MACHINE LEARNING FOR
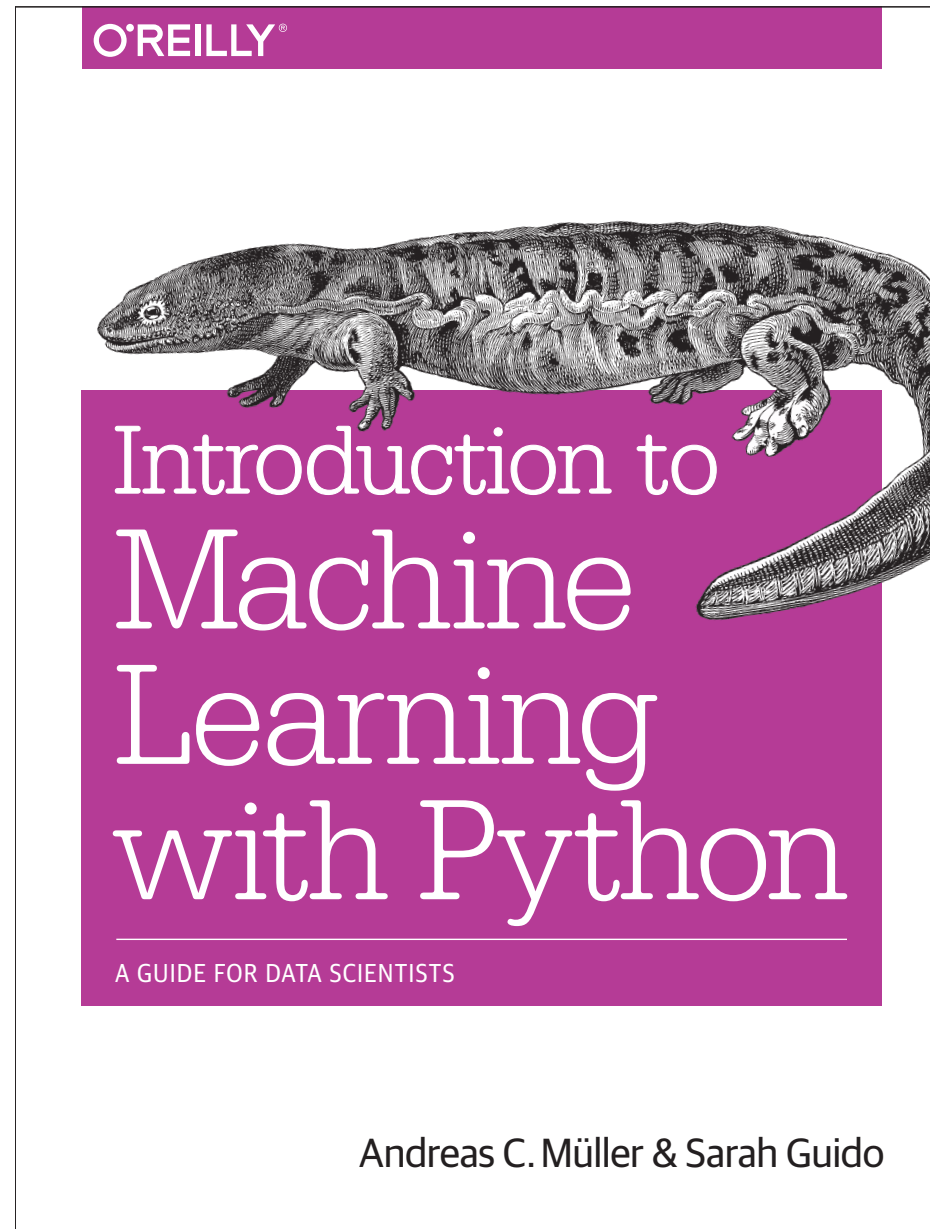## PREDICTIVE PROBLEMS

12.10.2017

Anton Akusok, Leonardo Espinosa Leal
Specialisation Studies at Arcada UAS

ARCADA

# Let's do Machine Learning!

- You will learn to run Machine Learning in Python

- Use **scikit-learn** library with existing ML methods

- Focus on solving analytical problems,
  not proving why machine learning works

- Explanation and understanding of methods without formulas!

Course Book

O'REILLY®

Introduction to Machine Learning with Python

A GUIDE FOR DATA SCIENTISTS

Andreas C. Müller & Sarah Guido

Course Book

- Andreas C. Müller, Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists*, O'Reilly Media, September 2016.

- Accompanying website (code, etc.):
  https://github.com/amueller/introduction_to_ml_with_python

- My copy of the book available on Itslearning
  (please don't distribute!)

- You will need Python version 3 (i.e. from Anaconda),
  scikit-learn (`pip install --upgrade scikit-learn`)
  and custom package **mglearn** (`pip install mglearn`)
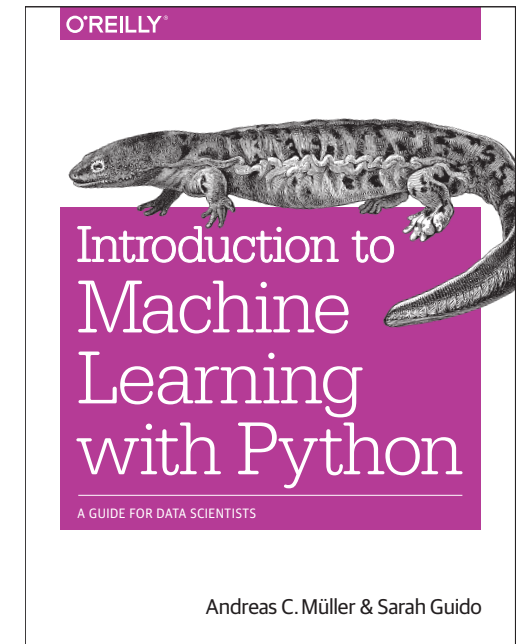
New: Flipped Classroom

- Course book focused on self-study, no point reading it at lectures…

- You get grade points for reading the book!
  Half the grade comes from questionaries on reading material

- Other half from tinkering with Python code from the book and applying it to the course dataset   (*the book has all the code*)

New: Flipped Classroom

- Meetings at Arcada instead of lectures

- Starts with introduction to the topic and examples

- Use time to solve programming homework with help from a lecturer *and your peers*

- Read the material at home to get the second half of grade

- Can pass the course without visiting Arcada at all…

Meetings

- Thu 12.10 F249   Introduction to ML, model evaluation
- Fri   13.10 F249   Supervised and unsupervised methods

- Thu 26.10 B518   Feature extraction, distributed processing
- Fri   27.10 F249   Model structure selection, pipelines

- Thu 09.11 F143   Deep Learning
- Fri   10.11 F249   Project presentations, ELM toolbox demo

Workload and Grading

- *50% grade*
  Solve programming homeworks and submit Jupyter file
  (make sure it runs beginnig-to-end without errors)

- *50% grade*
  Read book or paper, answer questions on Itslearning

- *50% grade (optional)*
  Make a course project and present in on the last meeting

Course Project (optional)

For people who want challenge and have spare time

- Use your own data, or take our large text dataset

- Must use distributed processing with Dask

- Must include extensive model selection

- Can do in small groups of 2-3 people

OR finish Big Data Analytics project that will improve your grade for the previous course

Machine Learning

- Machine Learning (ML) is a part of Artificial Intelligence that actually works.

- Machine Learning — Who are you without your hype?

  - ML fits a mathematical function to the given data samples, in such a way that the function value can be computed for new data samples.

Validation of Machine Learning Predictions

Machine Learning calculates a number
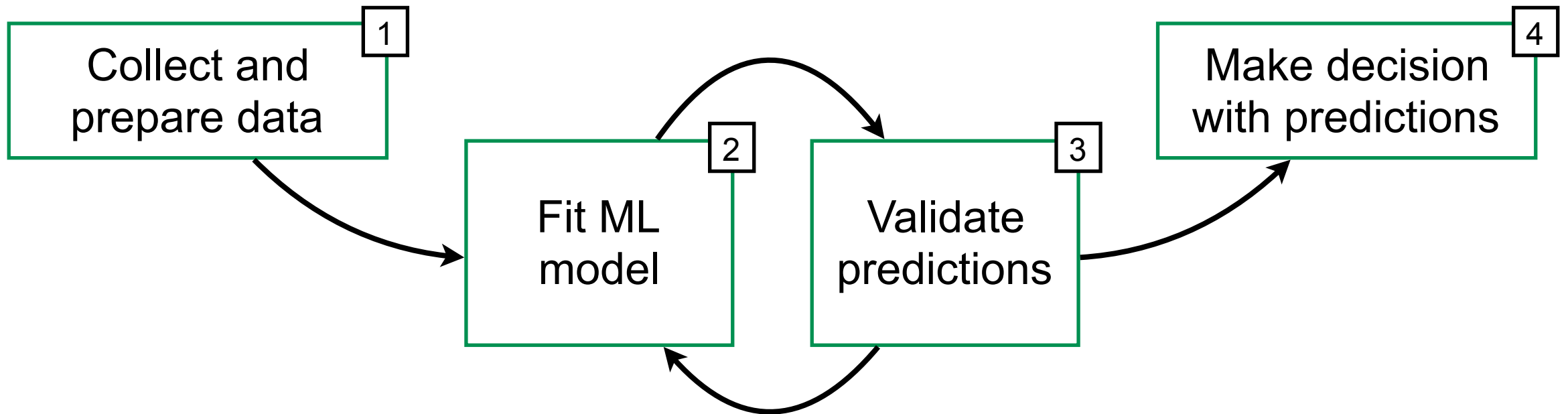
We solve a problem by predicting some real quantity

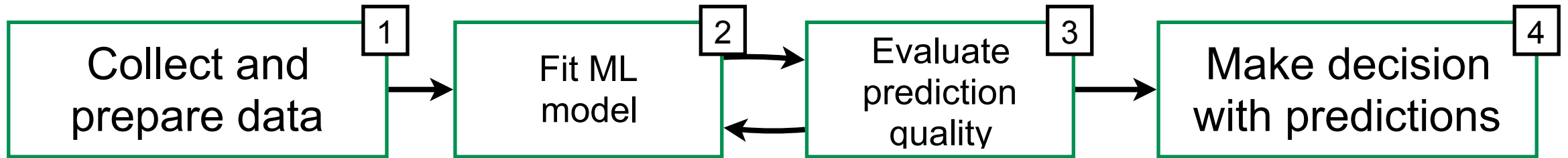$$\alpha = 3.2$$

sea shell is 5 years old

Validation tells how close are the computed numbers to the real quantities

Gerenalization tells the same for new, previously unseen data samples

Machine Learning for Predictive Problems

Machine Learning for Predictive Problems

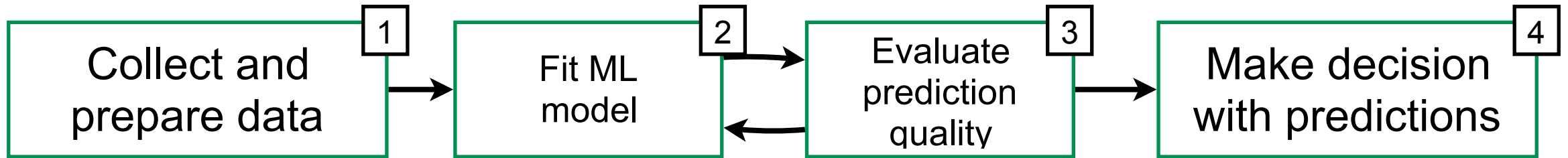| Collect and prepare data [1]	→	Fit ML model [2]	→ ←	Evaluate prediction quality [3]	→	Make decision with predictions [4]

1. The part that puts intelligence to Machine Learning.

An "intelligent" model is defined by a good dataset — with all possible scenarios, accurate labels, and tools for human experts to easily collect more data in the future. A better data always beats the better model.
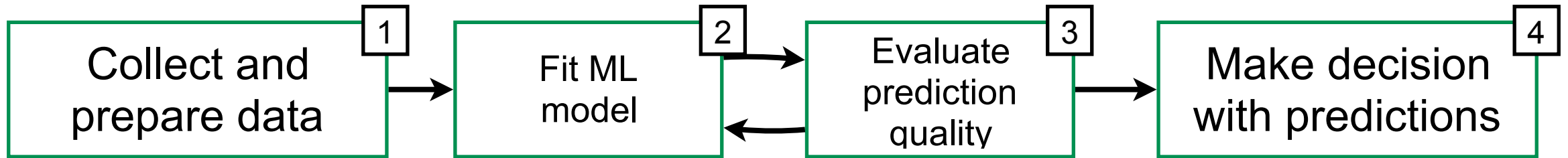
Big Data helps with this part.

Machine Learning for Predictive Problems

| Collect and prepare data | 1 | → | Fit ML model | 2 | → | Evaluate prediction quality | 3 | → | Make decision with predictions | 4 |

2. Part that puts intelligence to Machine Learning

Easy programming exercise — useful models are already coded. This course teaches to use them in Python language.

Machine Learning for Predictive Problems

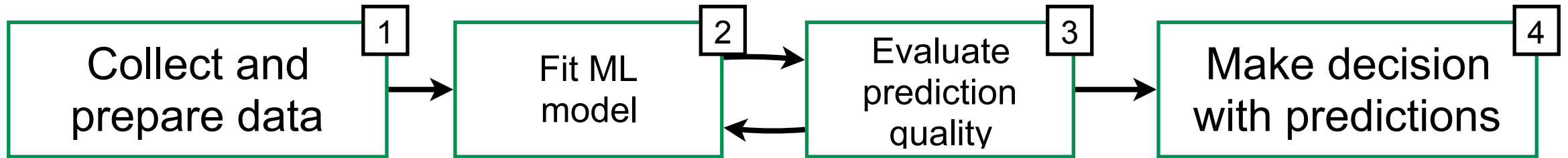| Collect and prepare data [1]	→	Fit ML model [2]	→	Evaluate prediction quality [3]	→	Make decision with predictions [4]

3. Connects mathematical function to a real-world quantity

ML calculates a number. We must check how similar is that number to the quantity that we want to predict.

The difference between a number and a true quantity is called an "*error*". The unknown difference for future data samples is called "*generalisation error*". A model that calculates numbers similar to the actual values on the future data samples can "*generalise*" well.

Machine Learning for Predictive Problems

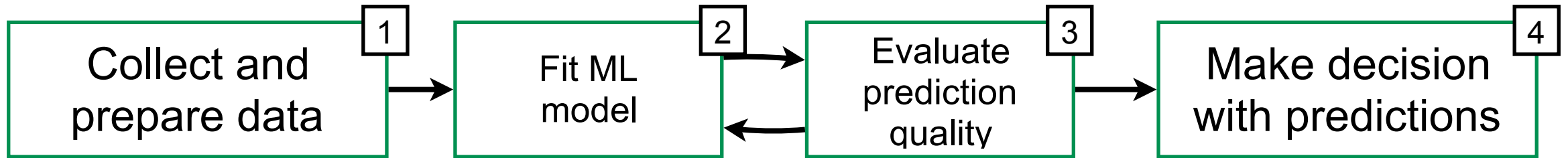| Collect and prepare data | 1 | → | Fit ML model | 2 | → | Evaluate prediction quality | 3 | → | Make decision with predictions | 4 |

3. Connects mathematical function to a real-world quantity

Important thing to remember:

— *Find the true generalisation error (on new data), not the lowest error!*

Machine Learning for Predictive Problems

| Collect and prepare data [1] | → | Fit ML model [2] | → ← | Evaluate prediction quality [3] | → | Make decision with predictions [4] |

4. The art of connecting mathematical functions to actual people, problems and revenue.

This is the art of an expert in you application field — yourself!

The process involves many people from different areas (ML, management, sales, services). Can fail even if ML works:
no access to private data, consequences of wrong predictions, etc…

Plan of Remaining Courses

• Predictive Analytics:
      Learn tasks (2) and (3)

• Descriptive Analytics:
      Learn task (1)

• Analytical Service Development:
      Try task (4) on a real problem