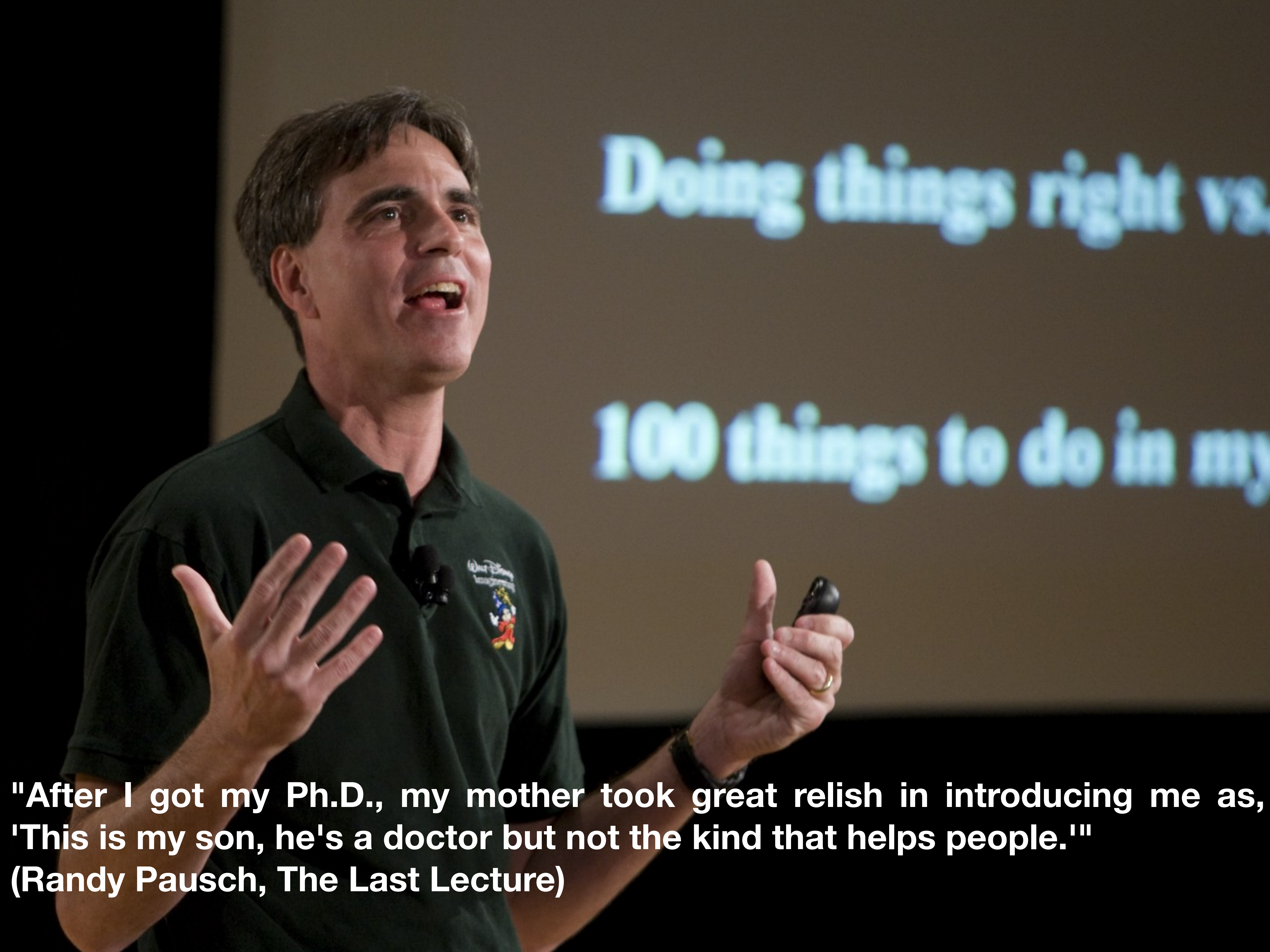
I Lost 25 Pounds Thanks to Python: Personal Data Analytics Using Pandas and Numpy

Jack Bennett

I Lost 25 Pounds Thanks to Python: Personal Data Analytics Using Pandas and Numpy

Jack Bennett, PhD



"After I got my Ph.D., my mother took great relish in introducing me as, 'This is my son, he's a doctor but not the kind that helps people.'"
(Randy Pausch, The Last Lecture)

I'm not your doctor

- "I am a doctor, but not the kind of doctor who helps people."
- I do not have an MD degree or medical training. This is not medical advice or health advice.
- Formal version: *Please talk to your own doctor and other medical professionals before making any changes to your diet, exercise, medications*.
- Informal version: *Body-hack at your own risk!*
- **However ... I do intend this talk to help you and be valuable for you.**

tl;dr

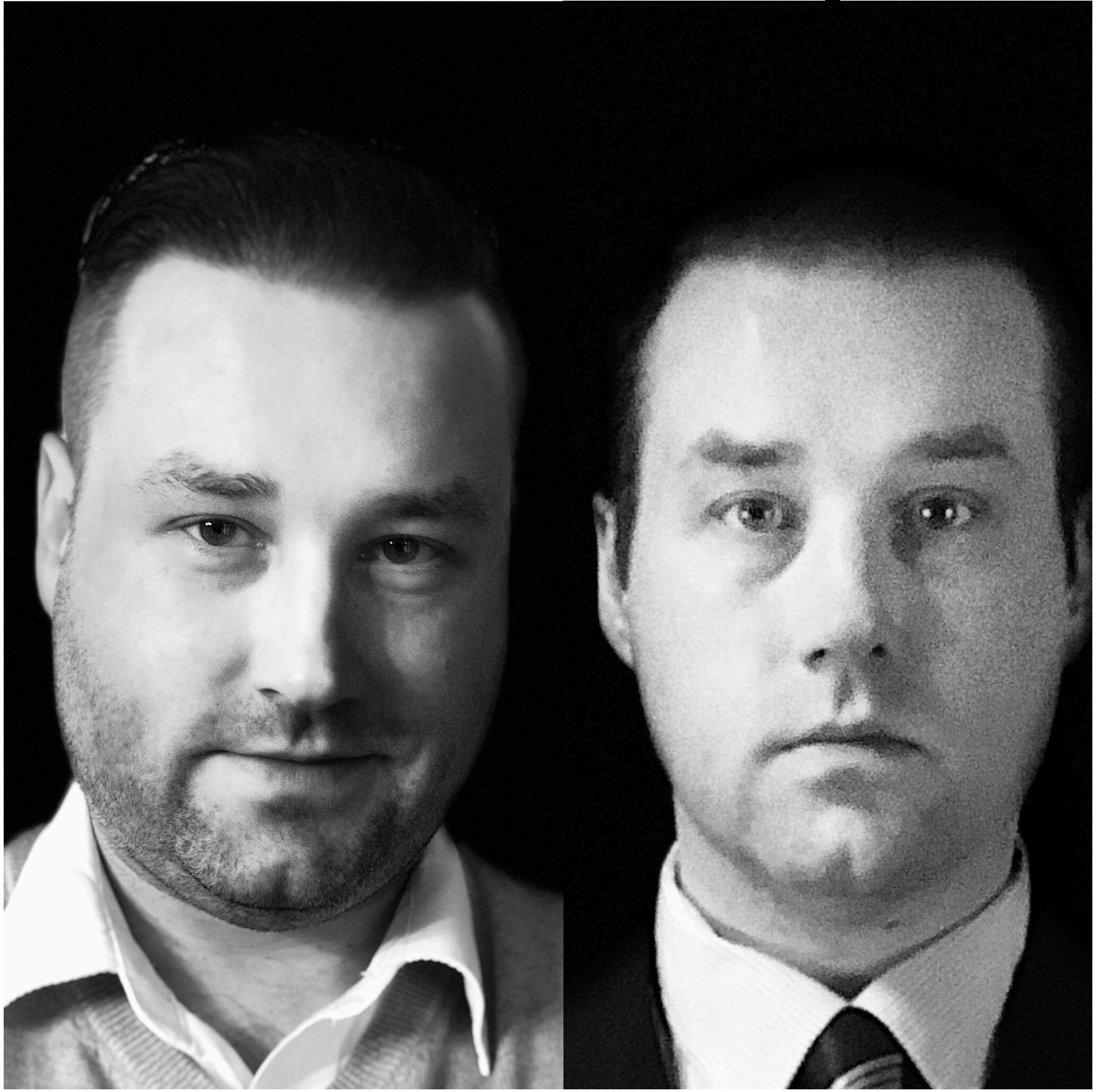
- Choosing a "data-driven goal" that is **(1) easily measurable** and **(2) important to you** increases your chances of success

- You don't need to gather a lot of data or complicated data to create significant results. Keep things simple!

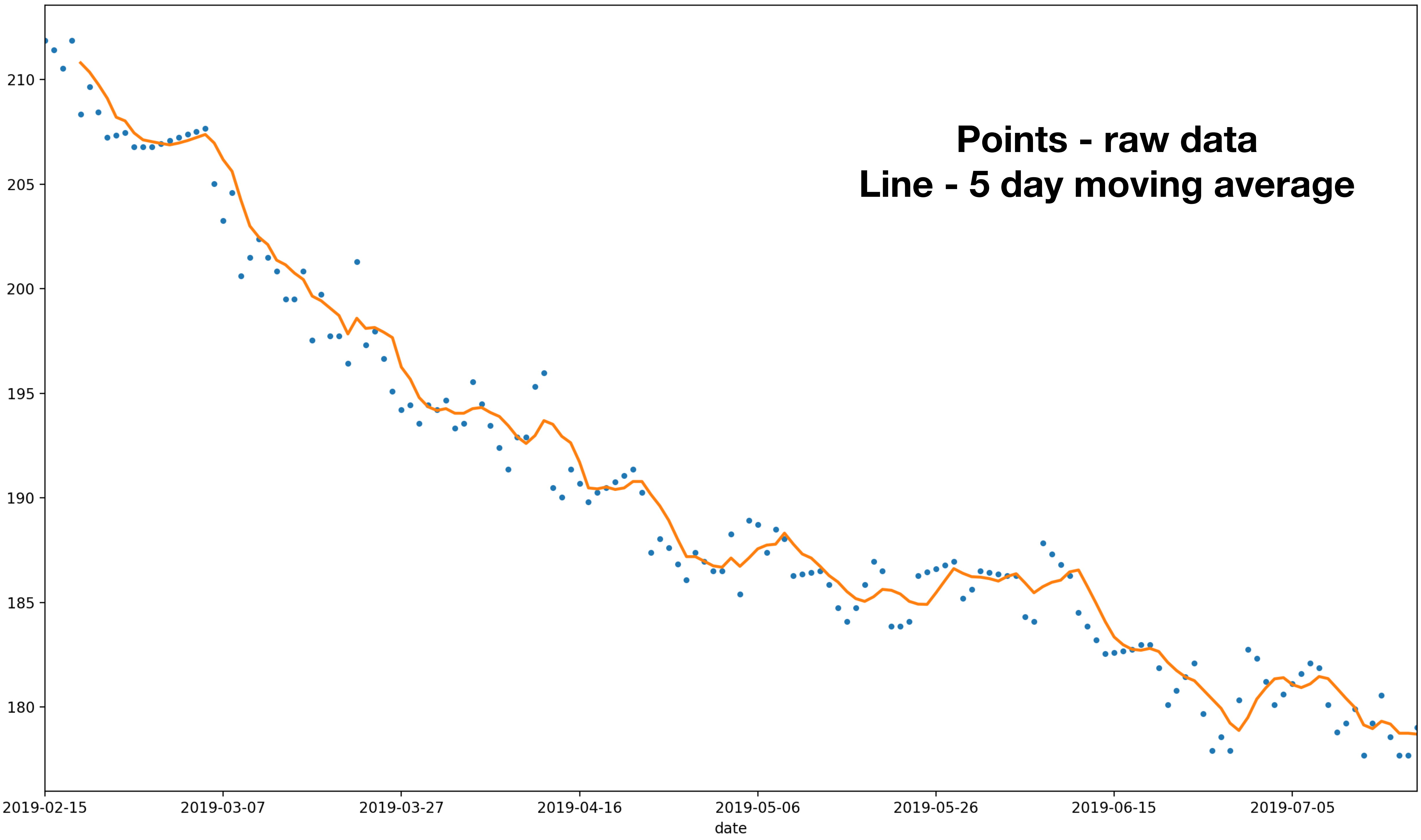
- Automating the routine and repetitive parts of your workflow makes it easier to gather, process, and interpret your data.

- Python provides a powerful, fun, and easy to use set of tools to do exploratory data analysis 🐍 🎉

Before and after pictures



Interpolated raw data and five-day trailing moving average



Think generally about the principles in this talk

- If body weight isn't interesting for you, consider these principles in terms of personal data that **does** matter to you.

- Examples:
  - savings account 💰
  - minutes spent meditating per day 🙏
  - sales calls 📞
  - words written per day on your novel 📚
  - whatever **you** (1) care about and (2) can measure! 👊

Why might you want to do this?

- **Learning** 🎓 - Great way to learn and practice with useful, real-world data science tools on small data sets.
- **Personally relevant** 🙋‍♀️ - Investigate data sets that are personally important to you.
- **Drive habit changes** 💪 - Use the situational awareness created by data to create personal change.

Results = Psychology + Data + Personal Relevance

- How to use data science to help yourself reach a goal:
  - Pick something that you personally care about ❤️
  - Identify and collect data 📈
  - Analyze data to inform your decision making ("close the feedback loop").
- You don't need **"a lot"** of data. You need **meaningful and actionable** data. (All the data in this talk: < 200 points.)

Think like a scientist

- **"All models are wrong. Some are useful."**
- Example:
  - body weight is an imperfect—but potentially useful, and easy to measure—proxy for metabolic health (e.g. diabetes, metabolic syndrome, heart disease, etc).
  - It's expensive and complex to directly measure visceral body fat (DEXA scan, etc).
- Where is your model effective? Where is your model ineffective? Where does it work? Where does it break down?
- Control what you can, and estimate what you can't control (weight measurement example)

Think like a linguist

- Case study: goal is **"losing weight"**

- "Losing" is usually associated with something bad or unwanted. Let's use "reducing"—a more neutral word.

- Are all parts of the body identical? Usually when people refer to "*~~losing~~ reducing weight*", they mean body fat, not brain or bone or muscle. Let's be specific about body fat if that's what we really mean.

- **Revised goal: "Reducing body fat in a healthy and permanent way."**

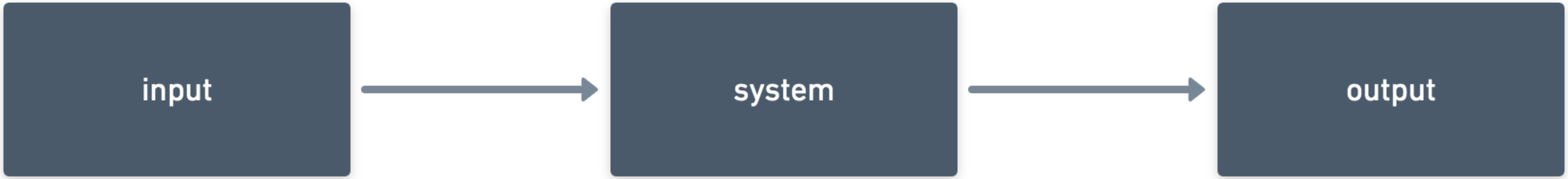
Think like a psychotherapist

- Case study: goal is ~~**"losing weight"**~~ **"reducing body fat in a healthy and permanent way"**

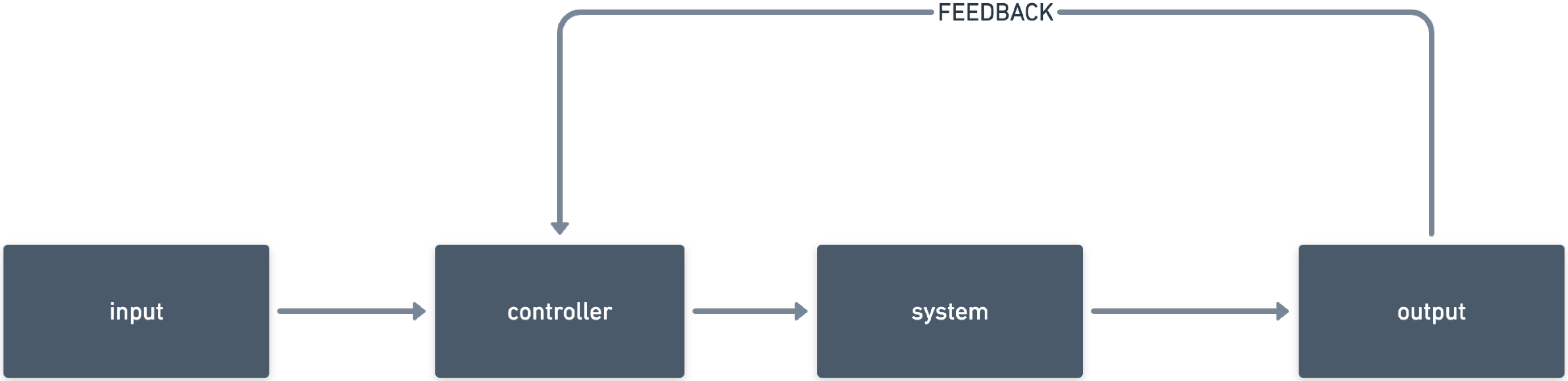
- Be precise -> what do you **really** want? Ask "Five Whys".

- Drill down to the real motivation and the real goal.

Open-loop system (uncontrolled)



Feedback control loop



Feedback control loop

- We use **measurement** of the **output** (body weight) to change the **inputs** that influence the measurement.

- This process **closes the feedback loop**.

- Now it's possible to run experiments -- change the **inputs** and see if and how they change the **outputs**!

Inputs and outputs

- **Inputs:**
  - **food:** composition, quantity, timing
  - **sleep:** quantity and quality
  - **exercise:** type and quantity
- **Outputs:**
  - **body weight**
  - **body composition** (muscle, bone, fat, water, etc)

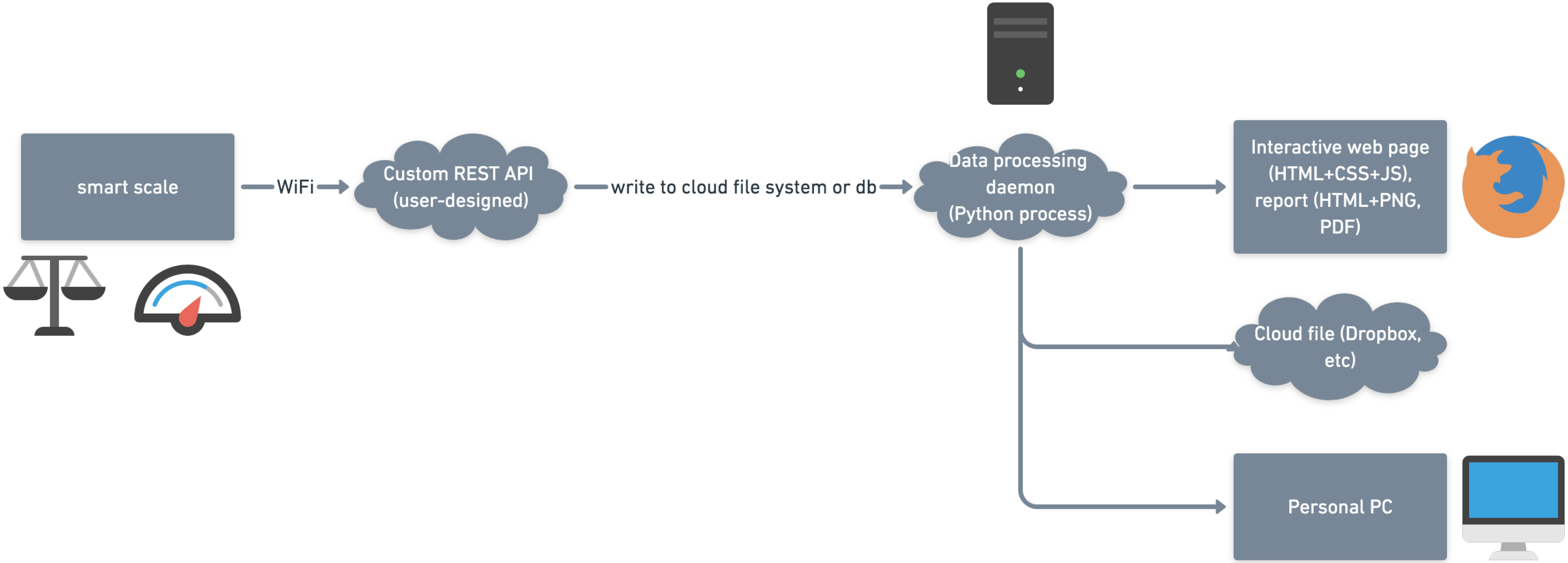
Some thoughts about measurements and data

- How precisely can you measure something? ("My weight is 198.717583 lbs")
- Does "I increased/reduced half a pound last week" mean anything at all?
- **Instrument error** - your scale is imperfect ⚖️
- **Unwanted but real variations of the measurement** - e.g. some days you are more hydrated than others 🚰. This is an uncontrolled feature of the system.
- **True and meaningful variation in data over time** - the actual signal (the thing you actually want to measure) 📈
- Think carefully about the story emerging from your data - what is meaningful and what isn't?

Ideal data analysis pipeline

- Data moves from smart device to computer or cloud under full user control (e.g. device writes to an HTTP API on my own server)
- Aggregate data set on cloud server and/or user's PC
- Analyze data in Python (IPython, NumPy, Pandas, etc)
- Render results in desired location (web, document, etc)
- Unfortunately, we do not have this. 🙁

Ideal data analysis pipeline



smart scale

WiFi

Custom REST API (user-designed)

write to cloud file system or db

Data processing daemon (Python process)

Interactive web page (HTML+CSS+JS), report (HTML+PNG, PDF)

Cloud file (Dropbox, etc)

Personal PC

Actual data analysis pipeline

- Smart device (scale) writes to proprietary smartphone app via Bluetooth (semi-manual)

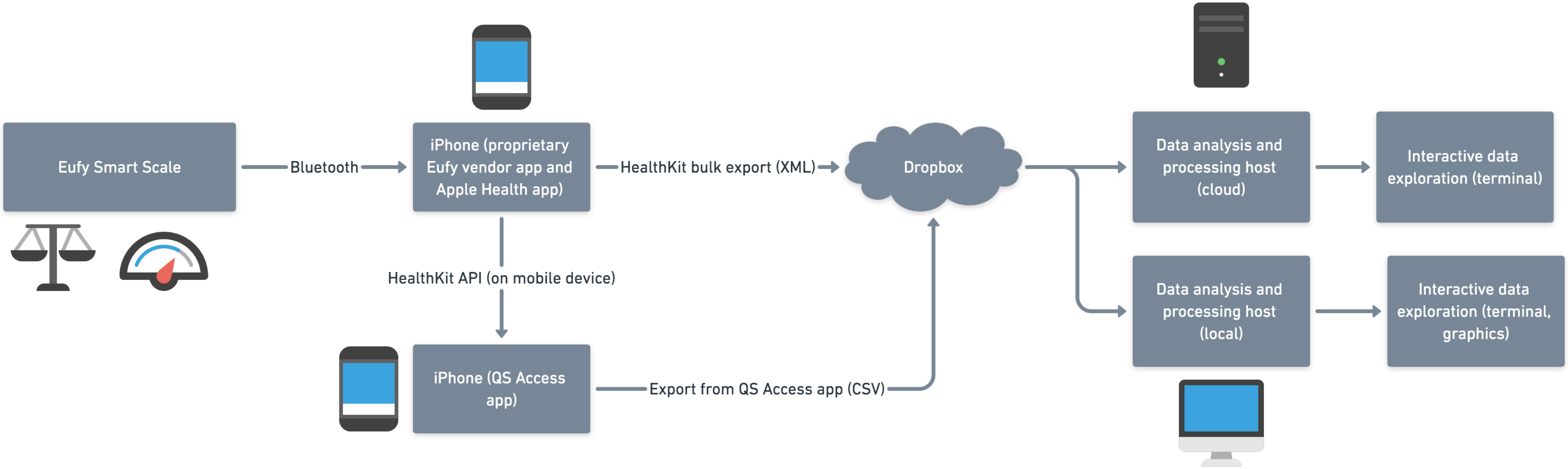
- Export all HealthKit data (Apple Health) or on-device third-party HealthKit client app (QS Access) **(manual step)**

- Save/transfer to a convenient location from mobile device (I use Dropbox).

- Process data. Numerical analysis, graphics, etc. Automated or interactive.

- Once it's in our hands or on your server, we can do what we like with it. 😀🎉 However, it's harder to get there than we want.

Actual data analysis pipeline



Eufy Smart Scale

Bluetooth

iPhone (proprietary Eufy vendor app and Apple Health app)

HealthKit bulk export (XML)

Dropbox

HealthKit API (on mobile device)

iPhone (QS Access app)

Export from QS Access app (CSV)

Data analysis and processing host (cloud)

Interactive data exploration (terminal)

Data analysis and processing host (local)

Interactive data exploration (terminal, graphics)

Pandas, NumPy, Matplotlib

- **NumPy:** array data structures and tools to work with arrays
- **Pandas:** statistical and analytical structures for data tables and time series
- **Matplotlib:** plotting library for data visualization
- The main applications of these libraries are for numerical analysis, exploratory/interactive data analysis, etc.
- Not enough time to teach you all this here... but I hope I can spark your interest with a quick demo. The rabbit hole goes very deep 🐰🕳️

System configuration for exploratory data science

- Use a **virtualenv** to define and install dependencies (i.e. 3rd party libraries).

- Libraries include: **requests; dropbox; numpy; pandas; matplotlib; ipython.**

- Advantage to using your own desktop/laptop: **interactive graphics**.

- Otherwise it doesn't matter if you use a cloud VPS (remote VM), a local VM, or your local computer itself.

Very useful idiom:

`import IPython; IPython.embed()`

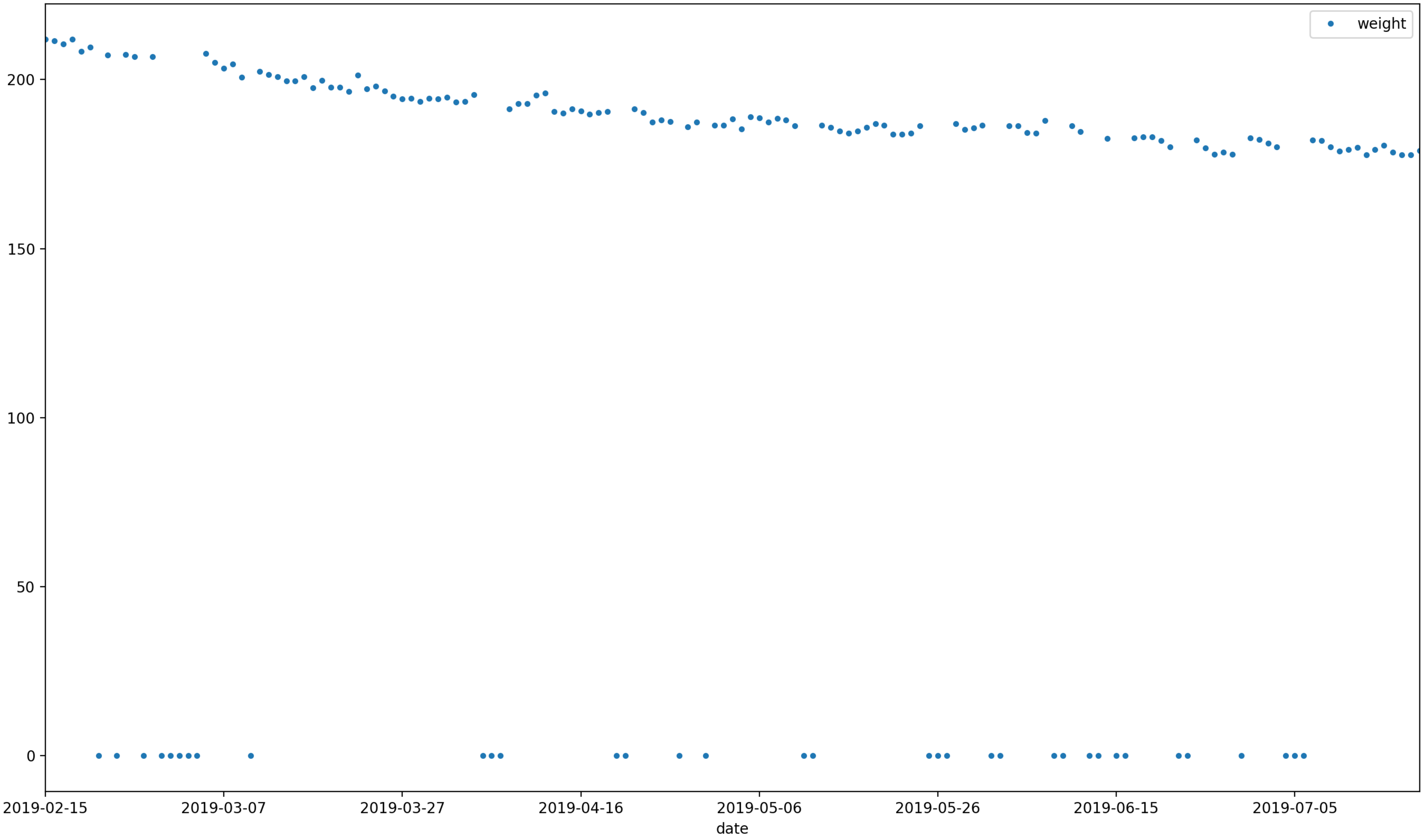
- This statement operates like a debugger. It is incredibly useful and versatile.

- It drops you into an interactive IPython REPL session, with all the class, function, variable, and other definitions active at the point of the statement!

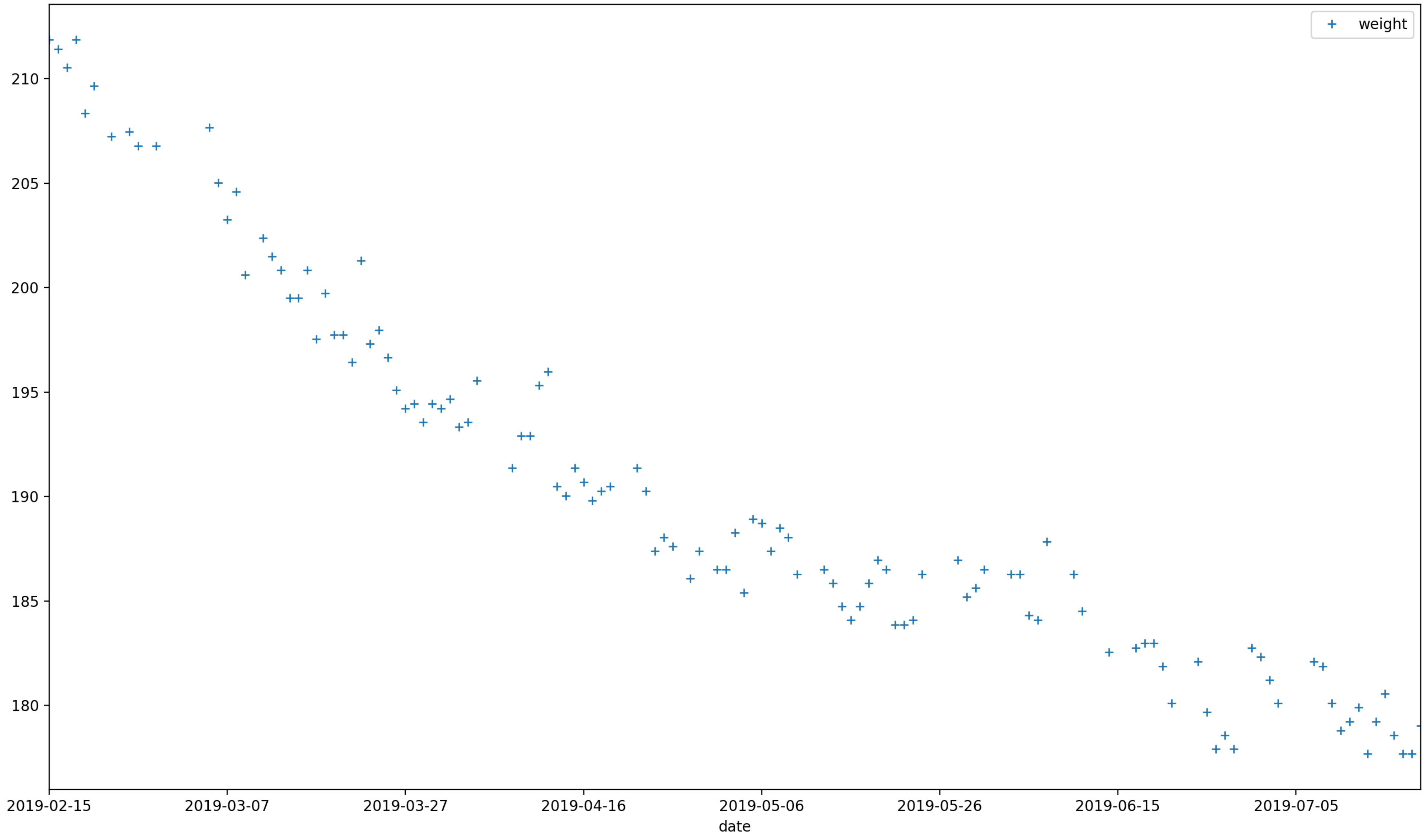
- A great way to use this interactively is to have a Python script set up all your data structures, functions, and so forth, and then execute this statement.

- When you exit the REPL, execution continues at the point right after the statement.

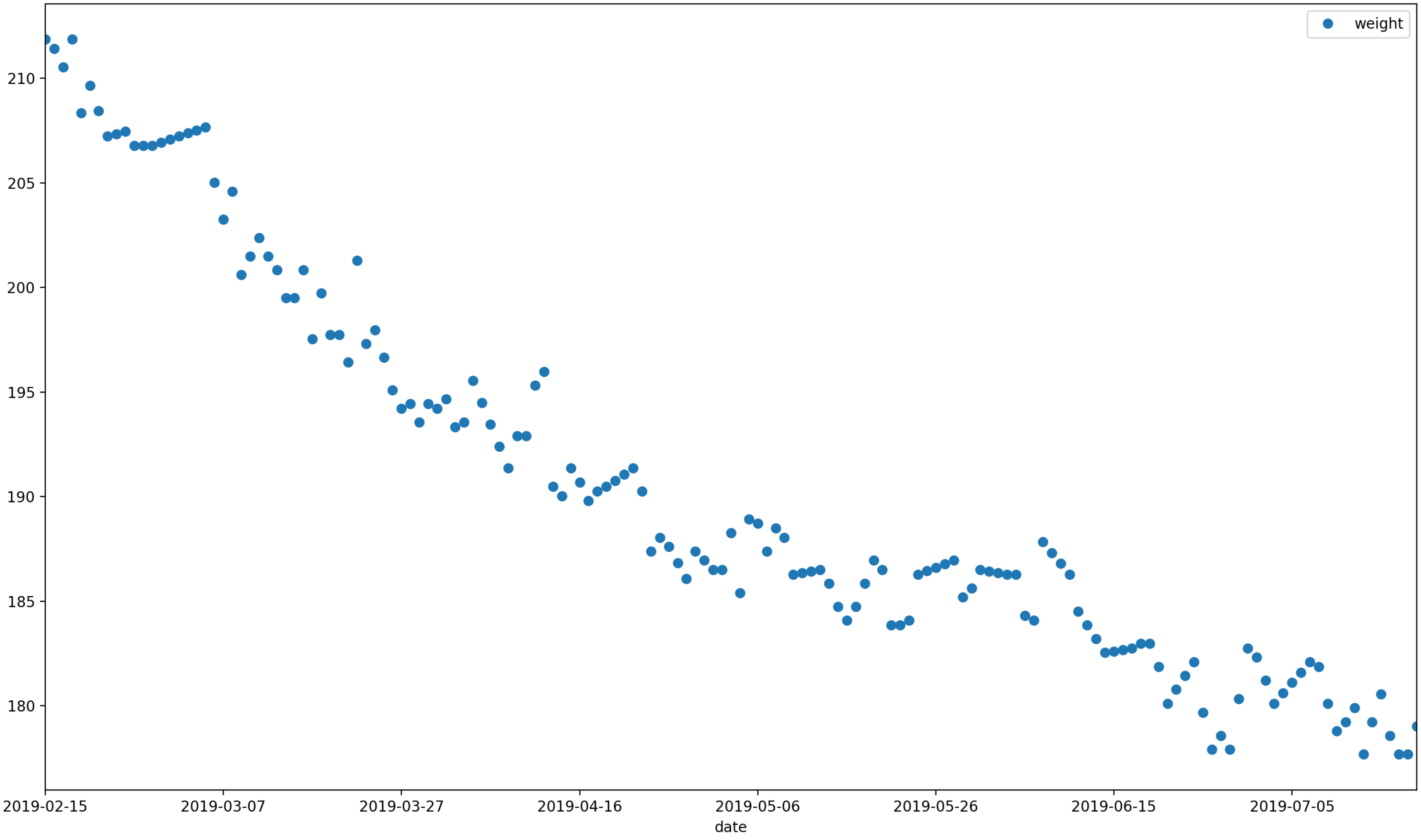
Raw data, including false zero values from QS Access app



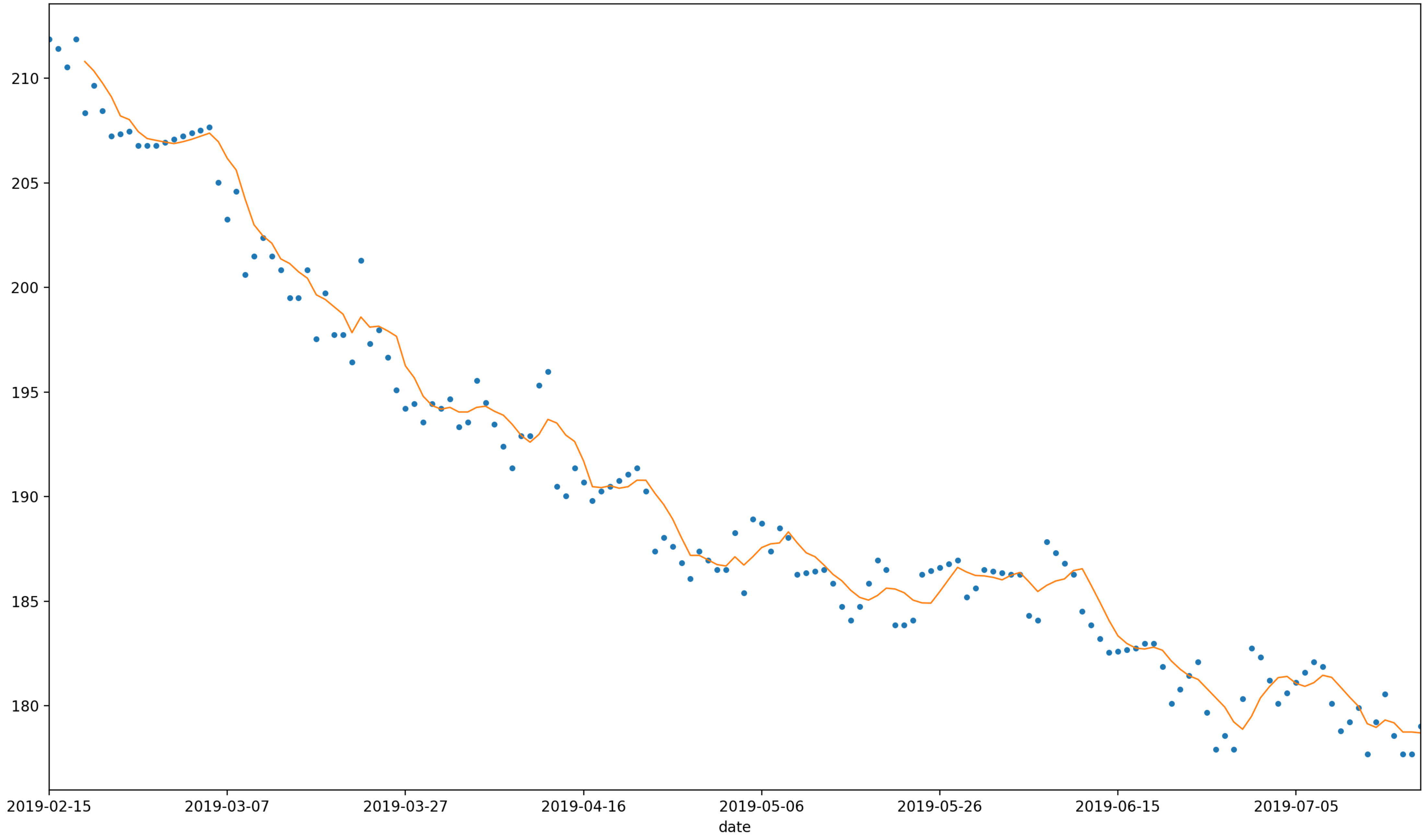
Raw data, with zeroes removed (NaN)



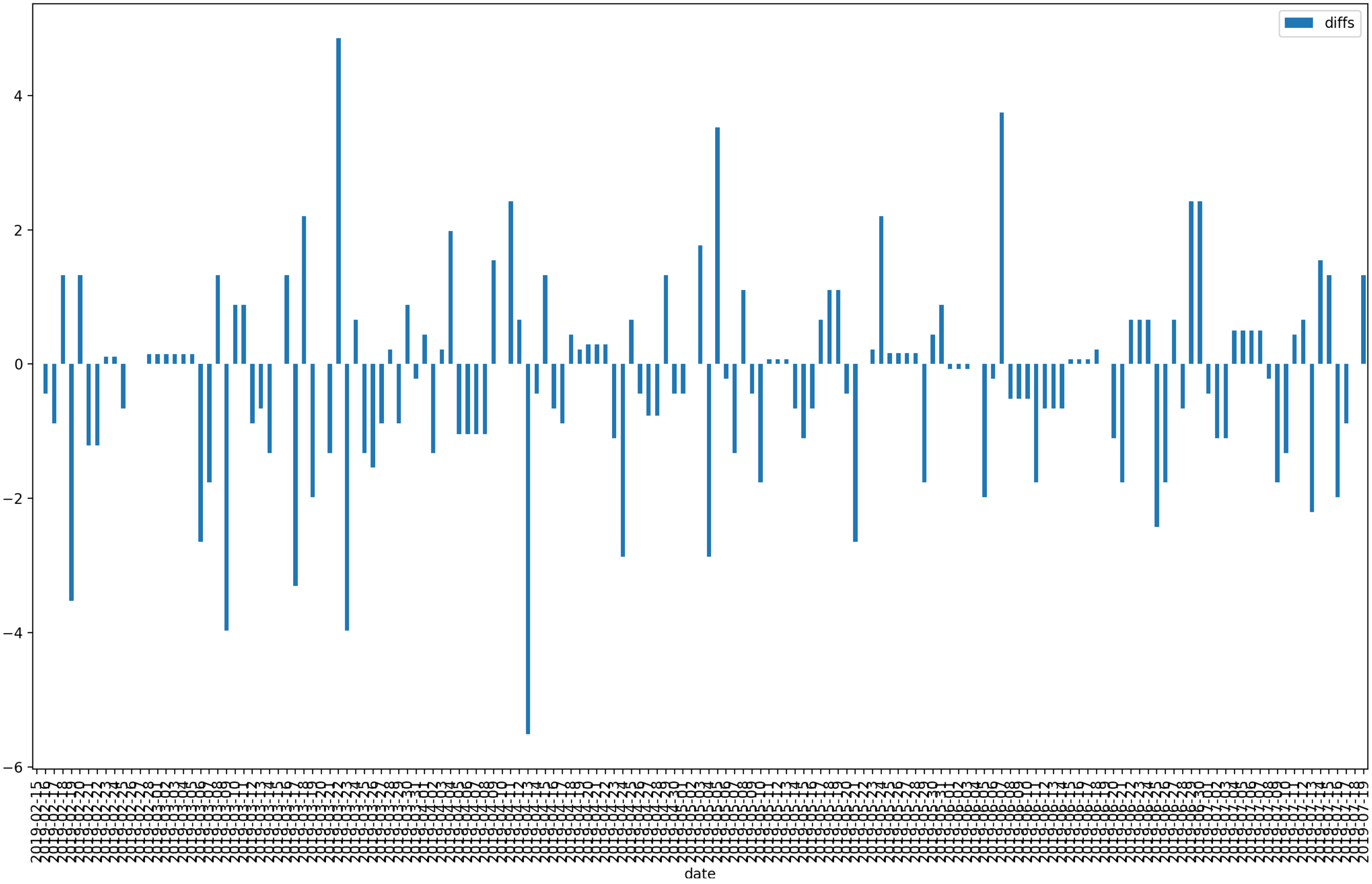
Missing data replaced with linear interpolation between real points

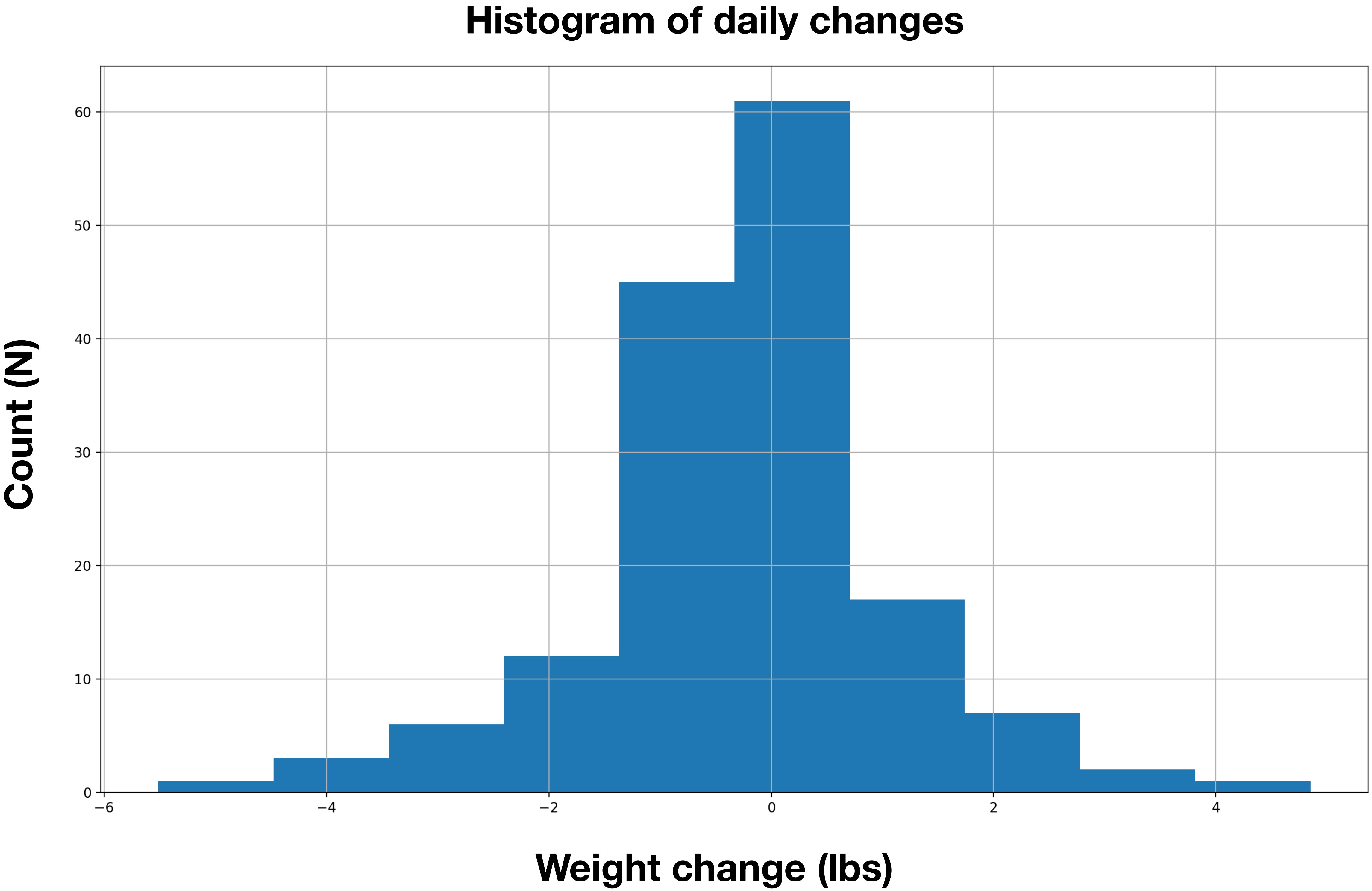


Raw + linearly interpolated data; 5 day moving average line fit



Difference plot, showing change per day ("return series")




Histogram of daily changes
Count (N)
Weight change (lbs)
0
10
20
30
40
50
60
−6
−4
−2
0
2
4

Let's look at some Python code

Summary and Conclusion

- Choosing a "data-driven goal" that is **(1) easily measurable** and **(2) important to you** increases your chances of success

- You don't need to gather a lot of data or complicated data to discover create significant results. Keep things simple! (Correlations example)

- Automating the routine and repetitive parts of your workflow makes it easier to **gather, process, interpret, and understand** your data.

- Python provides a powerful and easy to use set of tools to do exploratory data analysis 🐍🎉

Future extensions

- Automated reporting - publish daily report (web, PDF) and email to user

- Tracking different habits with different data sources - other iPhone or Android APIs, Apple or Samsung watch, web APIs, other personal smart devices, etc.

Resources

- The code samples I presented https://github.com/ajbennett/pyohio_demo_2019

- "10 minutes to pandas" https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html

- Good general introduction (the later lectures specialize in Quantitative Finance) https://www.quantopian.com/lectures