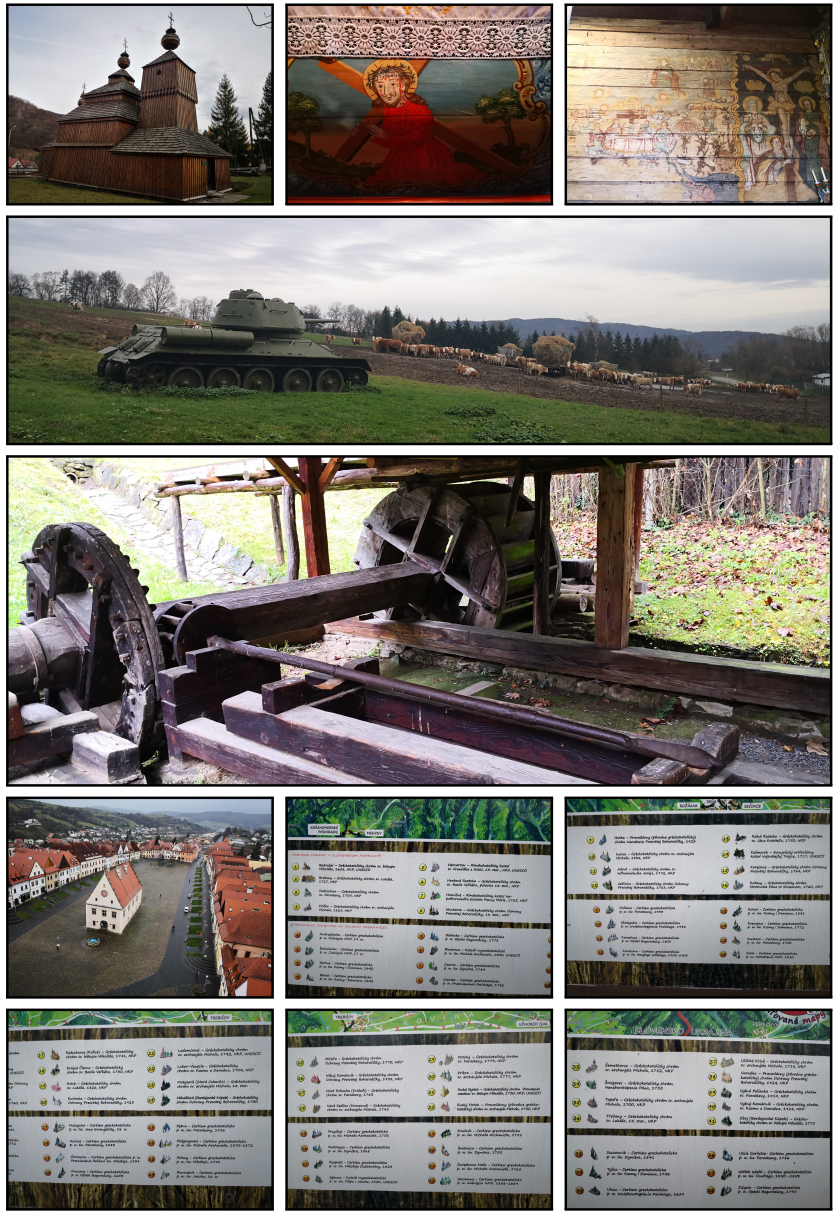


VÝUKA STATISTIKY PRO INFORMATIKU A MANAGEMENT V ÉŘE DATOVÉ VĚDY

STATISTICS IN INFORMATICS AND MANAGEMENT EDUCATION IN THE DATA SCIENCE ERA

Hana Skalská

Adresa: Fakulta informatiky a managementu, Univerzita Hradec Králové, Rokitanského 52, 500 03 Hradec Králové

E-mail: hana.skalska@uhk.cz

Abstrakt: Technologie zvyšují nároky na znalosti absolventů v informatických a manažersko-informatických oborech. Rozšiřuje se spektrum datových typů, které se analyzují. Zvyšuje se rozsah datových souborů, řeší se nové typy problémů. Jsou nové možnosti vizualizace dat i výsledků z nich získaných, zkracuje se doba mezi záznamem dat a prezentováním výsledků, softwarové produkty často automatizují kroky od získání dat po návrhy využití výsledků. Změny souvisí s většími nároky na znalosti statistiky. Analýza dat je často spojená s novými typy statistických problémů, které je vhodné nově zařadit nebo zdůraznit ve výuce statistiky. Příspěvek vychází z přehledu vývoje datové vědy, souvisejících typů úloh a popisuje návrh změn v sylabech výuky statistiky pro obory aplikovaná informatika a informační management na Fakultě informatiky a managementu Univerzity Hradec Králové.

Klíčová slova: Datová věda, výuka statistiky, nestatistické obory, statistika, aplikovaná informatika, informační management.

Abstract: Outstanding progress in technology, mainly in programming languages, visualization techniques, and communication tools resulted in increased popularity of Data Science field. This multidisciplinary field is closely related to statistics. The paper starts with mentioning several important changes in technology which influenced the practice of data analysis and contributed to data driven and model driven approaches to decision support. Then it reports a range of statistical problems important in data science applications which should be emphasized in statistical courses for non-statistical studies. Suggestions for changes in the content of statistical courses offered at the Faculty of Informatics and Management, University of Hradec Kralove, cover statistical problems appearing in big data statistical analysis, modelling, and visualization.

Keywords: Data Science, education in statistics, non-statisticians, statistics, applied informatics, information management.

1. Úvodní východiska

Absolventi oborů aplikovaná informatika (AI) a informační management (IM) FIM UHK pracují s technologiemi, které se v posledním desetiletí výrazně zdokonalily a generují množství dat, pro jejichž využití jsou nutné další technologie. Nápadité aplikace využívají výhod konektivity a možností generovaných dat, ovlivňují zájmy a preference studentů, kteří vidí budoucnost i ekonomické možnosti v technologiích a jejich užití (Google, Facebook, Twitter, Amazon, Heureka, Uber, Airbnb a mnoho dalších). Neuvědomují si, že za viditelnými výsledky aplikací se skrývají teoretické a obecnější znalosti, které jsou potřebné pro návrh a vytvoření aplikace, její aktualizaci, nebo pro řešení následných problémů. Statistické a matematické metody, které jsou na pozadí většiny těchto řešení, často objevují až během naší výuky.

Technologie se dříve prosazovaly v úlohách, které se nedají řešit analyticky. Ovlivnily tak chápání rolí informatiky a statistiky, které se svými přístupy k analýze mohly jevit jako dva rozdílné světy. Postupně se ukázalo, že poznatky pravděpodobnosti a statistiky jsou nezbytné k vysvětlení problémů pozorovaných při ryze informatickém přístupu k některým typům úloh, které vycházejí z dat. Také ve statistice vznikaly obory, které závisí na intenzivním využití počítačů (Bayesovské přístupy, bootstrap inference, MCMC, nebo optimalizační algoritmy) a tvoří oblast výpočetní statistiky.

Obor datová věda (DS), Data Science, jehož současná podoba je výsledkem masivního rozvoje technologií a možností dat, vychází z obou přístupů. Stěžejní pro DS jsou informatika, technologie a inženýrství, statistika a aplikační oblast. Statistické přístupy tvoří součást metodologie DS a jsou zahrnuty v klíčových učebnicích oboru DS, například [2], [5], [12].

Článek vychází z rešerše, která stručně mapuje vliv technologií na analýzu dat, až po současné pojetí DS. Zaměří se na typické úlohy, které může řešit většina firem. Cílem článku je odvodit okruh problémů, které jsou pro současné časté typy analýzy dat důležité a je vhodné zařadit je do statistických sylabů na Fakultě informatiky a managementu Univerzity Hradec Králové (FIM UHK).

2. Datová věda

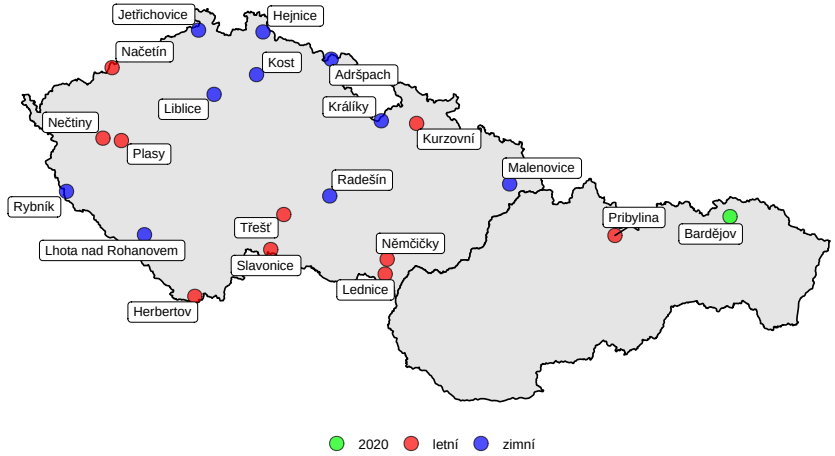
Datová věda je označení pro proces vyhledávání nových znalostí a informací z dat obecně libovolného typu a rozsahu. Koncept DS není pro statistiku novým pojmem. Odborné vysvětlení a použití termínu „analýza dat“ zavádí ve statistice Tukey [16], který nazval proces analýzy a využití dat učením z dat a charakterizoval jej jako vědu. Tyto myšlenky dál rozvíjel [17] a jejich opod-

POZVÁNKA NA KONFERENCI ROBUST 2020
INVITATION TO THE ROBUST 2020 CONFERENCE

Redakce

Konference ROBUST je nejvýznamnějším setkáním statistiků pořádaným Českou statistickou společností ve spolupráci s JČMF. Poprvé se konala v roce 1980 a dále každý sudý rok jako letní či zimní škola.

Kdy: 7. června (neděle) – 12. června (pátek) 2020
Kde: Bardějov, Slovensko (foto J. Vencálková; další strany J. Antoch)
Více informací: `https://www.karlin.mff.cuni.cz/~antoch/`
Registrace: `https://robust.nipax.cz/`



statněnost zdůvodnil rozvojem možností kvantifikovat děje, snahou získávat o dějích stále více informací pomocí dat, vývojem počítače a dalších zařízení, která umožní práci s rozsáhlými typy dat a pokroky statistické teorie.

V prvním období rozvoje počítačů se však za vědu o datech považovala spíše softwarová a technická řešení, spojená s analýzou velkých datových souborů. Formální vznik výpočetní statistiky, která je součástí vědeckého přístupu k analýze dat, je spojený se založením International Association for Statistical Computing (IASC) v roce 1977. Dalším vývojovým stupněm v technikách analýzy dat byla oblast vyhledávání znalostí z databází, odborně uznaná vznikem společnosti Knowledge Discovery in Databases (KDD) v roce 1989. V roce 2003 vychází časopis Journal of Data Science a profesní označení Data Scientist bylo odbornými společnostmi definované v roce 2005.

Vývoj kopíruje více vlivy technologických změn než sjednocování pohledů na statistický a informatický přístup k analýze dat. Koncept, kterým Tukey v již zmíněném článku [16] vysvětluje úlohu statistiky a statistického modelování v analýze dat, vyvolal řadu diskusí a dosud je inspirativní. Například Breiman [3], v návaznosti na názory o modelování z dat, které vyjádřil Tukey, mluví o dvou kulturách analýzy dat (inferenční úloha versus prediktivní úloha), Peng [14] polemizuje s jeho názory na fomulaci cíle explorace dat.

V současnosti vede konektivita mobilů, počítačů a sítí k množství generovaných dat. Využití dat k popisu a vysvětlení dějů předpokládá použití vědecky korektních metod analýzy dat. Donoho, více než 50 let od článku [16], shrnuje odezvy na článek, ve kterém Tukey označill analýzu dat vědou. Donoho [7] také charakterizuje současný stav a uvažuje o pohledu na analýzu dat v budoucnosti. Vychází z vývoje technologií i aplikací DS a popisuje proces DS šesti kroky, z nich každý by se nějakým dílem měl objevit ve výuce:

- *Získání dat a explorace*: Data z různých zdrojů, očištění, řešení anomálií a chybějících hodnot, posouzení eventuálních podmnožin dat, explorace.
- *Reprezentace dat*: Transformace, přestrukturování, matematické vyjádření některých typů (obraz, zvuk).
- *Výpočty*: Nezbytná znalost několika programovacích jazyků, techniky výpočtů v klastrech, cloudech, vytváření opakovatelných postupů.
- *Modelování na základě dat*: Model, který generuje sledovaný proces nebo model z dat, který umožní predikovat budoucí stav.
- *Vizualizace a reprezentace* dat i výsledků analýzy.
- *Vědecké zhodnocení procesu DS*: Sledování vlastností modelů pomocí zvolených metrik a jejich vyhodnocování. Uchování dokumetace jed-

notlivých typů analýz tak, aby v budoucnu bylo možné provést meta analýzu.

Specialisté skupiny KDnuggets (`https://www.kdnuggets.com/`) shrnují oblasti v analýze dat, které byly nejvíce ovlivněny technologickým pokrokem:

1. *Typy a formáty dat.* Text, obraz, zvuk, objekty, geografická data, sítě.
2. *Velké rozsahy datových souborů*, distribuované zdroje dat.
3. *Nové typy úloh.* Podobnost dokumentů, identifikace jedinců, analýza sítí, klasifikace prvků, tisíce atributů měřených na malém počtu prvků souboru.
4. *Problémy při analýze.* Příprava nenumerických dat, popis a analýza dat senzorů a analyzátorů, stovky až tisíce atributů v sadě, rozhodnutí o statistické významnosti u simultánních testů nebo při tvorbě modelů.
5. *Možnosti vizualizace výsledků* na webu nebo formou interaktivních grafů.
6. *Zkracování doby* mezi záznamem údajů a prezentováním výsledků, nároky na technologická řešení a na správnost postupů (informace z dat může být bezprostředně využitá), data products jako cíl analýzy, virtualizace.
7. *Nestrukturovaná data.* Databáze, jejich výkon, konektivita.
8. *Vyhledávací nástroje*, služby poskytované (sdílené) přes Internet, analýza efektivity a GDPR.
9. *Vývoj softwaru*, softwarových platforem, jazyků (R, Python, Java), virtualizace, růst požadavků na vzdělání, možnosti sebevzdělání (Coursera).

Kroky DS zahrnují: Datový management, vizualizaci dat, strojové učení, matematiku, statistické programy, programování a statistiku. V praxi jsou pro DS používané open source technologie samotné, nebo jsou integrované do komerčních informačních systémů.

DS tedy má velké požadavky na znalosti absolventů všech informatických a manažersko-informatických oborů, kteří se do některé fáze procesu zapojí. Požadavky souvisí s novými typy aplikací, využitím dat různého typu (strukturovaných i nestrukturovaných), datovými soubory velkého rozsahu. DS je obor multidisciplinární a pro vzdělávání ukazuje také nutnost přípravy na komunikaci a spolupráci specialistů různých oborů.

Termínem *data products* se označují aplikace, které integrují procesy získání a přípravy dat i statistické algoritmy inference a modelování [1]. Pro data, generovaná aktuálním procesem, jsou výsledky analýzy ve tvaru, který

Momentka z konference

Místo konání – zámek Křtiny a přilehlý barokní kostel Jména Panny Marie

Účastníci konference naslouchají křtinské zvonkohře

Exkurze v jeskyni Výpustek

usnadní interpretaci a využití. Možná statistická úskalí musí být ošetřena při návrhu aplikace a měl by je v principu znát i uživatel výsledků. Pokud data products generují současně i data pro meta analýzy, jedná se o procesy datové vědy [7], jejich platnost a spolehlivost lze ověřovat a analyzovat.

3. Statistické vzdělávání

Cílem výuky je rozvíjet statistické uvažování a znalosti, které se předpokládají při návrhu a provedení analýzy dat i při vyhodnocení výsledků. Pochopení principu a potřebné znalosti jsou důležité pro rozpoznání rizika nesprávného nebo nevhodného užití statistických metod, chybné interpretace výsledku nebo posouzení vhodnosti navržených postupů přípravy dat [11].

Podíl statistiky na DS je zřejmý. Obsah kurzů pro nestatistiky a forma výuky mají kromě nových znalostí a pochopení principů vést k zájmu o statistiku a získání znalostí, které umožní jejich budoucí rozšíření. Tomuto cíli se lze přiblížit vyváženým obsahem a formou výuky. Limitem jsou preference studentů, jejich očekávání a zájem vstřebat řadu pojmů a postupů.

Obsah, forma výuky i přístup vyučujících jsou důležité. V literatuře se obsahu i metodám výuky statistiky věnuje pozornost ve všeobecném vzdělávání [4] i ve vyšším vzdělávání [18]. Přesto výsledky experimentů a studií ukazují, že studenti nestatistických oborů často považují statistiku za obtížnou a někdy stresující, přestože vyučující se snaží využívat nové výukové metody, rozpoznat obtíže a předcházet stresujícím momentům [9]. Přehled výzkumů a šest doporučení pro předcházení stresu u studentů uvádí Chew [10].

Motivačním momentem pro naše absolventy může být i společenský zájem o DS (například [15]) a tím o datové specialisty. Uplatnění našich absolventů v DS je reálné díky technologiím, kterým se učí rozumět a ovládat je. Výhodou jejich zapojení do týmu, který navrhuje a vytváří náročnější aplikace, nebo pro realizaci jejich samostatných řešení, budou i znalosti statistických principů. Naše kurzy mají rozvíjet tyto znalosti a podpořit zájem o statistiku zařazením vybraných statistických témat, která jsou vyučovaná v kurzech pro DS specialisty a analytiky [6], [8], [13], nebo vycházejí z častých typů úloh. Zaměřujeme se na problémy, které souvisí s modelováním, analýzou rozsáhlejších dat a na možnosti softwaru.

4. Výuka statistiky pro AI a IM na FIM UHK

Ve studijních plánech AI a IM jsou tři povinné kurzy statistiky, každý je zakončený zkouškou a mají dotaci 26 + 26 + 13 hodin (přednášky + cvičení + práce) za semestr. Úvodní kurz Pravděpodobnost a statistika (PSTA) se vy-

učuje ve 3. semestru Bc. studia, v navazujícím studiu v 8. semestru je Aplikovaná statistika (APSTA), v 9. semestru Statistické modely a data (STMOD), tento kurz vznikl nedávno transformací z kurzu STOMO (Stochastické modelování). Veřejně dostupné aktuální informace ke kurzům jsou na webu. Názvy kurzů ve STAGu (`https://www.stag.uhk.cz`) obsahují vždy zkratku katedry KIKM, například KIKM/STMOD.

Vybraná témata statistiky jsou částí státní závěrečné zkoušky v navazujícím studiu. Otázka zkoušky se volí z okruhu, který souvisí s obsahem diplomové práce. Výuka je takto zavedená od roku 1997 s průběžnými aktualizacemi obsahu, techniky a softwaru na učebnách a výukových podkladů.

Ve cvičeních APSTA a STMOD pracujeme s IBM SPSS Statistics, který udržujeme v aktuální verzi. Licenční server IBM SPSS umožňuje mix uživatelů, kteří pracují se softwarem v síti a uživatelů na virtuálních desktopech v prostředí VMware a je tak dostupný studentům i vyučujícím (do počtu licencí) nepřetržitě.

Všechny kurzy mají opory v LMS prostředí BlackBoard, kde jsou prezentace k přednáškám, návody na cvičení, ukázky příkladů, příklady k použití softwaru, syntax pro SPSS nebo R a podobně. Toto výukové prostředí umožňuje přístup uživatelům interní počítačové sítě, nebo uživatelům se speciálním přihlašovacím jménem a heslem. Studenti dostávají přístup na kurzy, které si v daném akademickém roce zapsali.

Původní kurz Stochastické modelování (STOMO), který byl z části zaměřený na vytváření obecných modelů a jejich simulace, byl právě s přihlédnutím k nastupujícím trendům v analýze dat, záměnou přibližně jedné třetiny obsahu, převedený na současný kurz STMOD. V kurzu STOMO byla redukovaná témata generování náhodných čísel a obecných typů modelů a simulací. V současném STMOD byla nahrazena metodikou analýzy dat (CRISP), částečně problémy přípravy dat a aplikačním rozšířením témat regrese a časových řad, která navazují na předchozí kurz APSTA.

Vznikla i lepší návaznost v řešení úloh stejným softwarem, se kterým se pracuje v APSTA předchozího semestru. Pomocí IBM SPSS Statistics se řeší úlohy pro data, která jsou součástí instalace SPSS nebo otevřených databází, například UCI depozitáře (`https://archive.ics.uci.edu/ml/index.php`). Ve cvičeních je prostor pro ukázky přípravy dat, segmentace, transformace, pro návrhy řešení a následné vysvětlení výsledků modelování. Nově jsou do STMOD zařazeny klasifikační metody LDA a CART. Důraz je kladený na předpoklady LDA a jejich ověření, na princip metody CART, možnosti validace klasifikačních modelů, porozumění výstupům ze softwaru a aplikacím těchto metod. Stručně je v kurzu také uvedena metoda bootstrap odhadu intervalu spolehlivosti střední hodnoty (aplikace v R). Další témata se neměn-

Posluchači při úvodní přednášce Jakuba Fischera

Halina Šimková mluví o forenzní analýze DNA

na pochopenie, výber vhodnej metódy, aplikovanie a najmä správnu interpretáciu výsledkov.

Účelom ani cieľom nie je spomenúť všetkých autorov i príspevky, ktoré odzneli, tak len skonštatujem, že to stálo za to a niektoré príspevky nájdete publikované v Informačnom bulletine Českej štatistickej spoločnosti. Diskusia prebiehala vo veľmi priateľskej a srdečnej atmosfére, akú som doteraz nikde inde nezažila... Ďakujem za to všetkým zúčastneným a predovšetkým usporiadateľom. Konštruktívne a zanietené rozhovory na vážne i nevážne témy pokračovali i po oficiálnom konci programu konferencie. Padol i návrh na nové logo STAKAN, ktoré verím, že sa stane skutočnosťou. Zúčastnení vedia a ostatní nech sa nechajú prekvapiť. Príjemným prekvapením od organizátorov konferencie bola návšteva kvapľovej jaskyne Výpustek bez kvapľov.

Záverom môžem skonštatovať, že som sa cítila ako v rodine, medzi svojimi i keď som bola na tomto podujatí prvýkrát. Určite však nie posledný raz. Našla som viaceré inšpirácie i odpovede na mnohé otázky, ale stále som na ceste k objavovaniu. Záverom len konštatovanie smerom k štatistickej analýze. Keď ju miluješ, nie je čo riešiť.

Účastníci konference STAKAN 2019 – Křtiny

nila (Markovovy řetězce a jejich aplikce, model obnovy, model populačního procesu).

Ze zkušeností s výukou lze říci, že naši studenti neodmítají statistiku, ačkoliv ji považují za obtížný předmět. Zejména v posledním kurzu STMOD (dříve STOMO) si často volí náročnější témata pro seminární práce. Při anonymním elektronickém hodnocení výuky, do kterého se zapojuje necelá čtvrtina z 80–120 zapsaných studentů (hodnotit mohou od konce semestru po celé zkouškové období), tak tito převážnou většinou považují zařazení kurzů APSTA a STMOD do studijních plánů za vhodné. Skóre 5 nebo 4 (naprosto souhlasí nebo souhlasí) uvádí vždy zhruba 80–85 % studentů, nesouhlasí 1–2 studenti (skóre 1), žádní studenti neuváděli, že naprosto nesouhlasí (skóre 0).

5. Závěr

Rešerše, jejíž část zde byla představená, ujasnila roli statistiky pro nestatistiky v dnešním světě DS a přispěla k jasnějšímu názoru na vhodné změny obsahu kurzů APSTA i STMOD. Výuka by měla zahrnout i problémy, které se dosud ve výce zmiňovaly okrajově nebo vůbec, ale jsou časté v aplikacích. V kurzech APSTA a STMOD bude vhodné:

1. Vysvětlit obecný problém testování hypotéz, sílu testu a effect size.
2. Zmínit problém p-hodnoty v sadě testů a riziko false discovery rate (FDR).
3. Pro různé typy reziduí vysvětlit jejich užití v regresním modelu.
4. Věnovat se problému výběru nezávisle proměnných do modelu.
5. Důsledněji vést studenty k práci s literaturou v angličtině.
6. Cvičení zaměřovat více na vysvětlení a řešení problémů přípravy dat, transformace a zpětné transformace, segmentování a vizualizace dat.
7. Ve cvičeních opakovat stěžejní poznatky, například popisné charakteristiky pro různé typy znaků, jejich využití, vlastnosti, vizualizaci.

Poděkování

Děkuji anonymním recenzentům za cenné náměty a připomínky.

Literatura

[1] Bengfort, B., Kim, J. (2016): *Data Analytics with Hadoop: An Introduction for Data Scientists*, O'Reilly Media, Inc., pp. 269.
URL: http://shop.oreilly.com/product/0636920035275.do

[2] Blum, A., Hopcroft, J., and Kannan, R. (2018): *Foundations of Data Science*, online, pp. 479, 2018.
URL: https://www.cs.cornell.edu/jeh/book.pdf

[3] Breiman, L. (2001): Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* **16**(3), 199–231.
URL: https://projecteuclid.org/euclid.ss/1009213726

[4] Carver, R., Everson, M., Gabrosek, J., et al. (2016): Guidelines for Assessment and Instruction in Statistics Education (GAISE) Reports. American Statistical Association.
URL: https://www.amstat.org/education/gaise/

[5] Das, S. R. (2016): *Data Science: Theories, Models, Algorithms, and Analytics.* Published by S. R. Das, online, pp. 462.
URL: https://srdas.github.io

[6] De Veaux, R. D., Agarwal, M., et al. (2017): Curriculum Guidelines for Undergraduate Programs in Data Science. *Annu. Rev. Stat. Appl.* **4**, 15–30. URL:
https//doi.org/10.1146/annurev-statistics-060116-053930

[7] Donoho, D. (2017): 50 Years of Data Science. *Journal of Computational and Graphical Statistics* **26**(4), 745–766.
URL: https://doi.org/10.1080/10618600.2017.1384734

[8] Erickson, T., Wilkerson, M., Finzer, W., and Reichsman, F. (2019): Data Moves. *Technology Innovations in Statistics Education* **12**(1), pp. 25.
URL: https://escholarship.org/uc/item/0mg8m7g6

[9] Garfield, J. B., Ben-Zvi, D. (2008): *Developing Student's Statistical Reasoning. Connecting Research and Teaching Practice.* Springer, Science+Businees Media B.V., 2008.
URL: https://www.springer.com/gp/book/9781402083822

[10] Chew, P. K. H., Dillon, D. B. (2014). Statistics anxiety update: Refining the construct and recommendations for a new research agenda. *Perspectives on Psychological Science* **9**(2), 196–208.
URL: https://doi.org/10.1177/1745691613518077

SPOMIENKY NA STAKAN 2019
MEMORIES OF STAKAN 2019

Helena Fidlerová

E-mail: helena.fidlerova@stuba.sk

Rozhodnutie, prečo ísť na konferenciu STAKAN organizovanú 11. – 13. októbra 2019 Českou štatistickou spoločnosťou (ČStS) a Slovenskou štatistickou a demografickou spoločnosťou (SŠDS), bolo veľmi spontánne a jednoznačné, keď som sa dozvedela, že tohtoročnou témou konferencie bude výuka štatistiky pre nematematické odbory na stredných a vysokých školách. Počet odborných konferencií z oblasti štatistiky a pravdepodobnosti je veľký, ale toto bolo to, čo som už dlho hľadala. Výučbu štatistických metód na ústave priemyselného inžinierstva a manažmentu MTF STU Trnava zabezpečujem na našom pracovisku s prestávkami od roku 2003. Dúfala som, že nájdem odpovede na moje otázky, ako učiť milovanú i nenávidenú štatistiku.

Seminár sa konal v krásnom prostredí zámku Křtiny, kam to bolo zo všetkých kútov Slovenska i Česka blízko. Odborný seminár bol naplnený príspevkami širokého spektra odborníkov z oblasti teórie, výuky a uplatnenia štatistiky, predovšetkým z českých a slovenských vysokých škôl.

V úvode nás srdečne privítali predseda ČStS Ondřej Vencálek a predsedníčka SŠDS doc. Iveta Stankovičová, a hneď som vedela, že som na správnom mieste. Pozvané prednášky boli plné charizmatických osobností ako jsou profesor Ryozo Miura z Hitotsubashi University s témou *My experience of Teaching Statistics with a software JMP at Hitotsubashi University* a profesor Jakub Fischer z Vysokej školy Ekonomickej v Prahe s témou *Ekonomická a matematická statistika – vzájemné p(r)otkávání nejen ve výuce.* Následne odzneli i skúsenosti s výučbou na rôznych typoch škôl v Čechách, na Slovensku, ale i ďalekom Ománe (J. Mačutek, I. Stankovičová, Z. Šulc, M. Zikmundová a ďalší). Do pekelnej reality nás vniesol Tomáš Fürst s jeho vystúpením na tému *Perverzní incentivy a špatná statistika: Pohled z pekla.*

Výnimočné pre mňa bolo i osobné stretnutie s ľuďmi, ktorých publikácie boli pre mňa prvou barličkou pri poznávaní tajov štatistiky, ako pani prof. Ing. Hana Řezanková, CSc. Viaceré príspevky mali špecializovaný štatisticky obsah (P. Martinková, H. Šimková, T. Hajdúková, I. Waczulíková, J. Antoch). Skloňovala sa i Hejného metóda v príspevku Ota Přibylu a Pavla Hejného.

Široko bola diskutovaná problematika i úskalia používania rôznych štatistických softwarov pri výuke štatistiky, či už platených verzií alebo freewarov. Zhoda panovala v názore, že študenti rôzneho zamerania sa musia zamerať

```
library(png)
mujpng<-readPNG("tux/ossconf-2020-upravene.png")
unique(as.raster(mujpng[,,1:3]))
```

Výstup je:

```
[1] "#FFFFFF" "#4D4D4D" "#426BAA" "#37A9E9" "#FEFEFE" "#FEFFFF" "#999999"
```

U našeho konkrétního obrázku linuxového tučňáka Tuxe v logu konference OSSConf jsme zjistili, že tam máme dvě barvy (`#426BAA`, `#37A9E9`), dvě šedé (`#4D4D4D`, `#999999`), bílé pozadí (`#FFFFFF`) a na první pohled neviditelné, řekněme „parazitní“, chceme-li přehlédnuté, téměř bílé pixely: šedé a barevné (`#FEFEFE`, `#FEFFFF`).

Nyní už lze barvy oddělovat (například chceme mít co barva to jeden polygon), filtrovat (nechceme např. bílé pozadí ani padesát odstínů šedi), měnit (nahradit jednu barvu za jinou), spojovat (např. spojit barvy textů, nebo vše spojit do jedné barvy) ap.

9.3. Pracovní zobrazení obrázku

Zkusíme si jednoduchý filtr: odstranit bílé a téměř bílé pixely a zobrazit si pracovní náhled obrázku. Zároveň pro načtení více R knihoven otestujeme knihovnu `pacman`. Zmíněné otevřené problémy v R jsou tímto uzavřené.

```
#library(raster); library(sf); library(spex); library(ggplot2)
library(pacman) # alternativní postup načtení více knihoven
pacman::p_load("raster","sf","spex","ggplot2")
data<-raster("tux/ossconf-2020-upravene.png")
data[data>200] <- NA # ořez bílých a téměř bílých pixelů
vysledek<-polygonize(data)
#plot(vysledek) # je to pomalé, ale použitelné
#pdf("pracovni-nahled.pdf") # uložení pro bulletinek
ggplot()+geom_sf(data=st_as_sf(vysledek))
#dev.off() # uzavření ukládání pdf
```

[11] Ioannidis, J. P. A. (2005): Why Most Published Research Findings Are False. *PLoS Med.* (8)(2): e124, 696–701.
URL: https://doi.org/10.1371/journal.pmed.0020124

[12] Leskovec J., Rajaraman A., and Ullman J. (2014): *Mining of Massive Datasets.* Cambridge University Press, pp. 467.
URL: https://doi.org/10.1017/CBO9781139924801

[13] Mikroyannidis A., Domingue J., Phethean C., et al. (2018): Designing and Delivering a Curriculum for Data Science Education across Europe. In: Auer M., Guralnick D., Simonics I. (eds) *Teaching and Learning in a Digital World.* ICL 2017. Advances in Intelligent Systems and Computing. Vol. 716, 540–550. Cham: Springer.
URL: https://doi.org/10.1007/978-3-319-73204-6_59

[14] Peng, R. (2019): Tukey, Design Thinking, and Better Questions.
URL: https://simplystatistics.org/2019/04/17/tukey-design-thinking-and-better-questions/

[15] The Royal Society (2019): *Dynamics of data science skills: How can all sectors benefit from data science talent?*
URL: https://royalsociety.org/topics-policy/projects

[16] Tukey, J. W. (1962): The Future of Data Analysis. *Ann. Math. Statist.* **33**(1), 1–67.
URL: https://projecteuclid.org/euclid.aoms/1177704711

[17] Tukey, J. W. (1977): *Exploratory Data Analysis.* Addison-Wesley Publishing Company, pp. 688, 1977.
URL: https://doi.org/10.1002/bimj.4710230408

[18] Zieffler, A., Garfield, J., Alt, S., et al. (2008): What Does Research Suggest About the Teaching and Learning of Introductory Statistics at the College Level? A Review of the Literature. *Journal of Statistics* **16**(2), pp. 26.
URL: https://doi.org/10.1080/10691898.2008.11889566

PF2020! ANEB KDYŽ NESTAČÍ ANI R, ANI TEX
PF2020! OR WHEN NEITHER R NOR TEX ARE SUFFICIENT

Pavel Stříž

E-mail: pavel@striz.cz

Abstrakt: Článek zmiňuje úvahy a kroky vedoucí k vysázení loga PF2020! Logo je vysázené z šestiúhelníkových nálepek, více o kolekcích v GitHub repozitářích hex-stickers (Rstudio) a BiocStickers (Bioconductor). Hlavní zdroj inspirace bylo logo z konference useR! 2018, které vytvořil Mitchell O'Hara-Wild za použití jeho R skriptu hexwall.

Klíčová slova: R, ImageMagick, hexwall, raster, gglogo, magick, tidyverse, ggplot2, sf, TEX, Lua, Bash, GraphicsMagick, png, pacman.

Abstract: The article describes a thinking process and a problem-solving approach of creating a PF2020! logo. It's typeset of hexagon stickers, see Rstudio's hex-stickers and Bioconductor's BiocStickers repositories in GitHub. The main inspiration came from the useR! 2018 logo which was created by Mitchell O'Hara-Wild using his hexwall R script.

Keywords: R, ImageMagick, hexwall, raster, gglogo, magick, tidyverse, ggplot2, sf, TEX, Lua, Bash, GraphicsMagick, png, pacman.

1. Problém typu word cloud

Word cloud / wordle / tag cloud se běžně řeší tak, že slova či obrázky se převedou do rastrových obrázků a následně se zkoumají přesahy na úrovni pixelů[1]. Dlouhodobě se zabýváme vektorovým řešením tohoto problému na úrovni TEXu. To lze zrealizovat např. přes METAPOST a příkaz `bisect`, ale to je vše „na dlouhé lokte". Proto nás zaujalo logo z konference userR! 2018, které má svou mřížku z šestiúhelníků, ale nálepky (angl. stickers) jsou sázeny jen uvnitř polygonů, v jejich případě mapy Austrálie. Návod lze nalézt na blogu autora, viz webová stránka `https://www.mitchelloharawild.com/blog/user-2018-feature-wall/`.

2. První dojmy

Zkusili jsme v tomto duchu připravit PF2020!, v R, co nejrychleji… Ale! Shrneme-li problémy: je potřeba doinstalovat knihovny v R a k tomu je po-

[1] `https://www.jasondavies.com/wordcloud/about/`

9. Závěrečné tipy: ještě jedno ohlédnutí za R

S určitým časovým odstupem si odpovídáme na otevřené otázky u R.

Při instalaci knihovny rgdal to chce novější verzi R než nabízí standardní linuxový repozitář (Xubuntu 18.04). Vyřešili jsme to takto.

V `/etc/apt/sources.list` jsme přidali:

```
deb https://cloud.r-project.org/bin/linux/ubuntu bionic-cran35/
```

Následovala instalace R:

```
$ sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys
    E298A3A825C0D65DFD57CBB651716619E084DAB9
$ sudo apt update
$ sudo apt upgrade
$ sudo apt install r-cran-base
```

V případě problémů se závislostmi užíváme místo známých programů `apt` či `apt-get` program `aptitude`. Hodilo se nám to při míchání 32 a 64bitových aplikací, resp. programů z různých linuxových distribucí.

V R následuje doinstalování knihoven, které použijeme. Ideální řešení je instalovat jednu knihovnu za druhou a sledovat případné chybové zprávy.

```
install.packages(c("png","spex","raster","rgdal","sf","pacman"))
```

9.1. Zjištění barvy konkrétního pixelu

Poněvadž jsme zápasili s datovými typy, zkusili jsme si získat barvu konkrétního pixelu a dostat jeho RGB složky, ve stylu GIMPu, ale už bez GIMPu.

Využili jsme knihovnu raster a funkci `brick` nebo `stack`, získali jsme složky RGB, převedli na šestnáctkové hodnoty a přidali znak `#`.

```
library(raster)
data<-brick("tux/ossconf-2020-upravene.png")  # nebo
#data<-stack("tux/ossconf-2020-upravene.png")
danaBarva<-paste("#", paste( as.hexmode( getValues(data,50,1)[50,] ), sep="",
    collapse=""), sep="")
danaBarva
```

Výstup je:

```
[1] "#426baa"
```

To už lze přímo použít pro parametr `colour`, např. `colour=danaBarva`.

9.2. Výpis všech barev v obrázku

Máme za sebou barvu jednoho pixelu, podívejme se, jak lze proces zautomatizovat a získat všechny jedinečné barvy v rastrovém RGB obrázku. Použijeme knihovnu png.

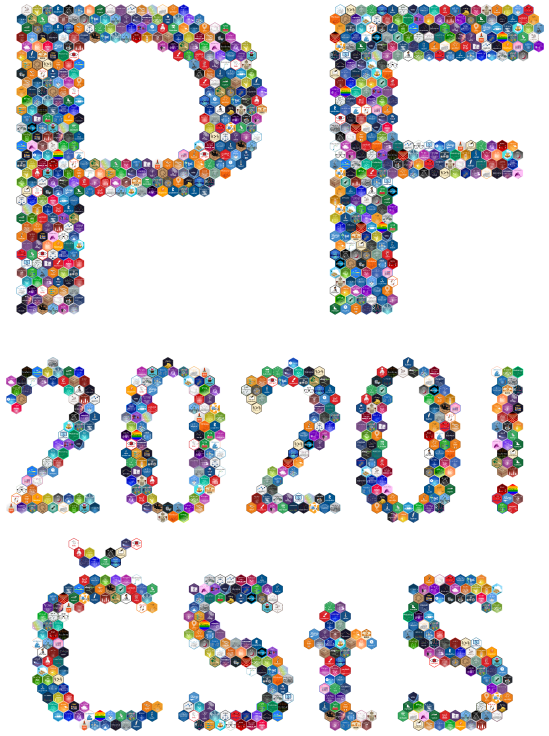
```
\begin{document}
\input ../zdrojak.tex
\end{document}
```

Nezbývá než si vše vysázet:

```
$ lualatex pf.tex; \
> lualatex pf.tex; \
> pdfcrop --hires --margins 0 pf.pdf; \
> gm convert -density 1800 pf-crop.pdf pf-crop.png
```

První dva řádky zajistí vysázení a absolutní umístění na straně. Třetí řádek nám ořeže ochrannou bílou zónu. Šikovný TEXista brzy zjistí, že by se to dalo zjednodušit na jeden běh TEXu bez nástroje `pdfcrop`. Necháváme otevřené pro badatele. Poslední řádek nám vygeneruje rastrový náhled.

Nezbývá než se pokochat vánočním dárkem, a doufat, že soubory pdf a png nebudou moc velké, nebude se to mezi svátky dlouho vykreslovat a do Nového roku se novoročenka celá zobrazí…



třeba mít nástroje a knihovny. Užili jsme Linux a omlouváme se uživatelům Microsoft Windows, že jsme testy nezkoušeli pod tímto operačním systémem. Instalovali jsme jednotlivé knihovny jednu po druhé a sledovali v R chybové zprávy. Tímto způsobem jsme se vrátili pod Linux a doinstalovali vždy to, co bylo potřeba. A opětovně instalovali konkrétní knihovnu v R. Jednalo se postupně o knihovny `libmagick++-dev`, `librsvg2-dev`, `libssl-dev`, `libogdi3.2-dev`, `libcurl4-openssl-dev`, `gdal-bin` a `libwebp-dev`. Je však možné, že jsme některé nástroje již měli nainstalované, proto není výčet úplný.

Funkce `hexwall` je závislá na knihovně magick, který nám pro velký počet souborů přestane fungovat. V našem případě kolem tisícího souboru. Nehledě na časovou náročnost práce s obřím rastrovým obrázkem v pozadí.

Další problém byl, že se nám za půlden testů zaplnil pracovní adresář `/tmp/` o 500 GB. Hledejme proto jiné nástroje.

Poslední problém byl, že jsme nechtěli mapu (geografická data), ale texty, případně užít obecný (černobílý) rastrový obrázek. Na tohle jsme se zaměřili.

3. Zdárné kroky s knihovnou gglogo

Při hledání převodu znaků do polygonů jsme objevili tuto knihovnu. Pracovali jsme v adresáři, kam jsme si uložili skript hexwall:

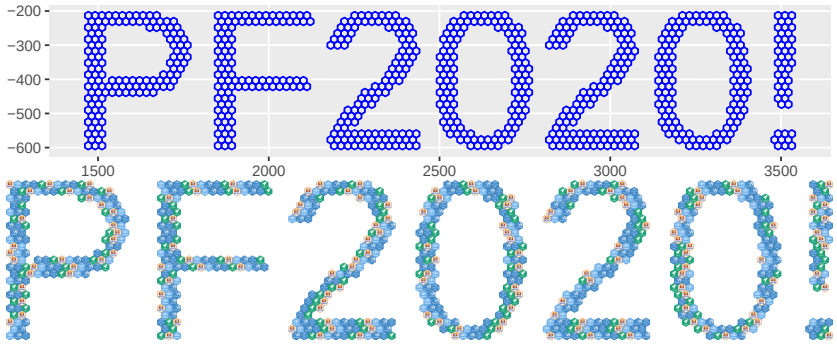
```
$ git clone https://github.com/mitchelloharawild/hexwall.git; \
> cd hexwall
```

Ačkoliv jsme zápasili s převody mezi datovými typy, zde je použitelný výsledek. Mezi pokusy jsme na vyčištění používali příkaz `rm(list=ls())`.

```
library(gglogo)
letter <- letterToPolygon("PF2020!", fontfamily="Helvetica", dim=c(5000,800))
library(raster)
Sr1 = Polygon(cbind(letter$x,letter$y))
Srs1 = Polygons(list(Sr1), "s1")
SpP = SpatialPolygons(list(Srs1), 1:1)
library(sf)
hex_points <- SpP %>% spsample(type = "hexagonal", cellsize = 20)
hex_points@coords

library(ggplot2)
library(tidyverse)
as_tibble(hex_points@coords)
aus_hex <- HexPoints2SpatialPolygons(hex_points, dx = 20)
#pdf("nahled.pdf")
ggplot() + geom_sf(data=st_as_sf(aus_hex), colour="blue", fill=NA)
#dev.off()
```

```
source("hexwall.R")
mojkovo <- hexwall("samplehex", sticker_width=20, coords=hex_points@coords,
     sort_mode="filename")
plot(mojkovo)
#image_write(mojkovo,"pf2020-pres-hexwall.png")
```



V této ukázce se nám nelíbilo, že je tam málo nálepek (5 souborů), nemůžeme užít zúženou mezeru za PF i netradičně před vykřičník, abychom poškádlili typografický svět, a chtěli jsme si zkusit výběr bez vracení s vracením (vysvětlíme). Pokusme se v další části článku tyto úkoly vyřešit.

4. První krok s knihovnou raster

Jistým vývojovým mezistupněm se nám stala „zakázka“, chceme-li experiment, pro organizátory konference OSSConf v Žilině, `http://ossconf.soit.sk/`. Vzali jsme logo roku 2019, zvětšili, v GIMPu zasáhli do roku, drobně jsme roztáhli cifry a vyčistili vyhlazování typické u obrázků na internetu (Colors→Threshold) plus úprava pixelů „zde, tu a támhle“. Tím jsme si zajistili obrázek přesně se čtyřmi barvami a bílým pozadím (v případě nutnosti průhledným).

Zkusili jsme načíst rastrový obrázek a do vzniklých polygonů vkreslit šestiúhelníky. Abychom se procvičili v R, zkusili jsme cyklus `for` přes barvy. Přes `data[]` jsme zjistili, že červená barva je v intenzitách 66, 55, 153 a 77. Příslušné RGB hodnoty jsme vyčetli v GIMPu. Zkušenější uživatelé R by jistě přišli na to, jak unikátní RGB hodnoty získat přímo v R a na tom postavit cyklus. Ručně to u 4 barev šlo, kdyby jich bylo víc, bylo by potřeba celý postup zautomatizovat.

Nepodařilo se nám hexa hodnotu barev vytáhnout z `data.frame` po spojení proměnných `hodnoty` a `barvy`. Zůstalo nám to jako otevřený problém.

Zde je výsledek našich snah.

zdrojový kód, který vysázíme. Navíc jsme si v Lua skriptu nastavili jednoduché měřítko. Výsledek našich snah by mohl vypadat takto:

```
math.randomseed(1)
soubory=io.open("soubory-pfhex.txt") -- seznam nálepek
obsah=soubory:read("*all")
soubory:close()
pngs={}
pngsfull={}

unicode.utf8.gsub(obsah, "([^\n]-)\n", function(s)
  table.insert(pngs,s) -- pracovní tabulka s nálepkami
  table.insert(pngsfull,s) -- neměnná tabulka všech nálepek
  end)
soubor=io.open("hexagony.csv") -- seznam nalezených souřadnic šestiúhelníků
obsah=soubor:read("*all")
soubor:close()
kam=io.open("zdrojak.tex","w") -- TikZový zdrojový soubor
kam:write("\\begin{tikzpicture}[remember picture, overlay]\n")
unicode.utf8.gsub(obsah, ",([^,]-),([^\n]-)\n", function(s,t)
  tos=s/1; tot=t/1 -- jednoduchá škála, je-li nutná
  pickup=math.random(1,#pngs) -- výběr bez vracení
  kam:write("  \\node[m] at ("..tos.."pt,"..tot.."pt)
    {\\includegraphics[width=\\maldimen]{"..pngs[pickup].."}};\n")
  table.remove(pngs,pickup) -- odebere vybrané logo
  if #pngs==0 then -- vrátit všechny nálepky, deep copy
    for k,v in pairs(pngsfull) do pngs[k] = v end -- for
  end -- if
end) -- unicode.utf8.gsub
kam:write("\\end{tikzpicture}\n")
kam:close()
```

Postupně se generuje soubor `zdrojak.tex`, první řádky vypadají takto:

```
\begin{tikzpicture}[remember picture, overlay]
  \node[m] at (62.795979916118pt,21.627203329949pt)
    {\includegraphics[width=\maldimen]{glue.png}};
  \node[m] at (64.295979916118pt,21.627203329949pt)
    {\includegraphics[width=\maldimen]{scmap.png}};
```

Náš poslední úkol je načíst si tento Ti*k*Zový kód a vysázet novoročenku, soubor `pf.tex`. To provedeme v TEXu. My jej měli ve složce `sazba`, aby se nám pomocné TEXové soubory nemíchaly s ostatními.

```
\documentclass[landscape]{article}
\pagestyle{empty}
\usepackage{tikz}
\graphicspath{ {../pfhex-output/} }
\newdimen\maldimen \maldimen=1.5pt % cellsize / škála v Lua souboru
\tikzset{inner sep=0pt, outer sep=0pt,
  m/.style={xshift=-100pt, yshift=-100pt, draw=none} }
```

7.3. Výběr bez vracení s vracením

Taková vánoční drobnost. U novoročenky jsme chtěli výběr bez vracení, ale nalezených šestiúhelníků jsme měli víc než nálepek. Místo nějaké formy stratifikovaného výběru jsme z pole hodnoty odebírali a jakmile bylo pole prázdné, vyplnili jsme si jej všemi dostupnými nálepkami znovu. Tím jsme zajistili, že jsou výběry náhodné, ale že se žádná nálepka neopakuje výrazně vícekrát než jiné. Základem Lua je práce s tabulkami, při jejich kopírování přebíráme jednotlivé položky (angl. deep copy), prosté „rovná se“ by nám nepomohlo.

Pojďme na školní ukázku. Z 52 čísel jich vybereme 117. Znak # nám zjistí aktuální velikost pole.

```
hodnoty={}; vsechny={}
for x=1,52 do
  table.insert(hodnoty,x)
  table.insert(vsechny,x)
end -- končí cyklus for
for k=1,117 do
  vyber=math.random(1,#hodnoty)
  io.write(hodnoty[vyber].." ") -- místo print
  table.remove(hodnoty,vyber)
  if #hodnoty==0 then -- deep copy
    for k,v in pairs(vsechny) do hodnoty[k]=v end
  end -- končí podmínka if
end -- končí cyklus for
```

Náš pokus by mohl dopadnout takto: 34 13 35 46 7 12 48 43 4 42 47 16 50 20 14 32 31 44 8 29 27 22 51 15 17 38 18 30 3 1 41 37 11 25 28 21 39 24 9 45 10 5 2 23 36 26 52 6 19 33 49 40 29 40 4 44 27 21 24 34 22 14 18 39 42 51 6 25 26 16 11 46 28 13 3 1 36 52 9 43 7 50 32 35 38 31 20 12 2 48 47 23 10 15 8 49 17 30 33 41 37 19 45 5 51 33 10 30 34 19 20 38 18 48 16 5 28.

Pozorný čtenář jistě brzy zjistí, že 39 hodnot se opakuje přesně dvakrát, 13 hodnot přesně třikrát. Obdoba by byla, když rozdáváme balíček karet, a jakmile jsme všechny karty rozdali, použijeme další balíček karet. Kdybychom rozdávali po 13 kartách, rozdali jsme karty 9 hráčům, spotřebovali bychom dva celé a jednu čtvrtinu balíčku žolíkových karet bez žolíků.

8. R + Lua + TEX

Bokem si uložíme z adresáře `pfhex-output` seznam nálepek:

```
$ ls *.png >../soubory-pfhex.txt
```

V R jsme provedli výpočty a získali jsme soubor `hexagony.csv`. Nyní si přes Lua skript tyto dva soubory načteme. Náš cíl je vygenerovat Ti*k*Zový


```
library(raster)
library(ggplot2)
library(sf)
data<-raster("tux/ossconf-2020-upravene.png")
hodnoty<-c(66,55,153,77)

barvy<-c("#426baa","#37a9e9","#999999","#4d4d4d")
mojkovo<-ggplot()
for (volim in 1:length(hodnoty)) {
  polygony <- rasterToPolygons(data, dissolve=TRUE,
    fun=function(x){x==hodnoty[volim]})
  hexiky <- polygony %>% spsample(type = "hexagonal", cellsize = 8)
  nahled <- HexPoints2SpatialPolygons(hexiky, dx = 8)
  mojkovo <- mojkovo + geom_sf(data=st_as_sf(nahled), colour=barvy[volim],
    fill=NA)
  }
mojkovo <- mojkovo + theme_void(); mojkovo
#pdf("2020-logo-ossconf.pdf"); mojkovo; dev.off()
```

5. Druhý krok s knihovnou raster

Připravili jsme si obecný rastrový obrázek. Začali jsme v TEXu ve složce `sazba` sázet soubor `pf2020.tex`:

```
\documentclass{article}
\pagestyle{empty}
\usepackage{graphicx}
\begin{document}
\centering\bf\sffamily
\resizebox{9.5mm}{!}{PF}\par2020!\par ČStS%
\end{document}
```

Získali jsme postupně pdf a poté tif soubor v linuxovém prostředí:

```
$ lualatex pf2020.tex; \
> pdfcrop --hires --margins 5 pf2020.pdf; \
> gm convert -density 300 -monochrome pf2020-crop.pdf pf2020-crop.tif
```

V R jsme užili tento skript:

```
library(raster)
library(tidyverse)
library(sf)
data <- raster("sazba/pf2020-crop.tif")
polygony <- rasterToPolygons(data, dissolve=TRUE, fun=function(x){x>0})
hexiky <- polygony %>% spsample(type = "hexagonal", cellsize = 1.5)
as_tibble(hexiky)
nahled <- HexPoints2SpatialPolygons(hexiky, dx = 1.5)
ggplot() + geom_sf(data=st_as_sf(nahled), colour="brown", fill=NA)
```

Zajistili jsem si tak, že můžeme pracovat s libovolným černobílým rastrovým obrázkem a libovolným rozlišením. V této chvíli jsme však neužili skript hexwall, ale vyexportovali jsme si nalezené souřadnice středů šestiúhelníků.

```
write.table(hexiky@coords, "hexagony.csv", sep=",", col.names=FALSE)
```

První řádky souboru `hexagony.csv` vypadaly takto:

```
"1",62.7959799161181,21.6272033299488
"2",64.2959799161181,21.6272033299488
"3",65.7959799161181,21.6272033299488
```

6. Sesypání šestiúhelníkových nálepek

Objevili jsme dva velké repozitáře s nálepkami, stáhli jsem si je přes:

```
$ git clone https://github.com/rstudio/hex-stickers.git; \
> git clone https://github.com/Bioconductor/BiocStickers.git
```

Nakopírovali jsme si png do nové složky `pfhex`, u hex-stickers to bylo rychlé. U BioStickers jsme použili pomocný skript v Lua, který nám naparsoval `README.md` a vytáhl si png soubory nálepek z podadresářů:

```
soubor=io.open("README.md")
obsah=soubor:read("*all")
soubor:close()
kam=io.open("spust.sh","w")
unicode.utf8.gsub(obsah, "<img src=\"([^\"]-%.png)\"", function(s)
  print(s)
  kam:write("cp "..s.." ../hexwall-master/pfhex\n")
  end)
kam:close()
```

Soubor jsme spustili přes `texlua` dostupný v TEXLive a vzniklý dávkový soubor přes `sh`.

Posledním úkolem bylo připravit si podklady. Připravili jsme si složku `pfhex-output` a spustili následující (šikovný administrátor si snadno převede do skriptu):


```
$ cd pfhex; \
> for file in `find -type f -iname \*.png -printf "%f\n"`; do \
>   echo $file; \
>   gm convert $file -transparent white ../pfhex-output/$file; \
> done
```

Tím jsme zajistili, že jsou soubory průhledné, nezlobí tam barevný profil ICC (ImageMagick) a můžeme operativně zasáhnout do velikosti obrázků přes parametr `resize`.

7. Luujeme

Nyní máme stavební kameny a opustíme R. V TEXu je práce s datovými soubory možná, my použijeme Lua skript a v TEXu budeme jen sázet výsledek. Je to obdoba generování HTML přes PHP či JavaScript. Kvůli čitelnosti však nemícháme Lua skript uvnitř TEXu, což je možné, ale oddělíme jej.

Lua jako skriptovací jazyk se stal neocenitelným pomocníkem v TEXovém světě. Příchod LATEXu3, nemluvě o ConTEXtu, sice umí mnohé, ale pro „obyčejné“ programátory je Lua jen jiná forma C++, Javy, JavaScriptu, Pythonu či Perlu. Ukažme si prvně střípky programování v Lua.

7.1. Výběr s vracením

Kdybychom chtěli vybrat 13 čísel od 1 do 52, můžeme to učinit takto:

```
for k=1,13 do
  print(math.random(1,52))
end -- končí cyklus for
```

Takový soubor bychom spustili přes `texlua` z TEXLive. Nabízí se možnost programu `lua`, např. z balíku `lua5.2`, případně `lua5.3` z balíku `lua5.3`.

7.2. Výběr bez vracení

Vytvoříme si prázdné pole a vyplníme jej hodnotami 1 až 52. Postupně volíme pořadí z pole a vypisujeme hodnotu na dané pozici. Tuto hodnotu pak z pole odebereme.

```
hodnoty={}
for x=1,52 do table.insert(hodnoty,x) end
for k=1,13 do
  vyber=math.random(1,#hodnoty)
  print(hodnoty[vyber])
  table.remove(hodnoty,vyber)
end -- končí cyklus for
```