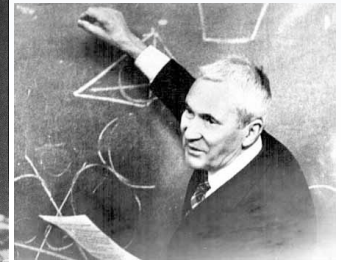
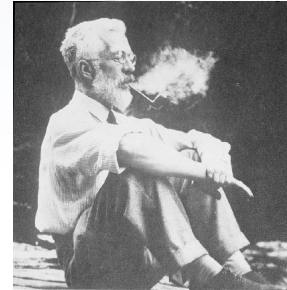
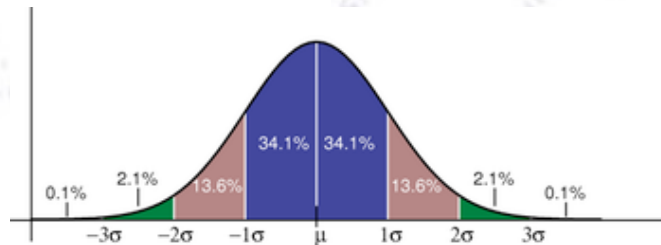


Applied Statistics

Course information 2025-26



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense!"

Applied Statistics 2025

...all the technical stuff!

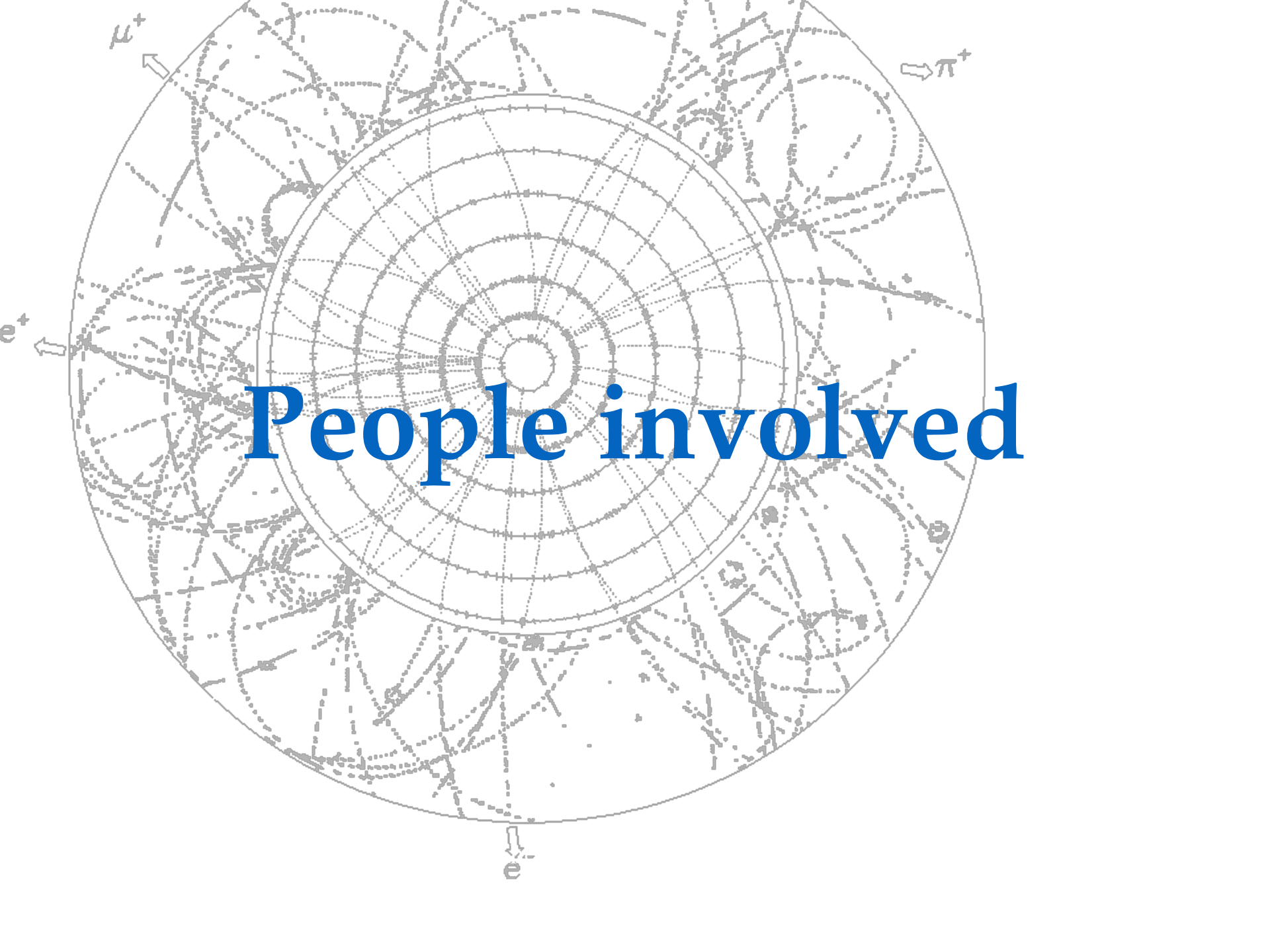
Technicals:

- Rooms and hours.
- Course structure and dates.
- Computers and software.
- Data sets.
- Literature.
- Curriculum.
- Problem set.
- Projects.
- Exam.
- Expectations.
- Goals.



The course webpage (central source of course information, bookmark or fail!):

<http://www.nbi.dk/~petersen/Teaching/AppliedStatistics2025.html>



People involved

Teachers

I have taught this course many times, and now have the honour of having **Mathias Heltberg** in it also. He has both had the course, been a TA, given some of the lectures, and used the course content in his research.

Also very importantly, we have **Nina, Janni, Gabriela, Preet, and Marc** with us as TAs.



Nina Nathanson



Janni Nikolaides



Gabriela Oliveira

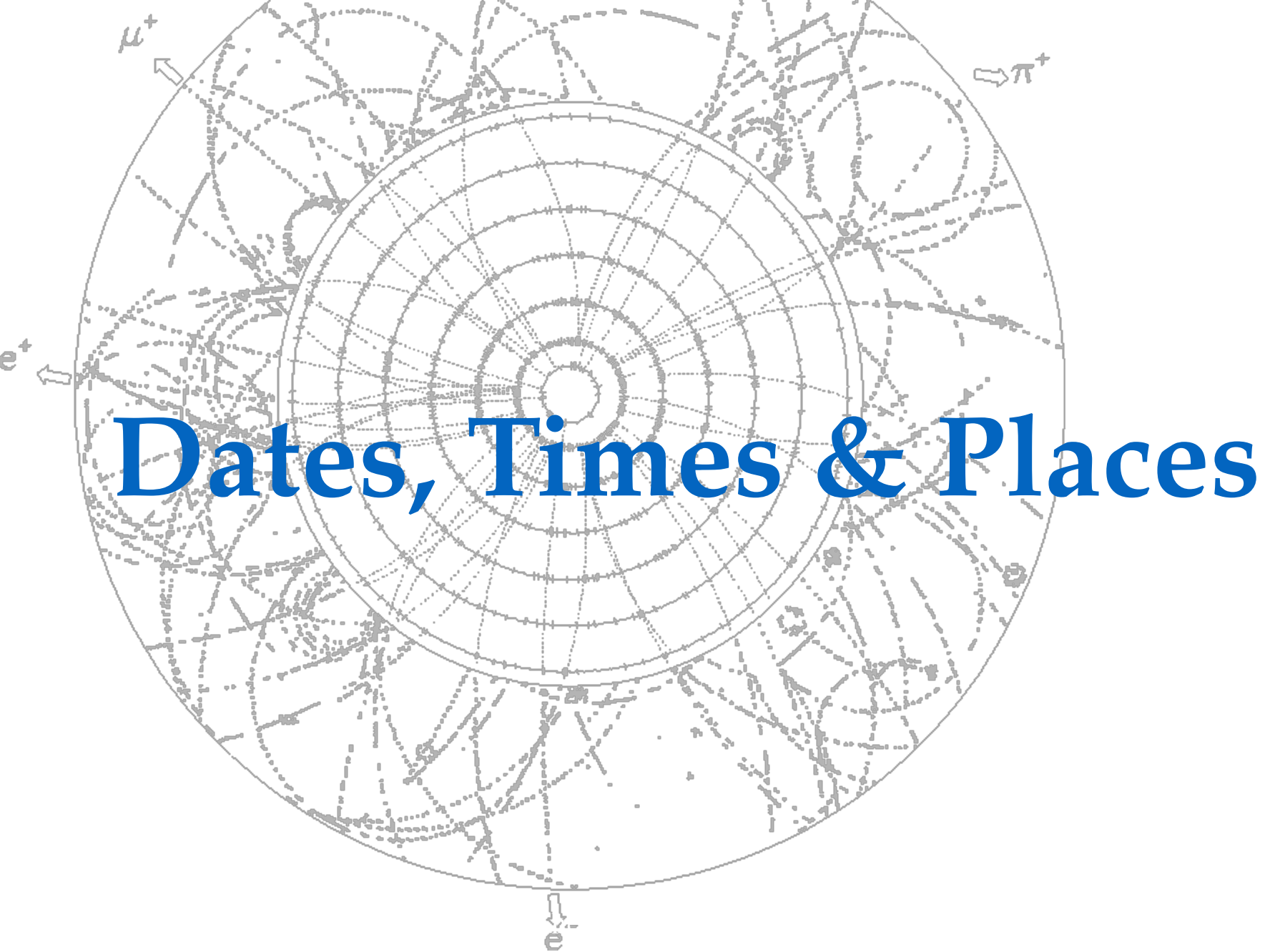


Preet Pati



Marc Jacquart

We all look forward to meeting all of you.

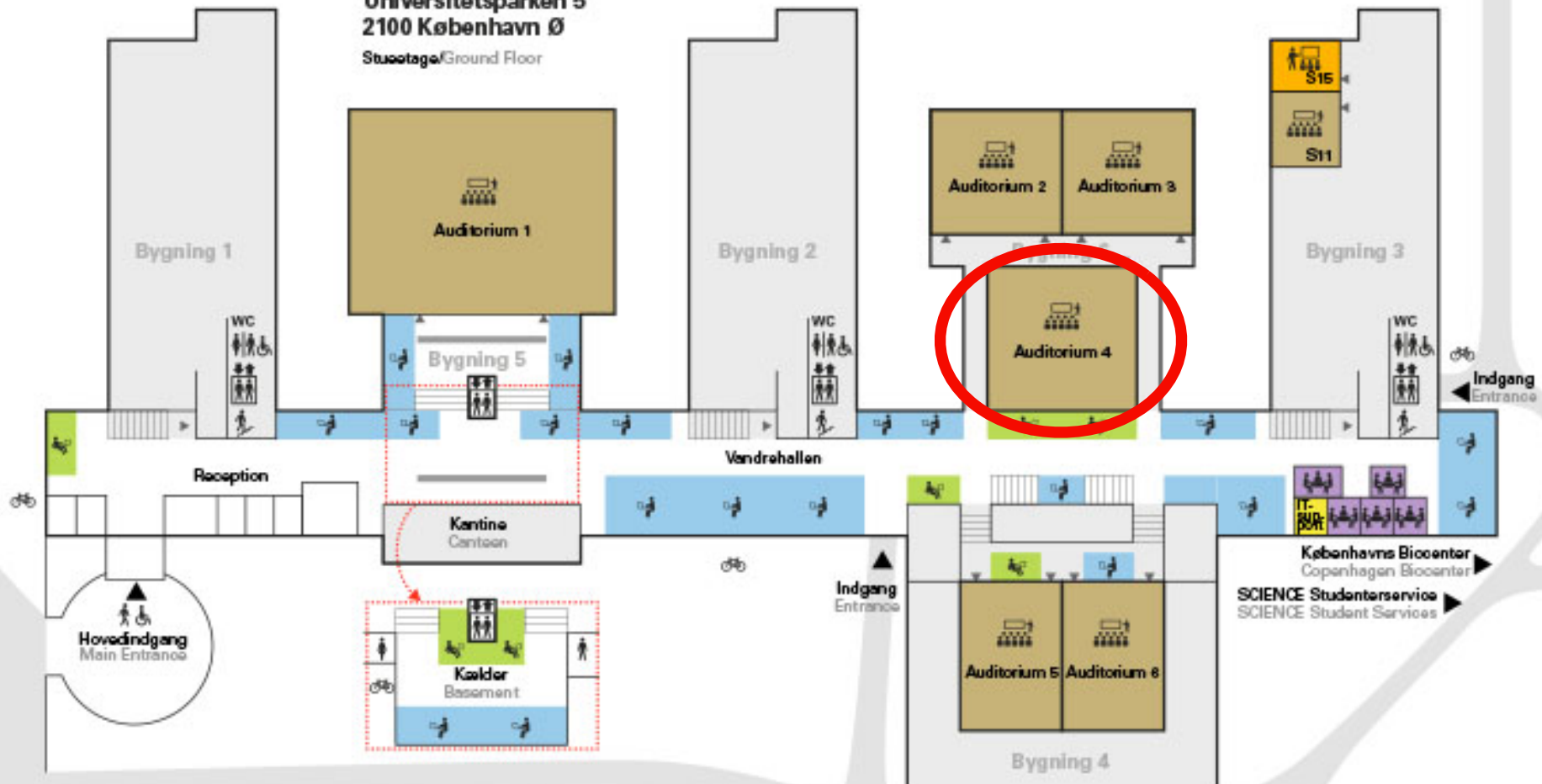


Dates, Times & Places

Lectures at HCØ

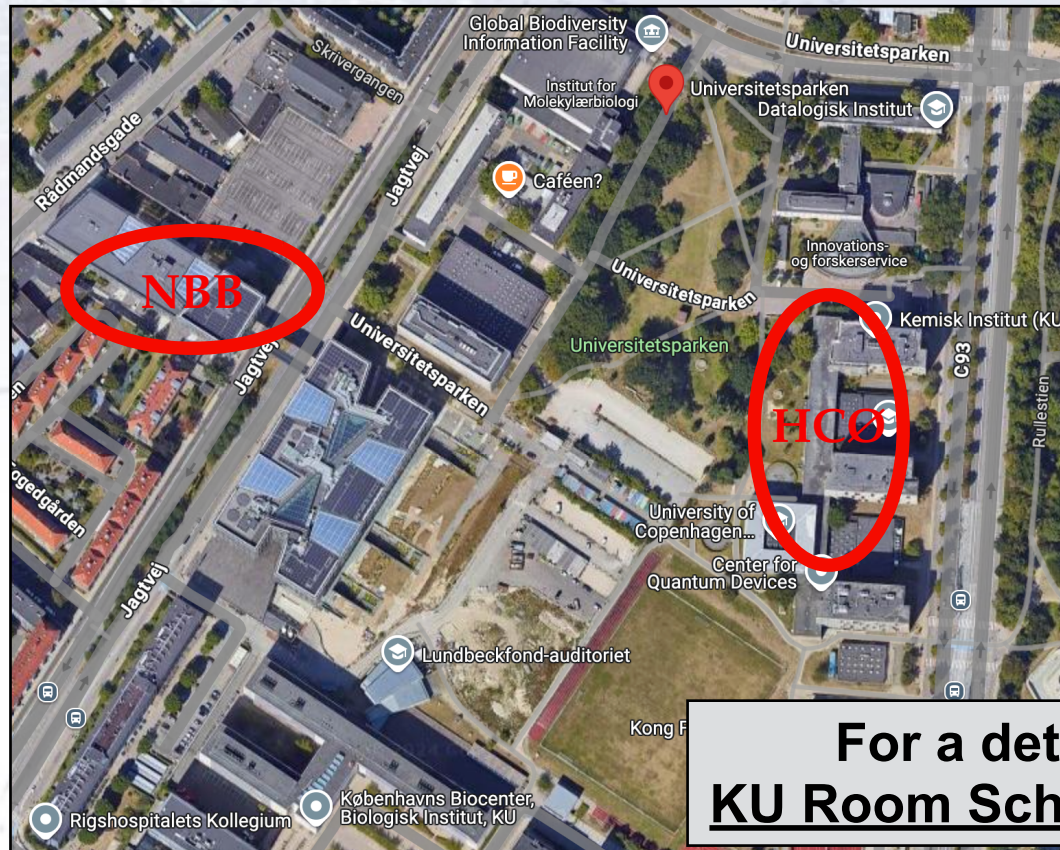
Lectures: Aud. 4

H.C. Ørsted Institut
Universitetsparken 5
2100 København Ø
Stueetage/Ground Floor



Exercises

Exercises will take place at HCØ, BioCenter, and DIKU.



**For a detailed view:
[KU Room Schedule Webpage](#)**

Mondays: NBB (2.3.H.142, 2.2.I.158, 2.2.H.142, 2.1.I.156, 1.2.B.015)

Tuesdays: HCØ (C103, A112, A111, Aud 07, Aud 04)

Fridays: NBB (2.3.H.142, 2.0.G.064/070, 1.01.F.70)

Note: This course does not use “hold” - you may do your exercises in any room you want!

Additional locations

Jagtvej

Mathias' Office (2nd floor)

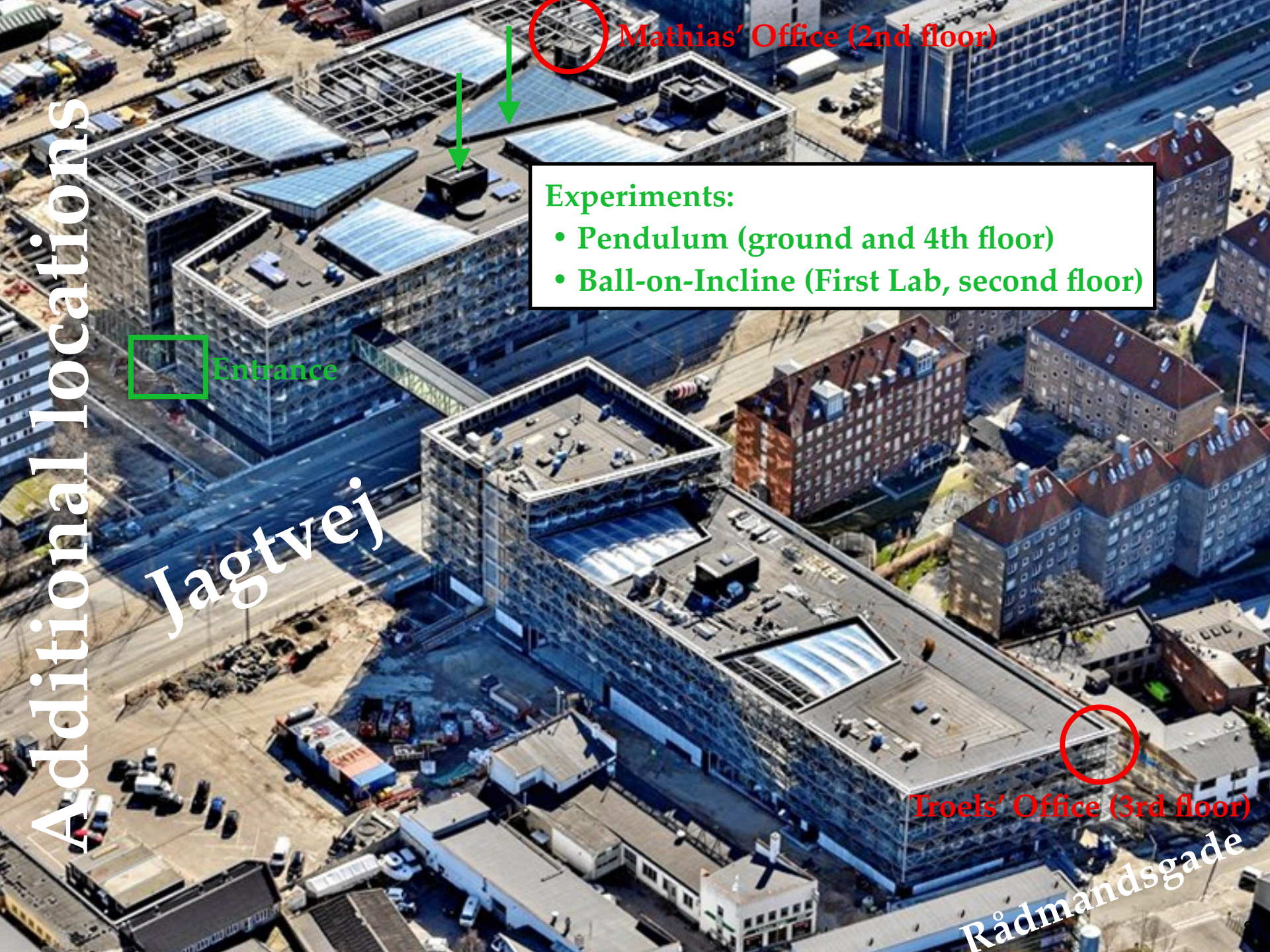
Experiments:

- Pendulum (ground and 4th floor)
- Ball-on-Incline (First Lab, second floor)

Entrance

Troels' Office (3rd floor)

Rådmandsgade



Additional locations

Blegdamsvej

Entrance to Auditorium A

For pre-course python help, and
measurement of lecture table.

Course dates & hours

Dates:

Block 2 (schedule B) will in 2025-26 consist of the following weeks:

Week 1: 17.-21. November

Week 2: 24.-28 November

Week 3: 1. - 5. December

Week 4: 8.-12. December

Week 5: 15.-19. December

No teaching: 22. December

Week 6: 2. January (on ML!)

Week 7: 5.-9. January

Week 8: 12.-13. January

Exam: 15.-16. January

Hours:

Following schedule B, but after the first 2.5 weeks, we will be using the morning hours 8:15 - 9:00 Monday and Friday for “self-studying”.

Monday:

8:15 - 10:00 Lectures

10:15 - 12:00 Exercises

Tuesday:

13:15 - 14:00 Lectures

14:15 - 17:00 Exercises

Friday:

8:15 - 10:00 Lectures

10:15 - 12:00 Exercises

Course dates & hours

Dates:

Block 2 (schedule B) will in 2025-26 consist of the following weeks:

Week 1: 17.-21. November

Week 2: 24.-28. November

Week 3: 1.-5. December

Week 4: 8.-12. December

Week 5: 15.-19. December

No teaching in December

Week 6: 2. January (On MLT.)

Week 7: 5.-9. January

Week 8: 12.-13. January

Exam: 15.-16. January

Hours:

Following schedule B, but after the first 2.5 weeks, we will be using the morning hours 8:15 - 9:00 Monday and Friday for “self-studying”.

Just to be clear:

**Even if the course can be followed online,
this is highly discouraged**

Tuesday:

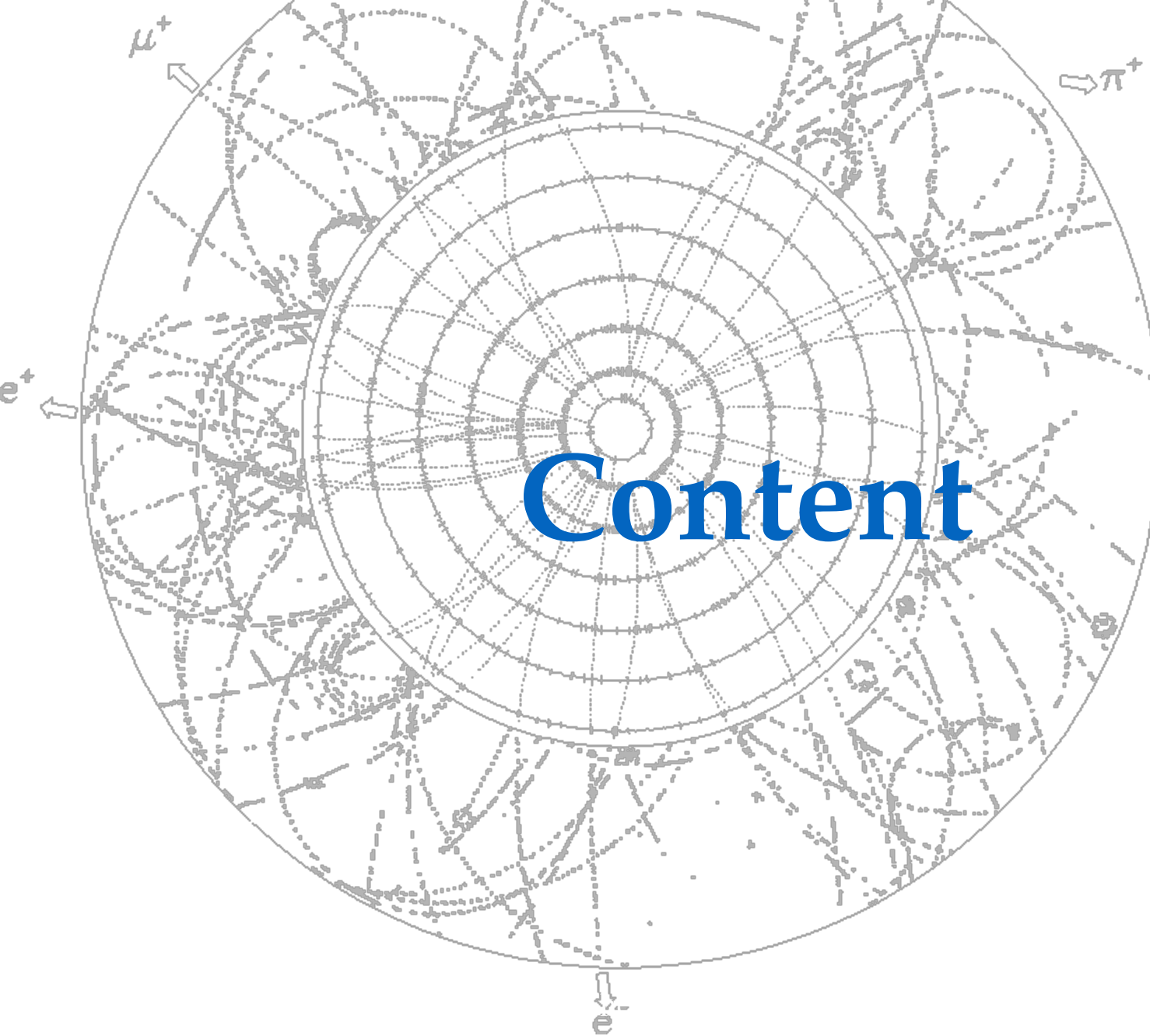
13:15 - 14:00 Lectures

14:15 - 17:00 Exercises

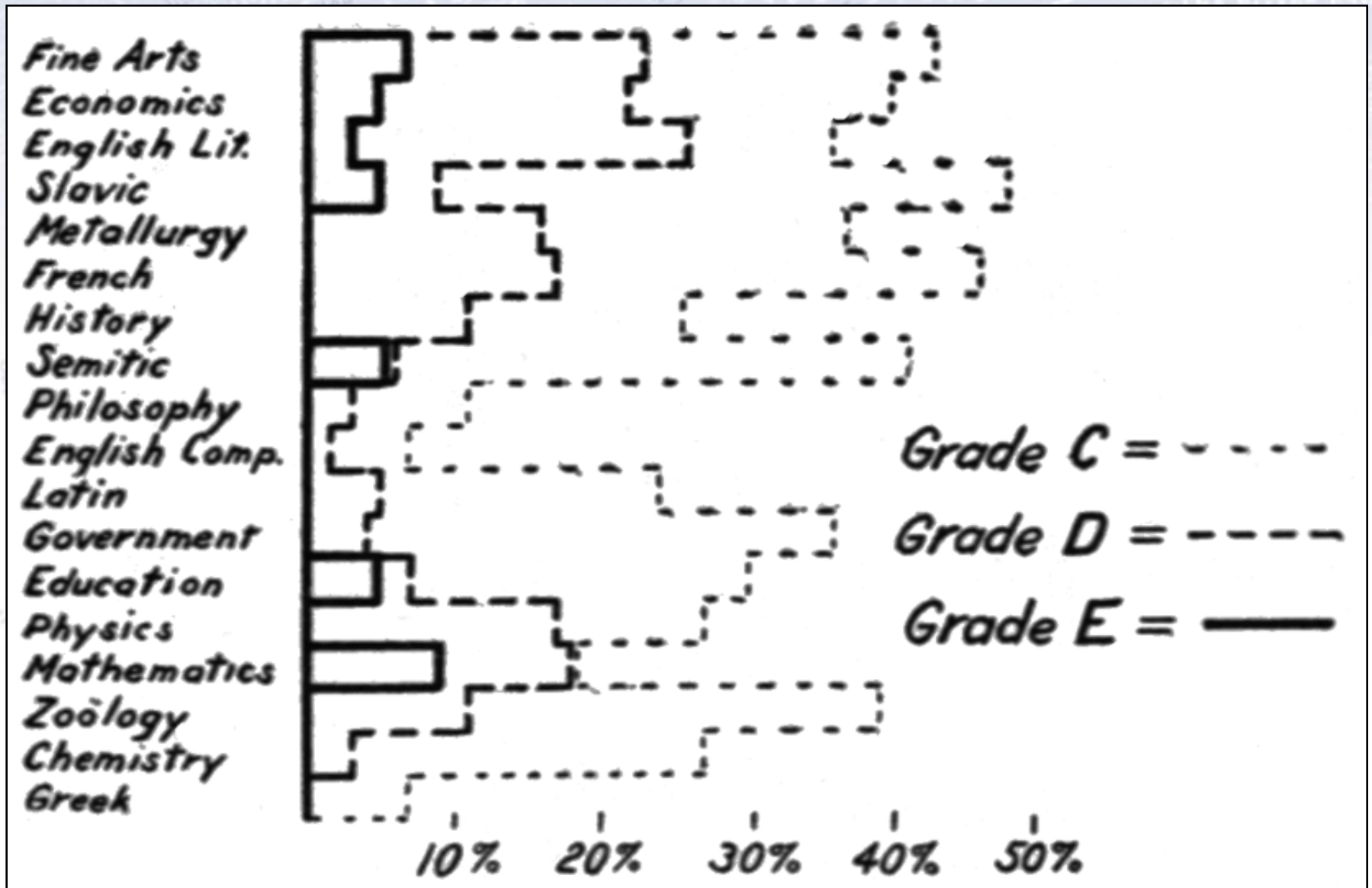
Friday:

8:15 - 10:00 Lectures

10:15 - 12:00 Exercises

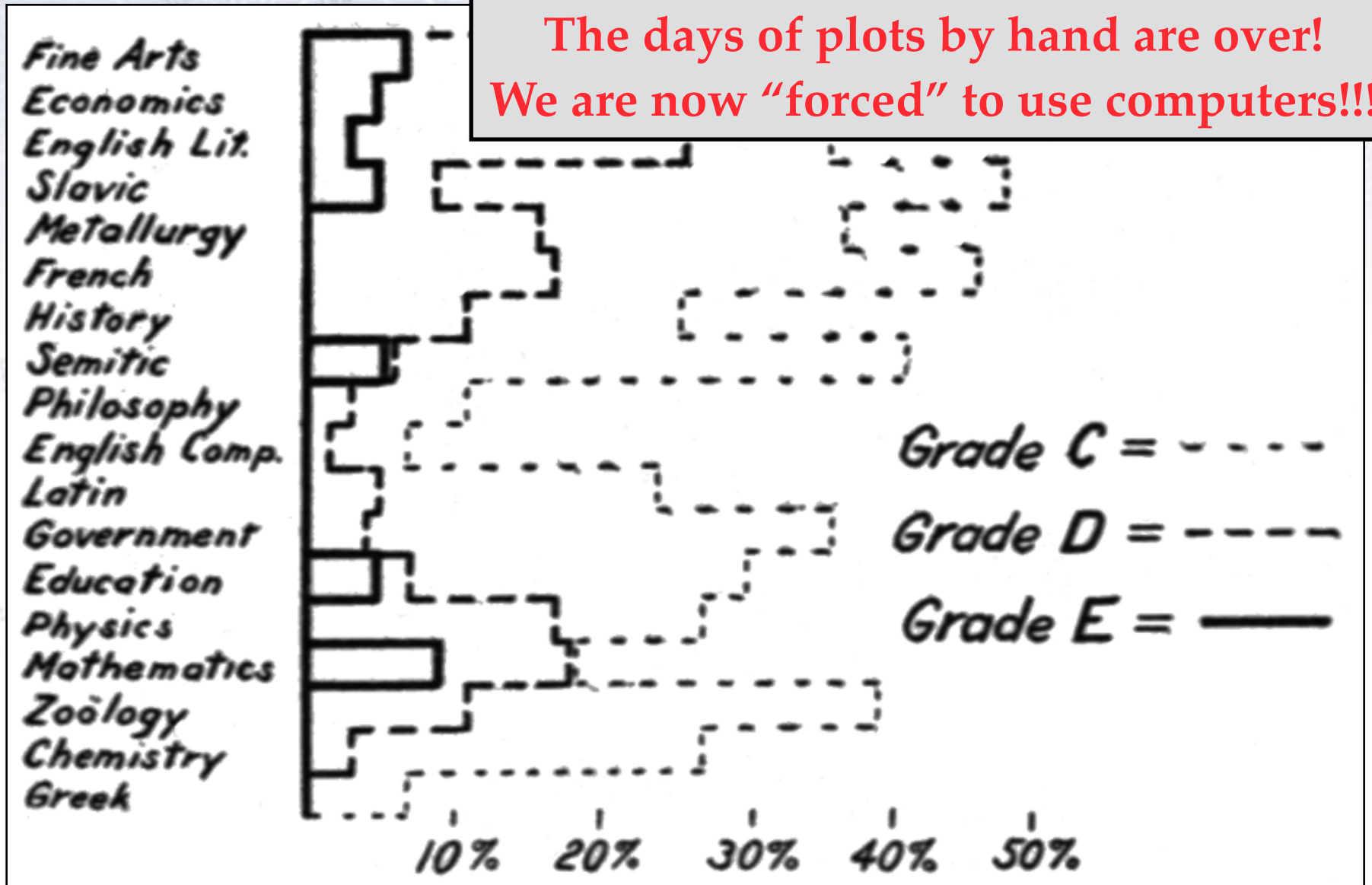


Computers and software



Computers and software

The days of plots by hand are over!
We are now “forced” to use computers!!!



Computers and software

The times are *way past* pencil and/or calculator stage!

Fast computers is the *only* answer to do (any serious) data analysis.

Operating system:	Linux/MAC OS/Windows
Programming:	Python - version 3.11+
Editor:	Jupyter Notebook (or own favorit!)

Python Packages used:

NumPy, Matplotlib, Pandas, iMinuit, SciPy, and maybe others.

Only iMinuit should possibly be “unknown” to many, but it is easy to install, and essential for fitting (SciPy’s “optimize” is the alternative).

Code repository used:

All code can be found on GitHub (webpage links there):

<https://github.com/AppliedStatisticsNBI/AppStat2025/>

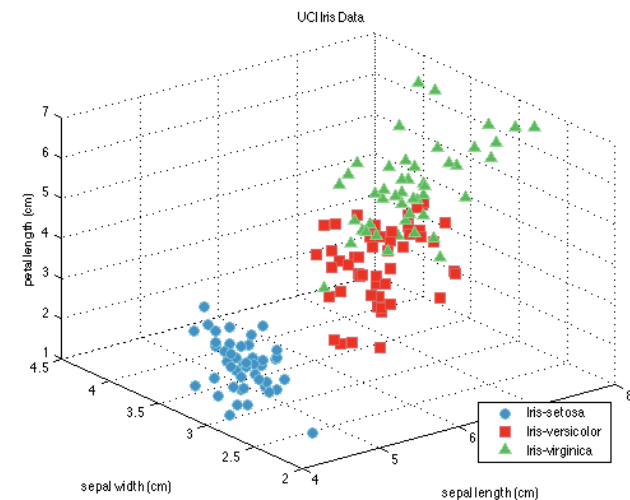
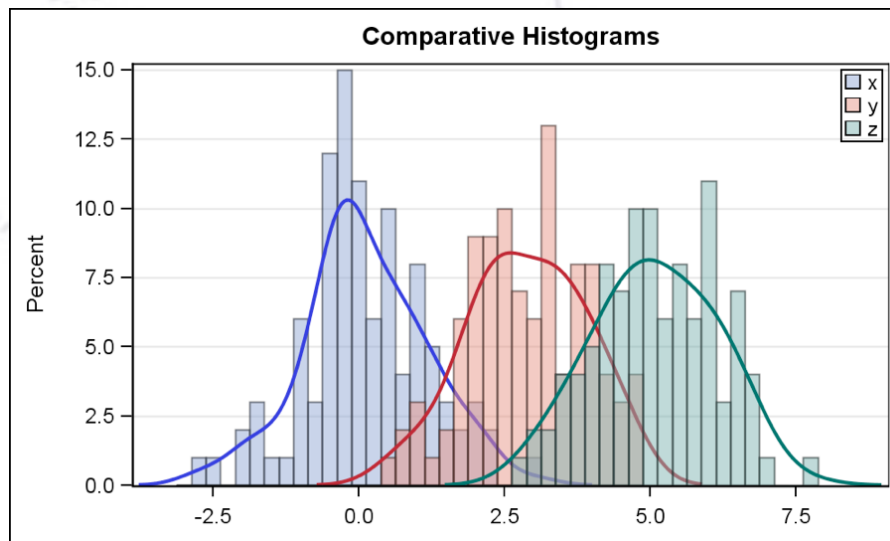
Note: You’re not “forced” to use Python, but we will only supply code in Python.

Data sets

In general, any data set can be used for this course! If you happen to have an interesting and illustrative one, bring it to me/class!

I've tried my best to search for a large variety of data sets, but this is not always easy. Publicly available data sets are often old / small / biased / etc.

As a result, one or two data sets are from my own field (particle physics). This is both due to my access to data here, but also because particle physics is one of the fields providing *billions of measurements*.



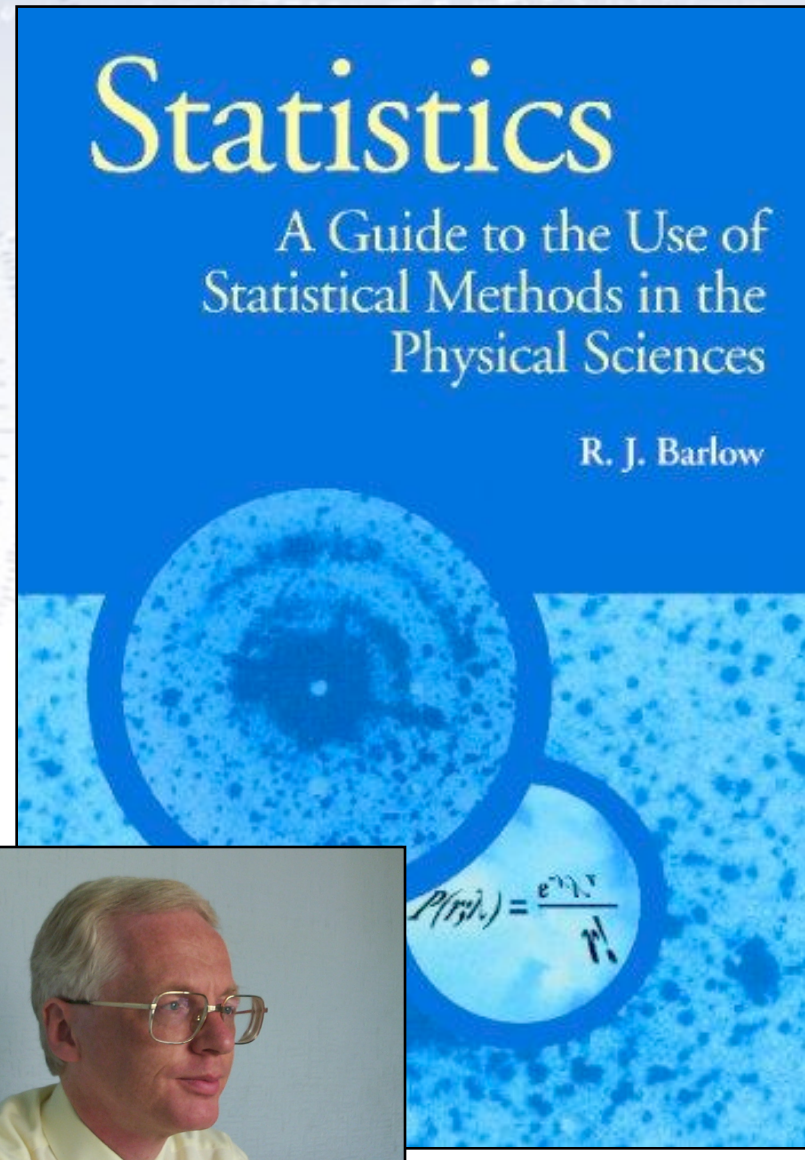
Literature

We use **Roger J. Barlow's "Statistics"**, as it is an accessible introduction to statistics with many examples, and the best overall book (I think).

If anything, it is lacking a bit on how to generate random numbers according to a specific PDF and on categorising events.

NOTE:

In addition to two other books (see next page), there is a great abundance of notes (e.g. from Particle Data Group), Wiki, fora, etc. on both statistics but especially also Python on the web, which I encourage you to use (with a proper critical mind).



Additional literature

Two additional great books are:

- P. R. Bevington: Data Reduction and Error Analysis
- Glen Cowan: Statistical Data Analysis

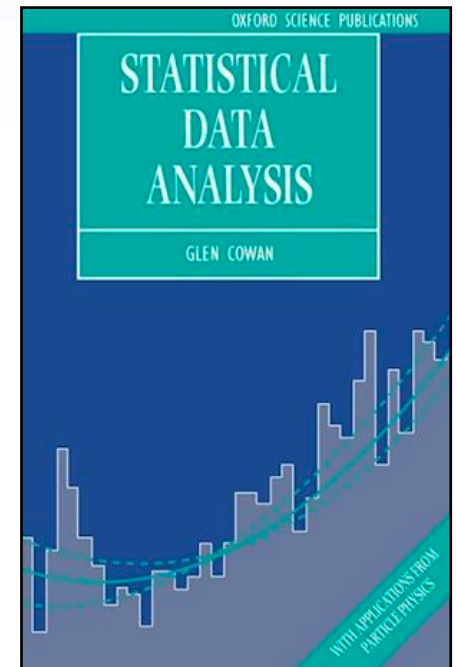
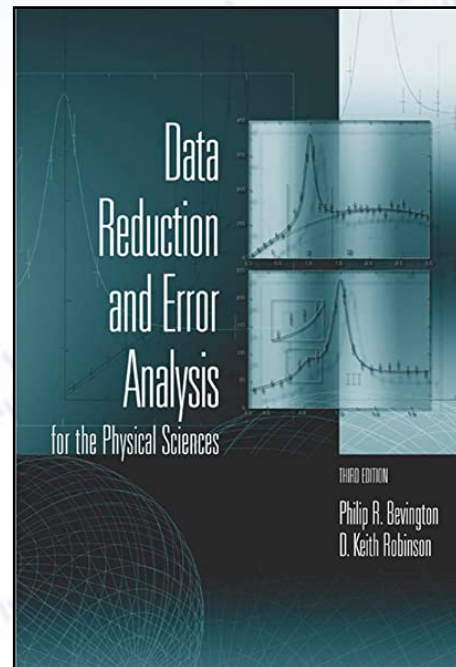
Bevington is a classic and very good basic introduction. If you don't understand something, try re-reading about it in Bevington.

Cowan is more “modern”, and for the slightly more advanced reader.

Great sections are:

- Producing random numbers
- Hypothesis testing

Links to electron versions of both books can be found on the course webpage.



Curriculum

The course will cover the following chapters in R. Barlow:

- Chapter 1 (All)
- Chapter 2 (All)
Exercises: All, except 2.5 and 2.9.
- Chapter 3 (Except 3.2.2, 3.3.2, 3.4.2, 3.5.2)
Exercises: All, except 3.7.
- Chapter 4 (All)
Exercises: All, except 4.10.
- Chapter 5 (Except 5.1.3, 5.3.2, 5.3.3 (formal part), 5.3.4, 5.5)
Exercises: 5.2
- Chapter 6 (Except 6.4.1, 6.7)
Exercises: All
- Chapter 7 (Except 7.3.1)
Exercises: All, except 7.1, 7.3, and 7.7.
- Chapter 8 (Except 8.4.4, 8.4.5, 8.5.1, and 8.5.2)
Exercises: All, except 8.6.
- Chapter 10 (All)

Core of Curriculum

The course will **focus mostly on** the following chapters in R. Barlow:

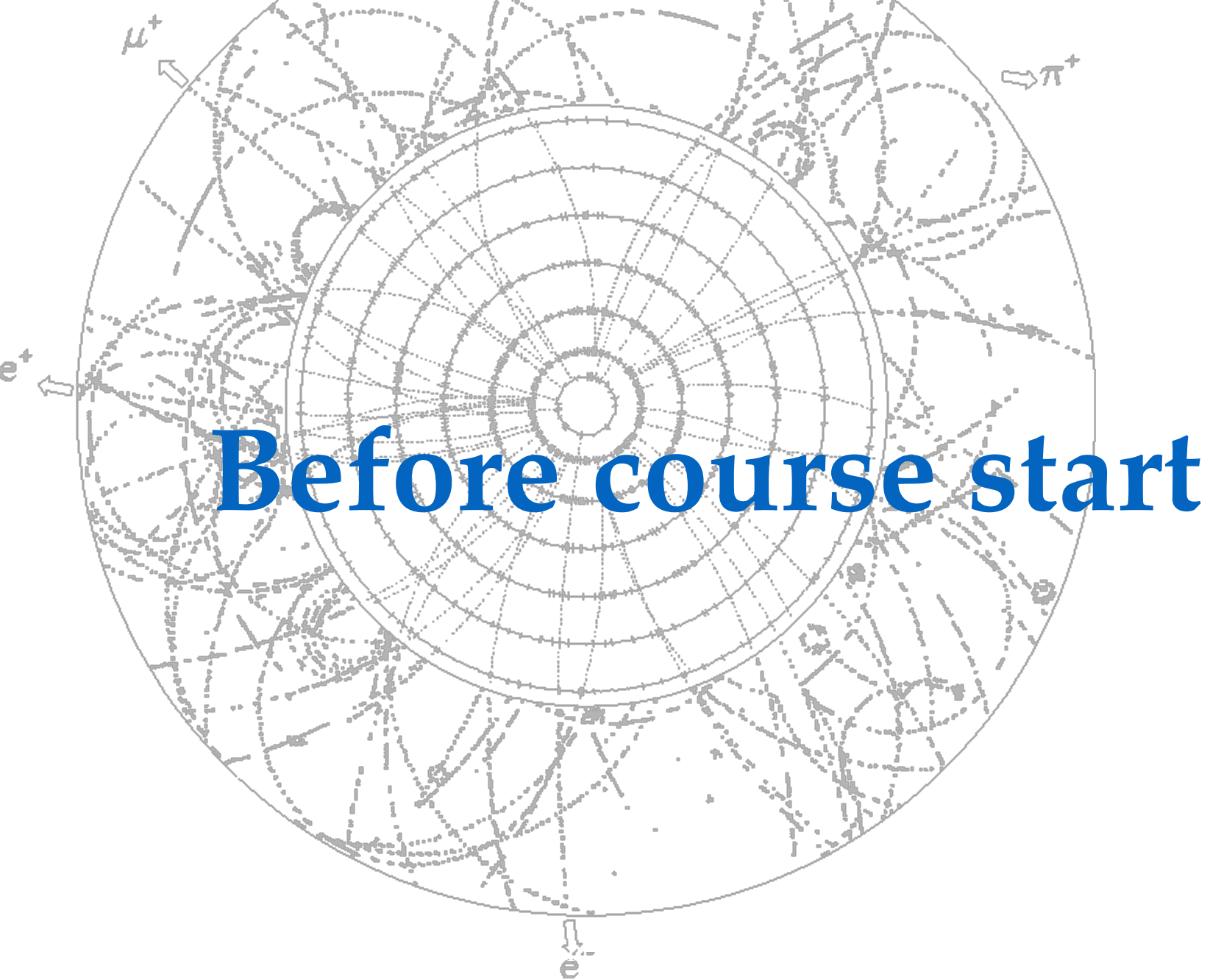
- Chapter 2: 2.1, 2.2, 2.3, 2.4.1, 2.4.2, 2.6
- Chapter 3: 3.1, 3.2, 3.2.1, 3.3, 3.3.1, 3.4.1, 3.4.7, 3.5.1
- Chapter 4: 4.1, 4.2, 4.3, 4.3.1, 4.3.2, 4.3.3
- Chapter 5: 5.1, 5.1.1, 5.1.2, 5.2, 5.6
- Chapter 6; 6.1, 6.2, 6.2.1, 6.2.2, 6.2.3, 6.2.4, 6.3, 6.4
- Chapter 8: 8.1, 8.2, 8.3, 8.4, 8.4.1, 8.4.2, 8.4.3

This is less than 80 pages, but... they do not only require reading!

They request understanding!!!

The plan is to go through most of curriculum in 4-5 weeks, spending the rest of the time on applying it.

It is through application that statistics is really understood.



Check list

In order for me to consider you inscribed in this course, you should make sure that you pass the following check list:

- **Have read the course information** (slides on course webpage).
Otherwise, you don't know what is going to happen.
- **Have filled in the questionnaire** (on course webpage).
Otherwise, we don't know what you know and don't know.
- **Have measured the length of the lecture table in Auditorium A***.
Otherwise, you haven't contributed to a common course dataset.
(This can also be done during the lectures of the project days).
- **Be registered on Absalon or accept invitation by me to be so**.
Otherwise, you won't get any of the general information I write out.
- **Be able to run Python on your own laptop and(/or) on ERDA**.
Otherwise, you can't follow the exercises or solve problems.

* NOTE: One should follow the instructions given in two slides!

NBI AUDITORIUM A:
ORIGIN OF QUANTUM MECHANICS
...AND WHERE MOST NOBLE PRIZE WINNERS IN PHYSICS HAVE BEEN.



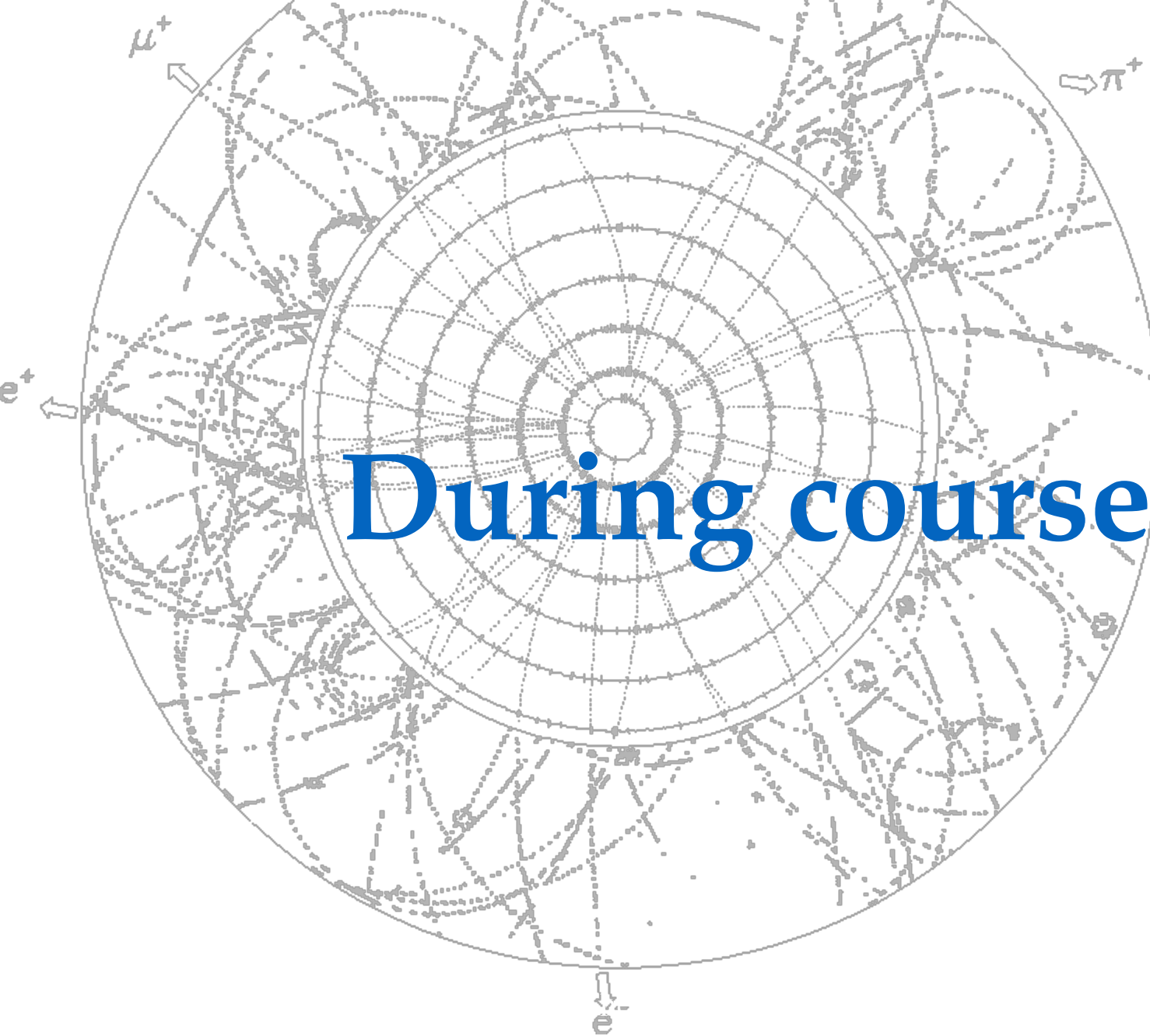
CHALLENGE (IN LATER EXERCISE):
DETERMINE THE LENGTH OF THE LECTURE TABLE
WITH A 30CM RULER AND 2M FOLDING RULE
...AND CALCULATE ITS UNCERTAINTY!

Exactly how to measure

Show up in Auditorium A (NBI, Blegdamsvej 17).

1. Say hello to the Teacher / TA in the Auditorium (if there), and get a slip of paper for the reporting of the measurements.
2. **First**, grab the **30cm ruler** and measure the length of the lecturing table. Write down the result to the millimeter. Do not round!
3. Think about what uncertainty you (gu)estimate this measurement has, and write that down too.
4. **Then**, measure the length again, now with the **2m folding rule**.
5. Again, also write down your estimate of the uncertainty.
6. Do all of the above (1-5) within 2 minutes, i.e. relatively fast!

Do NOT round your result, even if precision might be limited.
Do NOT change your results, even if you suspect/made a mistake.



Exercises

The exercises are (mostly) related to the topic of the lecture before it.

They are meant to:

- Make sure that you **understand** the lecture content, also the details of the math in it and *how to apply it*.
- Let you get **experience** with how and when the theory / principle / topic applies and works and also when it doesn't.
- Give you **confidence** in recognising certain cases and applying the fitting statistical approach next time you encounter the case type.
- Build up a **code repository** with the *relevant tools, packages, and algorithms that you know and trust*.

You don't hand in the exercises, and the questions are mainly suggestive.

You don't have to "solve it all" and there are often no unique solutions.

The best thing you can do is **sit down with peers** and go through the exercise and discuss the questions and their answers. And leave the exercise, when you feel that you're confident with the subject.

Code for Exercises

All code for exercises are located in the course GitHub repository:

<https://github.com/AppliedStatisticsNBI/AppStat2025>

Once set up (see instructions on GitHub page), you only need to do very few things for each exercise:

1. “git pull” - Gets the latest version of ALL code (incl. solution examples).
2. “cp x_original.ipynb x.ipynb” - Makes your own copy of the code (which is not overwritten, when you say “git pull” next time!)

We also provide “empty” code versions. They contain the introduction, the questions, and the learning points, **but not the essential code!**

This is for those, who would like to avoid my code, and write their own.

Code for Exercises

All code for exercises are located in the course GitHub repository:
<https://github.com/AppliedStatisticsNBI/AppStat2025>

The screenshot shows a Jupyter Notebook interface in a web browser. The browser address bar shows the URL: `localhost:8889/notebooks/Teaching/AppliedStatistics2023/AppStat2023_InstructorsFolder/Week1/CentralLimit/CentralLimit_original.ipynb`. The Jupyter Notebook header shows the title "CentralLimit_original" and the last checkpoint date "10/04/2023 (autosaved)". The notebook content is titled "Central Limit Theorem (CLT)" and includes a paragraph explaining the aim of the notebook, a paragraph about the Central Limit Theorem (CLT), and a "References" section. The "References" section lists three references: R. Barlow, G. Cowan, and a Wikipedia link. The "Author(s), contact(s), and dates:" section lists the author as Troels C. Petersen (NBI), the email as `petersen@nbi.dk`, and the date as 8th of November 2022. The notebook also shows a code cell with imports for numpy, matplotlib, seaborn, and Minuit, and a comment about setting the random seed for the random number generator (RNG).

Central Limit Theorem (CLT)

The aim of this notebook is to illustrate the Central Limit Theorem (CLT) through concrete examples.

When you add random numbers from different distributions, but with similar variance (or standard deviation), together and plot the distribution of these sums, you end up with a Gaussian distribution, as dictated by the CLT. The example also illustrates how widths (and therefore uncertainties) are added in quadrature, as one has to divide the sum by the square root of the number of random numbers that went into the sum in order to get a Gaussian of unit width (when using random numbers of unit width, i.e. $RMSE = \sigma = 1$).

References:

- **R. Barlow:** page 49 (and page 45 for Uniform distribution)
- **G. Cowan:** page 33
- **Wikipedia:** "[Central limit theorem](#)"

Author(s), contact(s), and dates:

- Author: Troels C. Petersen (NBI)
- Email: petersen@nbi.dk
- Date: 8th of November 2022

```
In [1]: import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from ipynbutils import Minuit
import sys

# Matlab like syntax for linear algebra and functions
# Plots and figures like you know them from Matlab
# Make the plots nicer to look at
# The actual fitting tool, better than scipy's
# Modules to see files and folders in directories
```

Here we set the random seed for the random number generator (RNG). This ensures reproducibility (the same results every time the notebook is restarted).

Project

In the second/third week of the course you will be working on the data analysis following two (simple?) experiments for about two weeks.

The project experiments will be in **First Lab** in NBB B1 ground floor (dividing class into two halves, the other half having lectures + exercises as normally) on:

- Friday the 28th of November 8:15-12:00.
- Monday the 1st of December 8:15-12:00.

This is your chance to fully do the statistics behind an experiment and play with real data to gain experience of what planning an experiment and detailed data analysis requires! This *will count 20% in your final grade!!!*

It will require the use of computers and modifications of some of the code you have been running.

You will be working in groups of 4-5 persons, and only one report (2-4 pages) is required from each group.

Real life problems/experiments will resemble this project!



Project

The project is an attempt at **precision measurement** of the Earth's gravitation locally at NBI, using only “simple” methods (OK - a little bit of cheating there).

You will be doing two separate experiments (both seen before by most):

- Simple pendulum.
- Ball rolling down an incline.

The goal is to **determine g in two ways and propagate the uncertainties** on these measurements. More on that (in time) on the webpages under “project”.

Project deadline: One report (in PRL style) per group only is to be handed in by **Saturday the 13th of December 22:00 on Absalon.**

Your group will be paired with another group to give each other feedback. We will of course grade projects internally.

In case you can't participate in person, you will be asked to do the pendulum experiment only, but working by yourself.



Problem set

During the course, I will give a larger problem set to be solved and handed in.

This will cover most of the curriculum covered at this point, and it *will count 20% in your final grade!!!*

It will require the use of computers and modifications of some of the code you have been running.

You are welcome (even encouraged) to work in groups, but **each student must hand in their own solution**, and you should **state your collaboration**.

It is due on **Saturday the 3rd of January 2026 by 22:00 on Absalon**.

Note:

The problem set is extensive, so I suggest that you start early.

The final exam will somewhat resemble this problem set!



Exam

Exam will be a **36 hour take-home exam** with a set of problems, which resembles the one previously given.

It will cover most of the curriculum, and it *will count 60% in your final grade!!!*

It will require the use of computers and modifications of some of the code you have been running.

You must work on your own!

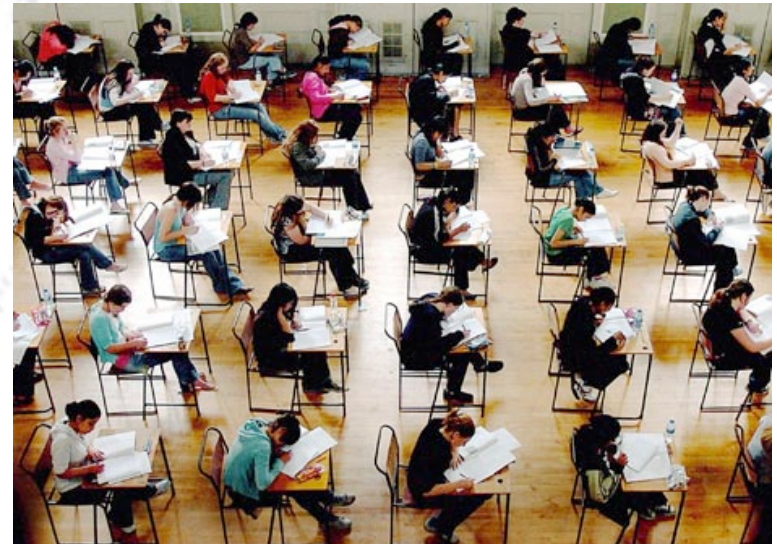
I will provide this 36 hour exam on:

Thursday the 15th of January 8:00am.

It will then naturally have to be handed in:

Friday the 16th of January before 20:00!

...on eksamen.ku.dk





Expectations

I want (read: insist) this course to be useful to all of you!

Therefore, please give me feedback (during the course, thanks!).

I will require much from you, but in return try to make statistics as interesting as possible.

The course requires self-disciplin and dedication.

We will of course do our best to inspire, help, and promote collaboration, but it is up to you, how much you want to learn/benefit from this course.

Course work can/should be done in collaboration with fellow students.

So please make small teams of peers, with whom you can discuss the many details of coding and the problems, challenges, and issues involved. This is you best way to **interact with peers, learning most, and not getting stuck.**

Problems?

If you experience problems in relation to Applied Statistics, whatever their origin and nature, then write me!

I may not be able to do anything about it, but I will try my best.

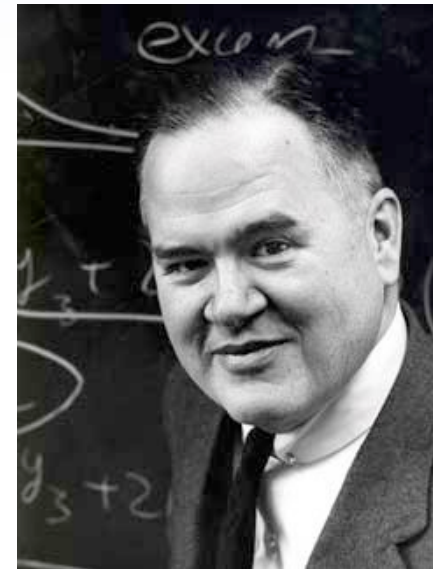
I consider myself fairly large, as long as I feel that this largeness is met by sincerity and will.

But... you need to write me in the first place! That is your responsibility.

Statistical practices

The famous statistician John Tukey (1915-2000) was quoted for wanting to teach:

- The **usefulness and limitation of statistics**.
- The importance of having methods of statistical analysis that are robust to violations of the assumptions underlying their use.
- The need to amass experience of the behaviour of specific methods of analysis in order to provide guidance on their use.
- The importance of allowing the possibility of data's influencing the choice of method by which they are analysed.
- The need for statisticians to reject the role of “guardian of proven truth”, and to resist attempts to provide once-for-all solutions and tidy over-unifications of the subject.
- **The iterative nature of data analysis**.
- Implications of the increasing power, availability and cheapness of **computing facilities**.
- The training of statisticians.



"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." J. W. Tukey

Top 10

Most important things in applied statistics

1. Errors decrease with the **square root of N**
2. **ChiSquare** is simple, powerful, robust and provides a **fit quality** measure
3. **Binomial** distribution → **Poisson** distribution → **Gaussian** distribution
4. **Error propagation** is **craftsmanship** - **fitting** is an **art**
5. Error on a (Poisson) number, N : $\sqrt{N}$ on a fraction, $f=n/N$: $\sqrt{f(1-f)/N}$.
6. **Correlations** are important and needs consideration
7. Hypothesis testing of H_0 (null) and H_1 (alt.) is done with a test statistic t
8. The **likelihood** (ratio) is generally the optimal estimator (test)
9. Low statistics is terrible – needs special attention
10. Prior probabilities needs attention, i.e. Bayes' Theorem