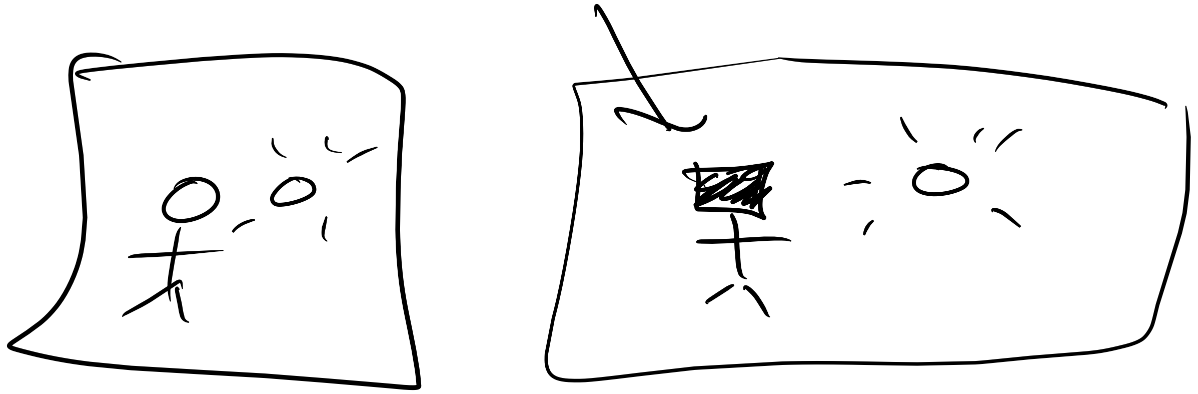
Machine learning landscape

Learning paradigm

Supervised (övervakad)

Selfsupervised →



semi-supervised

unsupervised (oövervakad)

reinforcement learning (självinlärning)

Task

Regression

Classification

Clustering

Dimensionality reduction (compression)

Anomaly detection

Novelty detection

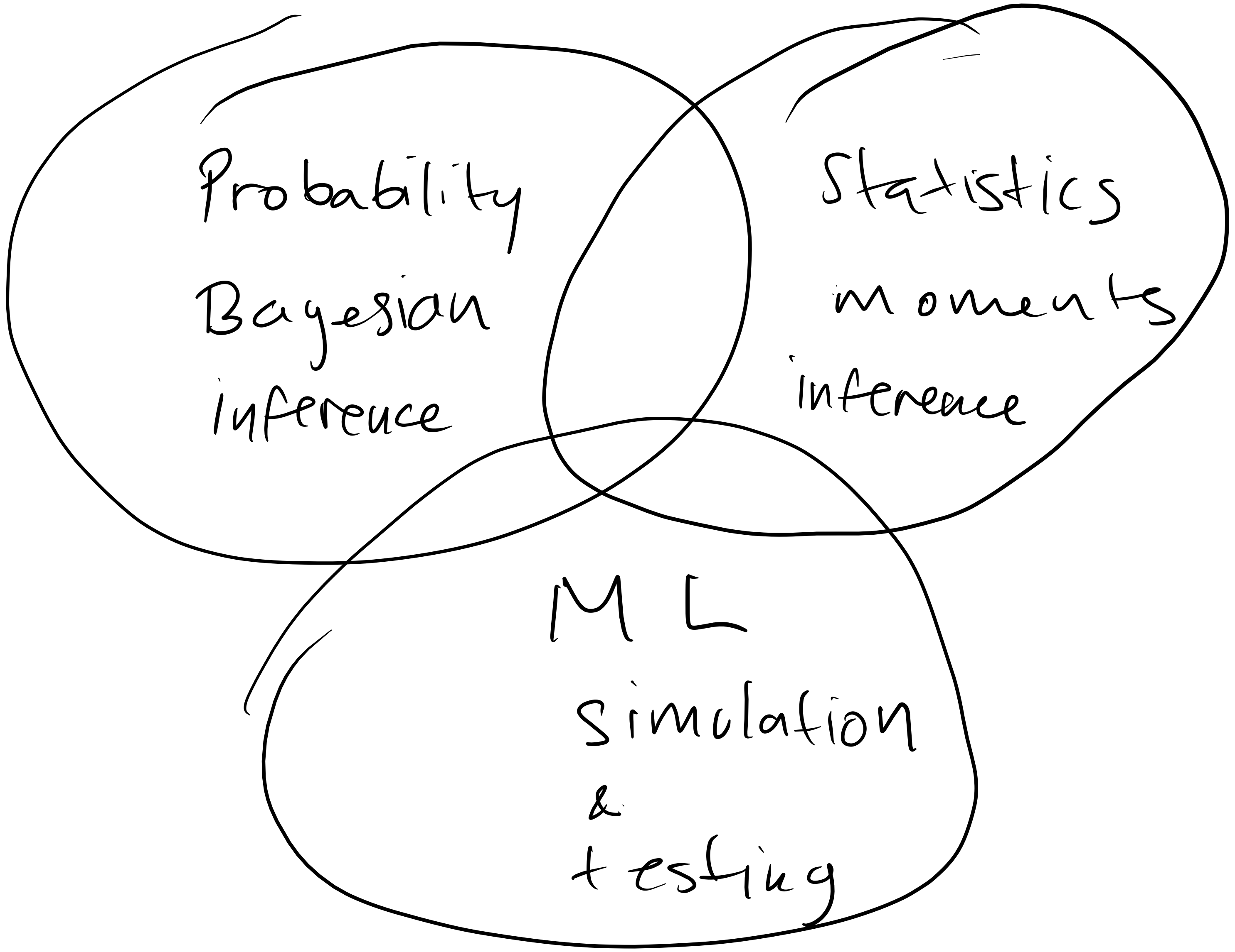
Training

- Batch
- Online

Model type

instance-based (non-generalising)

model-based



$(\mathbb{E}[X], \mathbb{E}[X^2], \text{etc})$

Allmän linjär regression (General Linear Models)

F-test

t-test

ANOVA, ANCOVA

I statistiken förenklade vi till en responsvariabel Y, men om vi har kategoriska kolonner kan vi se Y som flera $Y_1, Y_2, \ldots$

Linjär, Multipel, Polynomiell, Poisson, Logistisk ($[0,1]$)

Normalantagande "överallt" $Y = X\beta + E$

Generalized Linear Models

$$E[Y|X] = \mu = g^{-1}(\mathbb{X}\mathbb{B})$$

$$Var(Y|X) = V(g^{-1}(\mathbb{X}\mathbb{B}))$$

We need 3 things:

An exponential model for Y

A linear estimator $\eta = \mathbb{X}\mathbb{B}$

A link function g

Data	Distribution	Link function	
$(-\infty, \infty)$	normal	id	Linear Regression
$(0, \infty)$	exp., Gamma	neg. inv $X\beta = -\mu^{-1}$	(inverse normal)
$[0, 1, \ldots]$	Bernoulli	$X\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	logistic
$\mathbb{N}$	Binomial	$X\beta = \ln\left(\frac{\mu}{n-\mu}\right)$	binom. regression

Maximum Likelihood

(även Gibbs-sampling, en Monte-Carlo metod)

logit $\mu = \dfrac{1}{1+e^{-X\beta}}$ ←

Gradient descent