Yes**
…	…	…	…	**…**

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```

Πρόβλημα Φακών Επαφής

Age	Spectacle prescription	Astigmatism	Tear production rate	**Recommended lenses**
Young	Myope	No	Reduced	**None**
Young	Myope	No	Normal	**Soft**
Young	Myope	Yes	Reduced	**None**
Young	Myope	Yes	Normal	**Hard**
Young	Hypermetrope	No	Reduced	**None**
Young	Hypermetrope	No	Normal	**Soft**
Young	Hypermetrope	Yes	Reduced	**None**
Young	Hypermetrope	Yes	Normal	**Hard**
Pre-presbyopic	Myope	No	Reduced	**None**
Pre-presbyopic	Myope	No	Normal	**Soft**
Pre-presbyopic	Myope	Yes	Reduced	**None**
Pre-presbyopic	Myope	Yes	Normal	**Hard**
Pre-presbyopic	Hypermetrope	No	Reduced	**None**
Pre-presbyopic	Hypermetrope	No	Normal	**Soft**
Pre-presbyopic	Hypermetrope	Yes	Reduced	**None**
Pre-presbyopic	Hypermetrope	Yes	Normal	**None**
Presbyopic	Myope	No	Reduced	**None**
Presbyopic	Myope	No	Normal	**None**
Presbyopic	Myope	Yes	Reduced	**None**
Presbyopic	Myope	Yes	Normal	**Hard**
Presbyopic	Hypermetrope	No	Reduced	**None**
Presbyopic	Hypermetrope	No	Normal	**Soft**
Presbyopic	Hypermetrope	Yes	Reduced	**None**
Presbyopic	Hypermetrope	Yes	Normal	**None**

Σύνολο κανόνων

If tear production rate = reduced then recommendation = none

If age = young and astigmatic = no
and tear production rate = normal then recommendation = soft

If age = pre-presbyopic and astigmatic = no
and tear production rate = normal then recommendation = soft

If age = presbyopic and spectacle prescription = myope
and astigmatic = no then recommendation = none

If spectacle prescription = hypermetrope and astigmatic = no
and tear production rate = normal then recommendation = soft

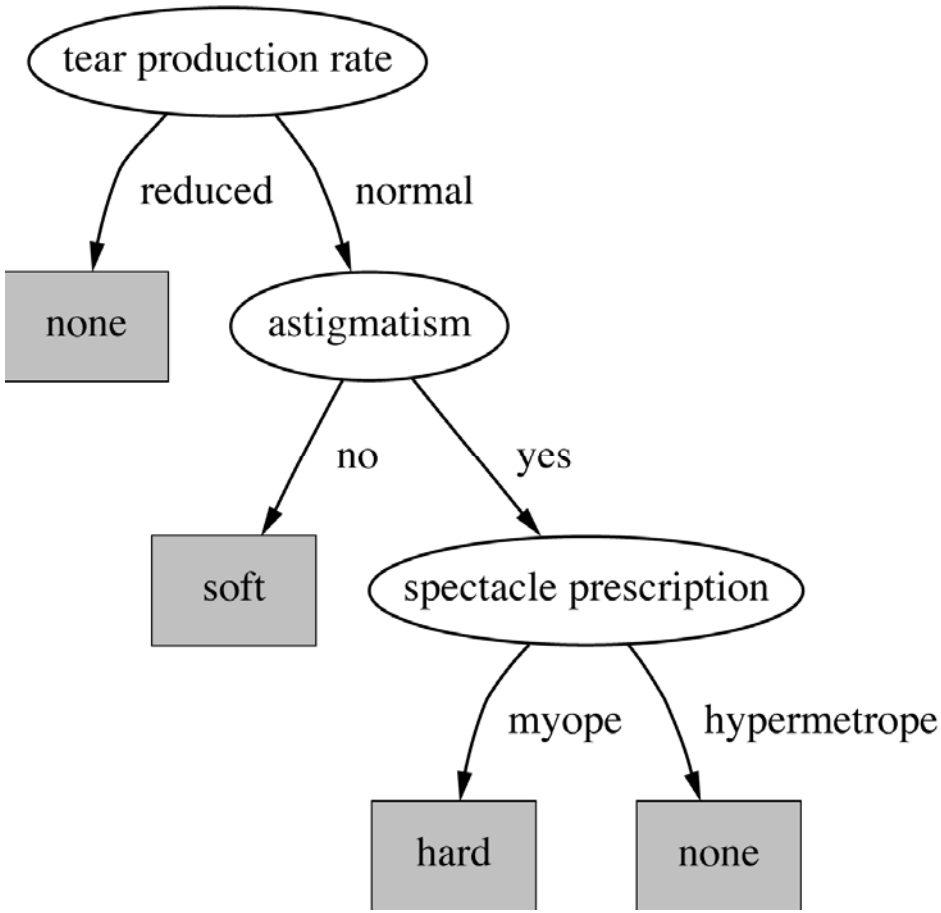
If spectacle prescription = myope and astigmatic = yes
and tear production rate = normal then recommendation = hard

If age young and astigmatic = yes
and tear production rate = normal then recommendation = hard

If age = pre-presbyopic
and spectacle prescription = hypermetrope
and astigmatic = yes then recommendation = none

If age = presbyopic and spectacle prescription = hypermetrope
and astigmatic = yes then recommendation = none

Δένδρο κανόνων



Πρόβλημα ταξινόμησης λουλουδιών Ίρις

	Sepal length	Sepal width	Petal length	Petal width	**Type**
1	5.1	3.5	1.4	0.2	**Iris setosa**
2	4.9	3.0	1.4	0.2	**Iris setosa**
…					
51	7.0	3.2	4.7	1.4	**Iris versicolor**
52	6.4	3.2	4.5	1.5	**Iris versicolor**
…					
101	6.3	3.3	6.0	2.5	**Iris virginica**
102	5.8	2.7	5.1	1.9	**Iris virginica**
…					

```
If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...
```

Πρόβλημα πρόβλεψης απόδοσης CPU

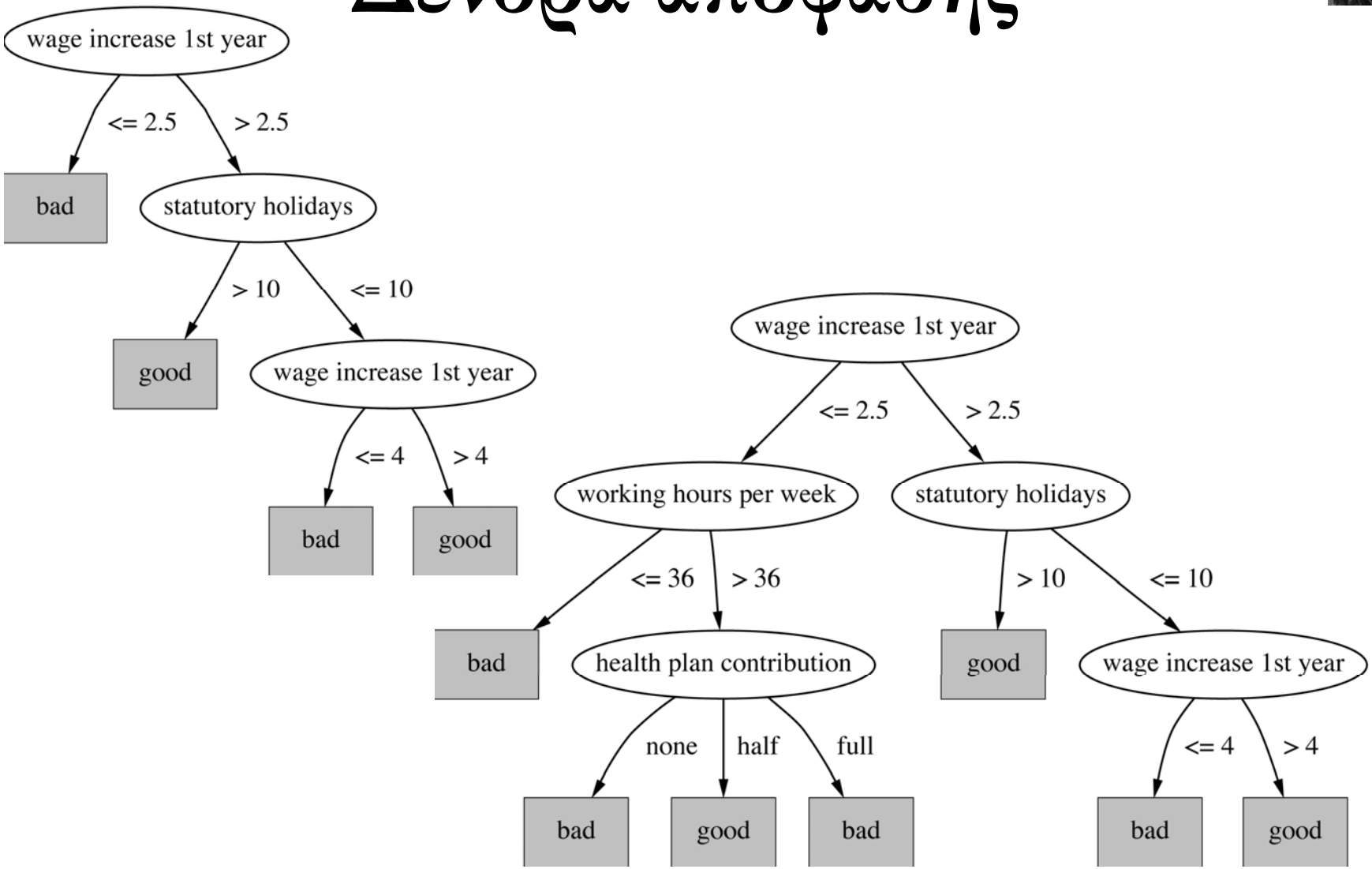
	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		**Performance**
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	**PRP**
1	125	256	6000	256	16	128	**198**
2	29	8000	32000	32	8	32	**269**
…							
208	480	512	8000	32	0	0	**67**
209	480	1000	4000	0	0	0	**45**

```
PRP =  -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
       + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

Πρόβλημα εργασιακών διαπραγματεύσεων

Attribute	Type	1	2	3	…	40
Duration	(Number of years)	1	2	3		2
Wage increase first year	Percentage	2%	4%	4.3%		4.5
Wage increase second year	Percentage	?	5%	4.4%		4.0
Wage increase third year	Percentage	?	?	?		?
Cost of living adjustment	{none,tcf,tc}	none	tcf	?		none
Working hours per week	(Number of hours)	28	35	38		40
Pension	{none,ret-allw, empl-cntr}	none	?	?		?
Standby pay	Percentage	?	13%	?		?
Shift-work supplement	Percentage	?	5%	4%		4
Education allowance	{yes,no}	yes	?	?		?
Statutory holidays	(Number of days)	11	15	12		12
Vacation	{below-avg,avg,gen}	avg	gen	gen		avg
Long-term disability assistance	{yes,no}	no	?	?		yes
Dental plan contribution	{none,half,full}	none	?	full		full
Bereavement assistance	{yes,no}	no	?	?		yes
Health plan contribution	{none,half,full}	none	?	full		half
Acceptability of contract	**{good,bad}**	**bad**	**good**	**good**		**good**

Δένδρα απόφασης



Ταξινόμηση σόγιας

	Attribute	Number of values	Sample value
Environment	Time of occurrence	7	July
	Precipitation	3	Above normal
…			
Seed	Condition	2	Normal
	Mold growth	2	Absent
…			
Fruit	Condition of fruit pods	4	Normal
	Fruit spots	5	?
Leaves	Condition	2	Abnormal
	Leaf spot size	3	?
…			
Stem	Condition	2	Abnormal
	Stem lodging	2	Yes
…			
Roots	Condition	3	Normal
Diagnosis		**19**	**Diaporthe stem canker**

Κανόνες

```
If leaf condition is normal
   and stem condition is abnormal
   and stem cankers is below soil line
   and canker lesion color is brown
then
   diagnosis is rhizoctonia root rot

If leaf malformation is absent
   and stem condition is abnormal
   and stem cankers is below soil line
   and canker lesion color is brown
then
   diagnosis is rhizoctonia root rot
```

Η σημασία της γνώσης πεδίου (domain knowledge):

Ισχύει ότι, εάν "leaf condition is normal", τότε "leaf malformation is absent", επομένως οι δύο κανόνες ταυτίζονται.

Εφαρμογές

- Τα προηγούμενα παραδείγματα είναι εξωπραγματικά απλά και έχουν περισσότερο διδακτικό παρά επιδεικτικό χαρακτήρα.
- Στην πράξη;
- Τυπικές εφαρμογές αλγορίθμων εξόρυξης δεδομένων είναι…

Παροχή δανείου

- Ερωτηματολόγιο για σχετικές οικονομικές και προσωπικές πληροφορίες
- Αποδοχή / απόρριψη αίτησης δανειοδότησης
- Συνήθης στατιστικές τεχνικές καλύπτουν αποτελεσματικά το 90% των περιπτώσεων
- Οι υπόλοιπες ασαφείς περιπτώσεις αξιολογούνται από ειδικούς
- Ωστόσο το 50% των αποδεκτών ασαφών περιπτώσεων δεν αποπληρώνουν
- Λύση: απόρριψη όλων των ασαφών περιπτώσεων
- Υποβέλτιστη, καθώς οι ασαφείς πελάτες είναι σημαντικοί από άποψη ρευστότητας για τον τραπεζικό οργανισμό

Παροχή δανείου

- 1000 παραδείγματα εκπαίδευσης ασαφών περιπτώσεων
- 20 χαρακτηριστικά
  - Ηλικία
  - Οικογενειακή κατάσταση
  - Διάρκεια εργασίας υπό τον ίδιο εργοδότη
  - Διάρκεια συνεργασίας με την τράπεζα
  - Άλλα δάνεια …
- Οι κανόνες που προέκυψαν αποδείχθηκαν σωστοί στο 70% των περιπτώσεων
  - Οι ειδικοί επιτύγχαναν μόνο στο 50%
- Οι κανόνες είναι ίσως χρήσιμοι για την απόκτηση γνώσης και την παροχή εξηγήσεων στους πελάτες

Πωλήσεις & Marketing

- Άμεσο Marketing: οι προσφορές προώθησης προϊόντος είναι συχνά κοστοβόρες και έχουν ένα πολύ χαμηλό –αλλά ιδιαίτερα προσοδοφόρο – ποσοστό απόκρισης
- Οι βάσεις δεδομένων καταναλωτών περιέχουν πλήθος στοιχείων αγοραστικής συμπεριφοράς
- Η άντληση κανόνων από τις βάσεις αυτές έχει πλήθος εφαρμογών στο άμεσο marketing
- Οι στοχευμένες προωθητικές ενέργειες κοστίζουν λιγότερο και αποδίδουν περισσότερα

Πωλήσεις & Marketing

- Ανάλυση καλαθιού αγοράς μέσω εύρεσης κανόνων συσχέτισης για την ανάδειξη ομάδων προϊόντων που συχνά αγοράζονται μαζί
- Τυπικά πραγματοποιείται σε βάσεις δεδομένων super market
- Μπορεί να αναδείξει πχ ότι
  - Οι πελάτες που αγοράζουν μπύρα συχνά προμηθεύονται και πίτσα
  - Κάθε Πέμπτη, οι πελάτες που αγοράζουν μπύρα αγοράζουν επίσης πάνες
- Τέτοιες πληροφορίες έχουν δυνητικά μεγάλη προστιθέμενη αξία
  - Αποδοτική αναδιάταξη προϊόντων, Διαχείριση αποθεμάτων…
- Συχνά τα δεδομένα αγοραστικής συμπεριφοράς που αποκτώνται μέσω προσωπικών εκπτωτικών καρτών είναι μεγαλύτερης αξίας από την παρεχόμενη έκπτωση

Η εξόρυξη γνώσης ως αναζήτηση & γενίκευση

- Επαγωγική μάθηση: εύρεση "αντίληψης" (concept, το αποτέλεσμα της μαθησιακής διαδικασίας) που περιγράφει τα δεδομένα
- Παράδειγμα: σύνολα κανόνων ως περιγραφική γλώσσα
- Εύρεση κανόνων: Το πρόβλημα εύρεσης κανόνων μπορεί να θεωρηθεί ως αναζήτηση σε έναν τεράστιο, πλην όμως πεπερασμένο, χώρο αναζήτησης
  - Καταγραφή όλων των πιθανών συνόλων κανόνων
  - Απόρριψη περιγραφών που δεν ταιριάζουν στα παραδείγματα
  - Επιλογή περιγραφών που εμπεριέχουν την επιθυμητή 'αντίληψη'

Η εξόρυξη γνώσης ως αναζήτηση & γενίκευση

- Έστω το προαναφερθέν πρόβλημα καιρού (διαφάνεια 19)
- Πιθανά σύνολα κανόνων: Άπειρα; **Όχι!**
- Ωστόσο, τεράστιος αριθμός
  - 4 x 4 x 3 x 3 x 2 = 288 πιθανοί κανόνες,
  - Αν κάθε σύνολο περιέχει το πολύ 14 κανόνες, όσα και τα παραδείγματα, προκύπτουν 2,7 x 10^{34} πιθανά σύνολα κανόνων
- Λύση: αλγόριθμοι περικοπής υποψηφίων συνόλων
- Πρακτικά προβλήματα:
  - Περισσότερες από μία λύσεις επιβιώνουν
  - Καμία λύση δεν επιβιώνει
    - Η γλώσσα περιγραφής που επιλέχθηκε είναι ακατάλληλη
    - Θόρυβος στα δεδομένα

Μεροληψία

- Υιοθετώντας την οπτική της εξόρυξης γνώσης ως αναζήτηση & γενίκευση, οι ακόλουθες αποφάσεις ανακύπτουν ως οι πλέον σημαντικές για τα συστήματα μηχανικής μάθησης
  - Γλώσσα περιγραφής πληροφορίας
  - Μέθοδος σάρωσης χώρου αναζήτησης
  - Μέθοδος αποφυγής υπερπροσαρμογής στα δεδομένα εκπαίδευσης
- 'Μεροληψία' (bias) κατά την αναζήτηση
  - Μεροληψία γλώσσας περιγραφής
  - Μεροληψία αναζήτησης
  - Μεροληψία αποφυγής υπερπροσαρμογής

Μεροληψία

- Μεροληψία γλώσσας περιγραφής
  - Είναι η γλώσσα περιγραφής οικουμενική ή θέτει περιορισμούς στο ποια γνώση μπορεί να αποκτηθεί;
  - Η γνώση πεδίου μπορεί να αποκλείσει a priori κάποιες περιγραφές από την αναζήτηση
- Μεροληψία αναζήτησης
  - Η αναζήτηση είναι ευρετική (πχ greedy / beam search)
  - Κατεύθυνση αναζήτησης (από το γενικό στο ειδικό / από το ειδικό στο γενικό)
- Μεροληψία αποφυγής υπερπροσαρμογής
  - Κριτήριο αποτίμησης, ισορροπεί ανάμεσα σε απλότητα και αριθμό λαθών
  - Διάφορες τεχνικές, πχ κλάδεμα προς τα εμπρός / πίσω

Ηθικά ζητήματα

- Ηθικά ζητήματα ανακύπτουν στις πρακτικές εφαρμογές εξόρυξης πληροφορίας από βάσεις δεδομένων
- Η εξόρυξη γνώσης χρησιμοποιείται συχνά για διακρίσεις
  - Για παράδειγμα, απόφανση περί δανειοδότησης, χρησιμοποιώντας δεδομένα όπως φύλο, θρησκεία, εθνικότητα με μη ηθικό ή και παράνομο τρόπο
- Εξαρτάται από την εφαρμογή
  - Για παράδειγμα, η χρήση των ίδιων δεδομένων σε ιατρικές εφαρμογές είναι αποδεκτή
- Ο αποκλεισμός των ευαίσθητων δεδομένων δεν είναι εύκολος
  - Για παράδειγμα, ο ταχυδρομικός κώδικας μπορεί να συσχετίζεται με μία συγκεκριμένη εθνικότητα

Ηθικά ζητήματα

- Ερωτήσεις υψηλής σημαντικότητας
  - Ποιος έχει πρόσβαση στα δεδομένα;
  - Για ποιο σκοπό έχουν συλλεγεί τα δεδομένα;
  - Τι είδους συμπεράσματα μπορούν να εξαχθούν νομίμως από αυτά;
- Πιθανές προειδοποιήσεις πρέπει να συνοδεύουν τα αποτελέσματα
- *Οι αλγόριθμοι εξόρυξης πληροφορίας είναι απλώς ένα εργαλείο, η αξιολόγηση και χρήση των αποτελεσμάτων τους είναι ζήτημα ανθρωπίνων αποφάσεων και όχι μηχανικής μάθησης*

Τέλος

Επόμενη διάλεξη:

Συνιστώσες Δεδομένων, Οπτικοποίηση & Εξερεύνηση