Vizualizace dat

Jak zobrazit data tak, aby se na to dalo dívat

# O co jde a k čemu to je

Před vypuštěním jakékoliv vizualizace dat do světa je dobré otestovat, zda je srozumitelná a pochopitelá.

Vizualizace je datový produkt a tak k němu můžete přistupovat. To znamená:

* Chápat, kdo jsou uživatelé
* Vcítit se do jejich situace
* Zjišťovat (nebo aspoň odhadovat), co potřebují
* Produkt dělat podle potřeb uživatelů
* Produkt před vypuštěním do světa otestovat
* Pak sbírat zpětnou vazbu a iterativně zlepšovat

Základním pojmem v této oblasti je tzv. prokletí poznání (*curse of knowledge*): uživatelé vaší vizualizace nevědí to, co vy, a pro vás je velmi těžké si představit, že to nevědí. Kroky, které v tomto návodu doporučujeme, vám kromě využití dobrých praxí v oboru vizualizace dat umožní také překonat toto prokletí [viz @xiongCurseKnowledgeVisual2020].

**Případová studie**

Tento návod jsme vyvinuli a testovali s týmem OVE na MMR

[přečíst případovou studii ⟶](../pripadove-studie/asz)

# Je to pro mě?

Tento návod vám pomůže, pokud platí některé z následujících tvrzení:

* Máme signály o tom, že naše vizualizace jsou pro uživatele našich výstupů těžko srozumitelné.
* Chceme si být jistější, že nepácháme žádné vizualizační hříchy.
* Občas máme pocit, že grafy do zpráv a prezentací dáváme spíš jako signál, že pracujeme z daty, než jako skutečně informativní prvek.
* Nejsme si při vytváření grafů a tabulek jistí, jestli využíváme vhodné formy zobrazení dat.
* Nedaří se nám v týmu shodnout na tom, jaké vizualizace používat na jaká data.

Pokud si jste docela jistí v kramflecích, ale rádi byste stavěli na aktuálním poznání o tom, jak funguje vnímání vizualizací, podívejte se do [literatury](#přehledy-dobrých-praxí-a-návody).

Pokud byste rádi rozšířili spektrum vizuálních forem, které pro vizualizaci dat využíváte, můžete se [inspirovat některým z mnoha zdrojů](#konkrétní-tipy-a-sbírky-příkladů) – nebo možná potřebujete využít nějaký jiný nástroj než Excel: [podívejte se na možnosti níže](#nástroje-pro-vizualizaci).

# Jak na to

## 1. Udělejte si obrázek o uživatelích vaší vizualizace

**Vytvořte karty uživatelů**

* Kdo to je, jaký\*á je, co dělá? (jméno, věk, role, vzdělání, postoje, …)
* Čeho potřebuje dosáhnout? (sepsat analýzu / přesvědčit paní X / vysvětlit téma T, …)
* Co ho\*ji trápí, čemu čelí, s čím se potýká (“Jsem busy” / “Nerad čtu data” / “Špatně vidím”)
* Co chce od vizualizace dat
	+ Forma: přehled vs. detail, standard vs. inovace, …
	+ Účel: Zaujmout, přesvědčit, prozkoumat, hledat vzorce…



## 2. Definujte sdělení

Podle typu dat, se kterými pracujete, a informace, kterou chcete předat, zvolte formu vizualizace

Zvažte:

* co je hlavní informace, kterou chceme předat?
* co s čím srovnáváme?
* co očekáváme, že bude uživatel s vizualizací dělat? (rychle si ji prohlédne, bude v ní sám hledat informace a vzorce, bude zkoumat jednotlivé hodnoty… – )

### Je potřeba tabulka, nebo graf?

**Kdy použít tabulku**

* Porovnat údaje (point in time, území, skupiny)
* Poskytnout hodnoty a měření (frekvence/počty, procenta/podíly, indexy)
* Poskytnout souhrnné údaje (součty, průměry)
* Zobrazit hodnoty velmi rozdílných velikostí (např. tisíce vs. miliony)

I tabulka je vizualizace: její rozložení, uspořádání a grafická úprava může usnadnit nebo zkomplikovat čtenáři život a navádět k různému čtení.

**Kdy použít graf**

* Zobrazit vzorce/pravidelnosti (např. sezonalita)
* Zvýraznit trendy (např. vývoj v čase) a rozdíly
* Ukázat vztahy (např. korelace)

### Grafy: čáry, body, sloupce, koláče

**Jaké typy grafů využít pro různé typy vztahů, srovnání a souvislostí?**

* Sloupcový
	+ rozdělení
	+ řazení
	+ variabilita, odlišnost
	+ velikost
	+ časová řada
* Čarový
	+ časová řada
	+ korelace
* Koláčový
	+ část celku
* Bodový
	+ korelace

**Další prvky grafů**

* Anotační vrstva: porozumění nebo zaměření na konkrétní momenty můžete pomoct tím, že do grafu umístíte šipky, popisky nebo návod na čtení
* Osy, škály, legendy: samy nesou informaci (např. řazení) a mohou usnadnit čtení (např. vhodně umístěná legenda nebo přímé označení hodnot)
* Nedatové prvky (vodítka, pozadí): mohou napomoci čtení, ale pokud nejsou vhodně zvoleny, mohou odvádět pozornost od informací
* Titulek a podtitulek
	+ Titulek = hlavní sdělení grafu: “Nárůst počtu uchazečů o zaměstnání”
	+ Podtitulek = popis datové sady, území a časového úseku: “Vývoj počtu uchazečů o zaměstnání v Praze v období 2010-2022”
* Zdroj

Celkové uspořádání grafu a formu jeho prvků tvořte tak, abyste minimalizovali nároky na tzv. pracovní paměť, tj. aby čtenář musel co nejméně přeskakovat mezi prvky grafu, srovnávat od sebe vzdálené hodnoty a podobně. Pamatujte, že lidem docela dobře jde rozeznat z grafu celkový trend. Naopak těžší je srovnávat jednotlivé hodnoty [@franconeriThreePerceptualTools2021]. Stavte grafy tak, abyste jim tyto úkoly usnadnili. Pokud je v datech jasný trend, měl by být poznat z “tvaru” grafu; pokud čekáte, že čtenáři budou srovnávat jednotlivé hodnoty, udělejte to tak, aby byly vedle blízko, podobné, dobře vizuálně porovnatelné.

## 4. Držte se dobrých praxí

Může k tomu použít třeba checklist níže

|  |
| --- |
|  Checklist dobrých vizualizačních praxí (rozbal kliknutím) |
| Tabulky**Sdělení*** ☐ je vhodnější tabulka nebo graf?
* ☐ je zvolena adekvátní úroveň detailu (např. setiny, halíře, časové rozlišení…)?
* ☐ uvádí tabulka příliš mnoho informací?

**Uspořádání*** ☐ hodnoty jsou graficky blízko k sobě
* ☐ hodnoty jsou uvedeny ve sloupcích
* ☐ zaokrouhlení odpovídá variabilitě proměnné a usnadňuje srovnání, pokud cílem není poskytnout přesné hodnoty

**Formátování*** ☐ číselné hodnoty jsou zarovnané doprava, popř. i na desetinnou čárku, textové doleva
* ☐ hodnoty <1 jsou uvedeny s nulou před čárkou
* ☐ přebytečné čáry jsou vypuštěny (často lze vypustit svislé)
* ☐ použité písmo má tabulární číslice (stejně široké)

**Seskupení, seřazení a hierarchizace*** ☐ hodnoty jsou seskupeny podle relevance
* ☐ hodnoty, které chceme srovnávat, jsou vedle sebe
* ☐ hierarchické (podmnožinné) vztahy jsou znázorněny odsazením
* ☐ seřazení je logické (např. čas, kraje…)

**Součty, souhrny a zaokrouhlení*** ☐ součty a souhry jsou dole nebo vpravo (! pokud hlavní sdělení: nahoře nebo vlevo)
* ☐ všechny hodnoty mají stejné zaokrouhlení

**Titulky a popisky*** ☐ titulek uvádí hlavní sdělení
* ☐ podtitulek uvádí dataset, časové období a území
* ☐ tabulka je očíslovaná
* ☐ pod tabulkou je uveden zdroj ve formátu [název datasetu / zdroj] s odkazem

**Záhlaví (headers)*** ☐ záhlaví popisují data v každé kolonce
* ☐ záhlaví uvádí jednotku (! pouze pokud odlišné, v opačném případě pouze v podtitulu)
* ☐ optional, podle využití: záhlaví se neopakují, pokud to není nutné

**Fonty*** ☐ jednotný font
* ☐ minimální velikost 12pt
* ☐ sans serif font (např. Arial)
* ☐ bold pouze pro zvýraznění, žádná kursiva

Grafy**Volba typu grafu*** ☐ účel/sdělení?
* ☐ tradiční/netradiční?
* ☐ podporuje ta srovnání, která bude čtenář chtít dělat?
* ☐ usnadňuje čitelnost? (např. small multiples)
* ☐ zobrazení nejistoty?

**Encoding*** ☐ barevné škály
	+ ☐ jsou zvoleny s ohledem na sdělení - např. divergentní škála pro zobrazení odchylek od průměru
	+ ☐ řazení např. barev ve složeném sloupci odpovídá přiřozenému pořadí
	+ ☐ kvalitativní barevná škála pro kvalitativní data, souvislá pro kvantitativní, odstupňovaná pro kardinální data, popř. odstíny pro podkategorie
	+ ☐ barev je přiměřené množství
	+ ☐ nejsou použity barvy špatně odlišitelné pro osoby s poruchou zraku
* ☐ velikost: body jsou zvětšeny svou plochou, ne průměrem
* ☐ u velkého množství bodů je použita průhlednost, aby byly vidět shluky
* ☐ barvy jsou využity cíleně, např.
	+ ☐ zvýraznění klíčové kategorie
	+ ☐ červená = problém/nízká hodnota
	+ ☐ šedá pro “Ostatní”
* ☐ datumová osa:
	+ ☐ různě dlouhé intervaly jsou zobrazeny odpovídajícími rozestupy

**Titulky a popisky*** ☐ text je dostatečně velký pro dané použití (dokument vs. prezentace)
* ☐ text názvu grafu sděluje hlavní informaci
* ☐ v titulku nebo podtitulku je popsána veličina, záběr a jednotka dat
* ☐ pro popisky hodnot je použito písmo s jednotnou šířkou číslic
* ☐ popisky hodnot na ose mají přiměřené zaokrouhlení / jednotku
* ☐ volitelně: popis osy Y může být v podtitulu (oproti otočenému textu u osy)

**Legenda*** ☐ řazení odpovídá grafu a/nebo přirozenému pořadí
* ☐ pokud lze, data jsou označena přímo (např. u čarového grafu), ne legendou

**Anotace*** ☐ anotace vysvětlují klíčové datové body, trendy, změny, anomálie aj.

**Nedatové prvky: osy, vodicí linky, řazení*** ☐ řazení kategorií na osách odpovídá přirozenému pořadí a/nebo podporuje čtení (např. kategorie Jiné/Ostatní je řazena poslední)
* ☐ řazení kategorií na koláčovém grafu usnadňuje srovnání a koláč začíná na 12:00
* ☐ jsou použity smysluplně, tj. tak, pomáhaly čtení grafu a nerozptylovaly
	+ ☐ přiměřený počet, tloušťka a barva vodicích linek
	+ ☐ přiměřený počet “zarážek” (ticks) na osách
	+ ☐ přiměřené množství popisků hodnot na osách bez přehnaného opakování např. roku v datumu
	+ ☐ nechybí/nejsou přeskočeny žádné popisky na kategorické ose

**Špatné praxe*** ☐ 3D grafy
* ☐ podseknutá osa u sloupcového grafu (nezačíná na nule)
* ☐ příliš roztažená osa Y u čarových grafů (není vidět variabilita)
* ☐ dvojitá svislá osa s různými rozpětími nebo jednotkami – zvlášť pokud navozuje falešný dojem vztahu mezi proměnnými měřenými na opačných osách
* ☐ koláčový graf s více než 4 kategoriemi
* ☐ popisky osy X svisle/našikmo: nenuťte své čtenáře kroutit hlavou
* ☐ červenozelená barevná škála: nekomplikujte život lidem s poruchami vnímání barev
 |

[Stáhnout checklist ve wordu](vizualizace-checklist.docx)

## 5. Otestujte srozumitelnost a pochopitelnost grafu

### Otestujte graf na sobě:

* *Jak bych popsal/a graf nebo tabulku po telefonu?*

### Testujte graf **na ostatních:**

* Čtení: *Ukážu ti graf, komentuj prosím průběžně nahlas, jak ho postupně čteš, jak nad ním přemýšlíš, čemu nerozumíš, co si odvozuješ.*
* Sdělení: *Projdi si, prosím, graf a zkus identifikovat, co je jeho hlavní myšlenkou, sdělením.*

Pokuste se taky ukázat někomu graf na pouhých 5 vteřin a zeptejte se, co je jeho hlavní myšlenkou.

# Tipy & triky a na co si dávat pozor

Většinu pastí jsme zachytili v checklistu – použijte jej.

Pár rad pro typické situace:

1. Chci ukázat hodně dat, např. časovou řadu pro více entit
* rozložte graf do více panelů (tzv. small multiples, viz @schwabishEconomistGuideVisualizing2014)
1. Nevím, jaké vybrat barvy
* podívejte se, jestli vaše organizace nemá sestavené barevné schéma např. pro vizualizace
* pro kvantitativní škály lze doporučit palety viridis (můžete využít např. [tento generátor](https://waldyrious.net/viridis-palette-generator/))
* pro kvalitativní může pomoct nástroj [ColorBrewer](https://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3)
1. Nevím, jaký typ grafu použít

Použijte některou z pomůcek pro výběr grafu:

* [Financial Times Visual Vocabulary](https://ft-interactive.github.io/visual-vocabulary/) + [PDF](https://journalismcourses.org/wp-content/uploads/2020/07/Visual-vocabulary.pdf) + [více](https://github.com/Financial-Times/chart-doctor/blob/main/visual-vocabulary/README.md)
* [Chart Chooser](https://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf)
* [From data to Viz](https://www.data-to-viz.com/#explore)
* [Which visualisation? Stephen Franconeri](https://experception.net/Franconeri_ExperCeptionDotNet_DataVisQuickRef.pdf)
1. Chci vyrobit graf, který Excel neumí

Můžete ho zkusit hacknout, neboli nějak to omezené obejít. Pomůže [některá z příruček](#v-excelu).

Anebo využít jiný nástroj. Máte dvě možnosti: využít nějaký mocný a flexibilní programovací jazyk (typicky Python nebo R) anebo naopak některý z [online nástrojů na tvorbu grafů](#nastroje).

# Další zdroje a nástroje

## ČP.edu: vizualizace dat

V listopadu 2022 jsme jako součást série ČP.edu organizovali workshop na téma vizualzizace dat. Mluvili na něm mj. [Kristína Pšorn Zákopčanová](https://twitter.com/kris_zak) (Fakta o klimatu, iRozhlas a PAQ Research) a Benedikt Kotmel z datové platformy Golemio. Pusťte si video:

<https://www.youtube.com/watch?v=PgmduPetX5k>

## Základní zdroje

### Příručky a učebnice

Na webu lze najít několik kompletních – přitom ale dobře použitelných – příruček.

Zcela volně dostupná je *Fundamentals of Data Visualization*, zaměřená na principy, postupy a dobré praxe [@wilkeFundamentalsDataVisualization2019]. Podobná, ale s příklady kódy v jazyce R je Data Visualisation in Practice od sociologoa Kierana Healyho [@healyDataVisualizationPractical2019].

Z běžných knih lze nejvíce doporučit *Data Visualisation Handbook* [@kirkDataVisualisationHandbook2019], zaměřenou na vizualizaci jako pracovní proces; naopak Alberto Cairo v *The Truthful Art* [@cairoTruthfulArtData2016] nabízí celkové pojednání o problematice komunikace čísel a vychází ze zkušenosti novináře a akademika. Podobně šířeji téma pojímá i David Spiegelhalter v knížce *The Art of Statistics* [@spiegelhalterArtStatisticsLearning2020].

#### v Excelu

Základní rady najdete v knihách od Stephanie Evergreen [@evergreenEffectiveDataVisualization2019] a Cole Knaflic [@knaflicStorytellingDataData2015].

Průvodce využitím různých typů grafů je @schwabishBetterDataVisualizations2021.

Specificky pro Excel lze doporučit knihy of Jorge Camoese [@camoesDataWorkBest2016] a Jonathana Schwabishe [@schwabishDataVisualizationExcel2023].

|  |
| --- |
| **Kde knížky vzít**Přístup k mnoha z knih v tomto seznamu lze získat ve fyzické i online formě prostřednictvním Národní technické knihovny. |

Praktické tipy k postupu designu evaluací z pohledy UX nabízí Chris Lysy na svém webu a blogu [Fresh Spectrum](https://freshspectrum.com/), kde najdete i několik volně dostupných příruček, např. @lysyUXEvaluationData2021.

### Přehledy dobrých praxí a návody

Aktuální systematický přehled dobrých praxí vyvozených z dostupné literatury od kolektivu špiček v oboru lze najít v @franconeriScienceVisualData2021. Praktický komentář k přehledu nabízí @schwabishPracticeVisualData2021 ([PDF](https://journals.sagepub.com/doi/pdf/10.1177/15291006211057899?casa_token=EJhaEmgaHVcAAAAA:QHk1Qfmiql9QSCf1QxvpBUQnQItOQ0lrUK_Kl-WtMknLPaifxQ9CARMVOVUu8DxW0rosWUYWQbs)).

Dobré tipy “co dělat a co ne” lze najít ve starším paperu Jonathana Schwabishe [@schwabishEconomistGuideVisualizing2014].

Dále lze doporučit několik online příruček dobrých praxí:

* Renomovaný praktik Maarten Lambrechts sestavil výborného [průvodce vizualizací](https://data.europa.eu/apps/data-visualisation-guide/) na webu Evropské komise [@euacademyDataVisualisationGuide2023]
* Doporučit lze i přehledy dobrých praxí od britské statistické společnosti [@krauseBestPracticesData2023] a státní služby [@analysisfunctioncentralteamDataVisualisationCharts2022]
* Britský statistický úřad Office of National Statistics udržuje vlastní [standard vizualizací](https://service-manual.ons.gov.uk/data-visualisation/about).

Praktické tipy pro design tabulek nabízí @schwabishTenGuidelinesBetter2020 ([PDF](https://www.cambridge.org/core/services/aop-cambridge-core/content/view/74C6FD9FEB12038A52A95B9FBCA05A12/S2194588820000111a.pdf/div-class-title-ten-guidelines-for-better-tables-div.pdf))

Vizualizace má i své etické a sociální rozměry: pokud chcete zapracovat na tom, aby vaše vizualizace neposilovaly nerovnosti a stereotypy, využijte Do No Harm Guide [@schwabishNoHarmGuide2021].

## Nástroje pro vizualizaci

Pokud vám Excel nestačí a nechcete se pouštět do složitějších technologií, nabízí se několik volně dostupných nástrojů:

* [Datawrapper](https://www.datawrapper.de/) má výborný výchozí design, zvládne spoustu standardních grafů rychle a chytře. Stačí překopírovat data z Excelu a pak ladit. Používají ho často redakce. A má skvělý systém na tvorbu map, který umí pracovat s většinou českých územních jednotek, nemluvě o Evropě a světě - takže pokud potřebujete mapu, nemusíte sahat po ArcGISu nebo QGISu. Umí interaktivitu, nápovědy a další.
* [RAW Graphs](https://www.rawgraphs.io/) je rychlý vizualizátor, který nabízí typy grafů, které v Excelu nenajdete. Vložte data přes schránku a zkoušejte. Vše se děje jen ve vašem prohlížeči.
* Tableau je zdarma jen zčásti, podobně PowerBI.

## Konkrétní tipy a sbírky příkladů

Jak hacknout Excel a dostat z něj skvělé vizualizace? Poradí [tato videa](https://www.youtube.com/playlist?list=PLfv89tPxlTiX1N3xMTvkLNkp-I-8fNJDU), podobně k [Tableau](https://www.youtube.com/playlist?list=PLfv89tPxlTiVKTO5jVnx_HJCyYT_IhmvQ).

Hodně se dá naučit z tohoto [výběru zaměřeného na drobné, ale mocné prvky vizualizací](https://visualisingdata.com/2016/03/little-visualisation-design/).

Skvělý je blog [Lisy Muth](https://lisacharlottemuth.com/articles.html), mj. texty o [barvách](https://blog.datawrapper.de/colors-for-data-vis-style-guides/), [anotacích](https://blog.datawrapper.de/text-in-data-visualizations/), a [písmech](https://blog.datawrapper.de/fonts-for-data-visualization/) v grafech.

Pro inspiraci se vyplatí sledovat

* datového reportéra Financial Times Johna Burn-Murdocha [twitter](https://twitter.com/jburnmurdoch) | [web](https://www.ft.com/john-burn-murdoch)
* vizuální sloupek *The Economist* [Graphic Detail](https://www.economist.com/graphic-detail)
* [datové novináře Českého rozhlasu](https://www.irozhlas.cz/zpravy-tag/datova-zurnalistika)

## Pokročilé zdroje a odborná literatura

Pokud si chcete vybudovat systematičtější přístup k vizualizaci dat, doporučuji seznámit se s konceptem tzv. gramatiky grafických prvků - *grammar of graphics*. Asi nejlépe je vysvětlena v [online příručce Evropské komise](schwabishNoHarmGuide2021), praktické provedení ukazuje [tento tutoriál](https://www.cedricscherer.com/2019/05/17/the-evolution-of-a-ggplot-ep.-1/); technické pojednání je v [online knize](https://ggplot2-book.org/) k vizualizačnímu balíčku [ggplot2](https://CRAN.R-project.org/package%3Dggplot2) v prostředí jazyka R.

## Odkazovaná literatura