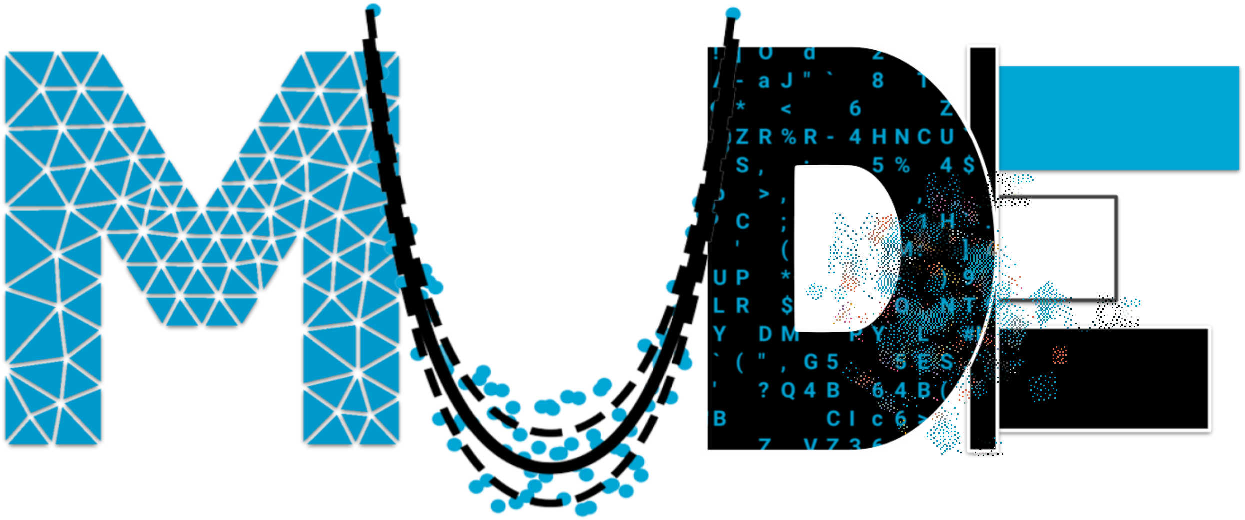
Welcome to...



Modelling, Uncertainty, and Data for Engineers

WEEK 2 !

Today's Lecture

- Recap of last week (activities)
- Demonstration (software)
- In-class activity: modelling concepts and an uncertain future
- Recap of last week (topics) + a quick look ahead
- Week 1.2 Topic: Uncertainty Propagation

What you should have done last week

- Find MUDE Website, online textbook, MUDE Files. Review course policies.
- Wednesday in-class session:
  - README.md instructions
  - Workshop (WS 1.1): Able to use conda and VS Code on your computer
  - Programming Assignment (PA 1.1): can run a notebook

  >>> Not critical to install everything exactly as in book; as long as you can activate `mude-base`, run a notebook

  >>> If you are still having issues, come to office hours!
- Friday in-class session:
  - You have a group assigned in Brightspace and met them Friday
  - Group Assignment (GA 1.1): nothing has to be turned in; solution online via MUDE Files page
  - BuddyCheck: released Friday, due Monday before 11:00. Just click the link and do it (more info on Wed)

  >>> No group? Not sure who is in your group? Other questions? Email MUDE-CEG@tudelft.nl

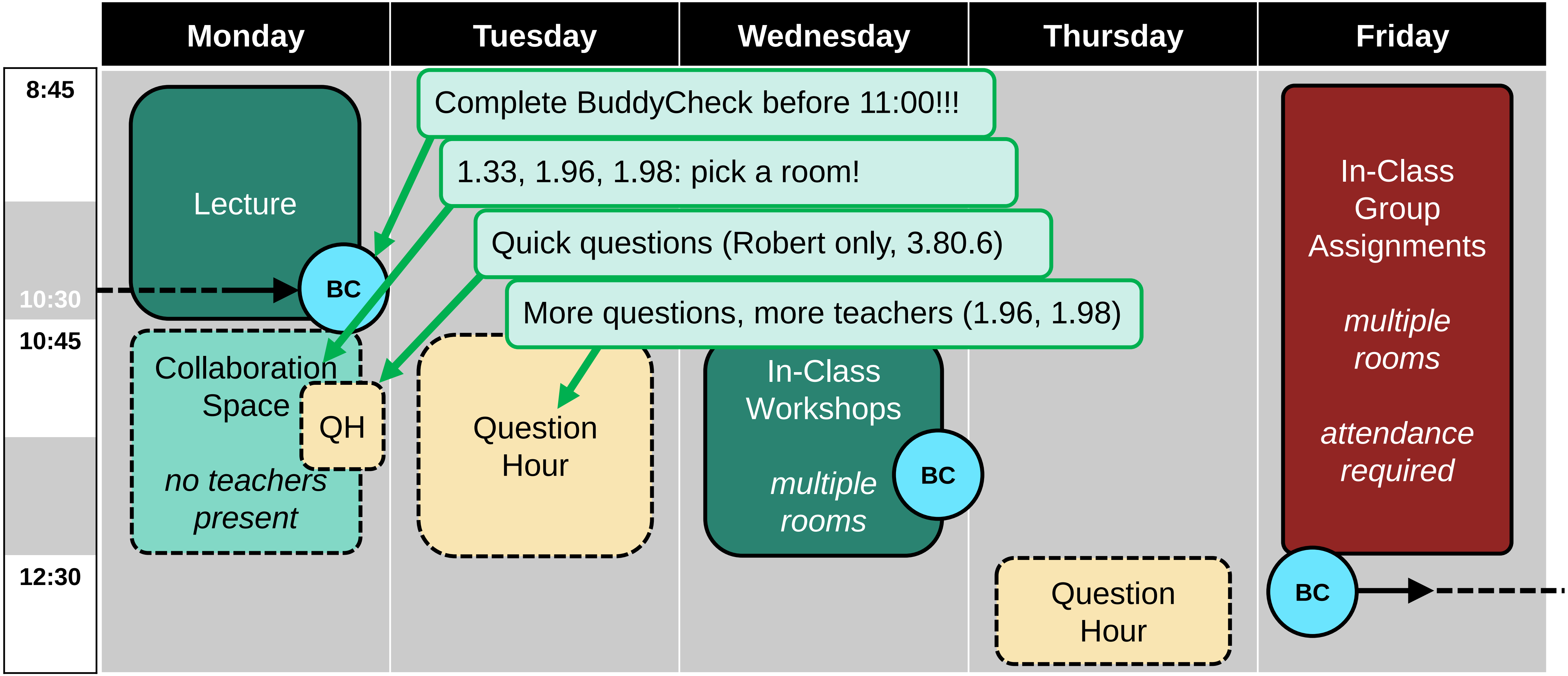
Demonstration

- Python, conda, etc
- VS Code
- Jupyter Notebooks

Demonstration

- MUDE Website, Book, Files
- Setup of computer and Python
  - Working directory setup
  - Checking conda: `conda --version`
  - Listing environments: `conda env list`
  - Looking at file `environment.yml`
- Some fun conda stuff
  - `conda list`
  - `conda info`

- Setup of VS Code
  - Opening a new folder / workspace
  - File viewer, Extensions
  - Terminal: setting defaults, checking conda
  - Activating an environment
- Jupter Notebooks
  - Overview of structure
  - Cell types
  - Running a cell, selecting environment
  - Restarting a kernel
  - Saving outputs (or not)
- Some fun with AI?



Monday

Tuesday

Wednesday

Thursday

Friday

8:45

10:30

10:45

12:30

Lecture

BC

Complete BuddyCheck before 11:00!!!

1.33, 1.96, 1.98: pick a room!

Quick questions (Robert only, 3.80.6)

More questions, more teachers (1.96, 1.98)

Collaboration Space

QH

no teachers present

Question Hour

In-Class Workshops

multiple rooms

BC

Question Hour

In-Class Group Assignments

multiple rooms

attendance required

BC

Programming Assignment: any time during the week, but... **Finish before Friday!**

BC = BuddyCheck: opens Fri (closes Mon); review results Wed with group

Question Hours (optional): Mon 11.00-12.00, Tue 10:45-12:30, Thu 12:30-13:30

In-Class Activity

Useful links:

- mude.citg.tudelft.nl/2024/book/modelling/gof.html
- mude.citg.tudelft.nl/2024/files/Week_1_2/

- We will use the file `In_Class_Activity` (located in MUDE Files, Week 1.2)

→ Read directly via the browser; can also download zip and run notebook (not needed in class)

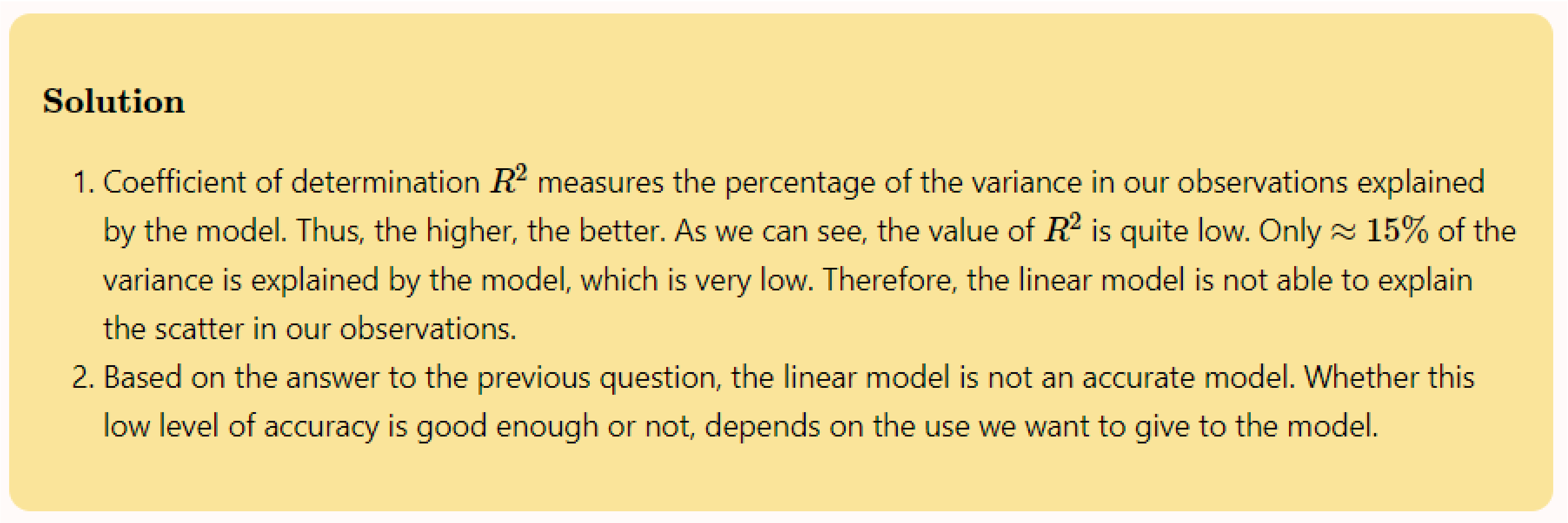
- Part 1-3: import data, preliminary analysis, create a model. **Validation: Goodness of Fit**.
- Part 4: confidence intervals. **Validation: would you use it?**.
- Part 5: a non-linear model. **Validation: Goodness of Fit**. **Which model is better?**

Read quickly over Parts 1-3 (stop at Part 4), then we will answer these questions (together):

General: what is the "model?"
3.1: What is the coefficient of determination? Is the model accurate?
3.2: Is the visual GoF consistent?
3.3: What is RMSE telling us in general? What does it tell us about the model?
3.4: What about *rbias*?

Task 3.1

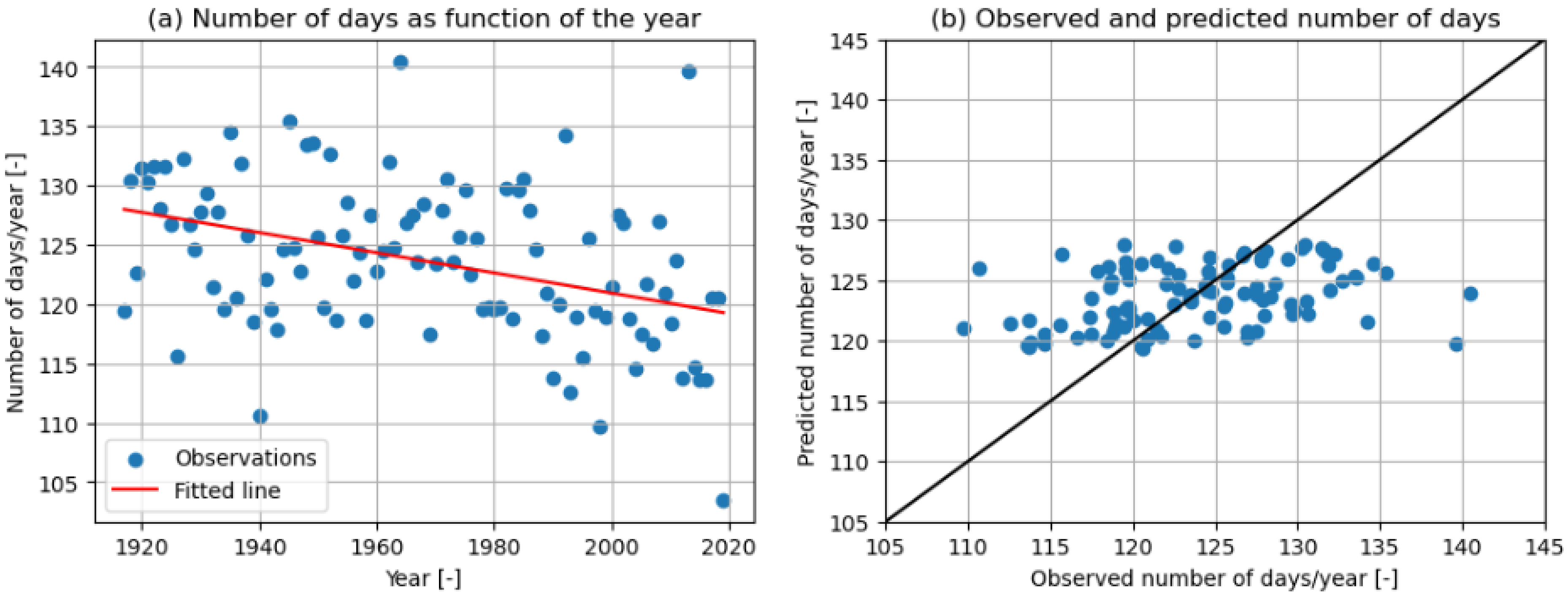
```
Coefficient of determination R^2 = 0.151
Intercept q = 291.239
Slope m = -0.085
```



Solution

1. Coefficient of determination R^2 measures the percentage of the variance in our observations explained by the model. Thus, the higher, the better. As we can see, the value of R^2 is quite low. Only $\approx 15\%$ of the variance is explained by the model, which is very low. Therefore, the linear model is not able to explain the scatter in our observations.
2. Based on the answer to the previous question, the linear model is not an accurate model. Whether this low level of accuracy is good enough or not, depends on the use we want to give to the model.

Task 3.2



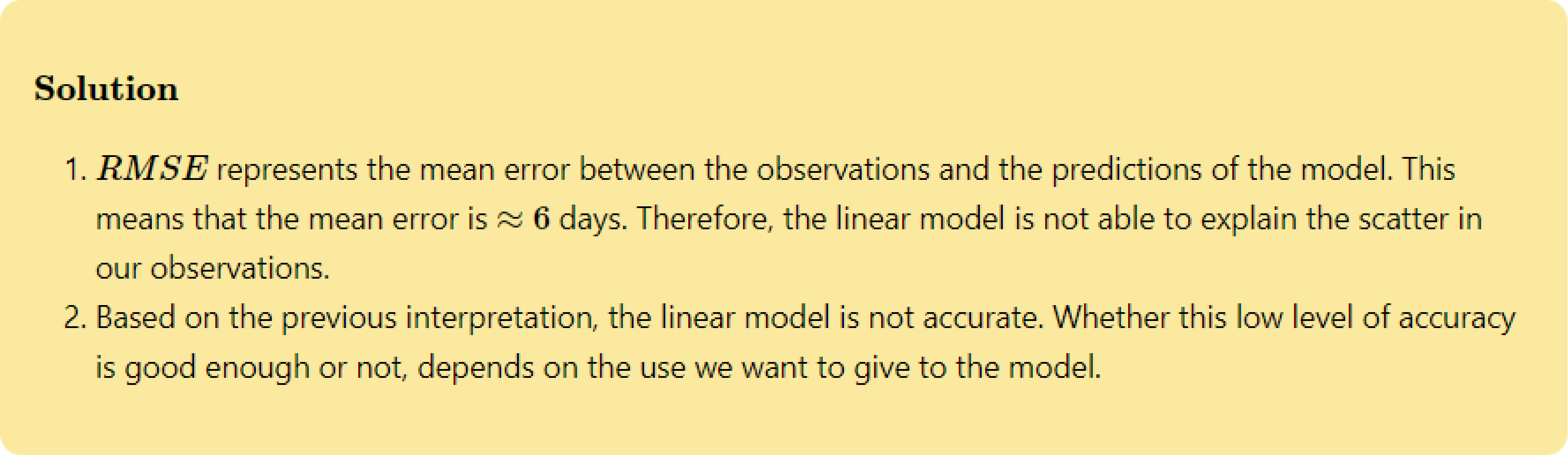
Solution

In the plot (a), we observe that the observations have a high scatter around the fitted line, and visually it looks consistent with the R^2 value found above.

In the plot (b), we compare the observations with the predictions of the model. The perfect fit would correspond to the 45-degrees line in black. Thus, the model has a poor performance as we already quantified using the coefficient of determination. Both results are aligned.

Task 3.3, 3.4

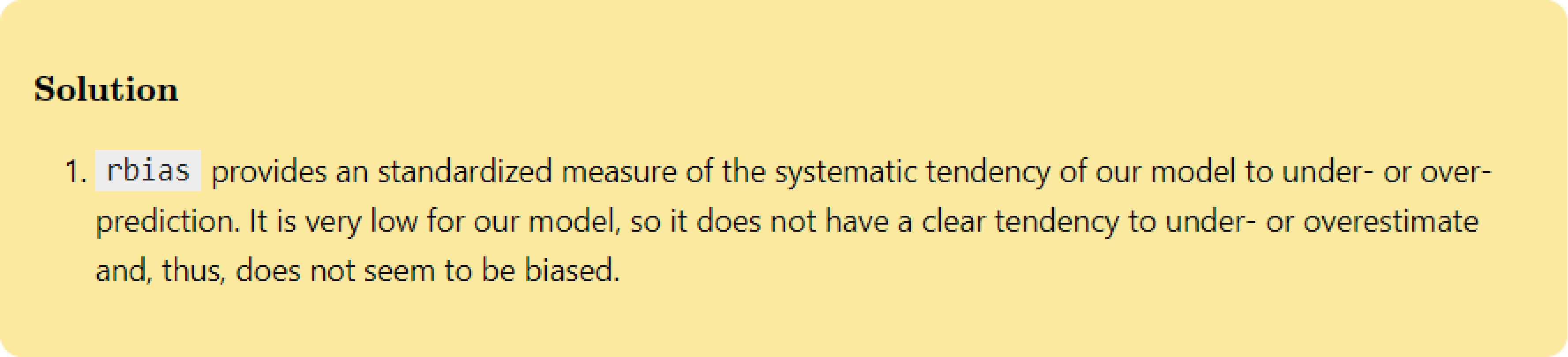
```
RMSE = 6.003
```



Solution

1. $RMSE$ represents the mean error between the observations and the predictions of the model. This means that the mean error is $\approx \mathbf{6}$ days. Therefore, the linear model is not able to explain the scatter in our observations.
2. Based on the previous interpretation, the linear model is not accurate. Whether this low level of accuracy is good enough or not, depends on the use we want to give to the model.

```
rbias = 0.002
```



Solution

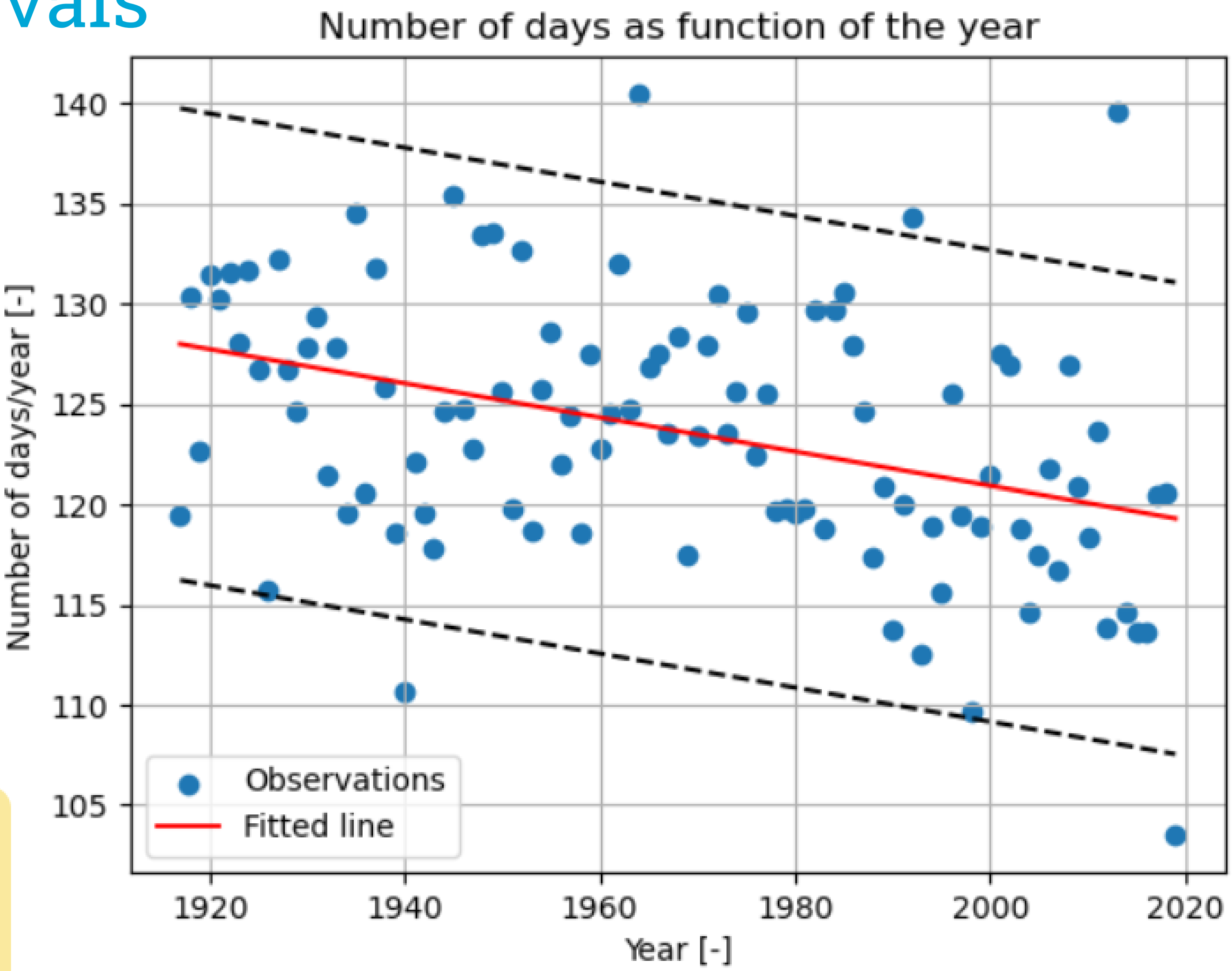
1. `rbias` provides an standardized measure of the systematic tendency of our model to under- or over-prediction. It is very low for our model, so it does not have a clear tendency to under- or overestimate and, thus, does not seem to be biased.

Part 4: Confidence Intervals

- Study the code and try to identify what the analysis accomplishes.
- What can you conclude from the figure?
- Would you bet $3? What about $3000?

Part 4: Confidence Intervals

- What can you conclude from the figure?
- Would you bet $3? What about $3000?



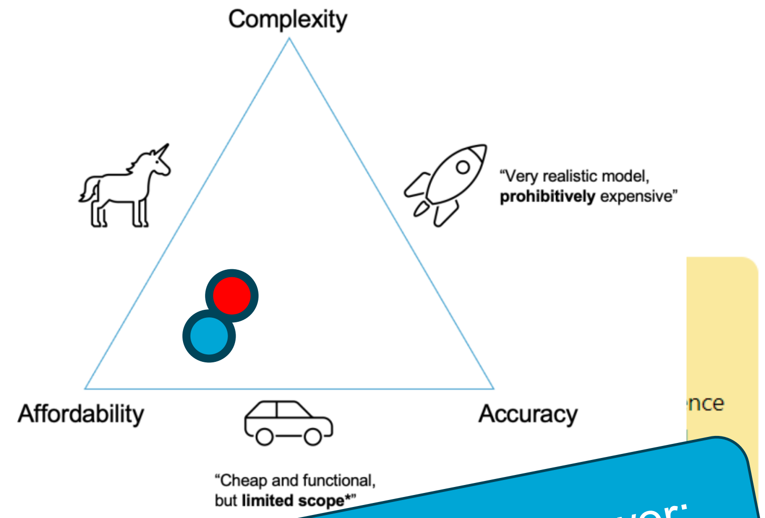
Solution: If you consider that you need to place a bet with not only the day but also the hour and minute at which the ice would break, the model is not accurate enough. You can see that the confidence interval spans almost 20 days!

Part 5: Non-linear Models

- Study the code and try to identify what the analysis accomplishes.
- What is the “model?”
- 5.1: is the quadratic model better than the linear one?

Part 5: Non-linear Models

- What is the “model?”
- 5.1: is the quadratic model better than the linear one?

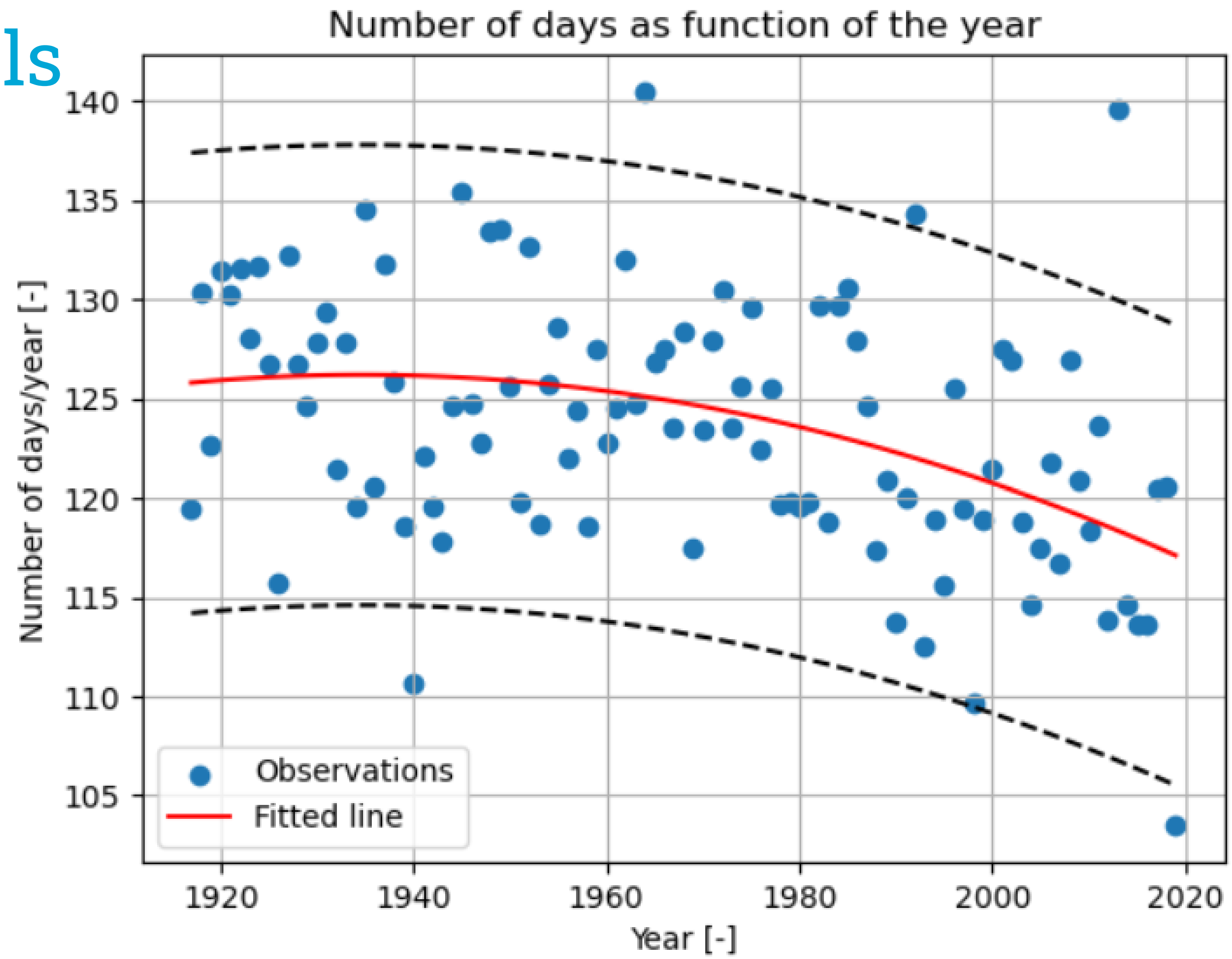

Complexity
"Very realistic model, **prohibitively** expensive"
Affordability
Accuracy
"Cheap and functional, but **limited scope**"

Solution: No! The GOF m
complicated (e.g., it incluc

At the same time we can … nce
bounds. Moreover, the 'm
number of days. This varia
(e.g., number of heatwaves). The question is then: …
related explanatory variable to get a b…
whether you are int…

…ation Theory in week 1.4.

See solution on MUDE Files…short answer:
No! (added complexity not worthwhile)

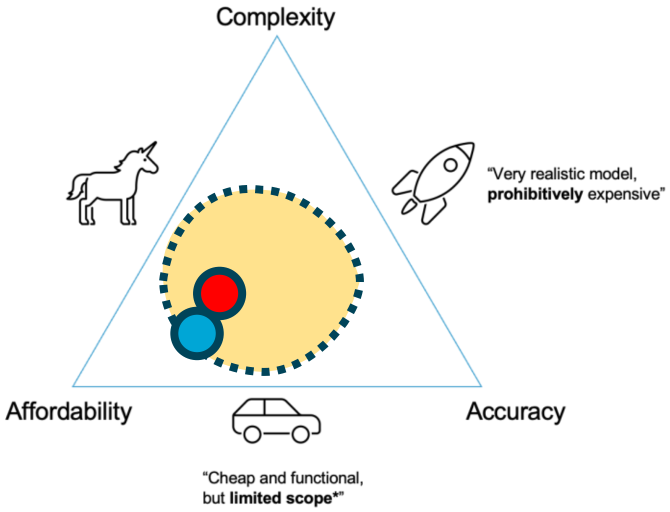


● Linear Model

● Quadratic Model

Were our models good enough?

	No. Parameters	RMSE	R^2	rbias
Line	2	6.00	0.15	0.002
Parabola	3	5.92	0.18	0.002

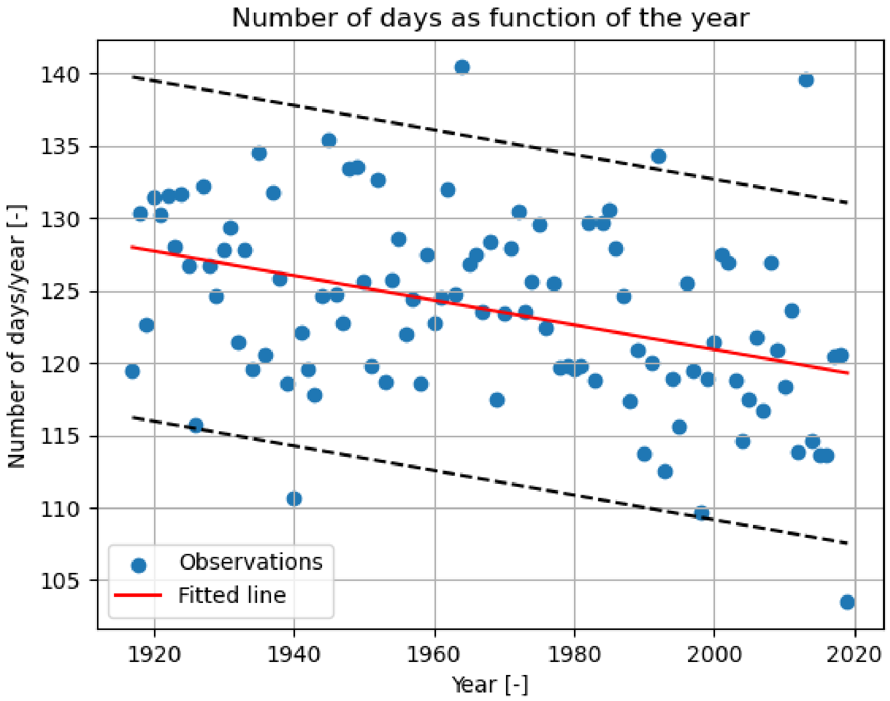


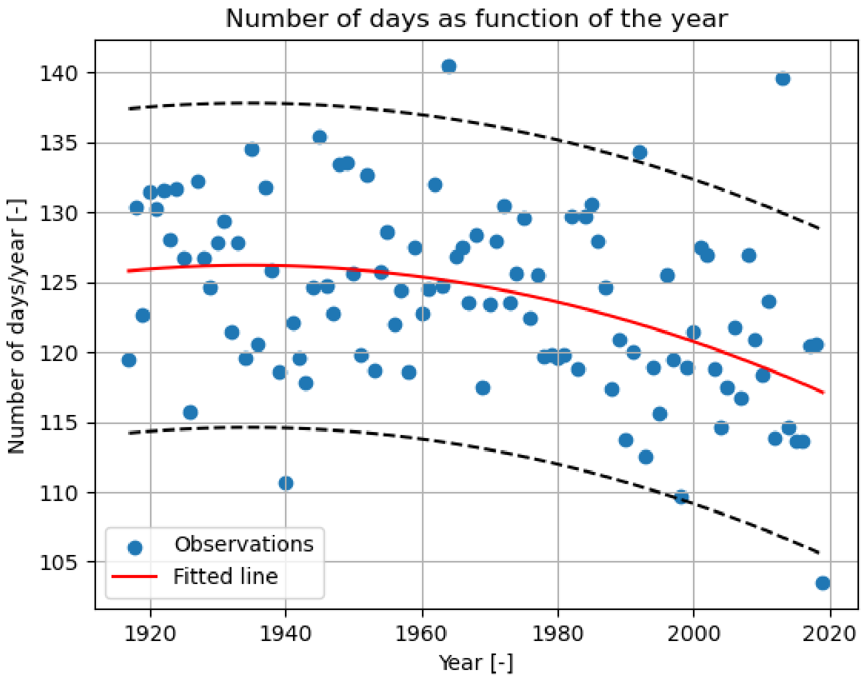
Complexity
"Very realistic model, **prohibitively** expensive"
Affordability
Accuracy
"Cheap and functional, but **limited scope***"

● Linear Model

● Quadratic Model

◌ Let's go here!

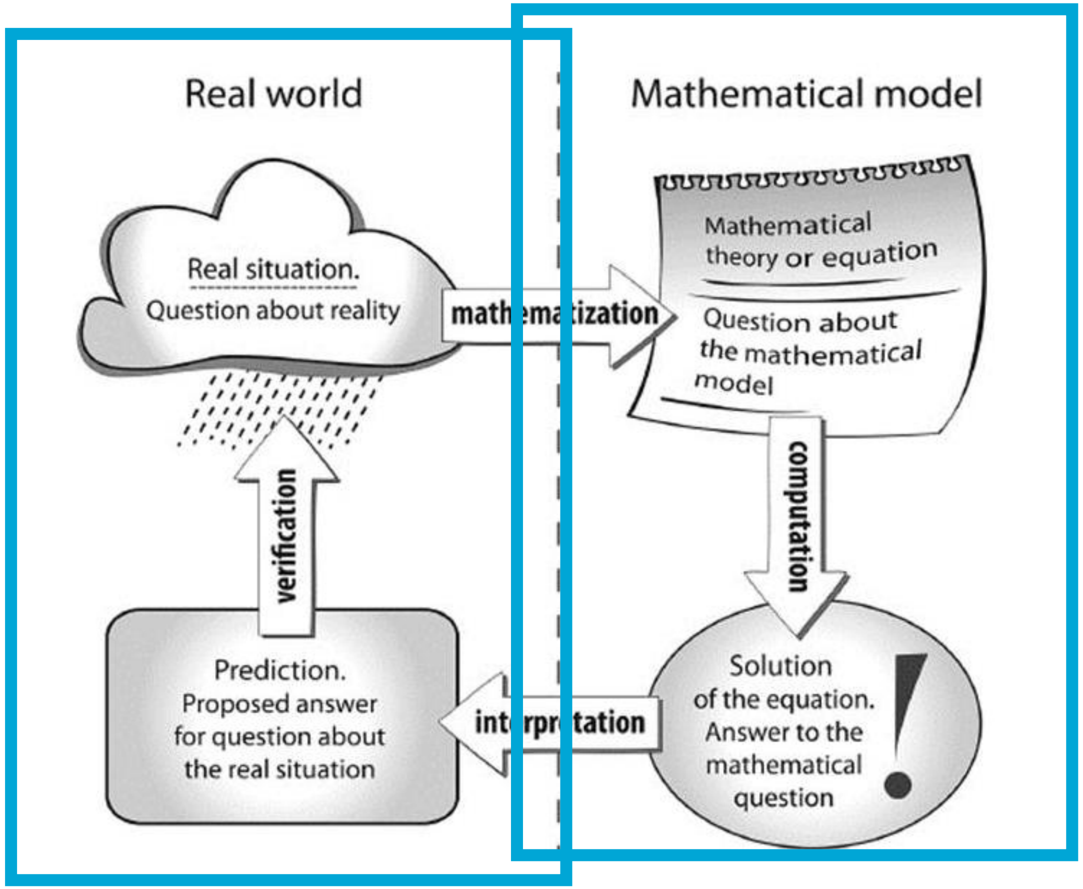




Recap of Week 1 Topics

1. GA 1.1, Task 1: affordability-complexity-accuracy diagram
2. GA 1.1, Task 2, 3: easy to connect variables; process harder
3. In-Class Activity: simple model is easy, not accurate
4. Real life is an iterative approach:

   data / phenomenological / mechanistic


Real world
Mathematical model
Real situation.
Question about reality
mathematization
Mathematical theory or equation
Question about the mathematical model
computation
Solution of the equation. Answer to the mathematical question
interpretation
Prediction. Proposed answer for question about the real situation
verification

Modelling, Uncertainty, and Data for Engineers

Where we are going next

- Next few weeks (2-4): data-driven, with a hint of phenomenological
- Weeks 5-6: mechanistic
- Week 7: stochastic (probabilistic)
- Week 8: combination!

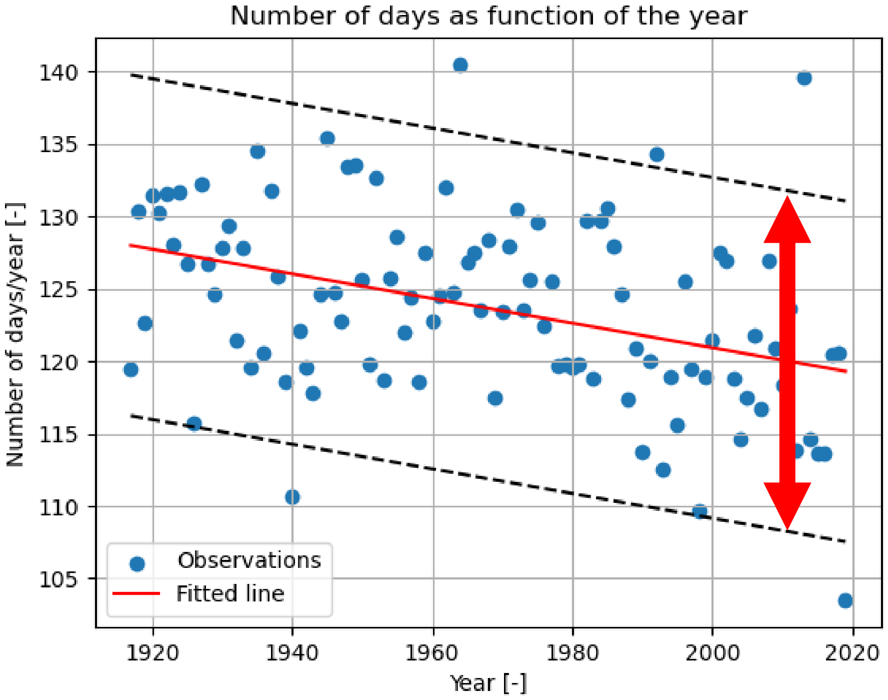
- Linear: useful for learning, understanding
- Non-linear: reality!

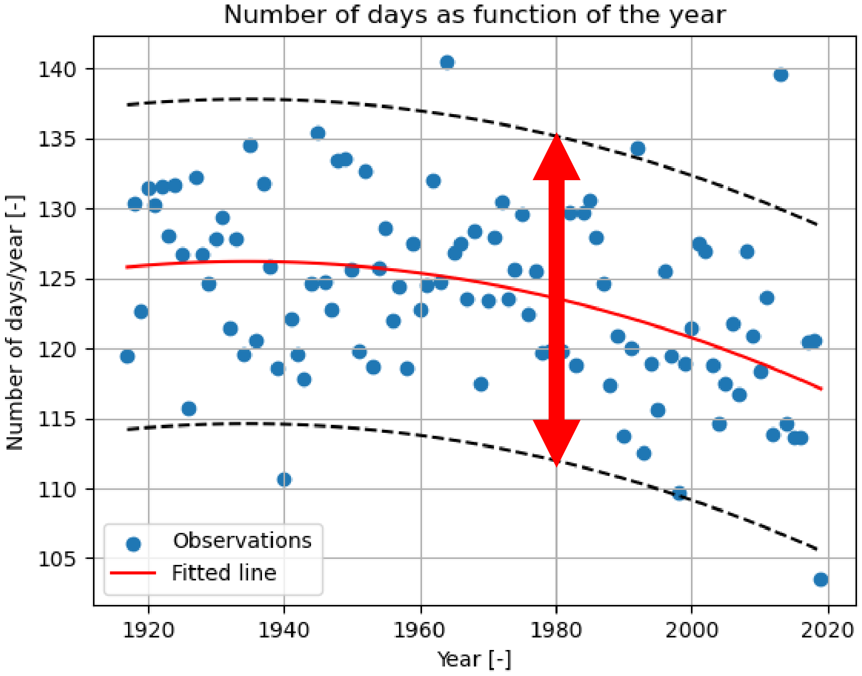
- First, we need a few tools to help quantify uncertainty (Week 2).

  Let’s revisit last week to see what we need…

Can we make some sense of the large CI?

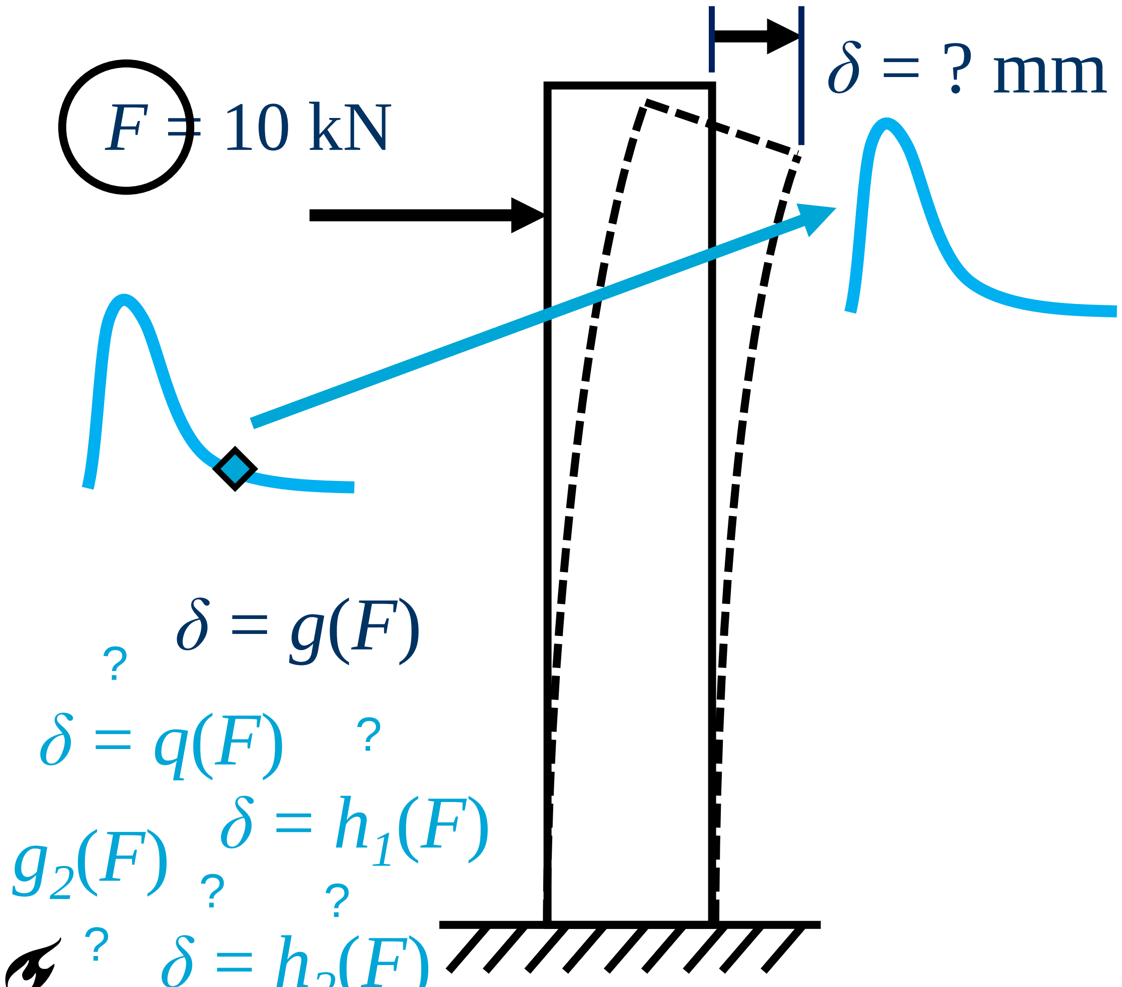
	No. Parameters	RMSE	R^2	rbias
Line	2	6.00	0.15	0.002
Parabola	3	5.92	0.18	0.002





Uncertainty – What is it?

Deterministic design (pre-MUDE) – parameters given

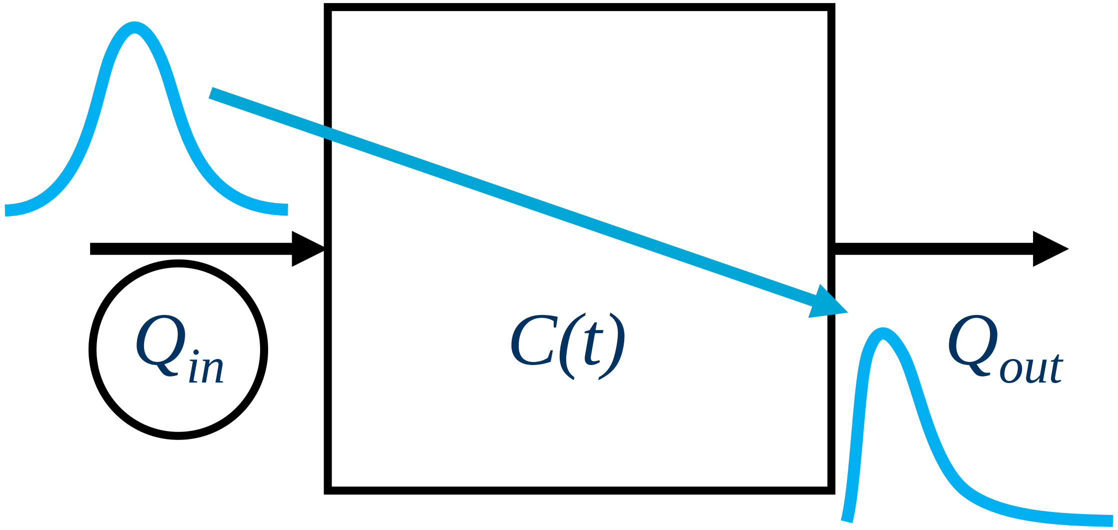


$\delta = g(F)$?

? $\delta = q(F)$?

$\delta = g_2(F)$? $\delta = h_1(F)$?

? $\delta = h_2(F)$?



$Q_{out} = g(Q_{in}, C(t))$

? ? ? $Q_{out} = h(Q_{in}, C(t))$?

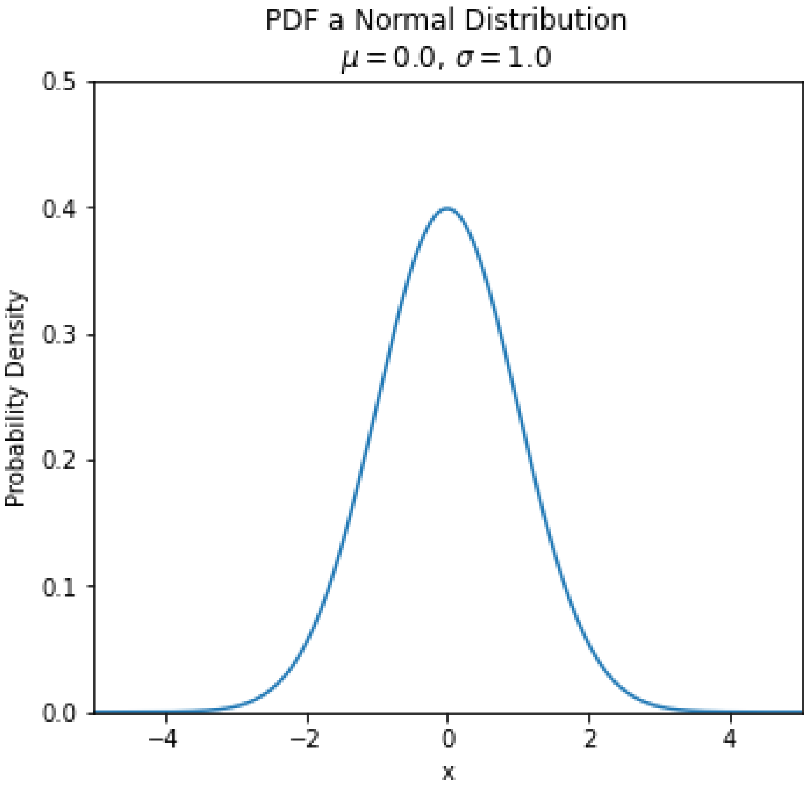
$Q_{out} = g_2(Q_{in}, C(t))$? ?

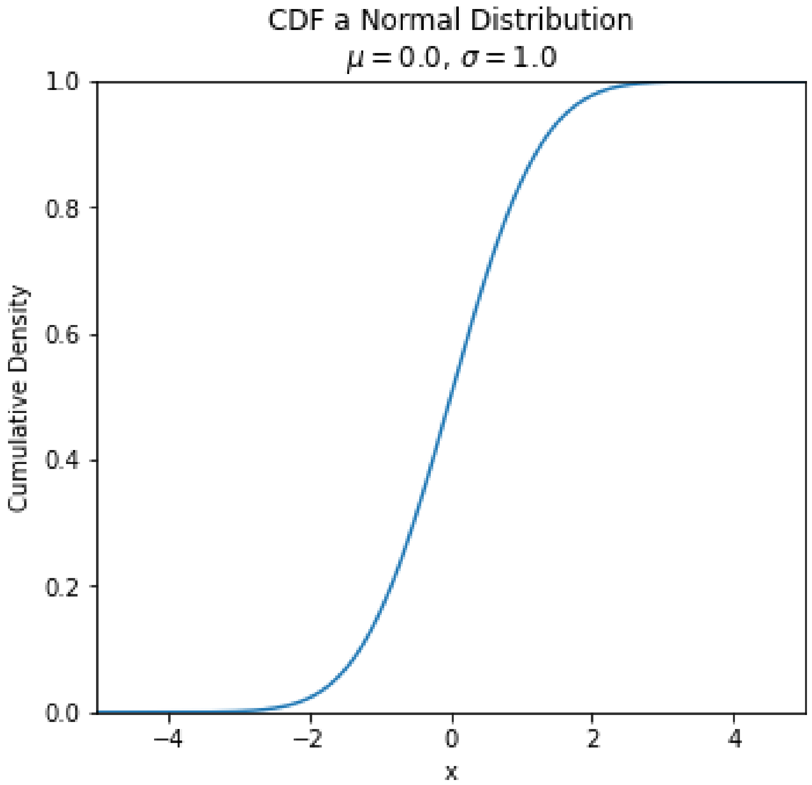
Uncertainty – What is it?

- “Uncertainty Quantification” … have you seen this term?
    - We avoid it in MUDE: often presented as a “black box”; assumptions can be unclear / used inconsistently.

Instead, we will focus on theoretical concepts: fundamentals of probability and statistics

- We will rely on:
    - Expectation, Variance
    - Covariance / Dependence
    - Probability Distributions: continuous, parametric
    - Set Theory





Uncertainty Classification – Introduction

Aleatoric

- intrinsic phenomenon; typically associated with variations that occur in nature

Epistemic

- lack of knowledge; often called model uncertainty

Error

- deficiency in any stage of modelling/simulation not due to lack of knowledge

Break-up Day Uncertainty

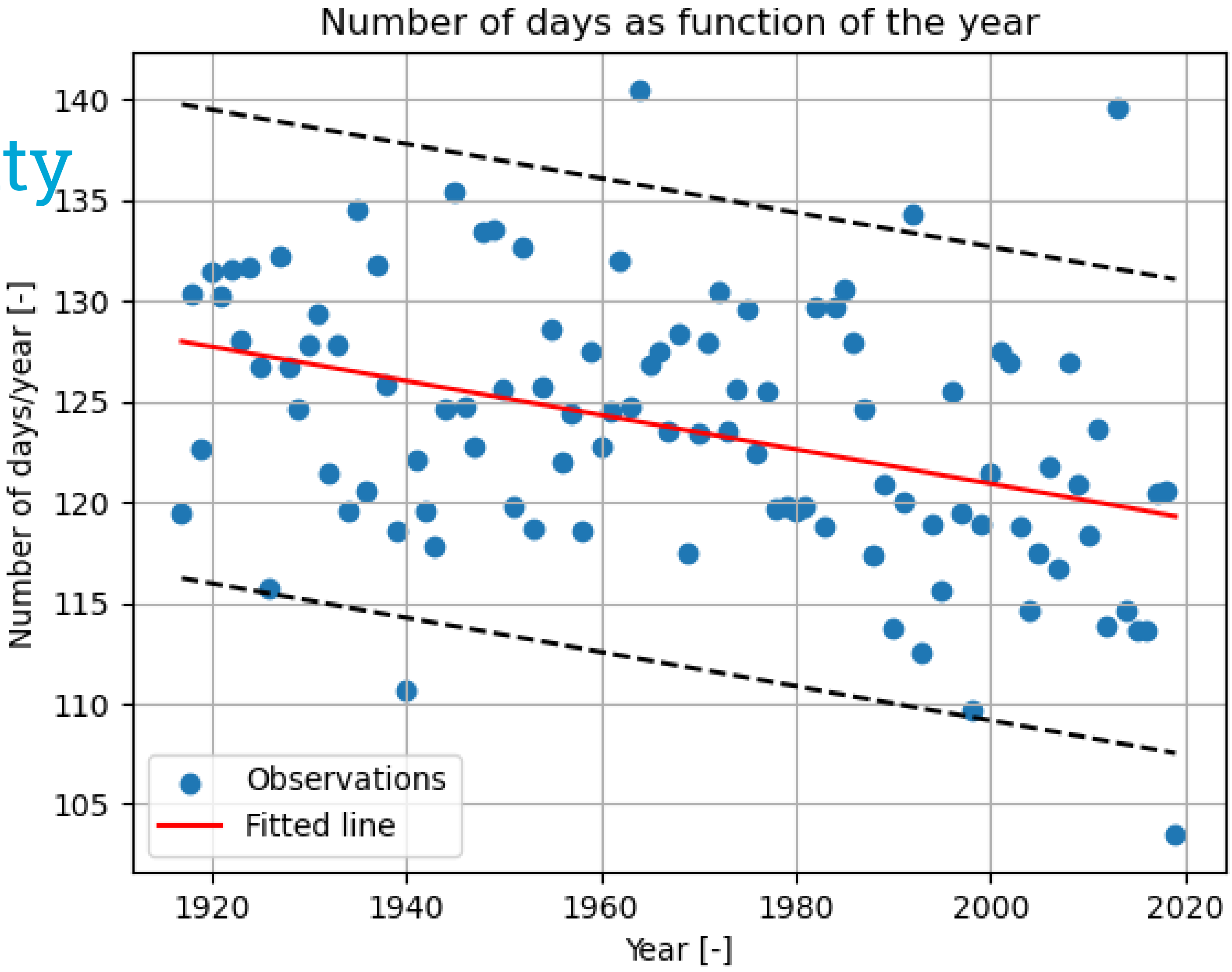
- Days, D, as a function of year, Y:

D = q(Y)   >>>   D = m*Y + b

>>> Use a distribution to represent uncertainty

Sources of uncertainty?

- Variation of natural phenomena
  - Aleatoric
- Model uncertainty?
  - Epistemic; not relevant
- Observations?
  - Error (accuracy / precision of historic points)
  - Epistemic (“model” of break-up / definition)

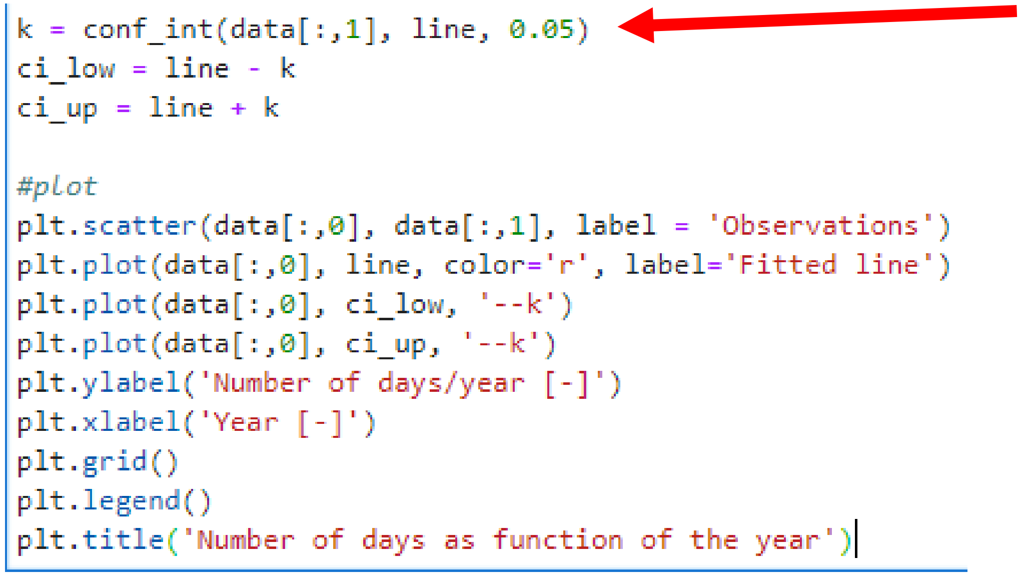


Can our confidence interval quantify this?

Confidence Intervals

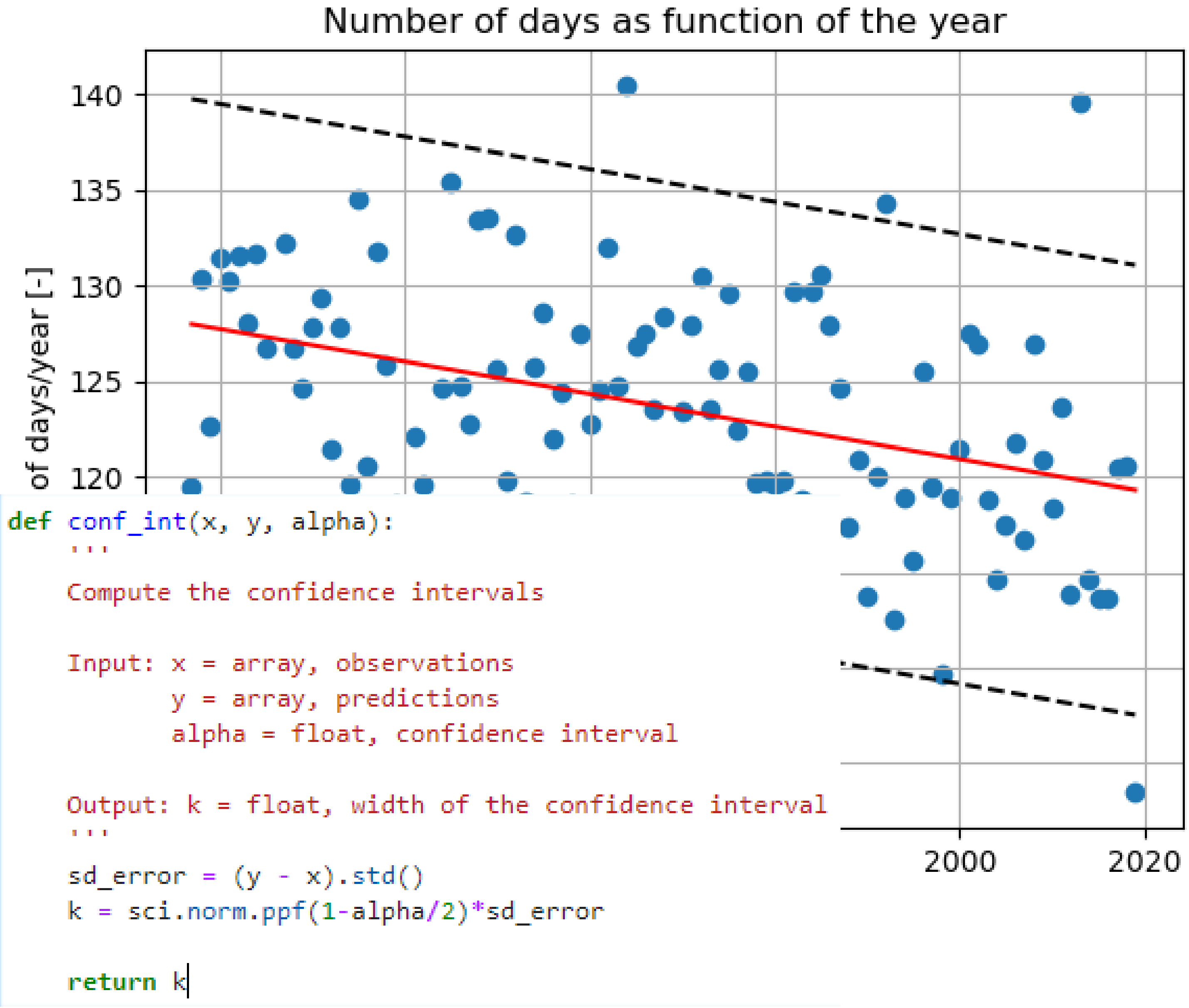
- Days, D, as a function of year, Y:

  D = q(Y)   >>>   D = m*Y + b



```
k = conf_int(data[:,1], line, 0.05)
ci_low = line - k
ci_up = line + k

#plot
plt.scatter(data[:,0], data[:,1], label = 'Observations')
plt.plot(data[:,0], line, color='r', label='Fitted line')
plt.plot(data[:,0], ci_low, '--k')
plt.plot(data[:,0], ci_up, '--k')
plt.ylabel('Number of days/year [-]')
plt.xlabel('Year [-]')
plt.grid()
plt.legend()
plt.title('Number of days as function of the year')
```



```
def conf_int(x, y, alpha):
    '''
    Compute the confidence intervals

    Input: x = array, observations
           y = array, predictions
           alpha = float, confidence interval

    Output: k = float, width of the confidence interval
    '''

    sd_error = (y - x).std()
    k = sci.norm.ppf(1-alpha/2)*sd_error

    return k
```

What is k?

Confidence Intervals

- What did we do?
  - Assume that the uncertainty in D has the Normal distribution D~N(μ,σ)
  - Mean defined by best estimate of D
  - Standard deviation defined by the standard deviation of error (constant over D)
  - Find region D ± ΔD such that probability is 95%: >>> P(D-ΔD ≤ D ≤ D+ΔD) = 0.95
- k is the distance that defines this region

Uncertainty type: Error

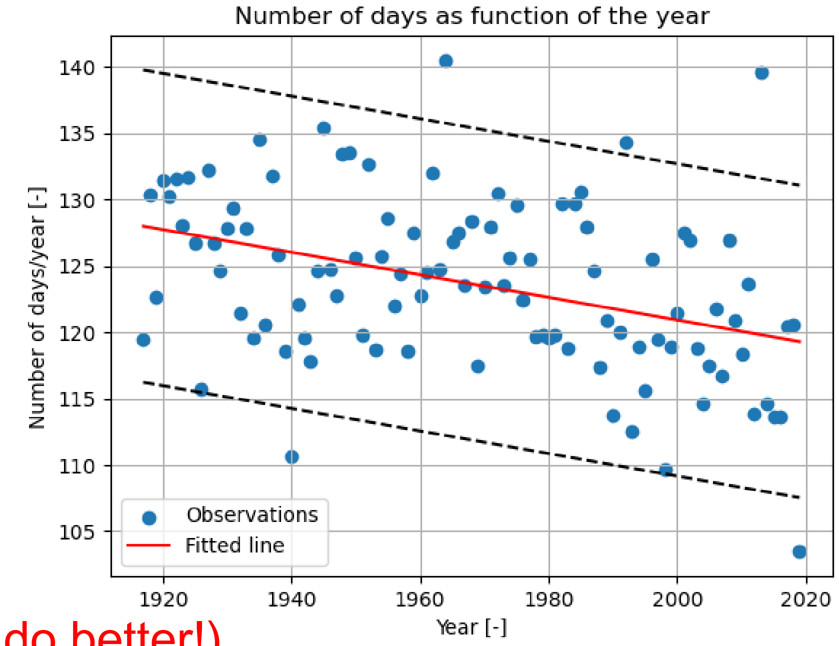
- What does this Normal distribution of D mean? (the interpretation part)
  - Uncertainty in the “true” value of D
  - Uncertainty in our model of D = q(Y)
  - High variance of past observations, D_i

Answer: mostly Error, but implicitly “most of the above”

Model has 2 parameters (m, b)

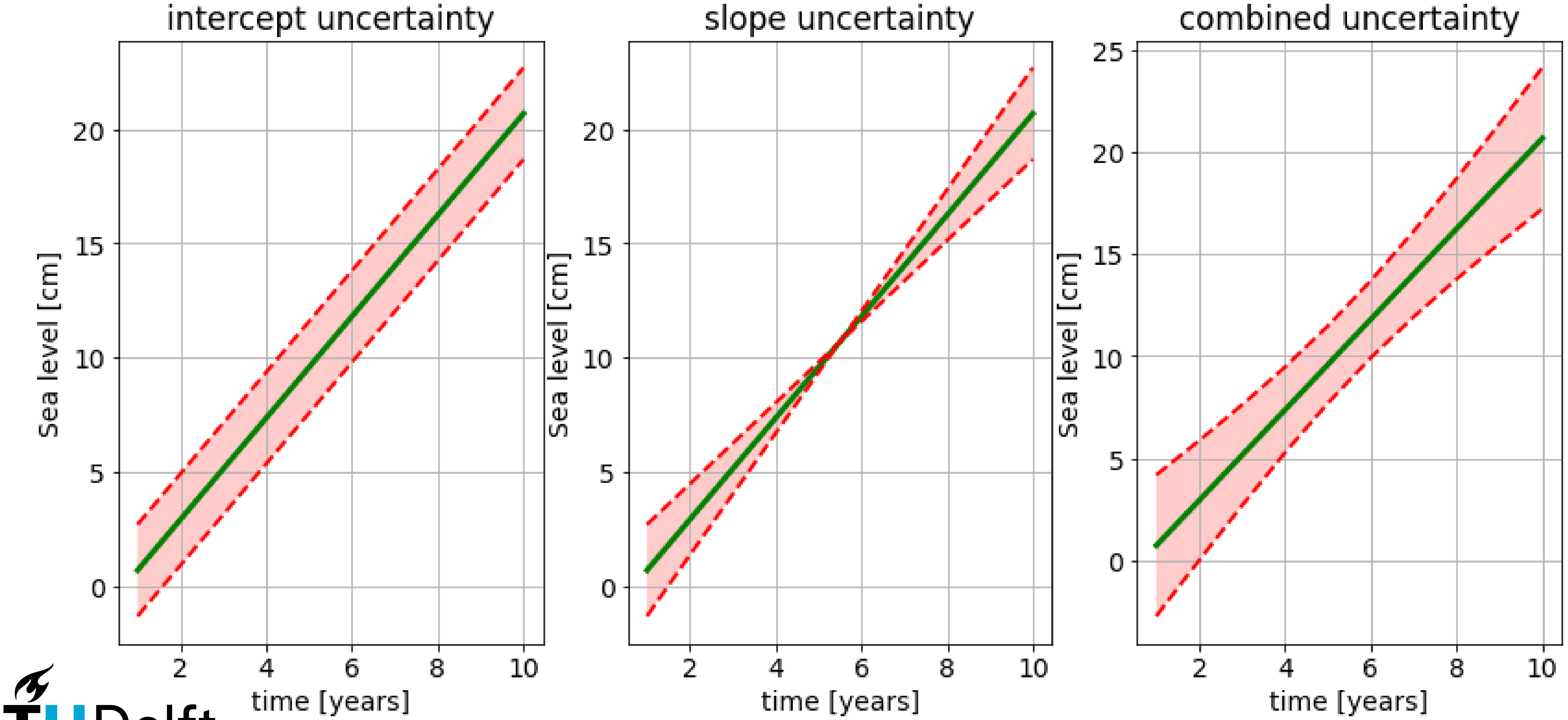
Unclear what contributes to σ_D

Variance not uniform in Y

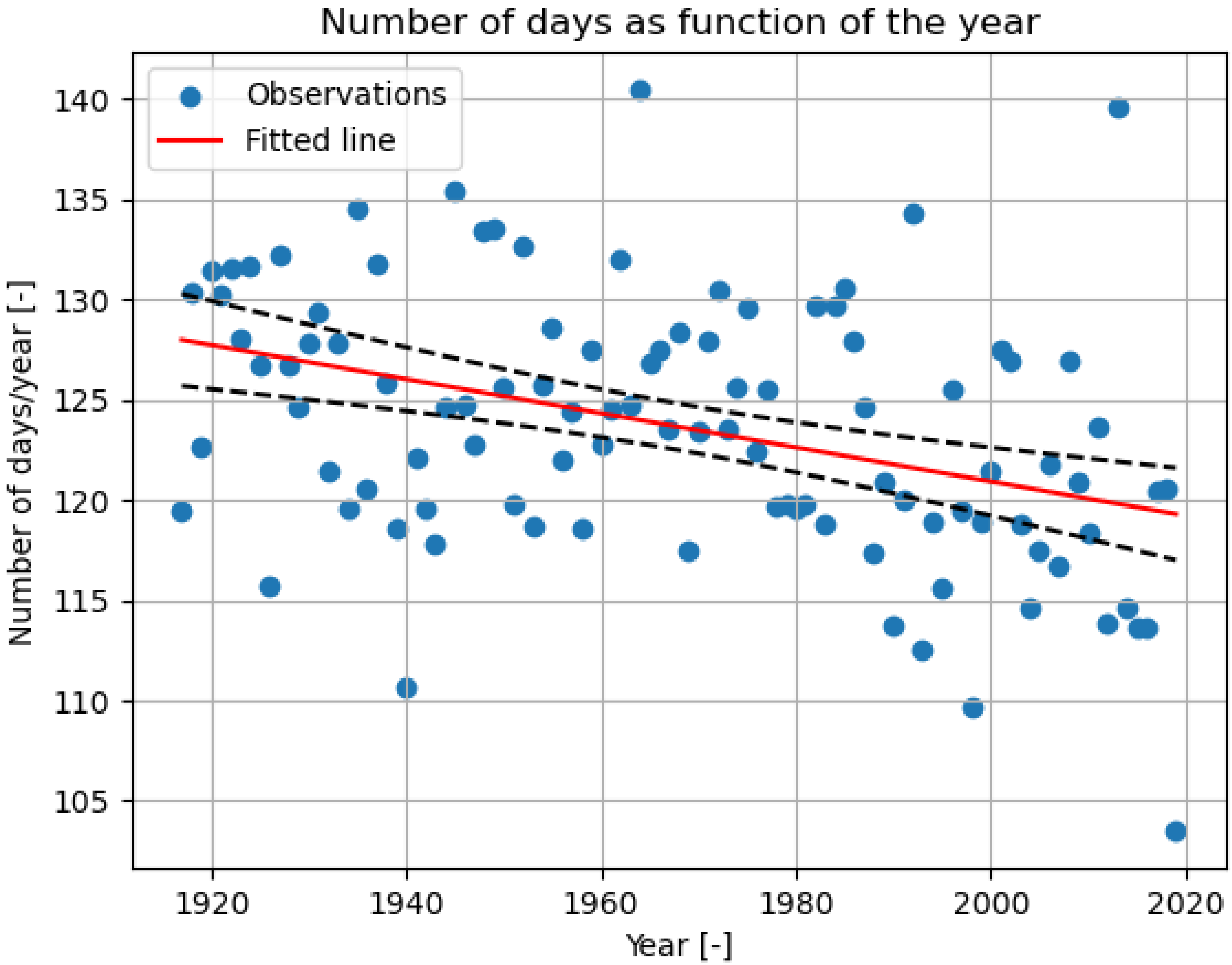


This approach = a simplification (not incorrect…but we can do better!)

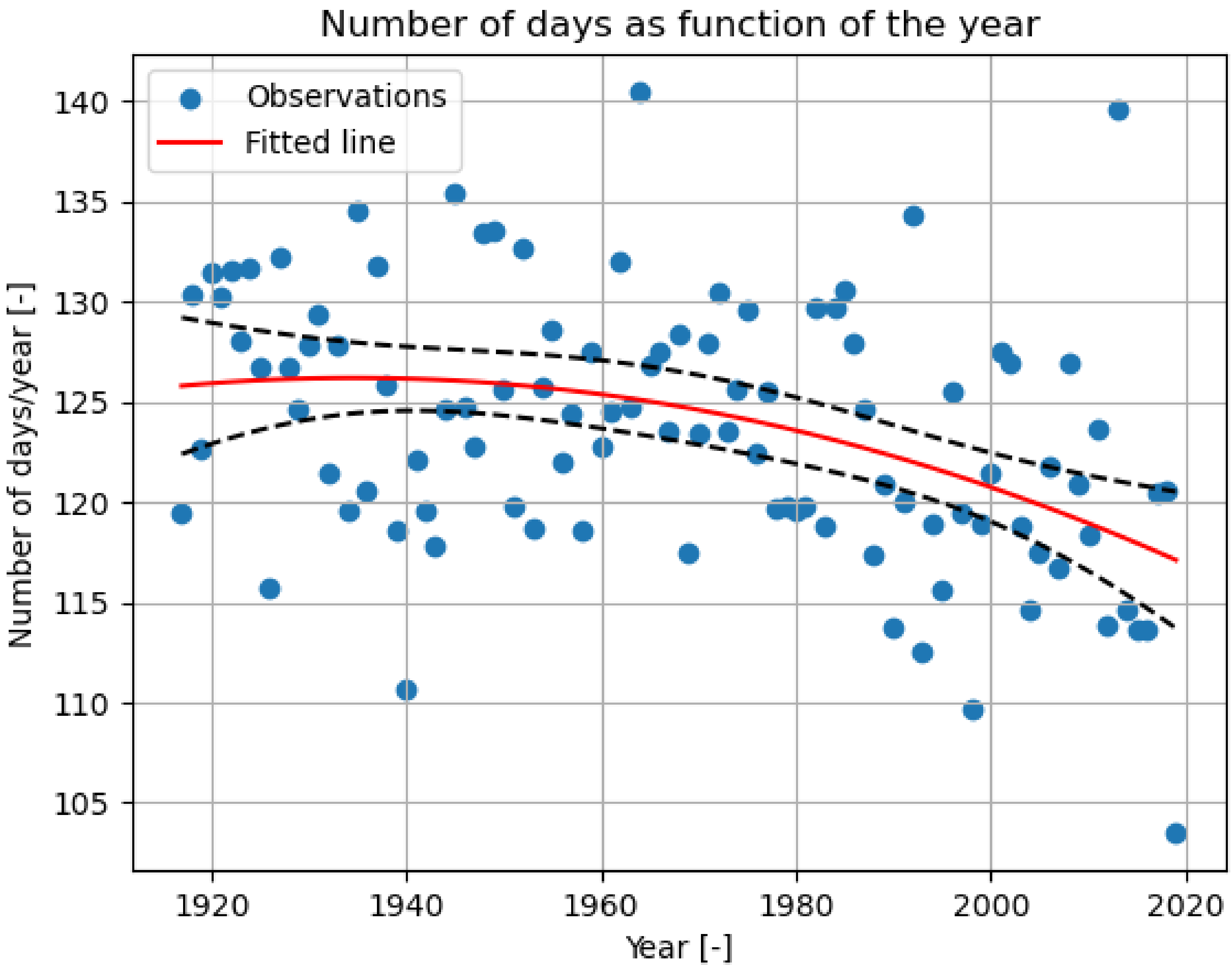
Confidence Intervals – What we will do next



Confidence Intervals – What we will do next



Confidence Intervals – What we will do next



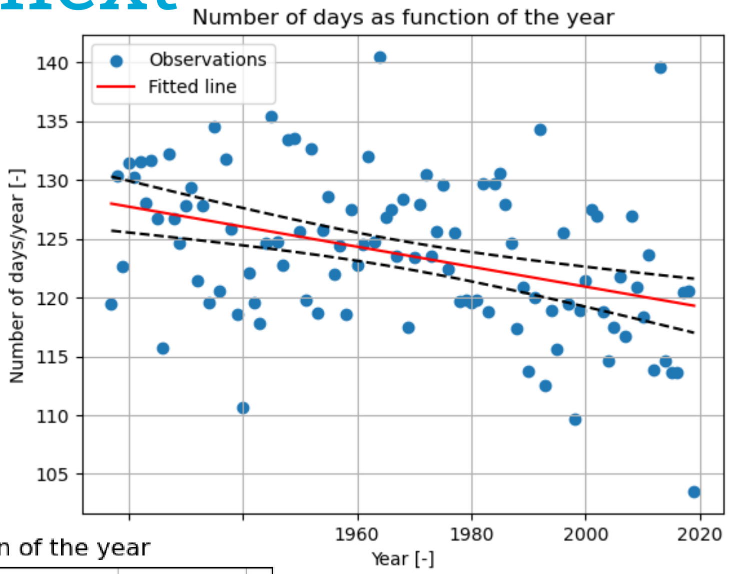
Confidence Intervals – What we will do next

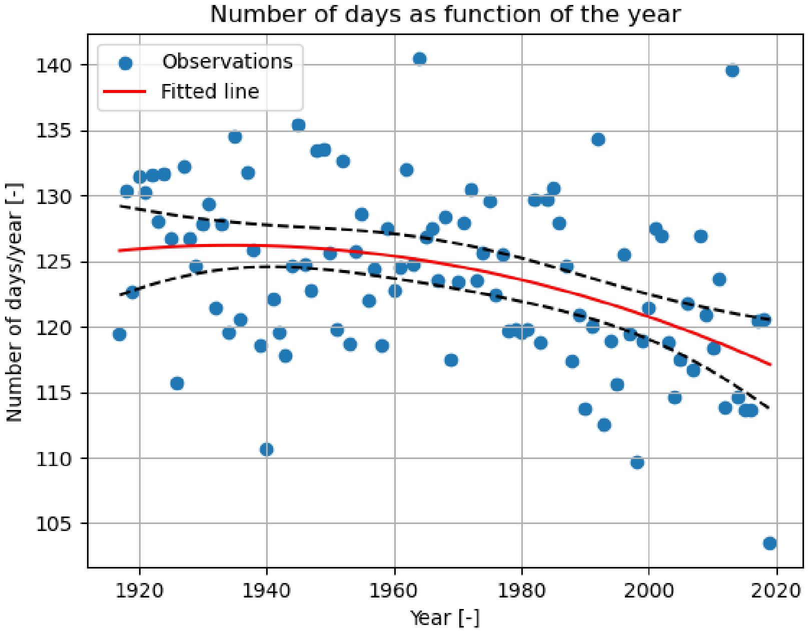
Propagation of Uncertainty

- Considers imprecise measurements (μ, σ for each point)
- Estimate μ, σ for output, D, as a function of inputs, Y

>>> epistemic uncertainty, error

In week 7, 8 we will focus more on aleatoric uncertainty





Week 2 Topic: Propagation of Uncertainty

- Chapter 2 in textbook
  - Videos & PDF Slides: supplementary to the text.
  - Interactive page, multivariate normal: try it!
- Fundamentals chapters in textbook (background knowledge / reference material)

- This week in Programming (PA 1.2): numpy, linear algebra, matplotlib, statistics; "data" handling

>>> PA 1.2 online (MUDE Files)

>>> Struggling with programming?

Week 2 Programming

Review of previous slide (Confidence)Intervals

Highlighting concepts you should know, or recognize soon!

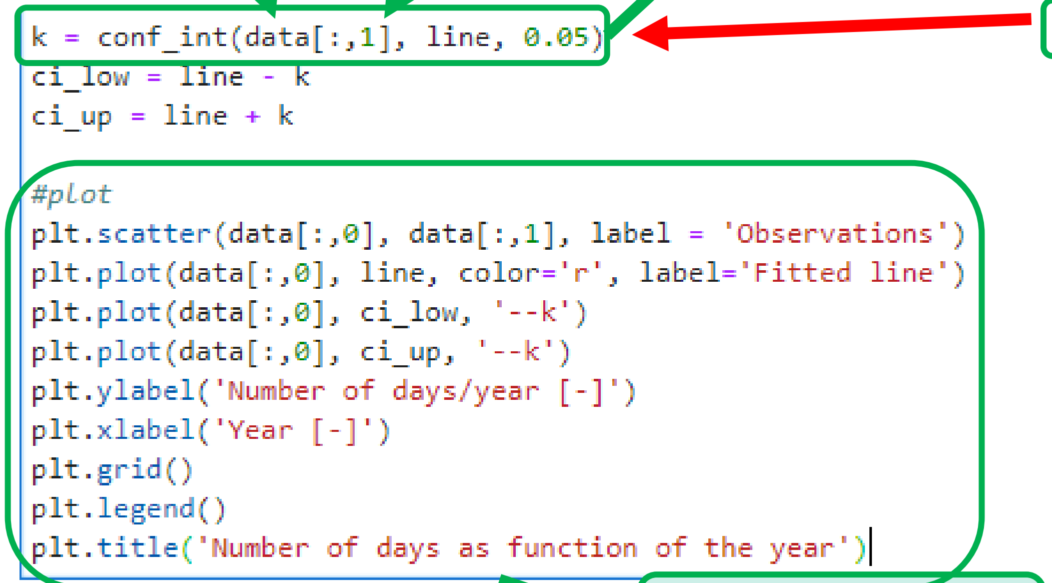
- Days, D, as function of year, Y:

>> D = m*Y + b

Numpy `ndarray`

Indexing/slicing

Function

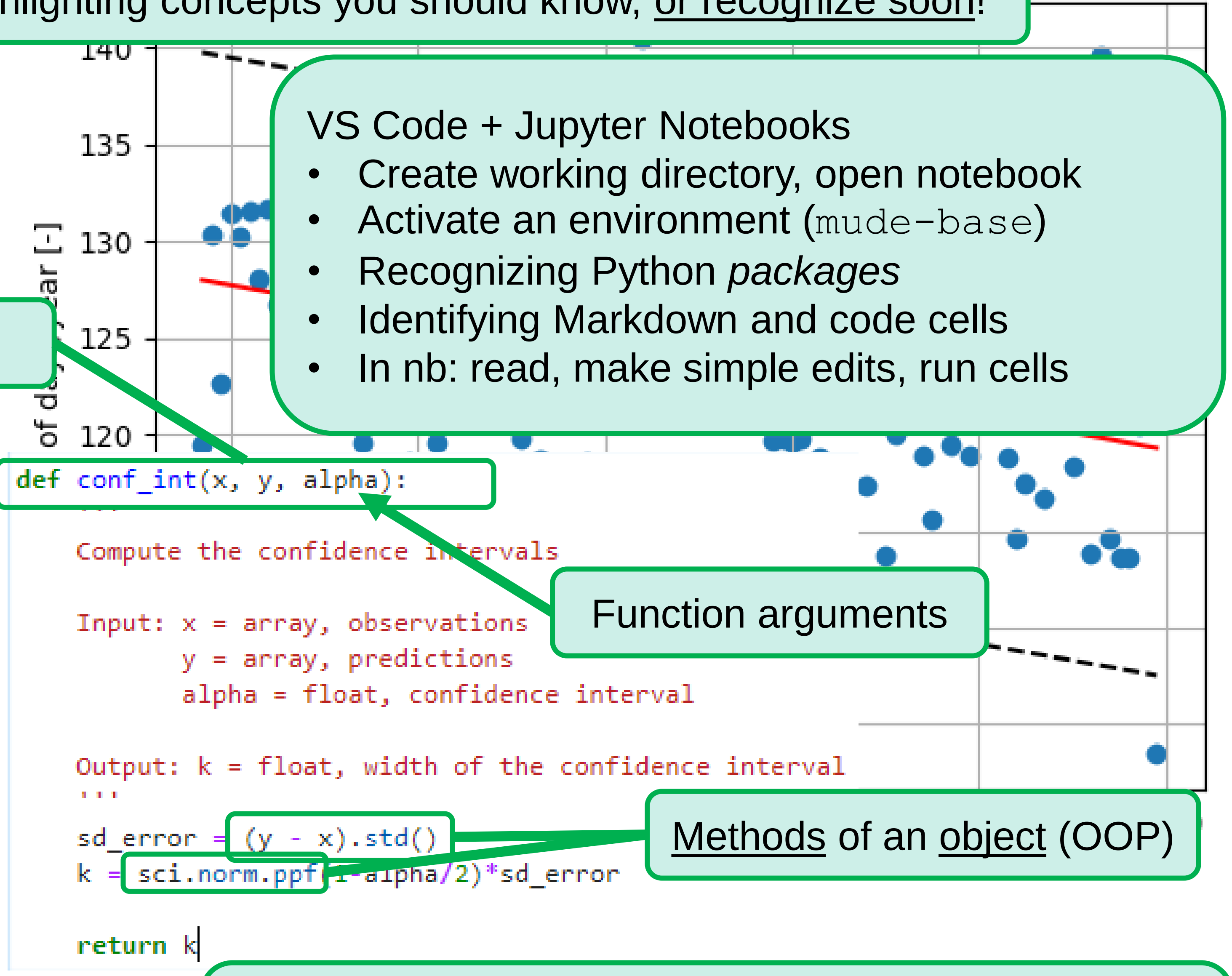


```
k = conf_int(data[:,1], line, 0.05)
ci_low = line - k
ci_up = line + k

#plot
plt.scatter(data[:,0], data[:,1], label = 'Observations')
plt.plot(data[:,0], line, color='r', label='Fitted line')
plt.plot(data[:,0], ci_low, '--k')
plt.plot(data[:,0], ci_up, '--k')
plt.ylabel('Number of days/year [-]')
plt.xlabel('Year [-]')
plt.grid()
plt.legend()
plt.title('Number of days as function of the year')
```

Matplotlib Plot

What is k?



```
def conf_int(x, y, alpha):
    '''
    Compute the confidence intervals

    Input: x = array, observations
           y = array, predictions
           alpha = float, confidence interval

    Output: k = float, width of the confidence interval
    '''

    sd_error = (y - x).std()
    k = sci.norm.ppf(1-alpha/2)*sd_error

    return k
```

VS Code + Jupyter Notebooks

- Create working directory, open notebook
- Activate an environment (`mude-base`)
- Recognizing Python *packages*
- Identifying Markdown and code cells
- In nb: read, make simple edits, run cells

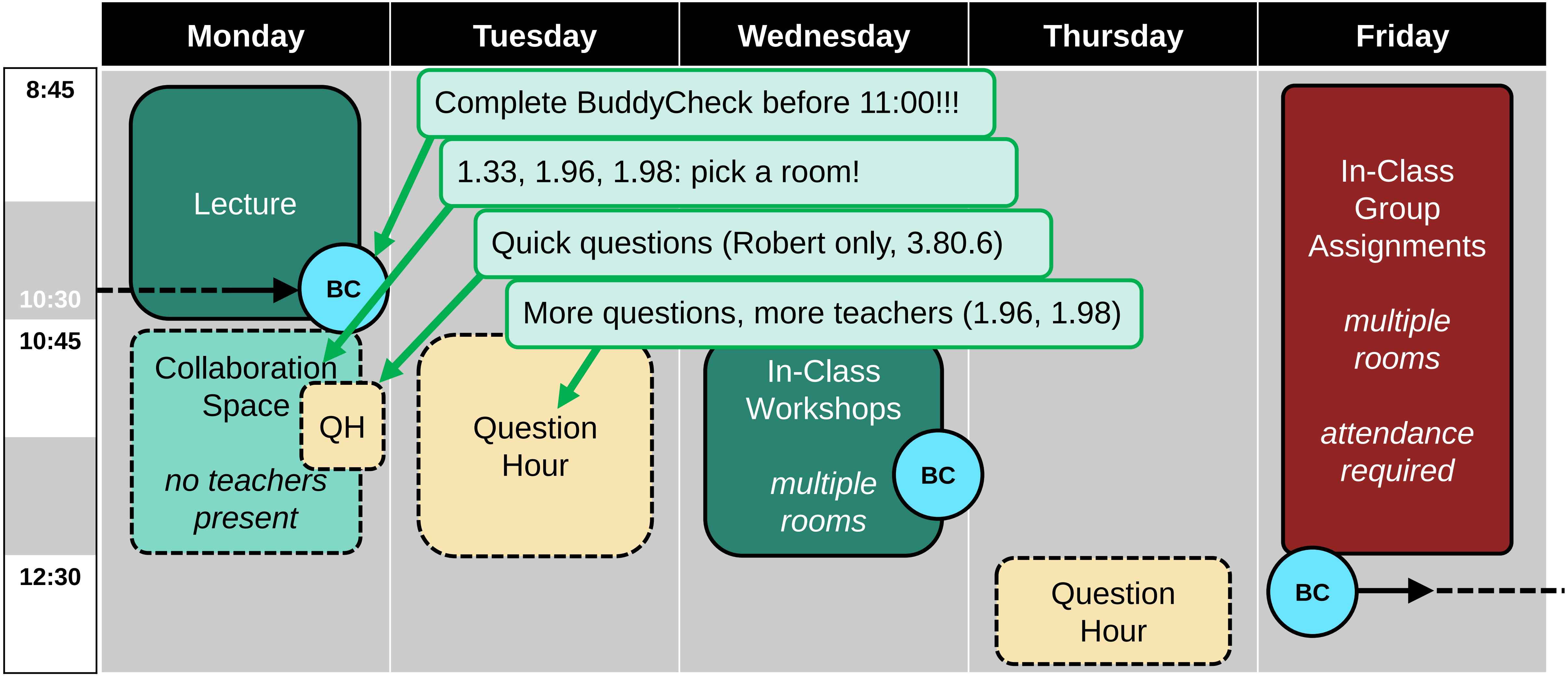
Function arguments

Methods of an object (OOP)

Lost?

- Come to office hours!
- Refer to teachbooks.github.io/learn-python

TU Delft



Monday

Tuesday

Wednesday

Thursday

Friday

8:45

10:30

10:45

12:30

Lecture

BC

Collaboration Space

QH

no teachers present

Question Hour

In-Class Workshops

multiple rooms

BC

Question Hour

In-Class Group Assignments

multiple rooms

attendance required

BC

Complete BuddyCheck before 11:00!!!

1.33, 1.96, 1.98: pick a room!

Quick questions (Robert only, 3.80.6)

More questions, more teachers (1.96, 1.98)

Programming Assignment: any time during the week, but... **Finish before Friday!**

BC = BuddyCheck: opens Fri (closes Mon); review results Wed with group

Question Hours (optional): Mon 11.00-12.00, Tue 10:45-12:30, Thu 12:30-13:30

TU Delft

MUDE

Closing

Delta Quick Reaction Student Squad ^

Apply now!

- Complete BuddyCheck (NOW!)
- Read the book (Chapter 2)
- Work on Programming Assignment (PA 1.2)

- Struggling with Programming? Don’t miss question hours!

>>> Don't worry about setting up VSC with Copilot now
(we will help you later)

>>> CLI: Note that PowerShell is different than Command Prompt