Proč nestrukturovaná data?

Prof. Ing. Zdeněk Molnár, CSc

VŠE Praha

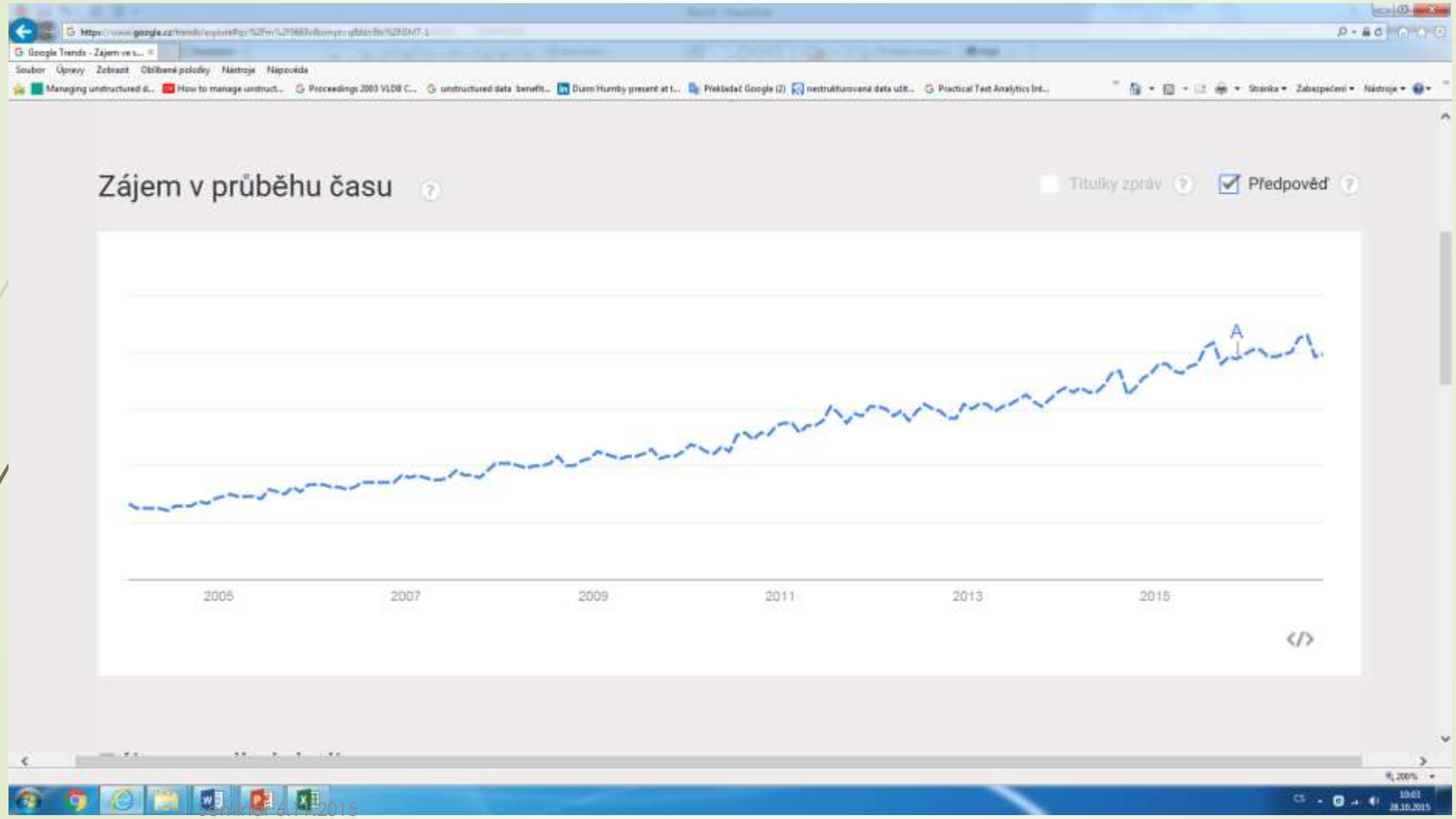
zdenek.molnar@vse.cz

Seminář 6.11.2015

Geneze tématu

- Zhruba před 10-ti lety intensivním zájmem o disciplínu **Competitive Intelligence.** Ta se zaměřuje převážně na vytěžování a analýzu externích nestrukturovaných dat (naslouchání co se kde děje, co se píše o mých produktech, o konkurenci a obchodních partnerech apod.)
- Čím dál tím intensivnější výskyt pojmů nestrukturovaná data a Big Data
- Růst významu sociálních sítí pro business
- Před 2 roky založení kompetenčního centra „Nestrukturovaná data“ při KIT na VŠE Praha

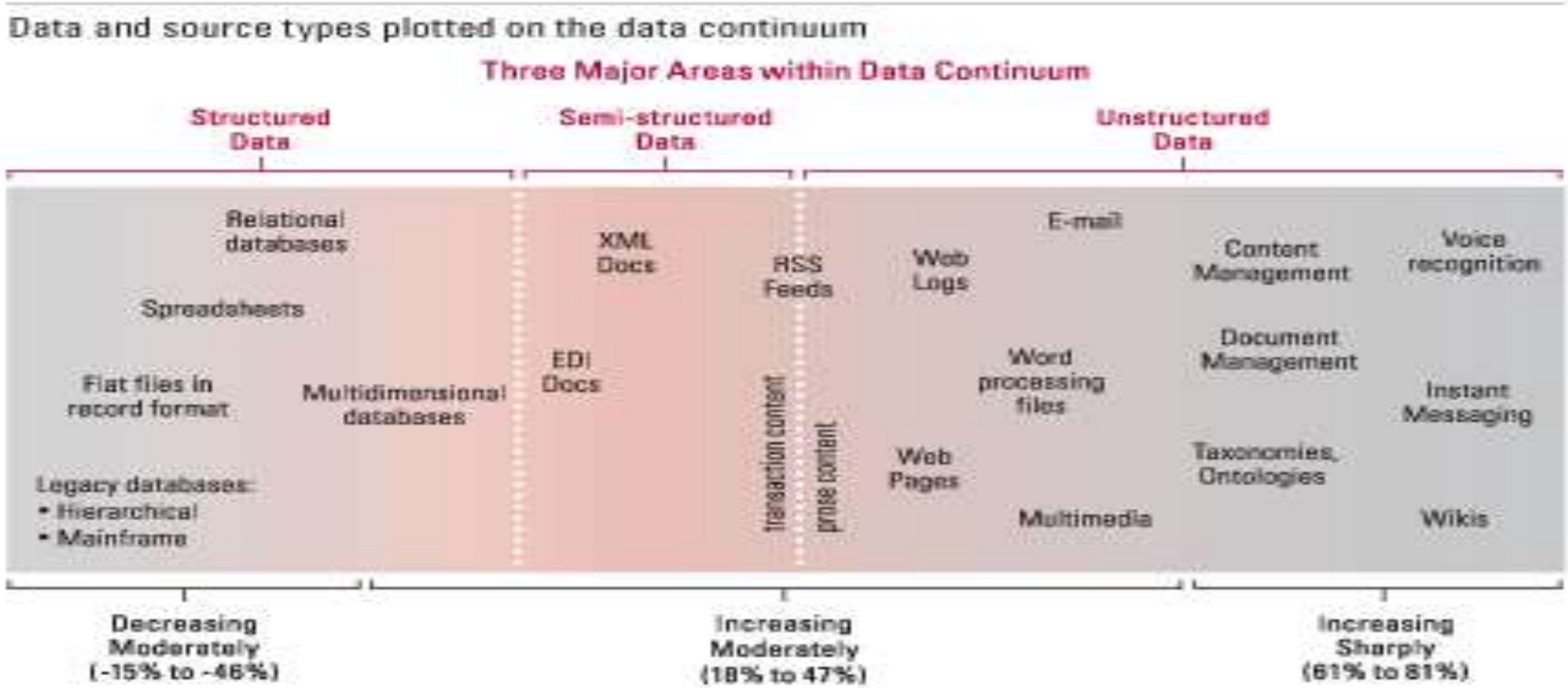
Google trends – nestrukturovaná data



Strukturovaná versus nestrukturovaná data

- Různé studie pak ukazují, že více než osmdesát procent všech dat v podniku má nestrukturovanou formu a že informační pracovníci tráví dnes téměř čtvrtinu svého času právě vyhledáváním informací potřebných pro různá, většinou strategická, rozhodnutí, přičemž čas pro jejich získání a analýzu je kritický pro daný byznys
- Strukturovaná data jsou reprezentována čísly, tabulkami, atributy a pod. a jsou výsledkem nějaké transakce. Stejné typy dat se vyskytují téměř ve všech typech transakcí a liší se jen hodnotami, kterých daný typ dat nabývá. Jsou disciplinovaná, „dobře se chovají“, jsou predikovatelná a opakovatelná
- Nestrukturovaná dat jsou dvojího typu: textová nestrukturovaná data a netextová nestrukturovaná data (tvary, barvy, zvuky, obrazy).
- Textová nestrukturovaná data ještě můžeme rozlišit na „uniformní“, (někdy nazývané jako semistrukturovaná - formuláře, spredsheety apod.) a volné texty (e-maily, zápisy, lékařská hlášení, webové stránky, blogy, apod.)

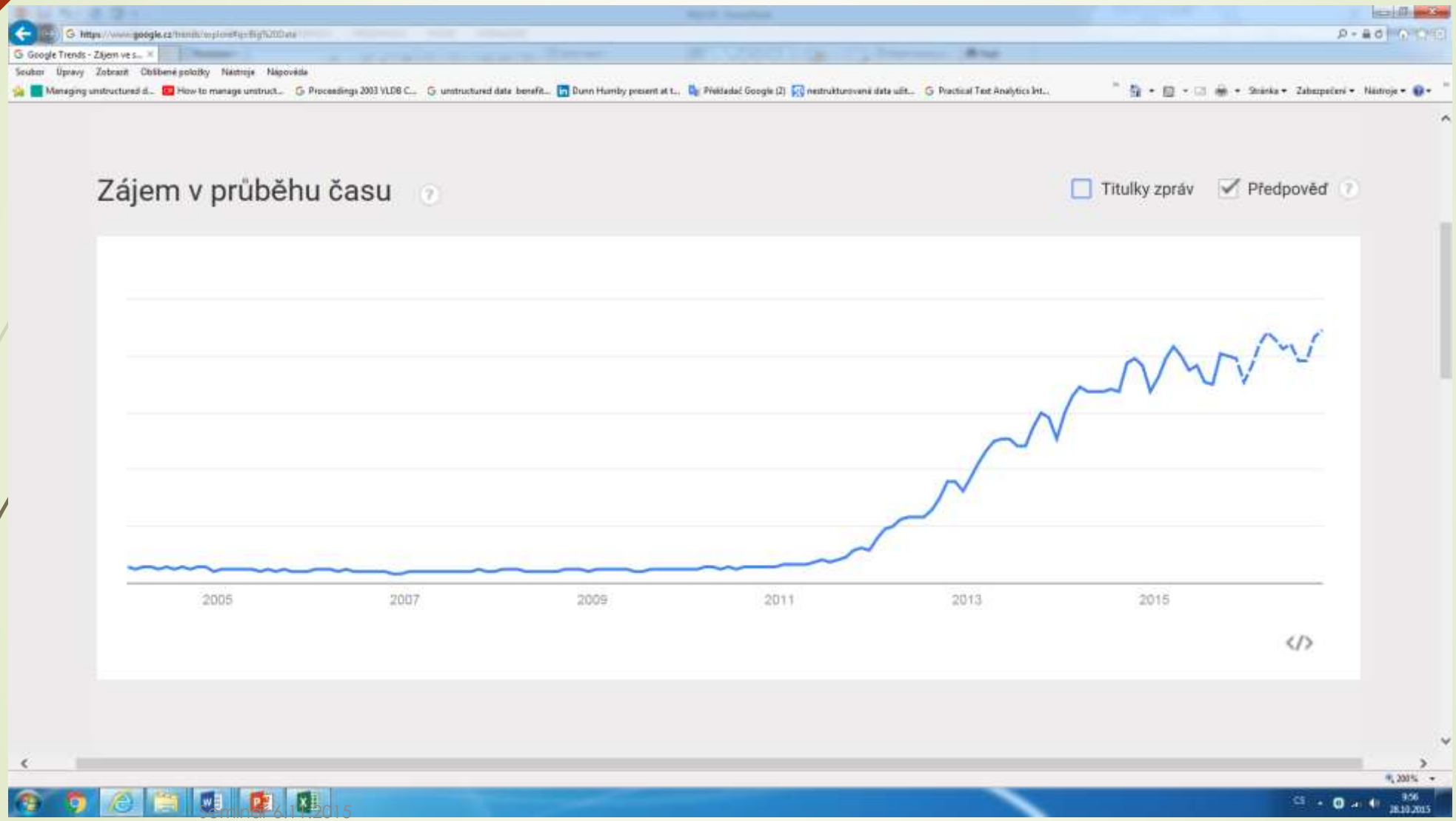
Philip Russom z The Datawarehouse Institute prezentoval v r. 2007 toto dělení s prognózou, která se naplňuje



Vznik Big Data a mylná představa o nich

- Nástup webu, mobilních zařízení a dalších technologií zapříčinil zásadní změnu charakteru dat a způsobu jejich využití.
- Data již nejsou centralizovaná, vysoce strukturovaná a snadno zvládnutelná, ale více než dříve jsou volně strukturovaná (pokud mají vůbec nějakou strukturu), vysoce distribuovaná a mají vzrůstající objem.
- V souvislosti s pojmem Big Data vzniká mylná představa, že se jedná jen o „hodně dat“. Pravda je ta, že objem dat je jasnou charakteristikou Big Data, ale Big Data jsou také o tom, že odkrývají problematiku strukturovaných a nestrukturovaných dat jak interních tak externích a to zejména z pohledu jejich analýzy v požadovaném čase.
- *Vědecko-technické výpočty se provádějí také nad ohromným objemem dat a neříká se jim Big Data.*

Google trends – Big Data



Nejčastěji uváděné charakteristiky Big Data

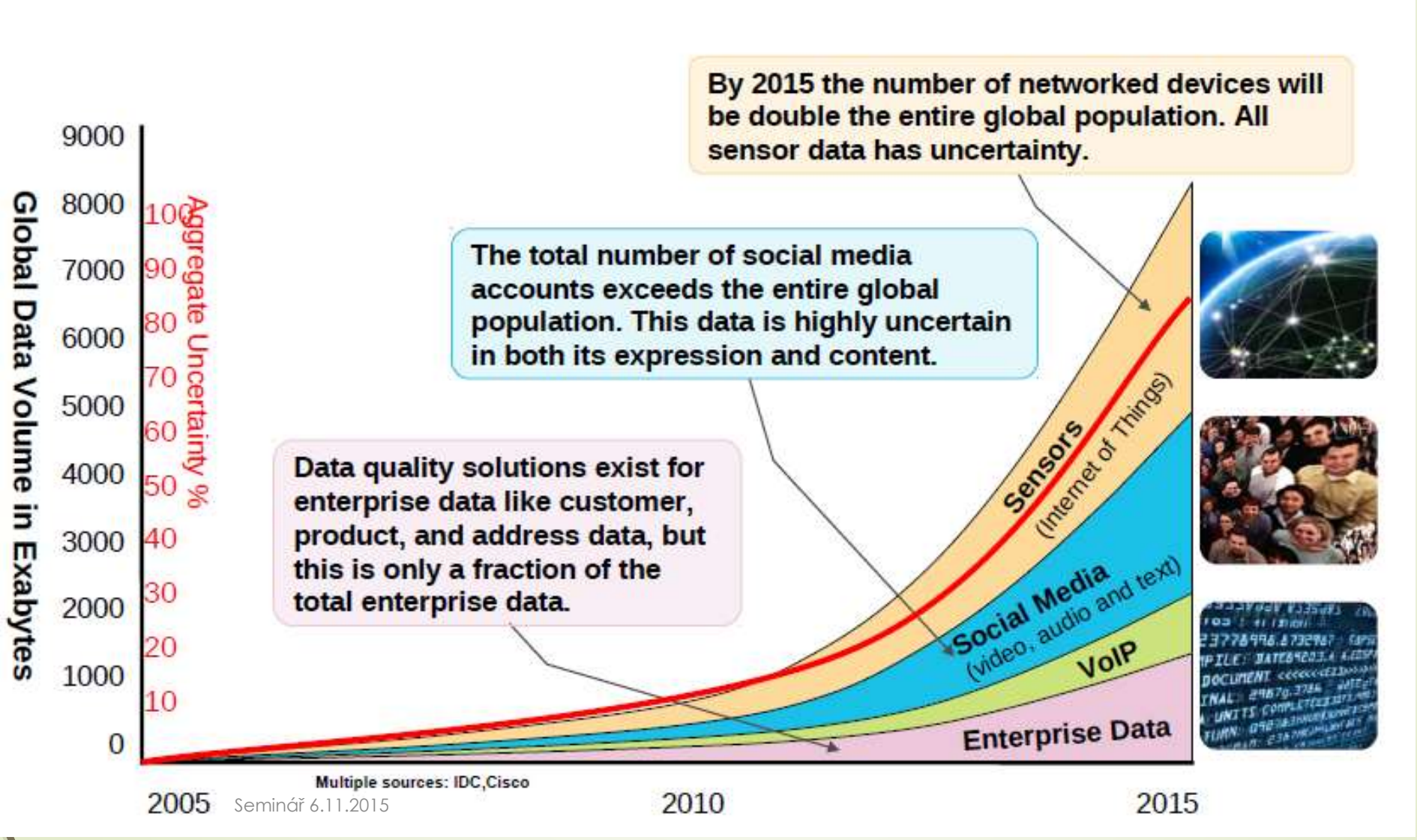
Volume – množství dat vznikajících v rámci provozu firem roste exponenciálně každý rok,

Velocity – rychlost s jakou data vznikají a potřeba jejich analýzy v reálném čase vzrůstá díky pokračující digitalizaci většiny transakcí, mobilním zařízením a vzrůstajícímu počtu internetových uživatelů

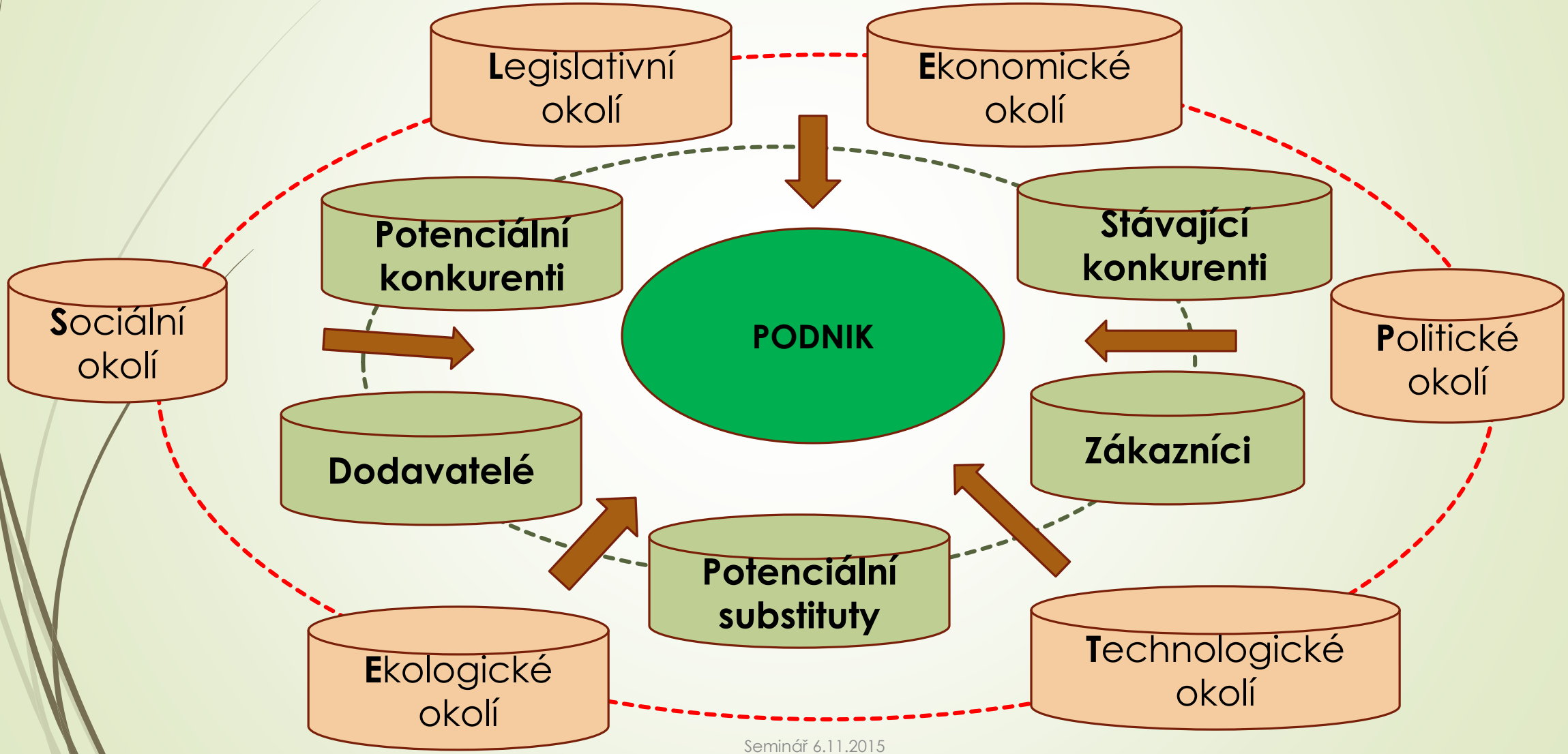
Variety – různorodost typů dat vzrůstá, například nestrukturované textové soubory, semi-strukturovaná data (XML), data o geografické poloze, data z čidel, videa apod.

Veracity - nejistá věrohodnost dat v důsledku jejich nekonzistence, neúplnosti, nejasnosti a podobně. Vhodným příkladem mohou být údaje čerpané z komunikace na sociálních sítích. Je nutné je čistit, vzájemně propojovat a korelovat jinak se snadno vymknou kontrole

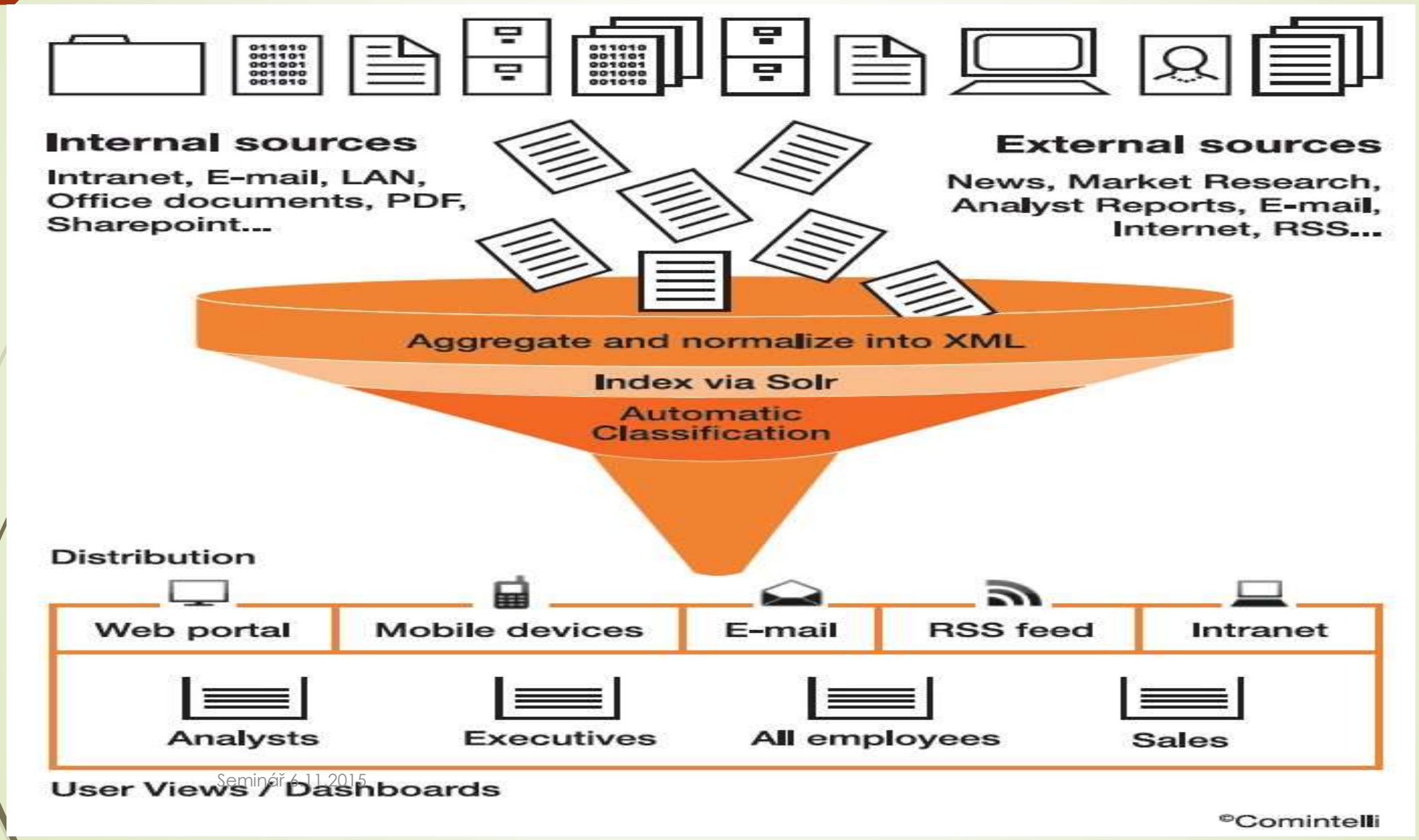
S rostoucím objemem dat roste i jejich neurčitost a nejistota



Roste i různorodost a komplexnost dat, která musíme zkoumat



Zpracování Big Data je o systematickém a účelovém spojování interních dat s externími a strukturovaných s nestrukturovanými



Zpracování Big data je proces sběru, organizování a analýzy s cílem pochopení významu a souvislostí těchto dat, které nám umožňuje nalézt odpovědi na otázky, které byly dříve nezodpověditelné nebo nepoložitelné.

Co se stalo?	Proč se to stalo?	Co se to děje?	Proč se to děje?	Co se asi stane?	Co bych s tím měl udělat?

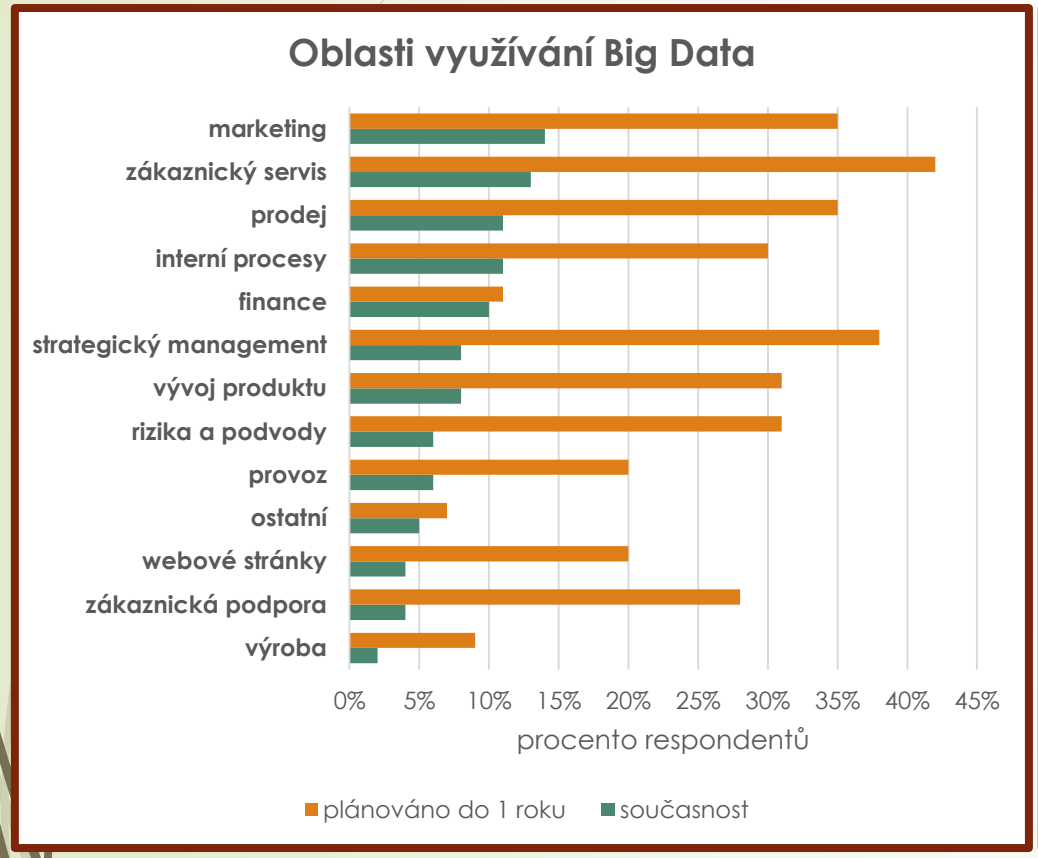
Cílem je predikce současných a budoucích hrozeb a příležitostí
Od deskriptivní analytiky přes prediktivní analytiku k preskriptivní analytice.

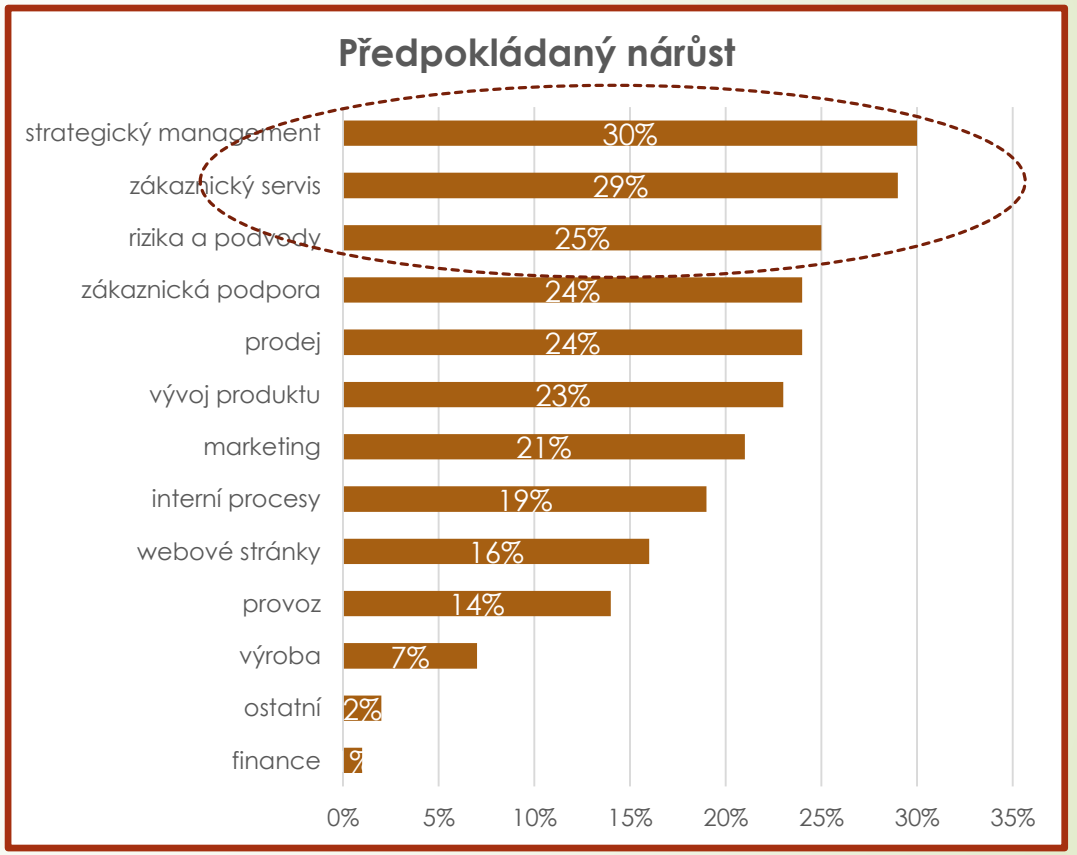
„Data jsou stejně jako ropa. Je to cenná surovina, ale pokud není rafinovaná, nelze jí použít. Proto, Big dat nejsou o tom, co máme, ale co chceme skutečně vědět a co s tím budeme dělat. " (Dunn Humby)

Průzkum využívání Big Data

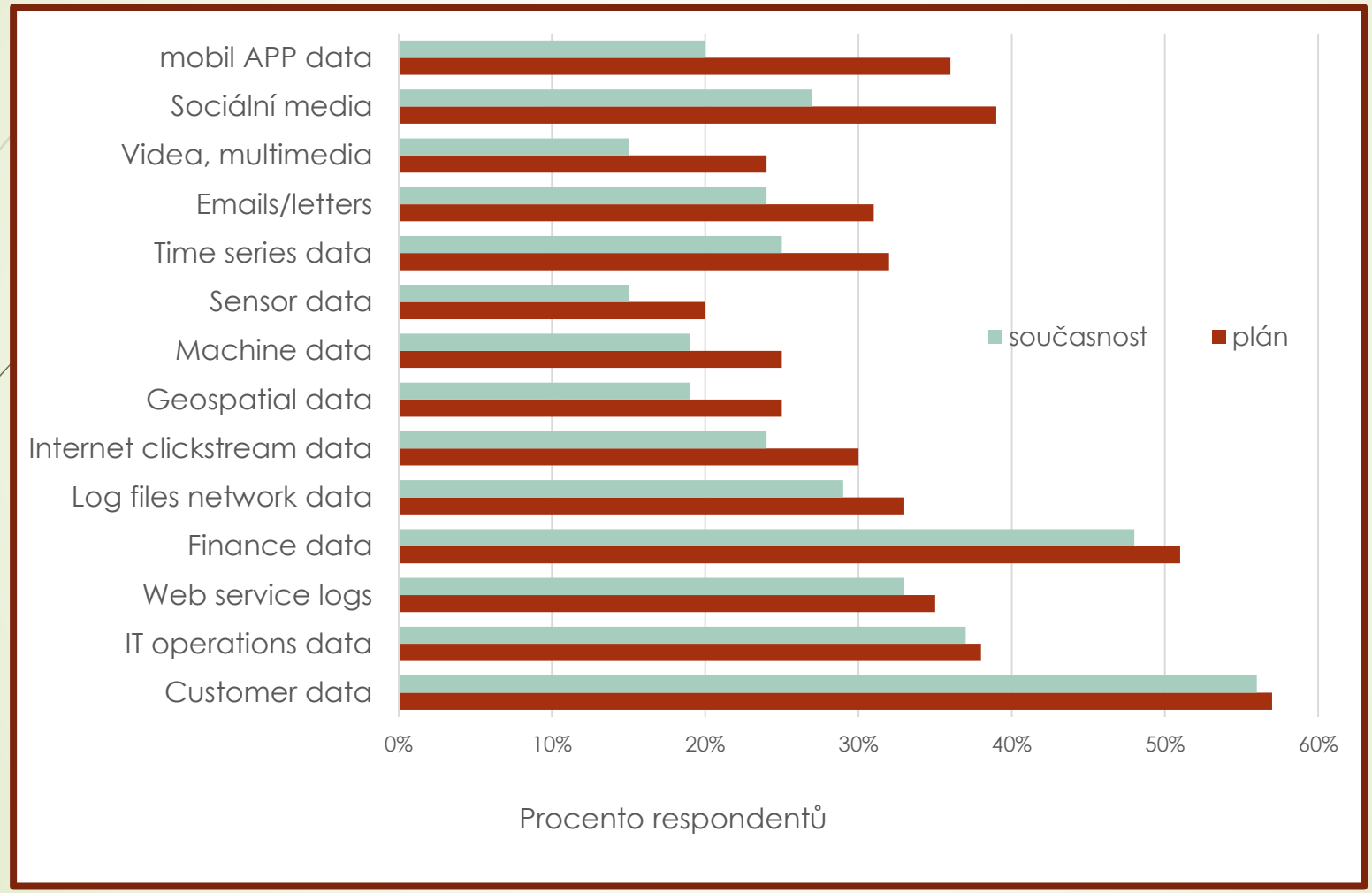
- Společnost TechTarget provedla v dubnu a květnu 2015 průzkum toho, jak společnosti analyzují Big Data. Zodpovědělo téměř 1000 respondentů ze Severní Ameriky, Evropy, Indie a Afriky.
- Z toho více jak jedna třetina měla více jak 5 tis. zaměstnanců
- 24% respondentů měla speciální Big Data programy, u 43% respondentů byla problematika Big Data řešena v rámci všeobecné IT strategie a 33% nemělo žádný projekt na Big Data
- 40% respondentů byli IT profesionálové, 20% analytici a pracovníci BI, 20% business konsultanti a 20% dodavatelé SW
- Výsledky jsou dostupné na:
- http://searchbusinessanalytics.bitpipe.com/data/demandEngage.action?resId=1435260274_463

Oblasti využívání a dosažené zlepšení

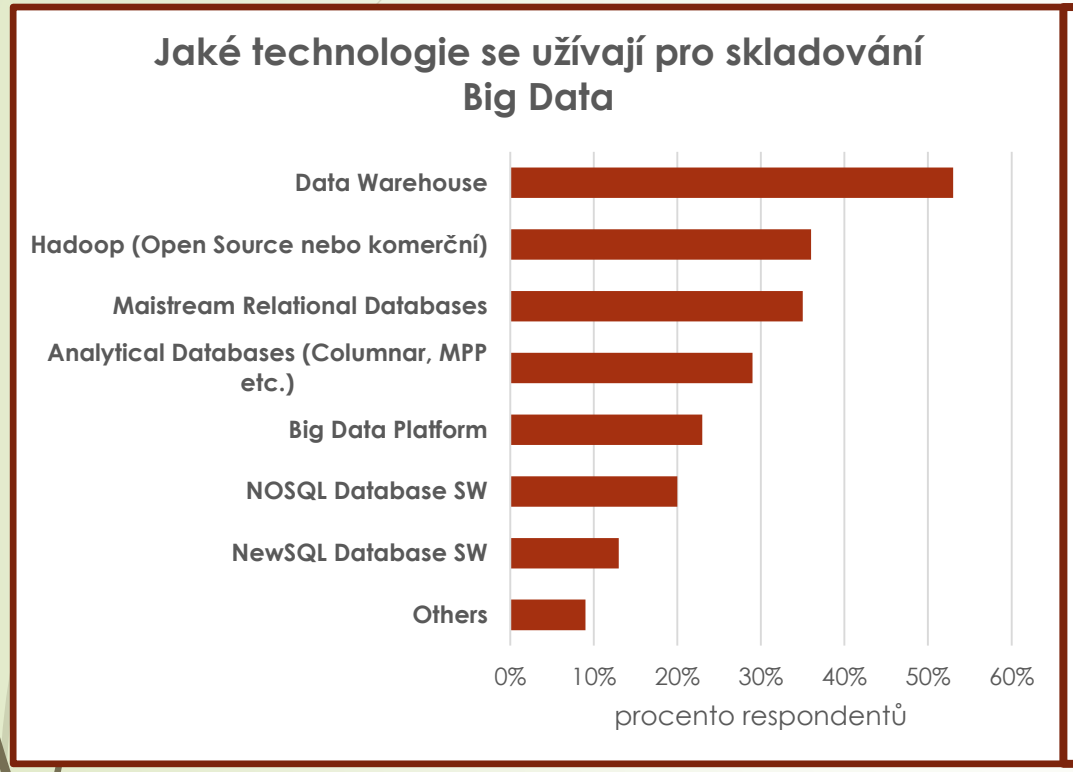


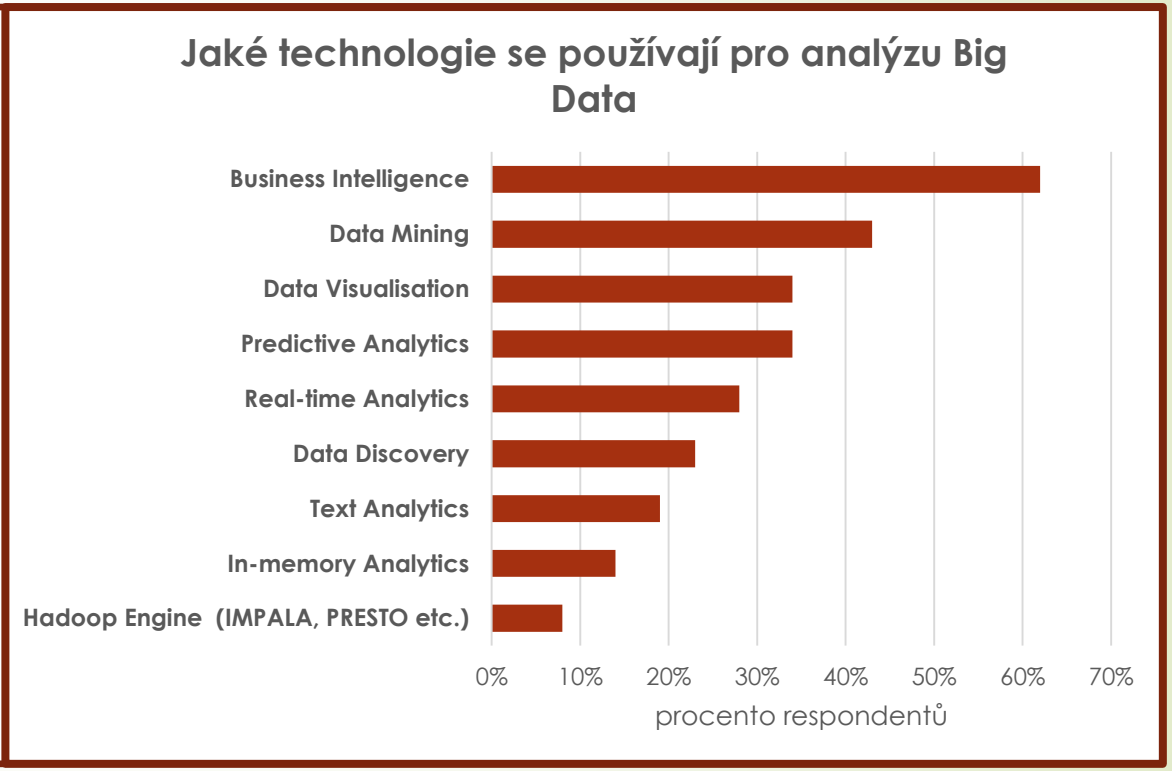


Zdroje analyzovaných dat

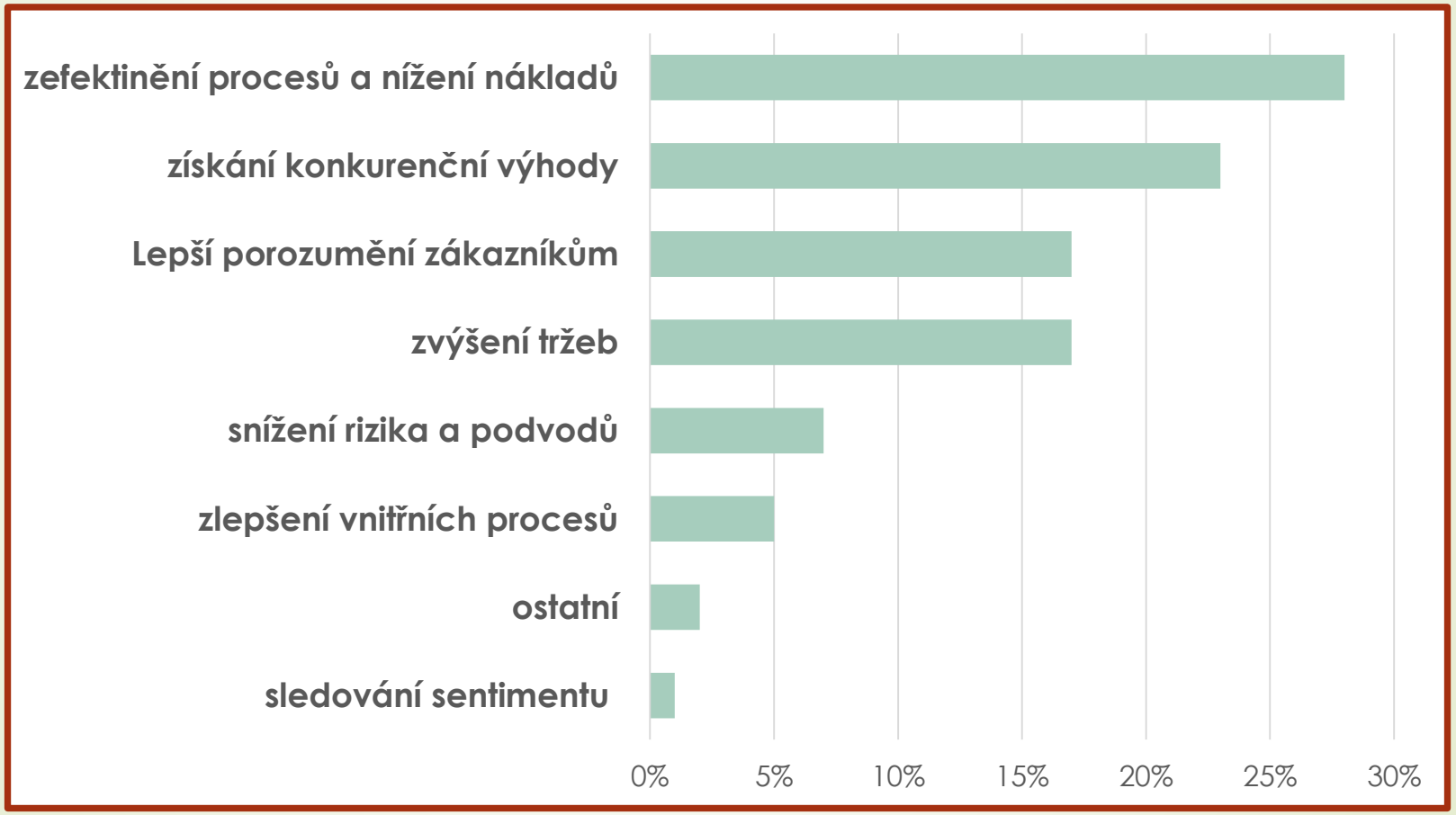


Big Data a technologie





Průměrná dosažená zlepšení aplikací Big Data



Děkuji za pozornost