Big data vs. Don Corleone

aneb Hollywoodské filmy očima datového analytika

Absolventská středa, 7.4.2021

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

BizMachine

Máte doplňující otázky?

Eliška Přichystalová, HR, R&R

eliska.prichystalova@bizmachine.com

+420 731 020 506

Martin Lucký, spoluzakladatel, prezentující

martin.lucky@bizmachine.com

+420 606 620 419

BizMachine Základní informace

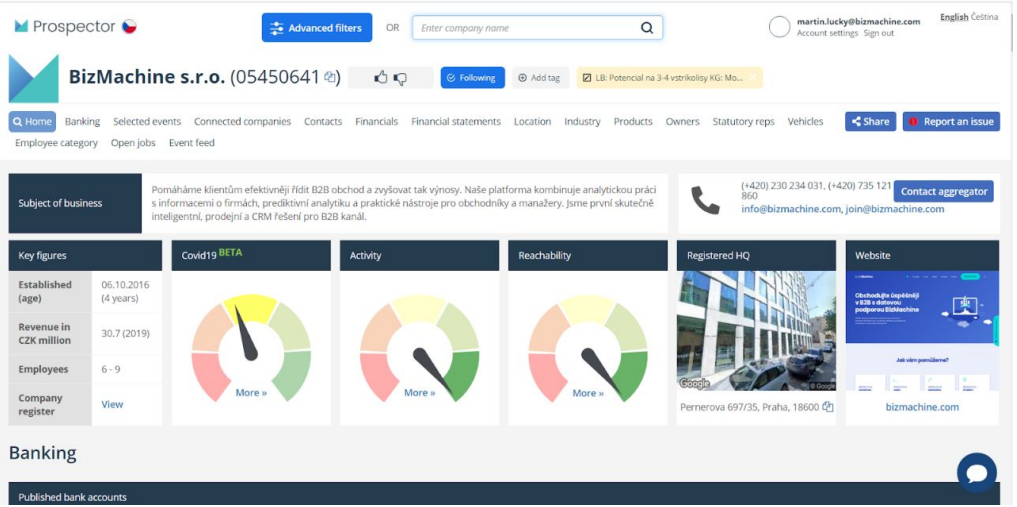
Založena:	říjen 2016
www:	www.bizmachine.com
LinkedIn:	LinkedIn
Sídlo:	Prague, CZE

Základní data 2020

Výnosy:	CZK ~43 mil
Zisk:	CZK ~4 mil
Tým:	~40 (>2/3 tech*)
Zákazníci:	>200 společností

Od roku 2016 pravidelně sbíráme a zpracováváme přes 100 zdrojů dat o firmách při aplikaci širokého spektra analytických metod.

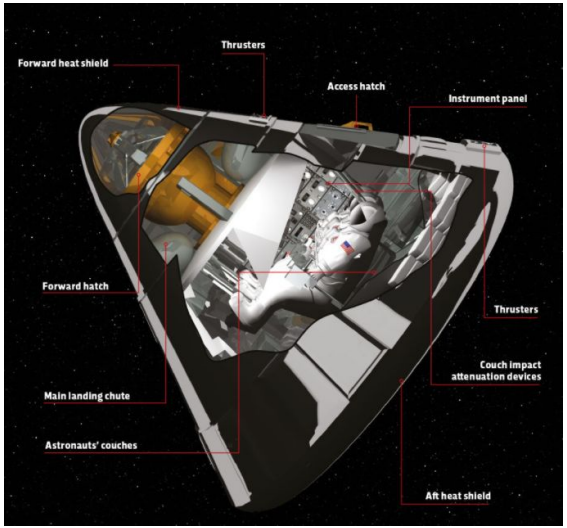
Pro více informací, navštivte https://bit.ly/BizMachine



BizMachine

* Data, vývoj, produkt

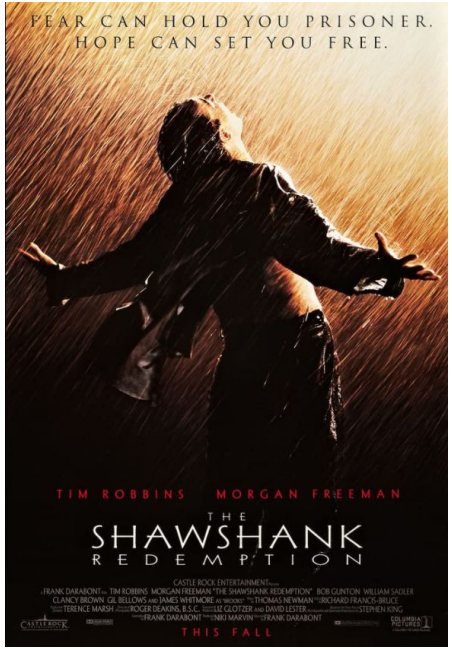
20.7.1968 - první člověk na Měsíci

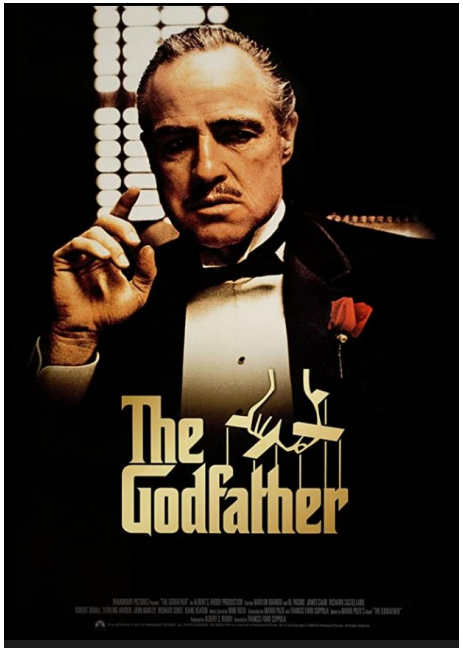


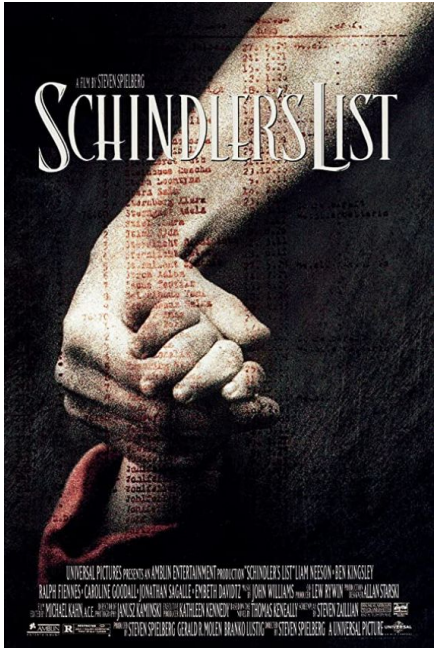
14.12.1972 - Eugene Cernan, zatím poslední člověk na Měsíci

A dál…?

BizMachine







Velká americká studia natočila přes 8 000 hollywoodských filmů a lidstvo vytvořilo zhruba 170 zetabytů (trilionů gigabytů) dat dostupných na internetu

BizMachine

Zdroje: Odhad na základě Hollywood's dominance of the movie industry: How did it arise and how has it been maintained? by Jonathan Derek Silver 2007; IDC/statista

Trocha teorie na začátek: co znamenají “big data”

“Big data” (neboli velká data) jsou data v **obrovských objemech**, která jsou generována tak **velkou rychlostí**, že je tradičními metodami nestíháme analyzovat. Velká data jsou velmi **různorodá** a sestávají z datasetů s velmi **rozdílnými hodnotami**. V originálu: (i) volume, (ii) velocity, (iii) variety and (iv) veracity.
Andreas Kaplan
Pár čísel pro kontext:

1. Každý člověk v roce 2020 vygeneruje přes 1,5 MB dat za sekundu
2. Podle IDC jsme jen v roce 2019 vytvořili zhruba 40 trilionů gigabytů dat dostupných na internetu
3. Každý den se na Google provede více než 3,5 miliardy vyhledávání
4. Podle výzkumu NewVantage Partners z roku 2018 mezi Fortune 1 000 společnostmi, 97.2 % z nich aktivně investuje do velkých dat a AI
5. Přes 80 % dat je nestrukturovaných

Andreas M. Kaplan

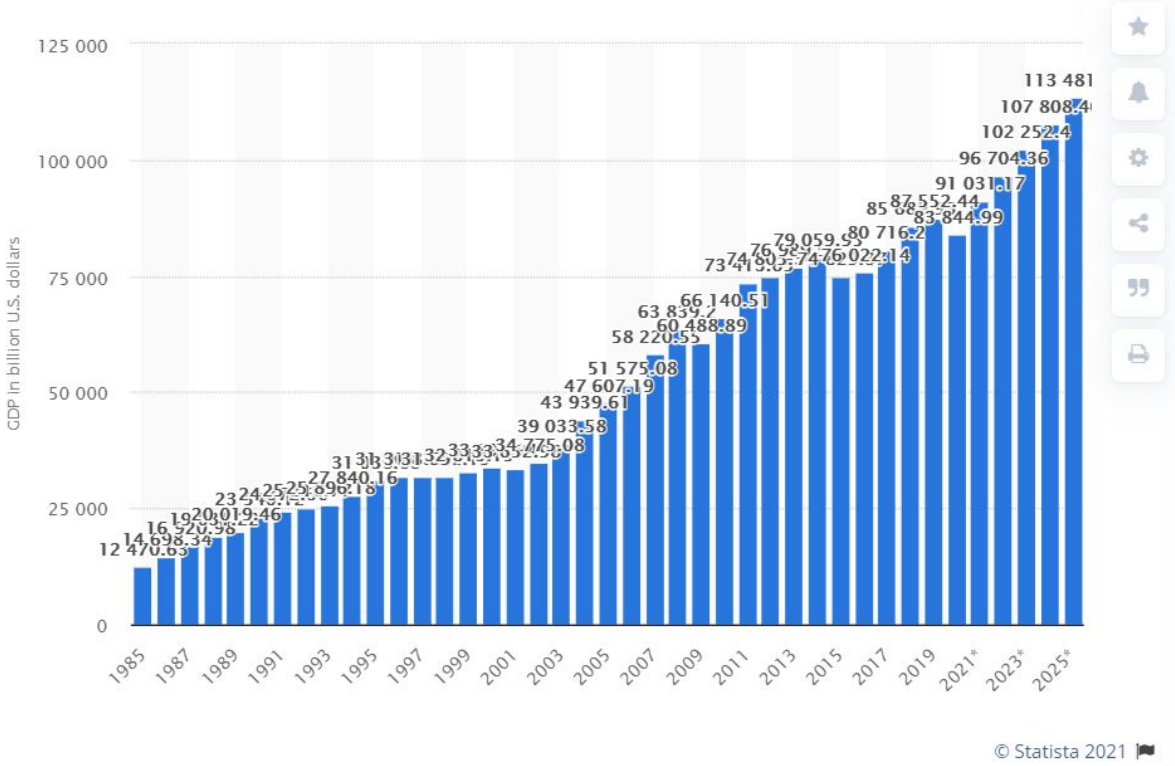
Andreas M. Kaplan

Born	October 5, 1977 (age 43) Munich
Nationality	German
Institution	ESCP Business School
Field	Marketing
Alma mater	École Nationale d'Administration, ESCP, HEC Paris, Ludwig Maximilian University of Munich; Sorbonne; University of Cologne

https://en.wikipedia.org/wiki/Andreas_Kaplan

Zdroje: Forbes, IDC, NewVantage Partners, Domo inc

Světové HDP

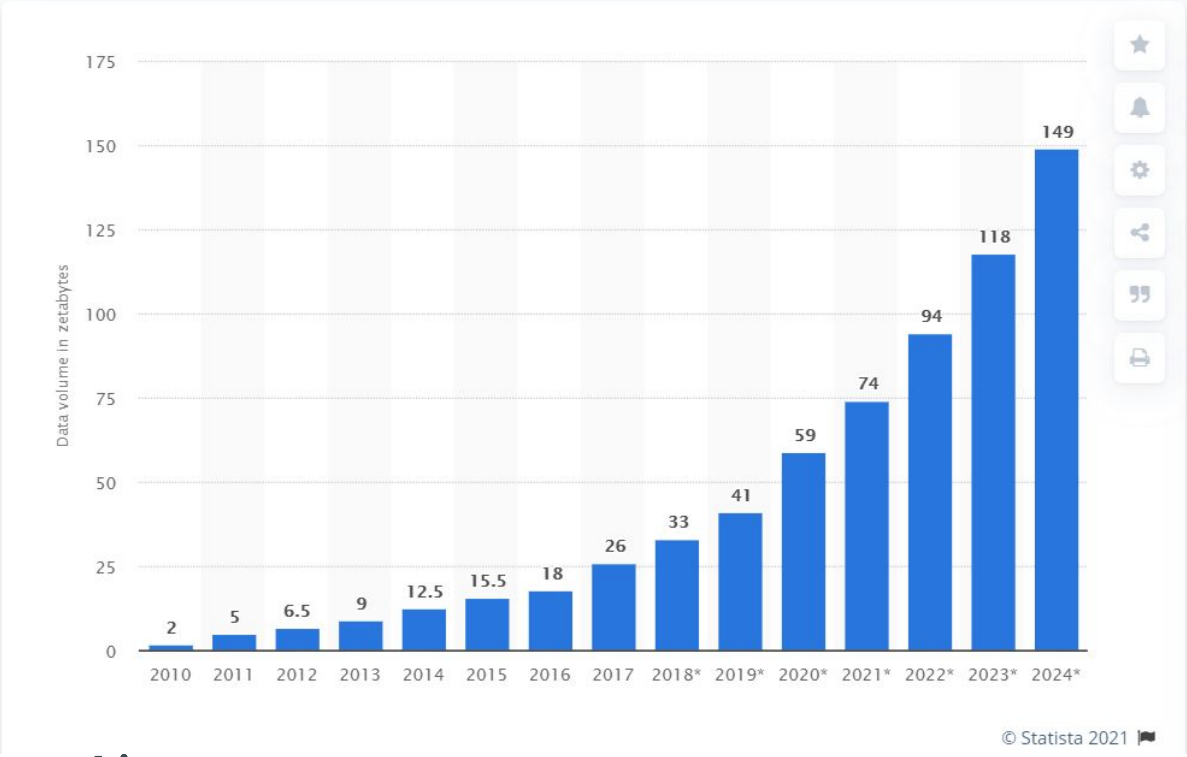


- +15 % od roku 2015 do roku 2019
- +2-3 % ročně typický růst v posledních dekádách

BizMachine

Zdroje: Světová banka, OSN

Objem vytvořených a zpracovaných dat v letech 2010 až 2024



- +165 % od roku 2015 do roku 2019
- +25 % ročně předpoklad na následující roky
- V roce 2022 předežene objem dat v zetabytech světové HDP v trilionech USD

BizMachine

Zdroj: IDC

Moderní trendy v oblasti velkých dat

1. Data a informace se stávají v bezprecedentním množství všeobecně dostupnými
2. Zpracování dat, jejich automatizace a pokročilé analytické metody se stávají dostupnými
3. Díky tomu dochází k praktickému využití velkých dat

… a to vše by proměnilo tvář některých známých blockbusterů…

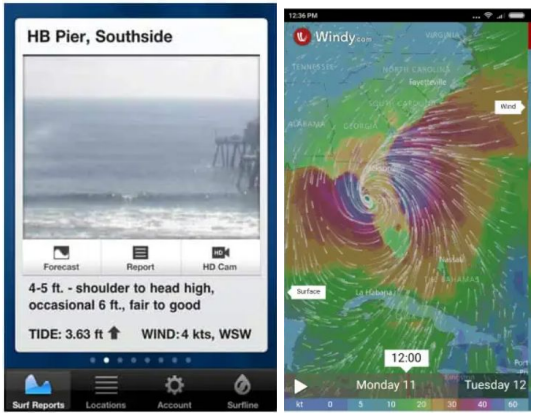
Point break (Bod zlomu), 1991, 2 h 2 min

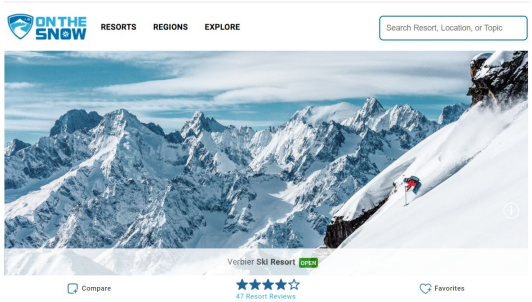
"Agent FBI Johny Utah se tajně vydá mezi kalifornské surfaře, aby vystopoval takzvané Exprezidenty – ozbrojenou bandu bankovních lupičů. Potká se s výborným surfařem Bodhim, jehož drogou je adrenalin. Rád riskuje. **Utah hledá pravdu, Bodhi největší vlnu.** Jsou natolik ponoření do vzájemného soupeření, že svým vášním téměř obětují zdravý rozum, hrdost a život…"
ČFSD

BizMachine

Point break, 1991, 2 h 2 min

- Desítky aplikací jako Surfline, Magic seaweed, On the snow
  - Kombinace “oficiálního” obsahu (pevné kamery, předpověď počasí, informace ze středisek) a obsahu vygenerovaného uživateli (komentáře, trasy, tipy, videa)
  - Náklady typicky 5-10 USD za měsíc (premium access)
- Velká šířka i hloubka informací integrovaných do úzce vymezeného případu použití (use case)

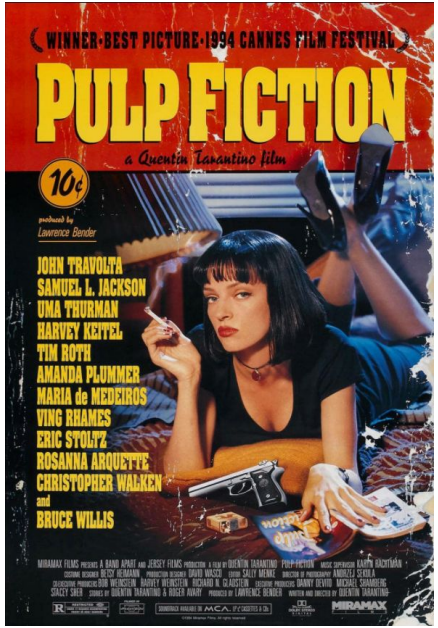




Forrest Gump, 1994, 2 h 22 min

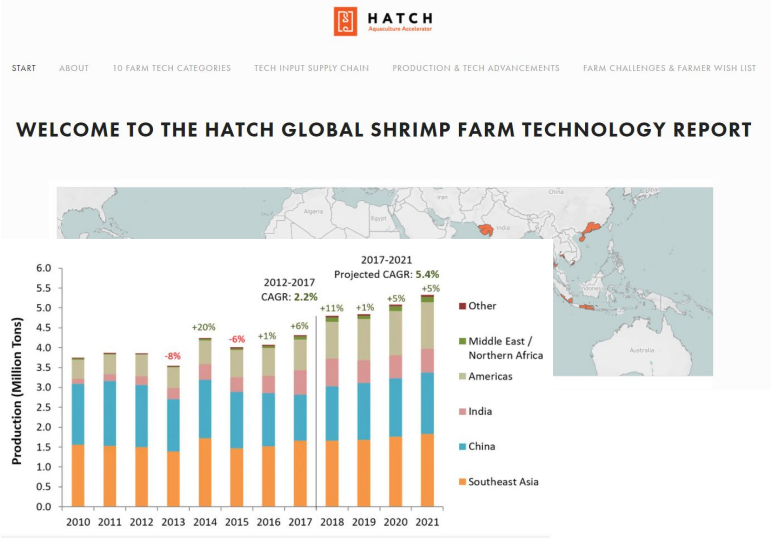
"...hrdinou je prosťáček Forrest Gump, typický obyčejný muž, který od dětství dělal, co se mu řeklo...
...Svůj život spojuje s ... přítelem z vojny, černochem Bubbou, **který dá směr jeho úvahám o lovu krevet** a s poručíkem Taylorem, jemuž ve Vietnamu zachrání život."
ČFSD

Pulp Fiction, 1994, 2 h 34 min



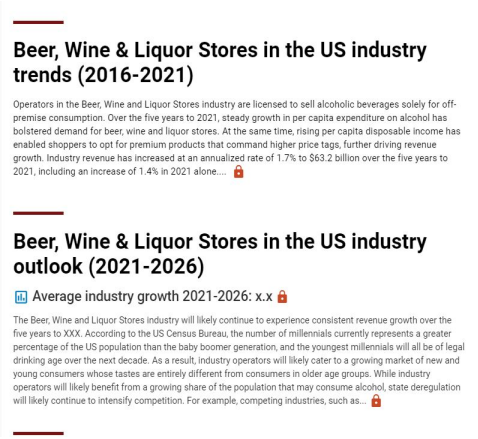
"Pumpkin v bistru vysvětluje Honey Bunny, že krámy s **chlastem už dělat nebudou, protože** si každý chce hrát na hrdinu a **je z nich málo peněz.** Nechá si dolít kafe a poté zasvětí svou milou, že od teďka budou dělat restaurace a kavárny..."
film-arena.cz

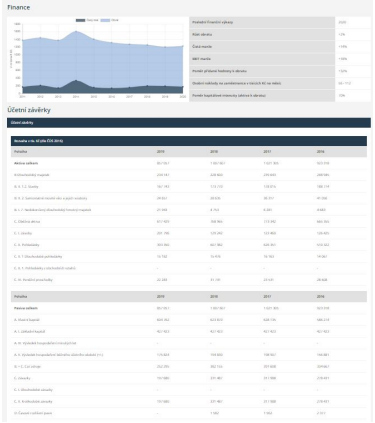
Forrest Gump, 1994, 2 h 22 min



- ~100 knih na amazon.com o chovu krevet
- 2 jednoduše dohledatelné detailní studie o odvětví (nepočítáme akvakulturu obecně)
- Desítky studií na jednotlivá témata

Pulp Fiction, 1994, 2 h 34 min





- Přes 45 tis. obchodů, ~200 tis. zaměstnanců v U.S.
- Pravidelně aktualizované reporty
- Detailní finanční výkazy k firmám

BizMachine

Forrest Gump, 1994, 2 h 22 min

- Desítky napojených veřejných zdrojů (finanční trhy, veřejné rejstříky, agregované statistiky lokálních statistických úřadů, atp.), proprietární průzkumy (osobní dotazování, digitální průzkumy)
- Integrované odvětvové znalosti (provozní, technické, socio-ekonomické, ekologické)
- Bez user-generated obsahu, menší šíře variant, globální dostupnost
- Relativně jednoduchá interpretace (de-facto pouze vizualizace)

Pulp Fiction, 1994, 2 h 34 min

- Velké objemy dat generovaných uživateli (hodnocení, návštěvnost, …)
- Technicky náročný sběr dat ze zdrojů (je jich jen pár, ale mohutných)
- Složitá párovací algoritmizace (náročný první krok interpretace)
- Vysoké požadavky na aktuálnost a stabilitu řešení v čase
- Vysoká výpočetní náročnost

Pulp Fiction, 1994, 2 h 34 min

Kavárna Zanzibar, Americká 152/15, 12000, Praha, CZ ●

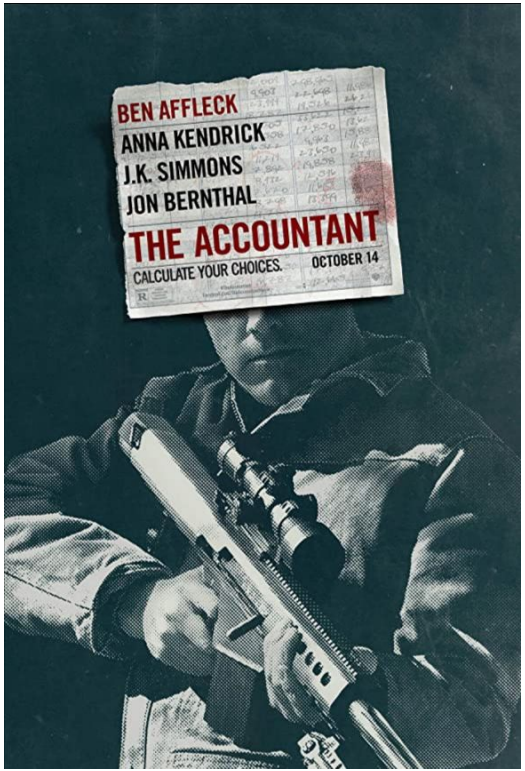
Sdílet | Nahlásit chybu

Kategorie:	Pohostinsví
Založeno (věk):	20.09.1995 (před 26 lety)
Web:	www.kavarnazanzibar.cz/,
Telefon:	(+420) 725 734 613,+420 725 734 613,(+420) 222 520 315
Email:	info@kavarnazanzibar.cz
BizMachine ID:	BMPCZ00172999AJ
Online profily:	RŽP - IČP: 1004413351 Google Places (4.4/5) Firmy.cz Restu.cz (4.3/5, 194 recenzí) Ceske-hospudky.cz Zomato.cz (4/5, 0 recenzí) Facebook.com (4.5/5, 60 recenzí, 1.6k lajků) Tripadvisor.com (5/5, 1 recenzí)
Typ outletu:	**Restaurace**
Velikost outletu:	
Přítomnost na webu:	7/7
Sociální skóre:	

Okres: Praha 2	HoReCa pořadí	Poř. v kategorii (Restaurace)	Hodnota
Hodnocení uživateli:	310/663 (▼15)	211/437 (▼6)	4.34/5 (▲0.06)
Popularita (počet hodnocení):	80/663 (▲1)	74/437 (– 0)	676 (▲57)

BizMachine

The Accountant (Zúčtování), 2016, 2 h 8 min

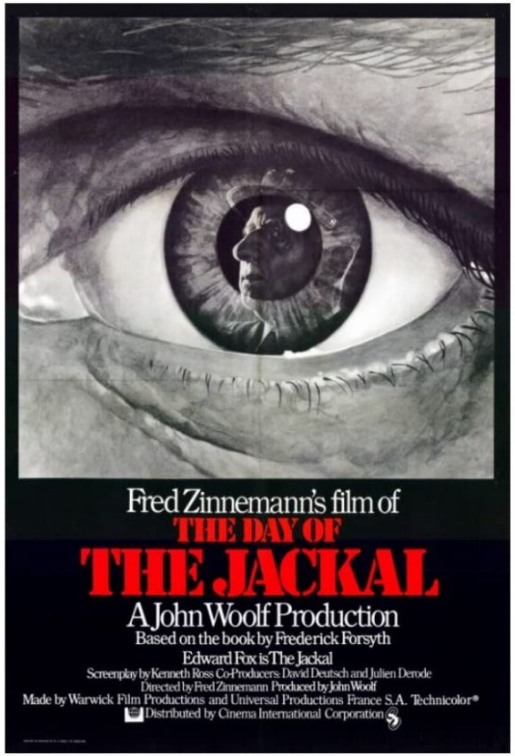


"Christian Wolff (Affleck) je matematický génius, který má lepší vztah k číslům než k lidem… Christian přijme legálního klienta: dobře zavedenou společnost zabývající se robotikou, ve které účetní (Anna Kendrick) zjistila účetní nesrovnalosti ve výši milionů dolarů. Ale s tím, jak Christian **rozkrývá účetní záznamy a přibližuje se k odhalení pravdy,** začíná v jeho okolí stoupat počet mrtvých."
ČSFD

The Godfather (Kmotr), 1972, 2 h 55 min

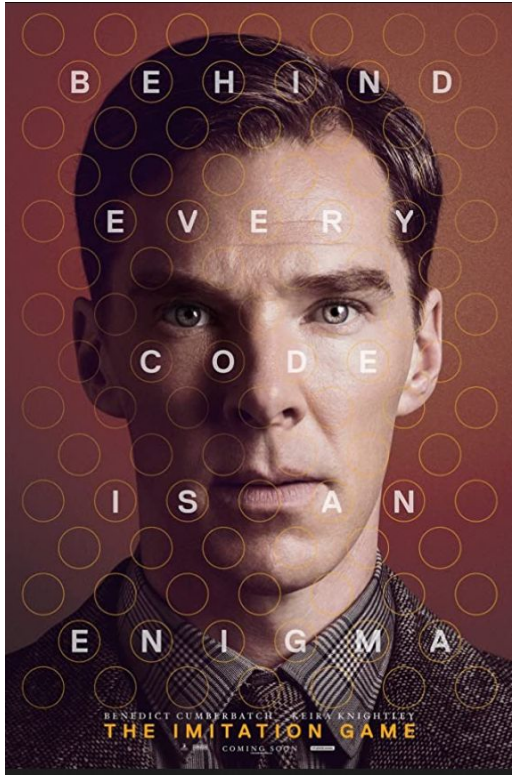
"Příběh newyorské mafiánské rodiny Corleonů, jíž tvrdou rukou šéfuje starý don Vito (Marlon Brando), se odehrává v letech 1945 - 55 a zaznamenává prudký rozmach organizovaného zločinu po skončení druhé světové války..."
ČSFD

The day of the Jackal (Den Šakala), 1973, 2 h 23 min



"Kultovní politický thriller, v němž tajná francouzská polovojenská organizace plánuje atentát na francouzského prezidenta Charlese De Gaulla…
…vyšetřovateli Lebelovi se chytře daří **skládat střípky vedoucí k odhalení** Šakalovy identity. Složitý děj využívá paralelní střihové skladby mezi detaily Šakalových příprav na atentát a Lebelovým úsilím najít vraha dřív, než bude příliš pozdě."
ČSFD

The imitation game (Kód Enigmy), 2014, 1 h 54 min



"V zimě roku 1952 britská policie dostala hlášení o vloupání do domu profesora matematiky a kryptoanalytika Alana Turinga…
… Téměř nikdo ze zúčastněných nevěděl, že jde o válečného hrdinu, který se spolu s týmem podílel na **prolomení kódu legendárního německého šifrovacího stroje Enigma.** The Imitation Game vypráví příběh matematického génia, který díky svému úžasnému talentu pomohl zkrátit 2. světovou válku a zachránil tak tisíce lidských životů."
ČSFD

The Accountant, The Godfather, The day of the Jackal

- Sběr informací je jednoduchý ("stačí" načíst data z finančních výkazů, stáhnout informace o vlastnících a statutárních členech nebo, získat data z rezervačních systémů)
- Klíčová je ale následná analýza dat (stavová i v čase)
  - Sestavení relevantní datové sady
  - Vím co hledám, nebo to nejdřív potřebuji zjistit?
- K tomu potřebujeme učební sady (learning sets) a předpřipravené metody
  - https://www.clarifai.com/
  - https://scale.com/
  - https://www.dataquest.io/blog/15-python-libraries-for-data-science/
  - https://networkx.org/

Saving private Ryan (Zachraňte vojína Ryana), 1998, 2 h 49 min

“Film vypráví příběh kapitána americké armády Johna Millera (Tom Hanks) při vylodění v Normandii 6. června 1944 a v následujících dnech. Po dobytí pláže Omaha je jednotka kapitána Millera vyslána **najít vojína** Ryana (Matt Damon), jehož tři bratři již padli, a přivést ho zpět k jeho matce.”
wikipedia.org

Saving private Ryan, 1998, 2 h 49 min

- Personnel Recovery Device/recovery beacon (od roku 2019, i pro situace, kde nefunguje GPS signál)
- Combat Survivor Evader Locator (CSEL) systém (od r. 2009, pouze v dosahu GPS)
- Nett warrior (od r. 2011, používá Samsung Galaxy Note II a SW od USAF
- A spousta dalších technologií
  - Identify friend or foe strobes (IFF)
  - The Personal Role Radio (PRR)
  - Tactical and satellite comms (dedicated personnel or vehicle mounted)
  - “Eye in the sky” drones
- 29 GPS satelitů na oběžné dráze (~2 700 celkem funkčních, ~6 000 celkem)
- Přes 60 tis. letových hodin měsíčně velkých U.S. dronů

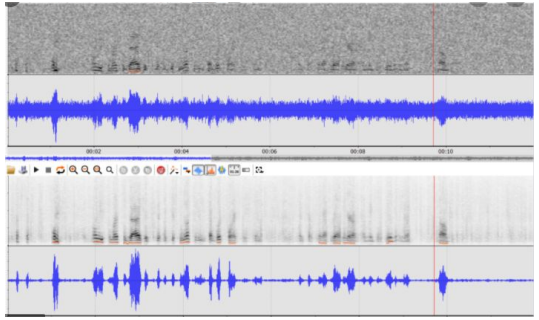
Sources: imdb.com, asc.army.mil, wikipedia.org, armytimes.com

Blade Runner, 1982, 1 h 57 min

“Počátkem 21. století dosáhla společnost Tyrell ve výzkumu robotů fáze Nexus. Její výrobky byly téměř identické s lidmi, nelišily se ani inteligencí. Lidé používali replikanty při nebezpečných výpravách nebo jako otroky při kolonizaci jiných planet. Po krvavé vzpouře replikantů v jedné z mimozemských kolonií byl robotům pod trestem smrti zakázán pobyt na Zemi. Na plnění tohoto nařízení dohlížely speciální jednotky Blade Runner, pověřené **vyhledáváním** a likvidací replikantů, kteří zákaz poruší.”
ČSFD

Blade Runner, 1982, 1 h 57 min

- Phonexia technologie zpracovává 52.5 milionů minut hovorů každý den
- V září 2019 využívalo AI pro sledování 75 zemí světa (včetně ČR), z toho 64 metodu rozpoznávání obličejů (face recognition). Do 63 zemí dodává tyto technologie Čína, USA do 32 zemí, Japonsko do 14.
    - V roce 2019 jsme se díky úniku dat dozvěděli, že Čína má ve své databázi i obličeje nemluvňat (nejmladší zjištěný záznam je pro 9 denní nemluvně)
    - Face recognition se používá i proti krádežím toaletního papíru ve Chrámu nebeského parku v Pekingu
- Termovize (thermal imaging): např. v UK nasadila všechna letiště v průběhu roku 2020 ve spojení s technologií rozpoznáváním obličejů
- USA od roku 2020 používají facial recognition na 20 největších letištích pro všechny mezinárodní cestující
- Nejnovější technologie zvládají identifikovat tvář s 99% přesností za 2 vteřiny



Co to všechno znamená

- Zdá se, že tentokrát se “revoluce velkých dat” skutečně koná (není omezena na pár vyvolených, ale je přístupná široké veřejnosti)
- Podobně jako u ostatních mladých odvětví lze očekávat:
  - Konsolidaci a vzhledem k omezenému lidskému kapitálu také rozevírání nůžek mezi “profíky” a “amatéry”
  - Postupnou produktizaci a radikální pokles jednotkových nákladů
  - Dočasně spoustu nekvalitních polotovarů
- Společnosti proto mají v zásadě 2 cesty, jak se k velkým datům postavit:
  - Do velkých dat skutečně zainvestovat tak, aby se z nich stala klíčová dovednost
  - Stát se “fast followers”, tedy sledovat, jaké produkty a služby jsou v této oblasti k dispozici a “naskočit na ně”. I tato strategie bude vyžadovat “spojku” s dodavateli, někoho, kdo je schopen kontroly kvality a hovoří stejnou řečí