



# Az ELKH Cloud előnyei a POLTEXT projekt példáján



Sebők Miklós (TK PTI),  
Kacsuk Zoltán (HdM IAAI)

# Research context

Intersection of two research projects at the Centre for Social Sciences, Institute for Political Science:

- **Text Mining of Political and Legal Texts (POLTEXT) Incubator Project** (Principal Investigator: Miklós Sebők)
- **Hungarian Comparative Agendas Project (CAP) project** (Project leaders: Zsolt Boda & Miklós Sebők)

# Research problem

- **Quantitative analysis** of qualitative data
  - Classifying articles according to policy topics
  - Topics: education to defense, Comparative Agendas Project
- **Gold standard:** double-blind human coding by well-trained researchers
- What if this is **unfeasible?**
  - Article counts of over 100.000
  - **Cost and training** of human coders for this scale



## Markó Béla szerint nem jó ötlet a küllhoni magyarok szavazása

• Sok Fidesz-szimpatizáns voksolt az RMDSZ-re Erdélyben.  
• A választók pragmatikusak – állítja. Interjú →13

ÚJSÁGOKBÓL KÉSZÜLT SZALÁMI NEWYORKBAN Rivalda →20

Családi gazdálkodás Elégedetlenség Fejér megyében

## Földfoglalás Kajászón



Félszázézer négyzetméter az száz futbalpályára FOTÓ BEVICKZY ZSOLT

Rab László

Az a kajásói gazdálkodás, amely a Nemzeti Földalapkezelő haszonbérleti pályázatán egyetlen négyzetméternyi területet sem nyert el, úgy döntött, hogy az egyik nyertes területén hetektárt művelésbe von.

Az engedéltségi akcióra azért szánták rá magukat, mert a 284 hektáron nyertes képviselői mádnárra vállalkoztak, *Kiss Árpád* a hatékony koronás öszmélet 2194 aranykoronával túllépte. Az ezt képviselő területet választották le Kajászón határában, és a hetven hektárt vasárnap beosztották.

A gazdák közbizhatóságát alapították, és mint mondták, nem térnek el a törvénytelen állapotok, az eltagadást nem tudják.

A helyiek a pályázatot kihirdetése után borszagadtak föl először, amikor kiderült, hogy a több mint 400 hektár

nyi állami földterületből csak az említett *Kiss Árpádnak*, illetve a felcsúti polgármester, *Mészáros Lőrinc* érdekeltségébe jutott húszéves bérletre a falu határában. Az itteni visszaszárogás *Ángyán István* volt vidékfejlesztési államtitkár is kitért nevévesztés tanulságában. Angyán a földterületek körül kialakított visszaszárogási mintáit mondott le tisztázta.

Visszatérve Kajászóra: a helyiek szerint a képviselői vállalkozó, *Kiss Árpád* a politikai kapcsolatait felhasználva jutott hozzá az állami föld bérletéhez, művelésre alkalmas gépet mincsenek, másval végzettségi mezőgazdasági munkákat.

A vasárnap akció a helyi gazdálkodás, a *Péter Szívó* és az *Öcsés* területén a Földterület és Vidékért Egyesület támogatásával történt. A pályázatot a Földalapkezelő, *Magyar Zoltán*, a mezőgazdasági bizottság tagja is.

A Kiss-féle földterület a kajásói

azért választották, mert meggyőződésük róla, hogy a nyertes pályázó a 284 hektárt úgy nyerte el, hogy a terület nagysága 8194 aranykorona értékű. Az NFA-pályázat kiírásában pedig az szerepel, hogy egy gázta nem nyelhet többet 6000 aranykoronánál. Míg az Magyarországon nemrég egy pályázatot értett vont vissza a földalap, úgy gondolták, Kajászón is vissza kell állítani a „jörvényes” állapotot. Többeszer keresnek már különben a kapcsolatot a hivatalossal, de eddig senki nem állt velük szóba.

– Az október 17-i helyi gazdálkodás döntése alapján veték a helyiek a kezükbe a földterületet – mondta a határban *Bóor Péter*, a *Péter Szívó* vezetője, és hozzátette, a jelenleg érvényben lévő földterületi veték alapul, amikor úgy döntöttek, hogy beosztják a 70 hektárt, a 6000 aranykoronán felül részt.

– A hetven hektárt az 700 ezer négyzetméter, száz darab futbalpá-

lya – mondta a négy gazdát tömörítő közbizhatóság nevében *Köcskői Csörgő Zoltán* gazdálkodó. Nem gondoltuk volna, hogy így kell érvényes szerezni a törvényességnek, s mindent az állam helyett kell megintenni.

A gazdálkodó szerint az elkészítés is vezette őket, amikor az engedéltséget választották. Szerintük az országban az összes földhaszonbérleti pályázatot vissza kellene vonni, amint a gazdák pedig, akik igazságos módon, rendes körülmények között nyertek – őket szövegek mutogatni a tévben – bírtalanítani kellene. Fejér megyében azonban igen sok lenne az az pályázatos száma, amely megőrt nagyon hiszen rengeteg benne.

Az érintett földterületeken egyébként két traktor vágta el a mélyszántást, másfelől két óra alatt végzett. Előzőleg szalagpal körbekerítették a tábla „fogható” részét, a kitraktó „Kajásói gazdák körélesbe vett terület”.

EGYRE TÖBB AZ ILLEGALIS BELEPŐ

Magyarország szegmens határain szep-tember végéig több mint ötszáz jogellenesen belepő embert fogtak el. A múlt év azonos időszakában még csak 3524-nél tartottak. Afgánok, macedonok, irakok, kosovóiak, szerbek próbálkoznak leginkább. →4



## Alaptalan pedofilvádak miatt távozott a BBC vezérigazgatója

George Entwistle csak nyolc hete lépett hivatalba. Azért kellett lemondania, mert a BBC2 egyik műsorában egy tory politikus tévesen vádolták gyermekmészárlással. →8

## A kormány már lemondott róla, mégsem lesz állami mobilcég?

Lapunk piaci forrásból úgy értesült, hogy a negyedik szövegező bírókötő három állami cég már nem szívesen finanszírozná a több százmilliárdos beruházást. →9

## ÖTÖDIK HASÁB Nem ász

Az Állami Számvevőszék (ÁSZ) Hagelmayer István-díjat kapta Rogán Antal, a parlament gazdasági bizottságának elnöke. Bőnyára fontolgatták, hogy a saját szabadságért, illetve az útieljárás play szabályait megnevezett tépkeelésért járó díjat is juttassanak neki.

A szervezet alapító elnökéről elnevezett díjat azok kaphatják, akik kimagasló teljesítményükkel hozzájárultak az ÁSZ jogállami és társadalmi súlyának kialakításához, a közpénzek hatékonyabb felhasználásához. A szervezetet létehez 1989. évi XXXVIII. törvény elfogadásának napja, október 28-a díjazás napja. Így Rogán Antal kitüntetése erkölcsileg érvénytelen. A Nemzeti Hivatal szerint hazánk 1994. március 19-én elvesztett állami rendelkezéseinek visszavonását 1990. május 1-től, az előző szabadon választott népképviselet megalakulásától számítják. E díjat tehát egy személynél rendszerint mások emlékére hozták létre. Maradjunk is ennyiben.

Rogán Antal ugyanis az egyik legelismertebb, legkompetensebb vezetője, ami – mivel a Belvárosban kevés a lakó, de sok az adóadó nagyszámú – teljesítménye ő az, aki a kötelező lakcímváltoztatás megújítását juttatta, hogy stat viszony az érdekelten választási regisztráción. Korábban a parlamenti vitát elfogó technikai, egyes képviselői indítvánnyal terjesztette be a diszkriminatív különadókat, és (Csor-Palkovics Andrásal együtt) a nemzeti és hazai időbenre kiváló médiumtervező-egység. Nem soroljuk további érdemeit. Hivatkozunk inkább az ünnepi alkalmából a hatalom hírlapjában, amely szerint a huszadik század egyik legnagyobb mértékű vezető értelmiséje után múlt utadon emlékeztünk van a helyi és nemzeti megújításra. Egyet most hátrahagynak. Alig várjuk a megindulást. (A szerk.)

## Vélemény



Válogatott keret 250 millió egy állami cégtől a Fradnak algha nevezhető privát adakozókedvnek – írja Tamás Ervin. →11

Érvek és ellenérvek Marnitz István szerint az energiaár ugyan nem piaci, de nehezen fizethető. →11



# An example



Complete text:  
„The government passed on the option – There will be no state-funded mobile carrier? Our business sources indicate that the three state-owned corporations with a stake in the firm would not want to invest multiple hundred billion HUFs in the fourth carrier.”



Policy code (major topic): 17, which stands for „Space, science, technology, and telecommunications”

# The project

- **Hungarian country project** of the Comparative Agendas Project ([cap.tk.mta.hu](http://cap.tk.mta.hu))
  - Media module – 3 daily newspapers
- For this pilot project:
  - Left-liberal **Népszabadság (NS)**: Over 50 000 front-page articles (1990-2014)
  - Centre-Right **Magyar Nemzet (MN)**: 35 021 articles (2002-2014)
- Hand-coding was unfeasible for our purposes
- Solution: **text mining + machine learning**

# A machine learning solution



- **Text as Data** – qualitative data is converted to quantitative (matrices)
- How to categorize articles into pre-defined classes: **Dictionary-based** or **supervised learning**
- For the latter a sufficiently **large human-coded *training/test set*** is needed

Part 1

# **CREATING A MACHINE CODED TRAINING SET FOR THE LEFT-WING DAILY NÉPSZABADSÁG (NS)**

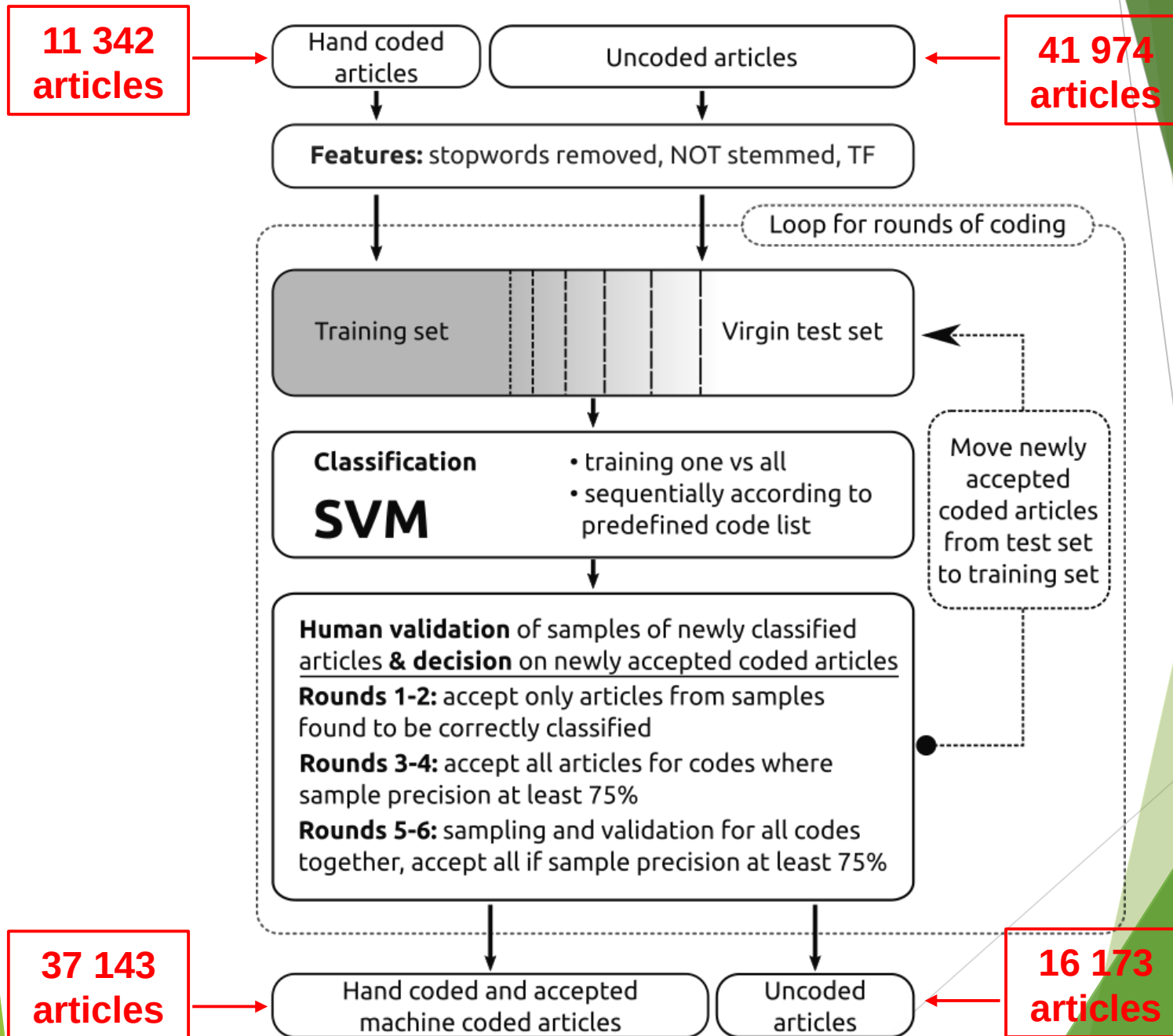
# The Hybrid Binary Snowball (HBS) process



- We need to keep human **coding costs as low as possible**, while extracting the largest possible gain per invested human coding hour
- We simplify multi-class classification by rephrasing it as a **series of pairwise comparisons**
- We apply a snowball method to **augment the training set with machine-classified observations**



# Coding NS articles



# Infrastructural bottlenecks 1: **Memory**

- Our desktop workstation had **only 32 GB RAM**
- Encountered **problems**:
  - Could not work on the **whole virgin data** set
  - Could not run **certain configurations**, for example: Term frequency - Inverse document frequency (Tf-Idf) weighting
- Even the **solutions were problems**:
  - Virgin data was partitioned up for processing
  - This would impact Tf-Idf weighting significantly
- **Real solution** going forward:
  - Using larger capacity single virtual instances or a **cluster in the cloud**

# Infrastructural bottlenecks 2: **Time**

- **Huge numbers of small operations** add up quickly
- If process runtimes become too long, project execution becomes unfeasible
- Solution: **parallelizing** the execution of operations

Part 2

# **USING THE CODED LEFT-WING DAILY ARTICLES TO TEACH THE ALGORITHM HOW TO CODE THE CONSERVATIVE DAILY MAGYAR NEMZET (MN)**

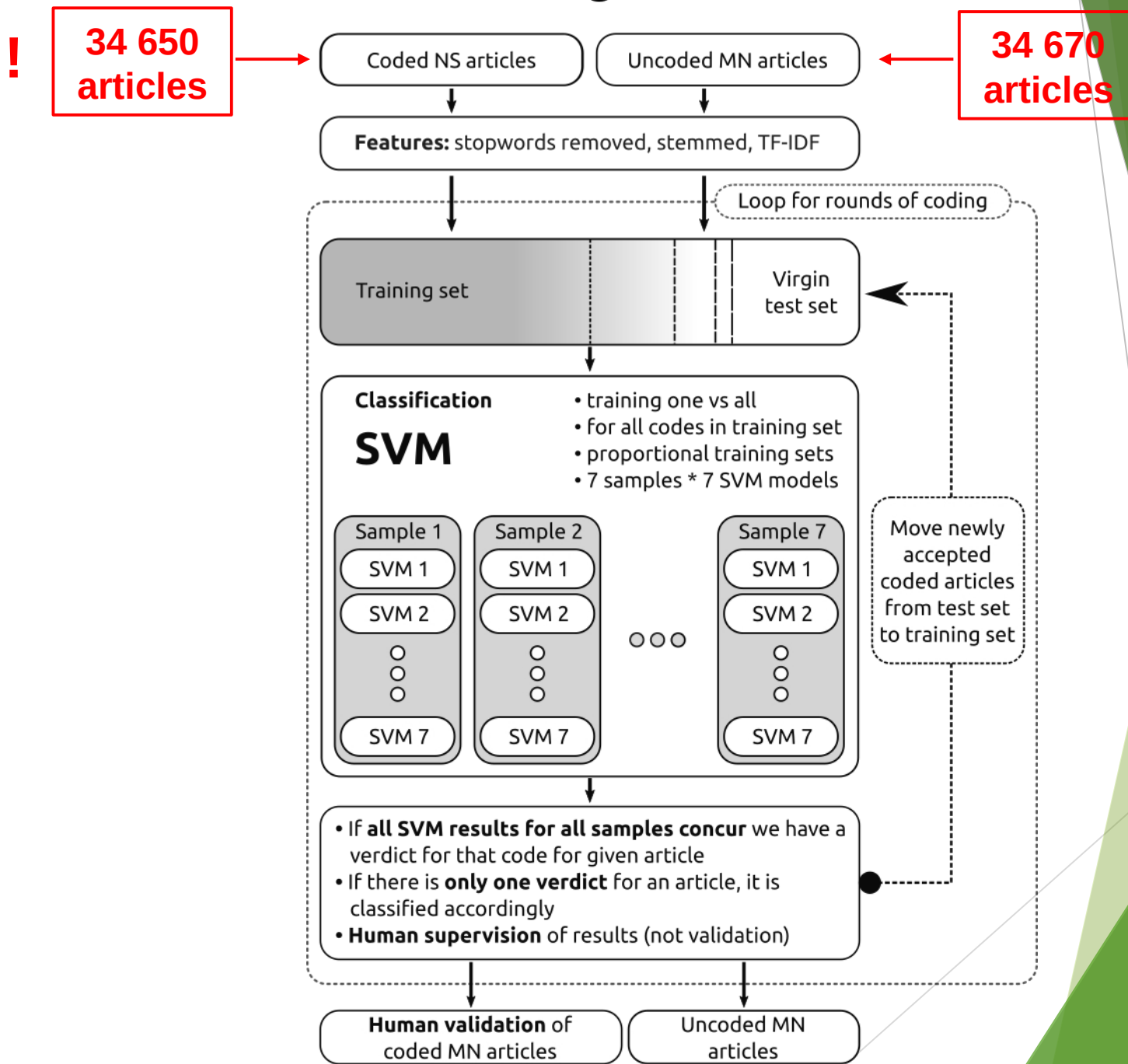
# Apache Spark cluster

- With the help of the Laboratory of Parallel and Distributed Systems at the Institute for Computer Science and Control (**SZTAKI LPDS**)
- Apache Spark cluster running on five virtual instances in the **SZTAKI ELKH Cloud**
- All five virtual instances had 8 virtual processors and 32 GBs of RAM each, and were running Ubuntu 16.04.
- Four instances acted as worker nodes and one as the master node of the Spark cluster. Each Spark session was running with 32 VCPUs (but **default parallelism set to 24**) and **96 GBs of RAM** total on the four worker nodes combined.

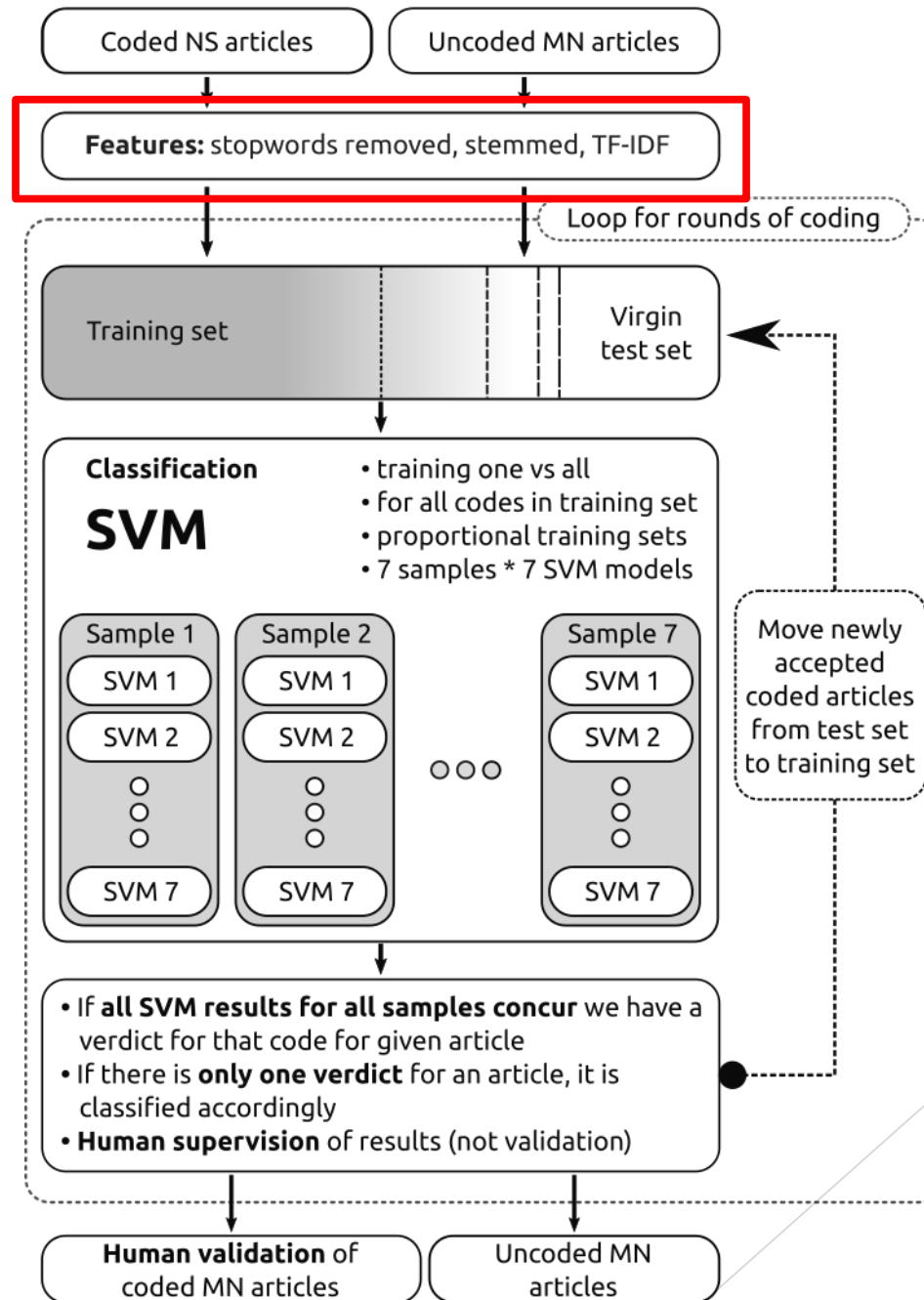
# Manifold increase in speed

- **Old desktop setup:** roughly **3 days** for a full round of coding (33 code categories)
- **Spark cluster:** ca. **30 minutes** for a full round of coding
- This increase in speed enabled:
  - **1) Rapid prototyping**
  - **2) Complex classification workflow**

# Coding MN articles

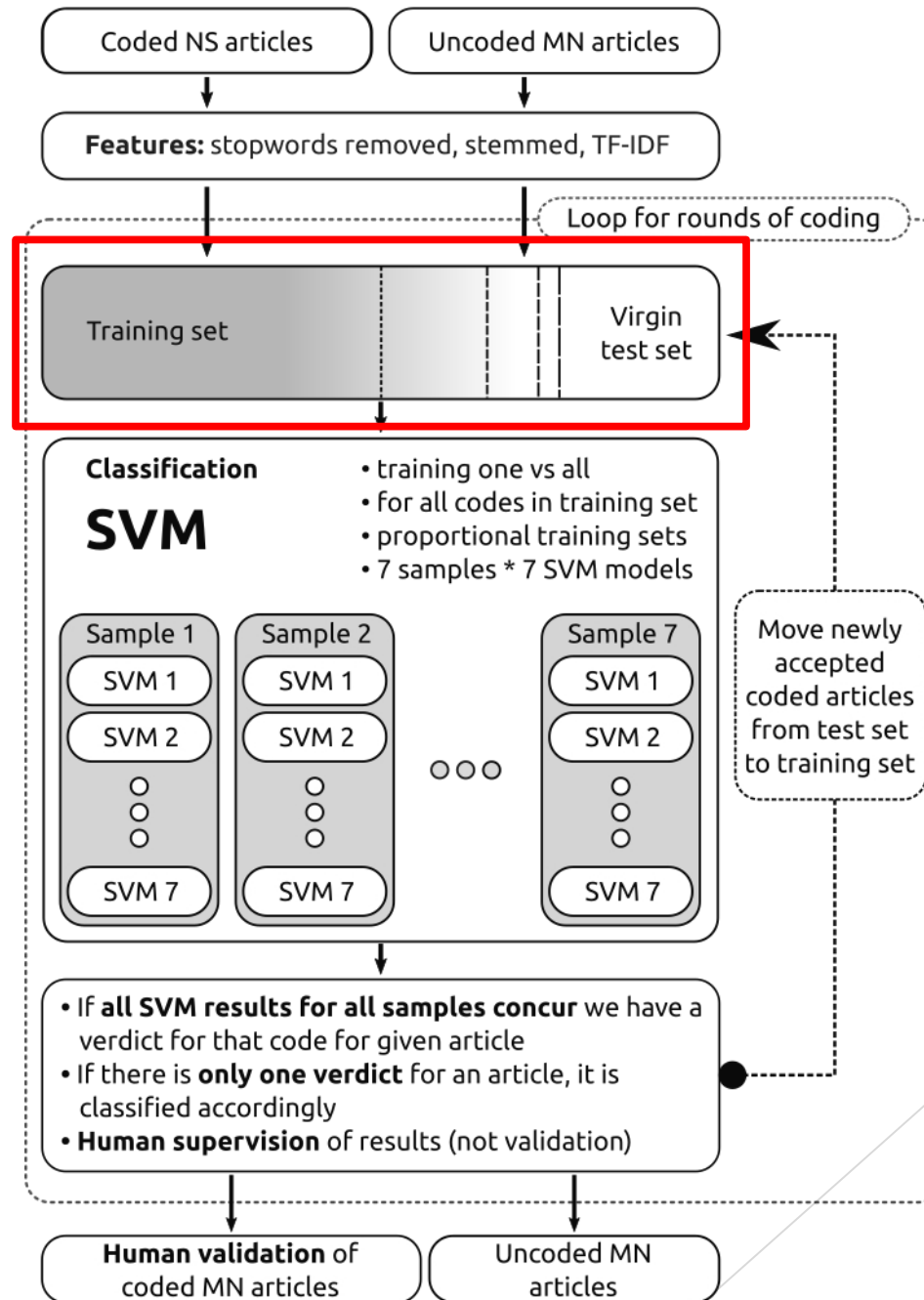


# Coding MN articles

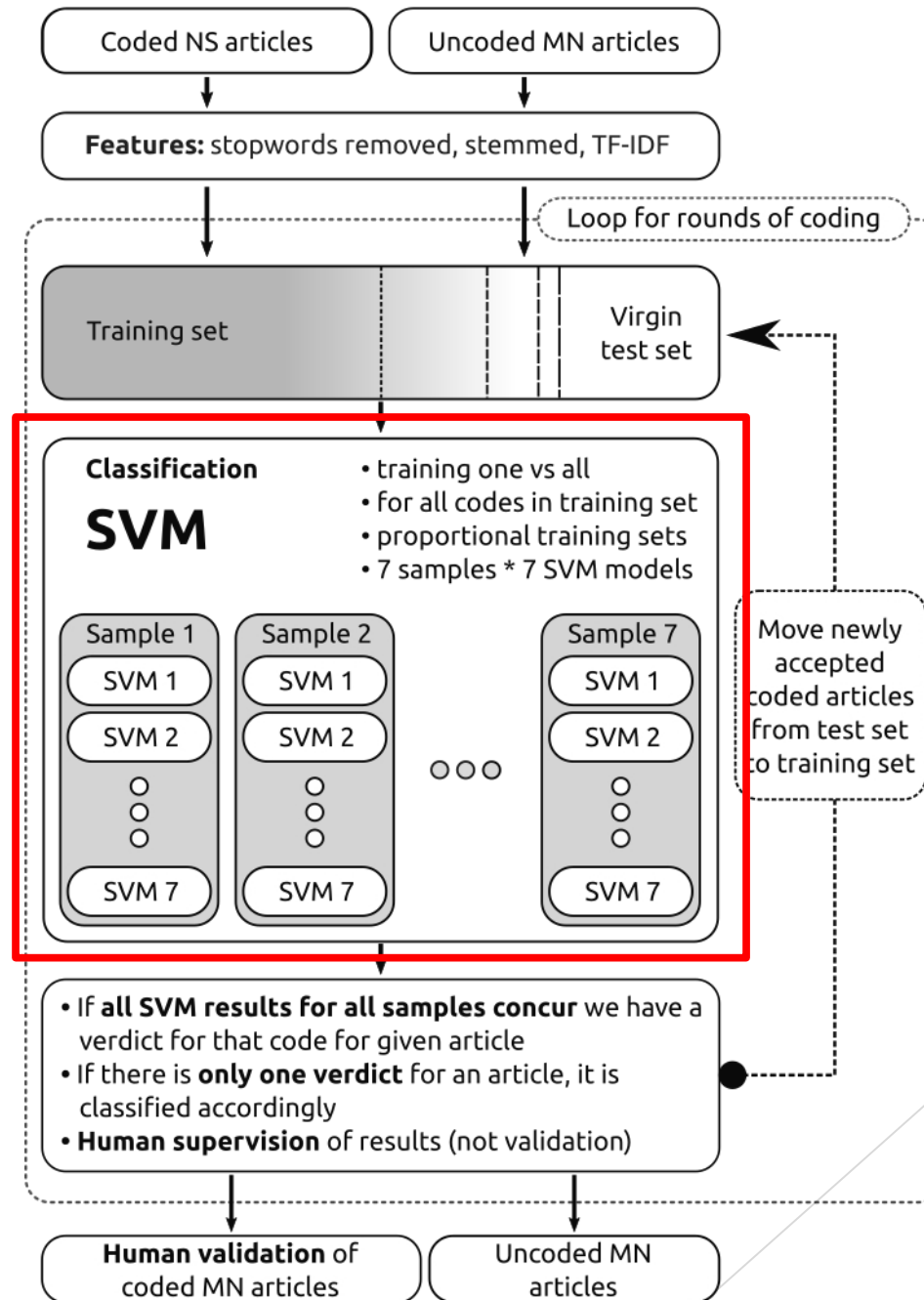




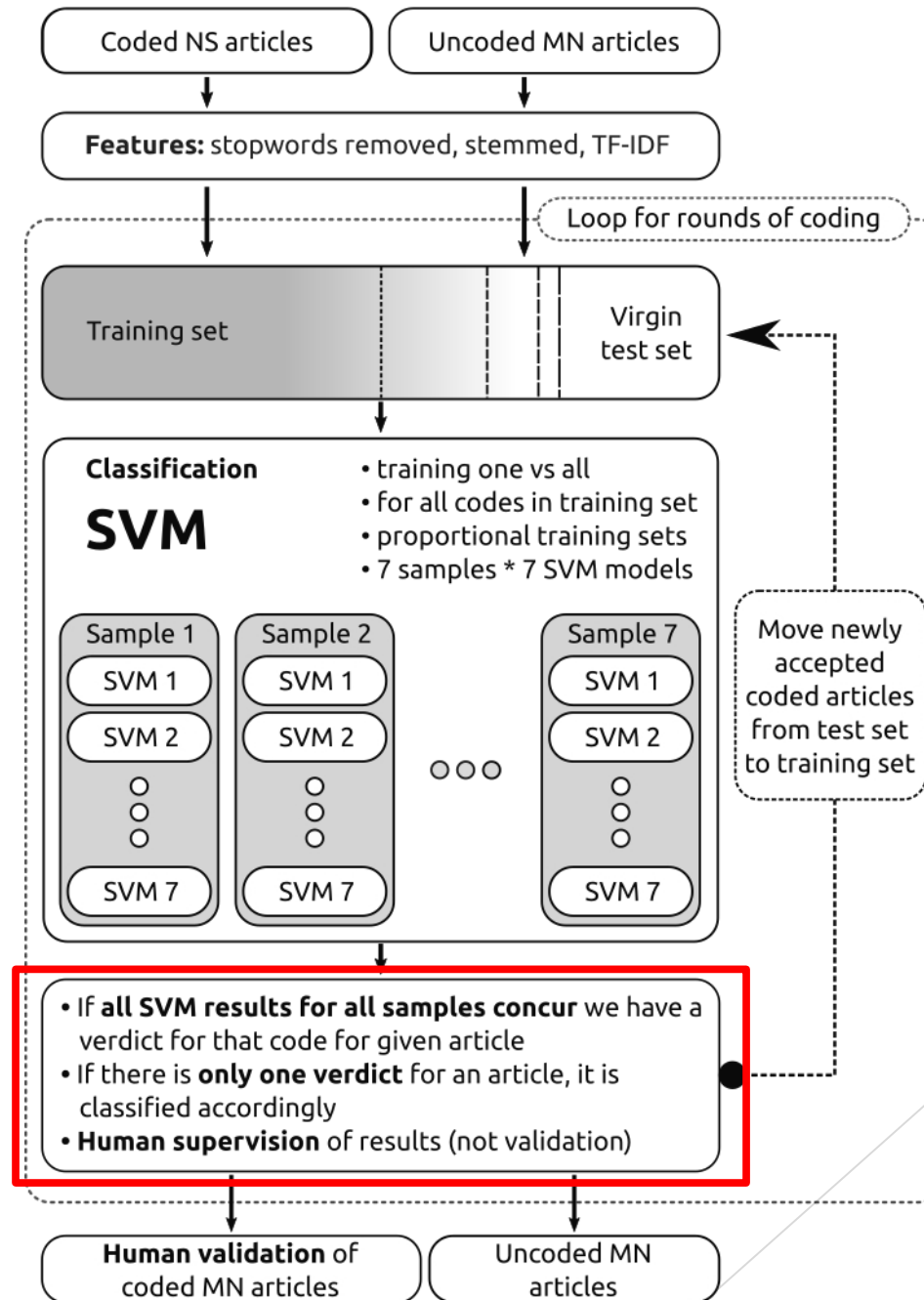
# Coding MN articles



# Coding MN articles



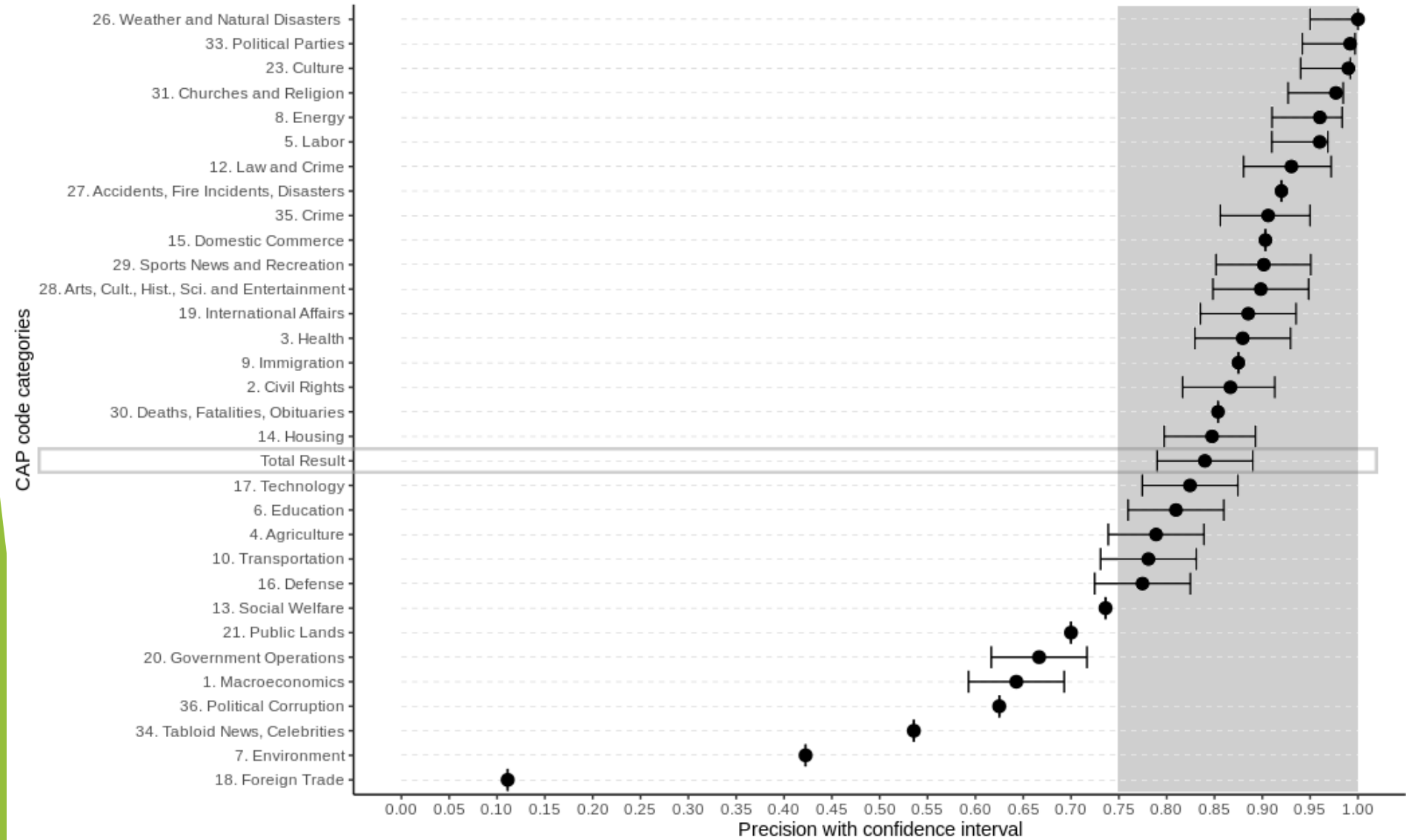
# Coding MN articles



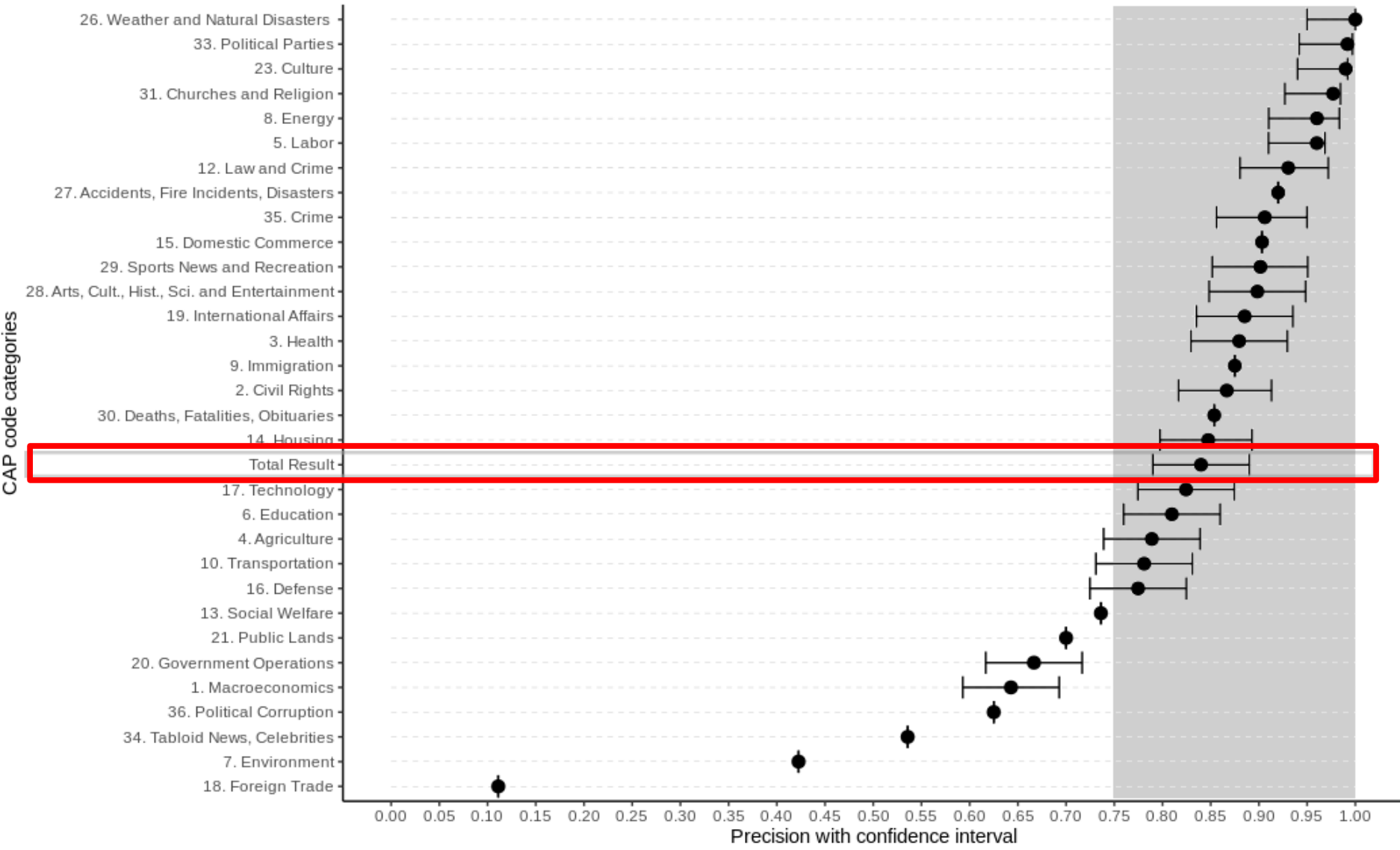
Major Topic	Coded Articles	Sample Size	Precision
1. Macroeconomics	3833	350	0.64
2. Civil Rights	207	135	0.87
3. Health	700	249	0.88
4. Agriculture	408	199	0.79
5. Labor	127	100	0.96
6. Education	436	205	0.81
7. Environment	71	71	0.42
8. Energy	542	226	0.96
9. Immigration	8	8	0.88
10. Transportation	459	210	0.78
12. Law and Crime	570	230	0.93
13. Social Welfare	72	72	0.74
14. Housing	168	118	0.85
15. Domestic Commerce	93	93	0.90
16. Defense	342	182	0.77
17. Technology	196	131	0.82
18. Foreign Trade	27	27	0.11
19. International Affairs	7617	366	0.89
20. Government Operations	1247	294	0.67
21. Public Lands	10	10	0.70
23. Culture	124	100	0.99
26. Weather and Natural Disasters	201	133	1.00
27. Accidents, Fire Incidents, Disasters	50	50	0.92
28. Arts, Cult., Hist., Sci. and Entertainment	677	246	0.90
29. Sports News and Recreation	385	193	0.90
30. Deaths, Fatalities, Obituaries	41	41	0.85
31. Churches and Religion	194	130	0.98
33. Political Parties	661	244	0.99
34. Tabloid News, Celebrities	28	28	0.54
35. Crime	339	181	0.91
36. Political Corruption	8	8	0.63
<b>Total Result</b>	<b>19841</b>	<b>4630</b>	<b>0.84</b>



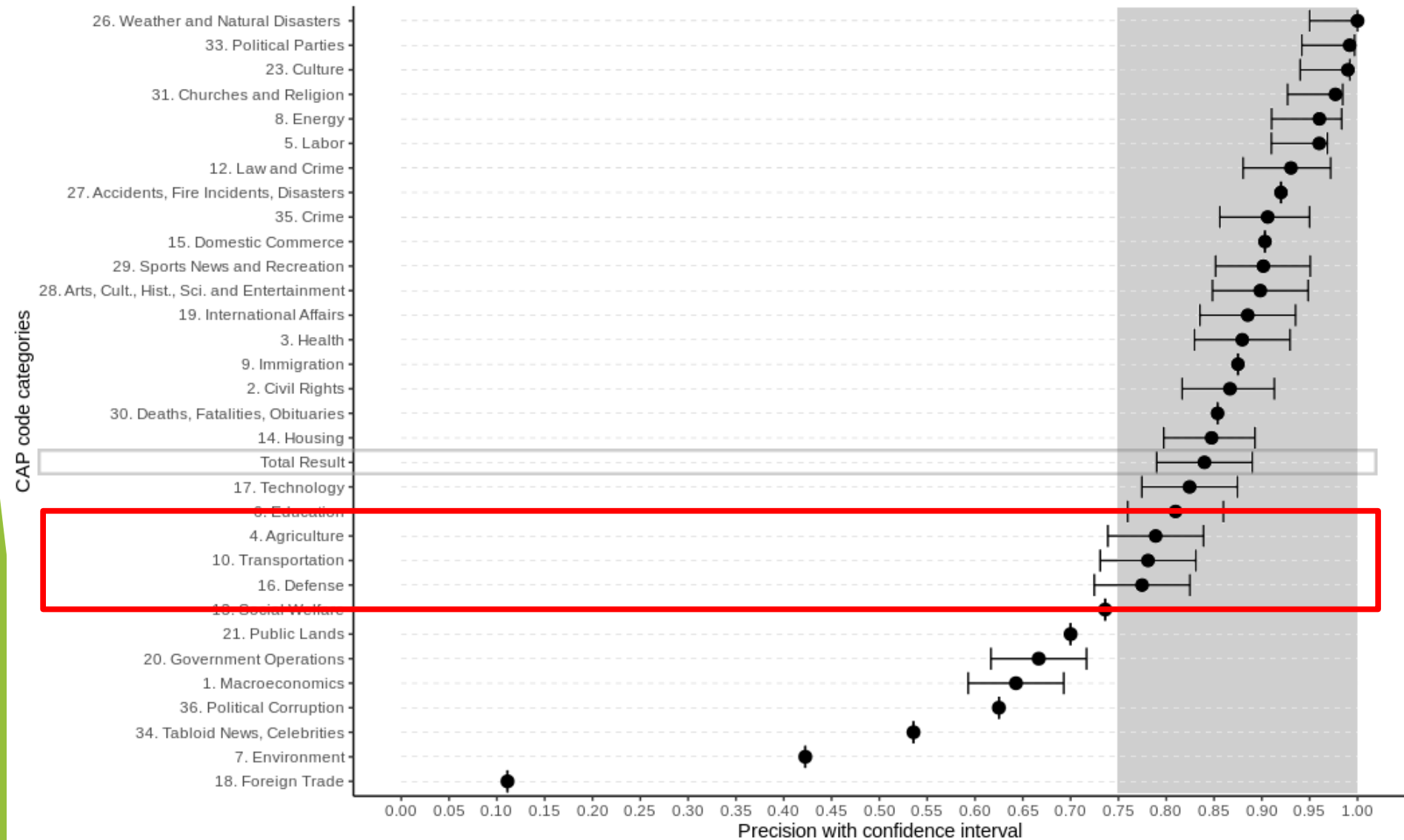
# Precision of MN corpus coding by CAP code category



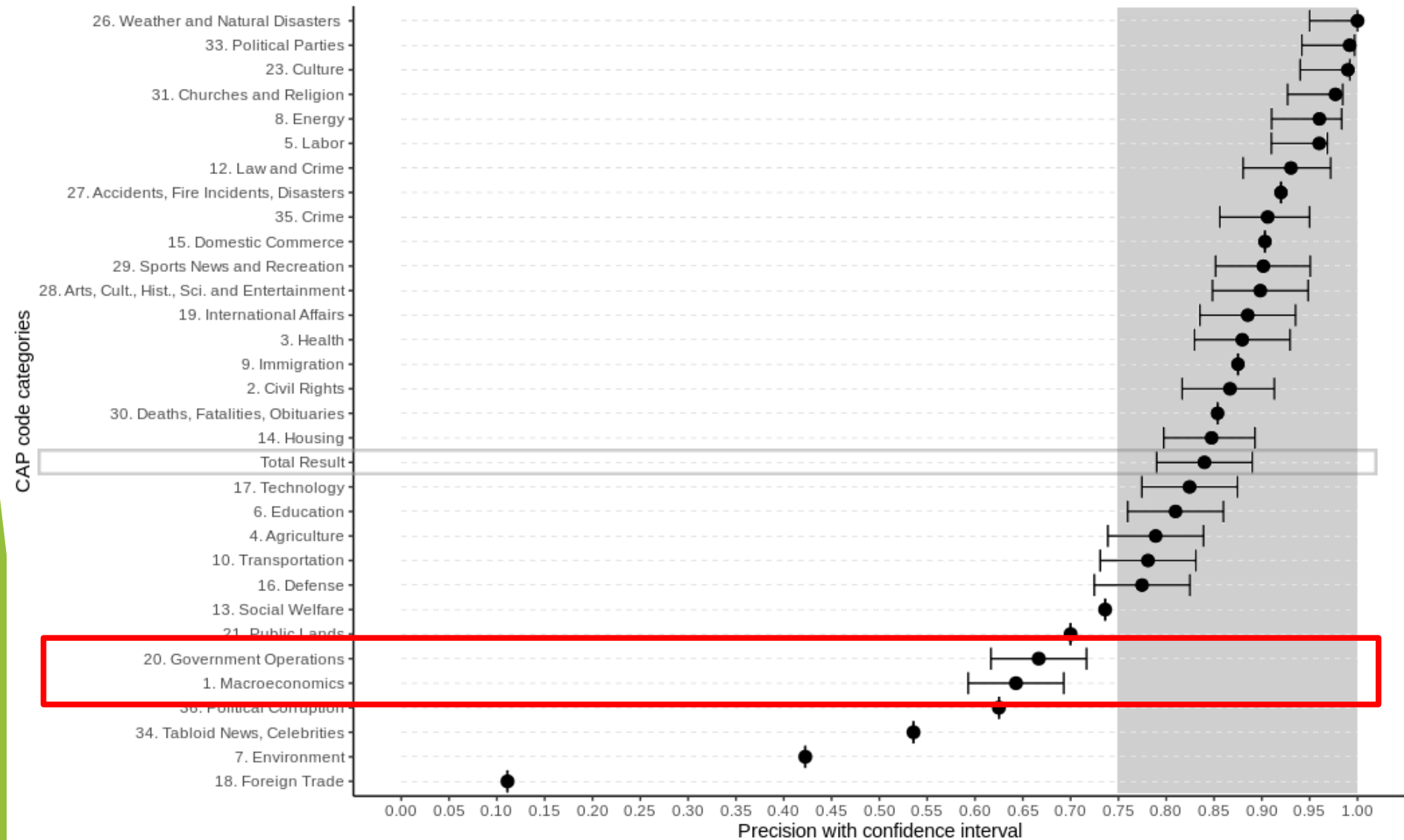
# Precision of MN corpus coding by CAP code category



# Precision of MN corpus coding by CAP code category

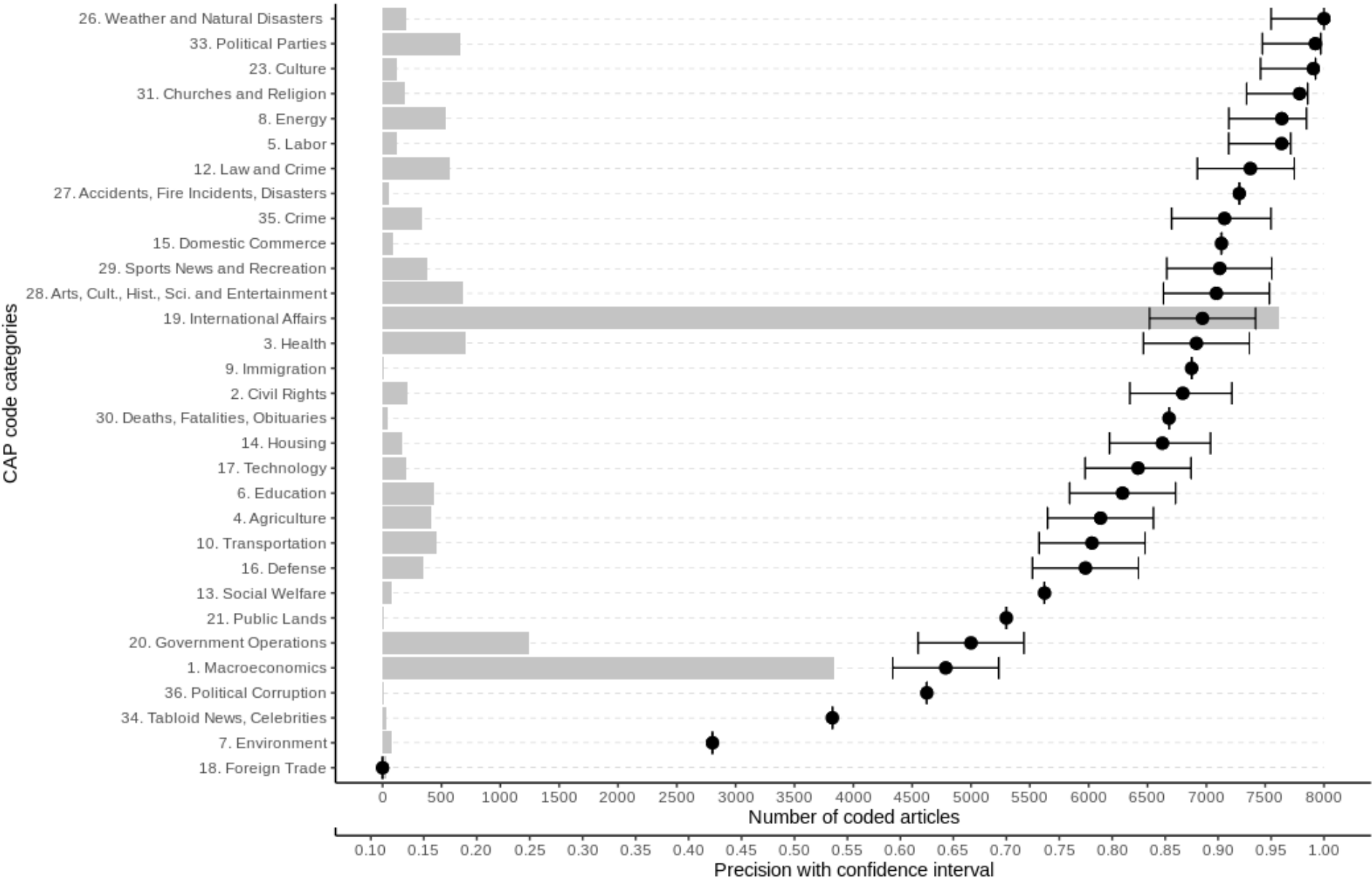


# Precision of MN corpus coding by CAP code category

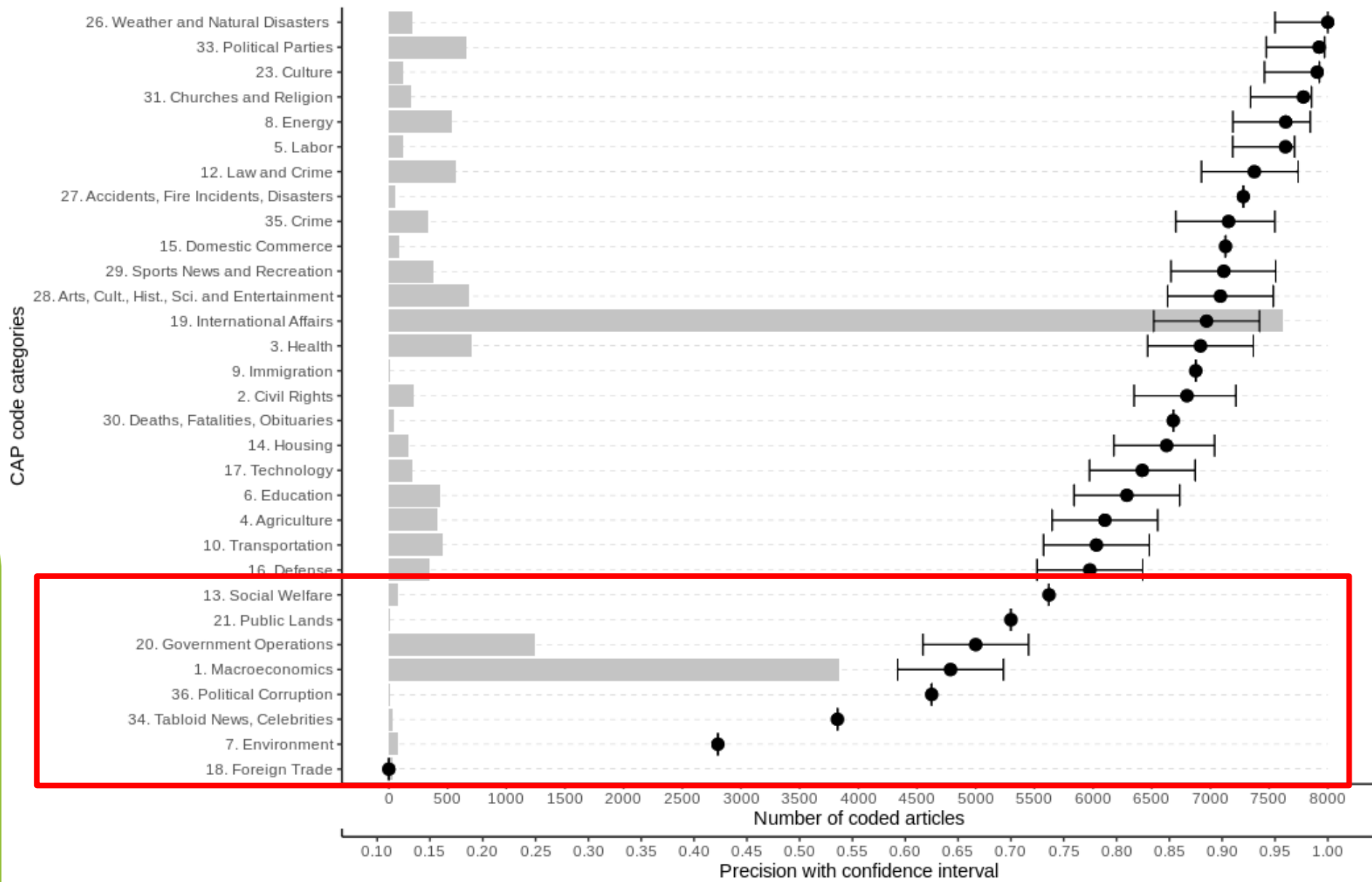




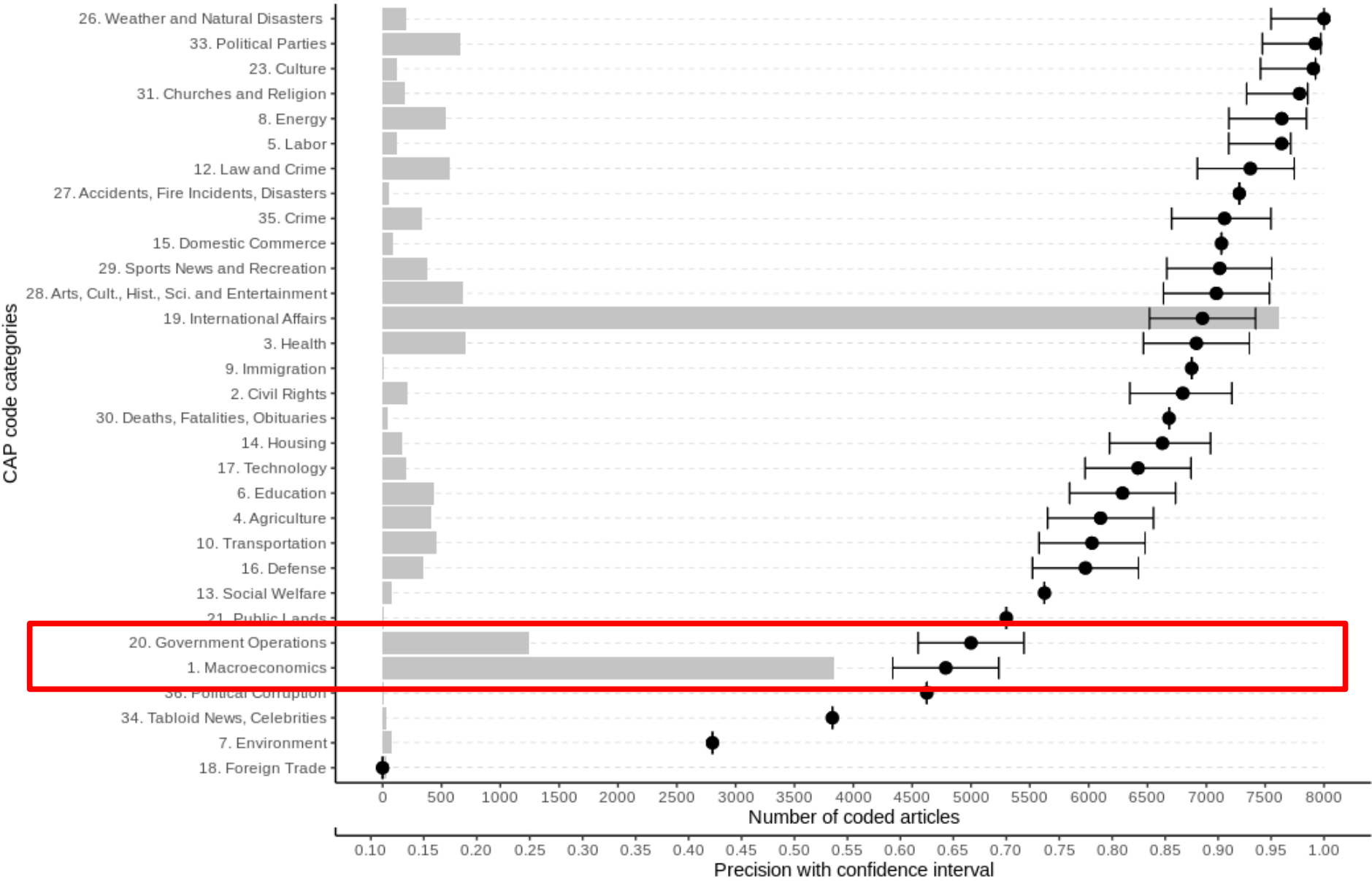
Precision and total number of coded articles of MN corpus by CAP code category



# Precision and total number of coded articles of MN corpus by CAP code category



Precision and total number of coded articles of MN corpus by CAP code category



Major Topic	Coded Articles	Sample Size	Precision
1. Macroeconomics	3833	350	0.64
2. Civil Rights	207	135	0.87
3. Health	700	249	0.88
4. Agriculture	408	199	0.79
5. Labor	127	100	0.96
6. Education	436	205	0.81
7. Environment	71	71	0.42
8. Energy	542	226	0.96
9. Immigration	8	8	0.88
10. Transportation	459	210	0.78
12. Law and Crime	570	230	0.93
13. Social Welfare	72	72	0.74
14. Housing	168	118	0.85
15. Domestic Commerce	93	93	0.90
16. Defense	342	182	0.77
17. Technology	196	131	0.82
18. Foreign Trade	27	27	0.11
19. International Affairs	7617	366	0.89
20. Government Operations	1247	294	0.67
21. Public Lands	10	10	0.70
23. Culture	124	100	0.99
26. Weather and Natural Disasters	201	133	1.00
27. Accidents, Fire Incidents, Disasters	50	50	0.92
28. Arts, Cult., Hist., Sci. and Entertainment	677	246	0.90
29. Sports News and Recreation	385	193	0.90
30. Deaths, Fatalities, Obituaries	41	41	0.85
31. Churches and Religion	194	130	0.98
33. Political Parties	661	244	0.99
34. Tabloid News, Celebrities	28	28	0.54
35. Crime	339	181	0.91
36. Political Corruption	8	8	0.63
Total Result	19841	4630	0.84



# Main contributions of HBS and the present study



- Enhance ML precision and recall by both **human input** (validation) and **workflow design** (one-vs-all classification, ensemble voting)
- Start working from a **limited training set**
- Able to **maximize ROI on human coding**
- **Move between (intra-domain) corpora**
- Take advantage of **cloud infrastructure** and **parallel processing** with Apache Spark

# Further work

- Implement a finishing step using **regular expressions to correct systematic errors**
  - “design” in *Environment*
  - “icerink” in *Public Lands*
- Testing the HBS approach on **further languages**
- Generalizing the method to **other domains** beyond media



ELKH Cloud

**Thank you for your attention!**

sebok.miklos@tk.mta.hu  
kacsuk.zoltan@tk.mta.hu

poltext.tk.mta.hu  
cap.tk.mta.hu

 tkpti