Bayesian Machine Learning for Scientific Research

Maximizing information from data

Ezequiel Alvarez
sequi@unsam.edu.ar
ICTP-SAIFR October 2024

Previous lecture

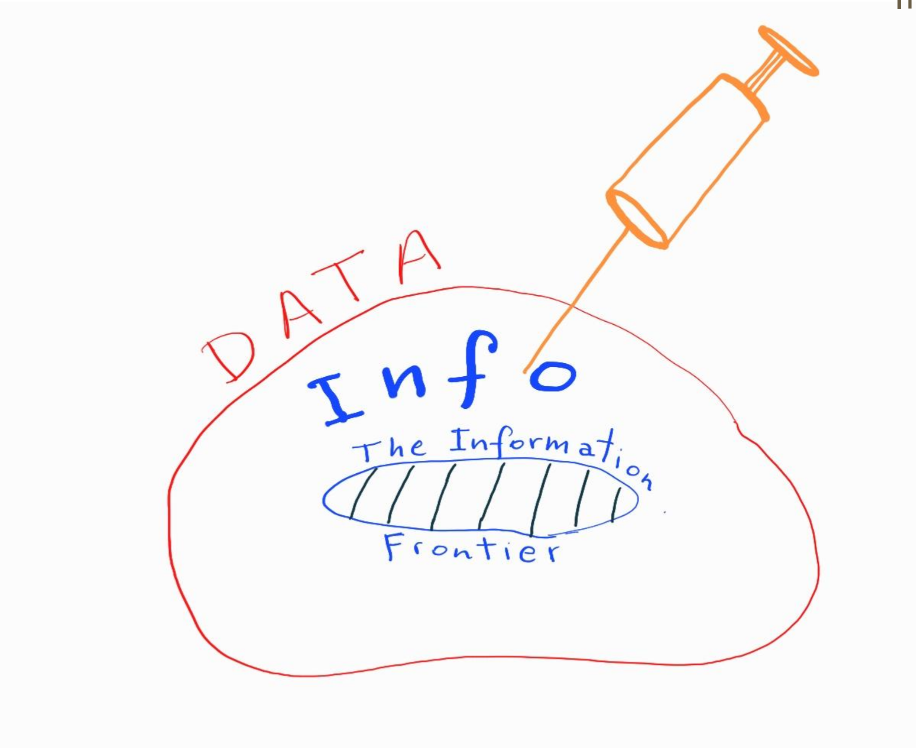
Previous lecture

Bayesian: assume data being sampled from a PDF, infer its parameters and learn the internal structure of the data

To learn the PDF of the data and then being able to assess, predict , generate, etc.

More scientific

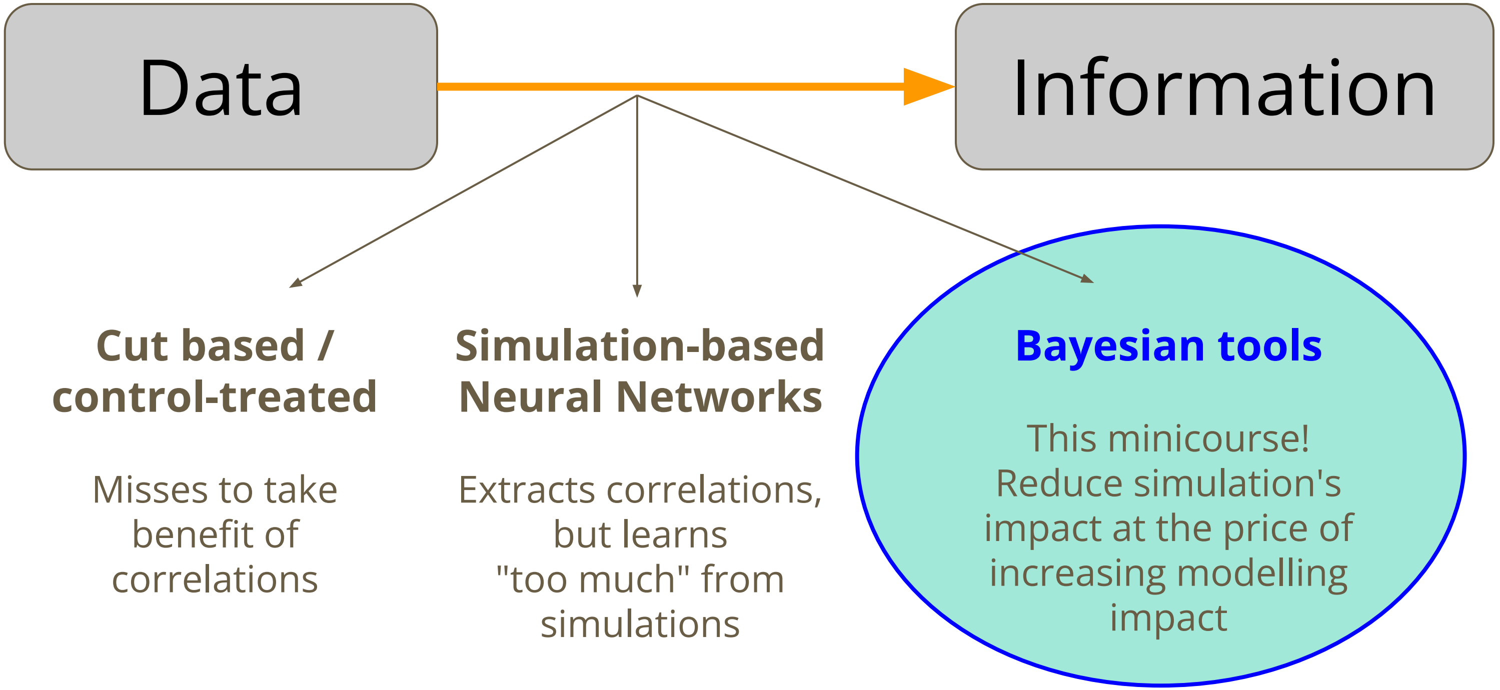
Previous lecture



Inject *catalysts*:

- Modeling
- Tools & techniques
- Prior info

Previous lecture

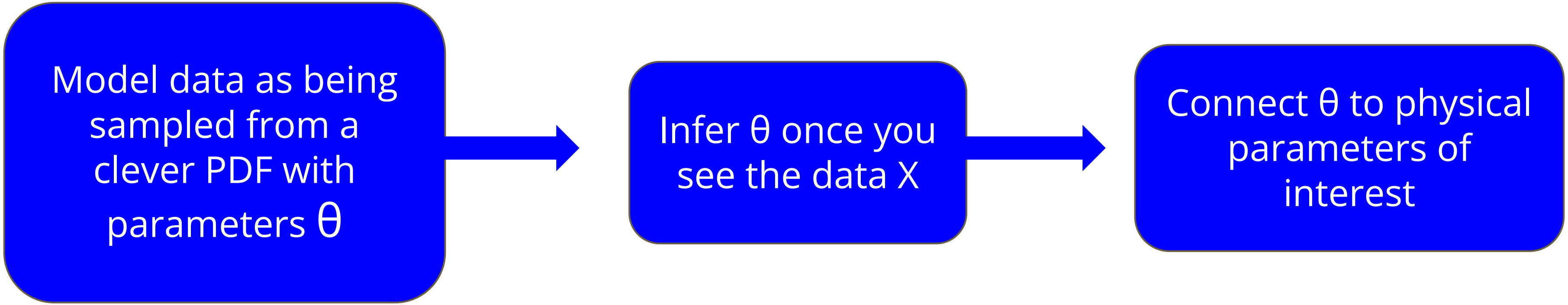


Previous lecture

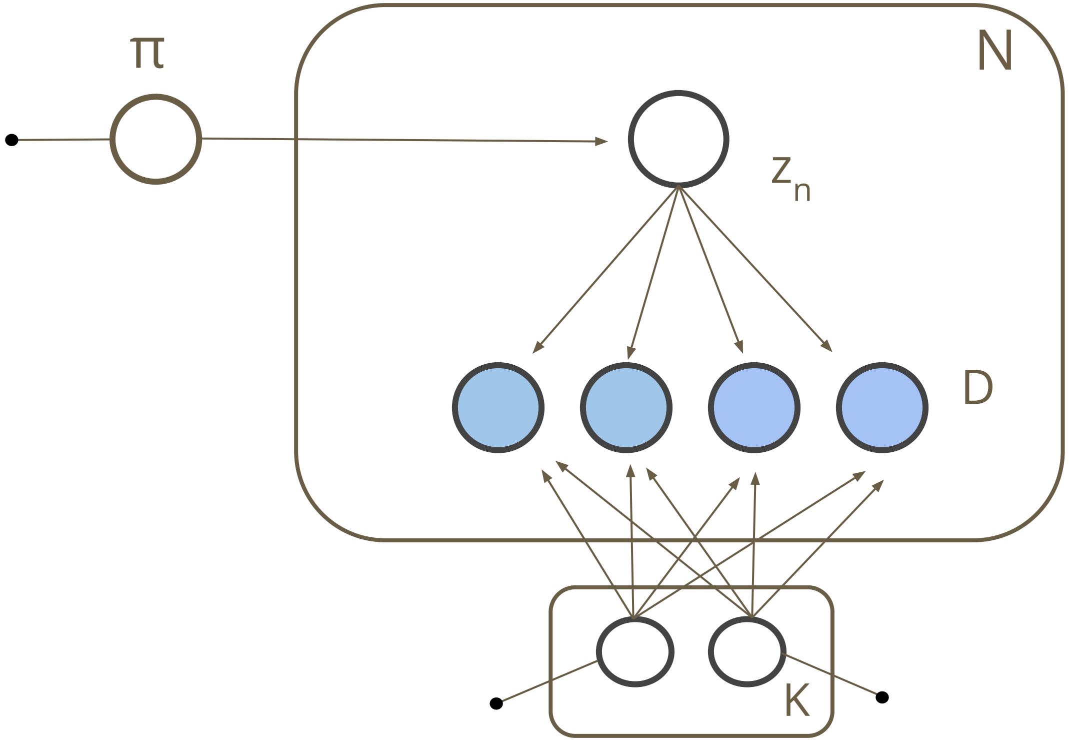
Bayes Theorem:

$$p(\theta|X) = \frac{p(X|\theta) * p(\theta)}{p(x)}$$

Our utility: X = data, θ=parameters

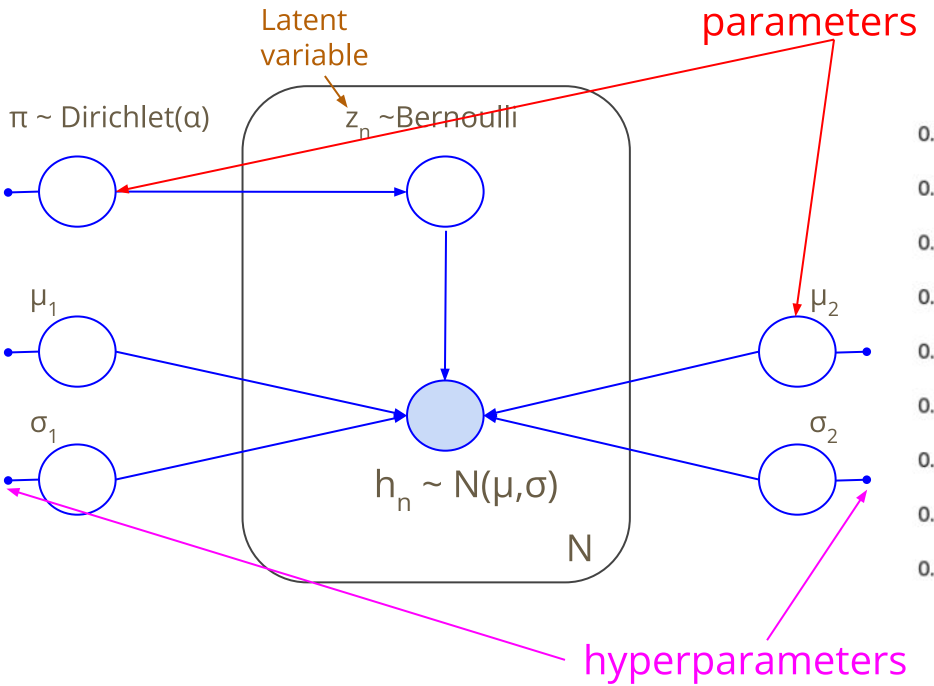


Previous lecture

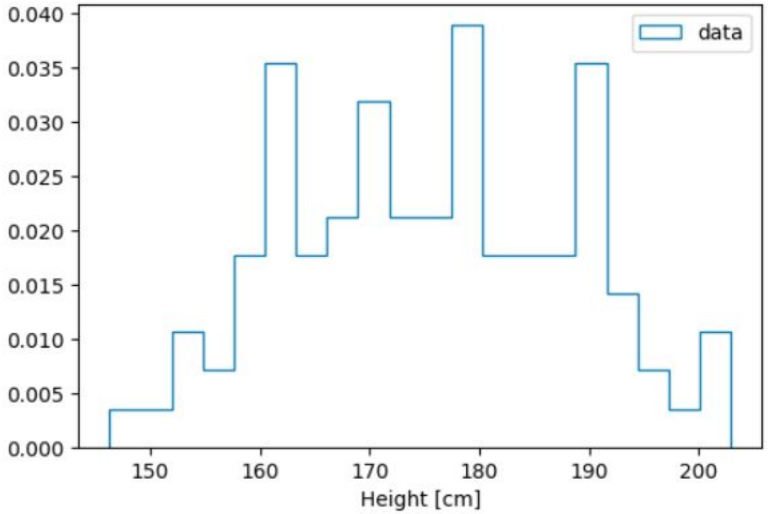


Mixture Model

Previous lecture

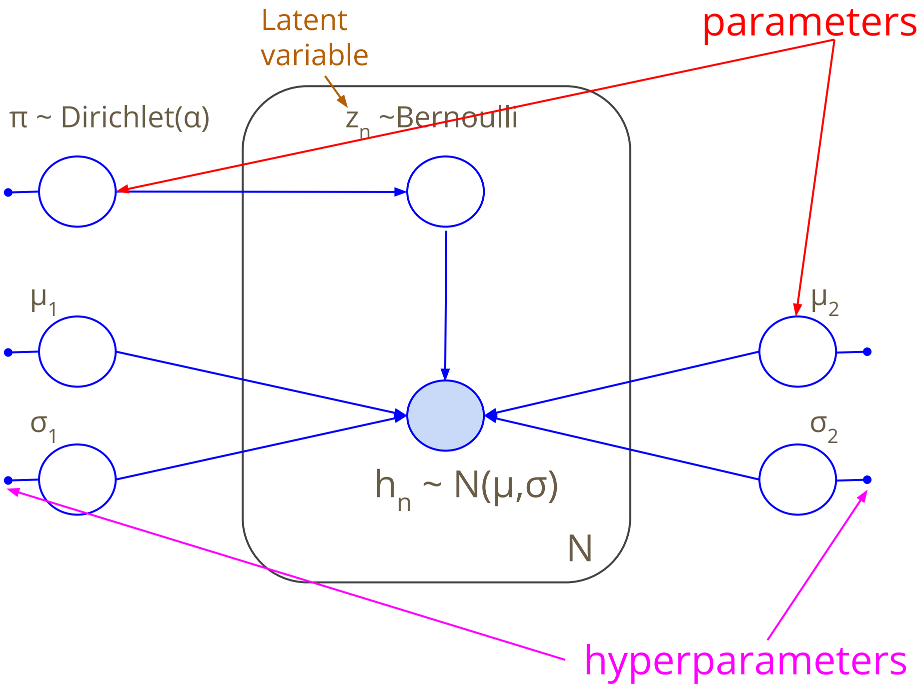


Graphical Model



1D Gaussian Mixture

Previous lecture



Graphical Model

$X : \{ h_n \}$

$\theta : \mu_{1,2}, \sigma_{1,2}, \pi, \{ z_n \}$

$P(x_n | \theta) = \pi \, N(h_n;\mu_1,\sigma_1) + (1-\pi) \, N(h_n;\mu_2,\sigma_2)$

Bayesian Machine Learning:
The Tool

Lecture 2

Summary

Art et métiers in proposing models

- Analytic Bayesian Inference: The box factory
- Linear Regression: River contamination
- Anomalies in collective motion: Clinical practices
- Mixture of Bernoulli: Unsupervised Learning handwritten

Connection to scientific research

Simple explicit Bayesian problem

The Box Factory

The box factory

Box OK: p
Box wrong: $1-p$

The box factory

Box OK: *p*
Box wrong: *1-p*

Experiment: k boxes ok out of n

X = k, n
Θ = p

The box factory

Box OK: *p*
Box wrong: *1-p*

Experiment: k boxes ok out of n

X = k, n
Θ = p

$$p(X|\theta) = \frac{n!}{k!\,(n-k)!} p^k \,(1-p)^{n-k}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Experiment: k boxes ok out of n

X = k, n
Θ = p

$$p(X|\theta) = \frac{n!}{k!\,(n-k)!} p^k\,(1-p)^{n-k}$$

$$p(\theta|X) = \frac{p(X|\theta)\,p_{prior}(\theta)}{p(X)}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Experiment: k boxes ok out of n

X = k, n
Θ = p

$$p(X|\theta) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

$$p(\theta|X) = \frac{p(X|\theta)\, p_{prior}(\theta)}{\int p(X|\theta)\, p_{prior}(\theta)\, d\theta}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Experiment: k boxes ok out of n

X = k, n
Θ = p

$$p(X|\theta) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

$$p(\theta|X) = \frac{p(X|\theta)\, p_{prior}(\theta)}{\int p(X|\theta)\, p_{prior}(\theta)\, d\theta}$$

=1

=1

(new factory!)

The box factory

Box OK: *p*
Box wrong: *1-p*

Experiment: k boxes ok out of n

X = k, n
Θ = p

$$p(X|\theta) = \frac{n!}{k!\,(n-k)!} p^k\,(1-p)^{n-k}$$

$$p(\theta|X) = \frac{p(X|\theta)\,p_{prior}(\theta)}{\int p(X|\theta)\,p_{prior}(\theta)\,d\theta}$$

=1

=1

$$p(\theta|X) = \frac{p^k\,(1-p)^{n-k}}{\int_0^1 p^k\,(1-p)^{n-k}\,dp}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Experiment: k boxes ok out of n

X = k, n
Θ = p

$$p(X|\theta) = \frac{n!}{k!\,(n-k)!} p^k \,(1-p)^{n-k}$$

$$p(\theta|X) = \frac{p(X|\theta)\, p_{prior}(\theta)}{\int p(X|\theta)\, p_{prior}(\theta)\, d\theta}$$

=1

=1

$$p(\theta|X) = \frac{1}{B(n,k)}\, p^k \,(1-p)^{n-k}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Experiment: k boxes ok out of n

X = k, n
Θ = p

$$p(\theta|X) = \frac{p(X|\theta)\, p_{prior}(\theta)}{\int p(X|\theta)\, p_{prior}(\theta)\, d\theta}$$

Today's posterior is tomorrow's prior!

$$p(\theta|X) = \frac{1}{B(n,k)}\, p^k\, (1-p)^{n-k}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Tomorrow: k' boxes ok out of n'

X = k', n'
Θ = p

$$p(\theta|X) = \frac{p(X|\theta)\, p_{prior}(\theta)}{\int p(X|\theta)\, p_{prior}(\theta)\, d\theta}$$

Today's posterior is tomorrow's prior!

$$p(\theta|X) = \frac{1}{B(n,k)}\, p^k\, (1-p)^{n-k}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Tomorrow: k' boxes ok out of n'

X = k', n'
Θ = p

$$p(\theta|X) \propto \frac{p^{k'}(1-p)^{n'-k'}\,p^{k}(1-p)^{n-k}}{\int_0^1 p^{k'}(1-p)^{n'-k'}\,p^{k}(1-p)^{n-k}dp}$$

Today's posterior is tomorrow's prior!

$$p(\theta|X) = \frac{1}{B(n,k)}p^{k}(1-p)^{n-k}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Tomorrow: k' boxes ok out of n'

X = k', n'
Θ = p

$$p(\theta|X) \propto \frac{p^{k'+k}(1-p)^{n'+n-k'-k}}{\int_0^1 p^{k'+k}(1-p)^{n'+n-k'-k}dp}$$

Today's posterior is tomorrow's prior!

$$p(\theta|X) = \frac{1}{B(n,k)} p^k (1-p)^{n-k}$$

The box factory

This only occurs in very few cases known as Conjugate Posterior

Box OK: *p*
Box wrong: *1-p*

Tomorrow: k' boxes ok out of n'

X = k', n'
Θ = p

$$p(\theta|X) \propto \frac{p^{k'+k}(1-p)^{n'+n-k'-k}}{\int_0^1 p^{k'+k}(1-p)^{n'+n-k'-k}dp}$$

Today's posterior is tomorrow's prior!

$$p(\theta|X) = \frac{1}{B(n,k)} p^k (1-p)^{n-k}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Tomorrow: k' boxes ok out of n'

X = k', n'
Θ = p

$$p(\theta|X) = \frac{1}{B(n'+n, k'+k)} p^{k'+k}(1-p)^{n'+n-k'-k}$$

Today's posterior is tomorrow's prior!

$$p(\theta|X) = \frac{1}{B(n,k)} p^{k} (1-p)^{n-k}$$

The box factory

Box OK: *p*
Box wrong: *1-p*

Tomorrow: k' boxes ok out of n'

X = k', n'
Θ = p

$$p(\theta|X) = \frac{1}{B(n'+n, k'+k)} p^{k'+k}(1-p)^{n'+n-k'-k}$$

- We have updated posterior with data
- Posterior conjugate → analytic!

The box factory

Box OK: p
Box wrong: $1\text{-}p$

Tomorrow: k' boxes ok out of n'

X = k', n'
Θ = p

$$p(\theta|X) = \frac{1}{B(n'+n, k'+k)} p^{k'+k}(1-p)^{n'+n-k'-k}$$

- We have updated posterior with data
- Posterior conjugate → analytic!
- Observe how data can erase priors!

($n', k' >> n, k$)

The box factory

Box OK: *p*
Box wrong: *1-p*

Tomorrow: k' boxes ok out of n'

X = k', n'
Θ = p

$$p(\theta|X) = \frac{1}{B(n'+n, k'+k)} p^{k'+k}(1-p)^{n'+n-k'-k}$$

- We have updated posterior with data
- Posterior conjugate → analytic!
- Observe how data can erase priors!

(*n', k' >> n, k*)

From now on and most of the times the update will be numerical (Stan, PYMC, etc)

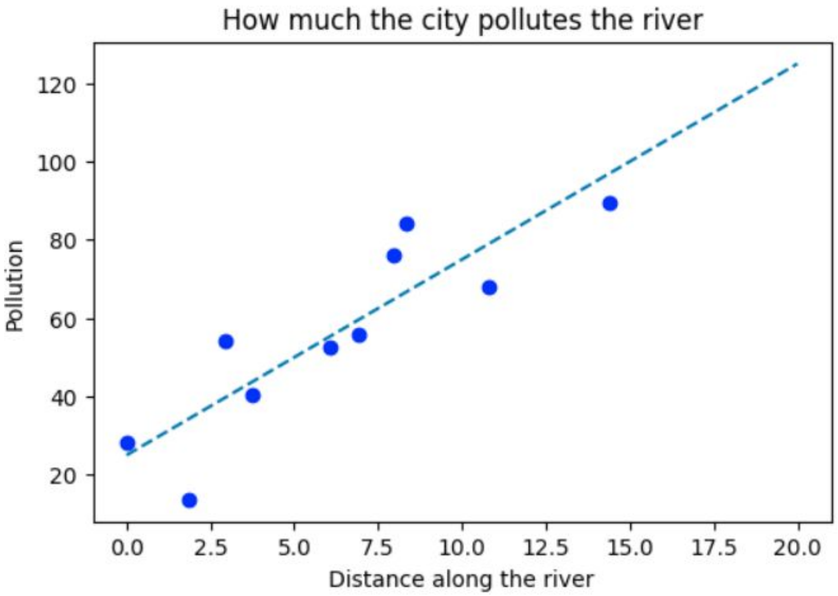
Linear Regression

A Hello World example!

Linear Regression

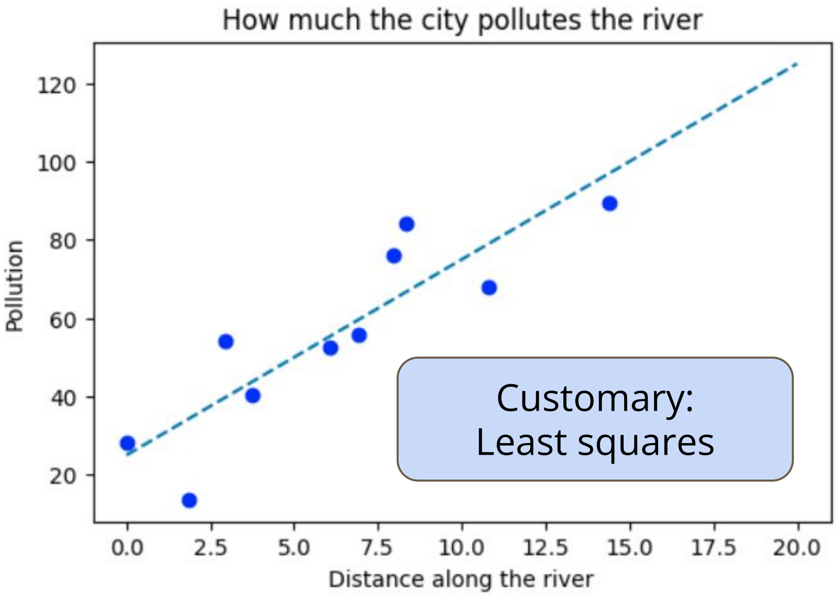
How much pollution comes from the city ?

Linear Regression



How much pollution comes from the city ?

Linear Regression



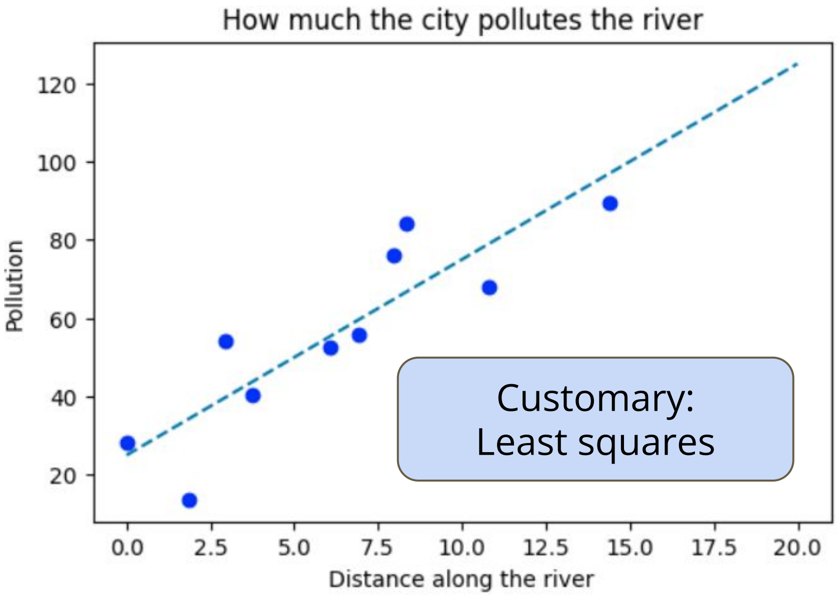
How much pollution comes from the city ?

Linear Regression

Bayesian

Model that *y* is sampled from a Normal around a (unknown) straight line:

y ~ normal(m * x + b, σ)

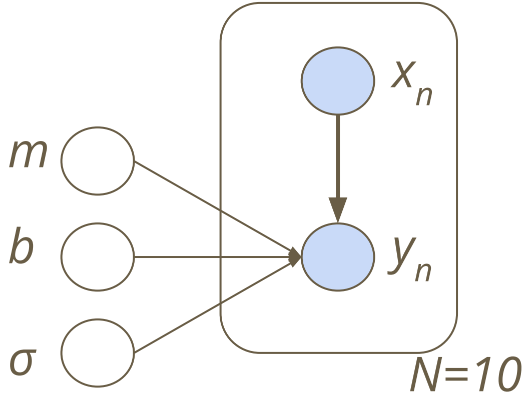


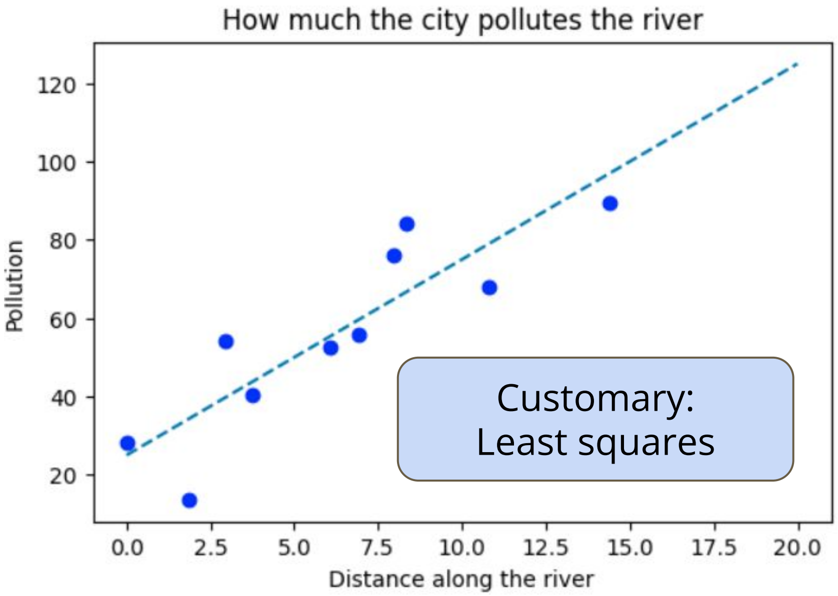
Linear Regression

Bayesian

Model that *y* is sampled from a Normal around a (unknown) straight line:

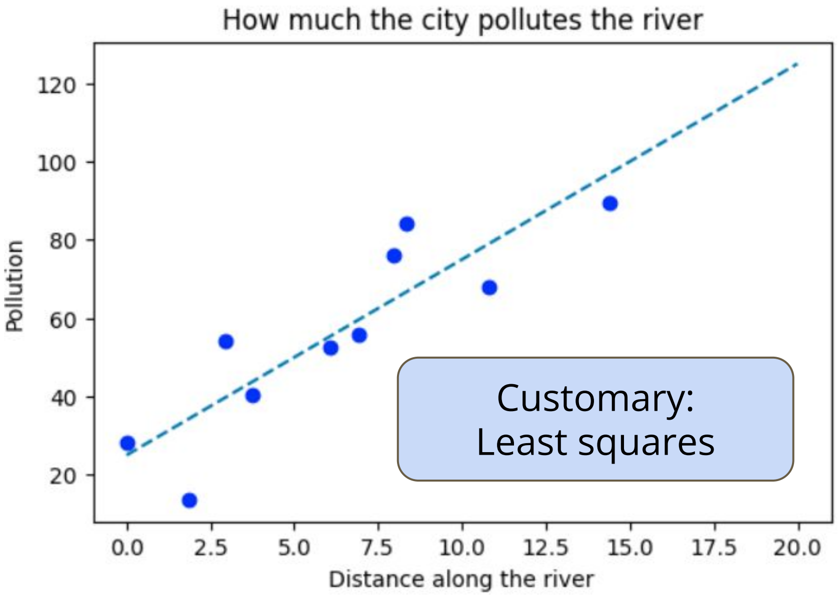
y ~ normal(m * x + b, *σ*)





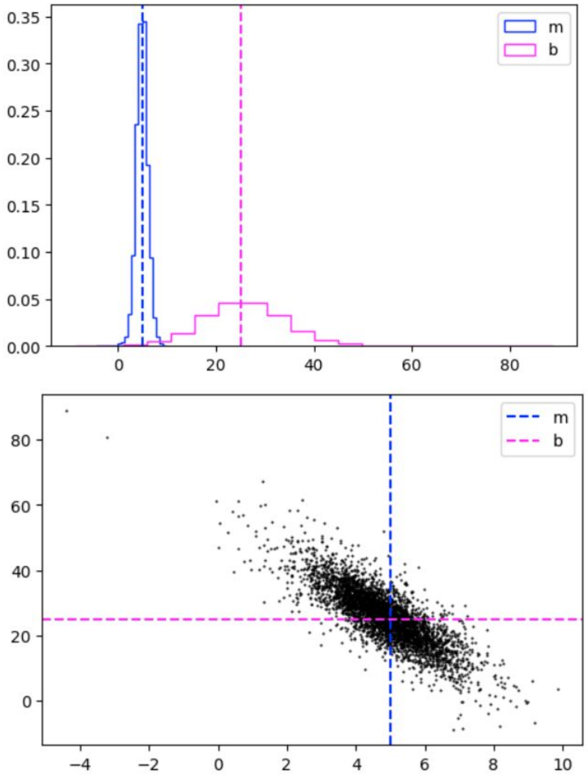
Linear Regression

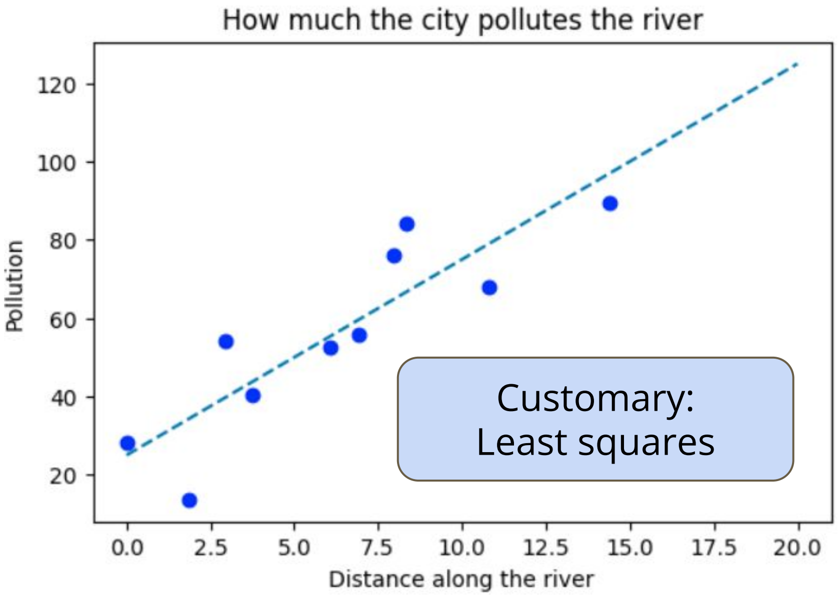
```
contamination_model="""
data {
 int <lower=0> N;
 vector[N] x;
 vector[N] y;
}

parameters {
  real m;
  real b;
  real<lower=0> sigma;
}
model {
  //b ~ normal(10,10);
  y ~ normal(m * x + b, sigma);
}
"""
```



Customary:
Least squares

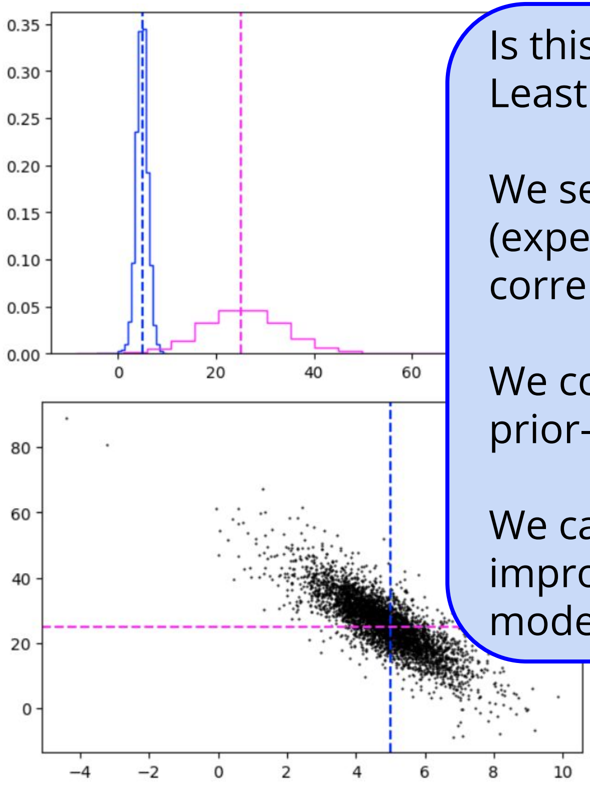
Linear Regression





Inference results

Linear Regression

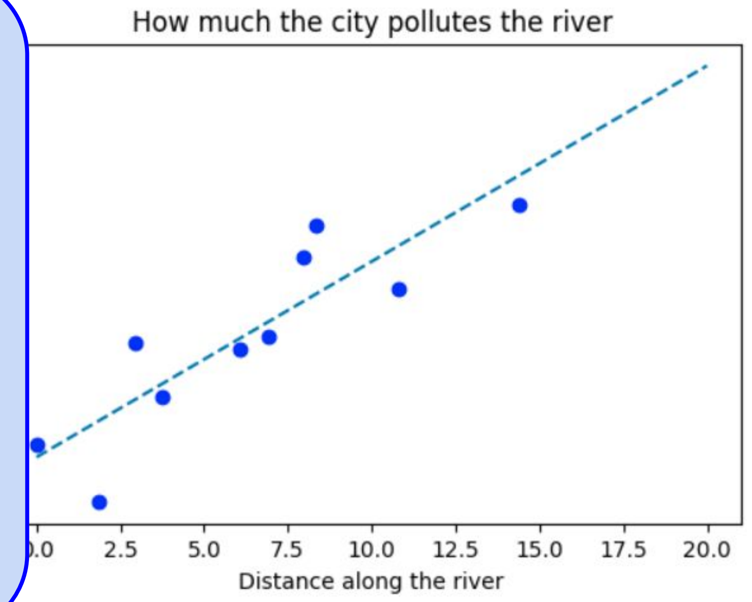


Is this better than Least Squares?

We see some (expected) correlation

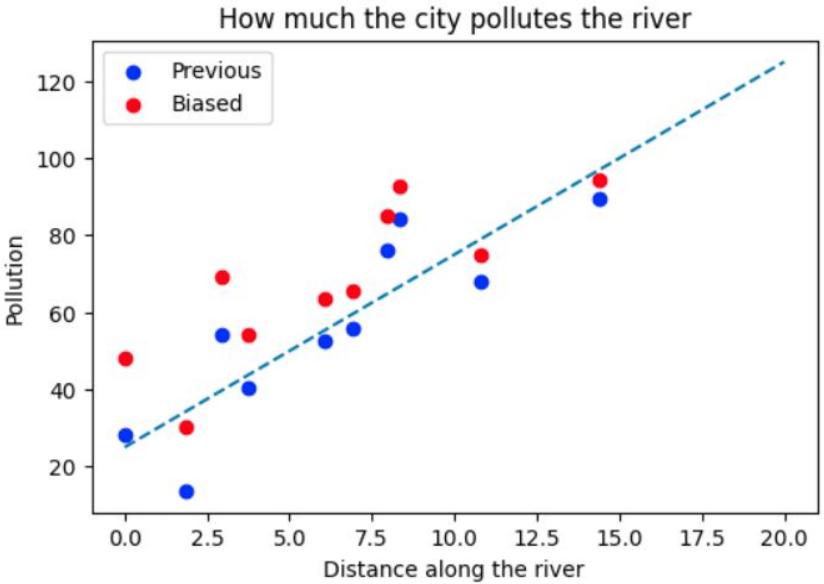
We could add prior-information

We can easily improve/adapt the model!



Inference results

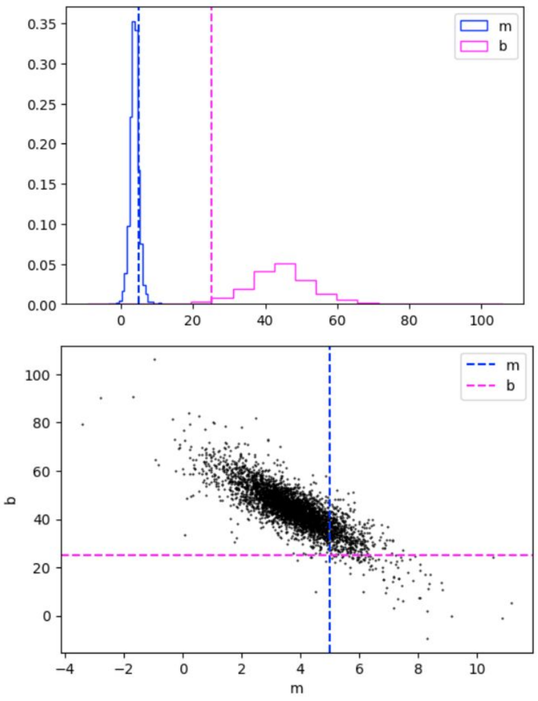
Linear Regression

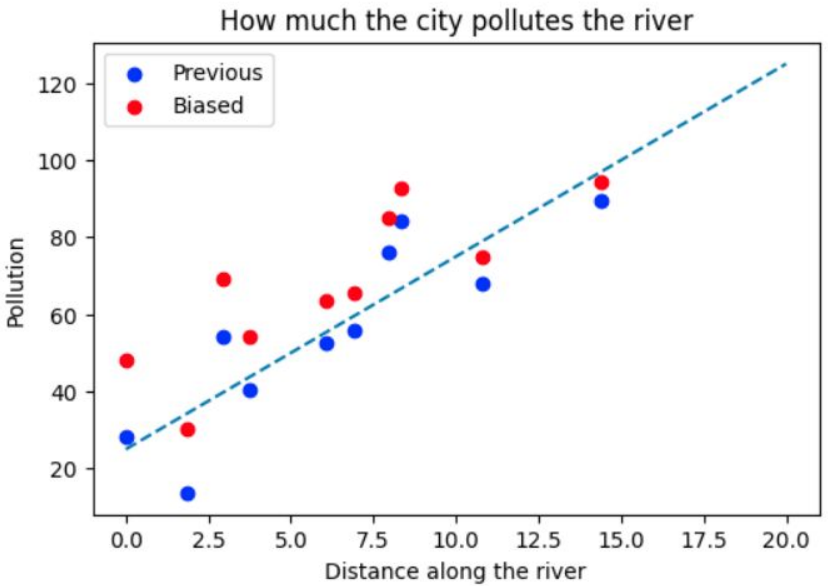


Expect more pollution from downtown

Linear Regression

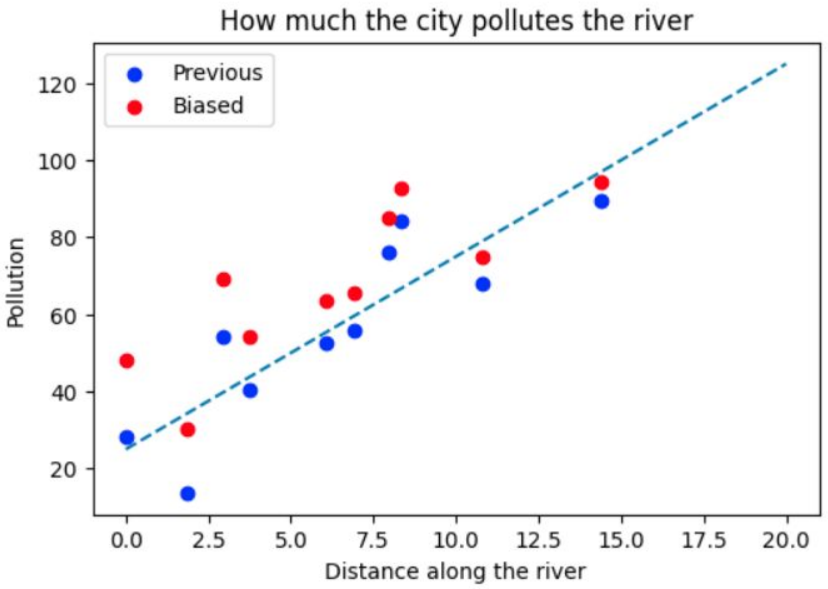
Wrong:
Biased data with unbiased model





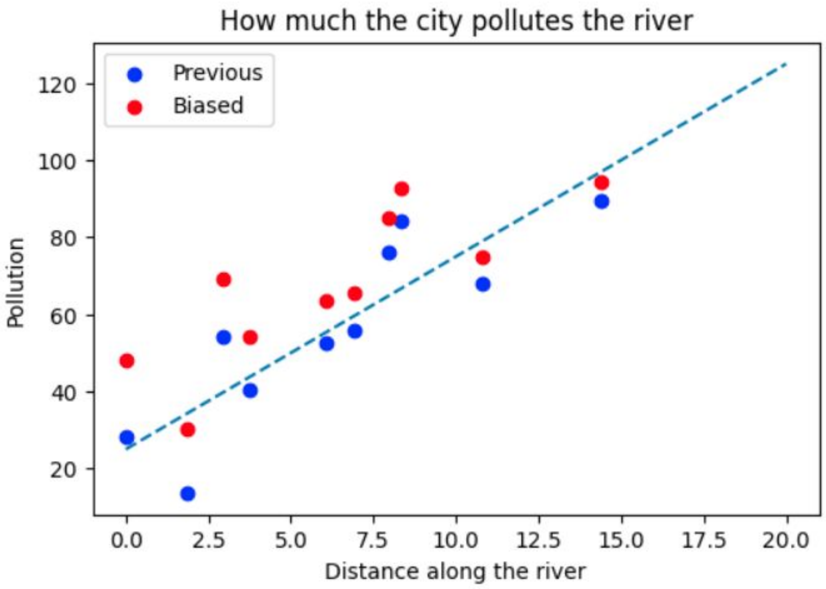
Linear Regression

```
contamination_model_with_bias="""
data {
 int <lower=0> N;
 vector[N] x;
 vector[N] y;

}

parameters {
  real m;
  real b;
  real<lower=0> sigma;
  real<lower=0> bias;
}
model {
   bias ~ cauchy(0,10);  // we are using a half cauchy
                         // because bias >= 0
   y ~ normal(m * x + b + bias*exp(-x/10), sigma);
}
"""
```



Linear Regression

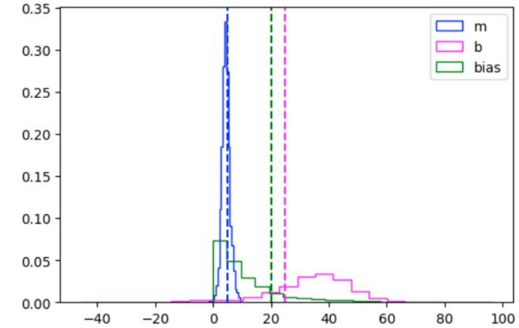
```
contamination_model_with_bias="""
data {
 int <lower=0> N;
 vector[N] x;
 vector[N] y;

}

parameters {
  real m;
  real b;
  real<lower=0> sigma;
  real<lower=0> bias;
}
model {
   bias ~ cauchy(0,10);  // we are using a half cauchy
                         // because bias >= 0
   y ~ normal(m * x + b + bias*exp(-x/10), sigma);
}
"""
```

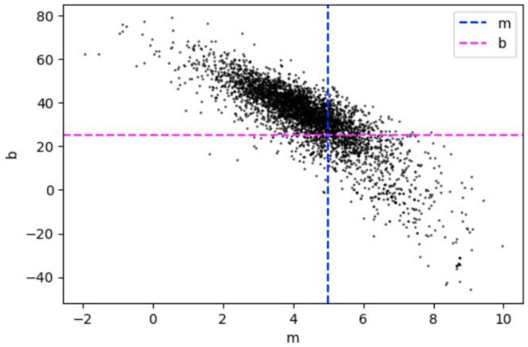


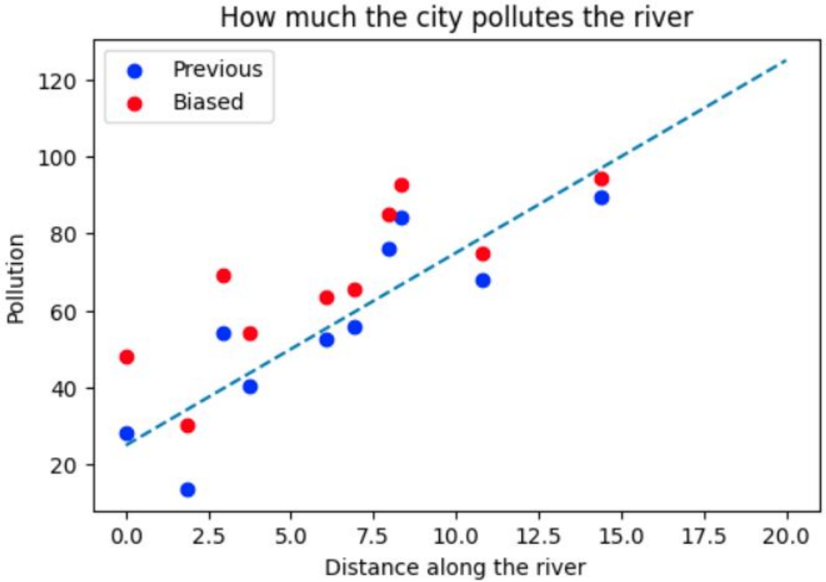
Very easy to adapt the model!

Linear Regression

Correct model!

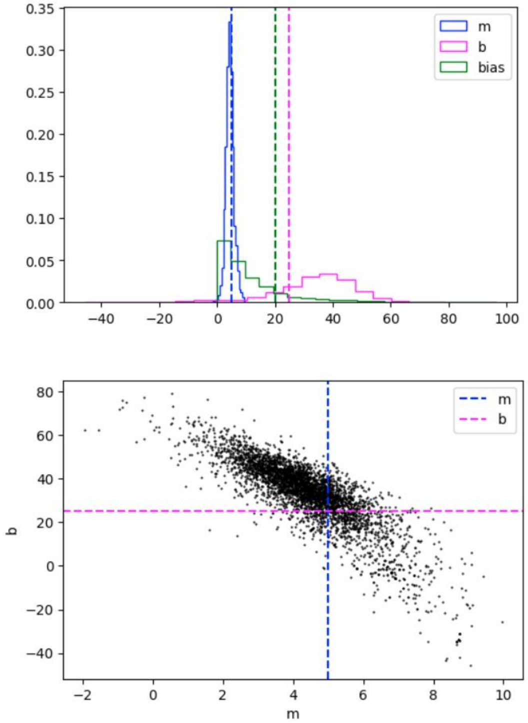


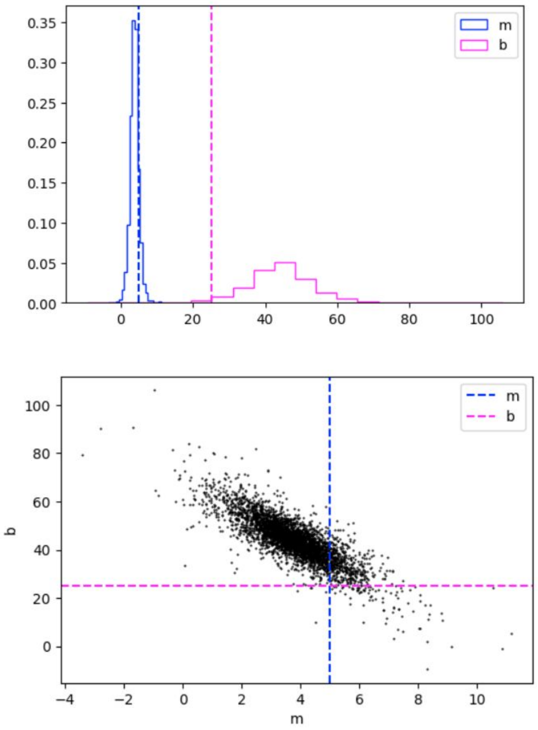




Linear Regression

Correct model!



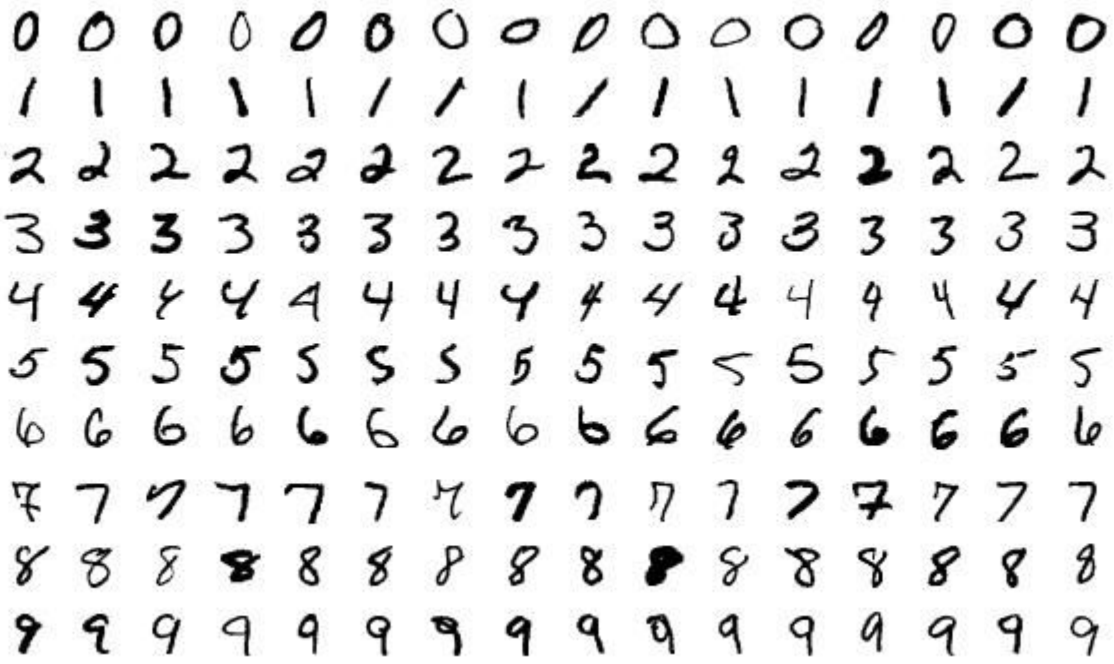


Wrong model!

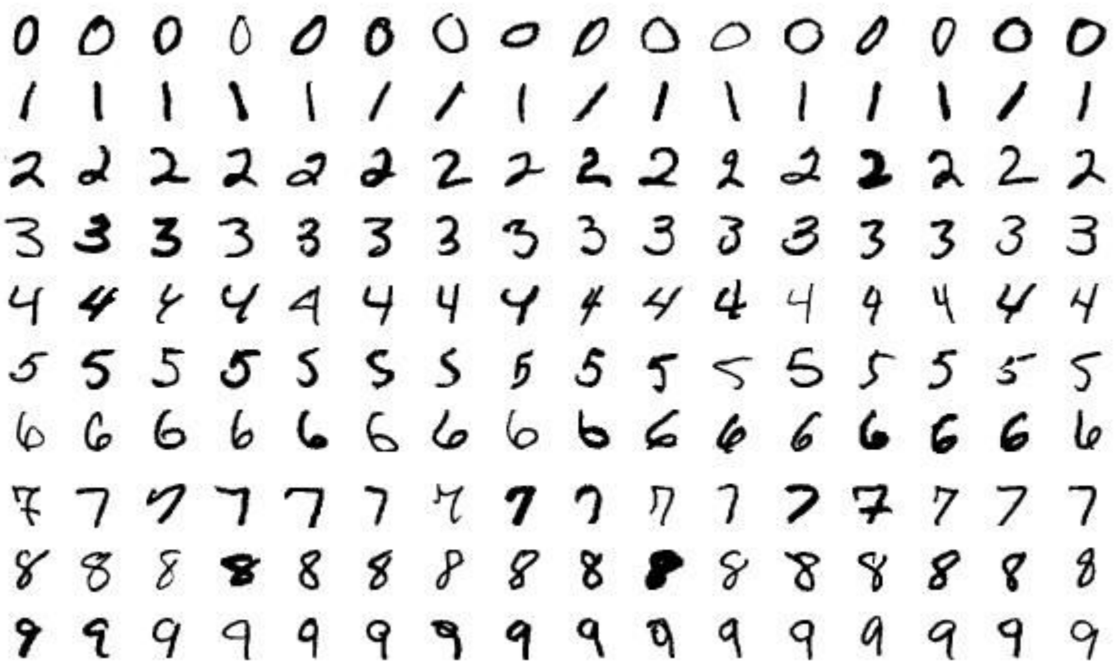
Finding anomalies in *collective motion*

Mixture of Bernoulli

Mixture of Bernoulli



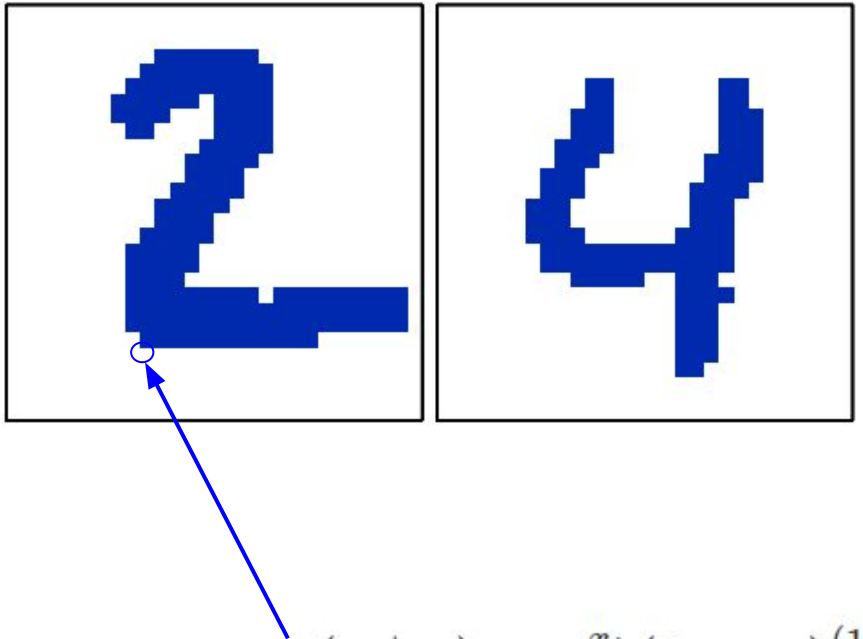
Mixture of Bernoulli



Model:

In each class (number), each pixel sampled from a to-be-inferred Bernoulli.

Mixture of Bernoulli



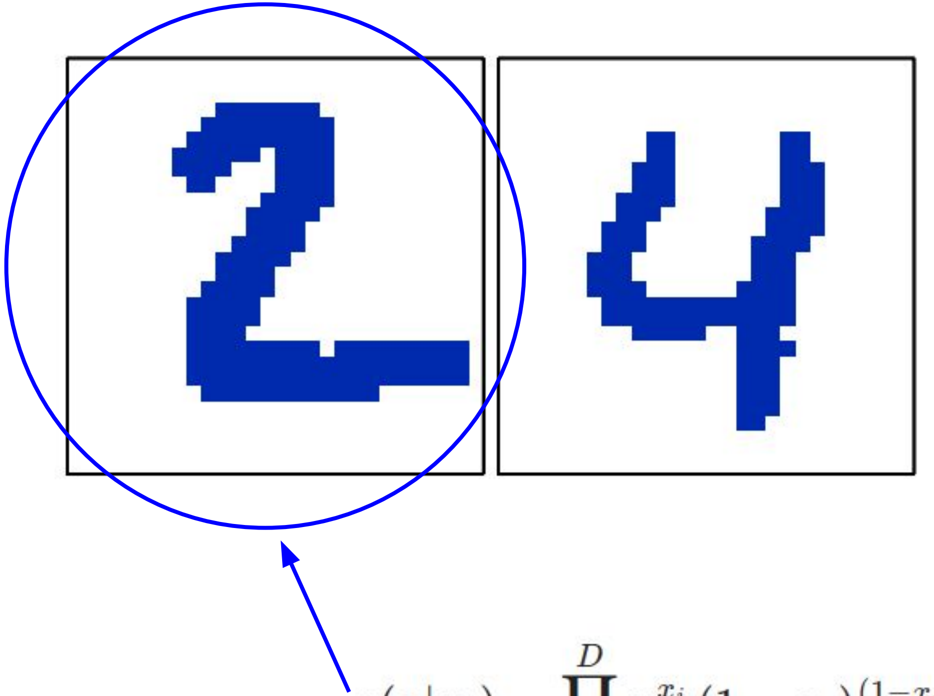
$$p(x_i|\mu_i) = \mu_i^{x_i}\,(1-\mu_i)^{(1-x_i)}$$

Model:

In each class (number), each pixel sampled from a to-be-inferred Bernoulli.

X_i = 0 or 1

Mixture of Bernoulli



$$p(x|\mu_i) = \prod_{i=1}^{D} \mu_i^{x_i} (1-\mu_i)^{(1-x_i)}$$

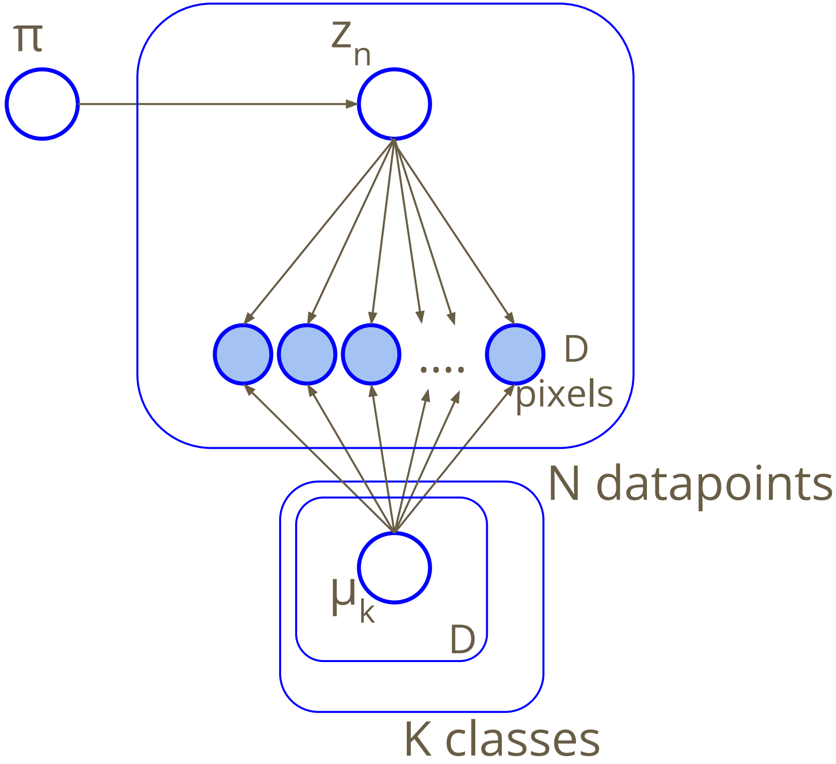
Model:

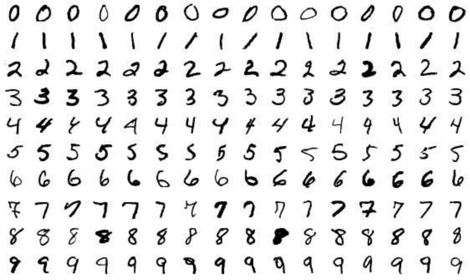
In each class (number), each pixel sampled from a to-be-inferred Bernoulli.

Model:

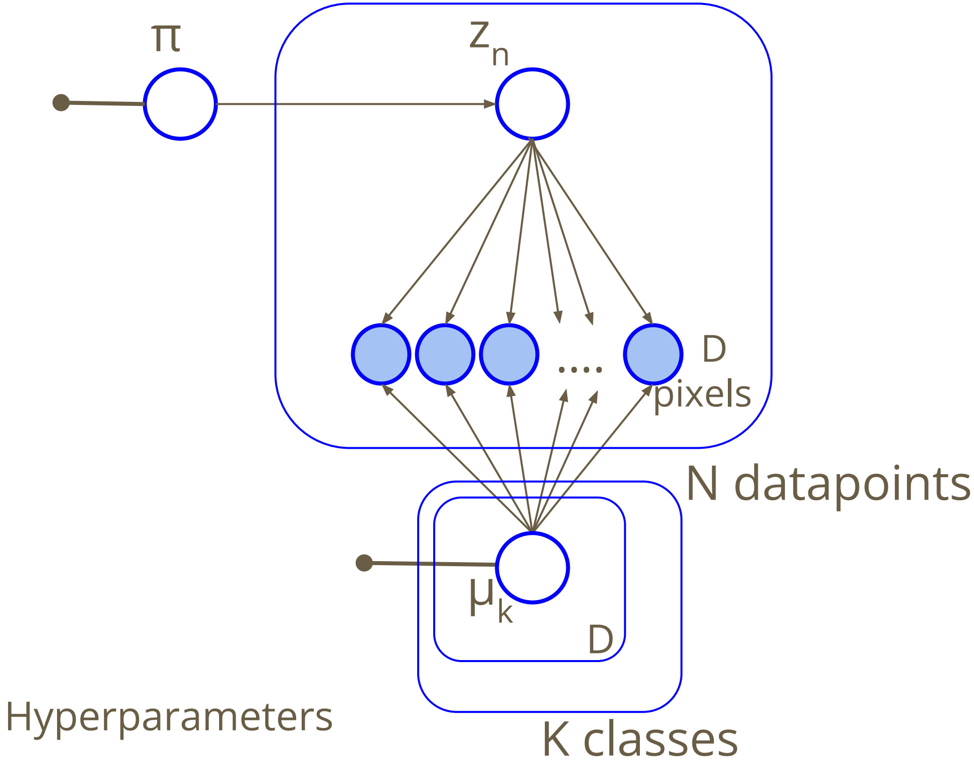
Conditionally independent

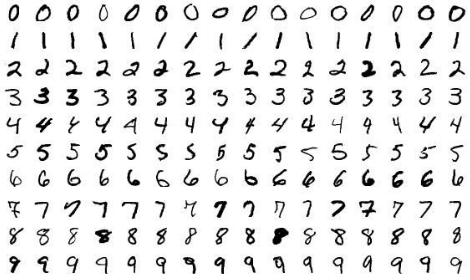
Mixture of Bernoulli



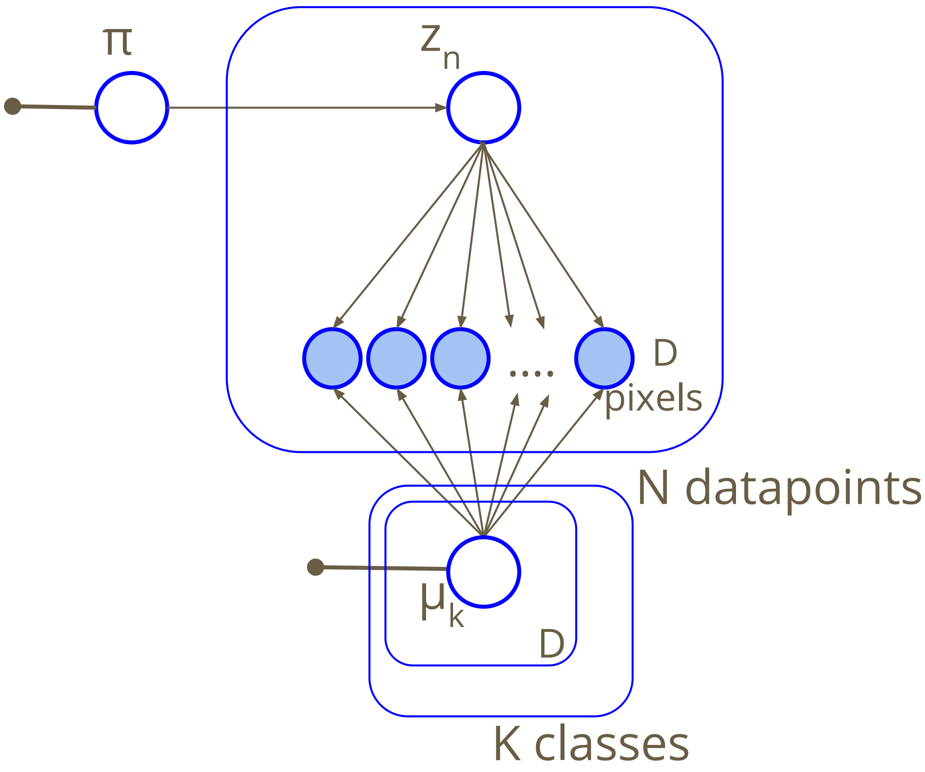


Mixture of Bernoulli



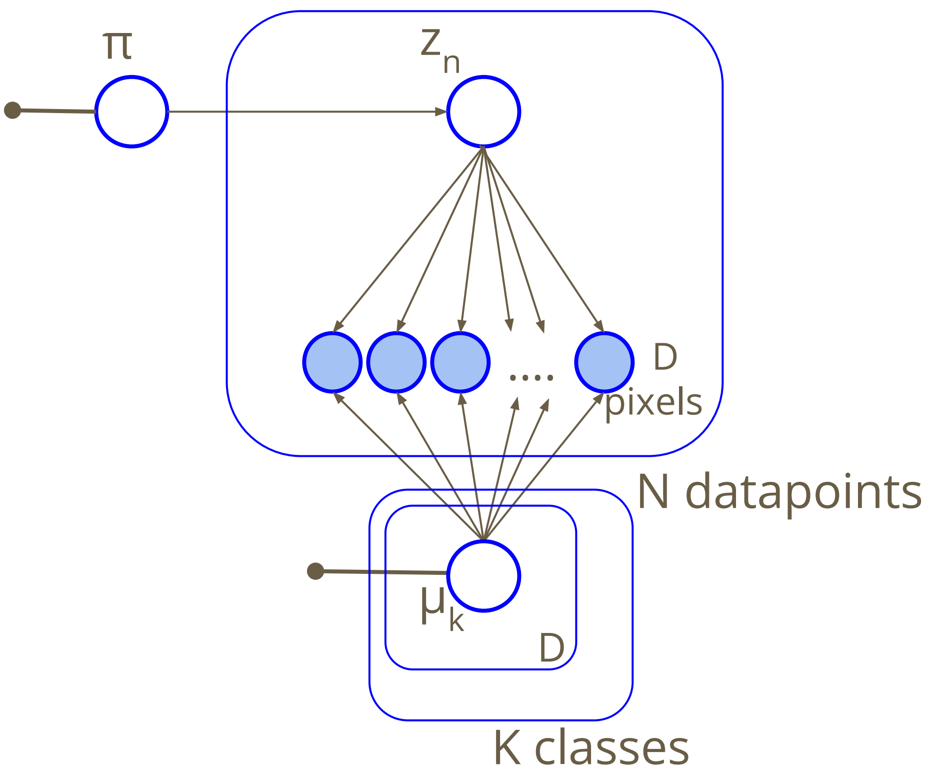


Mixture of Bernoulli



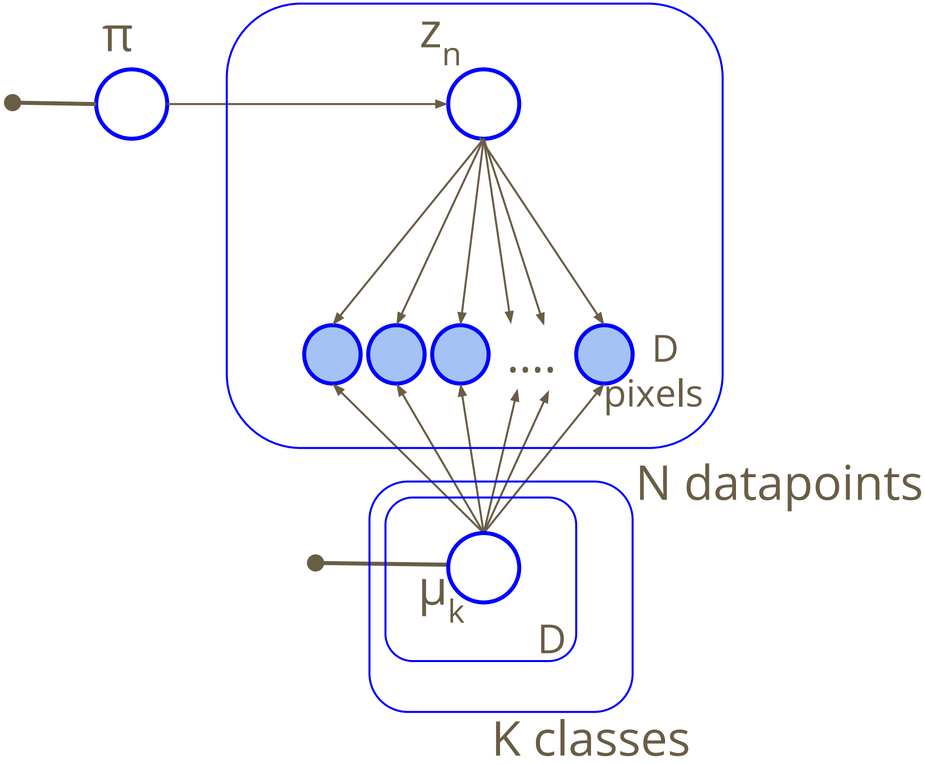
- Conditionally Independent (model is wrong, but works!)

Mixture of Bernoulli



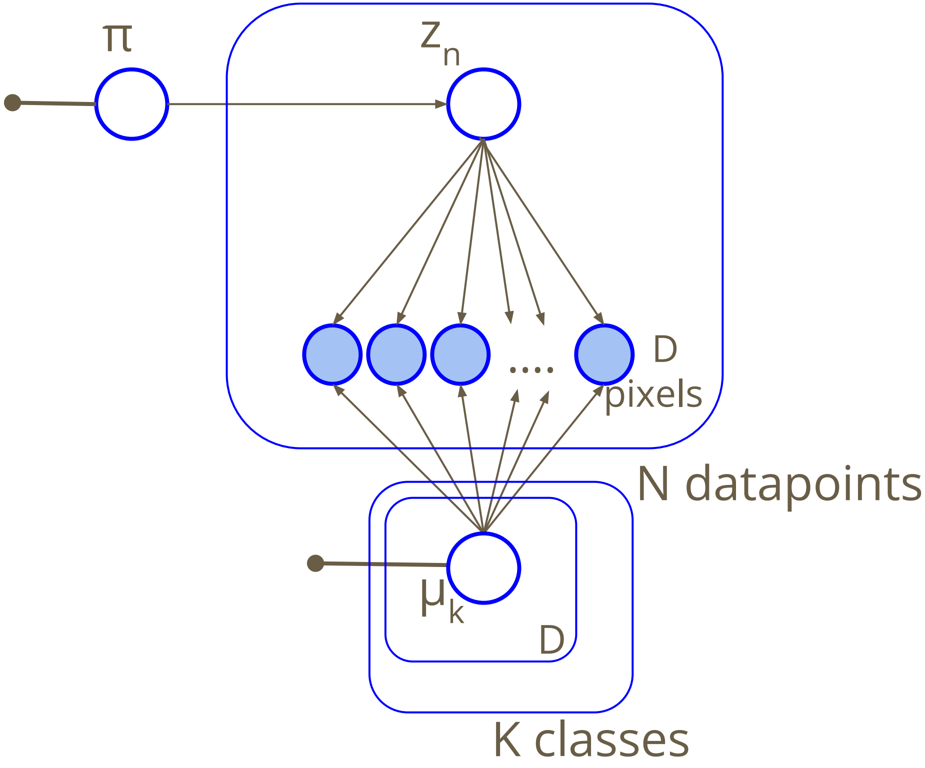
- Conditionally Independent (model is wrong, but works!)
- 2^D for ~*D*K* unknowns

Mixture of Bernoulli



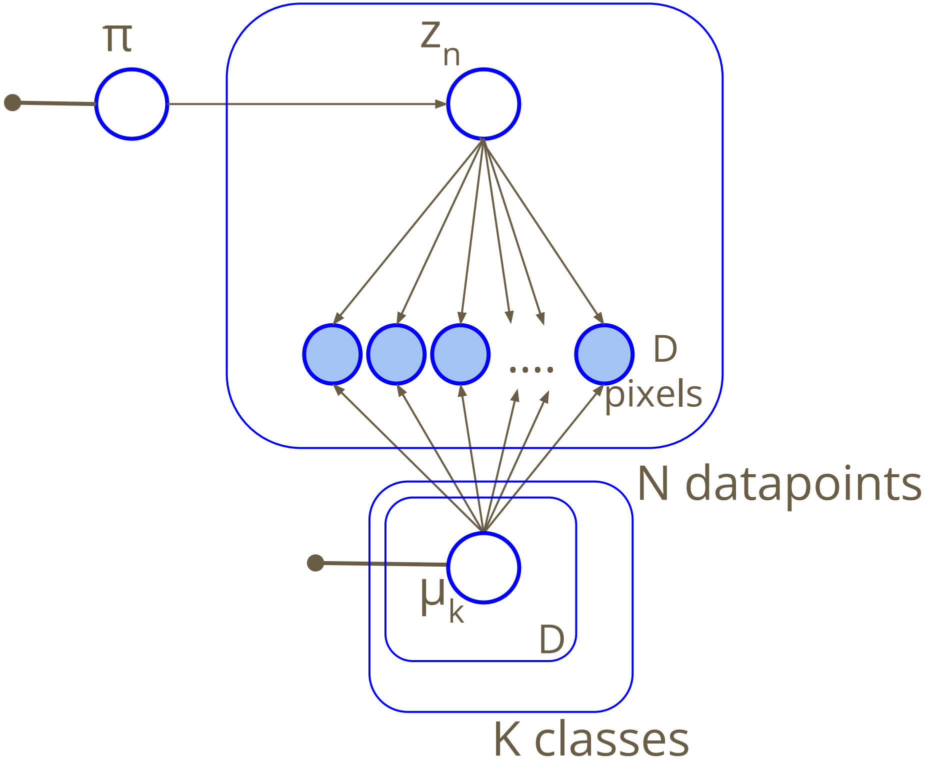
- Conditionally Independent (model is wrong, but works!)
- *2^D* for *~D*K* unknowns
- Dirichlet for π, Uniform for μ

Mixture of Bernoulli



- Conditionally Independent (model is wrong, but works!)
- 2^D for ~$D*K$ unknowns
- Dirichlet for π, Uniform for μ
- Learns with N digits

Mixture of Bernoulli

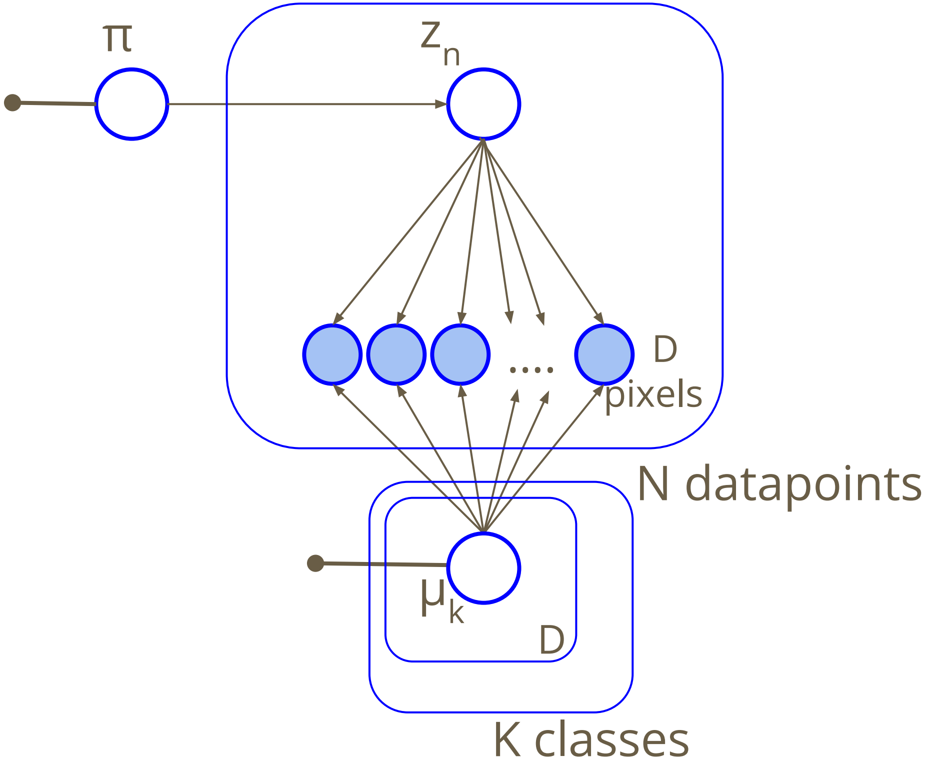


- Conditionally Independent (model is wrong, but works!)
- 2^D for ~$D*K$ unknowns
- Dirichlet for π, Uniform for μ
- Learns with N digits

Unsupervisedly learns digits probability

→ z_n

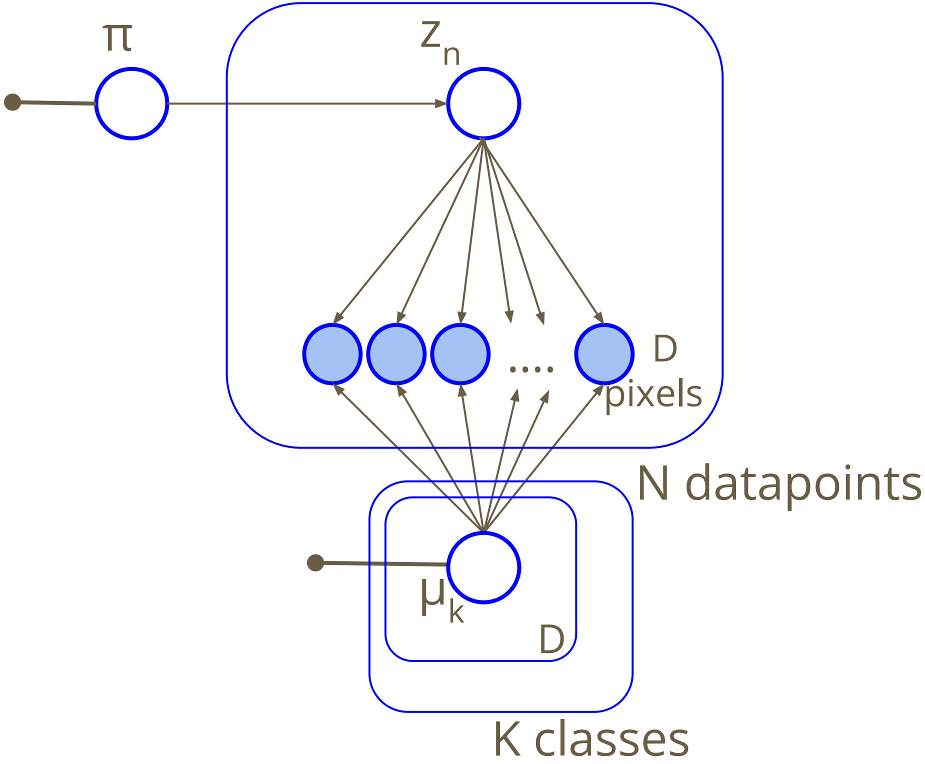
Mixture of Bernoulli



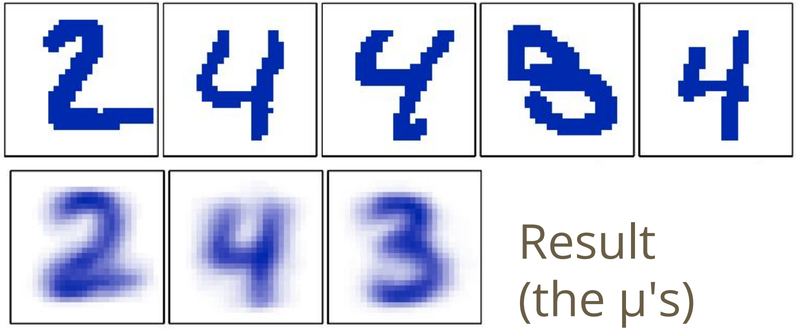
- Conditionally Independent (model is wrong, but works!)
- 2^D for $\sim D*K$ unknowns
- Dirichlet for π, Uniform for μ
- Learns with N digits

Scientifically: z_n is the probability of each class

Mixture of Bernoulli



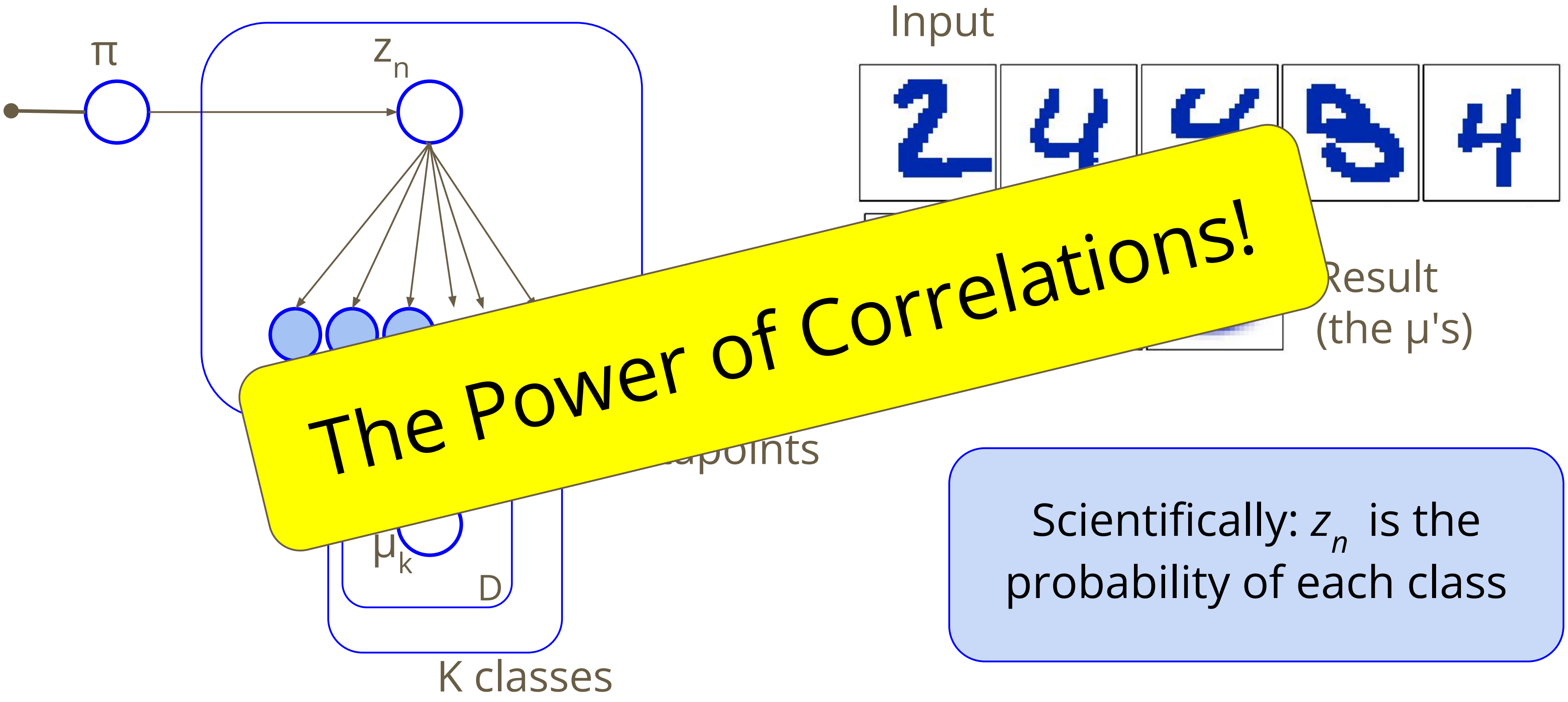
Input



Result
(the μ's)

Scientifically: z_n is the probability of each class

Mixture of Bernoulli



Scientifically: z_n is the probability of each class

Final remarks @ Lecture 2

Final remarks @ Lecture 2

We met the tool!

Final remarks @ Lecture 2

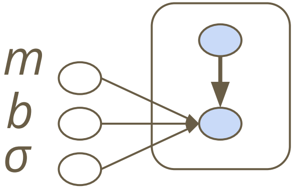
We met the tool!

$$p(\theta|X) = \frac{1}{B(n'+n, k'+k)} p^{k'+k}(1-p)^{n'+n-k'-k}$$

Final remarks @ Lecture 2

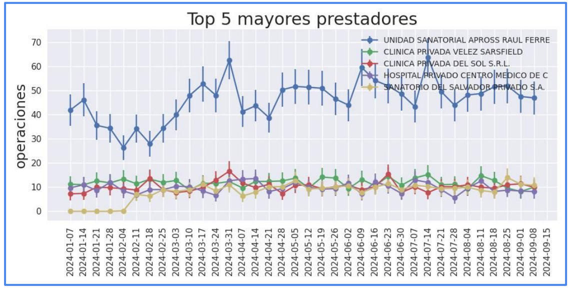
We met the tool!

$$p(\theta|X) = \frac{1}{B(n'+n, k'+k)} p^{k'+k}(1-p)^{n'+n-k'-k}$$

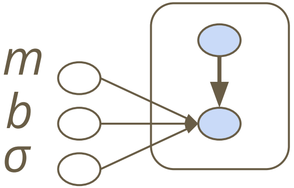


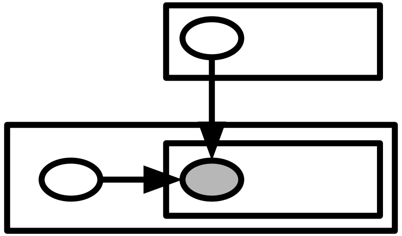
Final remarks @ Lecture 2

We met the tool!



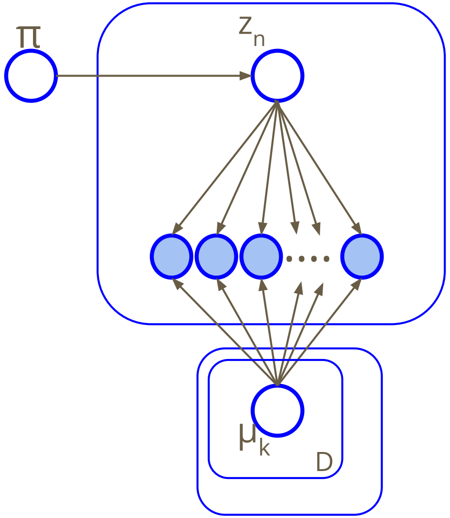
$$p(\theta|X) = \frac{1}{B(n'+n, k'+k)} p^{k'+k}(1-p)^{n'+n-k'-k}$$

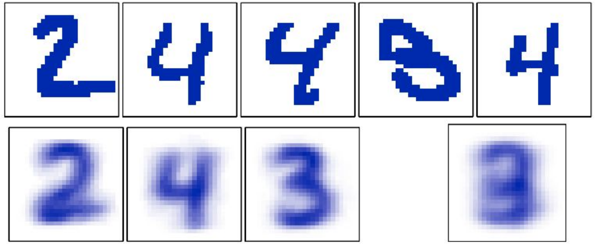




Final remarks @ Lecture 2

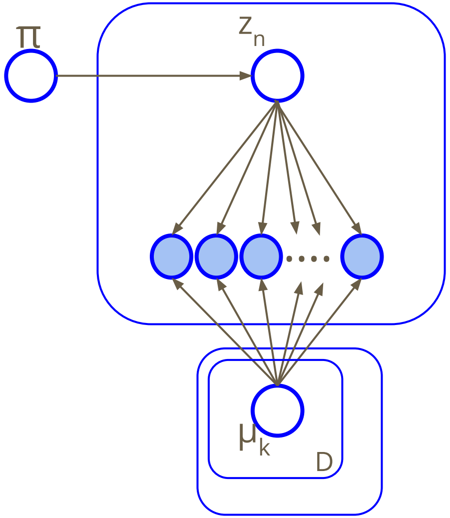
We met the tool!



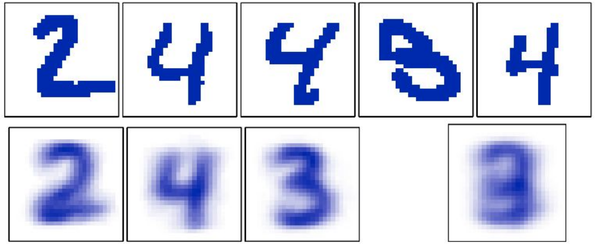


Final remarks @ Lecture 2

We met the tool!

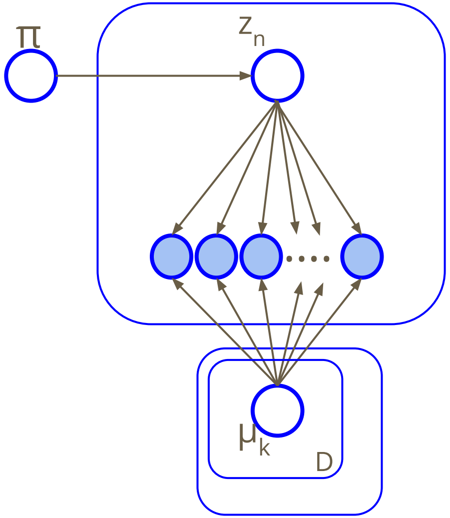


Power of Correlation
Power of Computing
Power of Modeling
Power of Bayes



Final remarks @ Lecture 2

We met the tool!



Power of Correlation
Power of Computing
Power of Modeling
Power of Bayes

Bayesian
Machine Learning!