Aktuálne trendy v **informatike**
Spracovanie dát (use cases)

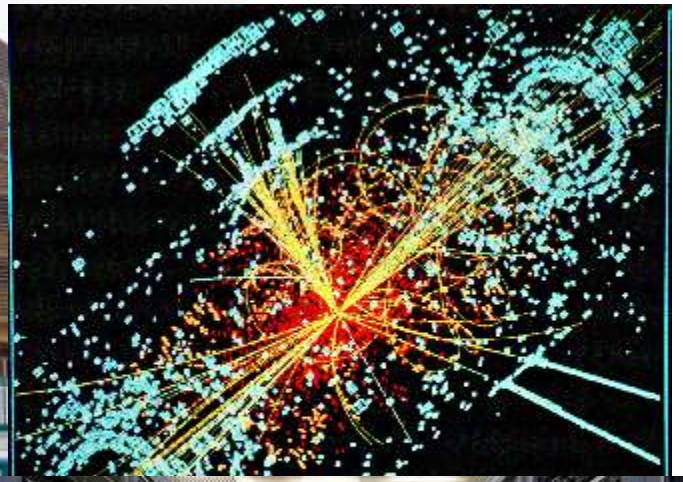
Genči
Katedra počítačov a informatiky
2017

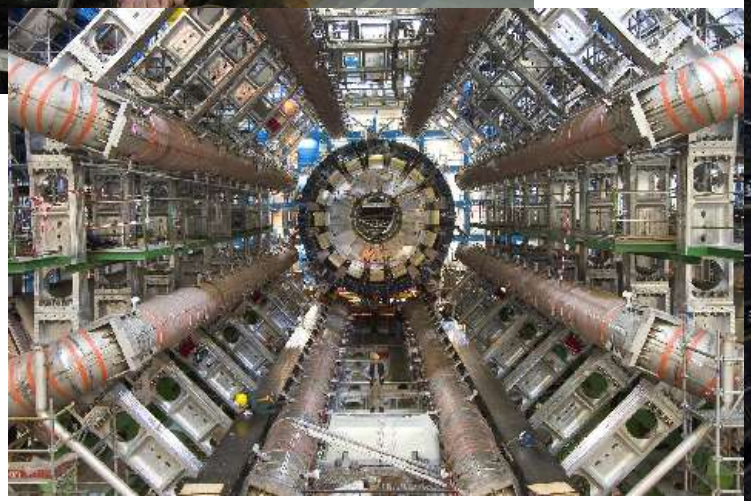
Galileo Galilei

Čo to vlastne robil?

CERN – LHC

Čo to vlastne robia?





Teda – čo vlastne robi-li/a?

- Naplánovali experiment
- Nazbierali dáta
- Vyhodnotili dáta (premenili ich na informácie)
- Urobili závery (získali znalosti)

Čo zvyčajne robíme dnes?

Zvyčajne ...

- Zbierame dáta (v nesmiernom rozsahu)
- Bez nejakého explicitného plánu

- Tieto dáta predstavujú informácie a obsahujú znalosti

Ešte „nedávno“

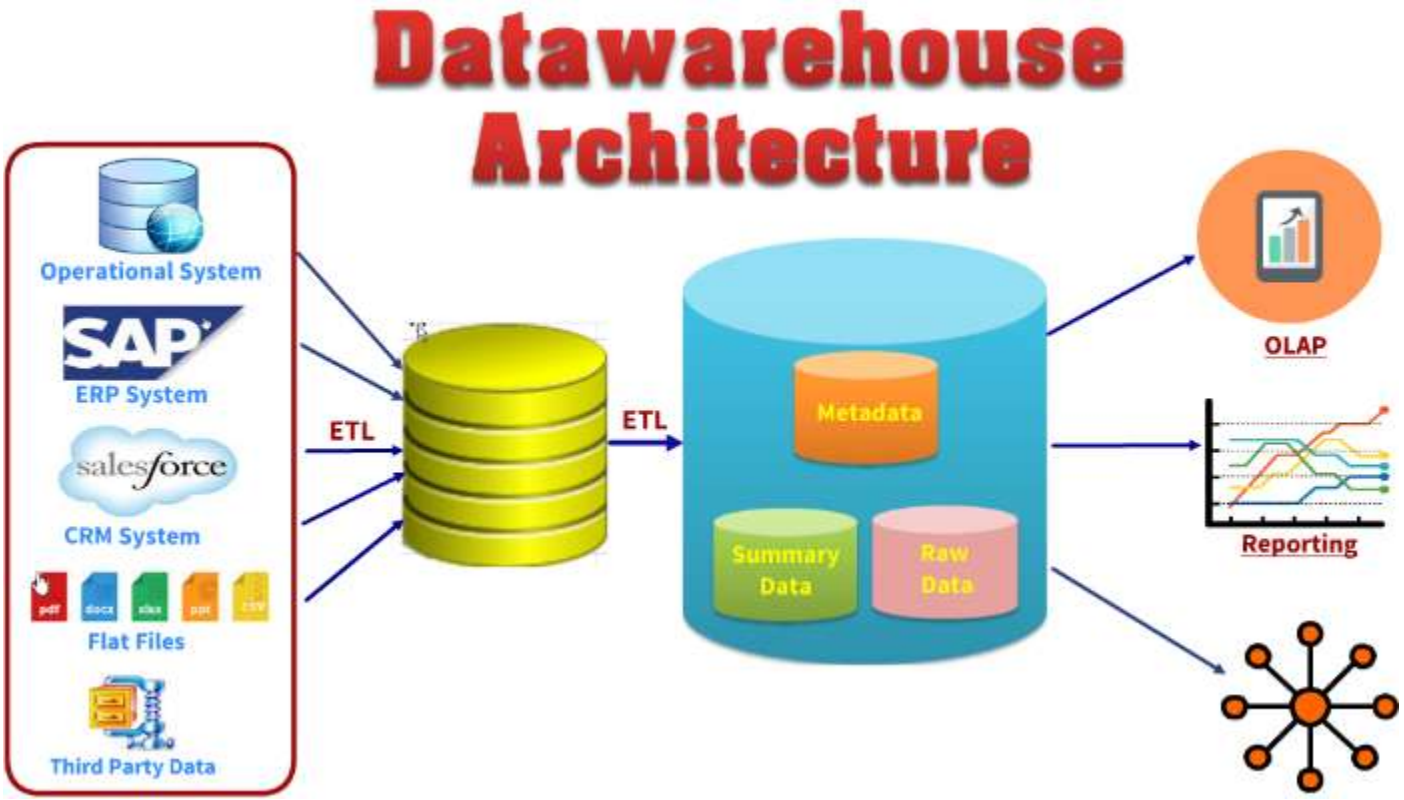
ERP (OLTP)


HUMAN RESOURCES
DEVELOPMENT
PURCHASING
ERP
DATA SERVICES
ENTERPRISE RESOURCE PLANNING
CRM
FINANCE ACCOUNTING
PROJECT MANAGEMENT

ERP (OLTP)


changes
information
management
solutions
needed
floor
plant
systems
citation
integrated
customization
business
planning
customer
process
ERP
data
equipment
integration
practices
costs
access
time
manufacturing
processes
database
Customization
migration
system
typically
software
customers
implementation

Data warehouse



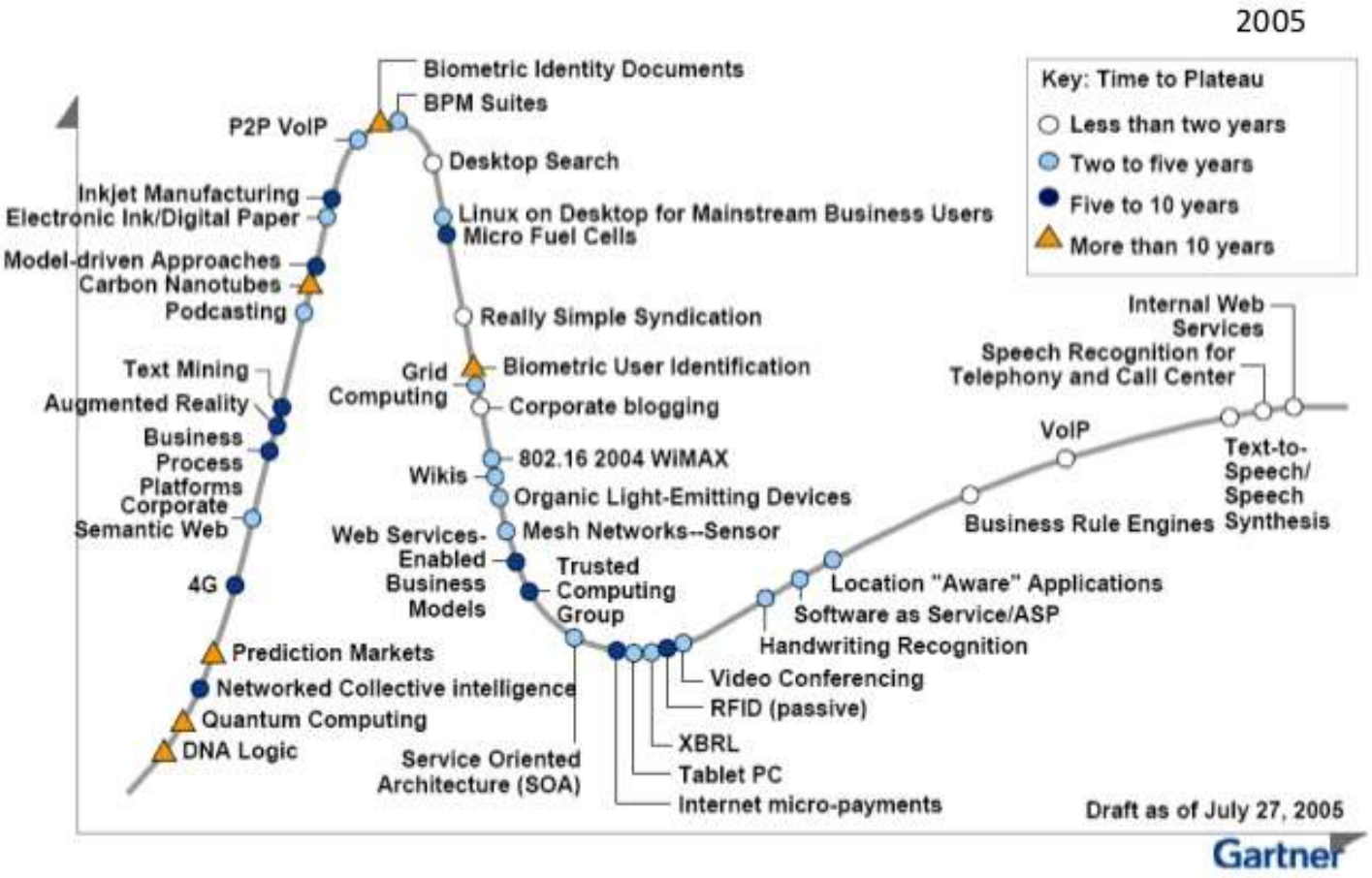

DATA
MINING
SET
PATTERNS
LEARNING
CUSTOMERS
ANALYSIS
MACHINE
BUSINESS
DECISION
PEOPLE
CUSTOMER
METHODS
INTELLIGENCE
DATABASE
INFORMATION
OFFERS
IDENTIFY
VISUALIZATION
E-MAIL
FINALLY
MARKETING
SOPHISTICATED
ALGORITHMS
DETECTION
OFFER
ARTIFICIAL
STRUCTURES
RULE
CITATION
NEW
SEND
WITHOUT
MIGHT
USE
STEP
GROUPS
TASK
MANAGEMENT
ALGORITHM
RECORDS
MODELS
KNOWN
WITHIN
FINAL
POTENTIAL
LEARNED
EXAMPLE
PROCESS
ASSOCIATION
NEEDED
MODELING
KNOWLEDGE
LARGE
TEST
TERM
MARKET
E-MAILS
BASKET
APPLICATIONS
PREDICT
LIKELY
MODEL
DETERMINE
TARGET
PRODUCT
PRESENT
USING
TRAINED
DISCOVERY
SYSTEM
MANY
PRE-PROCESSING
COMMON
HELP
INVOLVES
NUMBER

Business Intelligence

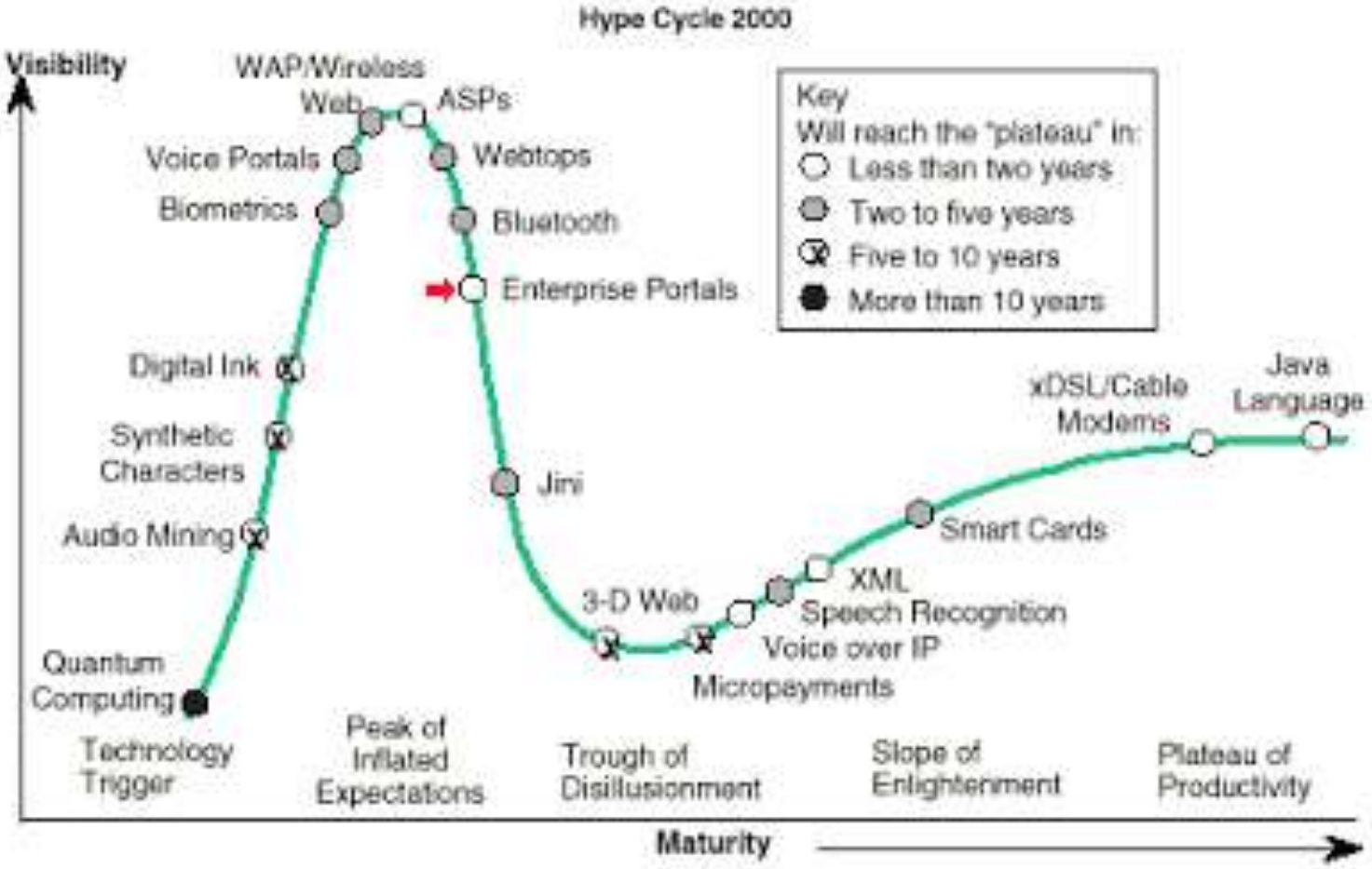
Gartner Hype Cycle
(kde je ERP, DM, BI?)

Source: Gartner (July 2016)

Gartner Hype Cycle 2005

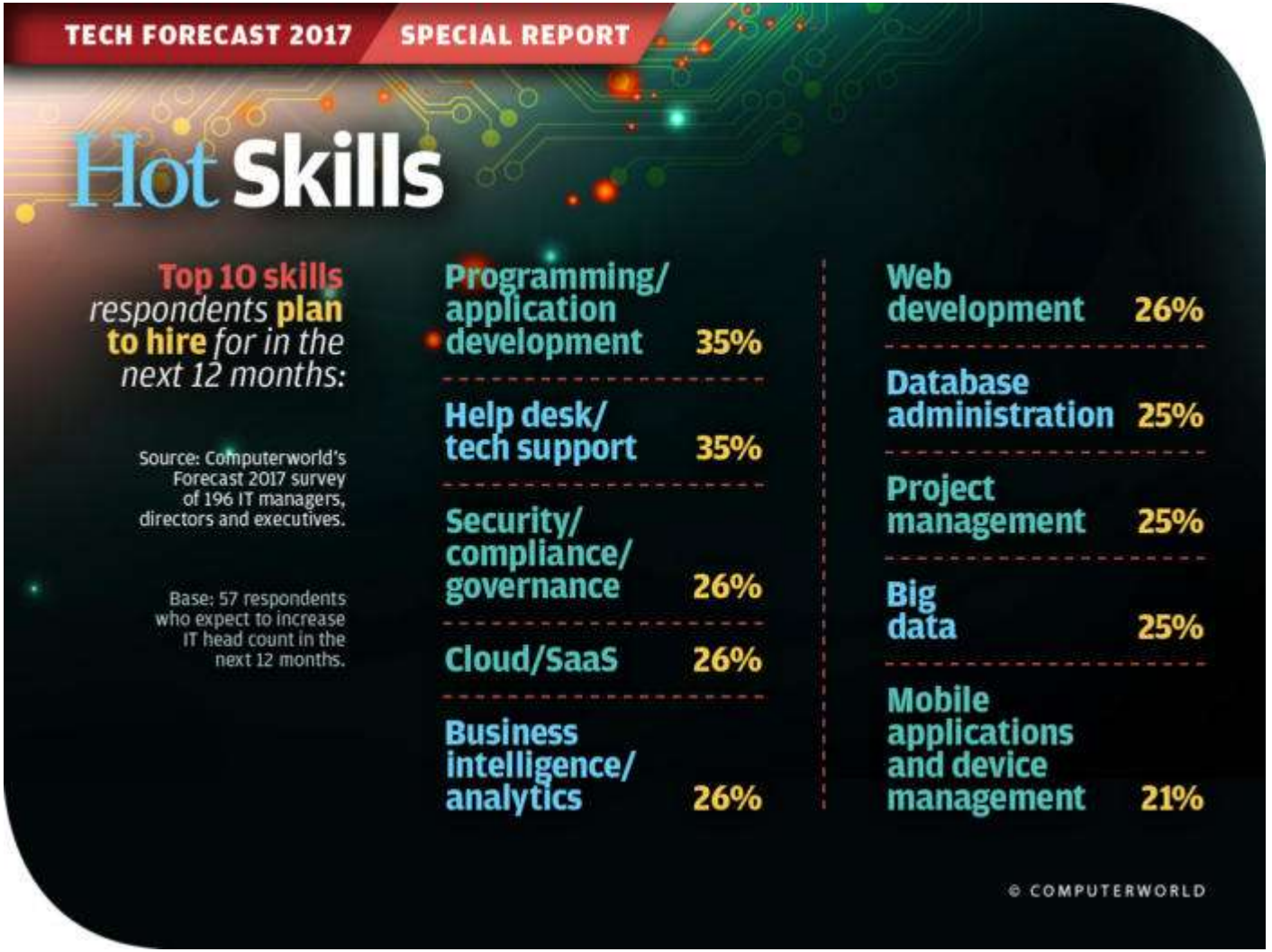


Gartner Hype Cycle 2000



Salary comparision

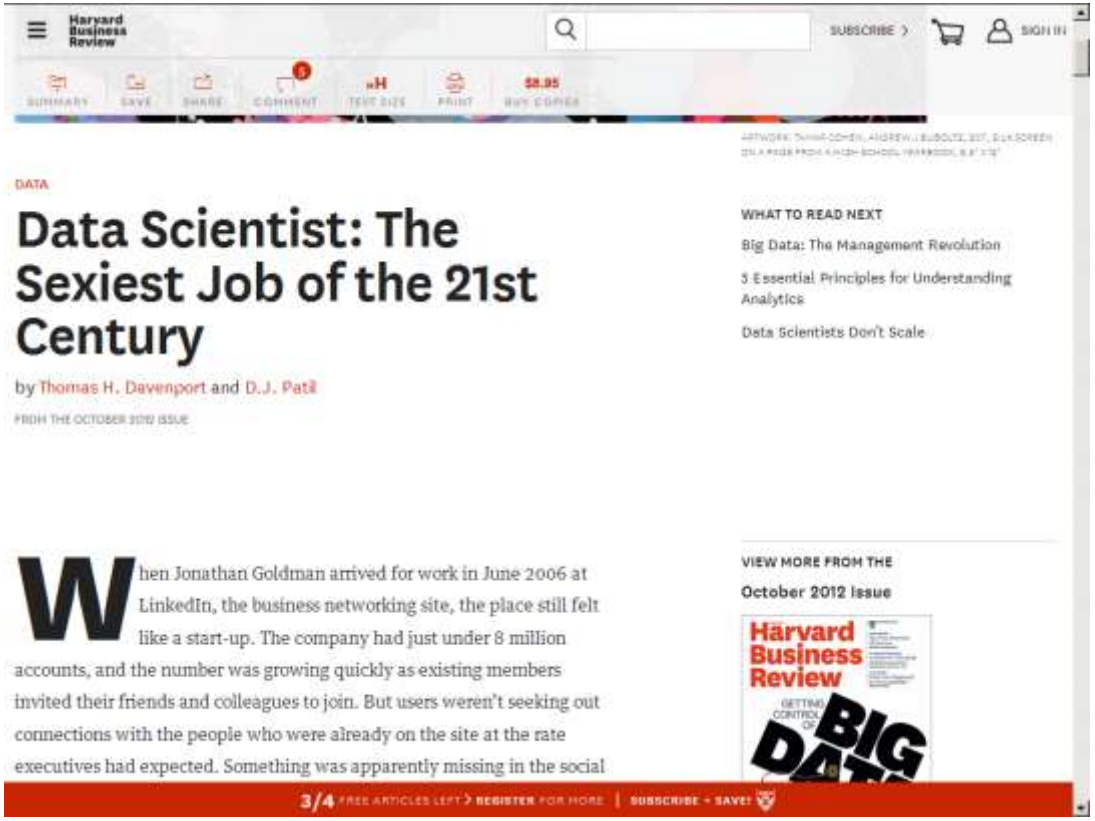
- http://www.computerworld.com/article/3169664/it-careers/13-tech-jobs-that-pay-200k-salaries.html
- http://www.computerworld.com/salarysurvey/breakdown/2016/joblevel/3



Zdroj: http://www.computerworld.com/article/3147427/it-skills-training/10-hottest-tech-skills-for-2017.html

Harward Business Review

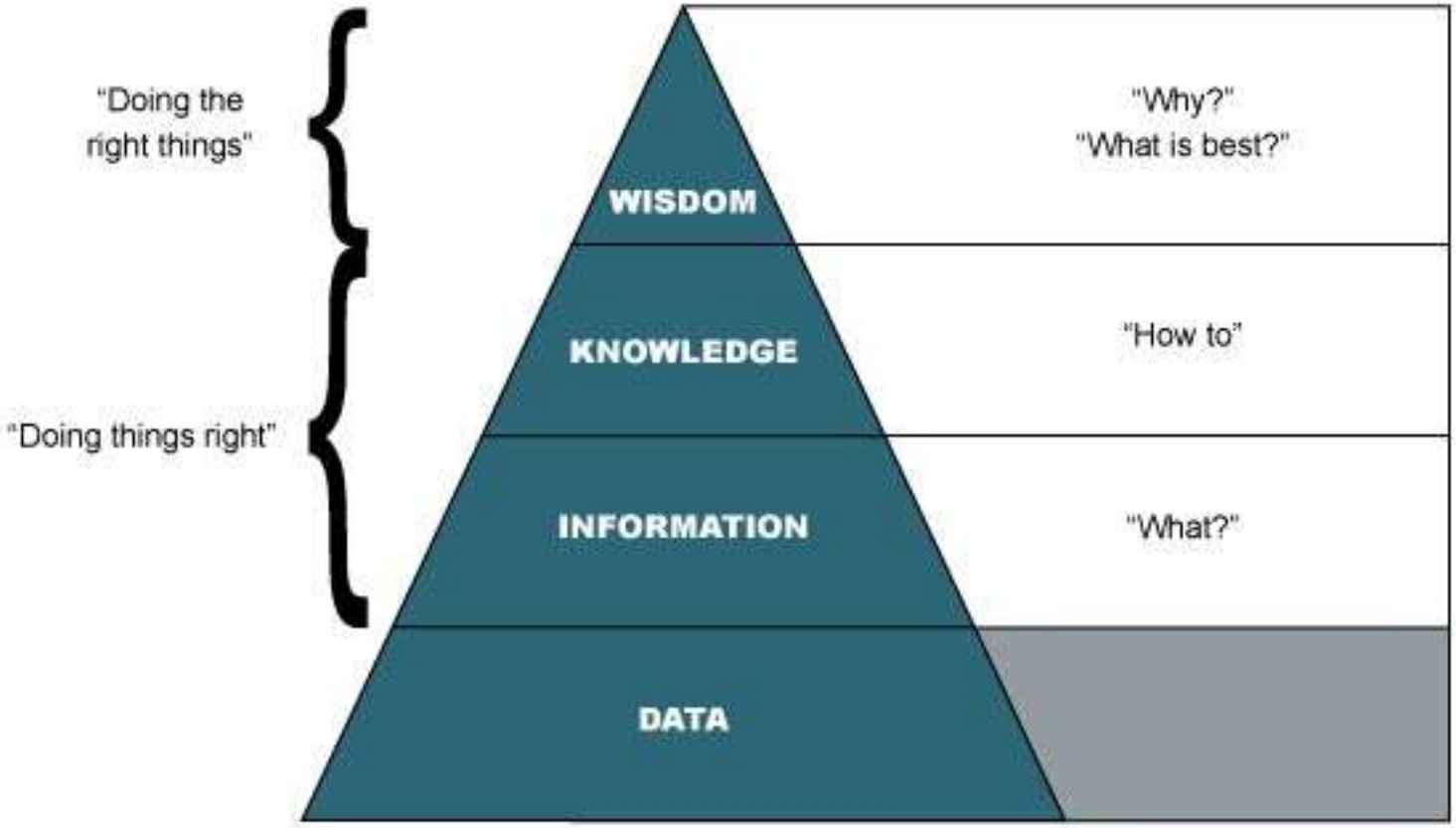
- **Sexiest Job of the 21st Century?**



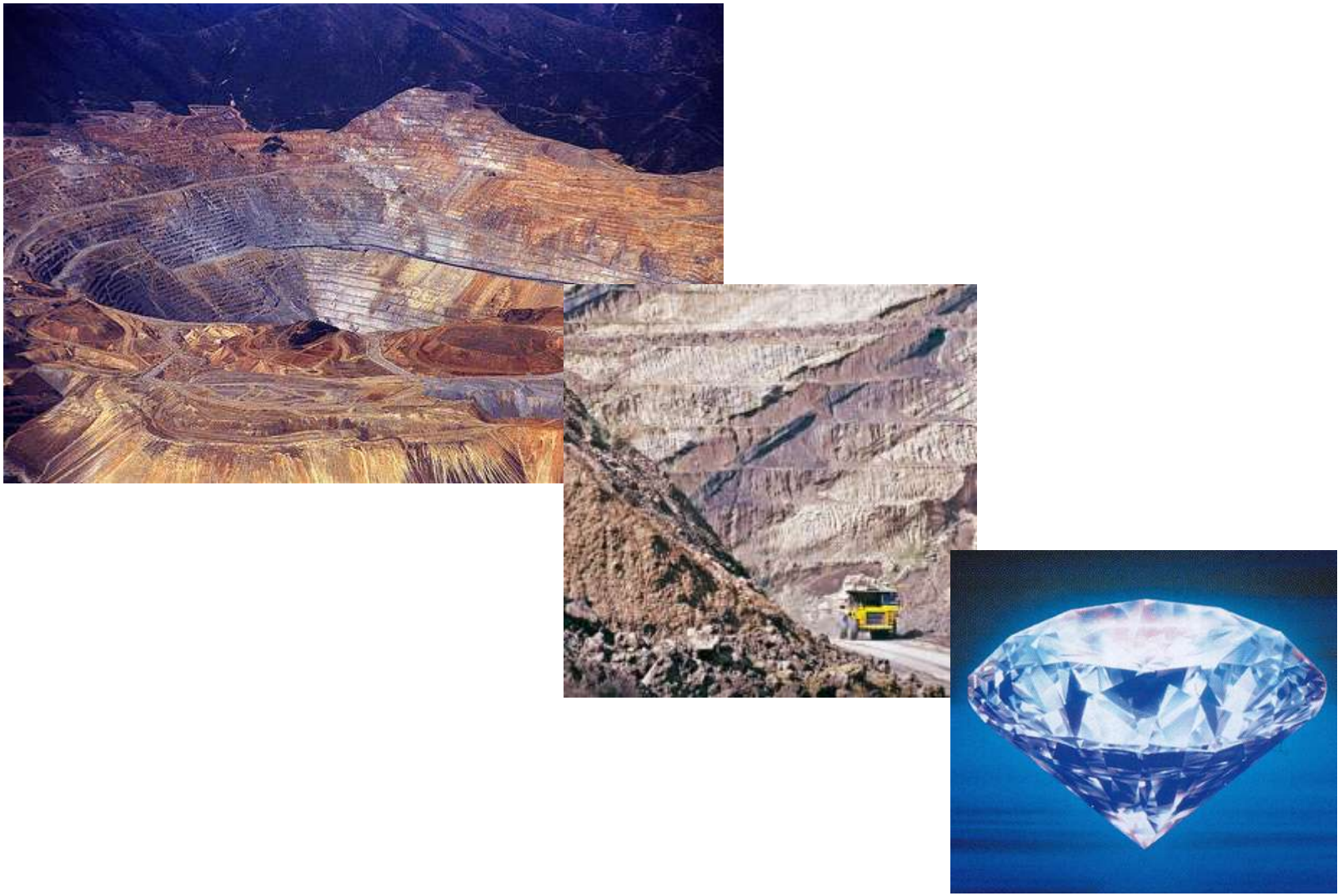
Wikipedia

- **Data Science** is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).

Podstata


"Doing the right things"
"Doing things right"
WISDOM
KNOWLEDGE
INFORMATION
DATA
"Why?"
"What is best?"
"How to"
"What?"

Podstata (analógia)



Audience Data Mining

Wikipedia (pokr.)

- Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression, computer programming, artificial intelligence, and high performance computing.


Data Science
Big Data (analytics)
Data Visualization
Data Engineering

Big Data

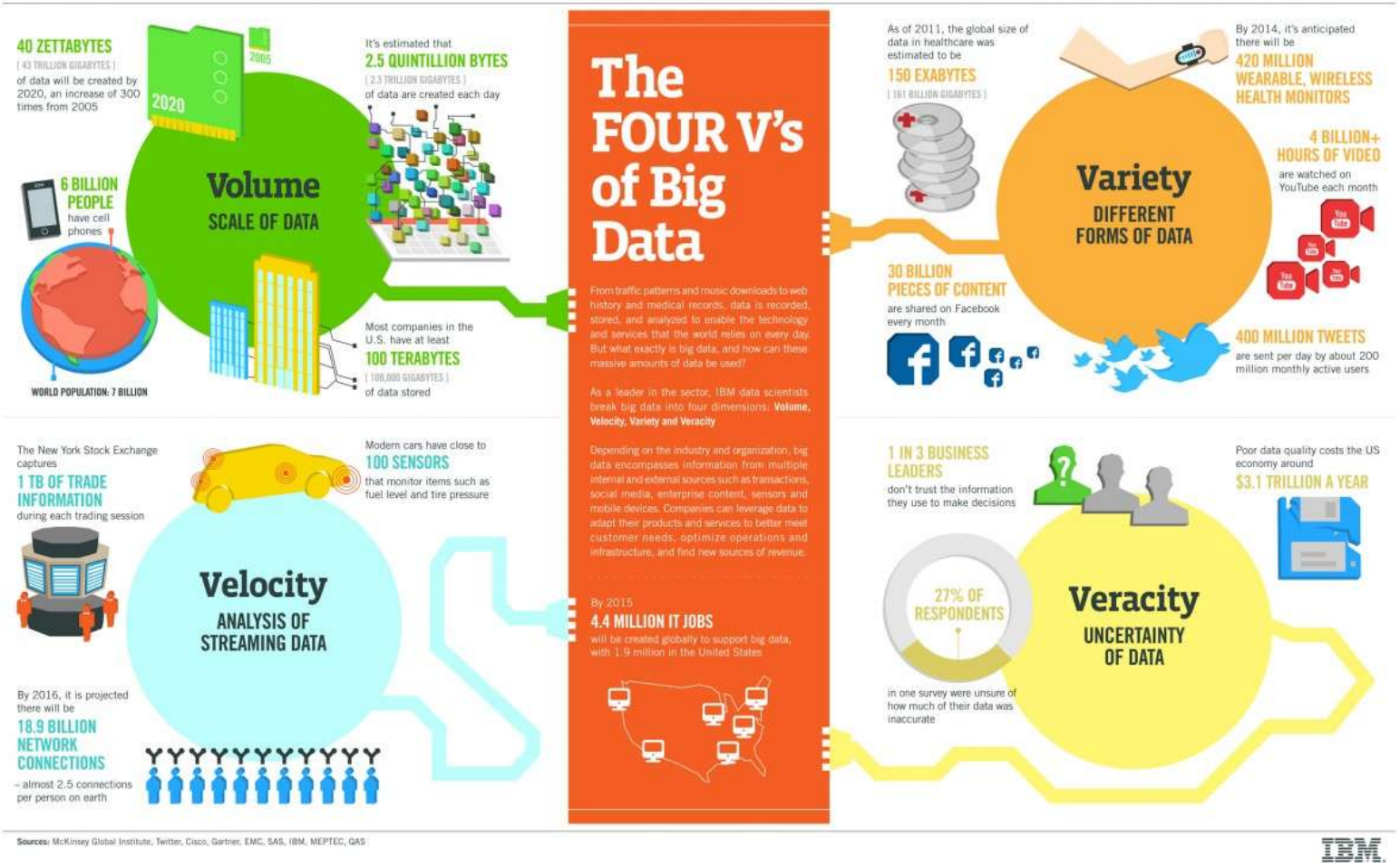
3V = **Volume**, **Velocity** (stream)

and **Variety** (structured, semi-structured and unstrucutred)

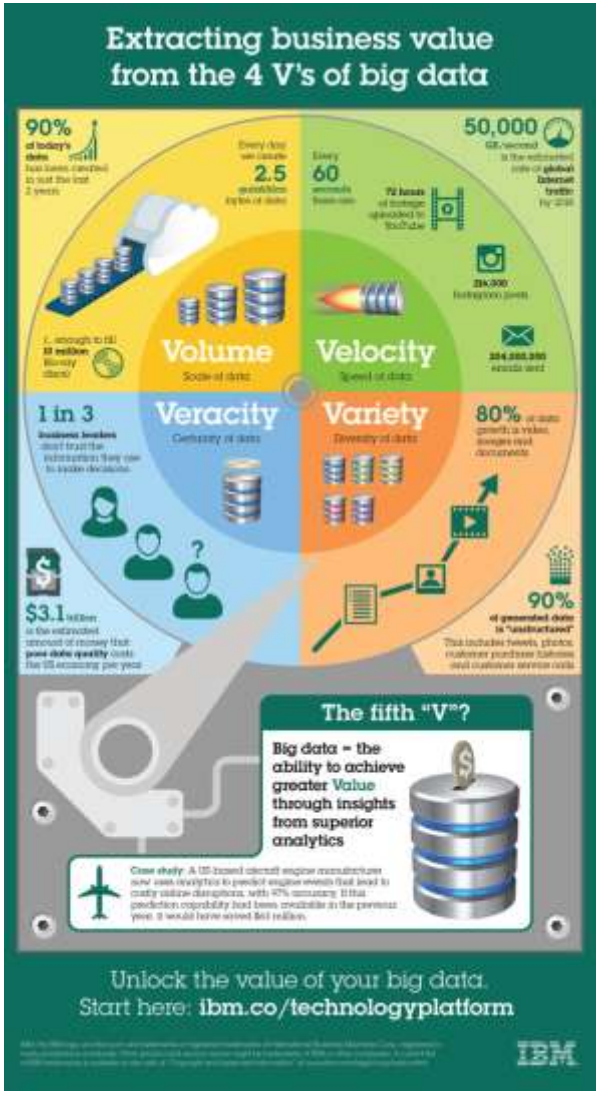
4V = 3V + **Veracity** (IBM)

5V = 4V + **Value** (ORACLE?)

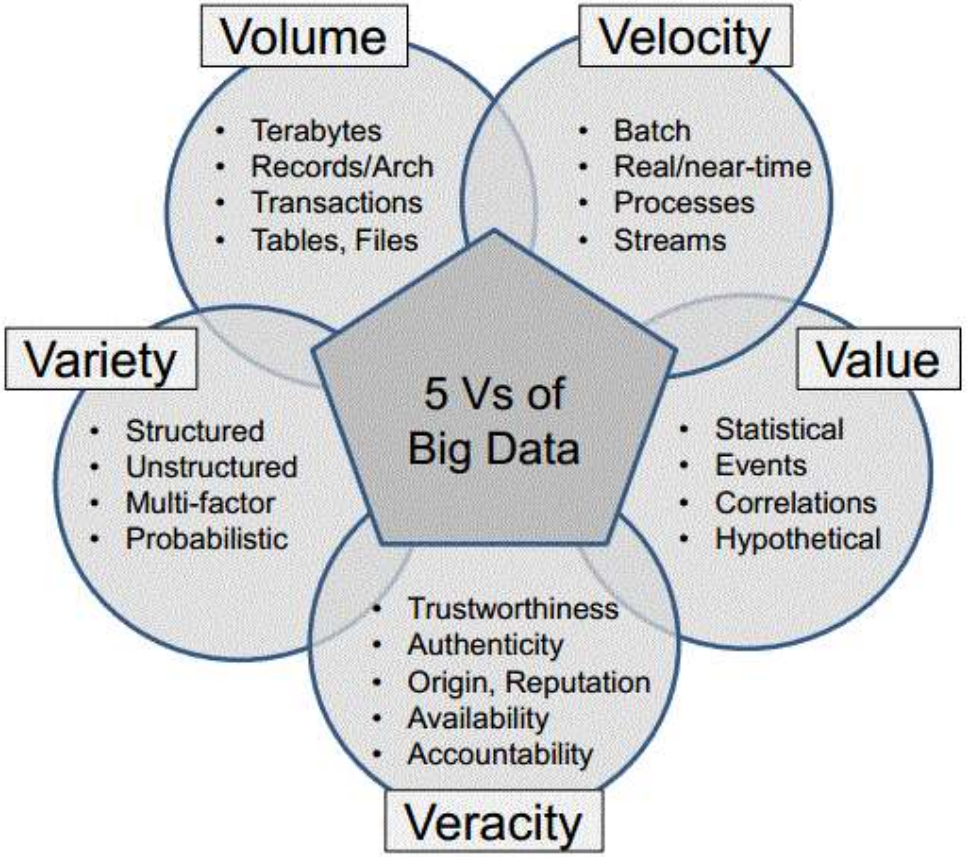
Big Data (IBM's 4V)



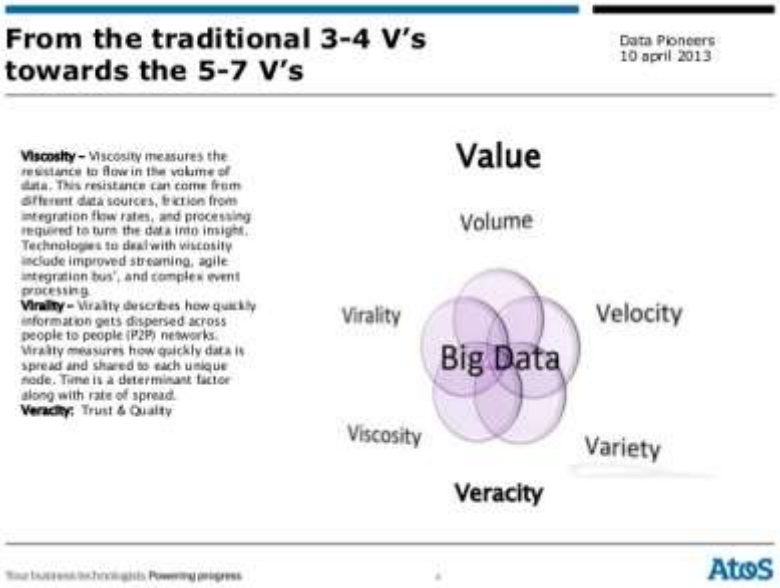
Big Data (IBM's 4V)

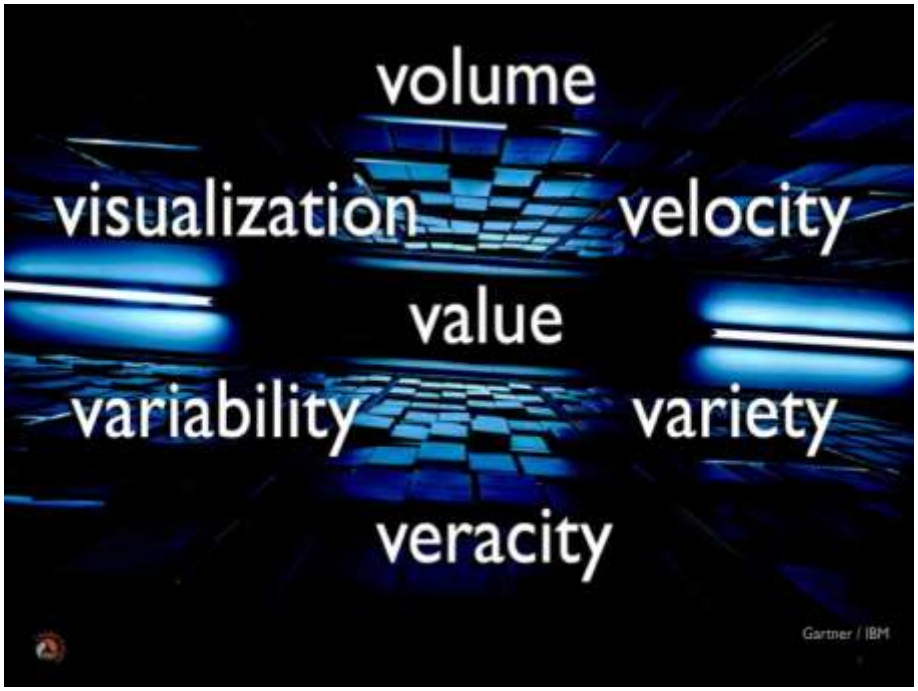


Big Data 5V


Volume
Terabytes
Records/Arch
Transactions
Tables, Files
Velocity
Batch
Real/near-time
Processes
Streams
Variety
Structured
Unstructured
Multi-factor
Probabilistic
5 Vs of
Big Data
Value
Statistical
Events
Correlations
Hypothetical
Trustworthiness
Authenticity
Origin, Reputation
Availability
Accountability
Veracity

Big Data 7V


From the traditional 3-4 V's towards the 5-7 V's
Data Pioneers
10 april 2013
Viscosity – Viscosity measures the resistance to flow in the volume of data. This resistance can come from different data sources, friction from integration flow rates, and processing required to turn the data into insight. Technologies to deal with viscosity include improved streaming, agile integration bus', and complex event processing.
Virality – Virality describes how quickly information gets dispersed across people to people (P2P) networks. Virality measures how quickly data is spread and shared to each unique node. Time is a determinant factor along with rate of spread.
Veracity: Trust & Quality
Value
Volume
Virality
Velocity
Big Data
Viscosity
Variety
Veracity
Atos


volume
visualization
velocity
value
variability
variety
veracity
Gartner / IBM

Big Data


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

NoSQL a NewSQL

- NoSQL (nie ~~**No** SQL~~ ale **Not only** SQL !!!)
  - Key value
  - Document Store
  - Column store
  - Graph

- NewSQL
  - class of modern relational database management systems that seek to provide the same scalable performance of NoSQL systems for online transaction processing (OLTP) read-write workloads while still maintaining the ACID guarantees of a traditional database system (https://en.wikipedia.org/wiki/NewSQL)

NoSQL CAP theorem

It is impossible for a distributed computer system to simultaneously provide more than two out of three of the following guarantees:

- **Consistency** - every read receives the most recent write or an error
- **Availability** - every request receives a (non-error) response – without guarantee that it contains the most recent write
- **Partition tolerance** - the system continues to operate despite an arbitrary number of messages being dropped (or delayed) by the network between nodes

Zdroj: https://en.wikipedia.org/wiki/CAP_theorem

Data Lakes

The method of storing data within a system or repository, in its **natural format**, that facilitates the collocation of data in various schemata and structural forms, usually object blobs or files.

Why natural format?

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

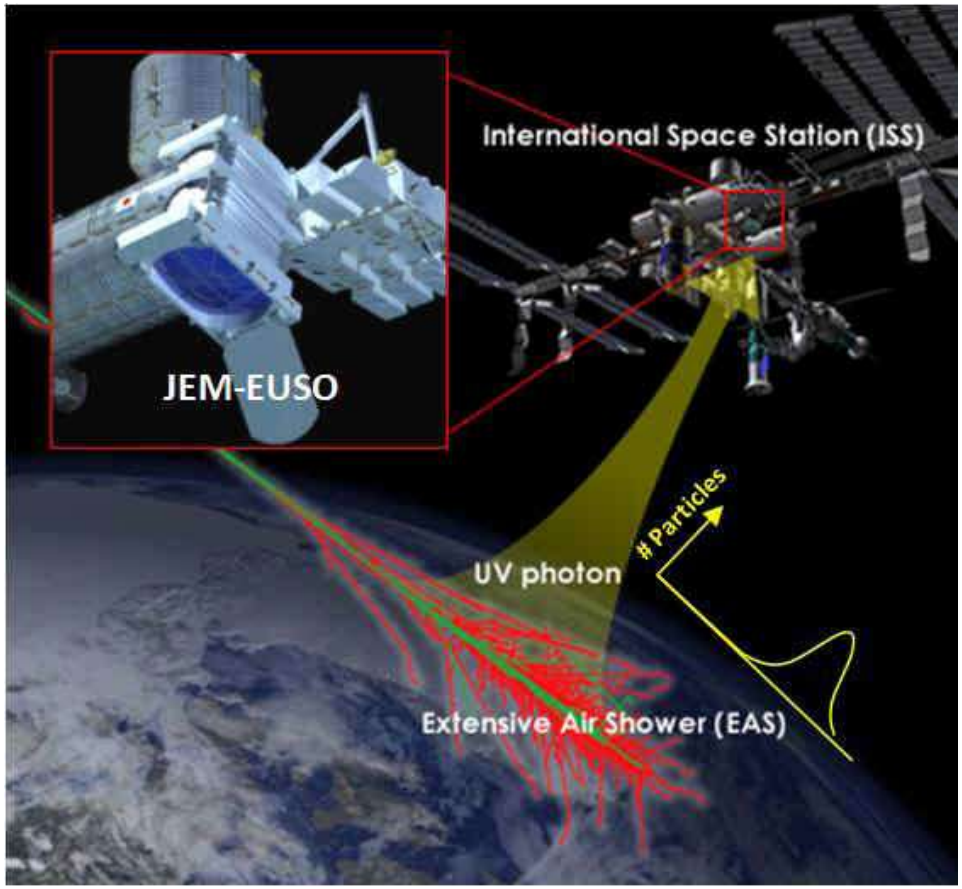
Príklady našej z praxe

Čo aktuálne robíme

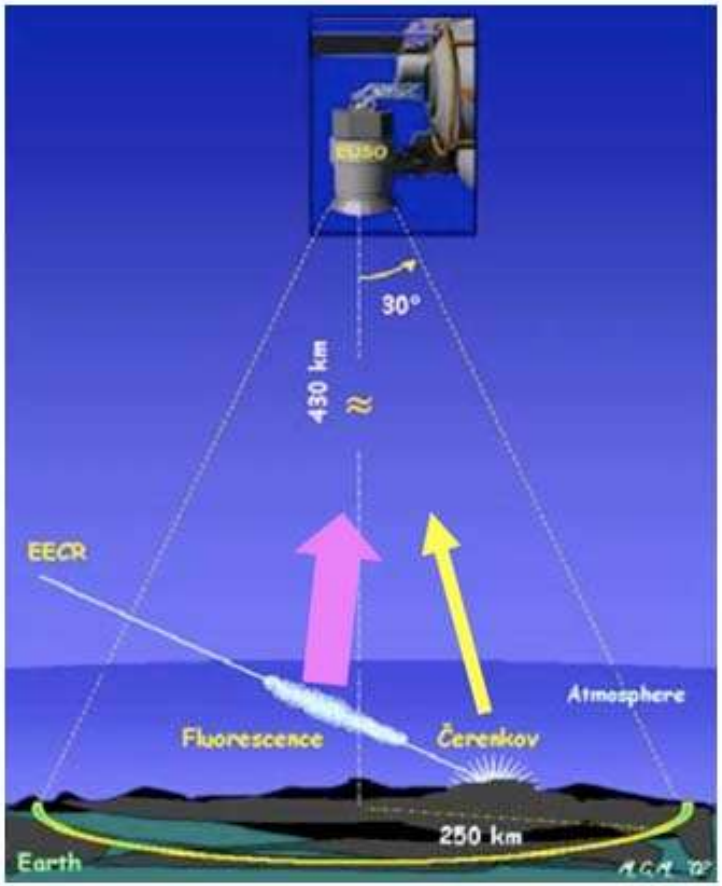
Kozmická fyzika

Stručne, prídu porozprávať zo SAV

JEM-EUSO


International Space Station (ISS)
JEM-EUSO
UV photon
Particles
Extensive Air Shower (EAS)

JEM-EUSO (pokr.)



Vplyv kozmického počasia na ...

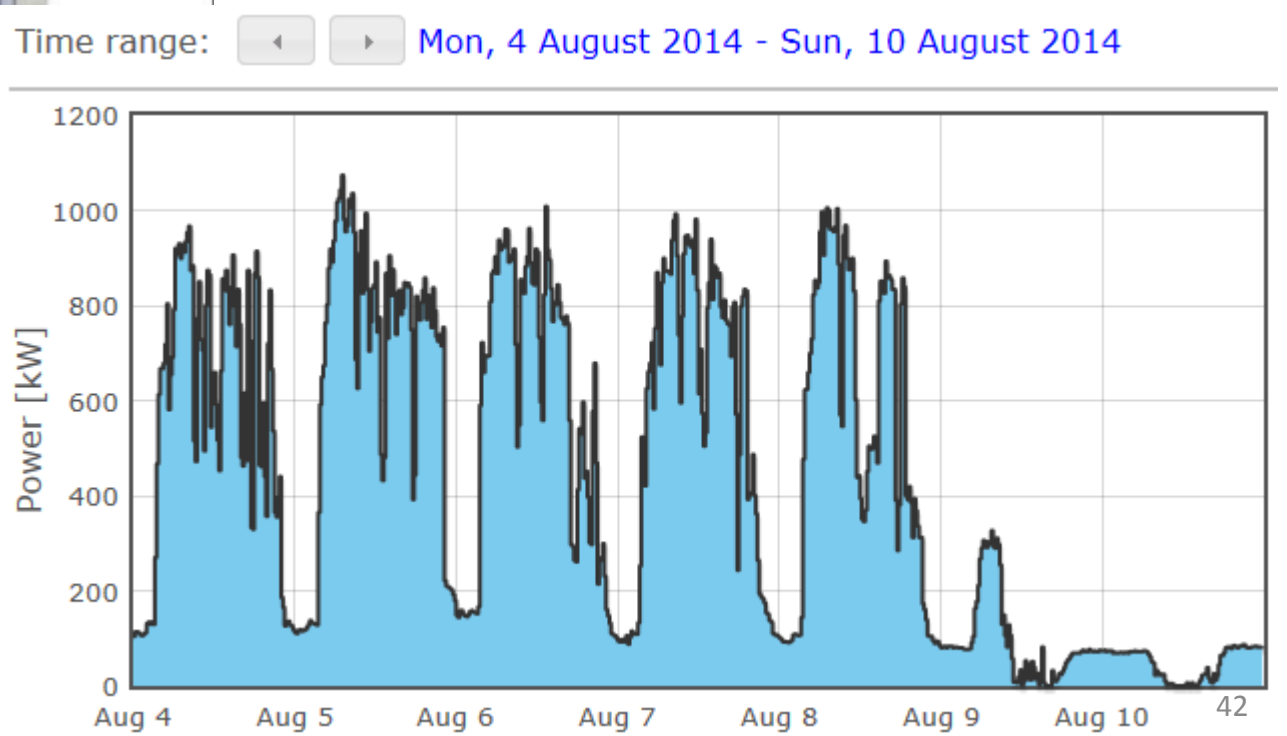
- sentiment obyvateľstva (globálny sentiment);
- výskyt diagnóz (infarkty, presnosť dát);
- výskyt udalostí
- ...

Inteligentné siete
Smart Grid (PowerEng+IT)

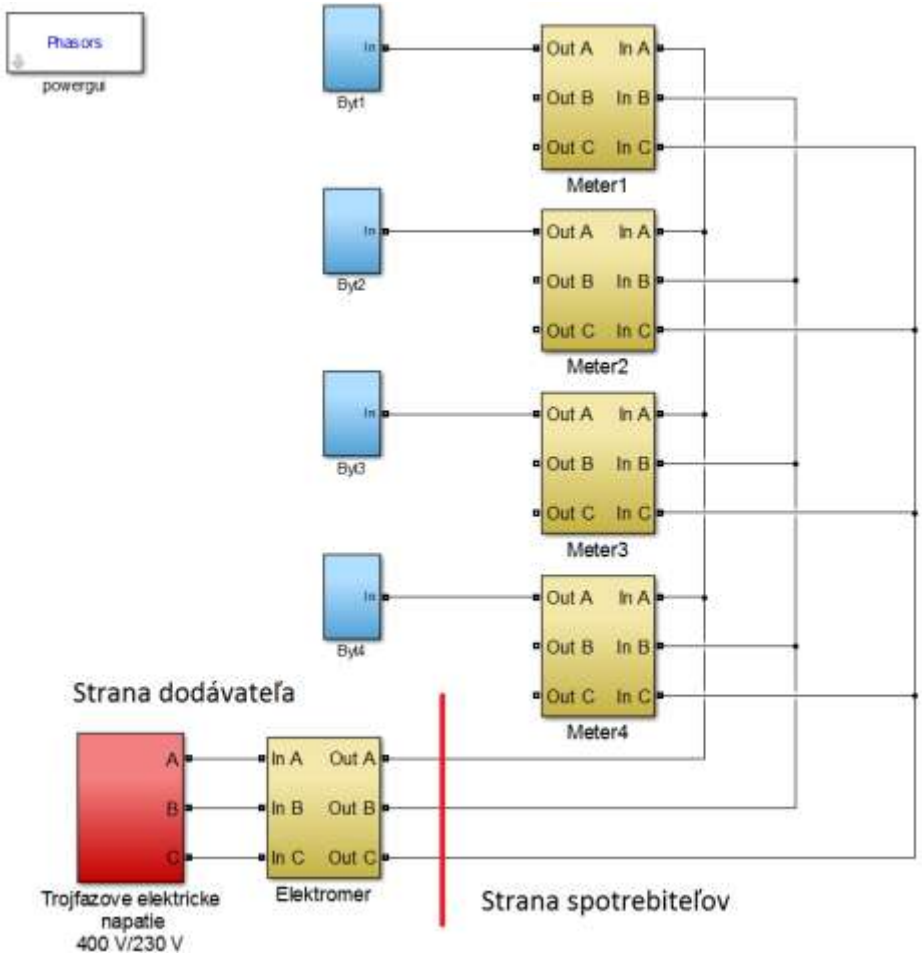
Čo sme zažili

Čo zažívame

Simulačné modely pre analýzu inteligentných sieti (Smart Grid)(1)

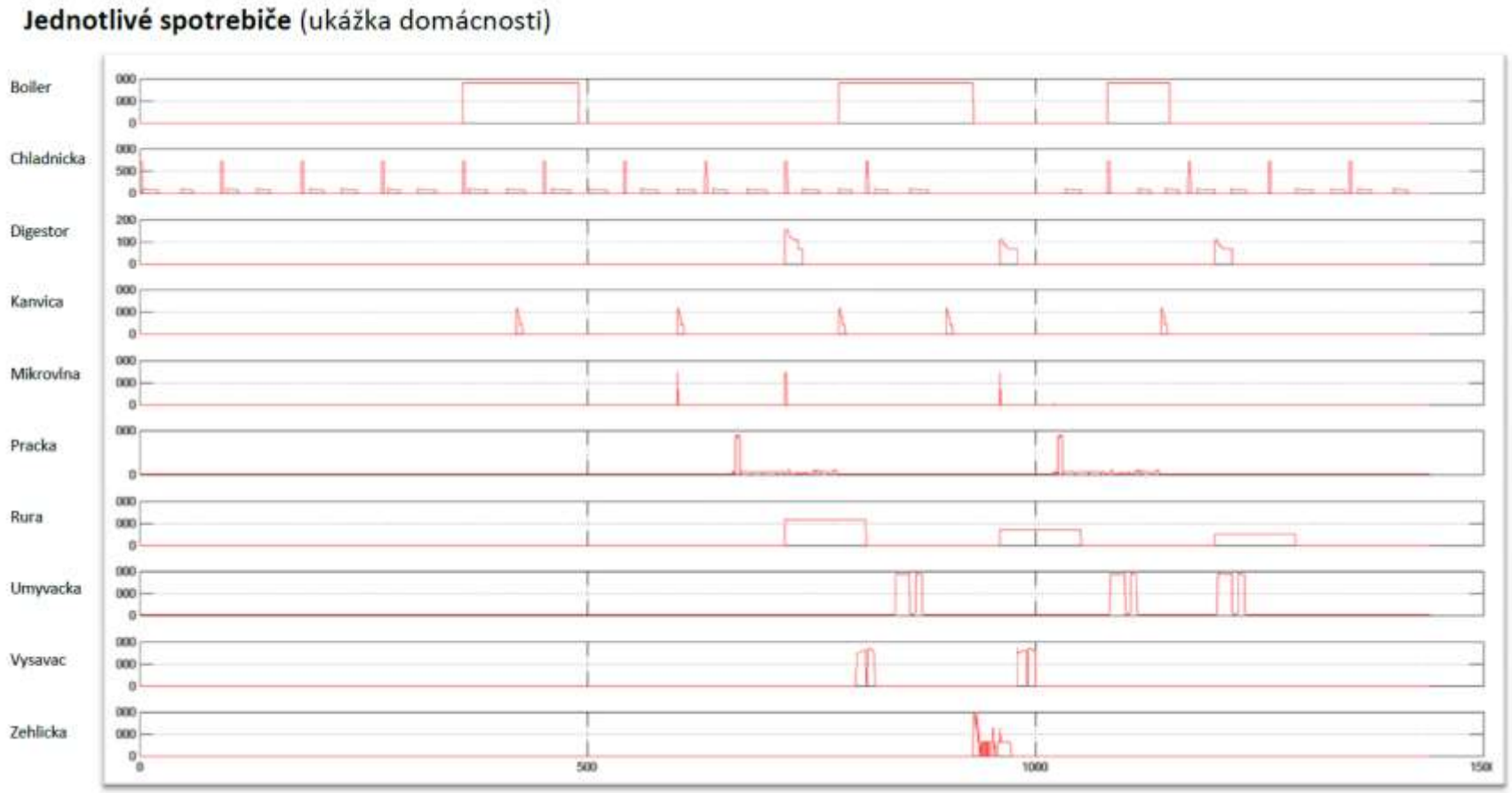


Simulačné modely pre analýzu inteligentných sieti (Smart Grid)(2)



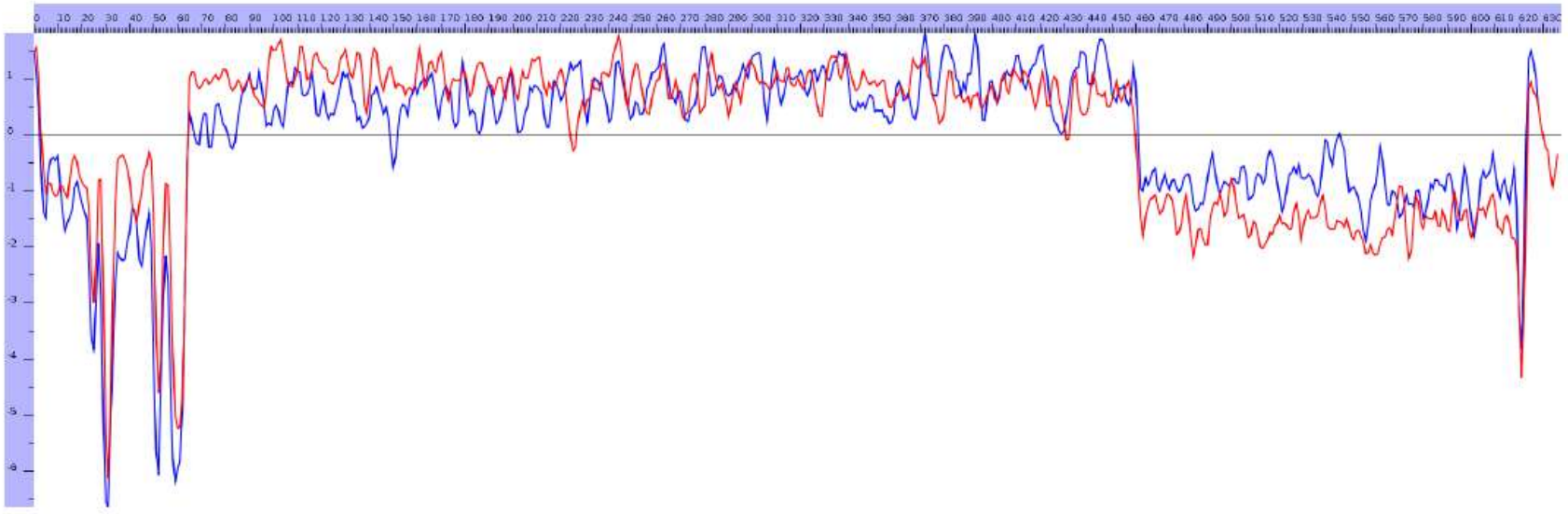
DP: Zsolt Zsóka: Simuláčné experimenty pre inteligentné siete. 2015

Simulačné modely pre analýzu inteligentných sieti (Smart Grid)(3)



DP: Zsolt Zsóka: Simuláčné experimenty pre inteligentné siete. 2015

Určenie filmu, podobnosti programov (1)



Obrázok 8 – 6 Dva profily spotreby pre rovnaké video s množstvom tmavých scén

DP: Tomáš Mak: Identifikácia prehrávaného filmu na základe profilu spotreby elektrickej energie televízneho prijímača. 2016

Určenie filmu, podobnosti programov (2)

		Philips 32PHH4100/88				
	Nahrávky	A	B	C	D	E
Tesla S 7090 TSP2	A	0,48	0,23	-0,01	-0,25	0,23
	B	0,12	0,53	0,00	-0,07	0,12
	C	-0,19	-0,09	0,37	0,29	-0,16
	D	-0,06	-0,07	0,11	0,44	0,02
	E	0,10	0,12	0,09	-0,17	0,55

		Samsung UE40D5003				
	Nahrávky	A	B	C	D	E
Tesla S 7090 TSP2	A	0,12	-0,20	0,01	-0,05	0,04
	B	0,00	0,15	0,06	-0,04	-0,07
	C	-0,07	0,26	0,00	0,05	-0,09
	D	-0,08	-0,03	0,00	0,07	0,02
	E	-0,04	-0,15	-0,12	-0,12	-0,02

DP: Tomáš Mak: Identifikácia prehrávaného filmu na základe profilu spotreby elektrickej energie televízneho prijímača. 2016

Ďalšie úlohy

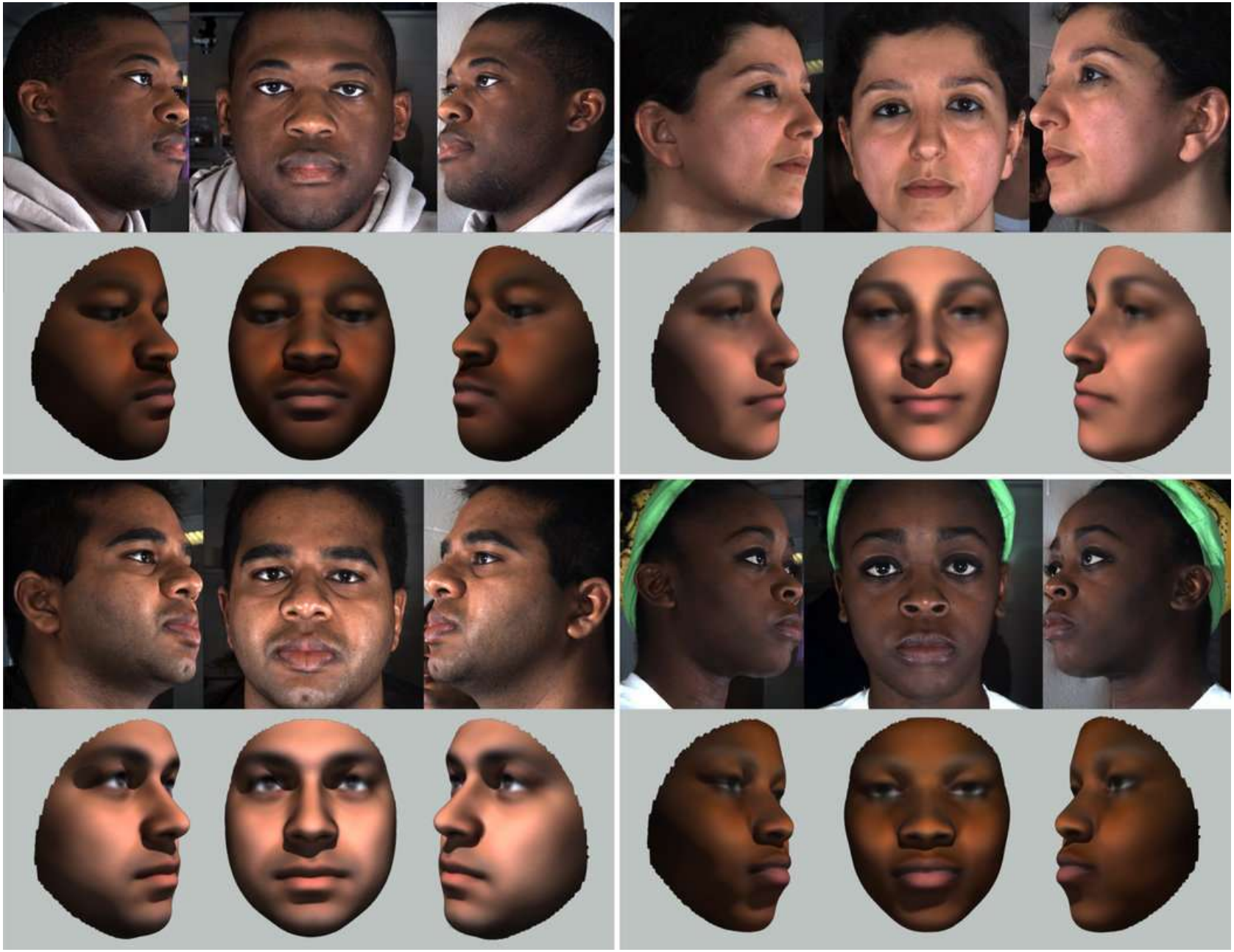
Security

Bioinformatika

Genomika

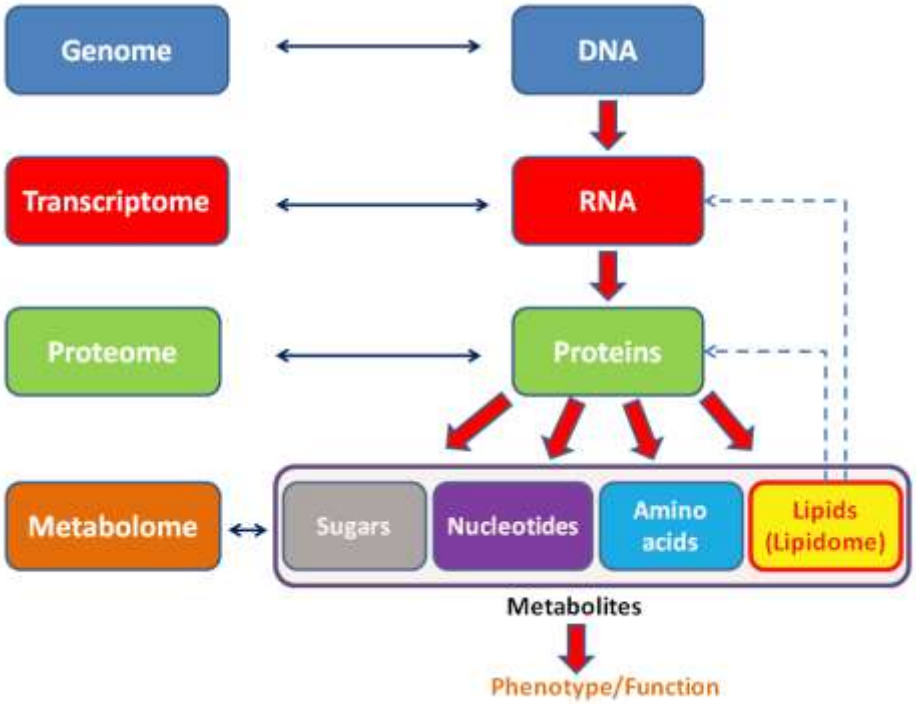
Čo napr. (už?) vieme získať z genómu?

- **Modeling 3D Facial Shape from DNA**
- https://www.newscientist.com/article/mg22129613-600-genetic-mugshot-recreates-faces-from-nothing-but-dna/

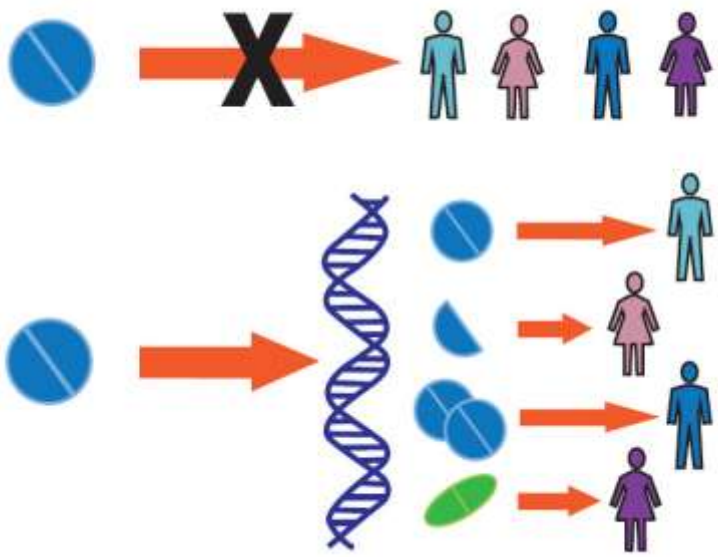


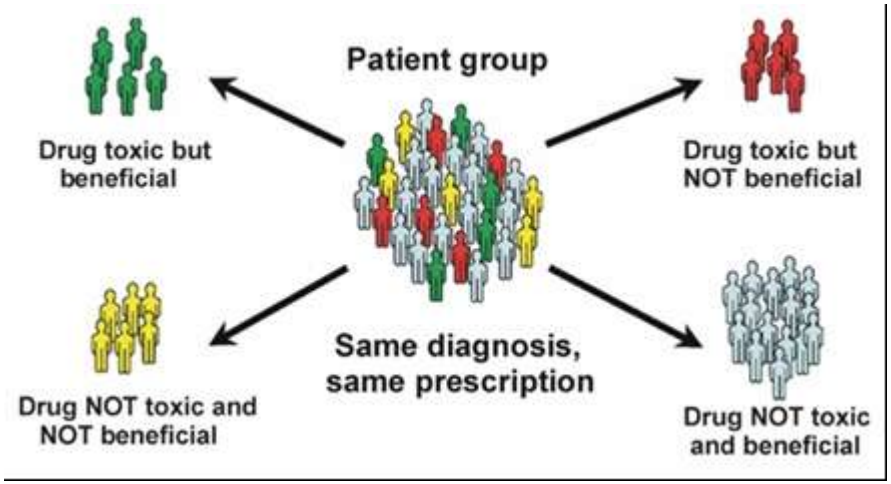
-omics

- https://en.wikipedia.org/wiki/Omics
  - Genomics
  - Transcriptomics
  - Proteomics
  - Lipidomics
  - Metalobomics
  - …



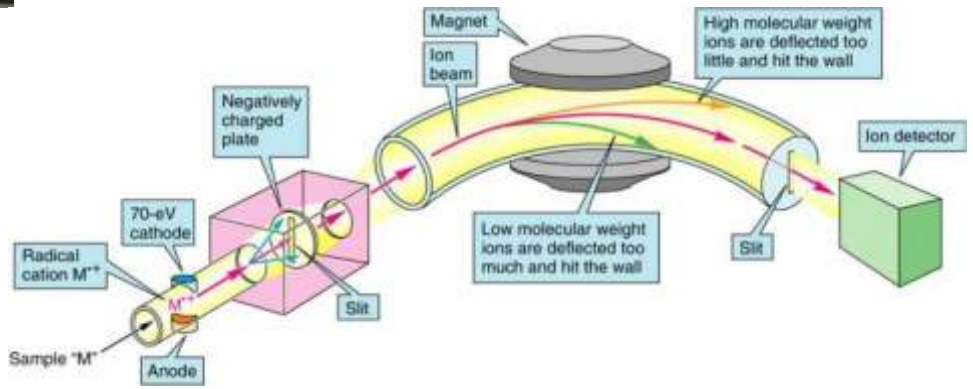
Načo je to dobré?



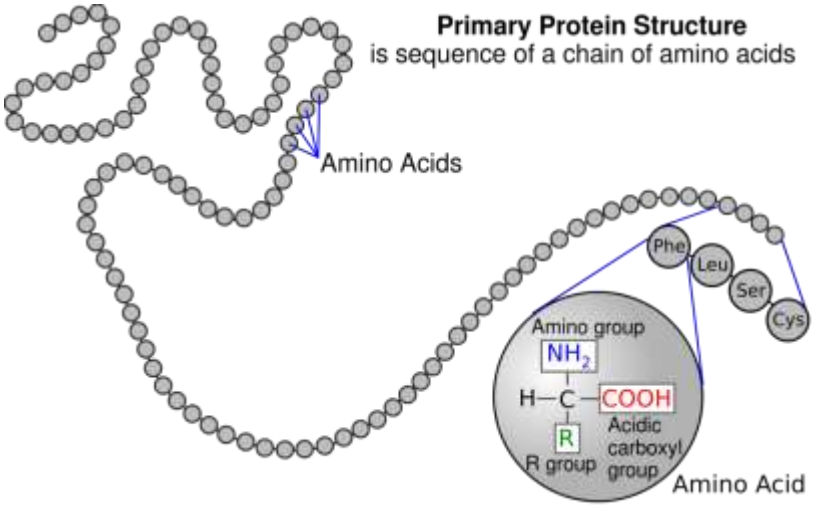


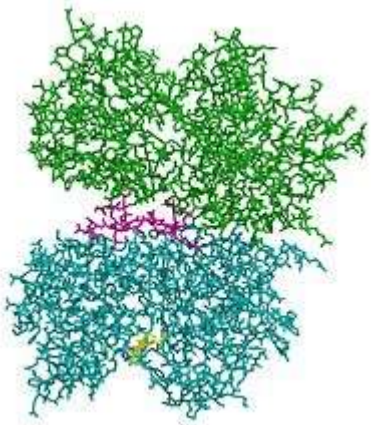
Personalizovaná medicína
Data mining v proteomike

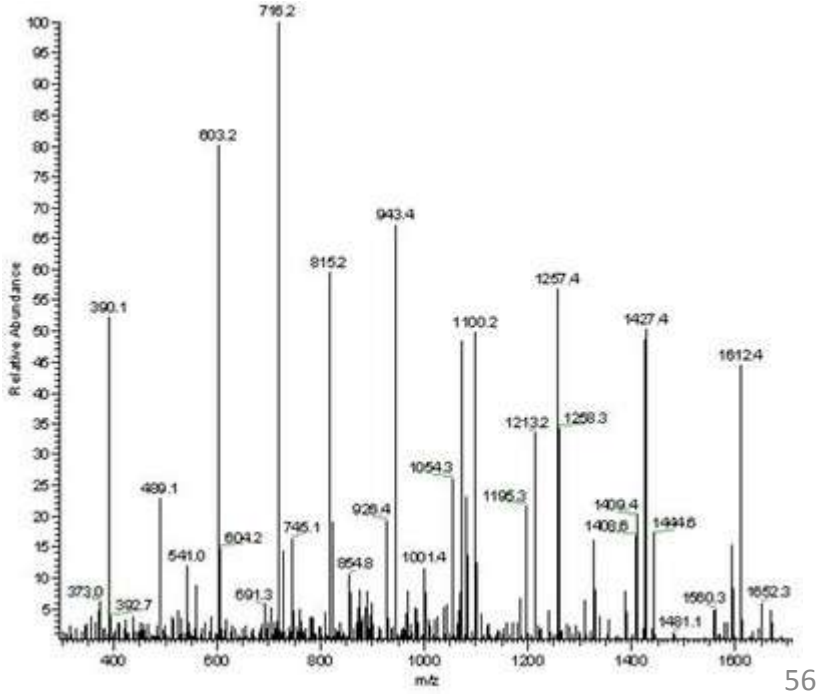
Hmotnostná spektrometria



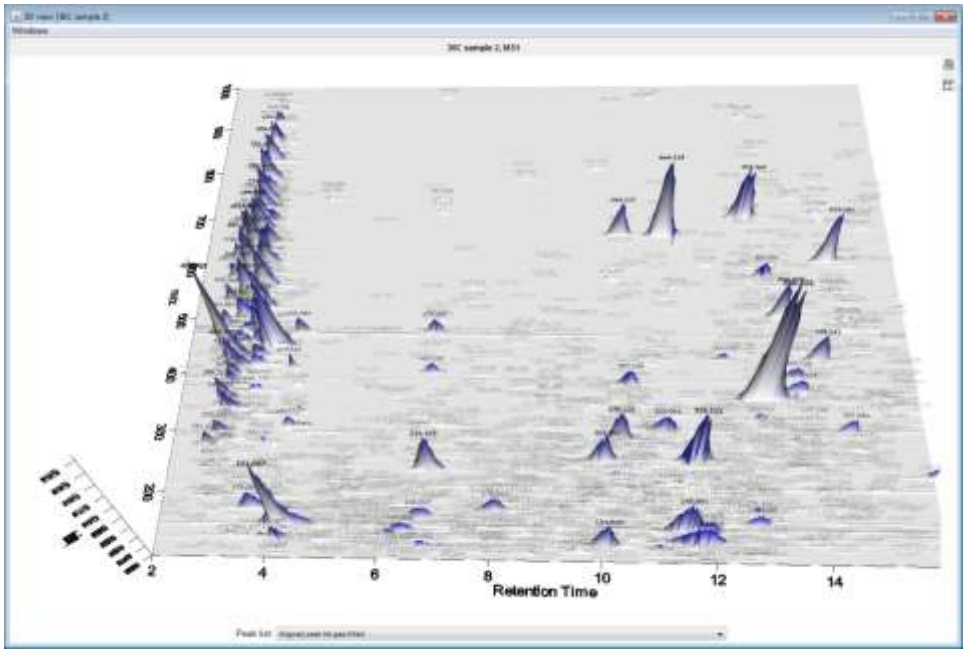
Proteíny a MS







„Biomarkery“

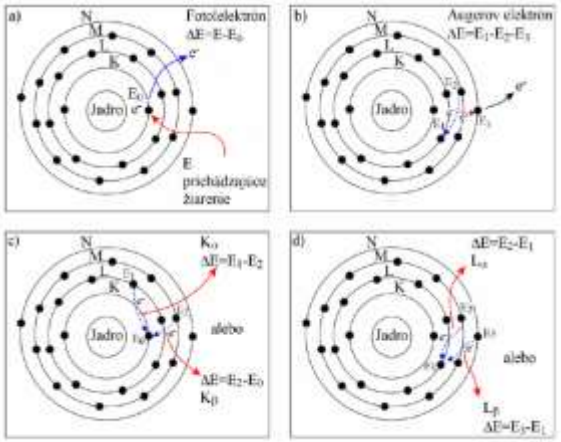


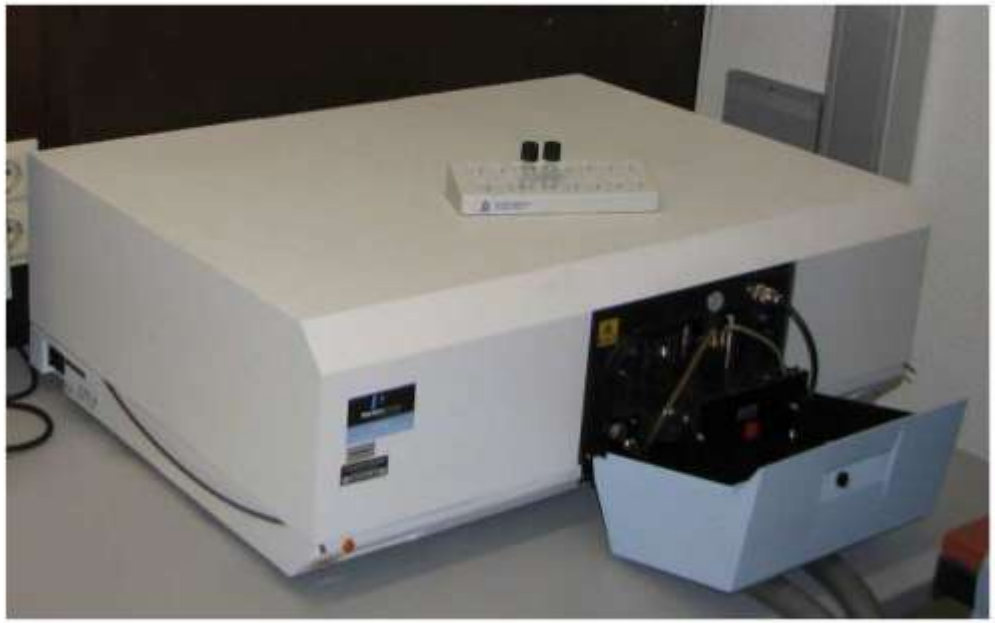
Reálny život

- Podpora spracovania dát – data engineering

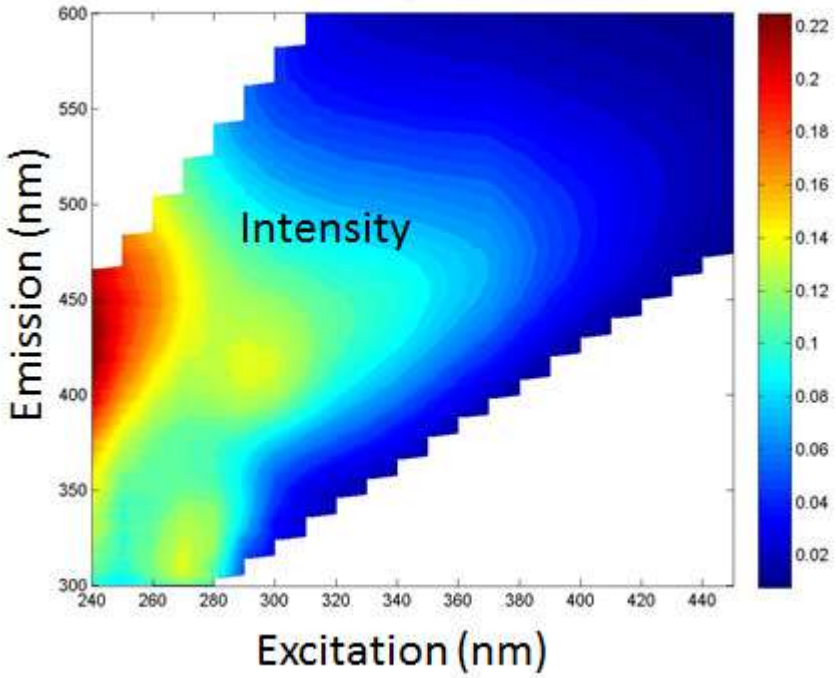
Spracovanie fluorescenčných spektrálnych matríc biologického materiálu

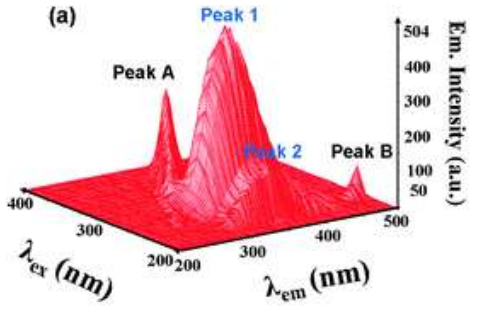
Fluorescencia

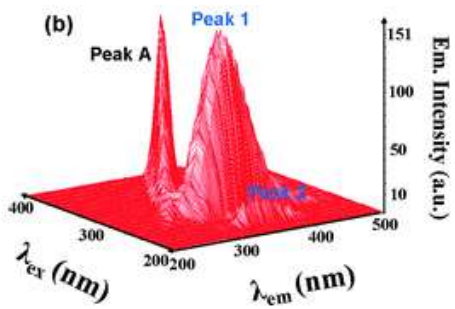


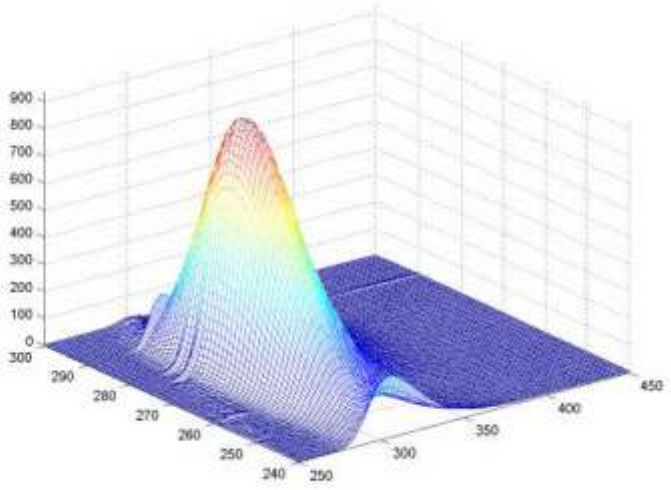


Spektrá - biomarkery









Počítačová lingvistika
Computational linguistics

Vymedzenie pôsobnosti

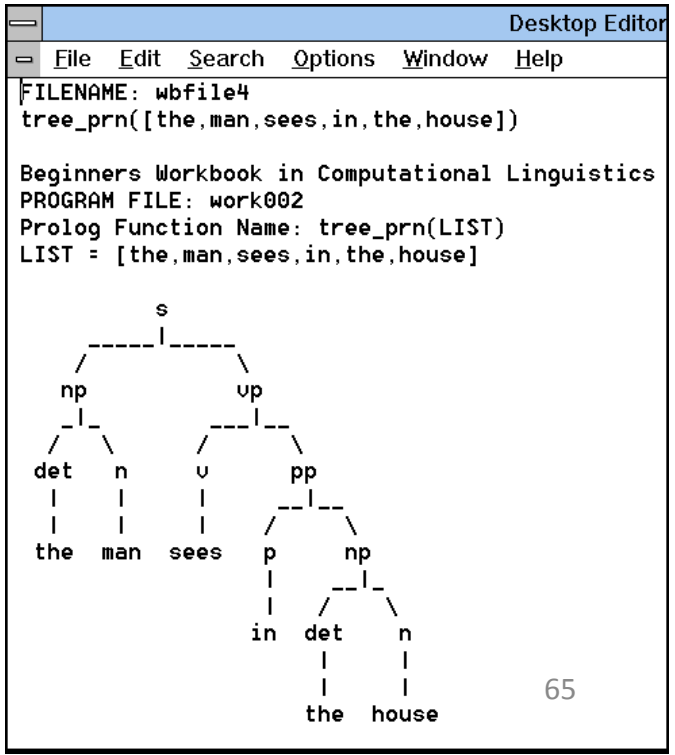
Aplikácia informačných technológií na spracovanie prirodzeného jazyka

Úlohy

- strojový preklad (asi top doména)
- dolovanie informácií (text mining)
- korektory pravopisu, gramatiky (len prípady)
- budovanie lingvistických zdrojov pre SK (morfologická DB, retrográdny, MWE, WordNetSK)
- Určenie autorstva textu
- …

Úlohy 2

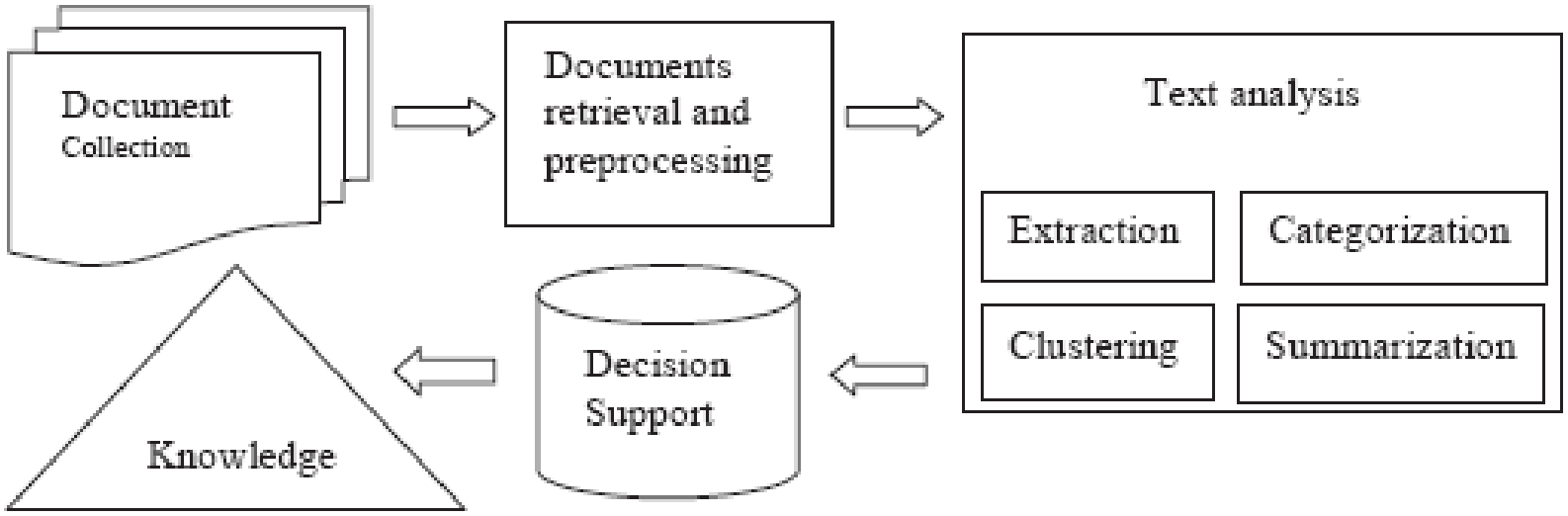
- Lexikálna analýza
  - Lematizácia
  - Steming
  - Identifikácia entít (napr. názvy)
- Syntaktická analýza
- Sémantická analýza
- …



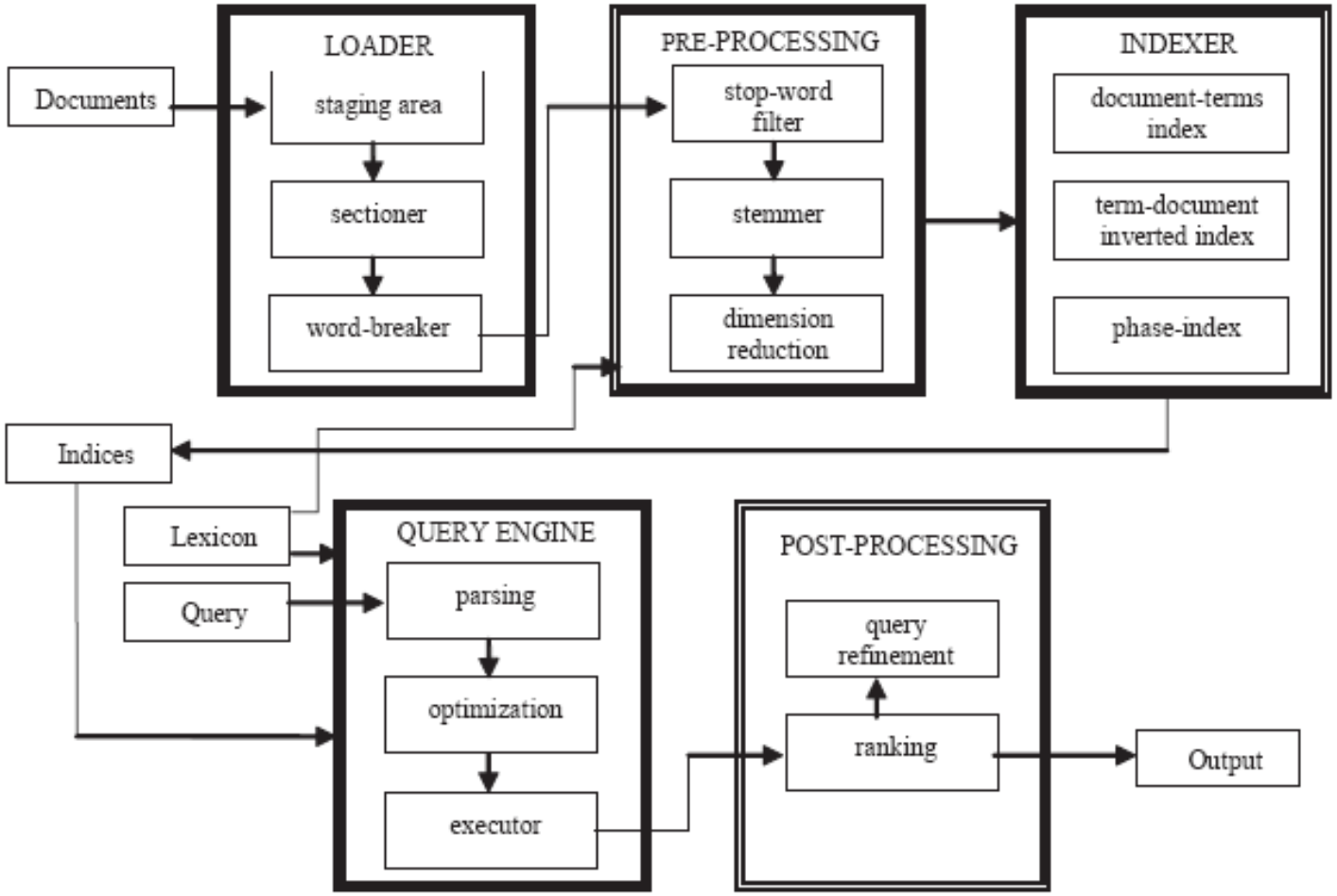
Témy

- Budovanie morfologickej databázy:
  - harvesting dát
  - určenie morfologických atribútov
- Slovníky (koreňových morfém, retrográdny)
- Kontrola pravopisu a gramatiky
- Viacslovné pomenovania (MWE)
- WordNetSK
- Prepis numerálov
- Hodnotenie krátkych odpovedí
- Určenie autorstva textu
- Adaptácia nástrojov pre slovenčinu
- Analýza sentimentu („nálady“, postoja)
- ...

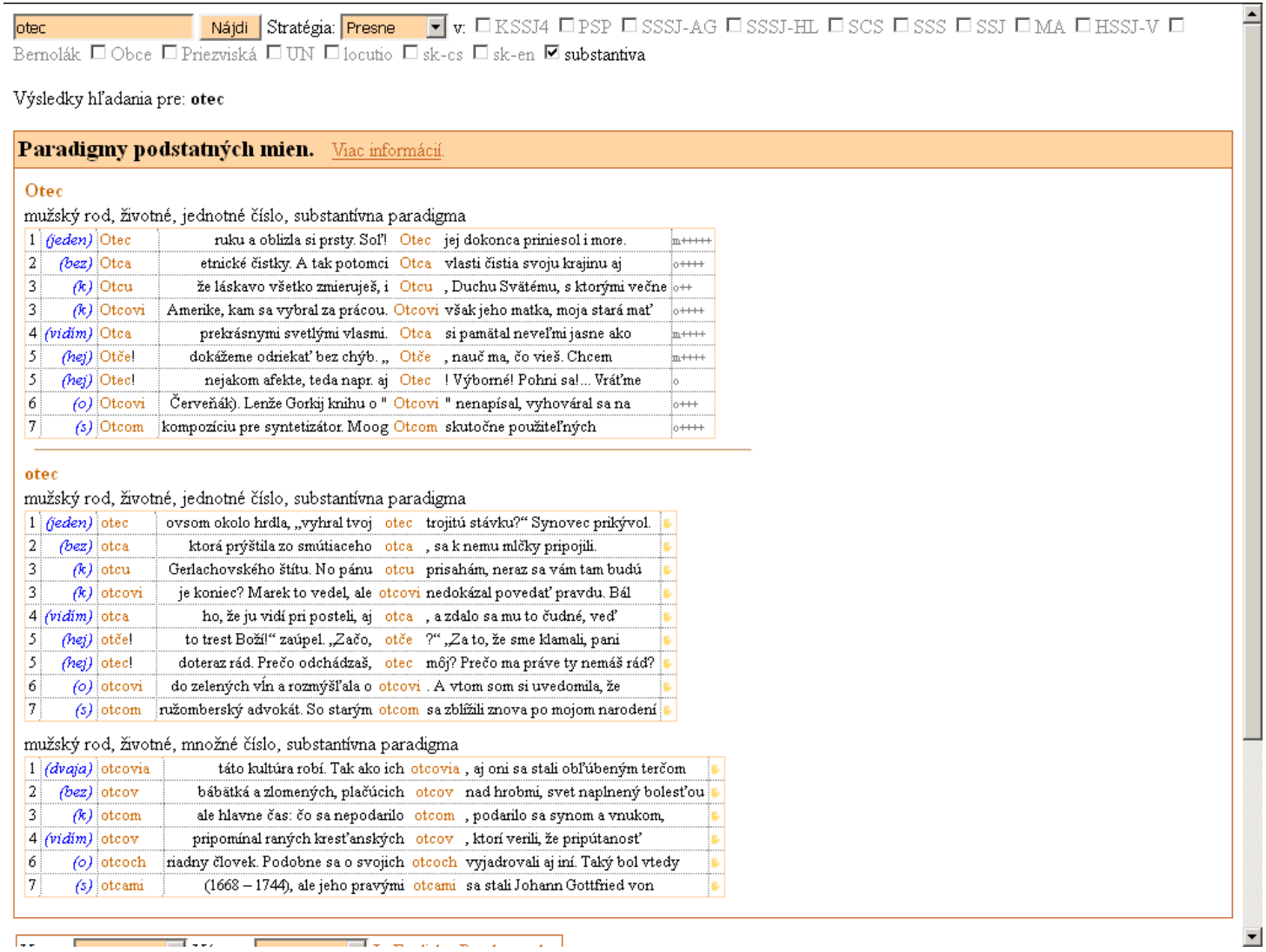
Schéma aplikácie text mining-u



FULL-TEXT SEARCH (FTS) ENGINES



Morfologická databáza



Morfologická databáza (2)

```
abakus abakus SSis1
abakus abakusu SSis2
abakus abakusu SSis3
abakus abakus SSis4
abakus abakus SSis5
abakus abakuse SSis6
abakus abakusom SSis7
abakus abakusy SSip1
abakus abakusov SSip2
abakus abakusom SSip3
abakus abakusy SSip4
abakus abakusy SSip5
abakus abakusoch SSip6
abakus abakusmi SSip7
```

```
zvľúdnieť->VKdpa+       zvľúdnieme
zvľúdnieť->VKdpb+       zvľúdniete
zvľúdnieť->VKdpc+       zvľúdnejú
zvľúdnieť->VKdsa+       zvľúdniem
zvľúdnieť->VKdsb+       zvľúdnieš
zvľúdnieť->VKdsc+       zvľúdnie
zvľúdnieť->VLdpah+      zvľúdneli
zvľúdnieť->VLdpbh+      zvľúdneli
zvľúdnieť->VLdpcf+      zvľúdneli
zvľúdnieť->VLdpci+      zvľúdneli
zvľúdnieť->VLdpcm+      zvľúdneli
zvľúdnieť->VLdpcn+      zvľúdneli
zvľúdnieť->VLdsaf+      zvľúdnela
zvľúdnieť->VLdsai+      zvľúdnel
zvľúdnieť->VLdsam+      zvľúdnel
zvľúdnieť->VLdsan+      zvľúdnelo
zvľúdnieť->VLdsbf+      zvľúdnela
zvľúdnieť->VLdsbi+      zvľúdnel
zvľúdnieť->VLdsbm+      zvľúdnel
zvľúdnieť->VLdsbn+      zvľúdnelo
zvľúdnieť->VLdscf+      zvľúdnela
zvľúdnieť->VLdsci+      zvľúdnel
zvľúdnieť->VLdscm+      zvľúdnel
zvľúdnieť->VLdscn+      zvľúdnelo
zvľúdnieť->VMdpa+       zvľúdnejme
zvľúdnieť->VMdpb+       zvľúdnejte
zvľúdnieť->VMdsa+       zvľúdnej
```

Transkripcia numerálov

Vstupný text:

Márii Kvopkovej olympijský limit ušiel tesne, lebo skončila na 40. mieste (+25,20) a zo 78 pretekárok, ktoré sa postavili na štart, bolo potrebné zdolať polovicu.

Medzivýsledok:

Márii Kvopkovej olympijský limit ušiel tesne, lebo skončila na <cislovka typ=radove tvar=samostatne rod=s cislo=s pad=5>40.</cislovka> mieste (<cislovka typ=desatinne tvar=samostatne rod= pad=>+25,20</cislovka>) a zo <cislovka typ=cele tvar=samostatne rod=ž pad=2>78</cislovka> pretekárok, ktoré sa postavili na štart, bolo potrebné zdolať polovicu.

Výstupný text:

Márii Kvopkovej olympijský limit ušiel tesne, lebo skončila na štyridsiatom mieste (plus dvadsaťpäť celých dvadsať stotín) a zo sedemdesiatich ôsmich pretekárok, ktoré sa postavili na štart, bolo potrebné zdolať polovicu.

BP: Ondrej DZURJUV: Transkripcia čísel na text. 2009

Transkripcia numerálov (2)

Vstupný text:

Diskusia: 8 - posl. 27.4.2009 08:16 od: Martin

Medzivýsledok:

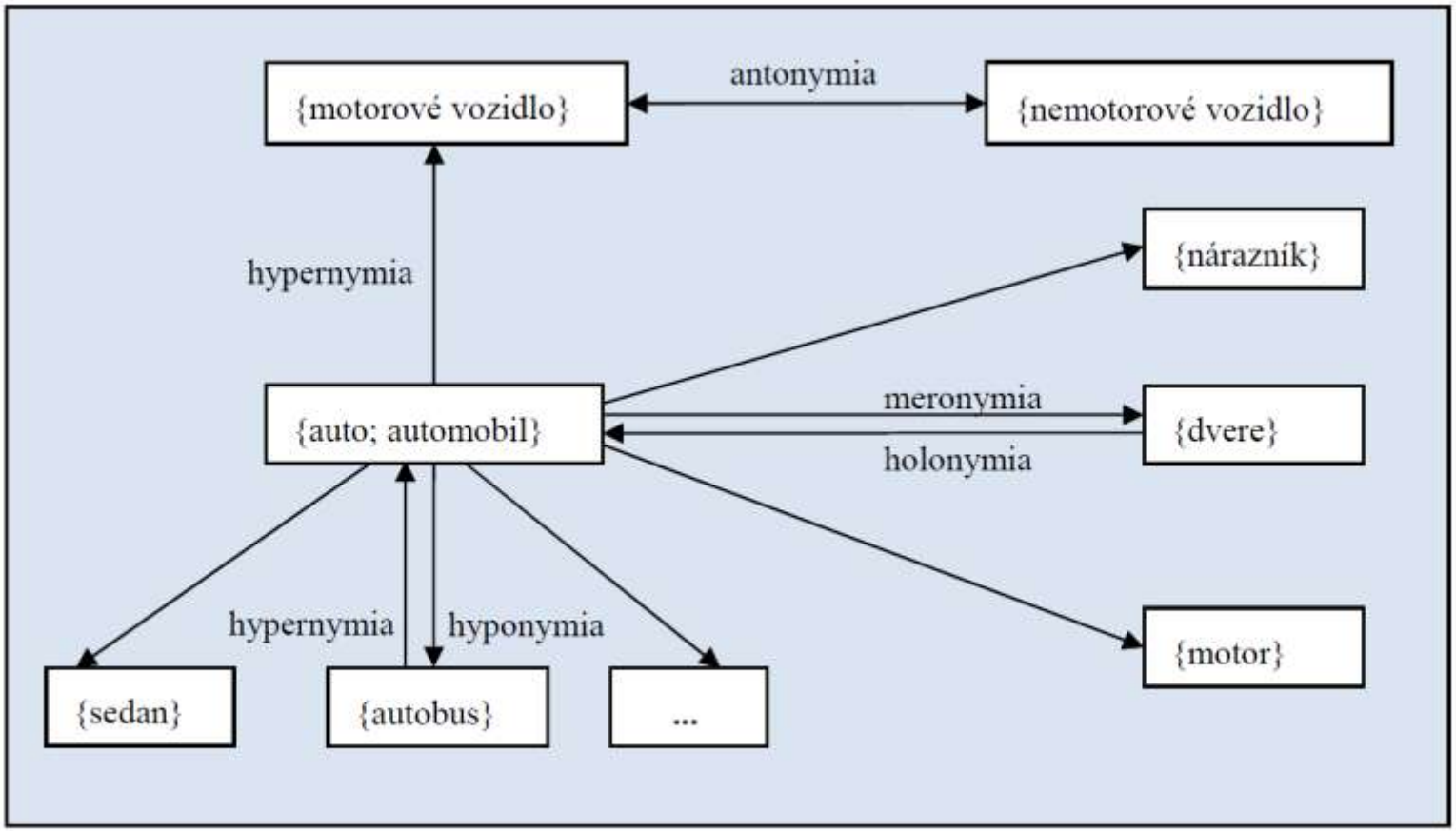
Diskusia: <cislovka typ=cele tvar=samostatne rod= pad=>8</cislovka> - posl. <cislovka typ=datum tvar=samostatne rod=m pad=>27.4.2009</cislovka> <cislovka typ=cas tvar=samostatne format=HHMM rod=ž pad=>08:16</cislovka> od: Martin

Výstupný text:

Diskusia: osem - posl. dvadsiaty siedmy apríl dvetisícdeväť osem hod. šestnásť min. od: Martin

BP: Ondrej DZURJUV: Transkripcia čísel na text. 2009

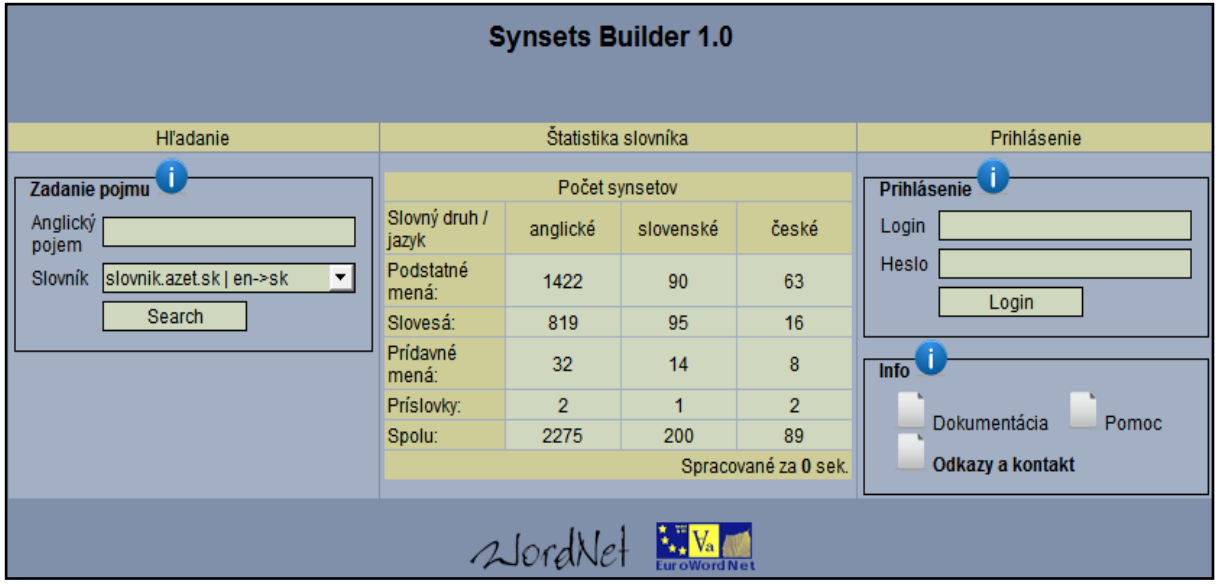
WordNet SK



Obr. 1: Základné vzťahy nachádzajúce sa vo WordNet-e

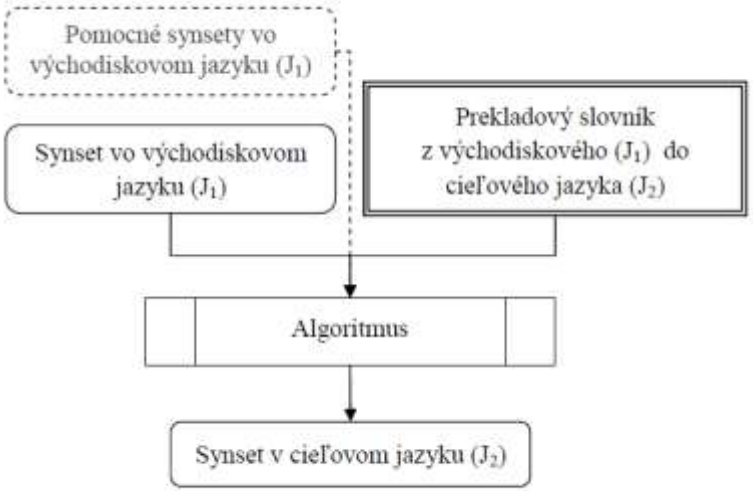
DP: Ondrej DZURJUV: Počítačová lingvistika - Generovanie synonymických radov medzi jazykmi. 2011

SynsetBuilder

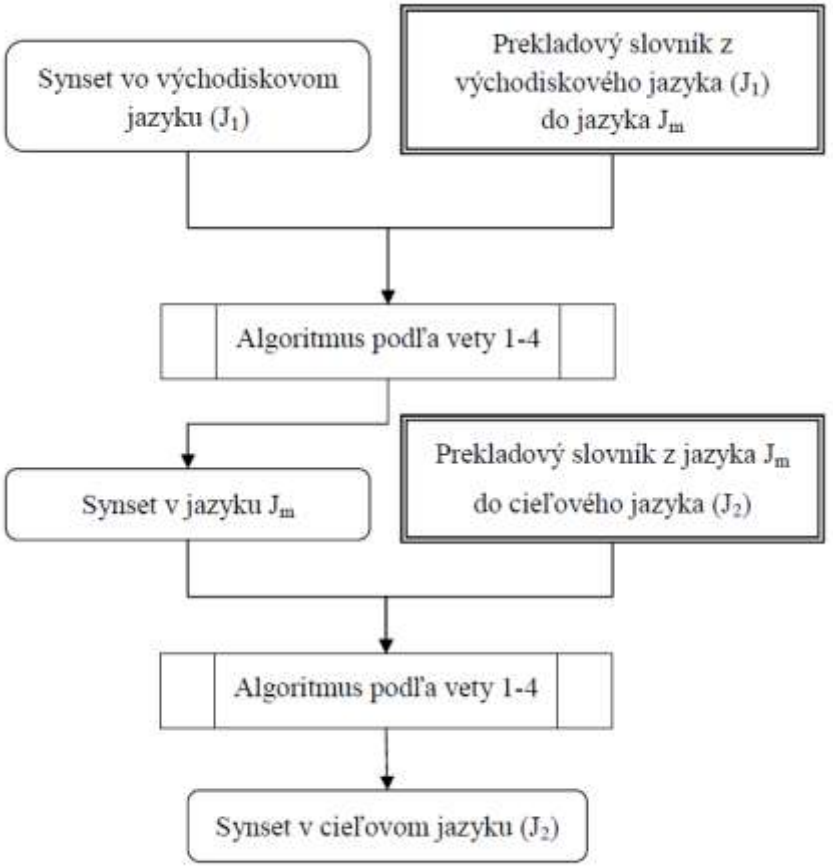


DP: Ondrej DZURJUV: Počítačová lingvistika - Generovanie synonymických radov medzi jazykmi. 2011

SynsetBuilder - princíp



Obr. 5: Všeobecný postup pri priamom budovaní synsetov v cieľovom jazyku



Obr. 7: Postup pri budovaní synsetov v cieľovom jazyku cez medzijazyk

DP: Ondrej DZURJUV: Počítačová lingvistika - Generovanie synonymických radov medzi jazykmi. 2011

SynsetBuilder - výsledky

Slovo car – vyplnené 2 synsety z 5:

02853224: car; auto; automobile; machine; motorcar *4-wheeled motor vehicle; usually propelled by an internal combustion engine; "he needs a car to get to work"*	vozidlo; automobil; motorové vozidlo; auto; automobilový; osobný automobil
02854760: car; railcar; railway car; railroad car *a wheeled vehicle adapted to the rails of railroad; "three cars had jumped the rails"*	železničný vozeň

Slovo computer – vyplnené 2 synsety z 2:

02971359: computer, computing machine, computing device, data processor, electronic computer, information processing system *a machine for performing calculations automatically*	počítač
09257296: calculator, reckoner, figurer, estimator, computer *an expert at calculation (or at operating calculating machines)*	kalkulant; počítač; kalkulátor

Slovo information – vyplnené 3 synsety z 5:

06225142: information, info *a message received and understood*	správa; informácia
07949563: data, information *a collection of facts from which conclusions may be drawn; "statistical data"*	informačný
05479334: information *knowledge acquired through study or experience or instruction*	informácia, vedomosť, znalosť

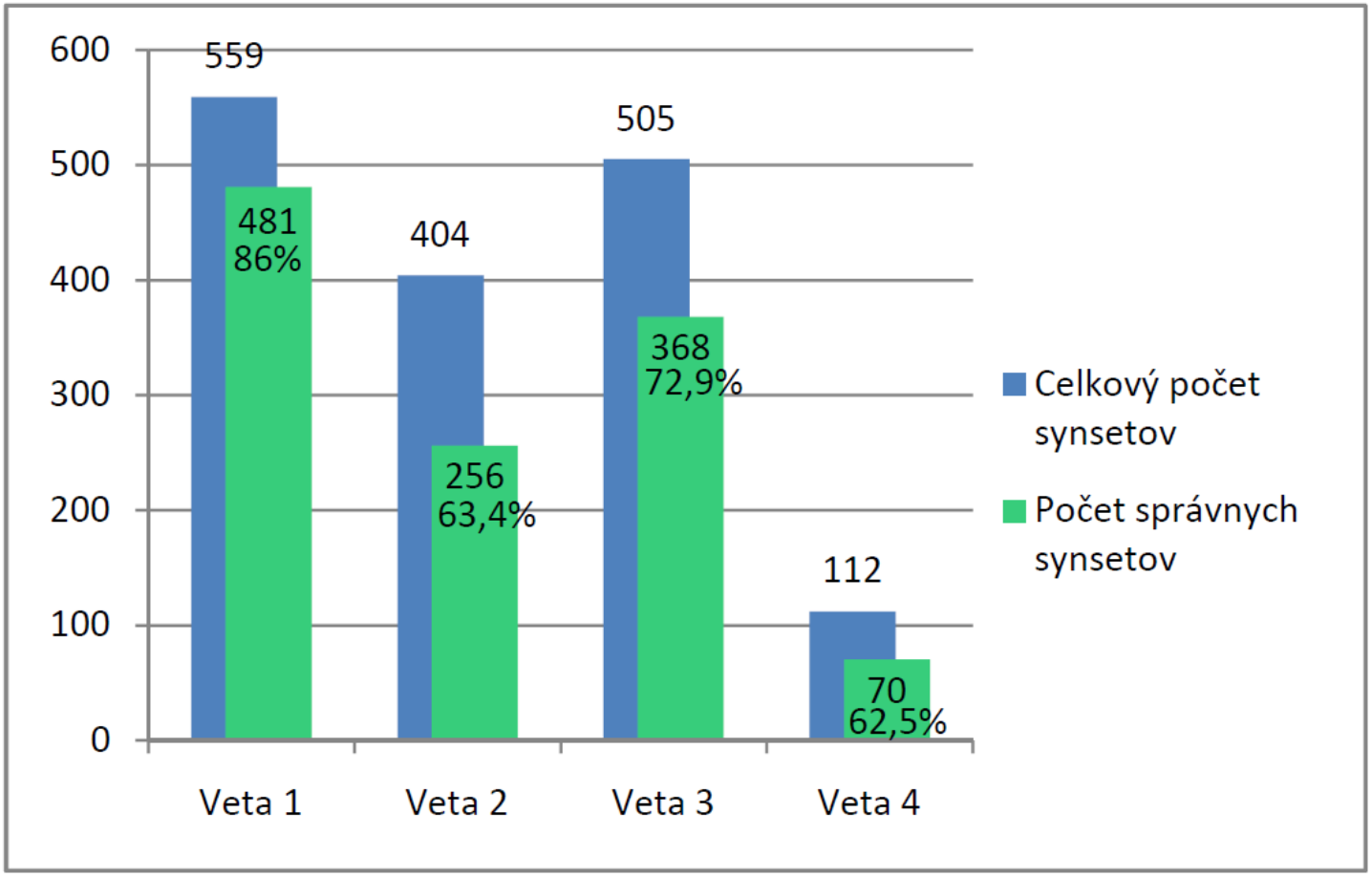
Synset Builder - štatistiky

Tab. 10: Úspešnosť celkového generovania slovenských synsetov

	Všetky	Podstatné mená	Prídavné mená	Slovesá	Príslovky
Počet EN synsetov vo WordNet-e	117 659	82 115	18 156	13 767	3 621
Počet EN synsetov, s nagenerovaným slovenským synsetom	40 521 (34,4%)	26 787 (32,6%)	6 859 (37,8%)	5 839 (42,4%)	1 036 (28,6%)
Počet SK synsetov vytvorených vetou 1	10 267 (8,7%)	5 705	2 175	2 109	278
Počet SK synsetov vytvorených vetou 2	30 243 (25,7%)	20 510	6 059	2 715	959
Počet SK synsetov vytvorených vetou 3	11 533 (12%)[3]	8 192	-	3 341	-
Počet SK synsetov vytvorených vetou 4	1 917 (1,4%)[3]	1 348	-	569	-

DP: Ondrej DZURJUV: Počítačová lingvistika - Generovanie synonymických radov medzi jazykmi. 2011

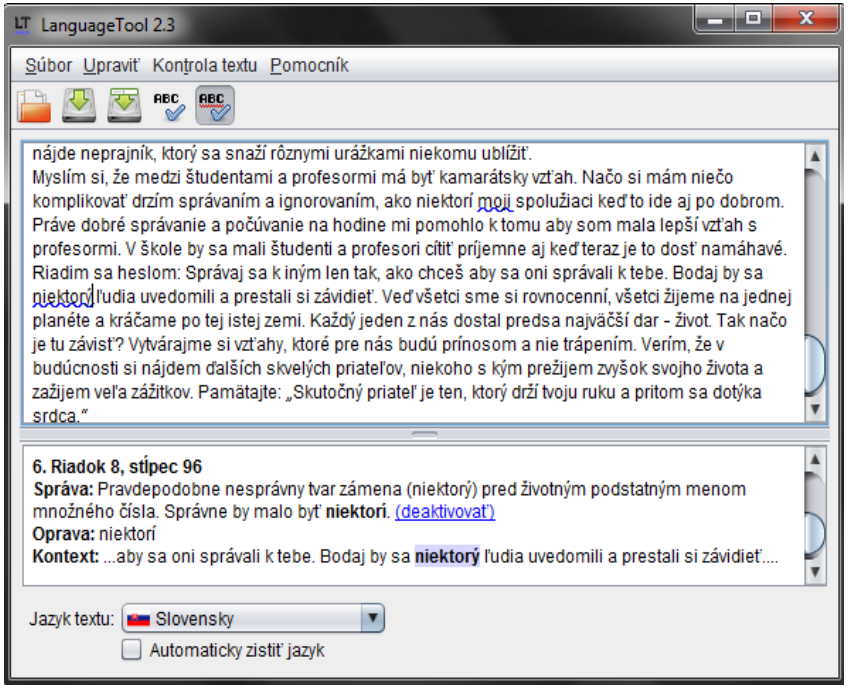
Synste Builder - štatistiky



Obr. 11: Pomer správnych slovenských synsetov k ich celkovému počtu vo vzorke

DP: Ondrej DZURJUV: Počítačová lingvistika - Generovanie synonymických radov medzi jazykmi. 2011

Kontrola gramatiky



DP: Mária Andóová: Kontrola gramatiky v slovenských textoch, 2014

Určenie autorstva textu

	Blaha	Chren	Hudacky	Kanik	Matovic	Micovsky	Mikus	Poliacik	Tomanova	Vasecka
Blaha	9.74	14.93		10.50	21.45	19.15	12.58	18.25	10.91	11.76
Chren	10.80	11.52	7.32		13.18	13.42	9.48	10.09	7.64	10.92
Hudacky	10.32	13.78	7.80	6.88	12.97	10.03		11.02	11.43	7.32
Kanik	8.52	11.79	7.33	5.28	12.88	10.36		10.52	10.87	6.39
Matovic	18.22	18.27	15.75	14.29	9.82		11.30	9.57	20.29	14.35
Micovsky	10.83	12.21	9.28	7.43	12.05	7.45		8.97	15.10	8.38
Mikus	11.28	13.53	9.62	8.43	11.63	7.20	5.10	9.08	12.98	7.91
Poliacik	11.60	10.01	8.43		9.31	9.15	6.01	6.67	10.24	11.31
Tomanova	6.95	12.68	7.33	7.09	16.97	14.19	7.43	13.83	10.80	
Vasecka	9.74	12.63	9.38		19.48	17.93	12.41	15.78	9.55	12.27

Obrázok 2 Frekvencia čiarok, politické rozpravy

	Blaha	Chren	Hudacky	Kanik	Matovic	Micovsky	Mikus	Poliacik	Tomanova	Vasecka
Blaha		15.63	14.42	11.34	12.70	11.39	12.01	11.69	9.91	10.66
Chren	12.19	7.87	14.56	10.33	10.01	12.58	13.37	12.54	11.38	
Hudacky	13.81	13.35		12.79	8.90	12.84	9.92	13.58	12.22	11.03
Kanik	9.56	12.84	12.88		8.78	10.93	9.60	10.98	10.57	9.42
Matovic	12.13	13.49	12.92	10.64		10.24	10.44	9.47	12.65	10.69
Micovsky		12.30	12.95	8.81	9.25		10.61	8.91	8.20	9.22
Mikus	10.48	13.35	8.61	10.48		9.90		11.05	11.35	9.43
Poliacik	10.37	10.36	13.38	11.00	9.00	10.08	10.46		10.83	
Tomanova	10.10	12.39	10.83		9.57	11.33		10.83	9.30	
Vasecka		11.68	13.38	9.99	11.28	11.53	11.83	10.45		

Obrázok 4 Dĺžka viet, politické rozpravy

	Blaha	Chren	Hudacky	Kanik	Matovic	Micovsky	Mikus	Poliacik	Tomanova	Vasecka
Blaha		4.71	5.20	4.36	4.65	4.97	4.16	4.65	5.15	
Chren	4.18		4.58	6.15	4.75	5.11	4.23	4.42	4.49	
Hudacky	6.32	5.22		7.31	5.35	5.10	5.05		6.43	5.48
Kanik	4.51	4.30	6.17		5.24	4.63	5.57	4.98	6.46	
Matovic	5.62	4.34	5.31	6.47			4.76	4.61	6.32	4.45
Micovsky	4.75	4.21	5.28	5.25			4.95	4.05	5.98	
Mikus	4.84	4.02	4.51	6.41	4.51	4.83		4.19	4.77	
Poliacik	4.49		5.67	5.32	4.41	4.31	3.32		4.41	
Tomanova	4.33		5.52	7.19	6.48	7.21		5.23		4.79
Vasecka		3.87	5.71	5.35			4.84	4.41	4.88	

Obrázok 6 Stop slová, politické rozpravy

n =	Blaha	Chren	Hudacky	Kanik	Matovic	Micovsky	Mikus	Poliacik	Tomanova	Vasecka
Blaha		4.18	1.94	2.95	5.31	4.24		4.50		
Chren										
Hudacky										
Kanik										
Matovic										
Micovsky										
Mikus										
Poliacik										
Tomanova										
Vasecka										

Obrázok 8 Výskyt najfrekventovanejších slov, politické rozpravy

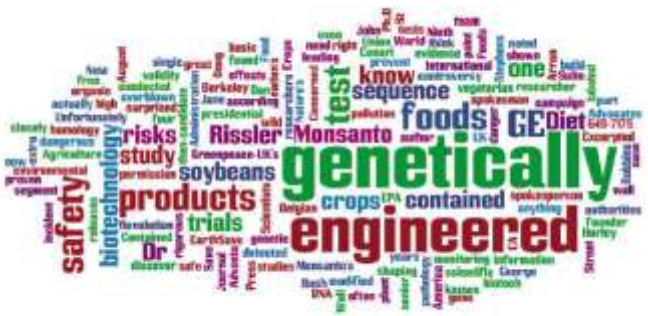
BP: Michal Pavlík: Určovanie autorstva textu. 2015

Triedenie

- Podľa kódu
- Podľa normy
- Retrográdne

Vizualizácia

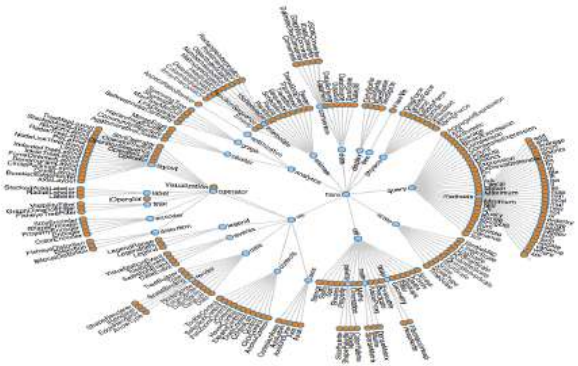
Príklady vizualizácie



Obr. 1 Príklad vytvorenia oblaku slov pre dokument [5].

Obr. 2 Príklad vizualizácie pomocou bublinového grafu [5].

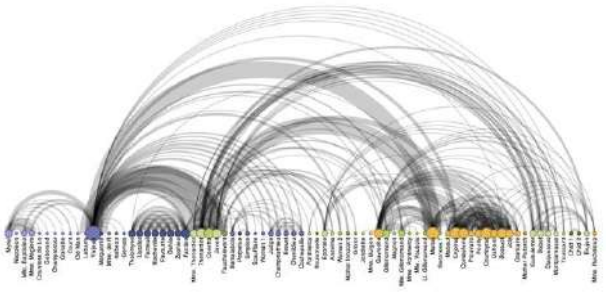
Obr. 3 Príklad stromového diagramu zobrazujúci postupnosti slov v dokumente [6].



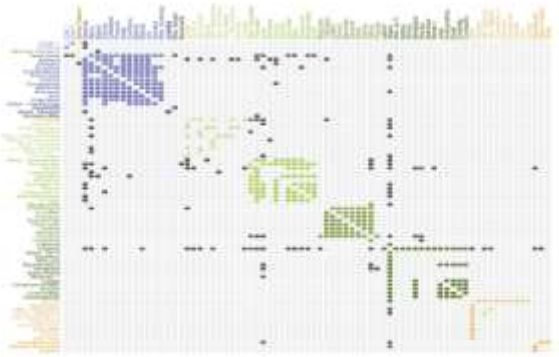
Obr. 4 Príklad vizualizácie lúčovitého stromu [7].

BP: Matúš Verčimák: Interaktívna vizualizácia dát - texty a dokumenty. 2016

Príklady vizualizácie (2)



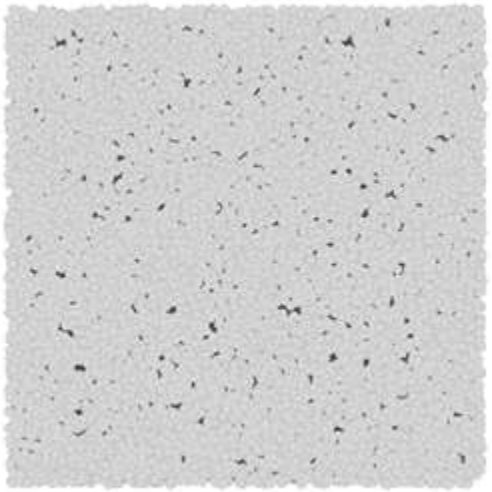
Obr. 5 Príklad vizualizácie slov v kapitole dokumentu pomocou oblúkového diagramu [8].

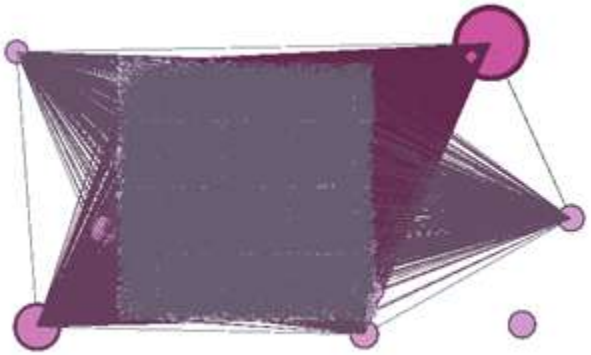


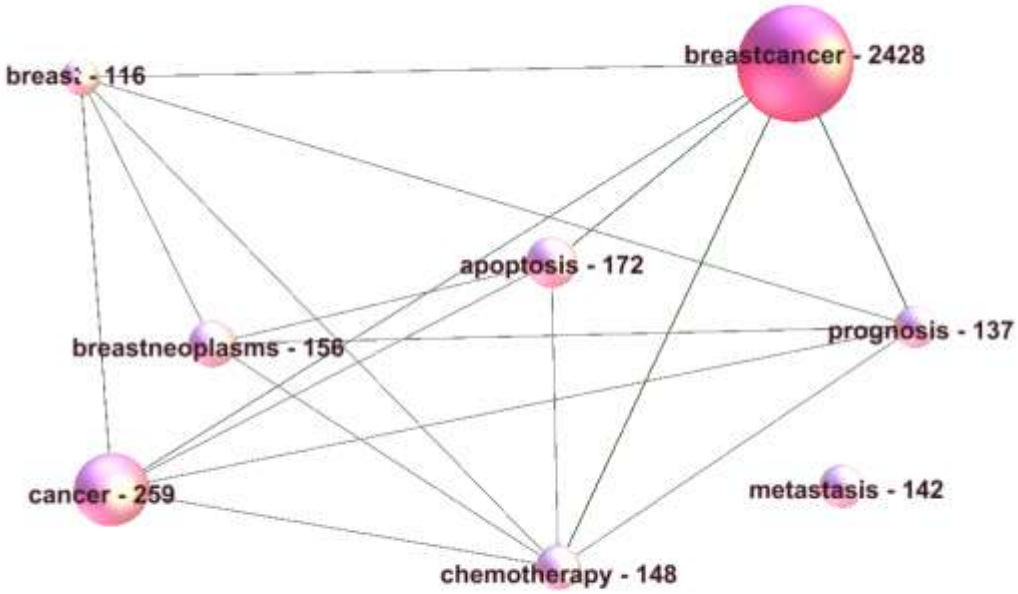
Obr. 6 Príklad maticovej vizualizácie slov v dokumentoch [9].

BP: Matúš Verčimák: Interaktívna vizualizácia dát - texty a dokumenty. 2016

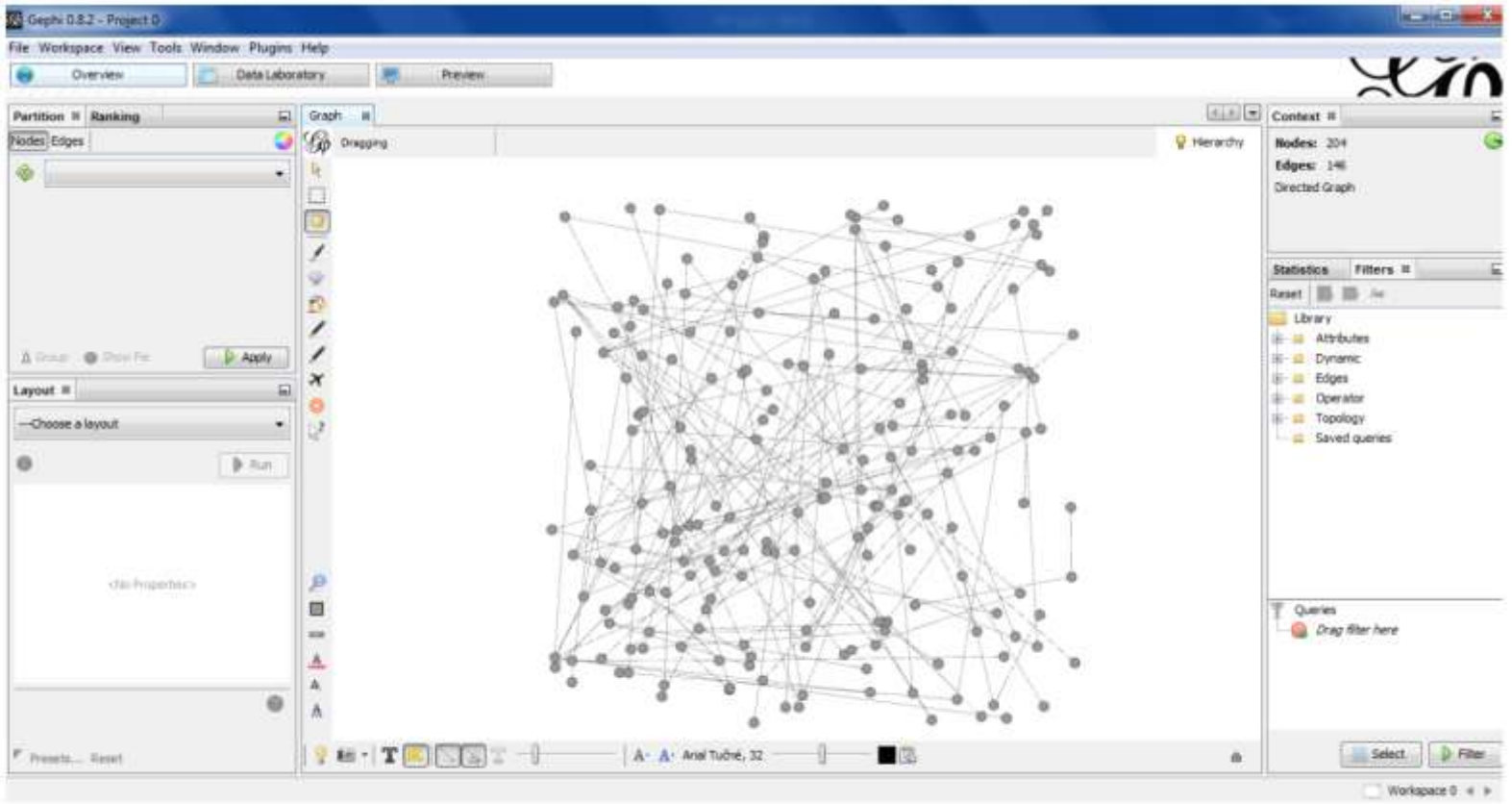
Vizualizácia kľúčových slov (1)







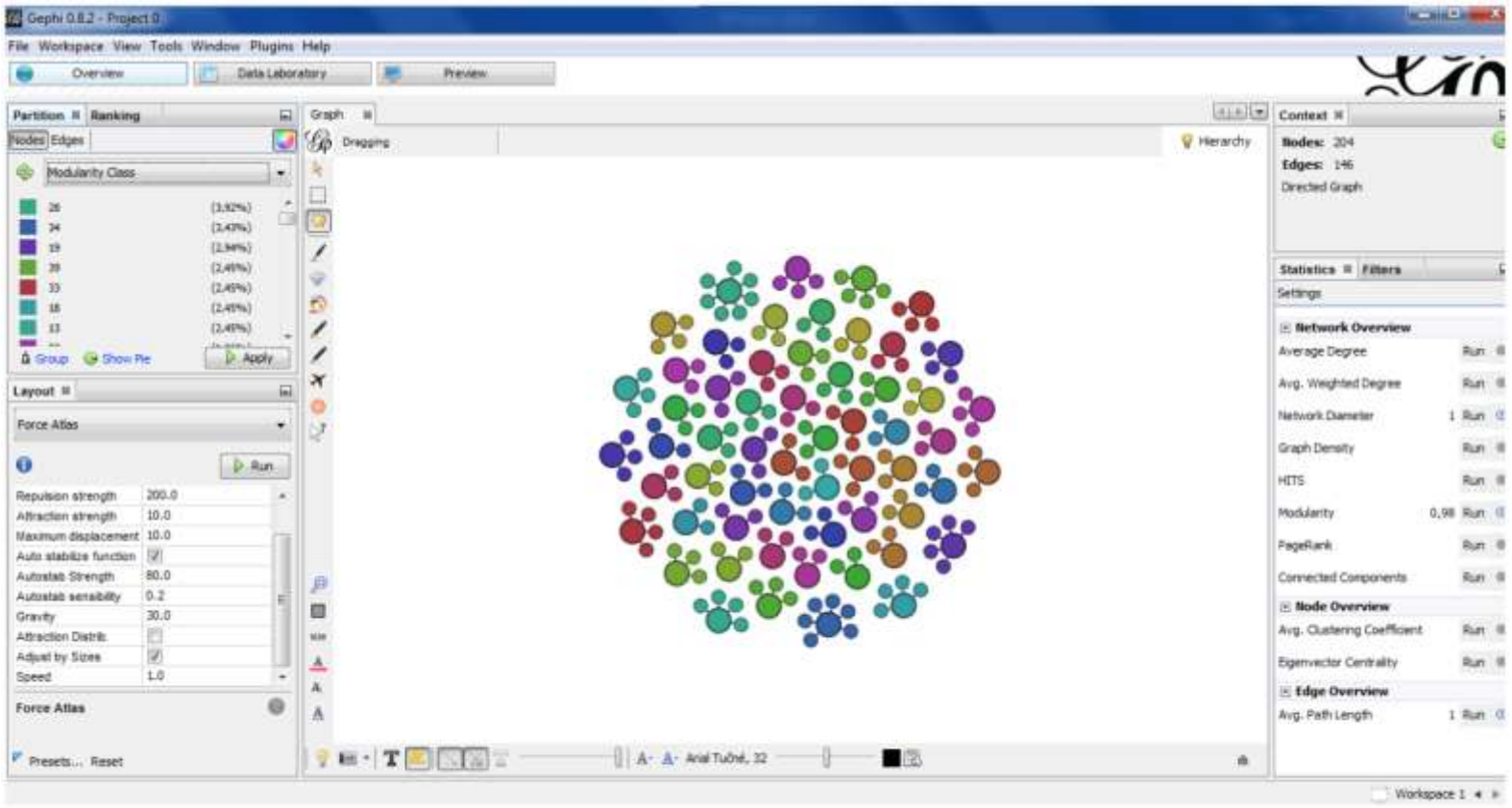
Vizualizácia kľúčových slov (2)



Obr. 15 Vzhľad grafu po importovaní dát bez akýchkoľvek ďalších úprav.

BP: Matúš Verčimák: Interaktívna vizualizácia dát - texty a dokumenty. 2016

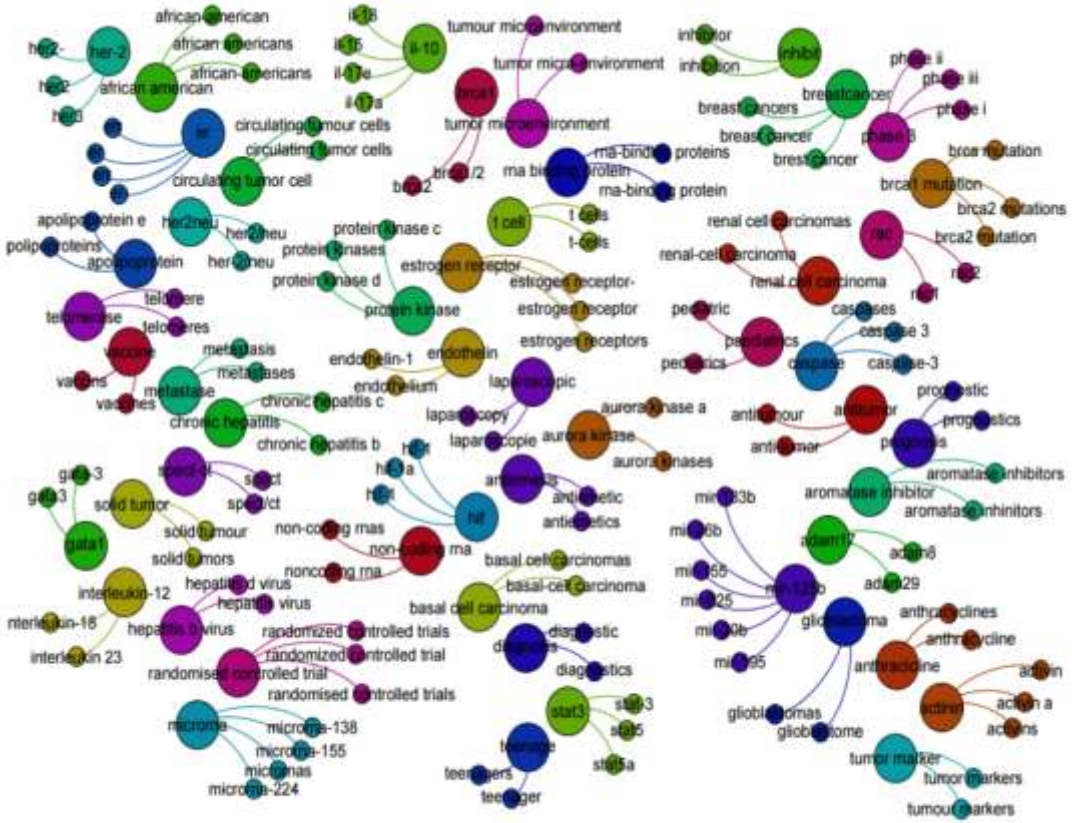
Vizualizácia kľúčových slov (3)



Obr. 18 Farebne odlíšené uzly grafu

BP: Matúš Verčimák: Interaktívna vizualizácia dát - texty a dokumenty. 2016

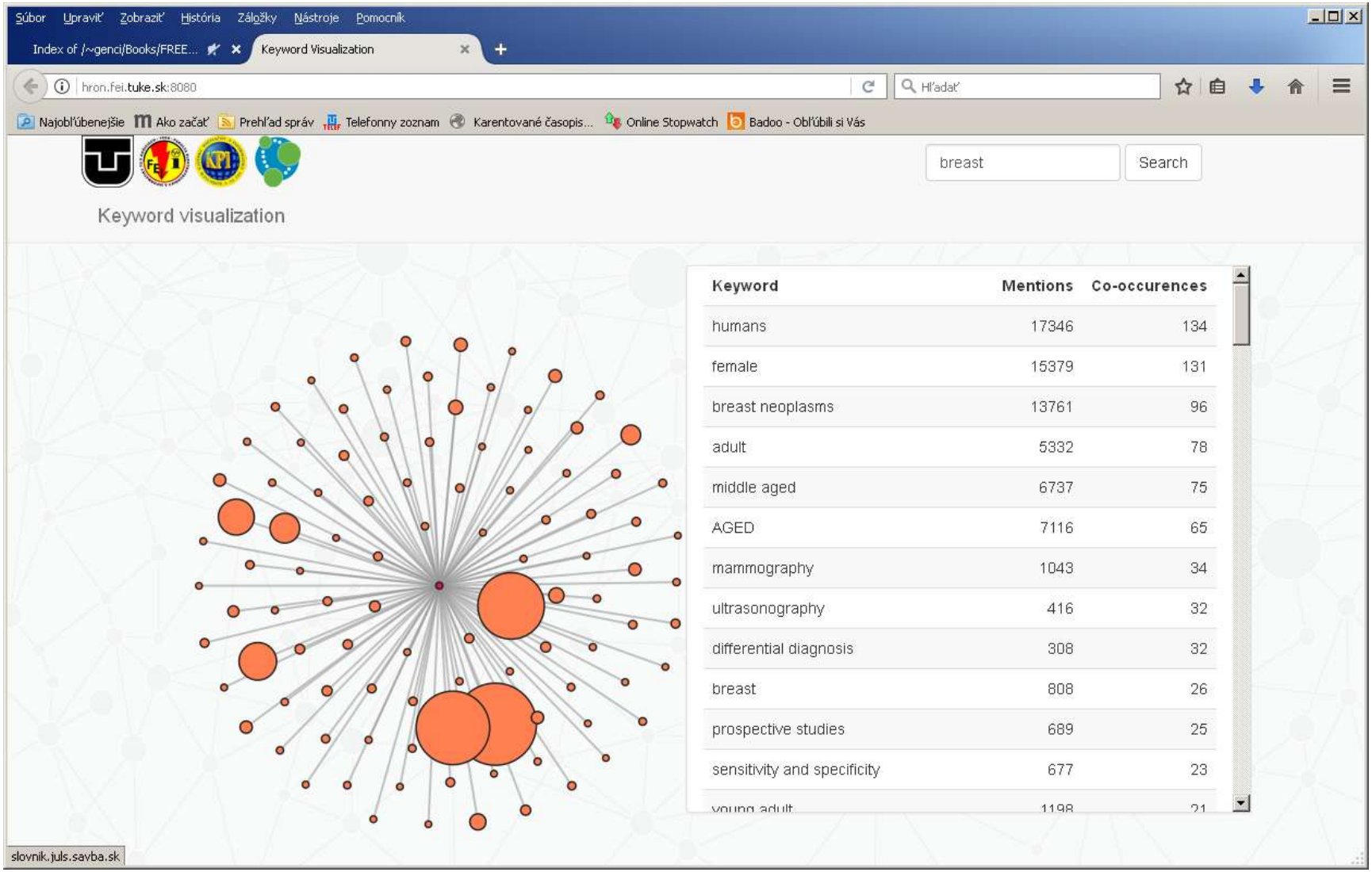
Vizualizácia kľúčových slov (4)



Obr. 19 Vzhľad výsledného grafu po finálnych úpravách

BP: Matúš Verčimák: Interaktívna vizualizácia dát - texty a dokumenty. 2016

Vizualizácia kľúčových slov (5)



Ján Magyar: Vizualizácia kľúčových slov. Work in progress

Viacslovné pomenovania, MWE

- Slovné konštrukcie, ktoré majú iný význam ako slová z ktorých sa skladajú. Problém (niekedy?) pri prekladoch:
  - starý otec -> old father (grandfather)
  - vysoká škola -> high school (university)
  - dať pokoj -> give rest (give peace)
- Problém identifikácie:
  - **daj** mi už dnes, prosím ťa, konečne **pokoj**

Viacslovné pomenovania
(korpusové štatistiky)

Tab. 4 Manuálne spracované údaje pre spojenia obsahujúce slovo otec

bigram	*f(xy)*	*f(x)*	*f(y)*	*PMI*	*AJ-slovník*
nebeský otec	1047	3651	93192	2,741057718	N
svätý otec	10885	44369	93192	2,730414679	A
starostlivý otec	226	1271	93192	2,488156216	N
starý otec	11162	77050	93192	2,420370161	A
milujúci otec	215	1897	93192	2,280584256	N
vzorný otec	96	1191	93192	2,191590743	N
nebohý otec	206	3241	93192	2,084992497	N
duchovný otec	706	14789	93192	1,936980839	A
otec biskup	1111	93192	31464	1,833809557	N
hrdý otec	208	10757	93192	1,70948726	N
otec arcibiskup	310	93192	16478	1,578521493	N
nešťastný otec	92	8674	93192	1,316993534	N
vlastný otec	404	58045	93192	1,305078563	N
šťastný otec	138	33018	93192	1,056189777	N
dobrý otec	318	144632	93192	0,873237917	N

DP: Matej Katuščák: Identifikacia viacslovných pomenovaní pre slovenský jazyk. 2015

Viacslovné pomenovania

(dáta v Webu)

Bigram	Frekvencia	VBI	T-skóre	Dice
stainless steel	383129	5.84485	608.205	0.940286
obchodné podmienky	446433	5.2125	650.136	0.809389
e mail	402408	4.28945	601.916	0.494587
made of	331733	4.5548	551.457	0.513679
práva vyhradené	153738	6.94351	388.909	0.865647
obchodné podmienky	446433	5.2125	650.136	0.809389
stainless steel	383129	5.84485	608.205	0.940286
e mail	402408	4.28945	601.916	0.494587
made of	331733	4.5548	551.457	0.513679
nákupný košík	179520	5.84878	416.346	0.598913

Tabuľka: Bigramy - prých 5 má najvyššiu hodnotu súčinu hodnôt, posledných 5 hodnotu T-skóre

DP: Jozef Kaščak: Identifikacia viacslovných pomenovaní. 2016

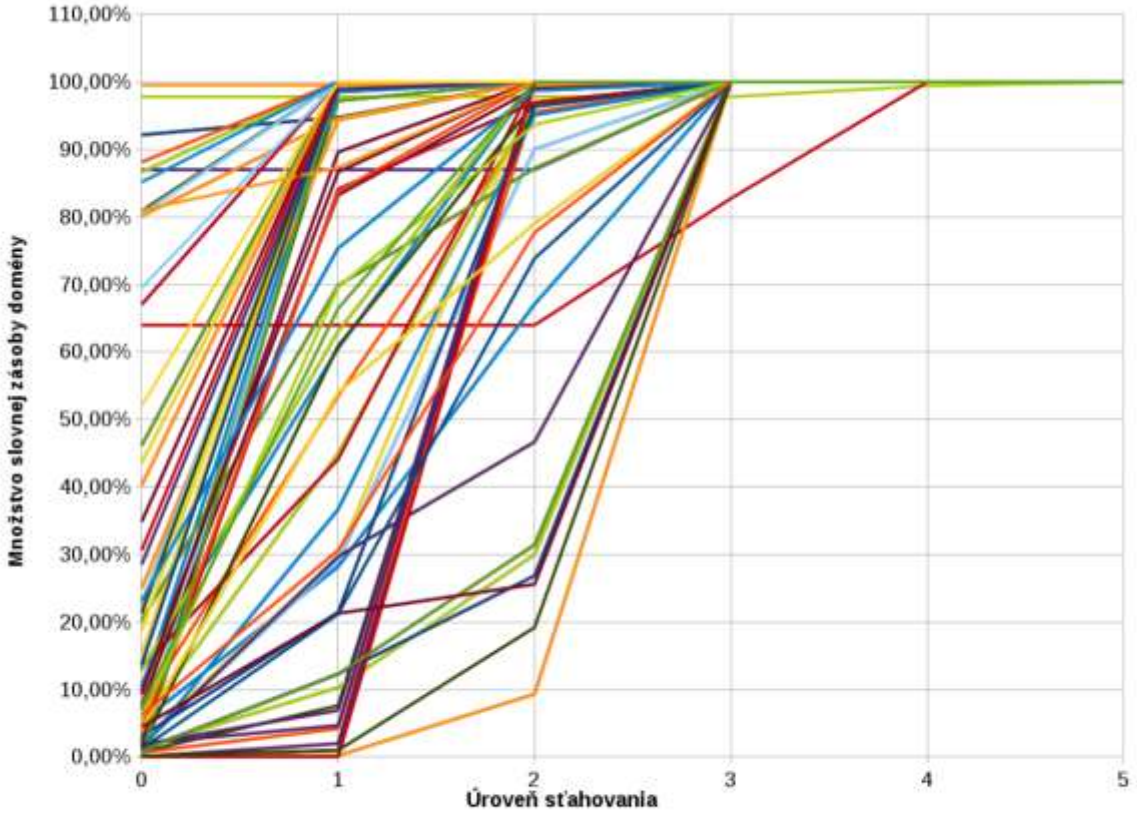
Viacslovné pomenovania
(dáta v Webu)

Trigram	Frekvencia	VBI	T-skóre	Dice
najkrajšej slovenke rozhodovali	3613	1.94527	60.1042	0.95355
petru kvitovú povedie	3237	2.19628	56.8911	0.937355
abdél hakim awyan	2365	3.28054	48.6292	0.979161
receiving personalized recommendations	3178	2.10449	56.3701	0.907395
pomlérun pomlérun pomlérun	2160	3.45609	46.4739	0.96
všetky práva vyhradené	151825	-10.7569	386.805	0.539609
ochrana osobných údajov	86795	-9.84355	293.167	0.491108
vytvorené systémom www	79429	-10.6806	279.196	0.271958
farebný uv gél	72089	-8.67889	267.49	0.596628
aisi stainless steel	72130	-11.8715	262.016	0.238011

Tabuľka: Trigramy - prých 5 má najvyššiu hodnotu súčinu hodnôt, posledných 5 hodnotu T-skóre

DP: Jozef Kaščak: Identifikacia viacslovných pomenovaní. 2016

Viacslovné pomenovania
(dáta v Webu)



Obr.: Množstvo slovnej zásoby k úrovni sťahovania

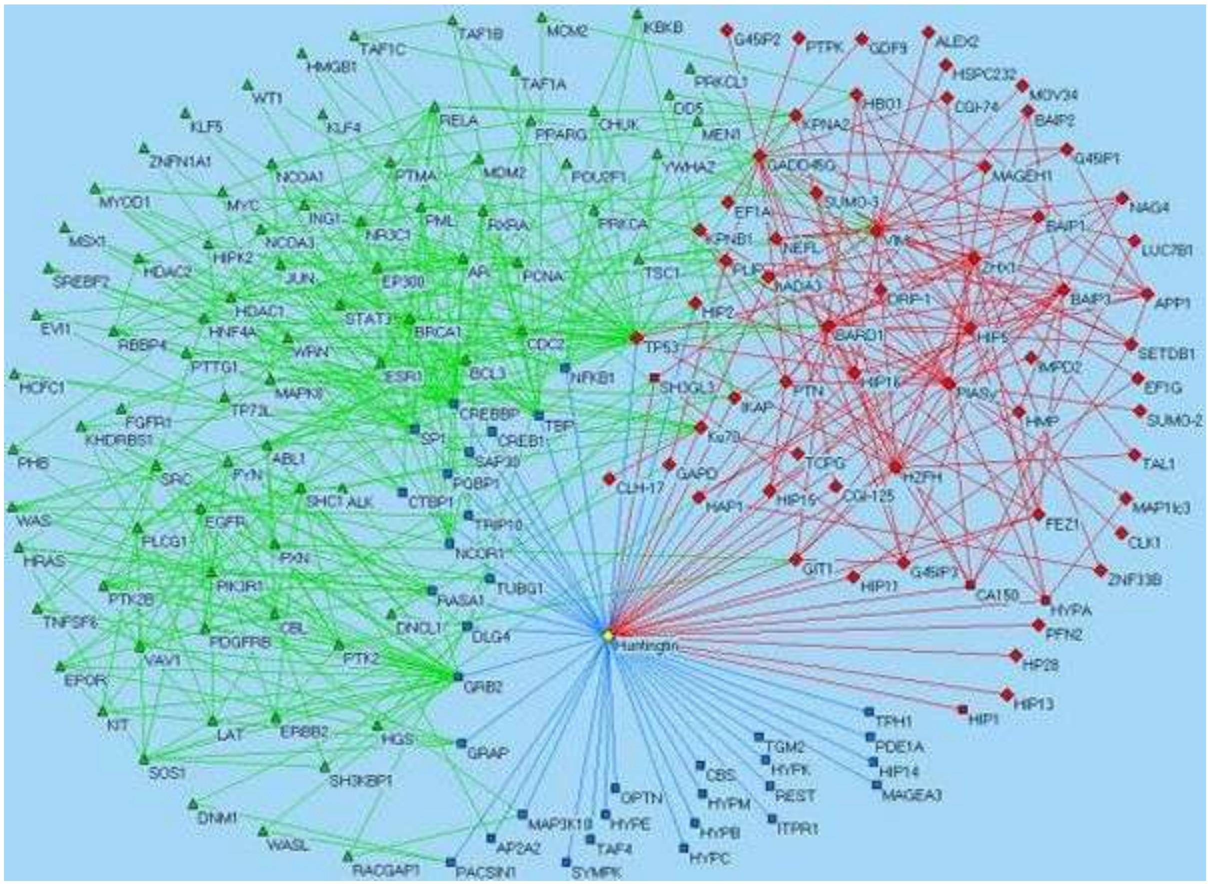
DP: Jozef Kaščak: Identifikacia viacslovných pomenovaní. 2016

NLP a Bioinformatika

Budovanie „informačných“ modelov

- Spracovanie publikácií
  - breast cancer od 1.1.2014
    - v databáze Pubmed – 336 508 článkov
    - Sciencedirect.com – 72 753 článkov
    - PlosOne – 64 035 článkov
- Modely interakcií proteínov
- Signálne dráhy

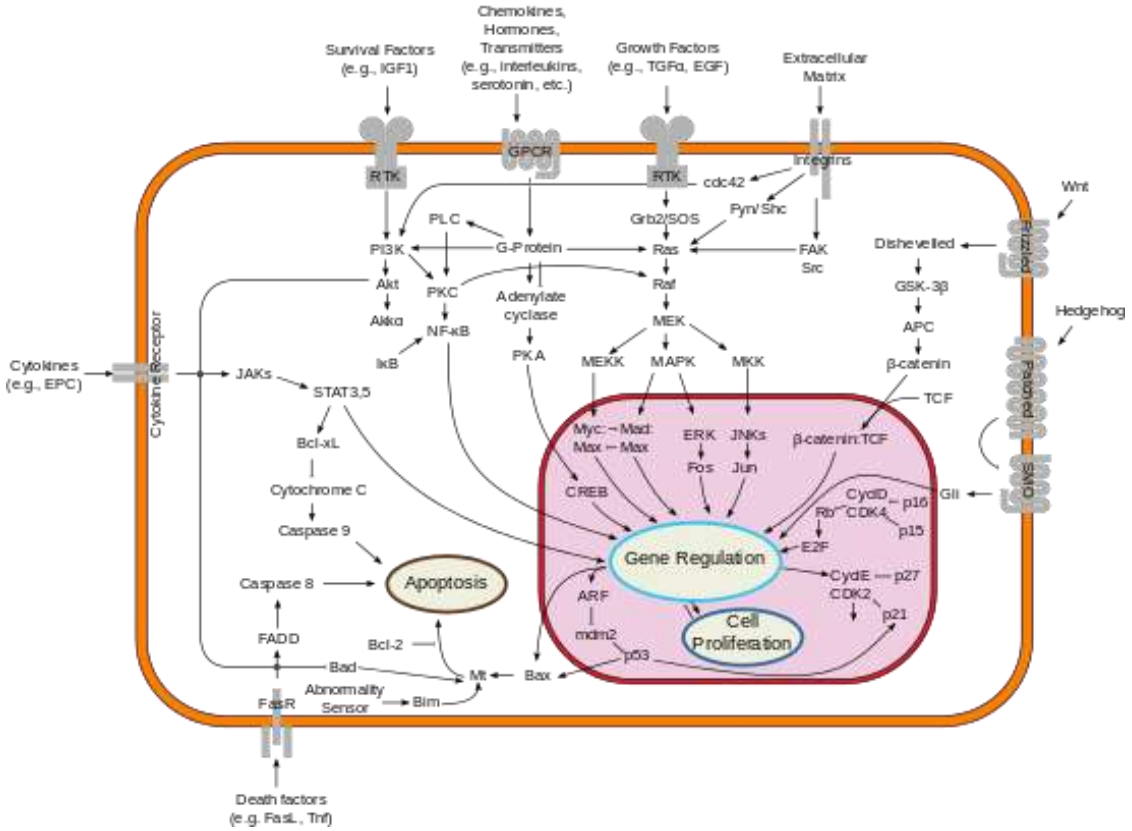
PPI



TAF1B
MCM2
KBKB
TAF1C
HMGB1
TAF1A
PRKCL1
WT1
KLF5
KLF4
RELA
PPARG
CHUK
DD5
MEN1
ZNFN1A1
NCOA1
PTMA
MDM2
POU2F1
YWHAZ
MYOD1
MYC
ING1
NR3C1
PML
RXRA
PRKCA
MSX1
HIPK2
NCOA3
HDAC2
SREBF2
JUN
EP300
AR
PCNA
TSC1
HDAC1
STAT3
EVI1
HNF4A
BRCA1
CDC2
TP53
RBBP4
WRN
PTTG1
ESR1
BCL3
NFKB1
HCFC1
MAPK8
CREBBP
TBP
FGFR1
TP73L
KHDRBS1
SP1
CREB1
ABL1
SAP30
PHB
SRC
FYN
PQBP1
SHC1
ALK
CTBP1
WAS
EGFR
TRIP10
PLCG1
PXN
NCOR1
HRAS
PIK3R1
TUBG1
PTK2B
RASA1
TNFSF6
CBL
DNCL1
DLG4
PDGFRB
VAV1
PTK2
EPOR
GRB2
KIT
LAT
ERBB2
HGS
GRAP
SOS1
SH3KBP1
DNM1
WASL
RACGAP1
PACSIN1
MAP3K10
AP2A2
TAF4
SYMPK
OPTN
HYPE
CBS
HYPM
HYPB
HYPC
TGM2
HYPK
REST
ITPR1
TPH1
PDE1A
HIP14
MAGEA3
Huntingtin
G45IP2
PTPK
GDF9
ALEX2
HSPC232
HBO1
CGI-74
MOV34
BAIP2
KPNA2
GADD45G
MAGEH1
G45IP1
SUMO-3
EF1A
NAG4
BAIP1
KPNB1
NEFL
VIM
LUC7B1
ZHX1
PUF
NADA3
DRP-1
BAIP3
APP1
HIP2
BARD1
HIP5
SETDB1
IMPD2
SH3GL3
PTN
HIP1K
PIASy
EF1G
IKAP
HMP
SUMO-2
Ku70
TCPG
HZFH
TAL1
CLH-17
GAPD
MAP1
HIP15
CGI-125
MAP1lc3
FEZ1
CLK1
GIT1
G45IP7
HIP11
CA150
ZNF33B
HYPA
PFN2
HP28
HIP13
HIP1

Protein signal path

chemical or physical signal is transmitted through a cell as a [series of molecular events]



Zdroj: https://en.wikipedia.org/wiki/Signal_transduction

Ďakujem za pozornosť!

Otázky?