MUNI
FI

IT pro prevenci a detekci neetického jednání

Mgr. Tomáš Foltýnek, Ph.D.

foltynek@fi.muni.cz

Osnova dnešní přednášky

- Opakování: Autorská práva a Creative commons
- Prezentace: Lucián Prodan – Open source licence

- IT pro prevenci a detekci podvodného jednání
  - Certifikace s využitím blockchainu
  - Detekce plagiátorství
  - Proctoringové systémy
  - Detekce textu vygenerovaného umělou inteligencí

- Dilemma game: Chyby v datech

Opakování: Commons

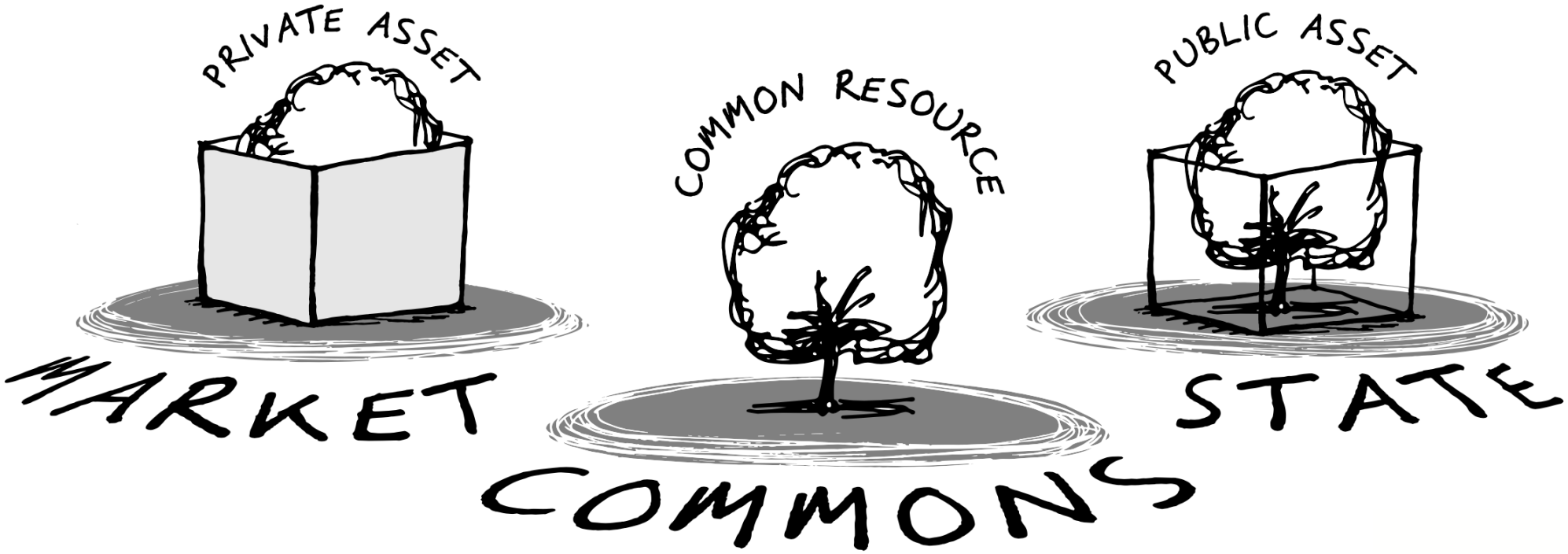


Fig. 3. How the market, commons, and state conceive of resources.

Obrázek převzat z knihy Made with Creative Commons (Paul Stacey and Sarah Hinchliff)

Opakování: Dílo

- Co je **dílo**, tj. předmět autorskoprávní ochrany?
  - „…dílo literární a jiné dílo umělecké a dílo vědecké, které je jedinečným výsledkem tvůrčí činnosti autora a je vyjádřeno v jakékoli objektivně vnímatelné podobě včetně podoby elektronické, trvale nebo dočasně, bez ohledu na jeho rozsah, účel nebo význam…
- „idea-expression dichotomy“
  - „Dílem podle tohoto zákona není zejména námět díla sám o sobě, denní zpráva nebo jiný údaj sám o sobě, myšlenka, postup, princip, metoda, objev, vědecká teorie, matematický a obdobný vzorec, statistický graf a podobný předmět sám o sobě.“

Obrázek: Midjourney painting Midjourney

Opakování: Oprávněné užití díla

Do práva autorského nezasahuje ten, kdo

a) užije v odůvodněné míře výňatky ze zveřejněných děl jiných autorů ve svém díle,

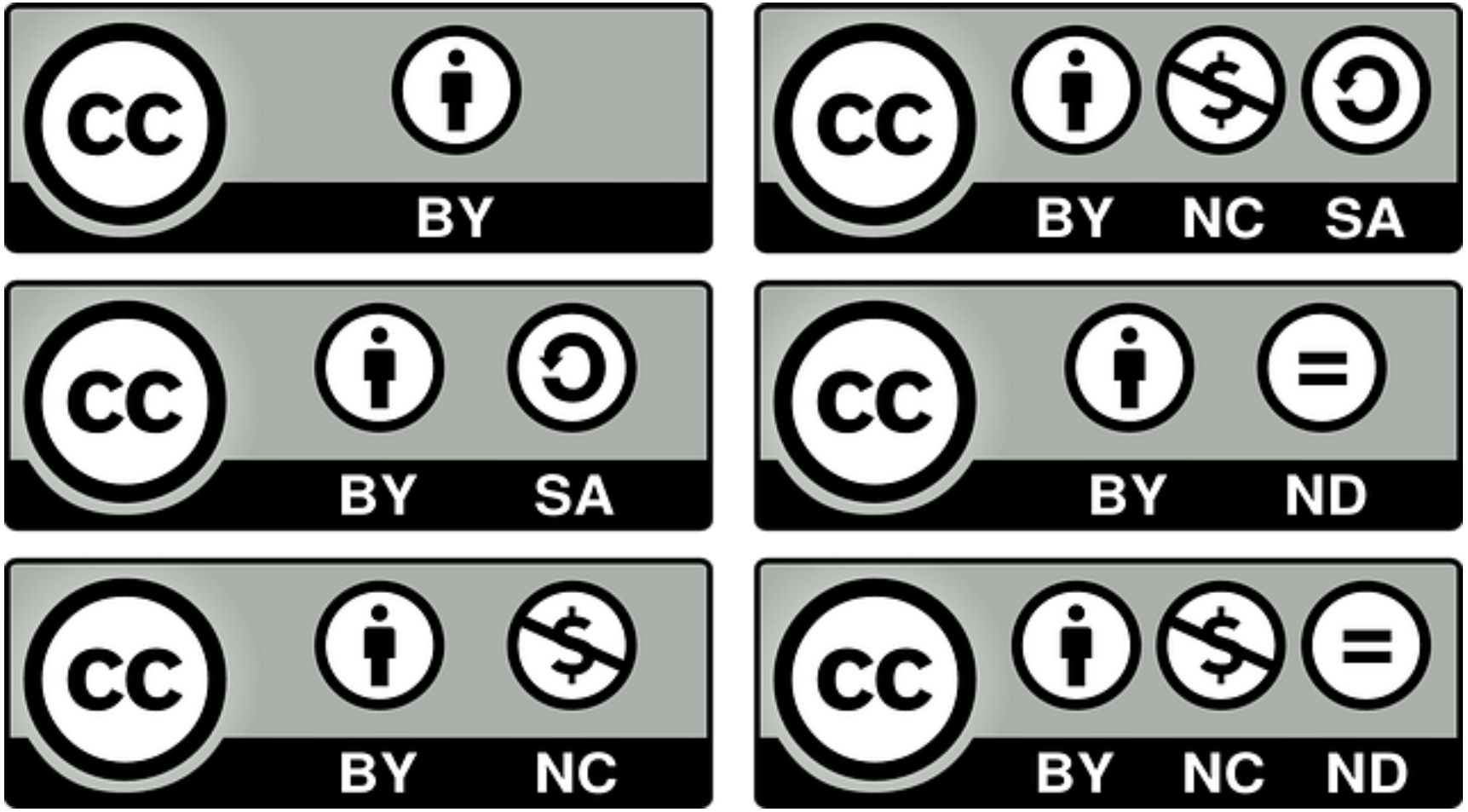
b) užije výňatky z díla nebo drobná celá díla pro účely kritiky nebo recenze vztahující se k takovému dílu, vědecké či odborné tvorby, a užití bude v souladu s poctivými zvyklostmi a v rozsahu vyžadovaném konkrétním účelem,

c) užije dílo při vyučování pro ilustrační účel nebo při vědeckém výzkumu (…) a nepřesáhne rozsah odpovídající sledovanému účelu

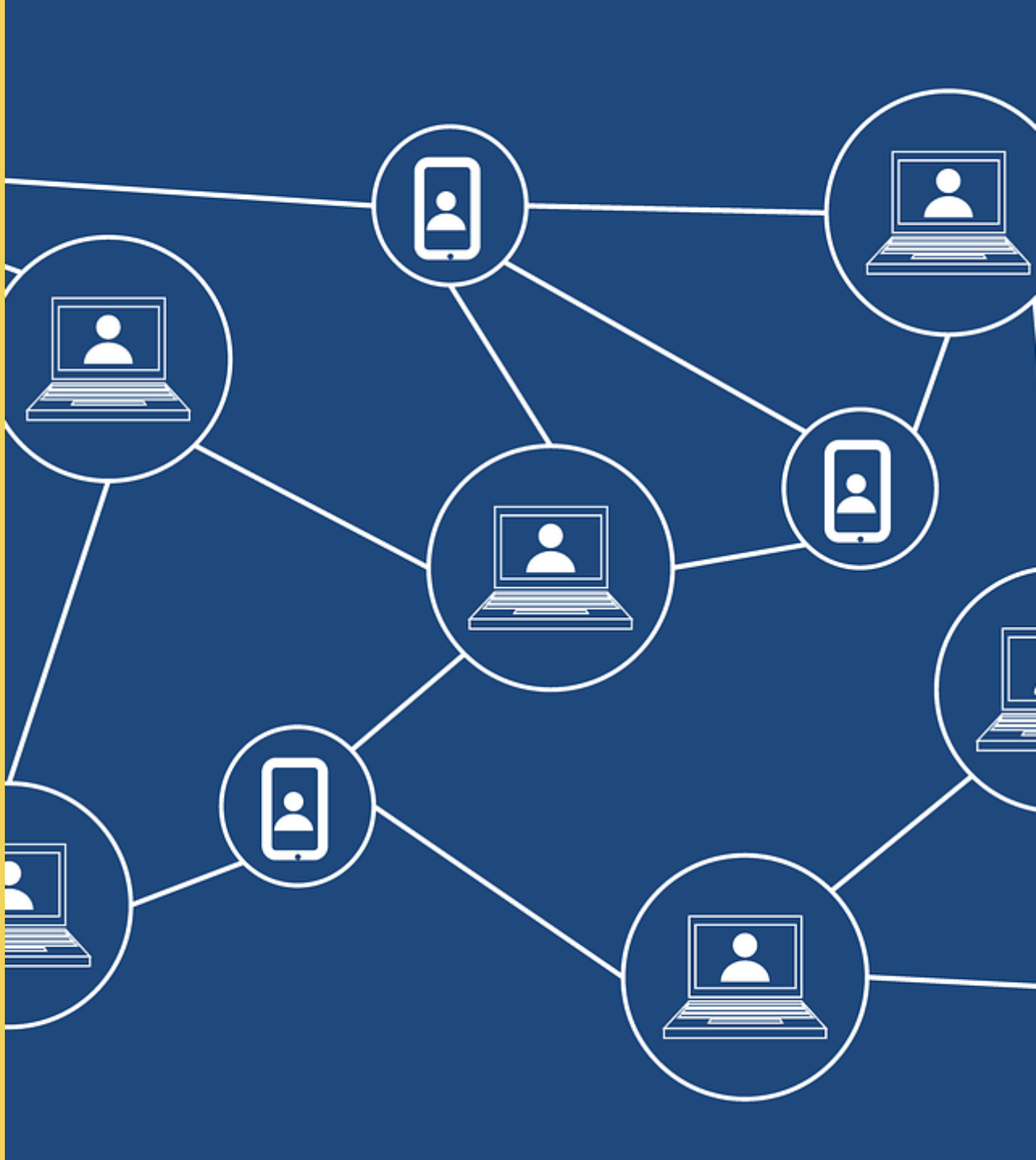
Vždy se musí uvést (je-li to možné)**:** jméno autora, název díla a pramen

§31 zákona č. 121/2000 Sb. (Autorský zákon)

Opakování: Druhy CC licencí



Využití blockchainu pro ověřování certifikátů



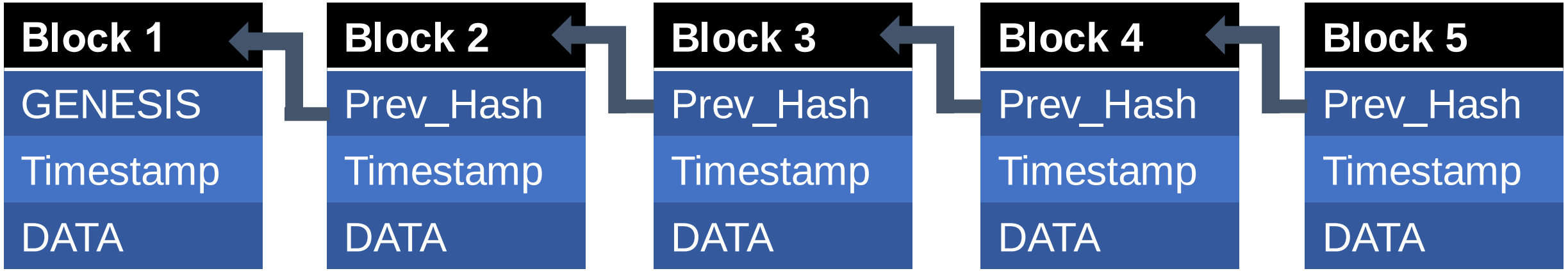
Falešné diplomy

- Falešné diplomy existujících univerzit
- „Pravé“ diplomy neexistujících univerzit
- Falešné „Transcript of records“ výměnných studentů
- Přibližně 30 % lidí uvádí nepravdivé údaje v CV

- Řešením je důvěryhodné potvrzení, že určitý dokument
  - Existoval (byl vytvořen) v daném čase
  - Byl vydán určitou institucí (osobou)
  - Tato instituce (osoba) měla právo dokument vydat
- Je třeba nemanipulovatelná databáze s časovými razítky

Základní princip blockchainu

- Každý blok obsahuje
  - Data
  - Časové razítko
  - Hash předchozího bloku
- Změna dat v jednom bloku → Změna hashe → změna všech násl. bloků
- Distribuovanost blockchainu + vhodný konsenzuální algoritmus zajistí neměnnost dat



Využití blockchainu

- Kdy má smysl použít blockchain?
  - Neměnná data
  - Hashovaná nebo šifrovaná data
  - Více účastníků
  - Vzájemná nedůvěra
- Kryptoměny
- Chytré kontrakty
- Logování informací
  - Logistika, pojišťovnictví, časová razítka,…
- Blockchain poskytuje důvěryhodnou platformu

Ověřování certifikátů pomocí blockchainu

- Hash údajů z certifikátu je uložený na blockchain
  - Spolu s časovým razítkem a digitálním podpisem vydavatele
- Kdokoliv může ověřit pravost
- Z hashe nelze zjistit osobní údaje

BLOCKCERTS
The Open Standard For Blockchain Credentials

BCdiploma
blockchain digital credentials

Certifaction

bloxberg

Detekce plagiátorství

Even popular ex-president of the United States has been accused of plagiarism. Before the elections in 2008, he used a speech which resembled notably an older speech of another politician. The other politician was Obamas's friend and both claimed afterwards that they prepared the speech together. Obama later apologized for not sufficiently admitting the co-authoritarianism of the other politician.

Let's stay in the politics for a while. In most of the following cases we are going to talk about plagiarism in dissertation theses. With dissertation theses students finish their doctoral studies, it is the highest level of university studies, where they gain degree PhD which is added behind their name.

We start in Germany. In 2009, Defense Minister Guttenberg, who is a star of German politics, appeared, has other sources of information, that he is the most popular German politician, and he is expected to have an amazing political career. But in 2011 he tries to describe your dissertation. It contains no minor plagiarism offenses, except for blows and conclusions that are not part of Guttenberg wrote almost nothing alone - he copied the absolutely low results of your work from various foreign texts. In fact, he wrote only the introduction and conclusion.

When this fraud occurred and appeared in the media, Mr. Guttenberg himself renounced his title. Ten taken from him at his university when she investigated the case. The title of Ph.D. it is not suitable for being against a government in which there has been a wave of resistance against former favorite ministers, who turned out to be a fraudster who had not resigned from Guttenberg as minister and then left the others aside. Thus, the dissertation described completely ended his promising career.

Karl zu Gutteberg is definitely not the only one of the German politicians. Like him, former Minister of Education Annette Schavan ended up. This September, the current Secretary of Defense Ursula von der Leyen was accused of plagiarism.

Plagiarism is also the dissertation thesis of the former President of Hungary, Pala Schmitt. In 2012, it turned out that almost the entire text is actually a translation of several previously published works in other languages. After the investigation, the President gave up his title and resigned as President. Interestingly, according to polls, the majority of the Hungarian public did not consider resignation necessary.

Vladimir Putin has also been accused of plagiarizing his dissertation - the core part of his work (and six pictures) is identical to an earlier work written by Pittsburgh University. Plagiarism has never been proven, and both the Russian President and his university rule it out.

Plagiarism was even the cause of the war. Before the outbreak of the Iraq war in 2003, the US used plagiarism as one of the evidence that Saddam Hussein is lying about the state of arms in Iraq. When the US wanted a declaration on the Iraqi arms program, the dictator took the UN report on Iraqi weapons, erased any criticism of its country from it, and simply copied the rest and sent it to the US as part of the required declaration. Why write something that others have already found out about us.

Definice plagiátorství

využití (myšlenek, obsahu, nebo struktury) jiného díla
bez řádného uvedení odkazu na zdroj
k získání určité výhody tam, kde se očekává původní dílo

the use of ideas, content, or structures
without appropriately acknowledging the source
to benefit in a setting where originality is expected

Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic Plagiarism Detection: A Systematic Literature Review. *ACM Comput. Surv.*, 52(6), 112:1--112:42. https://doi.org/10.1145/3345317

Tři úrovně detekce plagiátorství

Předpisy

IPPHEAE

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Nástroje

turnitin®

URKUND

PlagScan

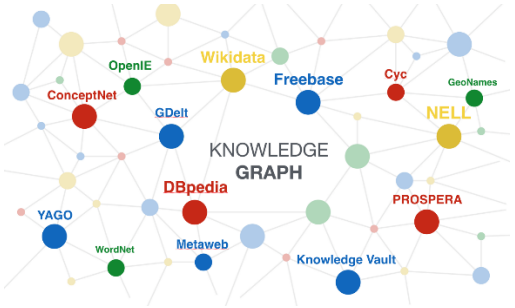
PLAGIARISM CHECK.ORG

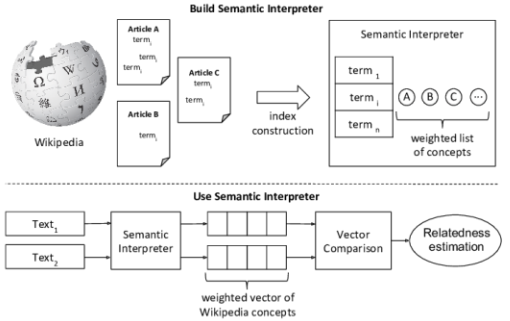
StrikePlagiarism.com

Metody

$$A = U \begin{pmatrix} w_1 & & \\ & w_2 & \\ & & w_3 \end{pmatrix} V^T$$

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} = \frac{\sum_{i=1}^{N} u_i v_i}{\sqrt{\sum_{i=1}^{N} u_i^2} \sqrt{\sum_{i=1}^{N} v_i^2}}$$





Detekce plagiátorství: Formulace problému

- Extrinsic plagiarism detection
  - Nalézt potenciální zdroje plagiátorství
  - Hledáme podobnosti mezi různými dokumenty

- Instrinsic plagiarism detection
  - Nalézt místa, kde se mění autorský styl
  - Hledáme různorodé části v rámci jednoho dokumentu

Detekce plagiátorství: Typologie metod

- Lexikální vrstva
  - Znakové nebo slovní n-gramy
  - Vektorové prostory
- Syntaktická vrstva
  - Slovní druhy, skladba věty
  - Syntaktické grafy
- Sémantická vrstva
  - Latentní sémantická analýza, explicitní sémantická analýza
  - Knowledge graphs
- Kombinace metod
  - Strojové učení

Přesnost metod

- Copy-paste ≈ 100 %
- Nahrazení synonym ≈ 90 %
- Přeskládání slov ≈ 90 %

Implementováno, využíváno

- Identifikace parafrází ≈ 80 %
- Sumarizace ≈ 75%
- Překladové plagiátorství ≈ 70 %

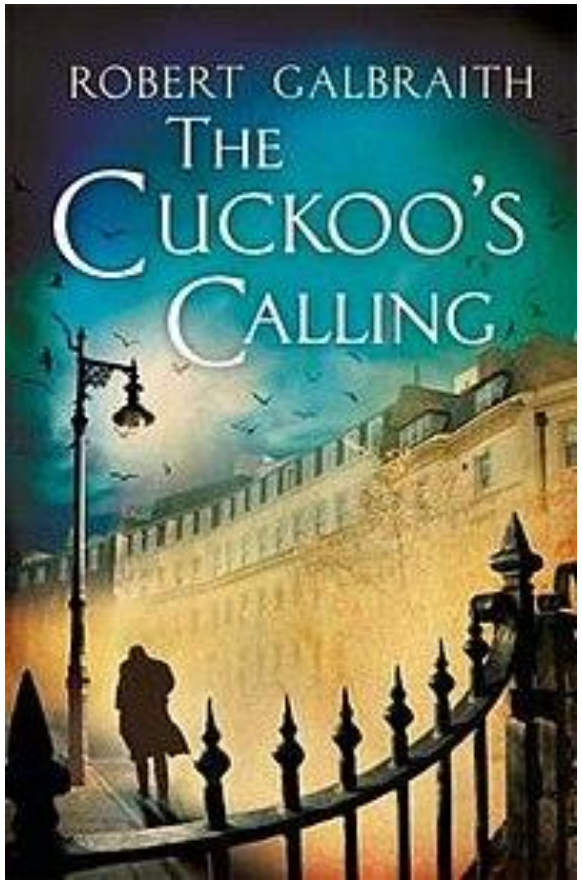
Probíhá vývoj

- Strukturní či myšlenkové plagiátorství ???

Obtížný problém

Rozpoznávání autorství

- Přesnost současných metod
  - Rozpoznání změny stylu ≈ 60 %
  - Seskupování podle autorství ≈ 60 %
  - Predikce mateřského jazyka ≈ 65 – 85 %
  - Rozpoznání pohlaví autora ≈ 80 %
  - Odhad věku autora ≈ 50 – 55 %
- Robert Galbraith: *The Cuckoo's Calling* (2013)
- Skutečná autorka: J.K. Rowling
  - Viz https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/



Detekce plagiátorství z Wikipedie

	Akademia	Copyscape	Docol©c	Dupli Checker	intihal.net	PlagAware	PlagiarismCheck.org	Plagiarism Software	PlagScan	DPV	StrikePlagiarism.com	Turnitin	Unicheck	Urkund	Viper
Copy-paste	3,6	4,4	4,6	2,0	1,6	1,5	4,4	4,6	3,6	4,3	4,9	4,9	5,0	5,0	4,5
Synonyms	3,0	3,9	2,9	1,0	0,8	1,1	3,6	4,4	1,8	3,8	3,9	4,1	3,3	4,6	2,6
Paraphrase	2,1	1,9	1,4	0,1	0,6	0,8	1,9	2,8	0,9	2,2	2,0	2,1	1,5	2,9	1,2

Maximum score: 5

Foltýnek et al. (2020): *Testing of Support Tools for Plagiarism Detection.* International Journal of Educational Technology in Higher Education, 17(46). DOI 10.1186/s41239-020-00192-4

TeSToP

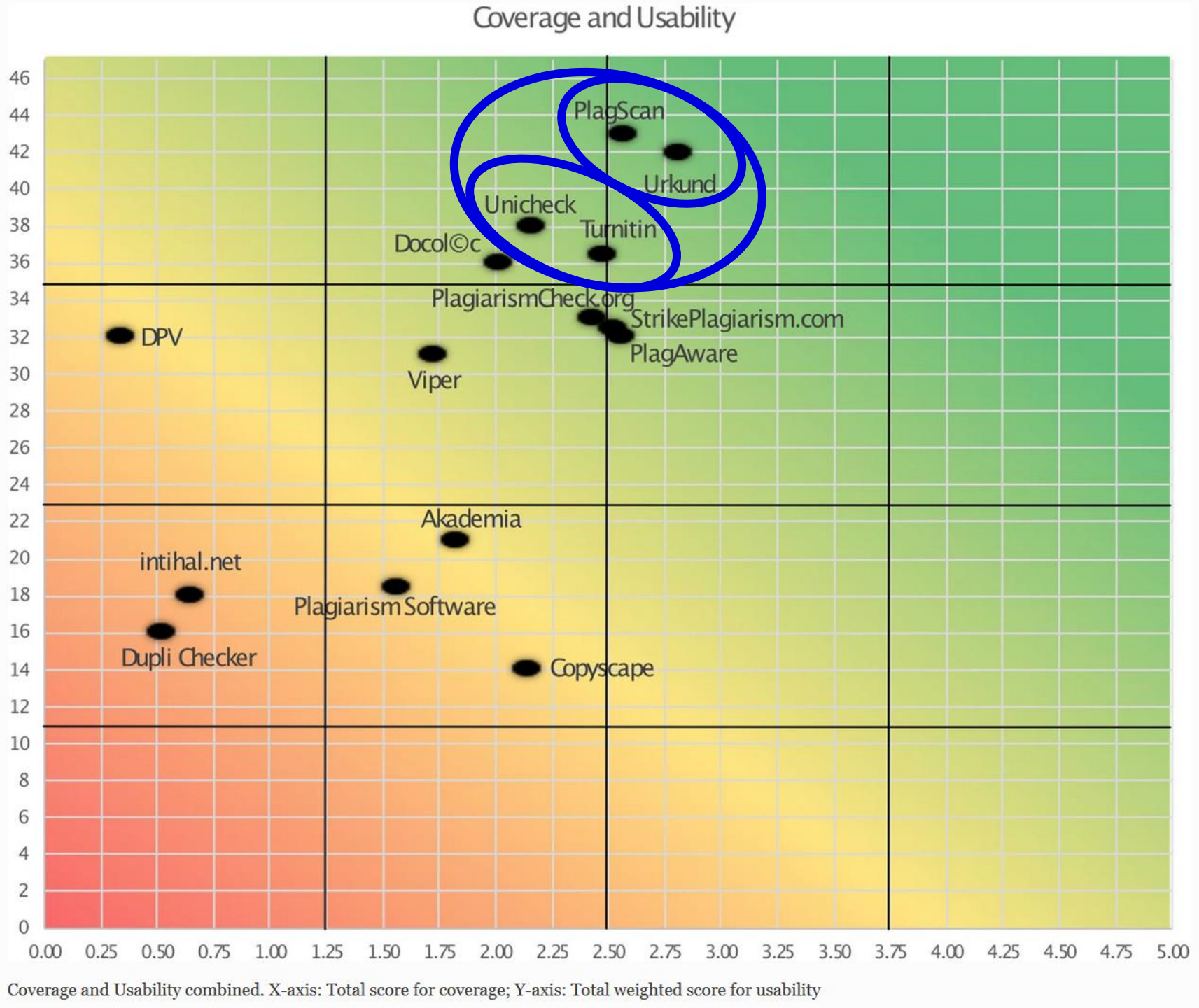
Testing of support tools for plagiarism detection

Osa X: Úspěšnost

Osa Y: Použitelnost

Žádný systém není dokonalý

Žádný systém nenajde vše

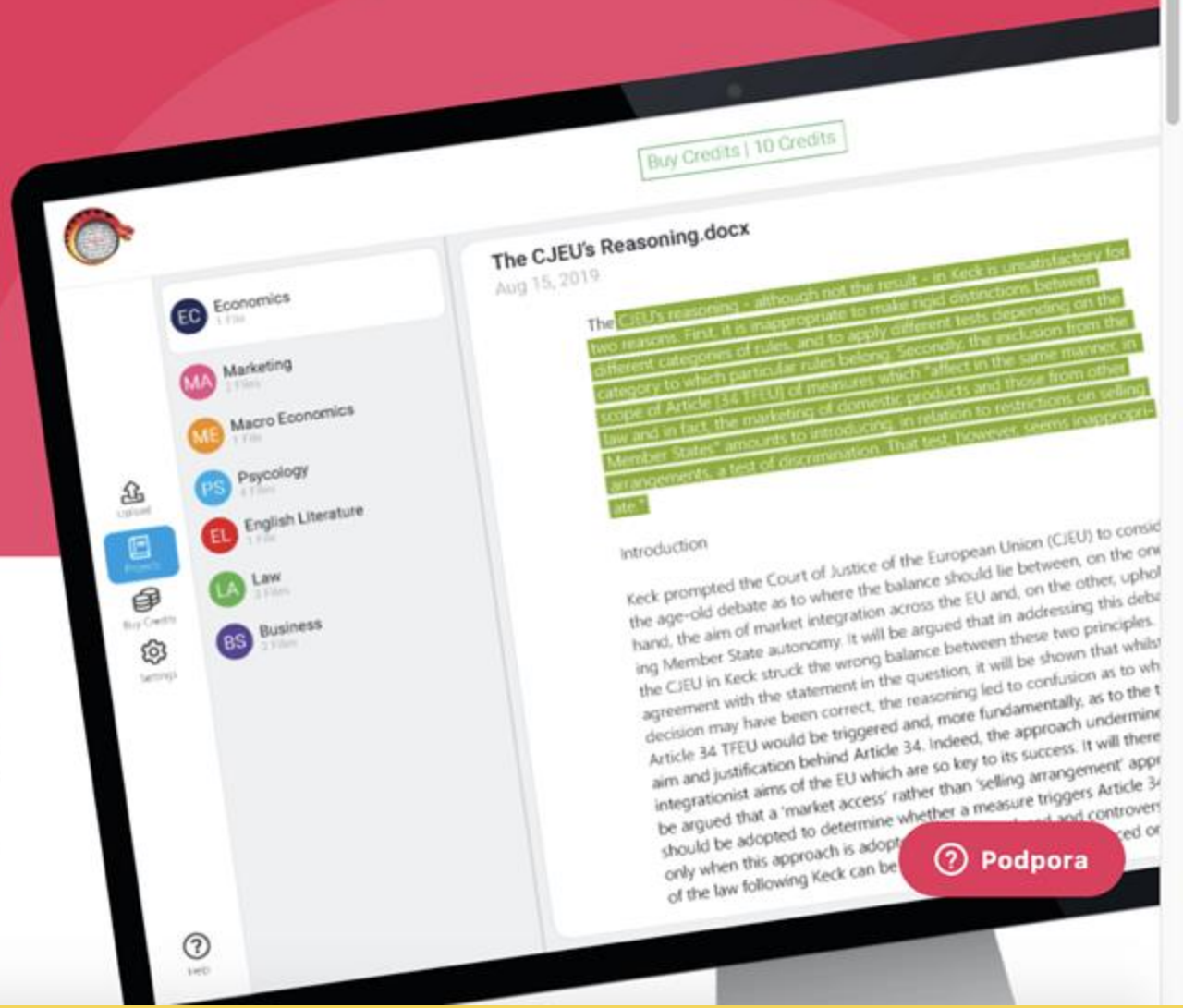


Coverage and Usability combined. X-axis: Total score for coverage; Y-axis: Total weighted score for usability

Viper Plagiarism Checker

Welcome to Viper - a leading plagiarism checker which, using its range of powerful features, will help you check for plagiarism and duplicate content in your work. From individual students to lecturers and institutions, Viper is the plagiarism checker of choice for thousands of people every month.

Register Sign in



Powerful features

With its easy-to-use interface and highly detailed scanning process, it only takes three simple steps for Viper to review your document for plagiarism and produce a full report.

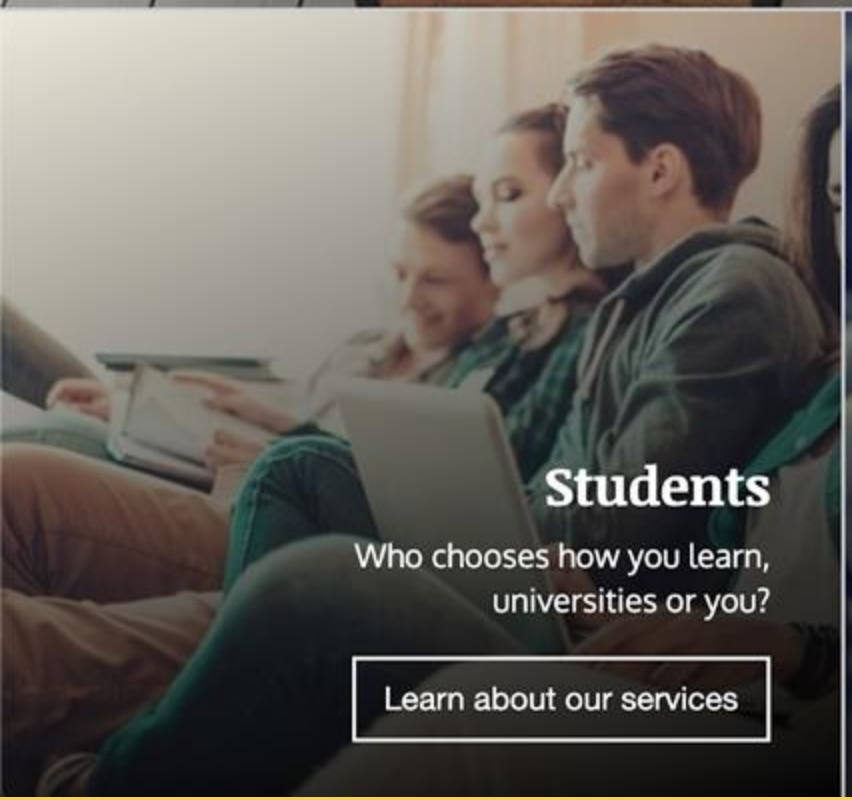
Check ✓

Simply select your document using the Viper online app and submit it to

All Answers Limited [GB] | allanswers.co.uk

Aplikace

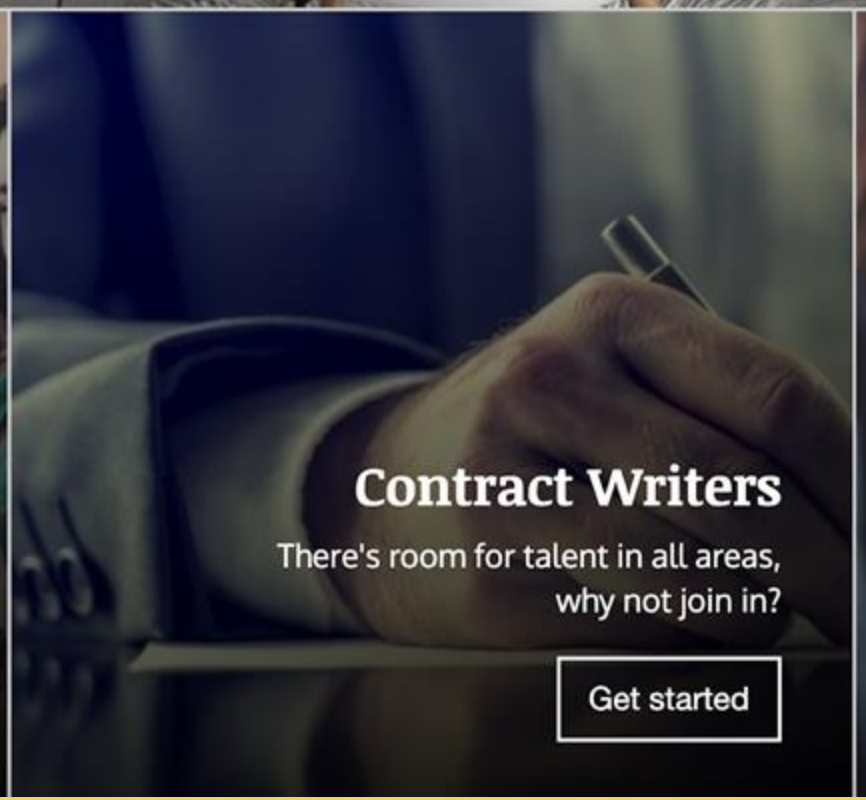
We help students achieve more



Students

Who chooses how you learn, universities or you?

Learn about our services



Contract Writers

There's room for talent in all areas, why not join in?

Get started

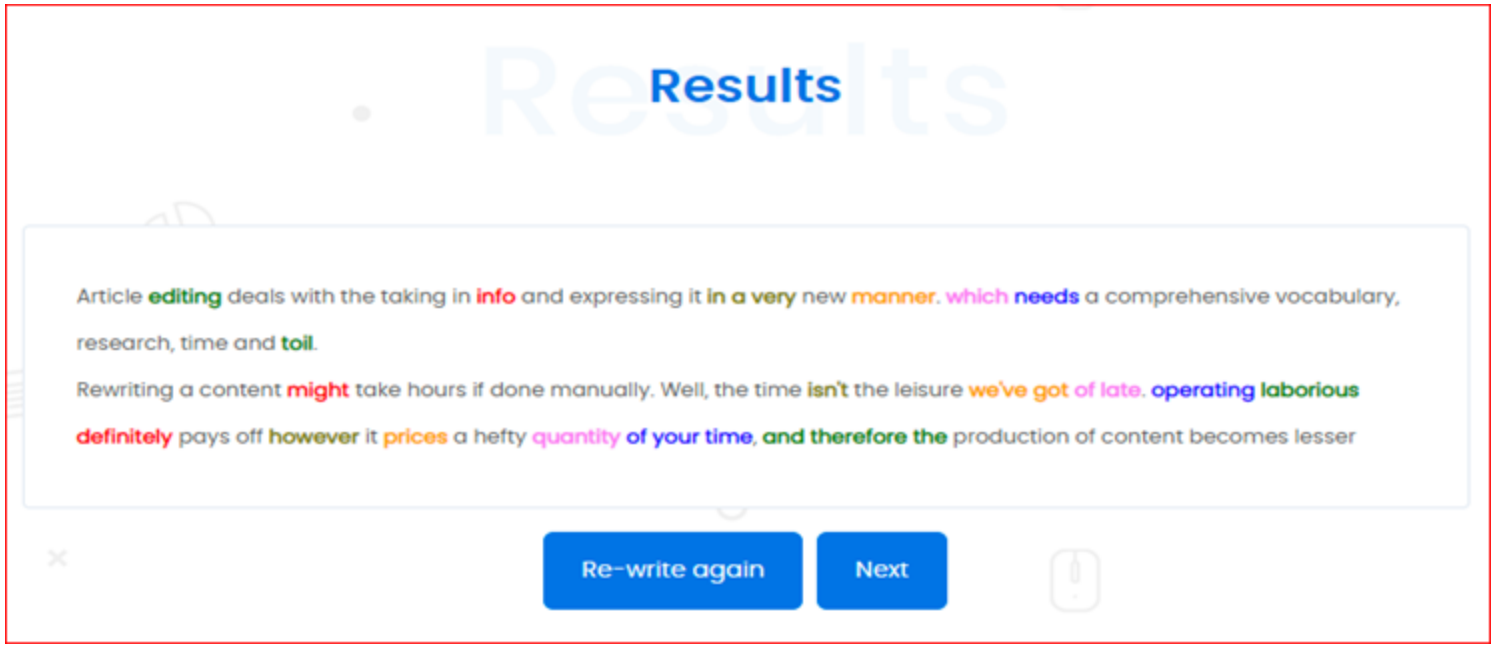
Internal Careers

Would you work for an exciting, dynamic company?

Apply today

Pozor na podezřelé nástroje zdarma!

- How does Viper use my essay/dissertation?
  - “When you scan your work for plagiarism using **Viper Premium** it will never be published on any of our study sites.”
- Některé nástroje jsou napojené na parafrázovací služby



Zdroj obrázku: https://www.duplichecker.com/article-rewriter.php

Proctoringové systémy

Big brother is watching you. Free photobank torrange.biz
https://torange.biz/fx/big-brother-watching-you-video-172966

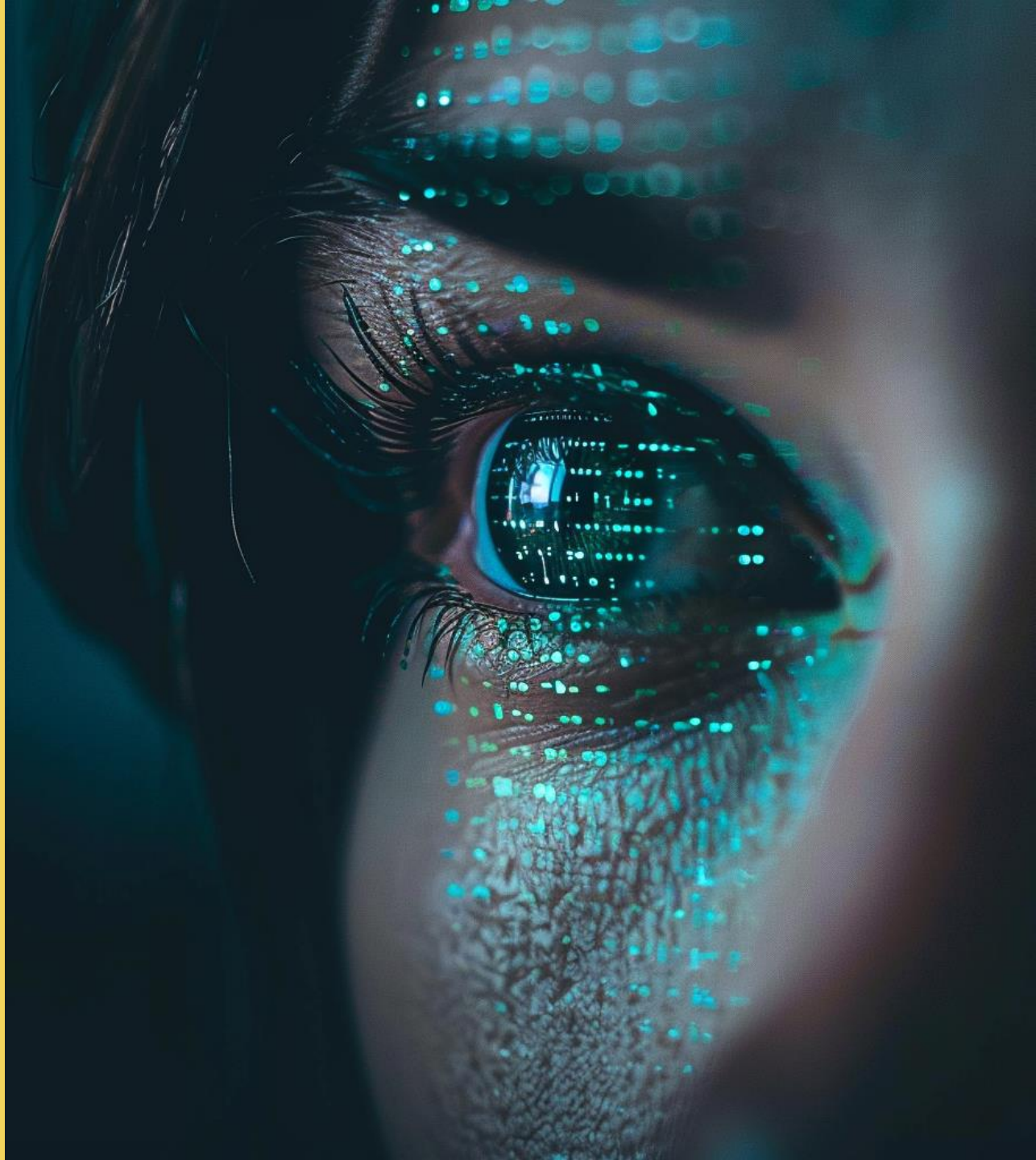
Proctoringové systémy

- Problém: Podvádění během (online) zkoušek
- Proctoringové systémy umožňují
  - Ověření totožnosti rozpoznáváním obličejů a průkazů totožnosti
  - Detekce a upozornění v případě, že
    - Zkoušený není přítomen
    - Je přítomna jiná osoba
    - Objeví se nepovolený předmět (mobilní telefon, kniha)
    - Jsou slyšet hlasy
  - Pravidelné snímání obrazovky
- Jak vnímáte využití takovýchto systémů?
  - Na univerzitě
  - V jiném prostředí (jazykové a další certifikace, najímání nových pracovníků,…)

Stanovisko MU k online proctoringu

- Jednoznačně nedoporučující
- Etické a koncepční důvody
  - Neslučitelnost s hodnotami MU – důvěra, respekt, důstojnost
  - Možnost obejít systém
  - Nepřiměřený zásah do soukromí
  - Problematické využívání umělé inteligence
- Právní důvody
  - Zásah do práv studentů → Nutnost souhlasu studentů
    - Jiná forma ukončení pro studenty, kteří odmítnou
  - Ukládání videozáznamů, GDPR
- Technické důvody
  - Nároky na technické vybavení (2 kamery)
  - Nároky na internetové připojení
  - Potřeba IT asistence pro studenty

Detekce textu vygenerovaného umělou inteligencí



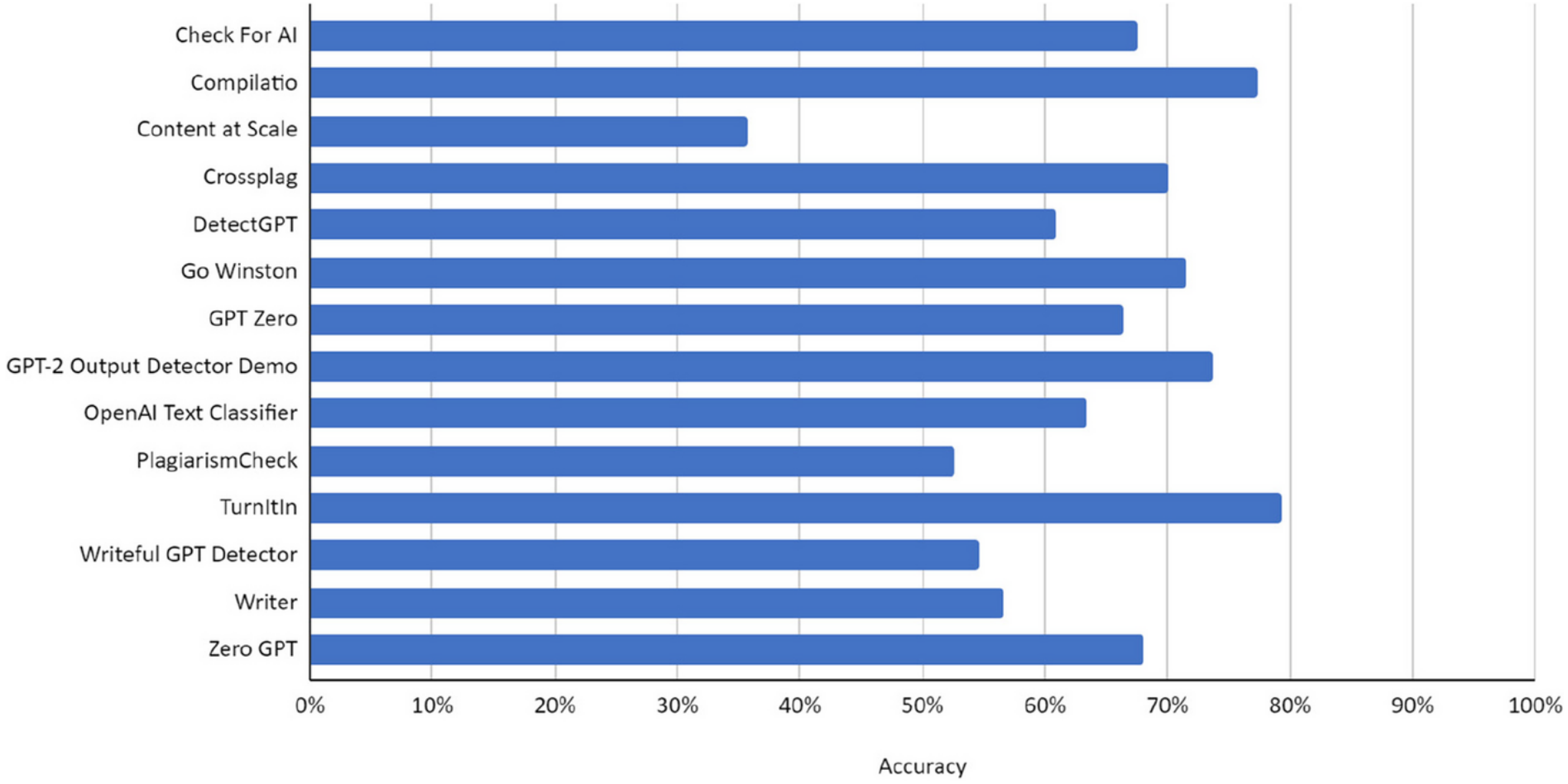
Testování detektorů AI-textů

- Jaro 2023
- 12 volně dostupných a 2 komerční nástroje
- 54 dokumentů v 6 kategoriích

  01-Hum: human-written
  02-MT: human-written + machine translation to English
  03-AI: AI-generated text
  04-AI: AI-generated text
  05-ManEd: AI-generated text + manual edits
  06-Para: AI-generated text + machine paraphrase

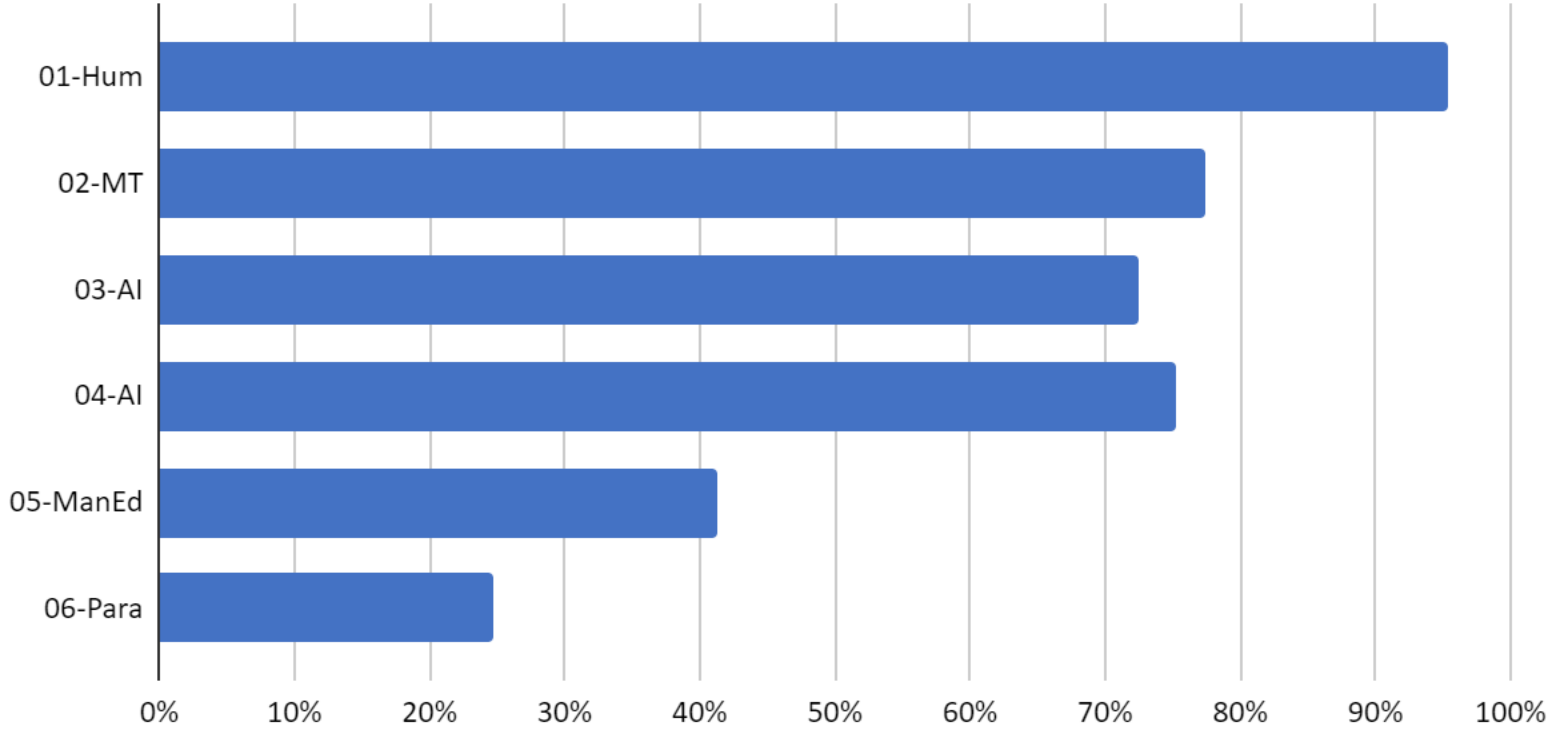
Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

Výsledky testu



Fungují nástroje na detekci textu vytvořeného umělou inteligenci?

- Nefungují ☺
- Zkreslení směrem k “napsán člověkem”
- I tak produkují falešně pozitiv výsledky
- Neposkytují důkaz
  - Nelze prokázat discipliární přestupek
  - Nemožnost obrany
- Text vytvořený AI a parafrázovaný AI je většinou klasifikován jako napsaný člověkem



Průměrná přesnost klasifikace (accuracy = správně klasifikované / všechny)
01-Hum a 02-MT: správně = napsané člověkem
03-AI, 04-AI, 05-ManEd, 06-Para: správně = vygenerované AI

Další experimenty

- Text vygenerovaný GPT-4 lze detekovat hůře než text vygenerovaný GPT-3.5
  - Obecně: čím lepší model, tím obtížnější detekce
- Text vygenerovaný s extra prompty lze detekovat hůře
  - Přidat občasné gramatické chyby do textu
  - Míchat krátké a dlouhé věty
  - Přepiš text tak, aby byl lingvisticky komplikovanější
- Detektor používá jazykový model ➔ Hůře detekuje texty vygenerované jiným modelem
  - Čím více různých modelů, tím obtížnější detekce

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). *GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education*. arXiv preprint

AI-texty nelze spolehlivě detekovat

- Detektory pracují s odlišnostmi v textu (perplexity, burstiness,…)
  - Pomocí promptu lze generátoru vnutit „více lidský“ styl psaní
  - Stovky videí na YouTube o tom, jak přelstít detektory
- U běžných znalostí není možné určit zdroj
- Jak hodnotit výsledek spolupráce AI + člověk?
- Generátor ve spojení s detektorem tvoří „neporazitelnou“ dvojici
  - Generátor může obměňovat text tak dlouho, dokud neprojde detektorem

Dilemma Game

Obrázky generovány nástrojem Midjourney /imagine people arguing ethical dilemmas

Dilema: Chyby v datech

V rámci svojí závěrečné práce zpracováváte data z datasetu, který je na fakultě běžně využíván.

Zjistíte, že data obsahují značné množství chyb (chybějící údaje, zjevně chybné hodnoty,…), kterými se dosud nejspíš nikdo nezabýval. Opravit chyby by vám zabralo půl roku a alternativní dataset neexistuje.

Vedoucí práce navrhuje držet se „běžné praxe“, tedy o chybách mlčet.

Co je správné v této situaci udělat?

A. Najdete si čas na důkladné prozkoumání problému, i kdybyste měl(a) odložit odevzdání práce.

B. Zajdete za vedoucím katedry či děkanem a požádáte o prověření všech výzkumných projektů, které dataset využívaly.

C. Změníte téma práce, abyste s těmito daty nemusel(a) pracovat.

D. Spojíte se s těmi, kteří používali data před vámi. Pokud budou chtít, abyste o chybách pomlčel(a), tak to uděláte.

Příští týden (2. května)

- Téma: Etika umělé inteligence

- Úkoly: Přečíst si
  - Ethics in AI: Introduction to the special issue
  - Scientists Built an AI to Give Ethical Advice, But It Turned Out Super Racist