Decision Trees, Random Forests, and the Genealogy of the Algorithmic Black Box

Matthew L. Jones
mjones@columbia.edu
@nescioquid

@nescioquid

NSA-GCHQ data mining

UK TOP SECRET STRAP1 COMINT
AUS/CAN/NZ/UK/US EYES ONLY

Reference: OPC-M/TECH.A/455 (v1.0, r206)
Date: 20 September 2011
Copy no:

HIMR Data Mining Research Problem Book

[redacted]

OPC-MCR, GCHQ

Summary
In this problem book we set out areas for long-term data mining research at the Heilbronn Institute for Mathematical Research starting in October 2011 and continuing for at least three years. The four areas are beyond supervised learning, information flow in graphs, streaming exploratory data analysis and streaming expiring graphs.

Friends of Random Forests

- "The NSA were very early adopters of Random Forests through direct contact with [redacted] via the NSA Statistical Advisory Group (NSASAG) [W31]."
  - steganography detection (Random Forest) [I74]
  - website classification (decision tree) [I36]
  - protocol classification (Random Forest and neural network) [W1]
  - spam detection (Random Forest) [I44]
  - payphone detection (Random Forest)

- [redacted] OPC-MCR-GCHQ. "HIMR Data Mining Research Problem Book," September 20, 2011.

Interpretability

- A problem with the use of Random Forests is that their decisions can not be simply and intuitively explained to an analyst. This black box nature can lower analyst trust in a prediction. [Redacted] (NSA R1) has been leading an effort to make Random Forests more interpretable. [I18].

  - [redacted] OPC-MCR-GCHQ. "HIMR Data Mining Research Problem Book," September 20, 2011.

Decision Trees

Number	Attributes				Class
	account	balance	employed	monthly expense	
1	bank	700	yes	200	accept
2	bank	300	yes	600	reject
3	none	0	yes	400	reject
4	other inst	1200	yes	600	accept
5	other inst	800	yes	600	reject
6	other inst	1600	yes	200	accept
7	bank	3000	no	300	accept
8	none	0	no	200	reject

"Training set"

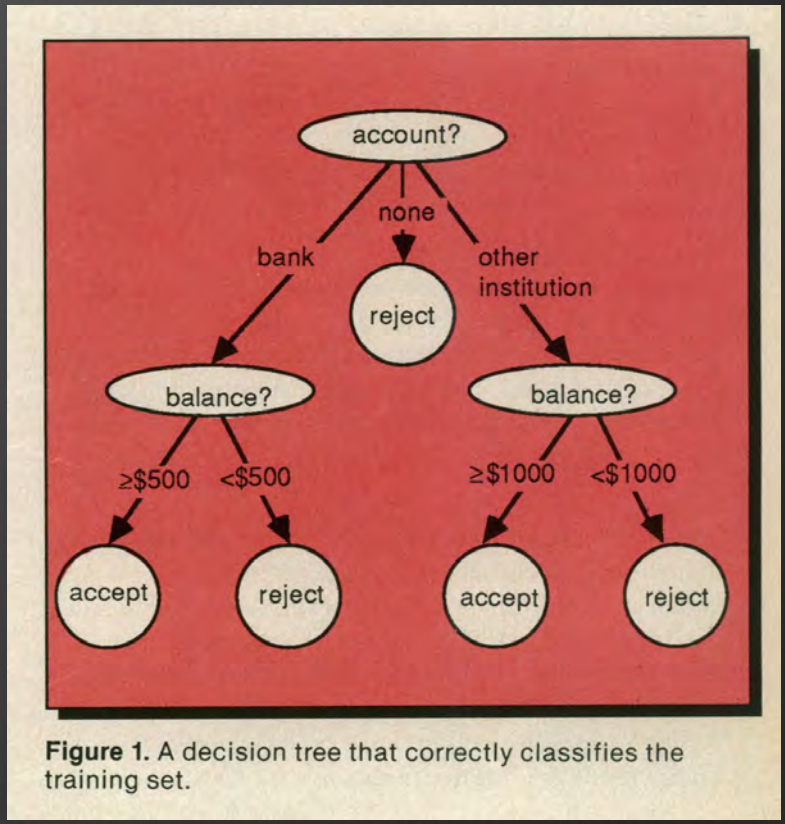


Figure 1. A decision tree that correctly classifies the training set.

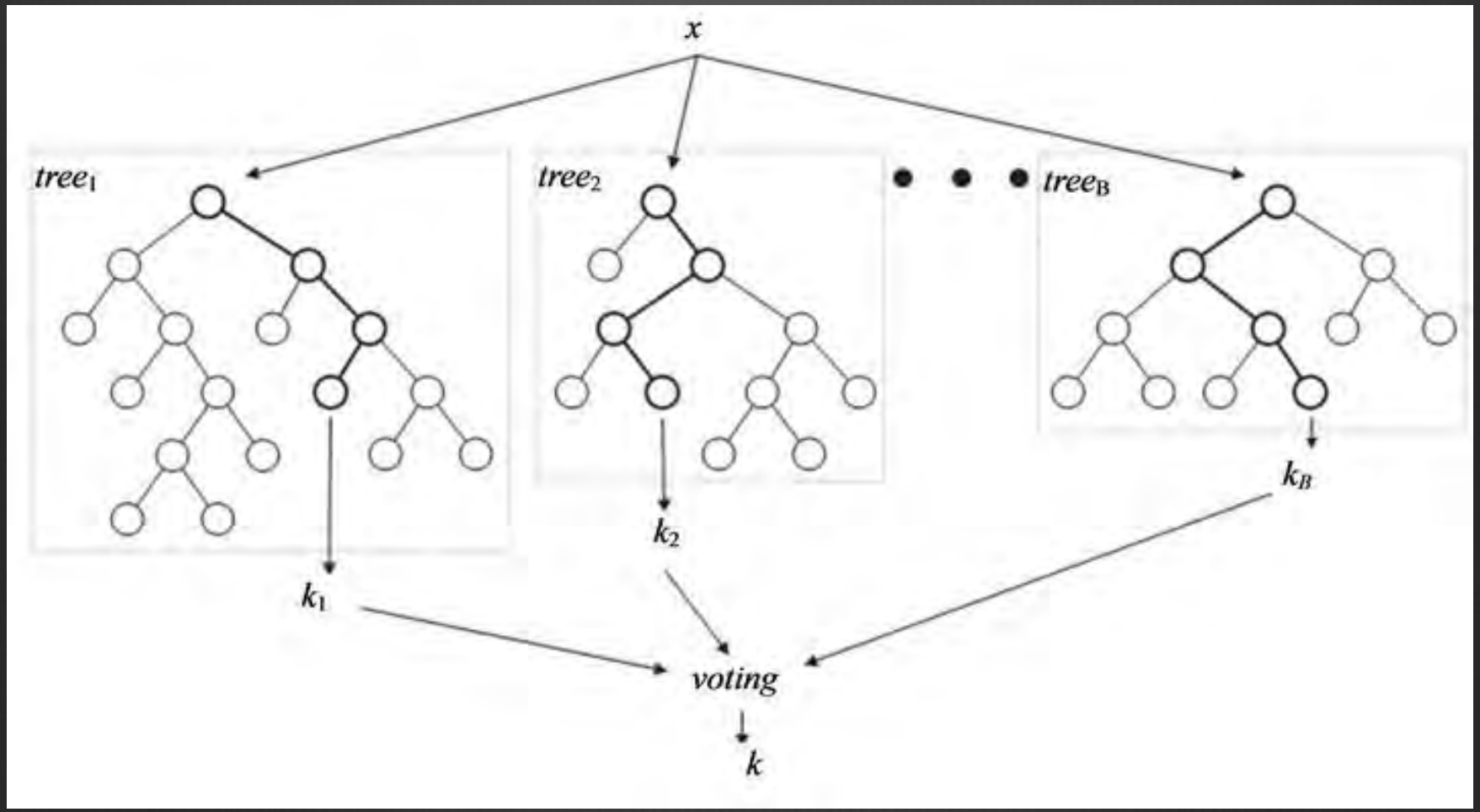
Carter, Chris, and Jason Catlett. "Assessing Credit Card Applications Using Machine Learning." *IEEE Expert* 2, no. 3 (September 1987): 71–79.

@nescioquid

Decision Trees

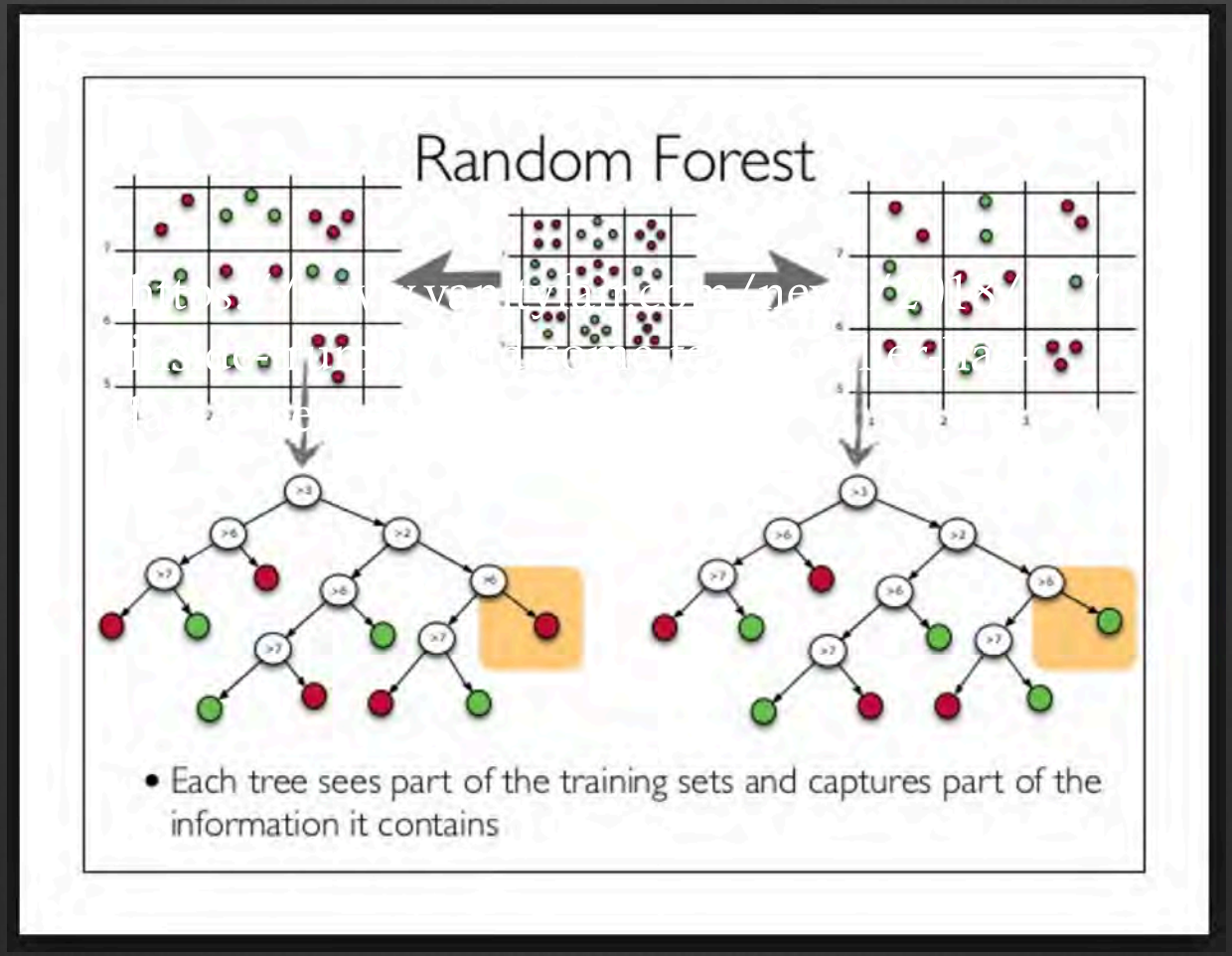
- Intuitive and highly interpretable
  - Crucial for numerous industries that must be able to justify their decision-making: credit, insurance, loans
- Great with heterogeneous data
- Moderately predictive
- Computable in polynomial time
- Extremely prone to overfit

Random forest



http://file.scirp.org/Html/6-9101686_31887.htm

Random forest



@nescioquid http://www.slideshare.net/0xdata/jan-vitek-distributedrandomforest522013

Forests

...forests are A+ predictors.

But their mechanism for producing a prediction is difficult to understand. Trying to delve into the tangled web that generated a plurality vote from 100 trees is a Herculean task. So on interpretability they rate an F.

Leo Breiman 2001

@nescioquid

Why so interesting?

- Predictive but not interpretable
- Excellent with *high dimensional, automatically produced data*
- Massively used in large technological systems

  if now waning a bit
- Helped secured *closure* on use of black box systems

Granularity and data ethics

the data sets that typically fall under the big data umbrella are about *people* — their attributes, their preferences, their actions, and their interactions

In other words, not only do these data sets document social phenomena, they do so at the granularity of *individual* people and their activities.

- Hanna Wallach, https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d#.in5ja1eeg

@nescioquid

@nescioquid

Decision Trees

Number	Attributes				Class
	account	balance	employed	monthly expense	
1	bank	700	yes	200	accept
2	bank	300	yes	600	reject
3	none	0	yes	400	reject
4	other inst	1200	yes	600	accept
5	other inst	800	yes	600	reject
6	other inst	1600	yes	200	accept
7	bank	3000	no	300	accept
8	none	0	no	200	reject

"Training set"

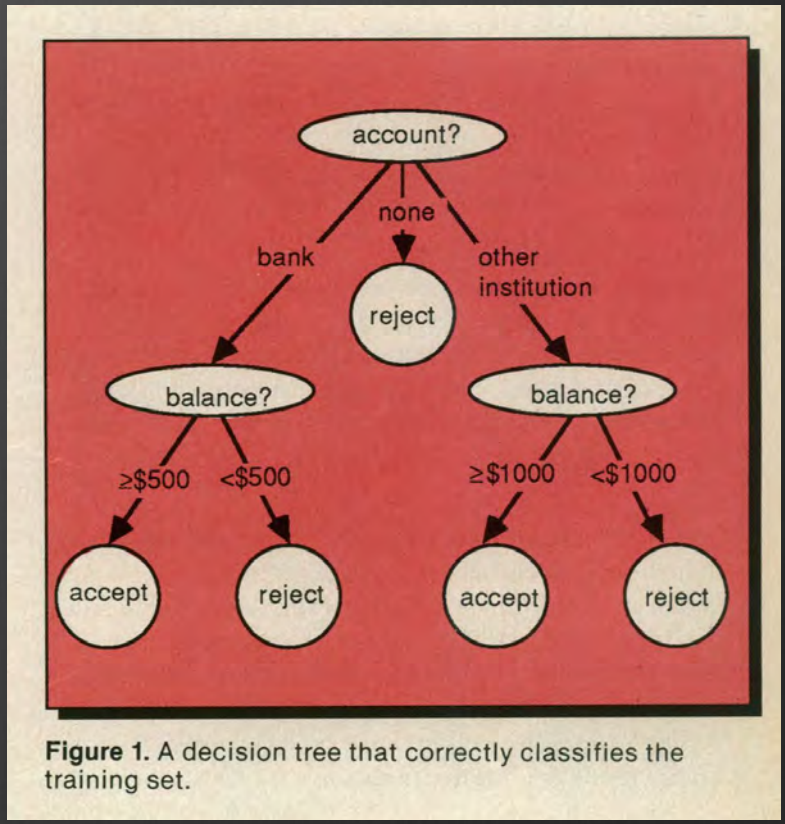
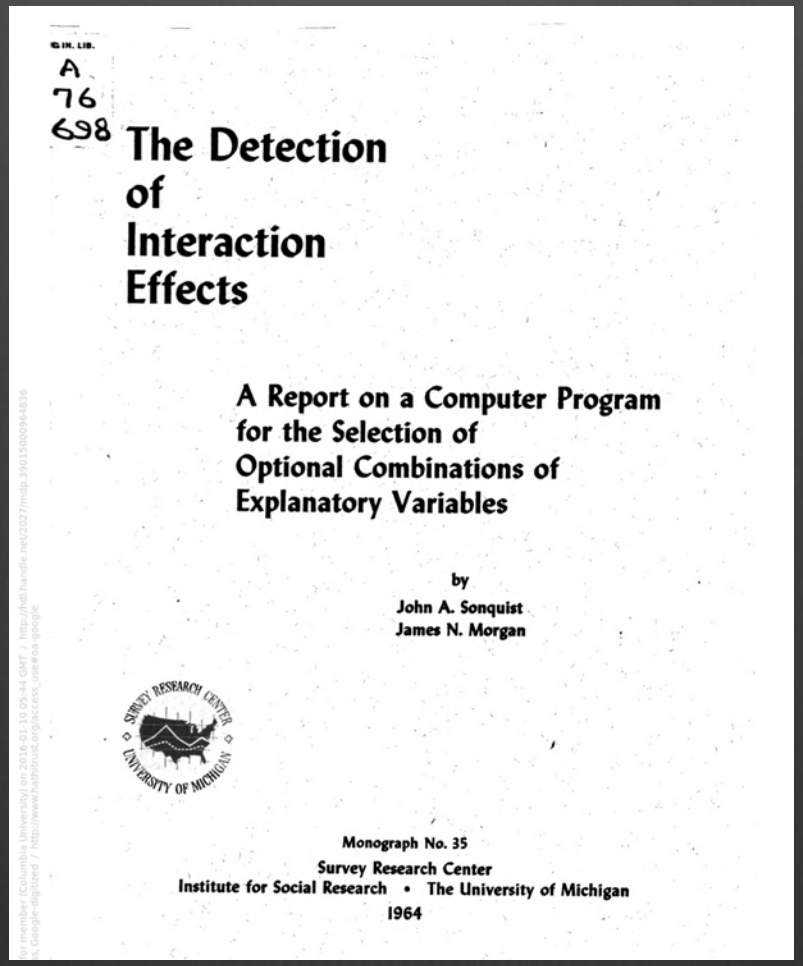


Figure 1. A decision tree that correctly classifies the training set.

Carter, Chris, and Jason Catlett. "Assessing Credit Card Applications Using Machine Learning." *IEEE Expert* 2, no. 3 (September 1987): 71–79.

@nescioquid

Institute for Social Research



The Detection of Interaction Effects

A Report on a Computer Program for the Selection of Optional Combinations of Explanatory Variables

by

John A. Sonquist
James N. Morgan

SURVEY RESEARCH CENTER
UNIVERSITY OF MICHIGAN

Monograph No. 35
Survey Research Center
Institute for Social Research • The University of Michigan
1964

“Automatic Interaction Detection” (AID)

The motivation for the development of the computer program described in this report comes from two sources. First, is the belief that the multivariate statistical techniques in common usage are often inadequate for the analysis of the rich body of data from a cross section sample survey, and second is the conviction that a large-scale digital computer can be used for more than just a high-speed adding machine.

Statistical Restraint

We have tried to break away from the habit of asking the question, "What is the effect of x on y when everything else is held constant?" This has been replaced with, "What do I need to know most in order to reduce predictive error a maximum amount?"

This is the type of question that might be asked by a research scientist working in a substantive area in which theory is not yet very precise. Once he receives an answer, he may well ask, "Now that I know this, what additional information would help to reduce predictive error still further?" and so on. He would certainly ask other questions as

Ubiquitous now, not then

- **1963-70s: laughed out of water**
- Decision trees "were called 'a recipe for learning something wrong'" says Dan Steinberg. "This was a death sentence, like a restaurant with *E. Coli.* Trees were finished."

Data mining as pejorative

- "... proceeding via a 'dustbowl' empiricism is dangerous at worst and foolish at best .... The purely empirical approach is particularly dangerous in an age when computers and packaged programs are readily available, since there is temptation to substitute immediate empirical analysis for more analytic thought and theory building."
  - Einhorn, "Alchemy in the Behavioral Sciences," 1972

@nescioquid

Revivifying decision trees

Two fold

Critiques of artificial reason

Empirical machine learning

Critique of Mathematical Statistics

Critique of mathematical stats

Decision trees: Stats

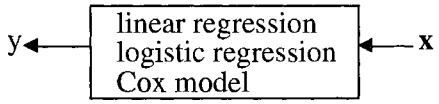
- Leo Breiman
  - Consultant for DOD contractor after leaving academia
  - Technology Services Corporation
- "We were working on prediction problems"
  - like next day ozone in the Los Angeles basin,
  - carbon monoxide levels on freeways,
  - but also things such as could we recognize the sender of handset Morse code—this was something we were doing for the spook agencies—
  - or could we recognize from sonar returns whether the other submarine was Russian or American?" (in Ohlsen, 188).

Decision trees: Stats

- Data-focused academic 1963 then late 1980s-1990s
- Breiman: "*Alice in Wonderland*. That is, I knew what was going on out in industry and government in terms of uses of statistics, but what was going on in academic research seemed light years away. It was proceeding as though it were some branch of abstract mathematics." (in Ohlsen, 196)
- Epistemic virtue: Predictive accuracy >> causal model

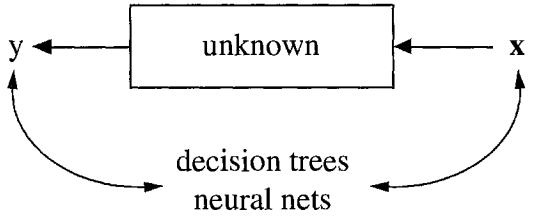
Two statistical cultures

"Data modeling culture"



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

"Algorithmic modeling culture"



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Approaching problems by looking for a data model imposes an a priori straightjacket that restricts the ability of statisticians to deal with a wide range of statistical problems.

Breiman, 2001, p. 204.

John Tukey and data analysis

Data analysis, […], must then take on the characteristics of a science rather than those of mathematics, specifically:

(1) Data analysis must seek for scope and usefulness rather than security.

(2) Data analysis must be willing to err moderately often […].

(3) Data analysis must use mathematical argument and mathematical results as bases for judgment rather than as bases for proofs or stamps of validity.

Tukey, "Future of Data Analysis," 1961, III:401

War time data work

Thanks to “war problems,” “it was natural to regard statistics as something that had the purpose of being used on data—maybe not directly, but at most at some remove. Now, I can’t believe that other people who had practical experience failed to have this view, but they certainly—I would say—failed to advertise it.”

Tukey interview, 85.

Low road of instrumental computational statistics

Vast archives of data

(cryptological in first instance)

Plus

Highly *instrumental* statistical approach

Not the sexy high theory cold war fields

Operations research

Game theory

Formal AI

Decision theory—Mathematical Statistics

SE[illegible]-AUTOMATIC IMAGERY SCREENING
RESEARCH STUDY AND EXPERIMENTAL INVESTIGATION

Volume I

Report No. 2 and 3

Philco No. V043-2 and 3

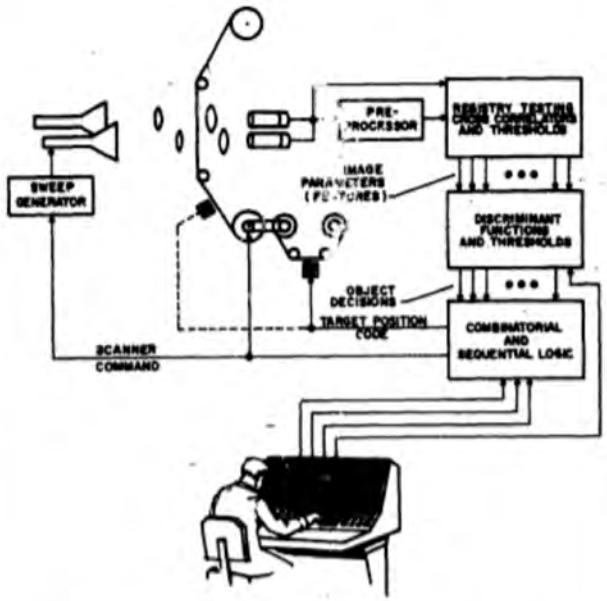
[illegible]al Corps Contract No. DA-36-039-SC-90742

U. ARMY ELECTRONICS RESEARCH AND DEVELOPMENT LABORATORY
FORT MONMOUTH, NEW JERSEY

Consolidation of Quarterly Report Numbers 2 and 3
1 September 1962 to 28 February 1963

DA Task No. 3A99-23-001-02

PHILCO
a subsidiary of Ford Motor Company
ADVANCED TECHNOLOGY LABORATORY
BLUE BELL, PENNSYLVANIA



Conceptual Block Diagram of Image Screening System

Ecologies of instrumental prediction on relatively high dimensional, messy data

US
USSR
France
Japan

@nescioquid

@nescioquid

Revivifying decision trees

Two fold

Critiques of artificial reason

Empirical machine learning

Critique of Mathematical Statistics

Empirical Machine Learning

Decision trees & Bottlenecks

"Human experts find it difficult to express their knowledge, or explain their actions, in terms of concise situation-action rules. If pressed to do so, they typically produce rules that are incorrect, or that have many exceptions. The articulation of specific intuitive knowledge into deterministic rules is a difficult, sometimes unrealistic, problem for human experts."

- Fayyad, U. M., K. B. Irani, J. Cheng, and Z Quin. "Machine Learning of Expert System Rules: Applications to Semiconductor Manufacturing." In *Collected Notes on the Workshop for Pattern Discovery in Large Databases (NASA Ames, January 14-15, 1991)*, 17–29. NASA Ames Research Center, 1991.

Knowledge Acquisition Bottleneck

"the central problem facing knowledge engineering today, the bottleneck of knowledge acquisition."

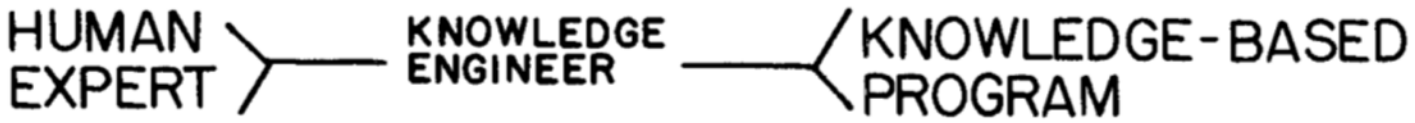
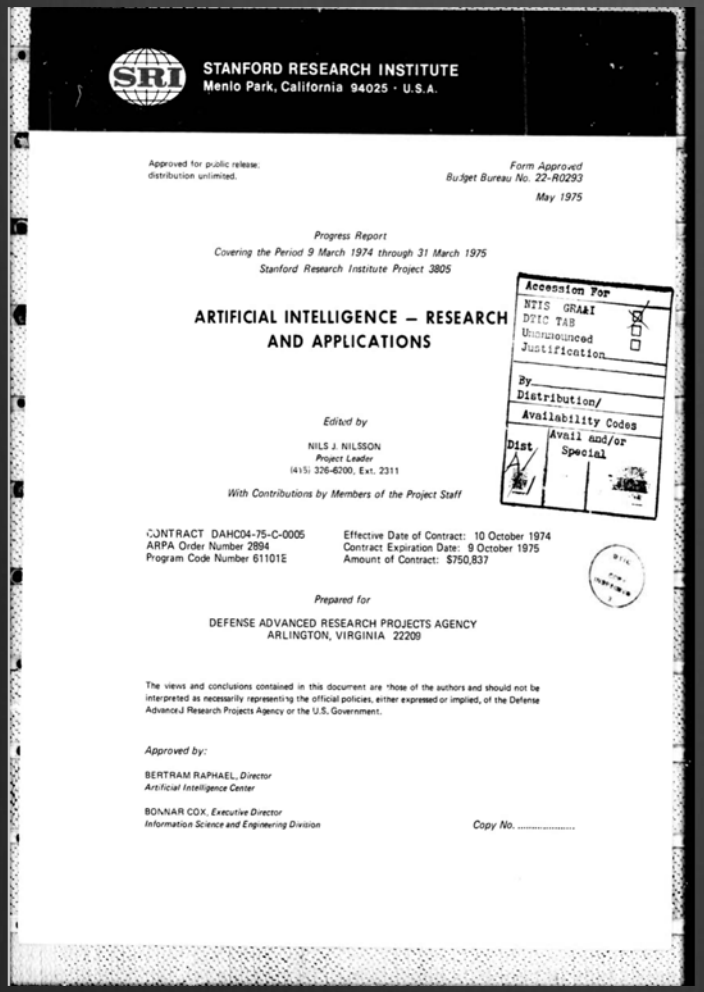


Figure 9-1: The bottleneck of knowledge acquisition is *transfer of expertise*. This comprises *(i)* the expert's difficulty in articulating what he knows, and *(ii)* the impedance mismatch between the concepts and vocabulary of the expert and the knowledge engineer.

Lenat, Douglas. "The Role of Heuristics in Learning by Discovery," 244

Getting expert knowledge



STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 · U.S.A.

Approved for public release;
distribution unlimited.

Form Approved
Budget Bureau No. 22-R0293
May 1975

Progress Report
Covering the Period 9 March 1974 through 31 March 1975
Stanford Research Institute Project 3805

ARTIFICIAL INTELLIGENCE – RESEARCH AND APPLICATIONS

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special

Edited by

NILS J. NILSSON
Project Leader
(415) 326-6200, Ext. 2311

With Contributions by Members of the Project Staff

CONTRACT DAHC04-75-C-0005
ARPA Order Number 2894
Program Code Number 61101E

Effective Date of Contract: 10 October 1974
Contract Expiration Date: 9 October 1975
Amount of Contract: $750,837

Prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
ARLINGTON, VIRGINIA 22209

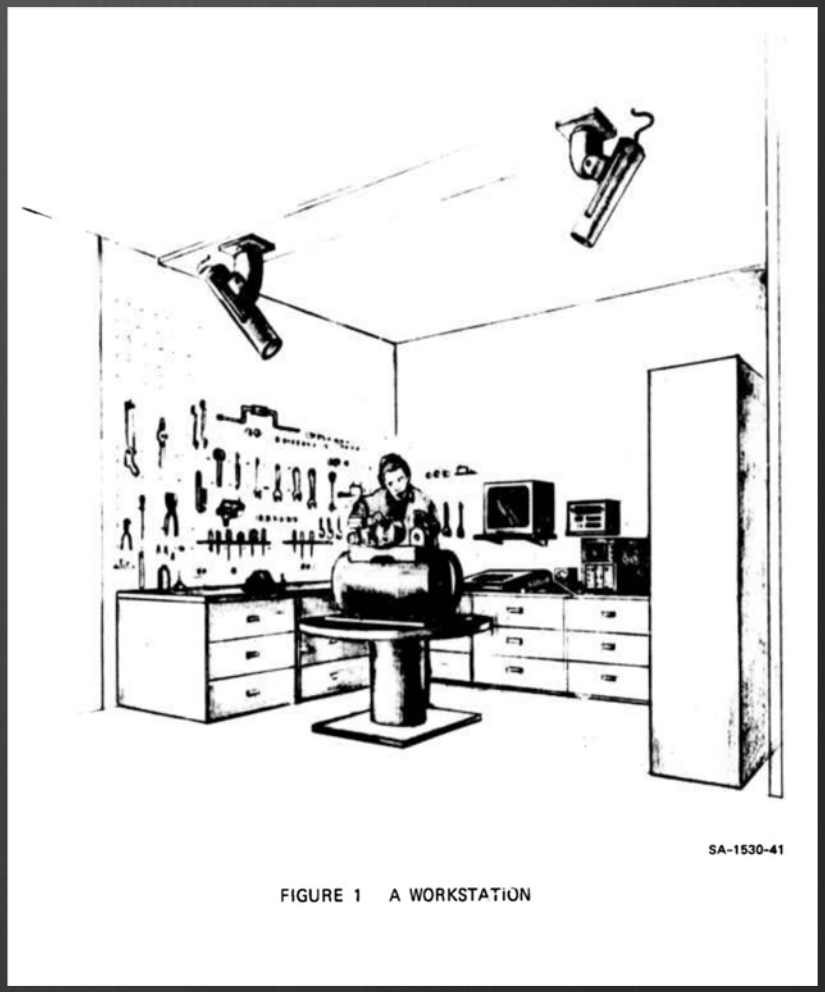
The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

Approved by:

BERTRAM RAPHAEL, Director
Artificial Intelligence Center

BONNAR COX, Executive Director
Information Science and Engineering Division

Copy No. .....



SA-1530-41

FIGURE 1 A WORKSTATION

@nescioquid

Getting expert knowledge

Expert: The pump pulley should be next.

Apprentice: Yes ... uh, does the side of the pump pulley with the holes face away from the pump or towards it?

E: Away from the pump.

A: All right.

E: Did you insert the key--that is, the half-moon shaped piece?

A: Yes, I did.

E: Be sure you check the alignment of the two pulleys before you tighten the set-screws.

A: Yes, I'm just now fiddling with that.

E: OK.

A: Tightening the Allen screw now.

E: OK, thank you.

.

SA-1530-41

FIGURE 1 A WORKSTATION

Machine learning

"Mastery is not acquired by reading books—it's acquired by trial-and-error and teacher supplied examples. This is how humans acquire skill. People are very reluctant to accept this. Their reluctance tells us something about the philosophical self image that we, as thinking beings, prefer. It tells us nothing about what actually happens when a teacher or a master trains somebody. That somebody has to regenerate rules from example to make them an intimate part of his intuitive skill."

Donald Michie, "Expert Systems Interview," *Expert Systems* 2, no. 1 (1985): 22.

Predict expertise

the machine learning technique *takes advantage of the data and avoids the knowledge acquisition bottleneck by extracting classification rules directly from data*. Rather than asking an expert for domain knowledge, a machine learning algorithm observes expert tasks and induces rule emulating expert decisions.

Keki B. Irani et al., "Applying Machine Learning to Semiconductor Manufacturing," IEEE Expert 8, no. 1 (1993): 41.

Example: Stellar Mapping

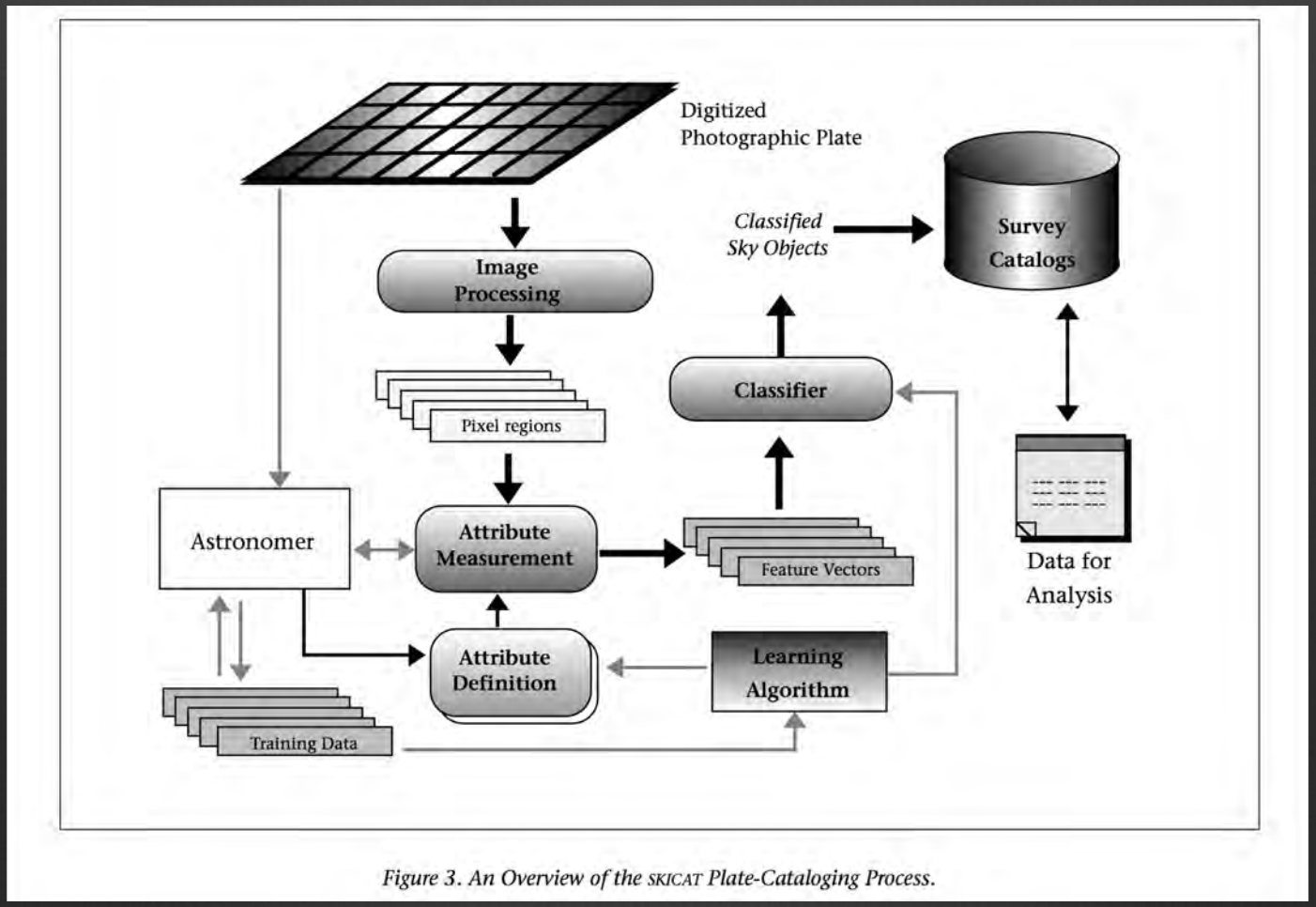


Figure 3. An Overview of the SKICAT Plate-Cataloging Process.

Induction from databases

Induction Over Large Data Bases

J. R. Quinlan
Basser Department of Computer Science
University of Sydney

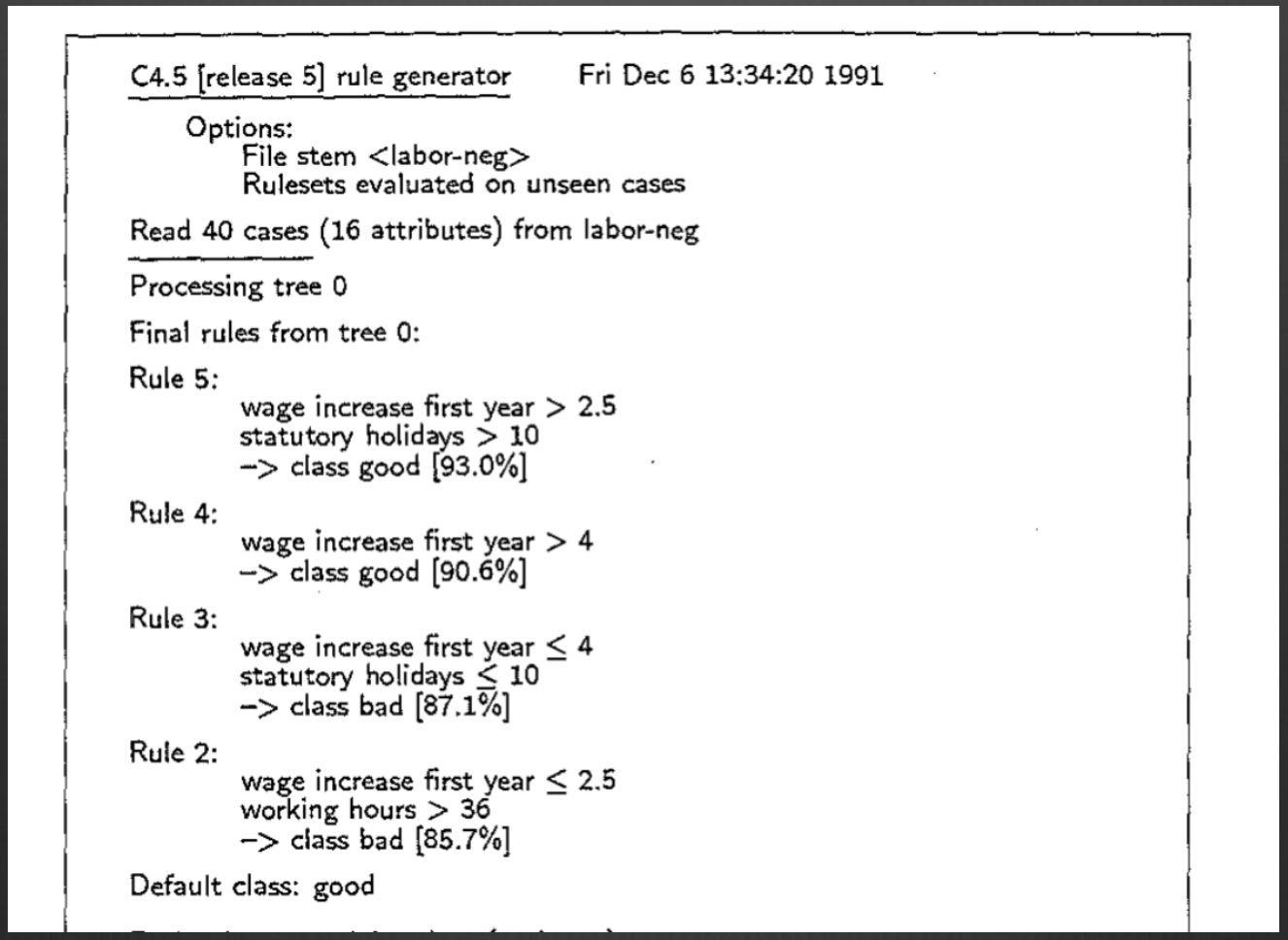
Abstract: Techniques for discovering rules by induction from large collections of instances are developed. These are based on an iterative scheme for dividing the instances into two sets, only one of which needs to be randomly accessible, These techniques have made it possible to discover complex rules from data bases *containing many thousands of instances. Results of several experiments using them are reported,*

Keywords: Induction, Inference, *pattern recognition, rule* **formation, concept learning, decision trees.**

ML: interpretability

- Not just prediction.
- "*ML requires that classifiers should not only classify but should also constitute explicit concepts, that is, expressions in symbolic form meaningful to humans and evaluable in the head.*"
    - (Feng and Michie, Machine Learning of Rules and Trees, 51, my italics)

ML rule sets



C4.5 [release 5] rule generator Fri Dec 6 13:34:20 1991

Options:
File stem <labor-neg>
Rulesets evaluated on unseen cases

Read 40 cases (16 attributes) from labor-neg

Processing tree 0

Final rules from tree 0:

Rule 5:
wage increase first year > 2.5
statutory holidays > 10
-> class good [93.0%]

Rule 4:
wage increase first year > 4
-> class good [90.6%]

Rule 3:
wage increase first year ≤ 4
statutory holidays ≤ 10
-> class bad [87.1%]

Rule 2:
wage increase first year ≤ 2.5
working hours > 36
-> class bad [85.7%]

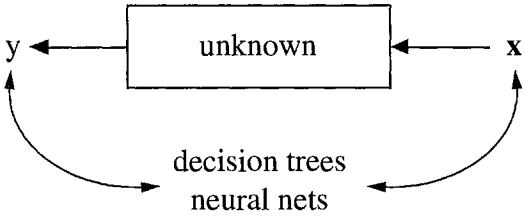
Default class: good

Two Algorithmic Cultures

Machine learning: comprehensibility

- Conversion to RULES: "C4.5 also contains a mechanism to re-express decision trees as ordered lists of if-then rules. […] There are substantially fewer final rules than there are leaves, and yet the accuracy of the tree and the derived rules is similar. Rules have the added advantage of being more easily understood by people." (Kohavi and Quinlan, 1999, 12)

Statisticians: prediction



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

@nescioquid

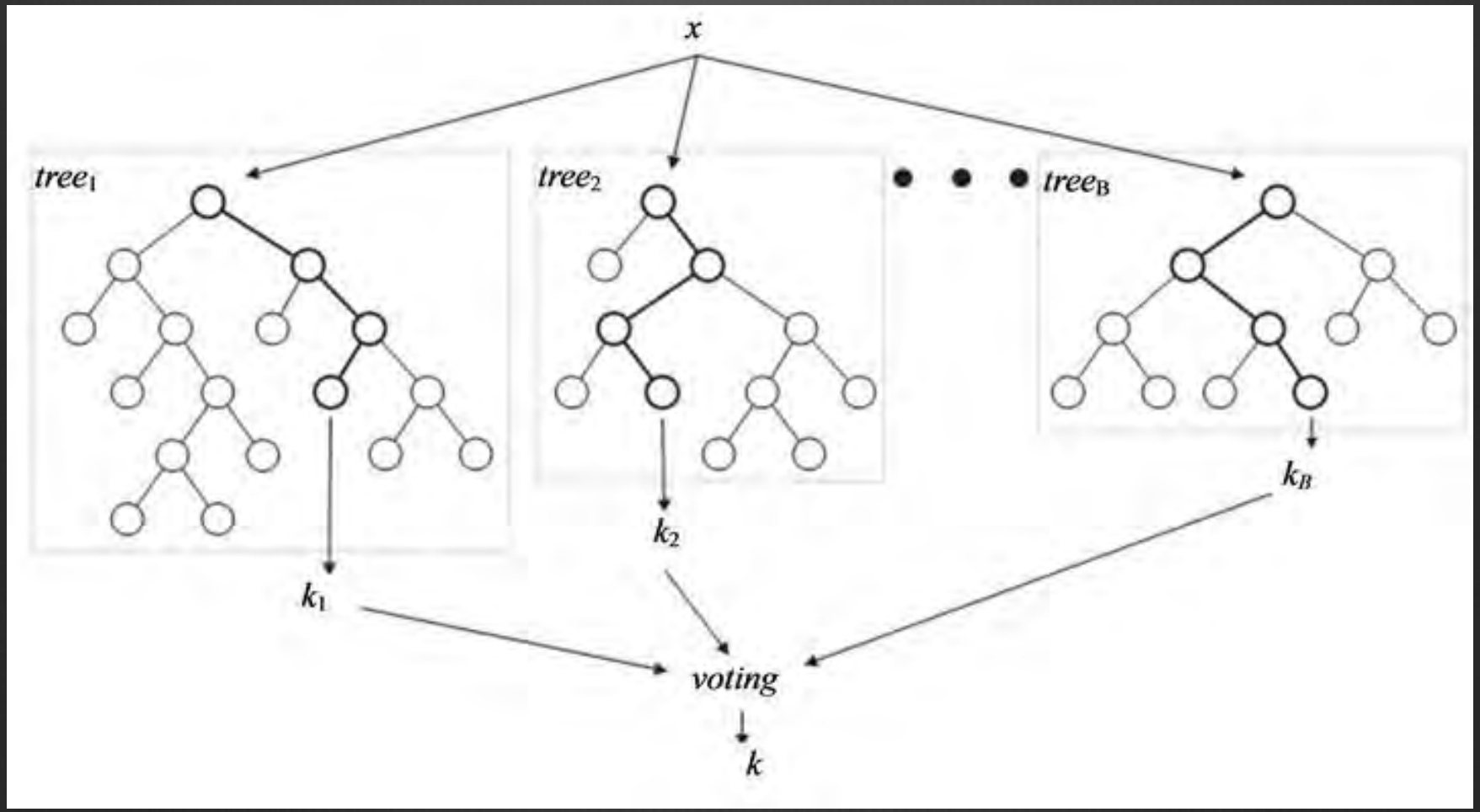
Ramification of Trees

- 1990s
  - trees among the major candidate algorithms for KDD
  - Simple to understand and track record
- Late 1990s
  - tree mining among multiple databases
- 2000s
  - Random Forests
  - Ensemble modeling

Ramification of Trees: Database values

- Major effort to make trees scale in late 1990s
- SPRINT- "Scalable PaRallelizable Induction of Decision Trees" (1996)
  - IBM Almaden lab
  - "With the recent emergence of the field of data mining, there is a great need for algorithms for building classifiers that can handle very large databases. . . . By eschewing the need for any centralized memory-resident data structures, SPRINT efficiently allows classification of virtually any sized dataset." (554)

Random forest



http://file.scirp.org/Html/6-9101686_31887.htm

Forests

...forests are A+ predictors.

But their mechanism for producing a prediction is difficult to understand. Trying to delve into the tangled web that generated a plurality vote from 100 trees is a Herculean task. So on interpretability they rate an F.

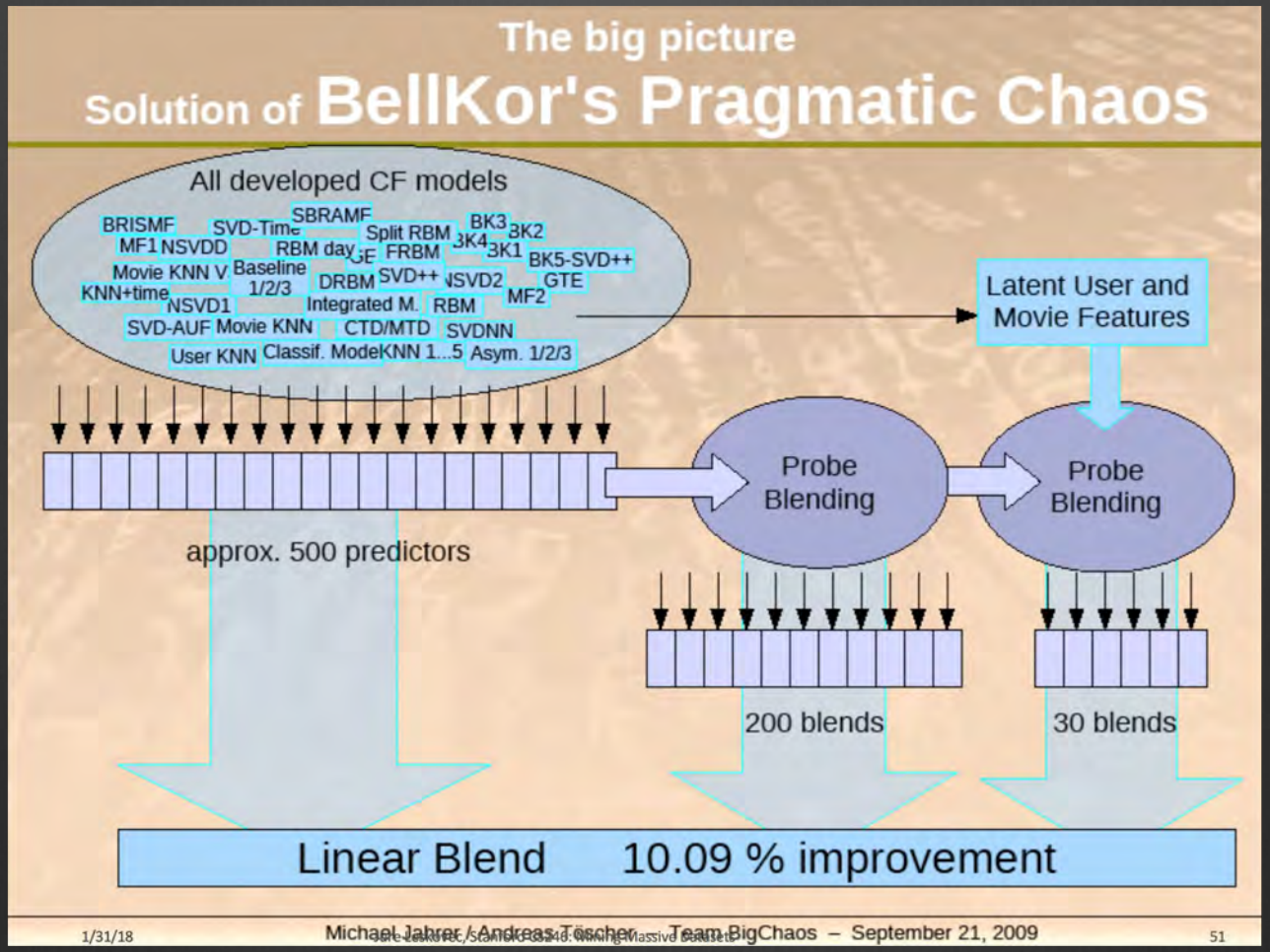
Leo Breiman 2001

@nescioquid

Netflix


NETFLIX
2009
DATE 09.21.09
PAY TO THE ORDER OF BellKor's Pragmatic Chaos
$ 1,000,000 00
AMOUNT ONE MILLION 00/100
FOR The Netflix Prize
Reed Hastings

Ensemble chaos



@nescioquid

NSA then and now

- "NSA Valued in the 1980s, Accuracy, Deep Knowledge, Thorough expertise, Productivity and Reputation […]."

- "NSA valued in the 2000s […] Speed-getting it 80 percent right now could make all the difference in saving lives. (Of course, if it were targeting information that would mean killing innocents 20 percent of the time.)"

    - redacted, "NSA Culture, 1980s to the 21st Century--a SID Perspective," *Cryptological Quarterly* 30, no. 4 (n.d.): 84.