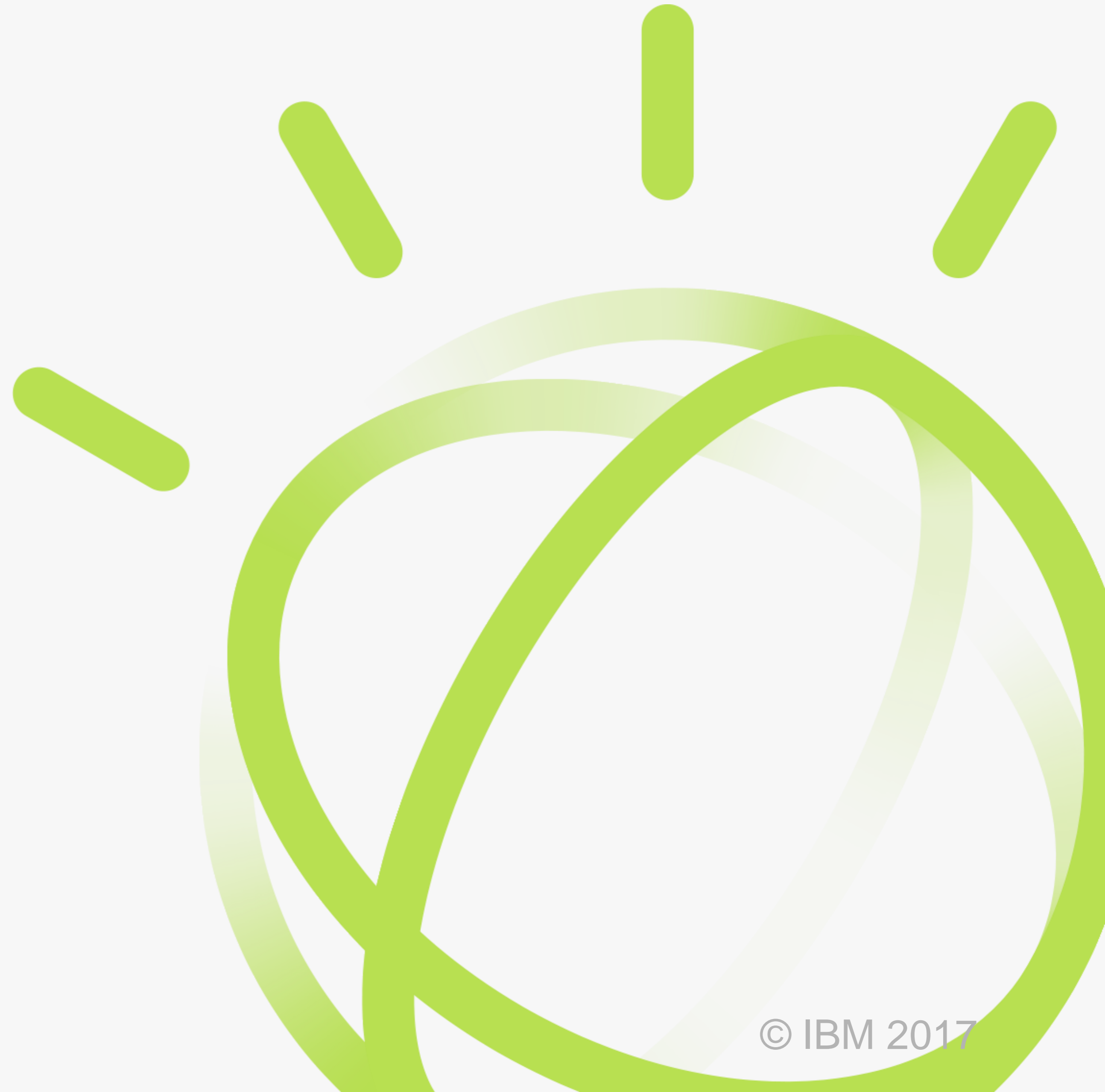
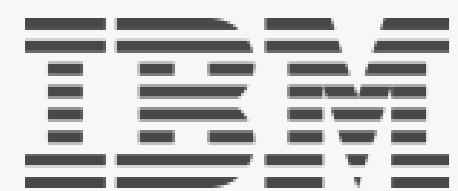


24.5.2017

Ennustemallien rakentamistekniikat

Antti Heino
IBM Advanced Analytics Team Lead



Agenda



Miksi koneoppiminen ja ennustemallit ovat tärkeitä terveydenhuollossa



Miten ennustemalleja rakennetaan



Esimerkkejä käyttötapauksista

Koneoppiminen ja tekoäly voivat vastata moniin terveydenhuollon haasteisiin



Mitä ovat koneoppiminen ja ennustemallit?

- **Koneoppimisen käyttö on valtavassa kasvussa, syynä ovat:**
 - Merkittävä kasvu datan määrässä ja laskentatehossa
 - Tarve automaattiselle oppimiselle datan perusteella, kaikkea ei pysty tarkasti ohjelmoimaan selkeiksi säännöiksi
- **Erilaisia lähestymistapoja koneoppimiseen**
 - **Supervised learning**
 - Koneelle annetaan opetusdata, johon on merkitty, mitkä lähtötiedot ovat johtaneet tiettyyn lopputulokseen
 - **Unsupervised learning**
 - Koneen annetaan itse löytää struktuuria ja säännönmukaisuuksia sille annetusta datasta
 - Pyrkimyksenä voi olla esimerkiksi jakaa datan tapaukset saman kaltaisiin joukkoihin
 - **Reinforcement learning**
 - Koneen annetaan toimia dynaamisessa ympäristössä, jossa sen pitää saavuttaa jokin päämäärä.
 - Kone hiljalleen pääsee lähemmäksi tavoitettaan kun se oppii mitkä tekijät ympäristössä auttavat sitä lähemmäksi tavoitetta ja mitä sen taas kannattaa välttää

MoleMap, Melanoma Institute Australia ja IBM



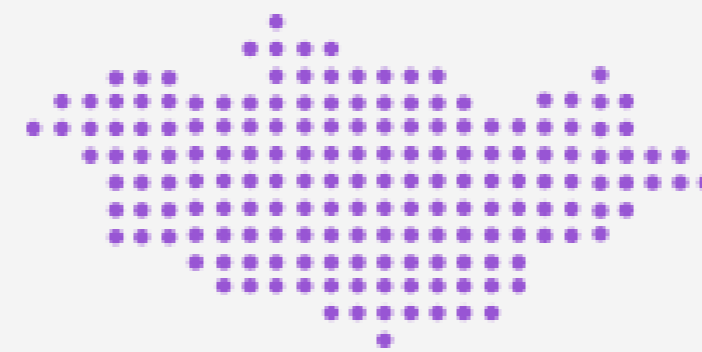
- Kuvat yksi runsas ja arvokas datalähde terveydenhuollossa
- Tutkimusyhteistyö, jossa konetta opetetaan tunnistamaan ihosyöpää käyttäen 40 000 tapauksen kuvapankkia
- Luomet kuvissa luokiteltu vaarallisiin ja harmittomiin sekä etenemisasteen mukaan
- Tällä hetkellä kuvat pystytään automaattisesti luokittelemaan 91% tarkkuudella
- Hyödyt
 - Parempi kyvykkyys seurata taudin etenemistä
 - Automatisoitavuus
 - Aikainen reagointi

Kone oppii tunnistamaan luomista erilaisia piirteitä



Colour

Watson analyses a skin mole and searches for the presence of any of the six suspicious colours, which range from tan to black. The more colours are identified, the higher the risk of melanoma.



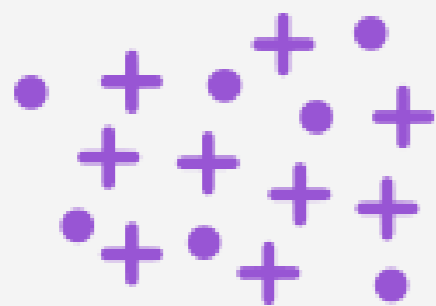
Border irregularity

Watson divides skin moles into 8 regions, and assesses the irregularity of their borders and assigns them a score. A total score is then calculated for the entire mole. The higher the score, the higher the likelihood of melanoma.



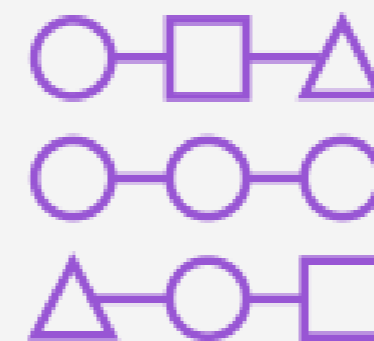
Asymmetry level

Watson analyses the major and minor axes of a skin mole to determine whether it is fully symmetrical, half symmetrical or fully asymmetrical. The more asymmetrical the mole, the higher the risk of melanoma.



Globule and network

Watson inspects the skin mole for the presence of any globular or network pattern, two patterns that are indicative of a potential melanoma, and calculates a percentage. The higher the percentage, the higher the risk.



Similar images

Watson searches the database for an image similar to the skin mole being analysed, and considers the prior diagnosis of this image.



Melanoma score

Watson scours through thousands of historical images in the database in a matter of seconds and uses an algorithm to assign a melanoma score to the skin mole being analysed. The higher the score, the higher the likelihood of melanoma.

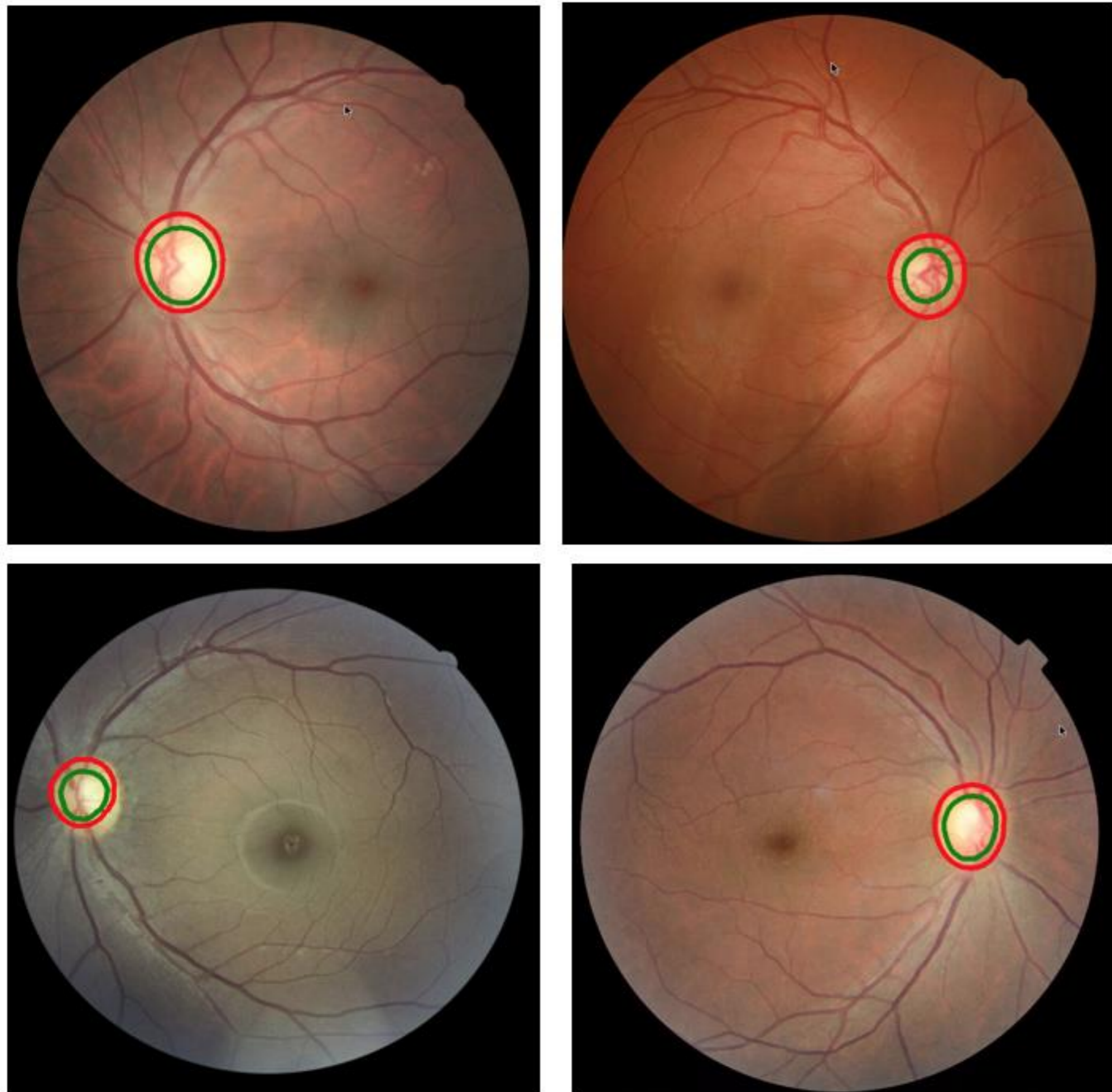
Ennustemallien rakentamisen vaiheet

- **Yleinen tavoite ennustemallien rakentamisessa on signaalin erottaminen satunnaisuuden keskeltä**
- **Löydetyn signaalin täytyy olla tilastollisesti merkittävä**
- **Feature extraction**
 - Muuttujien luominen
 - Automatisoidusti datan pohjalta (esim. eri keskustelujen aiheiden tunnistaminen käytettyjen sanojen avulla)
 - Asiantuntijoiden avulla (hyödyllistä myös automaattisten tulosten ymmärtämiseksi)
- **Regularization**
 - 2 ääripäätä
 - Konservatiivisuus, sama ennuste kaikille – sivuuttaa signaalin ja satunnaisuuden
 - Kompleksisuus, mallin opettaminen liian tarkaksi – ottaa huomioon signaalin ja satunnaisuuden
 - Ratkaisuna keskitie, jolla malli on riittävän yleistettävissä. Se huomio signaalit, mutta ei satunnaisia tapahtumia
- **Cross-validation**
 - Ennustemallin yleistettävyyden testaus datalla, jota ei ole käytetty sen opettamiseen

Algoritmi valittava tapauksen mukaan

- **Algoritmityyppi valitaan ennustettavan kohteen mukaan**
 - **Classification** = luokittelu kahteen tai useampaan luokkaan
 - **Regression** = jatkuvan arvon ennustaminen
 - **Clustering** = tapahtumien jaottelu samankaltaisiin ryhmiin
- **Päätöspuut (Random forest, CHAID, CRT etc.)**
 - Hyviä kun täytyy ymmärtää ennustemallin antamaa tulosta
- **Artificial neural networks**
 - Neuroverkko oppii muuttujien epälineaariset riippuvuussuhteet suoraan havaintoaineistosta
- **Deep learning**
 - Useiden virtuaalisten neuronitasojen hyödyntäminen neuroverkossa hyödyntämällä massiivista laskentatehoa
- **Support vector machines**
- **Bayesian networks**
 - Esim. oireiden perusteella verkosta voidaan laskea todennäköisyys eri taudeille
- **Association rule learning**
 - Yhteyksien havaitseminen, esim. ostoskorianalyysi
- **Similarity learning**
- **Reinforcement learning**

Silmänpainetauti – glaukooman tunnistaminen

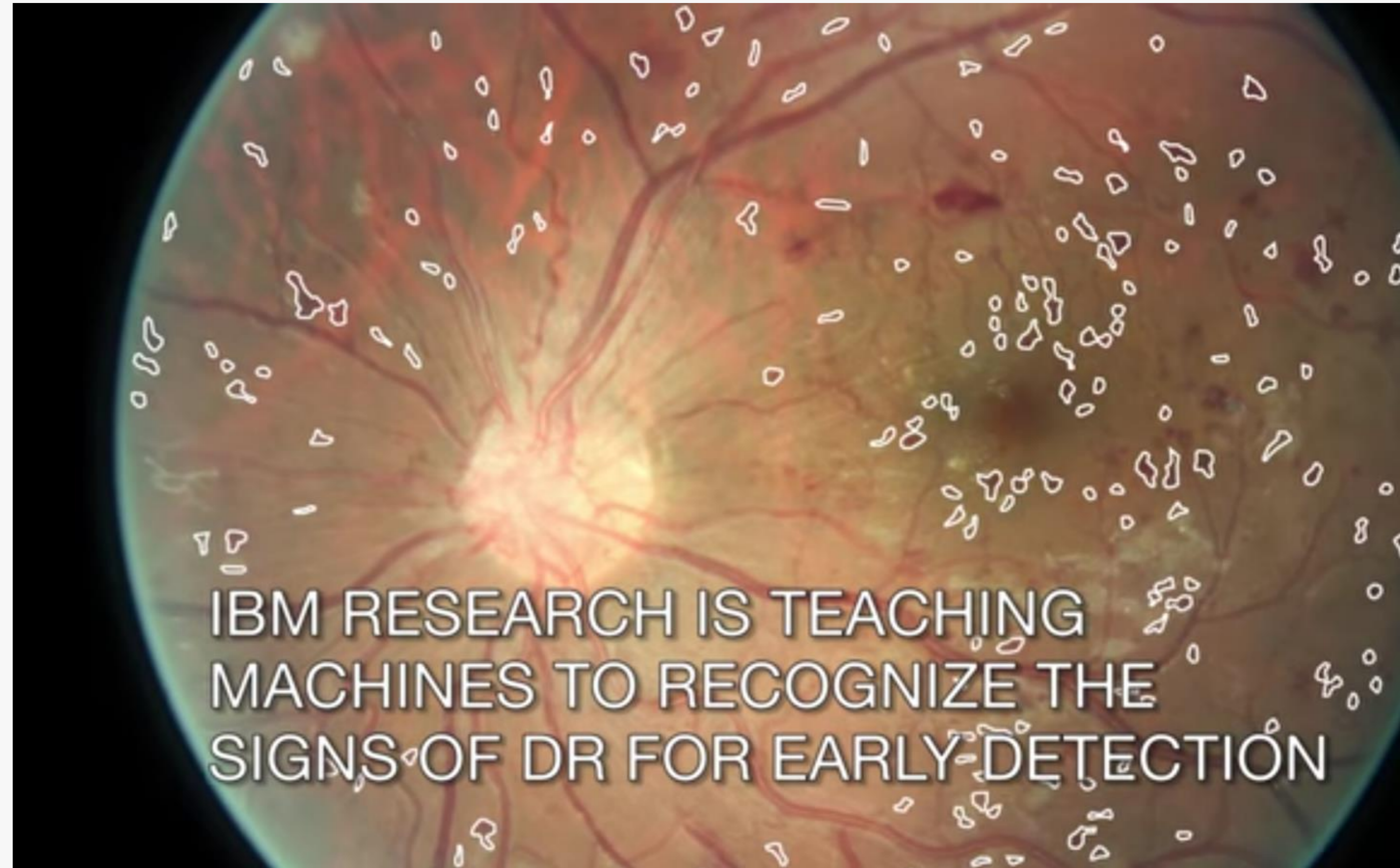


- IBM aloittanut 2015 tutkimustyön, jossa silmänpohjakuvia analysoitu koneoppimisen keinoin
- Käytössä ollut 88 000 kuvan datasetti, josta on pystytty analysoimaan glaukooman riskitekijöitä.
- Glaukooma on näköhermoa rappeuttava sairaus, joka on usein yhteydessä liian korkeaan silmänpaineeseen
- Yksi suurimmista sokeuden aiheuttajista
- Havaitaan yleensä myöhään ja sattumalta
- Koneellisesti pystytään tunnistamaan hyvin tarkasti (95%) esimerkiksi optic cupin ja discin eroaminen, joka on kohonneen silmänpaineen oire

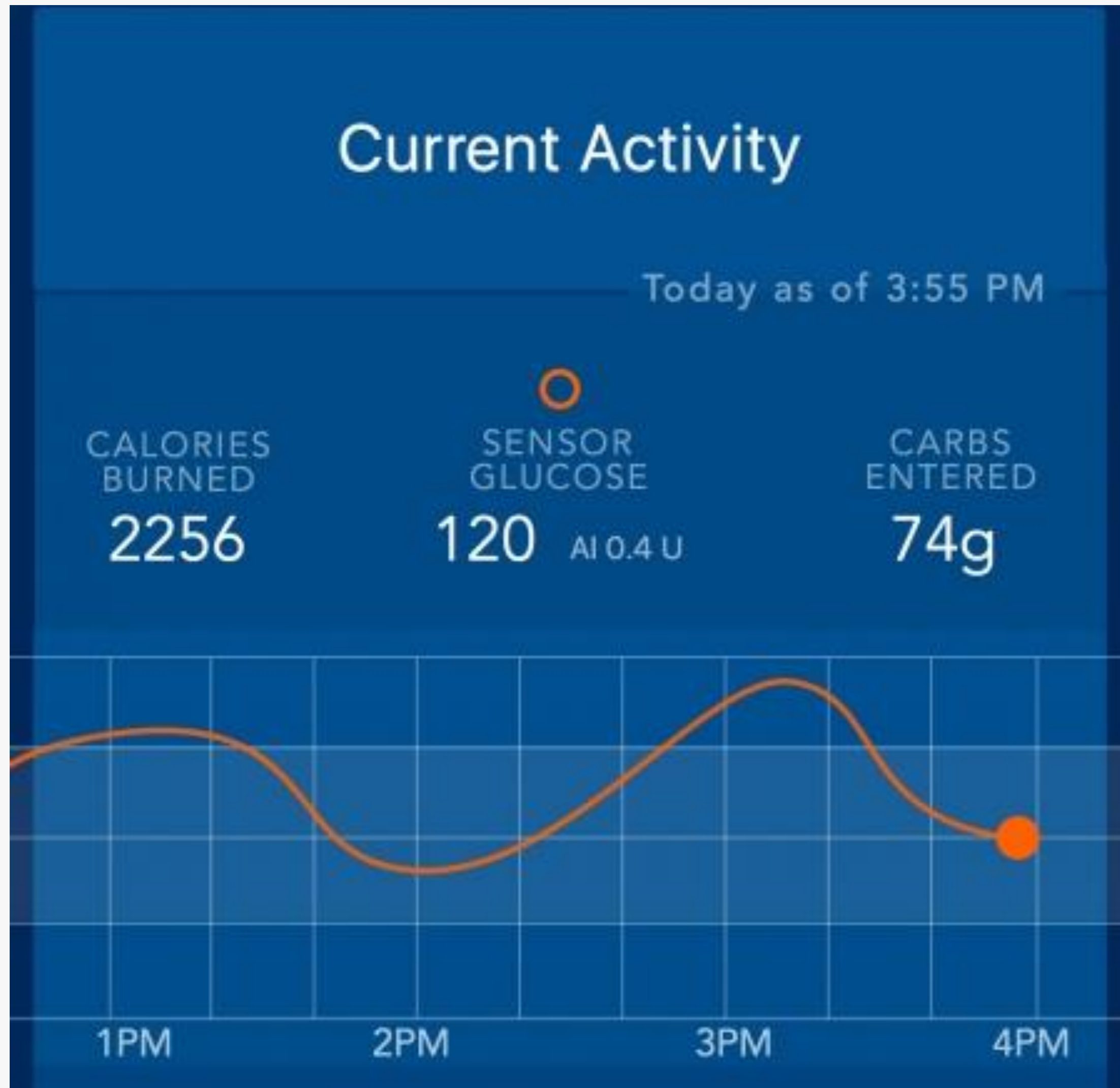


Diabeettisen retinopatian tunnistaminen

- Maailman 400 miljoonasta diabeetikosta noin kolmasosalle kehittyy diabeettinen retinopatia
- Voi johtaa sokeuteen jos jätetään hoitamatta
- Liian suuri sokerimäärä vaikuttaa verisuonten valkuaisaineisiin ja saa pienet suonet vuotamaan verkkokalvolle. Lisäksi se kasvattaa happiradikaalien määrään ja vaurioittaa myös silmän hermoston rakenteita.
- Deep learningiä ja neuroverkkoja hyödyntäen on voitu luoda luokittelija, joka 86% tarkkuudella tunnistaa DRn vakavuusasteen
- Tehty käyttäen 35 000 kuvan datasettiä



Medtronic – hypoglykemiaan ennustaminen



- N. 400 miljoonaa diabeteksen kanssa elävää maailmassa, joista koituu noin 600 miljardin kustannukset
- Medtronic on kerännyt insuliinipumpuistaan ja glukosimonitoreistaan 125 miljoonan potilaspäivän datan
- Yhdessä IBM:n kanssa luotu ennustemalli, joka pystyy ennustamaan hypoglykemiaa 3-4 tuntia etukäteen
- Analysoitu valtavaa datamassaa, josta voitu nähdä piirteitä tilanteita, jotka ovat johtaneet hypoglykemiaan
- Sugar iQ app

Sepsiksen ennustaminen keskosteholla – HUS & IBM

