TABLE OF CONTENTS

WORKSHOP PROGRAMME

WORKSHOP ON

Use of big data and AI in fighting corruption and misuse of public funds - good practice, ways forward and how to integrate new technology into contemporary control framework

organized by the Policy Department on Budgetary Affairs
for
the Committee on Budgetary Control

Tuesday, 23 February 2021
13:45-15:45
European Parliament, Brussels
Virtual meeting

DRAFT WORKSHOP PROGRAMME

Opening remarks and Introduction

13:45 -13:50	**Ms Monika HOHLMEIER, MEP** Chair Committee on Budgetary Control
13:50 - 13.55	**Mr Mikulas PEKSA, MEP** Rapporteur, MEP Greens

* * *

Presentations

13:55 -14:05 **Vítezslav Titl**,
Assistant professor of law & economics at Utrecht University, affiliated researcher at KU Leuven, specialised in big data in fraud prevention, especially concerning the monitoring of public procurement.

14:05 -14:15 **Mihály Fazekas**,
Assistant professor at the Central European University, director & founder of Government Transparency Institute.

* * *

Questions and answers

14:15-14:40 **Questions & answers**

* * *

Presentations

14:40 -14:50 **Beatriz Sanz Redrado**,
Director for Anti-Fraud Knowledge Centre, and
Rita di Prospero,
Head of Unit for intelligence and Operational Analysis (OLAF)

14:50 -15:00 **Petr Suchý**
Lead developer at Hlídač Státu

15:00 -15:10 **Willem Pieter De Groen**,
Senior researcher and head of unit at CEPS

* * *

Questions and answers

15:10 - 15:35 **Questions & answers**

* * *

Conclusions and Closing remarks

15:35 -15:40 **Mr Mikulas PEKSA, MEP,**
Rapporteur, Greens

15:40 - 15:45 **Ms Monika HOHLMEIER, MEP**
Chair Committee on Budgetary Control

BRIEFING

BRIEFING

Requested by the CONT committee

European Parliament

Big data technologies and Artificial Intelligence (AI)

Introduction

The quantity of numerical information ("data") is growing in parallel with the digitalisation of our social and economic activities. The ever-growing quantity of data has bought about the term **big data,** which refers to high-volume, high-velocity and high-variety information assets[1] that are too large and complex to be effectively stored or processed with traditional methods.[2] Together with the amount of data, the methods to process it have been evolving, supported by the growth in computing capacity, resulting in changing requirements for data storage and archiving. The available data, the technological advances and the new methods also enabled the field of artificial intelligence (**AI**) and machine learning (**ML**) to materialise.

Big data is considered to be the "new oil"[3] of AI. The main advantage of using AI/ML is that it can pore through vast quantities of data in a short time, and can enable users to detect anomalies more quickly and accurately, make predictions and/or find meaning in high volumes of data. For example, ML-equipped network monitoring systems can correlate events and segment data to identify unexpected situations and fix those that could jeopardize network performance, all before an outage occurs. Machine learning is predominant within the AI techniques. According to a World Intellectual Property Organization (WIPO) study, 89% of all patent fillings are mentioning this AI technique and 40% percent of all AI-related patents are based on machine learning. Within machine learning, the specific techniques currently revolutionising AI are deep learning and neural networks.[4]

1 Gartner IT Glossary definition. https://www.gartner.com/en/information-technology/glossary/big-data

2 EPRS Briefing : Artificial intelligence: How does it work, why does it matter, and what can we do about it?, European Parliament, EPRS, 2020, https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2020)641547

3 Agrawal, A., Gans, J., & Goldfarb, A. (2018). Prediction machines: the simple economics of artificial intelligence. Boston, MA: Harvard Business Press.

4 Technology Trends 2019 : Artificial Intelligence, WIPO study, 2019 https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf

Policy Department for Budgetary Affairs
Author: Camille Marie Ligia SIMOES
Directorate-General for Internal Policies
PE 689.929 - February 2021

EN

Definitions

Open data is data that anyone can access, use and share.[5]

Machine-readable data is information in a format that can be easily processed by a computer without human intervention.[6]

Linked data is structured data, which is interlinked, with other records so it becomes more useful through semantic queries.

There is no single definition of artificial intelligence, however, it can be broadly defined as intelligence demonstrated by machines. It is the ability of computers and algorithms to perform tasks and solve complex problems that would normally require (or exceed) the natural human intelligence, reasoning, and prediction power needed to adapt to changing circumstances[7]. Alternatively, it can be defined as the **"the ability of machines to execute tasks and solve problems in a human-way"**.[8] It also refers to a **scientific discipline** trying to give the computer or the machine the capacity to "perceive, memorize, learn, reason and use a language"[9] as human beings, using **algorithms** to process data.

AI is composed of several technologies including:

- **Algorithm** is a sequence of computer instructions, typically to solve a class of problems or to perform a computation.[10]
- **Data mining** refers to the use of machine-learning algorithms to find faint patterns of relationship between data elements in large, noisy, and messy data sets, which can lead to actions to increase benefit in some form (diagnosis, profit, detection, etc.).[11]
- **Artificial Neural Network (ANN)** is a learning process inspired by the neural structures of the brain. The network is a connected framework of many functions (neurons) working together to process multiple data inputs. The network is generally organized in successive layers of functions, each layer using the output of the previous one as an input.[12]
- **Fuzzy logic** is a decision-making approach which is not based on the usual "true or false" assessment, but rather on "degrees of truth" (where the "true" value ranges between completely true and completely false).[13]

5 https://www.europeandataportal.eu/elearning/en/module1/#/id/co-01
6 Open Data Handbook, https://opendatahandbook.org/glossary/en/terms/machine-readable/
7 AI and Big Data in Entrepreuneurship, Martin Obschonka & David B. Audretsch, 2019, https://link.springer.com/article/10.1007/s11187-019-00202-4
8 Yann LeCun, Professor and Founding Director of the New York University Center for Data Science and Director of artificial-intelligence research at Facebook, https://www.franceculture.fr/emissions/linvite-des-matins/lintelligence-artificielle-nous-veut-elle-du-bien
9 Jean Gabriel Ganascia, https://www.sciencesetavenir.fr/high-tech/intelligence-artificielle/jean-gabriel-ganascia-la-peur-de-l-intelligence-artificielle-est-infondee_125861
10 "The Definitive Glossary of Higher Mathematical Jargon — Algorithm" Math Vault. August 1, 2019.
11 https://www.sciencedirect.com/topics/mathematics/data-mining
12 Technology Trends 2019 : Artificial Intelligence, WIPO study, 2019, https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf
13 Technology Trends 2019 : Artificial Intelligence, WIPO study, 2019, https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf

- **Expert Systems** were among the first successful approach of AI. It is a computer system emulating the decision-making ability of a human expert[14] : knowledge are converted into rules.

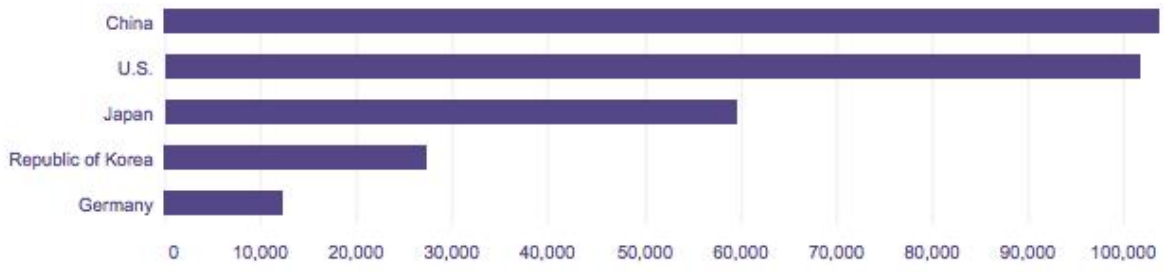
- **Machine Learning** (ML) is a technology developed with AI. It processes data to generate **prediction**. The application of ML includes for example automatic driving, image and voice recognition**[15].**

- **Deep learning** allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. It discovers intricate structure in large data sets to indicate how a machine should change its internal parameters[16]. The key aspect of deep learning is that layers of features are not designed by human engineers, they are learned from data using a general-purpose learning procedure.

Where is Europe in the race for AI?

Artificial Intelligence is "one of the most strategic technologies of the 21st century"[17]. According a study by the WIPO, 26 of the top 30 AI patent applicant worldwide are companies, while four are universities or public research organization. Leaders in AI patenting across different AI-related areas are Japan, the United States and China. While Japan filed the first AI patents, China has led the world in the number of first patent filings since 2014 followed by the USA (Figure 1). Together, these three patent offices account for 78% of total patent filings in this area of technology. [18]

Figure 1 - Number of first fillings by patent office



China and U.S. are the patent offices which are most frequently chosen as offices of first filing, followed by patent offices of Japan and South Korea.

Source : WIPO study

14 Jackson, Peter (1998). Introduction To Expert Systems (3 ed.)

15 https://www.bruegel.org/2020/02/the-dynamics-of-data-accumulation/

16 Nature Review, Deep learning Yann LeCun, Yoshua Bengio & Geoffrey Hinton, https://www.nature.com/articles/nature14539.pdf

17 Communication Artificial Intelligence for Europe, European Commission, April 2018, https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

18 Technology Trends 2019 : Artificial Intelligence, WIPO study, 2019, https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf

To remedy this situation, the President of the Commission Ursula von der Leyen proclaimed in her Agenda for Europe, that AI is a priority that *"will make the institution more agile and flexible, as well as more transparent in the way it works."* To foster innovation and development of AI, the European Commission published its Strategy on AI, announcing an increase of investment in AI research and innovation by at least €20 billion from now until the end of 2020. The Member States also signed a Coordinated Plan on AI (2018) in which they agreed to work together on the most important issues raised by AI, from ensuring Europe's competitiveness in the research and deployment of AI, to dealing with social, economic, ethical and legal questions. Moreover, the EC has established a High-Level Expert Group on AI to make recommendations on policy and investment, and set guidelines on the ethical development of AI. More recently, the Commission published the White Paper[19] on AI aiming to promote AI development in the global economy, avoid national fragmentation and prepare future legal framework. It provides suggestions for the harmonization of the European regulation and a global framework on AI, while investing in European companies."

The applications of AI

The top fields in which AI technologies are employed are telecommunications, transport, and life and medical sciences. In **telecommunication**, the main areas are computer networks and internet, radio and television, broadcasting, telephony, videoconferencing, and Voice over Internet Protocol (VoIP). **Transportation** applications encompass aerospace and avionics, autonomous vehicles, driver/vehicle recognition, transportation and traffic engineering. In **life and medical sciences**, the main uses are in bioinformatics, biological engineering, biomechanics, drug research, genetics and genomics, medical imaging, neuroscience and neurorobotics, medical informatics, nutrition and food science, physiological parameter monitoring and public health.

Beside these fields, **Big Data** technologies and AI are also used in the field of finance, amongst others in detecting fraud and money laundering. This is a field in which the EU budgetary authorities could make use of these new technologies. Indeed, to tackle corruption and money-laundering, the Council have also made recommendation to consider IT tools to improve existing monitoring systems[20].

Although European monitoring systems and procedures have improved in recent years, the number of actors potentially involved in the project implementation and the complexity of some processes can lead to corruption and fraud[21]. Indeed, a study published in January 2021 by the European Parliament highlights the difficulties of finding information on the ultimate beneficiaries of the European funds. They are partly caused by technical barriers such as data fragmentation between national and regional reporting systems, the machine

[19] White Paper on Artificial Intelligence - A European approach to excellence and trust, European Commission February 2020 https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=COM:2020:65:FIN&qid=1522050991135&from=EN

[20] Council Conclusions of 5 December 2019 on strategic priorities on anti-money laundering and countering the financing of terrorism "invites the Commission to further explore actions to enhance the Union's anti-money laundering and countering the financing of terrorism framework, including by considering whether some aspects could be better addressed through a regulation, and by exploring the opportunities and challenges in using technological innovation in combatting money laundering and countering the financing of terrorism." https://data.consilium.europa.eu/doc/document/ST-14823-2019-INIT/en/pdf

[21] Fraud and corruption in European structural and investment funds, OECD, 2019, https://www.oecd.org/gov/ethics/prevention-fraud-corruption-european-funds.pdf

readability and accessibility of data. They are also legal barriers, such as the limited information required by the systems making it impossible to identify ultimate beneficiaries[22].

Data-mining tools can indeed make monitoring system more efficient and able to detect fraud and mismanagement of public funds. The European Commission already developed ARACHNE, which is a risk scoring tool supporting managing authorities in their administrative controls with predictive models detecting fraud for the European Social Fund (ESF) and European Regional Development Fund (ERDF). It processes and analyses data of two million beneficiaries and crosses it with information from external databases that contains information on more than 210 million companies and 120 million people that are behind those companies. Nevertheless, for the time being, ARACHNE is used by Member States on a voluntary basis and access to its contents is limited to the managing authorities administering these funds.

The EU have also been developing a framework aimed to prevent money laundering and counter financing of terrorism (AML/CFT). In its 2019 communication[23], the Commission expresses the need for a high-quality and consistent supervision, and proposes conferring specific supervisory tasks to an EU body, interconnecting centralised bank account registries and strengthening mechanism to coordinate and support the work of the Financial Intelligence Units (FIUs)[24]. The Commission AML/CFT Action Plan states: *"there would be merits in building a more central capacity, based on IT tools, which should identify cross-border suspicious transactions and facilitate the identification of trends."*

Conclusion

Artificial Intelligence has spread out through the global marketplace and has a huge potential to improve working tools. It enables computers to detect patterns among billions of seemingly unrelated data points, improve forecasting, to share formatted data rapidly and could power many more autonomous applications assisting the decision-making processes. At the same time, concerns of abuse of personal data prompt the EU to understand **Big Data** technologies in order to create a regulatory framework that ensures that data is processed in a proper way while seizing the opportunity to improve monitoring and supervision systems in order to protect the EU's financial interests.

Administrator responsible: Andras SCHWARCZ Editorial assistant: Lyna PÄRT
Contact: Poldep-Budg@ep.europa.eu
This document is available on the internet at: www.europarl.europa.eu/supporting-analyses

22 Largest 50 beneficiaries in each EU Member State of CAP and Cohesion Fund, European Parliament https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CONT/DV/2021/01-25/Study_Largest50Beneficiaries_EN.pdf

23 Communication from the Commission to the European Parliament and the Council: Towards better implementation o fthe EU's anti-money laundering and countering the financing of terrorism framework, 2019 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52019DC0360

24 Communication from the Commission, on an Action Plan for a comprehensive Union policy on preventing money laundering and terrorist financing, 2020. https://ec.europa.eu/finance/docs/law/200507-anti-money-laundering-terrorism-financing-action-plan_en.pdf

BIOGRAPHIES OF SPEAKERS

Vítezslav Titl

Assistant professor of law & economics at Utrecht University, affiliated researcher at KU Leuven, specialised in big data in fraud prevention, especially concerning the monitoring of public procurement.

Vitezslav Titl is an Assistant Professor of Law & Economics Utrecht University School of Economics. He is also a member of Utrecht University Centre for Public Procurement, research/principal investigator of a Junior STAR grant at Charles University, and an affiliated researcher at KU Leuven and Free University of Brussels. Vitezslav obtained his PhD in economics from KU Leuven and during his doctoral studies held a visiting position at Bocconi University and Princeton University.

The research interest of Vitezslav Titl comprises of public economics, economics of corruption, public procurement markets and artificial intelligence. His research is funded by various national and international organizations such as the Weiss Fund at Harvard University, The Research Foundation – Flanders, and the Czech Science Foundation. He published in leading economic journals such as European Economic Review. He presented his research at top academic and policy-oriented institutions such as Bocconi University, The European Commission, The Organisation for Economic Co-operation and Development (OECD), The Parliament of the German-speaking Community of Belgium, and Princeton University. In 2019, he was a runner-up of the Young Economist of the Year in the Czech Republic and his academic articles were awarded prizes at the European Public Choice Society and Public Choice meetings. Besides the pure academic articles, he also publishes policy-oriented articles in Leuvense Economische Standpunten, IDEA CERGE-EI studies, and VoxEU columns.

Employment

11/2020 – present - Utrecht University, the Netherlands
Assistant Professor of Law & Economics (tenure track)
Affiliated Research at Utrecht University Center for Public Procurement
01/2021 – present - Charles University, the Czech Republic
Senior Researcher (part-time), Principal Investigator of a Junior STAR project.

Other Affiliation

1/2020 – present - KU Leuven, Belgium
Affiliated Researcher

Education

11/2016 – 09/2020 - KU Leuven, Belgium
PhD in Economics

Professional Experience

10/2016 – 02/2017 - The European Parliament, Brussels, Belgium
Robert Schuman Stagiaire at the Directorate for Budgetary Affairs

Selected Grants

The Role of Institutional Factors and Information in Public Procurement Markets, JUNIOR STAR Grant of the Czech Science Foundation, PI 2021-2025, (circa 227,000 EUR)
The Weiss Fund for Research in Development Economics, with Bruno Baranek and Leon Musolff, Harvard University, CO-PI 2020 (10,000 USD)

Professional Membership

European Economic Society, Royal Economic Society

Mihály Fazekas

Assistant professor at the Central European University, director & founder of Government Transparency Institute.

Education

2007	MSc Economics, and MA Pedagogy at Corvinus University of Budapest, Hungary
2009	Master of Public Policy, Hertie School of Governance, Berlin, Germany
2014	PhD in Political Sociology, University of Cambridge, Faculty of Human, Social, and Political Science, Cambridge, UK

Work Experience

2009-2010	associate analyst in the fields of institutional analysis, labour market, and education policy in the Evaluation and Audit Team of RAND Europe,
2012-2013	co-investigator in the project "Impact Assessments in Europe" (École Nationale d'Administration and University College London)
2008-2017	consultant on several projects and reviews at the Organization for Economic Co-operation and Development (OECD)
2014-2018	Scientific coordinator for the project DIGIWHIST (University of Cambridge): a large scale, Horizon2020 funded research project on public procurement data and corruption indicators across Europe and Caucasus
2010-2013	research fellow since 2013 Director Developing at Government Transparency Institute and a new corruption risk measurement framework in public procurement, legislation, and creating new indicators of collusive bidding
Since 2018	Assistant professor of public administration at Central European University – School of Public Policy

Grants

- Scientific coordinator for DIGIWHIST: The Digital Whistleblower, an EU funded project collecting procurement data and developing corruption indicators. Budget: 3 million EUR
- Co-principal investigator for a World Bank financed impact evaluation project looking at the impact of the e-procurement system on corruption in Bangladesh. Budget: 1 million USD
- Co-principal investigator for a DFID-funded research project looking at curbing corruption in development aid. Budget: 900,000 GBP

- Lead researcher for ANTICORRP at Corvinus University of Budapest, EU financed project assessing anticorruption policies in Europe. Budget: 300,000 EUR (of total 10.4 million EUR)

Selected policy impacts

- Estonian Ministry of Justice: It published its own calculations based on my data and methodology to inform local elections and highlight corruption risks in municipal administrations.
- European Investment Bank: It uses my data and indicators to target audits across Europe.
- Italian Anticorruption Agency: It changed its data collection processes as a result of my research pointing out systemic weaknesses.
- OECD-Sigma: It uses my indicators and data to benchmark EU accession countries since 2017.
- World Bank Latin America and Caribbean group actively promotes my work internally in the Bank and to client countries such as Peru, Brazil, and Paraguay organising workshops and developing online analytics tools.

Beatriz Sanz Redrado

Director for Anti-Fraud Knowledge Centre - CFE (Certified Fraud Examiner), CIA (Certified Internal Auditor), CRMA (Certification in Risk Management Assurance), DPAI (Diplôme professionnel d'audit interne)

Ms Beatriz Sanz Redrado is Director at the European AntiFraud Office since 2012, first in charge of Investigation Support and currently in charge of the General Affairs directorate as well as acting Director for the Anti-Fraud Knowledge Centre. Previously she was Vice President Audit and Chief Audit Executive in large private companies operating internationally in sectors varying from retail, defence, industry and financial services. Her responsibilities included creating and reshaping the internal audit structures, leading fraud and security investigations as well as implementing best practices in Corporate Governance.

Well renowned speaker on antifraud, governance, risk management and internal audit topics, Ms Sanz Redrado has been Chair of the Independent Advisory Oversight Committee (IAOC) of UN's WIPO (World Intellectual Property Organisation), ViceChair of the United Nations Joint Staff Pension board (UNJSPB) audit committee. Previously she has held various governance responsibilities as director of the Board of directors in the Institute of Internal Auditors' (IIA) and of the European Institute of Fraud Auditors (EIFA) and was the first qualified professional on the Audit, Accounting and Financial Ministerial Committee (CMACF) of the French Ministry of Defence.

Rita di Prospero

Head of Unit for intelligence and Operational Analysis (OLAF)

Dr Rita Di Prospero joined the European Commission in 2001. She is currently the Head of OLAF's Unit on 'Intelligence and Operational Analysis', a data driven service which offers accurate, objective, sound and timely analysis to feed into OLAF investigations. The Unit also develops and disseminates new knowledge to reinforce OLAF's internal and external stakeholders in the fight against fraud. Her previous functions in OLAF were of Head of Unit 'Investigation Workflow', Deputy Head of the Investigation Unit 'External Aid' and of the Unit 'Investigation Selection and Review'. Previously she worked for the Italian Customs Agency Antifraud Service, as well as at International Relations Office as National Coordinator of the EU Customs Cooperation Programme.

Dr Di Prospero has therefore an extensive experience in the antifraud investigation at national and international level, including leading an interethnic commission investigating fraud carried out at governmental level in postwar Bosnia Herzegovina. Dr Di Prospero also worked at the Directorate General for Migration and Home Affairs (DG HOME) where she headed the Migration Information Hub from 2015 to 2017 with the responsibility of drafting the 'Integrate Situational Awareness and Analysis' (ISAA) report within the EU 'Integrated Political Crisis Response' (IPCR) arrangements mechanism, and at the Directorate General for Taxation and Customs Union (DG TAXUD) where she worked at the conception of the New Computerised Transit System (NCTS).

Graduated (cum laude) in Political Science at 'La Sapienza' University, she holds a Master in Banking Legislation and Policy, and a PhD in Demography.

Petr Suchý

Software developer and IT specialist

Education

2009 University of Pardubice Bachelor in computer sciences

Work Experience

2019 September - Present, Hlídač Státu, DevOps
Development (.NET, C#), Datamining (MS-SQL, Elasticsearch) Docker

2017 December- 2019 August - Foxconn, Software architect,
DevOps, Microservice architecture, Docker, Kubernetes, Development C#

2015 - 2017 KBC ICT, IT Specialist
LifeCycle & configuration management, Reporting (MS-SQL, SCCM, SSRS), Scripting (Powershell, .NET)

2015 - 2017 ČSOB Pojišťovna, System analyst
Application manager, Bancassurance development (Oracle, PL/SQL) Datamining, Application management, IBM BPM

Skills

- Development C#
- Databases (MS-SQL, Postgre, MongoDB, Elasticsearch)
- DevOps (Docker, Kubernetes, Gitlab, RabbitMQ)
- Teaching (Docker, programming, microservices)

Projects

- https://www.hlidacstatu.cz/
- https://nasipolitici.cz/
- iGAS
- Orion

Willem Pieter De Groen

Senior researcher and head of unit at CEPS

Willem Pieter de Groen is a Senior Research Fellow & Heading the Financial Markets and Institutions Unit at the Centre for European Policy Studies (CEPS) in Brussels.

He has since joining CEPS in 2009 (co)-authored studies and coordinated projects on financial services, taxation and digitalisation. Willem Pieter has further extensive experience in conducting public policy evaluations (evaluations, impact assessments, cost-benefit analyses), surveys, interviews and data analysis.

He holds a bachelor in Economics from Utrecht University master in Finance from VU University in the Netherlands. Moreover, he is a visiting professor at the College of Europe.

BIBLIOGRAPHY

Bibliography of the speakers

- Dávid-Barrett, Elizabeth ; Fazekas, Mihály. "Anti-corruption in Aid-funded Procurement: Is Corruption Reduced or Merely Displaced?" World Development 132 (2020): 105000.
- Dávid-Barrett, Elizabeth ; Fazekas, Mihály. "Grand Corruption and Government Change: An Analysis of Partisan Favoritism in Public Procurement." European Journal on Criminal Policy & Research 26.4 (2020): 411-31.
- Bauhr, Monika ; Czibik, Ágnes ; Fine Licht, Jenny ; Fazekas, Mihály. "Lights on the Shadows of Public Procurement: Transparency as an Antidote to Corruption." *Governance (Oxford)* 33.3 (2019): 495-523.
- Titl, Vitezslav ; De Witte, Kristof ; Geys, Benny. "Political Donations and the Allocation of Public Procurement Contracts." *European Economic Review* 111 (2019): 443-59.
- Fazekas, Mihály ; King, Lawrence Peter. "Perils of Development Funding? The Tale of EU Funds and Grand Corruption in Central and Eastern Europe." Regulation & Governance 13.3 (2018): 405-30.
- Fazekas, Mihály ; Tóth, Bence. "The Extent and Cost of Corruption in Transport Infrastructure. New Evidence from Europe." *Transportation Research. Part A, Policy and Practice* 113 (2018): 35-54.
- Fazekas, Mihály ; Mungiu-Pippidi, Alina. "Red Tape, Bribery and Government Favouritism: Evidence from Europe." Crime, Law, and Social Change 68.4 (2017): 403-29.
- Fazekas, Mihály ; Kocsis, Gábor. "Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data." British Journal of Political Science 50.1 (2017): 155-64.
- Fazekas, Mihály ; Tóth, István János ; King, Lawrence Peter. "An Objective Corruption Risk Index Using Public Procurement Data." European Journal on Criminal Policy and Research 22.3 (2016): 369-97.
- Broms, Rasmus ; Dahlström, Carl ; Fazekas, Mihály. "Political Competition and Public Procurement Outcomes." Comparative Political Studies 52.9 (2019): 1259-293.

Further reading

- Velasco, Rafael B. ; Carpanese, Igor ; Interian, Ruben ; Neto, Octávio C. G. Paulo; Ribeiro, Celso C.. "A decision support system for fraud detection in public procurement" *International Transactions in Operational Research* 28 (2021): 27-47
- Decarolis, Francesco ; and Giorgiantonio, Cristina. "Corruption red flags in public procurement: new evidence from Italian calls for tenders." *Bank of Italy Occasional Paper* No. 544 (2020):
- Professor Giovanni Sartor, Francesca Lagioia, "The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence." European Parliament | Science and Technology Options Assessment (STOA), 2020.
- Boucher, Philip Nicholas. "Artificial Intelligence: How Does It Work, Why Does It Matter, and What Can We Do about It?" European Parliament | Science and Technology Options Assessment (STOA), 2020.

- Ageh, Kunmi. "When Artificial Intelligence Met Public Procurement: Improving the World Bank's Suspension and Debarment System with Machine Learning." Public Contract Law Journal 48.3 (2019): 565-91.
- Boucher, Philip Nicholas. "How Artificial Intelligence Works." European Parliament | Science and Technology Options Assessment (STOA), 2019.
- Reillon, Vincent. "Understanding Artificial Intelligence." European Parliament | Members' Research Service, 2018.
- Cleary, Duncan. "Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit." Electronic Journal of E-government 9.2 (2011): 132.

PRESENTATIONS

Presentation by Vítezslav Titl

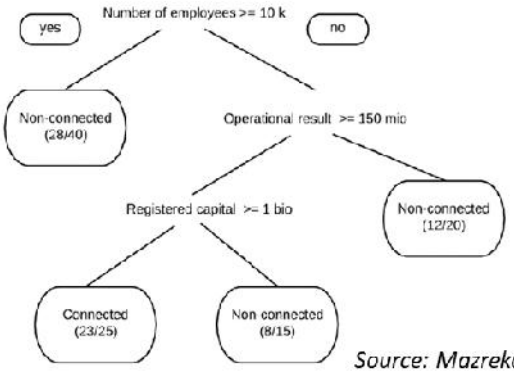
Utrecht University EUROPEAN PARLIAMENT, CONT WORKSHOP 22/02/2021

Machine learning and fraud prevention in public procurement

Dr. Vitězslav (Vita) Titl

Algorithm

- A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.



Utrecht University 18-2-2021 2

Source: Mazrekaj, Schiltz, Titl (2019)

Machine learning

- Algorithms that learn themselves.
- General idea: a machine learning algorithm learns on on a training dataset how to perform a task.
  - Subsequently, it can be used on new (test) data to perform the same task.
- In our case: the task could be to predict whether there is fraud/corruption in a particular public procurement contract or EU subsidy.

Utrecht University 18 2 2021 3

How can such algorithms be used?

General idea: instead of a random check (or tip), an algorithm can suggest a potentially fraudulent/corrupt procurement, subsidy allocation etc.

Proven high accuracy:

1. Corruption in public procurement (Ferwerda et al., 2017; Decarolis & Giorgiantonio, 2020)
2. Political connections of suppliers of public procurement (Mazrekaj, Schiltz and Titl, 2020).

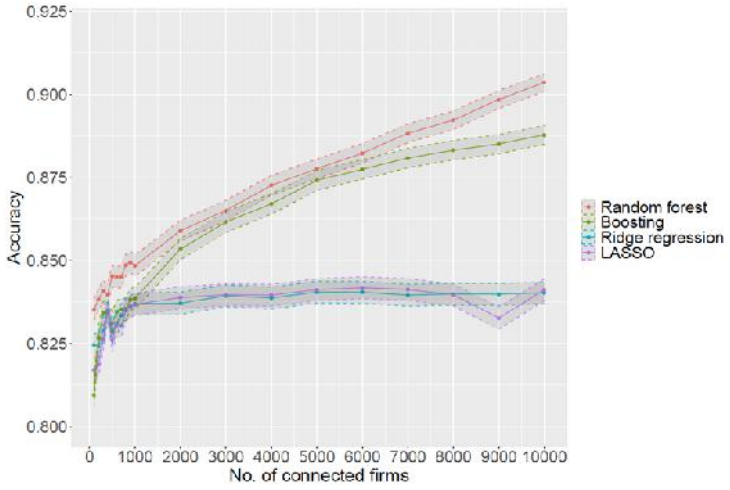
Utrecht University 18 2 2021 4

Do these algorithms predict well?

Various methods correctly predict in more than 80 % cases.

Also, if a firm is connected, there is more than 80 % probability that the algorithm discovers it.

A random draw would discover a connected firm in about 5 % of cases.



Source: Mazrekaj, Schiltz, Titl (2021)

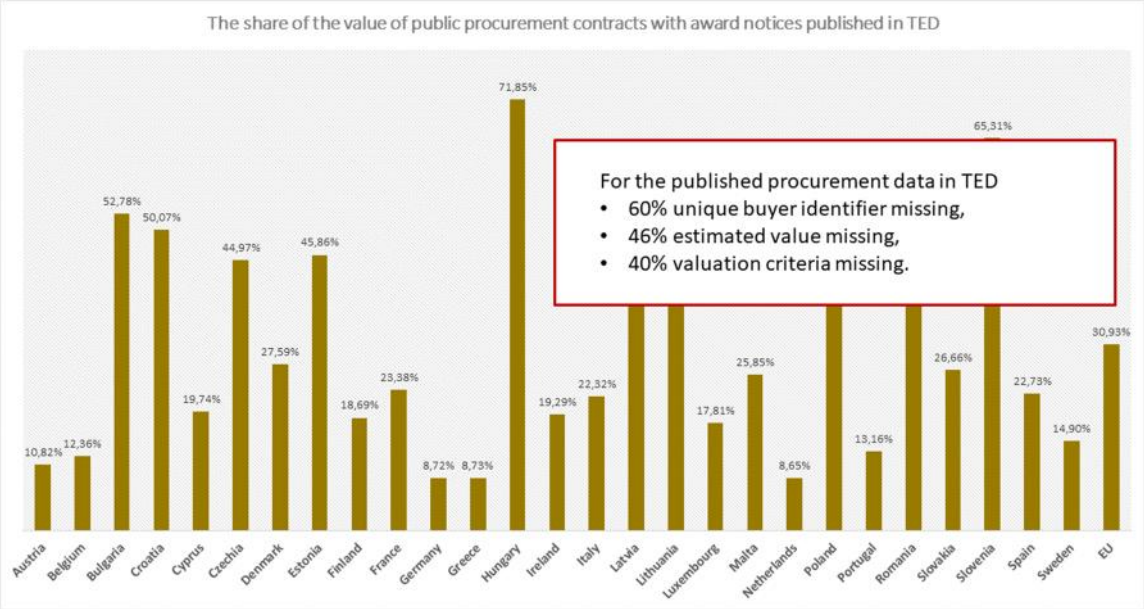
Utrecht University 18-2-2021 5

What do we need to use these algorithms?

- Good data – not many missing values and no clearly incorrect values in public datasets (see public procurement next slide).
- Interoperable data – no scanned PDFs, but simple API.
- Datasets interlinked (through unique identifies) – e.g. company registry to public procurement data.
- A sample of corrupt and non-corrupt cases (training data).

Technically, this is relatively easy. Software packages are available and often for free.

Utrecht University 18-2-2021 8



Source: Titl (2020)

Privacy issues

- The data can be **pseudo-anonymised** before they are merged and used for prediction.
  - *Only after a likely irregularity is found,* it can be communicated back to the data maintaining institution and the identity recovered.
- Before anything is communicated with the suspect, a manual check by an officer has to be done (to avoid scandals like the one in NL recently).

Utrecht University 18 2 2021 8

Presentation by Mihály Fazekas

Using Data Analytics for Detecting Corruption in Public Procurement

Mihály Fazekas
Central European University and
Government Transparency Institute
misi.fazekas@gmail.com

Use of big data and AI in fighting corruption and misuse of public funds. Committee on Budgetary Control, European Parliament.
23/2/2021

2021. 02. 18.

SCHOOL OF PUBLIC POLICY | CENTRAL EUROPEAN UNIVERSITY

Government Transparency Institute

UKaid

digiwhist

1

Main points

1. Corruption can be estimated in public procurement with tageted metrics
   - Range of red flags and validity testing methods are available
     → *See our IMF Anti-Corruption Challenge winning project*
   - BUT only if fundamental conditions are met: data scope, quality, etc.

2. Advanced analytics are for safeguarding the EU's financial interests

2021. 02. 18.

Government Transparency Institute

2

I. Measuring corruption risks

Corruption measurement steps: the case of public procurement

1. Goals of measurement: **Proxying corruption**
2. Specific definition of corruption: **High-level in PP**
3. Dictionary of corruption technologies: **long list**
4. Identify target population & data: **admin. data**
5. Tailoring and validation:
   - Modelling corrupt contracting: **parametrisation**
   - Indicator **validation**

Based on: Mungiu-Pippidi, Alina & Fazekas, Mihály (2020), How to define and measure corruption. In Alina Mungiu-Pippidi & Paul M. Heywood (eds.) A Research Agenda for Studies of Corruption. Ch. 2. Edward Elgar, Cheltenham.

Corruption definition (#2)

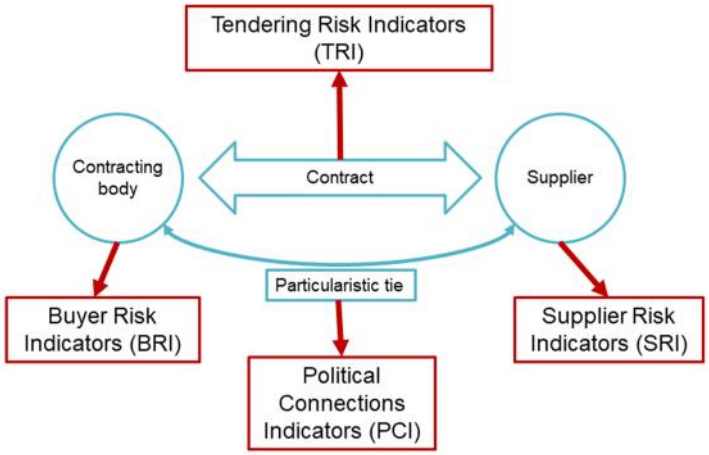
In public procurement, the aim of corruption is to steer the contract to the favored bidder without detection. This is done in a number of ways, including:

- ***Avoiding competition*** through, *e.g.*, unjustified sole sourcing or direct contract awards.
- ***Favoring a certain bidder*** by tailoring specifications, sharing inside information, *etc.*

See: World Bank Integrity Presidency (2009) Fraud and Corruption. Awareness Handbook, World Bank, Washington DC. pp. 7.

→ Note the difference from legal definitions

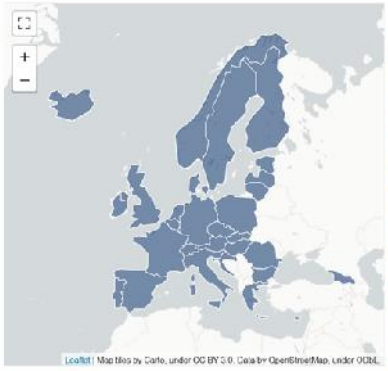
Conceptualizing public procurement corruption indicators (#3)



Source: Mihály Fazekas, Luciana Cingolani, & Bence Tóth (2018). Innovations in Objectively Measuring Corruption in Public Procurement. In Helmut K. Anheier, Matthias Haber, and Mark A. Kayser (eds.) Governance Indicators. Approaches, Progress, Promise. Ch. 7. Oxford University Press, Oxford.

Open data for public good: DIGIWHIST data maintained by the Government Transparency Institute (#4)

- Compiling and standardizing government contracting datasets is the hardest part of our work
- Over 30 million public tenders from 33 European countries+EC in a standardized format
- Containing validated transparency and corruption risk indicators



	Nr. of tenders		
Austria	74,355	Italy	5.6 Million
Belgium	92,398	Latvia	124,229
Bulgaria	346,120	Lithuania	185,223
Croatia	233,259	Luxembourg	7,569
Cyprus	7,891	Malta	9,425
Czech Republic	245,345	Netherlands	116,876
Denmark	43,963	Norway	249,563
Estonia	27,537	Poland	2.5 Million
EU Institutions	18,968	Portugal	798,267
Finland	56,926	Romania	12.7 Million
France	2.6 Million	Slovakia	127,825
Georgia	282,232	Slovenia	123,545
Germany	447,056	Spain	484,963
Greece	53,892	Sweden	110,468
Hungary	387,750	Switzerland	97,549
Iceland	3,486	United Kingdom	427,954
Ireland	37,632		

Source: https://opentender.eu/start

7

Corruption measurement methodology: parametrisation&validation (#5)

- A key **challenge for ML** in corruption measurement is the lack of learning dataset (i.e. both proven corrupt nad clean cases)
- Our ‚**Red flags**' methodology gets around this problem by looking for risky situations, for example when the tender is not advertised on the national procurement portal the probability of only one bidder bidding in a competitive market increases
- **Corruption Risk Index (CRI)** running between 0 and 1, with 1 indicating the highest risk, by combining indicative signs of corruption (i.e. 'red flags') to get a robust risk estimate

Source: GTI calculation, Mexico, federal procurement, 2017 2018

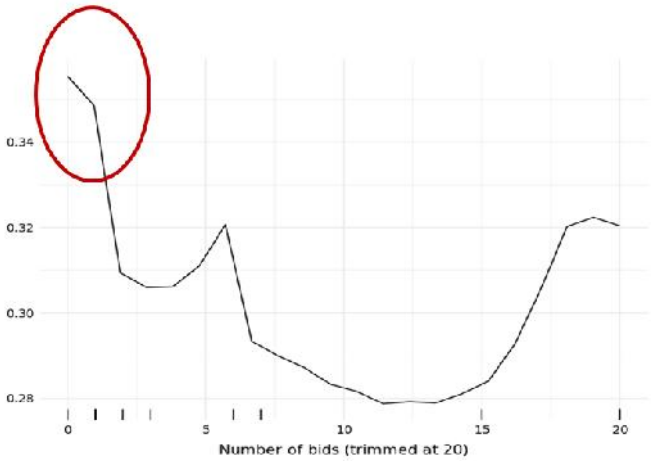
8

Machine learning: Validity testing and risk prediction

Indicator validity: Single bidding & bidder number

Proven mafia infiltration of municipal administration → central government dissolving municipality and taking over administration

✓ Random forest analysis also underpins indicator validity in the same sample of Italian municipal procurement contracts



Publication: Fazekas Mihaly, Sberna Salvatore, Vannucci Alberto (2021) The Extra-legal Governance of Corruption. Tracing the Organization of Corruption in Public Procurement. Under peer review.

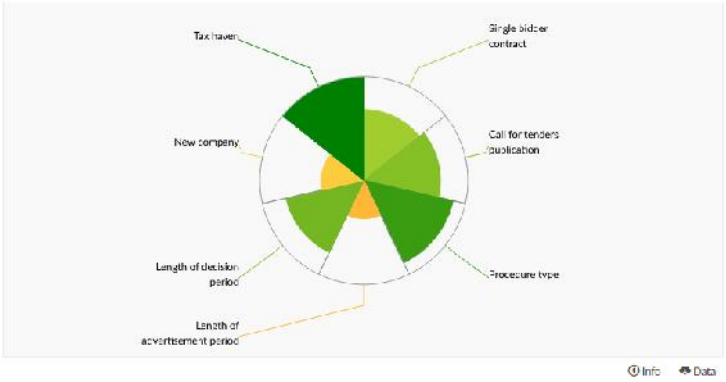
Government Transparency Institute

9

EU-wide red flags list: https://opentender.eu

Average Score of Integrity Indicators

7 validated and parametrized proxy indicators of high-level corruption in public procurement (tendering&supplier risks) for 30 million tenders throughout 2009-2020



2021.02.18. Government Transparency Institute 10

II. Using Big Data to safeguard EU Funds

2021.02.18. Government Transparency Institute 11

Uses of the indicators

1. Monitoring progress and trends
2. Preventing risks from arising
3. Investigate and punish wrongdoing

2021.02.18. Government Transparency Institute 12

Monitor risks in EU Funds:

Example of Hungary, 2009-2012

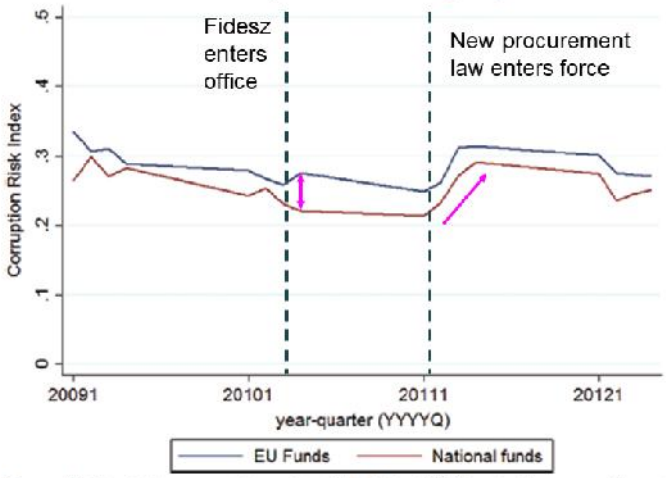


Figure G6 Quarterly average Corruption Risk Index (CRI) by funding source (unmatched samples), Hungary.

Source: Fazekas, Mihály, and King, Peter Lawrence, (2019), Perils of development funding? The tale of EU Funds and grand corruption in Central and Eastern Europe. Regulation & Governance, 13(3).

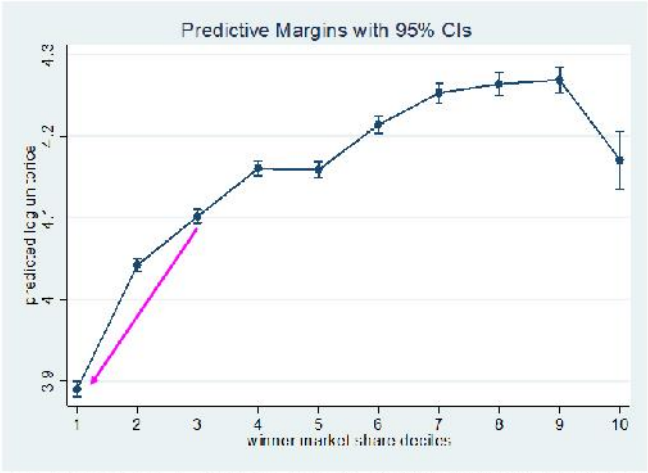
Risk prevention:

Lower concentration and increase competition

Decreaseing market concentration from the 3rd decile to the lowest decile

Decreases prices of standard goods by 20%

Federal Government of Brazil, 2014-2016



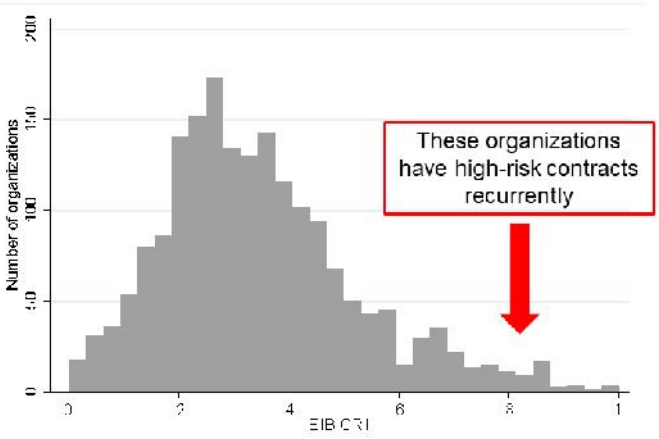
Source: De Oliveira, Alexandre; Fabregas, Abdoulaye; & Fazekas, Mihaly. (2019): Strategic Sourcing 2.0: Improving Fiscal Efficiency Using Big Data. Conference paper: "Public Procurement: Global Revolution IX" at the University of Nottingham, June 2019.

Supporting investigation:

Selecting organisations to review: European Investment Bank

Distribution of EIB-financed organizations by their composite red flag scores

This composite is the combination of red flags such as single-bidding, non-open procedures, short deadlines, extreme spending concentration etc.



Government Transparency Institute

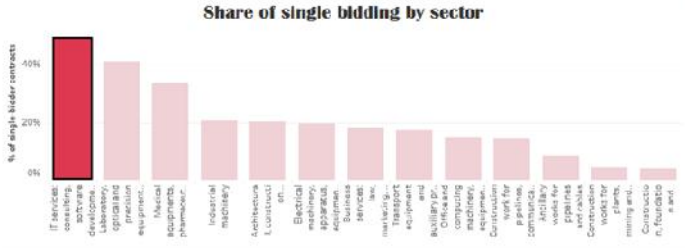
Supporting investigation:

Supplier risk ranking, EU funded contracts

Dashboard Slovakia: Companies' single bidding by sector

TOP-10 winners with highest share of single bidding contracts

Winner_name	% of contracts w single biddi..	Contract value
Vodohospodárske stavby a.s.	100%	8,527,800
Tenza Slovakia, spol. s r.o.	100%	42,739,334
Tecnotrade obrábecí stroje s.r.o., org..	100%	1,027,747,254
Dopravoprojekt, a.s.	100%	4,077,970
KPMG Slovensko, spol. s r.o.	83%	50,015,290
EuroForest, s.r.o.	71%	27,723,960
See & Go s.r.o.	60%	42,619,990
Dobrý den, s.r.o	50%	17,009,525
EXPO CENTER a.s.	33%	14,063,587
AutoCont SK a.s.	25%	9,554,241



Source: https://public.tableau.com/profile/directorate.general.for.regional.and.urban.policy#!/vizhome/Singlebidnon-open_proc/Singlebiddingoverview

Government Transparency Institute

16

Looking forward to the discussion!

Further resources

http://www.govtransparency.eu/

http://redflags.govtransparency.eu/

https://www.researchgate.net/profile/Mihaly_Fazekas/research

Selected further readings

Fazekas, M., & Kocsis, G. (2020). Uncovering High-Level Corruption: Cross-National Corruption Proxies Using Government Contracting Data. British Journal of Political Science.

Bauhr, Monika ; Czibik, Ágnes ; Fine Licht, Jenny ; Fazekas, Mihály. (2019) "Lights on the Shadows of Public Procurement: Transparency as an Antidote to Corruption." *Governance (Oxford)* 33.3: 495-523.

Dávid-Barrett, Elizabeth & Fazekas, Mihály (2019). Grand corruption and government change: an analysis of partisan favoritism in public procurement. European Journal of Criminal Policy and Research.

Fazekas, Mihály. (2019) Single bidding and non-competitive tendering procedures in EU Co-funded Projects. European Commission, Brussels.

Fazekas, Mihály, Ugale, Gavin, and Zhao, Angelina, (2019) Analytics for Integrity. Data-Driven Approaches for Enhancing Corruption and Fraud Risk Assessments. OECD, Paris.

Fazekas, Mihály ; King, Lawrence Peter. (2018) "Perils of Development Funding? The Tale of EU Funds and Grand Corruption in Central and Eastern Europe." Regulation & Governance 13.3: 405-30.

Fazekas, Mihály ; Tóth, Bence. (2018) "The Extent and Cost of Corruption in Transport Infrastructure. New Evidence from Europe." *Transportation Research. Part A, Policy and Practice* 113: 35-54.

Fazekas, M., Cingolani, L., & Tóth, B. (2016). A comprehensive review of objective corruption proxies in public procurement: risky actors, transactions, and vehicles of rent extraction: GTI-WP/2016:03. Government Transparency Institute. Budapest.

Presentation by Beatriz Sanz Redrado and Rita di Prospero

OLAF use of Big Data technologies to safeguard the EU's financial interests

Beatriz SANZ REDRADO
Director General Affairs, acting-Director Anti-Fraud Knowledge Center

Rita DI PROSPERO
Head of Unit - Intelligence and Operational Analysis

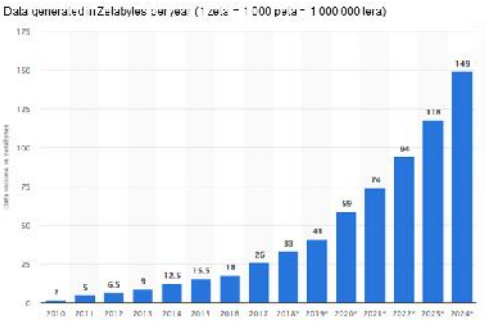
OLAF EUROPEAN ANTI-FRAUD OFFICE

The Data Landscape

In 2019 every 24 hours:

- 500 million tweets sent
- 294 billion emails sent
- 4 petabytes of data created on Facebook
- 4 terabytes of data created from each connected vehicle
- 65 billion messages sent on WhatsApp
- 5 billion searches made

Source: https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bcdf29f/



Source: https://www.statista.com/statistics/871513/worldwide-data-created/

In 2000 laptops had a hard disk average size of 40 Gb of data. Mobile devices today have an average of 60 Gb memory.

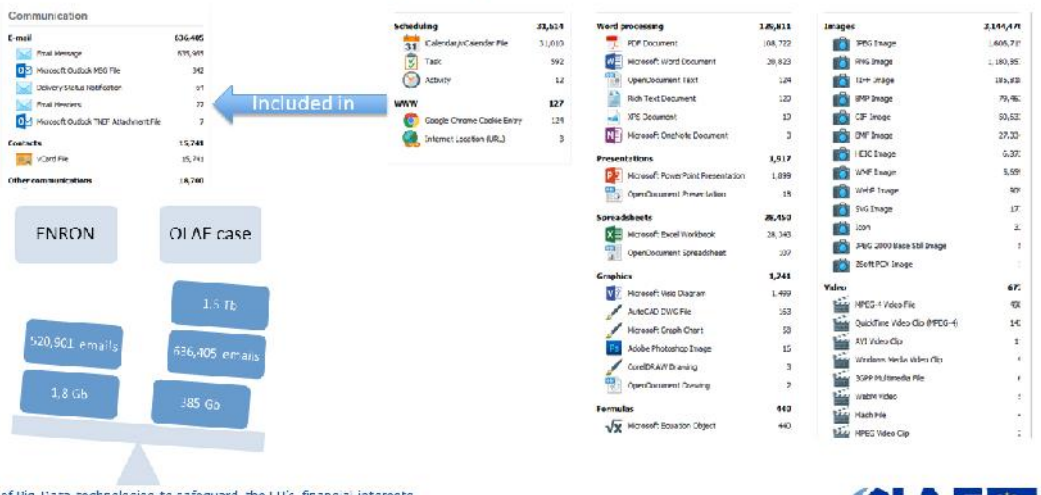
OLAF EUROPEAN ANTI-FRAUD OFFICE

Big Data & AI at OLAF

- Forensically acquired media linked to a specific case
- Closed (EC, MS) and commercial databases
- OSINT (Open Sources Intelligence)
- Building customised tools and systems

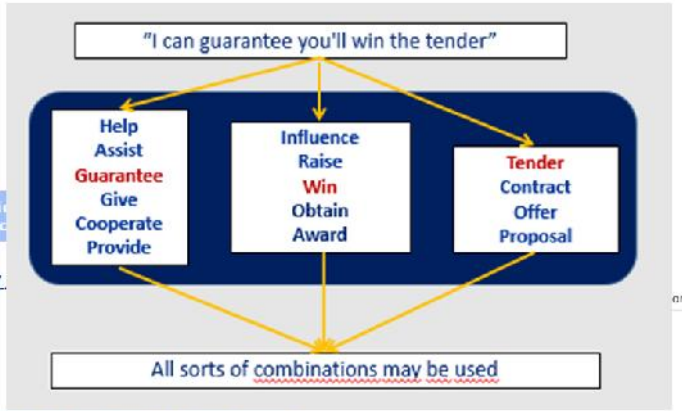
OLAF EUROPEAN ANTI-FRAUD OFFICE

Use of AI – Forensically acquired media



Use of AI – **data mining**

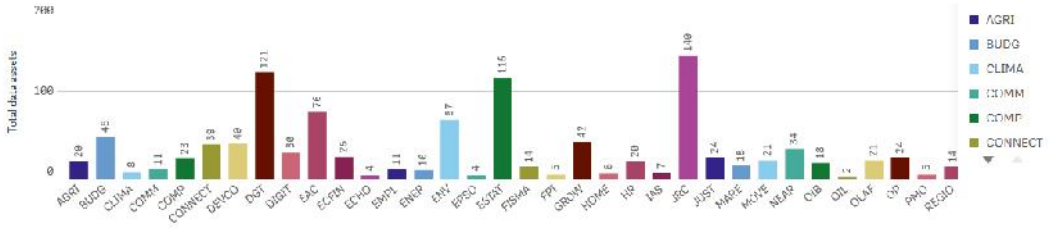
SAMPLE LIBRARY with possibly **dubious** and **suspicious language – beyond Keyword searches**



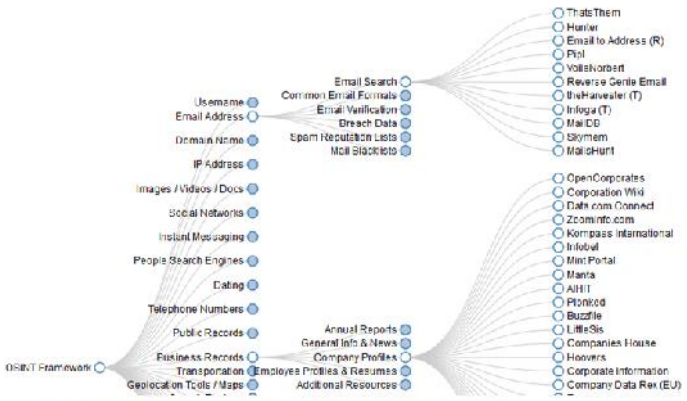
Use of AI – **data aggregation**
EU databases

1st wave DGs and services	2nd wave DGs & Services	Number of DGs reported so far	Total data assets reported
11	31	41	1.252

Number of data assets reported by DG/ Services



Use of AI – **data validation**
Open Sources Intelligence **OSINT Framework**



Use of AI - **revenue**

- **Investigation type**
  - Reiteration of Joint Customs Operation
  - Participants sent to OLAF data related to consignments originated from China
  - Almost 2 million consignment information
- **Data challenges**
  - Name of companies normalisation
  - Various formats for data
  - Language of the data (some companies were provided with Chinese characters)

Use of AI - **plagiarism**

- **Investigation type**
  - Allegation concerning plagiarism in relation to research documents
  - 25 000 documents received from audit company
- **Challenge**
  - 90% of the documents scanned as images, needing OCR
  - Comparison of the documents off-line
  - Automate the process of image comparison

CHALLENGES

- Achieve balance between applying data principles (expressed in Regulation 2018/1725) and the fact that AI is based on high volume of data
- Amount of resources dedicated to technical break-throughs
- Collecting the 'right' data and ensure no bias in use of AI

BENEFITS

- Decrease the time for operational analysis
- Enables the execution of complex tasks
- Automates processes

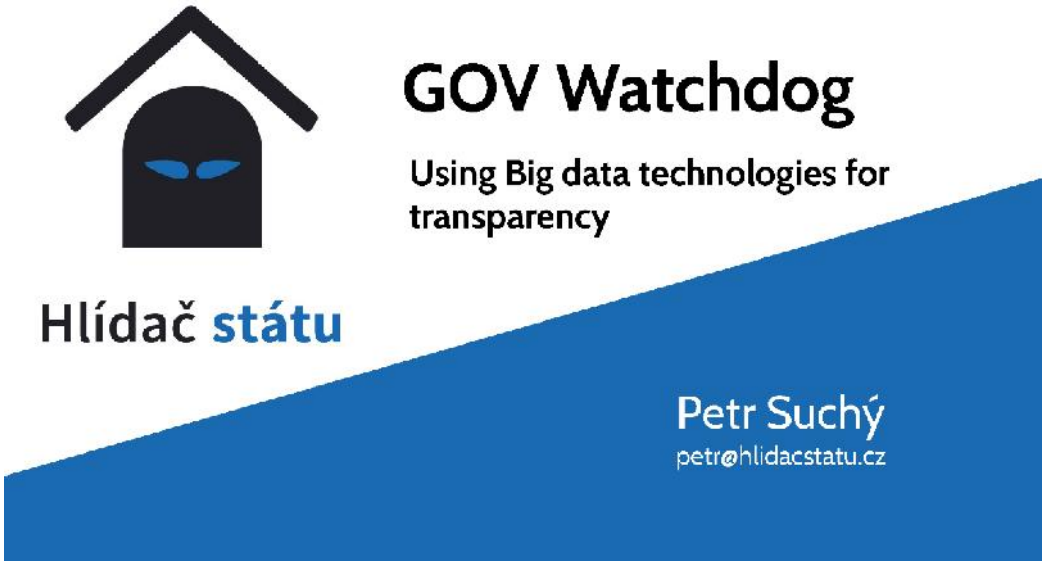
OLAF EUROPEAN ANTI-FRAUD OFFICE

Thank you for your attention

Q&A

OLAF EUROPEAN ANTI-FRAUD OFFICE

Presentation by Petr Suchý



About Hlidac Statu

- NGO founded in 2016 by Michal Bláha
- Started as a public contracts monitor
- Currently monitoring:
  - Public contracts
  - Public procurement
  - Subsidies
  - Insolvency
  - Sponsorship of political parties
  - ... and few other areas of lesser importance
- Publishing data analyses and news articles
- Processing daily hundreds of GB

Hlídač státu

Big data as a tool to fight corruption

- Publicly available contracts serve as a prevention
- Public checks
- Anomaly checks
- Ad hoc analysis according to certain criteria
  - Benchmark
  - Problematic
- Corruption patterns check

Hlídač státu

K-index (corruption index)

- Project supported by Active Citizens Fund
- Corruption risk factor indicator
- Public sector - Benchmarking
  - Ministries
  - Regional offices
  - Cities
  - Hospitals
  - Public transport companies
  - ...

Hlídač státu

Limitations and barriers to use Big Data

- Data quality
  - Completeness
  - Accuracy
  - Structuredness
  - Availability
- Documentation
  - Context
- Willingness to provide data
- Computer processable
  - Structured format, not plain text
- Legal limitations, GDPR

Hlídač státu

How could EP help to remedy limitations

- Uniform obligation to publish data
  - Clear structure and format (open data) requirements
- Unify methodology how to publish data
  - Attributes from different countries and offices are the same
  - Data with a clear meaning
- Legal obligation to publish basic registers (open data, limitless)
  - Companies
  - Contracts + invoices
  - Subsidies
- Some subsidies (EU rules in agriculture) are hidden after 2 years
  - It should be visible for longer periods

Hlídač státu

Nice to have

- ➔ Unified system with well documented API
  - Public procurement
  - Public contracts
  - Invoices
  - Subsidies

Hlídač státu

Thank you for your attention!

Hlídač státu

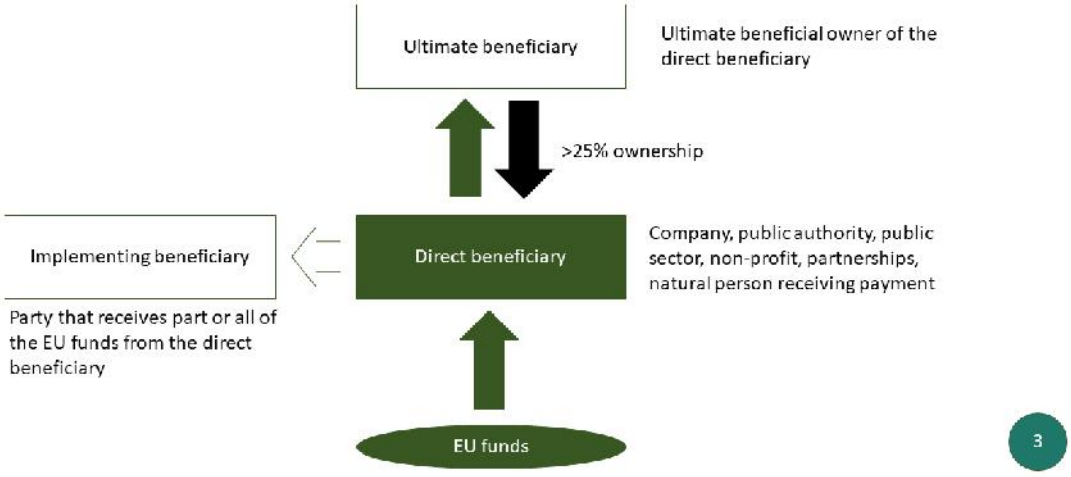
Presentation by Willem Pieter De Groen

Introduction

- CAP and Cohesion Policy is responsible for more than two-thirds of EU budget
- Currently only information on direct beneficiaries is publicly disclosed, more comprehensive information is reported to the regional and national payment agencies
- Information is fragmented into many regional, national, inter-regional and EU reporting systems
- *Information in this presentation is based on experiences with the use of existing public CAP and Cohesion Policy beneficiary information in the context of a study for the European Parliament*

2

Direct vs ultimate beneficiary



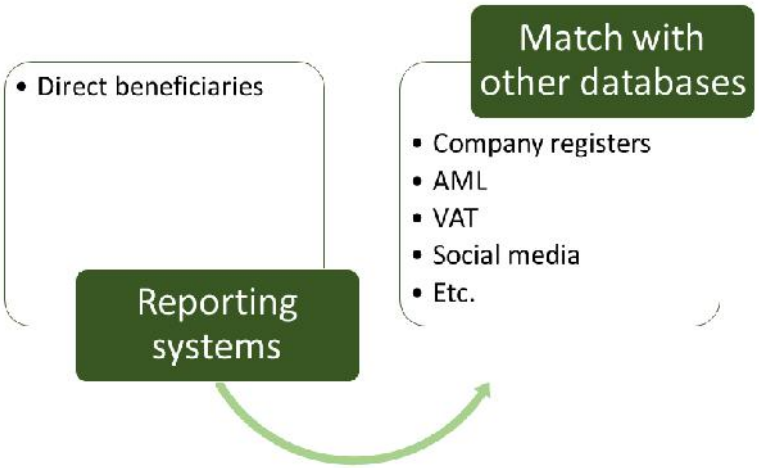
3

Main technical obstacles in accessing data

- Identifying cohesion policy reporting systems
  - More than 250 reporting systems
  - Overlap in reported beneficiaries
- Obtaining direct beneficiaries of some CAP reporting systems
  - Only about half of the reporting systems allow easy extraction
  - Search and display limitations (display of results, loading of page, CAPTCHAs, etc.)
- Different formats and indicators (e.g. address information)

5

Linking (cross-border) databases

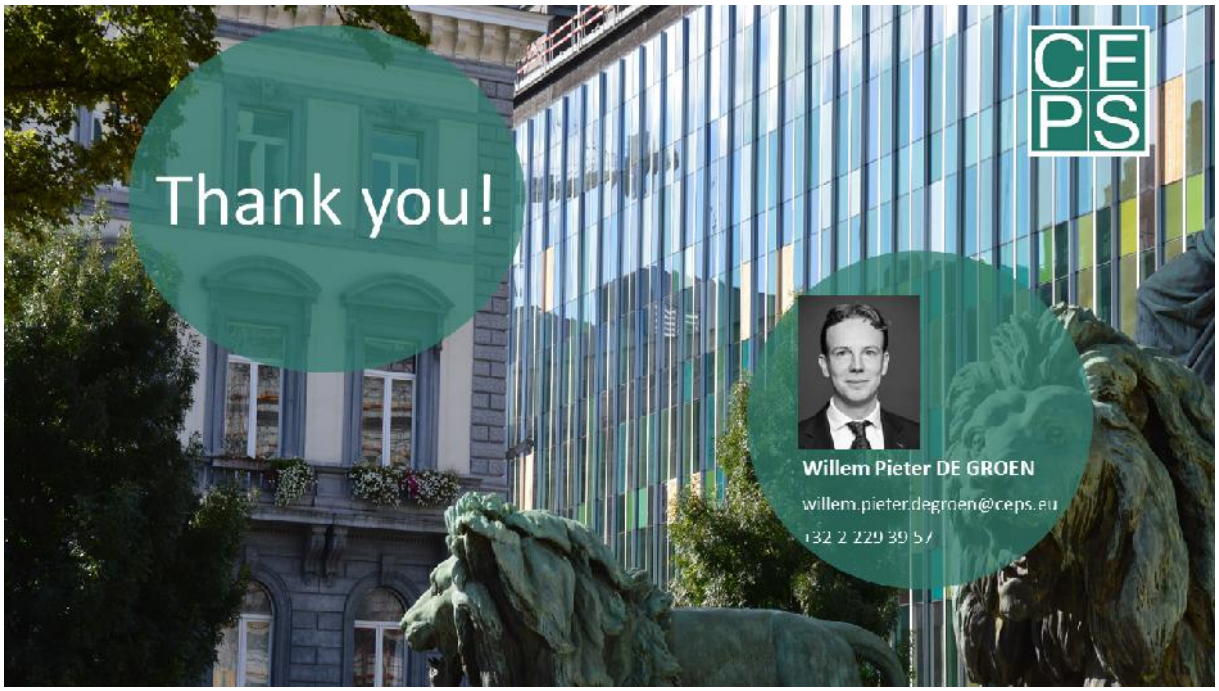


4

Main technical obstacles in linking data

- Company IDs and person IDs only privately reported
  - Single direct beneficiaries can have different identifiers
- Public reported information often insufficient
  - Names of beneficiaries are incorrectly/differently spelled
  - Names of beneficiaries are not fully disclosed
  - Name in combination with geographical location not unique
  - Databases covering public bodies and public sector entities are often not available, thus information on such beneficiaries cannot be properly linked
  - Company registers are not complete, while the use of other databases may be prohibited or restricted

5



NOTES

The Budgetary Control Committee (CONT) organised the workshop on 'Use of big data and AI in fighting corruption and misuse of public funds - good practice, ways forward and how to integrate new technology into contemporary control framework' on 23 February 2021. This document consists of the briefing on 'Big data technologies and Artificial Intelligence (AI)', biographies of the speakers and the PowerPoint slides of the presentations.