

# Oppivat koneet elämän arvoituksen jäljillä

Antti Honkela

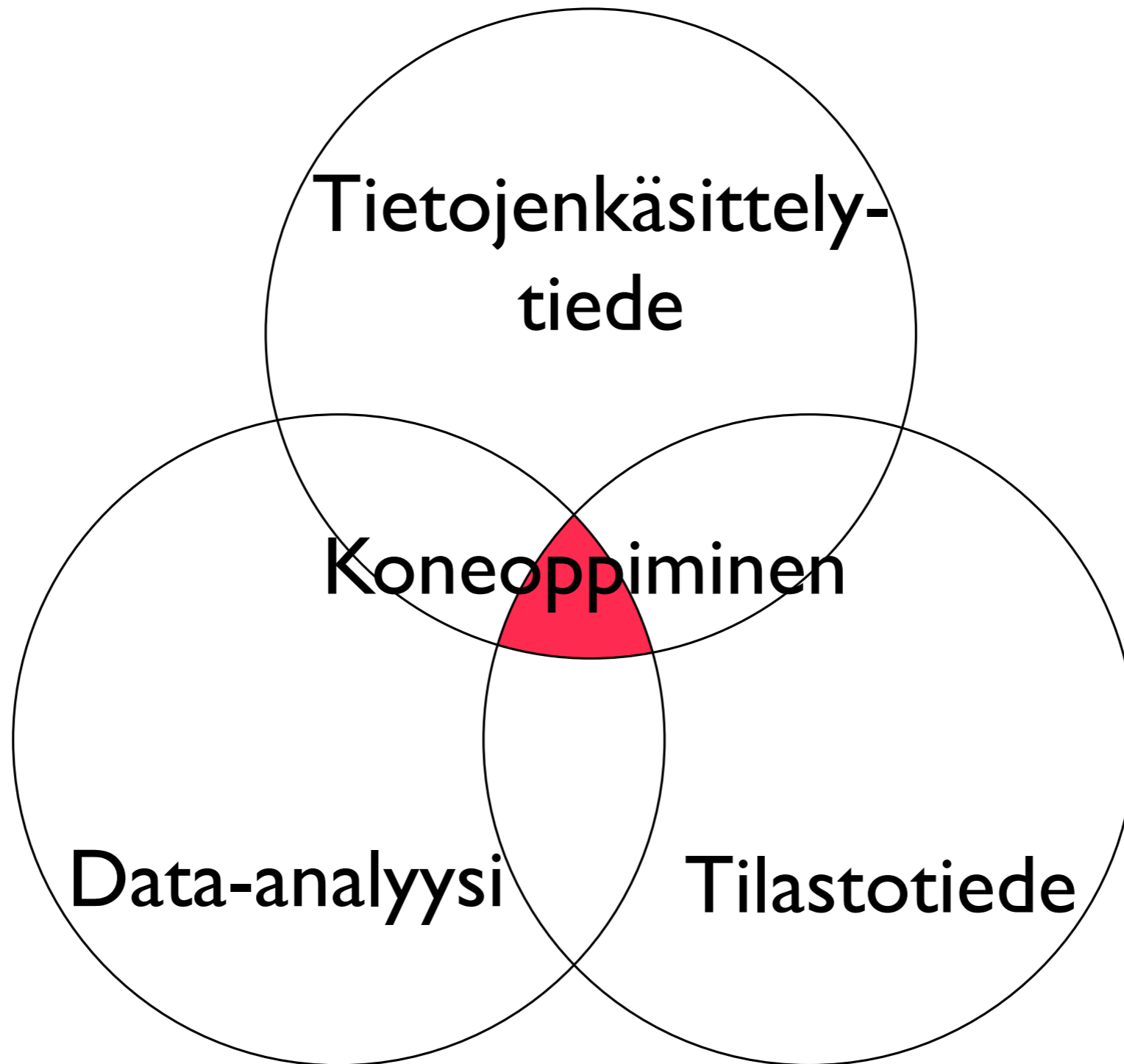
TKK / Tietojenkäsittelytieteen laitos

19.10.2009

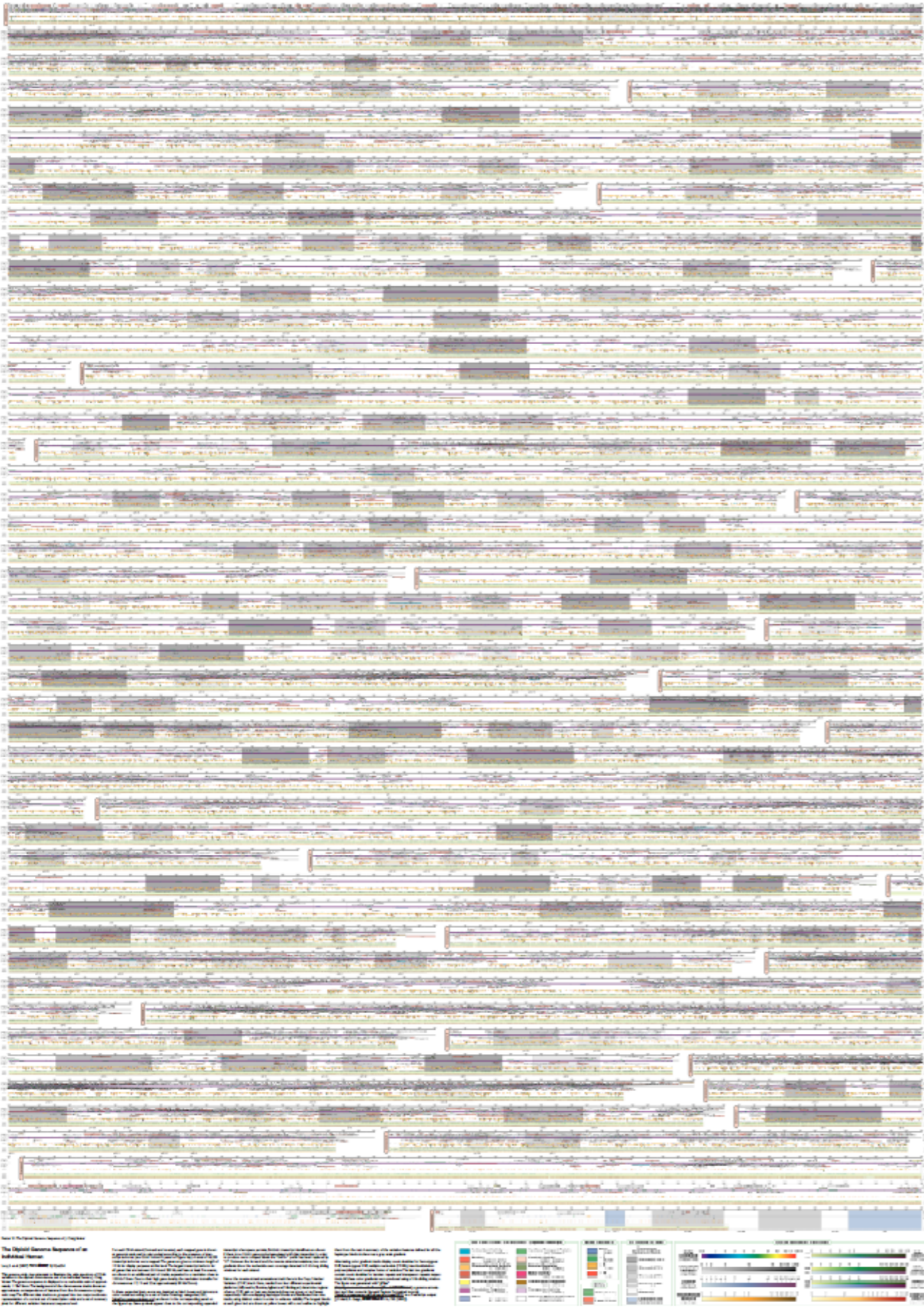
# Koneoppiminen

- Tekoälyn osa-alue, menetelmät perustuvat havainnoista oppimiseen
- Tavoitteena automatisoida data-analyysiä
- Pohjana usein abstraktien ongelmien ratkaiseminen, esim. luokittelun
- Läheinen suhde tilastotieteeseen

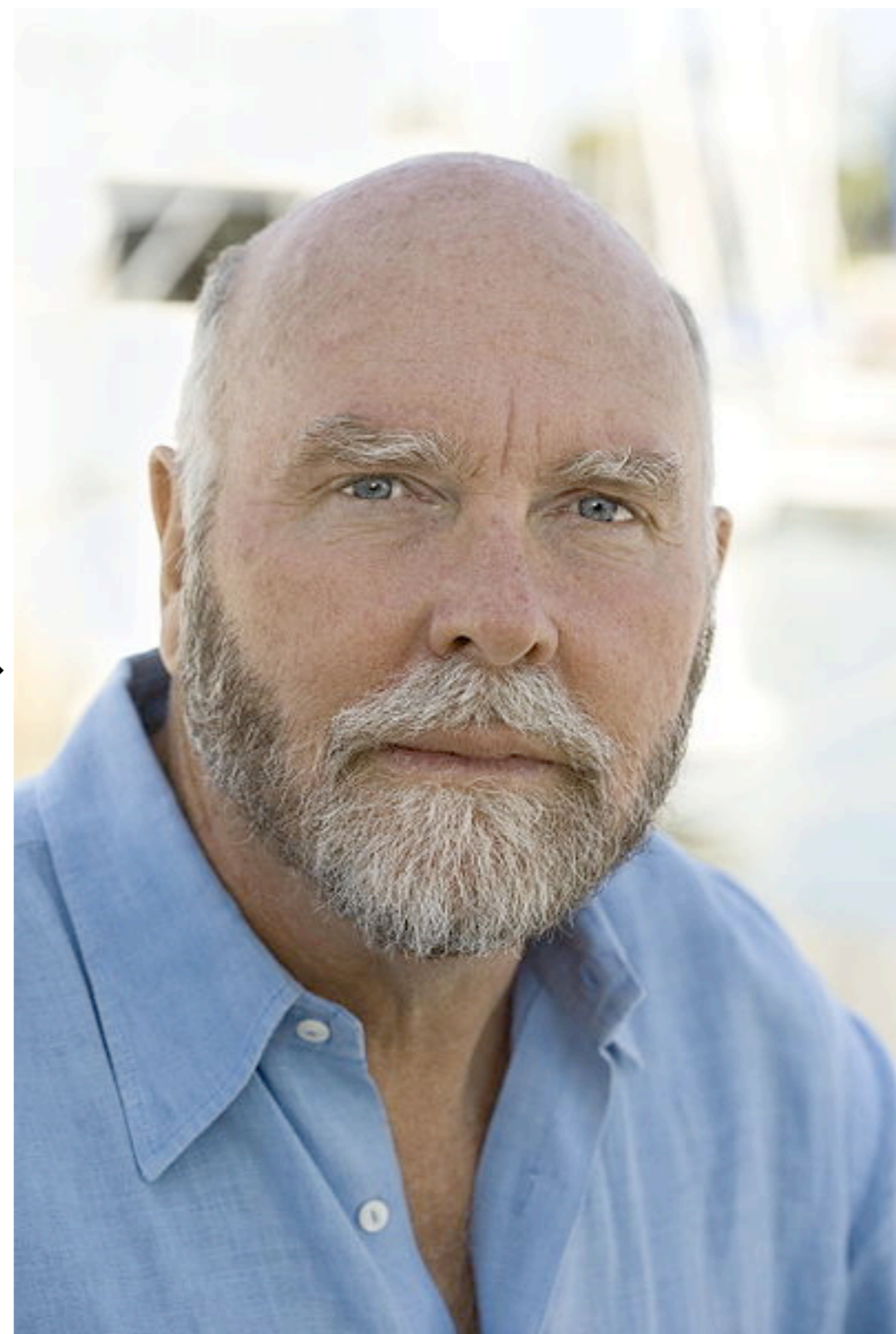
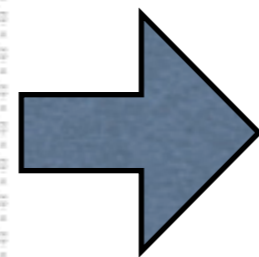
# Tieteiden rajapinnalla



# The Diploid Genome Sequence of J. Craig Venter



???



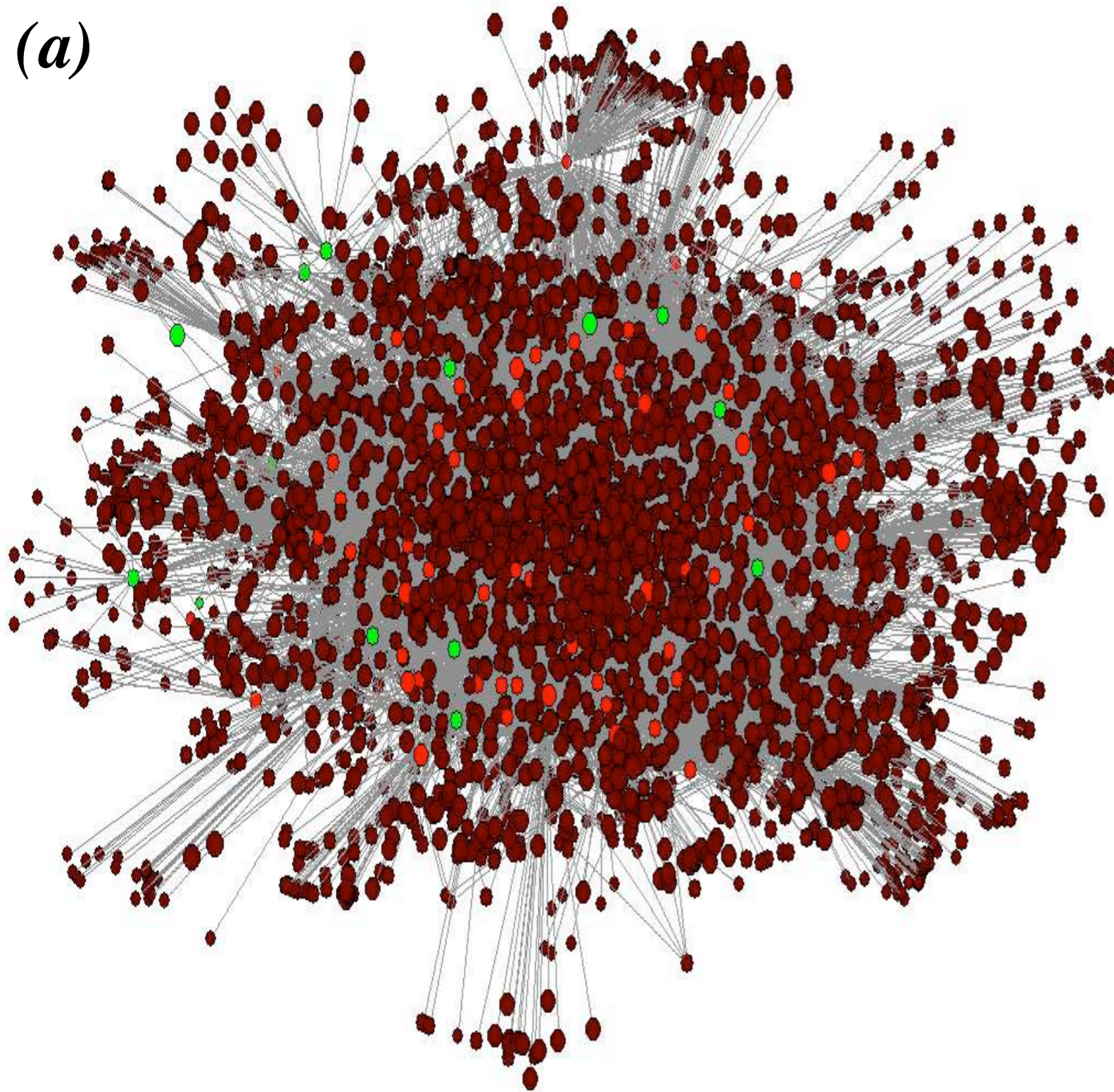
Images from Levy et al., PLoS Biology 5(10):e254 (2007); Gross, PLoS Biology 5(10):e266 (2007).

# Geenisäätely: faktoja

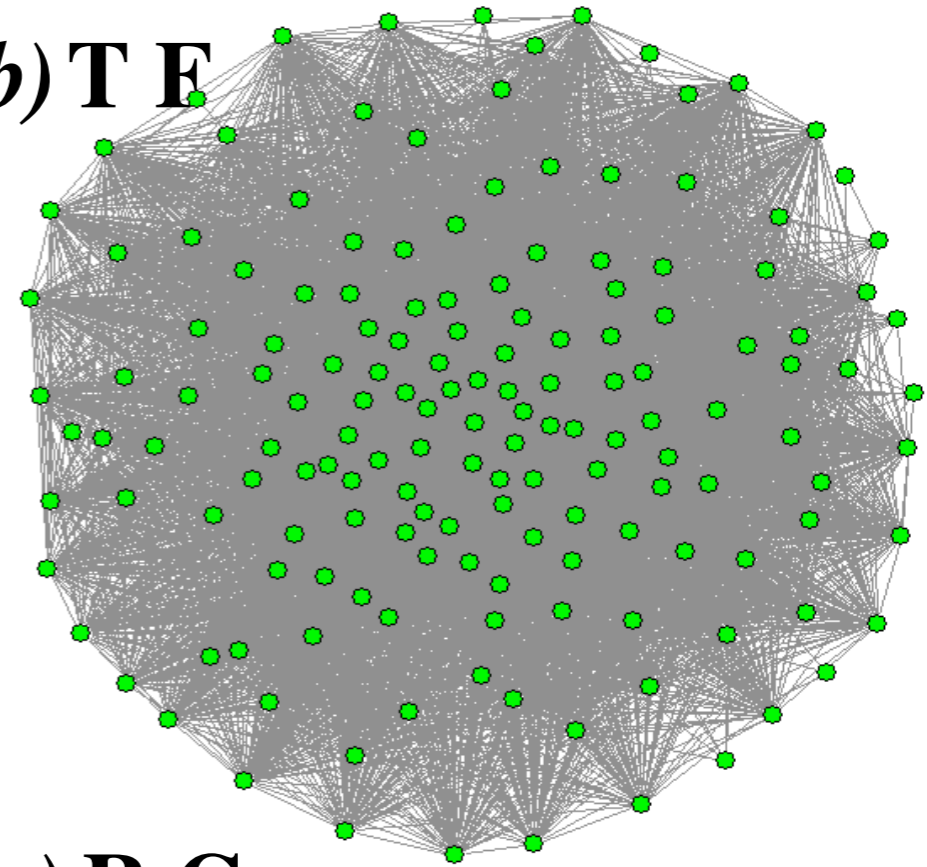
- Geenien aktivoituminen on sekä ajallisesti että paikallisesti lokalisoitunutta
- 98,5% ihmisen genomista ei koodaa proteiineja
- Yleisellä tasolla tunnetaan lukuisia säätelymekanismeja
- Esimerkiksi transkriptiosäätely

# *S. cerevisiae*

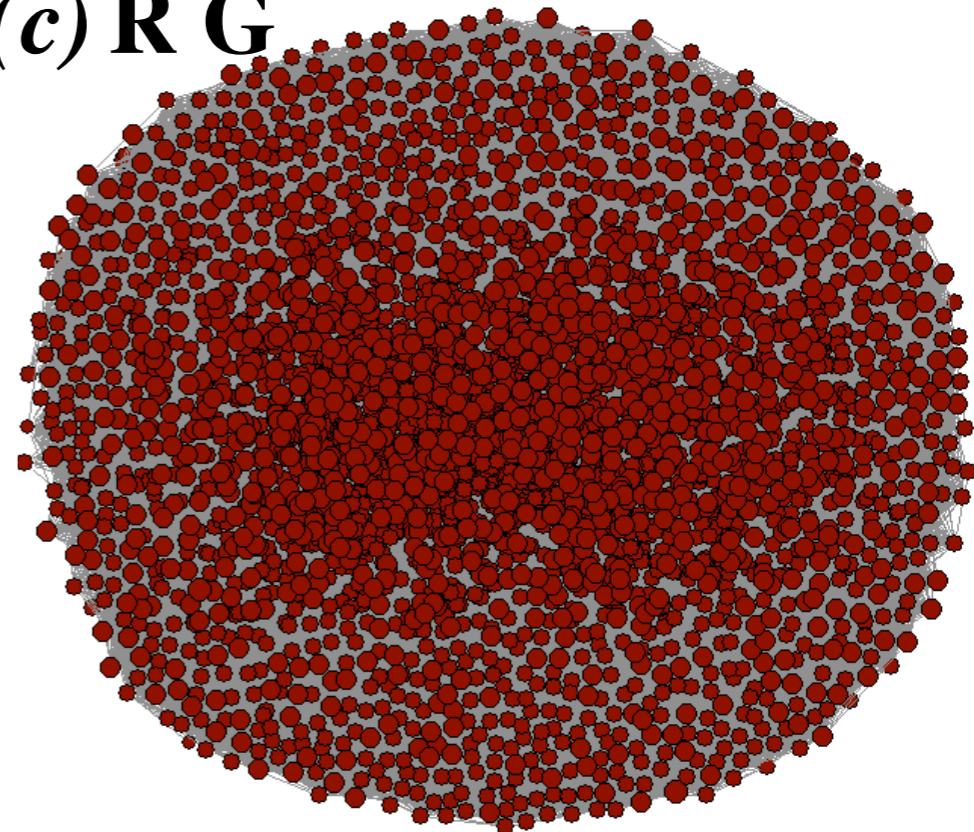
*(a)*



*(b)* T E



*(c)* R G



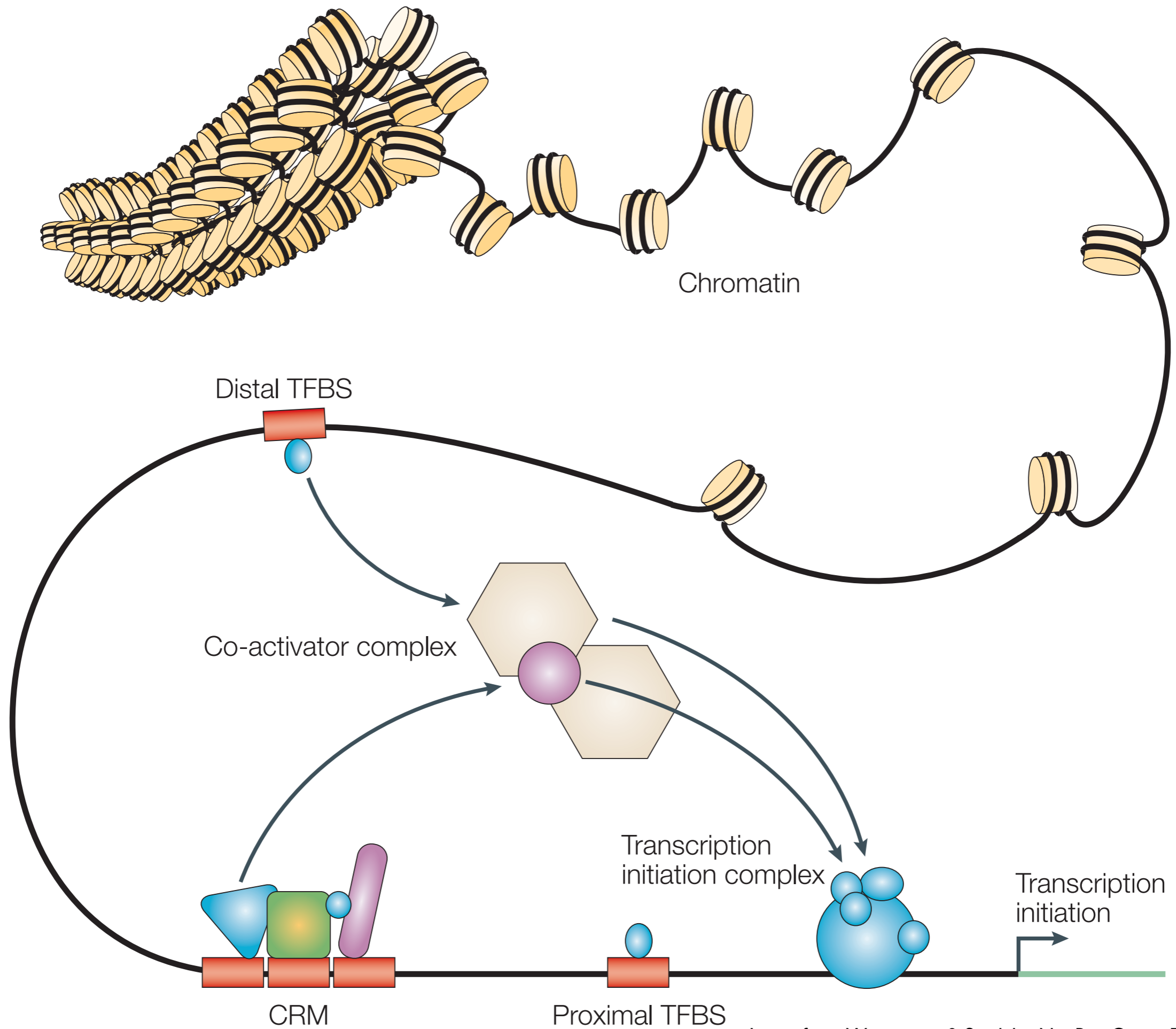


Image from: Wasserman & Sandelin. Nat Rev Genet. 5(4):276-87 (2004)

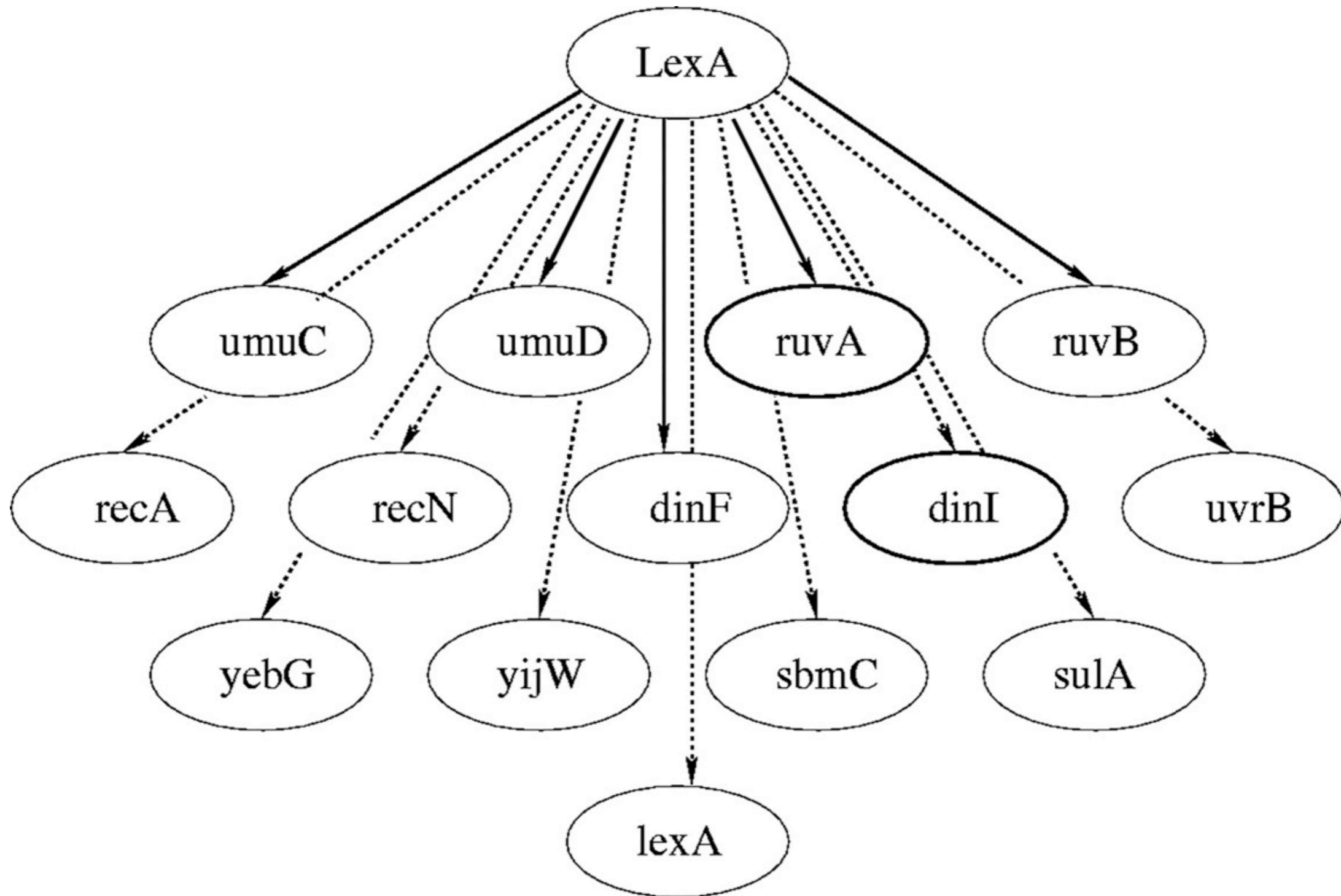
# Mallinnustehtävän haasteita

- Kaikkien geenien aktiivisuutta helppo mitata samanaikaisesti, proteiinien ei
- Mittaukset usein huomattavan kohinaisia
- Vähän mittauspisteitä suhteessa muuttujiin
- Paljon tuntemattomia vuorovaikutuksia

# Havaitsemattomien aineiden rekonstruktio

- Tavoitteena selvittää havaitsemattomien aineiden aktiivisuus biokemiallisessa systeemissä
- Esimerkkinä transkriptiofaktoriproteiinit, apuna niiden vaikutus kohdegeeneihin

# Säätelyverkkomotiivi



# Transkriptiomalli

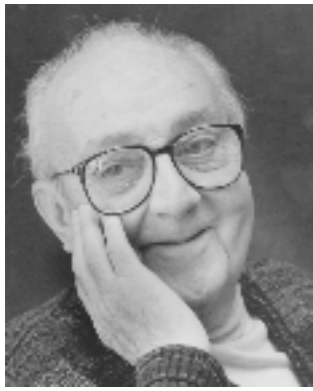
$$\frac{dx_j(t)}{dt} = B_j + S_j f(t) - D_j x_j(t)$$

Geenin aktiivisuuden muutos =

tausta-aktiivisuus

+ säätelyn vaikutus

– hajoamisen vaikutus



**“Essentially, all models  
are wrong, but some  
are useful.”**

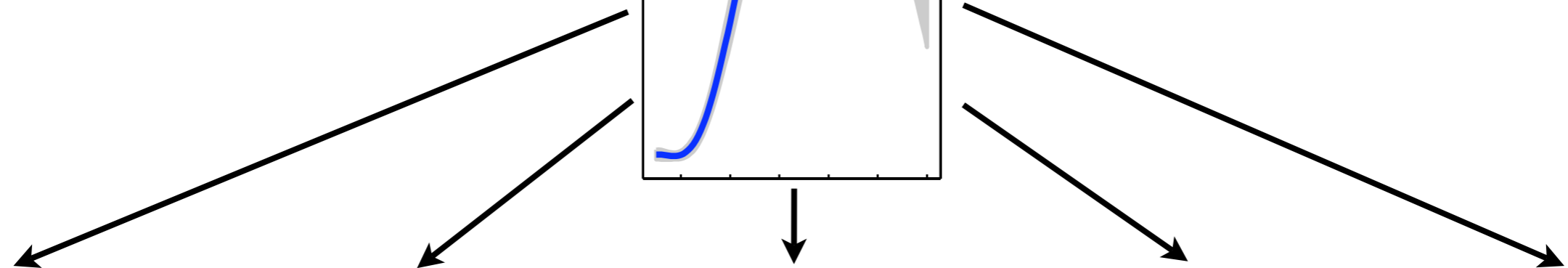
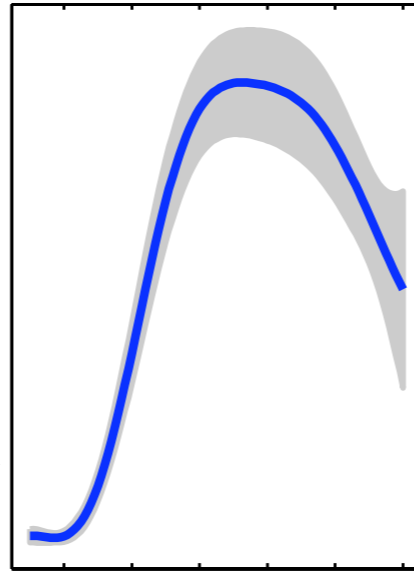
**– George Box (1919-)**

# Mallin sovittaminen

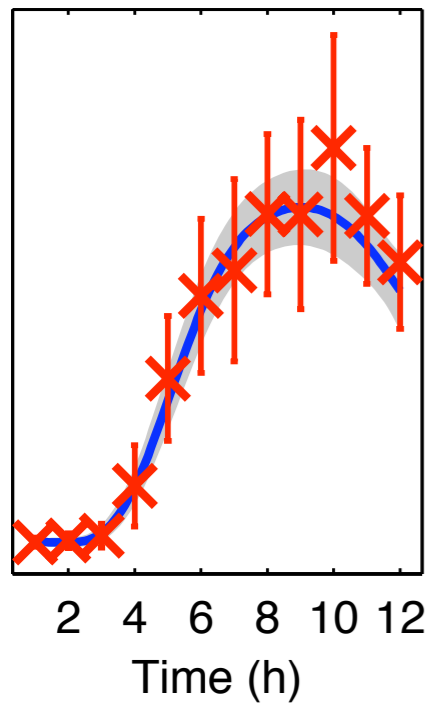
- Bayesiläinen päättely: priorista posterioriin
- Oletetaan tuntemattomille suureille Gaussin prosessia noudattava priorijakauma
  - Vastaava yhteisjakauma kaikille suureille
- Suurimman uskottavuuden menetelmään perustuva parametrien estimointi
- Tässä esimerkissä laskennallisesti tehokas

# Esimerkkimalli

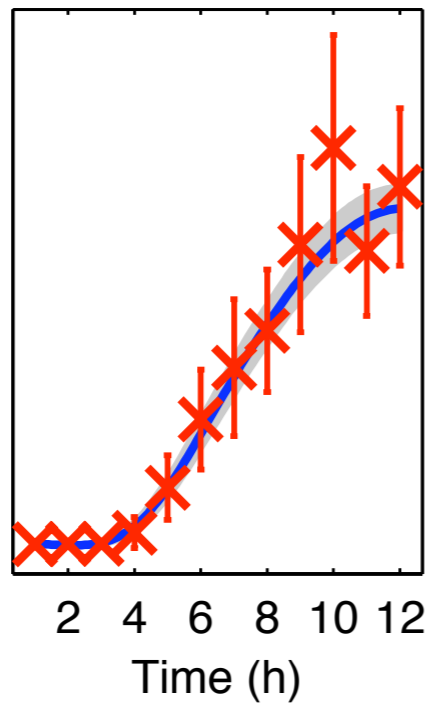
Inferred twi protein



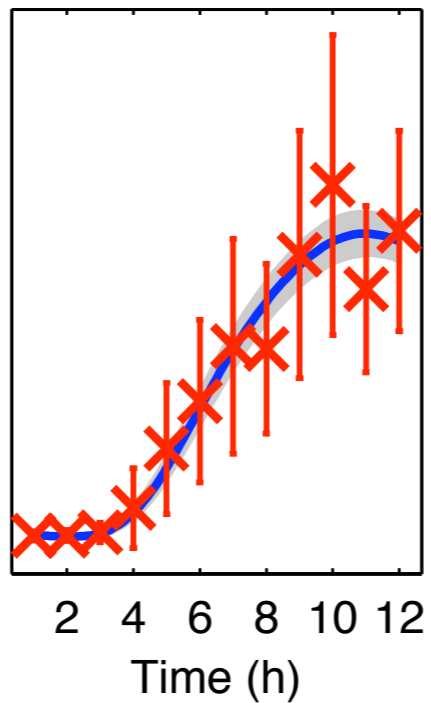
FBgn0033188 mRNA



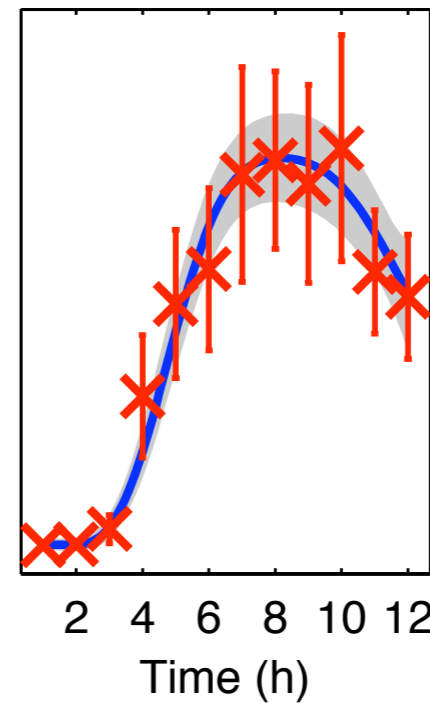
FBgn0035257 mRNA



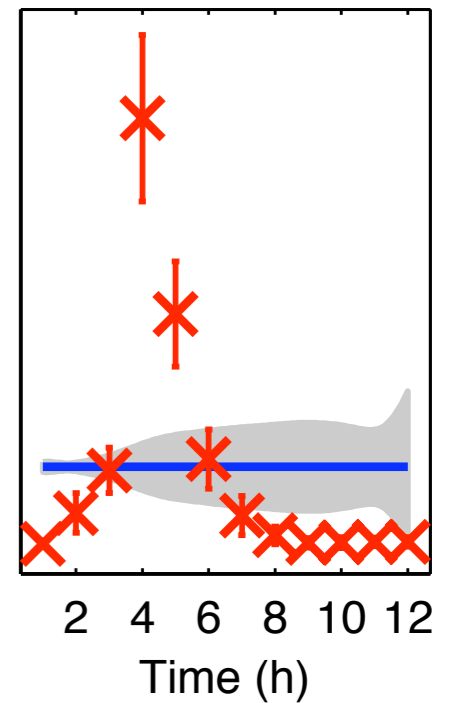
FBgn0011206 mRNA



FBgn0004646 mRNA



FBgn0039286 mRNA



# Sovelluksia

- Bioinformatiikassa eri tietolähteiden yhdistäminen olennaista
- Malli mahdollistaa ekspressioaikasarjojen käytön transkriptiofaktoreiden kohdegeenien tunnistamisessa
- Vaatii vain yksinkertaisia kokeita
- Ensimmäiset tulokset hyvin lupaavia

# Yhteenveto

- Koneoppimisessa pyrkimyksenä data-analyysin osittainen automatisointi
- Moderniin biologiaan liittyy paljon haastavia data-analyysitehtäviä