



732A54

Big Data Analytics

6hp

<http://www.ida.liu.se/~732A54>

Teachers

- Examiner: Patrick Lambrix (B:2B:474)
- Lectures: Patrick Lambrix,
Christoph Kessler,
Jose Pena,
Valentina Ivanova,
Johan Falkenjack
- Labs: Zlatan Dragisic,
Valentina Ivanova
- Director of studies: Patrick Lambrix



Course literature

- Articles (on web/handout)
- Lab descriptions (on web)



Data and Data Storage



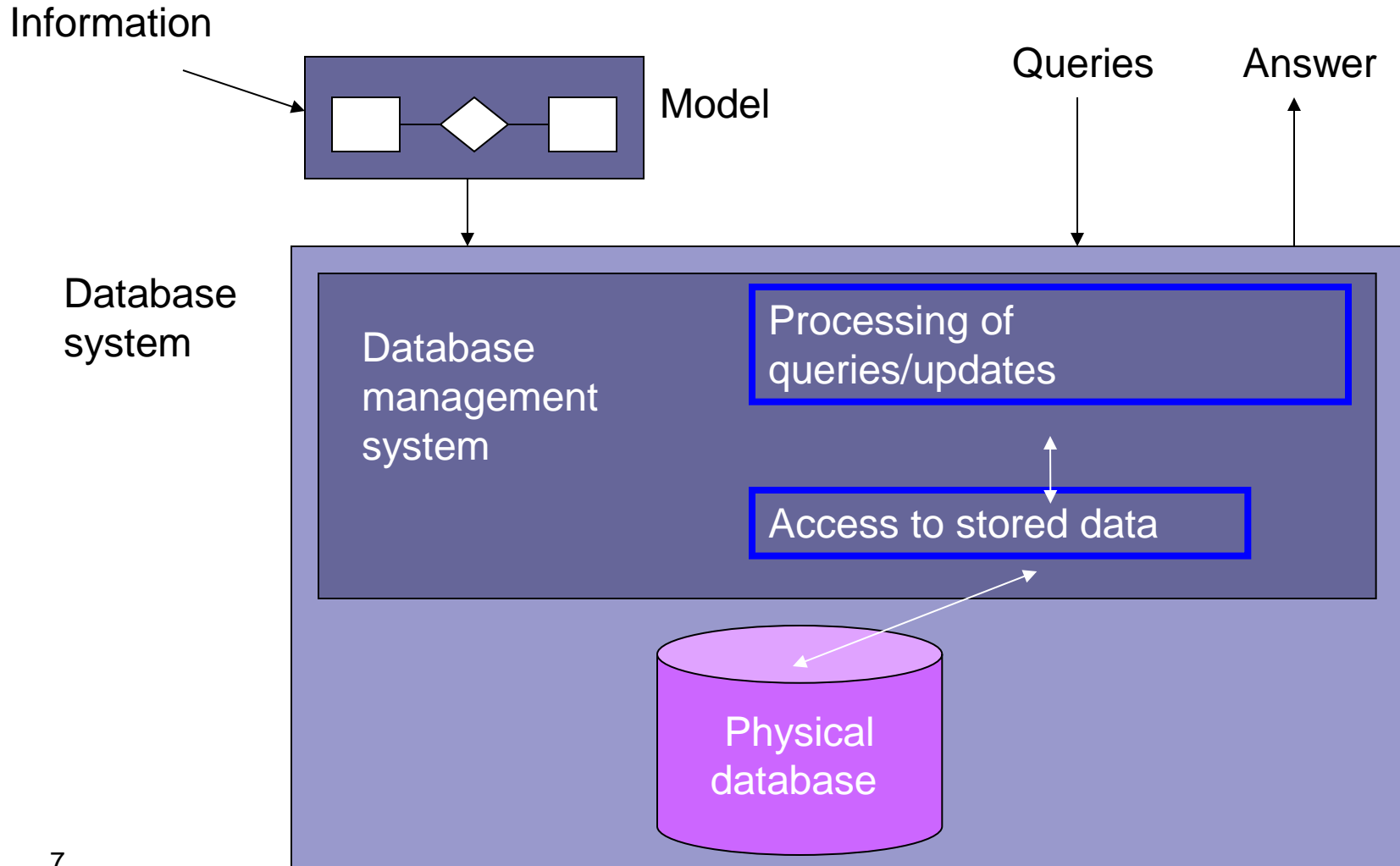
Data and Data Storage

- Database / Data source
- One (of several) ways to store data in electronic format
- Used in everyday life: bank, hotel reservations, library search, shopping

Databases / Data sources

- Database management system (DBMS): a collection of programs to create and maintain a database
- Database system = database + DBMS

Databases / Data sources





What information is stored?

- Model the information
 - Entity-Relationship model (ER)
 - Unified Modeling Language (UML)




What information is stored? - ER

- entities and attributes
- entity types
- key attributes
- relationships
- cardinality constraints

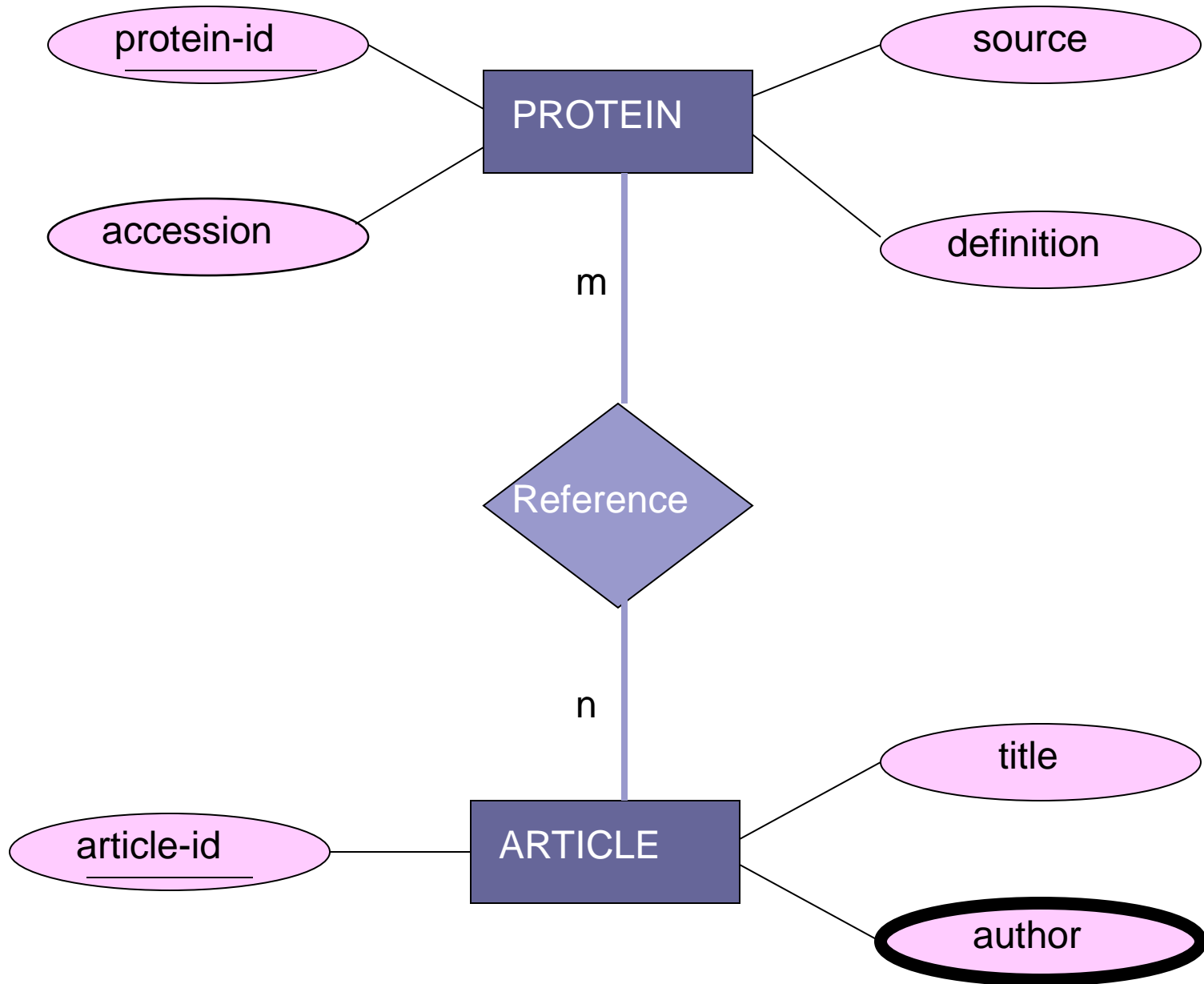
- EER: sub-types

1 tgctacccgc gcccgggctt ctggggtgtt cccaaccac ggcccagccc tgccacaccc
 61 cccgcccccg gcctccgcag ctcgcatgg gcgcgggggt gctcgtctg ggcgcctcg
 121 agccccgtaa cctgtcgtcg gccgcaccgc tccccgacgg cgcggccacc ggcgcgcggc
 181 tgctggtgcc cgcgtcgccg cccgcctcgt tgctgcctcc cgccagcgaa agccccgagc
 241 cgctgtctca gcagtggaca gcgggcatgg gtctgtgat ggcgtcatc gtgtgtctca
 301 tcgtggcggg caatgtgctg gtgatcgtgg ccatcgccaa gacgccgcgg ctgcagacgc
 361 tcaccaacct ctcatcatg tccctggcca gcgccgacct ggtcatgggg ctgctggtg
 421 tgccgttcgg ggccaccatc tgggtgtggg gccgctggga gtacggctcc ttcttctcg
 481 agctgtggac ctacgtggac gtgtgtgctg tgacggccag catcgagacc ctgtgtgtca
 541 ttgccctgga ccgtacctc gccatcacct cgcccttcg ctaccagagc ctgtgacgc
 601 gcgcgcgggc gcggggcctc gtgtgcaccg tgtgggcat ctgggccctg gtgtccttc
 661 tgccatcct catgactgg tggcgggcgg agagcgacga ggcgcgccgc tgctacaacg
 721 accccaagt ctgcgacttc gtcaccaacc gggcctacgc catcgctcgt tccgtagtct
 781 ccttctacgt gccctgtgc atcatggcct tcgtgtacct gcgggtgttc cgcgagggcc
 841 agaagcaggt gaagaagatc gacagctgc agcgcggtt ctcgggcggc ccagcgcggc
 901 cgccctcgcc ctgcacctc cccgtccccg cgccgcgcc gccgcccga ccccgccgc
 961 ccgcccgcgc cgccgccacc gcccgcctgg ccaacgggcg tgcgggtaag cggcgccct
 1021 cgcgcctcgt ggccctacgc gagcagaagg cgctcaagac gctgggcatc atcatgggcg
 1081 tcttcacgt ctgtggctg cccttcttc tggccaact ggtgaaggcc ttccaccgcg
 1141 agctggtgcc cgaccgctc ttgtcttct tcaactggct gggctacgcc aactcggcct
 1201 tcaaccccat catctactgc cgcagccccg acttcgcaa ggcttccag ggactgctt
 1261 gctgcgcgcg cagggctgcc cgccggcgcc acgcgacca cggagaccgg ccgcgcgct
 1321 cgggctgtct ggccccggcc ggacccccgc catcgcccg ggccgcctcg gacgacgag
 1381 acgacgatgt cgtcggggcc acgcccccgc cgcgcctgct ggagccctgg gccggctgca
 1441 acggcggggc ggcggcgac agcgactcga gcttgacga gccgtgccgc cccggcttcg
 1501 cctcggaatc caaggtgtag ggccggcgcc ggggcgcgga ctccgggcac ggcttccag
 1561 gggaacgagg agatctgtgt ttacttaaga ccgatagcag gtgaactcga agcccacaat
 1621 cctcgtctga atcatccgag gcaaagagaa aagccacgga ccgtgcaca aaaaggaaag
 1681 ttgggaagg gatgggagag tggctgctg atgtccttg ttg

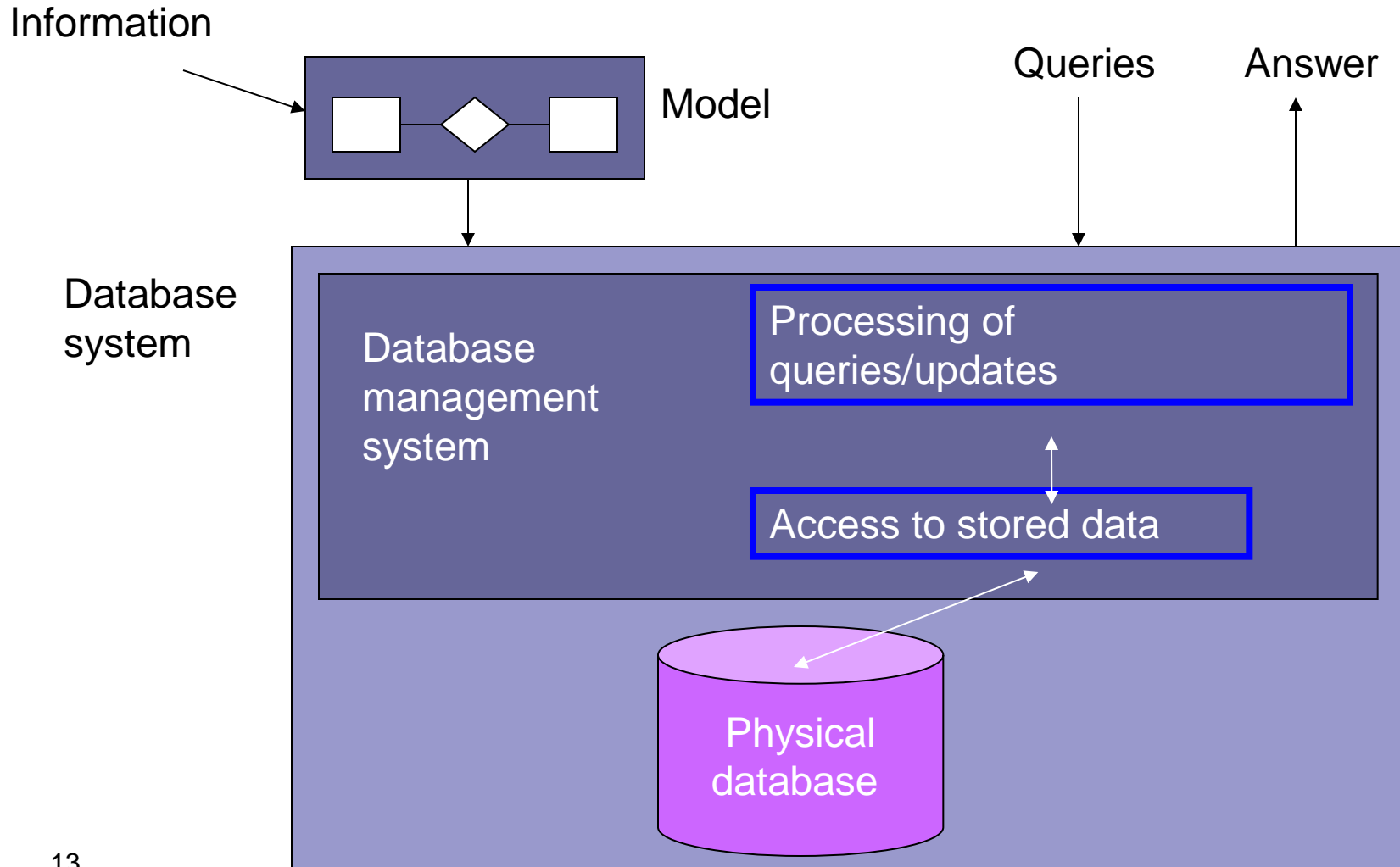


DEFINITION	Homo sapiens adrenergic, beta-1-, receptor
ACCESSION	NM_000684
SOURCE ORGANISM	human
REFERENCE	1
AUTHORS	Frielle, Collins, Daniel, Caron, Lefkowitz, Kobilka
TITLE	Cloning of the cDNA for the human beta 1-adrenergic receptor
REFERENCE	2
AUTHORS	Frielle, Kobilka, Lefkowitz, Caron
TITLE	Human beta 1- and beta 2-adrenergic receptors: structurally and functionally related receptors derived from distinct genes

Entity-relationship



Databases / Data sources

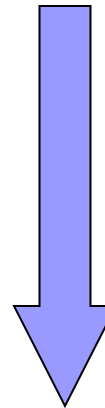


How is the information stored?
(high level)

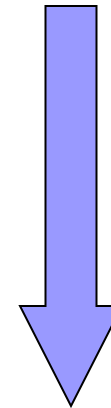
How is the information accessed?
(user level)

- Text (IR)
- Semi-structured data
- Data models (DB)
- Rules + Facts (KB)

structure



precision





IR - formal characterization

Information retrieval model: (D, Q, F, R)

- D is a set of document representations
- Q is a set of queries
- F is a framework for modeling document representations, queries and their relationships
- R associates a real number to document-query-pairs (ranking)

IR - Boolean model

	adrenergic	cloning	receptor	
Doc1	yes	yes	no	--> (1 1 0)
Doc2	no	yes	no	--> (0 1 0)

Q1: cloning and (adrenergic or receptor)

--> (1 1 0) or (1 1 1) or (0 1 1)

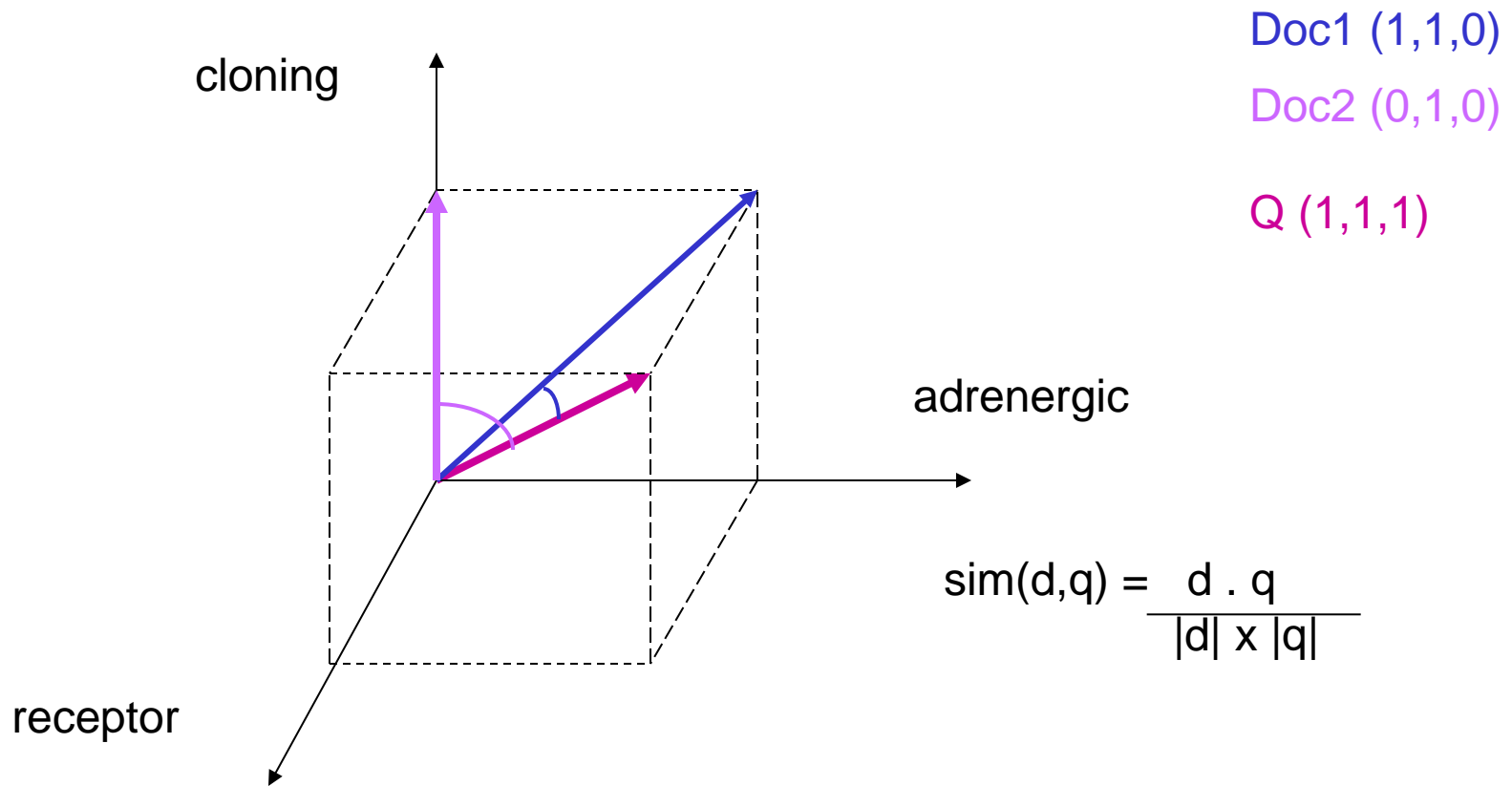
Result: Doc1

Q2: cloning and not adrenergic

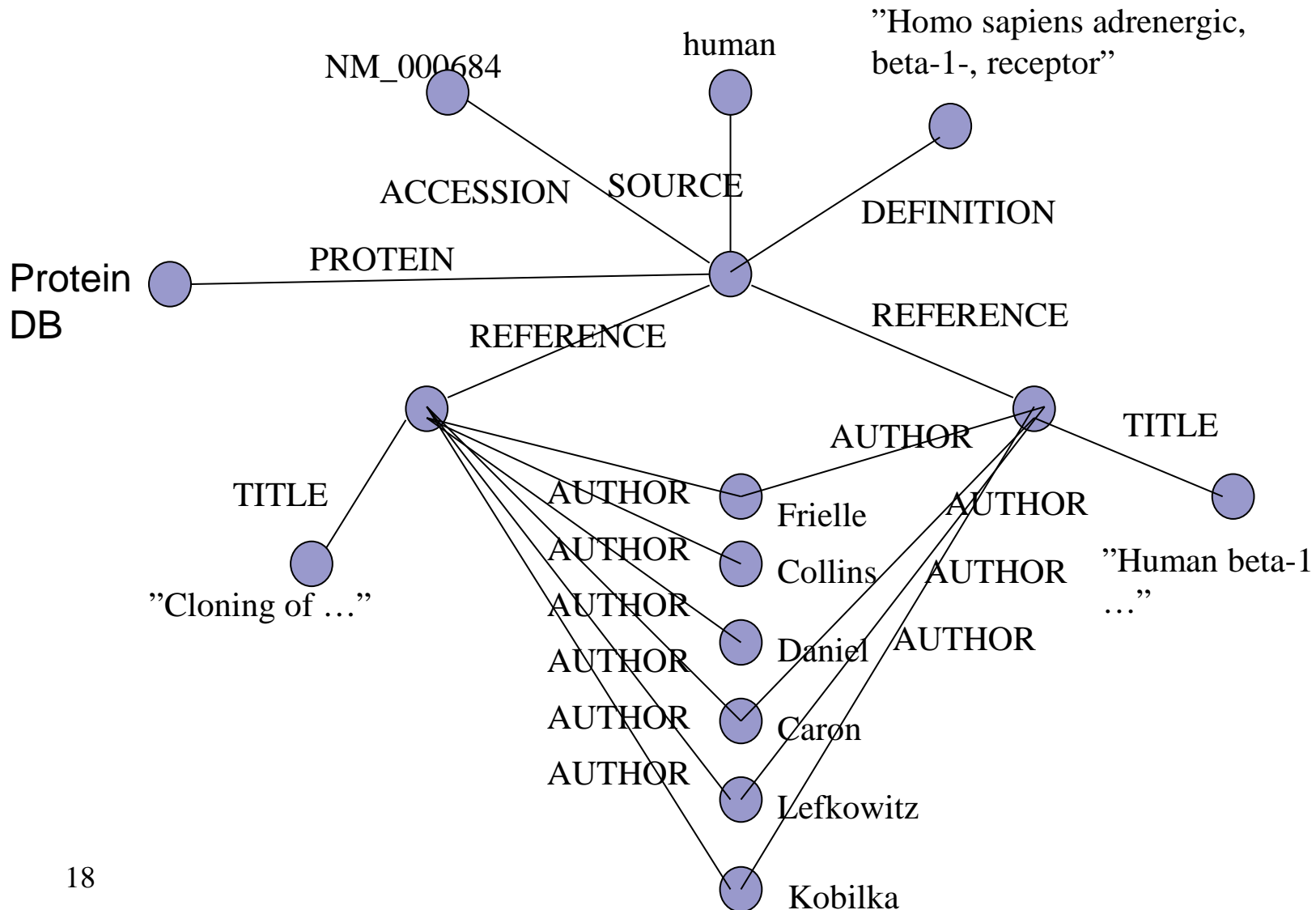
--> (0 1 0) or (0 1 1)

Result: Doc2

IR - Vector model (simplified)



Semi-structured data





Semi-structured data - Queries

```
select source  
from PROTEINDB.protein P  
where P.accession = "NM_000684";
```

Relational databases

PROTEIN

PROTEIN-ID	ACCESSION	DEFINITION	SOURCE
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human

REFERENCE

PROTEIN-ID	ARTICLE-ID
1	1
1	2

ARTICLE-AUTHOR

ARTICLE-ID	AUTHOR
1	Frielle
1	Collins
1	Daniel
1	Caron
1	Lefkowitz
1	Kobilka
2	Frielle
2	Kobilka
2	Lefkowitz
2	Caron

ARTICLE-TITLE

ARTICLE-ID	TITLE
1	Cloning of the cDNA for the human beta 1-adrenergic receptor
2	Human beta 1- and beta 2- adrenergic receptors: structurally and functionally related receptors derived from distinct genes

Relational databases - SQL

```
select source  
from protein  
where accession = NM_000684;
```

PROTEIN

PROTEIN-ID	ACCESSION	DEFINITION	SOURCE
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human

Evolution of Database Technology

- 1960s:
 - Data collection, database creation, IMS and network DBMS
- 1970s:
 - Relational data model, relational DBMS implementation
- 1980s:
 - Advanced data models (extended-relational, OO, deductive, etc.)
 - Application-oriented DBMS (spatial, temporal, multimedia, etc.)
- 1990s:
 - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
 - Stream data management and mining
 - Data mining and its applications
 - Web technology (XML, data integration) and global information systems
 - NoSQL databases

Knowledge bases

(F) source(NM_000684, Human)

(R) source(P?, Human) \Rightarrow source(P?, Mammal)

(R) source(P?, Mammal) \Rightarrow source(P?, Vertebrate)

Q: ?- source(NM_000684, Vertebrate)

A: yes

Q: ?- source(x?, Mammal)

A: x? = NM_000684

Interested in more?

- 732A57 Database Technology
(relational databases)
- TDDD43 Advanced data models and databases
(IR, semi-structured data, DB, KB)
- 732A47 Text mining
(includes IR)



Analytics



Analytics

- Discovery, interpretation and communication of meaningful patterns in data



Analytics - IBM

- What is happening? Descriptive
Discovery and explanation
- Why did it happen? Diagnostic
Reporting, analysis, content analytics
- What could happen? Predictive
Predictive analytics and modeling
- What action should I take? Prescriptive
Decision management
- What did I learn, what is best? Cognitive



Analytics - Oracle

- Classification
- Regression
- Clustering
- Attribute importance
- Anomaly detection
- Feature extraction and creation
- Market basket analysis

Why Analytics?

- The Explosive Growth of Data
 - Data collection and data availability
 - Automated data collection tools, database systems, Web, computerized society
 - Major sources of abundant data
 - Business: Web, e-commerce, transactions, stocks, ...
 - Science: Remote sensing, bioinformatics, scientific simulation, ...
 - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!

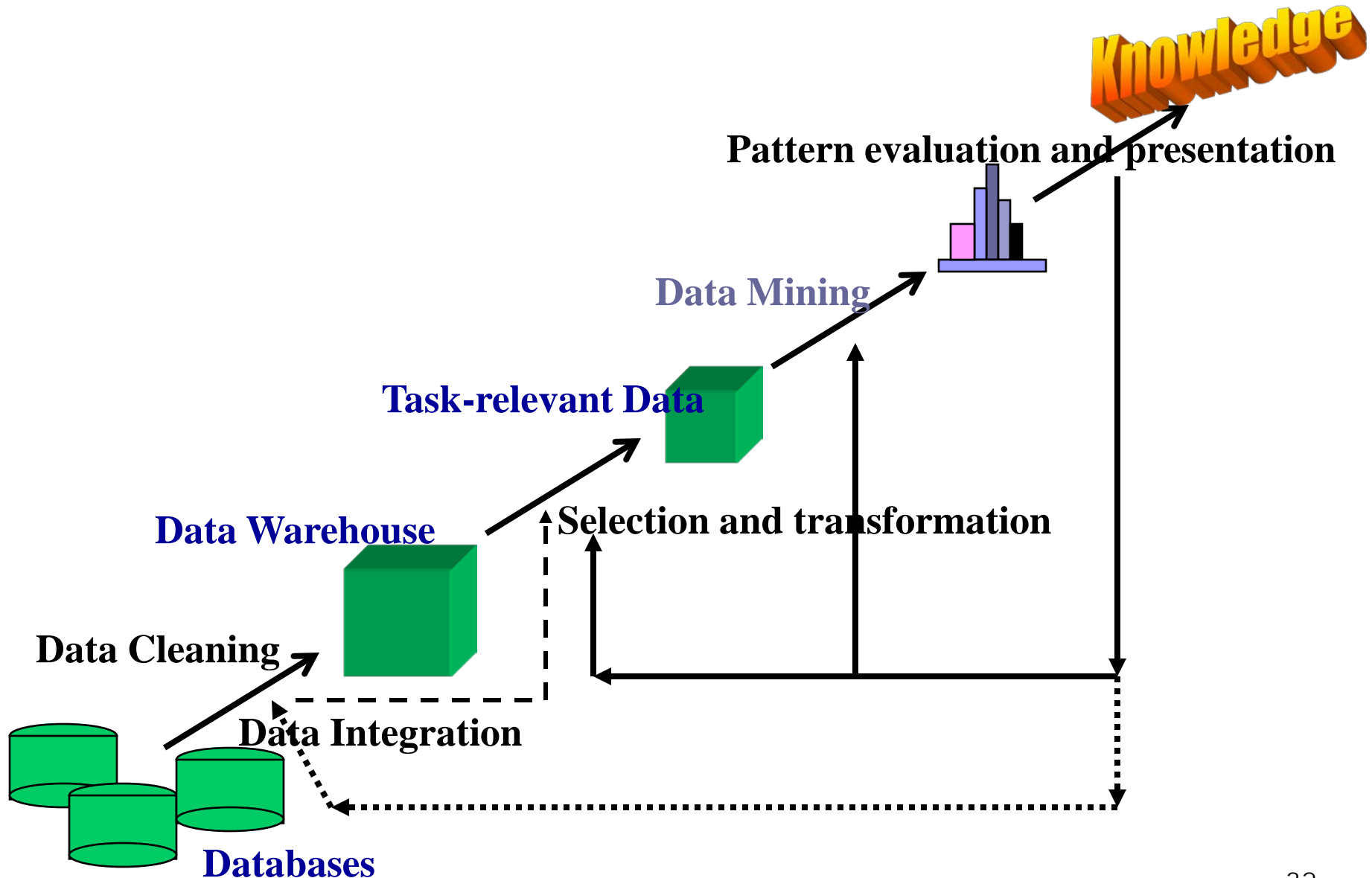
Ex.: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
 - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
 - Determine customer purchasing patterns over time
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
- Customer profiling—What types of customers buy what products (clustering or classification)
- Customer requirement analysis
 - Identify the best products for different groups of customers
 - Predict what factors will attract new customers
- Provision of summary information
 - Multidimensional summary reports
 - Statistical summary information (data central tendency and variation)

Ex.: Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
 - Auto insurance: ring of collisions
 - Money laundering: suspicious monetary transactions
 - Medical insurance
 - Professional patients, ring of doctors, and ring of references
 - Unnecessary or correlated screening tests
 - Telecommunications: phone-call fraud
 - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
 - Anti-terrorism

Knowledge Discovery (KDD) Process





Data Mining: Classification Schemes

- General functionality
 - Descriptive data mining
 - Predictive data mining

Data Mining – what kinds of patterns?

- Concept/class description:

- Characterization: summarizing the data of the class under study in general terms

- E.g. Characteristics of customers spending more than 10000 sek per year

- Discrimination: comparing target class with other (contrasting) classes

- E.g. Compare the characteristics of products that had a sales increase to products that had a sales decrease last year

Data Mining – what kinds of patterns?

- Frequent patterns, association, correlations

- Frequent itemset
- Frequent sequential pattern
- Frequent structured pattern

- E.g. $\text{buy}(X, \text{"Diaper"}) \rightarrow \text{buy}(X, \text{"Beer"})$ [support=0.5%, confidence=75%]

confidence: if X buys a diaper, then there is 75% chance that X buys beer

support: of all transactions under consideration 0.5% showed that diaper and beer were bought together

- E.g. $\text{Age}(X, \text{"20..29"})$ and $\text{income}(X, \text{"20k..29k"}) \rightarrow \text{buys}(X, \text{"cd-player"})$ [support=2%, confidence=60%]

Data Mining – what kinds of patterns?

- Classification and prediction

- Construct models (functions) that describe and distinguish classes or concepts for future prediction.

The derived model is based on analyzing training data – data whose class labels are known.

- E.g., classify countries based on (climate), or classify cars based on (gas mileage)
 - Predict some unknown or missing numerical values

Data Mining – what kinds of patterns?

- Cluster analysis
 - Class label is unknown: Group data to form new classes, e.g., cluster customers to find target groups for marketing
 - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
 - Outlier: Data object that does not comply with the general behavior of the data
 - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
 - Trend and deviation



Interested in more?

- 732A95 Introduction to machine learning
- 732A61 Data mining – clustering and association analysis



Big Data



Big Data

- So large data that it becomes difficult to process it using a 'traditional' system



Big Data – 3Vs

- Volume

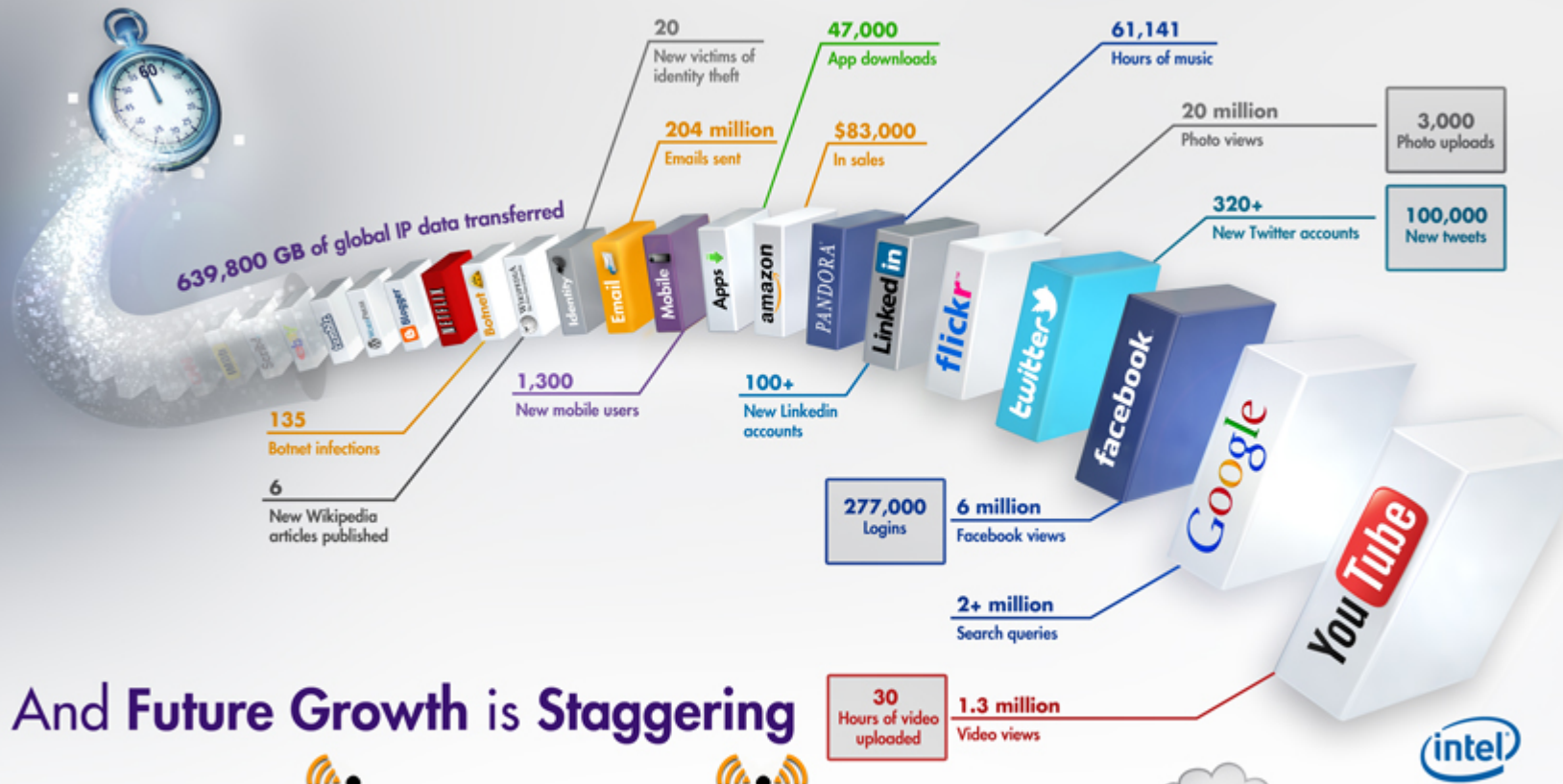
- size of the data



Volume - examples

- Facebook processes 500 TB per day
- Walmart handles 1 million customer transaction per hour
- Airbus generates 640 TB in one flight (10 TB per 30 minutes)
- 72 hours of video uploaded to youtube every minute
- SMS, e-mail, internet, social media

What Happens in an Internet Minute?



And Future Growth is Staggering



<https://y2socialcomputing.files.wordpress.com/2012/06/>

social-media-visual-last-blog-post-what-happens-in-an-internet-minute-infographic.jpg



Big Data – 3Vs

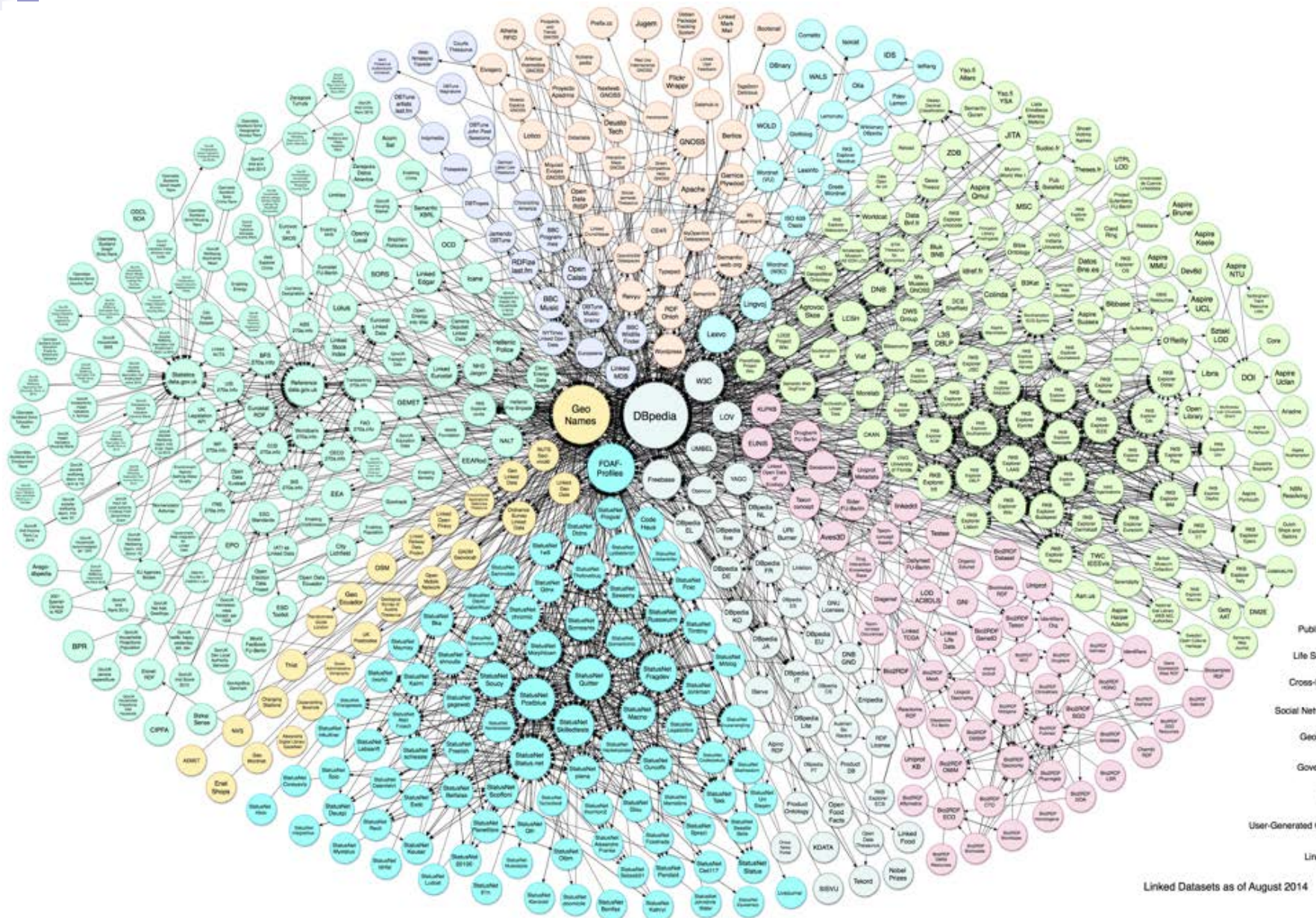
- Volume

- size of the data

- Variety

- type and nature of the data

- text, semi-structured data, databases, knowledge bases



Linked Datasets as of August 2014



Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak.
<http://lod-cloud.net/>



Linked open data of US government

Format (# Datasets)

<http://catalog.data.gov/>

- HTML (27005)
- XML (24077)
- PDF (19628)
- CSV (10058)
- JSON (8948)
- RDF (6153)
- JPG (5419)
- WMS (5019)
- Excel (3389)
- WFS (2781)



Big Data – 3Vs

- Volume


- size of the data

- Variety

- type and nature of the data

- Velocity

- speed of generation and processing of data



Velocity - examples

- Traffic data
- Financial market
- Social networks

40 ZETTABYTES

[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005



Volume SCALE OF DATA

It's estimated that 2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]
of data are created each day



Most companies in the U.S. have at least 100 TERABYTES [100,000 GIGABYTES] of data stored

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data,
with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



30 BILLION
PIECES OF CONTENT
are shared on Facebook
every month



Variety DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be

420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS

4 BILLION+
HOURS OF VIDEO
are watched on
YouTube each month



400 MILLION TWEETS
are sent per day by about 200
million monthly active users



The New York Stock Exchange captures

1 TB OF TRADE
INFORMATION
during each trading session



Velocity ANALYSIS OF STREAMING DATA

Modern cars have close to 100 SENSORS that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION
NETWORK
CONNECTIONS

— almost 2.5 connections
per person on earth



1 IN 3 BUSINESS
LEADERS

don't trust the information
they use to make decisions



Poor data quality costs the US
economy around

\$3.1 TRILLION A YEAR



27% OF
RESPONDENTS

in one survey were unsure of
how much of their data was
inaccurate

Veracity UNCERTAINTY OF DATA

Big Data – other Vs

- Variability
 - inconsistency of the data
- Veracity
 - quality of the data
- Value
 - useful analysis results
- ...

BDA system architecture

Specialized
services
for domain A

Specialized
services
for domain B

Big Data Services Layer

Knowledge Management Layer

Data Storage and Management Layer

BigMecs





BDA system architecture

- ☐ Large amounts of data, distributed environment
- ☐ Unstructured and semi-structured data
- ☐ Not necessarily a schema
- ☐ Heterogeneous
- ☐ Streams
- ☐ Varying quality

Data Storage and Management Layer



Data Storage and management

– this course

■ Data storage:

- ☐ NoSQL databases
- ☐ OLTP vs OLAP
- ☐ Horizontal scalability
- ☐ Consistency, availability, partition tolerance

■ Data management

- ☐ Hadoop
- ☐ Data management systems



BDA system architecture

- ☐ Semantic technologies
- ☐ Integration
- ☐ Knowledge acquisition

Knowledge Management Layer



Knowledge management – this course

- Not a focus topic in this course
- For semantic and integration approaches see TDDD43



BDA system architecture

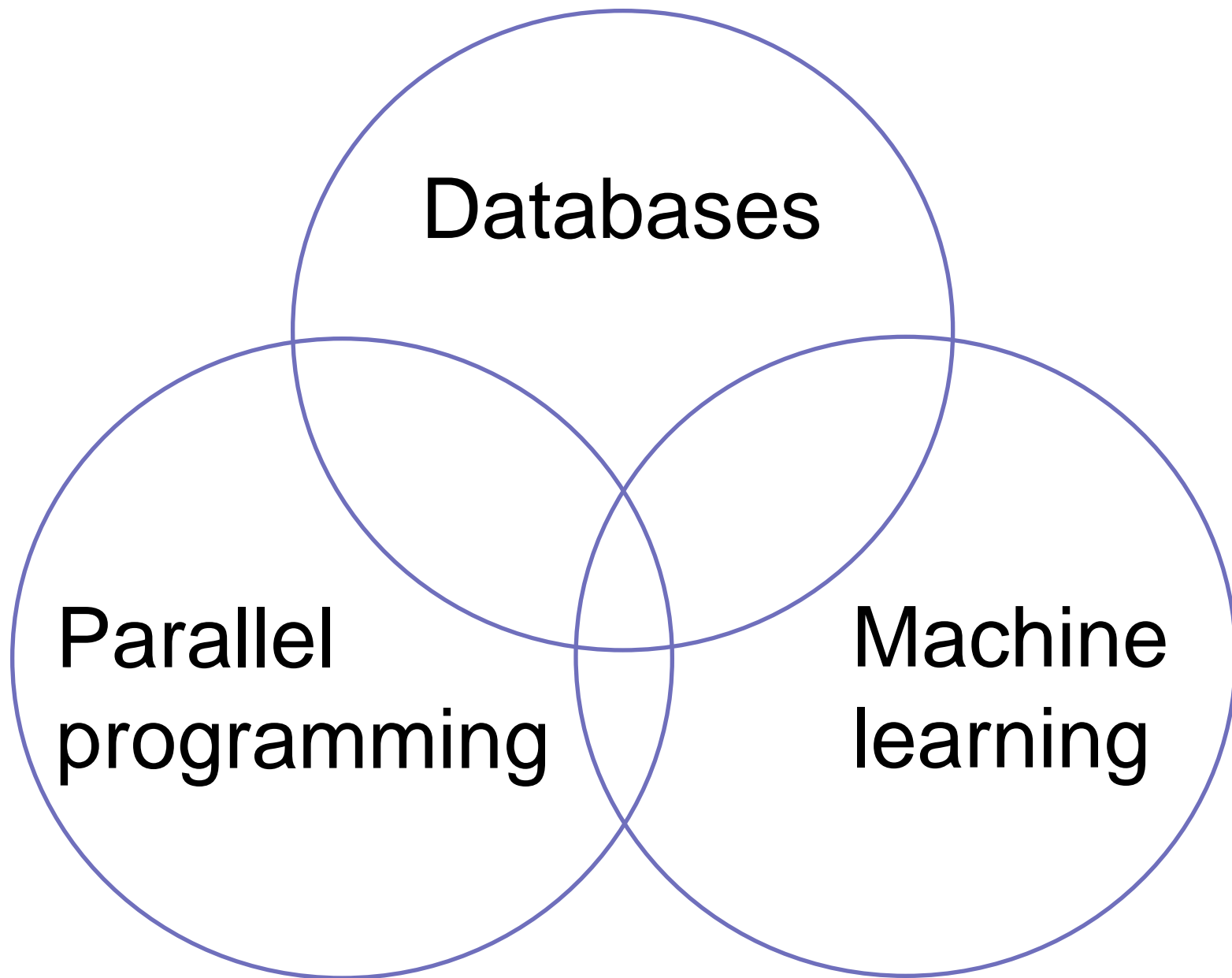
- Analytics services for Big Data

Big Data Services Layer



Big Data Services – this course

- Big data versions of analytics/data mining algorithms



- 
- 2016: course for SDM year 1 and year 2

→ some review/introduction lectures



Course overview

- Review
 - Databases (lectures + labs)
 - Python (lectures + labs)
- Databases for Big Data (lectures + lab)
- Parallel algorithms for processing Big Data (lectures + lab + exercise session)
- Machine Learning for Big Data (lectures + lab)
- Visit to National Supercomputer Centre



Info

- Results reported in connection to exams
- Info about handing in labs on web; strong recommendation to hand in as soon as possible
- Sign up for labs via web (in pairs)



Info

- Relational database labs require special database account
 - make sure you are registered for the course
- BDA labs require special access to NSC resources
 - fill out forms

Info

■ Lab deadlines:

- Final deadlines in connection to the exams;
no reporting between exams

- **HARD DEADLINE:** March exam

(No guarantee NSC resources available after April.)



Examination

- Written exam
- Labs



What if I already took ...?

What if I also take...?

- TDDD37/732A57 Database technology
 - RDB labs 1-2 in one of the courses, results registered for both
- 732A47 Text mining
 - Python labs in one of the courses, results registered for both



Changes w.r.t. last year

- New course

My own interest and research

- Modeling of data
 - Ontologies
- Ontology engineering
 - Ontology alignment
(Winner Anatomy track OAEI 2008 / Organizer OAEI tracks since 2013)
 - Ontology debugging
(Founder and organizer WoDOOM/CoDeS since 2012)
- Ontologies and databases for Big Data
- Former work: knowledge representation, data integration, knowledge-based information retrieval, object-centered databases
- <http://www.ida.liu.se/~patla00/research.shtml>



<https://www.youtube.com/watch?v=LrNIZ7-SMPk>