Learning Boolean Queries for Article Quality Filtering

Yin Aphinyanaphongs, M.S.
Constantin Aliferis, M.D. Ph.D.
Department of Biomedical Informatics
Vanderbilt University
MedInfo Symposium, September 8-11, 2004

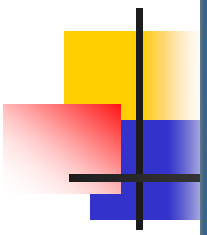
Information Overload

- The pace of research far overcomes the ability of modern health professionals to be up to date about all the recent research developments and current best practices.
- Increasingly, physicians are turning to electronic sources for their information needs.
- The final authority on what constitutes best medical practices and high quality knowledge is provided by the primary sources themselves

Previous Research

- Construction of Boolean queries optimized for sensitivity and specificity.
  - Treatment, Etiology, Prognosis, Diagnosis.
  - “Randomized Controlled Trial” [Publication Type] OR “Drug Therapy” [MeSH Subheading] OR “Therapeutic Use” [MeSH Subheading] OR “Random” [Textword]
- Our own work provided an alternate methodology, instead, using machine learning to identify the content specific, high quality articles.

Haynes, B., et al., *Developing Optimal Search Strategies for Detecting Sound Clinical Studies in MEDLINE.* JAMIA, 1994. **1**(6): p. 447-458.



NCBI

PubMed Clinical Queries

PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | OMIM | Books

About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby

Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central

Privacy Policy

Select from two filters to limit your retrieval. Choose either Clinical Queries or Systematic Reviews. Enter your search topic in the box below and click Go.

◉ Clinical Queries using Research Methodology Filters

This specialized search is intended for clinicians and has built-in search "filters" based largely on Haynes RB et al.. Four study categories are provided, and the emphasis may be more sensitive (i.e., most relevant articles but probably some less relevant ones) or more specific (i.e., mostly relevant articles but probably omiting a few). See filter table for details.

Indicate the category and emphasis below:

Category: ◉ therapy ○ diagnosis ○ etiology ○ prognosis
Emphasis: ○ sensitivity ◉ specificity

○ Systematic Reviews

This feature retrieves systematic reviews and meta-analysis studies for your search topic(s). For more information, see Help. Related sources are also provided.

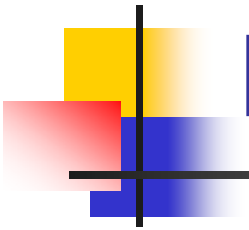
Enter subject search:

[ ] Go Clear

Note: If you want to retrieve everything on a subject area, you should not use this screen. The objective of filtering is to reduce the retrieval to articles that report research conducted with specific methodologies.

Previous Research

- Our own work addresses the problem of returning quality articles by running a suite of powerful classifiers on a suitable corpus. Models perform well with several areas for improvement.
  - Best performing machine learning models are not human-readable.
  - Best machine learning models can not be used in modern Boolean based search engines.



Hypothesis

- Is it possible to automatically construct Boolean queries from a corpus using machine learning techniques such that the Boolean queries have as good classification performance as the SVM models, and are the resulting Boolean queries human-readable, manageable, and simple for use in current search engines?

Gold standard corpus

- The ACP journal club.
- The ACP journal is a meta-publication that routinely reviews over a hundred journals for articles that meet its selection criteria.
- Selected articles are abstracted and cited in a secondary publication.

Purpose and Proc. ACP Journal, 1999. **131**(1): p. A15.

Treatment Selection Criteria

- The treatment criteria -ACP journal club
  - “Random allocation of participants to comparison groups.”
  - “80% follow up of those entering study.”
  - “Outcome of known or probable clinical importance.”

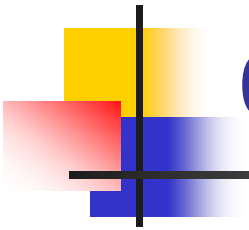
Purpose and Proc. ACP Journal, 1999. **131**(1): p. A15.

Etiology Selection Criteria

- In studies of etiology, good design is:
  - exploration of the relation between exposures and putative clinical outcomes.
  - prospective data collection with clearly identified comparison groups for those at risk for the outcome of interest (in descending order of preference from randomized controlled trial, quasi-randomized controlled trial, nonrandomized controlled trial, cohort studies with case-by-case matching or statistical adjustment to create comparable groups, to nested case-control studies.
  - masking of observers of outcomes to exposures (criterion assumed to be met if outcome is objective, i.e., all-cause mortality, objective test).

Purpose and Proc. ACP Journal, 1999. **131**(1): p. A15.

Steps for automated Boolean query construction.



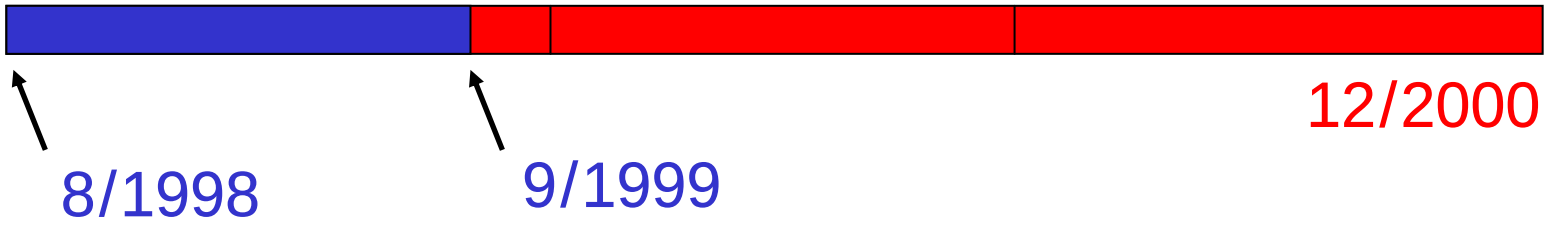
- Corpus Construction.
- Document Representation.
- Study Design
- Apply feature selection algorithms.
- Apply support vector/ decision tree classifiers.
- Evaluate the classifiers.

Select the Journals

- Age and ageing
- AMERICAN JOURNAL OF CARDIOLOGY
- AMERICAN JOURNAL OF EPIDEMIOLOGY
- AMERICAN JOURNAL OF GASTROENTEROLOGY
- AMERICAN JOURNAL OF MEDICINE
- AMERICAN JOURNAL OF PUBLIC HEALTH
- AMERICAN JOURNAL OF RESPIRATORY AND CRITICAL CARE MEDICINE
- ANNALS OF EMERGENCY MEDICINE
- ANNALS OF INTERNAL MEDICINE
- ANNALS OF MEDICINE
- ARCHIVES OF FAMILY MEDICINE
- ARCHIVES OF INTERNAL MEDICINE
- ARCHIVES OF NEUROLOGY
- ARTHRITIS AND RHEUMATISM
- BRITISH MEDICAL JOURNAL
- BRITISH JOURNAL OF GENERAL PRACTICE
- CANADIAN MEDICAL ASSOCIATION JOURNAL
- CANADIAN JOURNAL OF CARDIOLOGY
- CANADIAN JOURNAL OF GASTROENTEROLOGY
- Chest
- Circulation
- CLINICAL AND INVESTIGATIVE MEDICINE
- CRITICAL CARE MEDICINE
- Diabetes Care
- Gastroenterology
- Gut
- Heart
- Hypertension
- J Am Board Fam Pract
- JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
- JOURNAL OF THE AMERICAN GERIATRICS SOCIETY
- JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION
- JOURNAL OF CLINICAL EPIDEMIOLOGY
- JOURNAL OF FAMILY PRACTICE
- JOURNAL OF GENERAL INTERNAL MEDICINE
- JOURNAL OF INFECTIOUS DISEASES
- JOURNAL OF INTERNAL MEDICINE
- JOURNAL OF NEUROLOGY NEUROSURGERY AND PSYCHIATRY
- JOURNAL OF VASCULAR SURGERY
- JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION
- Lancet
- MEDICAL CARE
- MEDICAL JOURNAL OF AUSTRALIA
- NEW ENGLAND JOURNAL OF MEDICINE
- Neurology
- Pain
- Spine
- Stroke
- Thorax

Purpose and Proc. ACP Journal, 1999. **131**(1): p. A15.

Downloading and marking articles in the study period.



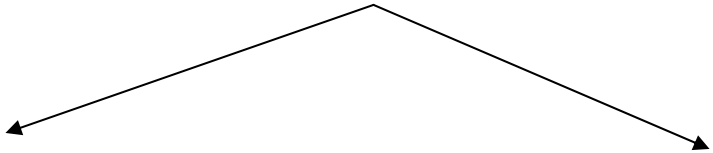
Downloaded all articles from the journals in the study period.

Review ACP Journal from 8/1998 to 12/2000 for treatment and etiology articles that are cited by the ACP.

Aphinyanaphongs, Y. and C.F. Aliferis. *Text Cat Models for Retrieval of High Quality Articles in Internal Medicine*. 2003

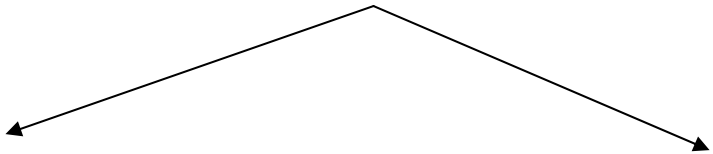
Corpus Composition for treatment and etiology categories.

For the study period from 8/98 – 9/99
15786 original articles



379 high quality Treatment articles

15407 non-high, treatment articles



205 high quality etiology articles

15581 non-high, etiology articles

Aphinyanaphongs, Y. and C.F. Aliferis. *Text Cat Models for Retrieval of High Quality Articles in Internal Medicine*. 2003

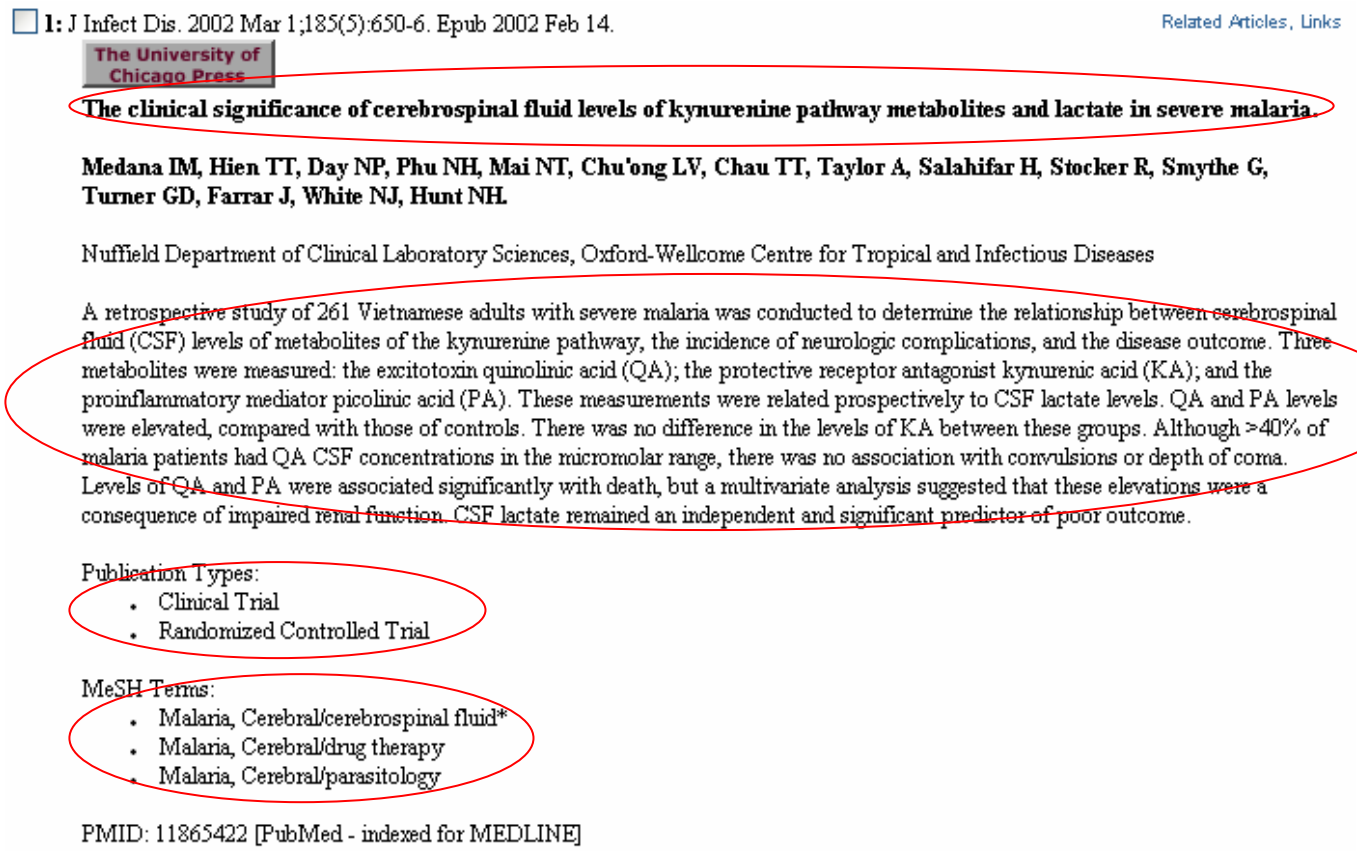
What words to use?

1

2

3

4



1: J Infect Dis. 2002 Mar 1;185(5):650-6. Epub 2002 Feb 14.

Related Articles, Links

The University of Chicago Press

The clinical significance of cerebrospinal fluid levels of kynurenine pathway metabolites and lactate in severe malaria.

Medana IM, Hien TT, Day NP, Phu NH, Mai NT, Chu'ong LV, Chau TT, Taylor A, Salahifar H, Stocker R, Smythe G, Turner GD, Farrar J, White NJ, Hunt NH.

Nuffield Department of Clinical Laboratory Sciences, Oxford-Wellcome Centre for Tropical and Infectious Diseases

A retrospective study of 261 Vietnamese adults with severe malaria was conducted to determine the relationship between cerebrospinal fluid (CSF) levels of metabolites of the kynurenine pathway, the incidence of neurologic complications, and the disease outcome. Three metabolites were measured: the excitotoxin quinolinic acid (QA); the protective receptor antagonist kynurenic acid (KA); and the proinflammatory mediator picolinic acid (PA). These measurements were related prospectively to CSF lactate levels. QA and PA levels were elevated, compared with those of controls. There was no difference in the levels of KA between these groups. Although >40% of malaria patients had QA CSF concentrations in the micromolar range, there was no association with convulsions or depth of coma. Levels of QA and PA were associated significantly with death, but a multivariate analysis suggested that these elevations were a consequence of impaired renal function. CSF lactate remained an independent and significant predictor of poor outcome.

Publication Types:

- Clinical Trial
- Randomized Controlled Trial

MeSH Terms:

- Malaria, Cerebral/cerebrospinal fluid*
- Malaria, Cerebral/drug therapy
- Malaria, Cerebral/parasitology

PMID: 11865422 [PubMed - indexed for MEDLINE]

Aphinyanaphongs, Y. and C.F. Aliferis. *Text Cat Models for Retrieval of High Quality Articles in Internal Medicine*. 2003

Document Preparation

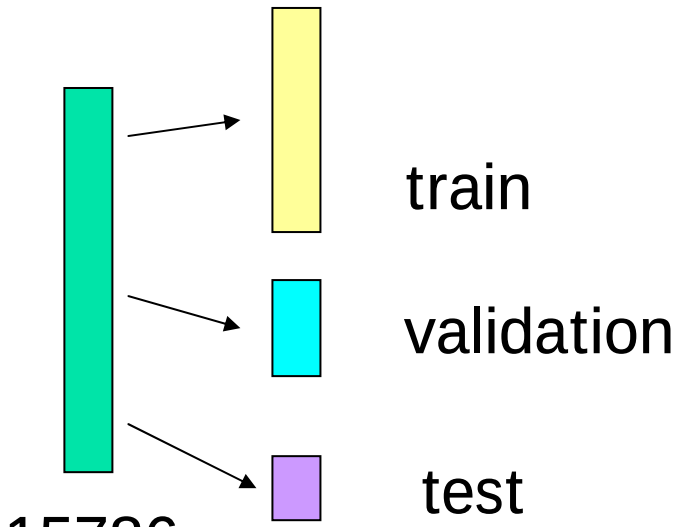
- "The clinical significance of cerebrospinal."

1. Representation
   - "The", "clinical", "significance", "of", "cerebrospinal"
2. Stop word removal
   - "Clinical", "Significance", "Cerebrospinal"
3. Porter Stemming (i.e. getting the roots of words)
   - "Clinic*", "Signific*", "Cerebrospin*"
4. Word encoding
   - All Words are encoded as binary variables.

Aphinyanaphongs, Y. and C.F. Aliferis. *Text Cat Models for Retrieval of High Quality Articles in Internal Medicine*. 2003

Study Design

- Build a model.
- Estimate the performance of the methodology.



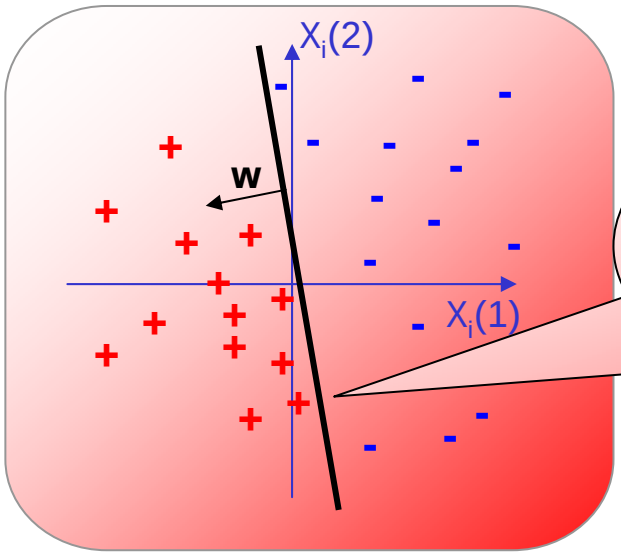
Treatment	Etiology
221+/8998-	123+/9349-
76+/3081-	41+/3116-
82+/3328-	41+/3116-

Word Selection

- All words.
- Haynes selected words.
- Linear and Approximate Polynomial Recursive Feature Elimination selected words.
- HITON-PC selected words.

Recursive Feature Elimination (RFE)

- The idea that features with low weight as determined by a support vector machine may be discarded.
  - Rank features according to $|w_i|$



if x(1)-axis is more informative than x(2)-axis, then

$$|w_{x(1)}| > |w_{x(2)}|$$

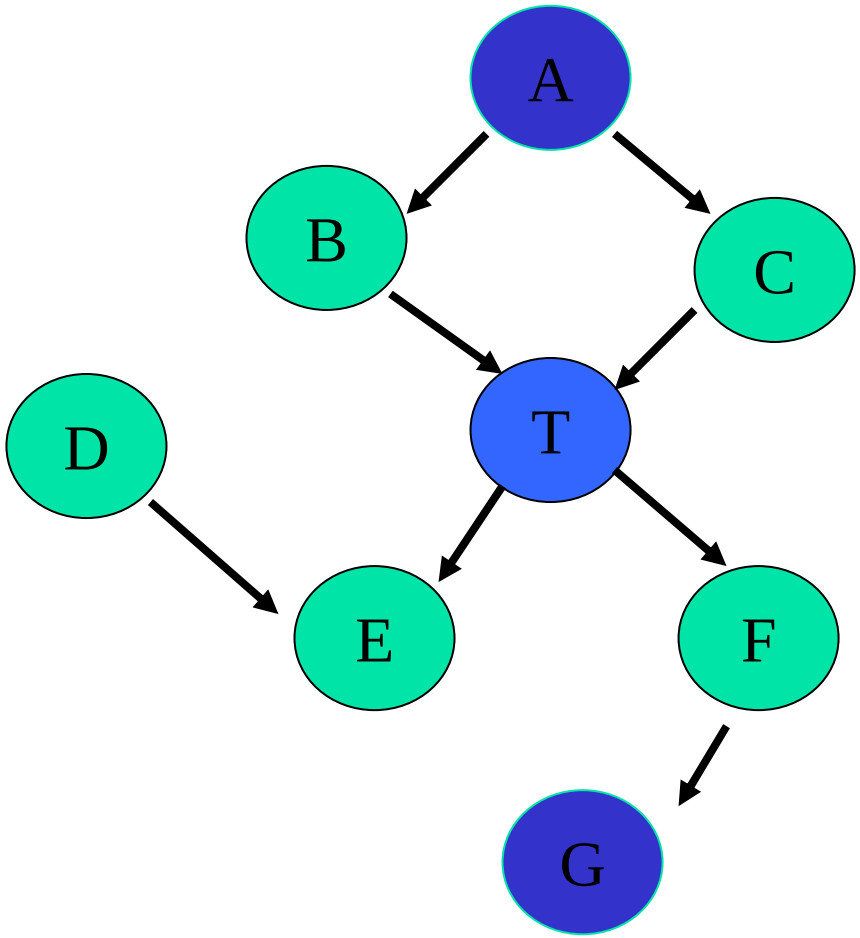
Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 2002. **46**: p. 389-422.

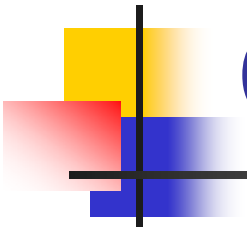
HITON-PC

- Algorithm that efficiently discovers the Markov Blanket (of a variable *T*)
- The MB is the set of variables MB, such that conditioned on MB every other variable becomes independent of *T*.
- Thus, the knowledge of the values of the Markov Blanket variables should render all other variables superfluous for classifying *T*.

C. F. Aliferis, I. Tsamardinos, A. Statnikov. "HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection" *In Proceedings of the 2003 American Medical Informatics Association (AMIA) Annual Symposium*, November 8-12, 2003, Washington, DC, USA, pages 21-25

The Markov Blanket

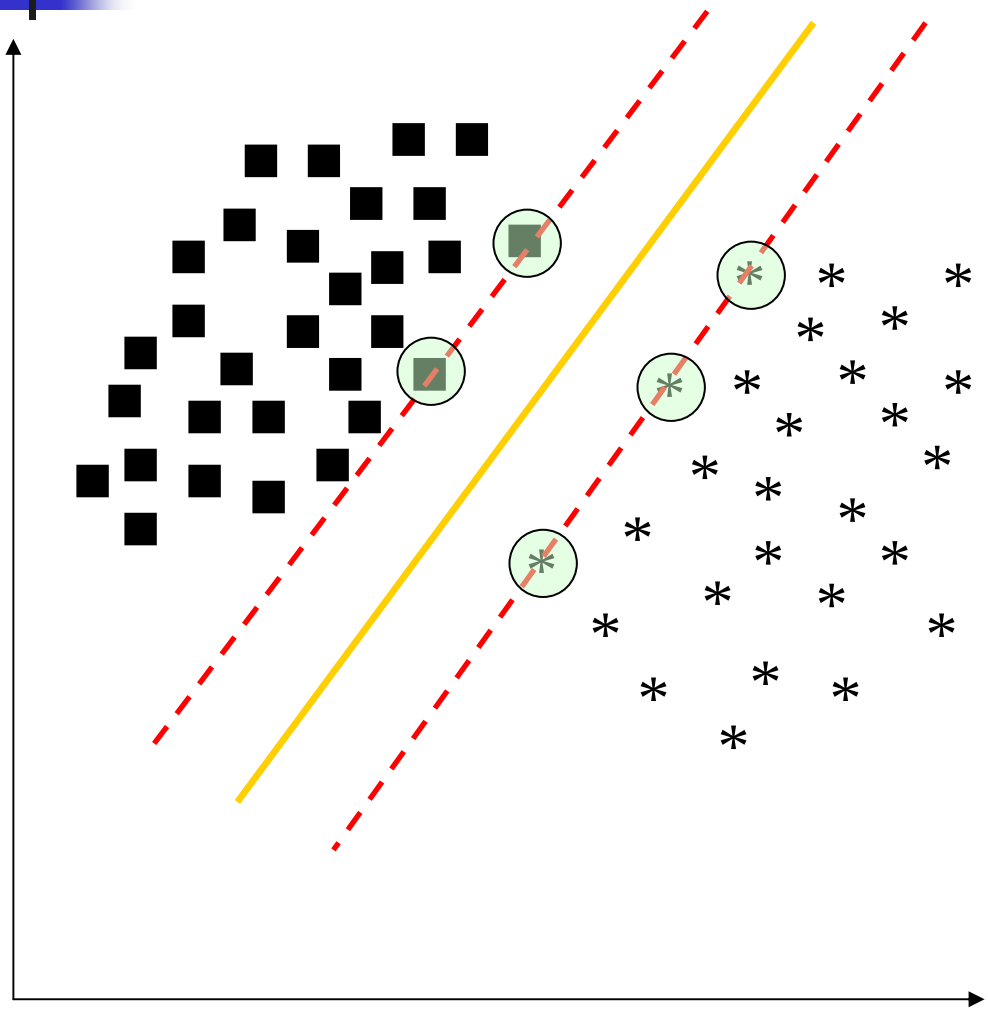




Classifier Algorithms

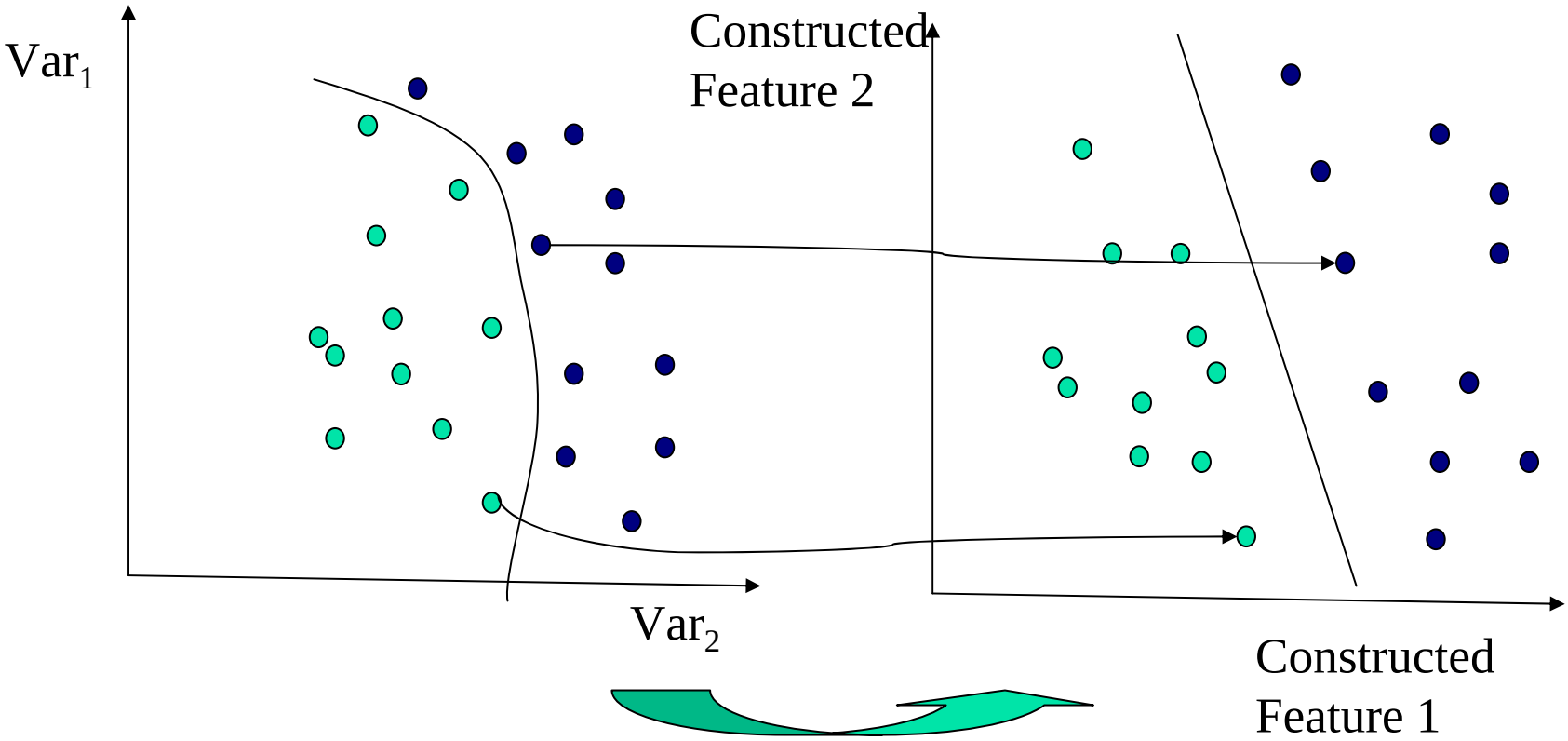
- Support Vector Machines
- Decision Trees

Linear Support Vector Machine



Burges, C., *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, 1998. **2**: p. 121-167.

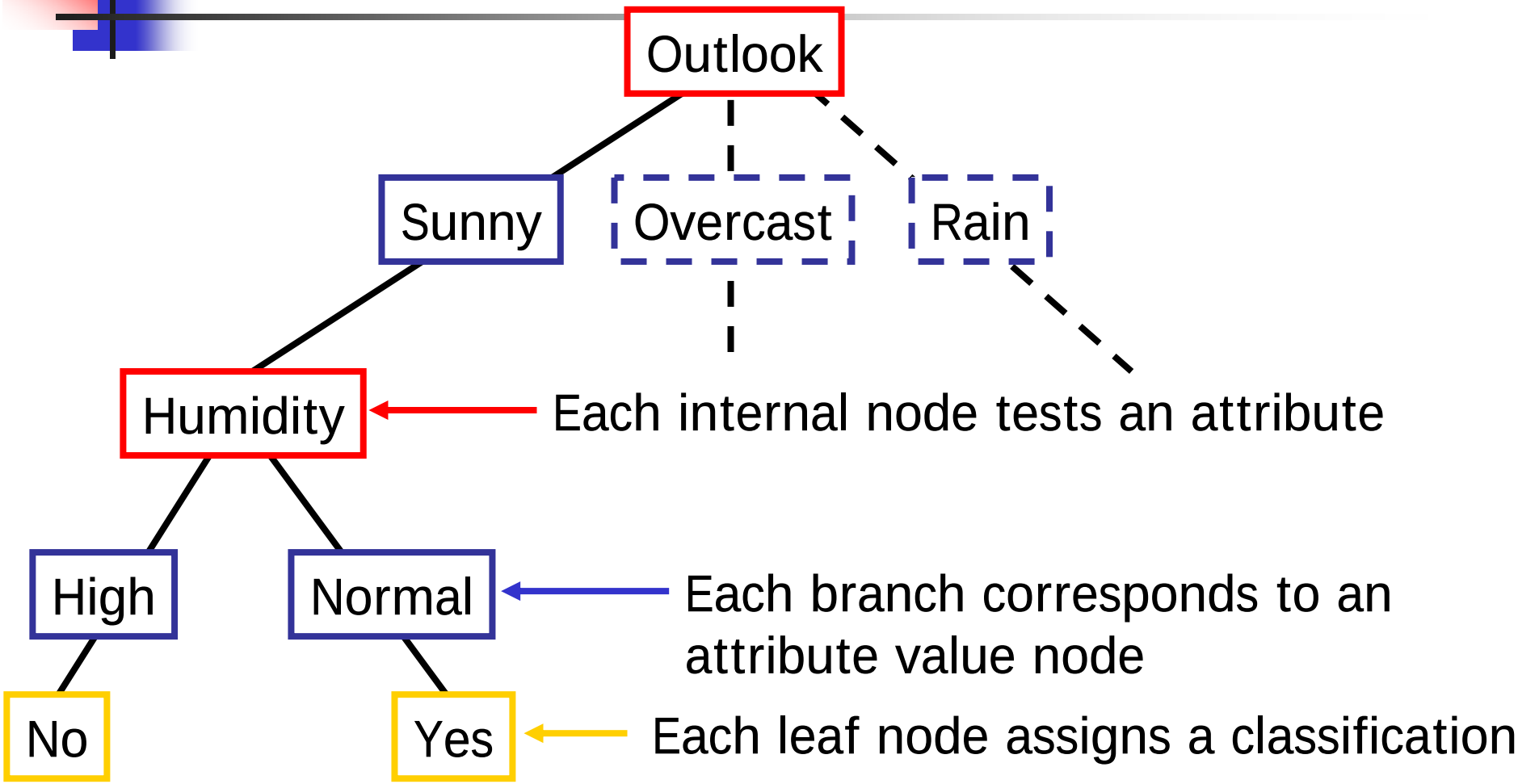
Non-linear Support Vector Machine



Find function Φ(x) to map to a different space

Burges, C., (1998).

Decision Tree for PlayTennis



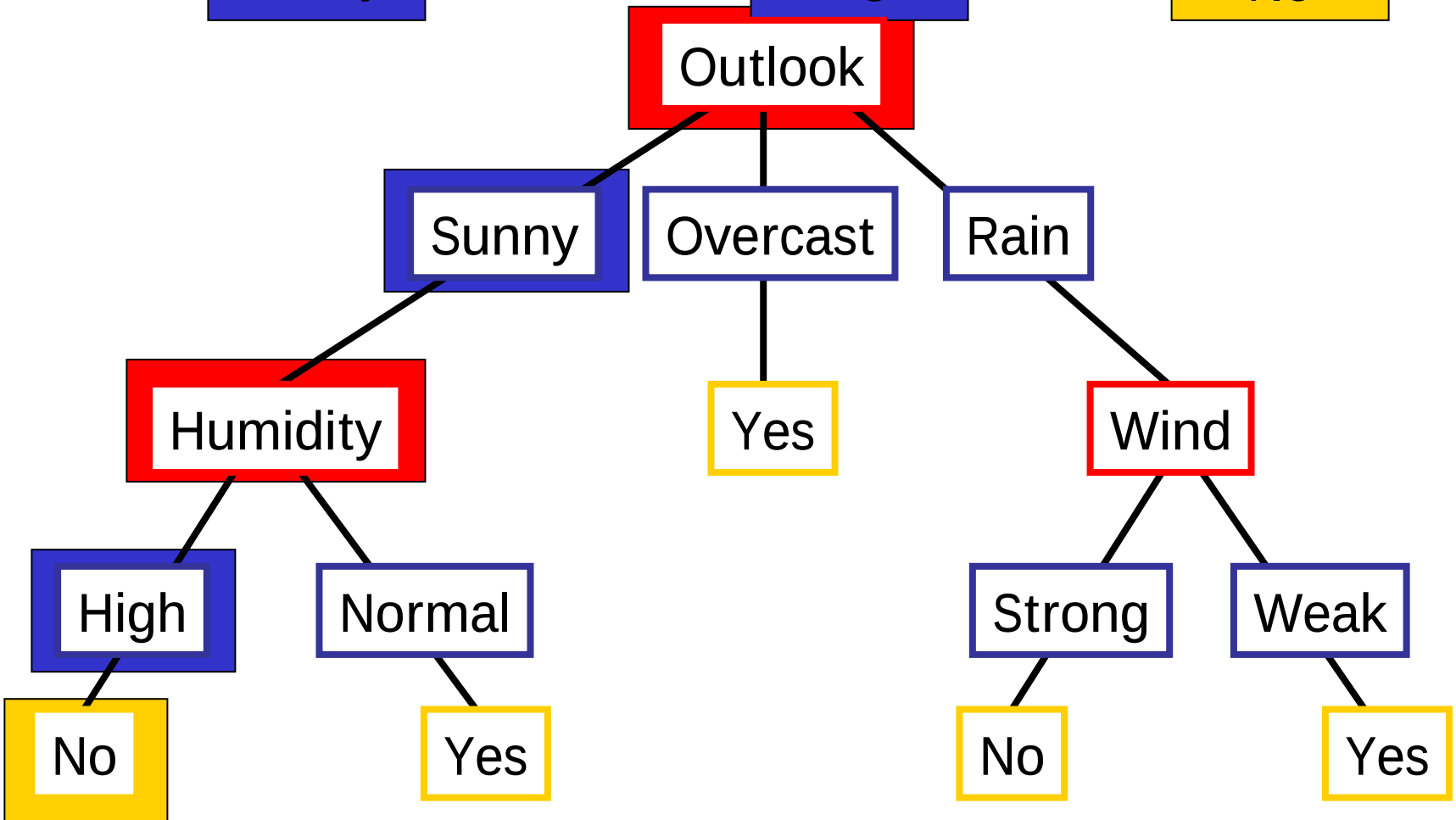
Mitchell '97 – Modified Chung Cheng University

Training Examples

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

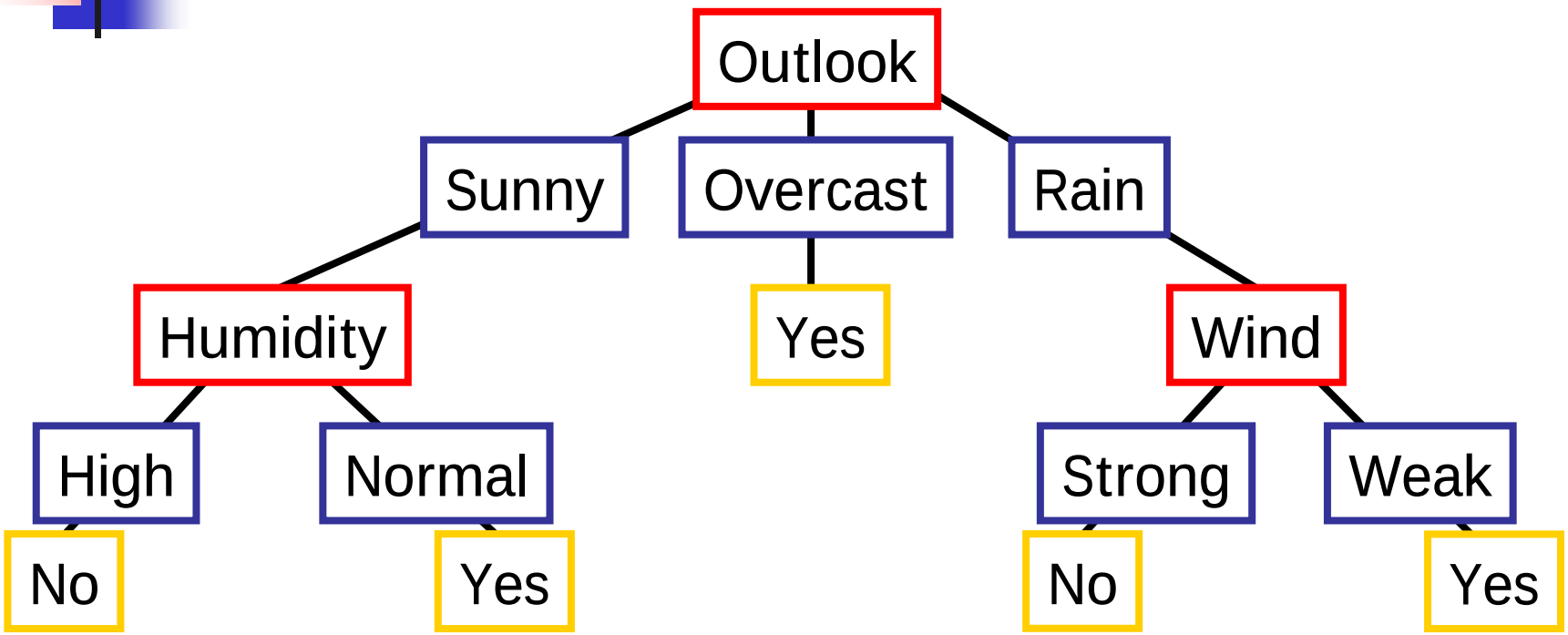
Decision Tree for PlayTennis

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No



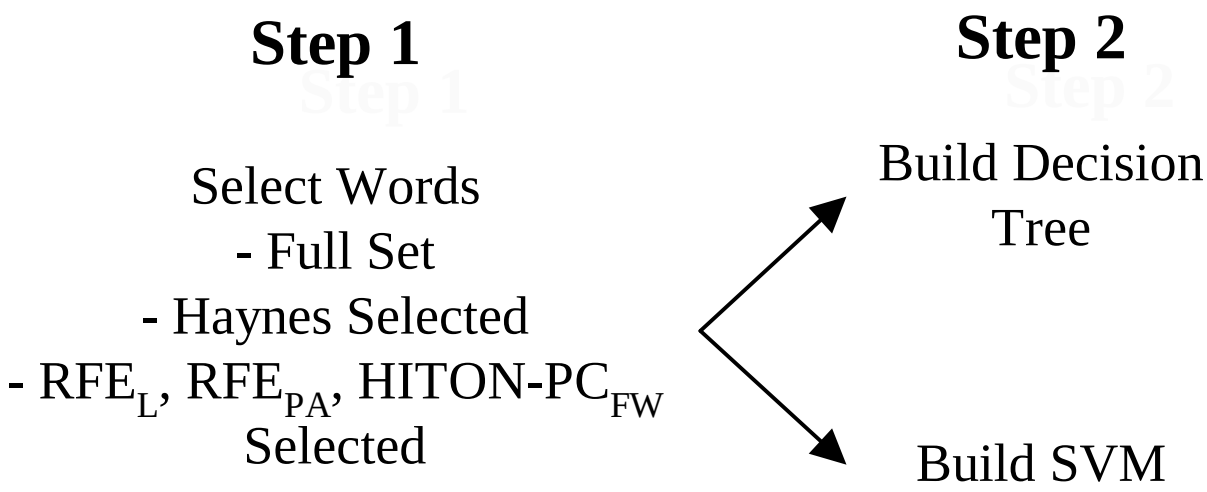
Decision Tree

- decision trees represent disjunctions of conjunctions



(Outlook=Sunny $\wedge$ Humidity=Normal)
$\vee$ (Outlook=Overcast)
$\vee$ (Outlook=Rain $\wedge$ Wind=Weak)

Experimental Design



Step 1 - Word Selection (Treatment)

HITON-PC$_{FW}$ (13 Features)*						0.92 AUC
RFE						
Features	**28000**	**1743**	**871**	**217**	**54**	**13**
RFE$_L$	0.95	0.85	0.96	0.97	0.86	#
RFE$_{PA}$	0.83	0.95	0.94	0.95	0.92	0.91

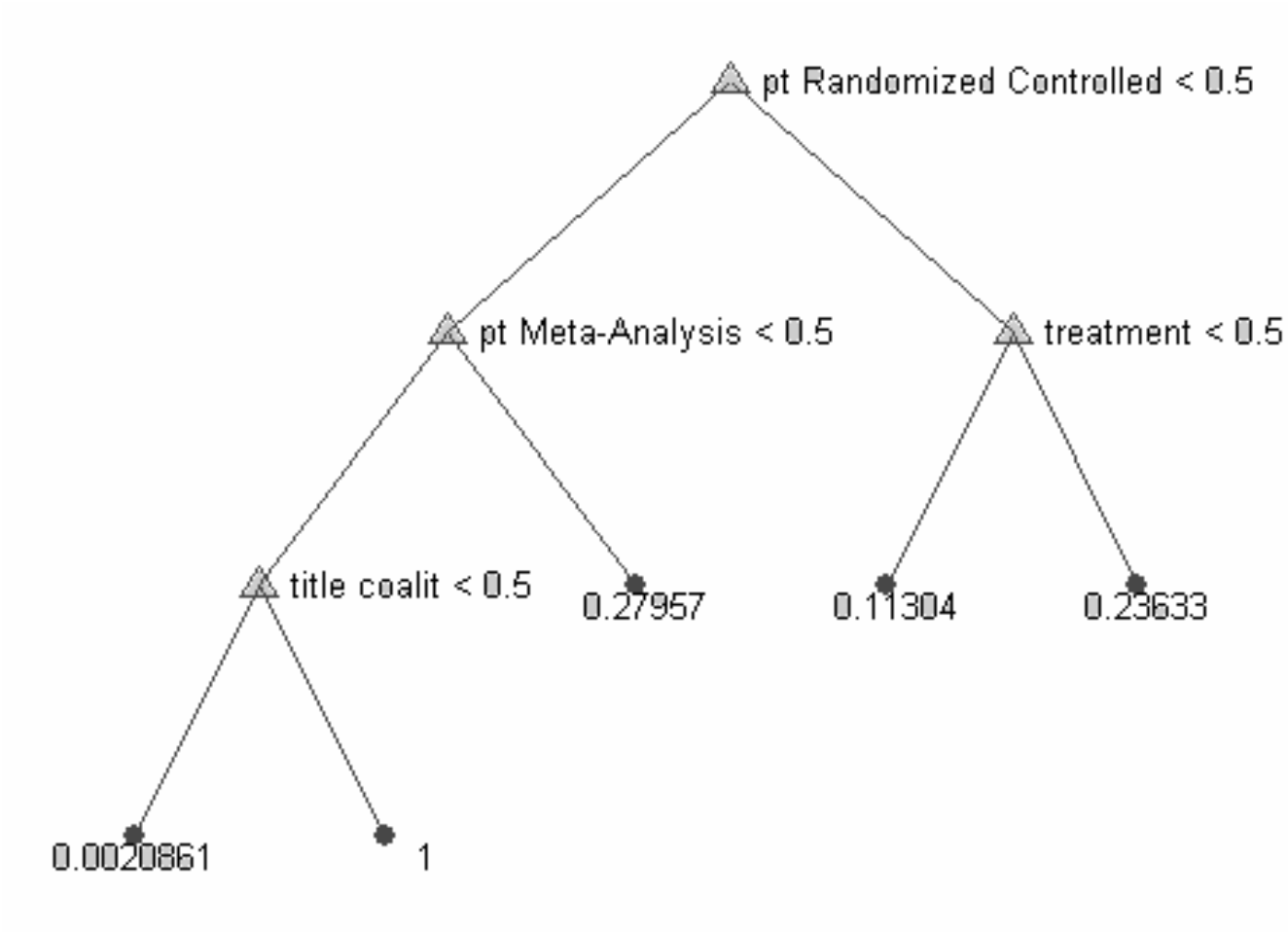
Step 2 – Support Vector and Decision Trees

Method	AUC	Words in pruned tree
Full Feature Set (27891 features)		
- SVMs	0.98	N/A
- DT	0.94	2
HITON-PC$_{FW}$ Feature Set (13 features)		
- SVM	0.95	N/A
- DT	0.95	4
Haynes Feature Set (747 features)		
- SVM	0.94	N/A
- DT	0.93	2

13 Selected Words (Treatment)

- Leagu [Textword]
- Coalit [Title]
- Catheterization, Central Venous: adverse effects [MeSH]
- Ribavirin: adverse effects [MeSH]
- Acupuncture Therapy [MeSH]
- Trandolapril [Textword]
- Pregnancy Complication: Prevention and Control [MeSH]
- Comprehens [Title]
- Meta-Analysis [Publication Type]
- Randomis [Textword]
- Trial [Textword]
- Randomized Controlled Trial [MeSH]
- Treatment [Textword]

Treatment Decision Tree

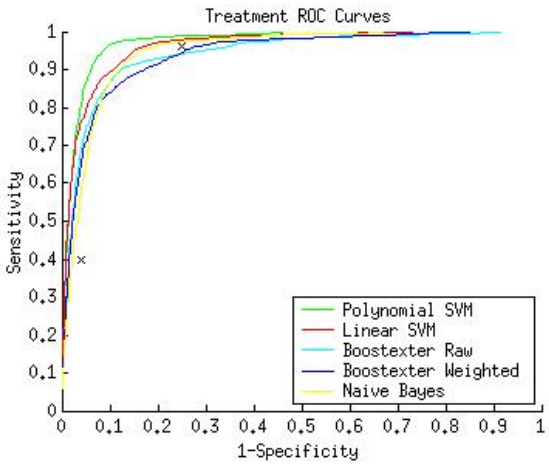


Decision Tree Boolean query

- (“Meta-analysis” [Publication-Type]) OR (“Randomized controlled trial” [Publication Type] and Treatment [TextWord]) OR (“Randomized controlled trial” [Publication Type])

Comparison to Clinical Query Filters

$$dist = \sqrt{(1-sens)^2 + (1-spec)^2}$$



Method	**Distance**
CQF filter – optimized for sensitivity	0.23
CQF filter – optimized for specificity	0.50
Full feature set/ decision tree	**0.11**
HITON features set/ decision tree	**0.11**
Haynes feature set/ decision tree	**0.11**

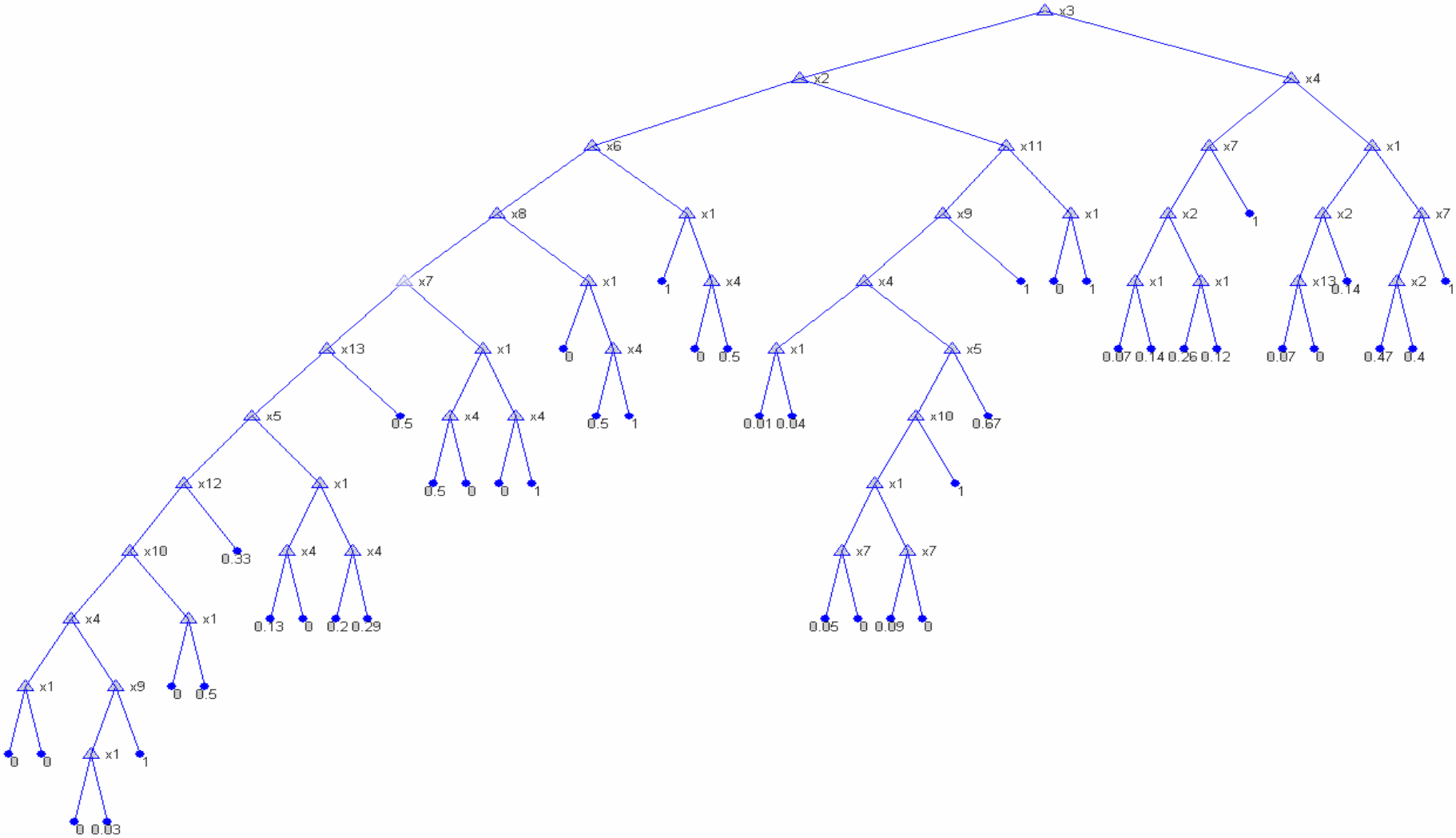
Etiology Category

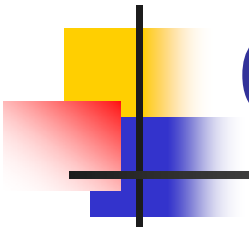
Category	Number of words	Support Vector Machine	Decision trees	Number of words by HITON	HITON with Support Vector Machine	HITON word selection with decision trees
Treatment	27891	**0.97**	**0.94**	13	**0.95**	**0.95**
Etiology	27891	**0.94**	**0.80**	13	**0.92**	**0.90**

13 Selected Words (Etiology)

- Associ [Textword]
- Risk Factors [MeSH]
- Mortal [Title]
- 95 [Textword]
- Meta [Title]
- Killip [Textword]
- Drinker [Textword]
- Phentermin [Textword]
- Sick role [MeSH]
- Autoimmunity [MeSH]
- Homocyst [Textword]
- Smoking Cessation [MeSH]
- Weather [MeSH]

Etiology Decision Tree





Conclusions

- We have presented a combined feature selection/decision tree method that can produce decision trees that perform as well as the best text classifiers.
- These decision trees are understandable, manageable, and amenable to validation by humans.
- These trees and queries are generated automatically from a corpus hence the process can be readily repeated many times in similar domains/tasks.
- The Boolean queries can be readily applied to existing search engines.

Acknowledgements

- NLM.
- MD/PhD Program at Vanderbilt University.
- Alexander Statnikov.