

Tiedon esitys tietokoneessa

Jyry Suvilehto

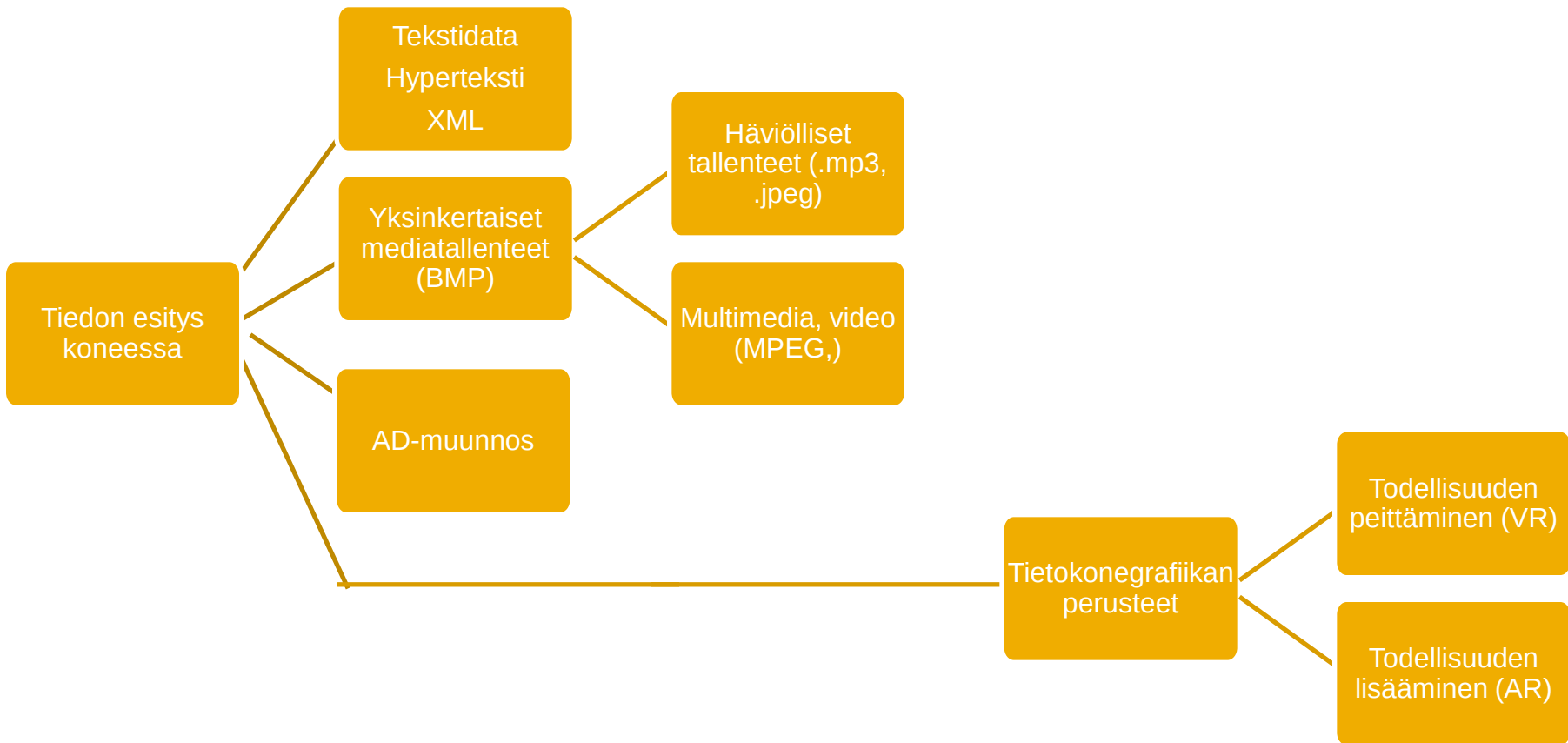
T-110.1100 Johdatus tietoliikenteeseen ja
multimediatekniikkaan
kevät 2011

Luennon sisältö

1. Kurssin loppupuolen rakenne
2. Tiedon binääriluonne
 - AD-muunnos
3. Tekstitiedostot
 - ASCII
 - XML
4. Binääritiedostot
 - Esimerkkinä BMP
 - Vähän väriavaruuksista
5. Välikoe

**HUOM! TÄMÄN LUENNON
UUSI SISÄLTÖ VASTA
SEURAAVAAN
VÄLIKOEeseen**

Multimediaosuus



5.3. Luento
• Jyry Suvilehto

18.3. Luento
• Petri Vuorimaa

25.3. Luento
• Timo Tossavainen

1.4. Luento
• Lauri Savioja

Synteesiosuus



8.4. EI
LUENTOA

16.4. Luento
Eero Hyvönen,
Eetu Mäkelä,
Osma Suominen
ja Kim Viljanen

23.04. Luento
Sakari Luukkainen

TIEDON BINÄÄRILUONNE

Bitti

- **Atomi** = pienin jakamaton osanen
 - Kreik
- Reaalimaailmassa on olemassa atomia pienempiä osasia
- Tietokoneissa **bitti** on ihan oikeasti pienin jakamaton osa
 - Jakamattomuudella järjestyttäviä seurauksia
- Bitti voi olla kahdessa tilassa, esim. Jännitteenä ilmaistuna
 - 0 on 0v
 - 1 on +5v
- Kun bitin arvo on 1, sanomme että bitti on päällä tai tosi, 0 vastaavasti on pois päältä tai epätosi
- Biteille on olemassa luonnollinen laskujärjestelmä, nk. Boolean logiikka
 - $1 \text{ and } 1 = 1$
 - $1 \text{ or } 0 = 1$
 - $0 \text{ and } 1 = 0$ etc.

Binääriaritmetiikka

- Monta bittiä peräkkäin laittamalla ja normaalilla yhteenlaskuoperaatiolla saadaan aikaiseksi algebrallinen rengas
- Formaalisti $\mathbb{Z}^+ \bmod 2^k$ kun käytössä on k bittiä
- Tarkemmin, kyseessä on äärellinen kunta eli ns. Galois-kunta
- *Peräkkäin asetetuille, sähköisesti esitetyille biteille on mahdollista Boolean logiikalla muodostaa sähköinen piiri joka tekee arbitäärisiä laskutoimituksia*
 - Claude Shannonin dippatyö

Lukujärjestelmämuunnos

- Ei kuulu tähän kurssiin varsinaisesti
 - Ei siis tulossa tenttiin, älkää hikoilko
- Kuuluu yleissivistykseen osata
- 13 in bin =
- 00011011 in dec =
- K bitillä voidaan esittää 2^k eri arvoa

Suurempia mittoja

- Bitit ryhmitellään usein **tavuiksi** (8 bittiä)
- Tietokoneen alijärjestelmät käsittelevät monen tavun mittaisia **sanoja** (tyypillisesti 16-64 bittiä)
- Monimutkaisemman konstruktion kautta on mahdollista määritellä desimaalilukuja, joita ovat kiinteän ja liukuvan pilkun luvut
 - Olennaista on se, että tarkkuus näilläkin äärellinen!

AD-muunnos

- Joskus tieto tarvitsee muuttaa analogisesta digitaaliseksi
- Suurin ongelma muodostuu äärellisyydestä
 - Joudutaan määrittelemään joku minimi, jota pienempi informaatio hävitetään
 - Aikasarjoissa näytteenottotaajuus, kuvissa tiheys (dots per inch)
- Alkuperäistä ei usein voida konstruoida takaisin tallennustusta
 - Joskus voidaan, ns. Nyquistin taajuus

Näytteenotto

- Yksiulotteisesta signaalista otetaan näytteitä ajanhetkillä $0, t, 2t, 3t$

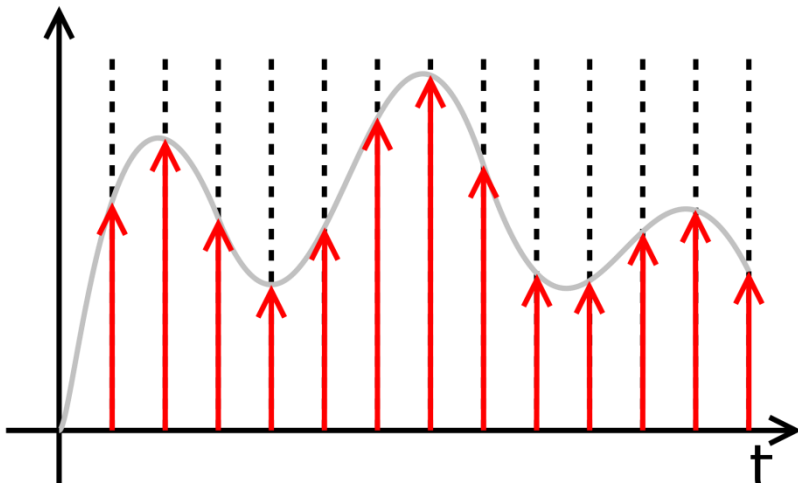
– t esim 1 ms

- Näytteenottotaajuus

$$f_s = \frac{1}{t}$$

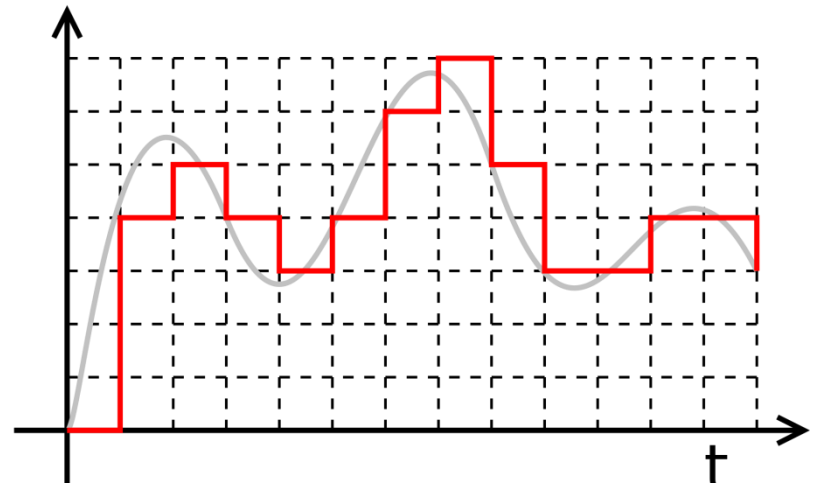
- Shannon-Nyquistin näytteenottoteoreema

– Signaali, jonka kaistanleveys on B voidaan rekonstruoida täysin jos $f_s > 2B$



Kvantisointi

- Yleensä ei ole järkevää käyttää arbitääristä tarkkuutta jokaiselle mittauspisteelle, vaan näytteet kvantisoidaan
 - Arvot pyöristetään tasaväliselle asteikolle
- Kvantisointi lisää virhettä selkeästi



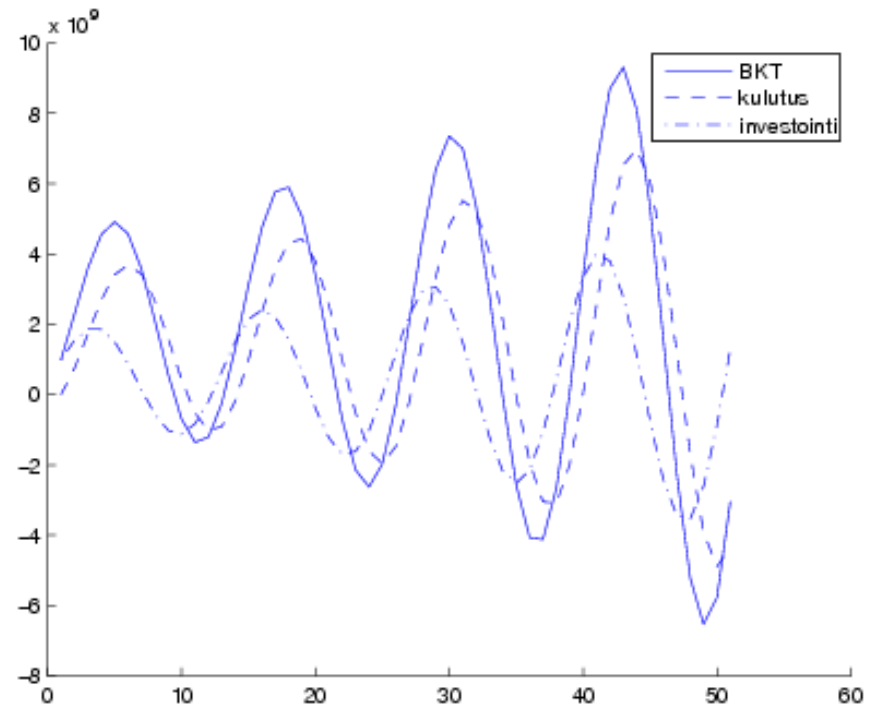
Edistyneitä AD-muunnoksia

- Tyypillisin esimerkki AD-muunnoksesta on yksiulotteinen suure ajassa (esim. Ääni)
- Kameran CCD-kenno suorittaa 2D-muunnoksen kahdessa tilaulottuvuudessa
- Videokamera suorittaa muunnoksen kahdessa tilaulottuvuudessa ja ajassa
- Perusperiaatteet ja ominaisuudet säilyvät

Toisinpäin

- Digitaalisesta analogiseksi muunnettaessa askeleet käännetään
- Toisaalta voi olla signaaleja, jotka on generoitu suoraan digitaalisessa muodossa
 - AD-muunnoksen ongelmia ei ole
- Tietokoneiden ensimmäisiä käyttötarkoituksia oli maailman simulointi
 - Jenkkiarmeijan tykistön ampuma-arvot
 - Atomin halkeaminen

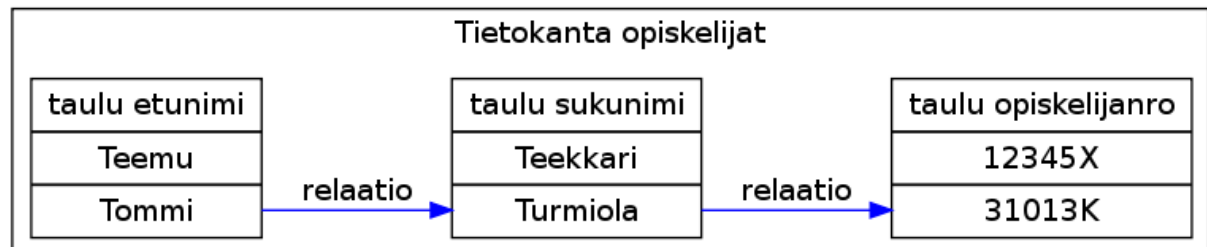
Simulointi



Tiedon jäsentämisen terminologiaa

- **Kenttä** on yksittäinen pala tietoa
 - opiskelijanumero tai nimi
- **Tietue** on kokoelma kenttiä
 - Yhdistetty opiskelijanumero, nimi ja ikä
- **Tiedosto** on kokoelma tietueita
 - Listaus kurssin opiskelijoiden opiskelijanumeroista, nimistä ja i'istä
- **Tietokanta** kokoelma kenttiä, joiden välillä on relaatioita

opiskelijat.csv		
Etunimi,	Sukunimi,	Opiskelijanumero
Teemu,	Teekkari,	12345X
Tommi	Turmiola	31013K



Tiedostotyypeistä

- Tiedostoja voidaan luokkiin monella tavalla
- Unixin *file*-työkalun filosofian mukaan on olemassa
 - Tekstitiedostoja
 - Binääritiedostoja
 - Ajettavia ohjelmia (binääritiedostojen erikoistapaus)
 - Ei käsitellä

TEKSTITIEDOSTOT

Teksti

- Luku- ja kirjoitustaito ovat aika vanhoja juttuja
- Syväälle yhteiskuntaan juurtunut vuorovaikutustapa
- Tietokoneiden tekstissä ensimmäinen ja laajimmalle levinnyt standardi on alunperin lennättimiä varten kehitetty **American Standard for Code for Information Interchange (ASCII)**
 - 7 bittiä koska haluttiin säästää
 - Englannin kirjoittamiseen tarvittavat aakkoset, erikoismerkkejä
 - Paljon lennättimen ohjaamiseen erikoistuneita ns. Non-printable merkkejä

Ei-tulostuvat ja tulostuvat merkit

DEC	HEX	Symboli	Kuvaus	Näppäimistöyöte	DEC	HEX	Merkki	DEC	HEX	Merkki	DEC	HEX	Merkki
0	00	NUL	Null-merkki		32	20	Välilyönti	64	40	@	96	60	'
1	01	SOH	Start of Heading	ctrl-A	33	21	!	65	41	A	97	61	a
2	02	STX	Start of Text	ctrl-B	34	22	"	66	42	B	98	62	b
3	03	ETX	End of Text	ctrl-C	35	23	#	67	43	C	99	63	c
4	04	EOT	End of Transmission	ctrl-D	36	24	\$	68	44	D	100	64	d
5	06	ENQ	Enquiry	ctrl-E	37	25	%	69	45	E	101	65	e
6	06	ACK	Acknowledgement	ctrl-F	38	26	&	70	46	F	102	66	f
7	07	BEL	BELL	ctrl-G	39	27	'	71	47	G	103	67	g
8	08	BS	Backspace	ctrl-H	40	28	(	72	48	H	104	68	h
9	09	HT	Horizontal Tab	ctrl-I	41	29	)	73	49	I	105	69	i
10	0A	LF	Line Feed	ctrl-J	42	2A	*	74	4A	J	106	6A	j
11	0B	VT	Vertical Tab	ctrl-K	43	2B	+	75	4B	K	107	6B	k
12	0C	FF	Form Feed	ctrl-L	44	2C	,	76	4C	L	108	6C	l
13	0D	CR	Carriage Return	ctrl-M	45	2D	-	77	4D	M	109	6D	m
14	0E	SO	X-On/Shift Out	ctrl-N	46	2E	.	78	4E	N	110	6E	n
15	0F	SI	X-Off/Shift In	ctrl-O	47	2F	/	79	4F	O	111	6F	o
16	10	DLE	Data Line Escape	ctrl-P	48	30	0	80	50	P	112	70	p
17	11	DC1	Device Control 1	ctrl-Q	49	31	1	81	51	Q	113	71	q
18	12	DC2	Device Control 2	ctrl-R	50	32	2	82	52	R	114	72	r
19	13	DC3	Device Control 3	ctrl-S	51	33	3	83	53	S	115	73	s
20	14	DC4	Device Control 4	ctrl-T	52	34	4	84	54	T	116	74	t
21	15	NAK	Negative Acknowledgement	ctrl-U	53	35	5	85	55	U	117	75	u
22	16	SYN	Synchronous Idle	ctrl-V	54	36	6	86	56	V	118	76	v
23	17	ETB	End of Transmit Block	ctrl-W	55	37	7	87	57	W	119	77	w
24	18	CAN	Cancel	ctrl-X	56	38	8	88	58	X	120	78	x
25	19	EM	End of Medium	ctrl-Y	57	39	9	89	59	Y	121	79	y
26	1A	SUB	Substitute	ctrl-Z	58	3A	:	90	5A	Z	122	7A	z
27	1B	ESC	Escape	ctrl-[	59	3B	;	91	5B	[	123	7B	{
28	1C	FS	File Separator	ctrl-\	60	3C	<	92	5C	\	124	7C	
29	1D	GS	Group Separator	ctrl-]	61	3D	=	93	5D	]	125	7D	}
30	1E	RS	Record Separator	ctrl-^	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit Separator		ctrl-_	63	3F	?	95	5F	_	127	7F	DEL

ASCII:n jälkeen

- Englannin kirjoittamiseen tarvittavat merkit eivät riitä kovin pitkälle
 - suuri määrä erilaisia merkistöjä
 - Länessä yleensä taaksepäin yhteensopivia ASCII:n kanssa
- Unicode –merkistö tullut pysyäkseen
 - 32 bittinen osoiteavaruus
 - Eri tapoja koodata sama merkistö, UTF-8, UTF-16, UTF-32

Tekstitiedostoformaatit

- Mikäli tekstitiedostossa on rakenne, voidaan tehdä ohjelma joka osaa lukea sitä
- Erittäin yksinkertainen esimerkki: Comma Separated Values

26, "Jyry", 191

23, "Teemu", 165

19, "Matti", 180

Tekstitiedostoformatit

- Tietynlaisen rakenteen odottaminen on rajoittavaa
- Voidaan määrittää rakenteen palaset ja sopia että jokainen palanen on merkattu
- Merkkauskieli

`<h1>Tämä on otsikkoa</h1>`

`<p>Tämä on kappale</p>`

Hyperteksti

- Teksti on luonteeltaan sekventiaalista
 - Kirjat on tarkoitus lukea alusta loppuun
- Lisätään mahdollisuus linkkien tekemiseen

```
<a href="#toinenosa">tästä eteenpäin</a>
```

```
...
```

```
<a name="toinenosa">
```

```
<h2>Toinen osa</h2>
```

- Miksi rajoittaa saman dokumentin (tai dokumenttikirjaston) sisälle linkkaamiseen?

```
<a href="http://toinen.resurssi.com/jotain/ihan/muuta">jotain ihan muuta</a>
```

HTML:n puutteita

- Tägit kertovat ladontatietoa
 - Taulukoita, otsikoita, kappalejakoja
- Tägejä äärellinen määrä
 - Ei laajennusmahdollisuutta
 - Ei yleinen ratkaisu
- Liberaalia sen suhteen, mitä hyväksytään
 - Ei koneiden ymmärrettävissä

XML

- e**X**tensible **M**arkup **L**anguage
- Vanhan SGML-standardin uudet vaatteet
- Kehittynyt irrallaan HTML:stä, ottanut oppia HTML:n virheistä
 - XHTML on XML-speksin mukainen HTML
- Perusidea on sama
 - `<tag attribute="example">`
 - This is tagged text
 - `<tag />`
- XML:ää pidetään välillä liiankin helppolukuisena
- Koneluettava, laajennettava

XML-dokumentin rakenne

- XML-dokumentin runko koostuu elementeistä (tag), joilla voi olla attribuutteja
- Elementtien välistä suhdetta kuvataan jollain tavalla
 - XML schema
 - **Document Type Description**

<!ELEMENT kirjasto
(hylly*)>

<!ELEMENT hylly
(kirja*)>

<!ELEMENT kirja (nimi,
kirjailija?, vuosi) >

<!ELEMENT kirja
(#PCDATA)>

<!ELEMENT kirjailija
(#PCDATA)>

<!ELEMENT vuosi
(#PCDATA)>

DTD ja sen mukainen dokumentti

```
<!ELEMENT kirjasto (hylly*)>
<!ELEMENT hylly (hyllynro, kirja*)>
<!ELEMENT hyllynro (#PCDATA)
<!ELEMENT kirja (nimi, kirjailija?, vuosi) >
<!ELEMENT kirja (#PCDATA)>
<!ELEMENT kirjailija (#PCDATA)>
<!ELEMENT vuosi (#PCDATA)>
```

DTD on sopimus dokumentin rakenteesta: ohjelma voidaan tehdä käsittelemään DTD:n mukaista dokumenttia.

Kaikille itseään kunnioittaville ohjelmointialustoille on jo olemassa XML-parsereita valmiina.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE kirjasto PUBLIC "KIRJASTO"
"http://www.cse.tkk.fi/u/jsuvileh/jotime/esim.dtd">
<?xml-stylesheet type="text/xsl" href="hylykko.xsl"?>
<kirjasto>
  <hylly>
    <hyllynro>1</hyllynro>
    <kirja>
      <nimi>Hobitti</nimi>
      <kirjailija>J.R.R. Tolkien</kirjailija>
      <vuosi>1937</vuosi>
    </kirja>

    <kirja>
      <nimi>Silmarillion</nimi>
      <kirjailija>J.R.R. Tolkien</kirjailija>
      <vuosi>1977</vuosi>
    </kirja>

  </hylly>

  <hylly>
    <hyllynro>2</hyllynro>
    <kirja>
      <nimi>1984</nimi>
      <kirjailija>George Orwell</kirjailija>
      <vuosi>1949</vuosi>
    </kirja>
    <kirja>
      <nimi>Suomen laki</nimi>
      <vuosi>2009</vuosi>
    </kirja>
  </hylly>
</kirjasto>
```

XSL ja XPath

- Mm. verkkoselaimet osaavat näyttää XML-dokumentteja
 - Muista, että XHTML on XML-kieli
- Xpath on tapa viitata dokumentin osiin
- XSLT-kuvauskielellä voidaan kuvata eri tapoja kuvata sama kerran määritelty data
 - Kuin CSS HTML:lle, paitsi äärettömästi voimakkaampi
- Demo
 - <http://www.cse.hut.fi/~jsuvileh/jotime/kirjasto.xml>
 - Katso myös esim.dtd ja hyllykko.xsl

BINÄÄRITIEDOSTOT

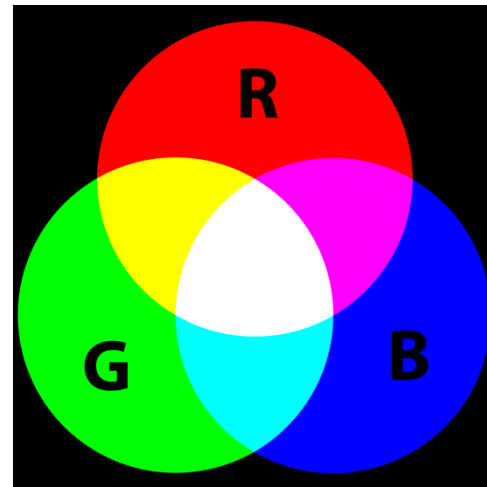
Binääritiedostot

- Binääritiedostoja ei yleensä ole tarkoitettu ihmisten luettavaksi
- Tarvitaan heksaeditori tai jotain muuta ikävää lukemiseen
- Binääritiedostot tarjoavat usein kompaktimman tavan esittää tietoa
 - ”vihreä” vs (0,255,0)

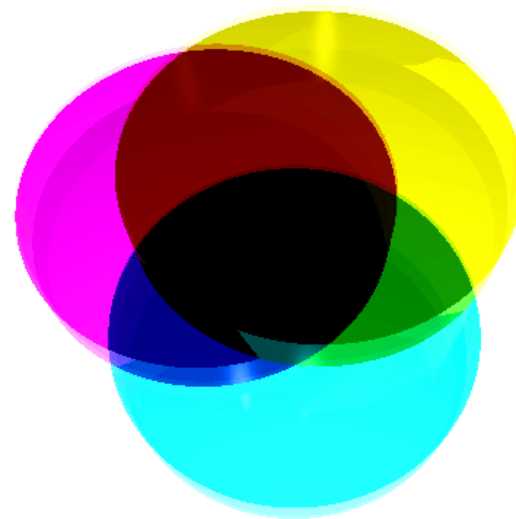
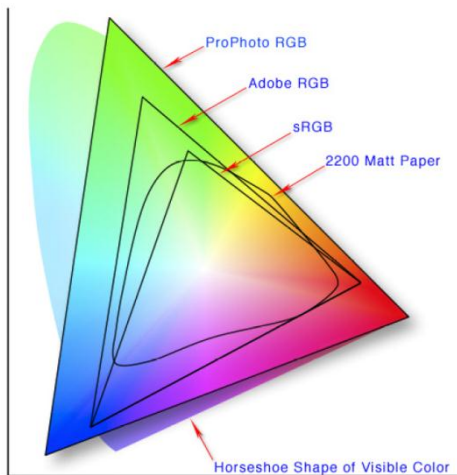
Värimallit

- Värejä voidaan yhdistää pääväreistä
- Värimallit voivat olla joko
 - Additiivisia (RGB), usein tietokonenäytöissä
 - Subtraktiivisia (CMYK), paperipainatuksessa
- Koko väriavaruutta on käytännössä mahdotonta näyttää tietyllä laitteella

Kuva: Mike Horvath/Wikipedia



Kuva: Mirsav Todorac/Wikipedia



BMP -perusidea

- Koska R, G ja B kuvaavat koko kuvan, ilmaistaan jokainen näistä n bitillä (2, 4, 6 bittiä) per pikseli
- 0 on musta, $2^n - 1$ on maksimi ao. väriä
- Lisäksi tallennetaan hieman otsaketietoja
 - Alkuun "magic number" kertomaan että kyseessä on BMP-kuva
 - Kuvan koko
 - Teknisyyksiä (helpottavat ohjelman tekoa)



Kuva voi antaa väärän käsityksen. BMP:ssä pikselin värit tallennetaan yhdessä. Ei kolmessa kuvassa.

Lena

- Miss Marraskuu 1972
- Käytetty siitä saakka esimerkkinä kuvankäsittelyssä
- Koska BMP-tiedostot tuppaavat olemaan kookkaita, käsittelemme vain Lenan vasenta silmää



Korkeus/leveys

Tavua/pikseli

Magic number

Osoitin kuvan alkuun

Ensimmäisenä alavasan pikseli
85 85 dc hex = 132 (G), 133 (B), 220 (R)
Tästä eteenpäin riveittäin ylöspäin.

BMP -huomioita

- BMP ei tässä tapauksessa pakannut tietoa mitenkään
- Kuvassa on redundanssia, joka voidaan poistaa häviöttömästi
- Kuvassa on paljon ihmissilmälle tarpeetonta informaatiota, voidaan poistaa häviöllisesti
- Näistä ensi kerralla

Yhteenvetoa

- Tietokoneet äärellisiä, jänniä kapistuksia
 - Mikäli kiinnostaa enemmän, T-106.1150
- Tietokone ja käyttöjärjestelmä
 - Yleissivistävä, ei liian raskas
 - Pitäisi mennä kaikkiin opintoihin
- Tiedostot voidaan jakaa karkeasti teksti- ja binääritiedostoihin
- XML on "human readable" sisällönkuvauskieli
- BMP on erittäin yksinkertainen binääriformaatti kuvien tallentamiseen

VÄLIKOE

Yleistä

- Välikokeessa vaaditaan
 - luennoilla (muutkin kuin kalvoissa) käsitellyt asiat
 - Harjoituksissa käsitellyt asiat
 - Sisälukutaitoa (yllättävän vaikeaa)
 - Kykyä soveltaa
- Kokeiden teko on kahden asian tasapainottamista
 - Osaamisen mittaaminen
 - Järkevä korjattavuus
- Pyrimme siihen, että välikoe on ymmärryksen, ei pelkän faktatiedon mittari

Tilastollinen tenttitehtäväänalyysi

- Tavallaan vanhat pätee
 - Toisaalta vain viime vuonna sama laatija
- Kiero opiskelija miettii laatijan preferenssejä ja mielenkiinnon kohteita
- Kiero opiskelija miettii kotitehtävissä esiintyneitä teemoja
 - Kannattaa ainakin vilkaista tänään julkaistavaa kierrosta

Välikokeen rakenne

- 5 kysymystä
 1. Sanaselitykset
 2. Väittämät
 - 3.-5. "normaalit" tehtävät

Sanaselitykset

- Sitä mitä voisi olettaa
 - Pari virkettä
 - Selityksessä pitää löytää *olennainen*
 - Random-faktat aiheesta eivät ole määritelmiä
 - Esimerkki ei ole määritelmä
 - Jos selität asiaa ja puutaheinää sekaisin, voitulla Op
 - Kahden termin tapauksessa olennaista selittää miten eroavat
 - Joskus olemassa määritelmä
- Signaalointi eli merkinanto
 - Paketti- ja piirikytkentäinen verkko
 - Asiakas-palvelin –malli
 - Vertaisverkko
 - Protokollapino
 - Palomuuuri
 - Yhteydellinen
 - Kanavointi
 - Reititys
 - Luottamuksellisuus
 - Eheyys
 - Saatavuus
 - Web 2.0
 - Push (tietoliikenteessä)
 - DHCP

Väittämät

- Soveltamista vaativat väittämät
 - Pelkästä Totta/Ei totta –vastauksesta 0p
 - Pisteet tulevat perustelusta
 - *Vastaus ei välttämättä ole yksikäsitteinen'*

- WLAN ja Ethernet varaavat median lähetystä varten samalla tavalla
- IP-osoite on globaalisti uniikki
- UDP-protokollan ominaisuudet sopivat TCP:tä paremmin puheen tosiaikaiseen siirtämiseen
- Pelkkä WW riittäisi kaikkien Internetin palveluiden toteuttamiseen, muut järjestelmät ovat vain historiallisia kehitysvaiheita
- TKK:n webnews hyötyisi merkittävästi RSS:n lisäämisestä
- Jos saat verkossa olevalta taholta varmenteen, tiedät keneen olet yhteydessä
- Jokainen puhelinverkon puhelu vaatii puhelua koskevan tiedon tallentamista verkon käyttämiin puhelinkeskuksiin
- NAT-muutos ratkaisee riittämättömien IP-osoitteiden ongelman eikä aiheuta mitään sivuvaikutuksia
- Tietynnumeroisen TCP- tai UDP-portin takana on aina oltava tarkalleen samantyyppinen ohjelmisto (esim. sähköposti- tai www-palvelin)

Normaalit tehtävät

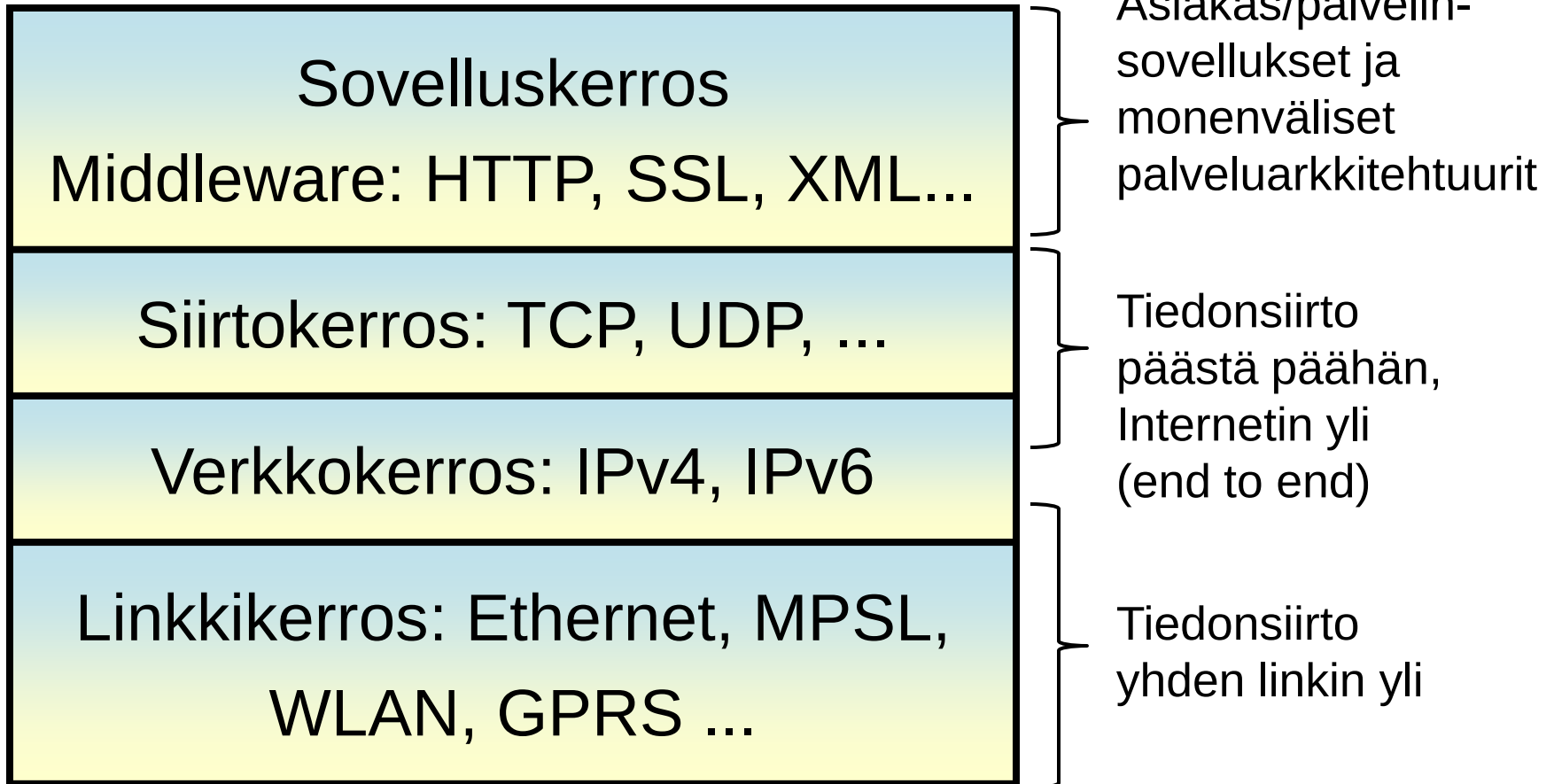
- Tarkastusteknisestä syistä suosimme jaettuja tehtäviä
- Pistemäärästä voi päätellä jotain vastauksen toivotusta pituudesta
- Olennainen on tärkeää, jaarittelusta ja kiertelystä ei saa pisteitä
 - Usean sivun asiaan liittymätön johdanto itse asiaan saattaa pahoittaa assarin mielen
- Selventäviä kuvia saa ja kannattaa piirtää, mutta jonkinnäköinen sanallinen selitysikin tarvitaan

Yleistä

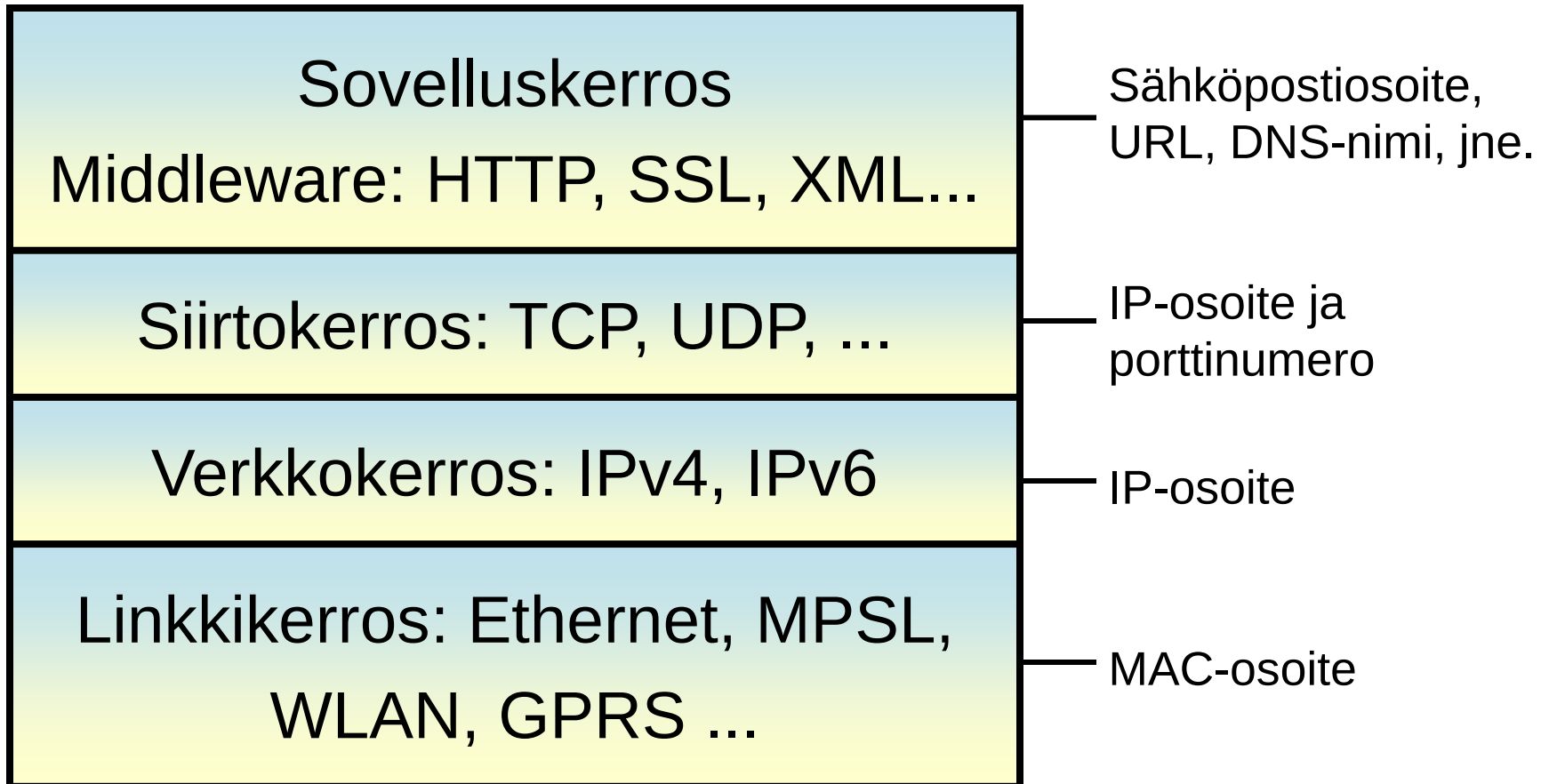
- Vastaukset pyritään korjaamaan mahdollisimman tasapuolisesti
- Käytännössä kuitenkin seuraavien asioiden vaikutusta mahdotonta eliminoida
 - Erittäin epäselvä käsiala
 - Hyppivä vastaus
 - Karkeat kielioppivirheet, slangi
- *Kirjoittakaa opiskelijanumero selvästi*

- Seuraavia protokollapinon kuvia voi käyttää kokonaisuuden hahmottamiseen tietoliikenneasioissa

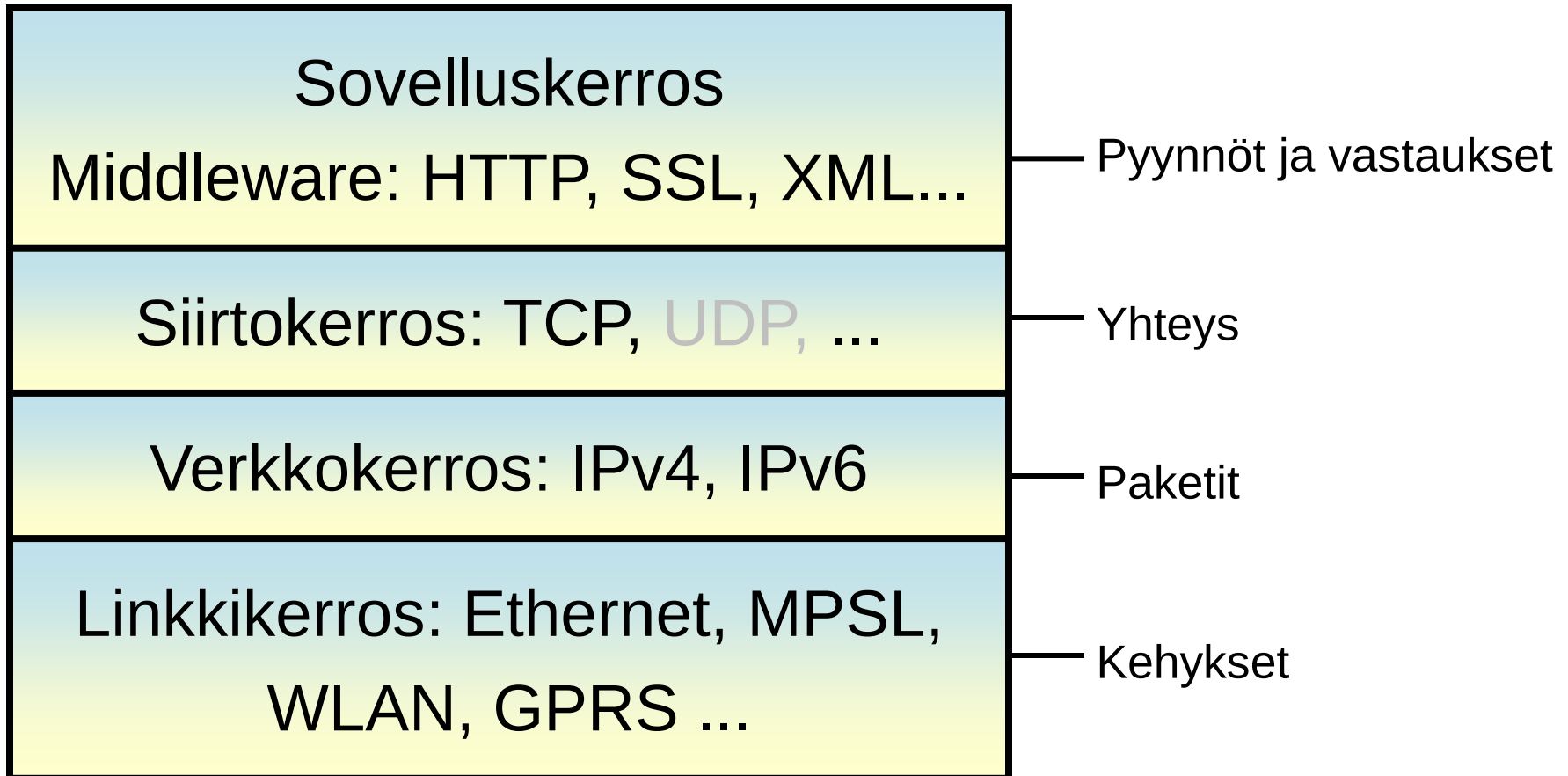
TCP/IP-protokollapino



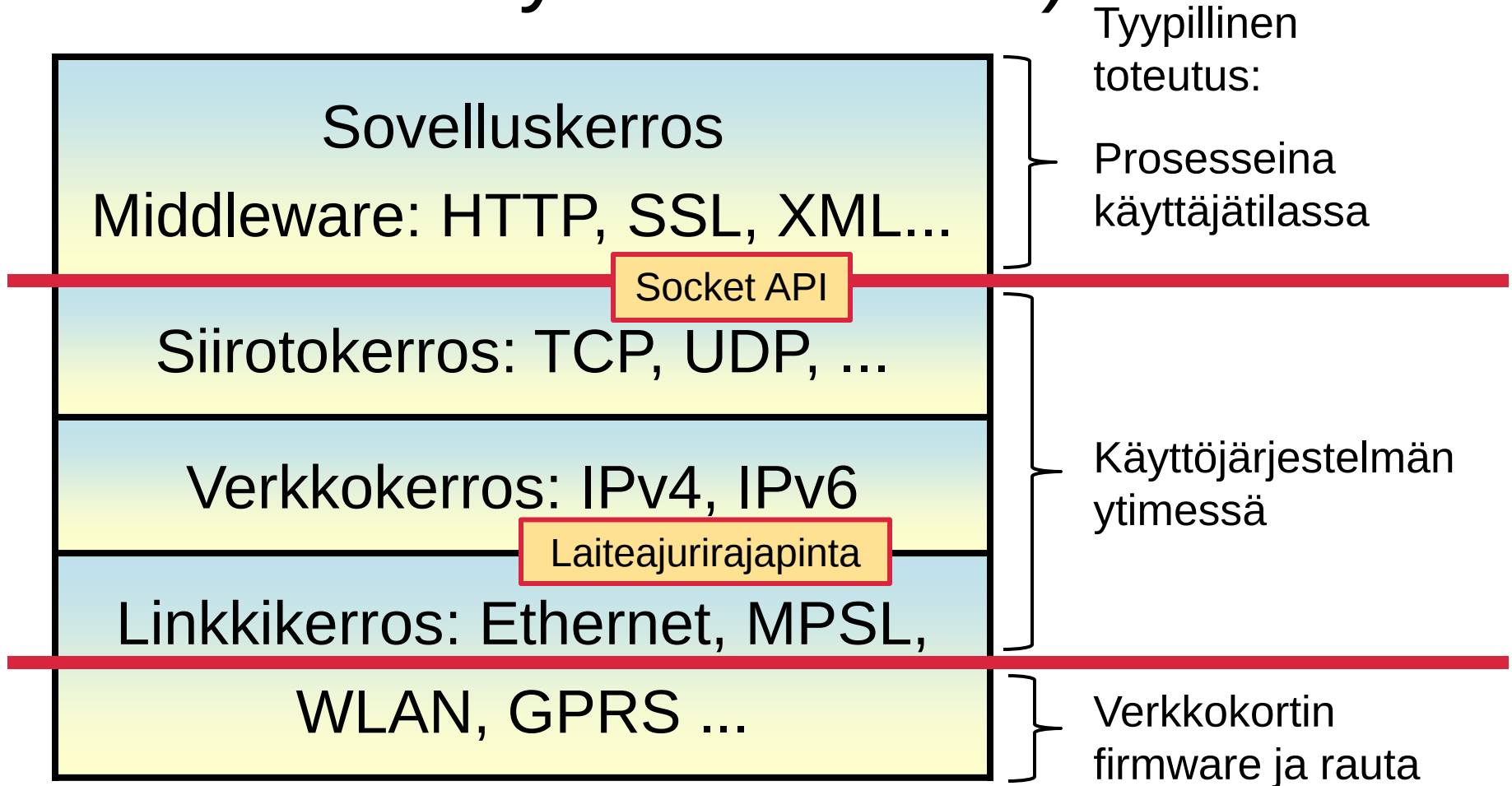
Nimiavaruudet



Viestien muoto



Toteutus ja rajapinnat (ei tarvitse täysin sisäistää)



Tiimalasimalli

