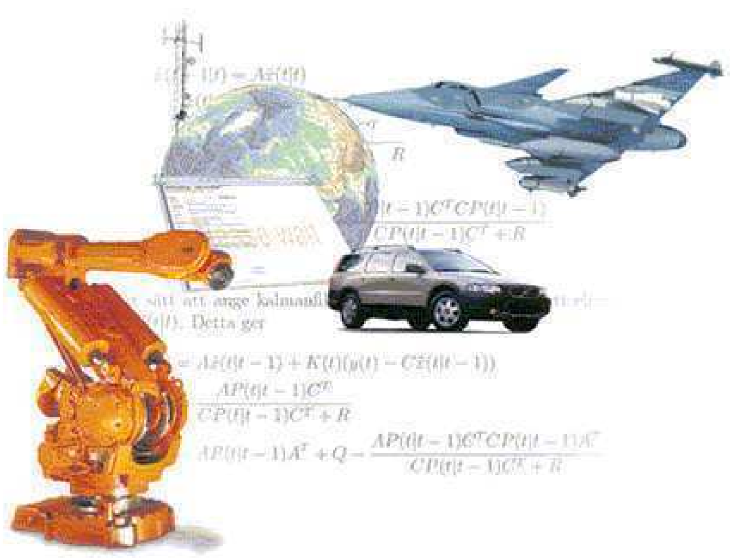
System Identification:
The Path from Data to Model

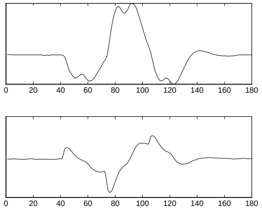


Lennart Ljung

Division of Automatic Control
Linköping University
Sweden

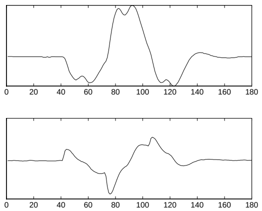
The Problem

Aircraft Dynamics:

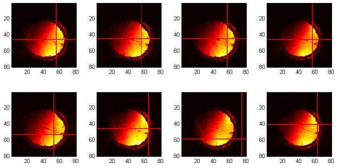


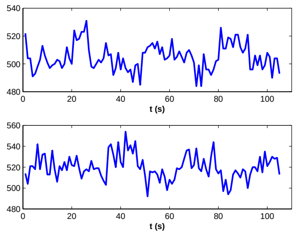
The Problem

Aircraft Dynamics:



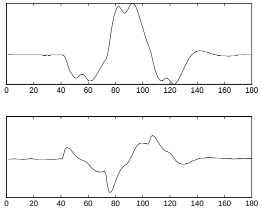
Brain Activity (fMRI):



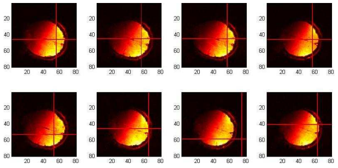


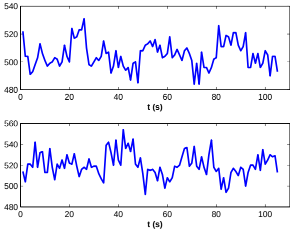
The Problem

Aircraft Dynamics:

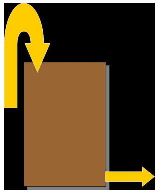


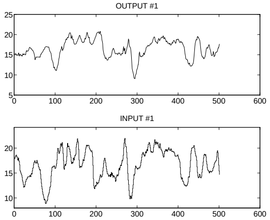
Brain Activity (fMRI):



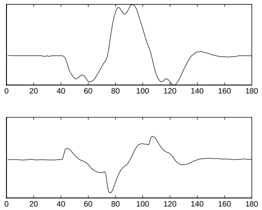


Pulp Buffer Vessel:

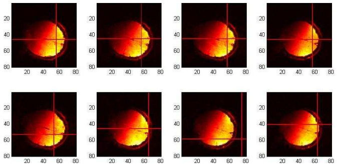


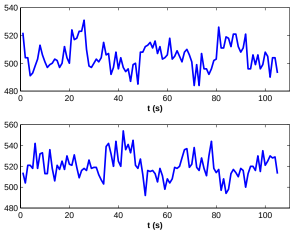


Aircraft Dynamics:

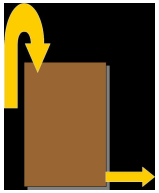


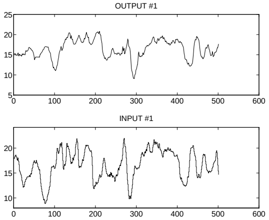
Brain Activity (fMRI):





Pulp Buffer Vessel:

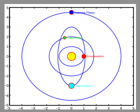




Industrial Engineering:

90 papers

Constructing (mathematical) models from data is a prime problem in many scientific fields and many application areas.



The Communities

Constructing (mathematical) models from data is a prime problem in many scientific fields and many application areas.
Many communities and cultures around the area have grown, with their own nomenclatures and their own “social lives”.

Constructing (mathematical) models from data is a prime problem in many scientific fields and many application areas.

Many communities and cultures around the area have grown, with their own nomenclatures and their own “social lives”.

This has created a very rich, and somewhat confusing, plethora of methods and approaches for the problem.

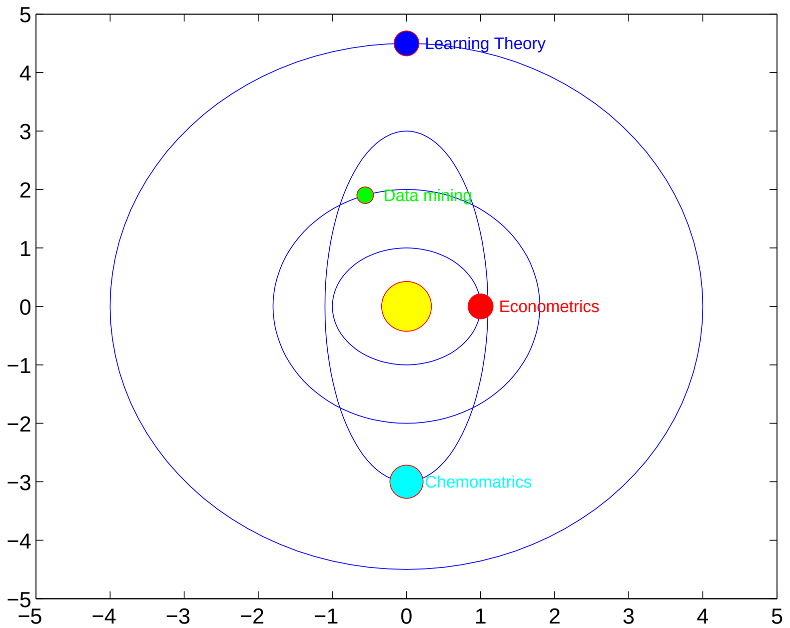
The Communities

Constructing (mathematical) models from data is a prime problem in many scientific fields and many application areas.

Many communities and cultures around the area have grown, with their own nomenclatures and their own "social lives".

This has created a very rich, and somewhat confusing, plethora of methods and approaches for the problem.

A picture: There is a core of central material, encircled by the different communities.



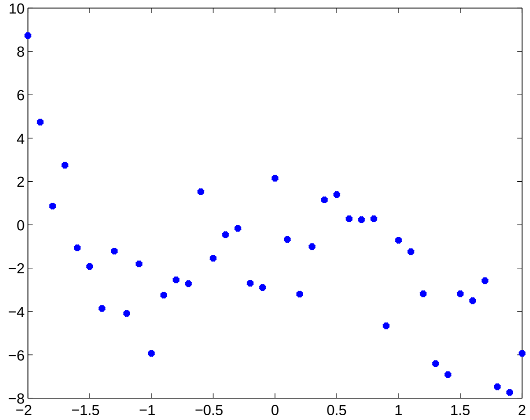
The Core

Central terms

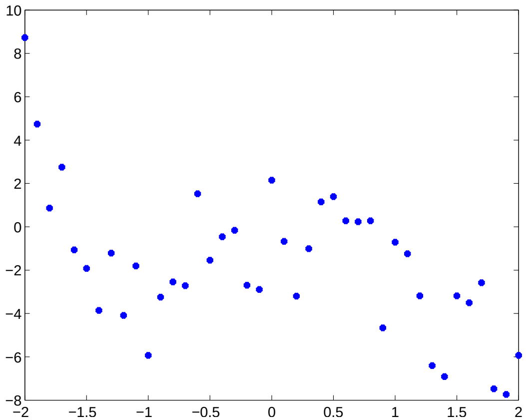
- Model $\mathfrak{m}$ – Model Class $\mathcal{M}$ – Complexity (Flexibility) $\mathcal{C}$
- Information $\mathcal{I}$ – Data Z
- Estimation – Validation (Learning – Generalization)
- Model fit $\mathcal{F}(\mathfrak{m}, Z)$

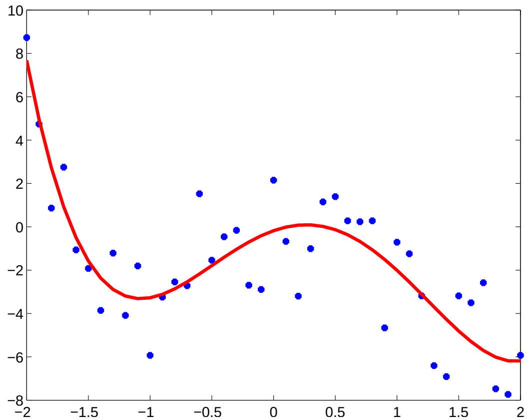
Estimation

information in data



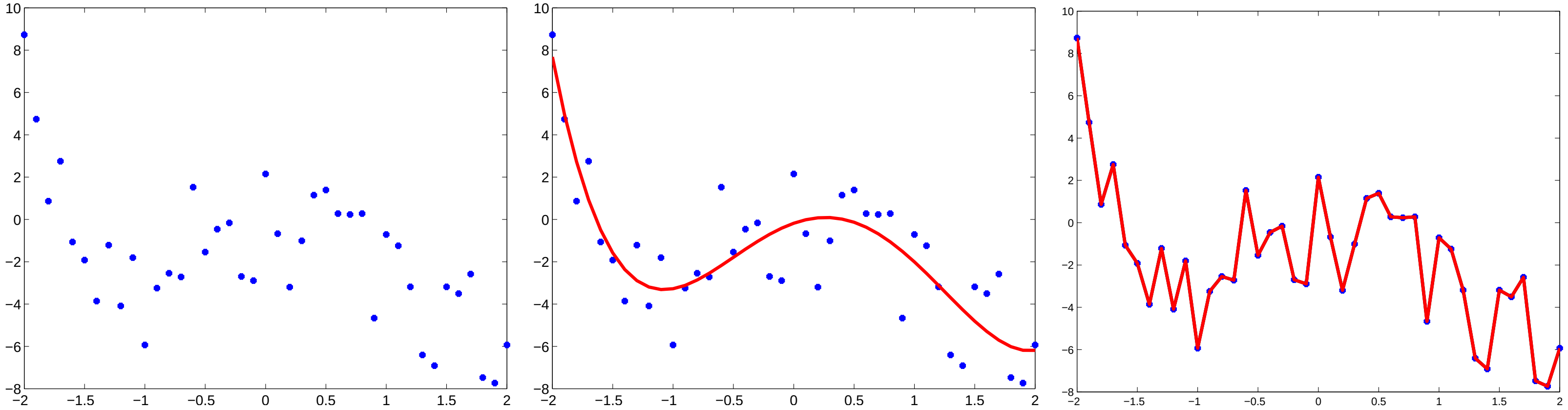
Squeeze out the relevant information in data





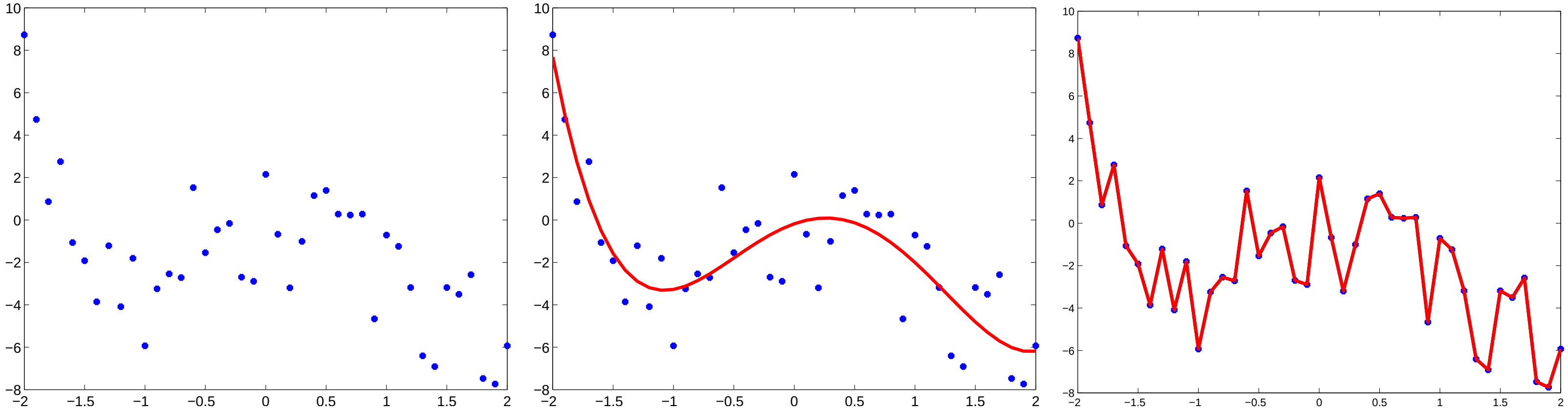
Estimation

Squeeze out the relevant information in data. (BUT NOT MORE!)



Estimation

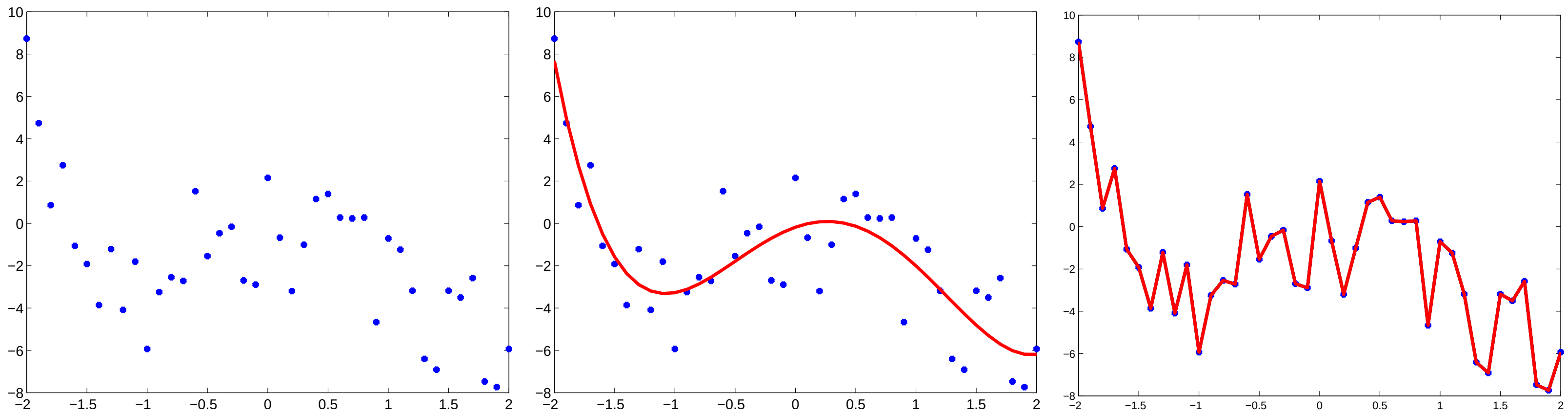
Squeeze out the relevant information in data. (BUT NOT MORE!)



All data contain Information and Misinformation (“Signal and noise”).

Estimation

Squeeze out the relevant information in data. (BUT NOT MORE!)



All data contain Information and Misinformation (“Signal and noise”).

So need to meet the data with a prejudice!

Estimation Prejudices

Nature is Simple!

Estimation Prejudices

Nature is Simple! (Occam's razor, Lex Parsimoniae...)

Estimation Prejudices

Nature is Simple! (Occam’s razor, Lex Parsimoniae...)

God is subtle, but He is not malicious (Einstein)

Estimation Prejudices

Nature is Simple! (Occam's razor, Lex Parsimoniae...)

God is subtle, but He is not malicious (Einstein)

So, conceptually:

$$\hat{\mathfrak{m}} = \arg\min_{\mathfrak{m}\in\mathcal{M}} \text{(Fit + Complexity Penalty)}$$

Estimation Prejudices

Nature is Simple! (Occam's razor, Lex Parsimoniae...)

God is subtle, but He is not malicious (Einstein)

So, conceptually:

$$\hat{\mathfrak{m}} = \arg\min_{\mathfrak{m} \in \mathcal{M}} \text{(Fit + Complexity Penalty)}$$

Examples:

- Search for a model in sets with a maximal Complexity
- The Akaike criterion
- Regularization

Fit to estimation data Z_e^N (N: Number of data points)

$$F(\hat{\mathfrak{m}}, Z_e^N) \qquad \text{("The empirical risk")}$$

Estimation and Validation

Fit to estimation data Z_e^N (N: Number of data points)

$$F(\hat{\mathfrak{m}}, Z_e^N) \qquad \text{("The empirical risk")}$$

Now try your model on a fresh data set (Validation data Z_v):

$$E\mathcal{F}(\hat{\mathfrak{m}}, Z_v) \approx \mathcal{F}(\hat{\mathfrak{m}}, Z_e^N) + f(\mathcal{C}(\mathcal{M}), N)$$

Estimation and Validation

Fit to estimation data Z_e^N (N: Number of data points)

$$F(\hat{\mathfrak{m}}, Z_e^N) \qquad \text{("The empirical risk")}$$

Now try your model on a fresh data set (Validation data Z_v):

$$E\mathcal{F}(\hat{\mathfrak{m}}, Z_v) \approx \mathcal{F}(\hat{\mathfrak{m}}, Z_e^N) + f(\mathcal{C}(\mathcal{M}), N)$$

f increases with the complexity $\mathcal{C}$ and decreases with N, so the more flexible the model set the worse fit to validation data.

Estimation and Validation

Fit to estimation data Z_e^N (N: Number of data points)

$$F(\hat{\mathfrak{m}}, Z_e^N) \qquad \text{("The empirical risk")}$$

Now try your model on a fresh data set (Validation data Z_v):

$$E\mathcal{F}(\hat{\mathfrak{m}}, Z_v) \approx \mathcal{F}(\hat{\mathfrak{m}}, Z_e^N) + f(\mathcal{C}(\mathcal{M}), N)$$

f increases with the complexity $\mathcal{C}$ and decreases with N, so the more flexible the model set the worse fit to validation data.

In words: If you have a model that describes the estimation data well, the fit will be (much) worse when you try it on validation data.

Fit to estimation data Z_e^N (N: Number of data points)

$$F(\hat{\mathfrak{m}}, Z_e^N) \qquad \text{("The empirical risk")}$$

Now try your model on a fresh data set (Validation data Z_v):

$$E\mathcal{F}(\hat{\mathfrak{m}}, Z_v) \approx \mathcal{F}(\hat{\mathfrak{m}}, Z_e^N) + f(\mathcal{C}(\mathcal{M}), N)$$

f increases with the complexity $\mathcal{C}$ and decreases with N, so the more flexible the model set the worse fit to validation data.

In words: If you have a model that describes the estimation data well, the fit will be (much) worse when you try it on validation data.

So don't be impressed by a good fit to estimation data in a flexible model set!

Bias and Variance

$\mathcal{S}$ – True system $\mathcal{M}$ – Model set $\hat{\mathfrak{m}}$ – Estimate

$\mathfrak{m}^*$ – Expected model $\mathfrak{m}^* = E\hat{\mathfrak{m}}$ Typically $\mathfrak{m}^* = \arg\min_{\mathfrak{m}\in\mathcal{M}} \|\mathcal{S} - \mathfrak{m}\|^2$

Then

Bias and Variance

$\mathcal{S}$ – True system $\mathcal{M}$ – Model set $\hat{\mathfrak{m}}$ – Estimate

$\mathfrak{m}^*$ – Expected model $\mathfrak{m}^* = E\hat{\mathfrak{m}}$ Typically $\mathfrak{m}^* = \arg\min_{\mathfrak{m}\in\mathcal{M}} \|\mathcal{S} - \mathfrak{m}\|^2$

Then

$$E\|\mathcal{S} - \hat{\mathfrak{m}}\|^2 = \|\mathcal{S} - \mathfrak{m}^*\|^2 + E\|\hat{\mathfrak{m}} - \mathfrak{m}^*\|^2$$

MSE $=$ B: BIAS $+$ V: Variance

$\mathcal{S}$ – True system $\mathcal{M}$ – Model set $\hat{\mathfrak{m}}$ – Estimate

$\mathfrak{m}^*$ – Expected model $\mathfrak{m}^* = E\hat{\mathfrak{m}}$ Typically $\mathfrak{m}^* = \arg\min_{\mathfrak{m}\in\mathcal{M}} \|\mathcal{S} - \mathfrak{m}\|^2$

Then

$$E\|\mathcal{S} - \hat{\mathfrak{m}}\|^2 = \|\mathcal{S} - \mathfrak{m}^*\|^2 + E\|\hat{\mathfrak{m}} - \mathfrak{m}^*\|^2$$

MSE $=$ B: BIAS $+$ V: Variance

Error: $=$ Systematic$+$Random

Bias and Variance

$\mathcal{S}$ – True system $\mathcal{M}$ – Model set $\hat{\mathfrak{m}}$ – Estimate

$\mathfrak{m}^*$ – Expected model $\mathfrak{m}^* = E\hat{\mathfrak{m}}$ Typically $\mathfrak{m}^* = \arg\min_{\mathfrak{m}\in\mathcal{M}} \|\mathcal{S}-\mathfrak{m}\|^2$

Then

$$E\|\mathcal{S}-\hat{\mathfrak{m}}\|^2 = \|\mathcal{S}-\mathfrak{m}^*\|^2 + E\|\hat{\mathfrak{m}}-\mathfrak{m}^*\|^2$$

MSE $=$ B: BIAS $+$ V: Variance

Error: $=$ Systematic$+$Random

As model complexity increases, Bias decreases and Variance increases

$\mathcal{S}$ – True system $\mathcal{M}$ – Model set $\hat{\mathfrak{m}}$ – Estimate

$\mathfrak{m}^*$ – Expected model $\mathfrak{m}^* = E\hat{\mathfrak{m}}$ Typically $\mathfrak{m}^* = \arg\min_{\mathfrak{m}\in\mathcal{M}} \|\mathcal{S} - \mathfrak{m}\|^2$

Then

$$E\|\mathcal{S} - \hat{\mathfrak{m}}\|^2 = \|\mathcal{S} - \mathfrak{m}^*\|^2 + E\|\hat{\mathfrak{m}} - \mathfrak{m}^*\|^2$$

MSE $=$ B: BIAS $+$ V: Variance

Error: $=$ Systematic$+$Random

As model complexity increases, Bias decreases and Variance increases

This bias/variance trade-off is at the heart of estimation.

Note that the model complexity that minimizes the MSE typically has a non-zero systematic error.

Bottom line: There is a theoretic best accuracy that can be achieved in estimation, independent of methods and computational effort. This bound depends on prior knowledge and data quality.

Bottom line: There is a theoretic best accuracy that can be achieved in estimation, independent of methods and computational effort. This bound depends on prior knowledge and data quality.

Formalization: Observe Y. Let its probability density function (pdf) be $f_Y(x, \theta)$
The (Fisher) Information Matrix is

$$\mathcal{I} = E\ell'_Y(\ell'_Y)^T, \qquad \ell'_Y = \frac{\partial}{\partial\theta} \log f_Y(x, \theta)$$

Bottom line: There is a theoretic best accuracy that can be achieved in estimation, independent of methods and computational effort. This bound depends on prior knowledge and data quality.

Formalization: Observe Y. Let its probability density function (pdf) be $f_Y(x, \theta)$
The (Fisher) Information Matrix is

$$\mathcal{I} = E\ell'_Y(\ell'_Y)^T, \qquad \ell'_Y = \frac{\partial}{\partial\theta}\log f_Y(x, \theta)$$

The Cramér-Rao inequality tells us that

$$\text{cov}\hat{\theta} \geq \mathcal{I}^{-1}$$

for any (unbiased) estimator $\hat{\theta}$ of the parameter.

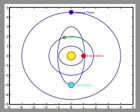
$\mathcal{I}$ is thus a prime quantity for Experiment Design.

The Communities Around the Core I

- Statistics, The Mother Area
  - Recent activities...
  - Bootstrap
  - Regularization to control complexity (LASSO, LARS,...)

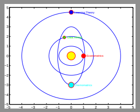
The Communities Around the Core I

- Statistics, The Mother Area
  - Recent activities...
  - Bootstrap
  - Regularization to control complexity (LASSO, LARS,...)
- Econometrics
  - Volatility Clustering (varying variance), GARCH
  - Common roots for variations (co-integration)



The Communities Around the Core I

- **Statistics, The Mother Area**
  - Recent activities...
  - Bootstrap
  - Regularization to control complexity (LASSO, LARS,...)
- **Econometrics**
  - Volatility Clustering (varying variance), GARCH
  - Common roots for variations (co-integration)
- **Statistical Learning Theory**
  - Convex Formulations, SVM
  - VC-dimensions

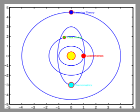


The Communities Around the Core I

- **Statistics, The Mother Area**
  - Recent activities...
  - Bootstrap
  - Regularization to control complexity (LASSO, LARS,...)
- **Econometrics**
  - Volatility Clustering (varying variance), GARCH
  - Common roots for variations (co-integration)
- **Statistical Learning Theory**
  - Convex Formulations, SVM
  - VC-dimensions

The Communities Around the Core II

- Chemometrics – Statistical Process Control
  - High-dimensional Data Spaces (Many process variables)



The Communities Around the Core II

- Chemometrics – Statistical Process Control
  - High-dimensional Data Spaces (Many process variables)
- Data Mining
  - The Internet!

The Communities Around the Core II

- Chemometrics – Statistical Process Control
  - High-dimensional Data Spaces (Many process variables)
- Data Mining
  - The Internet!
- Machine Learning

The Communities Around the Core II

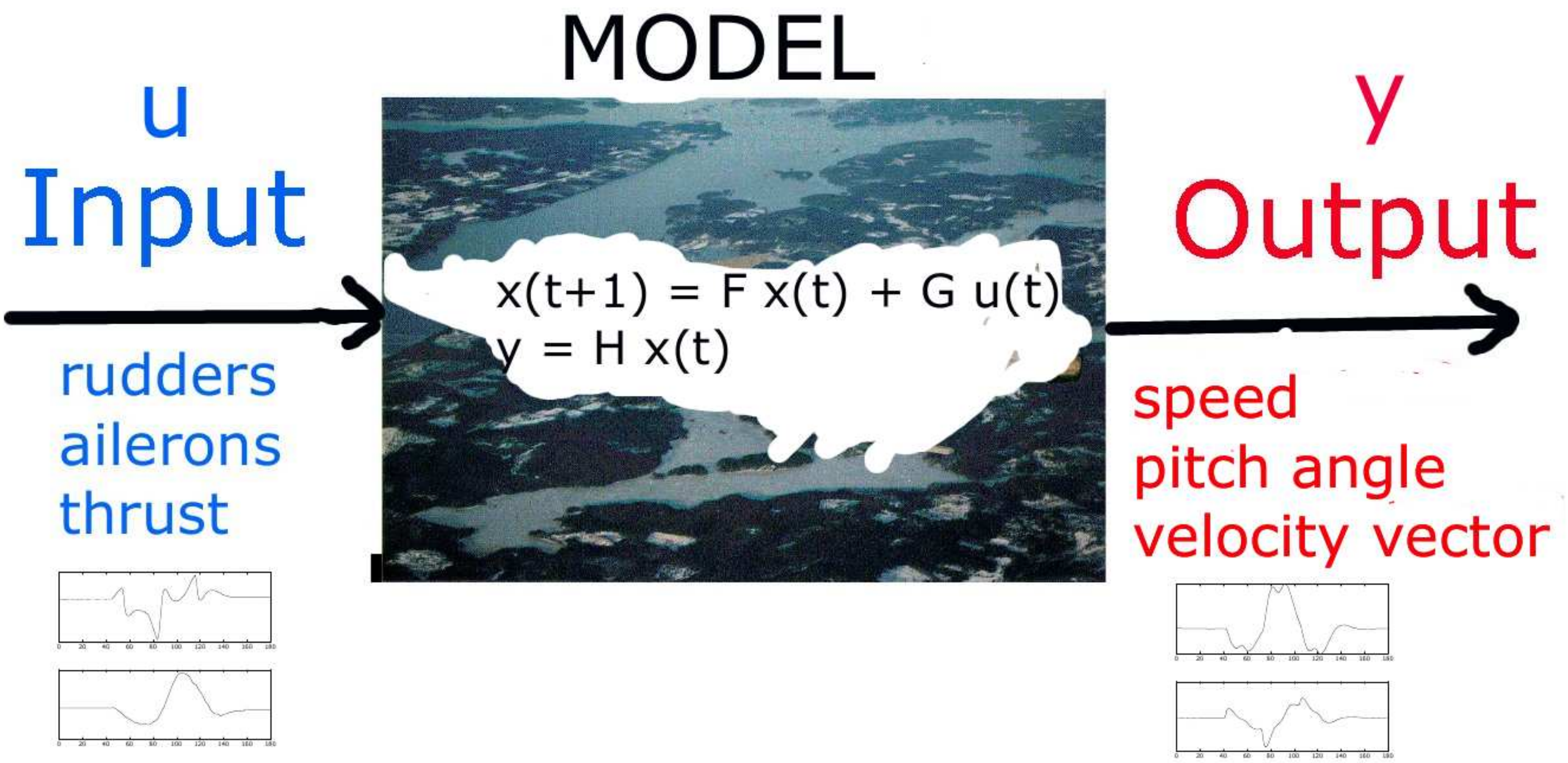
- Chemometrics – Statistical Process Control
  - High-dimensional Data Spaces (Many process variables)
- Data Mining
  - The Internet!
- Machine Learning
  - Grown out of artificial intelligence, more and more statistically oriented

The Communities Around the Core II

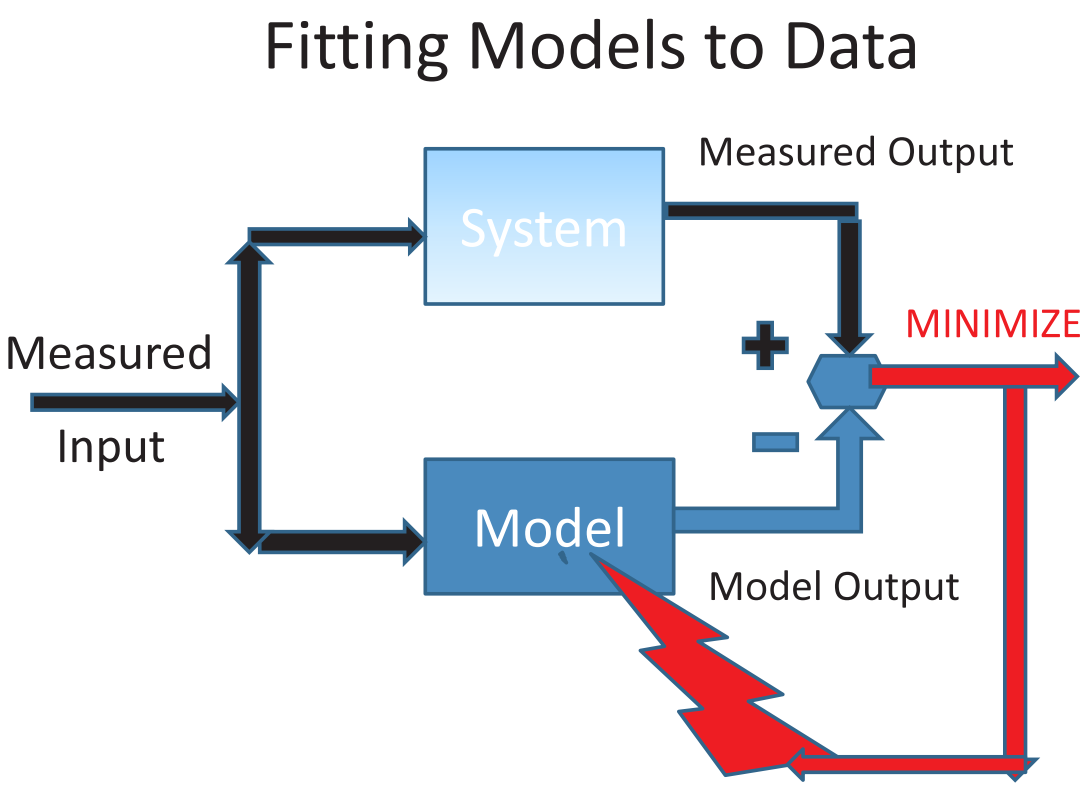
- Chemometrics – Statistical Process Control
  - High-dimensional Data Spaces (Many process variables)
- Data Mining
  - The Internet!
- Machine Learning
  - Grown out of artificial intelligence, more and more statistically oriented
- System Identification
  - Dynamical systems

System Identification: The System

System Identification: The Model



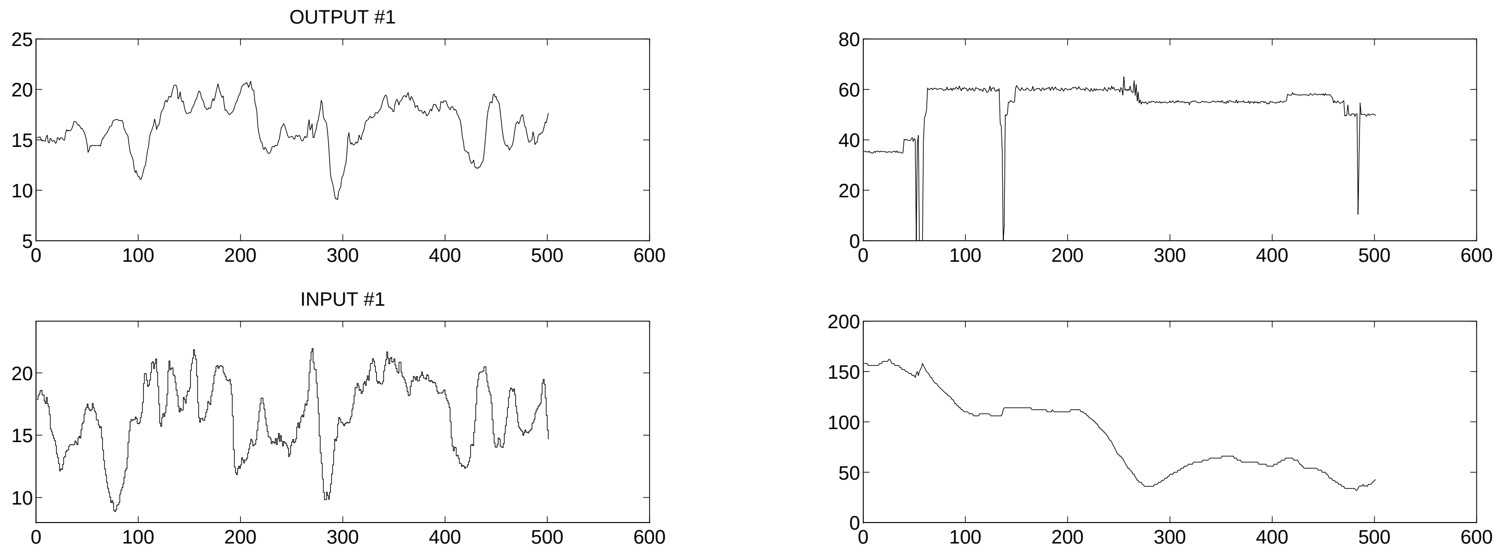
Let the model predict the next output and minimize the error in prediction:



How to come up with the class of models?

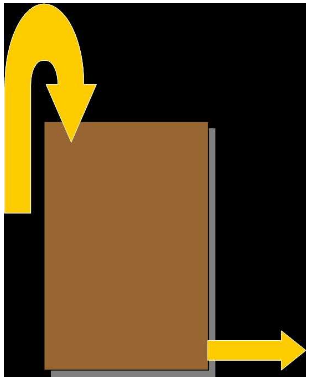
- Use physical insights/common sense together with standard flexible model classes:
- Color Coded: Black-box, Grey-box, White and Off-White Models.
- Let us look at a concrete example from Process Industry

Buffer Vessel Dynamics

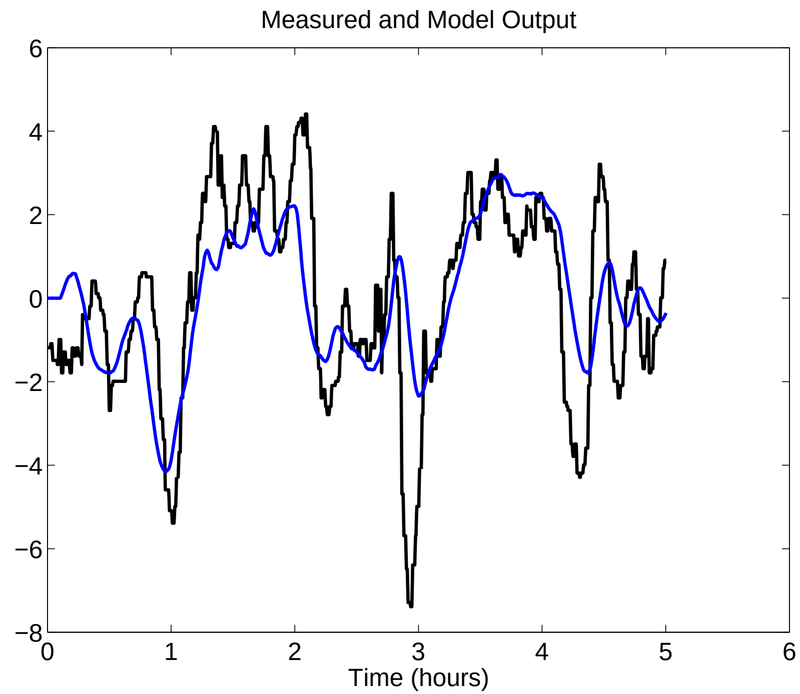


κ-number of outflow,
κ-number of inflow,

flow
volume



Model Based on Raw Data



Black line: κ-number after the vessel, actual measurements.

Blue line: Simulated κ-number using the input only and a process model estimated using the first 200 data points. $G(s) = \frac{0.818}{1+676s} e^{-480s}$

Now it's time to

Think: ....

Now it's time to

Think: ....

If no mixing in tank ("plug flow") a particle that enters the top will exit T seconds later, where

$$T = \frac{\text{Tank Volume}}{\text{Flow}} : \left[\frac{m^3}{m^3/s} = s \right]$$

Now it's time to

Think: ....

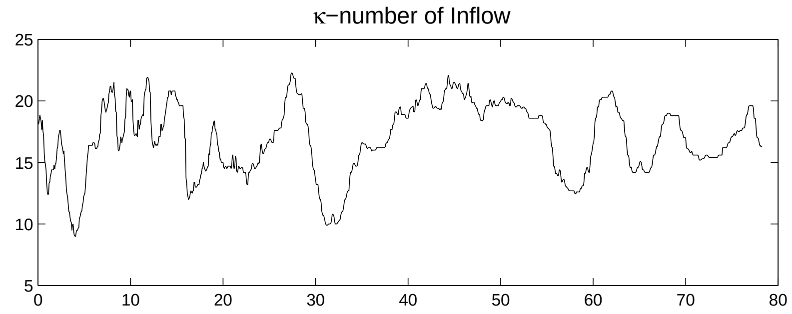
If no mixing in tank ("plug flow") a particle that enters the top will exit T seconds later, where

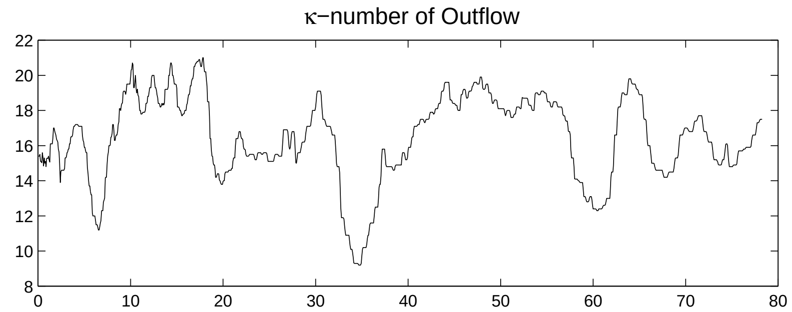
$$T = \frac{\text{Tank Volume}}{\text{Flow}} : \left[\frac{m^3}{m^3/s} = s\right]$$

But this "natural delay time" is time-varying: $T = T(t)$, since flow and volume changes

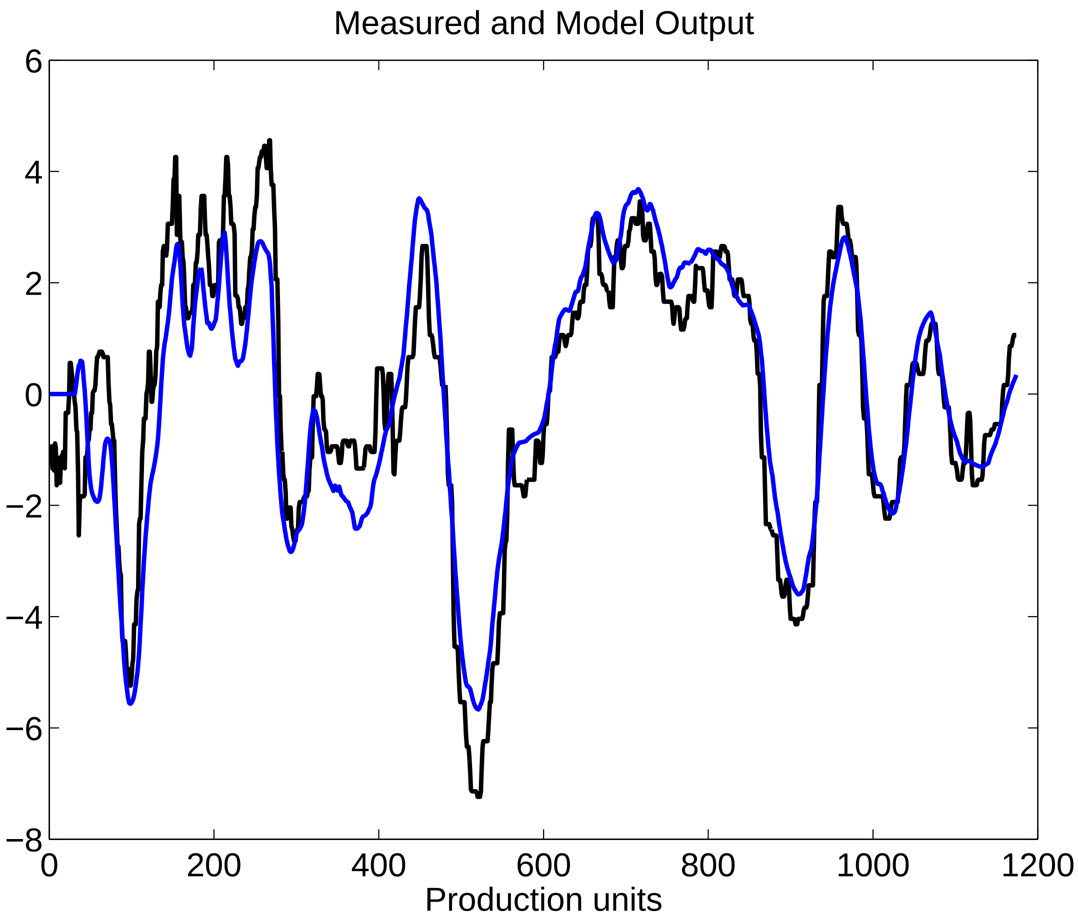
So: Resample Data!

```
z = [y,u]; pf = flow./level;
t = 1:length(z)
newt = interp1([cumsum(pf),t],[pf(1):sum(pf)]');
newz = interp1([t,z], newt);
```

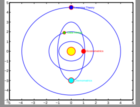




Semi-physical Model



$$G(s) = \frac{0.8116}{1+110.28s} e^{-369.58s}$$



System Identification - General Remarks

- Another satellite encircling the core.
- Deals with mathematical models of dynamic systems.

System Identification - General Remarks

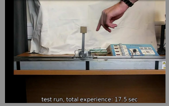
- Another satellite encircling the core.
- Deals with mathematical models of dynamic systems.

- Typical themes:
  - Useful model structures
  - Adapt and adopt the core's fundamentals: Prediction error methods
  - Experiment design (make $\mathcal{I}$ large)
    - with intended model use in mind ("identification for control")

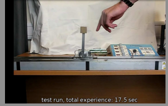
System Identification - General Remarks

- Another satellite encircling the core.
- Deals with mathematical models of dynamic systems.

- Typical themes:
  - Useful model structures
  - Adapt and adopt the core's fundamentals: Prediction error methods
  - Experiment design (make $\mathcal{I}$ large)
    - with intended model use in mind ("identification for control")

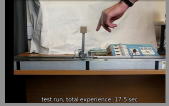


Machine Learning



Machine Learning

Rather From Data to Action

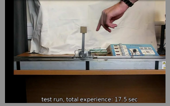


Machine Learning

Rather From Data to Action

Let us first watch a thought-provoking video (from Carl Rasmussen at Cambridge). It takes 1.5 minutes.

...



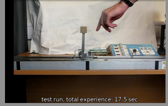
Machine Learning

Rather From Data to Action

Let us first watch a thought-provoking video (from Carl Rasmussen at Cambridge). It takes 1.5 minutes.
...

What is going on?

At time t, let the position of the pendulum be $y(t)$ and the control action (the force on the cart) $u(t)$.
Somehow we need to understand the mapping $u(t) \Rightarrow y(t)$.



Machine Learning

Rather From Data to Action

Let us first watch a thought-provoking video (from Carl Rasmussen at Cambridge). It takes 1.5 minutes.
...

What is going on?

At time t, let the position of the pendulum be $y(t)$ and the control action (the force on the cart) $u(t)$.
Somehow we need to understand the mapping $u(t) \Rightarrow y(t)$.

But the result of the action depends on where the pendulum is, so if

$x(t)$: pendulum's angle, angular velocity, cart position and cart velocity

we need to find a function $y(t) = f(x(t), u(t))$

Rather From Data to Action

Let us first watch a thought-provoking video (from Carl Rasmussen at Cambridge). It takes 1.5 minutes.
...

What is going on?

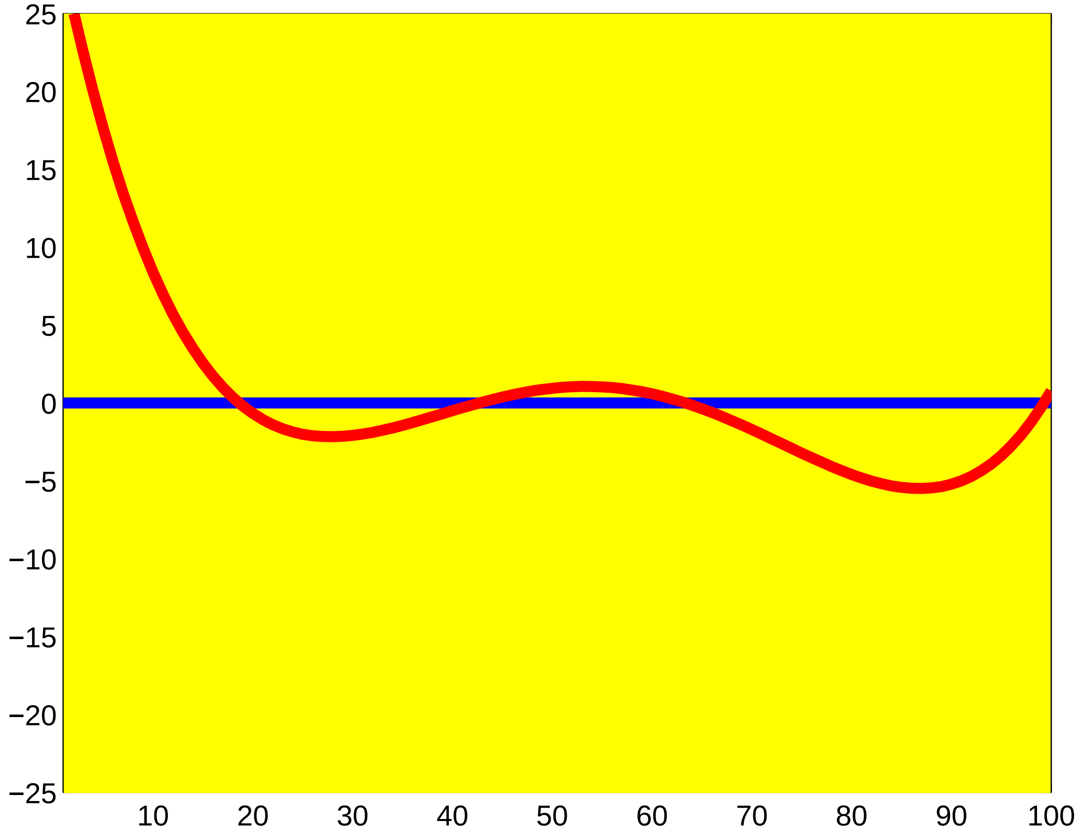
At time t, let the position of the pendulum be $y(t)$ and the control action (the force on the cart) $u(t)$.
Somehow we need to understand the mapping $u(t) \Rightarrow y(t)$.

But the result of the action depends on where the pendulum is, so if

$x(t)$: pendulum's angle, angular velocity, cart position and cart velocity

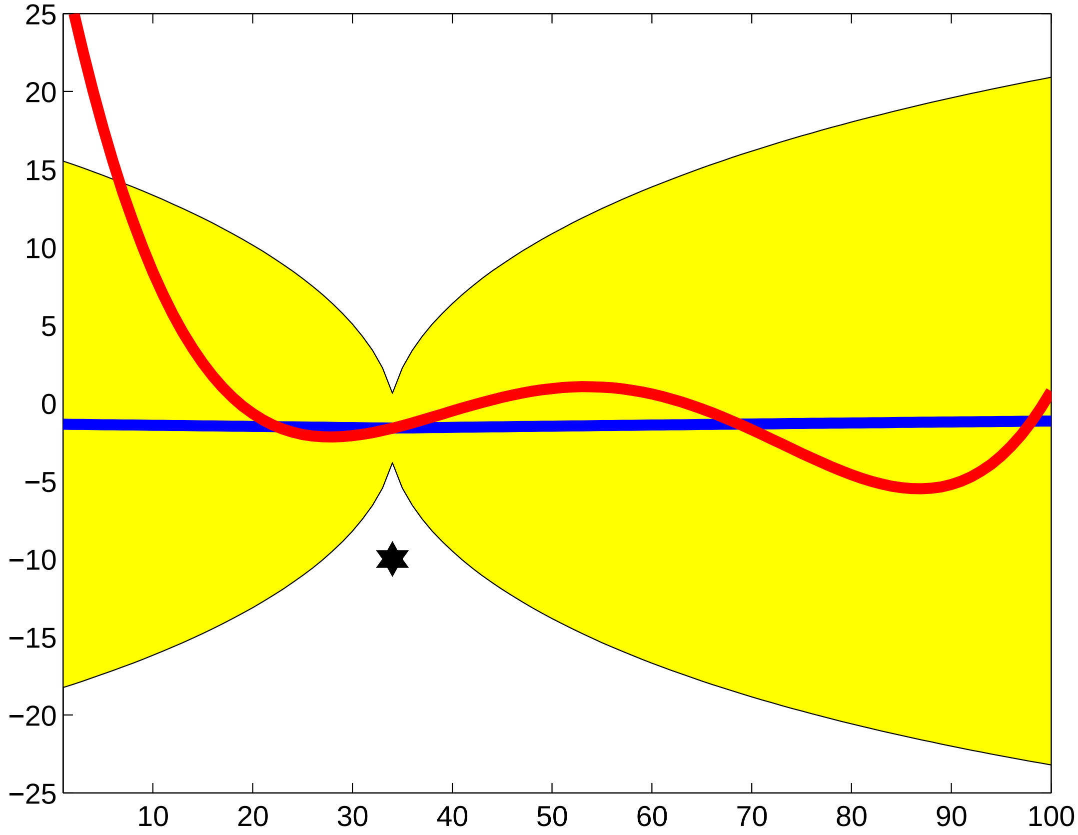
we need to find a function $y(t) = f(x(t), u(t))$

How to estimate a function f?

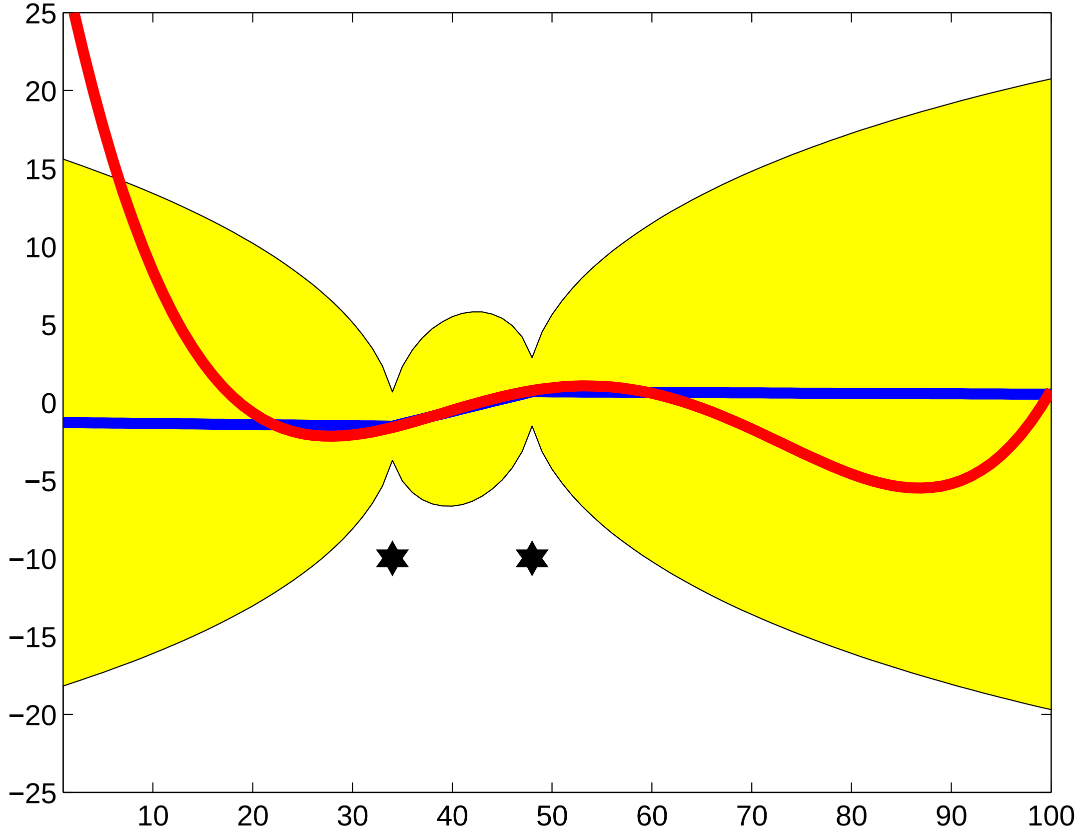


red: The unknown curve
blue: our current guess
yellow: our uncertainty

Learning by Bayes Theorem

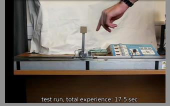


star: observation
blue: smooth curve with certain correlation
yellow: our uncertainty
The info from the measurement propagates due to the correlation

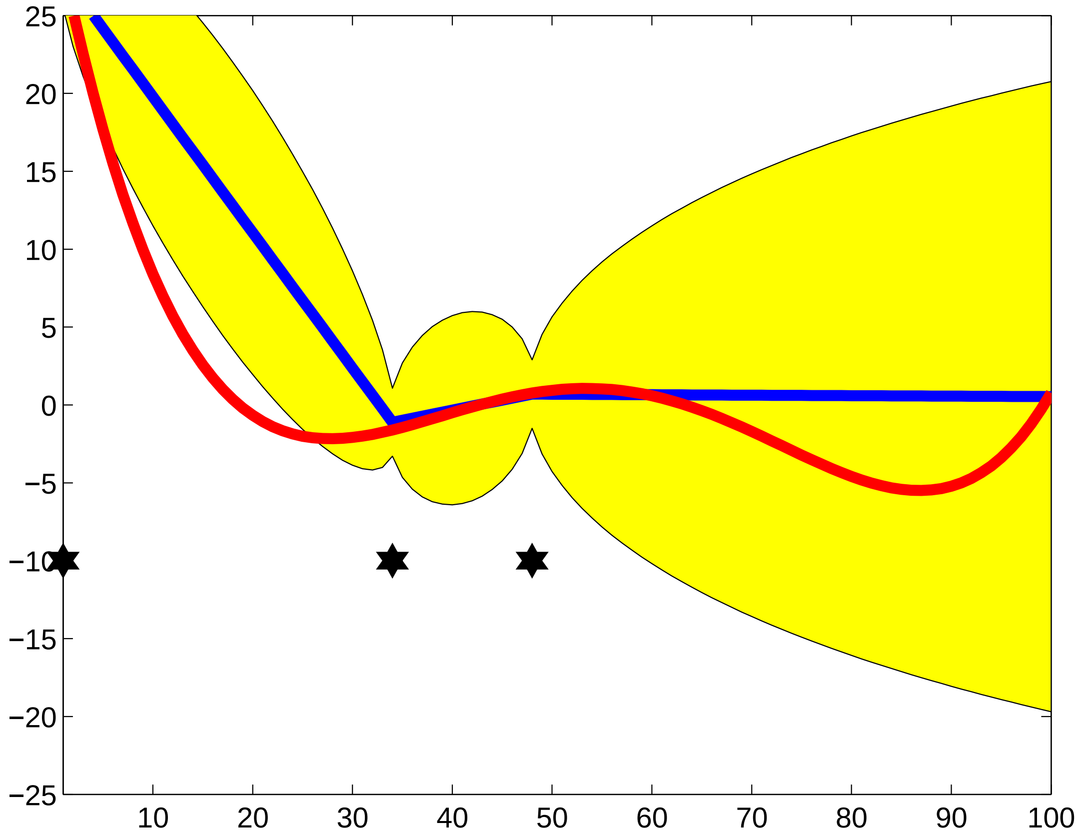


star: observation
blue: smooth curve with certain correlation
yellow: our uncertainty

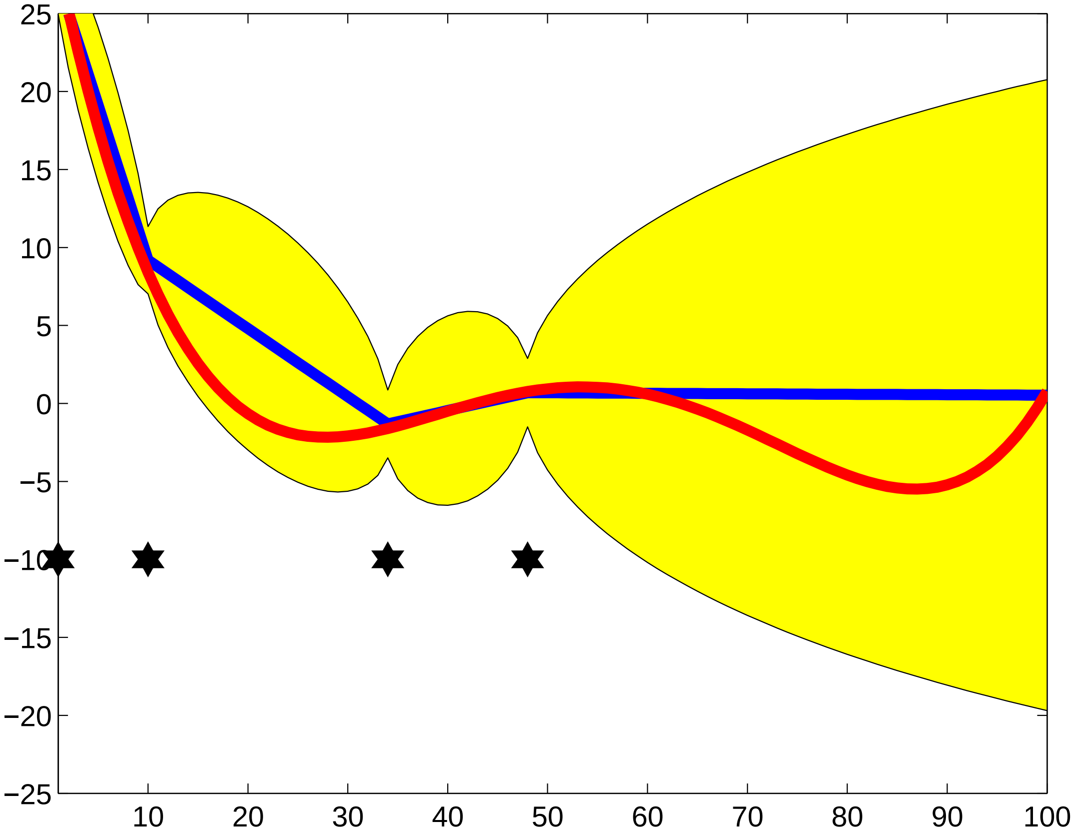
The info from the measurement propagates due to the correlation



Learning by Bayes Theorem - Kriging

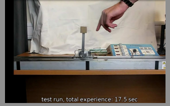


star: observation
blue: smooth curve with certain correlation
yellow: our uncertainty
The info from the measurement propagates due to the correlation

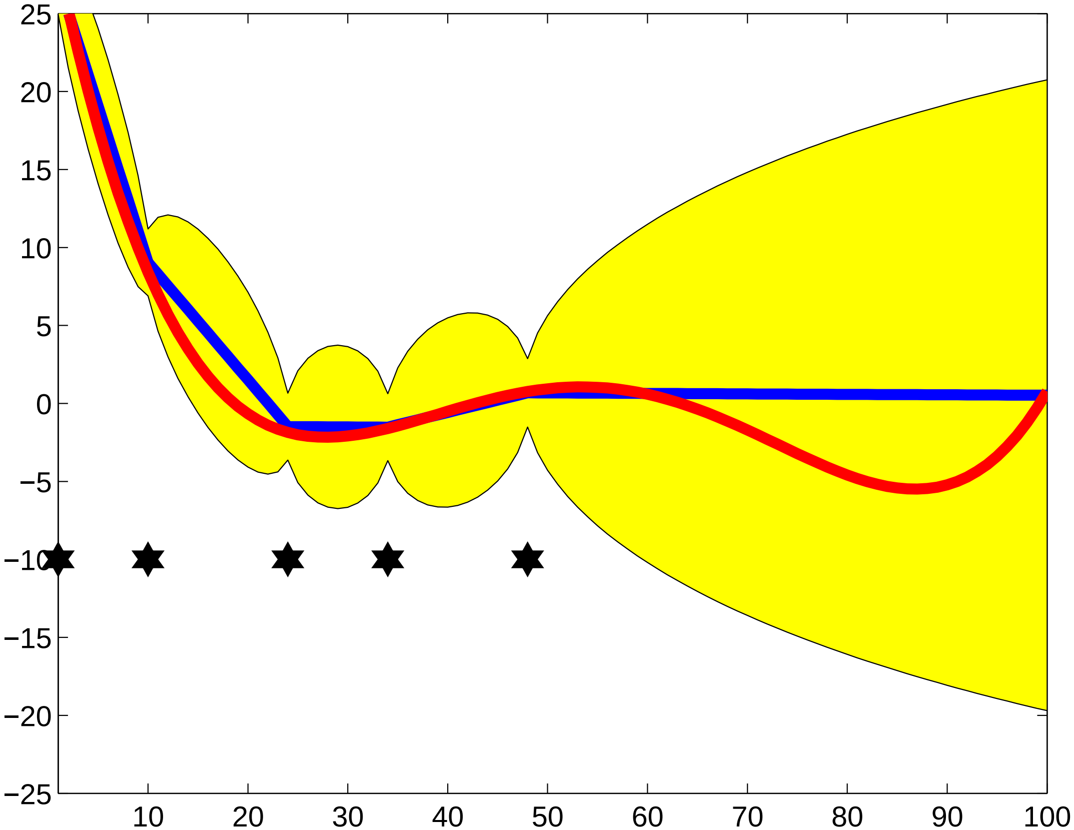


star: observation
blue: smooth curve with certain correlation
yellow: our uncertainty

The info from the measurement propagates due to the correlation

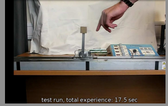


Learning by Bayes Theorem

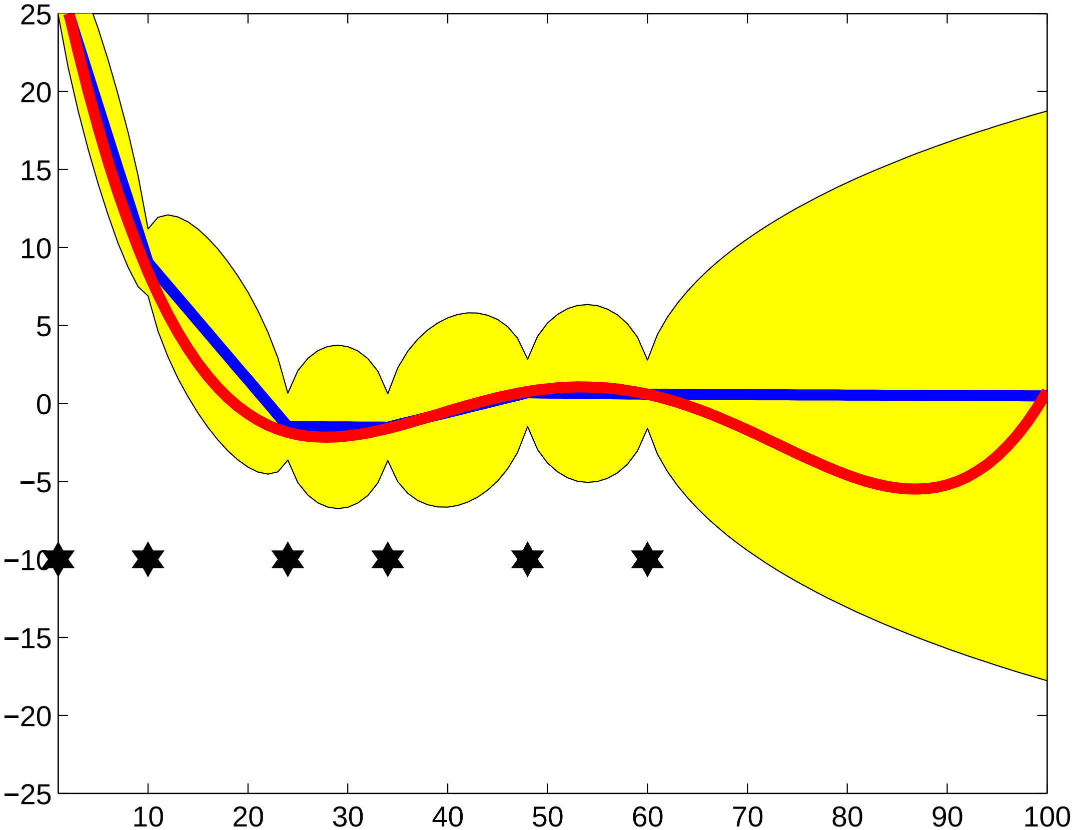


star: observation
blue: smooth curve with certain correlation
yellow: our uncertainty

The info from the measurement propagates due to the correlation



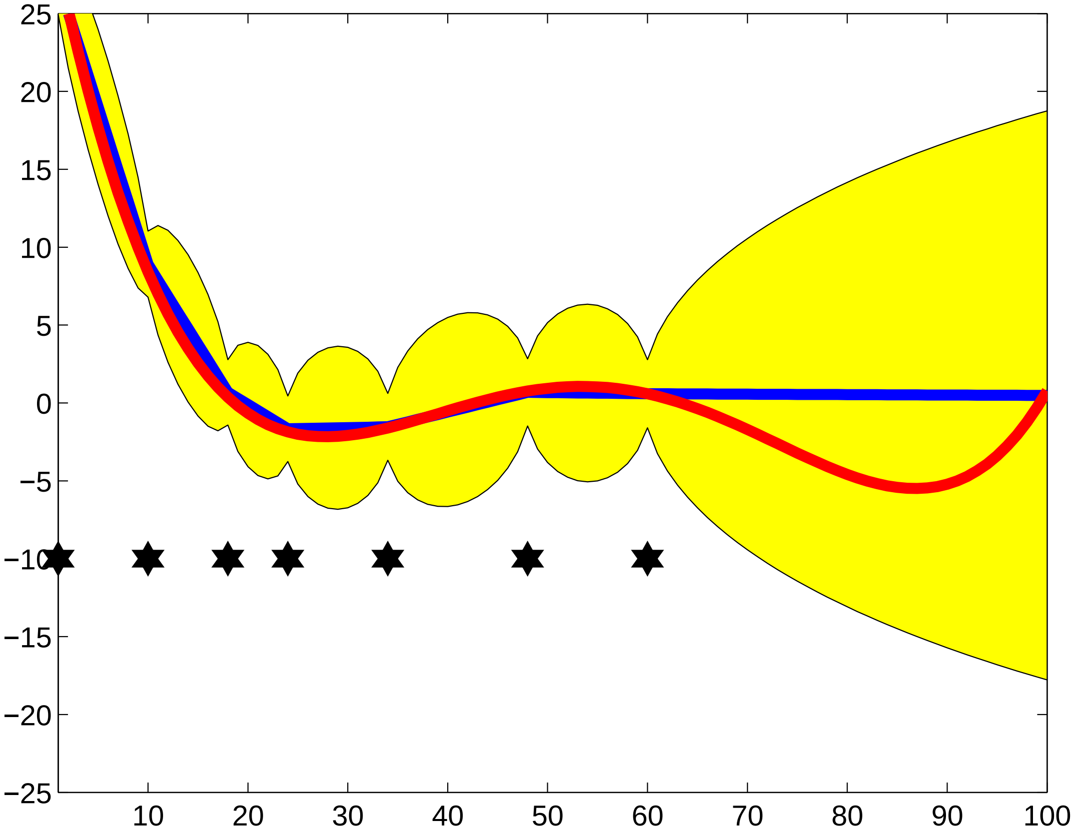
Learning by Bayes Theorem



star: observation
blue: smooth curve with certain correlation
yellow: our uncertainty

The info from the measurement propagates due to the correlation

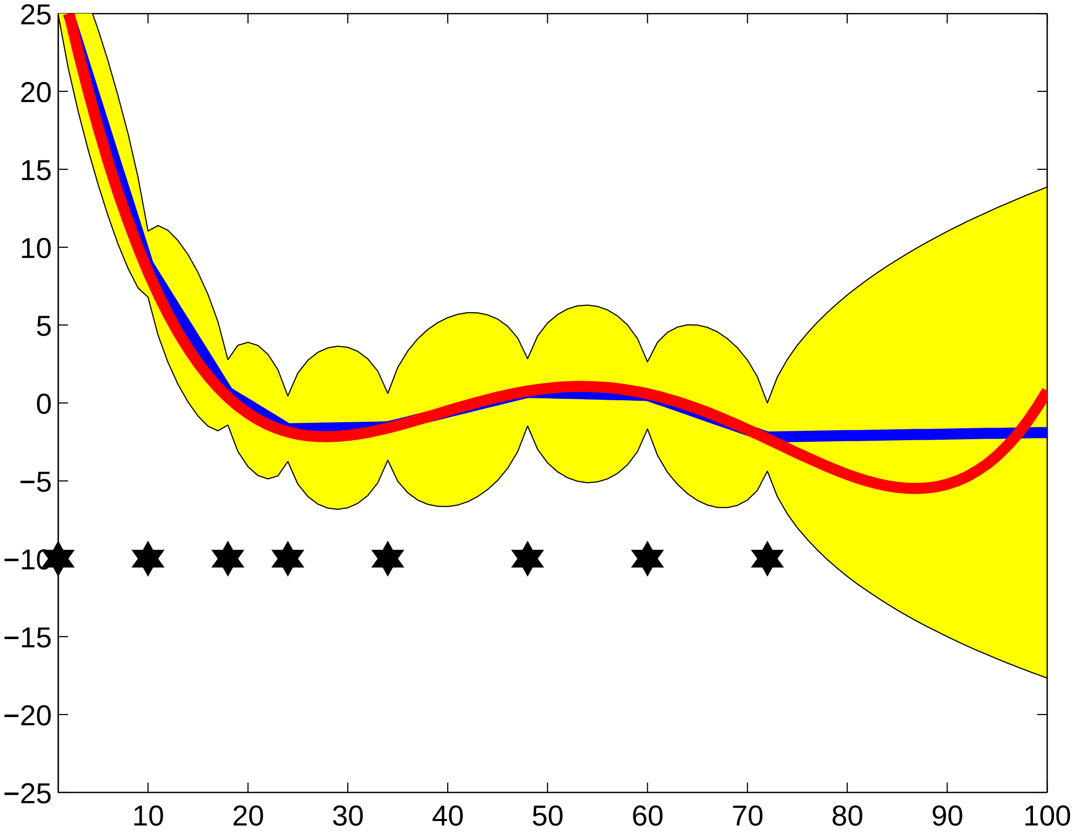
Learning by Bayes Theorem



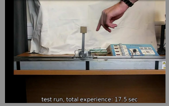
star: observation
blue: smooth curve with certain correlation
yellow: our uncertainty

The info from the measurement propagates due to the correlation

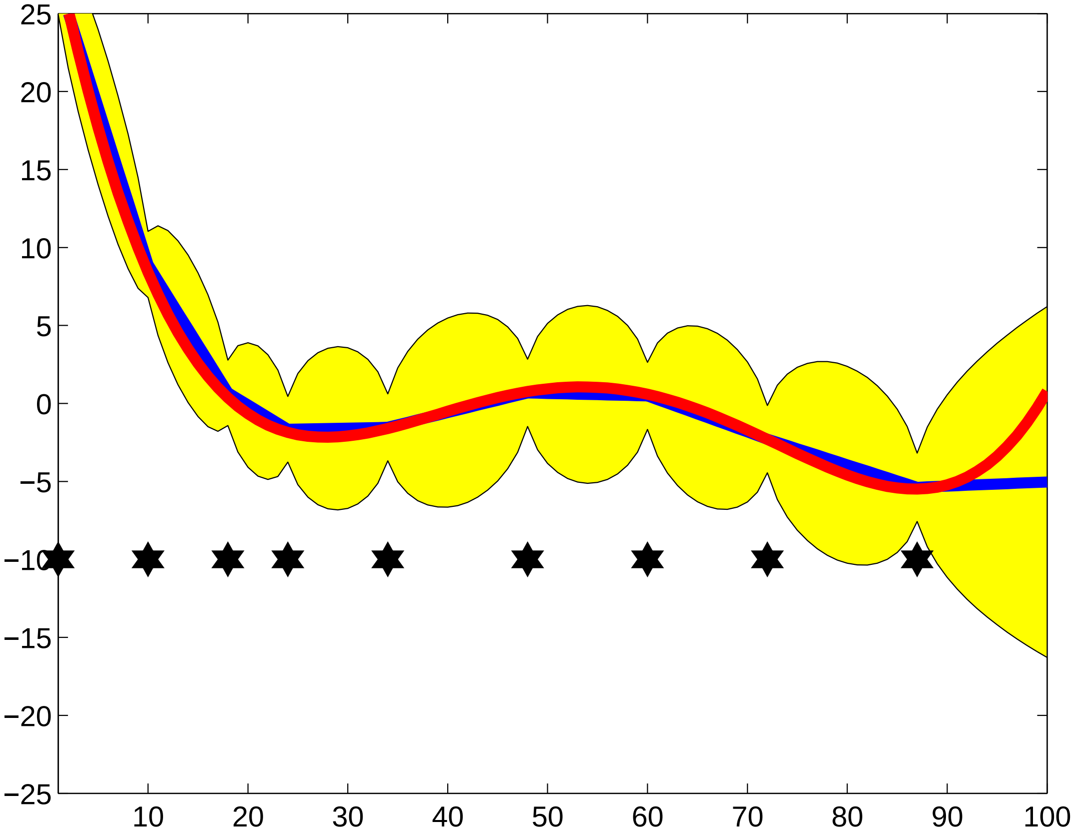
Learning by Bayes Theorem



star: observation
blue: smooth curve with certain correlation
yellow: our uncertainty
The info from the measurement propagates due to the correlation

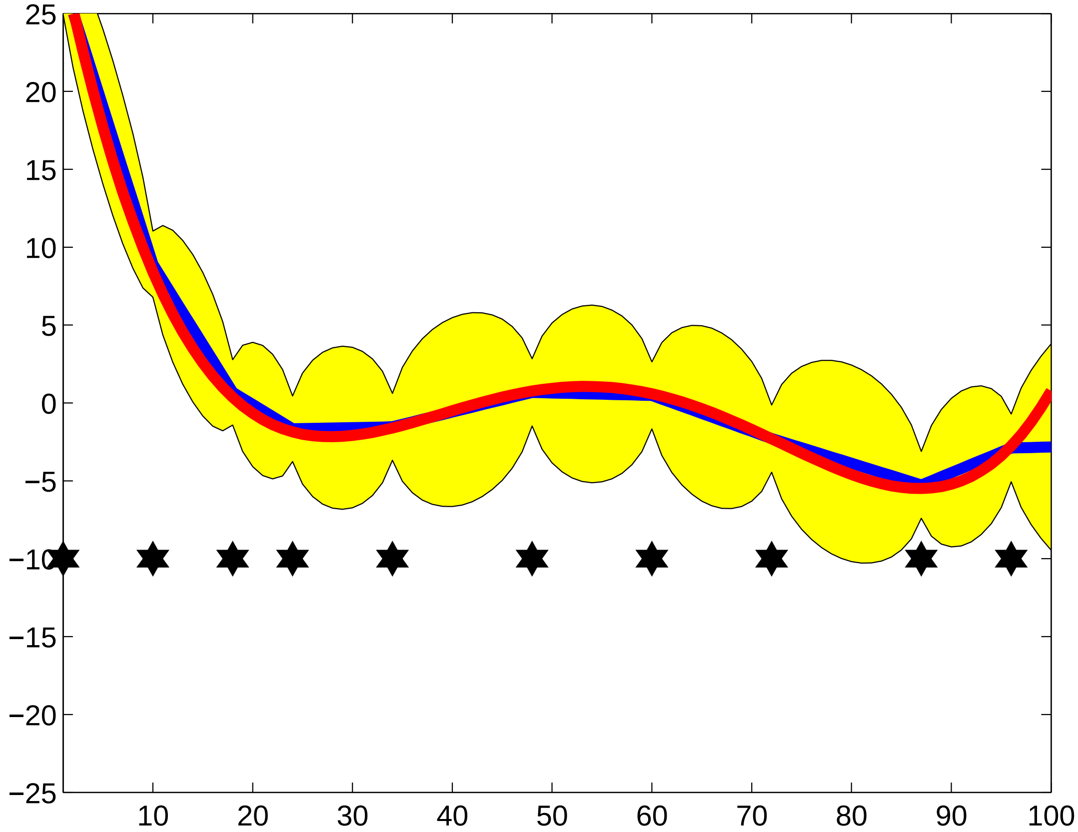


Learning by Bayes Theorem

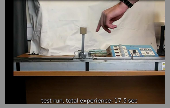


star: observation
blue: smooth curve
with certain correlation
yellow: our uncertainty

The info from the measurement propagates due to the correlation



We have a good estimate of the curve with 10 observations (which are pretty dense compared to the 880 observations in the 5-dimensional space of the pendulum)



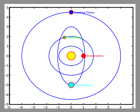
Connection to System Identification

BTW, finding the function f that predicts the next output(pendulum position) from the currect input (force on cart) and "state" (pendulum and cart movement) is the same problem as a prediction error method.

It is just that we have no parameterization of the prediction fuction, and don't want to introduce any physical knowledge, but just estimate it as a (smooth) curve.

Conclusions

- Very rich literature on building models from data ("The communities")



Conclusions

- Very rich literature on building models from data ("The communities")
- Relatively few leading principles ("The core")

Conclusions

- Very rich literature on building models from data ("The communities")
- Relatively few leading principles ("The core")
- Basic Principles are "classical"
  - Prediction error identification $\Leftarrow$ K.F. Gauss (1777-1855)
  - Pendulum – Prediction function learning $\Leftarrow$ T. Bayes (1702-1761)

Conclusions

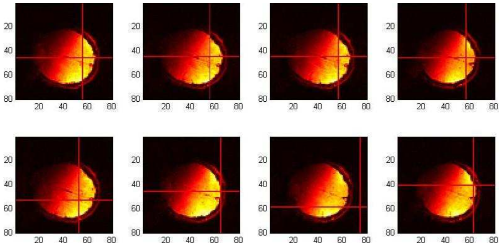
- Very rich literature on building models from data ("The communities")
- Relatively few leading principles ("The core")
- Basic Principles are "classical"
  - Prediction error identification $\Leftarrow$ K.F. Gauss (1777-1855)
  - Pendulum – Prediction function learning $\Leftarrow$ T. Bayes (1702-1761)
- Future:
  - Can't beat the theoretical limits
  - Can use (massively) more data
  - More "data mining" in model building
  - More contacts between the Communities

Methods – Bottom Line

Find function for predicting next observation!

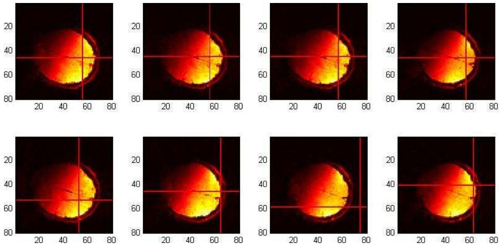
Find function for predicting next observation!

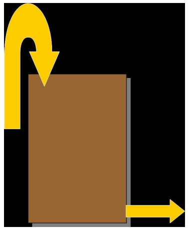
Find function for predicting next observation!



Methods – Bottom Line

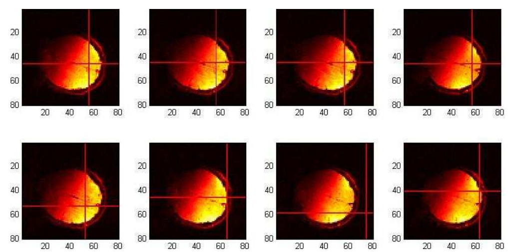
Find function for predicting next observation!

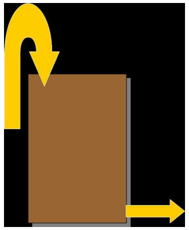


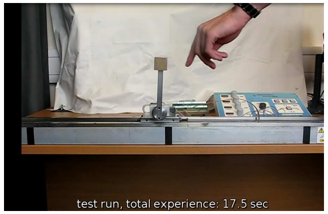


Methods – Bottom Line

Find function for predicting next observation!







Find function for predicting next observation!