Hogeschool van Amsterdam
Media, Creatie en Informatie

MACHINE LEREN VOOR E-DISCOVERY

Hans Henseler
Lector E-Discovery, HvA

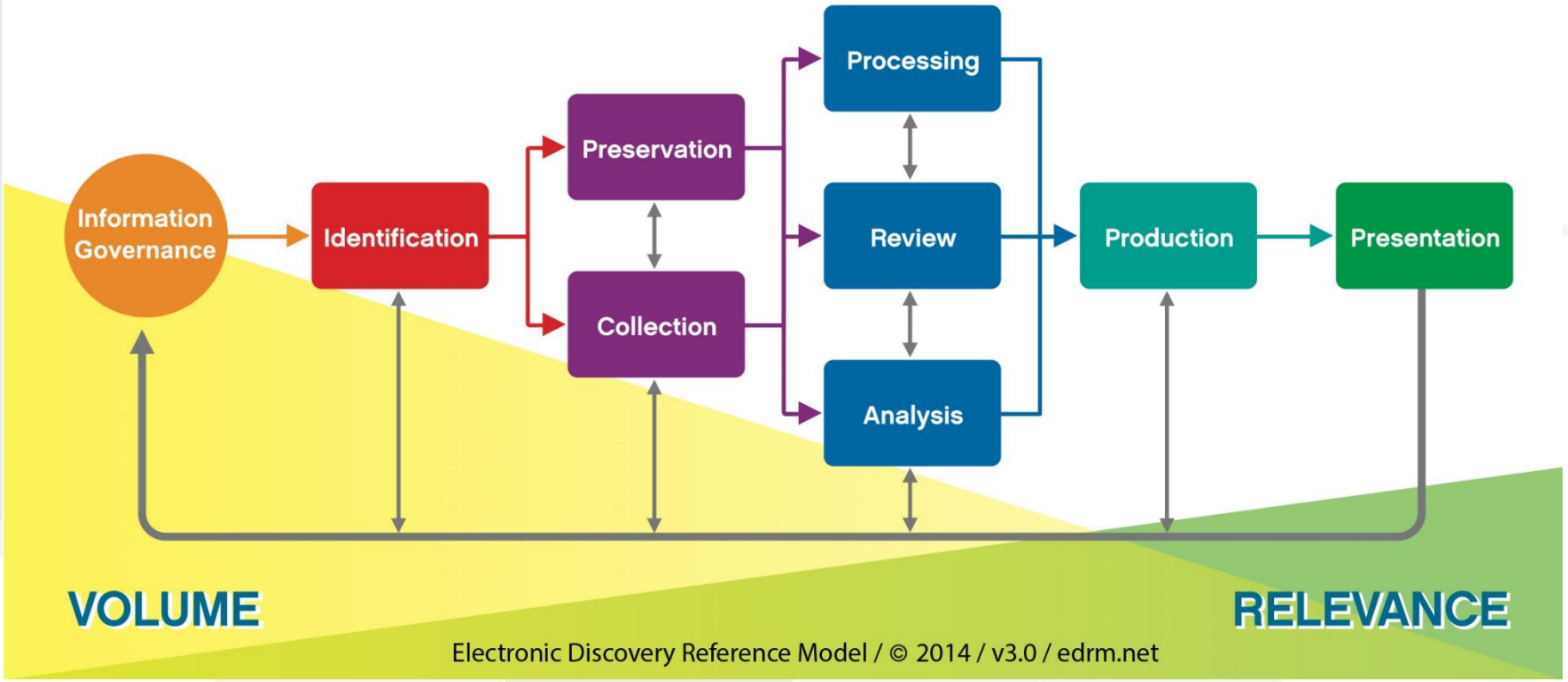
Symposium E-Discovery
Robotisering van Informatiemanagement

21 april 2016, Congrescentrum van de Gemeente Amsterdam

KENNISKRING E-DISCOVERY

Carla Bombeld
David van Dijk
Elmer Hoeksema
Annika Kuyper
Paul Rijnierse
Zaochun Ren (UvA)
David Graus (UvA)
Nina Kramer (eLaw)

Electronic Discovery Reference Model



EARLY CASE ASSESSMENT

Early case assessment (ECA) verwijst naar het inschatten van het risico (kosten in tijd en geld) om een juridische zaak aan te spannen of te verdedigen.

Grote organisaties worden met regelmaat verzocht electronische documenten beschikbaar te maken als gevolg van juridische discovery en disclosur verzoeken.

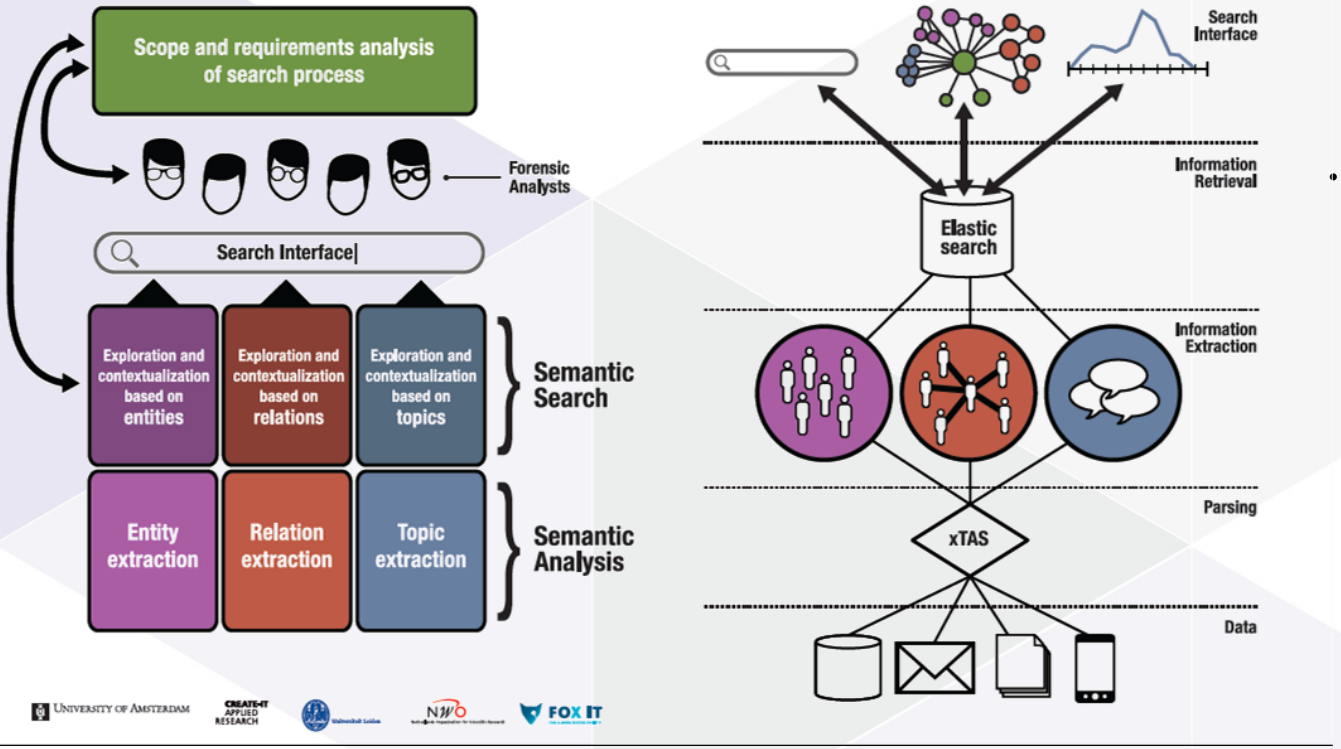
http://en.wikipedia.org/wiki/Early_case_assessment

EPLORATIEF ZOEKEN

- Onderzoekers weten niet precies waar ze naar op zoek zijn.
- Technologie kan gebruikers ondersteunen bij het verkrijgen van nieuwe inzichten in de data
- Bijvoorbeeld door interactieve zoekmogelijkheden aan te bieden
- Meta data helpt onderzoekers om zoekresultaten te filteren en in te zoomen op relevante delen.

SEMANTIC SEARCH IN E-DISCOVERY

Nina van der Knaap
David Graus
David van Dijk
Zhaochun Ren



Project met Prof. Maarten de Rijke, Universiteit van Amsterdam (UvA) met subsidie van NWO. 3 PhD studenten.

- Slimmer zoeken naar bewijs in een grote hoeveelheid digitale informatie, zoals email en documenten..
- Door semantisch te zoeken wordt er rekening gehouden met de betekenis van woorden in de contxt waarin ze gebruikt worden.

Lectoraat E-Discovery

https://ediscoverynl.dmci.hva.nl/

SEMANTIC SEARCH IN E-DISCOVERY

Relevant projects in the past years:

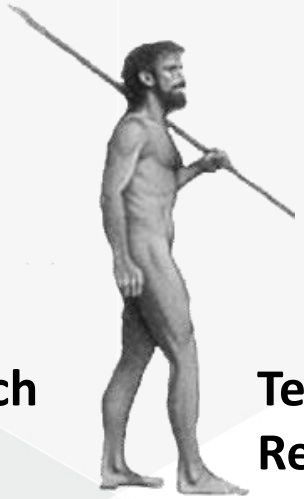
- Recipient recommendation (David Graus et al)
- Topic detection and clustering (Zaochun Ren et al)
- Semantic enrichment – YourHistory (David Graus et al)
- Who is involved (David van Dijk et al)
- Uforia Universal Forensic Indexer (Arnim Eijkhoudt et al)
- Time-aware Multi-viewpoint Summarization of Multi-lingual Social Tekst Streams (Zaochun Ren et al)
- Dynamic collective entity representations for entity ranking (David Graus et al.)
- Participation in Total Recall track from TREC (David van Dijk et al)

http://ilps.science.uva.nl/publications/

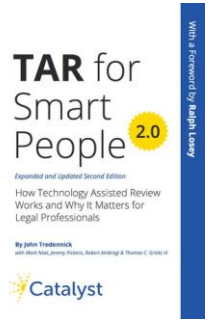
TECHNOLOGY ASSISTED REVIEW

Linear Review

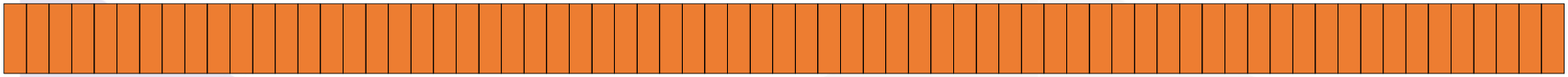
•Keyword Search
•Linear Review



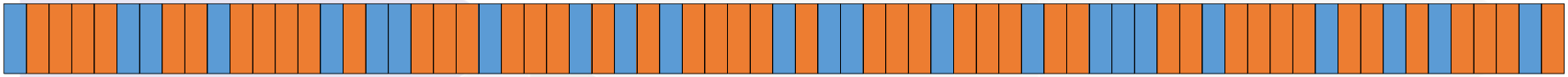
Technology Assisted Review (TAR)



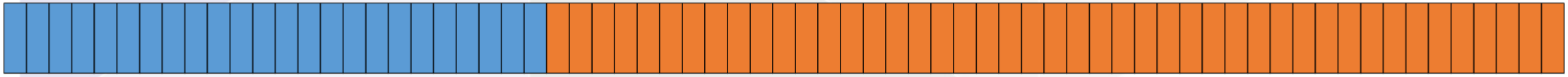
LINEAR VS PERFECT VS TAR



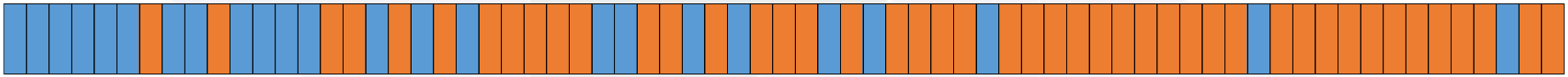
Linear review (random ranking):



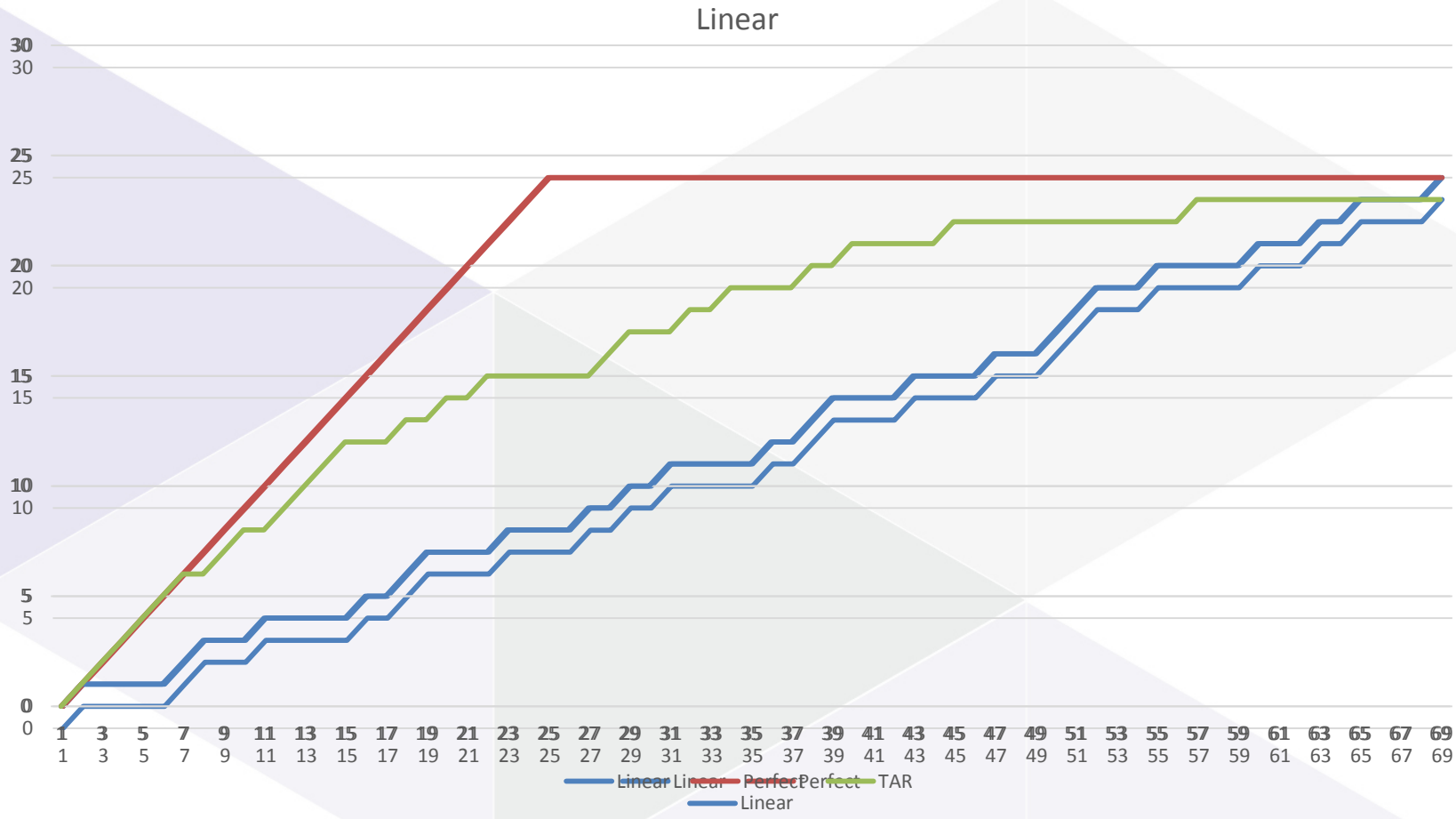
Perfect review (all relevant docs first):



Technology assisted review (improved ranking):



LINEAR VS PERFECT VS TAR

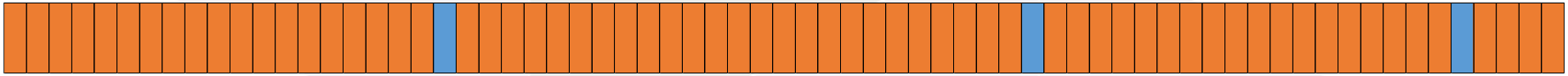


TTR: INSPIRED BY PREDICTIVE CODING (TAR 1.0)

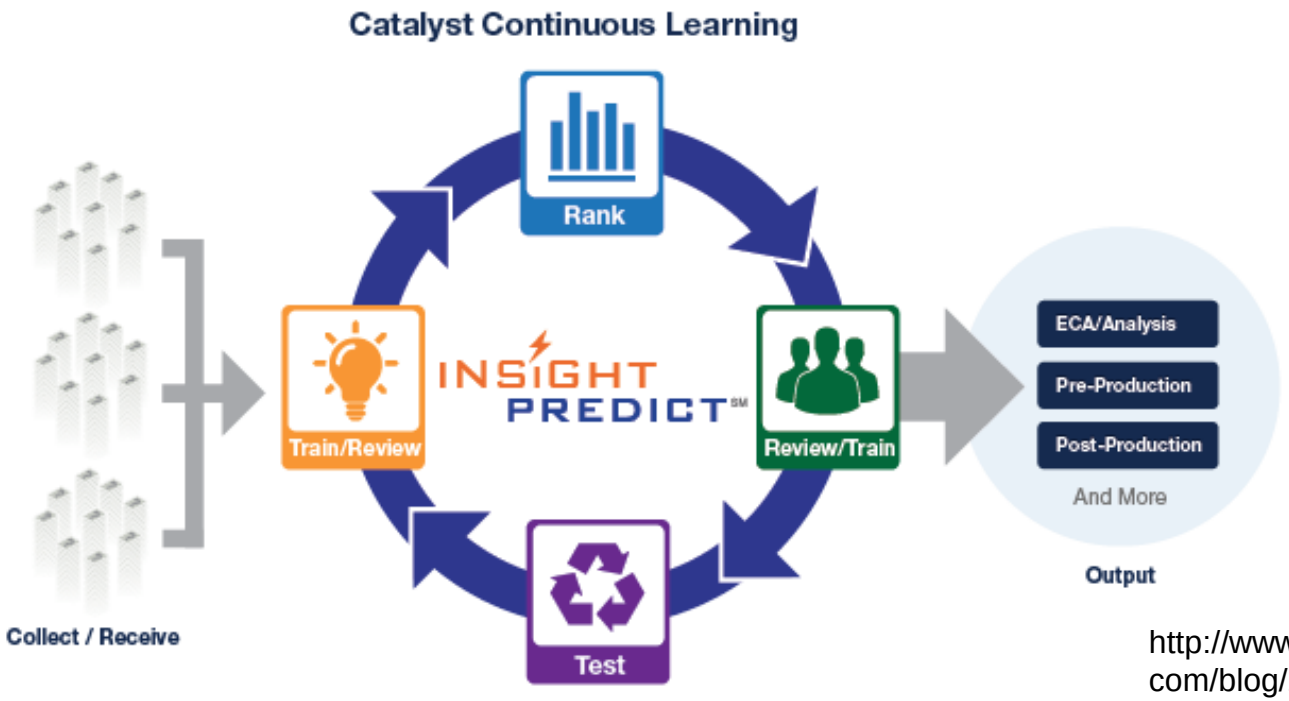
NADELEN VAN TAR 1.0

- Het machine leren model wordt maar één keer getrained
- Een senior advocaat moet de voorbeeld verzameling samenstellen:
  - Dat kost tijd
  - Omdat er een random selectie wordt gedaan zullen er relatief veel oninteressante documenten door de senior beoordeeld moeten worden
- Vaak komen documenten in batches binnen
- Het werkt niet met verzamelingen waarin weinig relevante documenten zitten

Verzameling met relatief weinig relevante documenten (low prevalence):



TTR: CONTINUOUS ACTIVE LEARNING (TAR 2.0)



http://www.catalystsecure.com/blog/2014/08/continuous-active-learning-for-technology-assisted-review-how-it-works-and-why-it-matters-for-e-discovery/

TREC 2015: TOTAL RECALL (TTR)

- Taak: Identificeer alle documenten in een corpus die relevant zijn voor een bepaald onderwerp, stuk voor stuk, met zo weinig mogelijk inspanning
- Doel: evalueer, door een gecontroleerde simulatie, methoden om een hoge recall (tegen de 100%) te krijgen met een mens in de loop.
- Hoe: testen worden uitgevoerd in een testomgeving waarin zoeksystemen getest kunnen worden zonder dat gevoelige data toegankelijk hoeven te worden gemaakt aan de deelnemers.
- De testomgeving werkt als een black box, zodat de deelnemers ook de zekerheid hebben dat hun eigen systemen niet makkelijk nagebouwd kunnen worden.

TTR: TEST COLLECTIONS

- Oefenmateriaal:
  - 20 Newsgroups Dataset (19K documents, three topics)
  - Reuters-21578 Test Collection (22K documents, four topics)
  - Enron dataset (724K emails, 2 topics)
- Competitie op eigen computers met downloads:
  - Jeb Bush emails (redacted) (290K emails,10 issues)
  - Illicit goods (465K web documents, 10 topics)
  - Local politics (902K articles, 10 topics)
- Competitie in de testomgeving van TRAC (op afstand):
  - Kaine email collection (402K emails, four categories)
  - MIMIC II Clinical dataset (32K patient ICU reports, 19 ICD codes)

Hogeschool van Amsterdam

Evangelos Kanoulas, UvA

Machine learning for E-Discovery