
BIG
DATA
USE
ONE
SOFTWARE
USING
GARTNER
MASSIVELY
WORKING
BIOLOGICAL
HUNDREDS
DISK
SHARED
TIME
INCLUDE
SYSTEMS
NETWORKS
INFORMATION
RECORDS
SEARCH
DEFINITION
PROCESS
WITHIN
TOOLS
SET
GENOMICS
PERFORMANCE
ALSO
RELATIONAL
ANALYTICS
SAN
PARALLEL
TOLERABLE
SIZE
NEEDED
PETABYTES
INTERNET
TECHNOLOGIES
DISTRIBUTED
CAPTURE
MANAGE
GROW
MANAGEMENT
SOCIAL
LARGE
EVERY
LARGER
CASE
STORE
PROCESSING
LOGS
MAY
BUSINESS
MOVING
TECHS
CAPACITY
PRACTITIONERS
COST
CONTINUES
CITATION
TERABYTES
RESEARCH
COMPLEX
DATABASES
STORAGE
DIFFICULTY
TARGET
ABILITY
SETS
EXAMPLES
APPLIED
CURRENT
ELAPSED
SENSOR
ARCHIVES

„Big data“ = velká data – obrovské objemy dat => vývoj nových metod zpracování datových souborů, obrazů

Vícebajtové jednotky

Desítková soustava			Dvojková soustava				
Hodnota	**Předpona SI**		**Hodnota**	**ISO/IEC**		**Paměť**	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte	TB	terabyte
1000^5	PB	petabyte	1024^5	PiB	pebibyte	—	
1000^6	EB	exabyte	1024^6	EiB	exbibyte	—	
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte	—	
1000^8	YB	yottabyte	1024^8	YiB	yobibyte	—	
1000^9	RB	ronnabyte	—				
1000^{10}	QB	quettabyte	—				

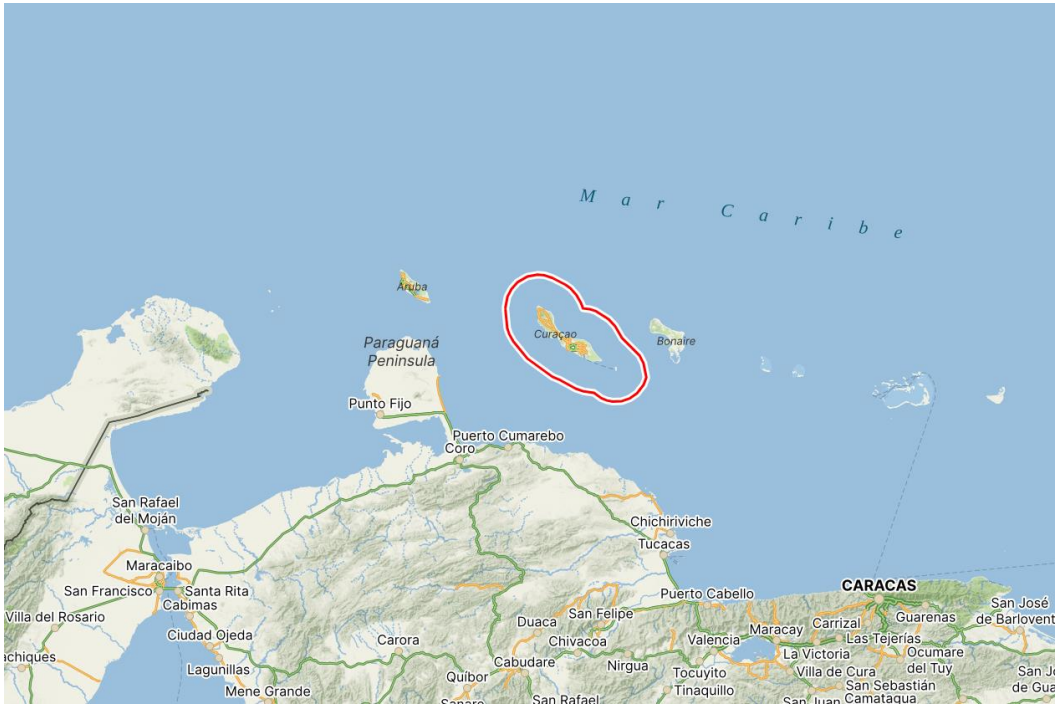
Jednotka	Přibližný ekvivalent
bit	proměnná typu Boolean - pravda (1), nepravda (0)
byte	(=8 bitů) základní písmeno latinské abecedy
kilobyte	krátký textový e-mail
	typická ikonka
megabyte	text knihy *Harry Potter a ohnivý pohár*
gigabyte	1 min MP3 audio záznamu
	CD kvalita nekomprimované audionahrávky alba skupiny Genesis: *The Lamb Lies Down on Broadway* 4 hod Skype/zoom; 1 hod Youtube
terabyte	největší HDD pro počítače v r. 2007
	animovaný TV seriál Avatar: Legenda o Aangovi, 61 epizod (1080p, 4:3) všechny rtg. snímky FN Motol 1 600 CD nebo 4,5 mld. knih
petabyte	2000 roků hudby ve formátu mp3 4000 fotografií každý den celý život kapacita jedné mozkové hemisféry
exabyte	celosvětový měsíční provoz na internetu v r . 2004 11 milionů 4K filmů data vygenerovaná na celé Zemi za 3 minuty
zettabyte	celosvětový roční provoz internetu v r. 2016

Průměrná spotřeba mobilních dat (za měsíc)

ČR	9.67 GB
Lotyšsko	52.7 GB
Finsko	30 GB
Průměr západní Evropy	20 GB
Čína	16 GB
Polsko	14 GB
Jižní Korea	6.91 GB
USA	4.45 GB

Kde je spotřeba mobilních dat největší?

Nejvíce – Curacao 131.3 GB



Celosvětový objem dat

celkový objem dat vytvořených, zachycených, zkopírovaných a spotřebovaných na celém světě (1 ZB = 10^9 TB)

Rok	Objem dat
2010	2 ZB
2015	12 ZB
2018	33 ZB
2020	59 ZB
2023	97 ZB
2024	149 ZB
2025	odhad 175 ZB

2023 - 5 miliard uživatelů internetu na světě (asi 63% světové populace) => generují obrovské množství dat prostřednictvím svých online aktivit

State of Data Quality 2022 se pouze 16 % společností – vysoká úroveň kvality dat, 27 % se k tomuto cíli blíží

Informační technologie a telekomunikace

- Internet: Každý den proteče internetem přibližně 8.2 EB (r. 2020)
- Sociální média: Facebook generuje více než 4 PB dat denně

Finanční sektor

- Bankovnictví: Velké banky mohou generovat až 2,5 TB dat denně z transakcí a dalších operací.

Zdravotnictví

- Elektronické zdravotní záznamy: V roce 2020 bylo odhadováno, že globální objem zdravotnických dat dosáhl 2 314 EB.

Marketing a e-komerce

- E-komerce: Amazon zpracovává přibližně 1,4 milionu transakcí za minutu během špičkových období, což generuje obrovské množství dat.

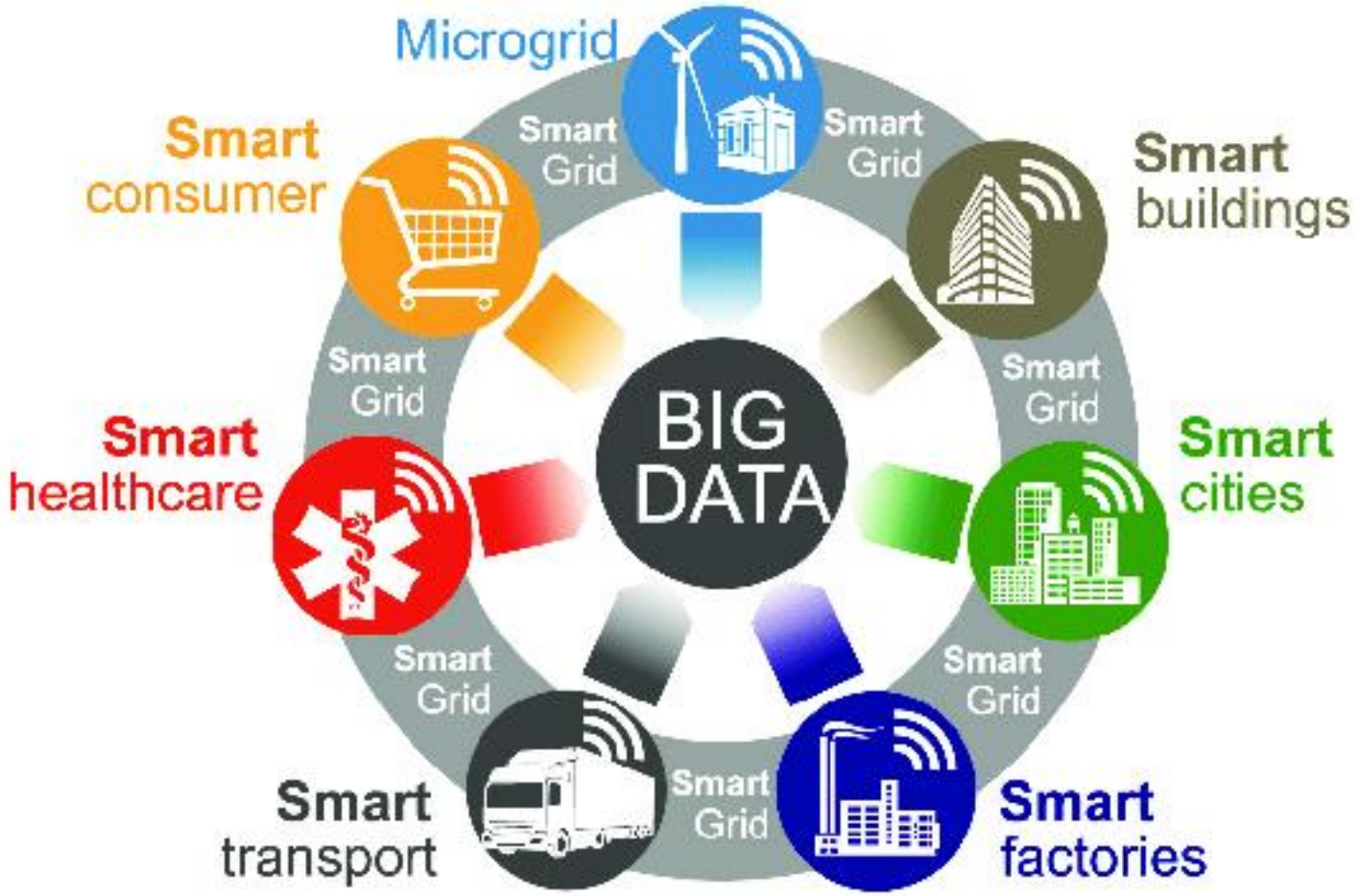
Výroba a logistika

- IoT zařízení: Průmyslové IoT zařízení generují přibližně 400 ZB dat ročně.

Vědní obory

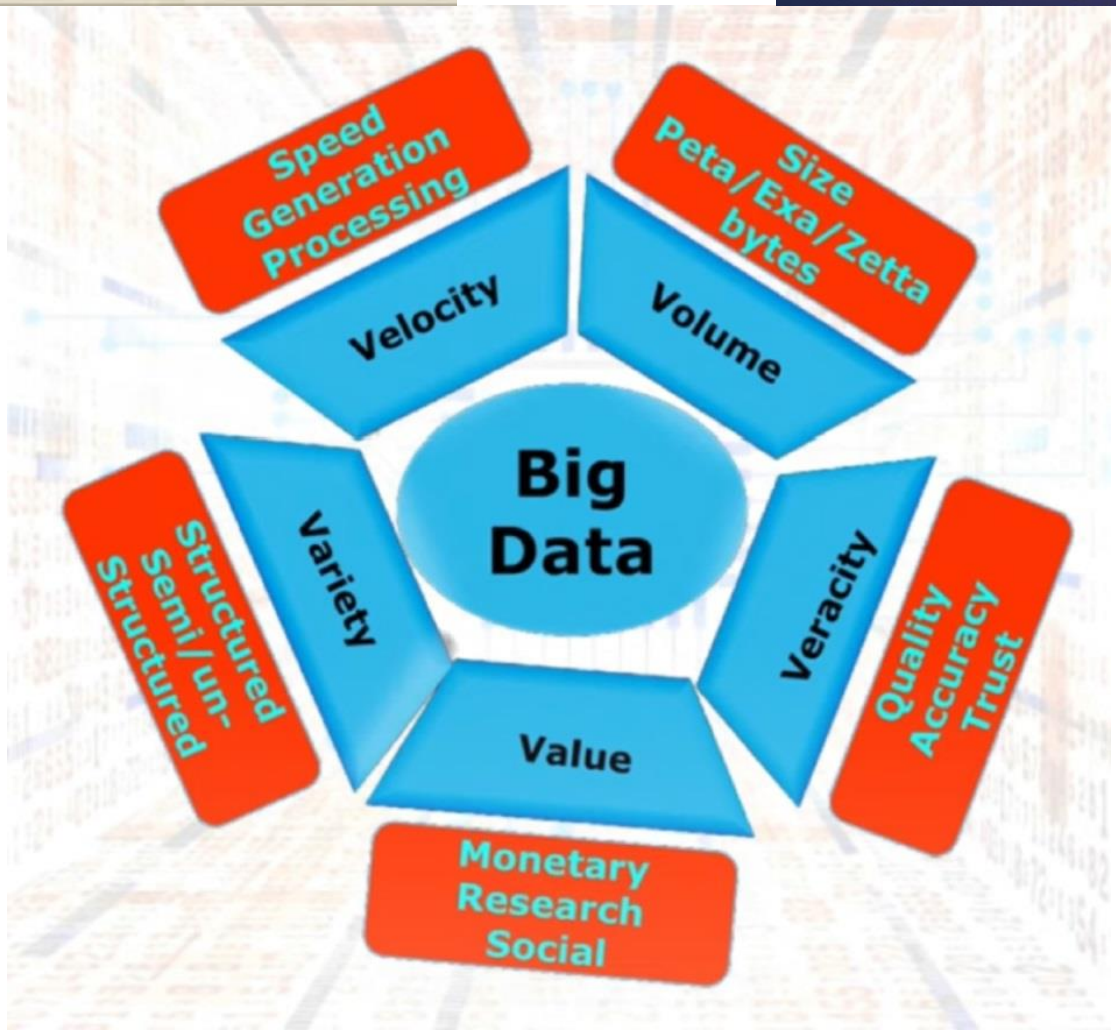
- Astronomie: Vera Rubin Observatory má produkovat přibližně 15 TB/noc
- Genomika: Sekvenování lidského genomu generuje přibližně 200 GB až 1 TB dat na jeden genom (1 genom- 3 miliardy párů bází DNA)
- Jaderná fyzika (CERN) – petabajty ročně
- Chemie, vývoj léků …

Tyto údaje ukazují, jak různé obory přispívají k celkovému objemu dat a jaké výzvy a příležitosti to přináší.


Microgrid
Smart consumer
Smart buildings
Smart healthcare
Smart cities
Smart transport
Smart factories
BIG DATA
Smart Grid
Smart Grid
Smart Grid
Smart Grid
Smart Grid
Smart Grid


THE BIG FIVE


VELKÁ PĚTKA
extroverze
neuroticismus
svědomitost
souhlasnost
otevřenost


Big Data
Velocity
Speed
Generation
Processing
Volume
Size
Peta/Exa/Zetta
bytes
Variety
Structured
Semi/un-
Structured
Veracity
Quality
Accuracy
Trust
Value
Monetary
Research
Social

5 „V“ ve velkých datech: Volume, Variety, Velocity, Veracity, Value

Volume (objem) - velké množství dat; v astronomii obrovské množství objektů a dat o nich, všechny uložené v rozsáhlých databázích. Streamovací služby jako Netflix nebo YouTube - videa, a data uživatelů (preference, historii vyhledávání, interakce) => pokročilé možnosti ukládání a zpracování

Variety (rozmanitost) - množství dat v mnoha formátech, strukturách včetně textu, obrázků, zvuku a videa => výzva pro efektivní získávání smysluplných poznatků; strukturovaná data = datové typy jako databáze jmen a čísel; nestrukturovaná data - datové typy jako je text, zvuk, obrázky a příspěvky na sociálních sítích.

Velocity (rychlost) - rychlost generování dat a jejich zpracování; sociální média - miliony příspěvků denně; dostupnost dat v reálném čase je klíčová

Veracity (věrnost) - zachování kvality dat, důležité pro přesnou analýzu a interpretaci

Value (hodnota) - strukturovaná data - mohou například odhalit číselné trendy a vzorce, nestrukturovaná textová data z příspěvků na sociálních sítích nebo zákaznických recenzí mohou odhalit pocity a názory, které řídí lidské chování.

Trocha hardwarové historie

1956 - 1. disk pro sálový počítač – 3.75 MB

1985-9 IQ 151 vnitřní paměť max 64kB + mg. páska

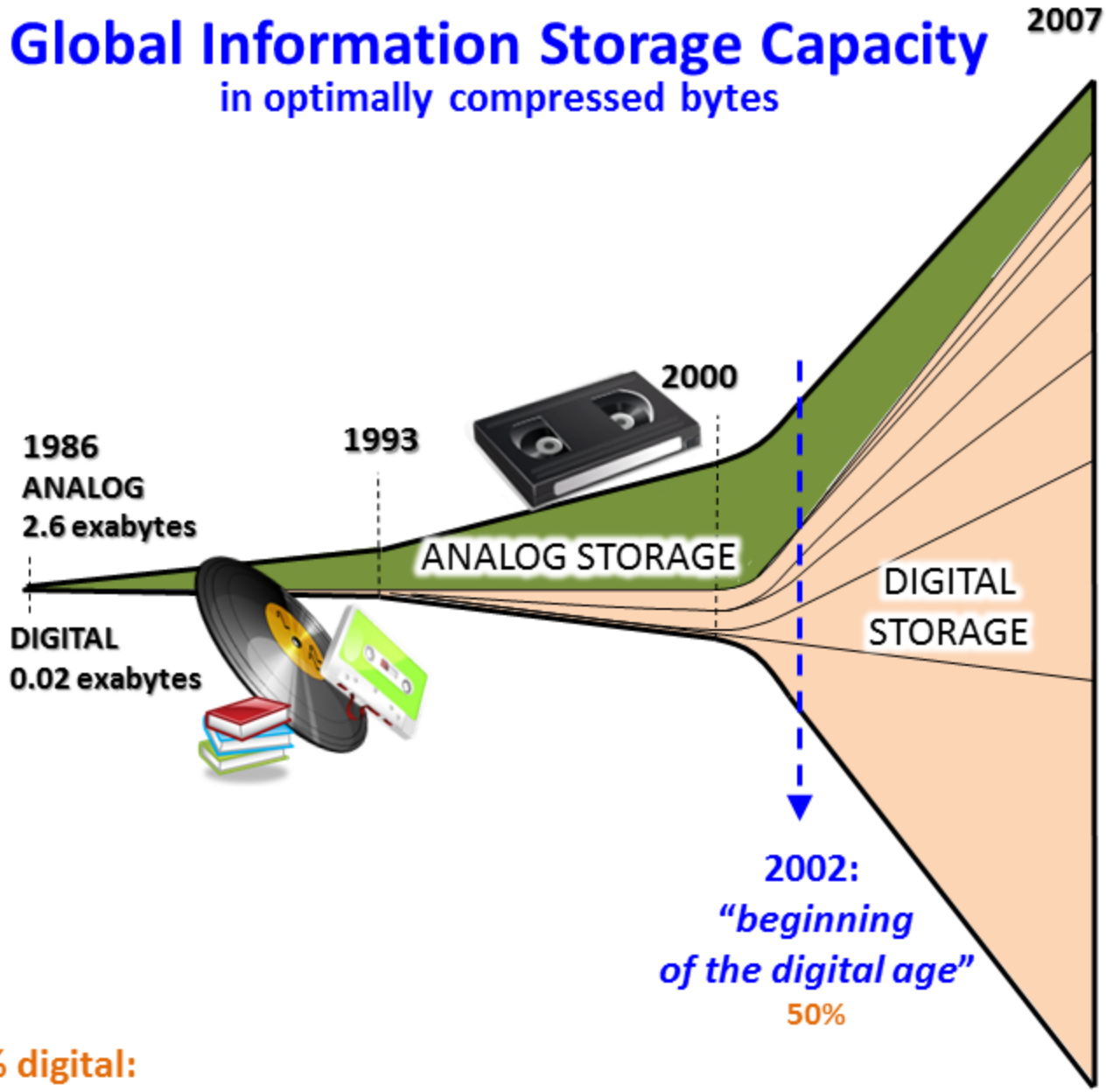
1986 PP06 - max 128 kB + diskety 360 kB

1973 – první osobní počítač MCM/70 – 8kB operační paměti

2024 – mobilní telefony 256-512 GB
– největší běžně prodejné disky 24 TB

Global Information Storage Capacity

in optimally compressed bytes



2007 **ANALOG**
19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % **ANALOG**

DIGITAL

- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %

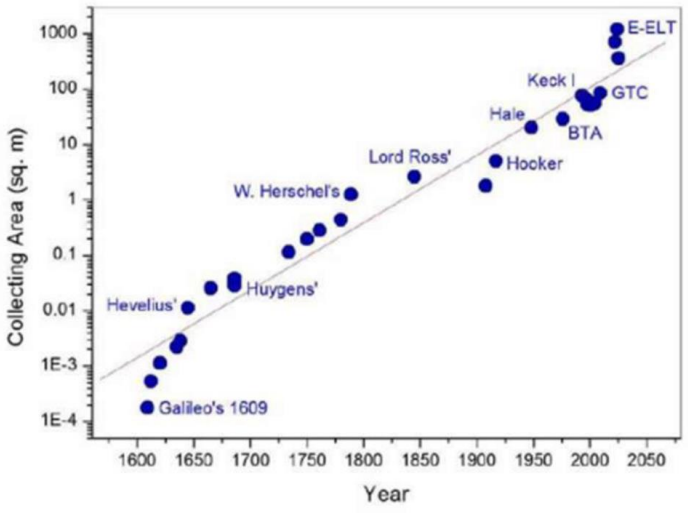
- PC hard disks: 44.5 %
  123 billion gigabytes

- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

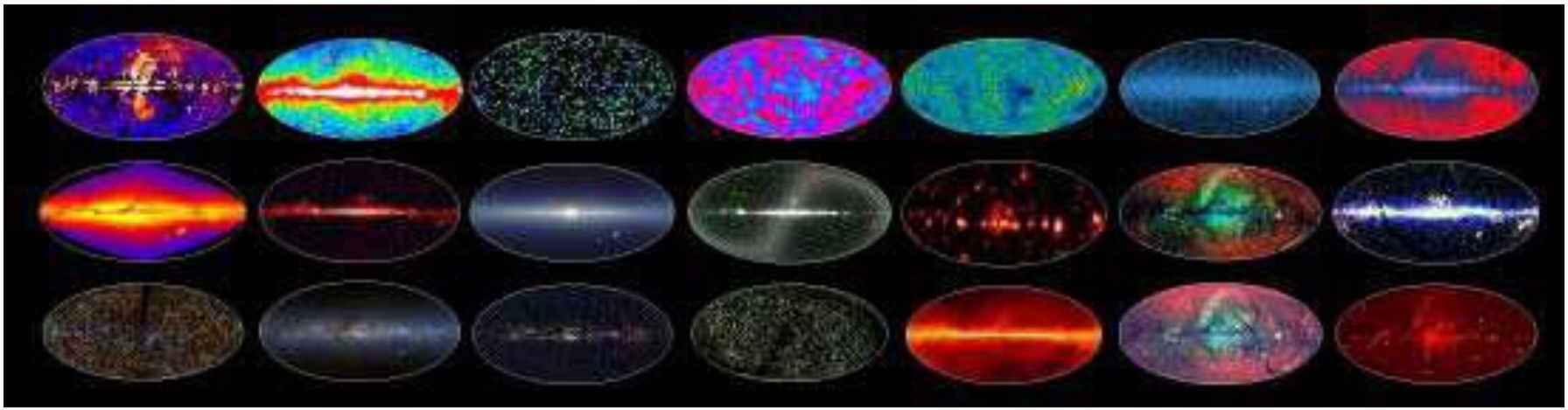
DIGITAL
280 exabytes

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Data v astronomii



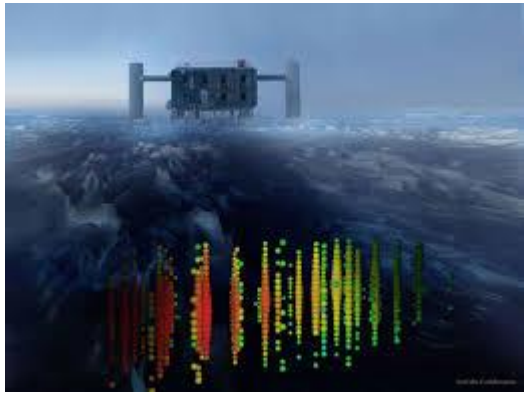
- astronomické přehlídky (zejména celooblohové a velkoplošné) - shromažďování pozorovacích dat => objevy nových nebeských objektů a jevů
- celé elektromagnetické spektrum (všechna okna do vesmíru)
- posun v našem chápání vesmíru; vznik holistická (celostní) představy o vesmíru



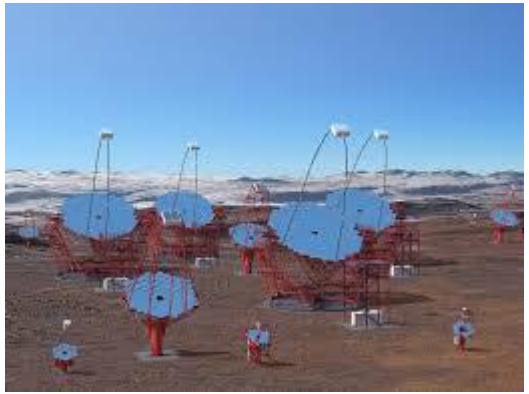
Přehlídky, projekty	**Zkratka**	**Obor**	**Množství dat**
Digitized First Byurakan Survey	DFBS	opt	400 GB
Digitized Sky Survey (založeno na Palomar Observatory Sky Survey)	DSS	opt	3 TB
Two Micron All-Sky Survey	2MASS	NIR	10 TB
Galaxy Evolution Explorer	GALEX	UV	30 TB
Sloan Digital Sky Survey	SDSS	opt	40 TB
GAIA	GAIA	nIR-nUV	200 TB
SkyMapper Southern Sky Survey	SkyMapper	opt	500 TB
Panoramic Survey Telescope and Rapid Response System	PanSTARRS	opt	∼40 PB
Vera Rubin Observatory, (od ledna 2025)	LSST	opt	∼200 PB
Square Kilometer Array, odhad	SKA	radio	∼4.6 EB

Název projektu	Denní objem dat	Obor činnosti
HST	1.5 GB	nIR – nUV
JWST	57 GB	mIR
ALMA	80 GB	mm
IceCube	100 GB	neutrina
LIGO	1.5 TB	gravitační vlny
Pierre Auger Observatory	3 TB	UV
Vera Rubin Observatory	25 TB	nIR – nUV
CTA	27 TB	gamma
Velký urychlovač CERN	několik TB	hadrony
EHT	80 GB	mm
SKA	1920 TB	mikrovlny

PanSTARRS – DR2 v 2019 1.6 PB dat
= 30 000x textový obsah Wikipedie

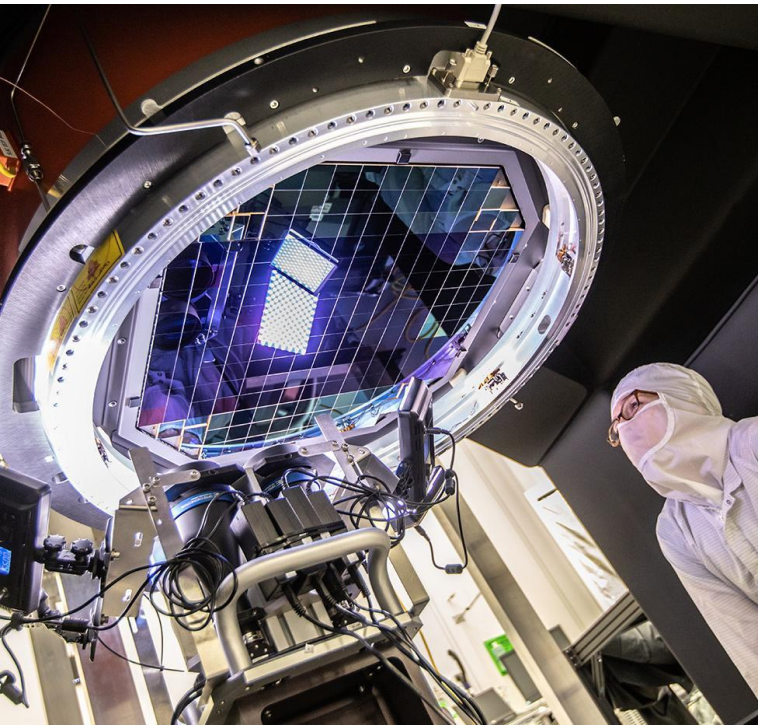


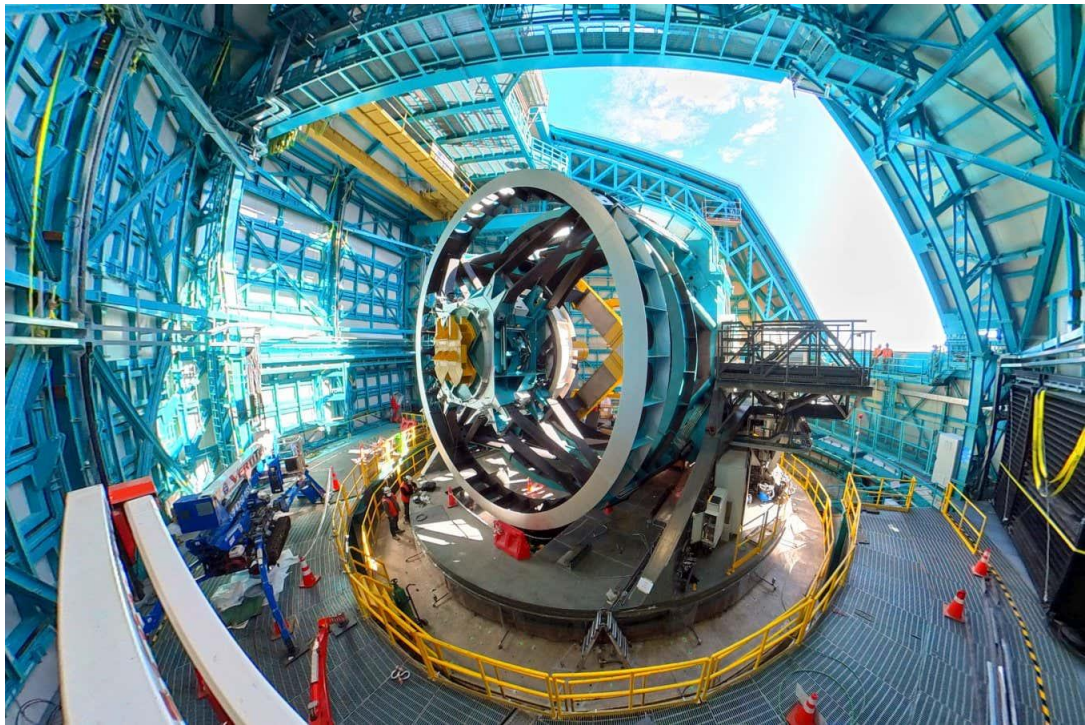

SMT
PV
SMA
JCMT
LMT
APEX
ALMA
SPT



Vera Rubin Observatory

- největší objektiv 8.4 m a CCD čip (3.2 Gpx) na světě
- začátek provozu - leden 2025
- 25 TB/noc
- 100 PB datový archiv
- 250 teraflops výpočetní síly
- zpracování každého snímku do 60 s od expozice (identifikace, transienty) => 10 milionů alertů





SKA – Square Kilometer Array

- SKA-mid (Afrika) ~ 200 antén
- SKA-low (Austrálie) ~ 130 000 antén
- datový tok 150 TB/s (antény-korelátor),
  resp. 2 TB/s (korelátor-> procesor)
- 80 000 km optických vláken (2x kolem Země)
- 2x150 Pflops - výpočetní výkon (TOP 10)
- archivované množství dat: 700 PB/rok, tj. 80 TB/hodinu
- celková spotřeba: 15 MW (před 10 lety odhad 100 MW)
- první světlo: 2027?
- soustavou teleskopů proteče za den stejné množství dat jako celosvětovým internetem

Výzvy exabajtové astronomie

Objem dat - obrovské soubory dat z observatoří => potíže při efektivním ukládání při zachování dostupnosti, spolehlivosti a dlouhodobé trvanlivosti

Analýza dat – složité a rozsáhlá data => nutné pokročilé algoritmy a techniky analýzy, odhalující vzorce a jevy, které tradičními metodami nejsou možné

Výpočetní výkon – zpracování obrovského objemu dat vyžaduje vysoce výkonné systémy => výpočetní zdroje, jejichž pořízení a údržba nákladná a logisticky náročná

Interdisciplinarita – bude třeba spolupráce více oborů (statistika, informatika a datová věda)

Rizika exabajtové astronomie

Riziko zahlcení – objem dat může převýšit analytické kapacity (lidí, ne strojů) => nedostatečné využití dat (-> občanská věda)

Zkreslení – algoritmy strojového učení mohou vést k nadměrnému přizpůsobení a zkreslení => zavádějící interpretace nebo „falešné objevy“

Opomenutí – sběr dat bez odpovídající analýzy => přehlédnutí důležitých výsledků a objevů (efekt „krabičky zlata“ -> občanská věda)

Vypnutí – náklady na ukládání, zpracování a údržbu velkých souborů dat jsou vysoké – dlouhodobé zatížení rozpočtu; rychlý technologický pokrok => zastarávání použitých nástrojů a vybavení => potřeba aktualizací, obměn => drahé

Nereprodukovatelnosti – reprodukovatelnost výsledků – základní princip výzkumu, vědy. Využití strojového učení může tento princip zpochybnit

Proměna astronomie

- Závislost na strojovém učení – větší zapojení strojového učení a umělé inteligence
- Interdisciplinarita – spolupráce s datovými vědci, statistiky, informatiky – vývoj algoritmů, statistických modelů a datových analýz
- Kreativita – větší koncentrace na modelování, interpretaci výsledků; větší zapojení virtuální a rozšířené reality; sběr dat a redukci - stroje
- Virtuální astronomie – práce astronoma u pc v kanceláři bez fyzického přístupu k dalekohledům
- Občanská věda – využití potenciálu občanské vědy, efektivnější zpracování, urychlení analýzy a objevování

Konec rutinních prací astronoma

V budoucnu astronomové budou více času trávit trénováním umělé inteligence než pozorováním hvězd.

Brzy bude největší výzvou nalezení způsobu, jak zapisovat data rychleji než přicházejí!

Big data nejsou problémem jen astronomie

Informační technologie a telekomunikace:
velké množství dat prostřednictvím internetu, mobilních zařízení a cloudových služeb.

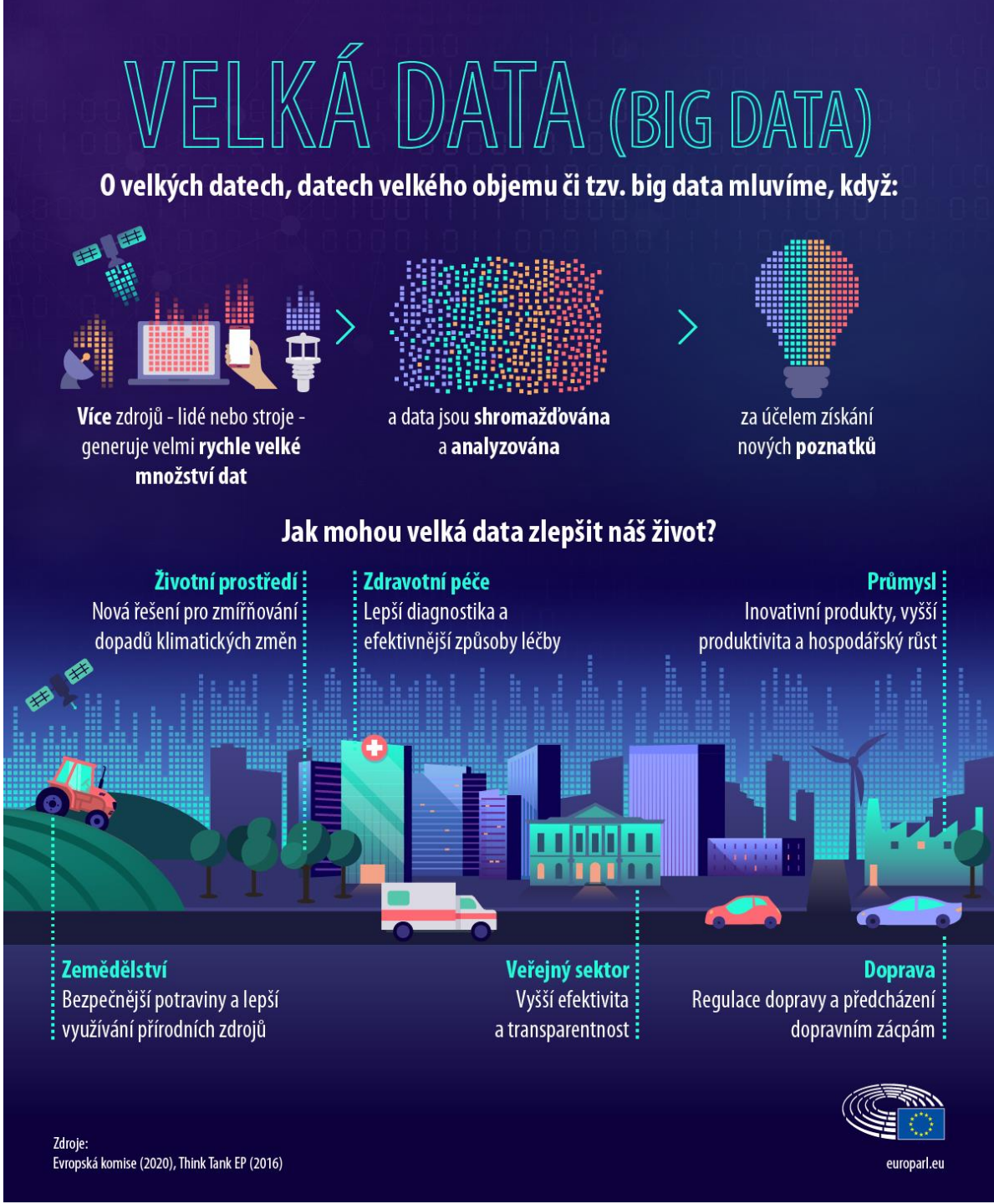
Finanční sektor:
Banky, pojišťovny a další finanční instituce shromažďují a analyzují data o transakcích, investicích a zákaznickém chování.

Zdravotnictví:
Elektronické zdravotní záznamy, lékařské výzkumy a biotechnologie.

Marketing a e-komerce:
Online prodejci a marketingové firmy shromažďují data o zákaznickém chování, preferencích a nákupních vzorcích.

Výroba a logistika:
Senzory a IoT zařízení v průmyslových procesech generují data o výrobě, údržbě a dodavatelských řetězcích.

(Data) Science



Vliv velkých dat na životní prostředí

- Spotřeba energie – nonstop energeticky náročný provoz; pokud není užita energie z obnovitelných zdrojů, přispívá to k emisím skleníkových plynů a změně klimatu.
- Elektronický odpad – rychlý růst objemu dat => rychlý technologický pokrok => častá obměna hardwaru => velké množství elektronického odpadu => možná kontaminace nebezpečnými látkami
- Využívání přírodních zdrojů - výroba elektronických zařízení a infrastruktury pro datová centra vyžaduje těžbu a zpracování surovin => možný negativní dopad na životní prostředí (odlesňování, znečištění vody a degradace půdy)

Možná řešení

- Využití obnovitelných zdrojů energie
- Efektivní chlazení – může snížit energetickou náročnost
- Recyklace a opětovné využití – snížení množství odpadu a potřeby nových surovin

Velká data jsou součástí života 21. století

- ✓ netvořte datový odpad, nekvalitní data
- ✓ využívejte jen kvalitní data
- ✓ data pokud možno ověřujte

ať v nich neutonete!