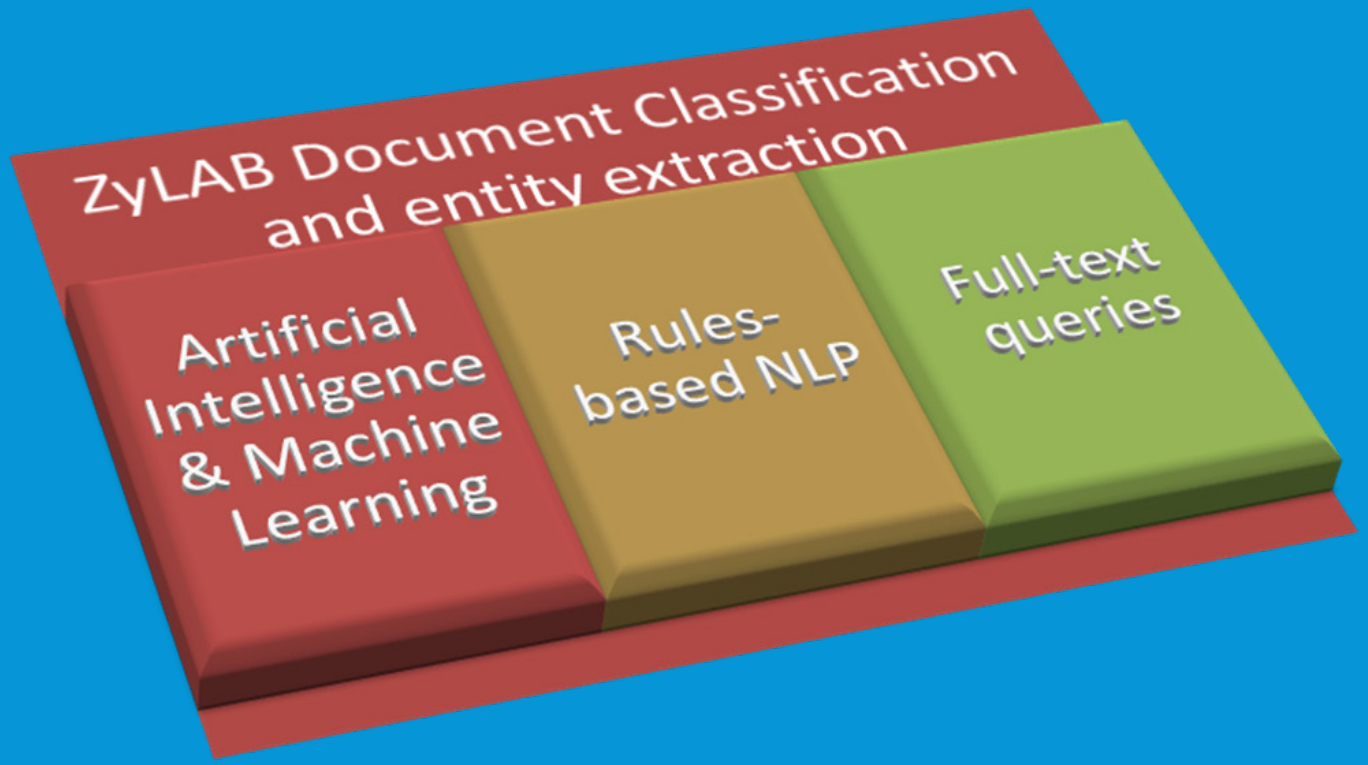
# ZyLAB Technology Assisted Review

## A Technical Guide

## INTRODUCTION

- Using techniques from Artificial Intelligence and Data Science in eDiscovery results in a true productivity revolution. It is however key to properly establish the defensibility of the entire process.
- To help you understand, trust and explain these technics, this eBook provides more technical insight in techniques that are used for review acceleration and Technology Assisted Review (TAR) in particular.
- This eBook follows the eBook “Artificial Intelligence for Your Daily Business” that provides a more basic overview of the different techniques.



## TOPICS THAT WILL BE DISCUSSED

- What is TAR?
- When do I need ZyLAB TAR?
- When does TAR not work for me?
- The Benefits of TAR.
- How does ZyLAB Technology Assisted Review (TAR) work?
- What are the differences with other TAR products in the market?
- Why and how is ZyLAB TAR defensible?
- How can one constantly measure the quality of ZyLAB's TAR?
- What is the best way to implement ZyLAB TAR in your law firm, in-house legal or internal investigation department?
- ZyLAB TAR Step by Step

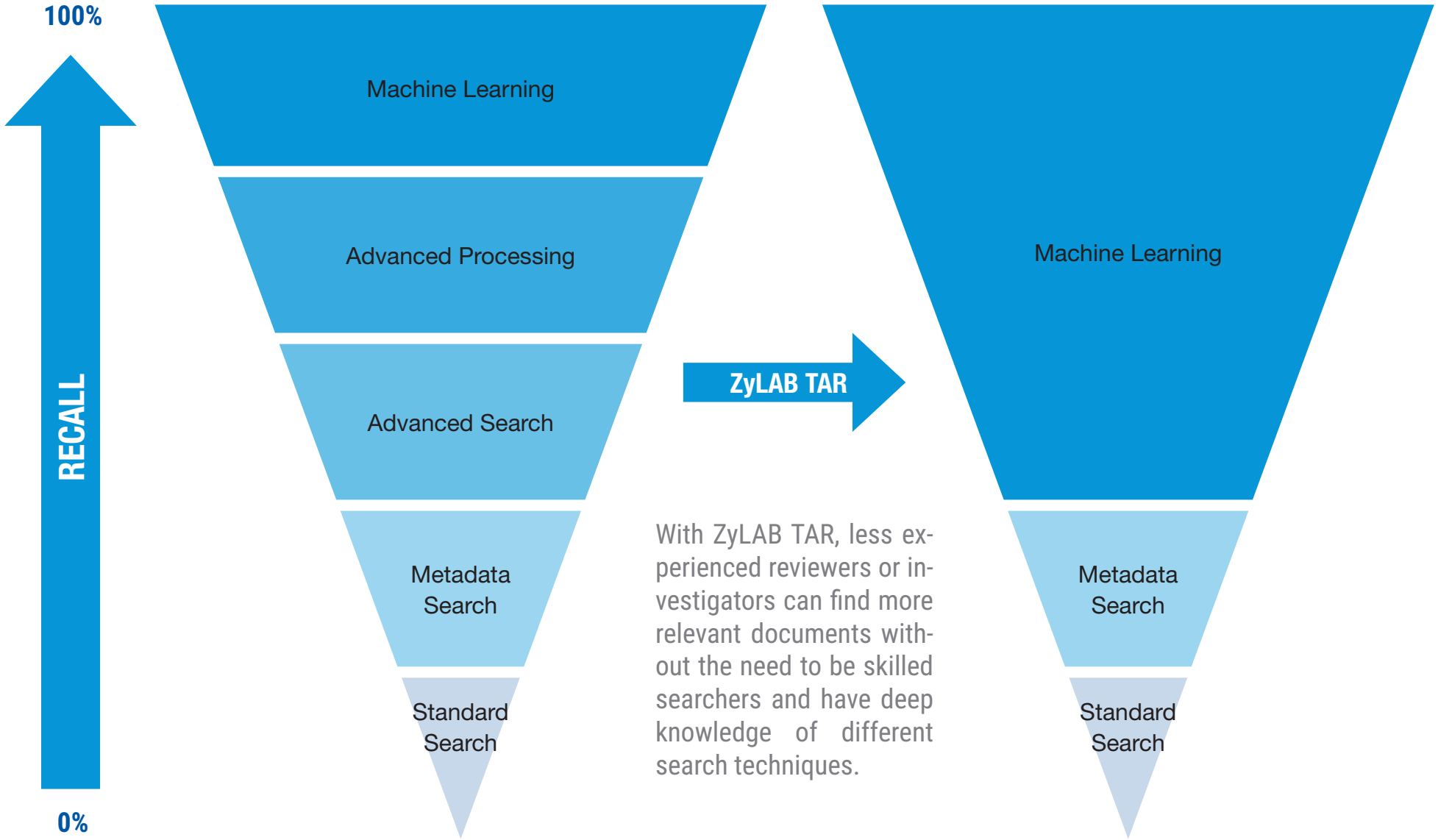
## WHAT IS TECHNOLOGY ASSISTED REVIEW (TAR)?

- TAR also known as Computer Assisted Review (CAR) or Predictive Coding, uses a series of algorithms to search and sort documents relevant for data investigation or eDiscovery. TAR also utilizes Machine Learning.
- With ZyLAB's machine learning, it is possible to teach the system to recognize specific document categories. This is done by providing the system with a number of positive and negative pre-labeled examples for each category.
- ZyLAB's machine learning uses the most advanced machine learning algorithms in combination with advanced search, statistical and semantic methods to represent the content of a document.

## WHEN DO I NEED ZYLAB TAR?

- eDiscovery: When you need to review a large data sets (>100.000 documents) to categorize documents into a number of categories such as responsive or not responsive for a set of claims or topics with a high degree of accuracy.
- M&A and Securitization: When you need to categorize a large set of documents into sets of conceptual categories, for instance the categories of a Virtual Data Room (VDR).
- Investigations: When you need to find as many relevant documents as possible in a large data set and you do not know exactly what words to look for or if your users do not have the skills to write complex search queries.

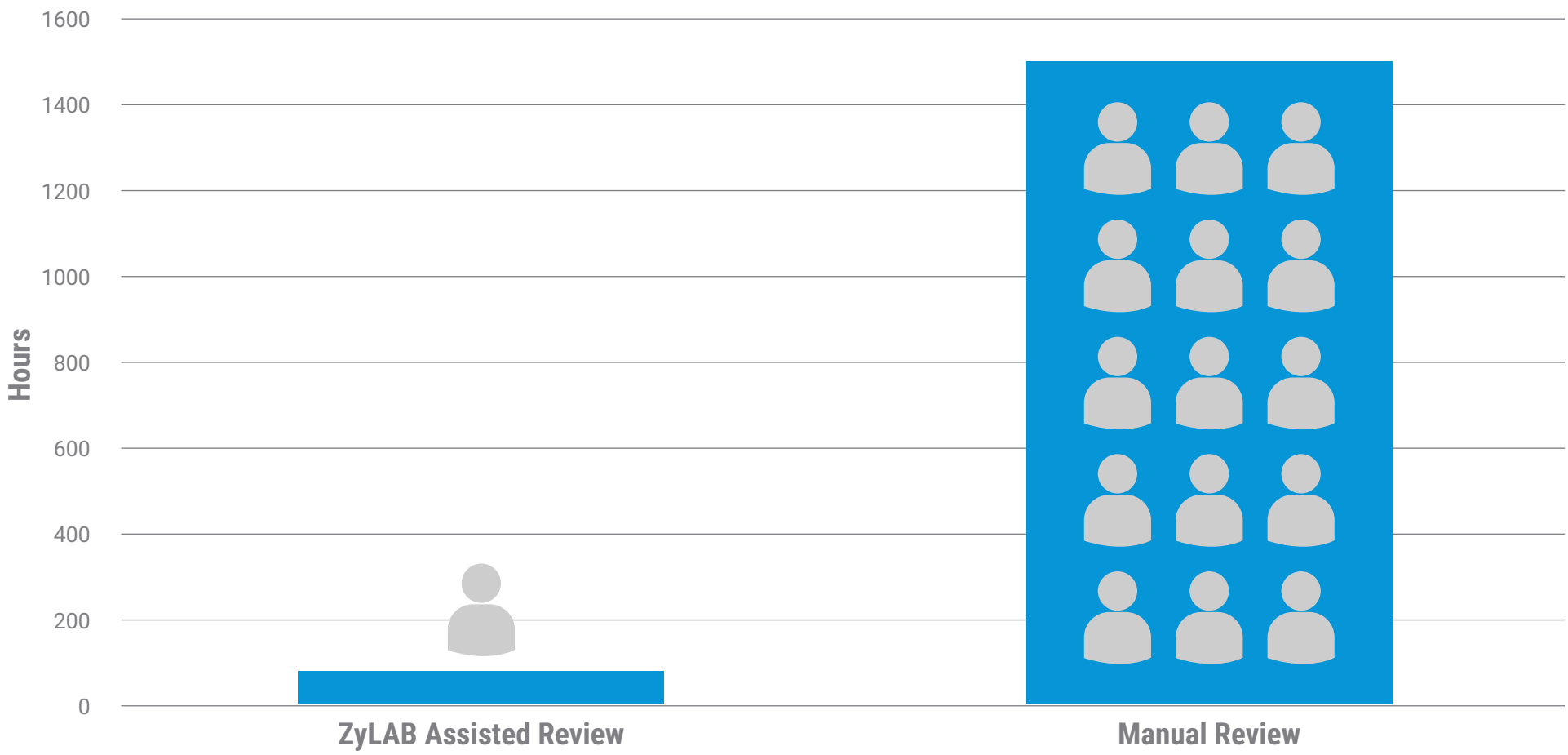
# TECHNOLOGY ASSISTED REVIEW HELPS TO BOOST SEARCH RECALL



## THE BENEFITS OF TAR

- Review more documents in less time with less resources.
- Saving tremendously on review cost.
- Have a higher quality review and more consistent tagging compared to human reviewers.
- Ability to measure recall and precision very accurate.
- Recall maximization: find 80% of all responsive documents quickly.
- Find relevant documents semi-automatically without depending on the search skills of end-users.
- Find without knowing exactly what you are looking for.
- Automatically classify documents in conceptual categories with a high degree of consistency and quality.

# HUGE ROI: TYPICALLY 15-20X MORE EFFICIENT



## WHEN DOES ZYLAB TAR NOT WORK FOR ME?

When the decision if a document is relevant or fits in a particular category is:

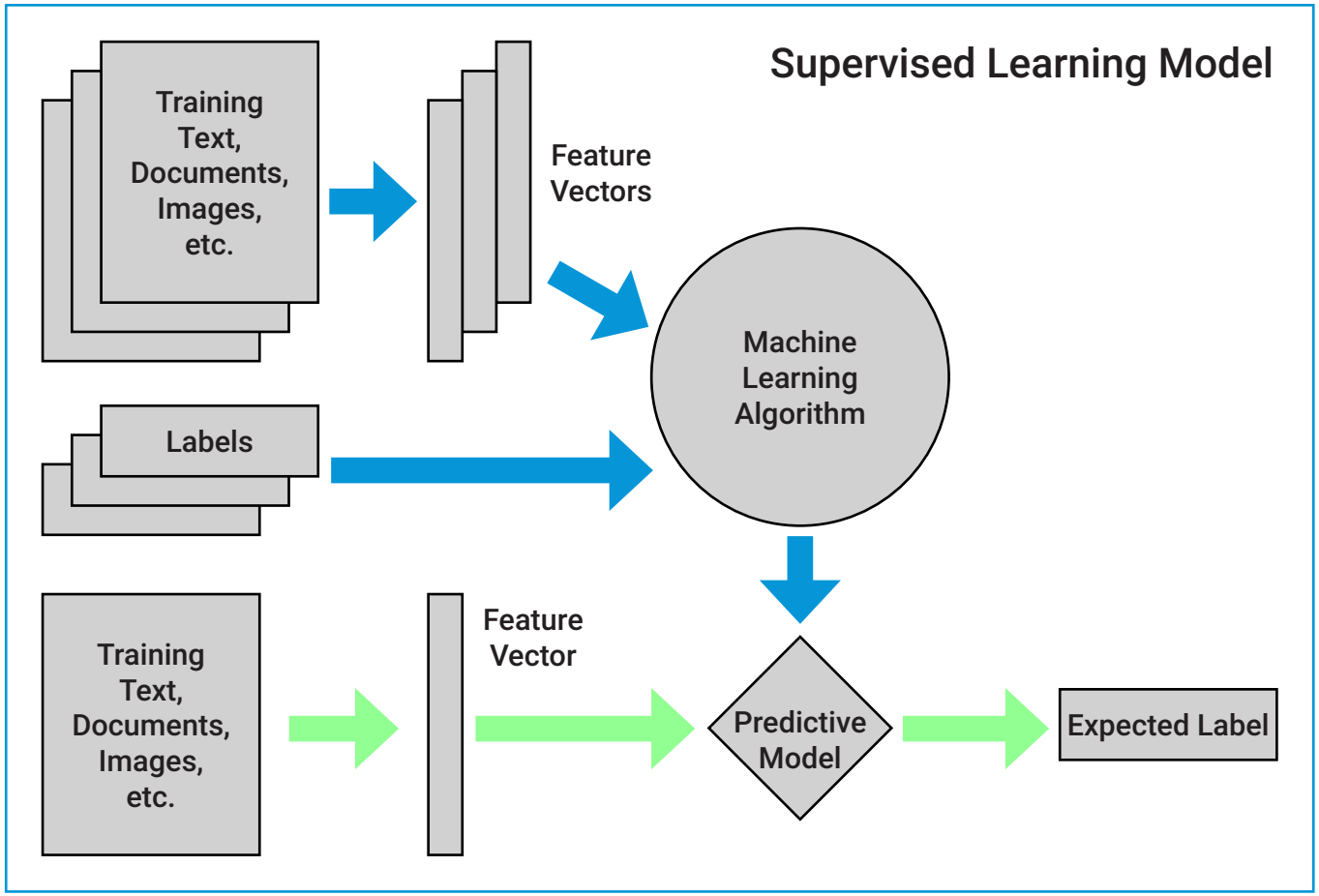
- Not depending on the document text;
- (Partly) based on numeric data and the size of numbers.

ZyLAB TAR works less well if you:

- Have to deal with different languages. Documents need to be sorted per language and TAR needs to be trained per language.
- Have to deal with very long documents.
- Have documents with very different language for similar documents.

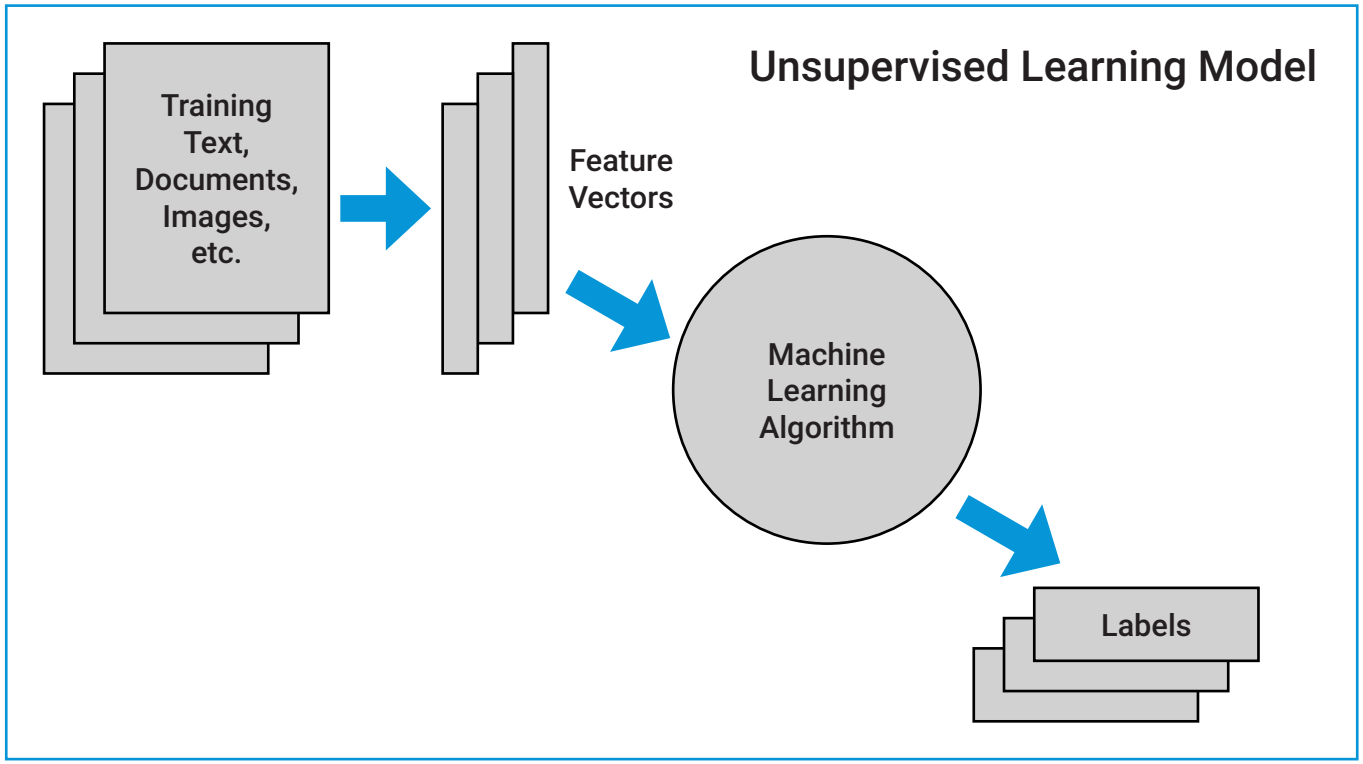
# HOW IS SUPERVISED MACHINE LEARNING USED FOR TEXT CLASSIFICATION?

- Documents are converted to feature vectors.
- Each set of documents has a label.
- The machine learning algorithm is trained to recognize the label (or category).
- After training, the classifier can be used to classify other documents and predict with a certain probability the label (category) of the document.



## WHAT ABOUT UN-SUPERVISED MACHINE LEARNING OR CLUSTERING?

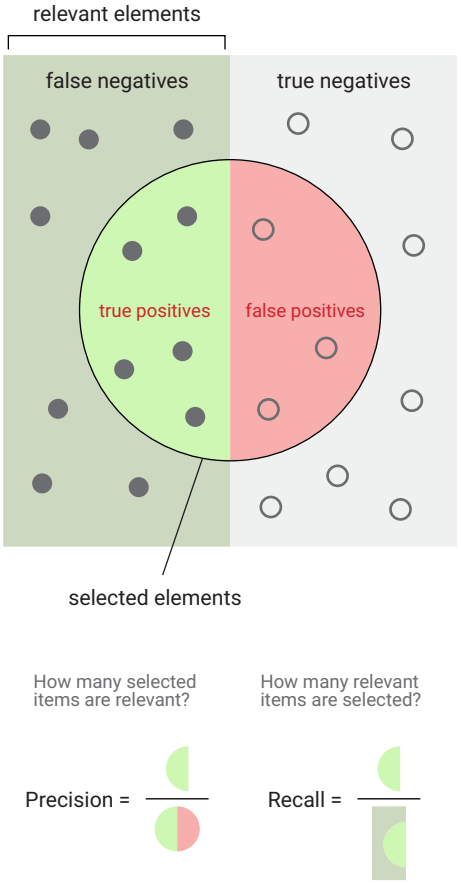
- Documents are converted to feature vectors.
- The unsupervised machine learning or clustering algorithm will recognize similar documents and concepts.
- These concepts can then be linked back to words which can be used as labels for the concepts or clusters.



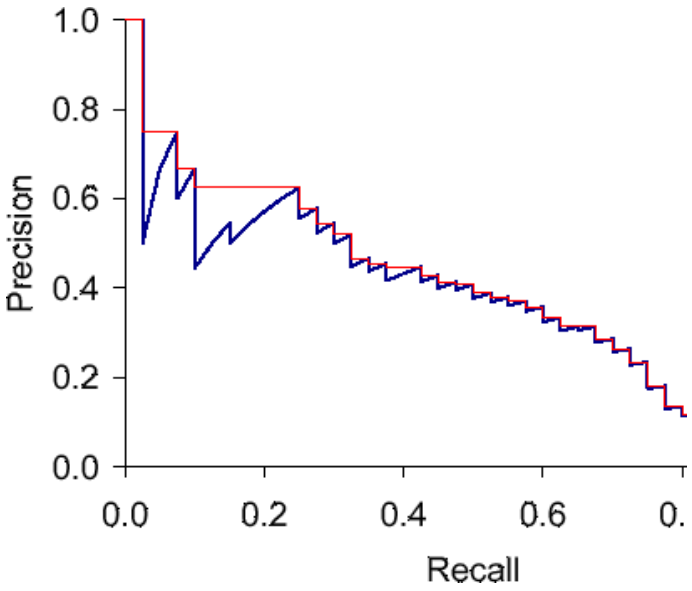
## TERMINOLOGY TO UNDERSTAND BEFORE WE CONTINUE

- ***Precision and Recall*** are measures used to indicate the quality of text-classification since the 1970's.
- ***10-fold cross validation*** is used to remove change from machine learning.
- ***Support Vector Machines (SVM)*** is a supervised machine learning technique.
- ***Non-Negative Matrix Factorization (NMF)*** is a unsupervised machine learning or clustering technique.

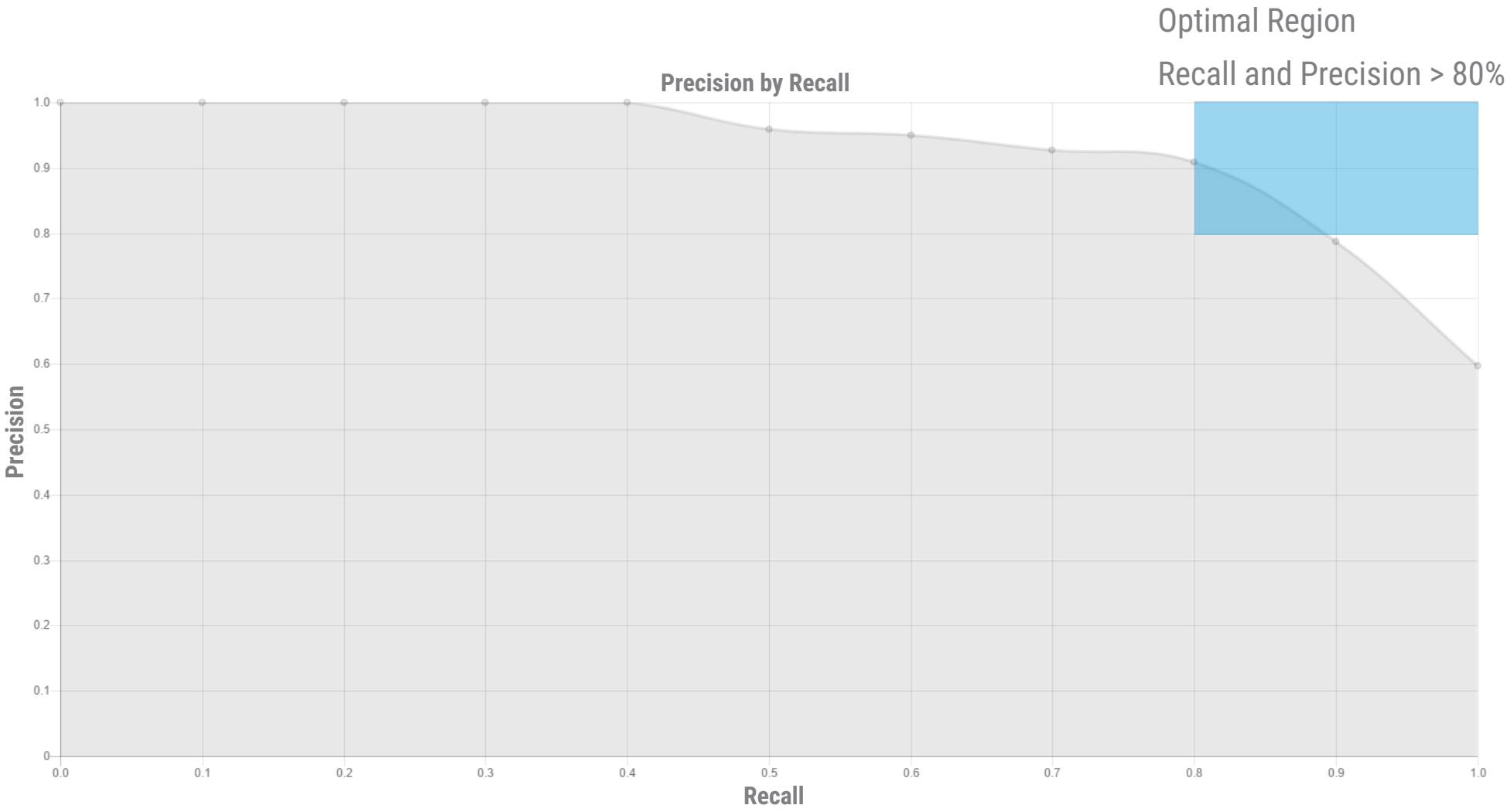
# WHAT ARE PRECISION/RECALL AND 11-POINT PRECISION/RECALL?




- Precision and recall are reverse proportional.
- Both can be manipulated and 100% precision or 100% recall are easy to obtain.
- The classifier returns a confidence value.
- We can use this confidence value to set a threshold to determine if a document is a match or not.
- When we have a ground truth: by changing this threshold, we can create 11 sets of recall from 0, 0.1, 0.2, … 0.9, to 1.0.
- For each of the recall values we can then measure the precision.
- We plot this in a so-called 11 points precision/recall graph.



## WHAT IS 11-POINT PRECISION/RECALL?

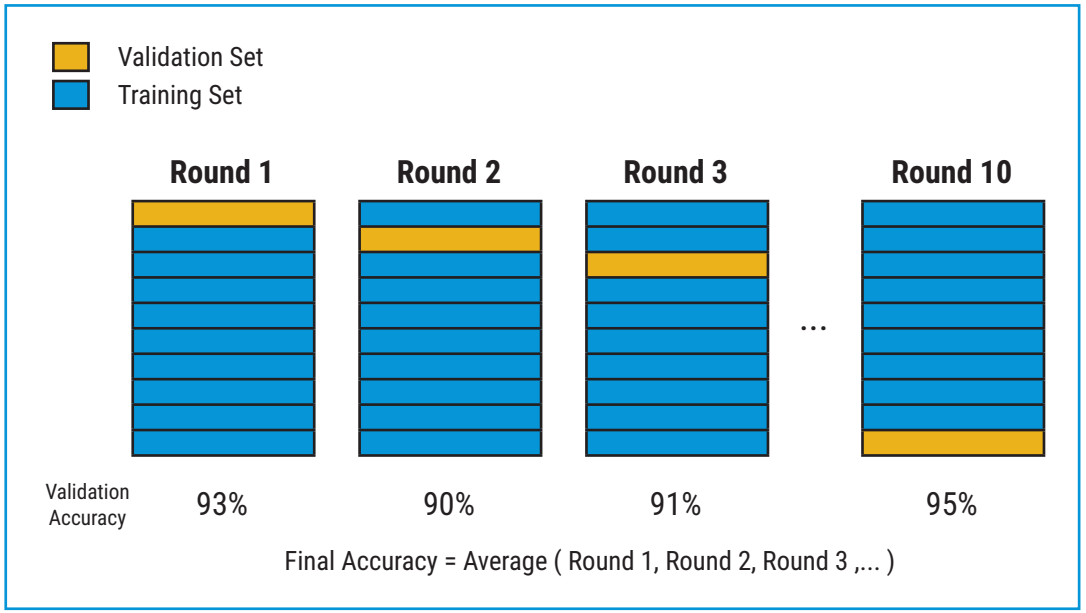



Goal is to have both values over 80%.

Humans perform at that level. Everything over 80% is often considered subjective.

# WHAT IS A 10-FOLD CROSS VALIDATION?

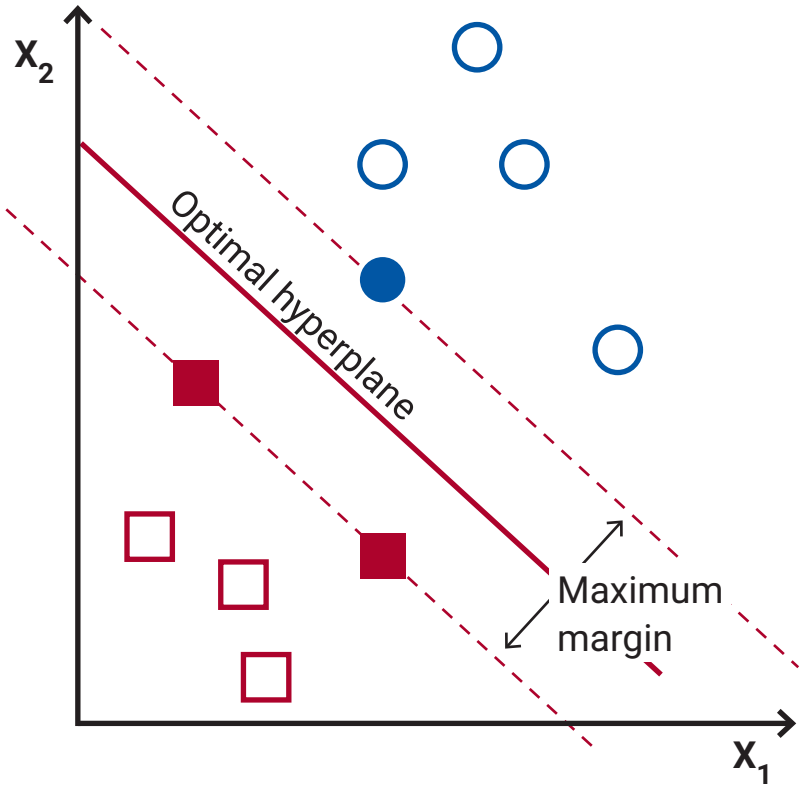
- To measure the performance of the classifier, cross validation can be used to reduce overfitting and make better use of the available data.
- With k-fold cross validation, the training data is randomly split into k subsets of equal size. k-1 subsets are used for training the classifier, the trained classifier is tested on the remaining subset. This is repeated k times using a different subset to test the classifier on. The results are averaged to remove change from the training process.



## WHAT IS A SUPPORT VECTOR MACHINE (SVM)?

A Support Vector Machine:

- Is the best performing algorithm for text classification which has consistently outperformed all other text-classification algorithms over the years.
- Has automatic feature selection, contrary to Bayes Classifiers.
- Is very robust against wrong training documents, contrary to decision trees.
- Needs a relatively small number of training documents (100-1000) compared to deep learning (easily needs millions of training documents).




A SVM finds the optimal hyperplane with a maximum margin to seperate within-class training documents from out-of-class training documents.

## WHAT IS A BINARY TEXT CLASSIFIER?

- A Binary Classifier can only detect if a document belongs to a category or class or not.
- The classifier returns a confidence value indicating how certain the classifier is that a document belongs to a class or not.
- For each category, a separate classifier is trained.
- When new documents are trained, they are matched against every classifier. Each classifier return a confidence value. A threshold value is used and only classifiers for which the confidence value is larger than the threshold are classified in that category.
- When the confidence value of multiple classifiers exceeds the thereshold, a document can get multiple category labels.

## WHAT IS TF-IDF?

- In information retrieval, tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- The tf-idf value increases proportionally to the number of times a word appears in the document, but is often offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.
- Nowadays, tf-idf is one of the most popular term-weighting schemes.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurences of $i$ in $j$

$df_i$ = number of documents containing $i$

$N$ = total number of documents

# WHAT IS A SEMANTIC DOCUMENT REPRESENTATION?

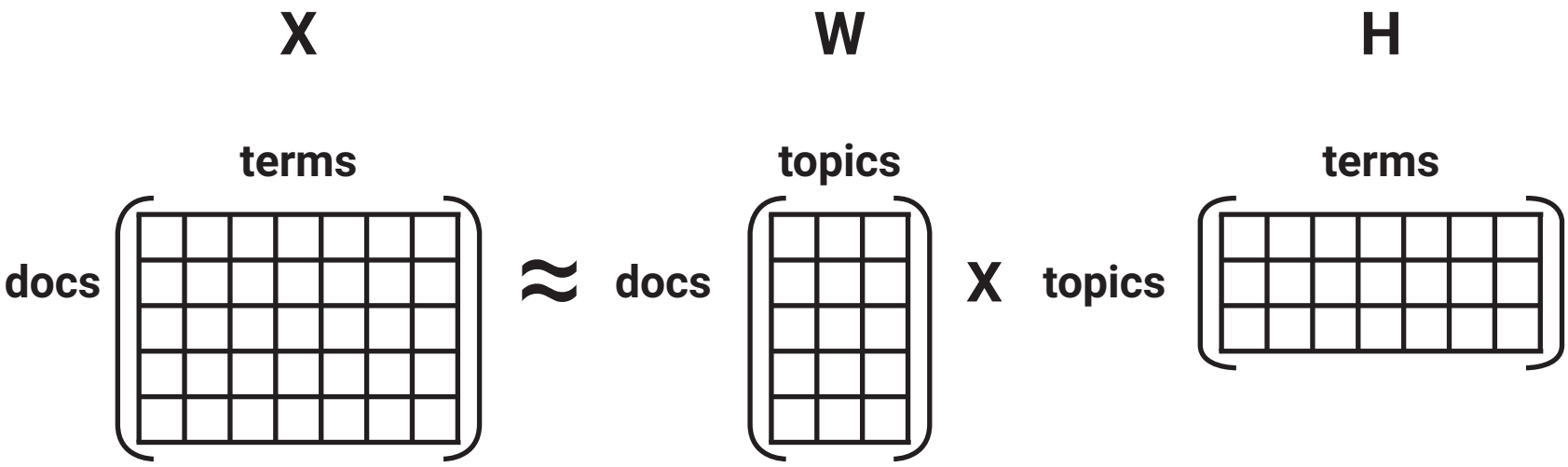
- Semantics are the branch of linguistics and logic concerned with meaning.
- Meaning can be obtained by extracting relevant entities, facts, patterns, events, but also sentiments and emotions.
- Semantic document representations focus on only these patterns instead of on the entire text.
- See US 9,235,812 Patent for more information on ZyLAB's semantic document representations for eDiscovery.
- Using semantic document representation results in more focussed and much faster machine learning.

| | |
|---|---|
| Language_Name | English |
| CITY | New Brunswick, WASHINGTON |
| COMPANY | J&J, Johnson & Johnson |
| COUNTRY | Greece, Poland, Romania, United Kingdom |
| CURRENCY | .02 USD, 21400000 USD, 48600000 USD, 59.47 USD, 70000000 USD |
| DATE | 04-08 |
| DAY | |
| NOUN_G | |
| ORGANIZ | |
| PEOPLE | |
| PERSON | |
| PLACE_F | |
| PRODUC | |
| PROP_MISC | Band-Aids, Food Program, Foreign Corrupt Practices Act, United Nations Oil |
| STATE | N.J. |
| TIME | 1:32 pm ET |
| TIME_PERIOD | 13 years, five years, six months, three years |
| YEAR | 2007 |
| Problem | "We went to the government to report improper payments and have taken full responsibility for these actions," said William Weldon, Chairman and CEO of J&J., Last month federal health regulators took legal control of the plant where millions of bottles of defective medication were produced, The charges against J&J were brought under the Foreign Corrupt Practices Act, which bars publicly traded companies from bribing officials in other countries to get or retain business, The company will pay $21.4 million in criminal penalties for improper payments and return $48.6 million in illegal profits, according to the govemment., The SEC says J&J agents used fake contracts and sham companies to deliver the bribes. |
| Sentiment | giving meaningful credit to companies that self-report, We are committed to holding corporations accountable for bribing foreign officials, what is honest |
| Request | make sure it complies with anti-bribery laws across its businesses |

| | |
|---|---|
| CITY | New Brunswick, WASHINGTON |
| COMPANY | J&J, Johnson & Johnson |
| COUNTRY | Greece, Poland, Romania, United Kingdom |
| CURRENCY | .02 USD, 21400000 USD, 48600000 USD, 59.47 USD, 70000000 USD |

## WHY NMF (NON-NEGATIVE MATRIX FACTORIZATION)?

Contrary to other popular clustering algorithms such as Latent Semantic Indexing (LSI), PLSA and LDA, NMF does not allow negative factorizations, which would essentially mean that one could have a negative occurrence of words in a document. Therefore, NMF clusters are more meaningful when used for text-clustering and concept search.



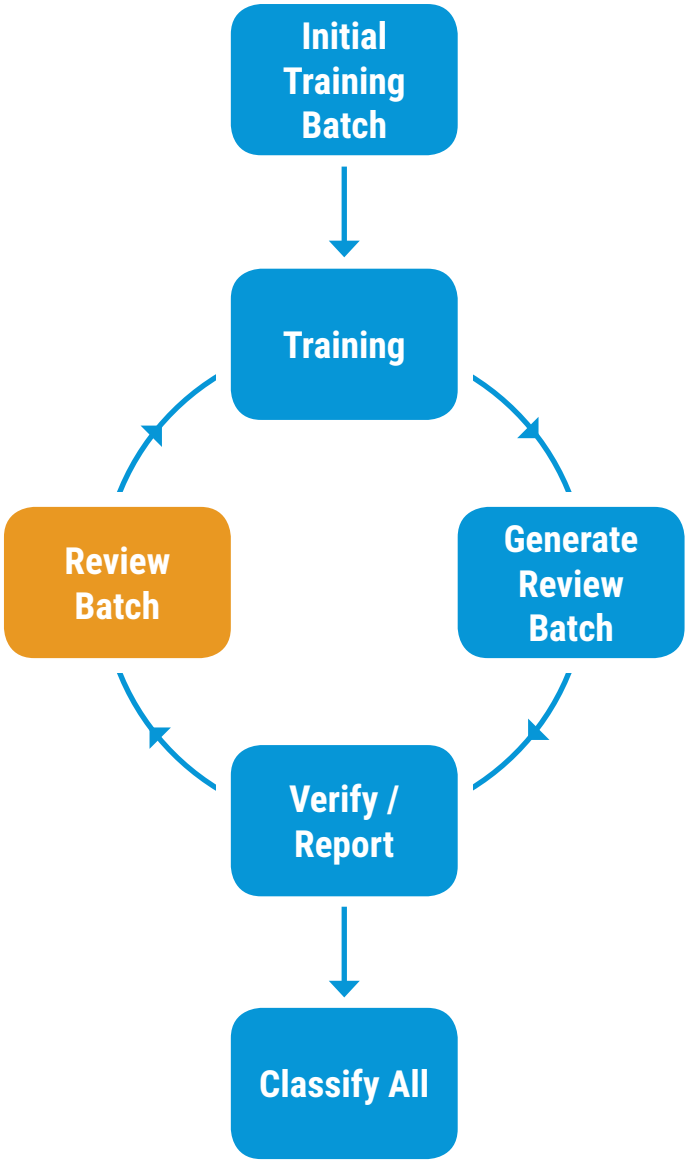
## USED TECHNOLOGIES BY ZYLAB TAR

| | |
|---|---|
| Protocols supported | Random Start (TAR 1.0), Search Start (TAR 2.0) and Start with Topic Modeling (TAR 3.0) or combine all methods (TAR 4.0) |
| Machine Learning and Topic Modeling Algorithms | Support Vector Machines (SVM) and Non-Negative Matrix Factorization (NMF) |
| Classifier type | Binary |
| Document Representation | Term Frequency–Inverse Document Frequency (TF-IDF) on full-text or on extracted semantic document features * (entities) |
| Evaluations | 11-point precision/recall measurements in combinations with 10-fold cross validation |

* Patented by ZyLAB under US 9,235,812 Patent

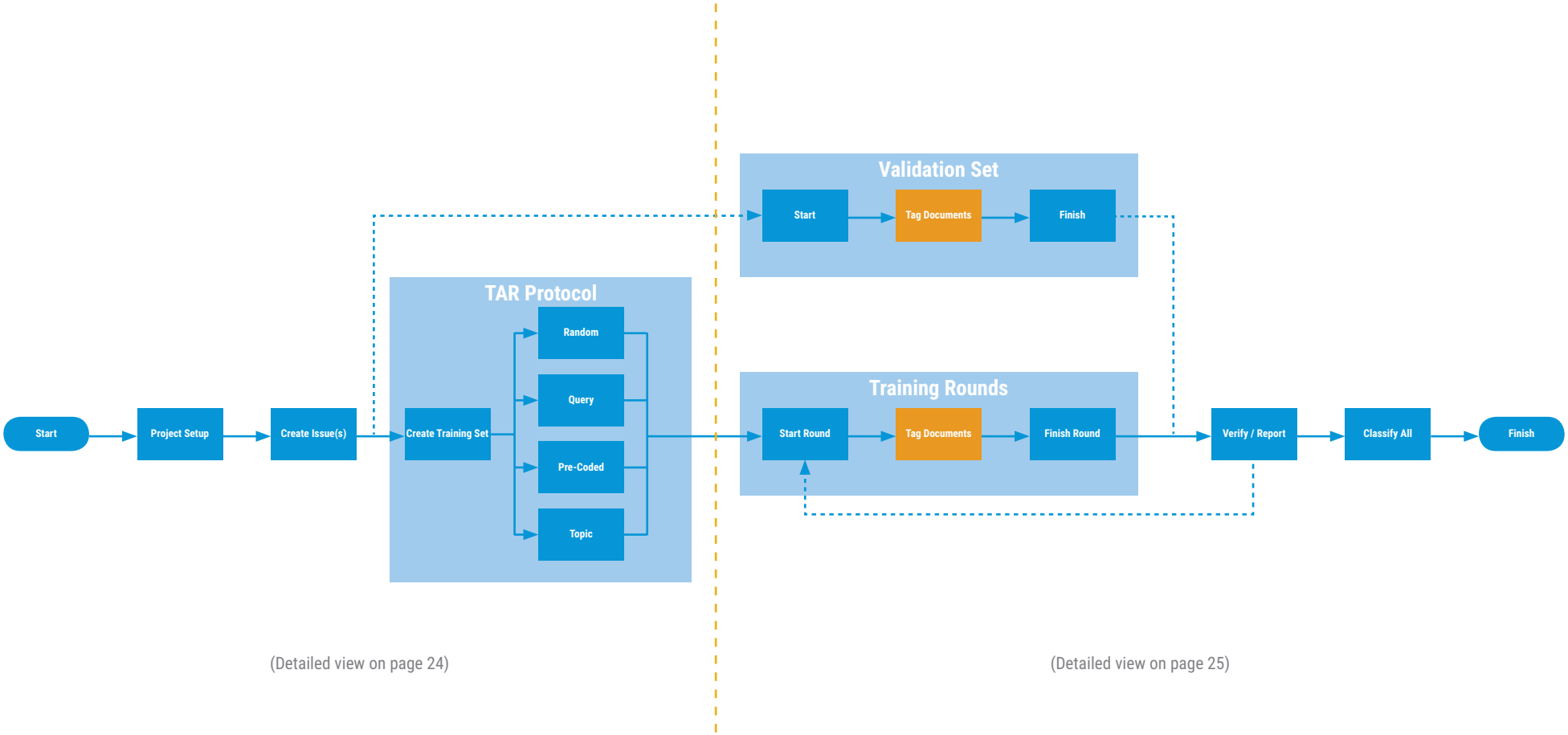
## HOW DOES ZYLAB TAR WORK?

1. Select a subset of documents in a matter by using full-text search or meta data selections.
2. Determine your topics of interest, which we call Issues.
3. For each of the issues, compile a small set of training documents. This can be done by using a validation set (TAR 1.0), full-text search (Tar 2.0), topic modeling (TAR 3.0) or any combination (TAR 4.0).
4. Train a classifier for each issue with 90% of the training documents. Test the classifier with the remaining 10% of the training documents.
5. Find more relevant document by matching the classifier against the rest of the data set.
6. Present the highest matching documents to the user for review.
7. After review the number of relevant documents will be larger and we can train a new (better) classifier in step 4.
8. Repeat this until a stop condition has been reached.
9. At the end, create a defensibility report, do additional sampling on the quality and classify all remaining documents automatically.




The TAR-process explained in a video: https://zylab.com/resources/videos/

# ASSISTED REVIEW WORKFLOW (1)

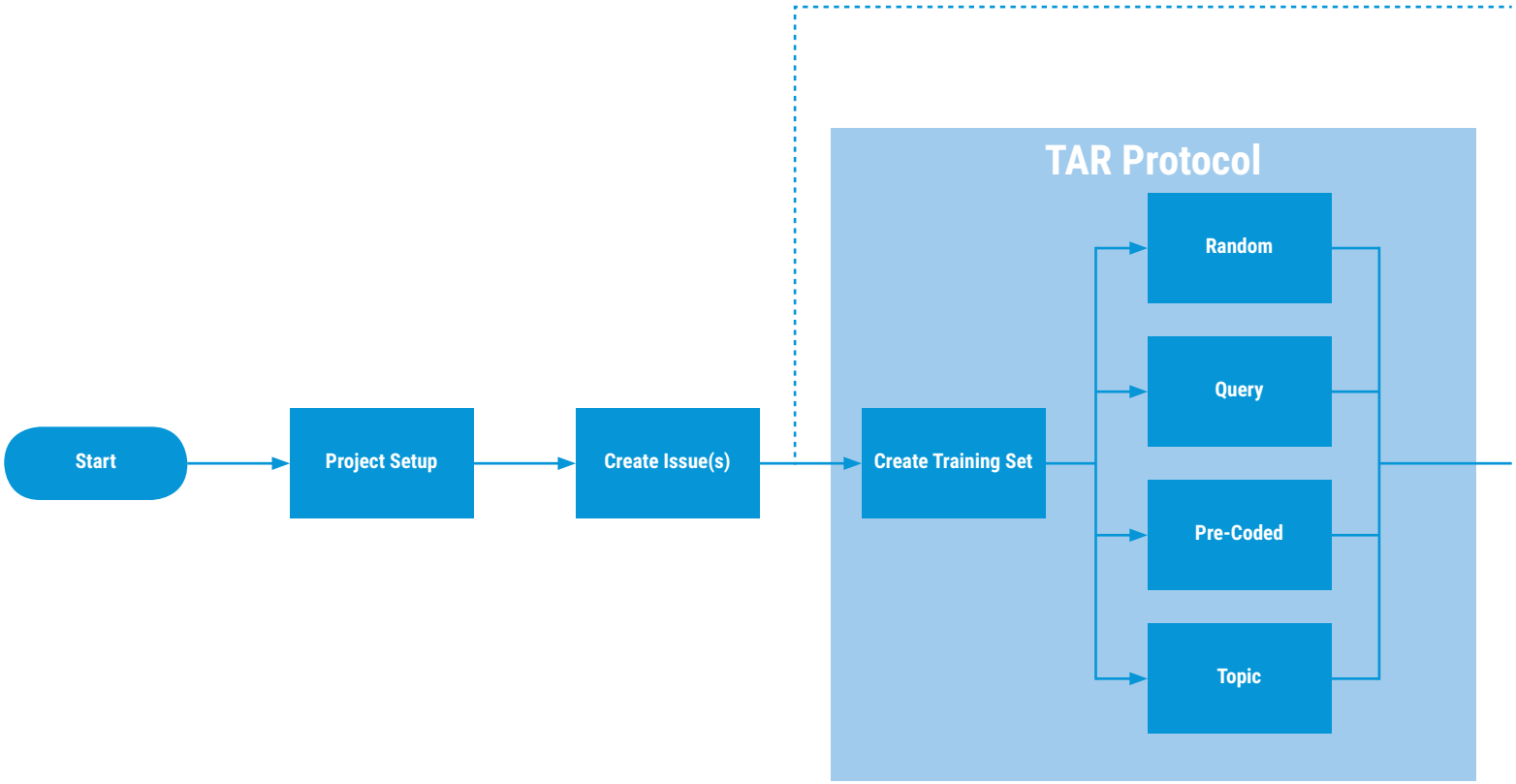



(Detailed view on page 24)

(Detailed view on page 25)

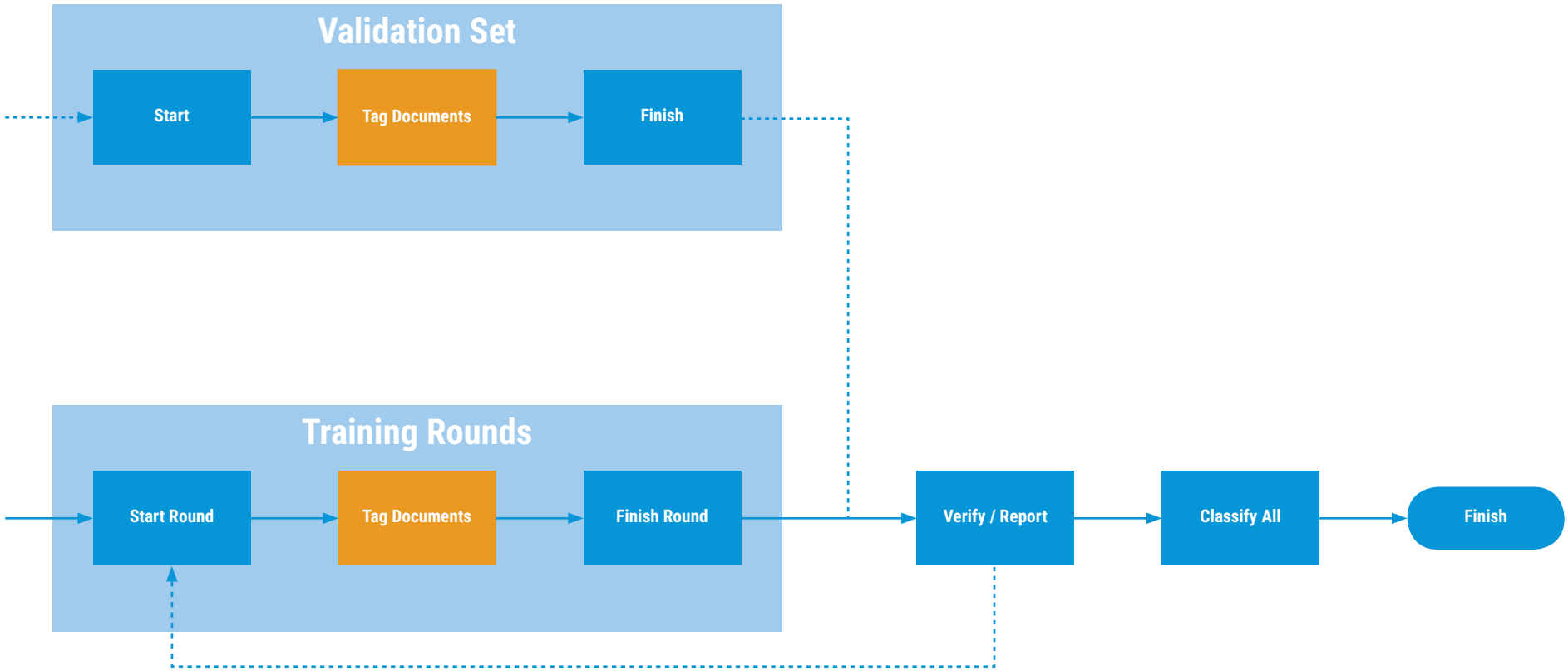
The TAR-process explained in a video: https://zylab.com/resources/videos/

# ASSISTED REVIEW WORKFLOW (2)



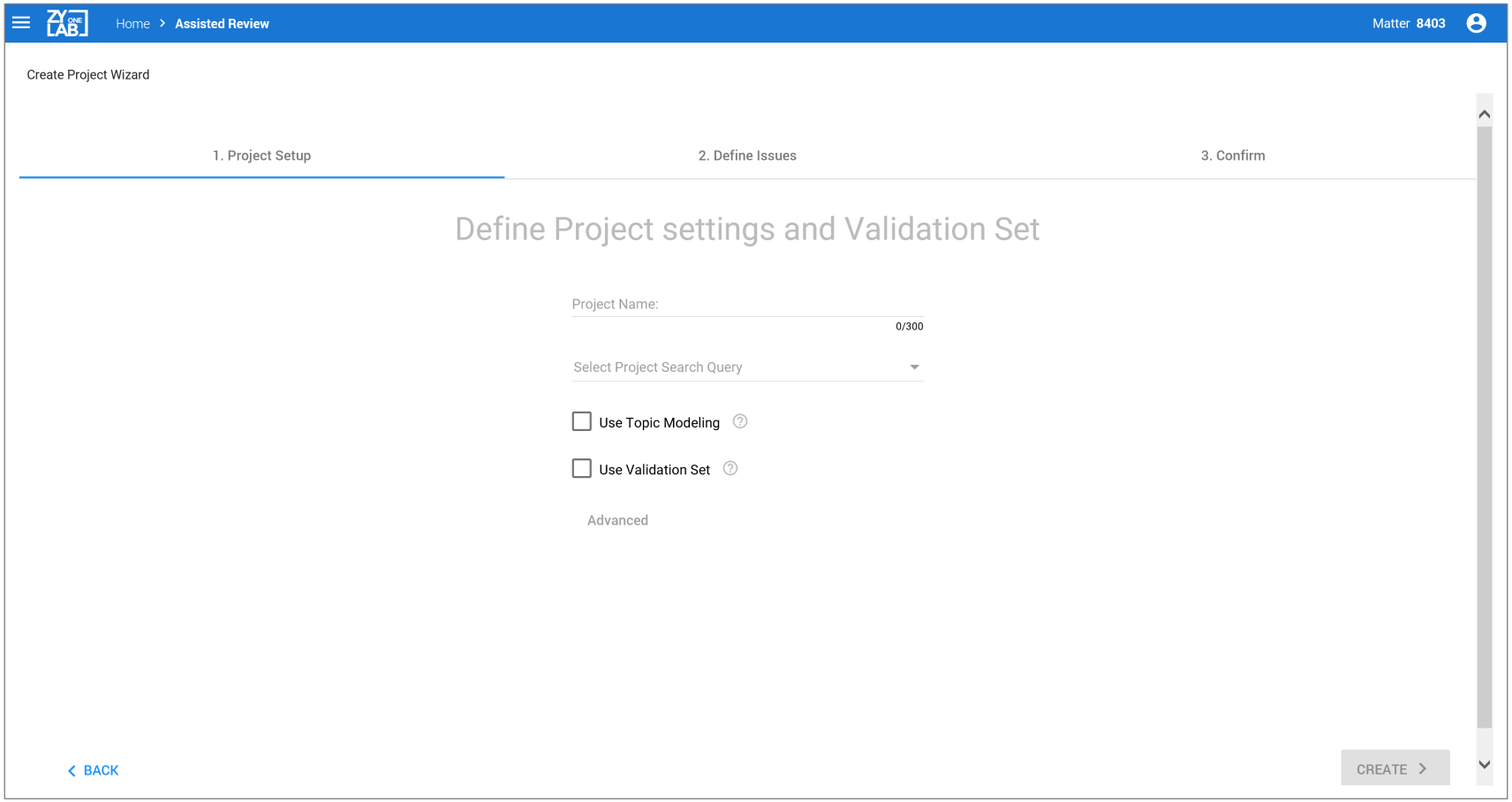

The TAR-process explained in a video: https://zylab.com/resources/videos/

# ASSISTED REVIEW WORKFLOW (3)

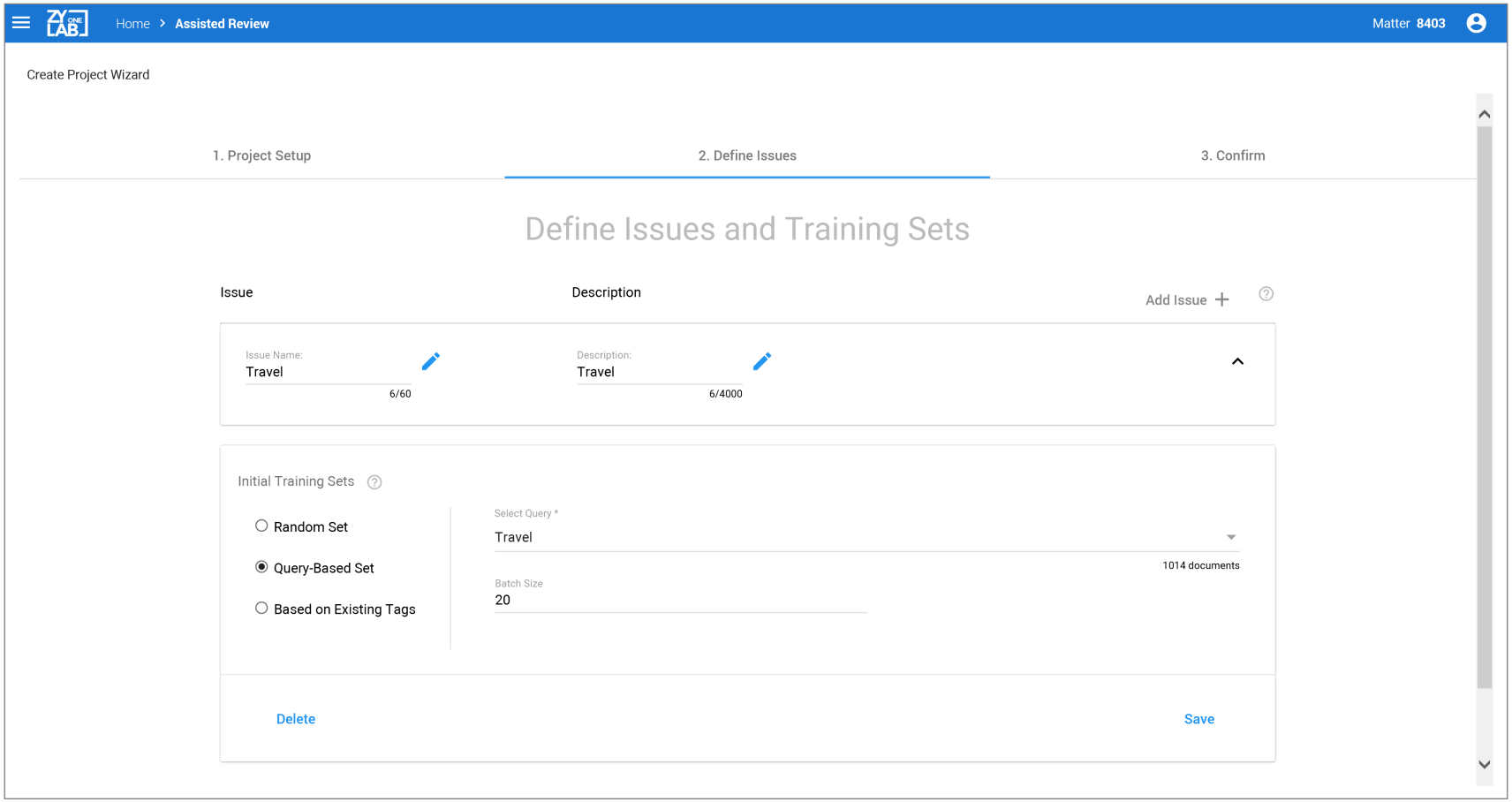



The TAR-process explained in a video: https://zylab.com/resources/videos/

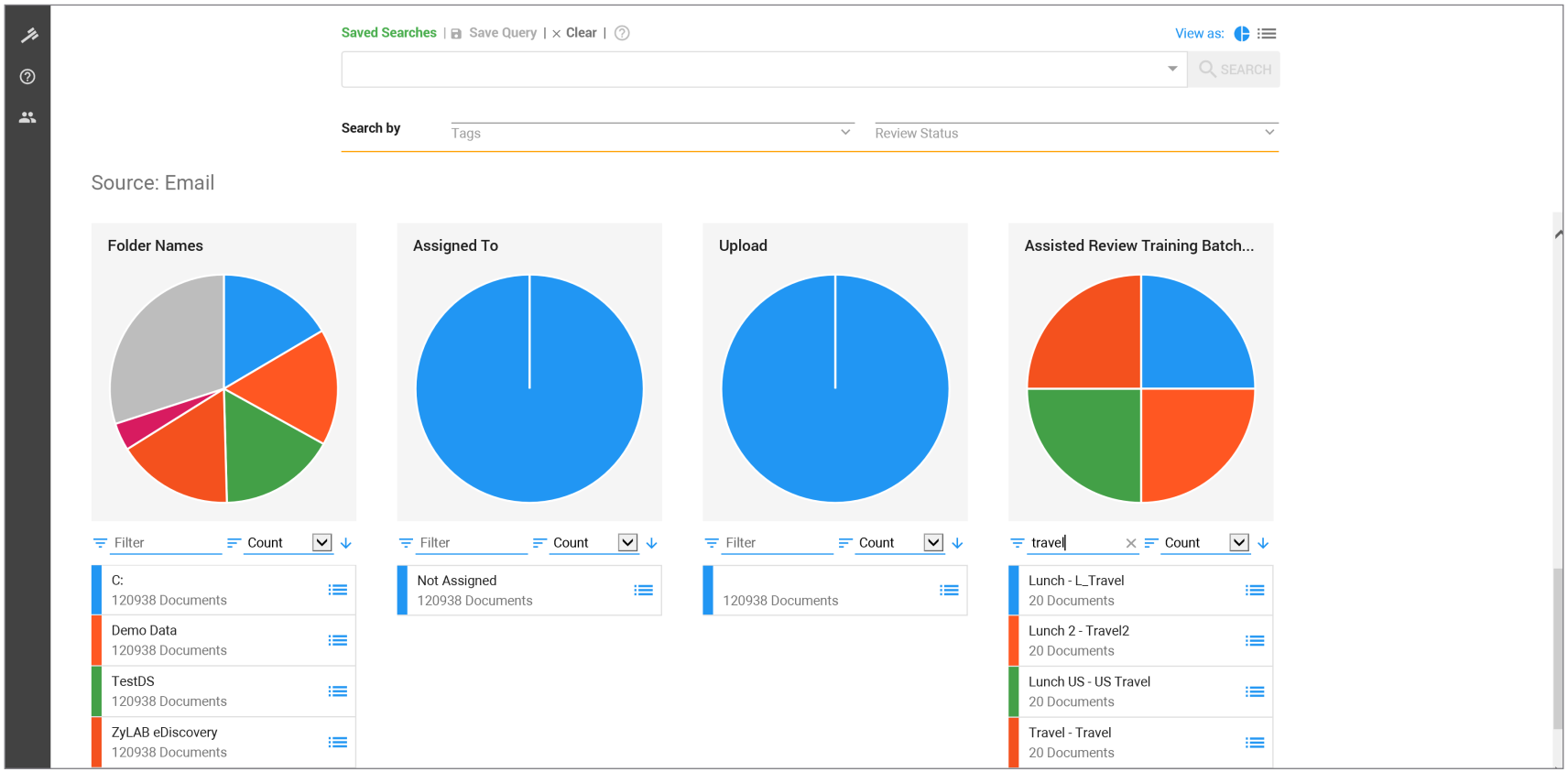
# DO IT YOURSELF: CREATE A PROJECT



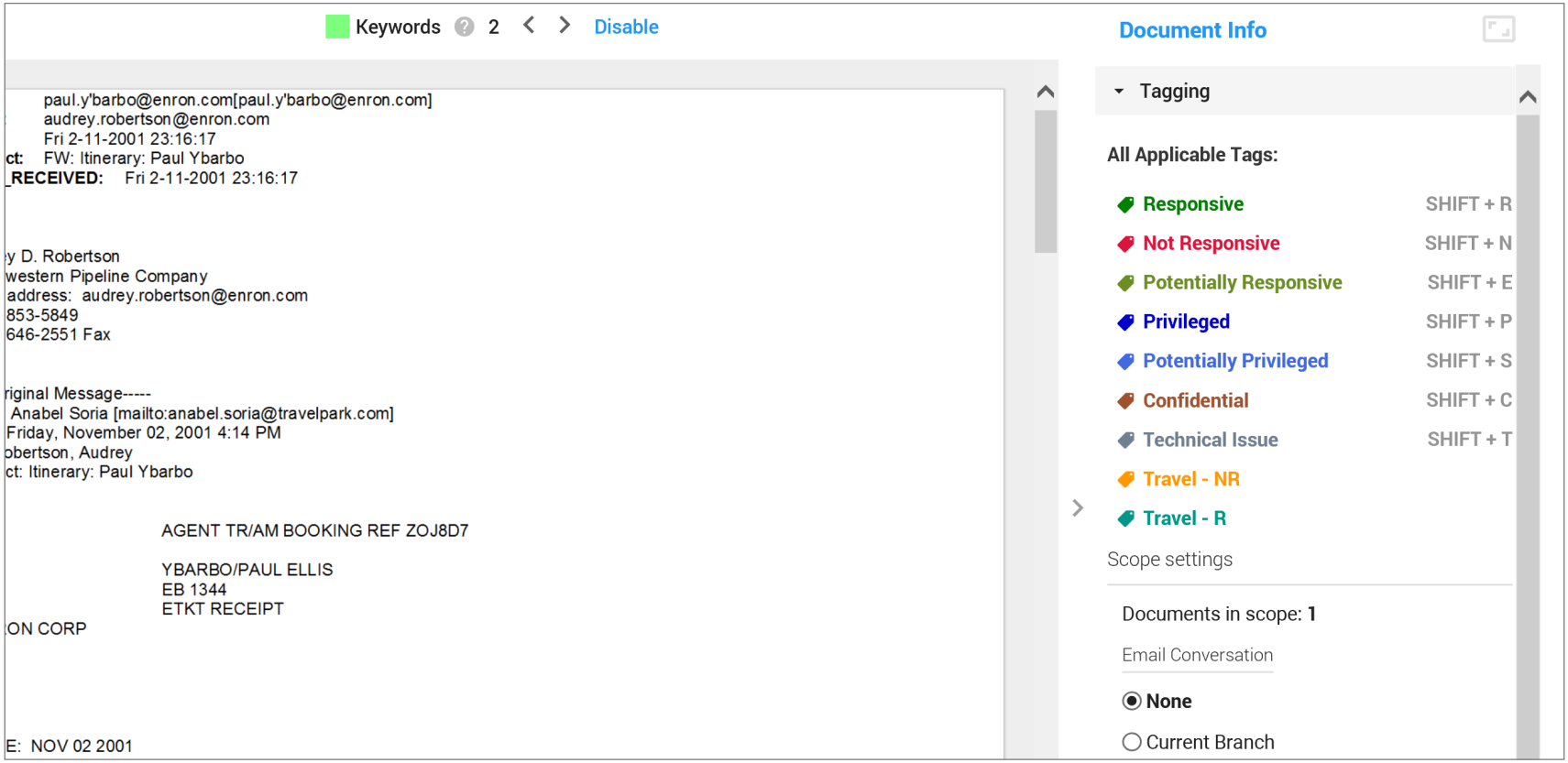
# DO IT YOURSELF: CREATE ISSUES



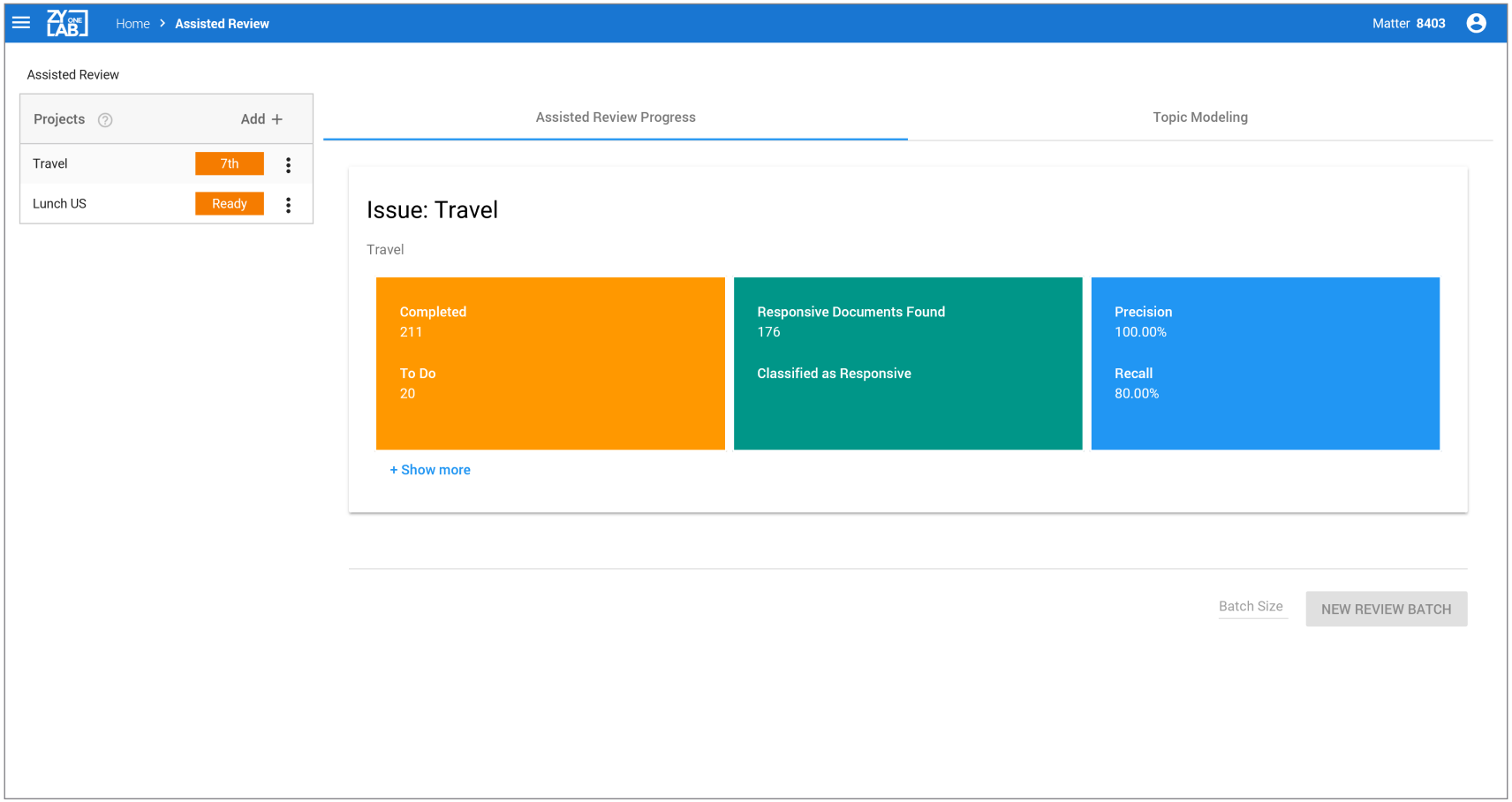
# DO IT YOURSELF: START TAGGING



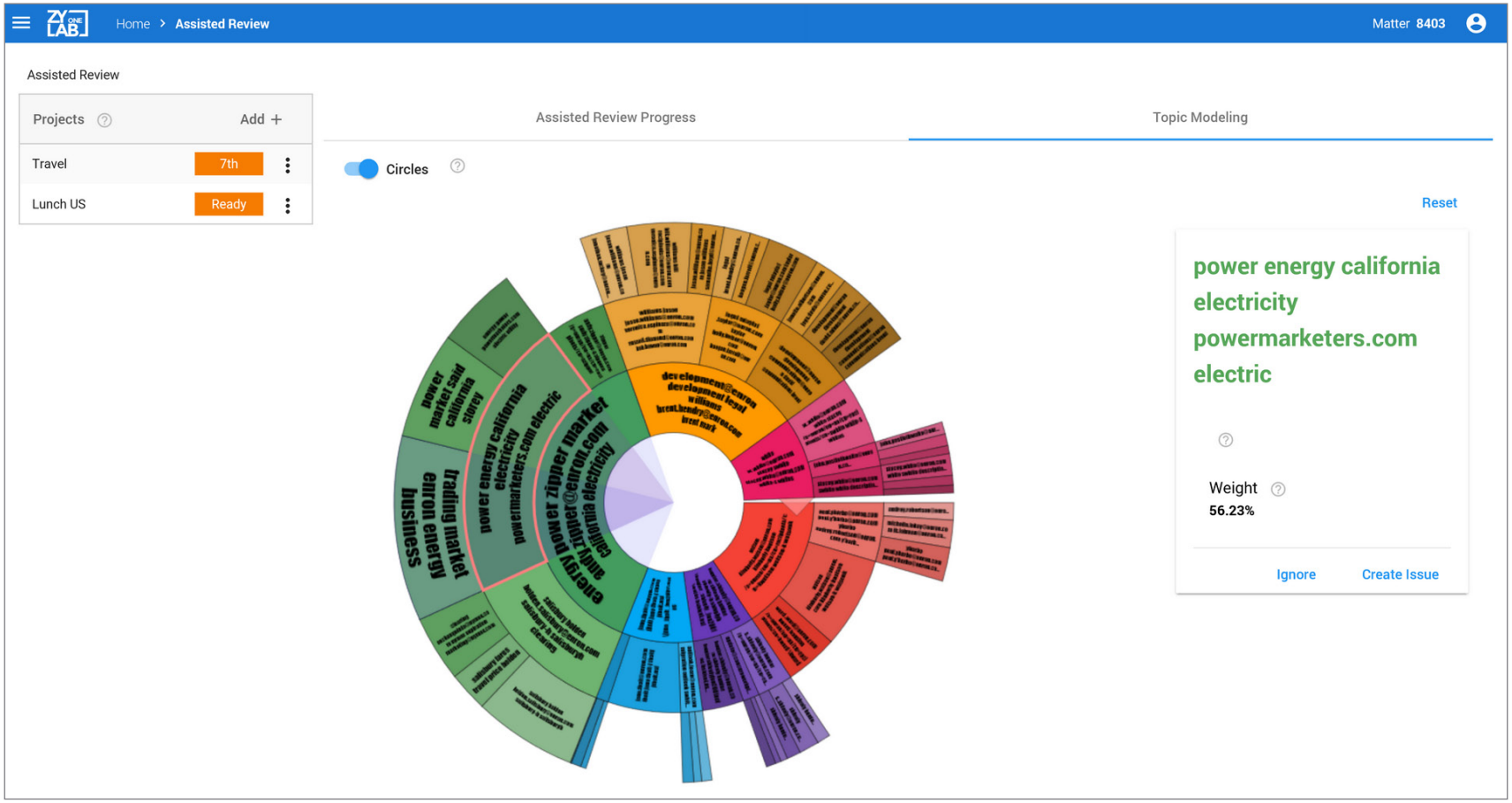
# DO IT YOURSELF: START TAGGING (2)



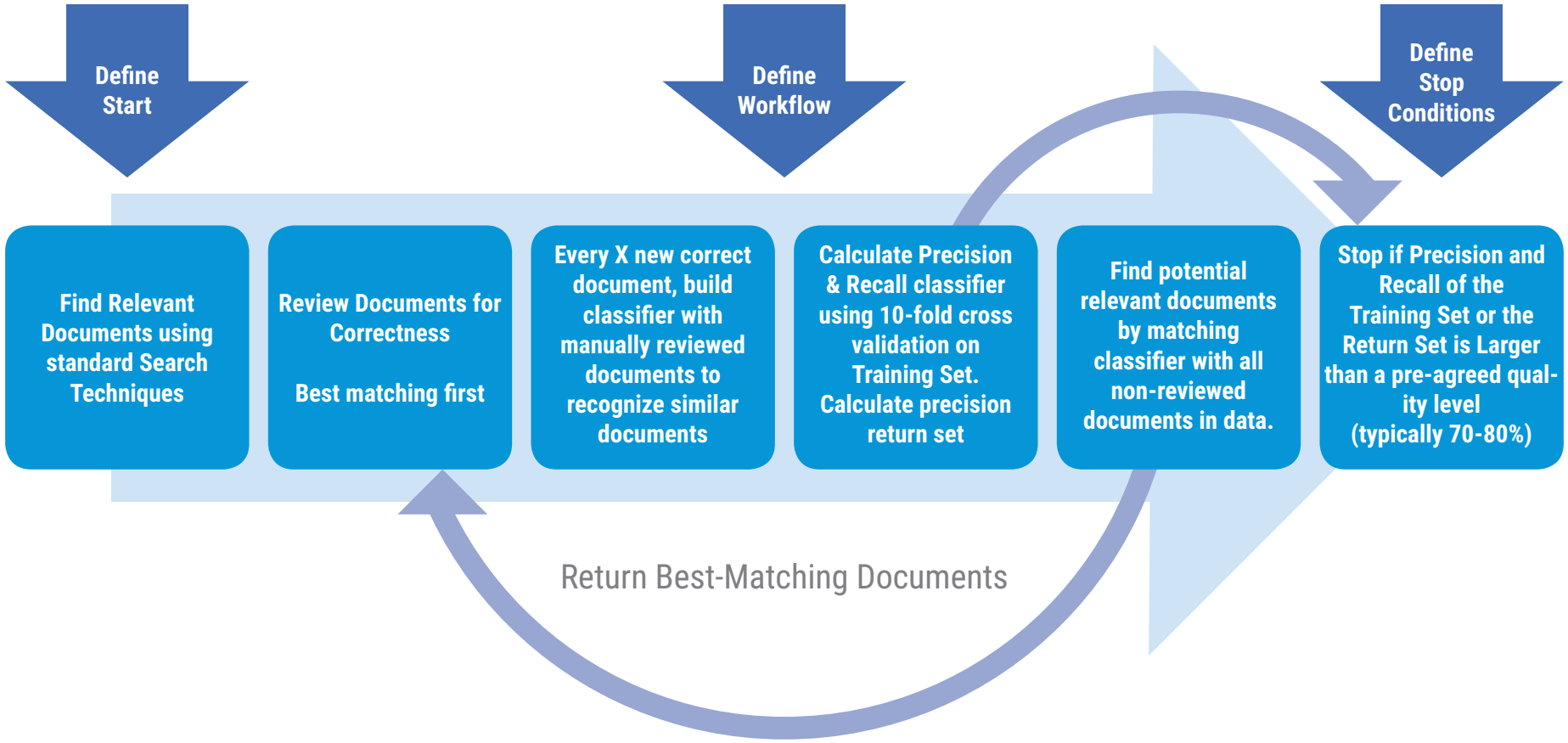
# DO IT YOURSELF: CREATE ISSUES



# CREATE ISSUES FROM TOPIC MODELING



# EXAMPLE OF A PROTOCOL



## ZYLAB SUPPORTS DIFFERENT TAR PROTOCOLS

- **TAR 1.0**: first TAR protocol. Starts with a random validation set which is reviewed for all issues. Used to estimate number of relevant document in entire text collection and to build the initial sets with training documents for each issue.
- **TAR 2.0**: Starts by using full-text search to build training sets per issue.
- **TAR 3.0**: Starts with a topic modeling, clustering or concept search process. Relevant topics are selected from the clusters. The most dominant documents per cluster are used to start the training process.
- **TAR 4.0**: Any combination of TAR 1.0 to 3.0.

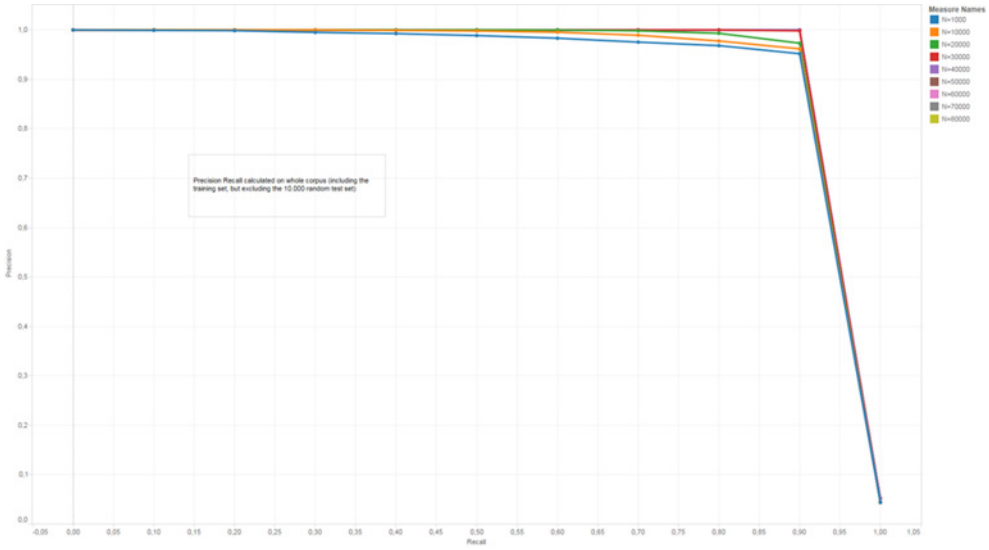
## WHAT IS A STOP CONDITION?

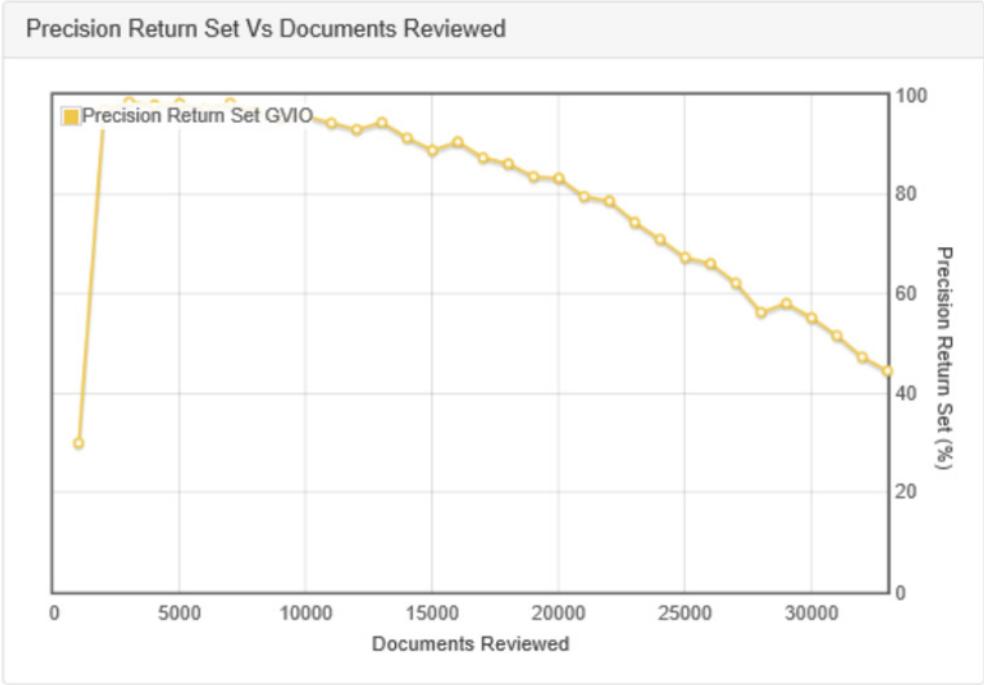
Classifier is good enough to classify the remaining documents automatically.

What is a definition of “good enough”: topic for negotiations.

**Some examples are:**

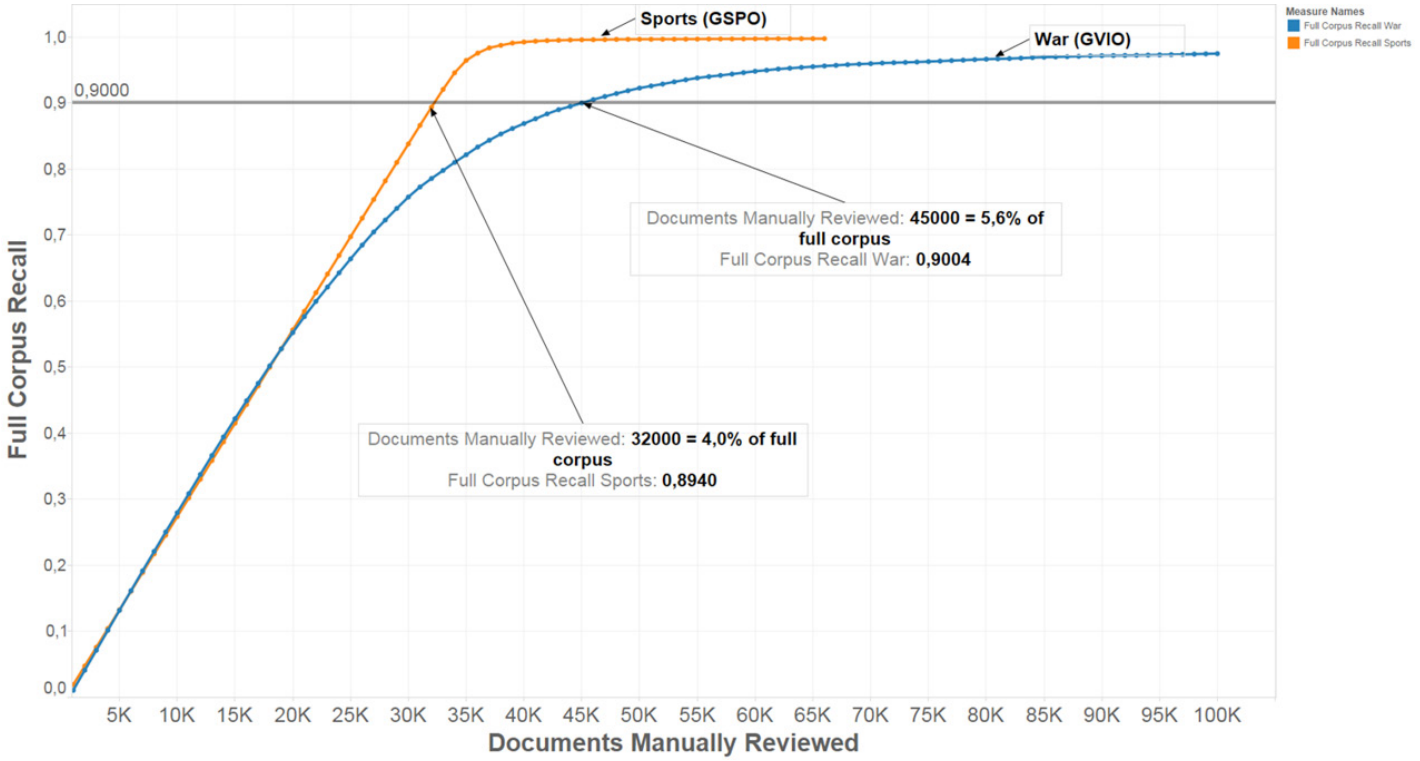
- Precision – recall classifier is structurally > 80% for both Precision and recall
- Precision of classification of new documents is > 80%
- Precision of classification is <10% after going to > 80% first.



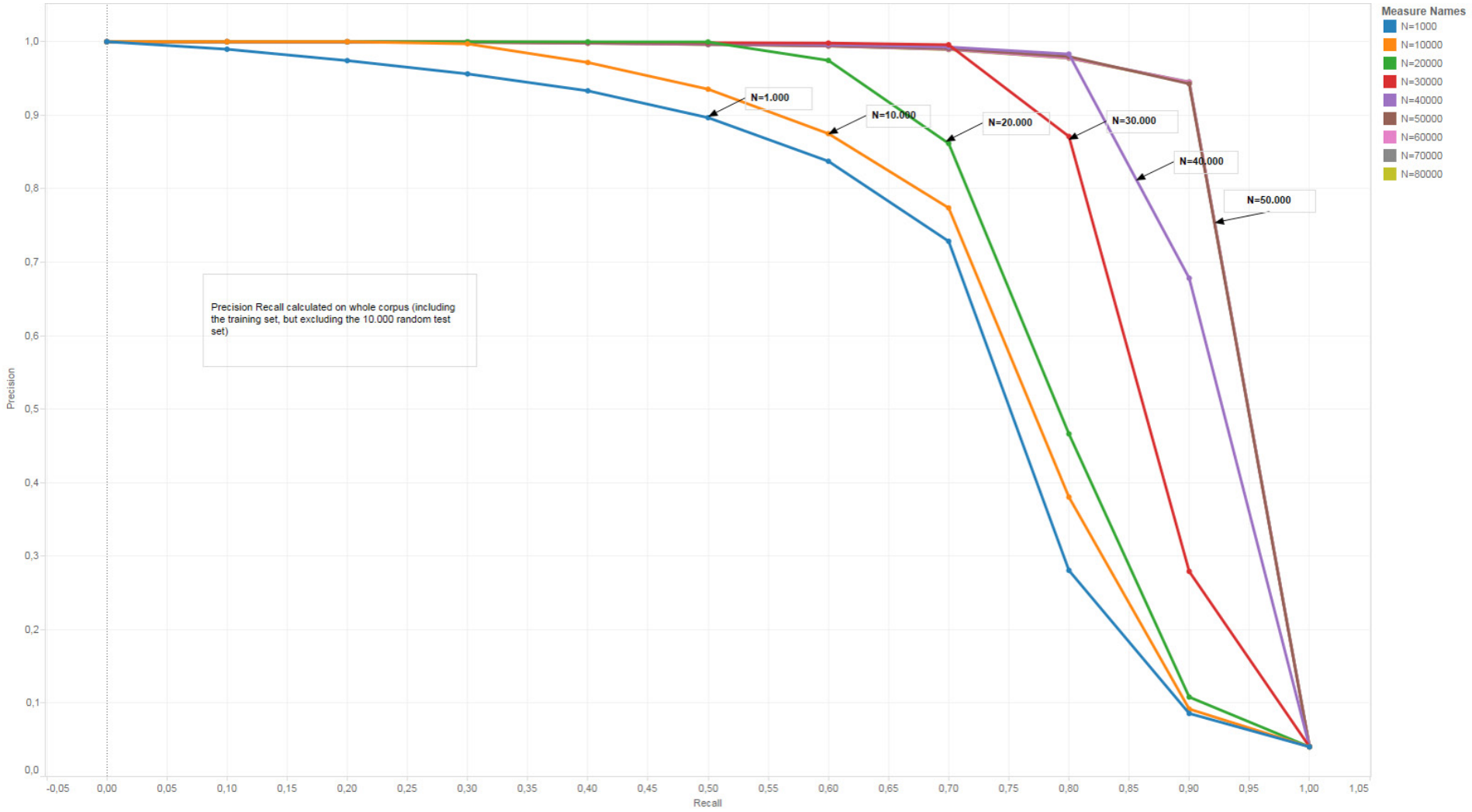


# SIMULATION OF CLASSIFYING REUTERS DOCUMENT SET

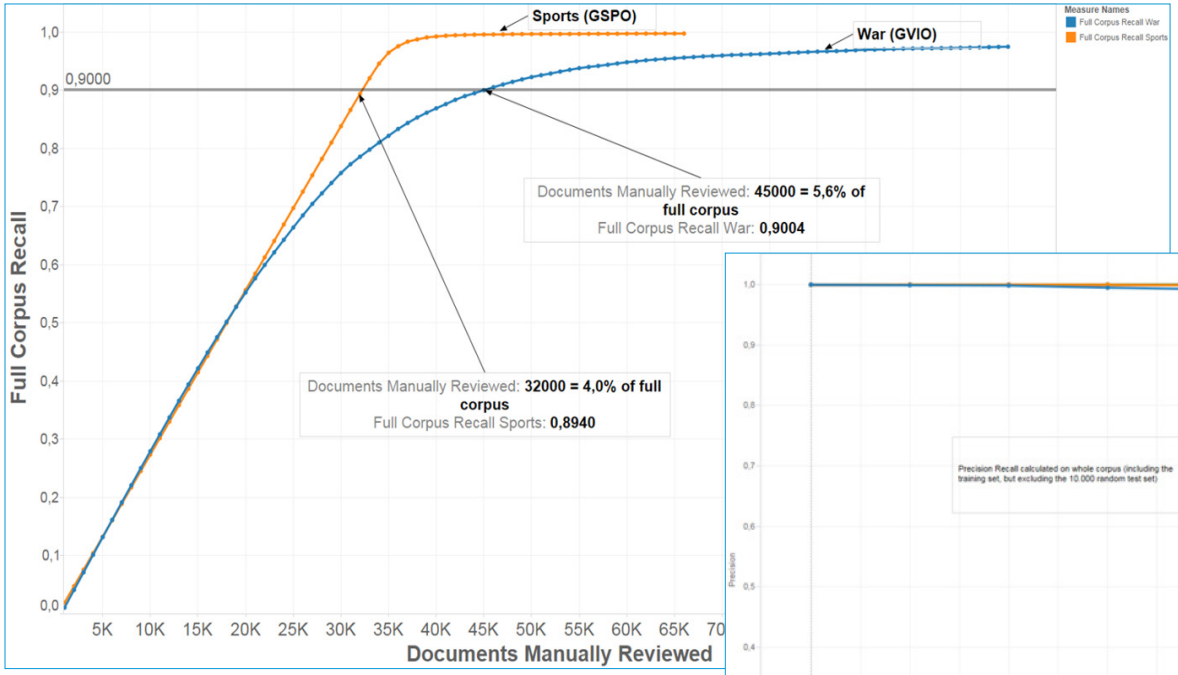
- 806.791 articles in total
- War, Civil War (GVIO): 32.615 articles (4,04%): 90% is found after reviewing only 45.000 documents, which is only 5.6% of full corpus.
- Sports (GSPO): 35.317 articles (4,38%): 90% is found after reviewing only 32.000 documents. This is only 4% of full corpus.



# EVOLUTION OF THE QUALITY OF A CLASSIFIER



# EXAMPLE OF A STOP CONDITION

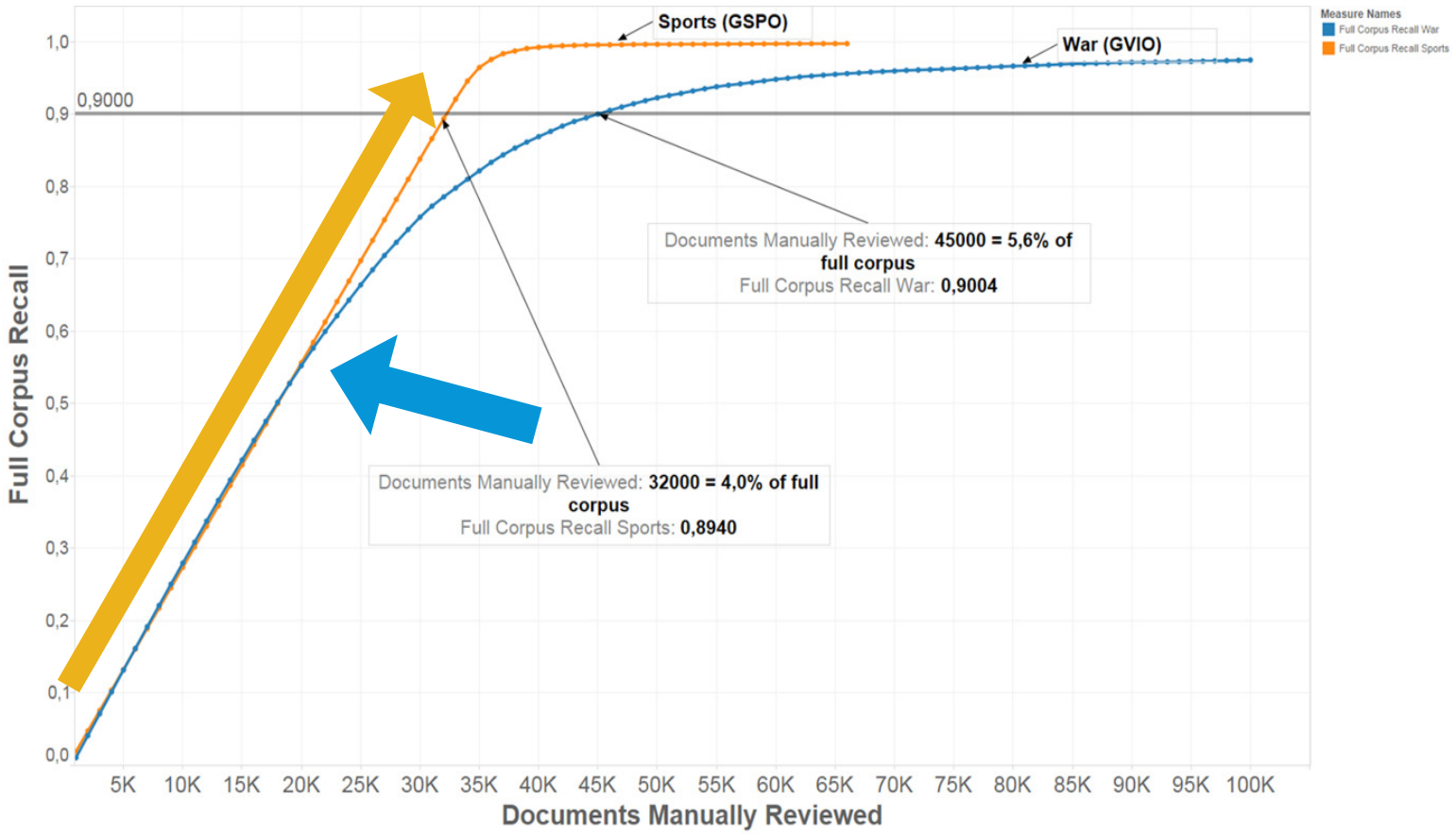




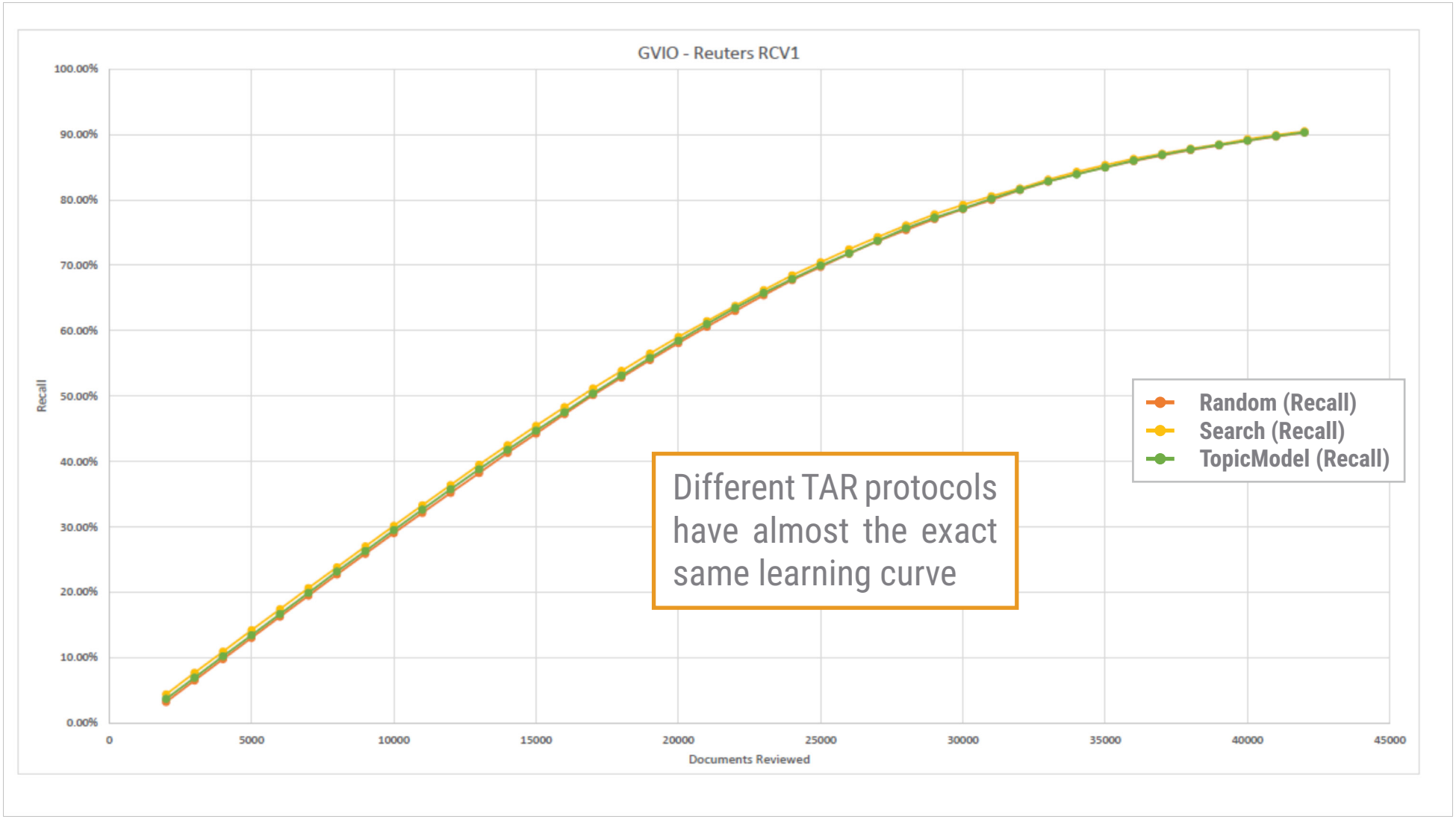
- There was no need to review 32.000 document for the SPORTS (GSPO) category, because both the precision and recall were well over 80% pretty much all of the process. There was also no decay in the slope of the learning progress contrary to the slope of the GVIO.
- We could have stopped reviewing after one training cycle (1.000) documents and find the rest of the responsive documents automatically.

# PREDICTING THE TIME NEEDED TO REACH A STOP CONDITION

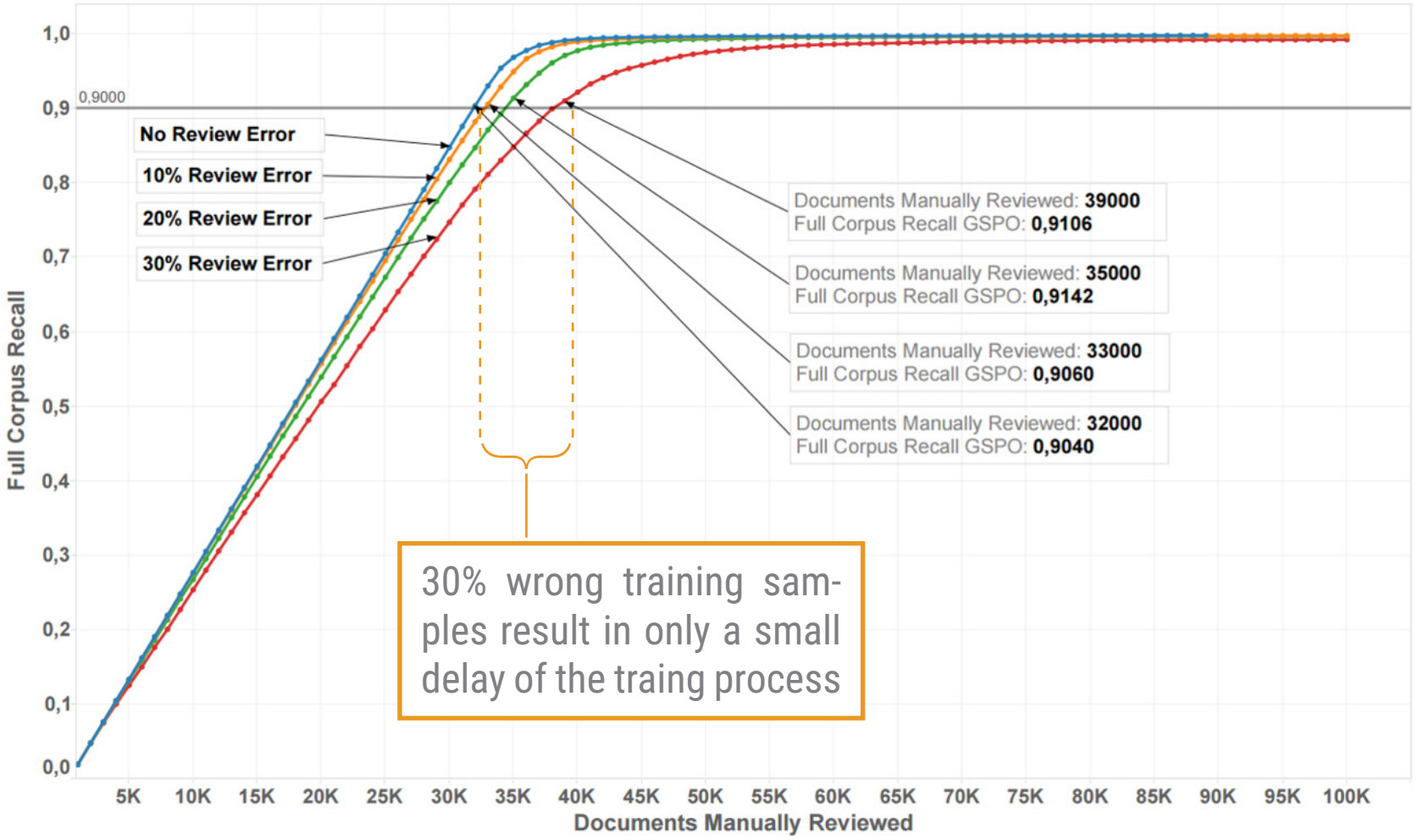
A change in the slope can be used to estimate the additional training cycles and this manual reviews in order to obtain certain recall percentages



# ARE THERE HUGE DIFFERENCES IN PERFORMANCE DEPENDING ON THE START CONDITION? NOT REALLY



# WHAT IF THE REVIEWER MAKES ERRORS: THIS IS NOT REALLY A PROBLEM



## WHY ZYLAB ASSISTED REVIEW (1)

- **Integrated**: ZyLAB's TAR is seamlessly integrated in the ZyLAB Legal Review and can be operated by non-technical users in the background of a review process based on standard general TAR settings, which can be pre-defined by project managers. However, once the TAR process runs, ZyLAB does provide deep insights to project managers and machine learning specialists into the details and progress of Machine Learning process and offers also Cost prediction and other useful analytics to decide when to stop manual review and rely on machine based classification for the remainder of the collection.
- **Protocols**: ZyLAB Supports all common TAR Legal protocols. This will provide you ultimate flexibility and compliance.
- **Technologies**: ZyLAB TAR combines different techniques for document classification and concept search: search-based, natural language processing, semantic, text-mining and AI machine learning. By combining these ZyLAB users can benefit from the advantages of all these different approaches.
  - Non-technical users can use search / Semantic / Text Mining-based classification, libraries of such queries can be re-used over projects or can be translated for other languages. One can also use these basic document classification techniques to harvest the low-hanging fruit and kick-start your review process.
  - Advanced users can use the AI machine learning to build classifiers based on document samples and benefit from the tremendous power and time savings of machine learning to reach very high recall levels and outperform human review speed and quality.

## WHY ZYLAB ASSISTED REVIEW (2)

- **Algorithms**: ZyLAB TAR uses the very best algorithms based on decades of independently evaluated scientific research. We are constantly improving our technology by working together with highly specialized universities and scientists.
- **Speed**: ZyLAB TAR uses advanced feature extraction and selection based on knowledge of statistics and natural language processing. By doing so, we can reduce the size of the data that we have to deal with to just the most relevant data, without lowering the quality of the classification. This allowed us to speed up the machine learning and topic modeling component of our TAR to unprecedented levels.

## WHAT ABOUT PRIVILEGED REVIEW?

- Privileged review is very hard. One single sentence in a document can determine if a document is privileged or not. This is hard to find with machine learning TAR.
- Privileged review has to be 100% correct. There is no room for errors.
- Documents can be privileged for many reasons.
- For now, searching for very specific privileged reasons (attorney-client communication, certain text phrases, names, keywords, regular expressions, patterns) can find potentially privileged documents. Rules-based TAR is most suited technology for finding privileged documents.
- For the larger part, this remains a manual review process of the responsive documents before they are disclosed.

**Query:**

(sendersdomain=?minterellison.com*.* OR sendersdomain=?kwm.com*.* OR sendersdomain=?allens.com.au*.* OR sendersdomain-=?freehills.com*.* OR sendersdomain=?claytonutz.com*.* OR sendersdomain=?mccarthy.ca*.* OR sendersdomain=?fidal.fr*.* OR sendersdomain=?noerr.com*.* OR sendersdomain=?boekeldeneree.com*.* OR sendersdomain=?loyensloeff.com*.* OR sendersdomain=?nautadutilh.com*.* OR sendersdomain=?stibbe.com*.* OR sendersdomains=?debrauw.com*.* OR sendersdomain=?houthoff.com*.* OR sendersdomain=?akd.eu*.* OR sendersdomain=?allenovery.com*.* OR sendersdomain=?cms-dsb.com*.* OR sendersdomain=?

**Name:** attorney client communication

**Shared With:** local\Domain Users ...

## WHY AND HOW IS ZYLAB TAR DEFENSIBLE?

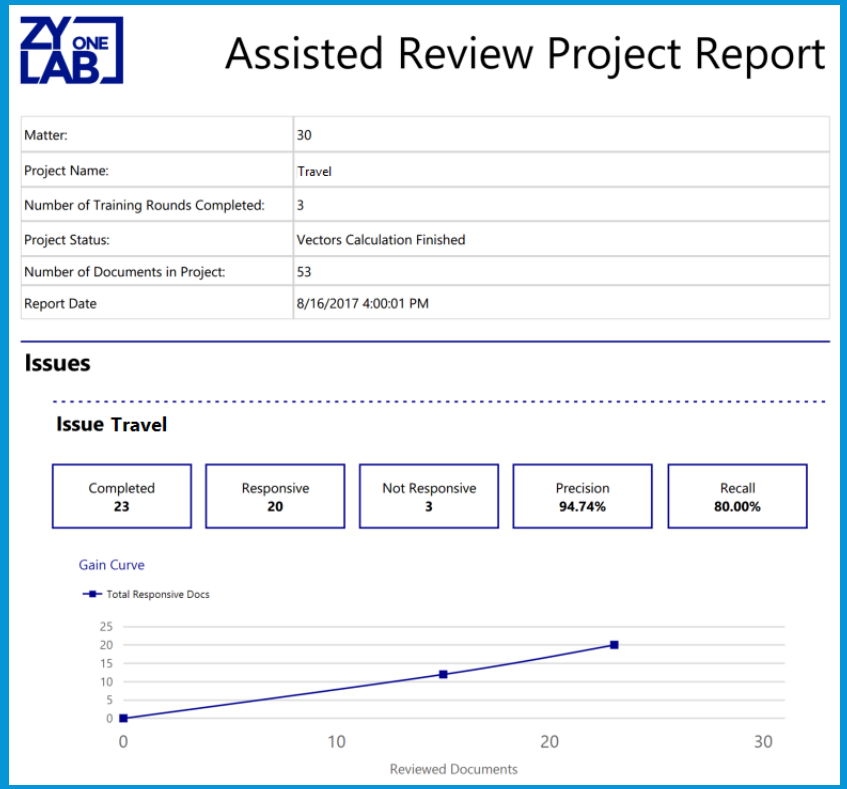
- It is important to properly establish the defensibility of the entire process. This is done by drawing up a so-called “defensibility report”. All the details of the training process are fixed in this report: which training documents, which users have reviewed, where, when, how long, how many training cycles were there, how did the quality of the classifiers develop, etc. All this information is saved in a report.
- Lastly, an independent test must be carried out into the quality of the classifier. This is done using a manual review process of a random selection (often 1-5%) of documents that the computer wants to classify in that category. And if these are also good enough, then the automatic classification of the rest of the documents may proceed.
- This test may be carried out at a later date as often as you want, to continually test the quality of the automatic classification process. Obviously it is very important to establish sound reporting of these tests.

## WHAT IS THE BEST WAY TO IMPLEMENT ZYLAB TAR IN YOUR LAW FIRM, IN-HOUSE LEGAL OR INTERNAL INVESTIGATION DEPARTMENT?

- In the same way that law firms check each other's work using random sampling, this should also be done for activities carried out automatically.
- By taking a random sample and getting specialists or senior lawyers to check them, the quality of the process can be monitored continually.
- So precisely the same quality standards can be used in the new situation.

## HOW ABOUT LEGAL RISK AND LIABILITY?

- Independent and continual validation of the results and defensibility of the automated process belong to the key elements of the entire process.
- The risk control and management of liability will therefore not be any different than with traditional processing of the projects.
- The task is also to thoroughly document the underlying steps and decision-making moments of the automatic process and establish this with an audit-trail and detailed reports. ZyLAB provides a range of automated support services for this purpose.



Assisted Review Project Report

| | |
|---|---|
| Matter: | 30 |
| Project Name: | Travel |
| Number of Training Rounds Completed: | 3 |
| Project Status: | Vectors Calculation Finished |
| Number of Documents in Project: | 53 |
| Report Date | 8/16/2017 4:00:01 PM |

**Issues**

**Issue Travel**

| Completed | Responsive | Not Responsive | Precision | Recall |
|---|---|---|---|---|
| 23 | 20 | 3 | 94.74% | 80.00% |

## ADDITIONAL REFERENCES


- Blair and Maron (1985). An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System. Communications of the ACM, 1985.
- Voorhees, Ellen M. (Editor), Harman, Donna K. (Editor), (2005). TREC: experiment and evaluation in information retrieval. MIT Press.
- Dan Regard and Tom Matzen (2013). A Re-Examination of Blair & Maron. DESI V Workshop, June 14, 2013, Position Paper.
- Jones, A. et al (2013). The Role of Metadata in Machine Learning for Technology Assisted Review. DESI V Workshop, June 14, 2013.
- Legal-TREC Research Program: http://trec-legal.umiacs.umd.edu/.
- Losey, Ralph (2013). Predictive Coding Narrative: Searching for Relevance in the Ashes of Enron. https://ralphlosey.files.wordpress.com/2013/04/predictive-coding-narrative_corrrected_3-21-13.pdf
- Maura R. Grossman & Gordon V. Cormack (2011), Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, XVII RICH. J.L. & TECH. 11 (2011)
- Maura R. Grossman & Gordon V. Cormack (2014), Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery. SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
- Manning, Raghavan & Schütze (2008). Introduction to Information Retrieval. Website: http://informationretrieval.org/. Cambridge University Press, 2008.
- Scholtes, J.C., Cann, T.H.W. van, and Mack, M. (2013). The Impact of Incorrect Training Sets and Rolling Collections on Technology-Assisted Review. International Conference on Artificial Intelligence in Law 2013, DESI V Workshop. June 14, 2013, Consiglio Nazionale delle Ricerche, Rome, Italy.
- Scholtes, J.C. and Cann, T.H.W. van (2013). Improving Machine Learning Input for Automatic Document Classification with Natural Language Processing. Benelux Artificial Intelligence Conference (BNAIC), 2013. Delft, the Netherlands, November 7-8, 2013.
- Tannenbaum, M., Fischer, A., and Scholtes, J.C. (2015). Dynamic Topic Detection and Tracking using Non-Negative Matric Factorization. Benelux Artificial Intelligence Conferenbce (BNAIC), 2015. Hasselt, Belgium, November 5-6, 2015.
- Smeets, J., Scholtes, J.C., Rasterfoff, C. and Schravemaker, M. (2016). SMTP: Stedelijk Museum Text Mining Project. Digital Humanities Benelux (DHBenelux), Luxemburg, June 2016.
- Smeets, J., Scholtes, J.C., Rasterfoff, C. and Schravemaker, M. (2016). SMTP: Stedelijk Museum Text Mining Project. Digital Humanities Conference, Krakow, Poland, July 2016.