University of Warsaw
MIM
Faculty
of Mathematics, Informatics
and Mechanics

Uniwersytet Warszawski

Data analysis and visualization (DAV)

Lecture 01

Łukasz P. Kozłowski

Warsaw, 2025

lukaskoz@mimuw.edu.pl

University of Warsaw
MIM
Faculty
of Mathematics, Informatics
and Mechanics

Uniwersytet Warszawski

Data analysis and visualization (DAV)

Lecture 01

https://www.mimuw.edu.pl/~lukaskoz/teaching/dav/

Łukasz P. Kozłowski

lukaskoz@mimuw.edu.pl

Warsaw, 2025

During the course, laboratories it is assumed that you use **Linux (Ubuntu/Debian**)

(no help for Windows, Mac users will be provided)

Rules, grading, etc.

The final score (grade) for the course depends on:

1) **Homeworks** - 50%,
2) **Project** (aka exam) – 50%,

The project: given the subject the student(s) will need to prepare the poster (static + animated) in which they:
- gather the data
- interpret the data
- present it using appropriate plots (static, interactive and animated)

To pass, you need at least 60% in both the syllabus and the project.

During last two-three laboratories students will present their topic in front of the class.

The homework: after (almost) each laboratory there will be home work which need to be sent before next unit. Total scores from home works will be scaled to 50% of the final grade

The final score (grade) for the course depends on:

1) **Homeworks** - 50%,
2) **Project** (aka exam) – 50%,

The project: given the subject the student(s) will need to prepare the poster (static + animated) in which they:
- gather the data
- interpret the data
- present it using appropriate plots (static, interactive and animated)

To pass, you need at least 60% in both the syllabus and the project.

During last two-three laboratories students will present their topic in front of the class.

The homework: after (almost) each laboratory there will be home work which need to be sent before next unit. Total scores from home works will be scaled to 50% of the final grade

During each laboratory there will be a list of presence
max 2 absences are allowed
(but this does not exempt you from submitting homework on time)

Lectures

1) Every Monday 14:15-16:00

2) The slides from the lectures will be provided at:

https://www.mimuw.edu.pl/~lukaskoz/teaching/dav/

and **Moodle:**

https://moodle.mimuw.edu.pl/course/view.php?id=2388
Key: k00Q6WbW

Homeworks

1) Frequently, there will homework after laboratory (it will be either the solutions from the laboratory itself or some extra exercise(s) similar to the one done during laboratories).

2) In order to pass given laboratory you need to send email with the solutions to lukaskoz@mimuw.edu.pl

(thus we will not use moodle for sending homoworks)

Homeworks

3) The email with homework need to have specific structure:

a) The subject: **DAV25_labN_hw_Surname_Name**
For instance:
DAV25_lab1_hw_Smith_John for homowork from lab1 by Smith John
DAV25_lab2_hw_Kozlowski_Lukasz for homowork from lab2 by Kozłowski Łukasz
DAV25_lab10_hw_Rodrigez_Jose for homowork from lab10 by José Rodríguez

Note underscores, lack of special letters and the order of the parts

b) no text body in the email

c) single standard attachment with the same name compressed with 7z
For instance:
DAV25_lab1_hw_Smith_John.7z
DAV25_lab2_hw_Kozlowski_Lukasz.7z
DAV25_lab10_hw_Rodrigez_Jose.7z

The content and the structure of the attachment is laboratory specific and it will be explained separately

Homeworks

Deadline

For the homework the deadline (for all groups) is
Saturday 23:59 CET (GMT+1 Winter and GMT+2 Summer) – **one week***

* You can send delayed homework a week later, but it will be awarded by handicap/offset by -50% of the score (homework send later will not be graded)

Remember that you need on average 60% to pass

For reference, last year, students could earn up to 1,200 points, with each week's workload worth approximately 100 points.

- optimize the size of files

- do not use special letters in file names e.g. Polish

- do not use Polish (everything should be in English)

- always add legends and descriptions for the plots follow the golden rule:

one plot, one (python) script

Thus:

- do not overuse jupyter

- do not overuse any web browser solutions follow the golden rule:

one plot, one (python) script

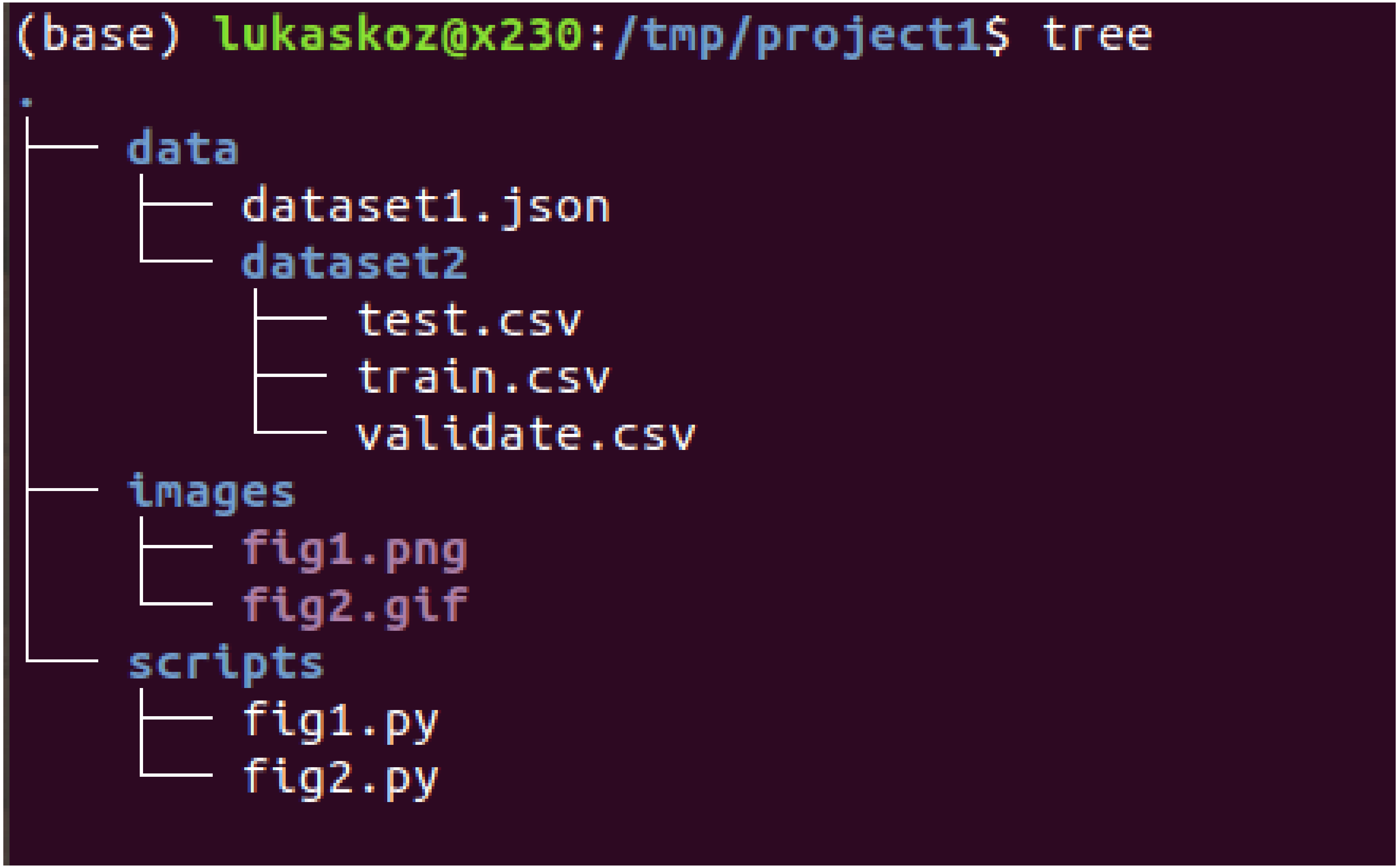
Thus:

- do not overuse jupyter

- do not overuse any web browser solutions

- make proper project structure

Be prepared to present everything as static file
(preferably pdf)



```
(base) lukaskoz@x230:/tmp/project1$ tree
.
├── data
│   ├── dataset1.json
│   └── dataset2
│       ├── test.csv
│       ├── train.csv
│       └── validate.csv
├── images
│   ├── fig1.png
│   └── fig2.gif
└── scripts
    ├── fig1.py
    └── fig2.py
```

- make proper project structure

Be prepared to present everything as static file
(preferably pdf)

DATA

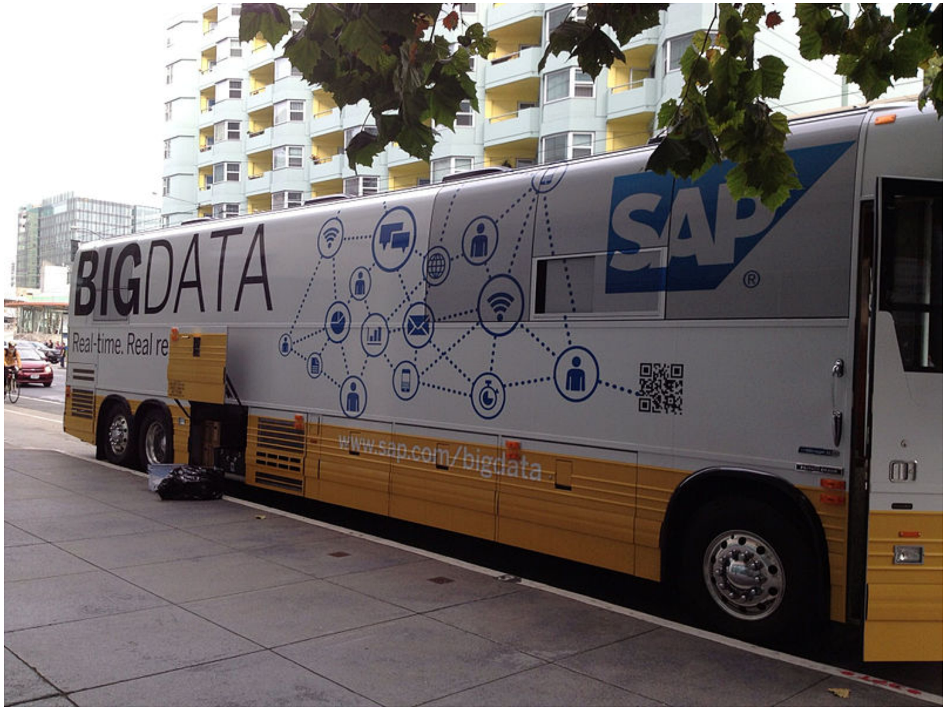
Data analysis and visualization

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making.

Wikipedia

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making.

Wikipedia



BIGDATA

We will be working with **Big Data** in **Bioinformatics**

This means that you will need to have solid background in basic biology or refresh it (there will no time to do it during the course)

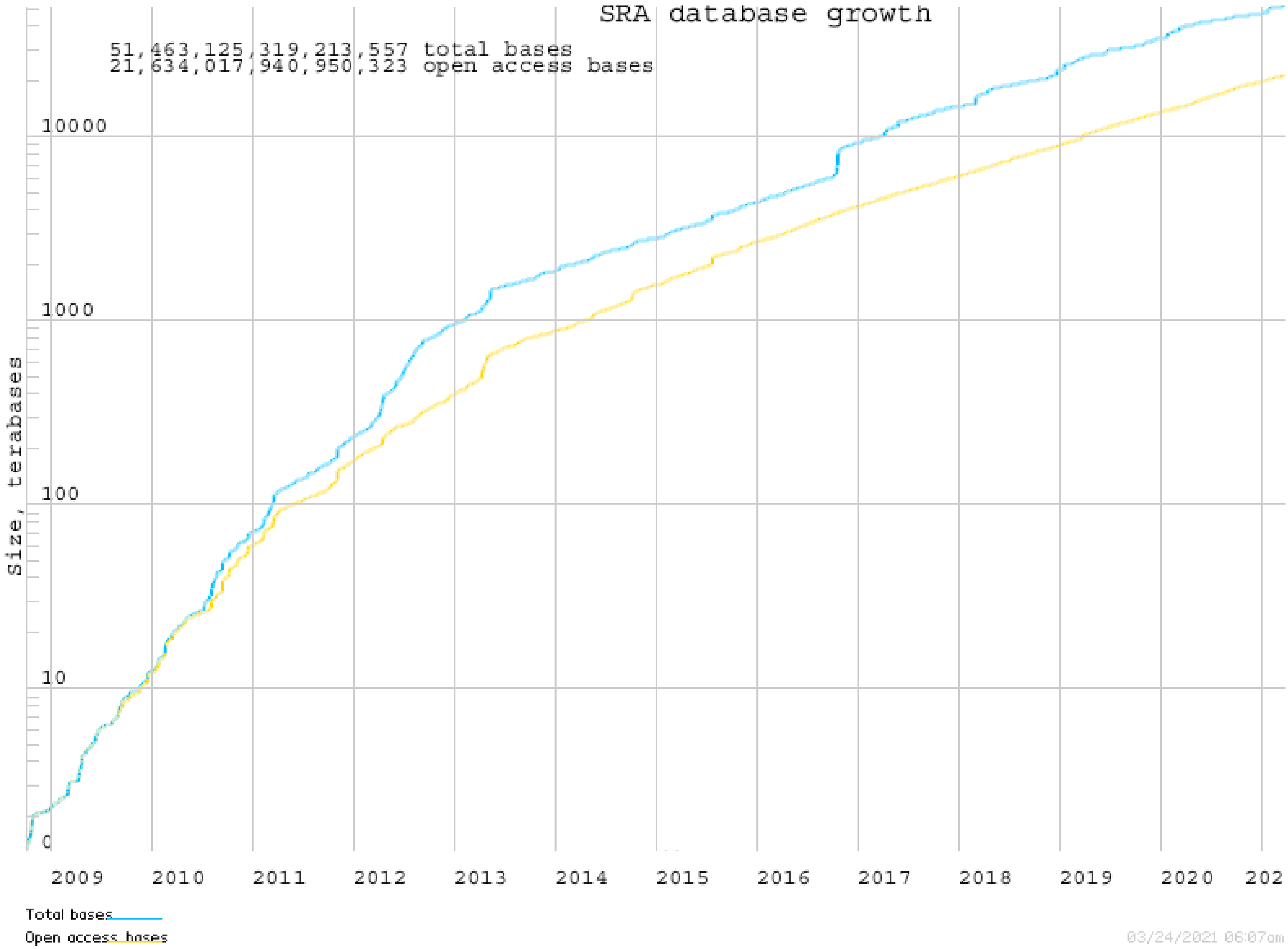
We will be working with **Big Data** in **Bioinformatics**

This means that you will need to have solid background in basic biology or refresh it (there will no time to do it during the course)

DIY (biology)

you need to know what is DNA, RNA, protein, what are their letters (ATGCN, AUTCN, VXCDBFMOLNYIQTGHWESKPAUR) and know basic facts from genetics (such as genetic code, gene, genome, proteome, cell structure, organism kingdoms, tree of life, etc.).

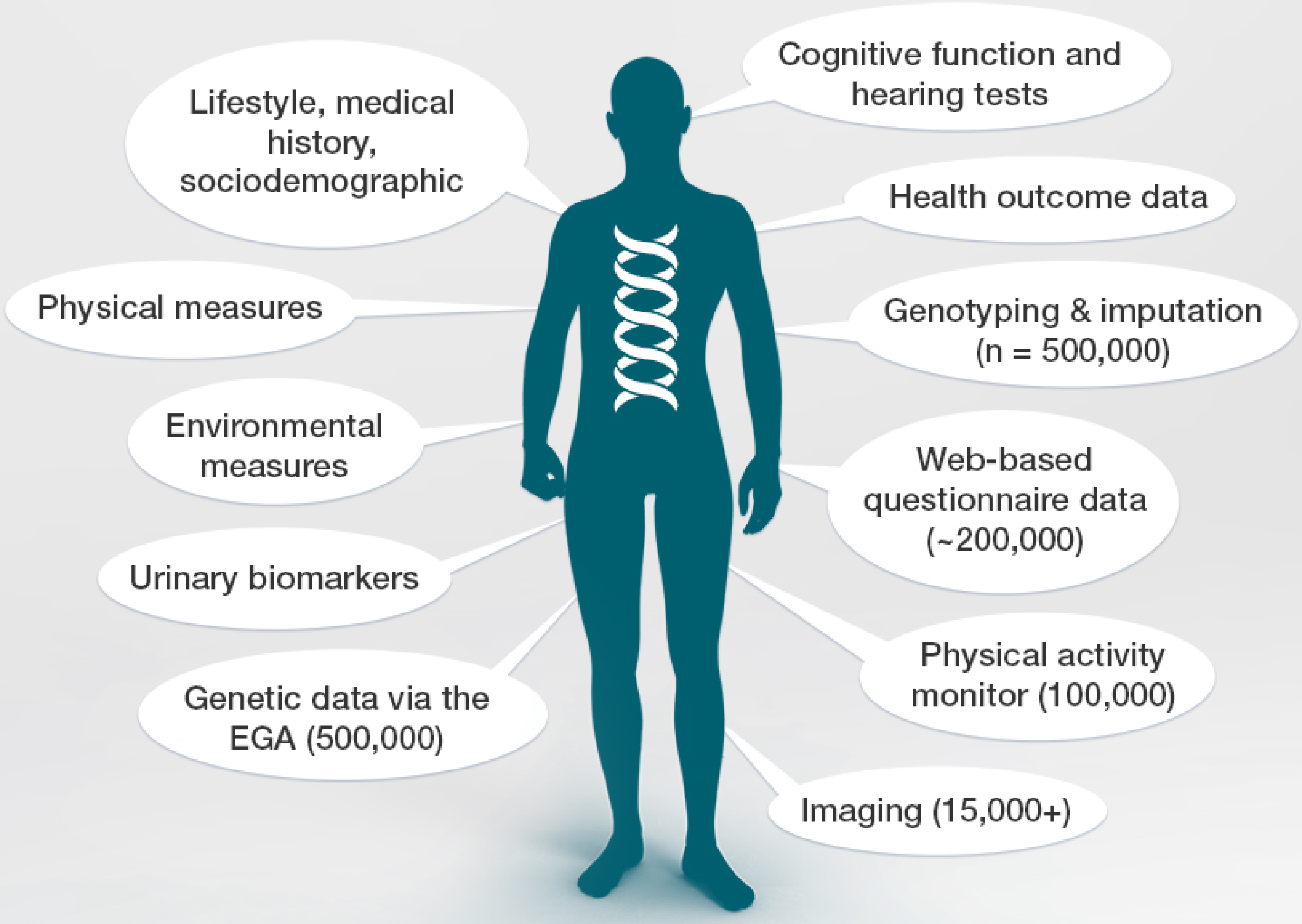
Sequence Read Archive (SRA)



biobank uk

Enabling scientific discoveries that improve human health

Data on UK Biobank participants



Lifestyle, medical history, sociodemographic

Physical measures

Environmental measures

Urinary biomarkers

Genetic data via the EGA (500,000)

Cognitive function and hearing tests

Health outcome data

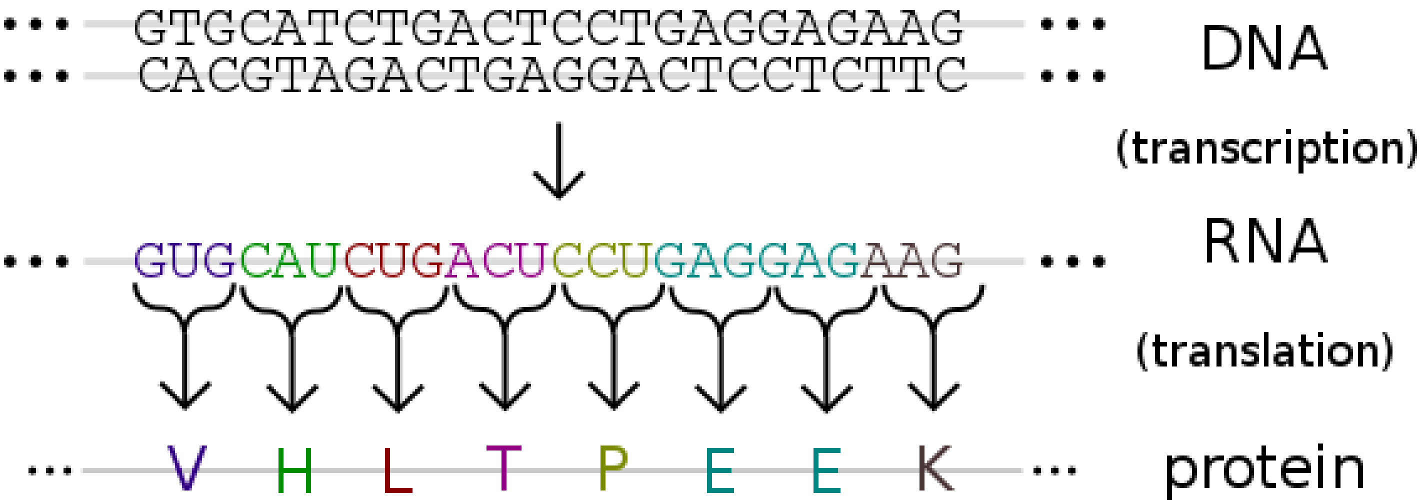
Genotyping & imputation (n = 500,000)

Web-based questionnaire data (~200,000)

Physical activity monitor (100,000)

Imaging (15,000+)

https://www.ukbiobank.ac.uk/



... GTGCATCTGACTCCTGAGGAGAAG ...

... CACGTAGACTGAGGACTCCTCTTC ...

DNA

(transcription)

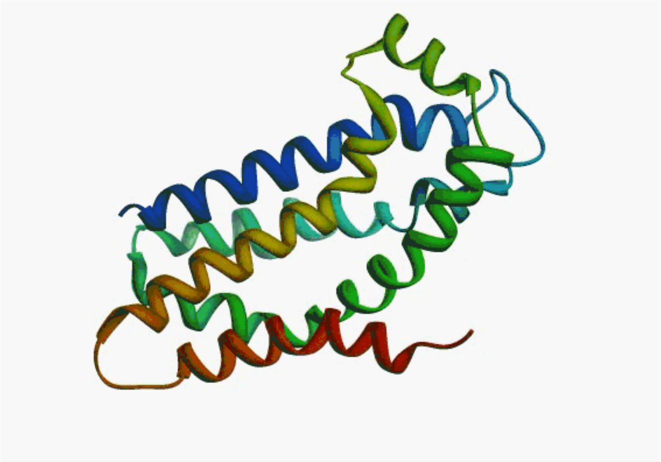
... GUGCAUCUGACUCCUGAGGAGAAG ...

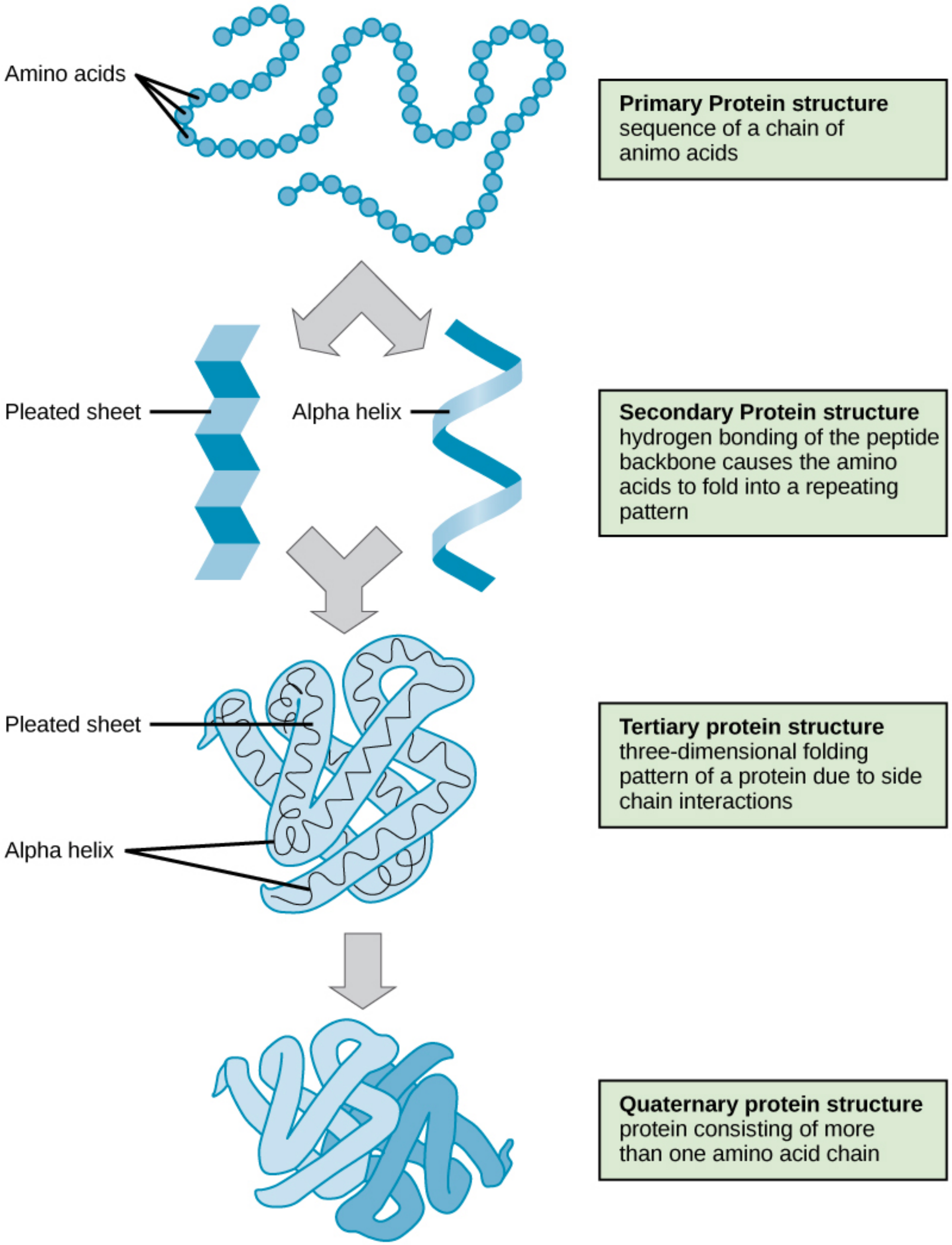
RNA

(translation)

... V H L T P E E K ...

protein



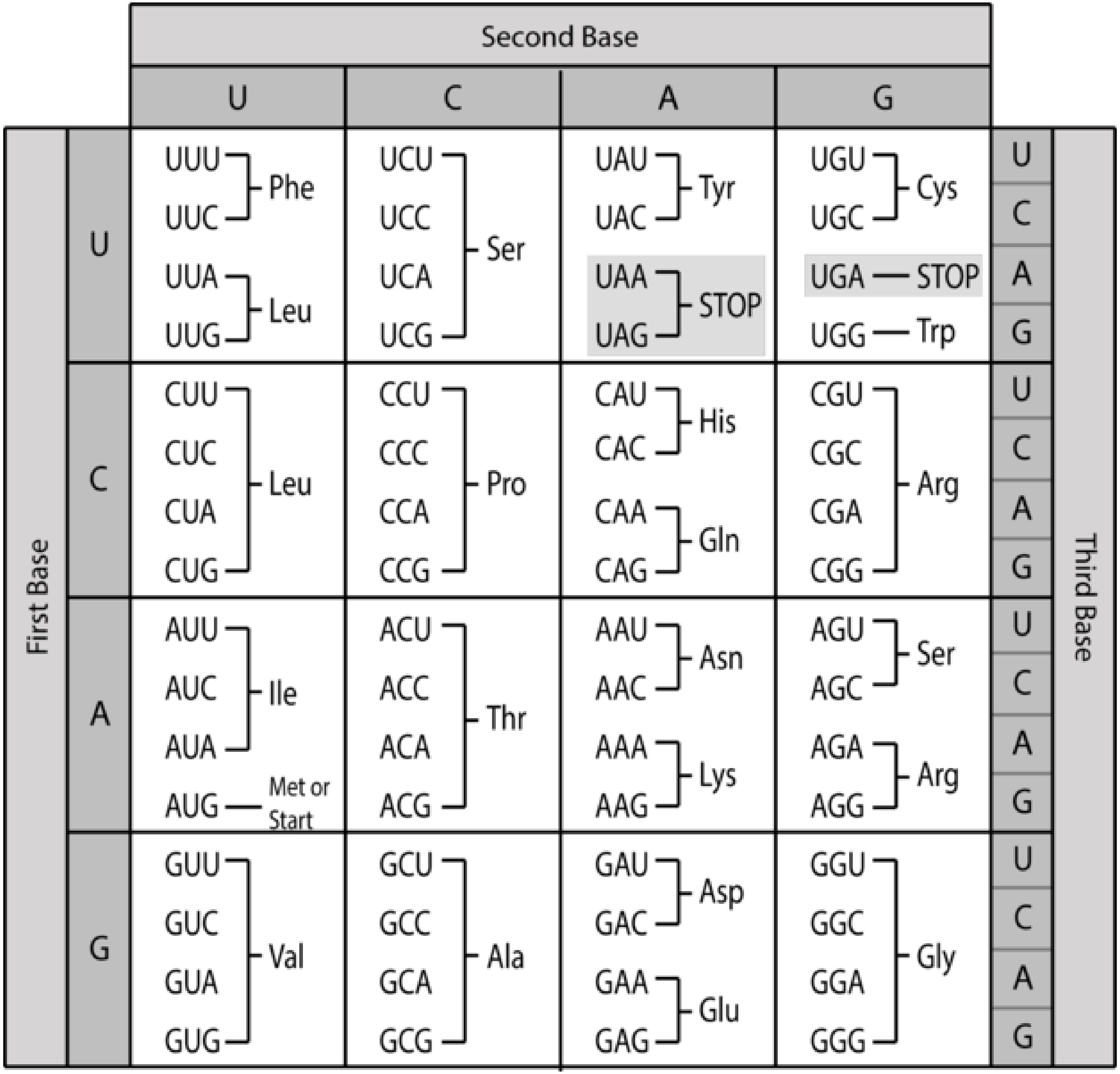


Primary Protein structure
sequence of a chain of
animo acids

Secondary Protein structure
hydrogen bonding of the peptide
backbone causes the amino
acids to fold into a repeating
pattern

Tertiary protein structure
three-dimensional folding
pattern of a protein due to side
chain interactions

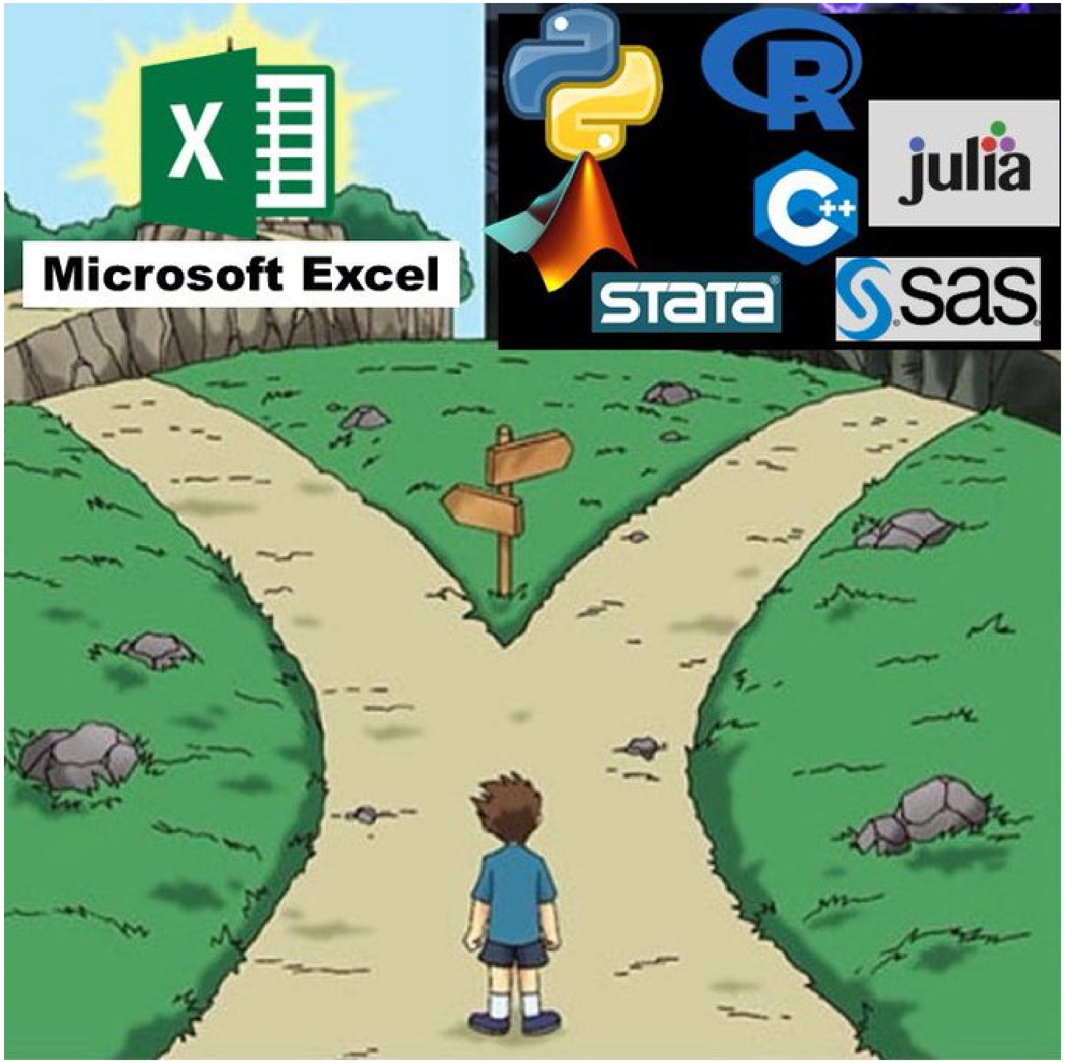
Quaternary protein structure
protein consisting of more
than one amino acid chain

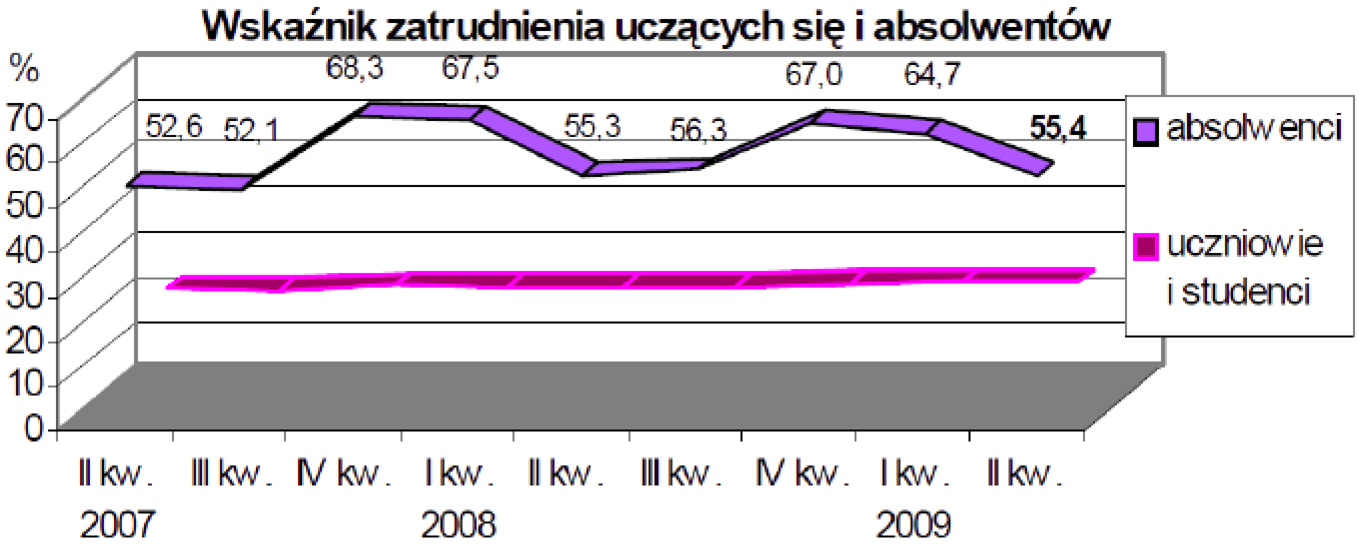


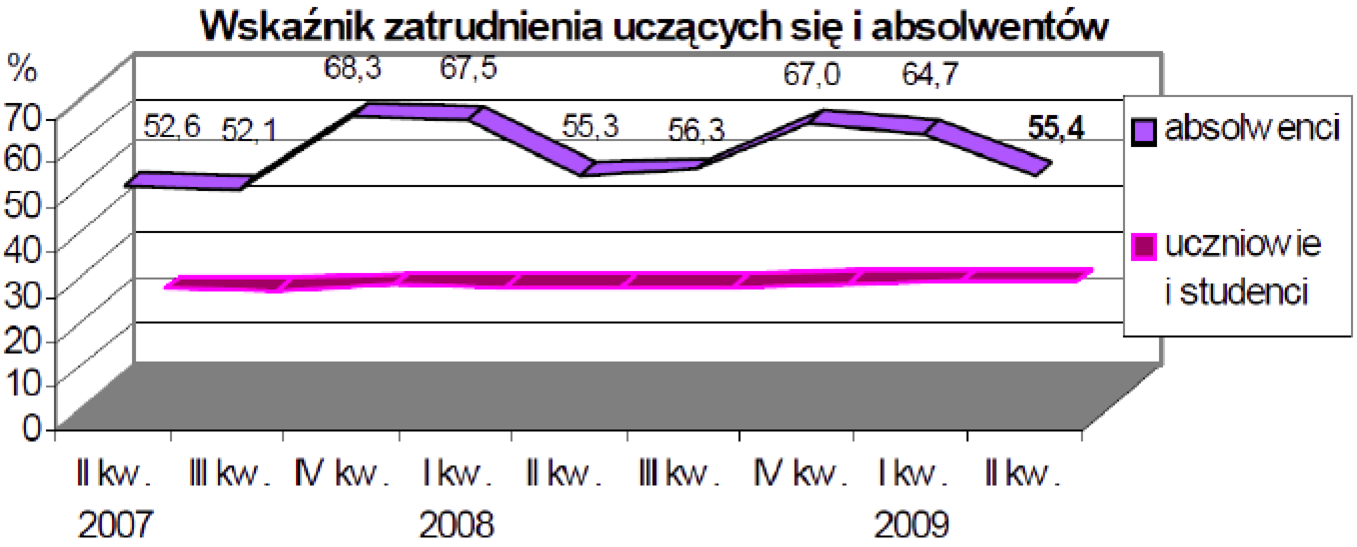
	Second Base								
First Base	U		C		A		G		Third Base
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met or Start	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Data formats

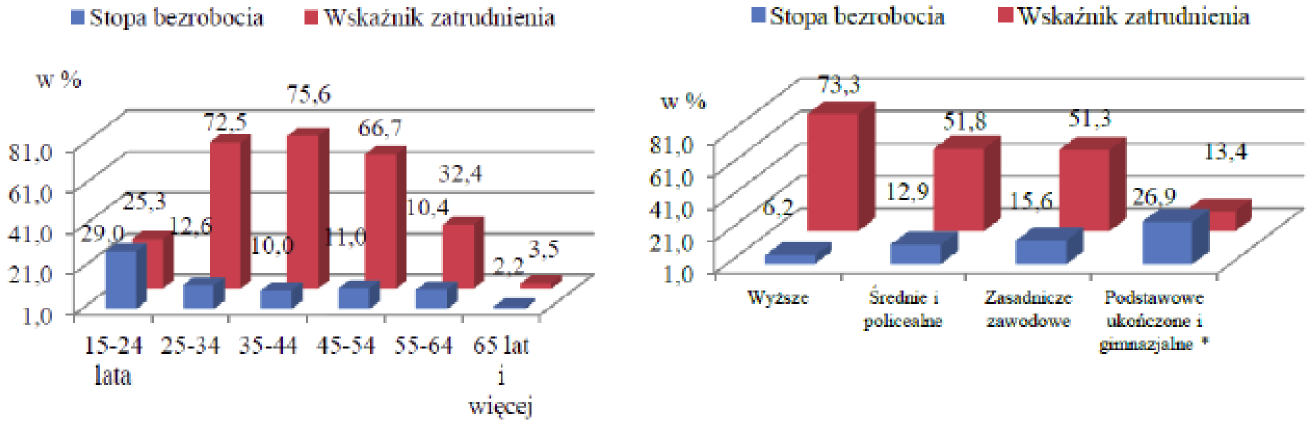
Data formats



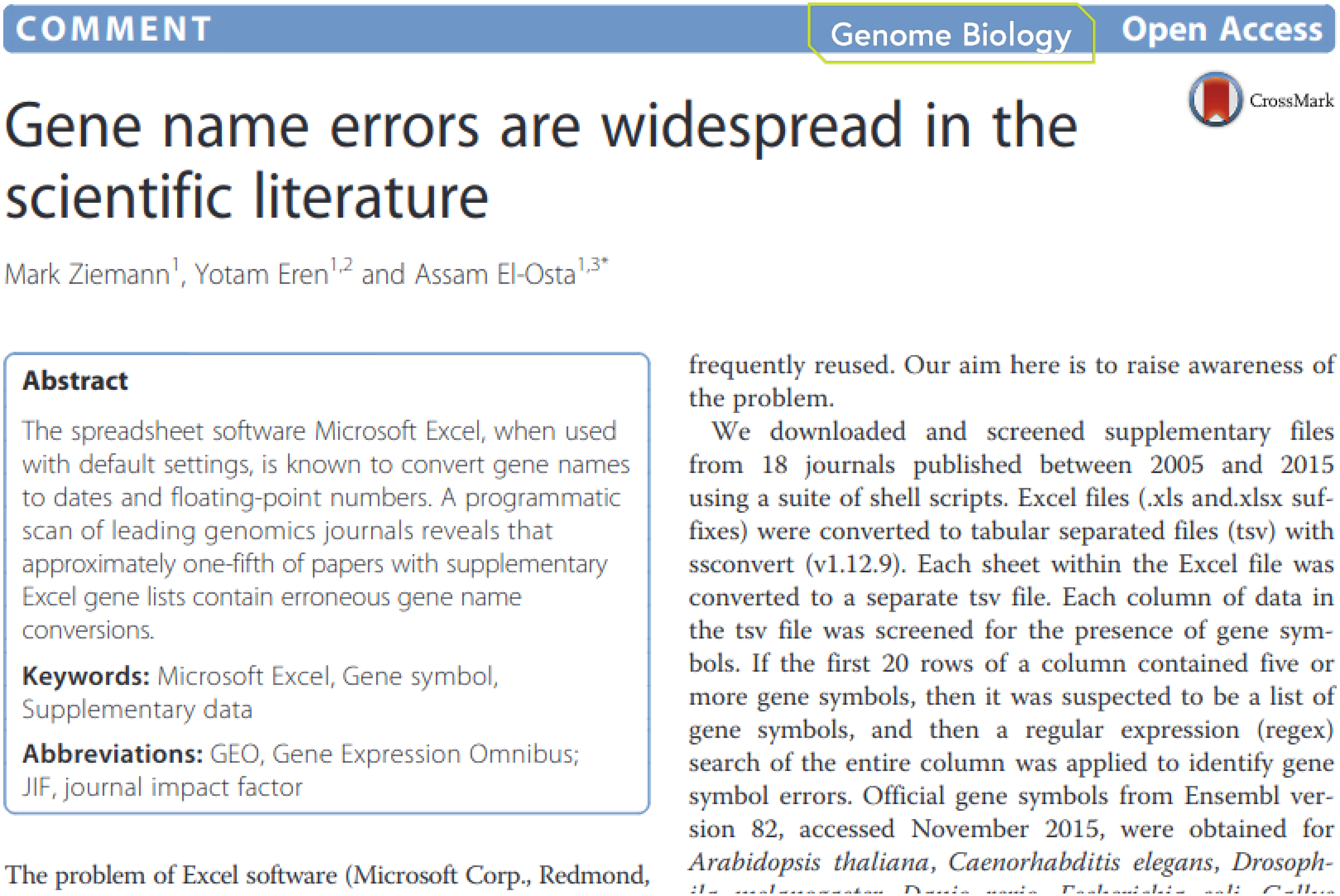

Wskaźnik zatrudnienia uczących się i absolwentów
%
70
60
50
40
30
20
10
0
52,6
52,1
68,3
67,5
55,3
56,3
67,0
64,7
55,4
absolw enci
uczniow ie i studenci
II kw. 2007
III kw.
IV kw.
I kw. 2008
II kw.
III kw.
IV kw.
I kw. 2009
II kw.


Wskaźnik zatrudnienia uczących się i absolwentów

Wykres 7. Wskaźnik zatrudnienia oraz stopa bezrobocia według grup wieku i poziomu wykształcenia w 2011 r.



* łącznie z wykształceniem podstawowym nieukończonym i bez wykształcenia szkolnego



COMMENT Genome Biology Open Access

CrossMark

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor

The problem of Excel software (Microsoft Corp., Redmond,

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene symbols. If the first 20 rows of a column contained five or more gene symbols, then it was suspected to be a list of gene symbols, and then a regular expression (regex) search of the entire column was applied to identify gene symbol errors. Official gene symbols from Ensembl version 82, accessed November 2015, were obtained for *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Escherichia coli*, *Gallus*

Data formats

Gene symbols:
SEPT2 (Septin 2) --> ‘2-Sep’
MARCH1 [Membrane-Associated Ring Finger (C3HC4) 1] --> ‘1-Mar’

RIKEN identifiers:
‘2310009E13’ --> ‘2.31E+13’

Gene symbols:
SEPT2 (Septin 2) --> ‘2-Sep’
MARCH1 [Membrane-Associated Ring Finger (C3HC4) 1] --> ‘1-Mar’

RIKEN identifiers:
‘2310009E13’ --> ‘2.31E+13’

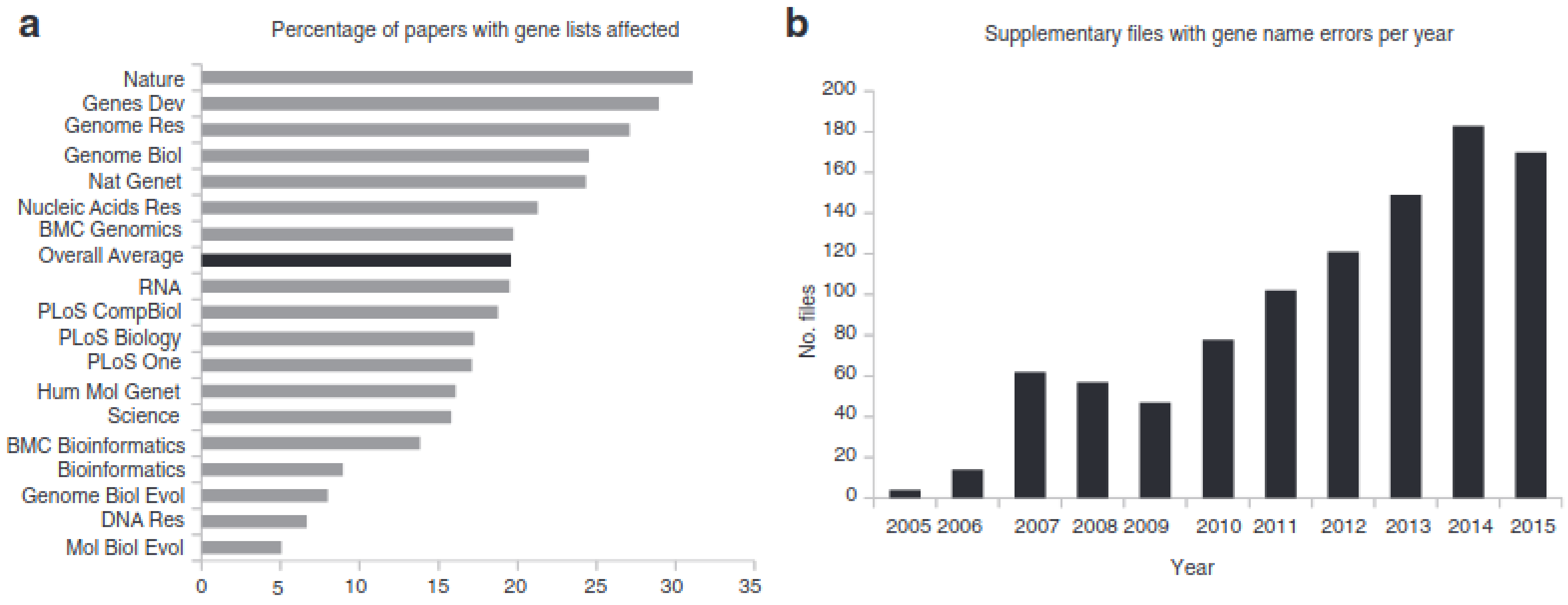
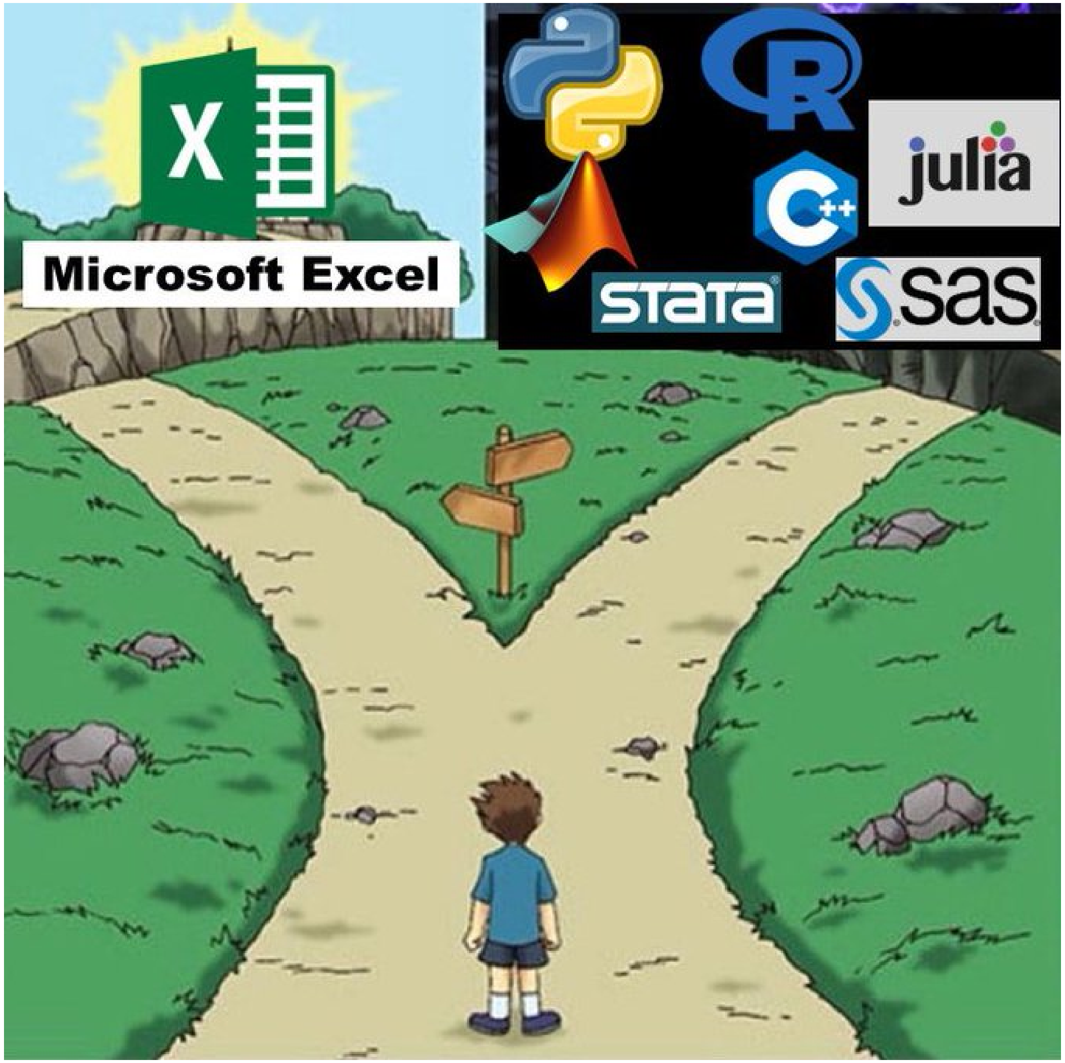


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year

Data formats

X
CSV
json
<>
XML



Programming language

This is not python programming course and it means that you should **know** how to program or you must learn the basics within short time

This is not python programming course and it means that you should **know** how to program or you must learn the basics within short time

1. Dive Into Python 3 (also in Polish, check on wikibooks)

2. Python Data Analysis, Ivan Idris, 2014

3. Python for Data Analysis, Wes MacKinney, 2013

python™

Python was conceived in the late 1980s

Guido van Rossum

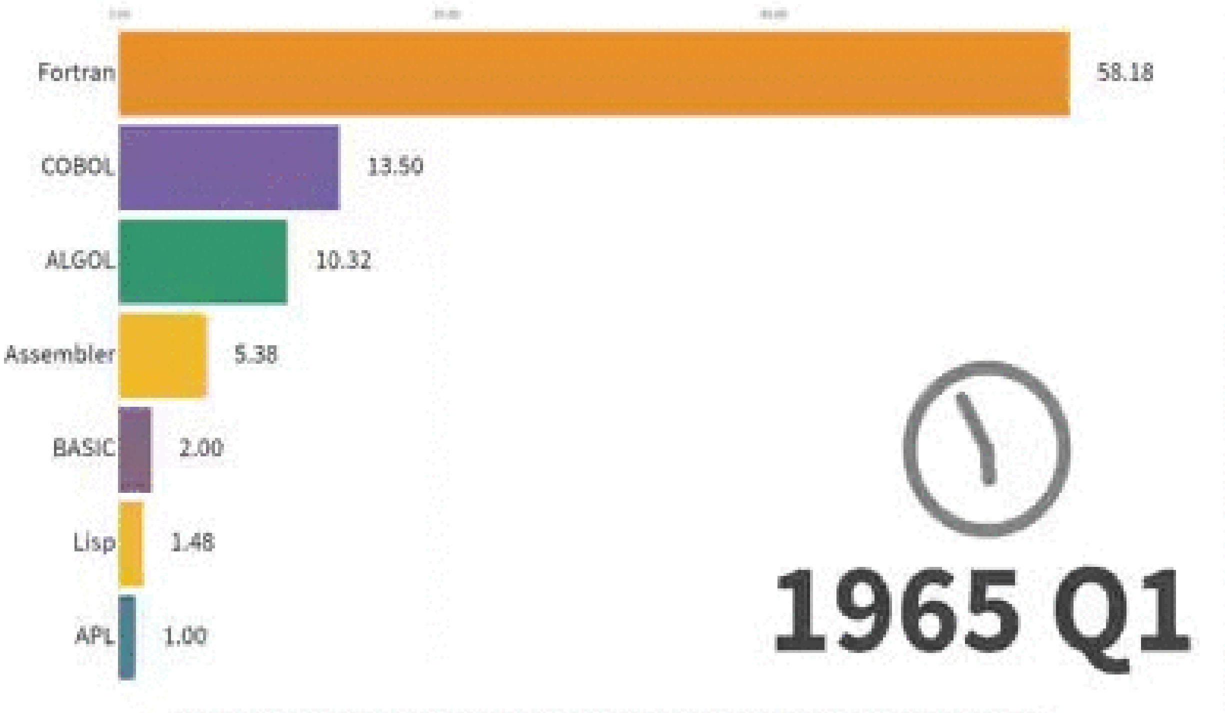
- scripting language (no compilation)

- uses whitespace indentation, rather

 than curly brackets or keywords,

 to delimit blocks

- we use **python3** instead **python2**



https://www.youtube.com/watch?v=Og847HVwRSI

Programming language

This is not python programming course and it means that you should **know** how to program or you must learn the basics within short time

We will use a number of python libraries useful for data analysis and visualization such as:

`jupyter, spyder, numpy, scipy, numba, pandas, dask, bokeh, holoviews, datashrader, matplotlib, sckit-learn, seaborn, plotly and many others`

We will use also external tools such as image-magick

Programming language

This is not python programming course and it means that you should **know** how to program or you must learn the basics within short time

We will use a number of python libraries useful for data analysis and visualization such as:

`jupyter``, spyder, `**`numpy`**`, `**`scipy`**`, numba, `**`pandas`**`, dask, bokeh, holoviews, datashrader, `**`matplotlib`**`, `**`sckit-learn`**`, seaborn, plotly and many others`

We will use also external tools such as **image-magick**

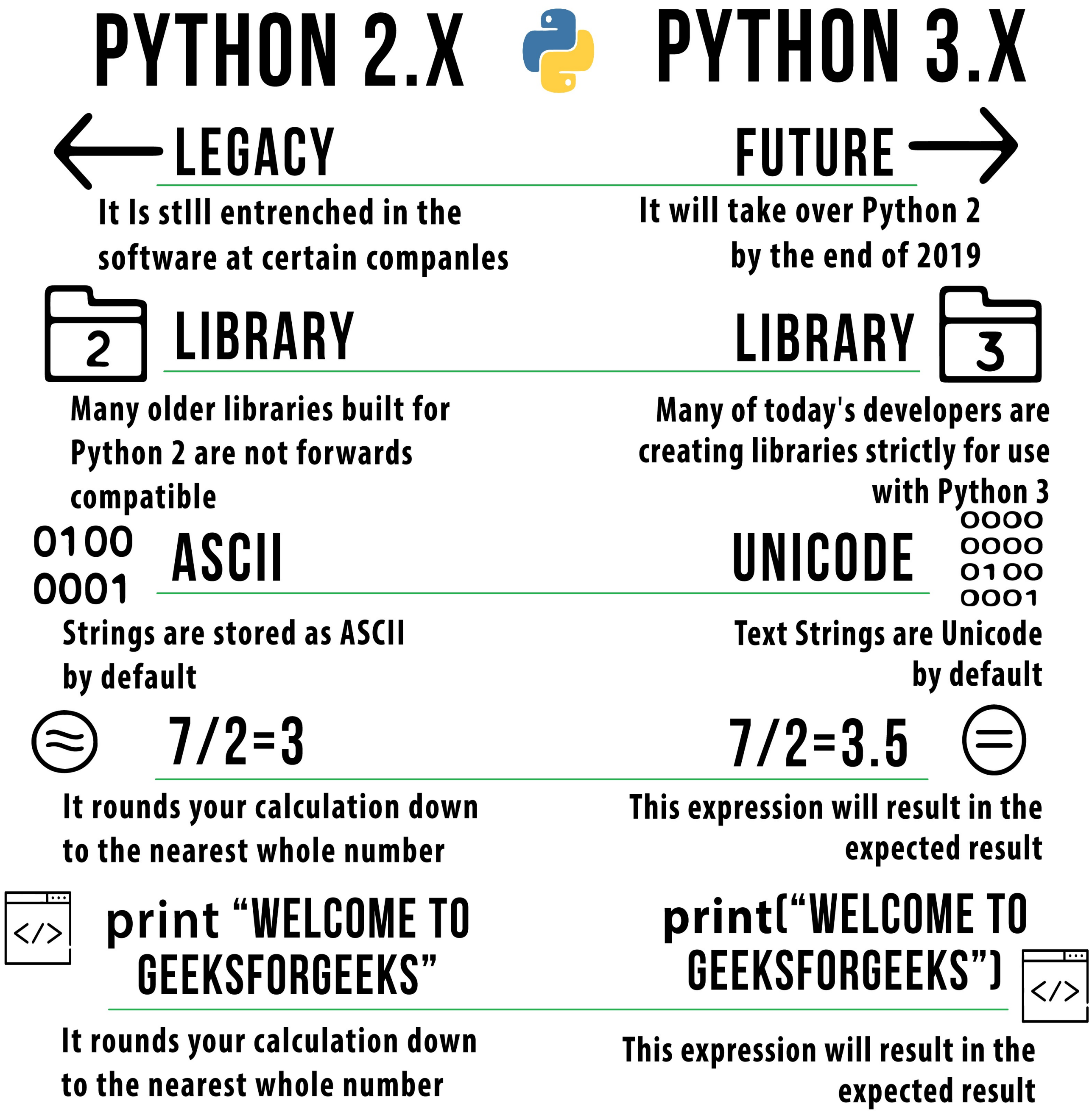
Programming language

This is not python programming course and it means that you should **know** how to program or you must learn the basics within short time

We will use a number of python libraries useful for data analysis and visualization such as:

jupyter, spyder, **numpy**, **scipy**, numba, **pandas**, dask, bokeh, holoviews, datashrader, **matplotlib**, **sckit-learn**, seaborn, plotly and many others

We will use also external tools such as **image-magick**



PYTHON 2.X

PYTHON 3.X

LEGACY

It Is stIll entrenched in the software at certain companles

FUTURE

It will take over Python 2 by the end of 2019

LIBRARY

Many older libraries built for Python 2 are not forwards compatible

LIBRARY

Many of today's developers are creating libraries strictly for use with Python 3

ASCII

Strings are stored as ASCII by default

UNICODE

Text Strings are Unicode by default

7/2=3

It rounds your calculation down to the nearest whole number

7/2=3.5

This expression will result in the expected result

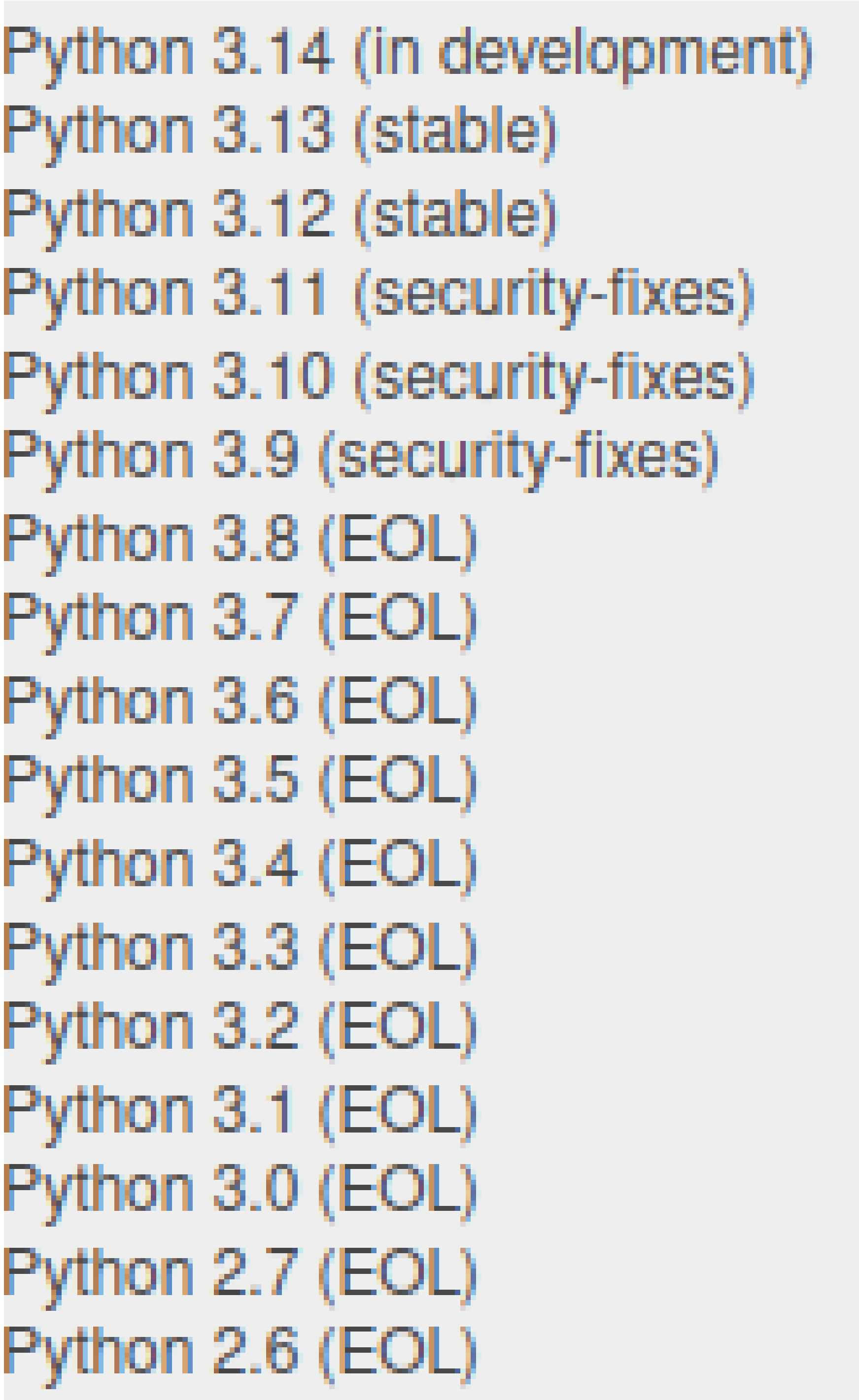
```
print “WELCOME TO GEEKSFORGEEKS”
```

It rounds your calculation down to the nearest whole number

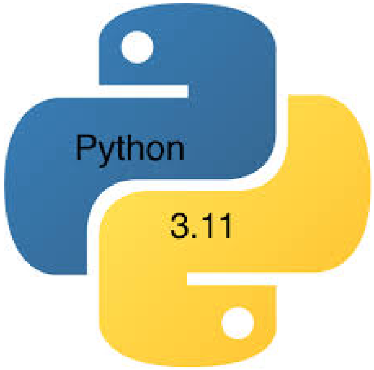
```
print(“WELCOME TO GEEKSFORGEEKS”)
```

This expression will result in the expected result

https://www.geeksforgeeks.org/python2-vs-python3-syntax-and-performance-comparison/



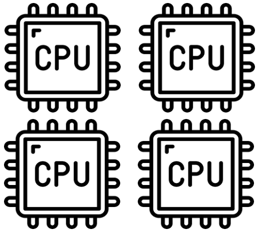
Python 3.14 (in development)
Python 3.13 (stable)
Python 3.12 (stable)
Python 3.11 (security-fixes)
Python 3.10 (security-fixes)
Python 3.9 (security-fixes)
Python 3.8 (EOL)
Python 3.7 (EOL)
Python 3.6 (EOL)
Python 3.5 (EOL)
Python 3.4 (EOL)
Python 3.3 (EOL)
Python 3.2 (EOL)
Python 3.1 (EOL)
Python 3.0 (EOL)
Python 2.7 (EOL)
Python 2.6 (EOL)



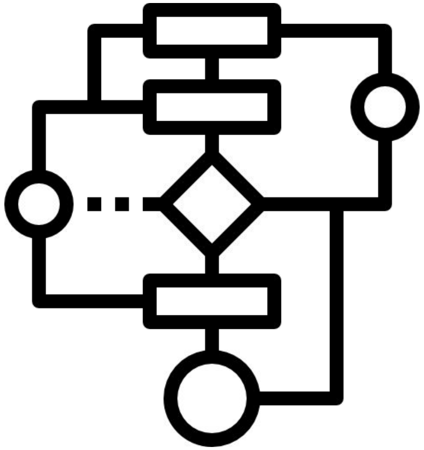
https://docs.python.org/3.13/
https://docs.python.org/3.12/whatsnew/3.12.html
https://docs.python.org/3.11/whatsnew/3.10.html#removed

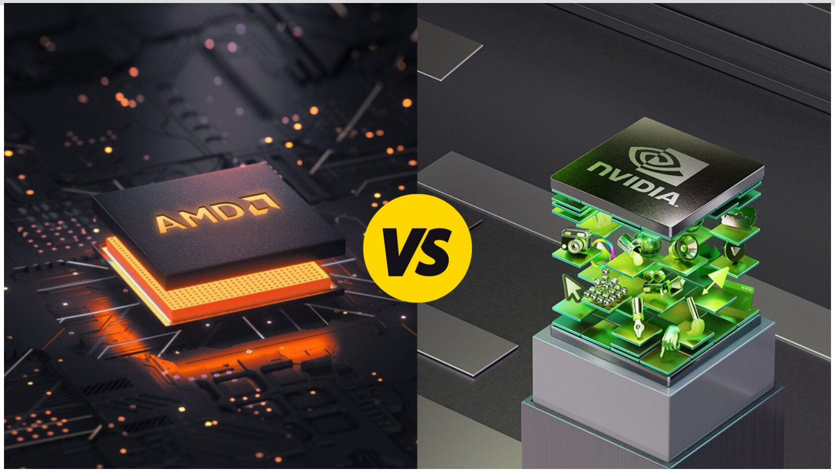
NLTK
Spark
theano
Numpy
K
Keras
TensorFlow
Pytorch
scikit learn
Scikit-learn
Pandas
mxnet

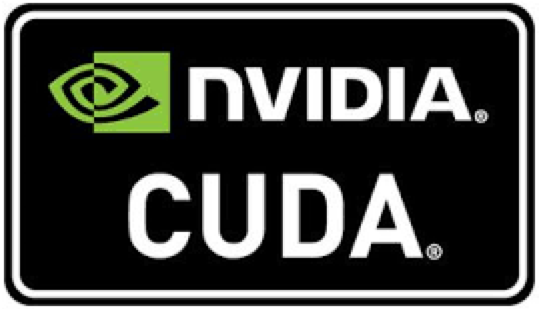
CPUs vs GPUs speed



100 CPUs = 1 GPU

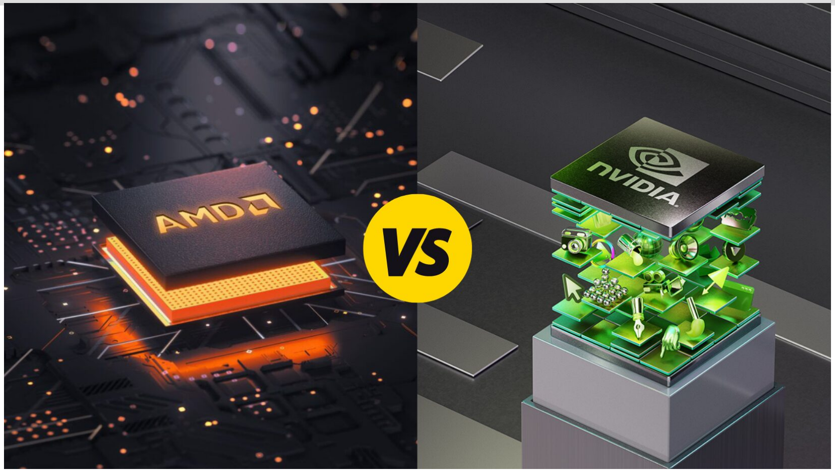


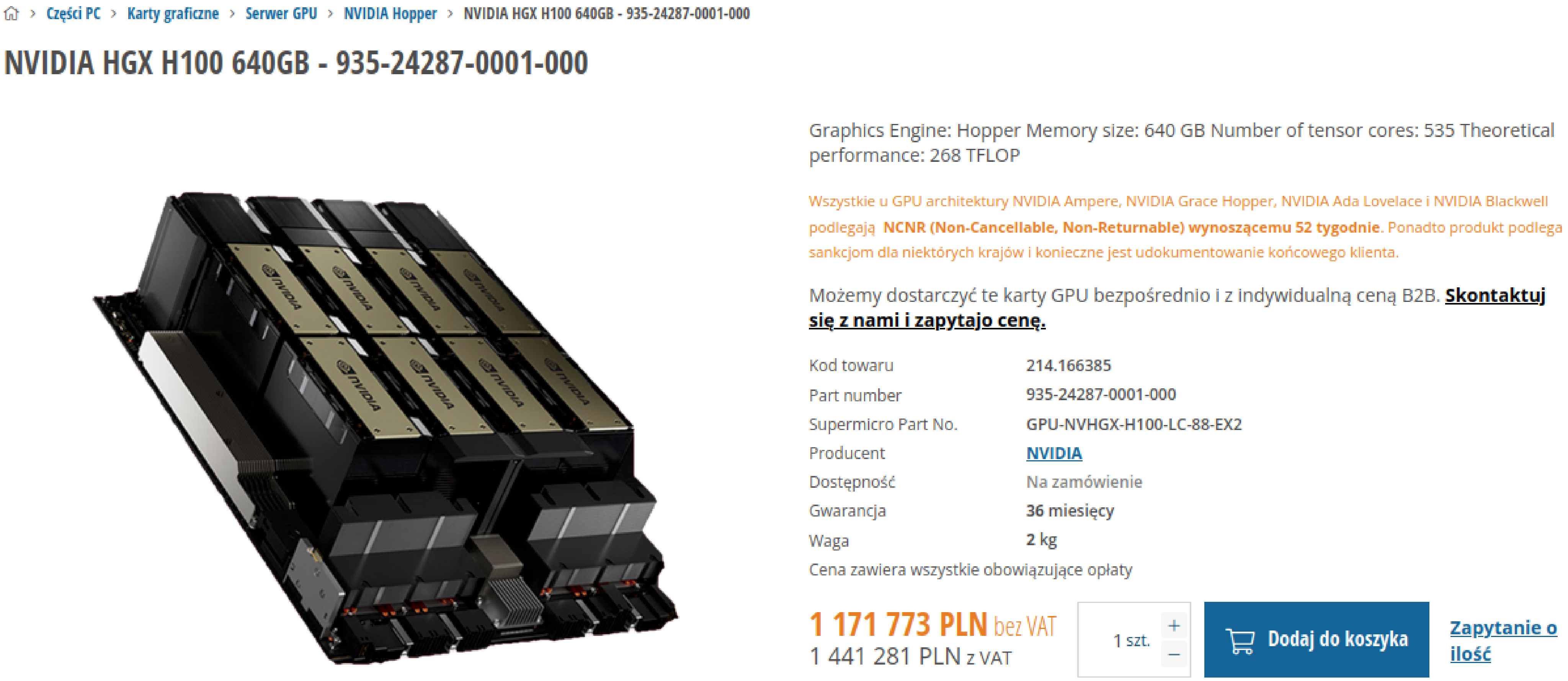




ROCm vs CUDA

https://thescimus.com/blog/rocm-vs-cuda-a-practical-comparison-for-ai-developers/





Części PC > Karty graficzne > Serwer GPU > NVIDIA Hopper > NVIDIA HGX H100 640GB - 935-24287-0001-000

NVIDIA HGX H100 640GB - 935-24287-0001-000

Graphics Engine: Hopper Memory size: 640 GB Number of tensor cores: 535 Theoretical performance: 268 TFLOP

Wszystkie u GPU architektury NVIDIA Ampere, NVIDIA Grace Hopper, NVIDIA Ada Lovelace i NVIDIA Blackwell podlegają **NCNR (Non-Cancellable, Non-Returnable) wynoszącemu 52 tygodnie**. Ponadto produkt podlega sankcjom dla niektórych krajów i konieczne jest udokumentowanie końcowego klienta.

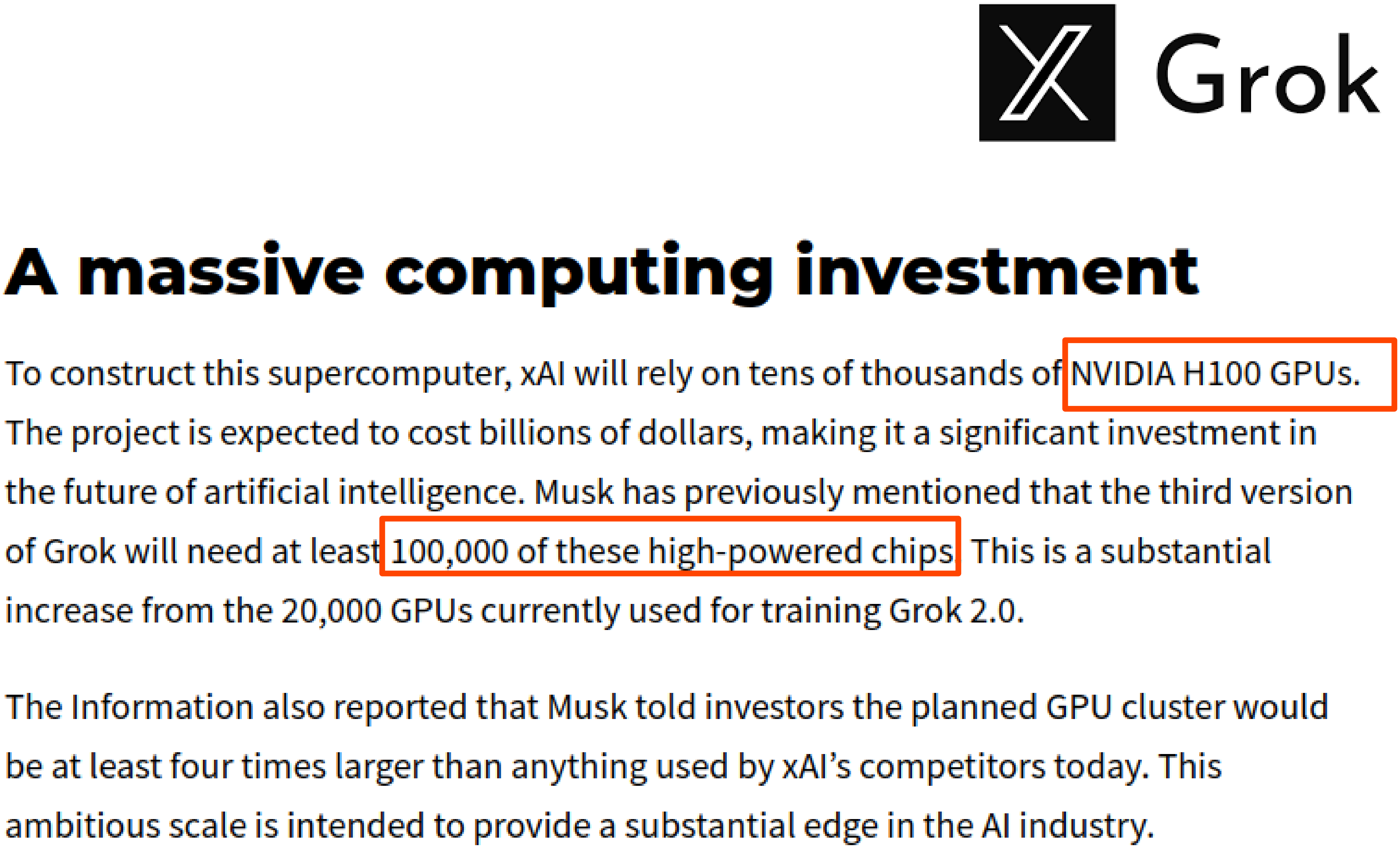
Możemy dostarczyć te karty GPU bezpośrednio i z indywidualną ceną B2B. **Skontaktuj się z nami i zapytajo cenę.**

Kod towaru	214.166385
Part number	935-24287-0001-000
Supermicro Part No.	GPU-NVHGX-H100-LC-88-EX2
Producent	NVIDIA
Dostępność	Na zamówienie
Gwarancja	36 miesięcy
Waga	2 kg

Cena zawiera wszystkie obowiązujące opłaty

1 171 773 PLN bez VAT
1 441 281 PLN z VAT

OpenAI
Meta AI
AI
Gemini
Grok

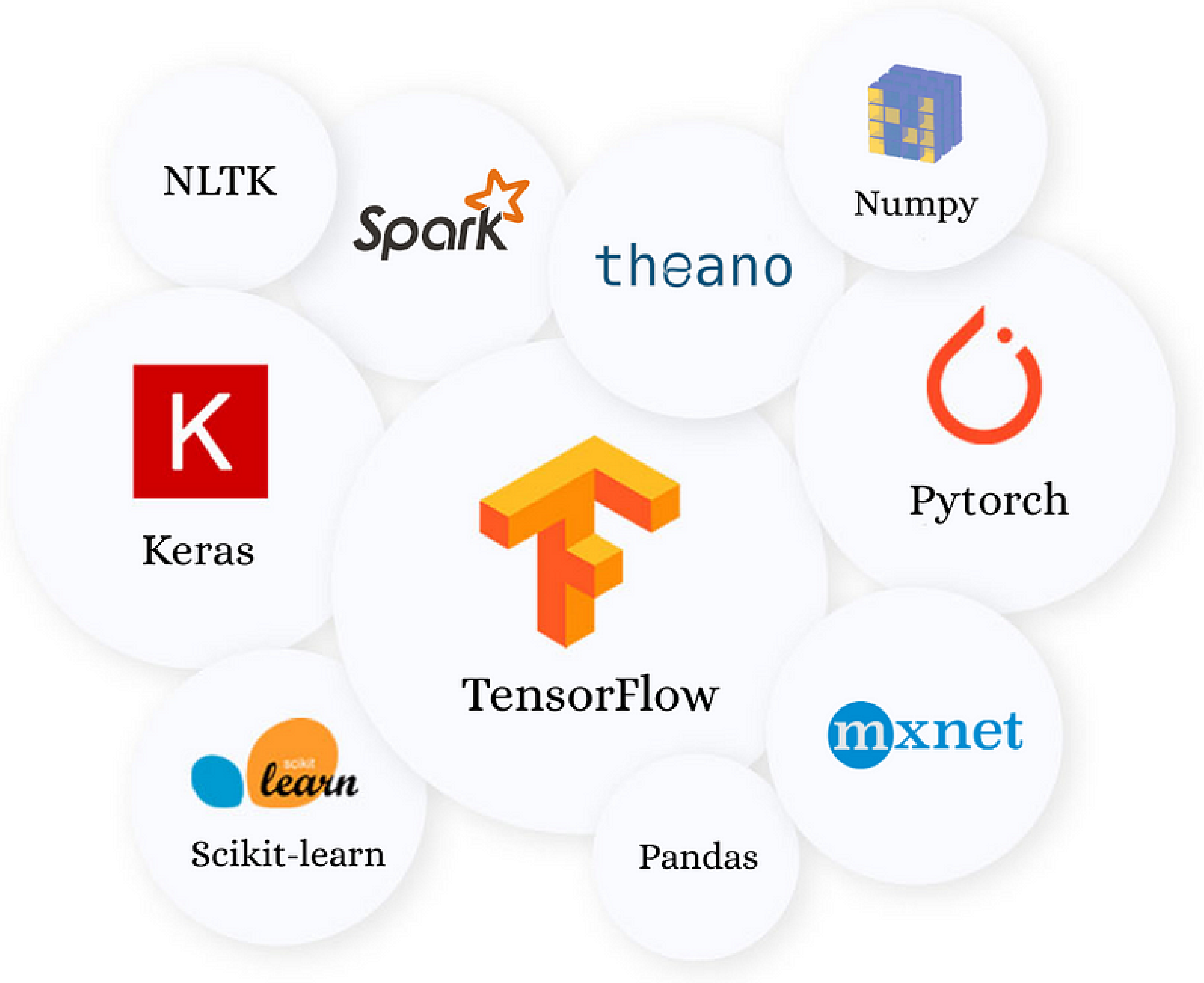


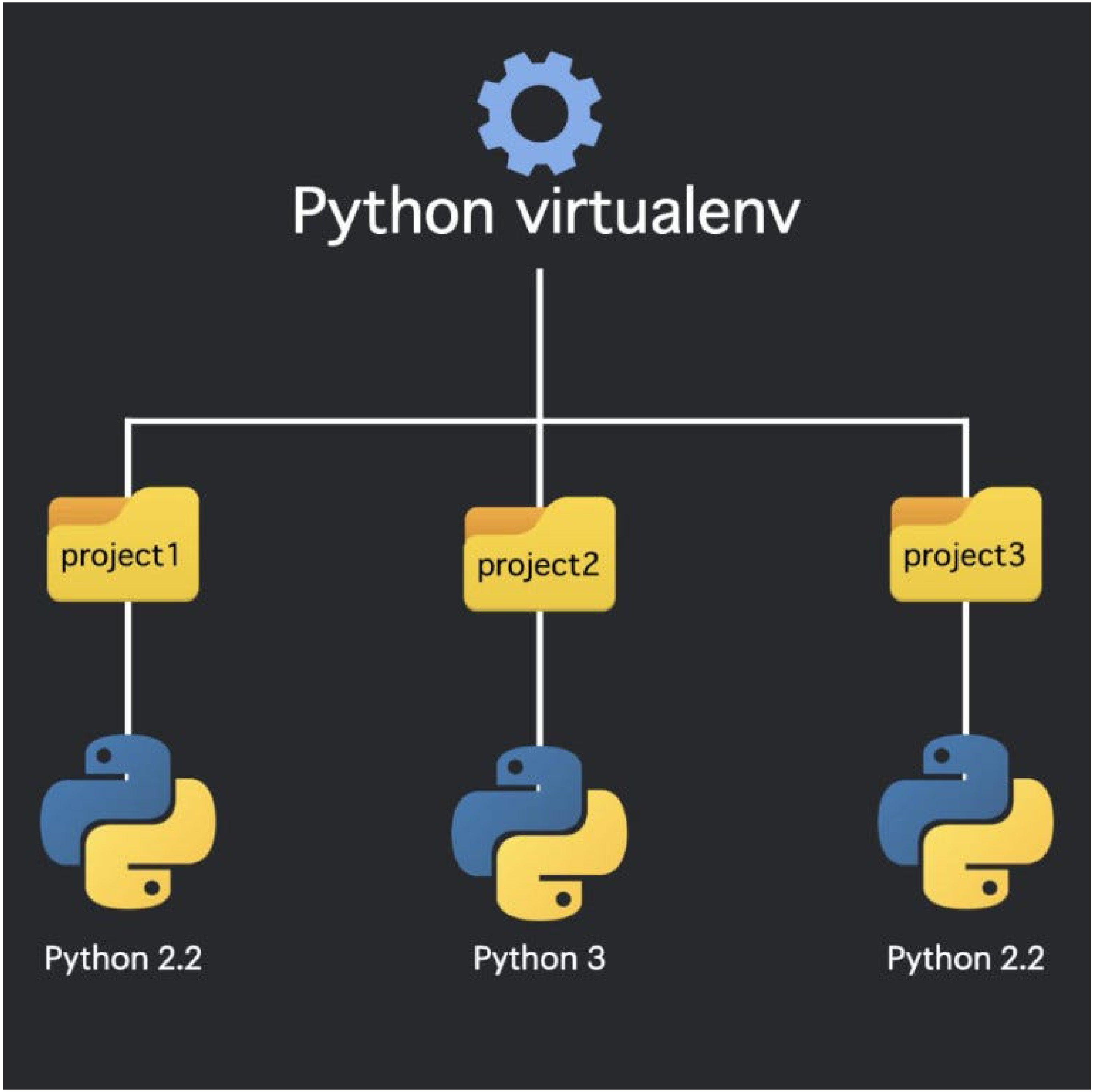
A massive computing investment

To construct this supercomputer, xAI will rely on tens of thousands of NVIDIA H100 GPUs. The project is expected to cost billions of dollars, making it a significant investment in the future of artificial intelligence. Musk has previously mentioned that the third version of Grok will need at least 100,000 of these high-powered chips. This is a substantial increase from the 20,000 GPUs currently used for training Grok 2.0.

The Information also reported that Musk told investors the planned GPU cluster would be at least four times larger than anything used by xAI's competitors today. This ambitious scale is intended to provide a substantial edge in the AI industry.

https://www.techedt.com/elon-musk-plans-a-new-xai-supercomputer-to-enhance-grok


NLTK
Spark
theano
Numpy
Keras
Pytorch
TensorFlow
mxnet
scikit learn
Scikit-learn
Pandas



docker

ANACONDA®

jupyter

We will use (and compare) different formats such as general ones (csv, xml, json, etc.) and domain specific

This means that you will need to write custom parsers* as well.

We will use (and compare) different formats such as general ones (csv, xml, json, etc.) and domain specific

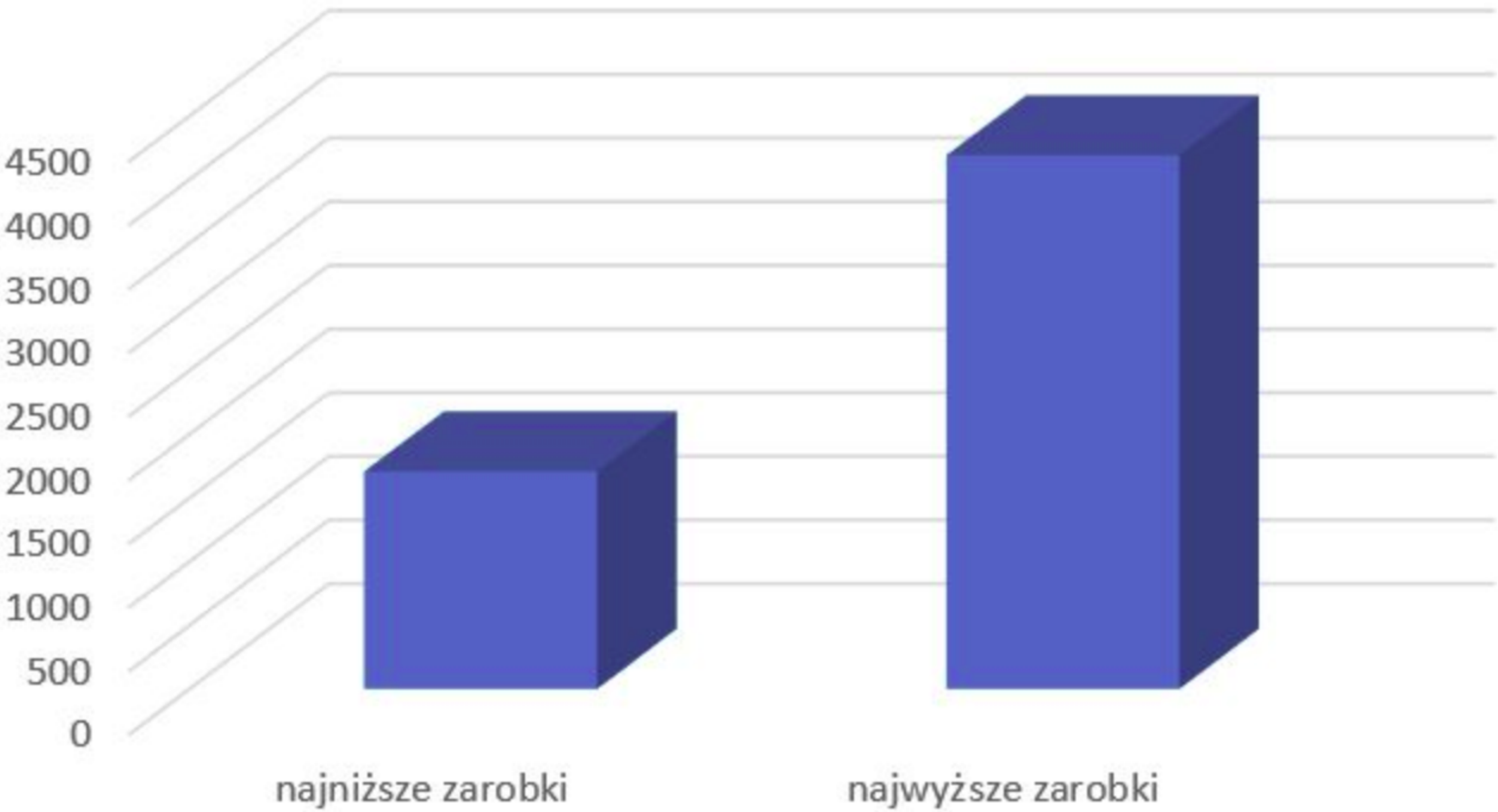
This means that you will need to write custom parsers* as well.

* during labs frequently you will be asked to implement simple functionalities even if some libraries are available

BAD PLOTS

Really bad plots

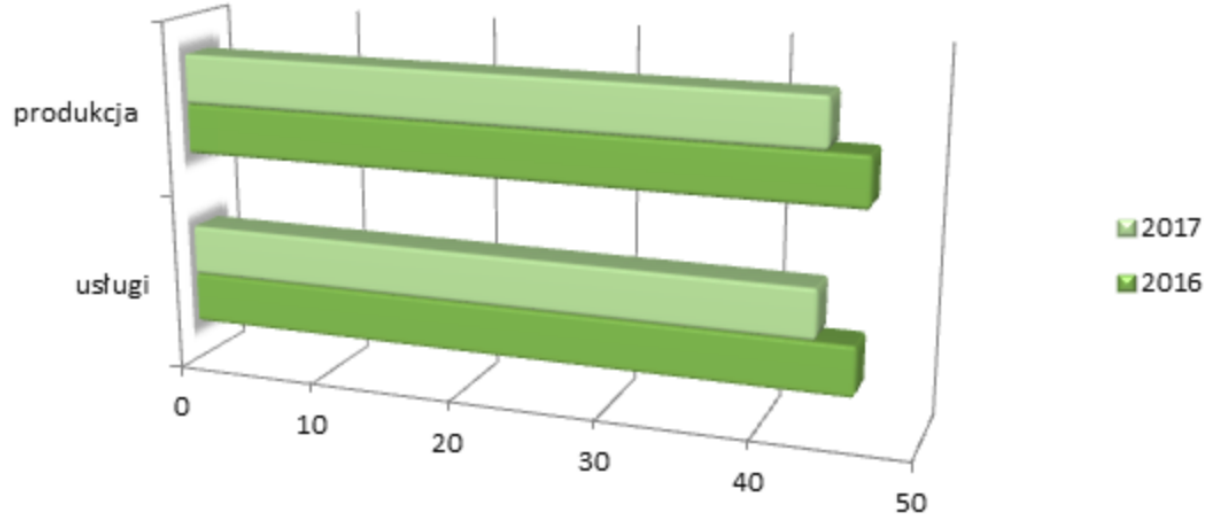
Wykres 2. *Zróżnicowanie dochodów najniższych i najwyższych wśród ankieterów*



Źródło:badania własne

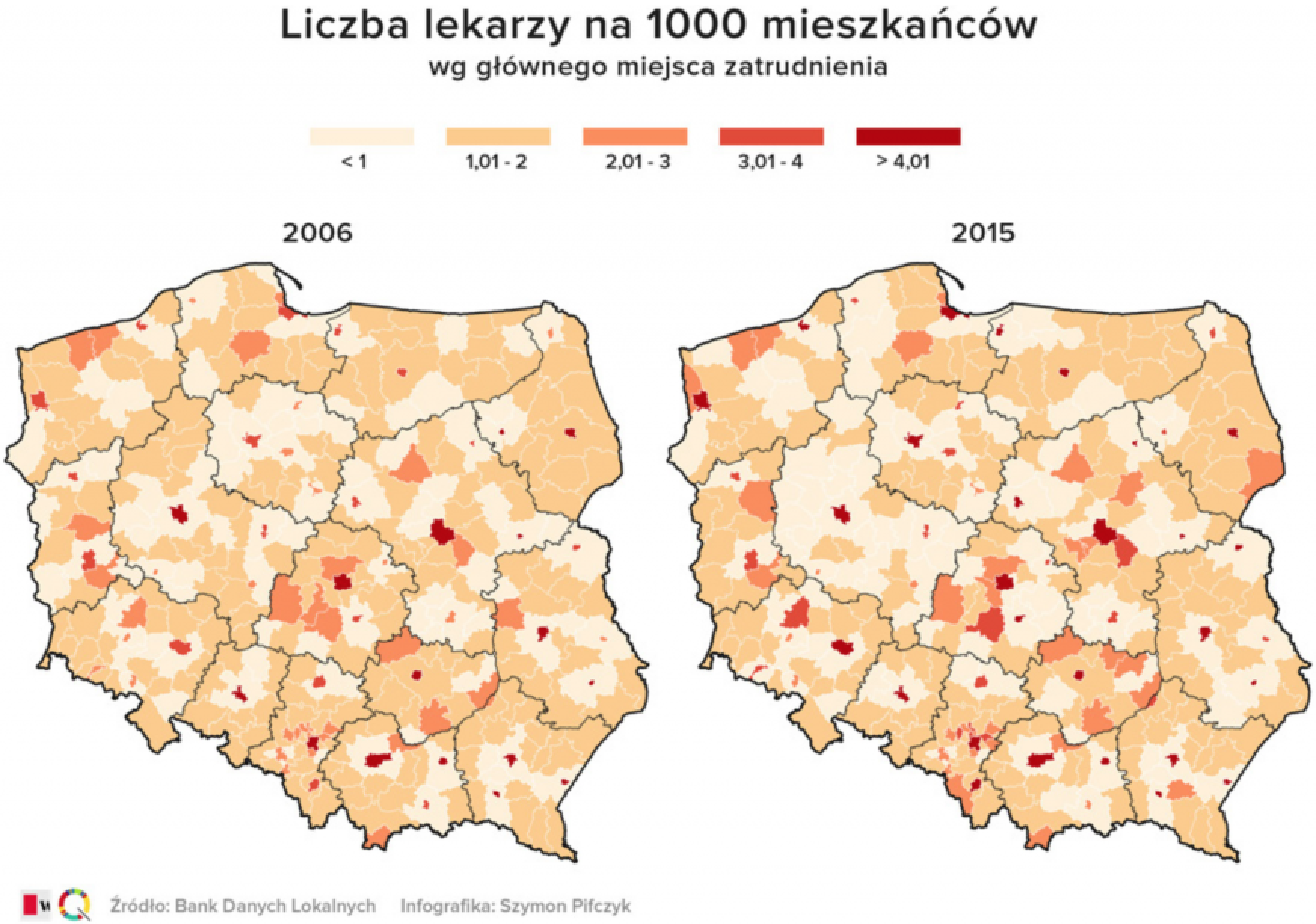
https://www.pielegniarkicyfrowe.pl/2018/07/08/zarobki-pielegniarek-i-poloznych-raport-stowarzyszenia/

Wykres 2. Absencja ogólna (w tym urlopy) w przedsiębiorstwach produkcyjnych i usługowych (w dniach)



Źródło: Raport WskaźnikiHR 2017 i 2018

https://rynekpracy.pl/artykuly/podsumowanie-raportu-wskaznikihr-2018



Liczba lekarzy na 1000 mieszkańców

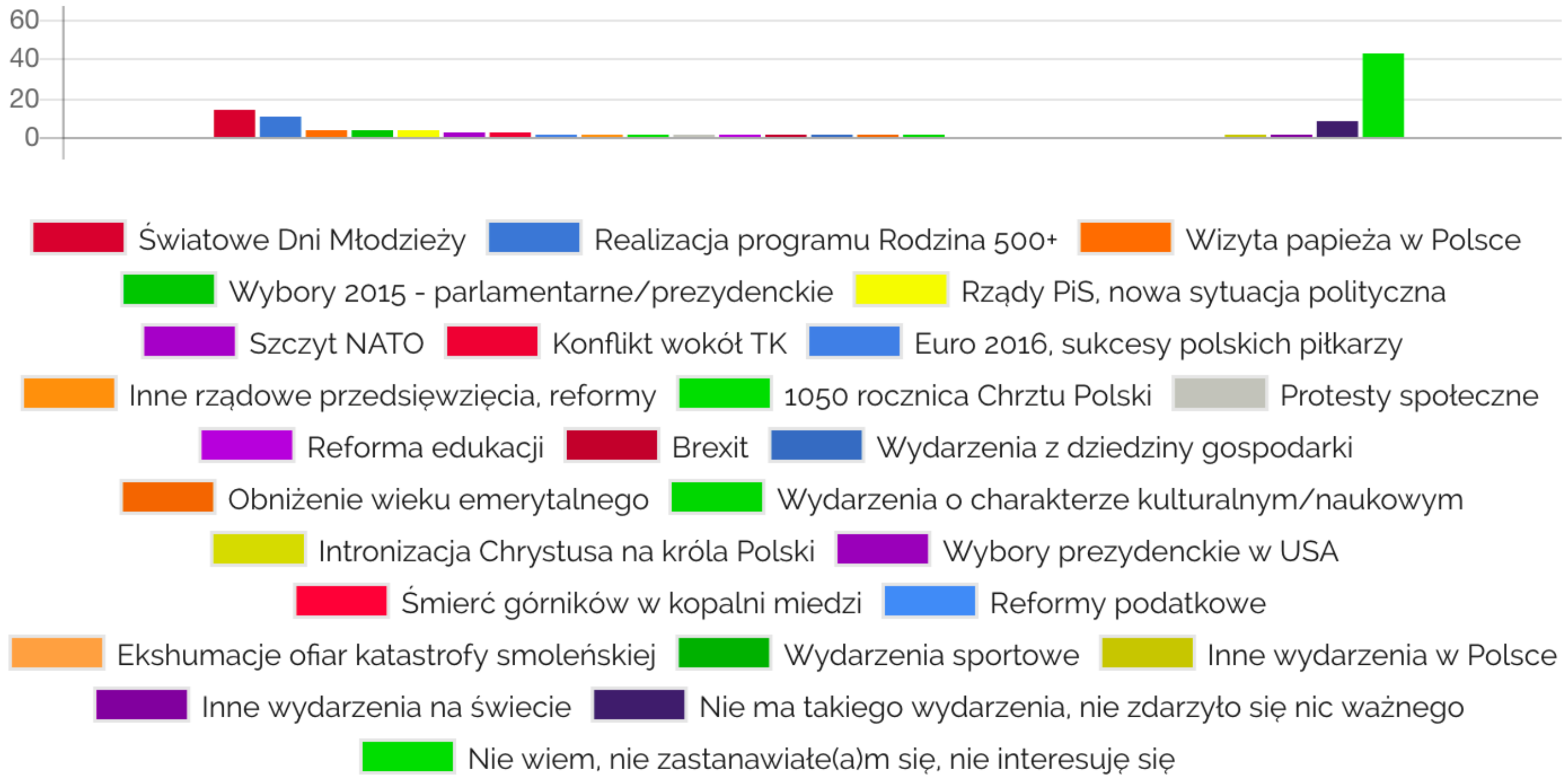
wg głównego miejsca zatrudnienia

Źródło: Bank Danych Lokalnych Infografika: Szymon Pifczyk

http://biqdata.wyborcza.pl/dlaczego-medycy-protestuja-bo-system-jest-na-skraju-wyczerpania

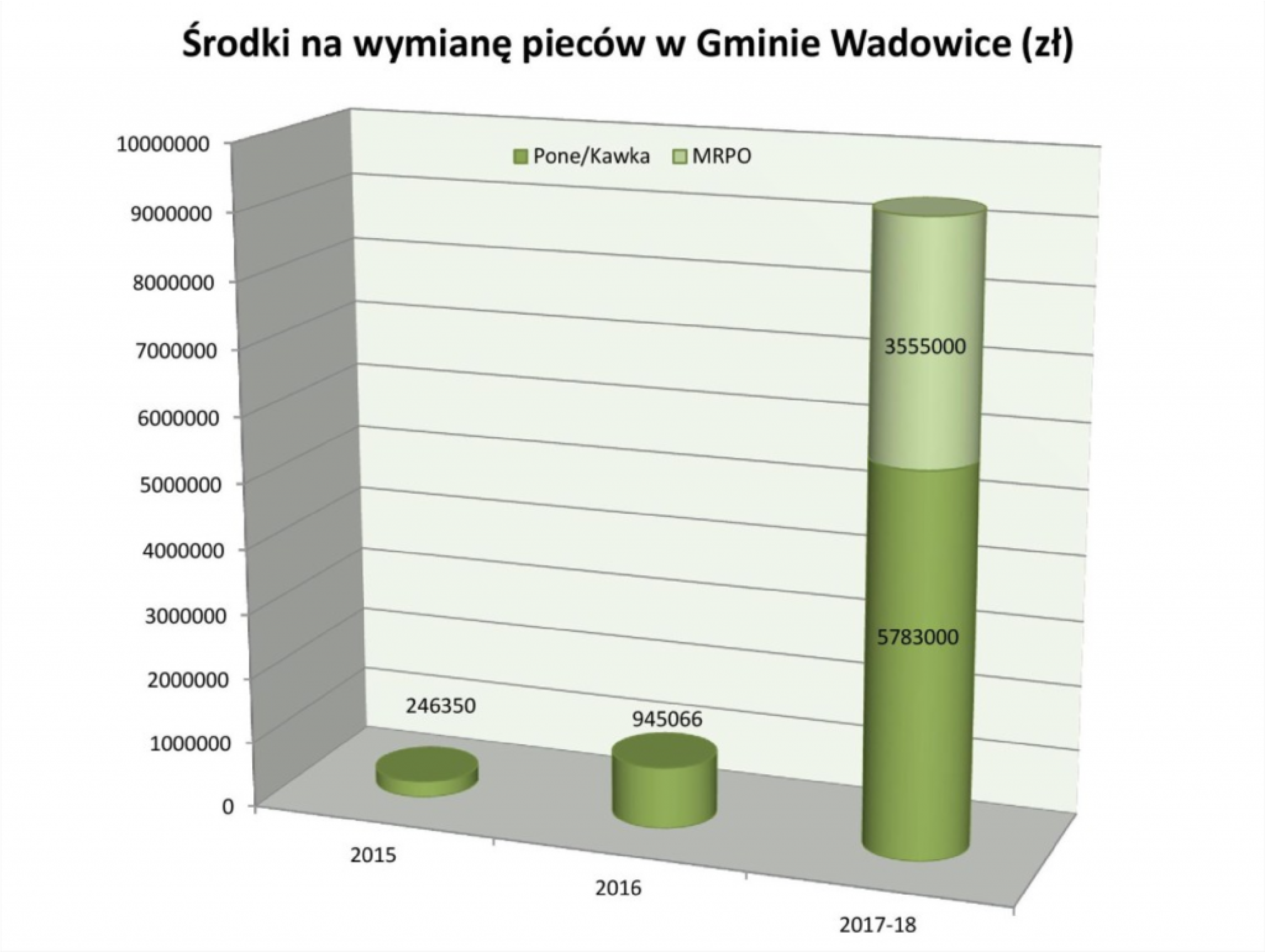
KTÓRE Z WYDARZEŃ MIJAJĄCEGO ROKU MOŻNA, PANA(I) ZDANIEM UZNAĆ ZA NAJWAŻNIEJSZE DLA POLSKI

Dane w proc. Pytanie miało charakter otwarty



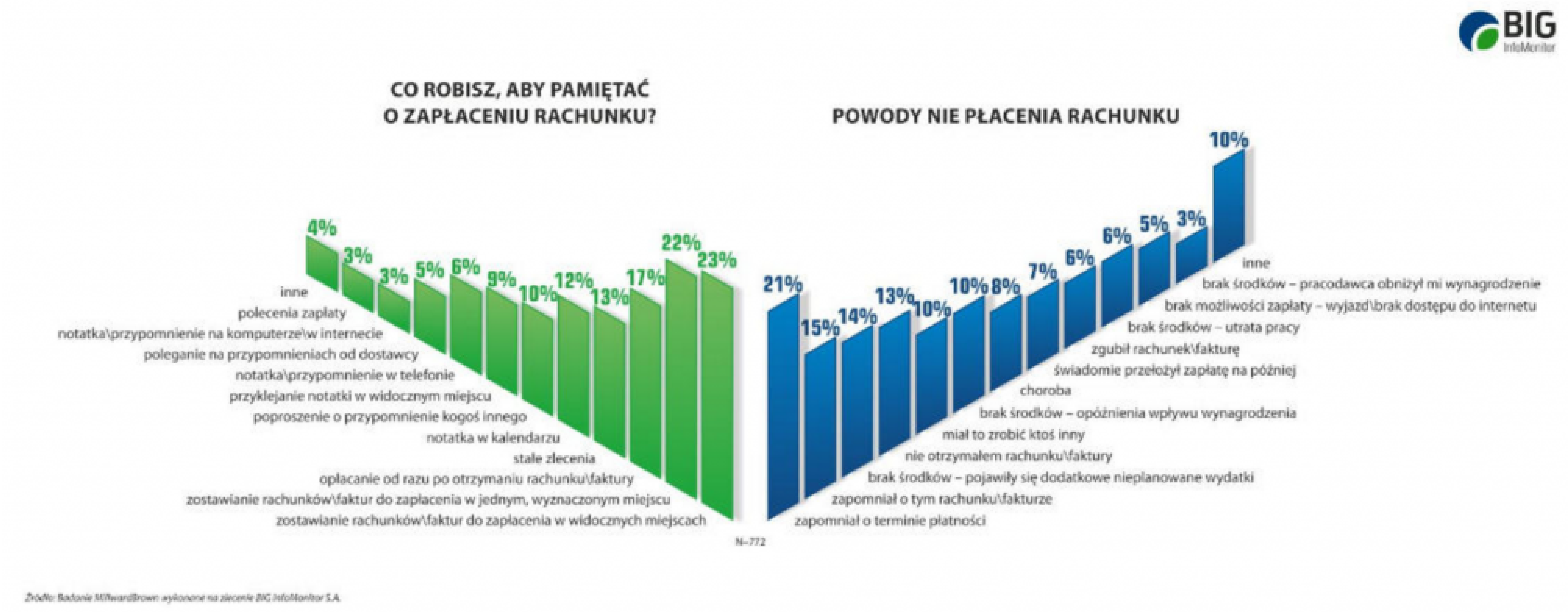
Na siódmym miejscu znalazł się konflikt wokół Trybunału Konstytucyjnego (2,1 proc.), na jedenastym protesty KOD i czarne marsze (1 proc.).

https://oko.press/bylo-najwazniejsze-2016-wedlug-polakow-sdm-500/



http://www.burmistrzklinowski.pl/2017/01/o-przyczynach-smogu/

Really bad plots

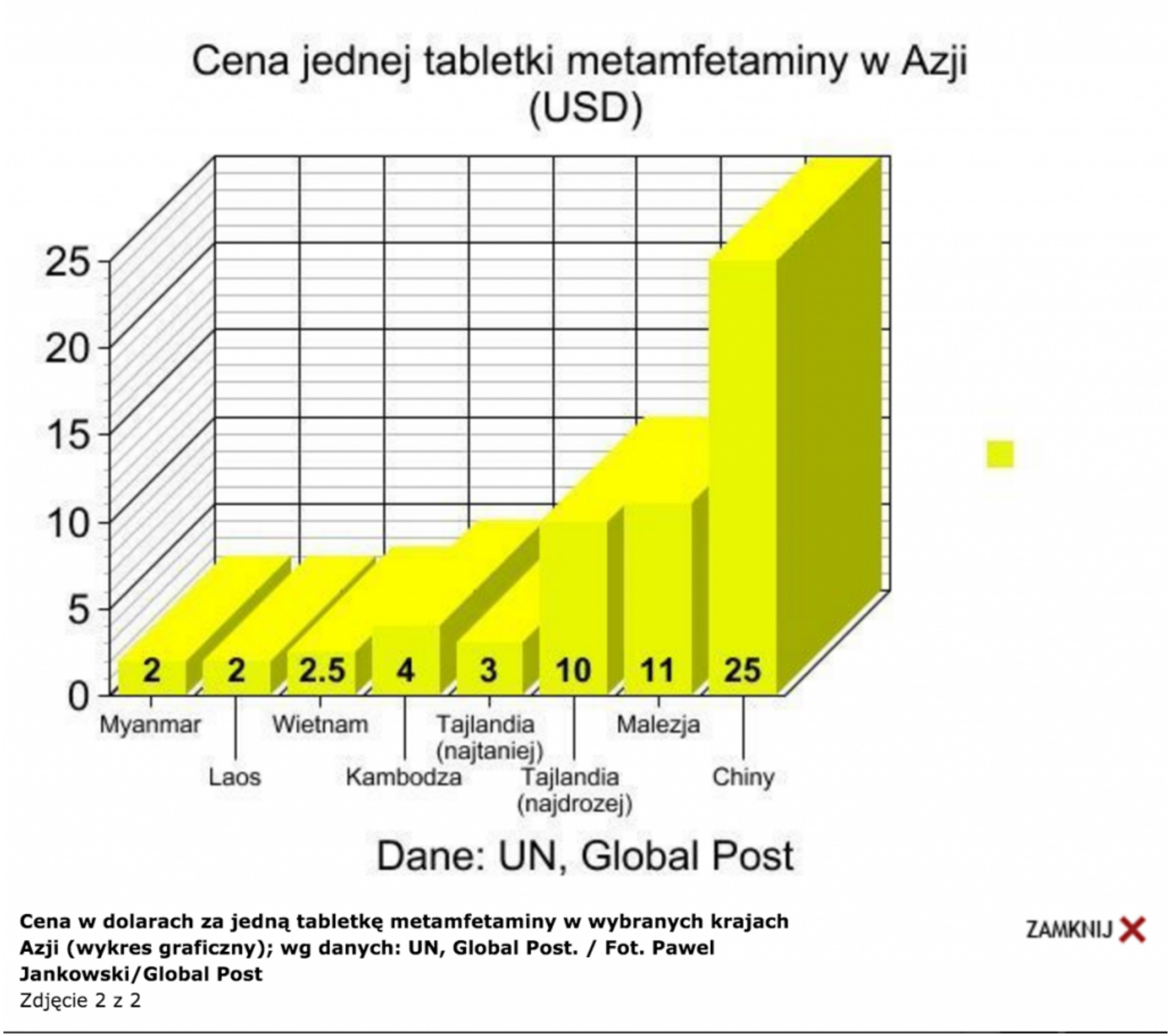


CO ROBISZ, ABY PAMIĘTAĆ O ZAPŁACENIU RACHUNKU?

POWODY NIE PŁACENIA RACHUNKU

Źródło: Badanie MillwardBrown wykonane na zlecenie BIG InfoMonitor S.A.

http://www.przeciwprzemocy.pl/big-infomonitor-800-tys-osob-nie-zaplacilo-rachunkow-na-3-mld-zl/

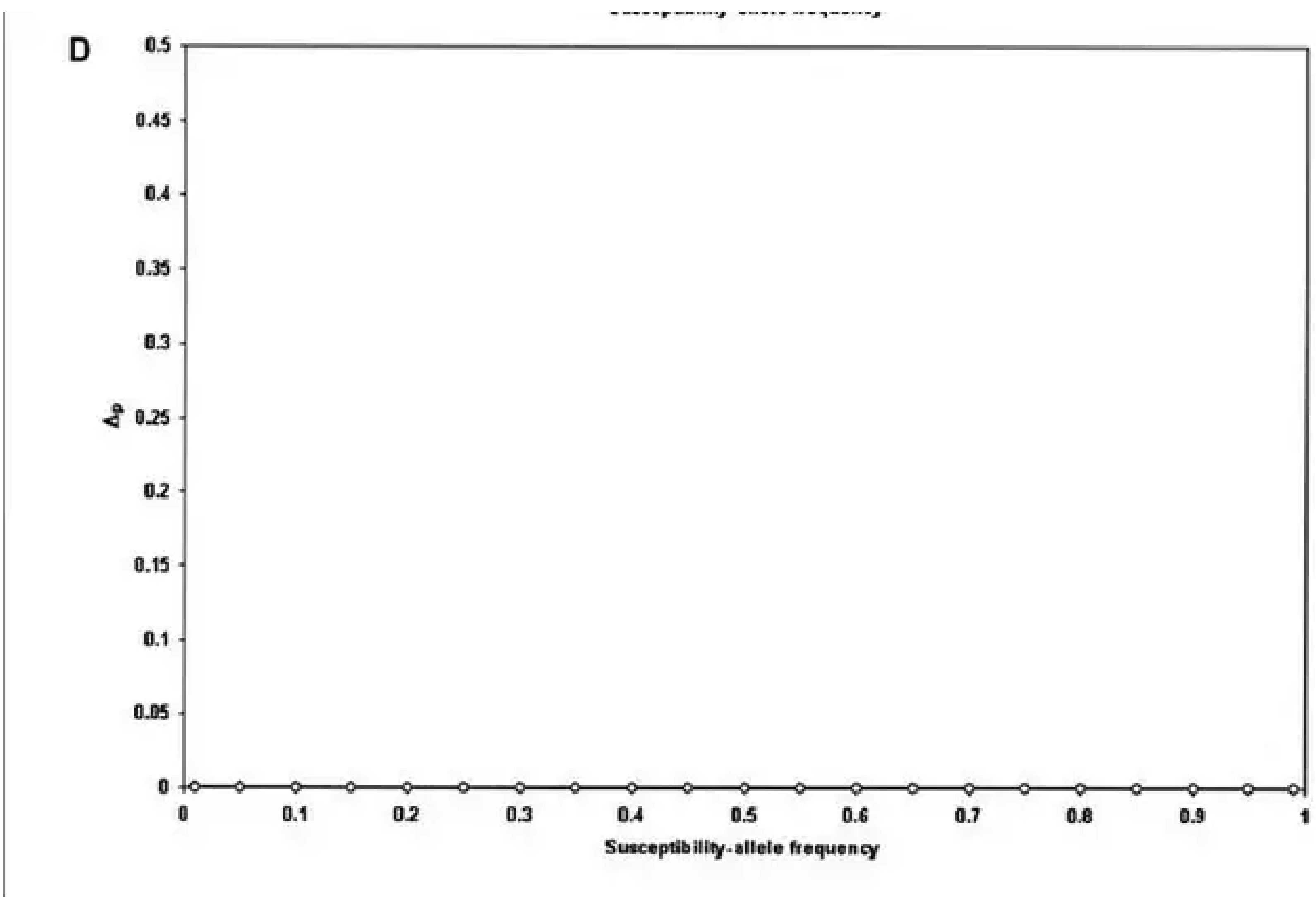


Cena w dolarach za jedną tabletkę metamfetaminy w wybranych krajach Azji (wykres graficzny); wg danych: UN, Global Post. / Fot. Pawel Jankowski/Global Post

Zdjęcie 2 z 2

http://www.wiadomosci24.pl/artykul/ceny_metamfetaminy_w_krajach_azji_sa_nizsze_od_hamburgerow_331503.html

Really bad plots



Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. American Journal of Human Genetics 76:967-986, Figure 1

Really bad plots

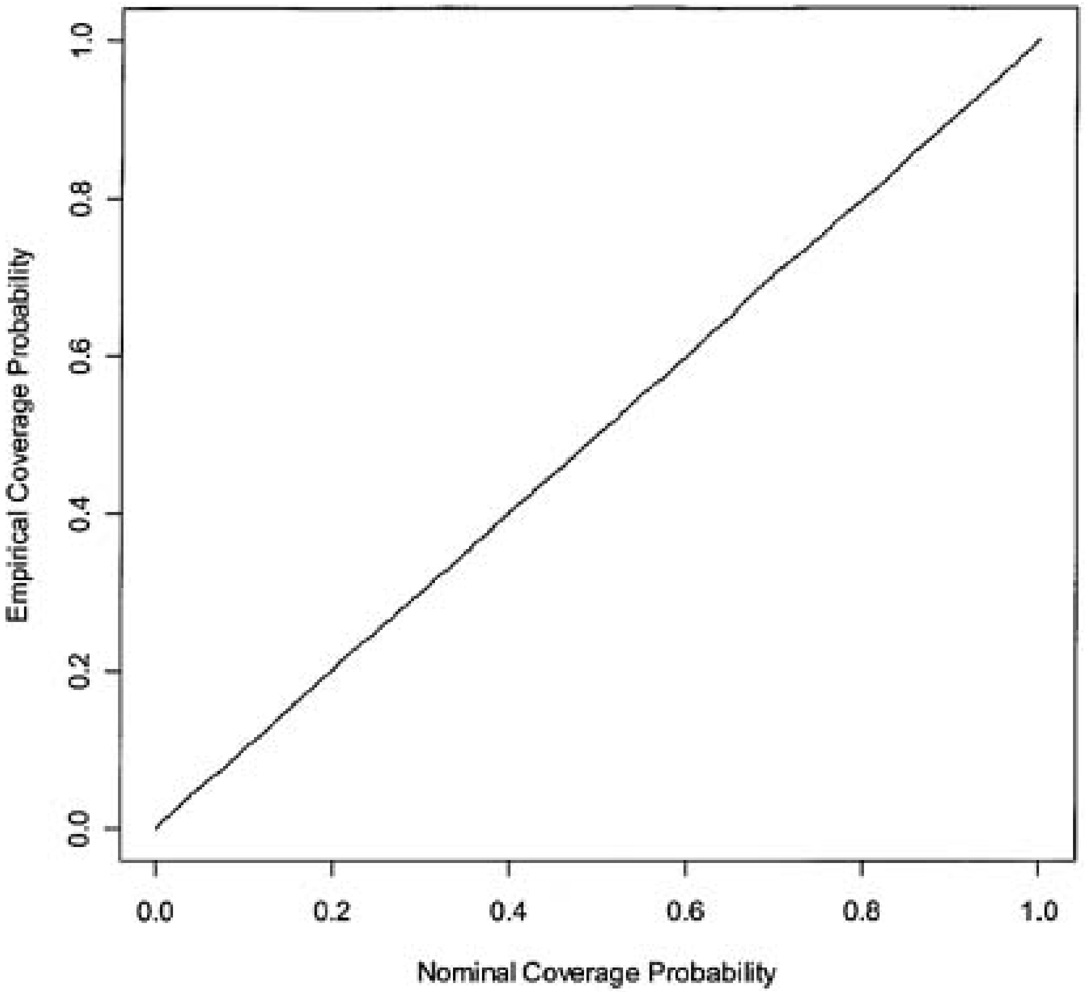
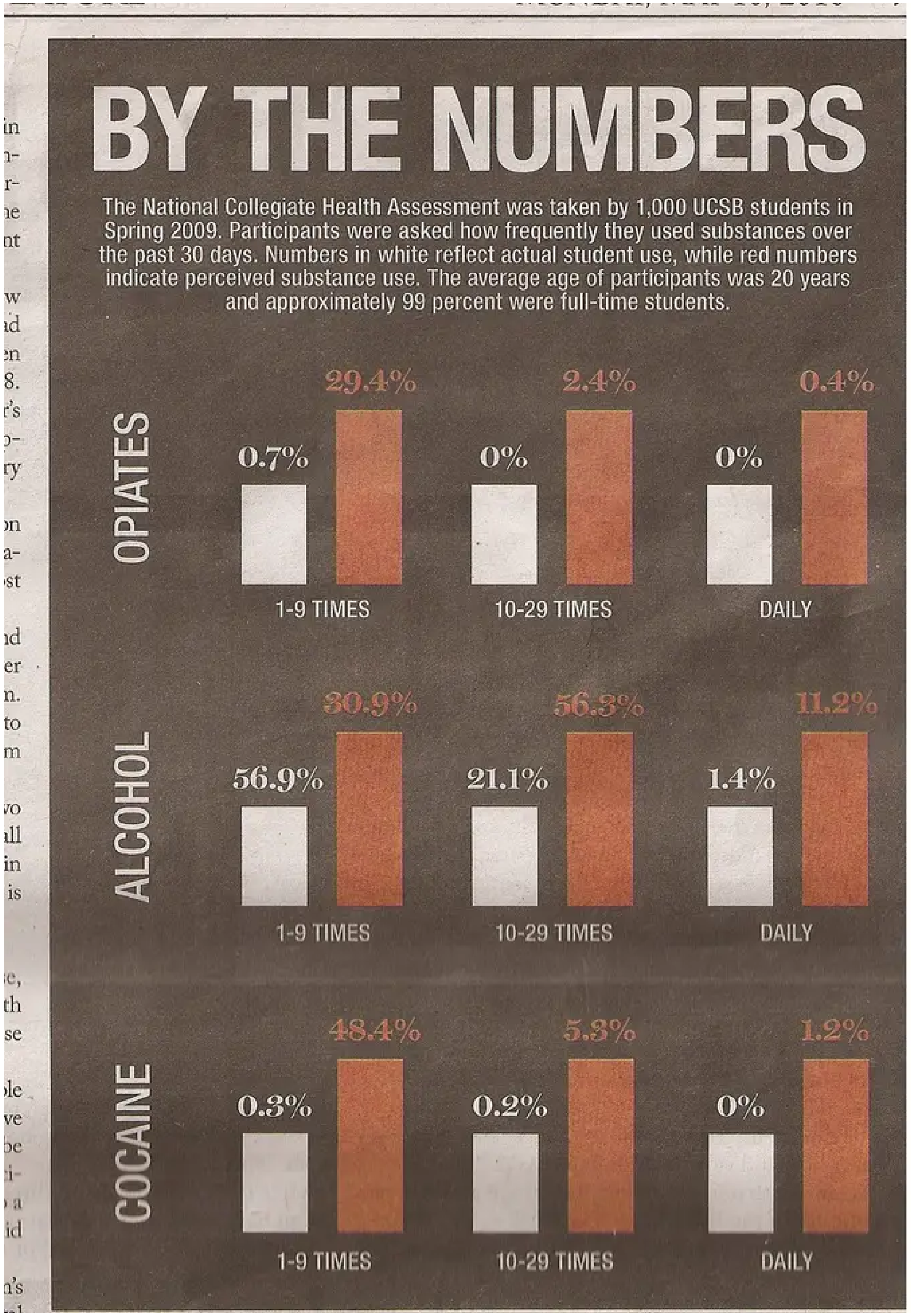


Figure 1 Empirical coverage of CIs for the relative-risk parameter β of haplotype 01100. Results are based on 10,000 simulated data sets with the same haplotype frequencies as the FUSION data. Haplotype 01100 has a multiplicative effect on disease risk, with $\beta = 0.35$.

Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. American Journal of Human Genetics 73:1316-1329, Figure 1

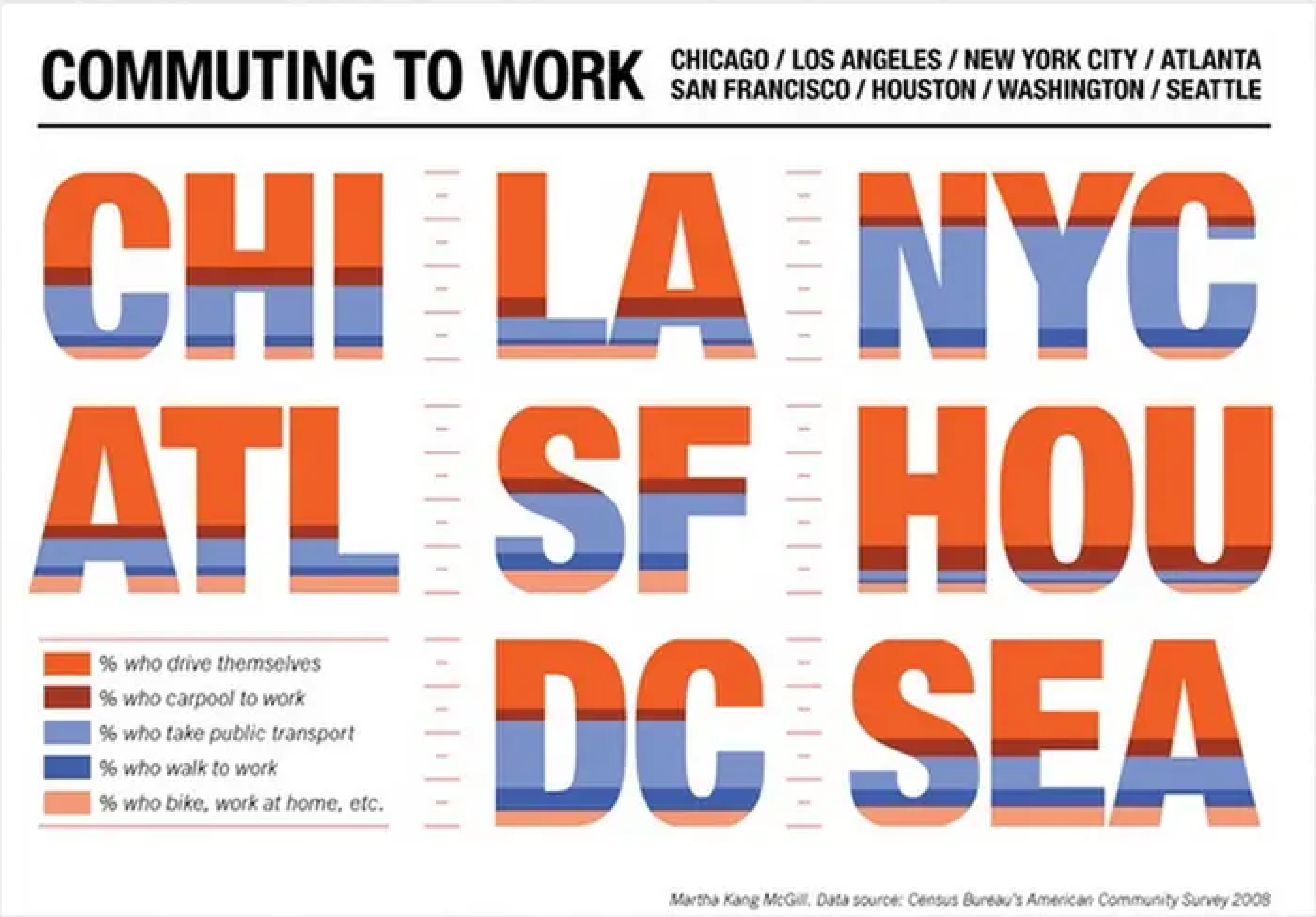


BY THE NUMBERS

The National Collegiate Health Assessment was taken by 1,000 UCSB students in Spring 2009. Participants were asked how frequently they used substances over the past 30 days. Numbers in white reflect actual student use, while red numbers indicate perceived substance use. The average age of participants was 20 years and approximately 99 percent were full-time students.

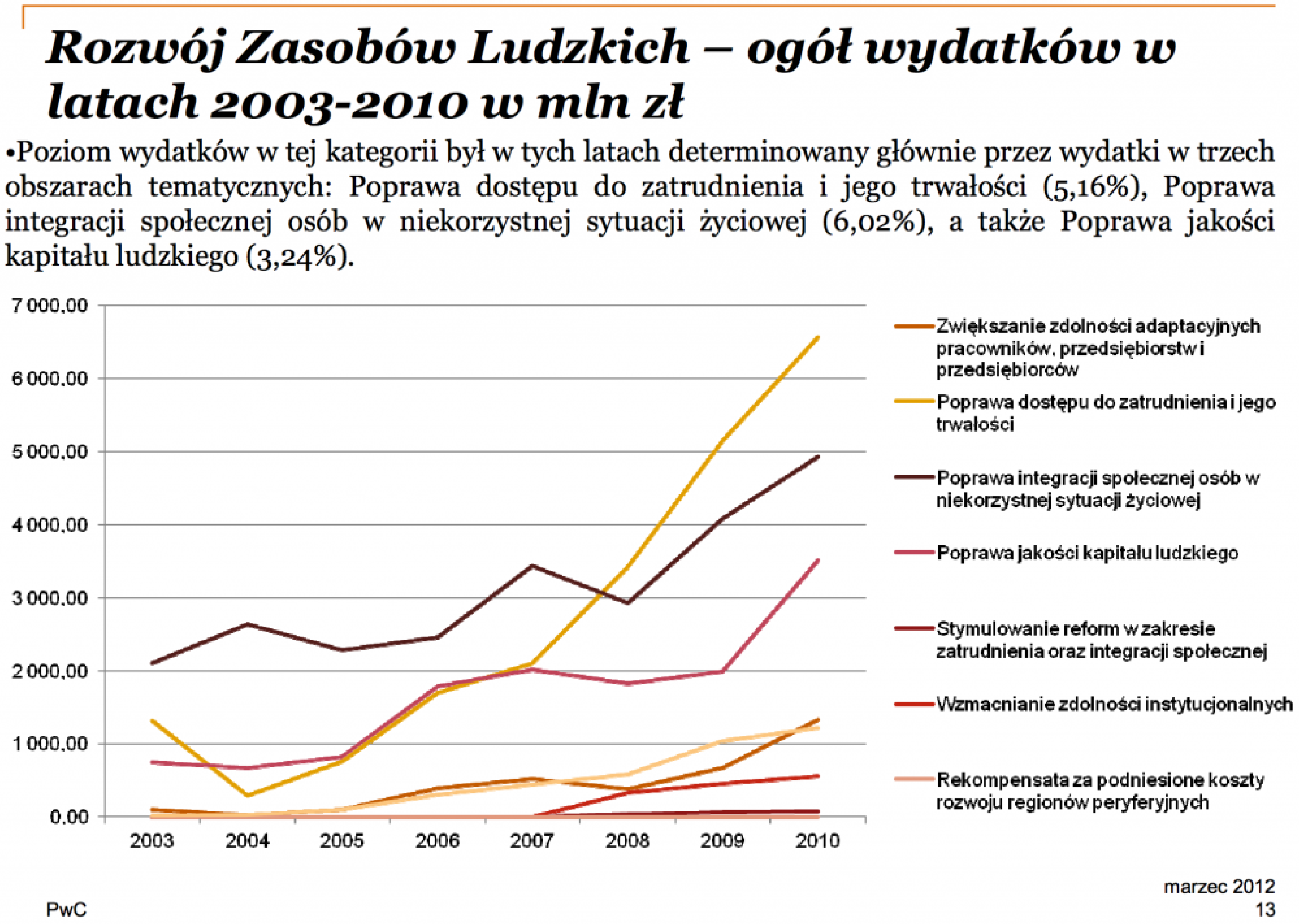
https://twitter.com/EagerEyes/status/13821850078

Really bad plots



http://eagereyes.org/criticism/march-chart-madness

Really bad plots



Rozwój Zasobów Ludzkich – ogół wydatków w latach 2003-2010 w mln zł

- Poziom wydatków w tej kategorii był w tych latach determinowany głównie przez wydatki w trzech obszarach tematycznych: Poprawa dostępu do zatrudnienia i jego trwałości (5,16%), Poprawa integracji społecznej osób w niekorzystnej sytuacji życiowej (6,02%), a także Poprawa jakości kapitału ludzkiego (3,24%).

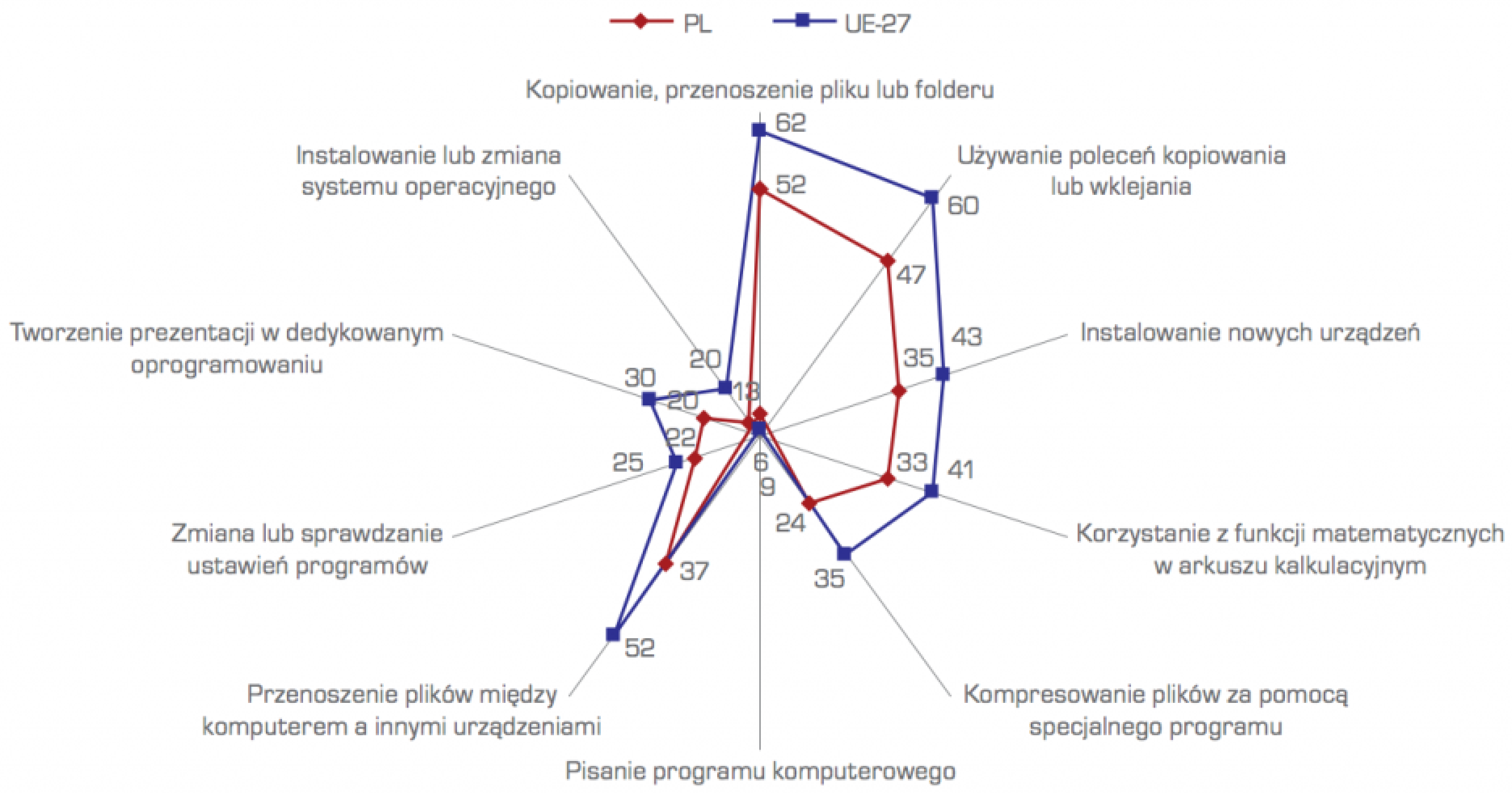
7 000,00
6 000,00
5 000,00
4 000,00
3 000,00
2 000,00
1 000,00
0,00
2003
2004
2005
2006
2007
2008
2009
2010
Zwiększanie zdolności adaptacyjnych pracowników, przedsiębiorstw i przedsiębiorców
Poprawa dostępu do zatrudnienia i jego trwałości
Poprawa integracji społecznej osób w niekorzystnej sytuacji życiowej
Poprawa jakości kapitału ludzkiego
Stymulowanie reform w zakresie zatrudnienia oraz integracji społecznej
Wzmacnianie zdolności instytucjonalnych
Rekompensata za podniesione koszty rozwoju regionów peryferyjnych

PwC

marzec 2012
13

http://www.mir.gov.pl/rozwoj_regionalny/Ewaluacja_i_analizy/Raporty_o_rozwoju/Documents/Prezentacja_13032012.pdf

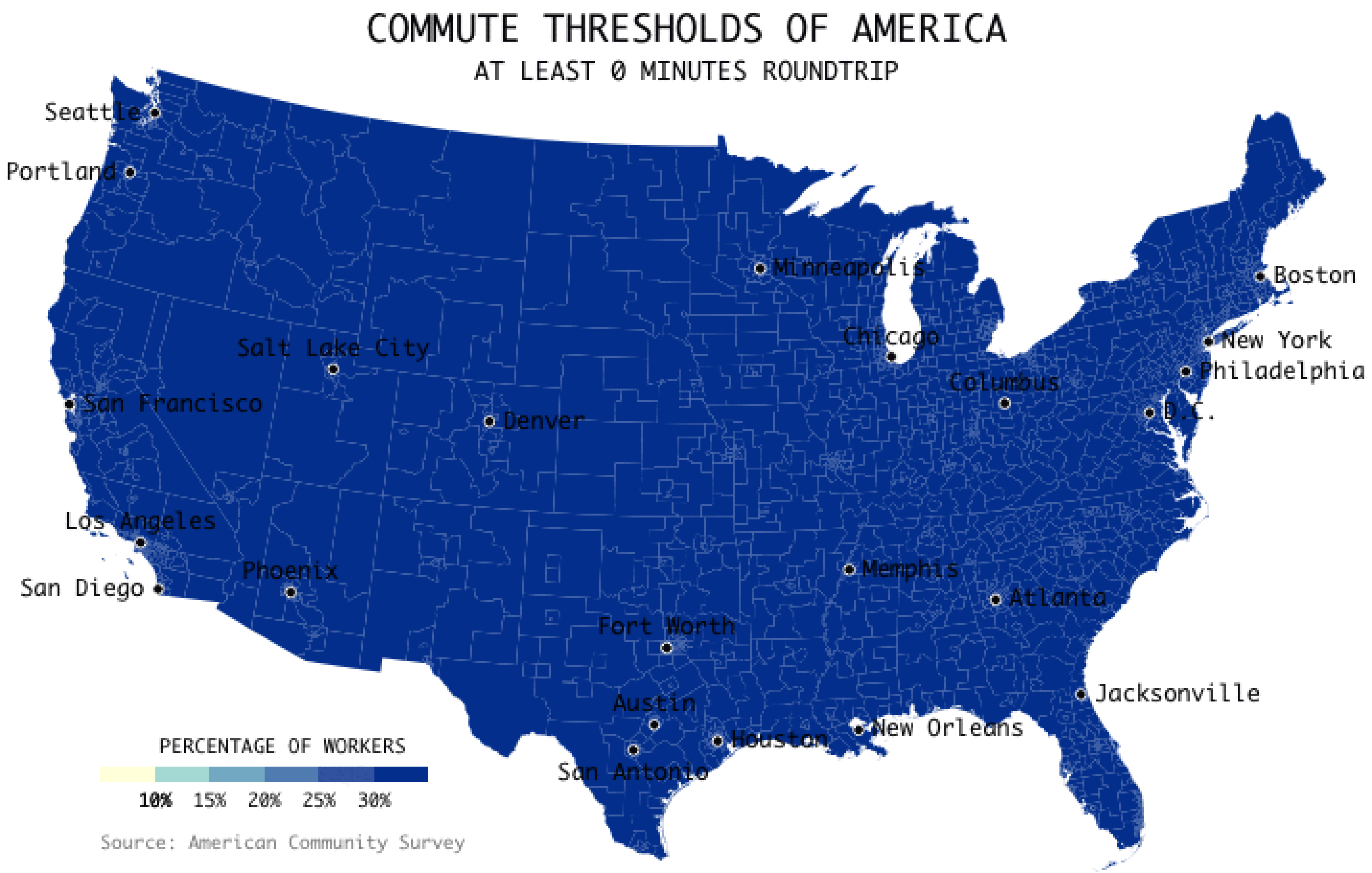
Wykres 8. Osoby wykonujące określone czynności związane z obsługą komputera w 2012 r. – Polska a średnia unijna (w %)



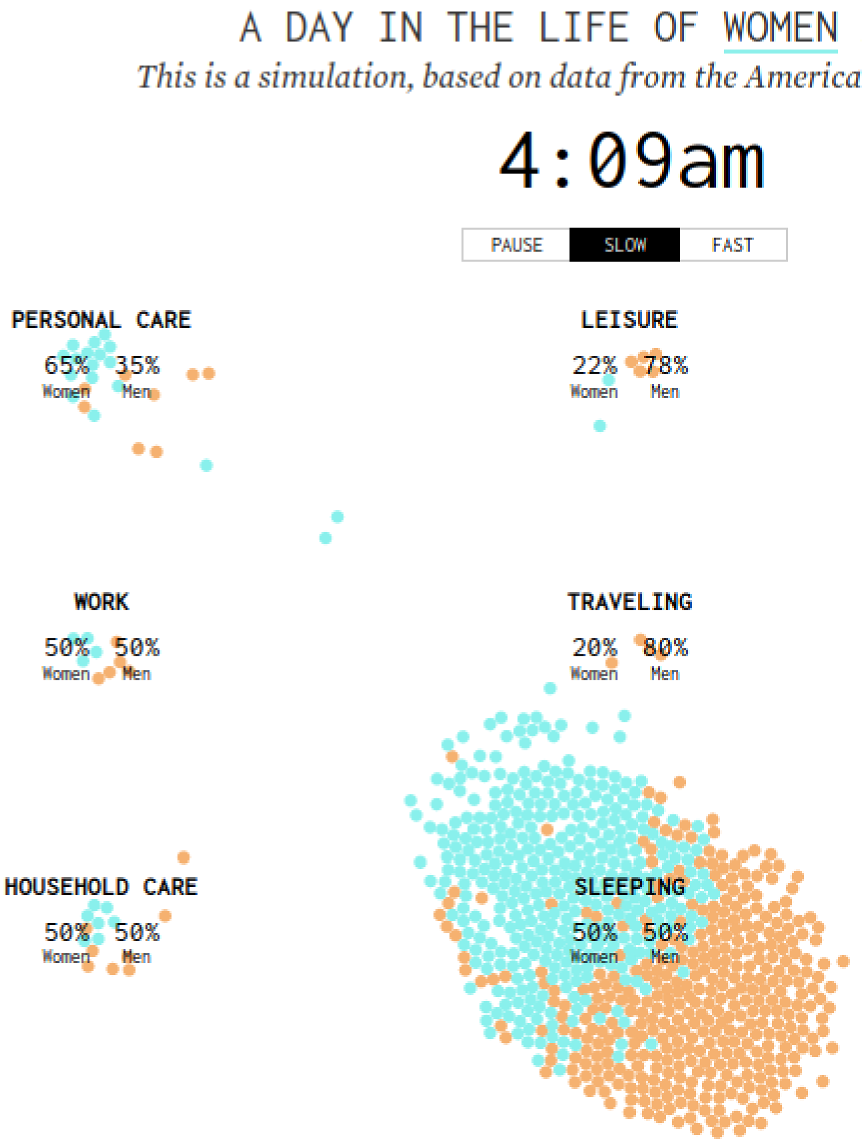
Opracowanie własne na podstawie danych Eurostatu.

,,Społeczeństwo informacyjne w liczbach” z MAC

HOW ABOUT NICE PLOTS ?



https://flowingdata.com/2019/10/16/how-much-commuting-is-too-much/



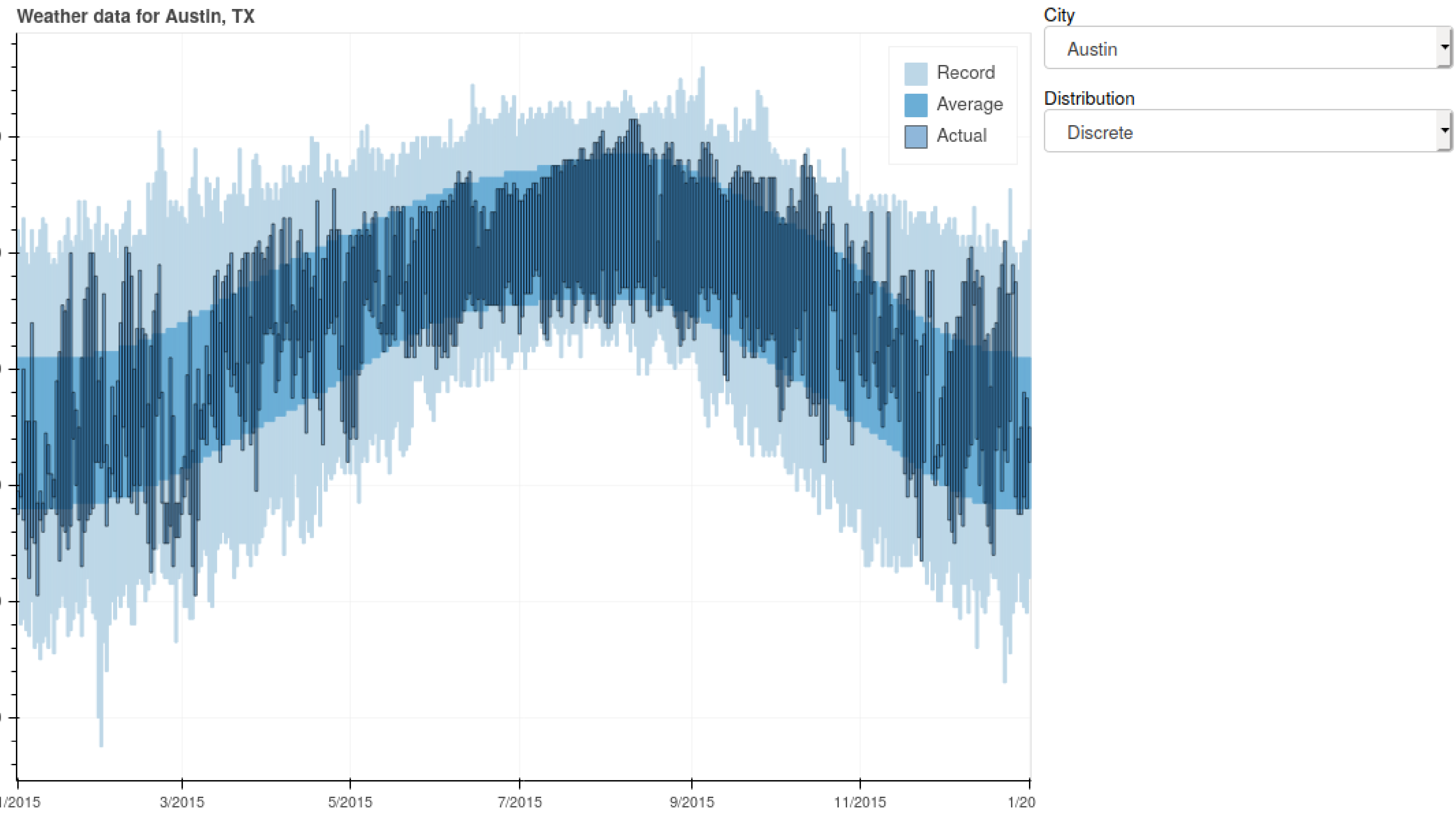
A DAY IN THE LIFE OF WOMEN AND MEN

This is a simulation, based on data from the American Time Use Survey.

4:09am

https://flowingdata.com/2019/03/06/women-men-timeuse/

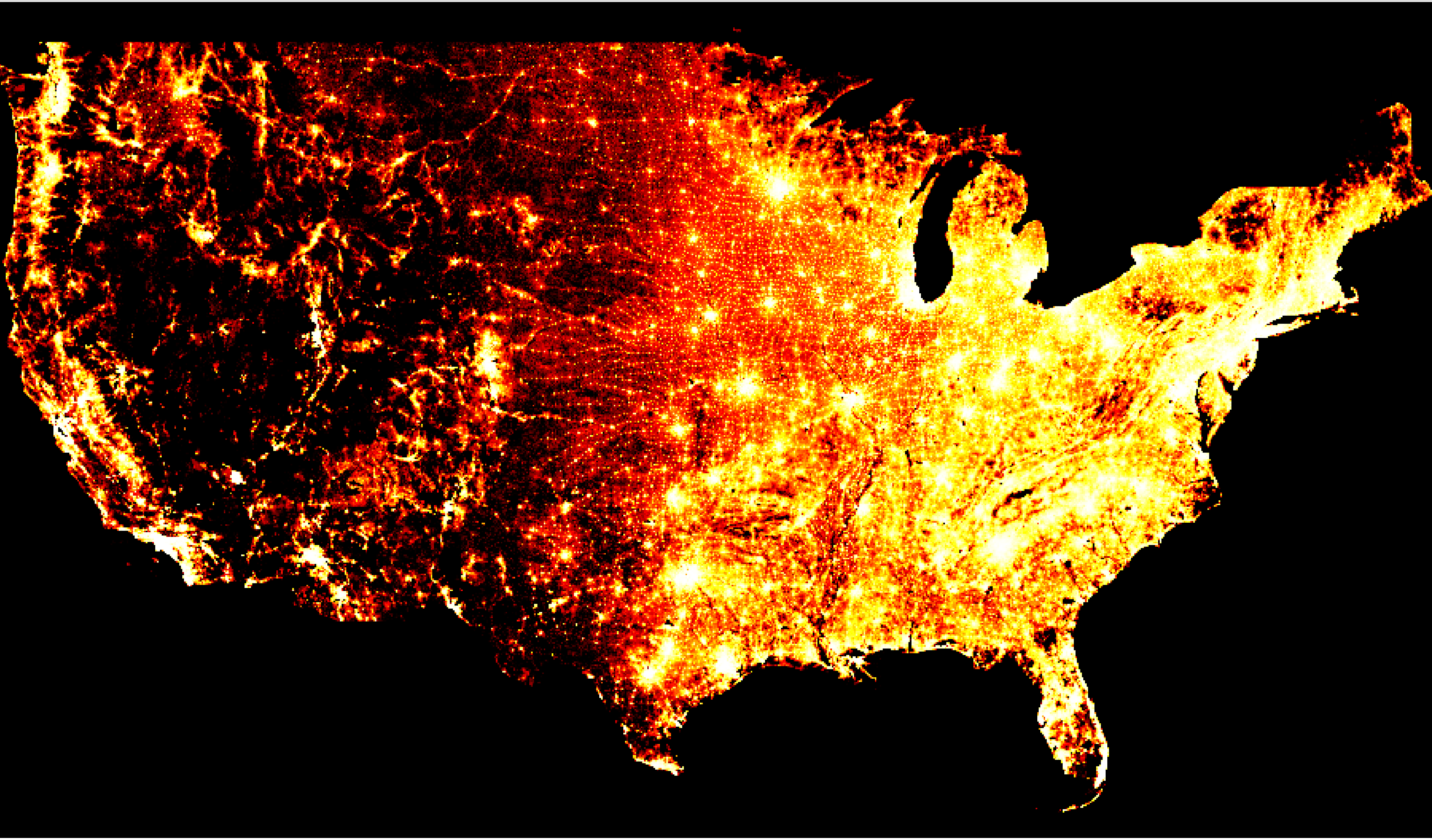
Really nice plots



City: Austin

Distribution: Discrete

https://demo.bokeh.org/weather

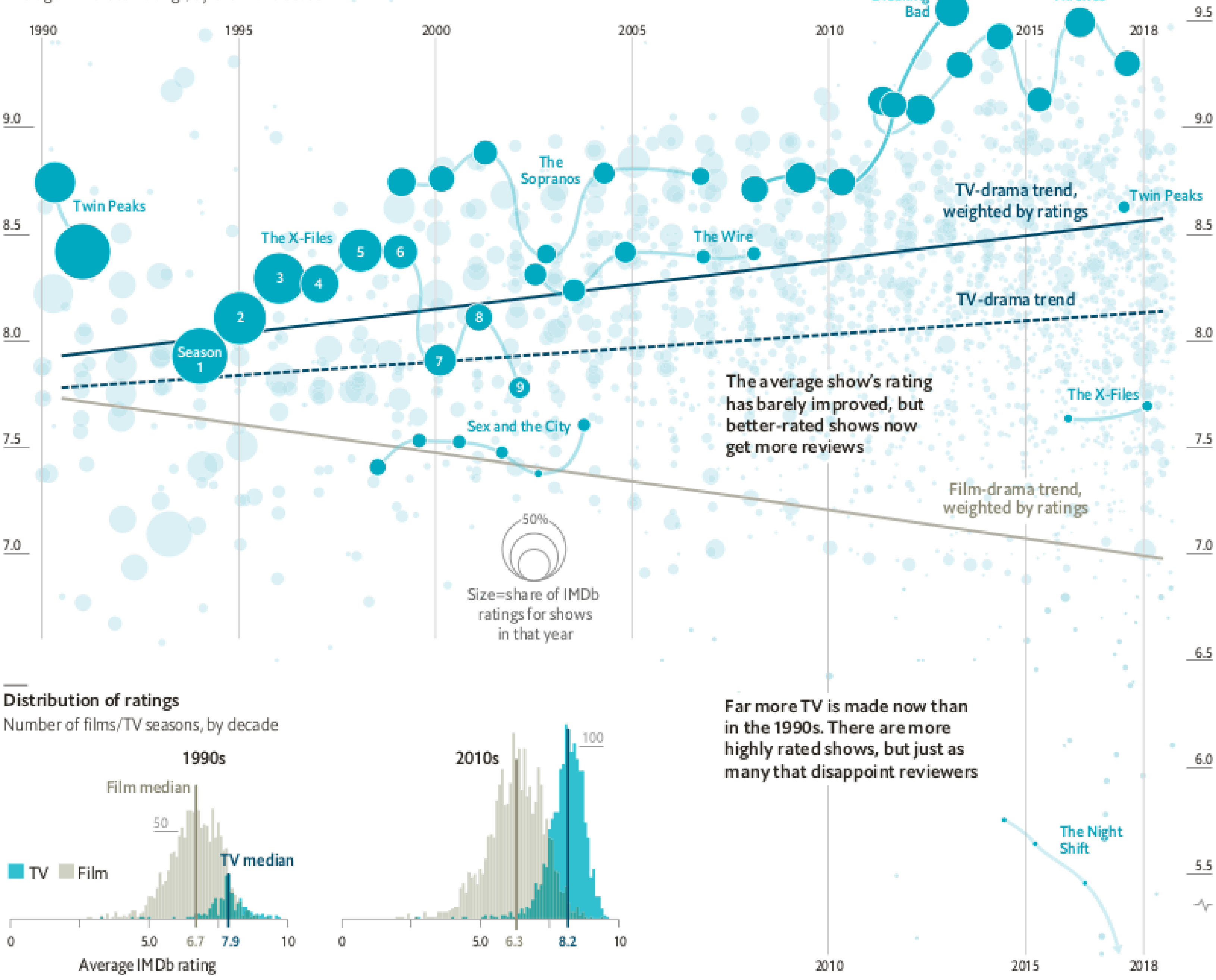


300M points
1 point = 1 person

https://datashader.org/

TV dramas shown in America

Average IMDb user ratings, by show and season*



The average show's rating has barely improved, but better-rated shows now get more reviews

Distribution of ratings

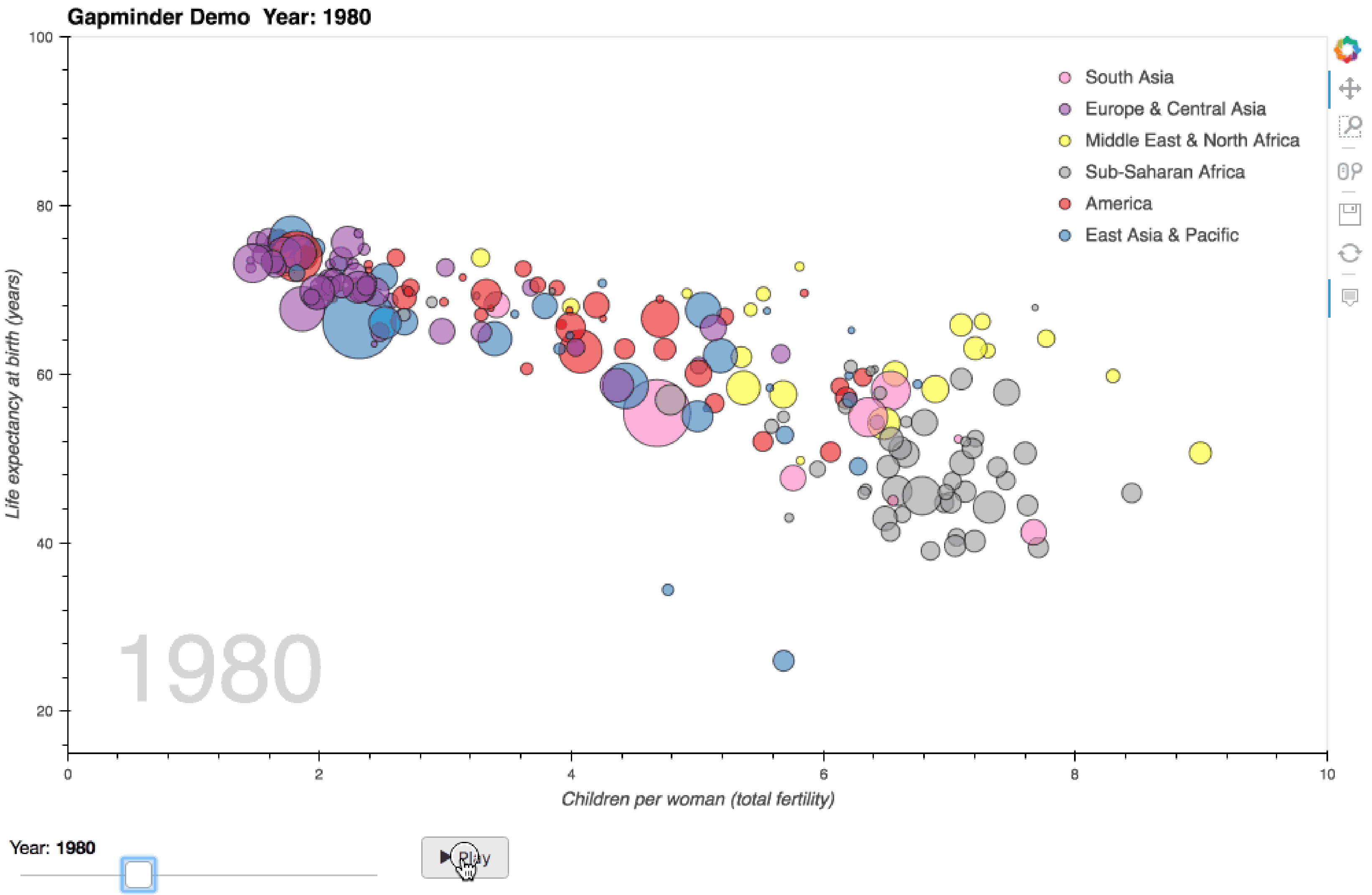
Number of films/TV seasons, by decade

Far more TV is made now than in the 1990s. There are more highly rated shows, but just as many that disappoint reviewers

*Seasons with at least 100 ratings on average

Sources: IMDb; OMDb; *The Economist*

Really nice plots



Year: 1980

Hans Rosling

1948-2017

https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen#t-72754

Really nice plots

Hans Rosling

1948-2017

https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen#t-72754

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

lukaskoz@mimuw.edu.pl