

## Co bude v dnešní přednášce

- Organizace předmětu atp.,
- proč se to učit, příklady využití,
- přelet nad probíranou látkou,
- studijní materiály: z čeho se učit a co už umět,

## 1 Organizace předmětu atp.

### Vyučující předmětu

- **Alexander Kovalenko** (cvičí a právě přednáší),
- **Pavel Kordík** (garant v pozadí, spoluautor sylabu, pečuje přednášky lidí z praxe)

### Zápočet

- **Zápočet bude postaven na vypracování domácích úkolů.**
- Úkoly budou během semestru **tři, každý za max. 15 bodů.**
- Zadání a instrukce k vypracování a odevzdání najdete na stránkách předmětu:

---

 [courses.fit.cvut.cz/BI-VZD/](https://courses.fit.cvut.cz/BI-VZD/) 

---

- Zadání prvního úkolu bude zveřejněno [zde](#).
- Vyučující Vám také může udělit až **10 bonusových bodů** dle svého uvážení ;).
- **K získání zápočtu je třeba získat alespoň 25 bodů.**

### Domácí úkoly

- Domácí úkoly budete vypracovávat v jazyce Python ve formátu Jupyter notebook (.ipynb).
- Zadání úkolu je notebook, ve kterém jsou podrobně popsány body zadání a který slouží k doplnění Vašeho kódu.
- Úkoly se budou odevzdávat přes fakultní [GitLab](#), přesné instrukce najdete v sekci [Domácí úkoly](#) na stránkách předmětu.
- Váš cvičící bude komentovat a vysvětlovat své hodnocení v rámci *issue* Vašeho repozitáře. Má ale také právo si vyžádat **osobní konzultaci nad úkolem, kde musíte být schopni Vaše řešení vysvětlit a obhájit.**

## Zkouška

- Zkouška bude mít **pouze ústní část**.
- U ústní zkoušky budou každému studentovi přiděleny **dvě otázky z předem zveřejněného seznamu**. Bude mít nějaký čas na vypracování písemné přípravy, nad kterou pak bude diskutovat postupně s dvěma přednášejícími.

## Předpokládaný FAQ

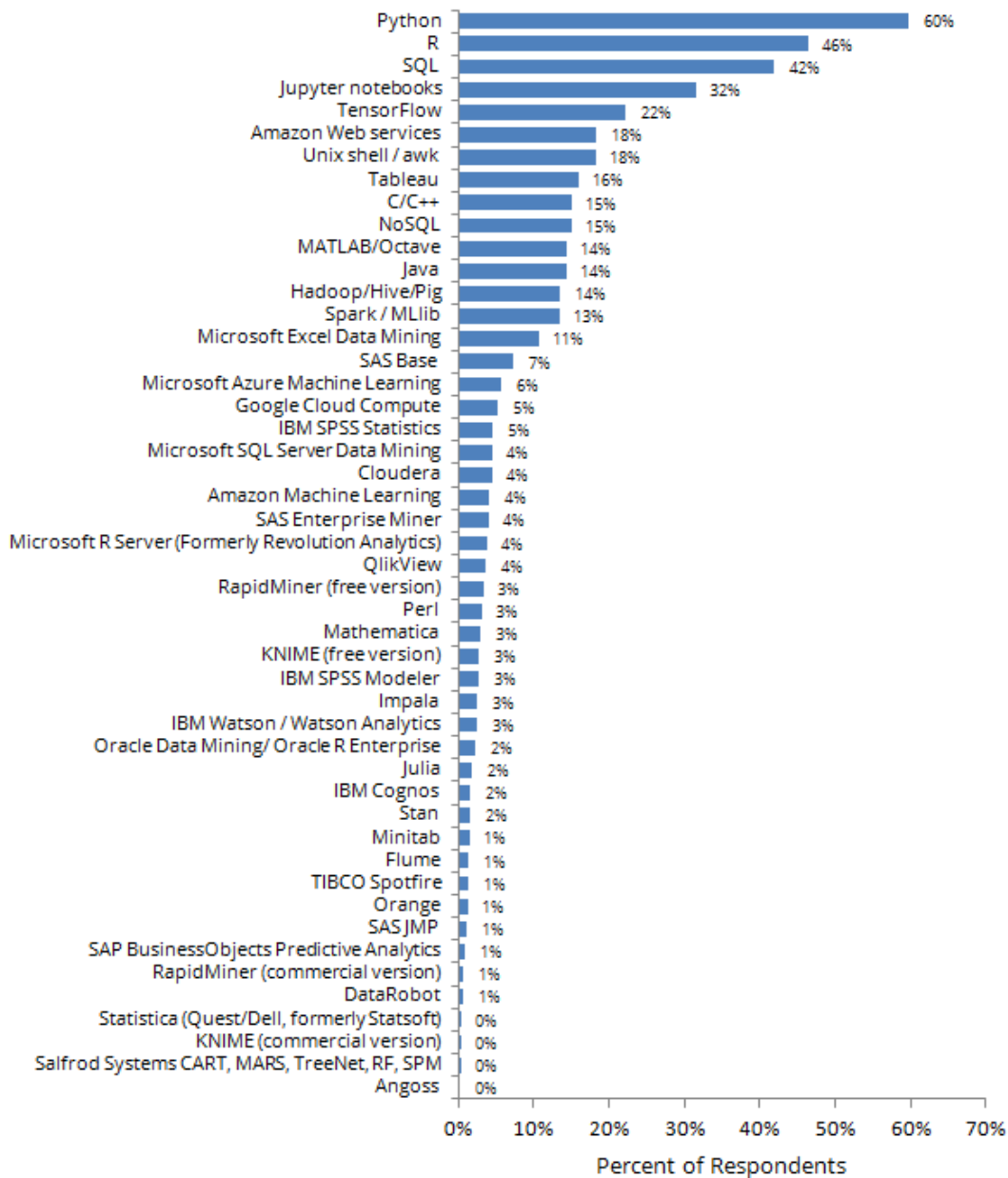
- ③ **Je u ústní části právo veta?** Po kratší diskuzi vyučujících jsme veto nezavedli, ale bylo to těsně. Můžete ovšem dostat 0 bodů ;).
- ③ **Jak dlouho bude zkouška trvat?** Budete pozvaní na konkrétní čas, na přípravu budete mít alespoň 20 minut, zkoušení pak trvá dalších max. 20 minut.

## O co v předmětu vlastně půjde

### O čem to všechno vlastně bude?

- Ve vší obecnosti: budeme se učit, jak z dat získat informace.
- Anglická *buzz words*, která popisují náplň tohoto předmětu, jsou **data mining** a **machine learning**; mezi těmito oblastmi je tenká a nezřetelná hranice.
- V tomto předmětu budeme používat jazyk **Python** a zejména balíčky, které se pro práci s daty používají.

## Data Science / Analytics Tools, Technologies and Languages Used in Past Year



Data are from the Kaggle 2017 The State of Data Science and Machine Learning study. You can learn more about the study and download the data here: <https://www.kaggle.com/surveys/2017>. Respondents were asked to indicate for work, which data science/analytics tools, technologies, and languages they used in the past year. A total of 10153 respondents answered the question.



Copyright 2018 Business Over Broadway

**O co v předmětu vlastně půjde**

**O čem to všechno vlastně bude?**

- Ve vší obecnosti: budeme se učit, jak z dat získat informace.

- Anglická *buzz words*, která popisují náplň tohoto předmětu, jsou **data mining** a **machine learning**; mezi těmito oblastmi je tenká a nezřetelná hranice.
- V tomto předmětu budeme používat jazyk **Python** a zejména balíčky, které se pro práci s daty používají.



**Příklad: prenatalní (1/4)**

- S prvním *datovým modelem* jste se pravděpodobně setkali ještě v prenatalním období Vašeho života.
- Při odhadování porodní váhy plodu pomocí ultrazvuku (tzv. SONO) se používá mj. následující model:

## Model pro porodní váhu (Shephard et al.)

$$\log_{10}(eFW) = -1.749 + 0.017 \cdot BPD + 0.005 \cdot AC - \frac{2.646}{1000} \cdot (BPD \cdot AC),$$

$eFW$  = odhadovaná hmotnost při narození,  $BPD$  = biparietal diameter (příčný průměr hlavy),  $AC$  = abdominal circumference (obvod břicha).

- Pro nás to bude **lineární regresní model** (6. přednáška), kde je
  - porodní váha  $eFW$  tzv. **vysvětlovaná proměnná** (angl. **target variable**),
  - $BPD$ ,  $AC$  a  $BPD \cdot AC$  jsou pak tři tzv. **příznaky** (angl. a často i česky **features**),
  - čísla 1.749, 0.017 atd. jsou takzvané **regresní koeficienty** (příp. váhy), obecně **parametry modelu**.

### Příklad: prenatální (2/4)

## Model pro porodní váhu (Shephard et al.)

$$\log_{10}(eFW) = -1.749 + 0.017 \cdot BPD + 0.005 \cdot AC - \frac{2.646}{1000} \cdot (BPD \cdot AC),$$

- **Jak se tento model používá?**
  1. Na ultrazvuku se změříte veličiny  $BPD$  a  $AC$ ,
  2. získaná čísla dosadíte do pravé strany vzorce,
  3. výsledek je odhad hodnoty  $\log_{10}(eFW)$  a tedy i kýžené porodní váhy  $eFW$ .
- **Kde se vzala ta divná čísla jako 2.646 apod.?** To se právě budeme učit v tomto kurzu: Jsou to parametry zvoleného modelu a ty se vždy nějakým způsobem odhadují na základě dostupných dat, to je tzv. **učení modelu**.

### Příklad: prenatální (3/4)

Jak asi pan Shephard et al. došli právě k tomuto modelu:

1. Z nějakého důvodu věřili, že příznaky  $BPD$  a  $AC$  jsou pro porodní váhu důležité. Nejspíše ale zkoušeli i jiné, ale ty buď nepřinášely velké zlepšení nebo se daly  $BPD$  a  $AC$  plně nahradit.
2. Pomocí různých testovacích procedur si jako model zvolili lineární regresní model s třemi příznaky a čtyřmi koeficienty:

$$\log_{10}(eFW) = w_0 + w_1 \cdot BPD + w_2 \cdot AC + w_3 \cdot (BPD \cdot AC).$$

3. Předchozí fázi, kdy hledáme „tvar“ modelu, se říká **ladění hyperparametrů** (angl. **hyperparameter tuning**).
4. Koeficienty  $w_i$  pak odhadli z dat o již narozených dětech, u kterých měli přesnou porodní váhu i prenatálně naměřené hodnoty  $BPD$  a  $AC$ .

### Příklad: prenatalní (4/4)

Zmíněná data, na kterých se model učil (a testoval), mohla vypadat nějak takto:

kid_id	<i>FW</i>	<i>BCD</i>	<i>AC</i>
1	číslo	číslo	číslo
2	číslo	číslo	číslo
⋮	⋮	⋮	⋮

**Pro zajímavost** (detaily v 6. přednášce):

- Označme sloupec pod *FW* jako vektor  $\mathbf{Y}$
- a vytvořme matici  $\mathbf{X}$  o čtyřech sloupcích, kde v každém řádku je vektor  $(1, BCD, AC, BCD \cdot AC)$ .
- Nejpoužívanější metoda pro výpočet koeficientů  $w_i$  je **metoda nejmenších čtverců**, ta nám říká, že

$$(w_0, w_1, w_2, w_3)^T = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

- Tento vzorec jsme získali vyřešením jistého optimalizačního problému (a.k.a. *hledání extrémů*), což je velice častý případ: **učení modelu = optimalizace!**

### Supervizované vs. nesupervizované učení

- Předchozí *prenatalní* příklad je typickou ukázkou **supervizovaného učení** (učení s učitelem, angl. **supervised learning**).
- Tím „učitelem“ jsou zde známé hodnoty porodních vah u dětí, což je veličina, kterou se snažíme pomocí modelu *predikovat* resp. pochopit, na čem závisí.
- Někdy takovou veličinu ale ani nemáme a prostě se v datech pokoušíme nějak vyznat a najít jejich skrytou strukturu.
- Takovým problémům se říká **nesupervizované učení** (učení bez učitele) a typickým příkladem je **clusterování** dat (téma 4. přednášky).

### Příklady nesupervizovaného učení

- Problém clusterování je velice obvyklý v praxi.
- Pokud máte například e-shop (nebo banku, nebo telefonního operátora), chcete se vyznat ve svých zákaznících, o kterých máte nasbíraná různá data (tzv. *customer segmentation*).
- Můžete tak hledat např. podmnožinu „nejlepších“ zákazníků, kterým má cenu věnovat speciální péči. Nebo naopak skupinu, která potřebuje k polepšení pomoci nějakou reklamní akcí (cílení reklamy je velký byznys).
- Do nesupervizovaného učení také (obvykle) spadá i **detekce anomálií** (angl. **anomaly detection**).
- Např. banka se snaží najít podezřelé transakce (fraud detection, ochrana proti zneužití karty, atp.).

## Další příklady: doporučovací systémy

- Dalším příkladem problému řešeného pomocí zkoumání dat je tzv. **doporučování** (angl. **recommendation**).
- Například: vlastníte-li e-shop (příp. internetový časopis, iTunes, Netflix atp.), snažíte se na základě dat o zákaznících a zejména zákazníkovi, který právě prohlíží Vaše stránky, odhadnout, co by si tak mohl ještě chtít koupit (přečíst, podívat, poslechnout) a to mu ukázat.

### Doporučeno přímo pro Vás

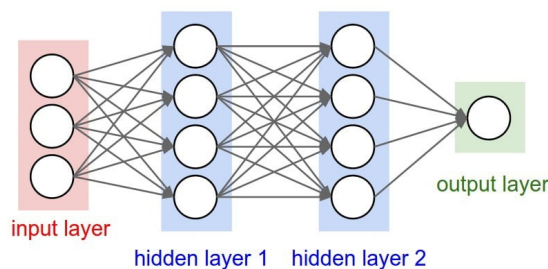


The screenshot displays a recommendation system interface with four product cards. Each card includes a product image, a star rating, the product name, and the price. The prices are shown in red text, with the original price crossed out for the Dell server.

Product	Rating	Price
Smart Cover iPad 2017 Charcoal Gray	★★★★★	1 149 Kč
Speck Balance Folio Black/Grey iPad 9.7" 2017	★★★★★	1 099 Kč
Sonos PLAY:5-2. bílý	★★★★★	15 490 Kč
Dell OptiPlex 3050 SFF	★★★★★	14 890 Kč <del>11 390 Kč</del>

## Další příklady: data bez jasných příznaků

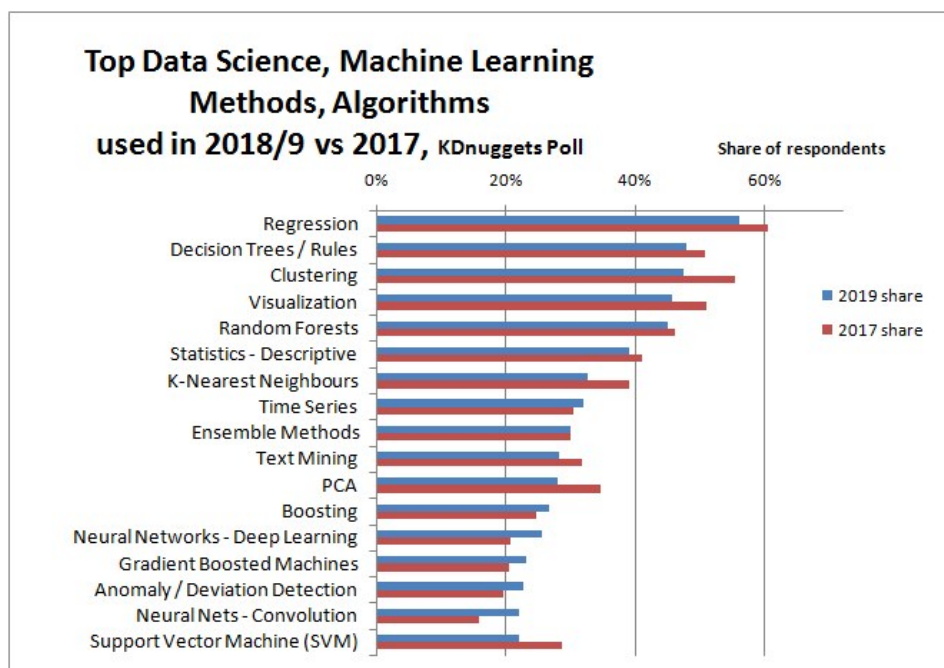
- Data ale vždy nemusí mít formu tabulky. Může se jednat o obrázky, videa, časové řady, dlouhé texty atp., ze kterých je těžké získat pro modely příznaky.
- V takovém případě si musíte dát práci a nějaké příznaky z dat vydolovat (tzv. **feature extraction**).
- Nebo použijete algoritmy a metody, které si příznaky vytvářejí samy automaticky.
- Mezi takové metody patří (čím dál populárnější) umělé **neuronové sítě** (angl. artificial **neural networks**, ANN)
- ANN se používají k všemožným úkolům (překlady, detekce objektů v obrázku videu, hraní GO, clusterování, detekci anomálií, ...).



## 2 Náplň předmětu

### Metody a algoritmy

- Cílem tohoto kurzu je naučit Vás základy dané problematiky, a tedy se zaměříme především na základní metody a algoritmy a na řešení problémů spojených s jejich použitím.
- Ovšem i úplně základní metody stále patří k nejpoužívanějším v praxi!



[zdroj: [www.kdnuggets.com](http://www.kdnuggets.com)]

### Témata přednášek (orientační)

1. Úvodní a přehledová přednáška
2. Rozhodovací stromy
3. Ensemble metody (náhodný les / Adaboost / Bagging / Bootstrapping)
4. Clustering (K-means, hierarchical clustering)
5. Lineární regrese
6. Logistická regrese
7. Redukce dimenzionality (SVD, PCA)
8. NLP = natural language processing
9. Vizualizace
10. *host z praxe???*
11. *host z praxe???*

### 3 Z čeho se učit a co už umět

#### Z čeho se učit

- Studijní materiály tohoto předmětu by Vám měly stačit pro jeho absolvování: jsou to prezentace k přednáškám (vč. jejich handout formátu), Jupyter notebooky ke cvičením a svým způsobem i úkoly.
- Probírané metody patří ke klasické látce a lze k nim najít nepřeborné množství videí a textů.
- Pěkným zdrojem je [dokumentace](#) ke knihovně `scikit-learn`, kterou budeme pro aplikování probíraných modelů především používat.
- Užitečným zdrojem (nejen) zajímavých datasetů je server [www.kaggle.com](http://www.kaggle.com). Tam najdete uživateli vytvořené příklady použití různých metod, tipy a triky, soutěže (i o docela slušné peníze) atp. **Silně doporučujeme se tam zaregistrovat.**



## Co už byste měli umět (1/2)

- Předpokládáme, že už umíte programovat, ale nepředpokládáme znalost Pythonu.
- Naučit se základy Pythonu, hlavně specializovaných knihoven, je součástí tohoto kurzu.
- O některých metodách už jste slyšeli v *BI-PST: Pravděpodobnost a statistika*, ale fakticky o látce nemusíte vědět nic ;).
- Předpokládáme znalost lineární algebry (*BI-LIN*) a matematické analýzy (*BI-ZMA*, *MI-MPI*), neb tabulky jsou matice a strojově učit znamená optimalizovat!!!
- Také předpokládáme základní znalosti pravděpodobnosti a statistiky z *BI-PST: Pravděpodobnost a statistika*

## Co už byste měli umět (2/2)



### Machine Learning in a nutshell

[zdroj: [xkcd.com/1838/](https://xkcd.com/1838/)]

Tak jsme si popovídali a teď začneme naostro!

## ChangeLog

Verze	Datum	Autor	Log
2.0	16.2.2021	KK	Úprava pro první běh BI-VZD.
1.0	2.10.2018	KK	Výchozí verze pro rok 2021/2022.