BRAVE NEW DIGITAL WORLD

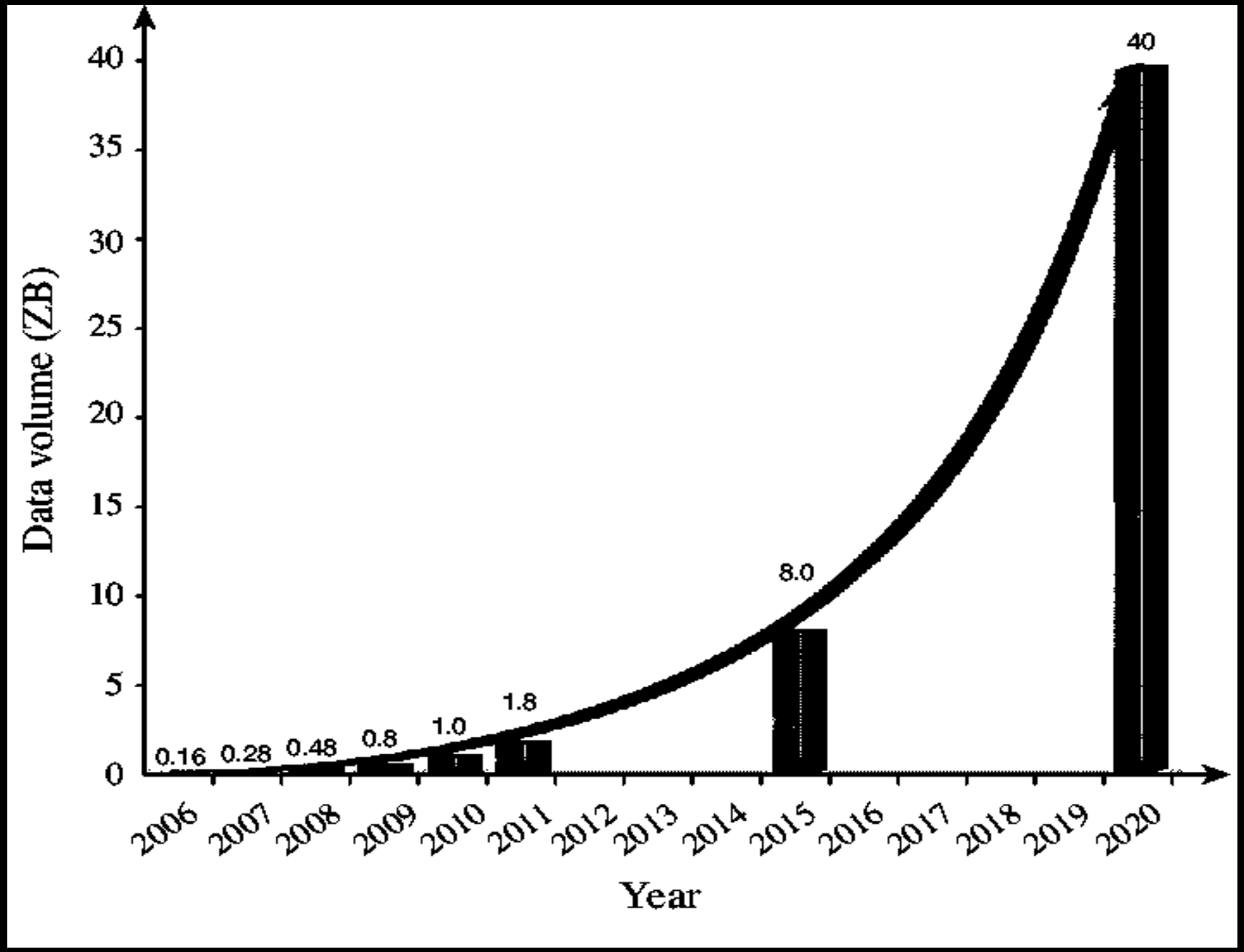
5. BIG DATA

JAK VELKÁ JSOU VELKÁ DATA

* globální objem dat se zdvojnásobí cca každé tři roky

* 95 mil. nových fotek a videí každý den na Instagramu, 450 tis. tweetů každou minutu, 5 nových fb profilů každou sekundu

* Large Synoptic Survey Teleskope každý den pořídí 28 TB dat; na Wikipedii je každou minutu provedeno 600 editací, v rámci činnosti LHC je každou sekundu možno zaznamenat 25 GB dat


Data volume (ZB)
40
35
30
25
20
15
10
5
0
0.16
0.28
0.48
0.8
1.0
1.8
8.0
40
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year

JAK VELKÁ JSOU VELKÁ DATA

* nejde ani tak o množství jako spíše o situaci
* 30‘s & 40‘s od stratifikovaných vzorků k vzorkům náhodným a nyní od vzorků k celkům
* změna měřítka způsobuje změnu stavu – kvantitativní změna iniciuje změnu kvalitativní
* velká data znamenají, že můžeme provádět některé operace, které nebyly v malém měřítku možné (resp. adekvátní)

ARCHEOLOGIE BIG DATA

* Mathew Maury jako inspektor skladů námořních map pomocí starých deníků „vyčísluje“ atlantický oceán a objevuje nové lodní cesty, později zavádí standardizovaný námořní záznam a nakonec vydává The Physical Geography of the Sea (1855)

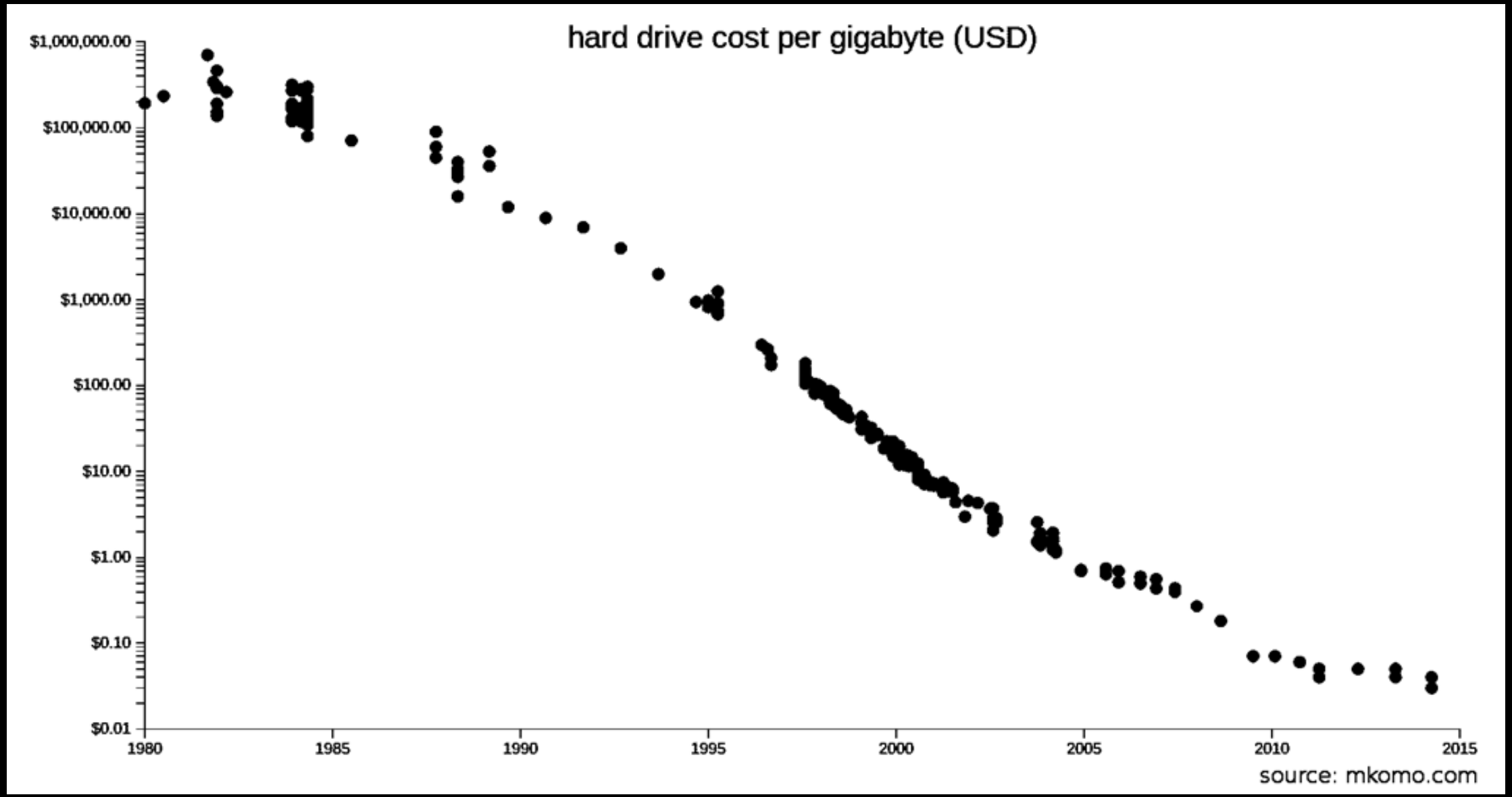
* Francis Galton vypracovává techniku průzkumu pomocí dotazníků, vytváří meteorologické mapy a představuje způsob klasifikace otisků prstů; do statistiky vnáší měření korelace - „pokud se vyskytuje jev A tak s pravděpodobností X se (ne)vyskytuje jev B“ (1888)

PROBLÉM STROJOVÉHO PŘEKLADU

* projekty Léona Dostera v období studené války, cílem je především rychlý překlad z ruštiny; kódování komplexního gramatického fundamentu se neukazuje jako dobrá cesta

* IBM Candide pracuje se záznamy jednání kanadského parlamentu a překlad určuje pomocí statistické pravděpodobnosti; jeho bází byly 3 mld. dobře přeložených vět

* Google od r. 2004 skenuje web a zaznamenává překlady kolísavé kvality; jeho báze má stovky mld. vět; výsledkem je služba translate


hard drive cost per gigabyte (USD)
$1,000,000.00
$100,000.00
$10,000.00
$1,000.00
$100.00
$10.00
$1.00
$0.10
$0.01
1980
1985
1990
1995
2000
2005
2010
2015
source: mkomo.com

BIG BUZZ DATA

* agnostický přístup + rozličná kvalita + korelace => např. se špatným počasím roste spotřeba sušenek s jahodovou příchutí, oranžová auta jsou v amerických autobazarech ta nejspolehlivější, vaše dcera je těhotná

* datafikace slov (books.google), datafikace polohy (Waze), datafikace interakcí (Facebook) nabízí řadu nových poznatků (např. šíření chřipky, pohyby cen nemovitostí, ceny letenek)

* „měření znamená vědění“ & „vědění znamená moc“

BIG BIZ DATA

* data netrpí rivalitní spotřebou, ITs přinášejí pasivní sběr (např. Analytics tracking code), hodnota nemusí nutně klesat

získávání hodnoty:

1) opakované použití (systém doporučení, open data)
2) slučování datových množin (vznik nových dat)
3) sběr s ohledem na vícero použití (street view, CCTV)

* podnikání s daty & podnikání se znalostmi

BIG WISE DATA

* analýza big data je jako rybaření – nevíme, co chytíme
* přichází konec vědecké metody a konec géniů?
* korelační analýza namísto „proč/jak“ říká „co“
* je toto omezení důrazu na kauzalitu adekvátní trade off nebo znamená příchod nové „doby temna“, kdy sice víme ale nerozumíme?

OD LSTIVOSTI …

*Amazon ví, co nakupujeme, Twiter co si myslíme, Facebook zná naše přátele, Google oblíbené weby, operátoři to, kde jsme a kdo je poblíž …

*„pokud je to zadarmo, produktem jste vy sami“

*inflace loginu – každý další je cennější, protože takto prohlubuje informační bázi

OD LSTIVOSTI ...

* spoustu našich HCI je datafikováno a obratem se ocitá v těch nejvýkonnějších výpočetních klastrech, které z nich vytěžují znalosti

* každá zachycená informace zhodnocuje ty ostatní

* takto vzniklá síť (model) je široce aplikovatelná

* dostatečná technologická úroveň spolu s adekvátním datovým fundamentem umožňují hlubší poznání a přesnější predikce

… K NÁZNAKŮM DICKOVSKÉHO SVĚTA

* Ne Orwell a 1984, ale Dickovy a Minority report a Adjustment Team

* Cambridge Analytica vypracovávala osobnostní model Big Five a na jeho základě pak na fb komunikovala behaviorální reklamu

* software PredPol předpovídá majetkovou trestnou činnost, automatizace řízení o podmínečném propuštění a práce Richarda Berga

… K NÁZNAKŮM DICKOVSKÉHO SVĚTA

* znamenají Big data konec svobodné vůle?

* operace á la Cambridge Analytica jsou kompatibilistickou manipulací – svobodná rozhodnutí sice činíme, ale kontext je vyfabrikován a přizpůsobený individuálním motivacím

* prediktivní kriminalistika a justice staví nad svobodnou vůli deterministický model – tvrdá determinace je oktrojována

STRATEGIE ZVLÁDÁNÍ

* odpovědnost držitelů dat – pouze vymezené typy operací

* zachovat instituci svobodného jednání – otevřené a certifikované algoritmy, jejichž výstupy by byly falzifikovatelné

* vznik protimonopolních a protikartelových zákonů regulujících datové barony