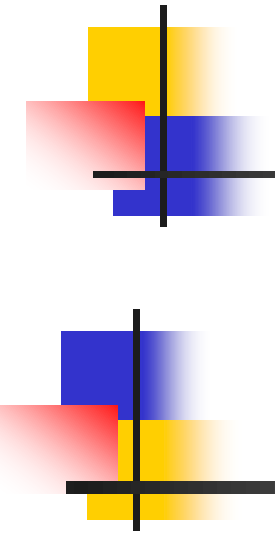
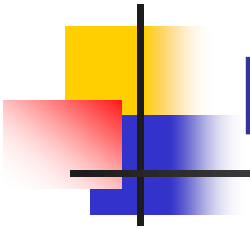


Propojená data na webu (motivační úvod)

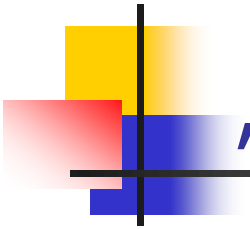
prof. Ing. Vojtěch Svátek, Dr.
Katedra informačního a znalostního inženýrství

ZS 2019

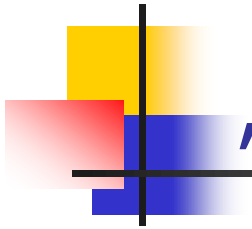


Propojená data na webu

- Jeden z frekventovaných termínů, které označují téměř totéž, např. v angličtině:
  - Linked (open?) data (on the web)
  - Web of data
  - Web of entities
  - Semantic web
  - Semantic technology ☹
  - Semantics ☹☹
  - Knowledge graphs !?



„Běžný“ vs. sémantický web

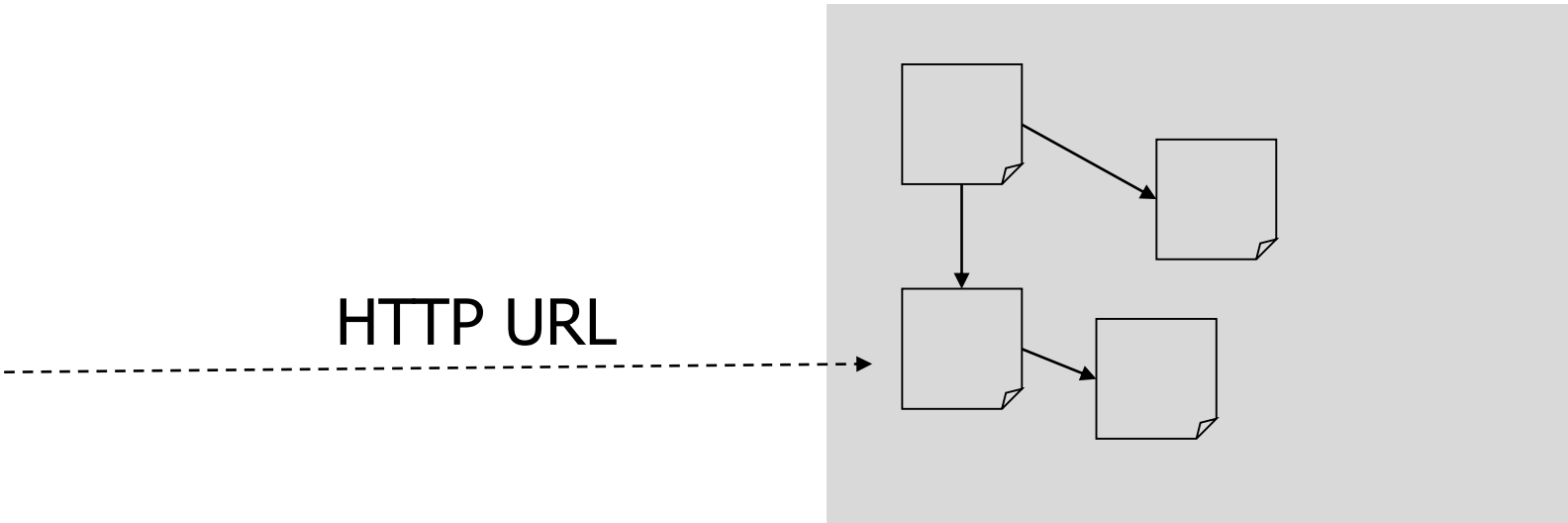


„Běžný“ vs. sémantický web

- „Běžný“ web
  - složený z dokumentů určených pro lidské uživatele
  - dokumenty jsou propojeny netypovanými odkazy

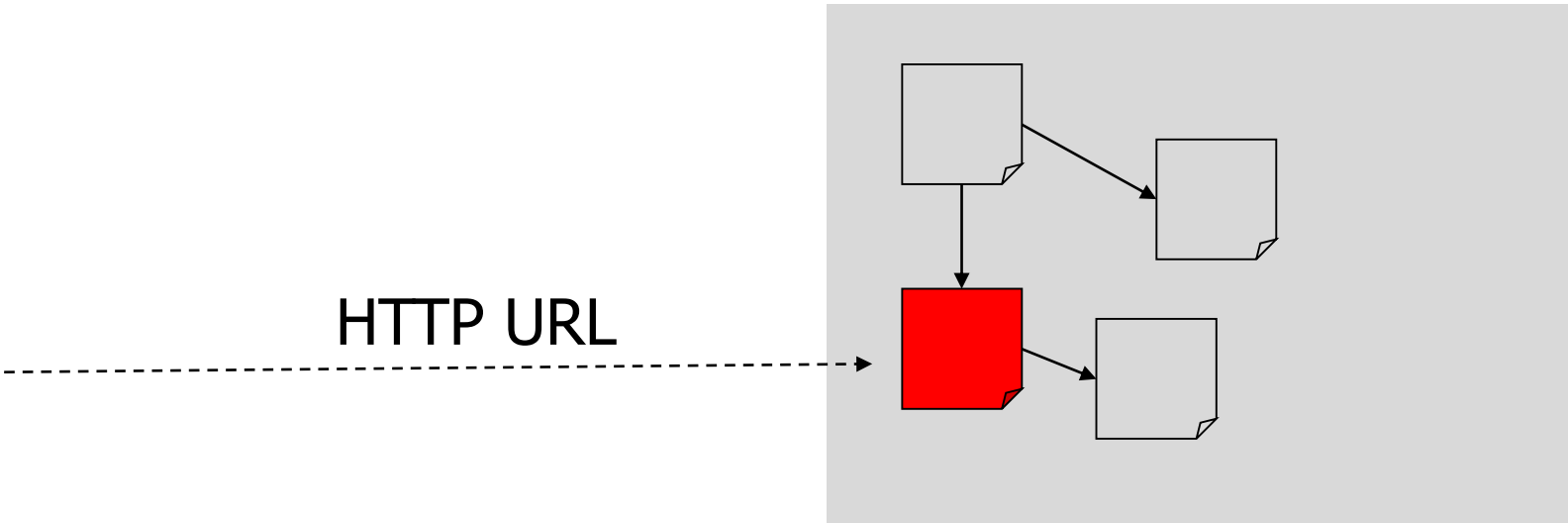
„Běžný“ vs. sémantický web

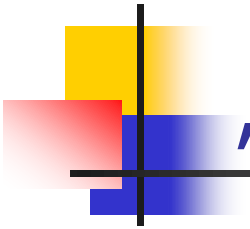
- „Běžný“ web
  - složený z dokumentů určených pro lidské uživatele
  - dokumenty jsou propojeny netypovanými odkazy
  - požadavek HTTP jde na server, ten vrací dokument



„Běžný“ vs. sémantický web

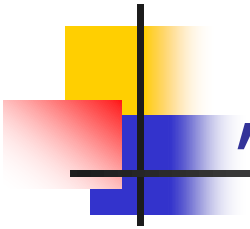
- „Běžný“ web
  - složený z dokumentů určených pro lidské uživatele
  - dokumenty jsou propojeny netypovanými odkazy
  - požadavek HTTP jde na server, ten vrací dokument





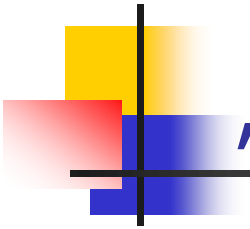
„Běžný“ vs. sémantický web

- Idea sémantického webu (T. Berners Lee)



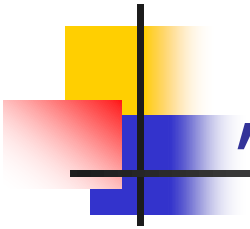
„Běžný“ vs. sémantický web

- Idea sémantického webu (T. Berners Lee)
  - propojit nejen dokumenty, ale i strukturovaná data



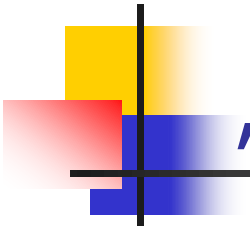
„Běžný“ vs. sémantický web

- Idea sémantického webu (T. Berners Lee)
  - propojit nejen dokumenty, ale i strukturovaná data
  - opatřit data strojově čitelným významem



„Běžný“ vs. sémantický web

- Idea sémantického webu (T. Berners Lee)
  - propojit nejen dokumenty, ale i strukturovaná data
  - opatřit data strojově čitelným významem
  - propojit je typovanými odkazy

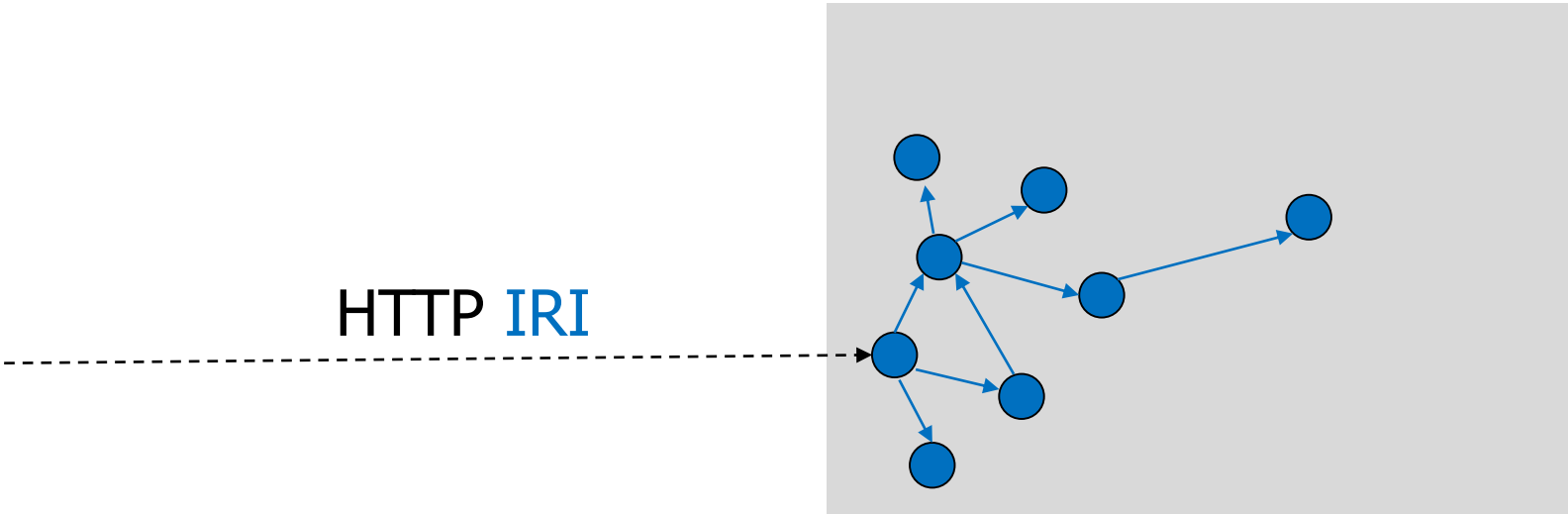


„Běžný“ vs. sémantický web

- Idea sémantického webu (T. Berners Lee)
  - propojit nejen dokumenty, ale i strukturovaná data
  - opatřit data strojově čitelným významem
  - propojit je typovanými odkazy
  - požadavek HTTP bude vracet množinu dat

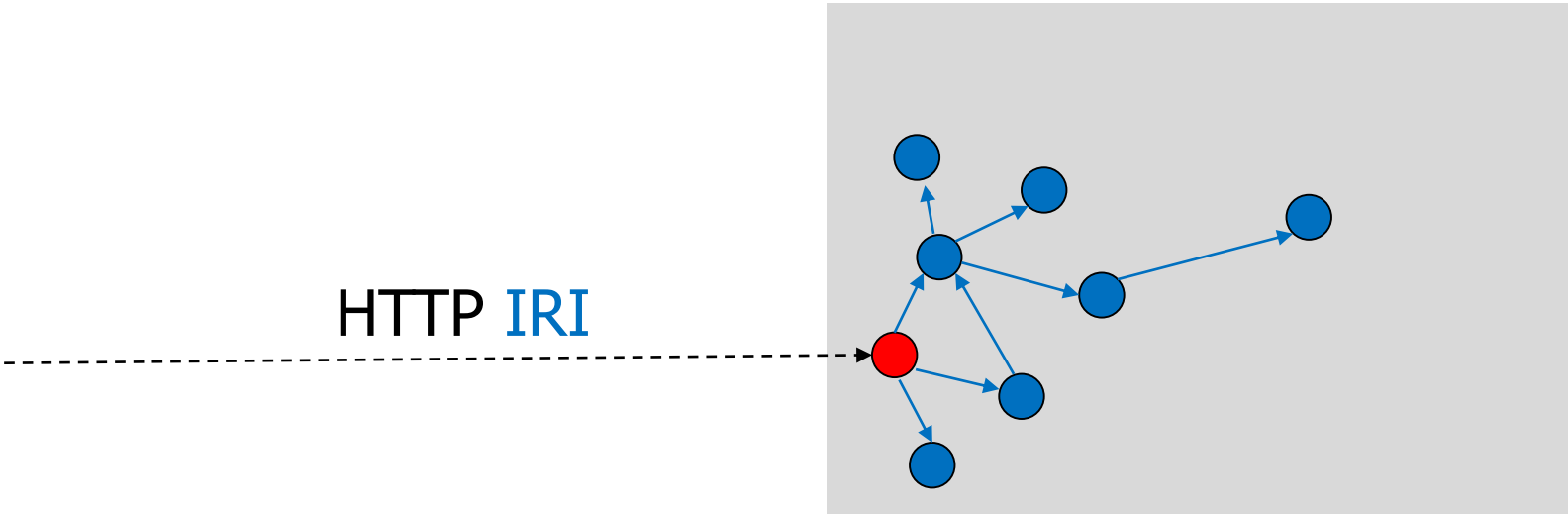
„Běžný“ vs. sémantický web

- Idea sémantického webu (T. Berners Lee)
  - propojit nejen dokumenty, ale i strukturovaná data
  - opatřit data strojově čitelným významem
  - propojit je typovanými odkazy
  - požadavek HTTP bude vracet množinu dat



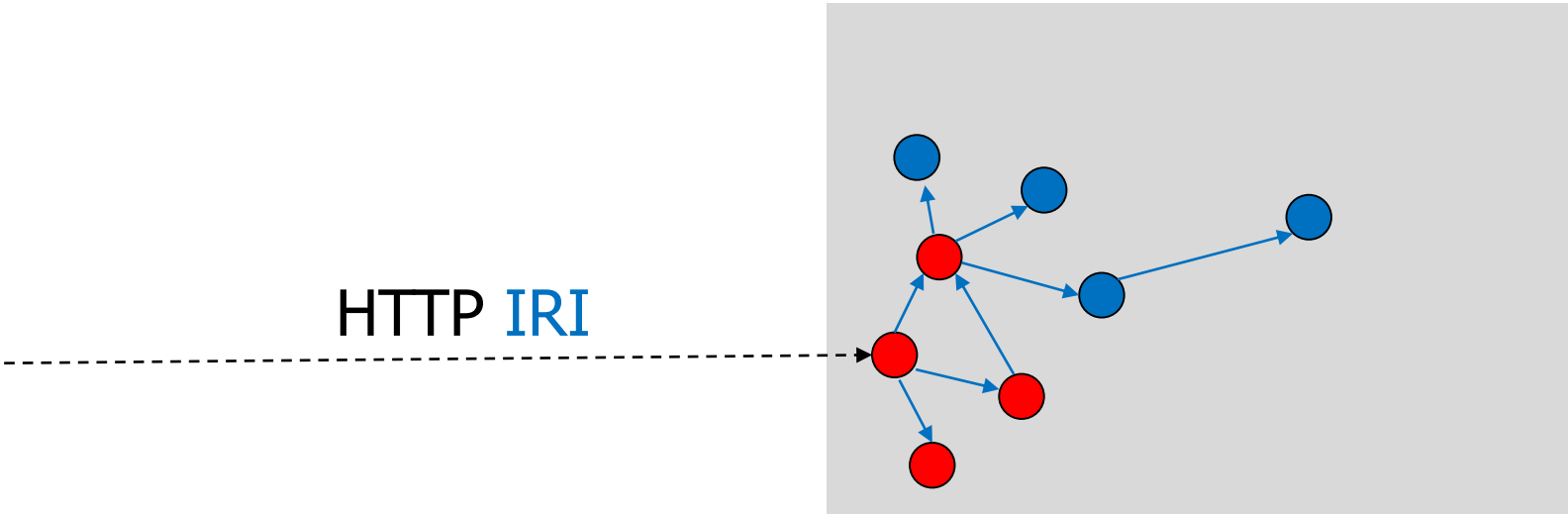
„Běžný“ vs. sémantický web

- Idea sémantického webu (T. Berners Lee)
  - propojit nejen dokumenty, ale i strukturovaná data
  - opatřit data strojově čitelným významem
  - propojit je typovanými odkazy
  - požadavek HTTP bude vracet množinu dat



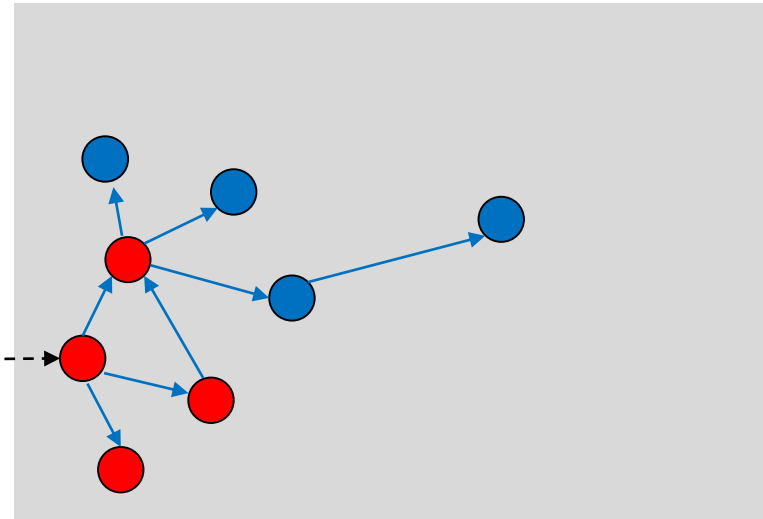
„Běžný“ vs. sémantický web

- Idea sémantického webu (T. Berners Lee)
  - propojit nejen dokumenty, ale i strukturovaná data
  - opatřit data strojově čitelným významem
  - propojit je typovanými odkazy
  - požadavek HTTP bude vracet množinu dat

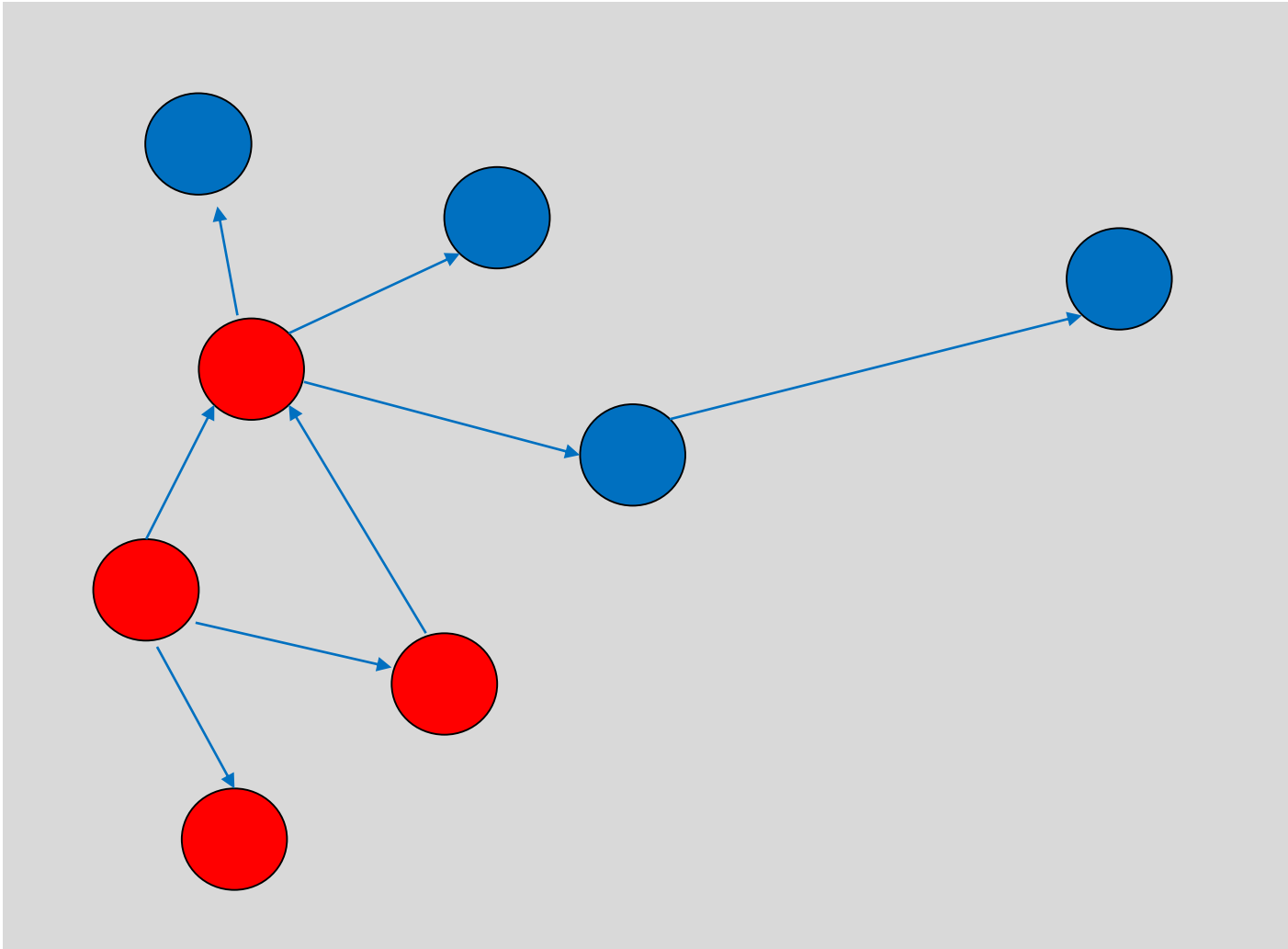


Strojově čitelný význam

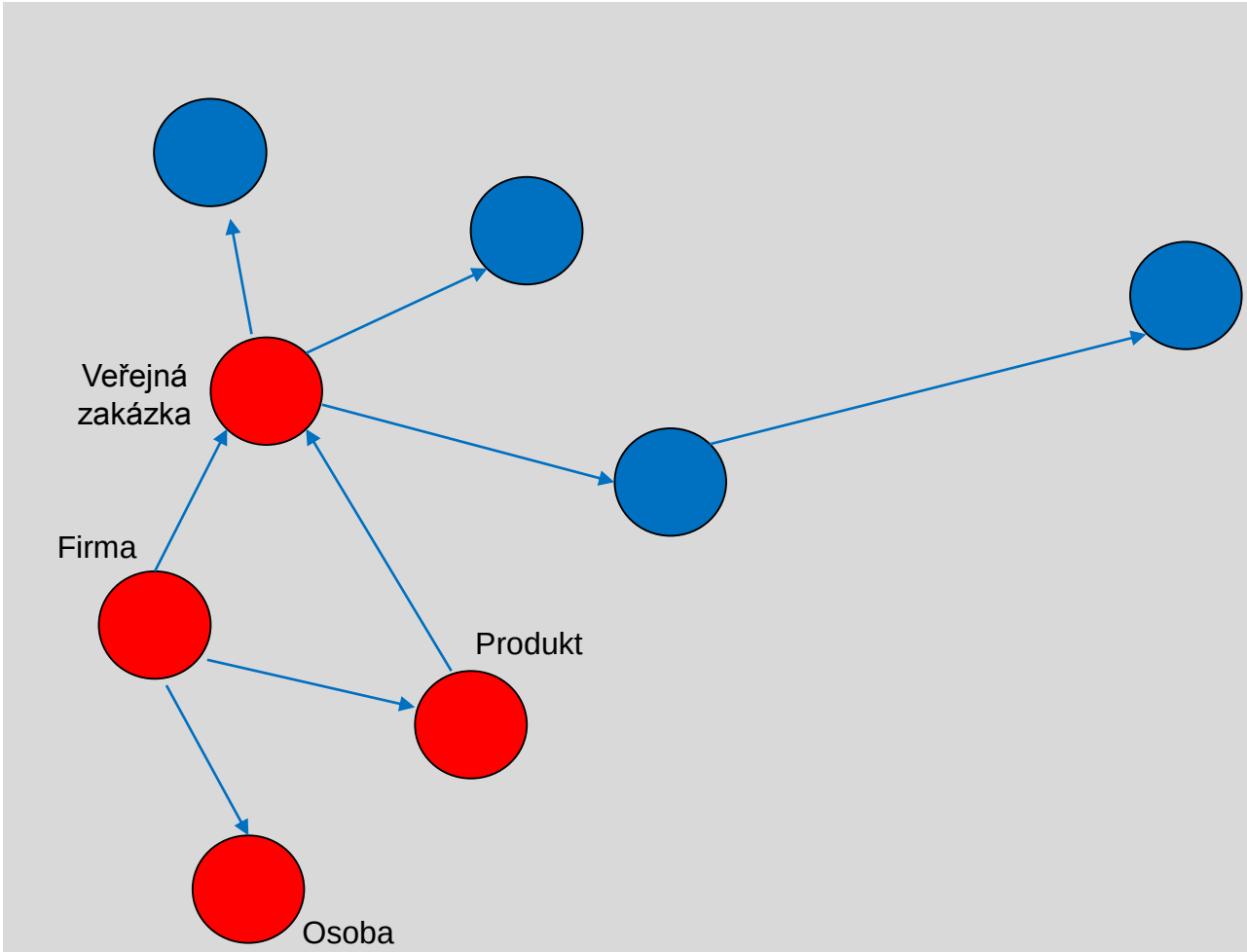
HTTP IRI



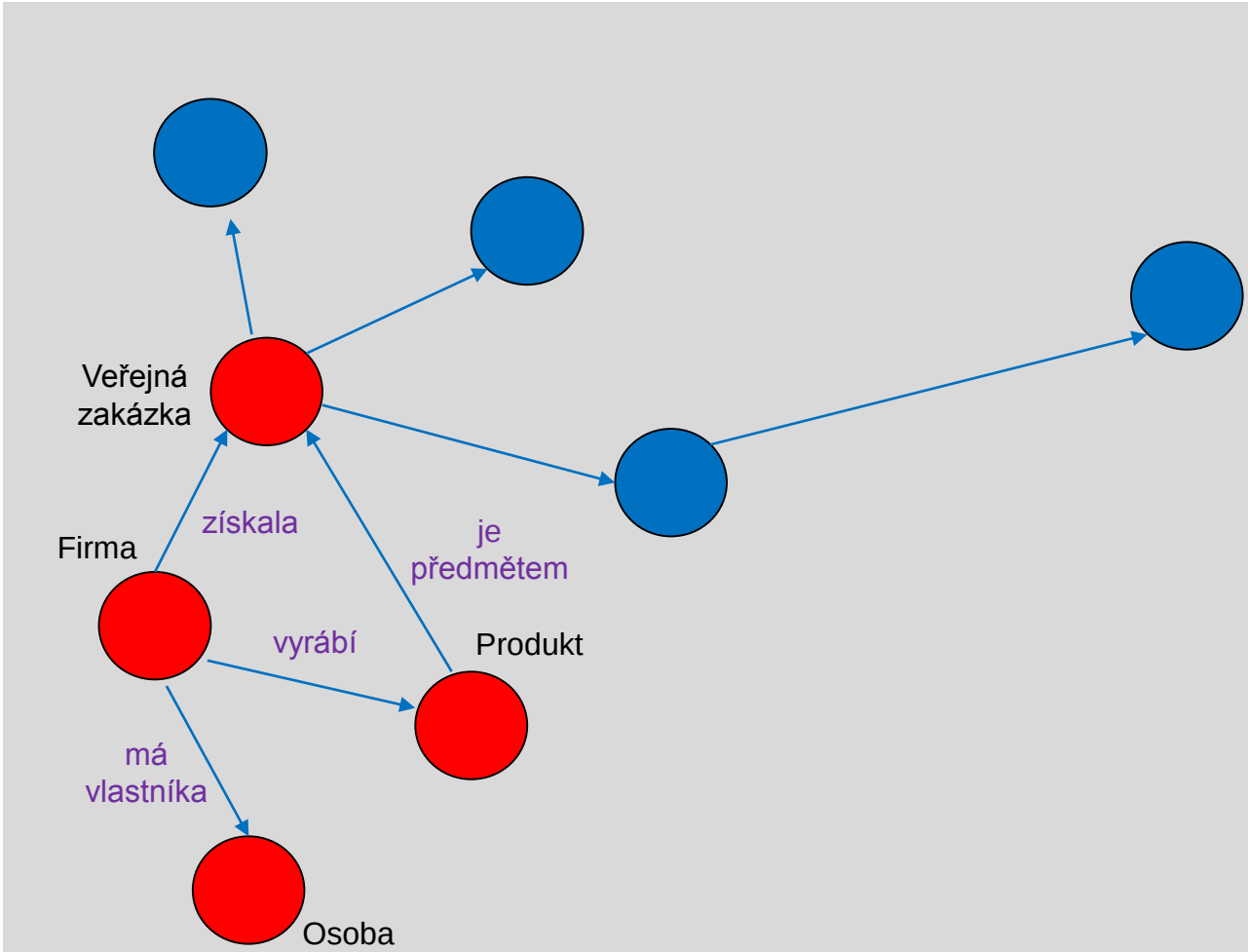
Strojově čitelný význam



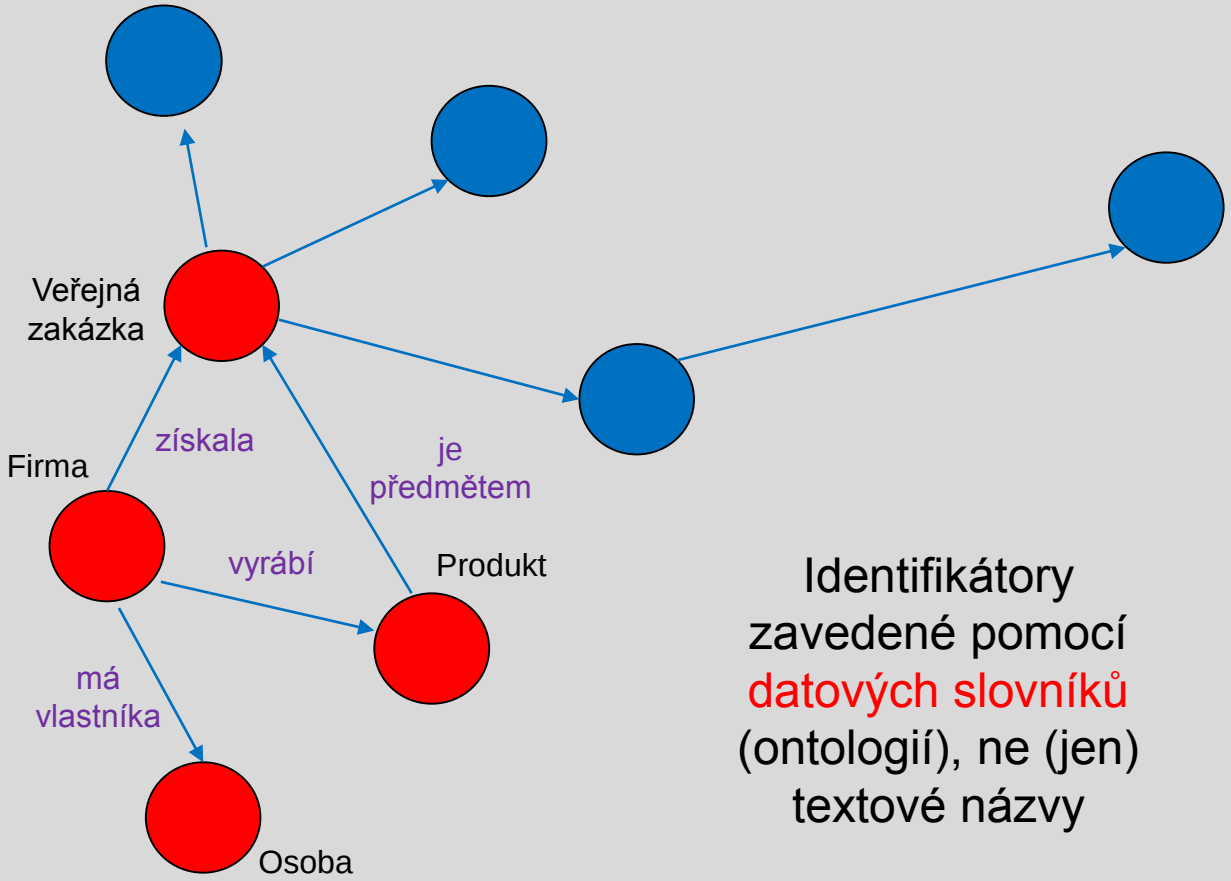
Strojově čitelný význam



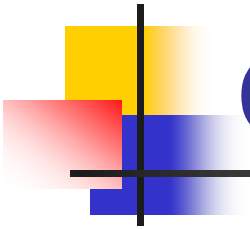
Strojově čitelný význam



Strojově čitelný význam

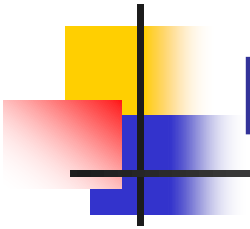


Identifikátory zavedené pomocí datových slovníků (ontologií), ne (jen) textové názvy



Cíl předmětu

- Naučit se data na sémantickém webu
  - **vytvářet a vystavovat**, včetně
    - návrhu datových schémat a (doménových) slovníků
    - extrakce dat, která nativně v RDF nejsou
  - **Zpracovávat**, tj. čistit, transformovat a propojovat
  - **využívat** v jednoduchých aplikacích



Probíraná témata

- **RDF**: jazyk pro reprezentaci propojitelných dat na webu v obecné grafové struktuře
- **SPARQL**: dotazovací jazyk pro RDF
- **Datové slovníky** (ontologie): popis významu RDF dat existujícími slovníky, i tvorba nových slovníků
- Proces **zpracování dat** („ETL“) vedoucí k jejich vystavení v RDF: extrakce, transformace (včetně čištění a propojování) a výsledné publikování
- Využívání RDF dat v **softwarových aplikacích**
- **Byznys modely** podporující využití RDF dat na webu

Dimenze „uvažování o tématu“

- Životní cyklus dat
  - zdrojová (ne-RDF) data – RDF schémata a slovníky – ETL – – dotazování pomocí SPARQL – využití v aplikaci
- Věcné domény dat
  - encyklopedická; produkty/služby; statisticko-ekonomická (jiné: biomedicína, akademická, knihovnictví, …)
- Myšlenková prostředí
  - Věda: výzkumné projekty, experimenty a publikace
  - Inženýrství: reprezentační jazyky a SW nástroje
  - Byznys: modely přínosů a nákladů při poskytování/využívání dat

Co předmět je a není?

- Nejde o desítky let ustálenou disciplínu, pro kterou by existovaly „nepohnutelné" osnovy (i když dobré knihy a příručky už vyšly!)
- Obsahem výuky není jedna přesně vymezená technologie (jazyk, metodika, software…)

ALE

- Jde o relativně volné propojení jazykových standardů, softwarových nástrojů, výzkumných projektů a iniciativ…
- „Správné odpovědi" na mnohé klíčové otázky zatím nikdo na 100% nezná, obor se rok od roku rychle vyvíjí
  - metafora „lesní školky" – něco vyroste, něco padne a pohnojí
- Vzhledem k tomu je u studentů velmi vítána vlastní iniciativa, zvědavost, zkoumání toho, co se nově objevuje jak na akademických konferencích, tak i např. v blogpostech

Prof. Anja Feldmann
Pražský informatický seminář, 27.4.2017

The Internet: A fascinating (research) object

Témata přednášky:
detekce výpadků sítě, predikce výkonu distribuovaných systémů, apod.

Prof. Anja Feldmann
Pražský informatický seminář, 27.4.2017

The Internet: A fascinating (research) object

Témata přednášky:
detekce výpadků sítě, predikce výkonu distribuovaných systémů, apod.

Internet, ale i web jako jeho nejviditelnější součást, jsou fascinujícími tématy výzkumu.

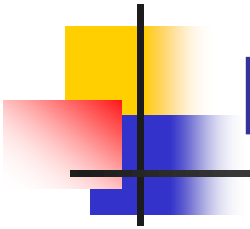
Prof. Anja Feldmann
Pražský informatický seminář, 27.4.2017

The Internet: A fascinating (research) object

Témata přednášky:
detekce výpadků sítě, predikce výkonu distribuovaných systémů, apod.

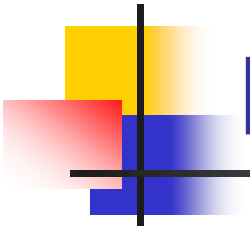
Internet, ale i web jako jeho nejviditelnější součást, jsou fascinujícími tématy výzkumu.

Neplatí to ještě více o spojení internetu/webu s možností předávat si přes něj strukturovaná data s popsaným významem?



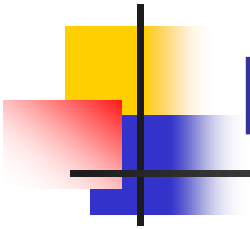
Možnosti uplatnění

- Univerzity
  - Dělá se v nějaké podobě skoro na všech
- Veřejná správa
  - Otevírání datových zdrojů státní správy, měst a obcí
- Výzkumná centra zahraničních korporací
  - 2 doktorandi KIZI „sdíleni“ s výzkumem v korporacích
- Tech startupy
  - Domácí (např. ContextMinds), ale zejména zahraniční: práce na dálku - Praha je „kompetenčním centrem“ pro PDW
- „Běžné“ (velké) firmy
  - **Ano** search engines, business info; pharma; automotive; …
  - **Ne** banky nebo SW firmy - spíš „mainstream“ technologie



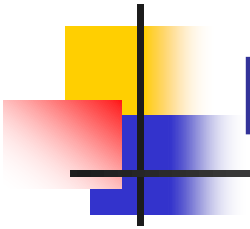
PDW, přednášející a katedra

- PDW spadá do tématiky pracovní skupiny SWOE („Semantic Web and Ontological Engineering“)
  - Jedna z pracovních skupin KIZI, viz https://kizi.vse.cz/english/research-groups/
  - Vedoucí V. Svátek a O. Zamazal
  - Skupina zaměřená primárně *vědecky*…
  - …spíše sekundárně na *spolupráci s praxí* a na (magisterskou) *výuku*; PDW je jediný takto zaměřený předmět povinný pro ZWT



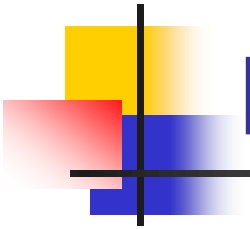
PDW, přednášející a katedra

- SWOE v kontextu ostatních pracovních skupin
  - Využití webu jako platformy pro sdílení dat
    - viz webové inženýrství - skupina WELT
    - v nové verzi programu ZWT společný „webový“ profil!
  - Využití strojového odvozování
    - viz umělá inteligence - skupina IIS
  - Příprava „vysokých“ (sémanticky popsaných) a „širokých“ (propojených) dat pro analytické využití
    - viz data mining/science - skupina DMKD
- Tj. skupina do značné míry průniková, lze propojit s kteroukoli z ostatních



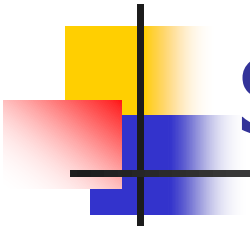
PDW, přednášející a katedra

- SWOE v kontextu ostatních pracovních skupin
  - Využití webu jako platformy pro sdílení dat
    - viz webové inženýrství - skupina WELT
    - v nové verzi programu ZWT společný „webový“ profil!
  - Využití strojového odvozování
    - viz umělá inteligence - skupina IIS
  - Příprava „vysokých“ (sémanticky popsaných) a „širokých“ (propojených) dat pro analytické využití
    - viz data mining/science - skupina DMKD
- *Kdo je tady spíš webový vývojář, nadšenec do AI, nebo datový analytik?*



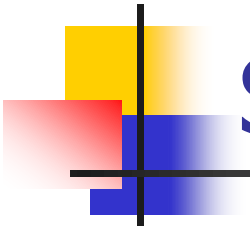
PDW, přednášející a katedra

- (Jen na okraj…) hlavní oblasti odborného zájmu přednášejícího
  - Ontologické inženýrství
    - úzká nika, v ČR převážně jen akademický výzkum
    - skupina SWOE zde patří k širší světové špičce oboru
  - Publikování a využívání propojených dat
    - i v ČR už řadu let populární téma, zvl. ve veřejné správě
    - zapojení do projektů EU, hlavní zdroj financování skupiny SWOE
    - **hlavní téma předmětu 4iz440**
  - Data a text mining (historicky – teď už spíš jen vedení DP, ev. DisP)
  - Zpracování přirozeného jazyka („samouk“)
  - Byznys informatika („zainteresovaný laik“)
    - plyne z pozice garanta doktorského oboru – sledování témat doktorandů všech 4 informatických kateder
    - hledání prostoru pro vědecká témata v této oblasti



Související předměty

- 5FI430 – Znalosti a ontologické inženýrství
  - M. Vacura, KFIL, povinný pro KI
  - Problematika ontologických modelů jako bohatší varianty datových slovníků – formální odvozování nad koncepty, filozofické ukotvení atd.
- 4IZ470 – Dolování znalostí z webu
  - V. Svátek, povinný pro ZWT (mnozí již absolvovali…)
  - Součástí „web miningu" je extrakce informací z webových textů – komplementární k převodu již strukturovaných dat do RDF
- 4IZ530 – Logické programování
  - Š. Sem, oborově volitelný pro ZWT a KI
  - Programování v jazyce Prolog; má řadu společných rysů s psaním dotazů ve SPARQL, lze v něm psát sémantické aplikace



Související předměty

- 4IZ??? – Grafová data a jejich vizualizace
  - M. Dudáš a V. Svátek
  - Nově připravovaný předmět (bude zřejmě otevřen už v LS 2020) spojuje dvě aktuální témata
    - Grafové databáze a dotazovací jazyky
    - Vizualizaci dat
  - Grafové databáze jsou na rychlém vzestupu
    - Viz např. https://www.stardog.com/blog/nasa-stardogs-going-to-mars/ ☺
  - Hlavní přístupy k reprezentaci dat (a dotazování do nich) v GDB jsou
    - RDF, s dotazováním pomocí SPARQL
    - LPG (labeled property graphs), s různými dotazovacími jazyky (Cypher, Gremlin, GraphQL)
  - Nový předmět bude pokrývat oba přístupy; v případě SPARQL bude přímo navazovat na 4IZ440

Pro hlubší zájemce

- Pracovní skupina KIZI **SWOE**
  - účastní se i studenti, zvl. v rámci DP

…weeding the semantic web garden

- Neformální výzkumný seminář **KEG** „Knowledge Engineering Group“
  - některé (1./2.) čtvrtky v semestru od 16 hodin zpravidla na 473NB (zasedačka FIS)
  - možno přijít bez předchozího přihlášení
  - viz http://keg.vse.cz/seminars
  - kdo chcete dostávat oznámení, napište!
- Aktivity mezi-institucionální iniciativy **OpenData.cz**: http://opendata.cz
  - vč. společného týmu KIT a KIZI, http://opendata.vse.cz/

Příklady zajímavých DP na SWOE (zpracování dat)

- Řízení kvality otevřených dat veřejné správy
  - ETL a využívání otevřených dat města Děčín
- Zpřístupnění dat o VŠ kvalifikačních pracech z Národního úložiště šedé literatury v podobě propojených dat
  - ETL vč. heuristické deduplikace osob
- ETL z české Wikipedie do znalostního grafu DBpedia
  - Vč. typologie chyb vyskytujících se ve znalostním grafu
- Webové aplikace pro vizualizaci profilů RDF datasetů
  - Pro farmaceutickou firmu Merck, integrováno do jejich DB
  - Grafová vizualizace nad nástrojem KIZI LODSight
- Využití propojených dat (DBpedia) ke tvorbě strategické znalostní hry

Příklady zajímavých DP na SWOE
(tvorba a management slovníků - ontologií)

- Ontologie přístupnosti budov
  - Využitelná službami pro handicapované
  - Prezentováno na konferenci SEMANTiCS v Lipsku
- Modelování událostí v ontologiích
  - Na základě DP (nominované do soutěže ACM SPY) vznikl příspěvek na CORE A konferenci „Formal Ontology in Information Systems“, Francie
- Párování entit ze standardní a lokální biomedicínské ontologie
  - Zadání od pharma firmy Merck, návazně přijetí na výzkumnou pozici v korporaci v kombinaci s doktorským studiem na FIS