

NÁSKOK  
DÍKY  
ZNALOSTEM

PROFINIT

# Big Data Science

Petr Paščenko

18. 12. 2018

# Osnova

1. Co je Data Science
2. Statistika
3. Strojové učení
4. Vizualizace dat
5. Data Science úlohy
6. Metodika Data Science Projektu
7. Role Big Dat v Data Science
8. Podobnosti a vztahy
9. Detekce podvodů v Internetovém bankovníctví

# Co je Data Science?

- › Data Science je spojení 4 disciplín
  - Statistika
  - Informatika
  - Strojové učení
  - Vizualizace



## › Základní otázka:

- Je za tím něco víc, nebo je to jen náhoda?

## › Práce s nahodilostí

- klasifikace, kvantifikace, simulace, predikce

## › Jiný pohled: entropie

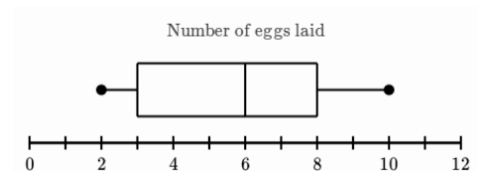
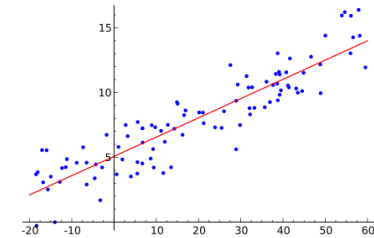
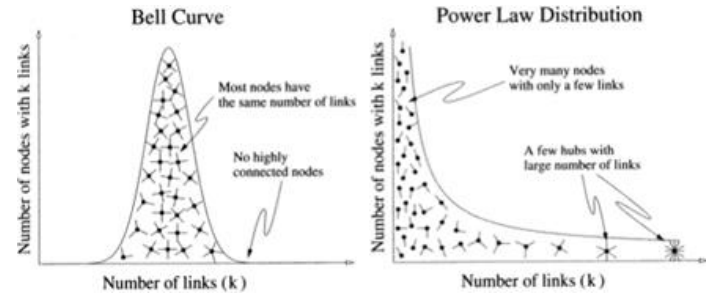
- Kolik informace je v datech? Jaký je podíl signálu a šumu?

## › Příklady statistických úloh

- Kolik osob můžu pustit do loďky / výtahu / letadla?
- Jsou muži chytřejší než ženy?
- Kdo by vyhrál volby, kdyby se konaly dnes?
- Má rodinné zázemí vliv na schopnost klienta splatit půjčku?
- Měl bych skočit z mostu po pozitivním HIV testu?

## › Klasická (frekventistická) statistika potřebuje velké soubory dat

## › Bayesianisté jsou s více daty přesnější



# Strojové učení

- › Základní cíl:
  - Napodobit lidský mozek v rutinních činnostech
- › Práce s nestrukturovanou informací
  - Obrázky, zvuky, videa, volné texty, sítě...
- › Co je pozitivní?
  - Víme, že to jde. Hledáme pouze cestu.
- › Příklady ML úloh
  - Strojový překlad mezi jazyky
  - Rozpoznávání obrázků (MNIST)
  - Porozumění mluvenému slovu
  - Autonomní robotika (auta, drony)
  - Analýza sociálních sítí
- › Nutnou podmínkou pro strojové učení jsou obrovské datové sady
- › Modely jsou v porovnání se statistikou podstatně komplexnější

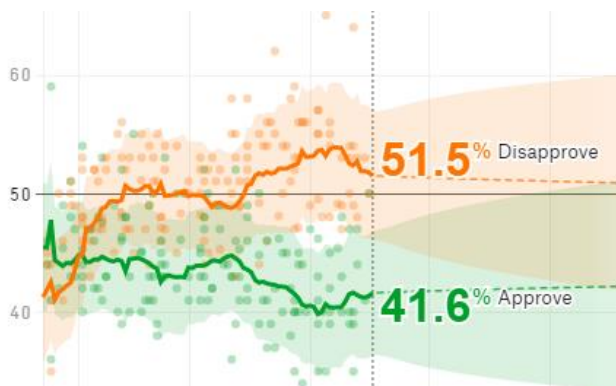
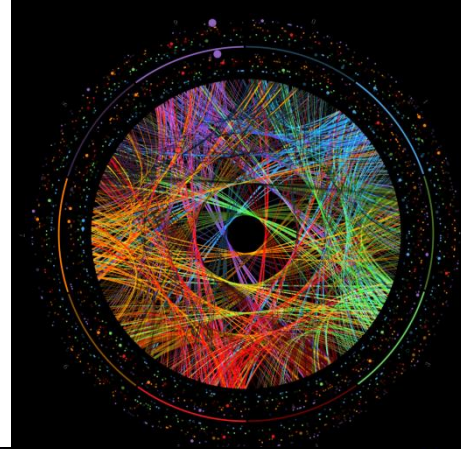




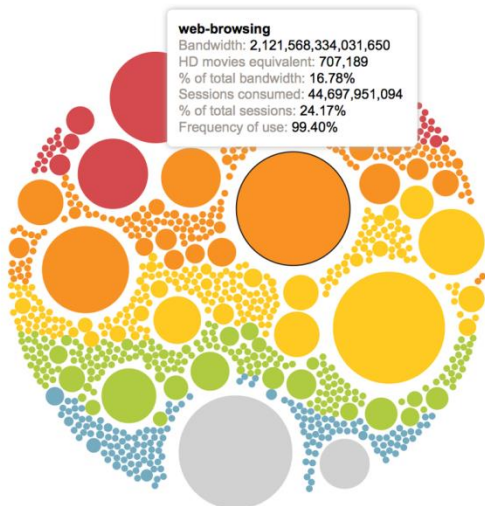
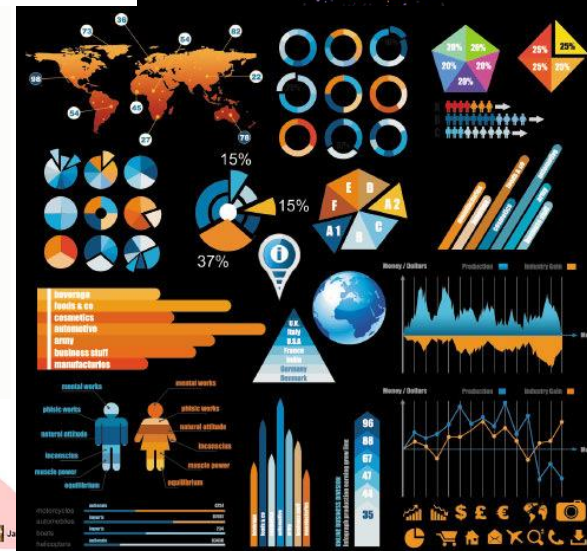
# Vizualizace

## › Základní úkol:

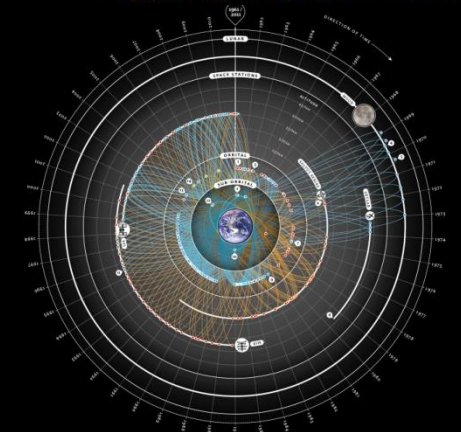
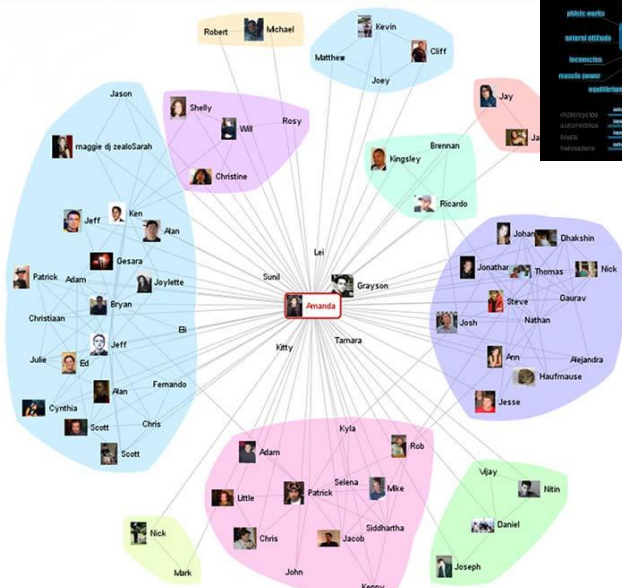
- Zprostředkovat komplexní multi-dimenzionální informaci člověku



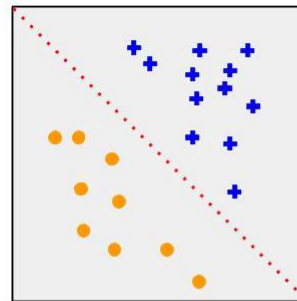
## The Wealth & Health of Nations



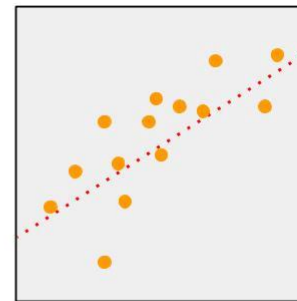
Size of circle indicates total bandwidth occupied



# Typologie DS úloh



Classification



Regression

## › Klasifikace

- Zařazení objektu do specifické diskrétní třídy
- Je na obrázku pes či kočka? Udělit nebo zamítnout hypotéku? Mám HIV?

## › Regrese

- Odhad konkrétní hodnoty (nebo intervalu) cílové proměnné
- Váha na základě výšky. Kolik km ještě ujedu? Jak daleko je objekt?

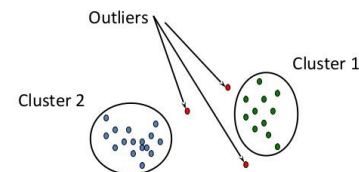
## › Klastrování / Segmentace

- Shlukování objektů a hledání typických zástupců jednotlivých skupin
- Zákaznické segmentace, sociální skupiny, funkční skupiny slov, atd.

## › Detekce odlehlých pozorování

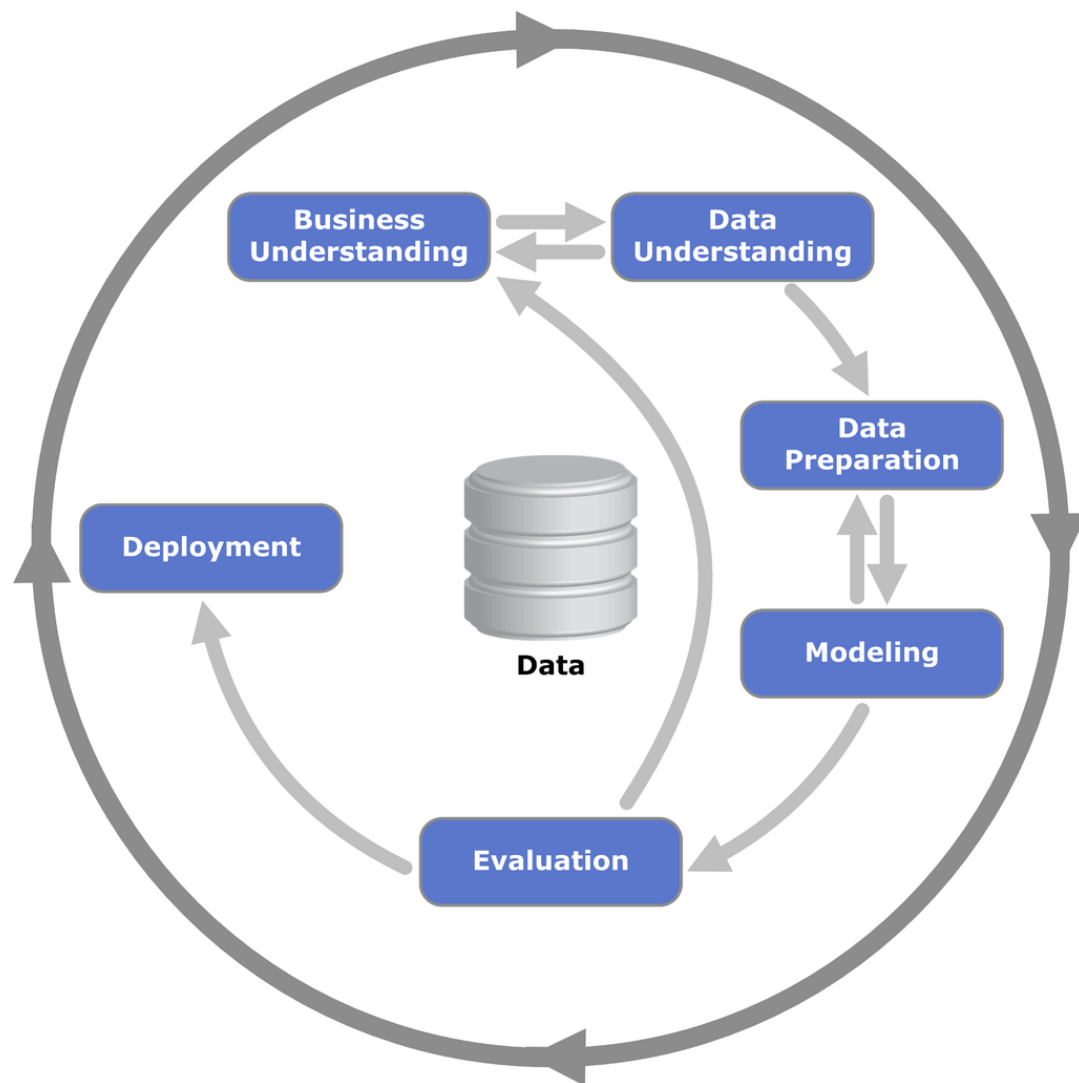
- Rozpoznání netypických objektů a kvantifikace jejich odlehlosti
- Diagnostika výrobků, bezpečnost a detekce fraudu

### OUTLIERS ANALYSIS



# Metodika pro Data Science CRISP-DM

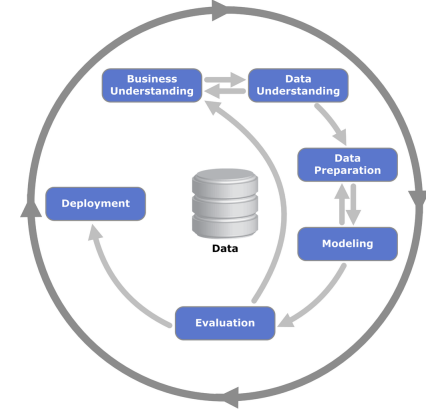
## Cross Industry Standard Process for Data Mining





# Metodika pro Data Science CRISP-DM

## Cross Industry Standard Process for Data Mining



### 1. Business Understanding

- Co zákazník potřebuje? Jaká k tomu má data? Jak se pozná úspěch?

### 2. Data Understanding

- Posbírání dat, exploratorní analýza, kvalita dat, první testy hypotéz.

### 3. Data Preparation

- Konstrukce datové sady pro modelování. Sestavení, transformace a výběr příznaků, redukce dimenzionality, atd.

### 4. Modeling

- Aplikace modelovacích technik, výběr modelu, kalibrace parametrů, testování výkonnosti modelu.

### 5. Evaluation

- Vyhodnocení úspěšnosti modelu vzhledem k věcným kritériím.

### 6. Deployment

- Produkční nasazení modelu v datovém workflow zákazníka. Vyřešení administrace, údržby, zaškolení, rekaliibrace...

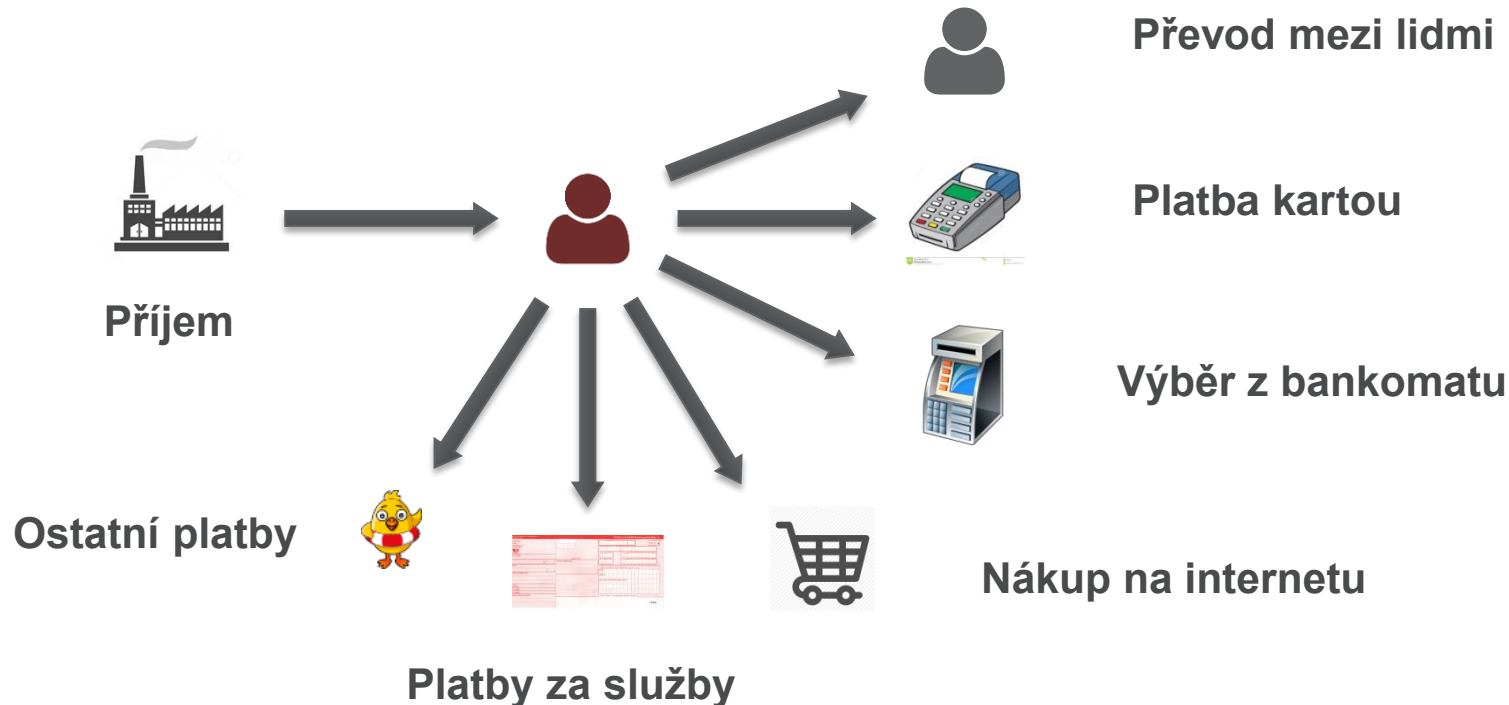
# Role Big Data technologií v Data Science

- › Více dat
  - Větší modelovací sada
    - více objektů, od výběru k celé populaci
  - Širší modelovací sadu
    - více příznaků, podrobnější příznaky (vteřinová měření atd.)
  - Delší historii
    - data za více předchozích let v plné granularitě
- › Větší výpočetní výkon
  - Pokročilé modely
    - možnost učit komplikované nelineární modely (konvoluční neuronové sítě)
  - Komplexnější příznaky
    - multimédia, sekvence, texty, atd.
- › Od příznaků k podobnostem
  - Dáme mu úvěr, protože: a) hodně vydělává, b) podobní lidé úvěry platí
- › Od podobností ke vztahům
  - A se zná s B, protože spolu chodí na oběd

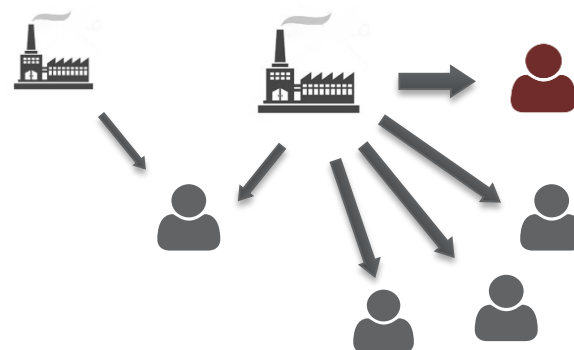
# Podobnosti a vztahy

# Analýza Finančních Transakcí pomocí BD

- › Vytváříme vyladěné modely pro retailové banky
- › Vstup – finanční transakce
- › Výstup – využitelné informace o klientovi, příznaky, události,
- › Cílem je obohatit stávající obchodní proces o novou znalost



# Salary detector



- › Vstup
  - Finanční transakce typu firma - klient
- › Výstup: Identifikované vztahy zaměstnavatel – zaměstnanec
- › Obchodní využití
  - Rizikové skóre, detekce událostí, podobnosti (c2c/b2b),...
- › Principy
  - Detekce transakčních vzorců, text mining, pokročilá statistika
- › Vysoká přesnost i pro
  - Krátké úvazky – délka nepřesahující 3 měsíce
  - Nestandardní úvazky (částečné úvazky, práce na živnost, atd.)
  - Firmy s malým počtem zaměstnanců



# Detekce domácnosti – Banka/Telco

## › Vstup

- Klientské transakce – banka (c2c, karetní operace,...)
- Informace ze sítě – telco (cdr, lokace, billing)
- Základní demografie (věk, pohlaví, adresa, příjmení,...)

## › Výstup

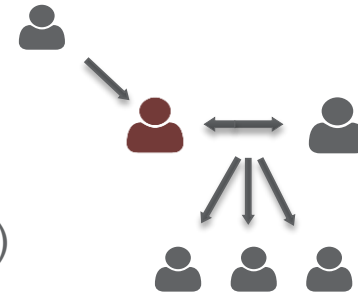
- Identifikace členů domácnosti a rodinných vztahů

## › Obchodní využití

- Rodinný marketing, robustní rizikové skóre,...

## › Principy

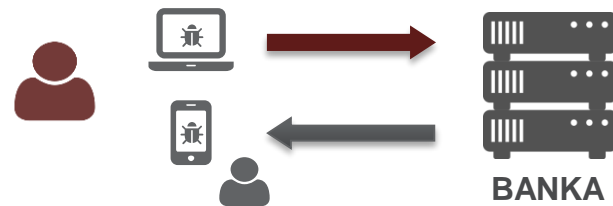
- Detekce transakčních vzorců, analýza interakcí, text mining





# Detekce Fraudu v Intenetovém bankovníctví

# Shrnutí úlohy



- › Detekce přístupu pod falešnou identitou s cílem vykrást účet
- › Scenář: podvodník překonal 2FA
- › Vstup
  - Běhová data z online bankovníctví (sekvence akcí v klientském sezení).
- › Výstup
  - Identifikovaná fraudulentní sezení
- › Principy
  - Velmi složitý problém, podíl fraudů cca 1:120 000
  - Vyžaduje vícero zřetězených sítí
  - Pokročilé statistické modely (detekce lokálních odlehlostí)
- › Nastavitelná přesnost (TP/FP), např.: TP: 50% for FP: 0.3%

# Klasifikační úloha

## › Vstupní data

- Akce klientů

ID	SESSION ID	DATETIME	ACTION	AMOUNT	RESULT
1234567890	vs3T ... dGpf	2015-04-03 13:03:58	112		0
1234567890	vs3T ... dGpf	2015-04-03 13:03:58	130		0
1234567890	vs3T ... dGpf	2015-04-03 13:04:14	1248		0
1234567890	vs3T ... dGpf	2015-04-03 13:04:14	120	12400	530
1234567890	vs3T ... dGpf	2015-04-03 13:07:21	530		0
1234567890	vs3T ... dGpf	2015-04-03 13:07:38	120	12400	0
1234567890	vs3T ... dGpf	2015-04-03 13:09:03	68		0

## › Příznakový vektor

- Statistiky session
- Délka, čas na akci, ...

## › Model

- Klasifikátor

## › Výsledek

- Ano / Ne



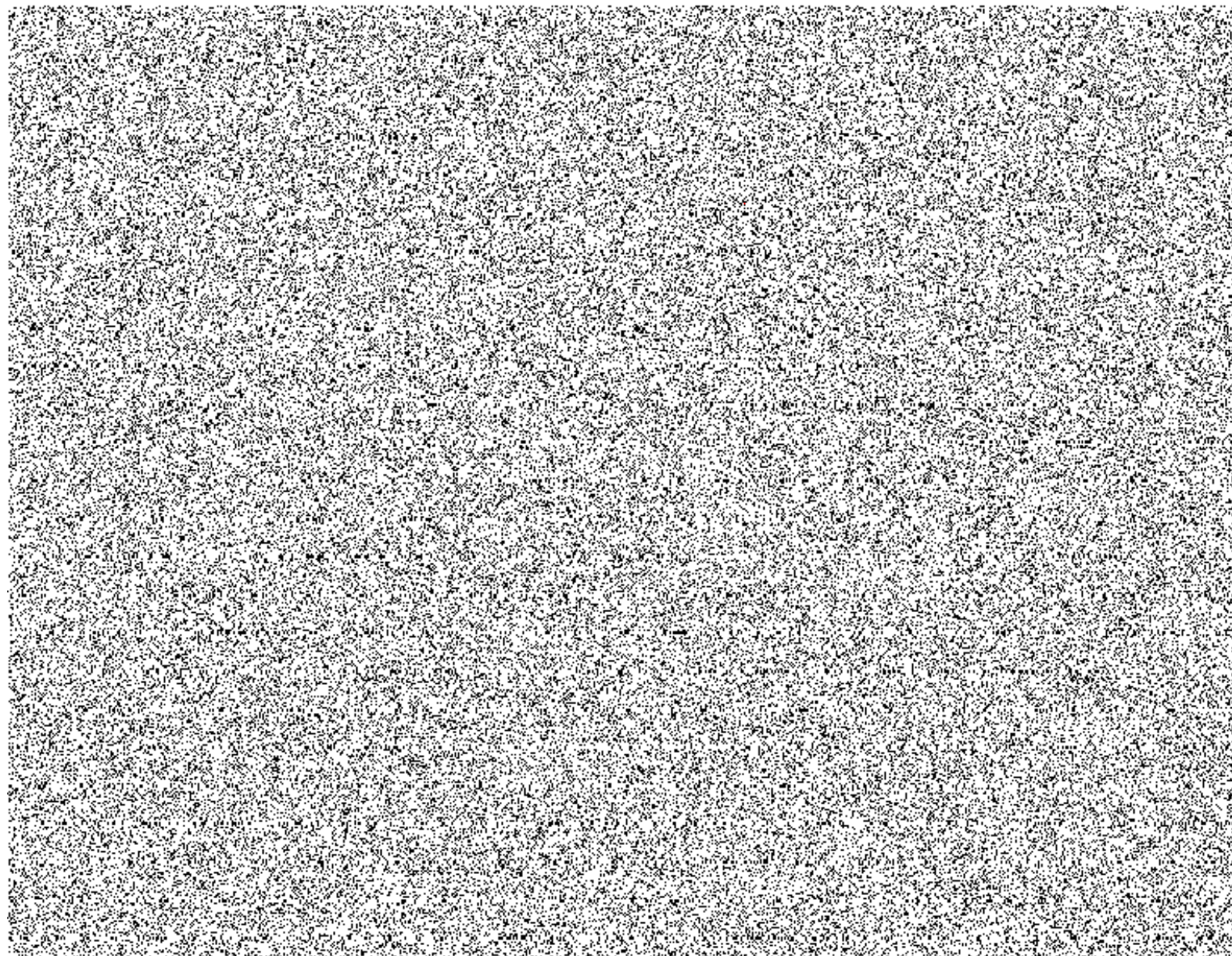
# Intenetové bankovníctví

- › Uvažujme banku s milionovou klientskou bází
- › Každý klient provede denně v průměru jednu návštěvu v IB
- › Denně průměrně 1 000 000 session
- › Z toho zhruba 12% session s platbou
- › 120 000 session s platbou
- › Denně v průměru 1 fraud
- › To není moc ;-)





1:120 000



# Co nefunguje

- › Klasifikátory učené z dat
  - Nevyvážené třídy
  - NE s úspěšností 99,999%

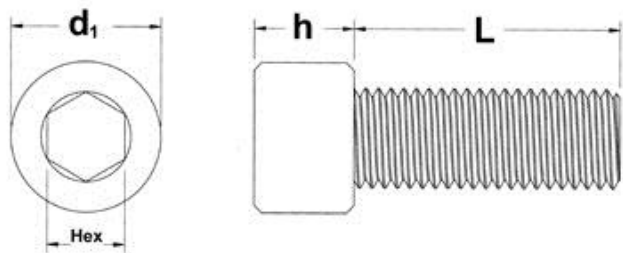


- › Popíšeme typický útok
  - Neexistuje typický útok
  - Příprava na minulou válku



# Detekce anomali

- › Podvodník se chová jinak než klient
- › Nevíme jak, ale jinak
- › Jak poznáte, že někdo nebo něco je divný?
- › Lidé nejsou šroubky



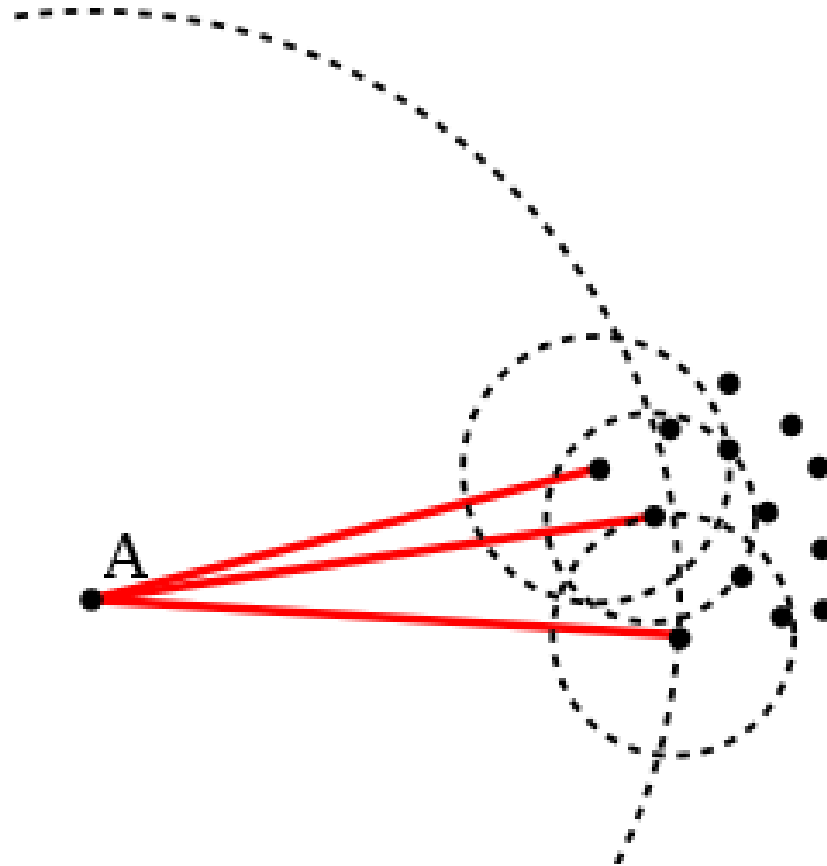
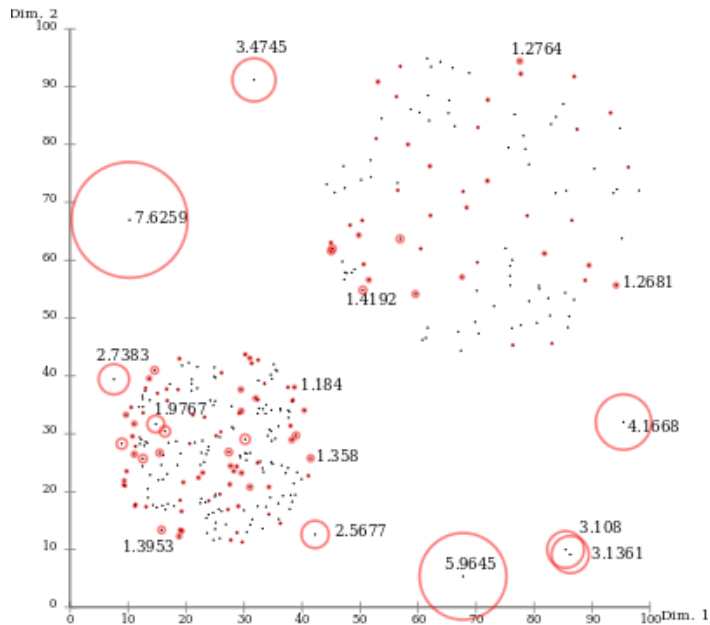
VS





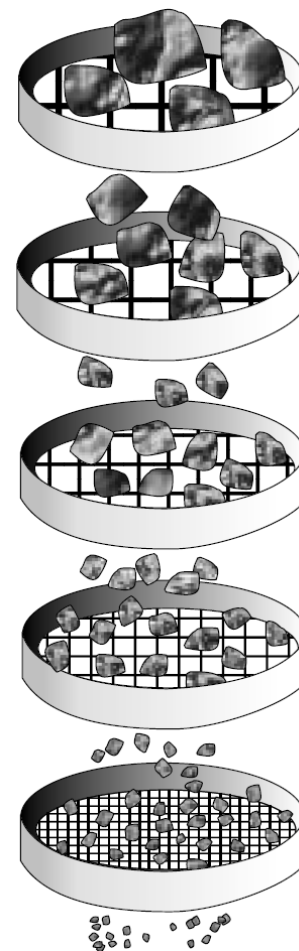
# Detekce lokálních anomalií

- › Jak si subjekt stojí vůči svému okolí



# Shrnutí postupu

- › Pro každou session spočteme příznakový vektor
  - Příznaky identifikovány na základě datové analýzy
- › Porovnání session s ostatními session daného klienta
- › Identifikace podezřelých session k prověření
- › Jak hodnotíme výsledek
  - True positive – kolik najdeme fraudů
  - False positive – kolik musíme prověřit session
- › Podstatnější je false positive
  - Limitovaná lidská kapacita
- › Zřetězení více sít
  - Jednoduché heuristiky
    - IP adresy, protiúčty
  - Detekce lokálních odlehlostí
- › Paralelizace





# Implementační realita

- › Pro každou session spočteme příznakový vektor
  - 100 session za sekundu
- › Porovnání session s ostatními session daného klienta
  - až 1000 předchozích session
  - to znamená načíst z databáze 100k řádků za sekundu
  - to znamená přenést po síti cca 20 MB za sekundu
  - to znamená spočítat 100M porovnání za sekundu
- › Překračuje kapacity konvenčního řešení
- › Úloha je naštěstí snadno paralelizovatelná
  - potřebujeme jen předchozí session daného klienta
  - distribuce záznamů podle klientského čísla
  - distribuované vyhodnocení – vrací se jen výsledek

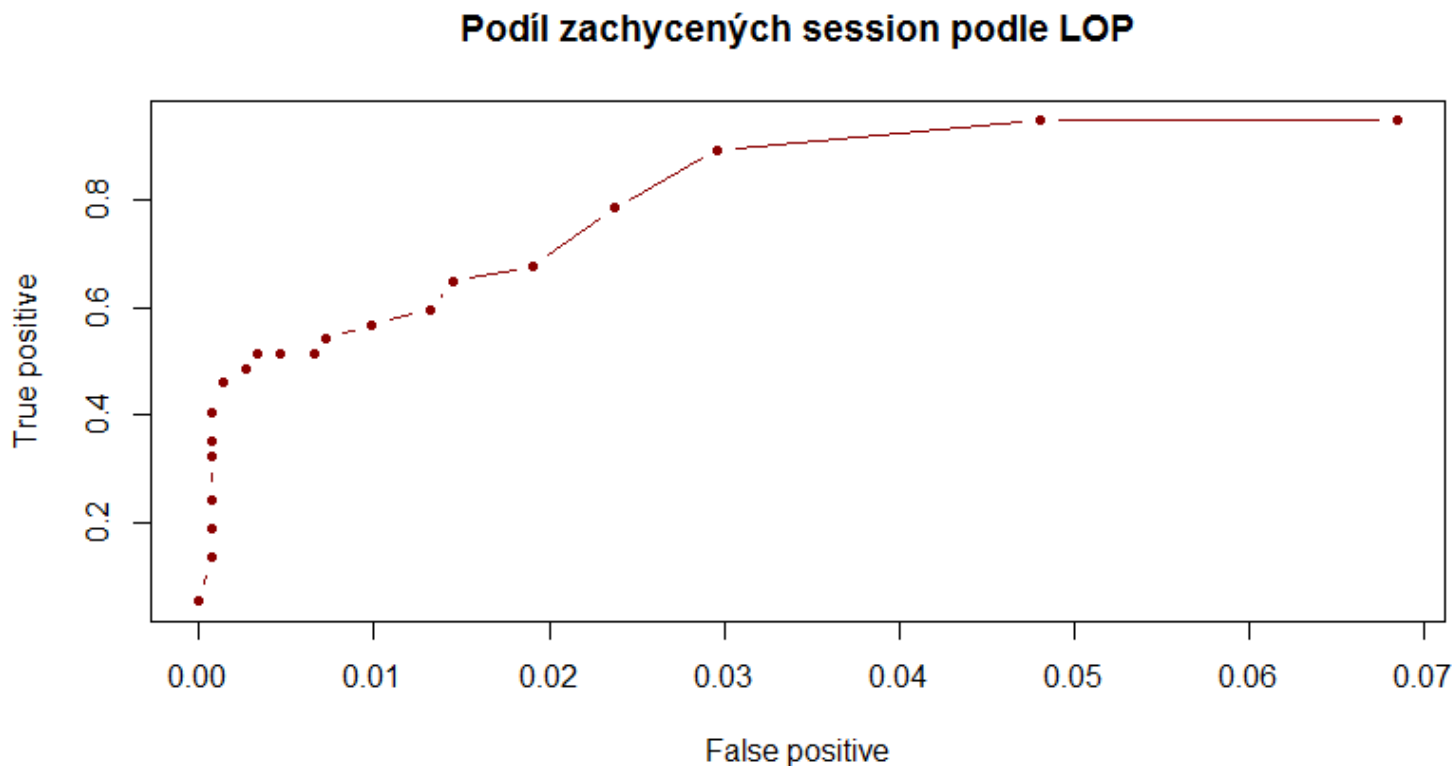


DWH

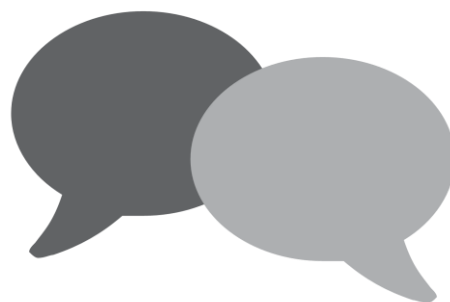


HADOOP + SPARK

# Výsledky



- › Pro nalezení 50% podvodů je třeba prošetřit cca 300 transakcí denně
  - Při 120 000 session s platbou denně



**Dotazy**