An empirical comparison of machine learning classification algorithms

&

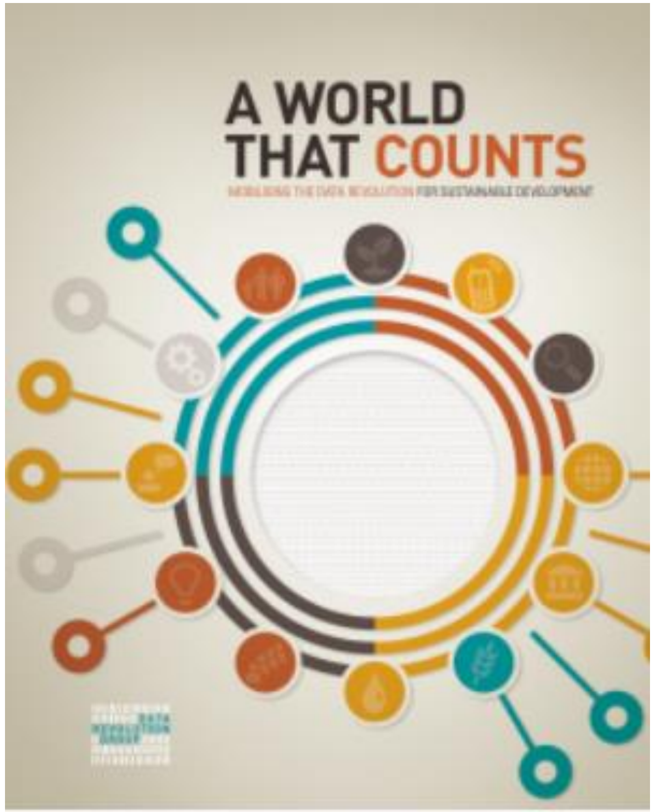
Topic Modeling
A quick look at 145,000 World Bank documents

Olivier Dupriez, Development Data Group

Slides prepared for DEC Policy Research Talk, February 27, 2018

The 2014 call for a Data Revolution

- Use data differently (innovate)
  - New tools and methods → A comparative assessment of machine learning algorithms
- Use different data (big data, …)
  - Text as data → Topic modeling applied to the World Bank Documents and Reports corpus



An empirical comparison of machine learning classification algorithms applied to poverty prediction

A Knowledge for Change Program (KCP) project

Documenting use and performance

- Many machine learning algorithms available for <u>classification</u>
- We document the <u>use</u> and <u>performance</u> of selected algorithms
- Application: prediction of household poverty status (poor/non-poor) using easy-to-collect survey variables
  - Focus on the tools → use “traditional” data (household surveys)
  - Not a new idea (SWIFT surveys, proxy means testing, survey-to-survey imputation, poverty scorecards; most rely on <u>regression</u> models)
  - Possible use cases: targeting; simpler/cheaper poverty surveys

Key question

NOT

"What is the best algorithm for predicting [poverty]?"

BUT

"How can we get the most useful [poverty] prediction for a specific purpose?"

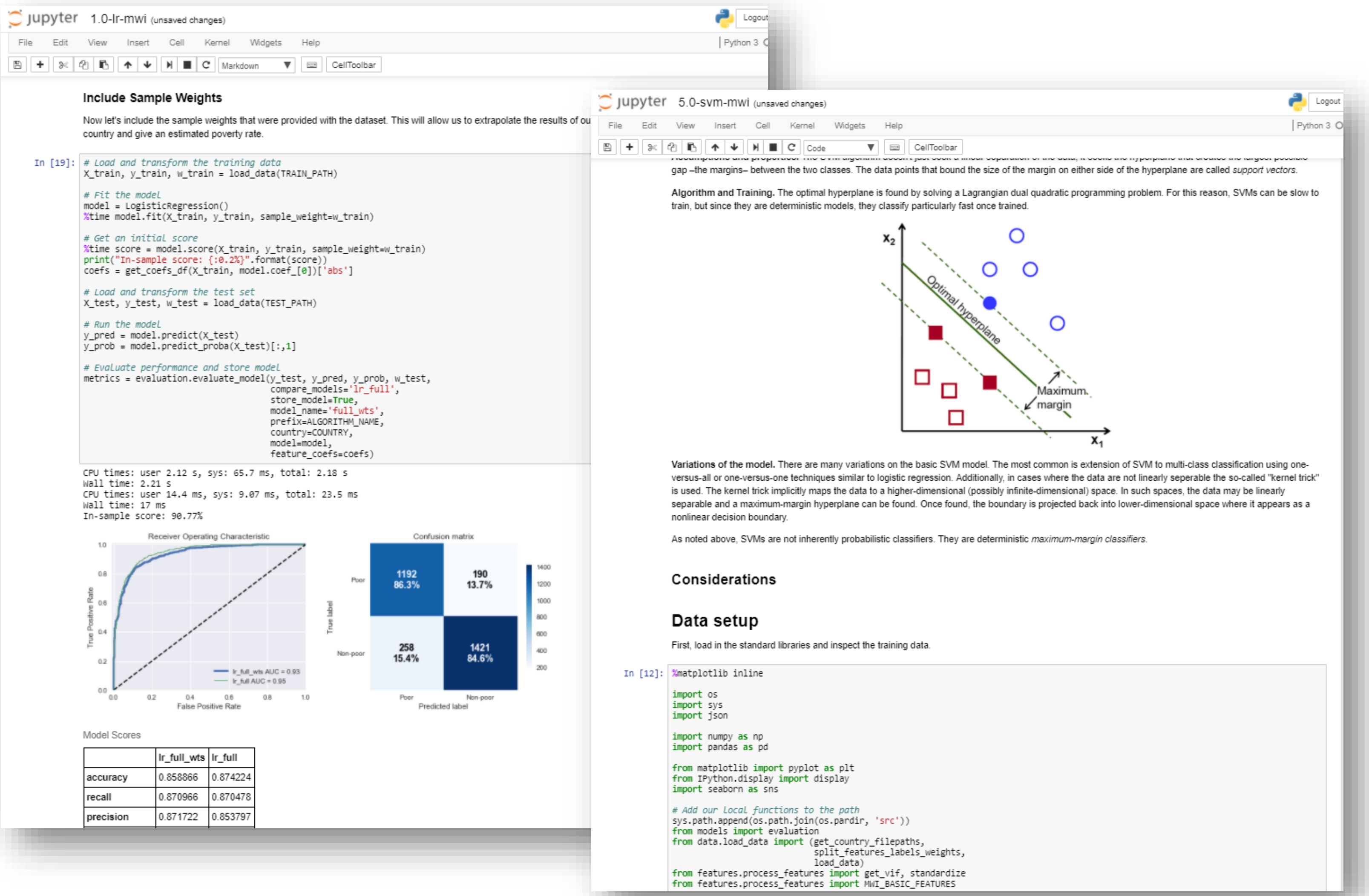
Approach

1. Apply 10 “out-of-the-box” classification algorithms
    - Malawi IHS 2010 – Balanced classes (52% poor ; 12,271 hhlds)
    - Indonesia SUSENAS 2012 - Unbalanced classes (11% poor ; 71,138 hhlds)
    - Data: mostly qualitative variables, including dummies on consumption (hhld consumed [*item*] – Yes/No). Did not try to complement with other data.
2. Challenge the crowd: data science competition to predict poverty status for 3 countries (including MWI)
3. Challenge experts to build the best model for IDN, with no constraint on method
4. Apply automated Machine Learning on IDN

Reproducible and re-purposable output

[Jupyter notebooks](#)

→ Reproducible script, output, and comments all in one file



Multiple metrics used to assess performance

		Predicted	
		Poor	Non poor
Actual	Poor	**True positive TP**	**False negative FN**
	Non poor	**False positive FP**	**True negative TN**

Accuracy: (TP + TN) / All

Recall: TP / (TP + FN)

Precision: TP / (TP + FP)

F1 score: 2TP / (2TP + FP + FN)

Cross entropy (log loss)

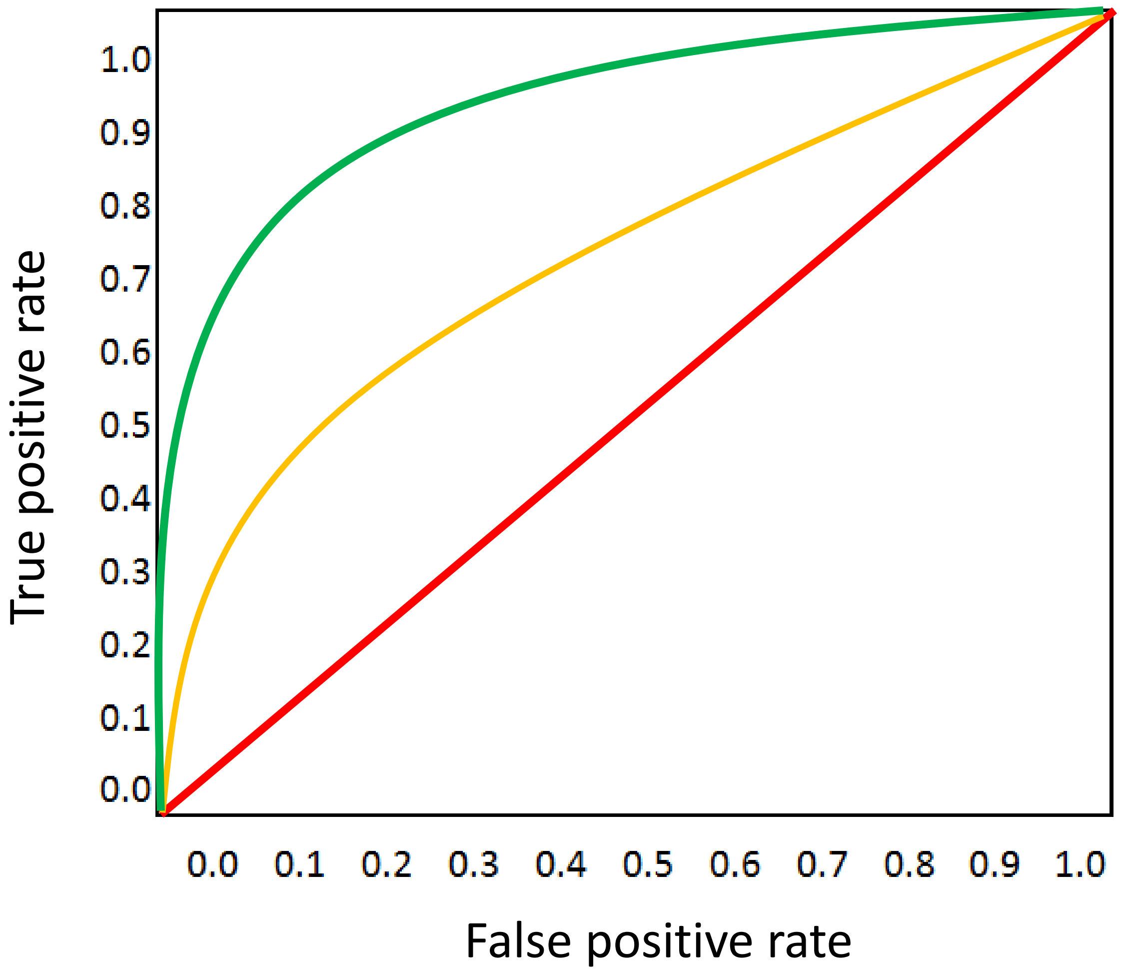
Cohen - Kappa

ROC – Area under the curve

(Calculated on out-of-sample data)

Area under the ROC curve (AUC)

- Plot the true and false positive rates for every possible classification threshold
- A perfect model has a curve that passes through the upper left corner (AUC = 1)
- The diagonal represents random guessing (AUC = 0.5)



http://www.dataschool.io/roc-curves-and-auc-explained/

10 out-of-the-box classification algorithms

Logistic Regression	CART Decision Trees
Linear Discriminant Analysis	Random Forests
K-Nearest Neighbors	eXtreme Gradient Boosting
Naive Bayes	Multilayer Perceptron
Support Vector Machines (SVM)	Deep Learning (Neural Networks)

With scaling, boosting, over- or under-sampling as relevant

Implemented in Python; scikit-learn for all except XGBoost

Results - Out-of-the-box algorithms (MWI)

Algorithm (no feature engineering) (Results for selected models)	Accuracy	Recall	Precision	f1	Cross entropy	ROC AUC	Cohen Kappa	Mean rank
Support Vector Machine (SVM) CV	0.874	0.894	0.878	0.886	0.287	0.949	0.758	5.000
Multilayer Perceptron CV	0.871	0.895	0.874	0.884	0.278	0.952	0.752	6.125
XGBoost selected features	0.872	0.892	0.877	0.884	0.289	0.949	0.754	7.375
SVM CV Isotonic	0.871	0.891	0.876	0.883	0.288	0.949	0.754	7.625
Logistic Regression – Weighted	0.873	0.892	0.879	0.885	0.301	0.944	0.734	7.750
XGBoost, all features CV	0.869	0.894	0.870	0.882	0.296	0.948	0.751	9.125
SVM Full	0.864	0.886	0.868	0.877	0.298	0.945	0.733	10.625
Logistic Regression Full	0.874	0.870	0.854	0.862	0.288	0.949	0.746	12.750
Random Forest, Adaboost	0.866	0.878	0.878	0.878	0.580	0.947	0.744	13.000
Decision Trees, Adaboost	0.866	0.878	0.878	0.878	0.353	0.941	0.737	13.000

No clear winner (best performer has a mean rank of 5)

Results - Out-of-the-box algorithms (IDN)

Algorithm (Results for selected models)	Accuracy	Recall	Precision	f1	Cross entropy	ROC AUC	Cohen Kappa	Mean rank
Logistic Regression	0.910	0.456	0.662	0.540	0.213	0.923	0.483	3.25
Multilayer Perceptron	0.909	0.543	0.619	0.579	0.496	0.923	0.548	4
Linear Discriminant Analysis	0.906	0.405	0.648	0.499	0.231	0.912	0.457	5
Support Vector Machine	0.902	0.208	0.782	0.329	0.204	0.932	0.312	5.125
K Nearest Neighbors	0.904	0.372	0.647	0.472	0.541	0.865	0.423	6.5
XGBoost	0.898	0.184	0.743	0.295	0.224	0.917	0.285	6.625
Naïve Bayes	0.807	0.603	0.322	0.420	1.893	0.828	0.238	7.25
Decision Trees	0.859	0.392	0.390	0.391	4.870	0.656	0.306	7.875
Random Forest	0.892	0.107	0.729	0.187	0.592	0.832	0.210	8
Deep Learning	0.884	0.000	0.000	0.000	0.349	0.896	0.000	9.5

No clear winner ; logistic regression again performs well on *accuracy* measure

Results – Predicted poverty rate (IDN)

Difference between predicted and measured poverty rate

Logistic regression	-3.1%
Multilayer perceptron	-0.4%
Support vector machine	-8.2%
Decision trees	0.0%
Random forest	-3.5%

Estimated on full dataset

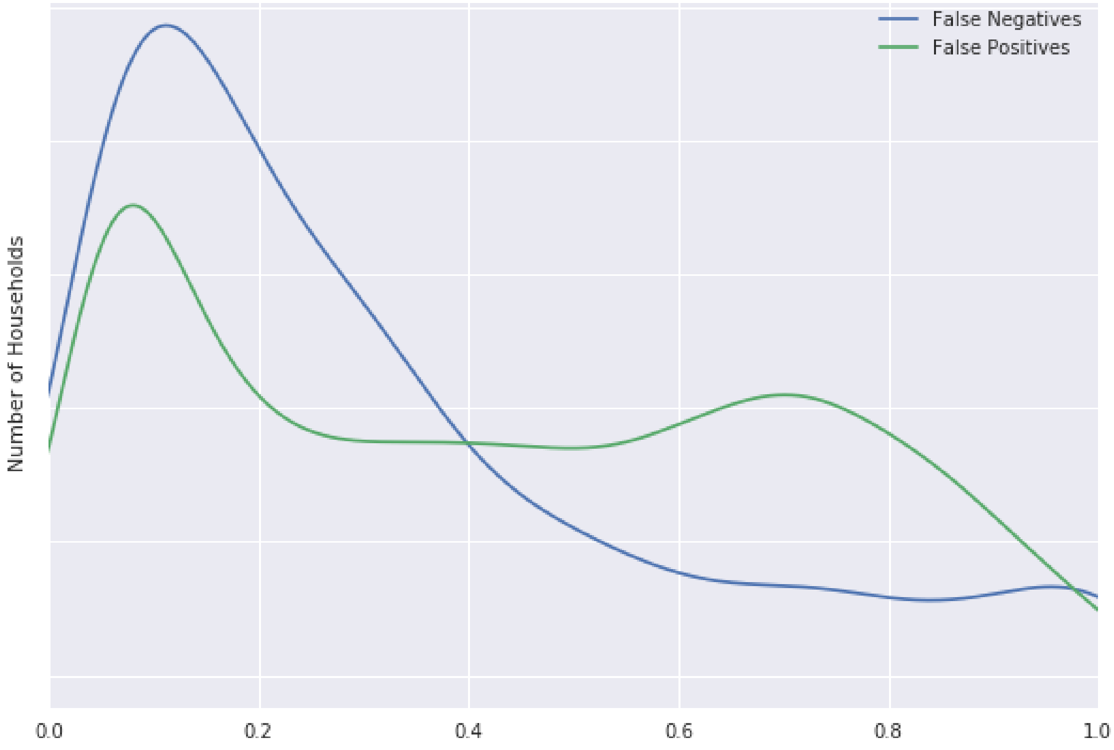
Not a very good model, but achieves quasi-perfect prediction of the poverty headcount (false positives and false negatives compensate each other)

→ A good poverty rate prediction is not a guarantee of a good poverty profile

Ensembling (IDN)

- Diversity of perspectives almost always leads to better performance
- 70% of the households were correctly classified by every one of the top 20 models
- 78% of poor households were misclassified by *at least* one model
- We take advantage of this heterogeneity in predictions by creating an *ensemble*

Inter-model agreement for misclassifications (IDN)

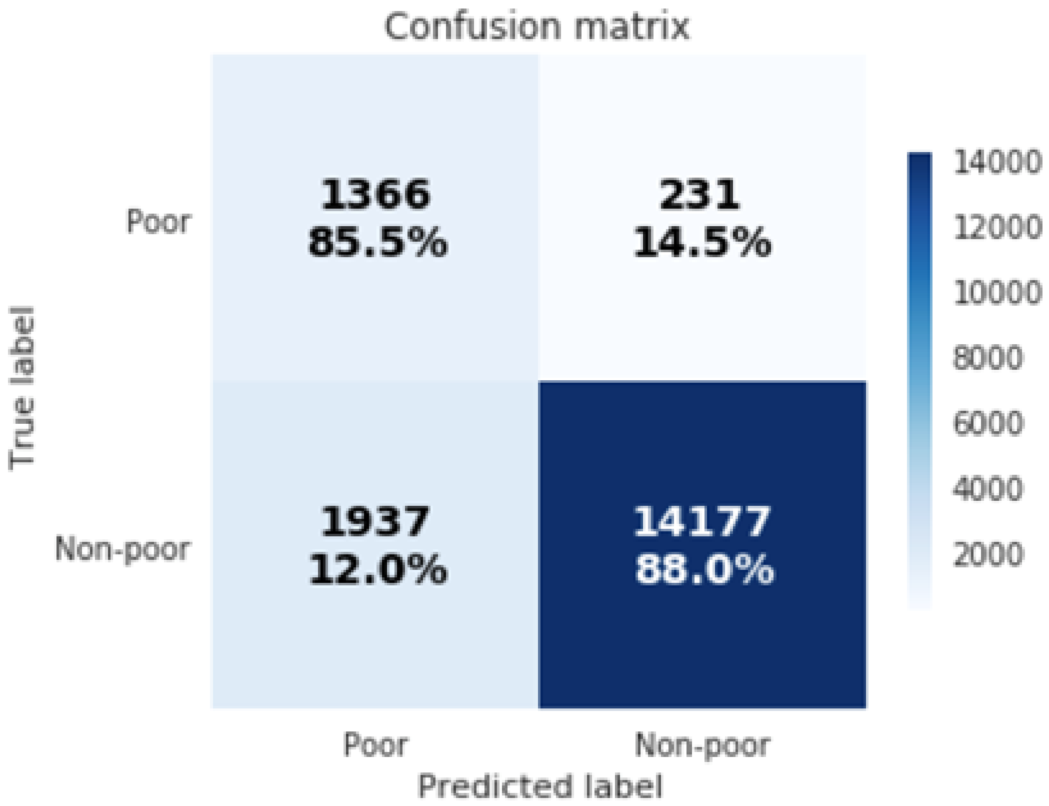


Fraction of top 20 models in error

Results: soft voting (top 10 models, IDN)

	ensemble_simple
accuracy	0.87759
recall	0.855354
precision	0.413563
f1	0.557551
cross_entropy	0.272985
roc_auc	0.945388
cohen_kappa	0.496322

(Max was 0.6)



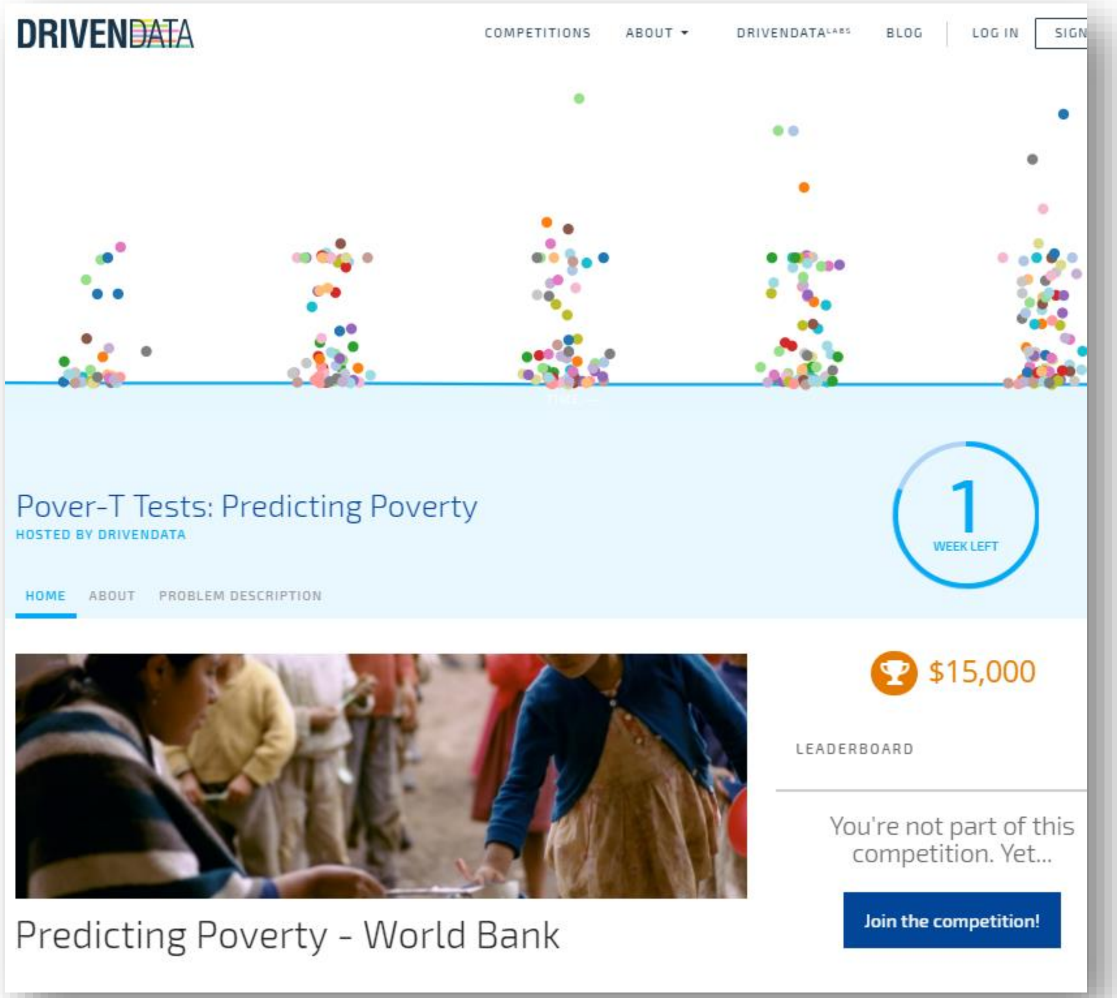
Major improvement in *recall* measure, but low *precision*

Error on poverty rate : +8.9%

Can the crowd do better?

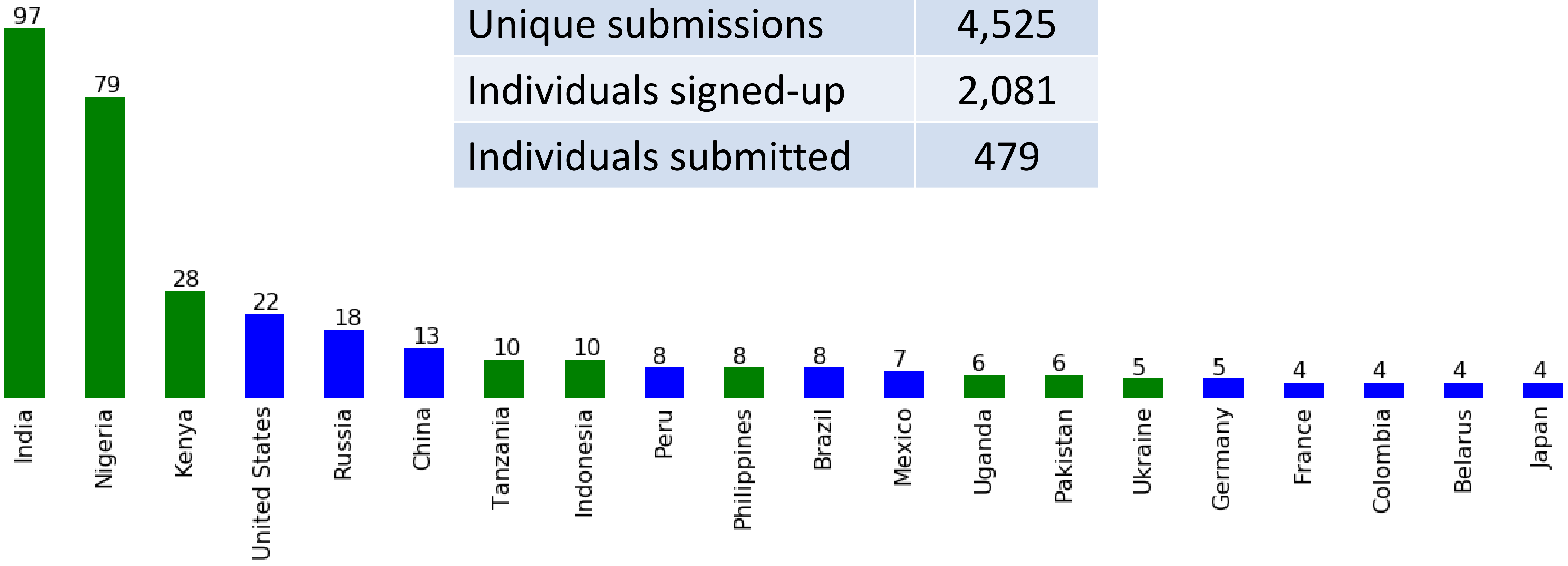
- Data science competition on DrivenData platform
- Challenge: predict household poverty status for 3 countries (including MWI)

Place	Prize Amount
1st	$6,000
2nd	$4,000
3rd	$2,500
Bonus	$2,500



Data science competition - Participation

As of February 22	Number
Unique submissions	4,525
Individuals signed-up	2,081
Individuals submitted	479



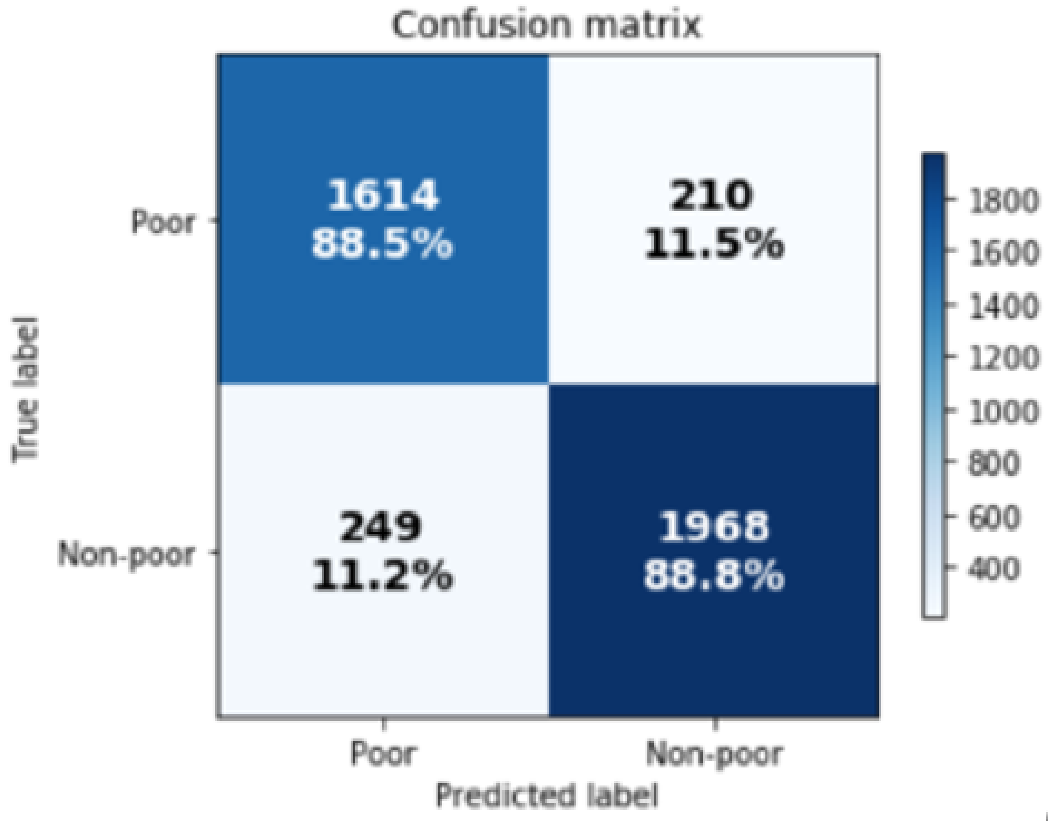
Distribution of registered participants by nationality (for those who provided this information at registration)

Results (so far) on MWI

Slightly better than the best of 10 algorithms

Good results on all metrics

	Score
accuracy	0.886414
recall	0.884868
precision	0.866345
f1	0.875509
cross_entropy	0.260754
roc_auc	0.958052
cohen_kappa	0.771093



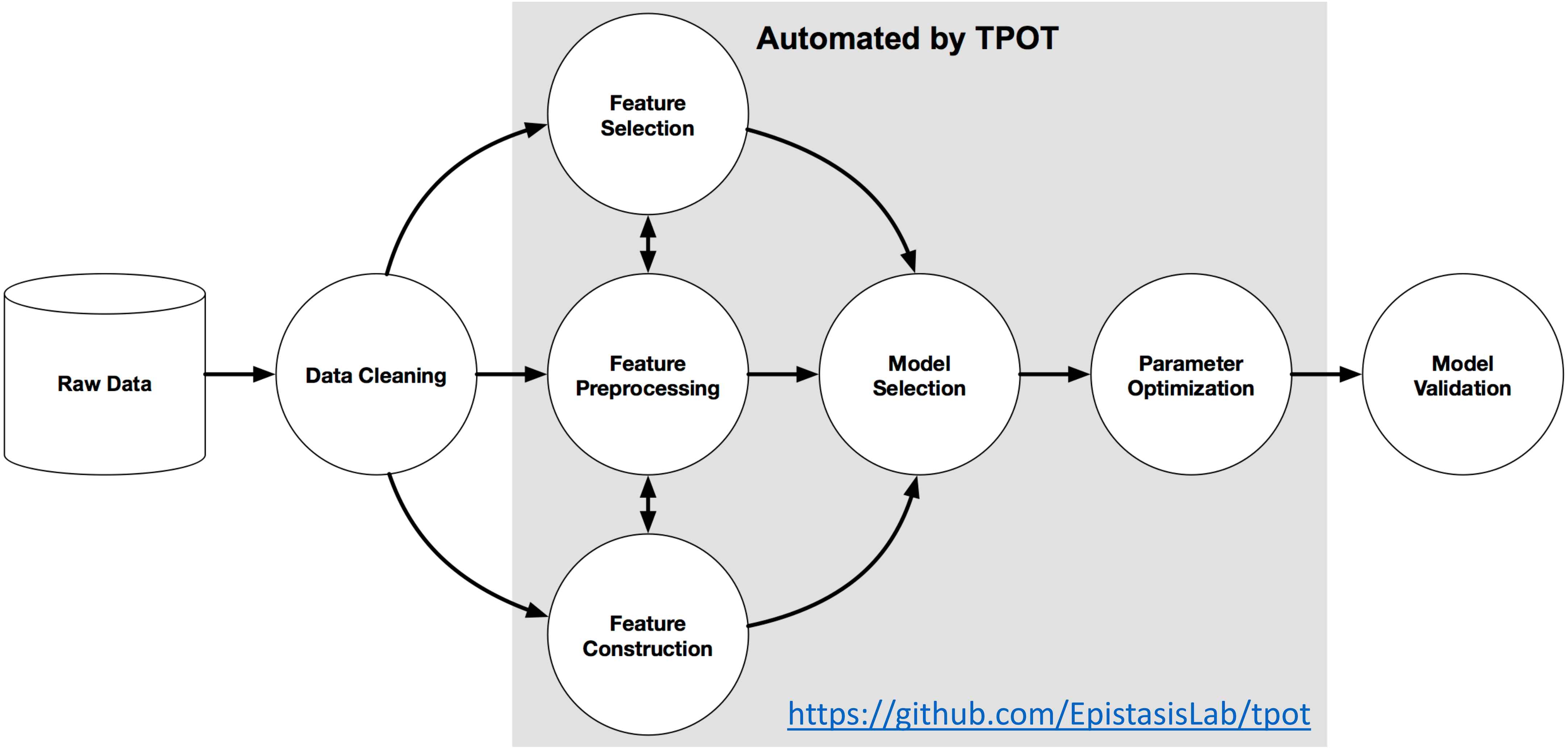
Experts – Advanced search for a solution (IDN)

- Intuition: a click-through rate (CTR) model developed for Google Play Store's recommender system could be a good option
  - High dimensional datasets of primarily binary features; binary label
- Combines the strengths of *wide* and *deep* neural networks
- But requires a priori decision of which interaction terms the model will consider → impractical (too many features to consider interaction between all possible pairs)
- Solution: Deep Factorization Machine (DeepFM) by Guo et al. applied to IDN

Automated Machine Learning (AutoML)

- Goal: let non-experts build prediction models, and make model fitting less tedious
- Let the machine build the best possible "pipeline" of pre-processing, feature (=predictor) construction and selection, model selection, and parameter optimization
- Using TPOT, an open source python framework
- Not brute force: optimization by genetic programming
- Starts with 100 randomly generated pipelines; select the top 20; mutate each into 5 offspring (new generation); repeat

Automated Machine Learning - TPOT



Automated Machine Learning applied to IDN

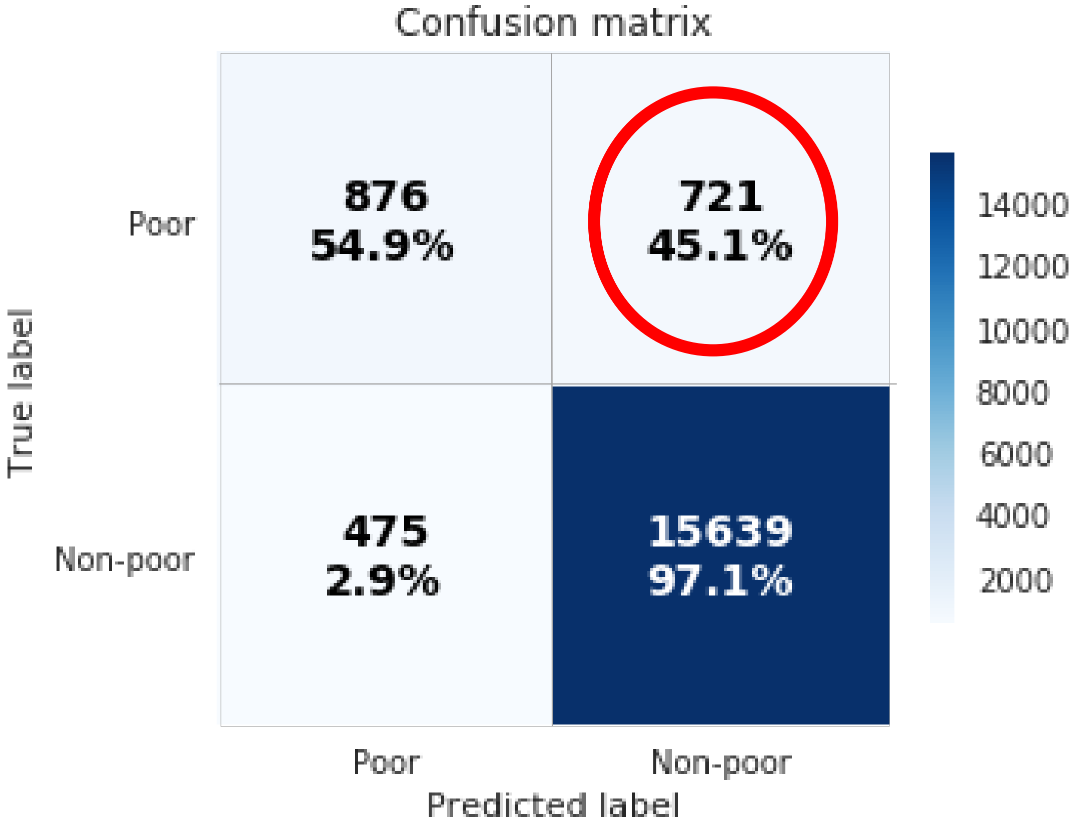
- A few lines of code, but a computationally intensive process (thousands of models are tested)
- ~2 days on a 32-processors server (200 generations)
- TPOT returns a python script that implements the best pipeline
  - IDN → 6 pre-processing steps including some non-standard ones (creation of synthetic features), then XGBoost (models assessed on *f1* measure)
- A counter-intuitive pipeline; it works, but not clear why

Results: DeepFM, TPOT, and some others (IDN)

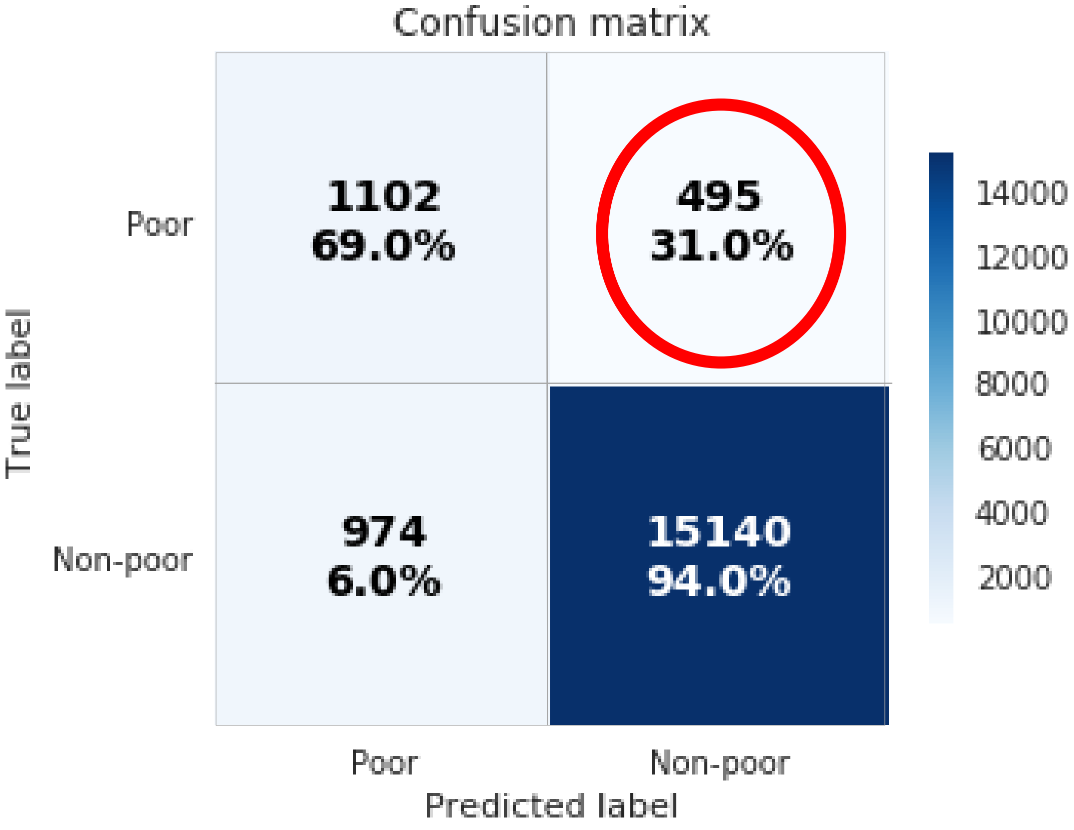
Algorithm	Accuracy	Recall	Precision	f1	Cross entropy	ROC AUC	Cohen Kappa	Mean rank
DeepFM	0.932	0.549	0.648	0.594	0.163	0.943	0.558	3.571
xgb_full_undersample_cv	0.833	0.893	0.400	0.552	0.376	0.932	0.448	4.143
lr_full_oversample_cv	0.853	0.838	0.431	0.569	0.347	0.926	0.471	4.714
nb_full_undersample_cv_isotonic	0.820	0.913	0.383	0.539	0.402	0.932	0.434	5.714
svm_full_undersample_cv	0.815	0.928	0.377	0.536	0.402	0.933	0.435	5.857
mlp_full_undersample_cv	0.819	0.904	0.380	0.535	0.391	0.930	0.434	6.714
rf_full_undersample_cv_ada	0.823	0.907	0.386	0.542	0.530	0.931	0.429	6.857
lr_l1_feats_oversample_cv	0.831	0.843	0.393	0.536	0.383	0.915	0.408	7.286
TPOT	0.917	0.690	0.531	0.600	0.622	0.815	0.555	7.571
lda_full_oversample_cv	0.814	0.887	0.372	0.524	0.425	0.922	0.408	9.286

- DeepFM is the best model on many metrics, but with an issue on *recall*
- TPOT is the best performer on *f1* and does well on *accuracy*, but overall it is far from the top performing models

DeepFM and TPOT – Confusion matrices



DeepFM

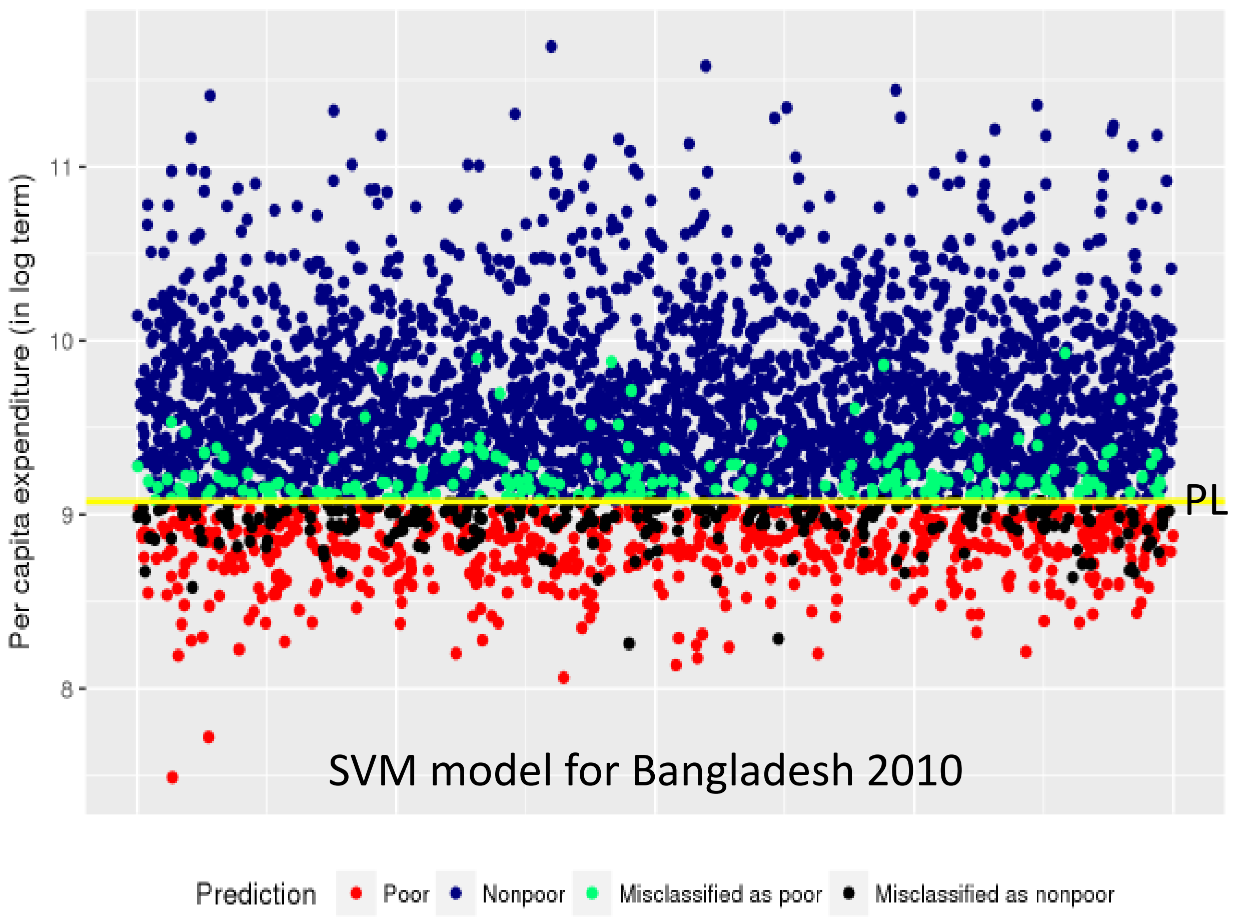


TPOT

Next steps

- Analysis of misclassifications
- Test robustness over time
- Assess impact of sample size
- Expand to regression algorithms
    - Complement existing and ongoing research

Distribution of predictions around the poverty line



Some takeaways

- ML provides a powerful set of tools for classification/prediction
  - Predicting poverty rates is challenging (we need better predictors more than we need better tools)
- Results should always be reported using multiple quality metrics
  - Different performance metrics are appropriate for different purposes
  - Good model = model “fit for purpose”
  - Quality has multiple dimensions (predictive performance, computational constraints, interpretability, and ease of deployment/maintenance/updating)
- Openness and full reproducibility must be the rule
  - Open data when we can ; open source software preferably ; open scripts always
  - Documented scripts published in GitHub (Jupyter Notebooks, R Markdown)
  - Need a metadata standard for cataloguing, and to foster meta-learning

Topic Modeling
A quick look at 145,000 World Bank documents

Improving data (and document) discoverability

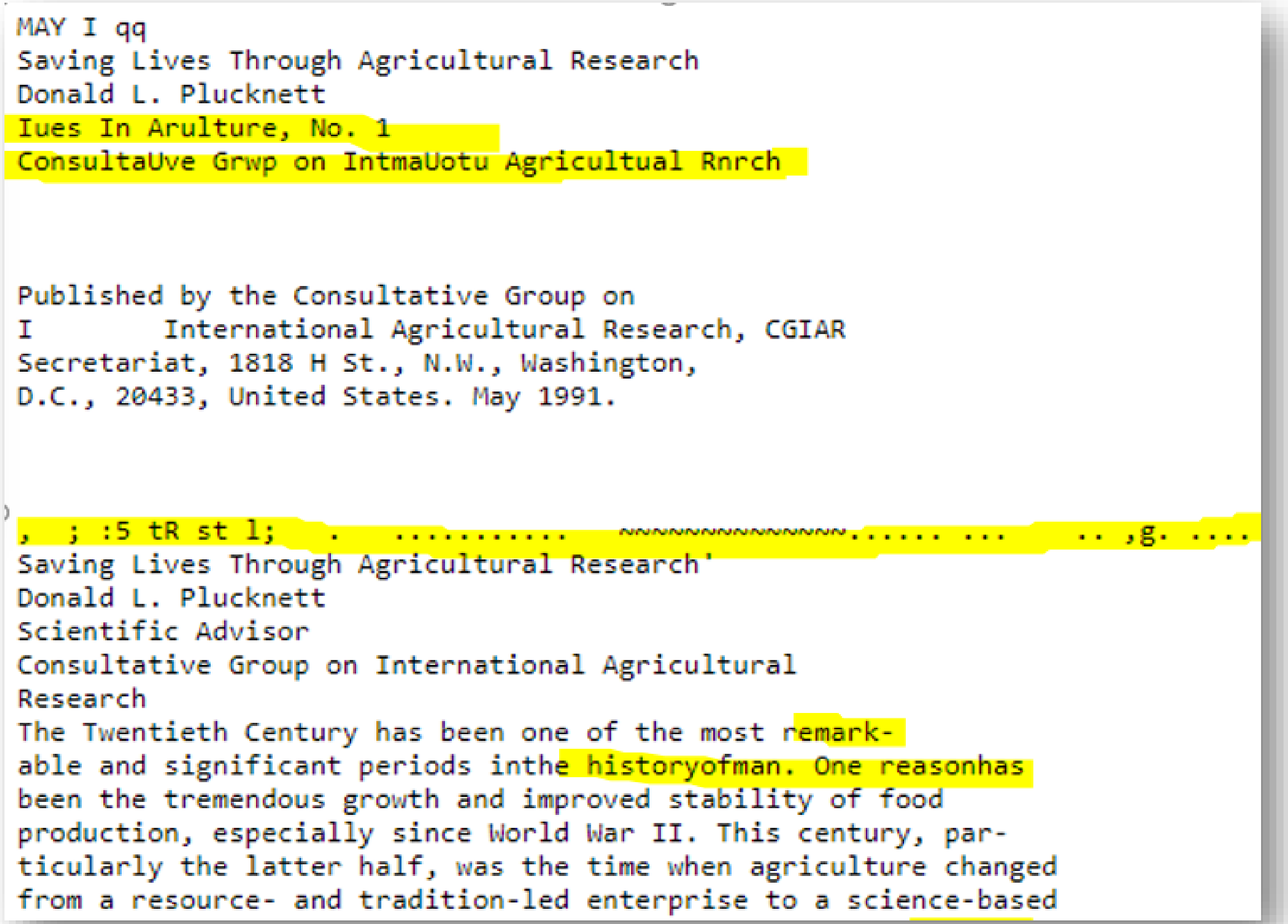
- Our data discovery solutions are not optimal
- E.g., searching “inequality” in the WB Microdata Library only returns 17 surveys
- Reason: relies on full-text search on survey metadata
  - “Inequality” not in survey metadata
- One solution: mine the analytical output of surveys (70,000+ citations)

Improving data (and document) discoverability

- What we want:
  - Fully automatic extraction of topics covered in these documents
  - An open source solution which does not require a pre-defined taxonomy (not a topic tagging system)
- One solution: Latent Dirichlet Allocation (LDA) algorithm
  - LDA topics are lists of keywords likely to co-occur
  - User-defined parameter for the model: number of topics
- Before applying it to survey citations, we tested it on the WB Documents and Reports - a well curated collection of > 200,000 documents openly accessible through an API

Preparing data

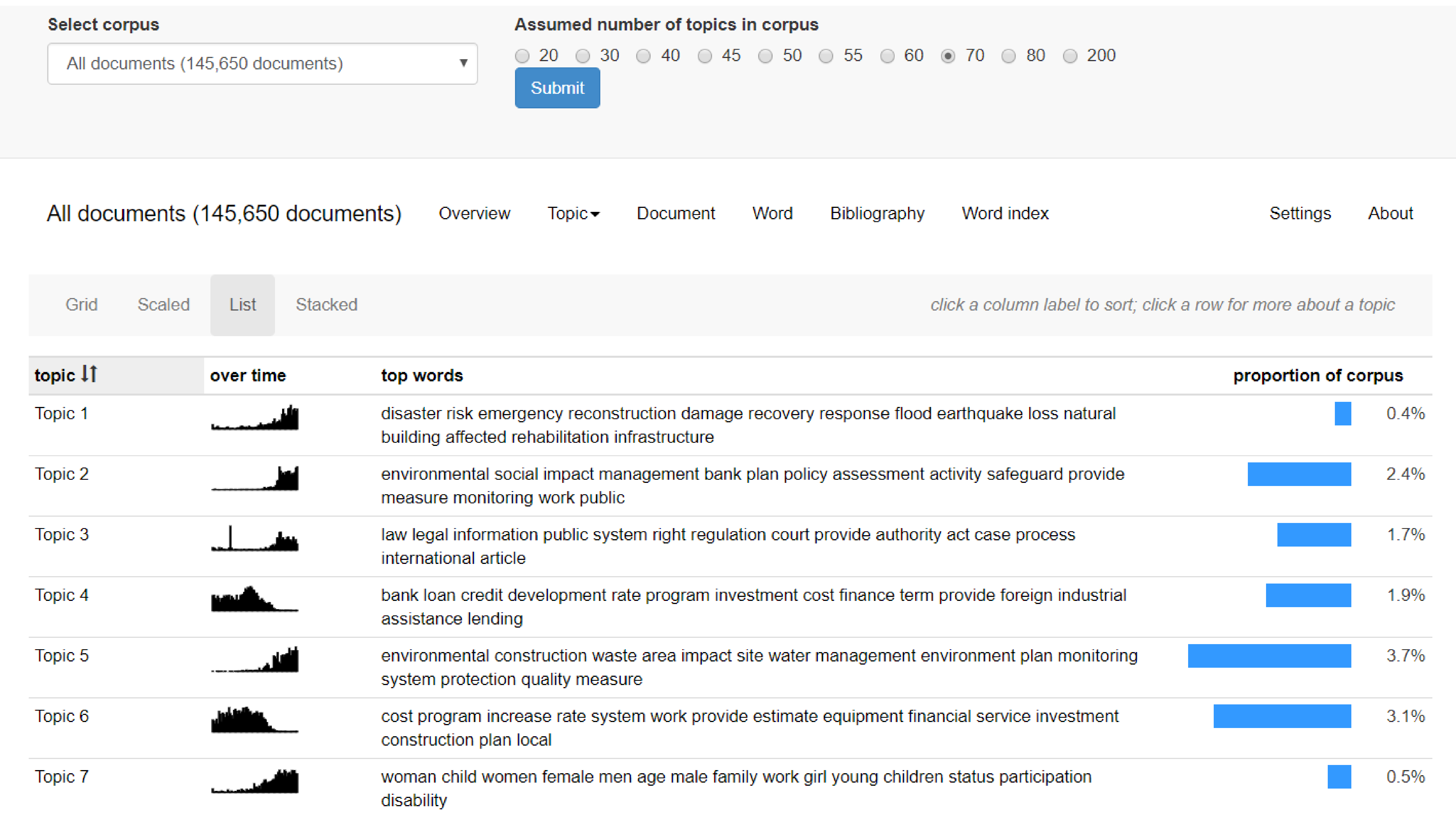
- Text is unstructured, sometimes messy data
- A “cleaning” process is required



```
MAY I qq
Saving Lives Through Agricultural Research
Donald L. Plucknett
Iues In Arulture, No. 1
ConsultaUve Grwp on IntmaUotu Agricultual Rnrch



Published by the Consultative Group on
I        International Agricultural Research, CGIAR
Secretariat, 1818 H St., N.W., Washington,
D.C., 20433, United States. May 1991.



,  ; :5 tR st l;     .    ..........    ~~~~~~~~~~~~~~......  ...     .. ,g. ....
Saving Lives Through Agricultural Research'
Donald L. Plucknett
Scientific Advisor
Consultative Group on International Agricultural
Research
The Twentieth Century has been one of the most remark-
able and significant periods inthe historyofman. One reasonhas
been the tremendous growth and improved stability of food
production, especially since World War II. This century, par-
ticularly the latter half, was the time when agriculture changed
from a resource- and tradition-led enterprise to a science-based
```

Preparing data - Procedures

- **We clean the text files** (Python, NLTK library)
  - Detect language; keep document if > 98% in English
  - Lemmatization (convert words to their dictionary form)
  - Remove numbers, special characters, and punctuation
  - Remove words that are not in the English Dictionary
  - Remove stop-words (“and”, “or”, “the”, “if”, etc.)
- **We obtain a clean corpus (145,000 docs ; ~ 800 million words)**
  - Generate a “bag of words” (documents/terms matrix)
- **We run the LDA model** (Mallet package)
  - Output published in a topic browser (adapted from dfr-browser)


Select corpus
All documents (145,650 documents)
Assumed number of topics in corpus
20 30 40 45 50 55 60 70 80 200
Submit
All documents (145,650 documents)
Overview
Topic
Document
Word
Bibliography
Word index
Settings
About
Grid
Scaled
List
Stacked
click a column label to sort; click a row for more about a topic
topic
over time
top words
proportion of corpus
Topic 1
disaster risk emergency reconstruction damage recovery response flood earthquake loss natural building affected rehabilitation infrastructure
0.4%
Topic 2
environmental social impact management bank plan policy assessment activity safeguard provide measure monitoring work public
2.4%
Topic 3
law legal information public system right regulation court provide authority act case process international article
1.7%
Topic 4
bank loan credit development rate program investment cost finance term provide foreign industrial assistance lending
1.9%
Topic 5
environmental construction waste area impact site water management environment plan monitoring system protection quality measure
3.7%
Topic 6
cost program increase rate system work provide estimate equipment financial service investment construction plan local
3.1%
Topic 7
woman child women female men age male family work girl young children status participation disability
0.5%

Topic 1

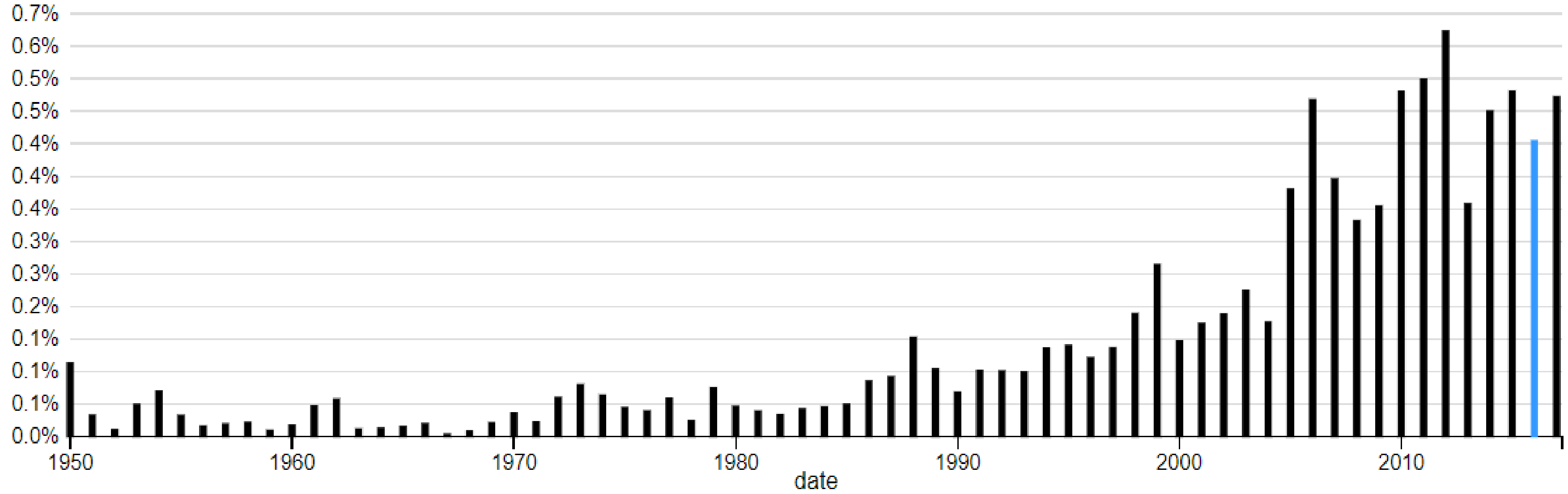
Top words

Word	Weight
disaster	
risk	
emergency	
reconstruction	
damage	
recovery	
response	
flood	
earthquake	
loss	
natural	
building	
affected	
rehabilitation	
infrastructure	
hazard	
human	
safety	

Conditional proportion of words in topic

Click a bar to limit to the documents it represents

clear selected



Top documents: 2016

Document	%	Tokens
[Anon]. “World Bank’s India disaster risk management program.” *090224b0840365f8* (January 2016): undefined.	40.2%	617
[Anon]. “Preparedness map for community resilience : earthquakes.” *090224b084adb9bf* (January 2016): undefined.	39.9%	1714
[Anon]. “Sri Lanka - Post-disaster needs assessment : floods and landslides.” *090224b084c0d0b8* (January 2016): undefined.	36.8%	18911
[Anon]. “Using the catastrophe risks Deferred Drawdown Option	36.3%	177

[Anon]. "World Bank's India disaster risk management program." *090224b0840365f8* (January 2016): undefined.

1534 tokens. (view publication)

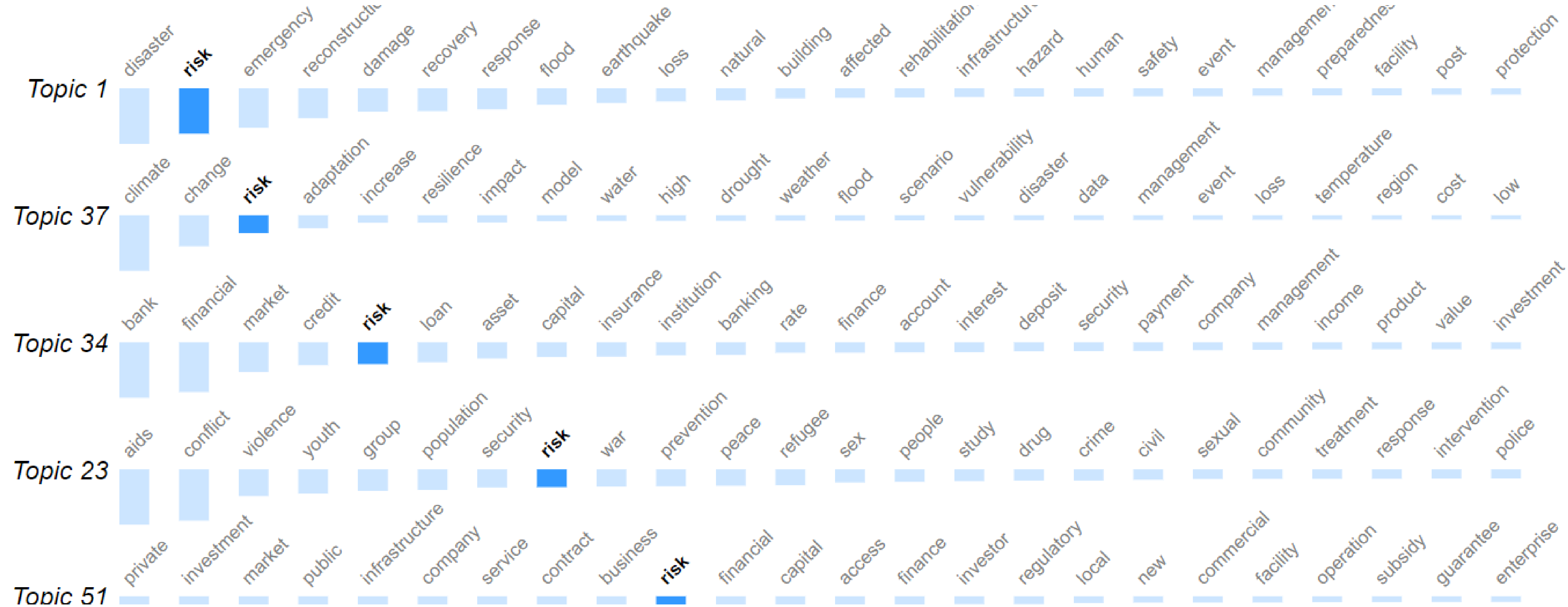
Topic	Top words	%	Tokens
Topic 1	disaster risk emergency reconstruction damage recovery response flood earthquake loss natural building affected rehabilitation infrastructure	40.2%	617
Topic 37	climate change risk adaptation increase resilience impact model water high drought weather flood scenario vulnerability	25.1%	385
Topic 29	development support program capacity activity level national service community local investment regional improve develop strategy	23.4%	359
Topic 40	agricultural farmer area agriculture land farm crop production irrigation rural extension rice food research input	2.9%	44
Topic 53	urban city municipal municipality area housing land infrastructure public development population service local planning cities	2.5%	39
Topic 69	state federal registration states register fee property inspection certificate panel lease federation payment office comments	2.2%	33
Topic 14	road maintenance transport highway roads construction traffic rehabilitation network study improvement bridge safety vehicle section	1.4%	22
Topic 59	line area impact power transmission site land substation	1.4%	22

All documents (145,650 documents) Overview Topic Document Word Bibliography Word index Settings About

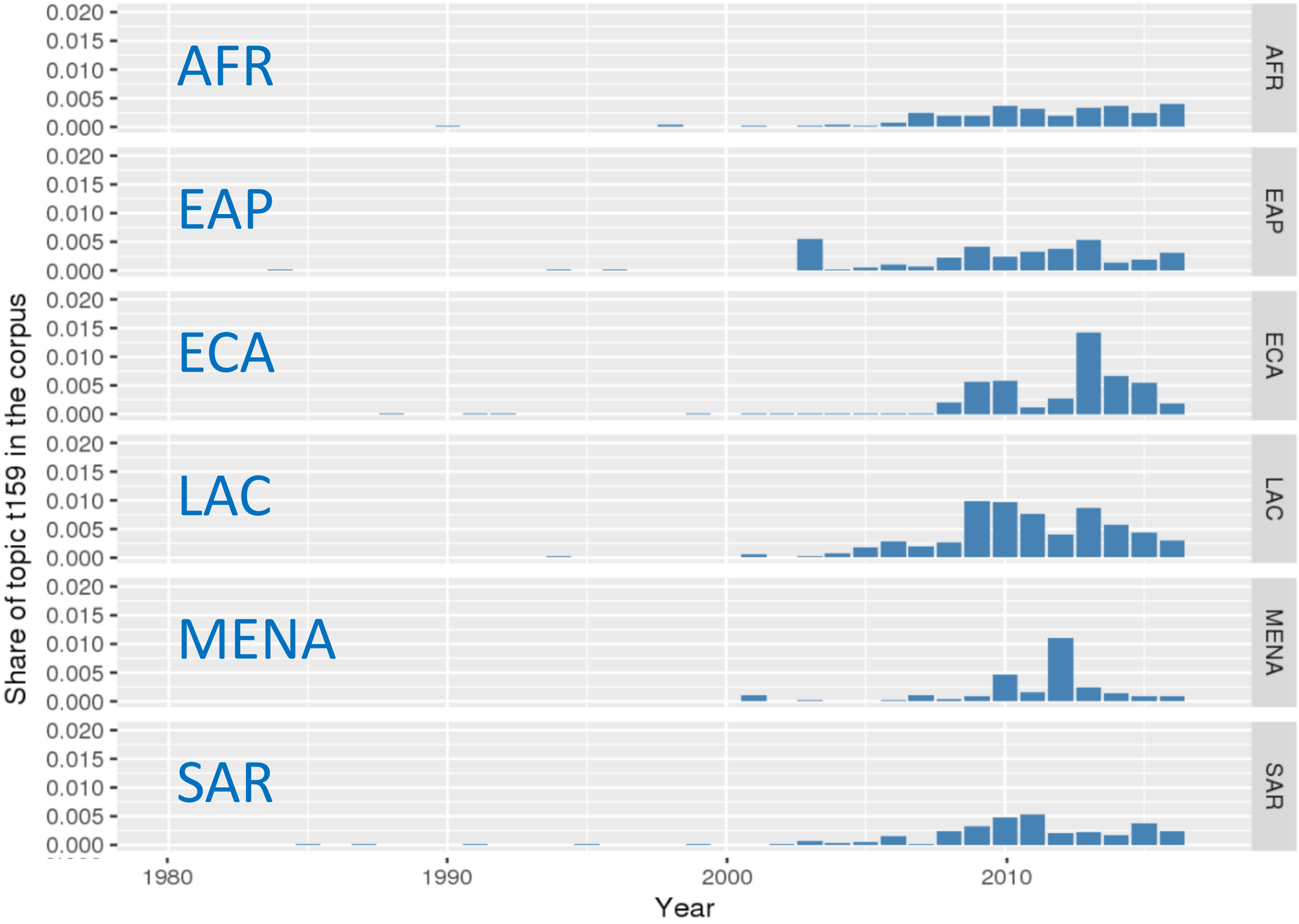
risk List topics

Prominent topics for *risk*

Click row labels to go to the corresponding topic page; click a word to show the topic list for that word.



Analysis: differences across regions

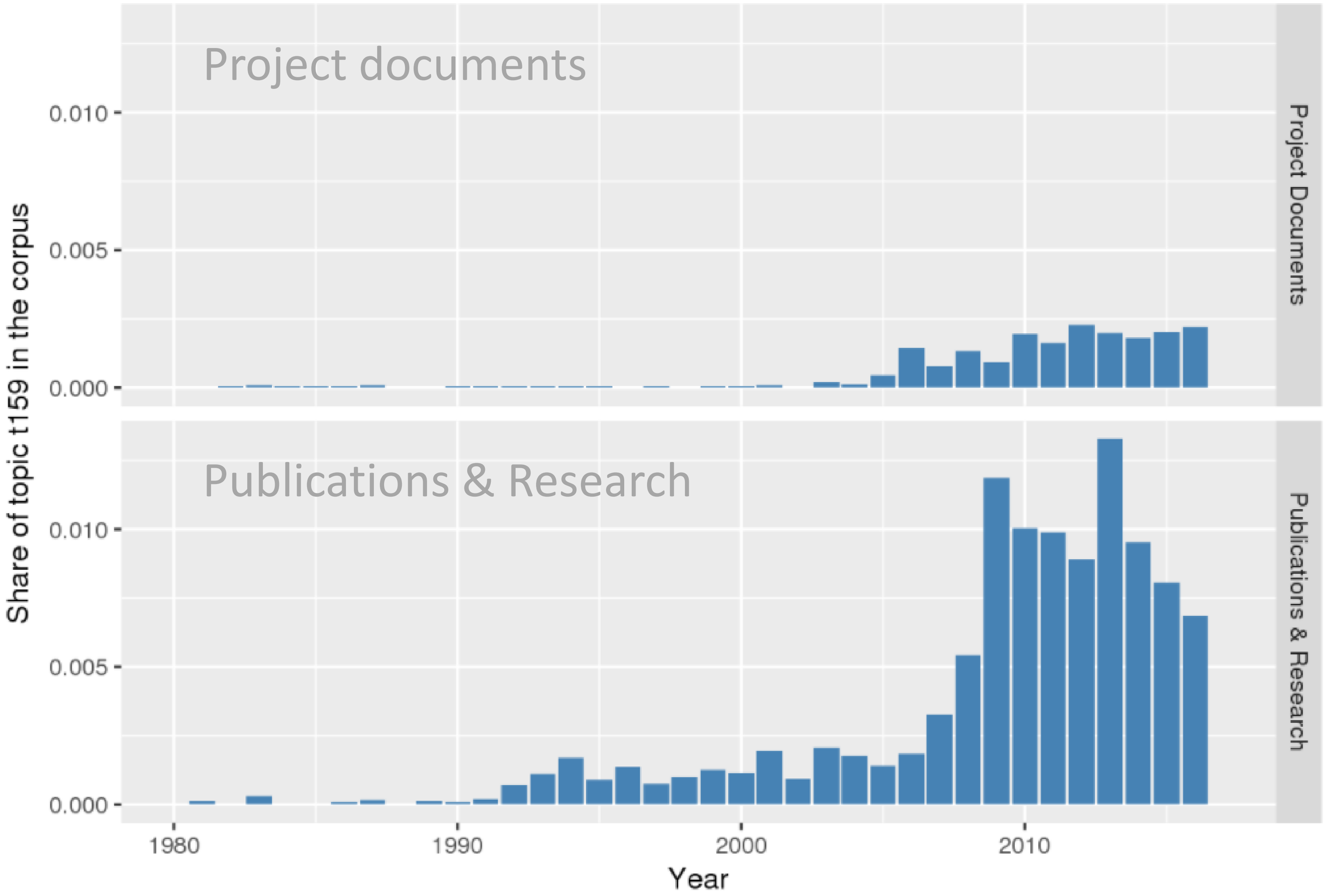


1980 – 2017

climate,
change,
adaptation,
increase,
impact,
resilience,
risk,
water,
vulnerability

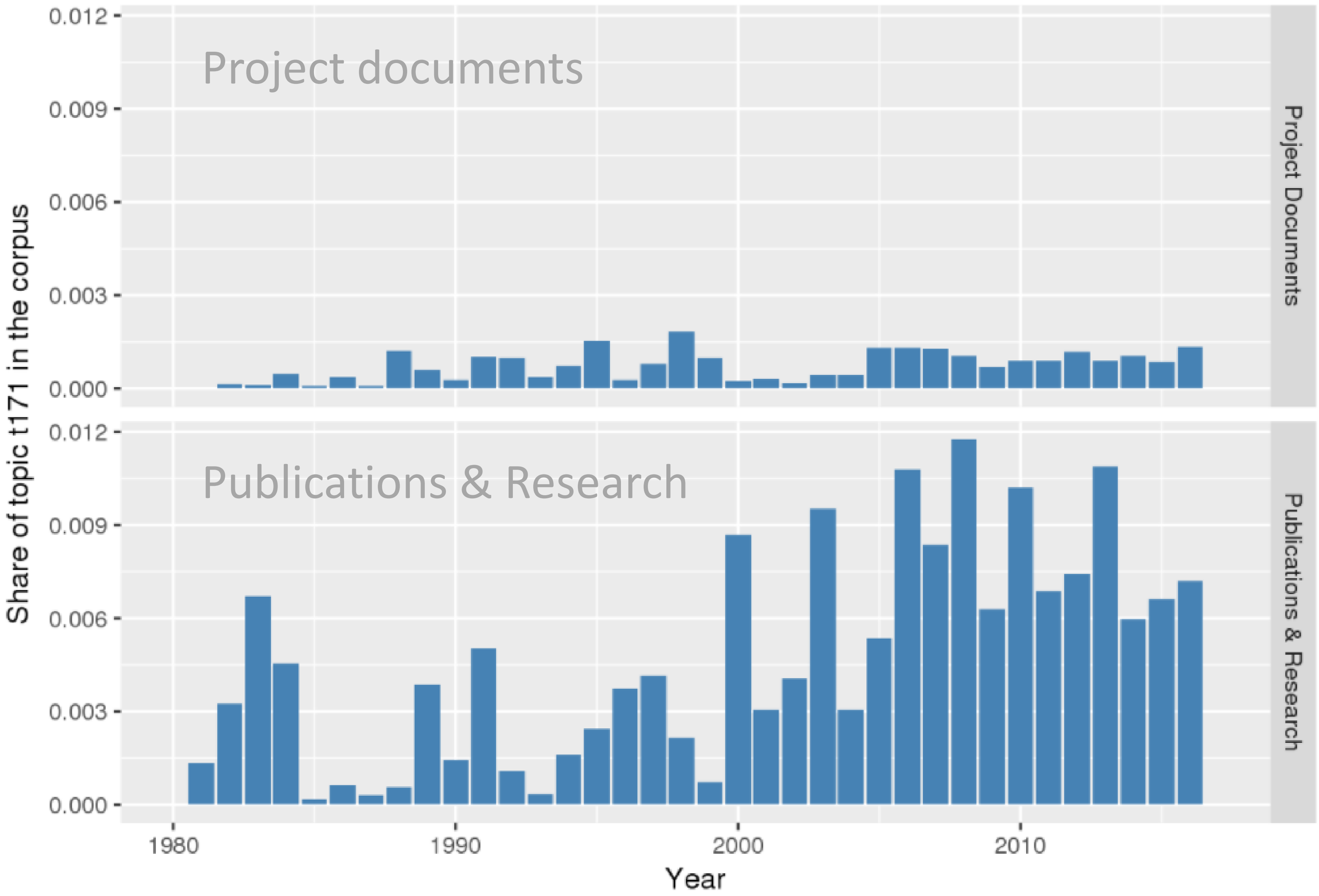
Analysis: differences across document types

climate,
change,
adaptation,
increase,
impact,
resilience,
risk,
water,
vulnerability



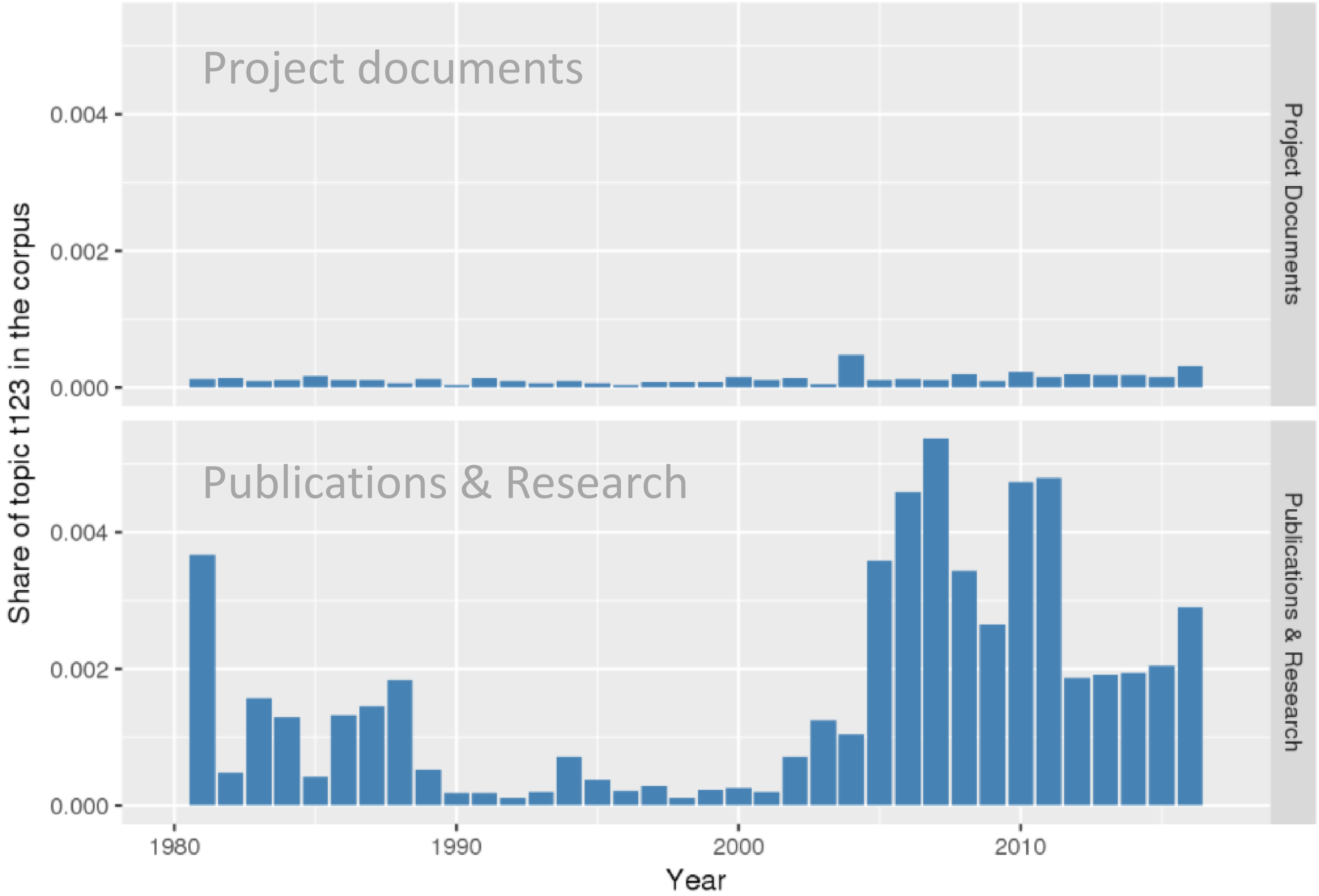
Analysis: differences across document types

technology,
innovation,
new,
development,
knowledge,
market,
economy,
competitiveness



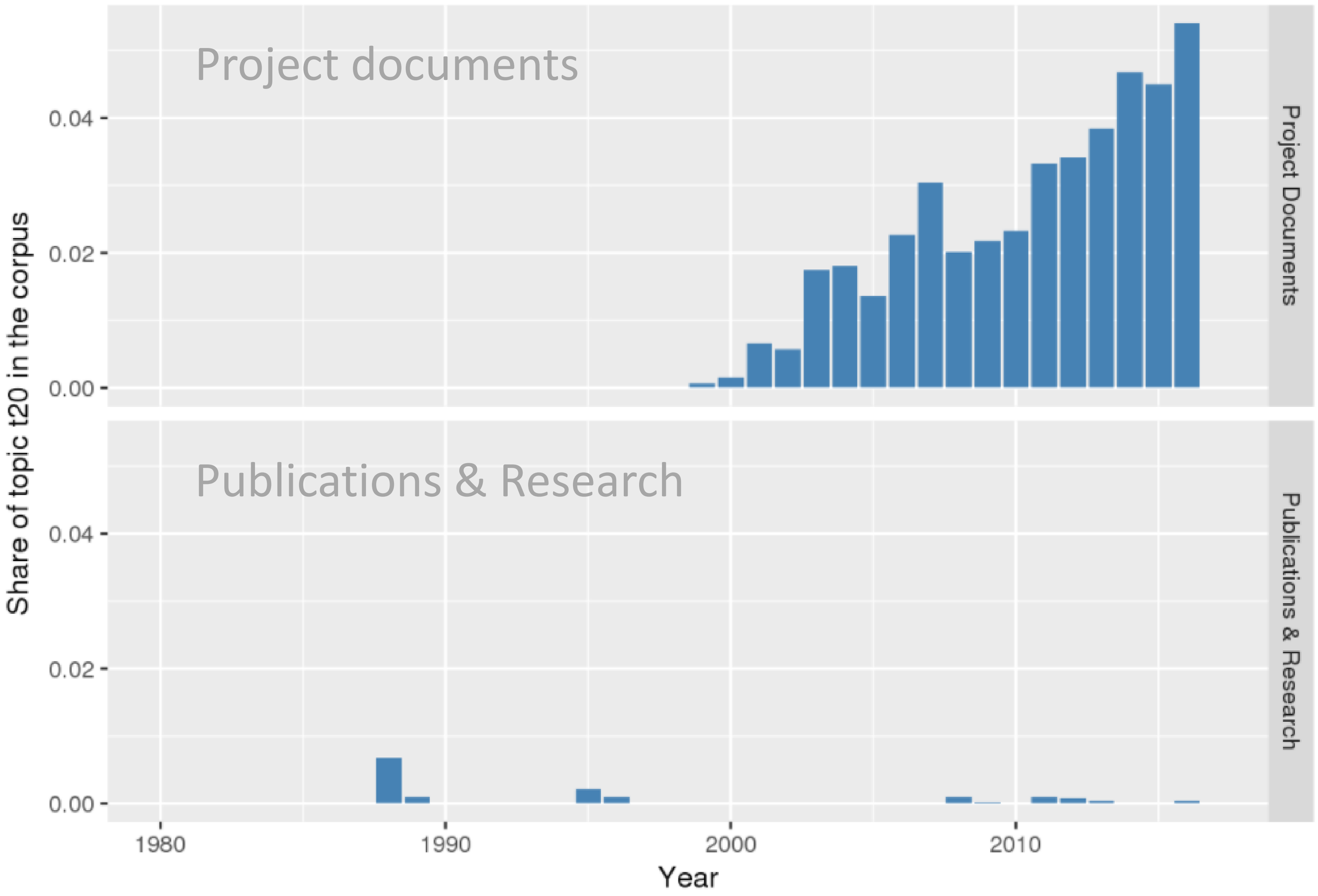
Analysis: differences across document types

migration,
migrant,
remittance,
international,
home,
return,
diaspora

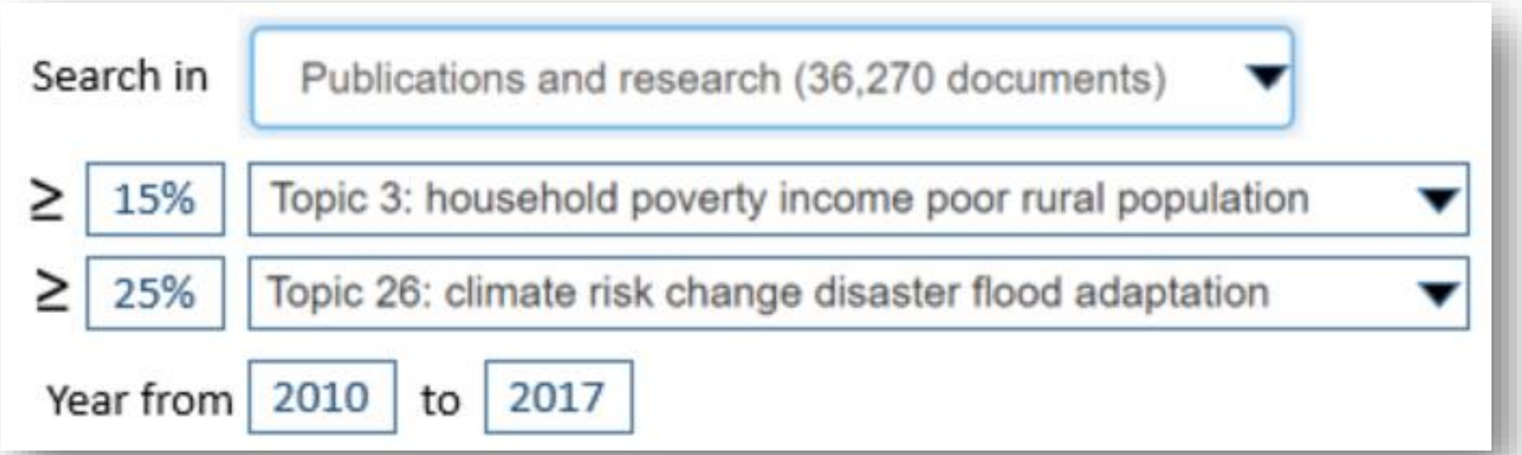


Analysis: differences across document types

land,
resettlement,
compensation,
affected,
policy,
area,
community



Finding documents based on topic composition



1. Model and methods for estimating the number of people living in extreme poverty because of the direct impacts of natural disasters
2. The Varying Income Effects of Weather Variation - Initial Insights from Rural Vietnam
3. Weathering storms : understanding the impact of natural disasters on the poor in Central America
4. The exposure, vulnerability, and ability to respond of poor households to recurrent floods in Mumbai
5. Climate and disaster resilience of greater Dhaka area : a micro level analysis
6. Why resilience matters - the poverty impacts of disasters
7. The poverty impact of climate change in Mexico

Finding closest neighbors

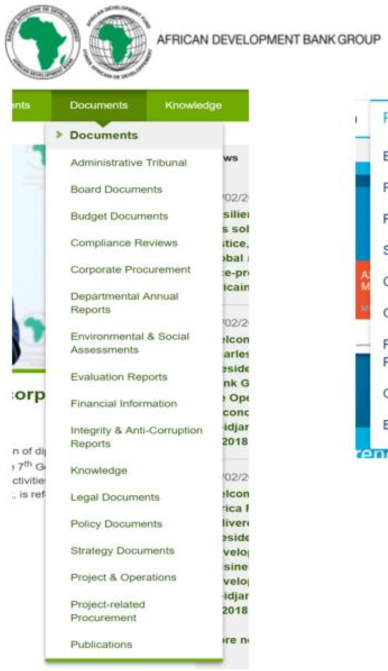
Upload or select a document, and find the N closest neighbors, e.g.:

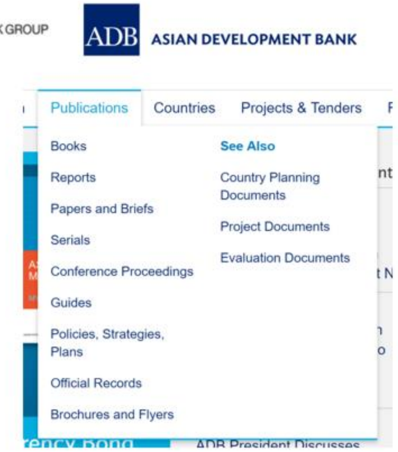
Monga, C. 2009. *[Uncivil societies - a theory of sociopolitical change]* Top 10

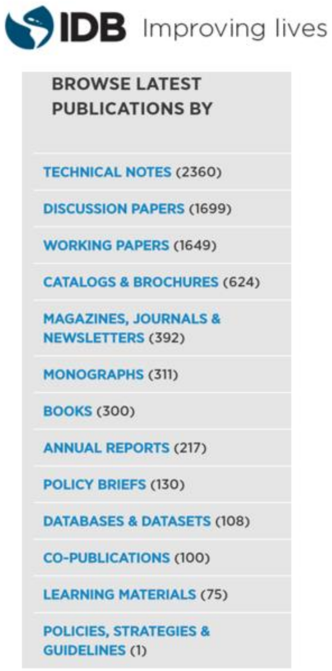
Inclusion matters : the foundation for shared prosperity
Representational models and democratic transitions in fragile and post-conflict states
How and why does history matter for development policy ?
Somalia and the horn of Africa
Limited access orders in the developing world :a new approach to the problems of development
Intersubjective meaning and collective action in 'fragile' societies : theory, evidence and policy implications
Equilibrium fictions : a cognitive approach to societal rigidity
The new political economy : positive economics and negative politics
The politics of the South : part of the Sri Lanka strategic conflict assessment 2005 (2000-2005)
Civil society, civic engagement, and peacebuilding

Expanding the corpus (not yet implemented)

A fully automated system collects documents from WB and other organizations, “cleans” them, extract topics, and update the browser and search UI







Web scraping (cron job; e.g. weekly run)

Process documents and infer topics

Publish in browser / search interface