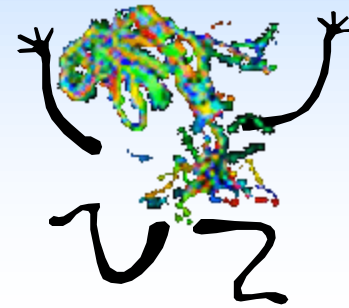


GET THAT PROTEIN!



Eller

TDDE49

Databaser och informationssäkerhet
för bioinformatik

<http://www.ida.liu.se/~TDDE49>

Lärare

- Examinator: Olaf Hartig
- FÖ: Olaf Hartig, Ulf Kargén, Patrick Lambrix
- LA: Sijin Cheng, Ulf Kargén
- projekt: Patrick Lambrix, Sijin Cheng
- databasadministration: Sijin Cheng

- studierektor: Patrick Lambrix

- Olaf/Sijin på engelska

Kurslitteratur

- Elmasri, R. and Navathe, S. B. Fundamentals of Database Systems, Addison Wesley.
- (Padron-McCarthy, webbkurs på Svenska)
- van Oorschot P: Computer Security and the Internet.
- Anderson R, Security Engineering.
- Adam Shostack: Threat Modeling.
- Lab + projekt: på hemsidan

Databaser

- Ett (av flera) sätt att lagra data i elektronisk format
- Används i det vardagliga livet: bank, bokning av hotell eller resa, sökning i biblioteket, handla
- nyare tillämpningar: multimediodatabaser, geografiska informationssystem, realtiddatabaser

Databaser

- databashanteringssystem (DBMS): en uppsättning program som tillåter en användare att skapa och underhålla en databas
- databassystem = databas + databashanteringssystem

Bioinformatik

- Kända sekvenser samlas i en stor databas. Insamlande och studier av sekvenser och jämförelser av sekvensernas uppbyggnad i olika organismer kallas bioinformatik. Forskningen inom bioinformatik är beroende av avancerad datalogi och matematik. (forskningsrådets strategidokument 2000)

Bioinformatik

- Bioinformatics: research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze or visualize data. (National Institutes of Health)

Bioinformatik

Ämnen på ISMB:

- protein structure and modeling
- sequence motifs, alignments and families
- networks and modeling
- gene structure, regulation and modeling
- sequence and phylogeny
- databases, information and knowledge management

TDDE49 Databaser och informationssäkerhet för Bioinformatik

- Denna kurs: fokus på biologiska databanker

Relation med andra kurser inom TB-programmet:

- förkunskaper: molekylärbiologi,
programmering
- bioinformatik

Årets ändringar i kursen

Ändringar 2022:

- Campus / Online (covid19)
- Ny lärare säkerhet

Ändringar 2021:

- Namnbyte
- Informationssäkerhet tillkommer
- Information retrieval, semi-strukturerad data, teoretiska delar av relationsdatabaser utgår.
- Nya labbar för databasdelen.
- Flipped classroom för databasföreläsningarna.

Årets ändringar i kursen

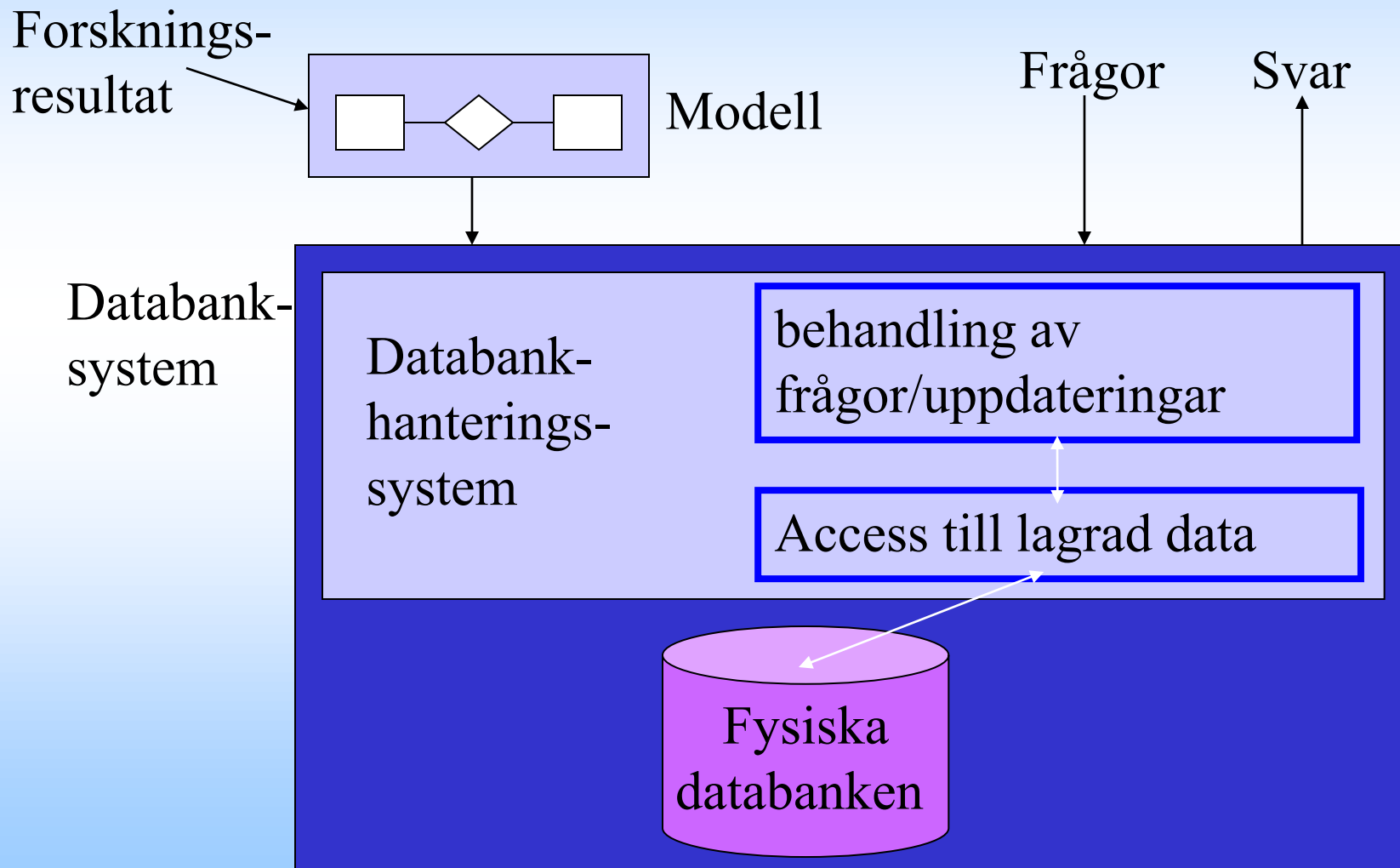
Tidigare ändringar:

- Separat inledning
- Genomgång projektbeskrivning
- Ändring inledning labbarna

Biologiska databanker

- biologisk data i elektronisk format
- exempel: SWISS-PROT/UniProt, EMBL, DDBJ, PDB, GENBANK, KEGG, ACEDB
- används dagligen i forskningen

Biologiska databanker



Frågeställningar

- Vilken information lagrar man?
- Hur lagras informationen? (hög och låg nivå)
- Hur accessar man informationen?
(användarnivå, systemnivå)
- Hur återställer man en databank efter crash?
- Hur kan flera användare accessa och uppdatera informationen samtidigt?

Personer

- databankadministratör
- databankdesigner
- användare ('end user')
- programmerare av tillämpningar

- DBMS designer
- utvecklare av verktyg
- operator, underhåll

1 tgetacccgc gcccgggctt ctgggggtgt cccaaccac ggcccagccc tgccacacc
61 cccgccccg gcctccgag ctggcatgg gcgcgggggt gctcgtctg ggcgcctccg
121 agcccgtaa cctgtctcg gccgaccgc tccccgacgg cgcggccacc gcggcgcggc
181 tgctgggtcc cegtcgccc cccgcctcgt tctgcctcc ccccagcga agccccgagc
241 cgtgtctca gcagtggaca ggggcatgg gtctgtgat ggcgctcacc gtgtgtca
301 tctggcggg caatgtgctg gtgatcgtg ccatcgcaa gacgccgcgg ctgcagaccg
361 tcaccaacct ctcatcatg tcctggcca gcgccacct ggtcatgggg ctgtgtgtg
421 tgccgttcg gccaccacc gtgggtggg gccgctggga gtacggctcc ttctctg
481 agctgtggac ctactggac gtgtgtgct gacggccag catcgagacc ctgtgtca
541 ttgccctga ccgtaccct gccatcacc ccccttcg ctaccagac ctgtgacc
601 gcgcgcgggc gcggggcctc gtgtgaccg tgtgggcat ctggccctg gtgtcctcc
661 tgccatcct catgactgg tggcggggc agagcgacga ggcgcgccg tctacaacg
721 acccaagtg ctgcactc gtaaccaacc ggcctaccg catgcctcg tccgtagt
781 cttctactg gccctgtg atcatggcct tctgtacct gcgggtgtc cgcgagccc
841 agaagcagg gaagaagac gacagctgc agcgcggtt cctcggcggc ccagcgcggc
901 cgcctcgc ctcgccctc cccgtcccc gcgccgcgc gccgccgga cccccgcgc
961 ccgccgcgc ccgccacc gcccgctgg ccaacgggcg tgcgggtaag cggcggcct
1021 cgcgcctcgt gccctaccg gagcagaagg cgctcaagac gctgggcatc atcatggcg
1081 tctcacgct ctgtggctg ccttctcc tggccaact ggtgaaggc tccaccgc
1141 agctggtgcc cgaccgctc ttgtctct tcaactggt gggctaccg aactggct
1201 tcaacccat catctactg cgcagcccc acttccgca gccctccag ggactgct
1261 gctgcgcgc cagggtgcc ccgccgcgc acgcacca cggagaccg ccgcgcct
1321 cgggtgtct ggcccggccc ggacccccg catgcccgg gccgcctcg gacgacgac
1381 acgacgatg cgtcggggc acgccccc gcgcctgct ggagccctg gccggctca
1441 acggcggggc ggcggcggc agcgactca gctggacga gccgtccc cccggttcg
1501 cctcggaat caaggttag ggcccgggc ggggcgcgga ctccgggac ggctcccag
1561 gggaacgag agatctgt ttaactaaga ccgatagc gtgaactca agcccaca
1621 cctcgtctga atcatccg gcaaagaga aagccacga ccgtgcaca aaaaggaa
1681 ttgggaagg gatgggag tggctgct atgtcctt 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

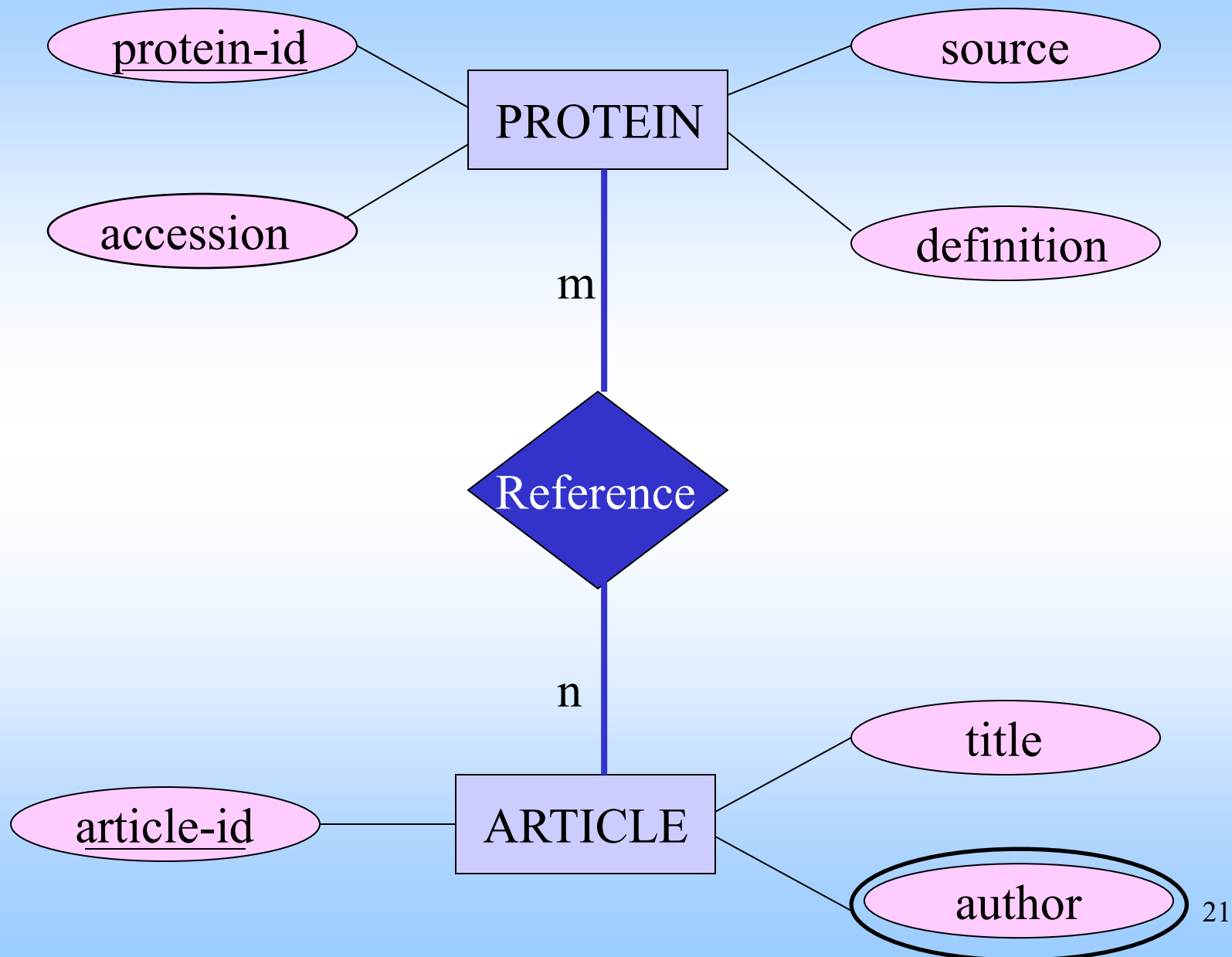
Vilken information lagrar man?

- Modell av verkligheten
 - Entity-Relationship modell (ER)
 - Unified Modeling Language (UML)

Entity-Relationship

- entiteter och attribut
- entitetstyper
- nyckelattribut
- relationer
- kardinalitetsvillkor

Entity-relationship



Hur lagras informationen? (hög nivå)

Hur accessar man informationen? (användarnivå)

- Text (IR)
- Semistrukturerad data
- Datamodeller (DB)
- Regler + Fakta (KB)

struktur



precision



Text - Information Retrieval

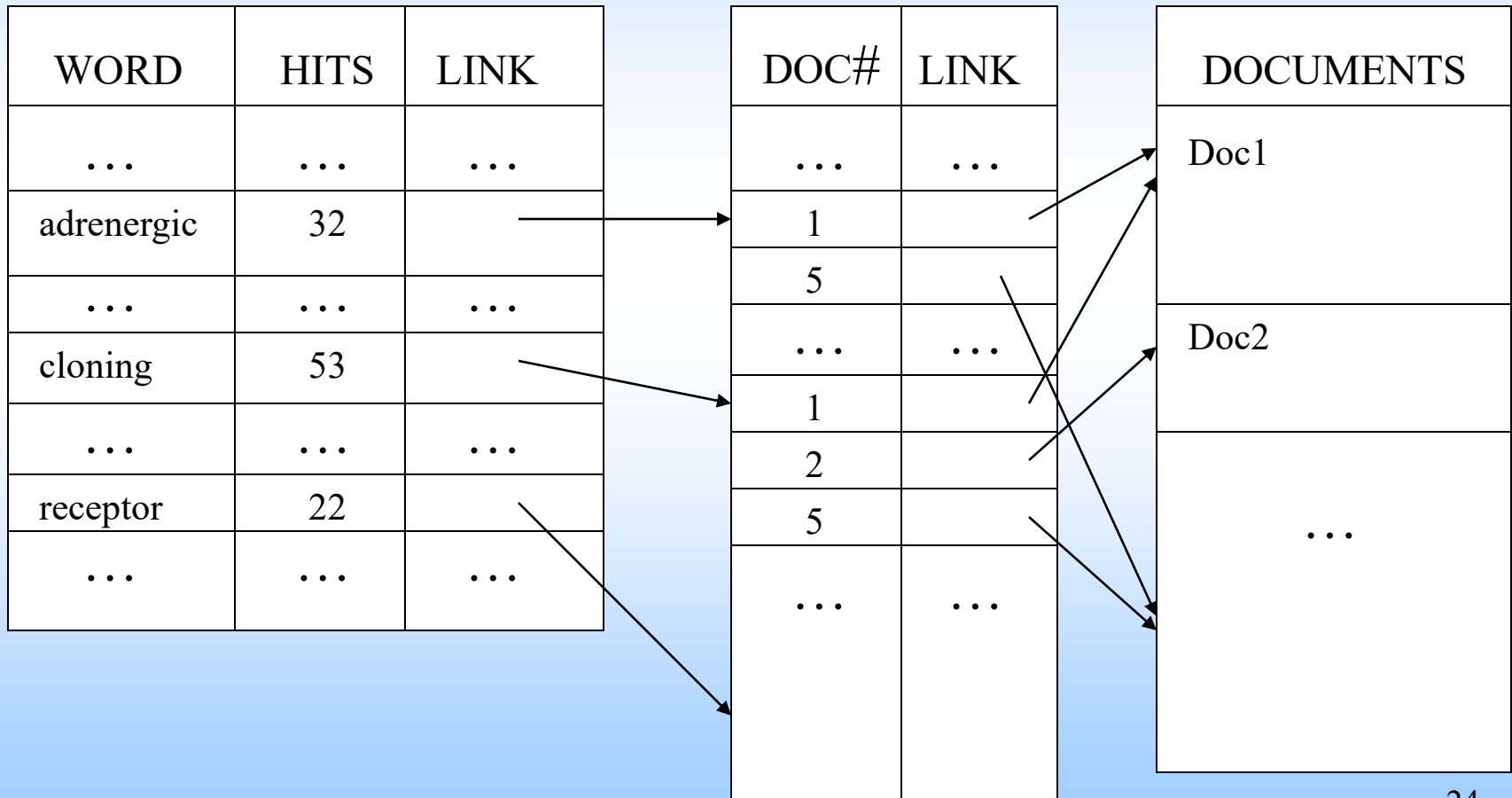
- sökning baseras på ord
- konceptuella modeller:
 - boolesk, vektor, probabilistisk, ...
- filmmodell:
 - flat fil, inverterad fil, ...

IR - Filmodell: inverterad fil

inverterad fil

anslagningsfil

dokumentfil

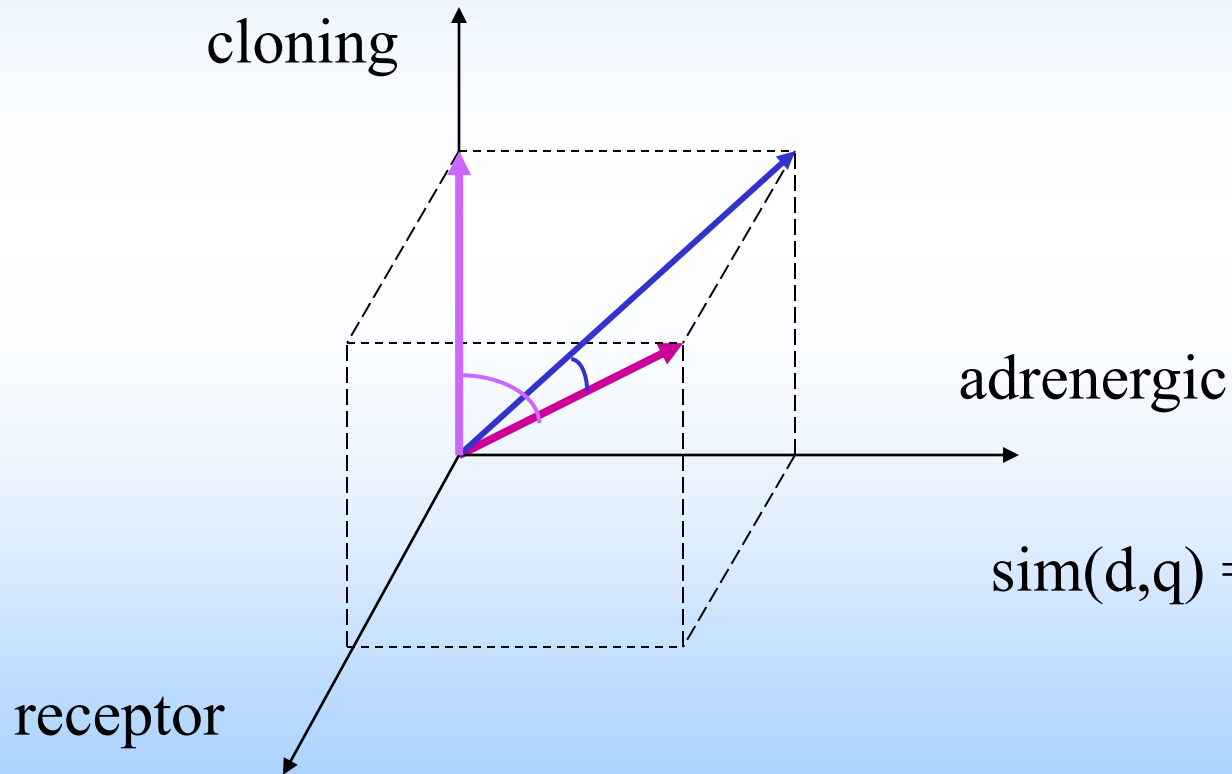


Vektormodellen (förenklad)

Doc1 (1,1,0)

Doc2 (0,1,0)

Q (1,1,1)



$$\text{sim}(d,q) = \frac{d \cdot q}{|d| \times |q|}$$

Databaser

- Relationsdatabaser:
 - modell: tabeller + relationsalgebran
 - frågespråk (SQL)
- Objektorienterade databaser:
 - modell: fortlevande objekt, meddelande, inkapsling, ärvning
 - frågespråk (t.ex. OQL)
- System: GDB (R), ACEDB (OO)

Relationsdatabaser

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

ARTICLE-ID	AUTHOR	TITLE
1	Frielle	Cloning of the cDNA for the human
1	Collins	Cloning of the cDNA for the human
1	Daniel	Cloning of the cDNA for the human
1	Caron	Cloning of the cDNA for the human
1	Lefkowitz	Cloning of the cDNA for the human
1	Kobilka	Cloning of the cDNA for the human
2	Frielle	Human beta 1- and beta 2-adrenergic receptors
2	Kobilka	Human beta 1- and beta 2-adrenergic receptors
2	Lefkowitz	Human beta 1- and beta 2-adrenergic receptors
2	Caron	Human beta 1- and beta 2-adrenergic receptors

Relationsdatabaser

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

SQL

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

Vilka kolumner?
Vilka tabeller?
Vilka rader?

PROTEIN

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

SQL

```
select title
from protein, article-title, reference
where protein.accession = NM_000684
and protein.protein-id
      = reference.protein-id
and reference.article-id
      = article-title.article-id;
```

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-TITLE

ARTICLE-ID	TITLE
1	Cloning of the ...
2	Human beta 1- ...

SQL

```
select title
from protein, article-title, reference
where protein.accession = NM_000684
and protein.protein-id
      = reference.protein-id
and reference.article-id
      = article-title.article-id;
```

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-TITLE

ARTICLE-ID	TITLE
1	Cloning of the ...
2	Human beta 1- ...

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

SQL

```
select title
from protein, article-title, reference
where protein.accession = NM_000684
and protein.protein-id
    = reference.protein-id
and reference.article-id
    = article-title.article-id;
```

REFERENCE

PROTEIN-ID	ARTICLE-ID
1	1
1	2

PROTEIN

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

ARTICLE-TITLE

ARTICLE-ID	TITLE
1	Cloning of the ...
2	Human beta 1- ...

PROTEIN

REFERENCE

PROTEIN-ID	ACCESSION	DEFINITION	SOURCE	PROTEIN-ID	ARTICLE-ID
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human	1	1
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human	1	2

SQL

```
select title
from protein, article-title, reference
where protein.accession = NM_000684
and protein.protein-id
      = reference.protein-id
and reference.article-id
      = article-title.article-id;
```

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-TITLE

ARTICLE-ID	TITLE
1	Cloning of the ...
2	Human beta 1- ...

PROTEIN-ID	ACCESSION	DEFINITION	SOURCE	REFERENCE		ARTICLE-TITLE
				PROTEIN-ID	ARTICLE-ID	
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human	1	1	Cloning of the ...
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human	1	2	Human beta 1- ...

SQL

```
select title  
from protein, article-title, reference  
where protein.accession = NM_000684  
and protein.protein-id  
      = reference.protein-id  
and reference.article-id  
      = article-title.article-id;
```

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-TITLE

ARTICLE-ID	TITLE
1	Cloning of the ...
2	Human beta 1- ...

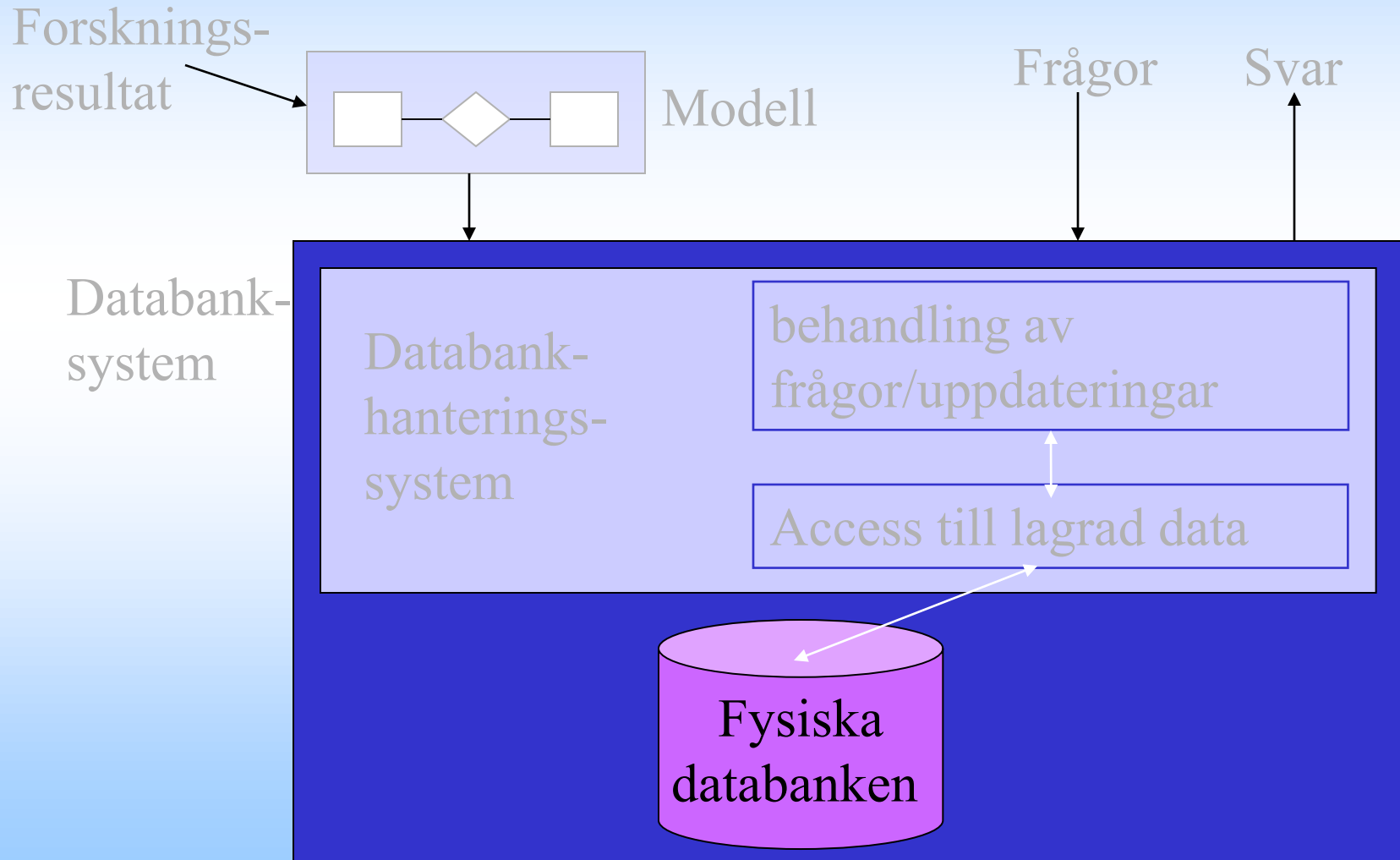
PROTEIN-ID	ACCESSION	DEFINITION	SOURCE	REFERENCE ARTICLE-TITLE			
				PROTEIN-ID	ARTICLE-ID	ARTICLE-ID	TITLE
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human	1	1	1	Cloning of the ...
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human	1	2	2	Human beta 1- ...

TITLE

Cloning of the ...

Human beta 1- ...

Hur lagras informationen? (låg nivå)



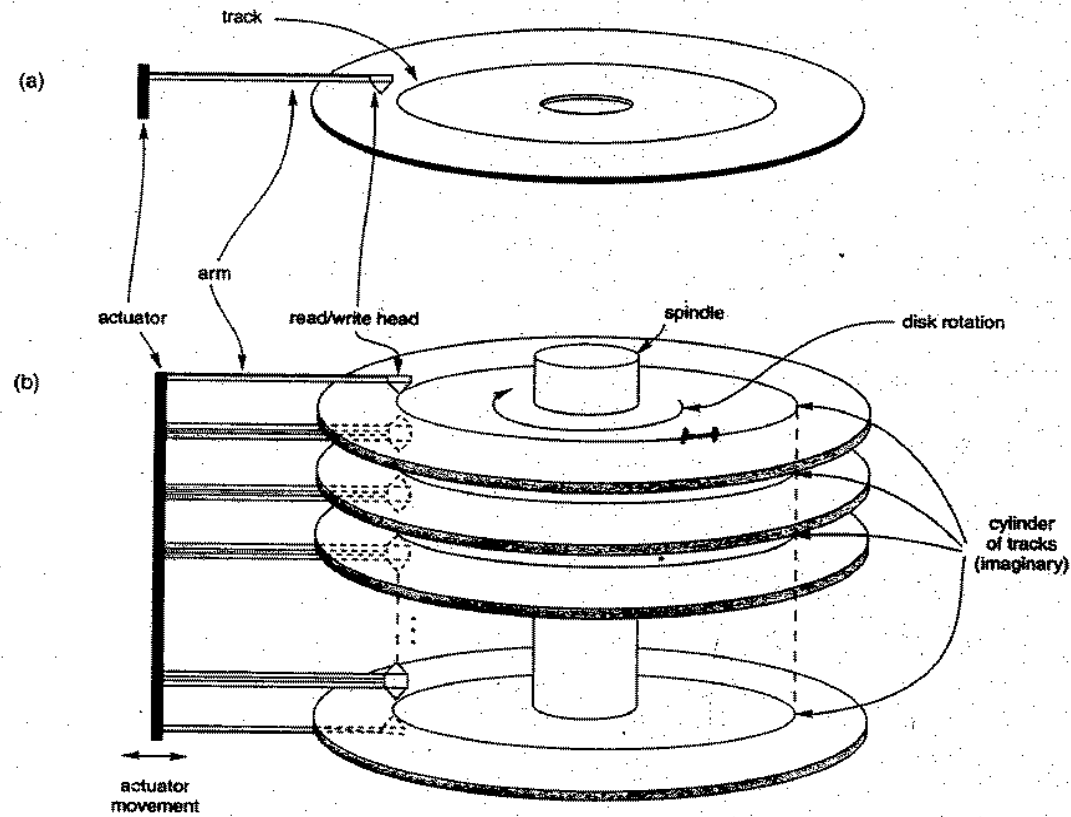
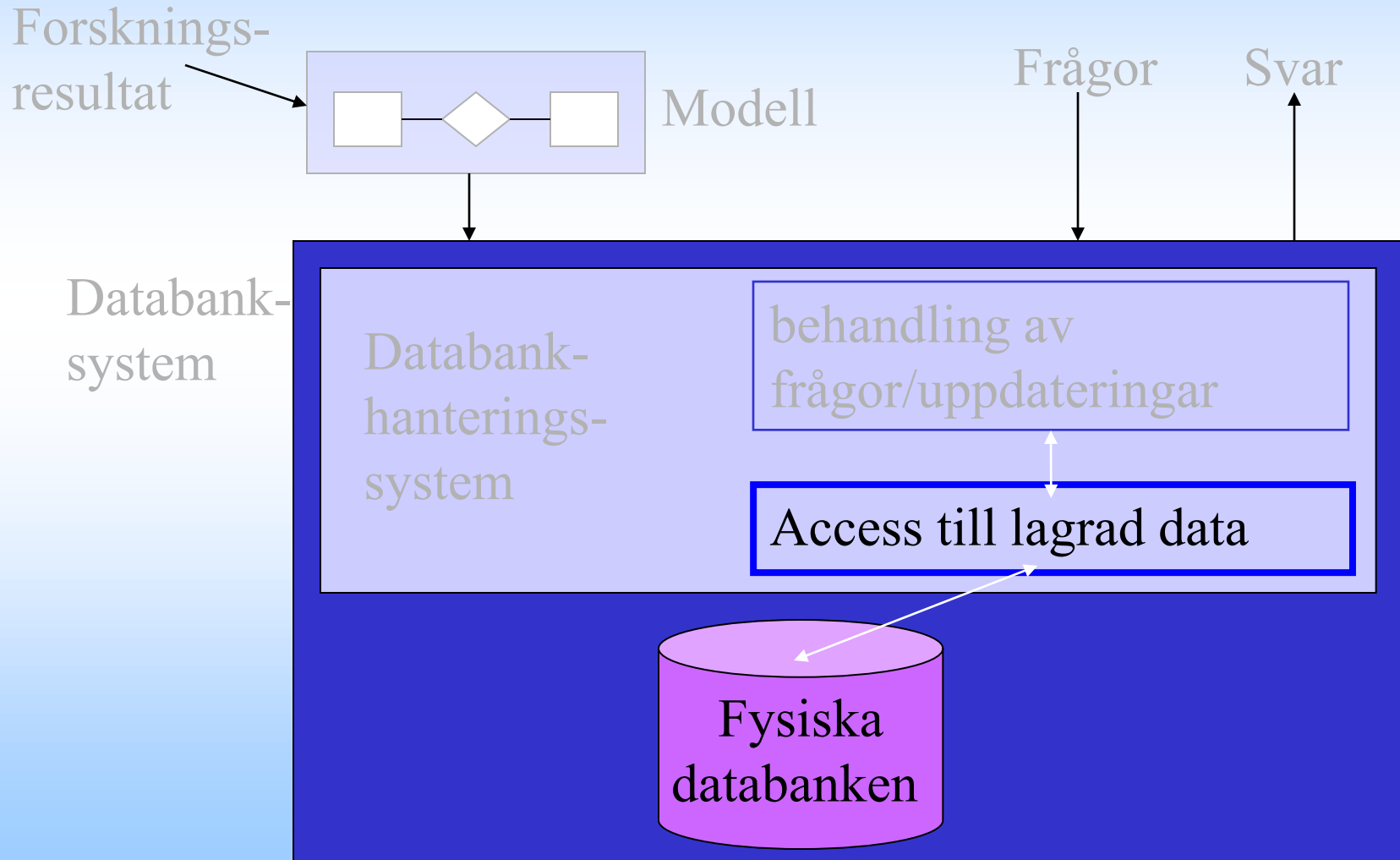


Figure 5.1 (a) A single-sided disk with read/write hardware. (b) A disk pack with read/write hardware.

Hur accessar man informationen? (systemnivå)

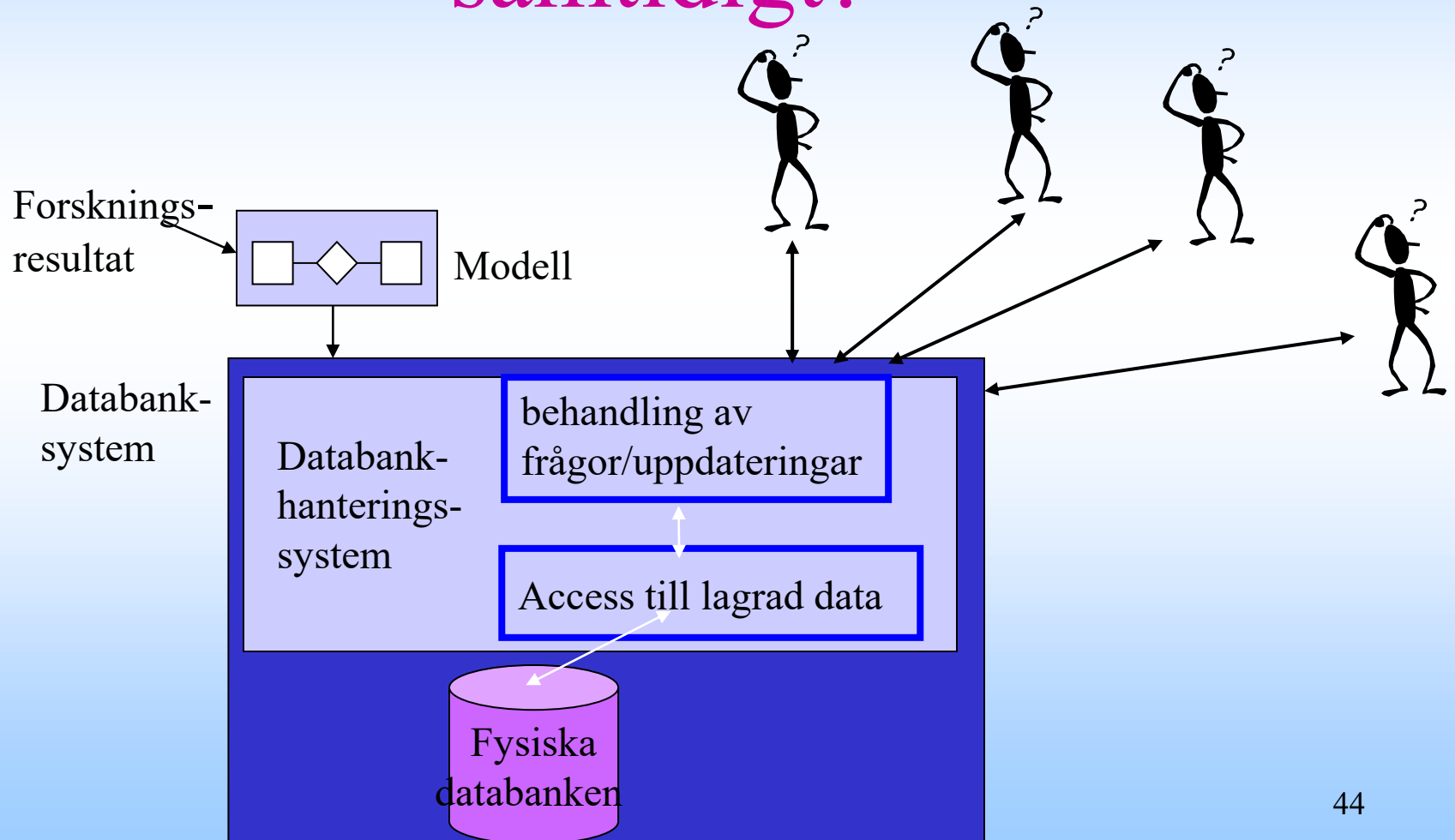


Hur återställer man en databank efter crash?

Återställning vid

- datorstop (system crash)
- systemfel
- samtidigthetsfel (flera användare)
- skivfel
- katastrofer

Hur kan flera användare accessa och uppdatera informationen samtidigt?



Flera användare

Administratör 1

TID

Administratör 2

Read(Antal-proteiner)

Antal-proteiner =
Antal-proteiner + 30

Write(Antal-proteiner)

Read(Antal-proteiner)

Antal-proteiner =
Antal-proteiner + 25

Write(Antal-proteiner)

Flera användare

DB



Antal-proteiner: 150

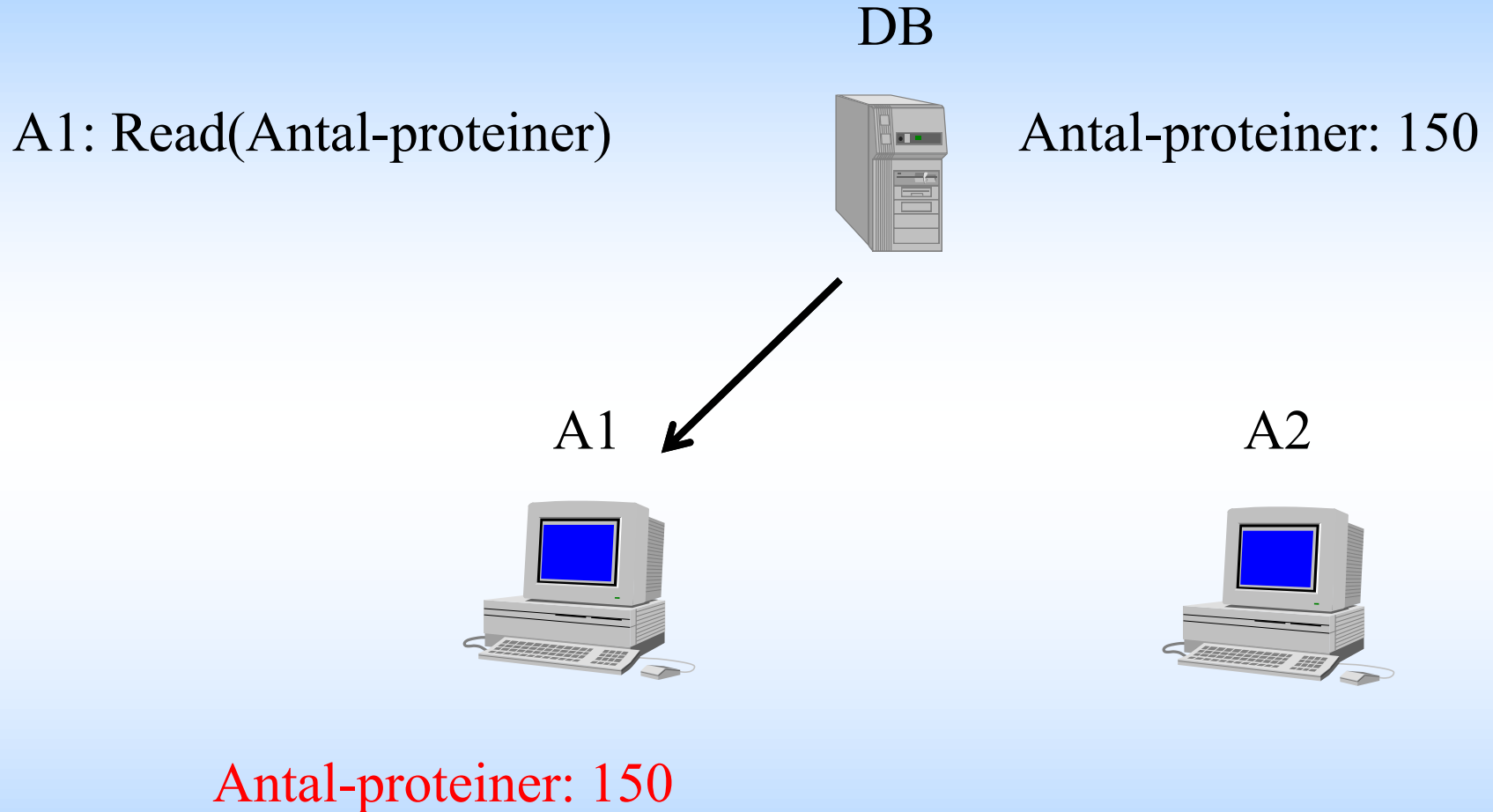
A1



A2



Flera användare



Flera användare

DB

A1: Antal-proteiner =
Antal-proteiner + 30

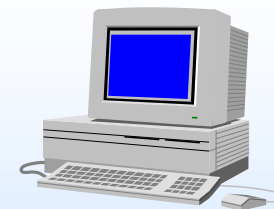


Antal-proteiner: 150

A1



A2



Antal-proteiner: 150 + 30

Flera användare

DB

A1: Antal-proteiner =
Antal-proteiner + 30

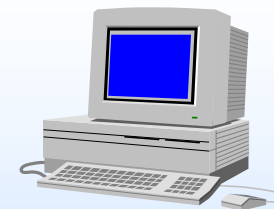


Antal-proteiner: 150

A1

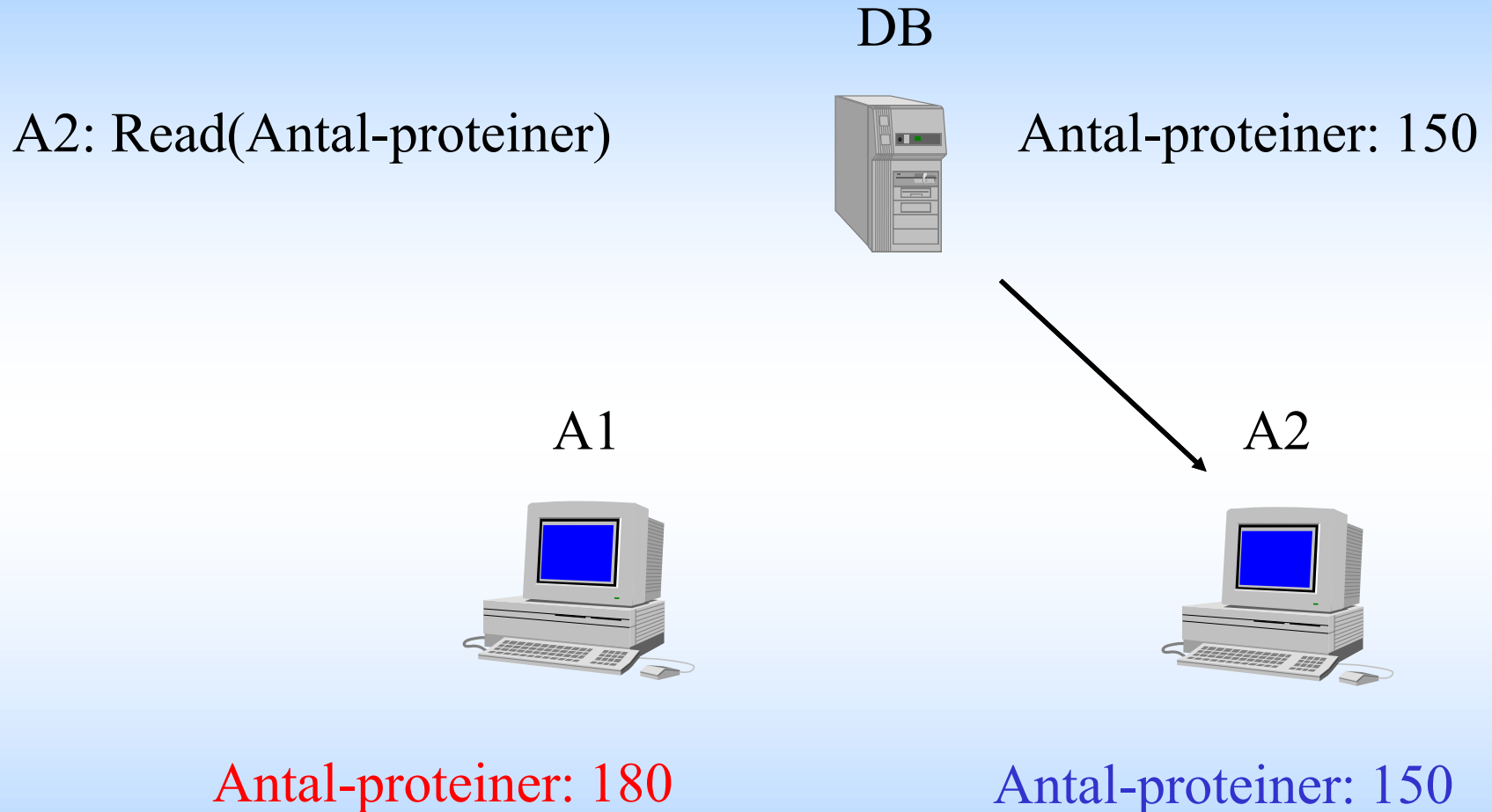


A2



Antal-proteiner: 180

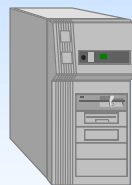
Flera användare



Flera användare

DB

A2: Antal-proteiner =
Antal-proteiner + 25



Antal-proteiner: 150

A1



Antal-proteiner: 180

A2

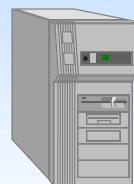


Antal-proteiner: 150 + 25

Flera användare

DB

A2: Antal-proteiner =
Antal-proteiner + 25



Antal-proteiner: 150

A1



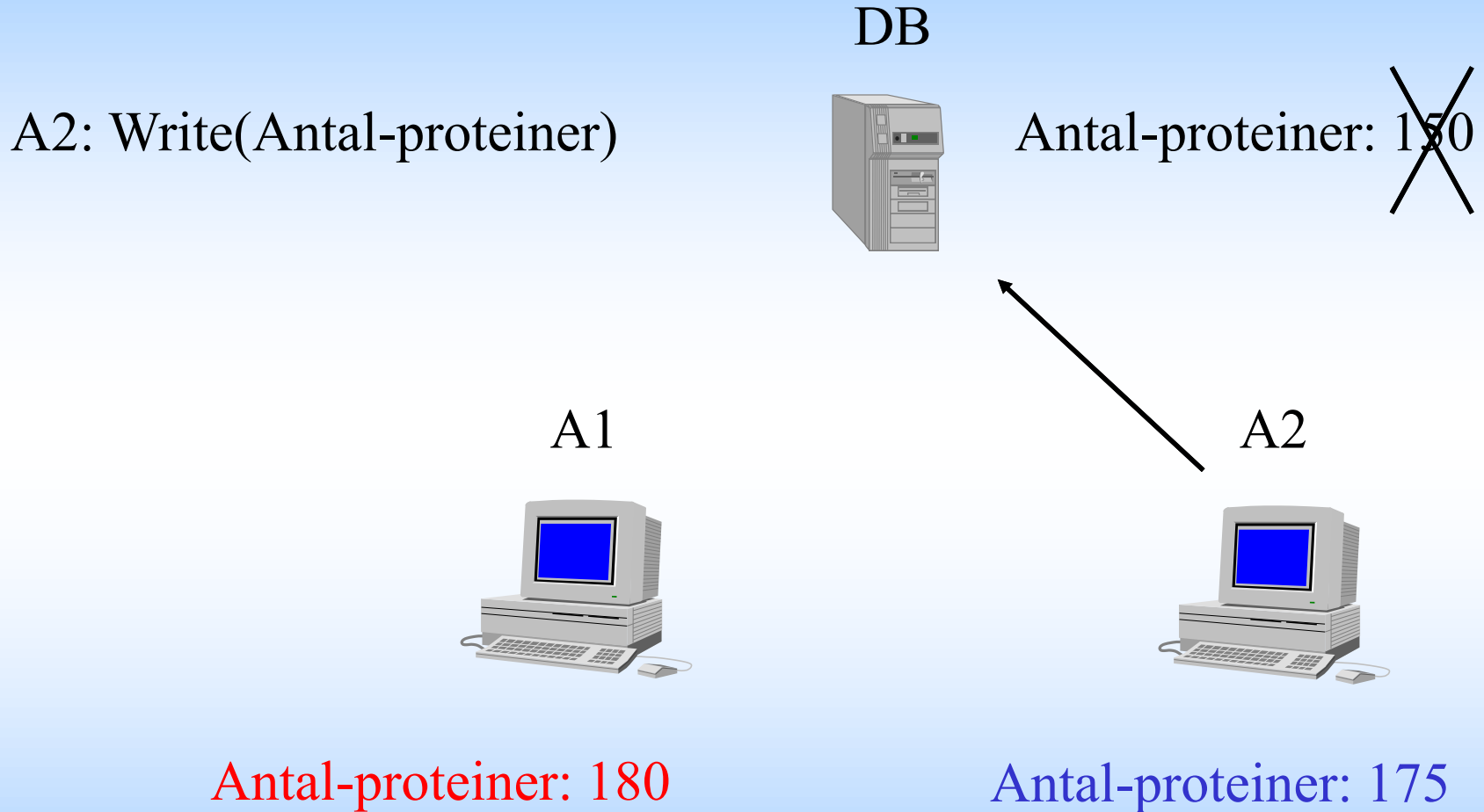
Antal-proteiner: 180

A2

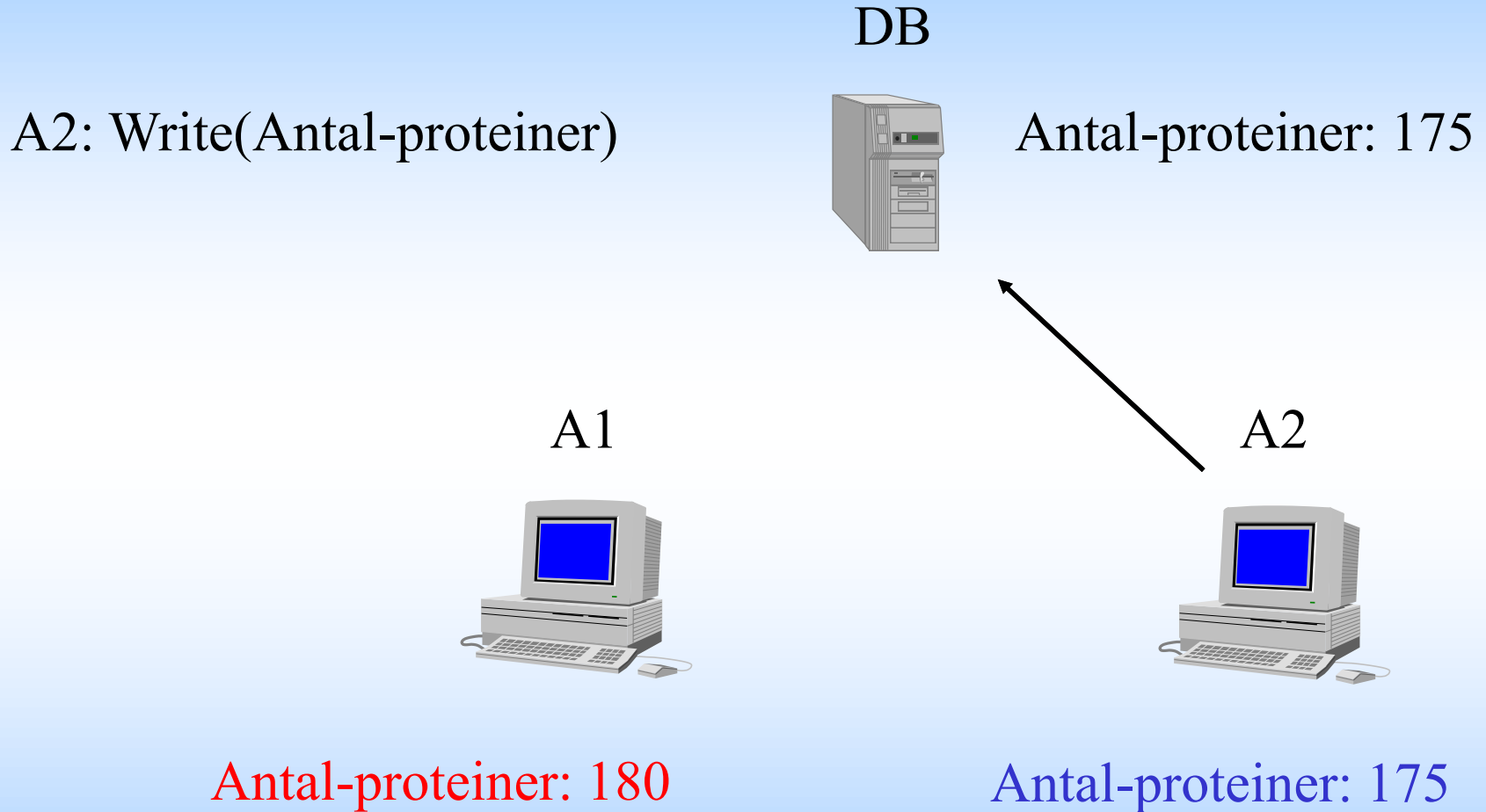


Antal-proteiner: 175

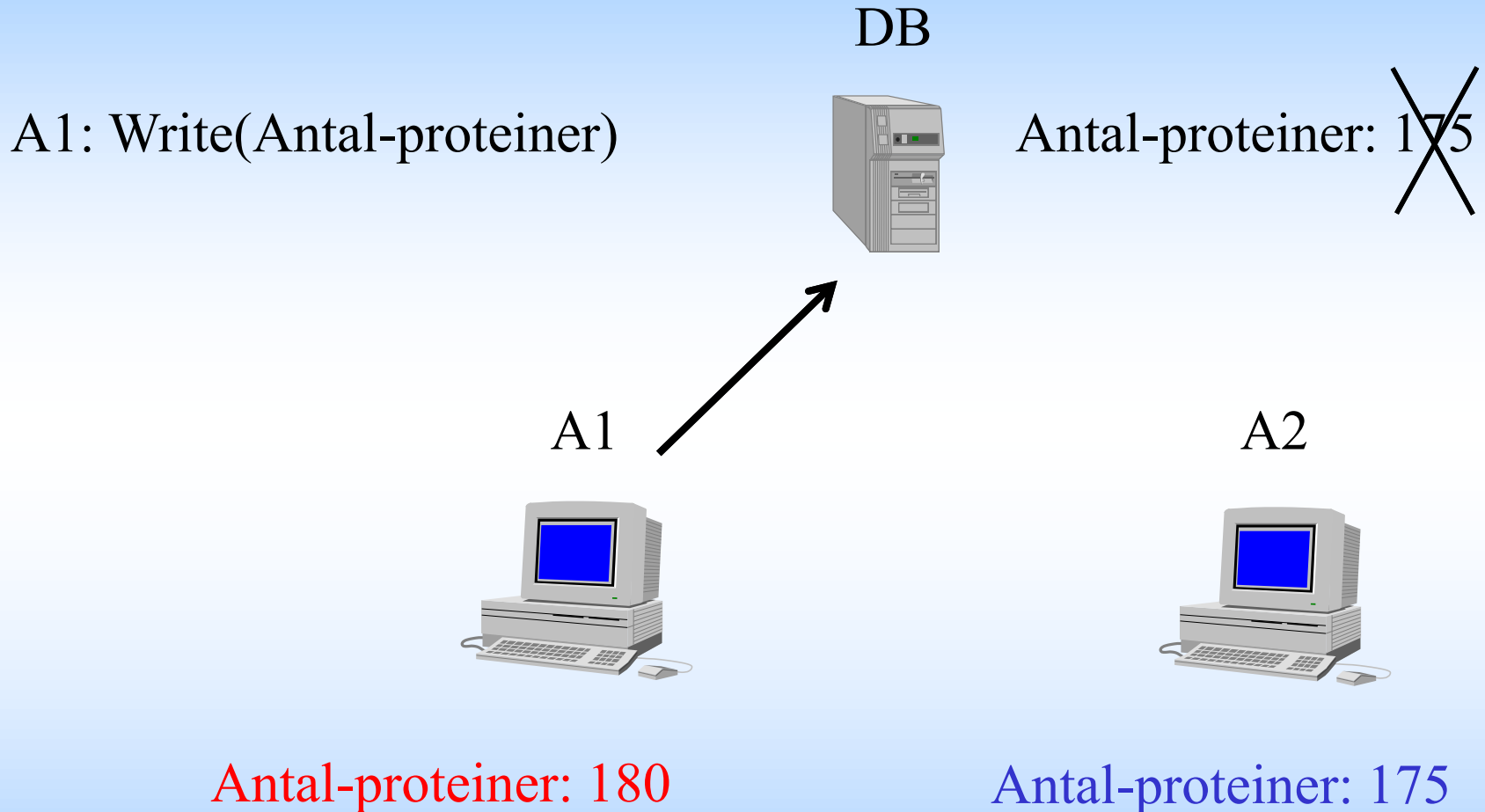
Flera användare



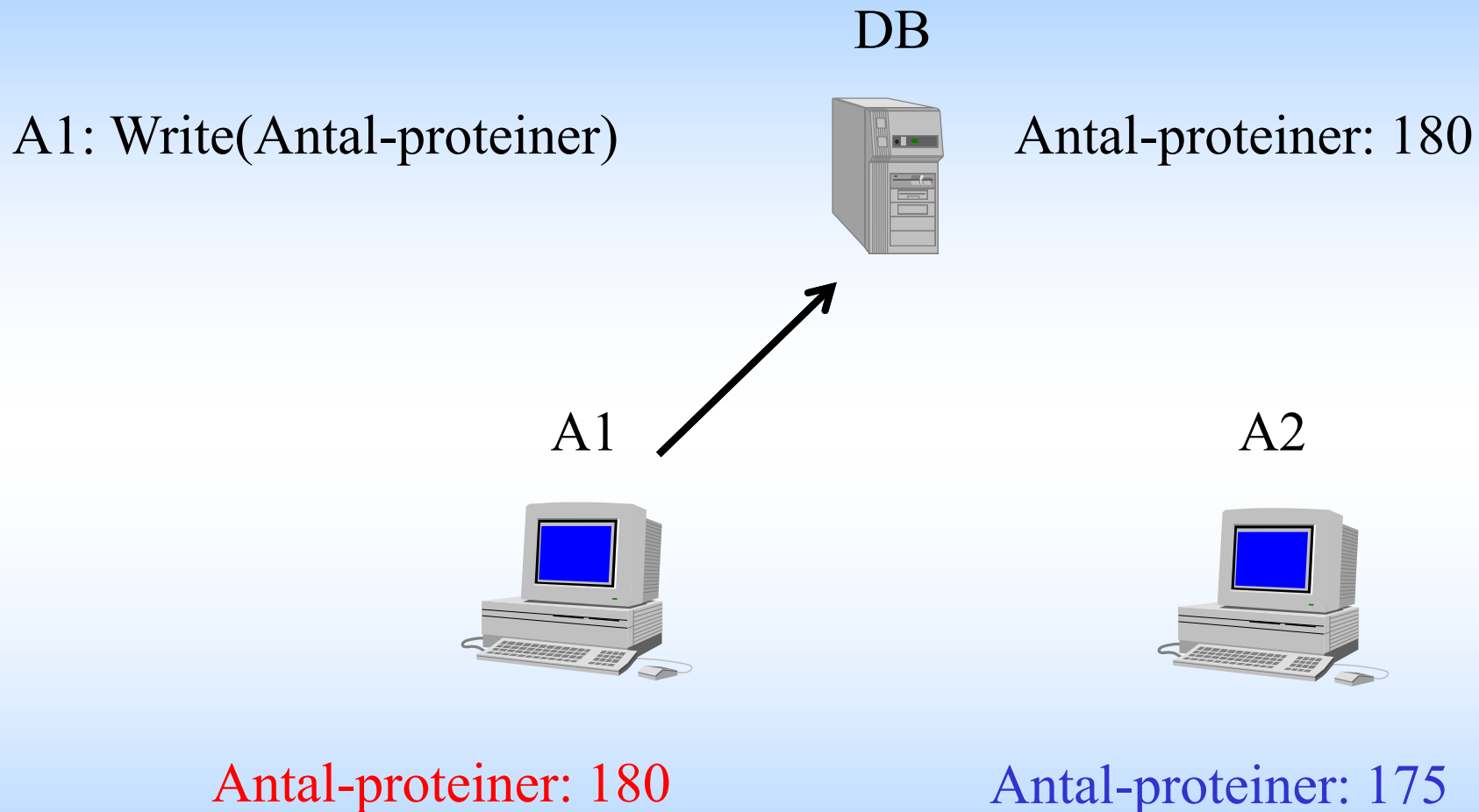
Flera användare



Flera användare



Flera användare



Flera användare

DB



Antal-proteiner: 180

Antal-proteiner: $150 + 30 + 25 = 205$

Informationssäkerhet

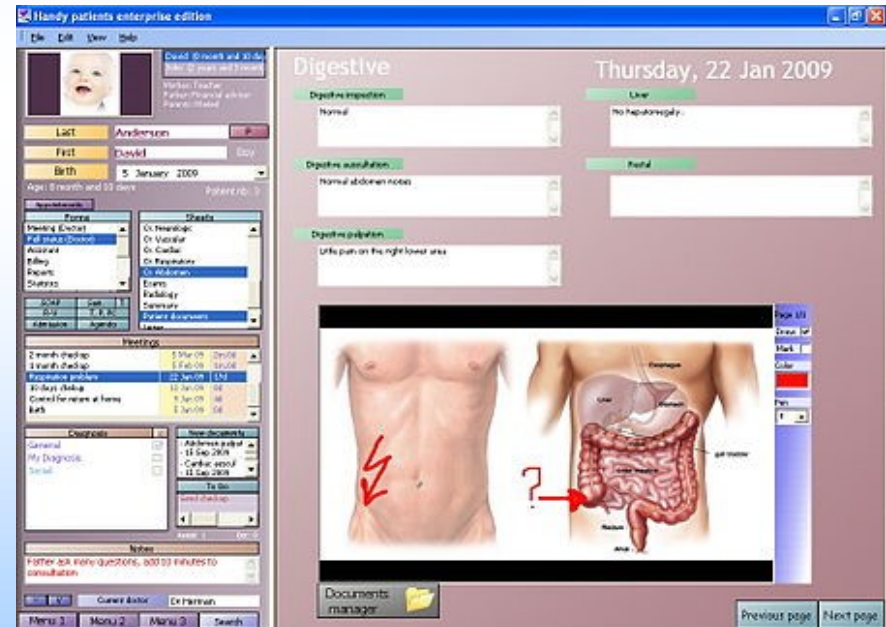
Sekretess

Enbart behöriga användare får ta del av informationen

Integritet

Korrekt och fullständig information

Tillgänglighet



Kursöversikt - FÖ

- Introduktion
- Relationsdatabaser och SQL
- Datamodellering, ER/EER diagram
- Att gå från EER diagram till relationsscheman

Kursöversikt - FÖ

- Informationssäkerhet
- Databassäkerhet

Kursöversikt - LA+projekt

- Lab1: Grundläggande SQL
- Lab2: Databasdesign och EER modellering
- Lab3: Avancerad SQL
- Lab4: Informationssäkerhet

Kursöversikt - LA+projekt

- 'Lab5': Projekt i bioinformatik
 - genomdatabas
 - proteindatabas
 - enzymdatabas
 - databas för biologiska reglersystem

Kursöversikt - LA+projekt

- Rapporteringsdeadline vid varje tentamenstillfälle
- behövs ett särskilt databaskonto
 - > automatisk vid registrering på kursen
 - databaskontona tas bort efter 1 år
- anmälan till laborationer via kurshemsidan senast 4 april

Examination

- tenta
- laborationsserie
- projekt

En kurs för TB

- Användning i senare kurser + arbete
- Unik och eftertraktad kompetens
 - Bio
 - Data
 - Förståelse av modellering + konsekvenser
(Hur modellera? Hur ställa frågor? Värfor går det långsamt? Varfor får man inget svar?...)

