The Structure of Predictive Coding

R.T. Oehrle & E.A. Johnson
presented by Eric Schwarz

DESI V Workshop
Roma
July 14, 2013

ERNST & YOUNG
Quality In Everything We Do

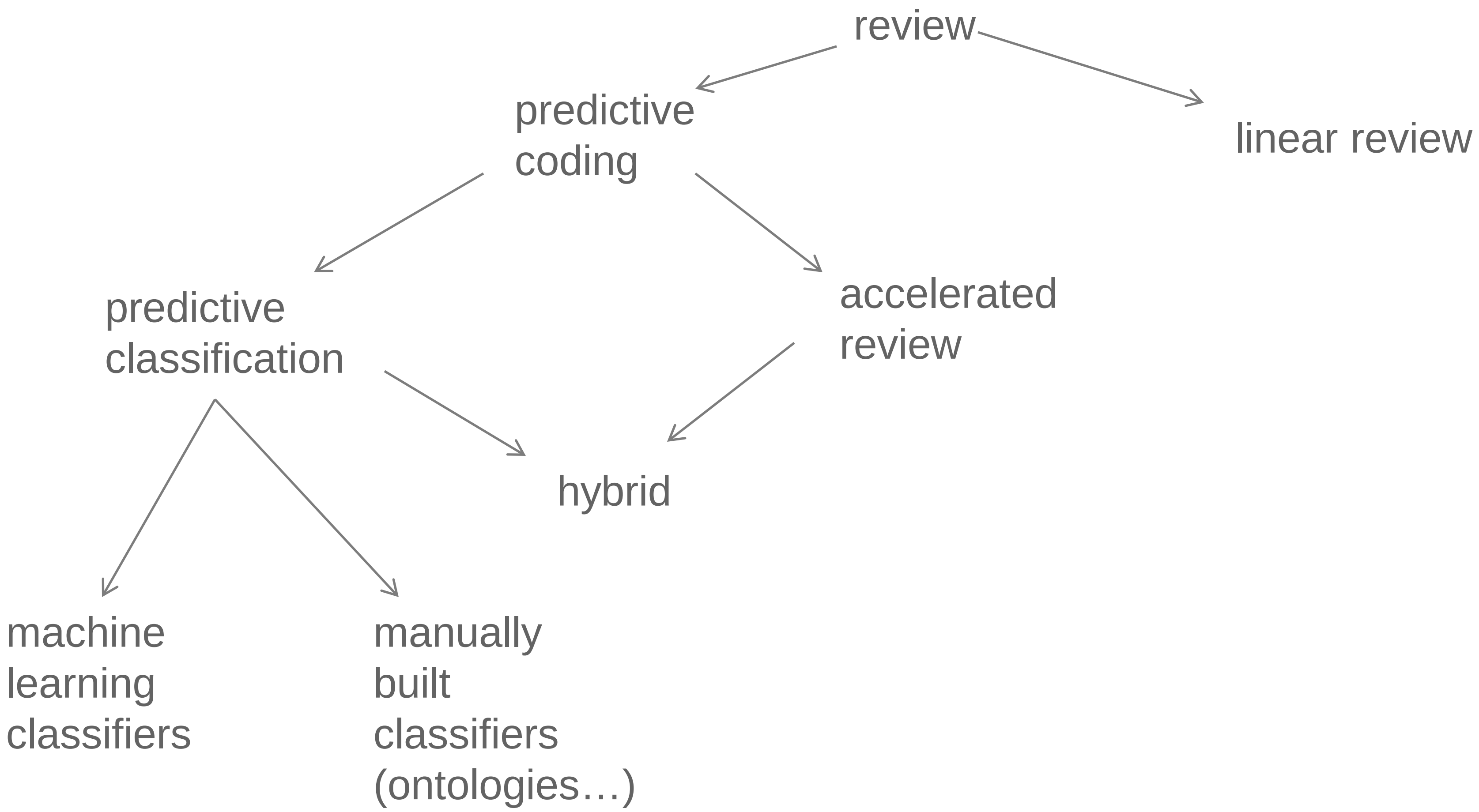
Why predictive coding?

- ESI volumes tending toward Big Data
- Empirical analysis of human review is not polishing its reputation
- Comparisons of human review and varieties of automated review has been favorable to the technology side
- Courts are looking to predictive coding as a solution to the problem of proportionality
- Core attribute of predictive coding:
  - mitigating the dependency of cost on volume
  - acceptable and improving quality

But what is predictive coding (PC), exactly?

- PC is TAR, but TAR need not be PC
  - Native review and native production are TAR, but not PC
  - Deduping by MD5 hash is TAR, but not PC (no coding!)
- Linear Review is neither TAR nor PC, but non Linear Review includes a variety of distinct forms (see below)
- PC can’t be characterized completely by its technological components, because the same technology can play different structural roles
- Coding (predictive or not) depends on the interplay between a document set and an RFP (subpoena, ...)
  - A single document set is likely to yield different productions in response to different RFP’s. (That is, productions are not determined by the document set alone.)

The basic landscape



review

predictive coding

linear review

predictive classification

accelerated review

hybrid

machine learning classifiers

manually built classifiers (ontologies…)

Structural distinctions

- Linear Review (LR) vs. PC
  - In LR, every doc is touched and coded by human review
  - In PC, not
- Accelerated Review (AR) vs. Predictive Classification
  - In AR (batch coding), every coding is based on human review, but not every doc is individually coded: sets of similar docs are coded.
  - Predictive Classification constructs a model of how each document is implicitly classified by the RFP
    - Model is based on a humanly coded sample
    - Projection across the document universe is tested and validated by sampling review
- Hybrid
  - Partial classification models (example: high precision non-responsive classifier, mixed with accelerated review of responsive)

Data structures, algorithms, process

- Data structures: choices for modeling the document set
  - document individuation, text, tokenization, indexing
  - vector models of documents
  - term-document matrix models (LSA)
  - topic models (PLSA, LDA)
- Machine learning algorithms
  - document *features* (based on data structures)
  - relevance *labels* (based on modeling the RFP)
  - choices for projection algorithms: linear regression, logistic regression, support vector machines (SVM's), classification and regression trees (CART's), ...
- Sampling: different choices (random, stratified, biased, ...) for different situations (initial sample, iterative sampling, validation of mixed populations,...)

The Structures of Predictive Coding

- Machine learning:
  - train/test/project on dataset only, without RFP: accelerated review
  - train/test/project on dataset AND RFP (via sampling review): classification
- Different structural configurations
- Each configuration offers an array of choices
- The same technology can play different structural roles
- Thus: predictive coding cannot be properly characterized on the basis of technology alone; structural configuration plays a distinguishing role
- (Further detail in the full paper)