Computational Methods for Linguists

Ling 471

Olga Zamaraeva (Instructor)
Yuanhe Tian (TA)
04/22/21

Reminders

- Respond to Blog 2
- Assignment 2 due April 27

Plan for today

- Text processing, continued:
  - NLTK module and its tokenizer
  - Modules setup, PYTHONPATH
  - Unicode
- New topic: Evaluation
  - Metrics
    - Accuracy
    - Precision and recall (time permitting, or next week)

Tokenization

- In Assignment 2:
  - We will simply split words by space
  - ...to make sure we can call string functions
- In real life:
  - Always use an off-the-shelf tokenizer **package**
  - e.g. NLTK module
    - ...which needs to be **installed** via **pip**
      - **pip** is an autoinstalled included in your python

NLTK tokenizer demo

Finding modules

```
export PYTHONPATH="${PYTHONPATH}:/path/to/your/project/"

# * For Windows
set PYTHONPATH=%PYTHONPATH%;C:\path\to\your\project\
```

Commands to add a project to Pyhonpath, in bach/batch
This is to be executed in command line, or to be added to e.g. bash_profile

- Keep modules **neatly** separately
- Add the path to the **folder** from which you want to be able to import a module to PYTHONPATH
  - **PYTHONPATH**: a list of paths for python interpreter to look for modules in
- Automatic installers will often add the path when installing
  - e.g. **pip**
  - …but not always. Then, need to **locate** the installed package **folder** and add its path to PYTHONPATH

```
└── myproject
    ├── mypackage
    │   ├── a.py
    └── anotherpackage
        ├── b.py
        ├── c.py
        └── mysubpackage
            └── d.py
```

Sample directory structure
https://towardsdatascience.com/how-to-fix-modulenotfounderror-and-importerror-248ce5b69b1c

Adding to PYTHONPATH in VS Code is confusing. Only do it if really needed. There are alternatives; VS Code is good for debugging, not generally running projects!

Adding to PYTHONPATH

```
export PYTHONPATH="${PYTHONPATH}:/path/to/your/project/"

# * For Windows
set PYTHONPATH=%PYTHONPATH%;C:\path\to\your\project\
```

Commands to add a project to Pyhonpath, in bach/batch
This is to be executed in command line, or to be added to e.g. bash_profile

- In VS Code:
  - For **debugging** mode: **launch.json** and **.env**
  - For non-debugging mode: **settings.json**
    - **find** settings.json with command+shift+P
  - All files go under your **.vscode** directory
  - see uploaded files on website for reference
- If just using command line:
  - PYTHONPATH=newpath:$PYTHONPATH
  - on Windows:
    - PYTHONPATH=newpath;$PYTHONPATH

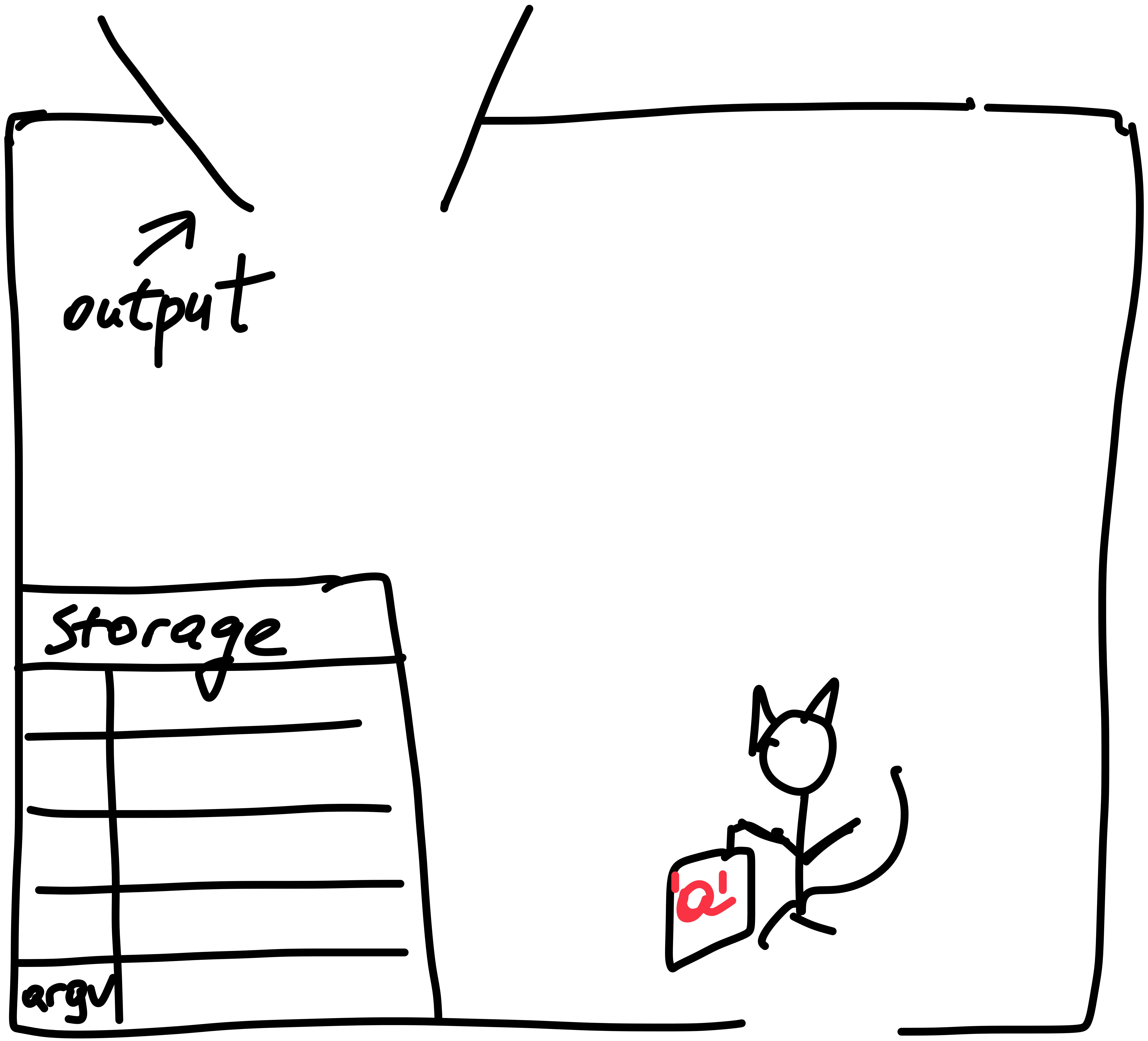
Modules demo

Encodings and the Unicode

Encodings

why they matter

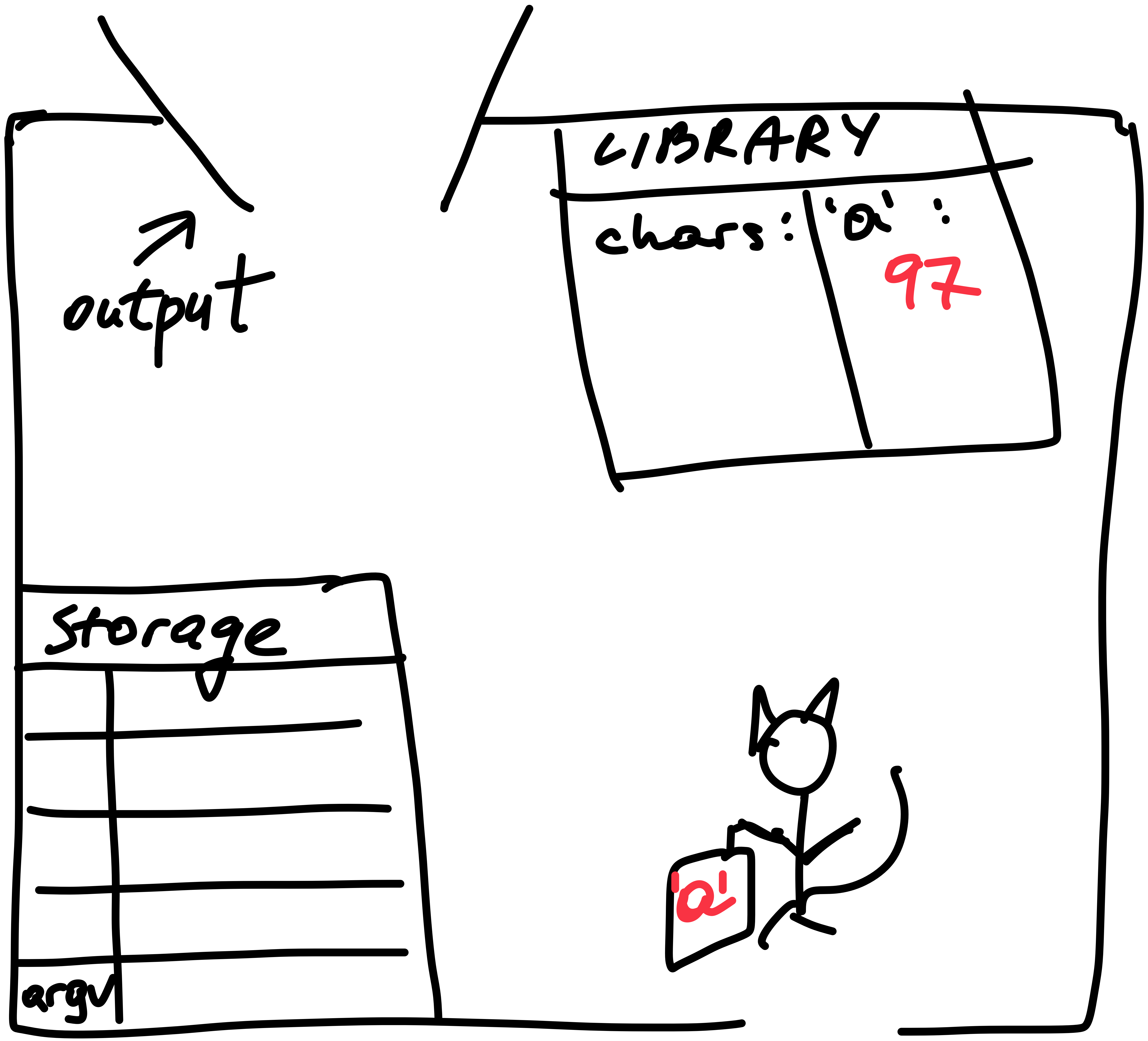
- The computer stores everything as numbers
  - including characters



Encodings

why they matter

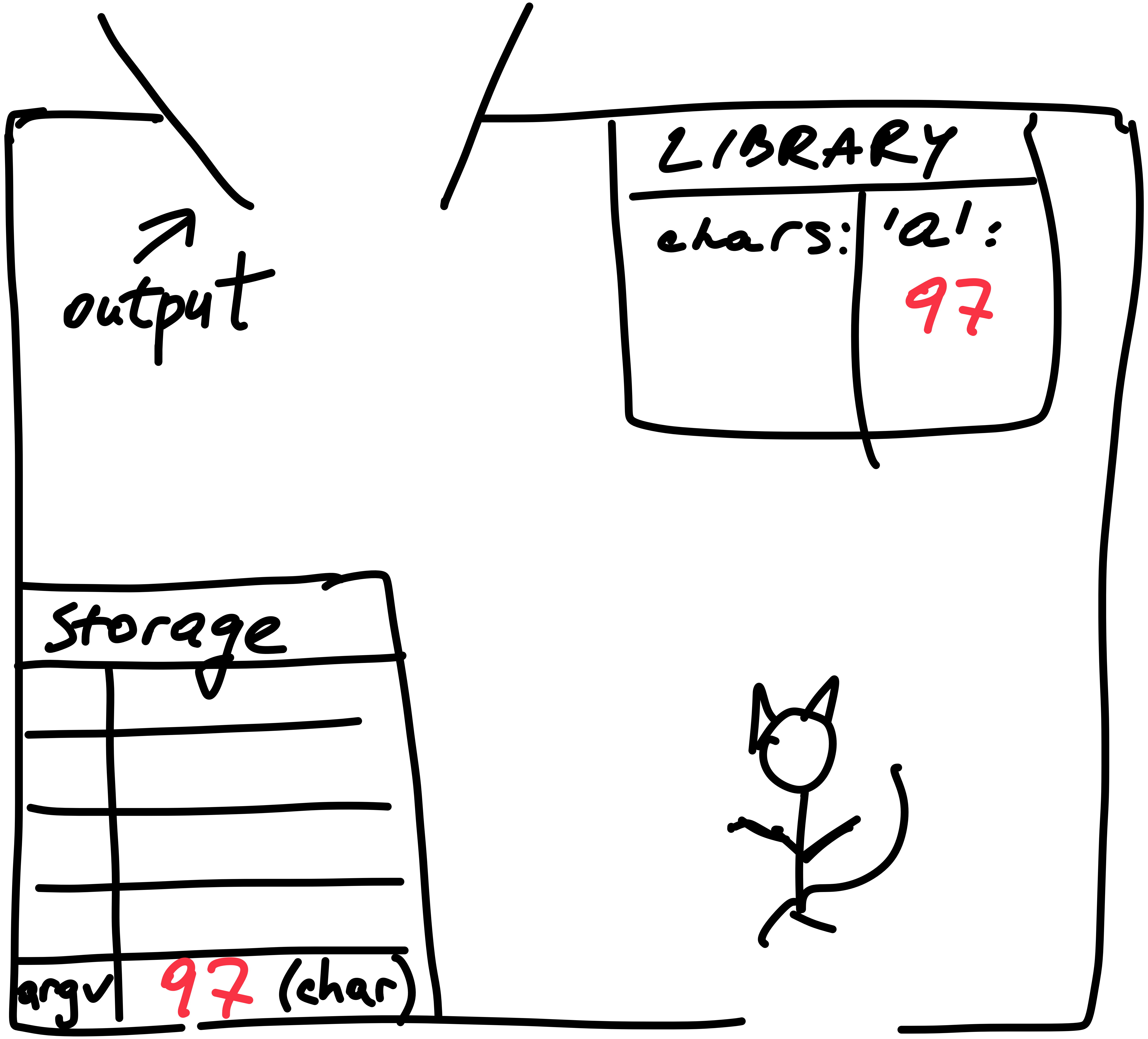
- The computer stores everything as numbers
  - including characters



Encodings

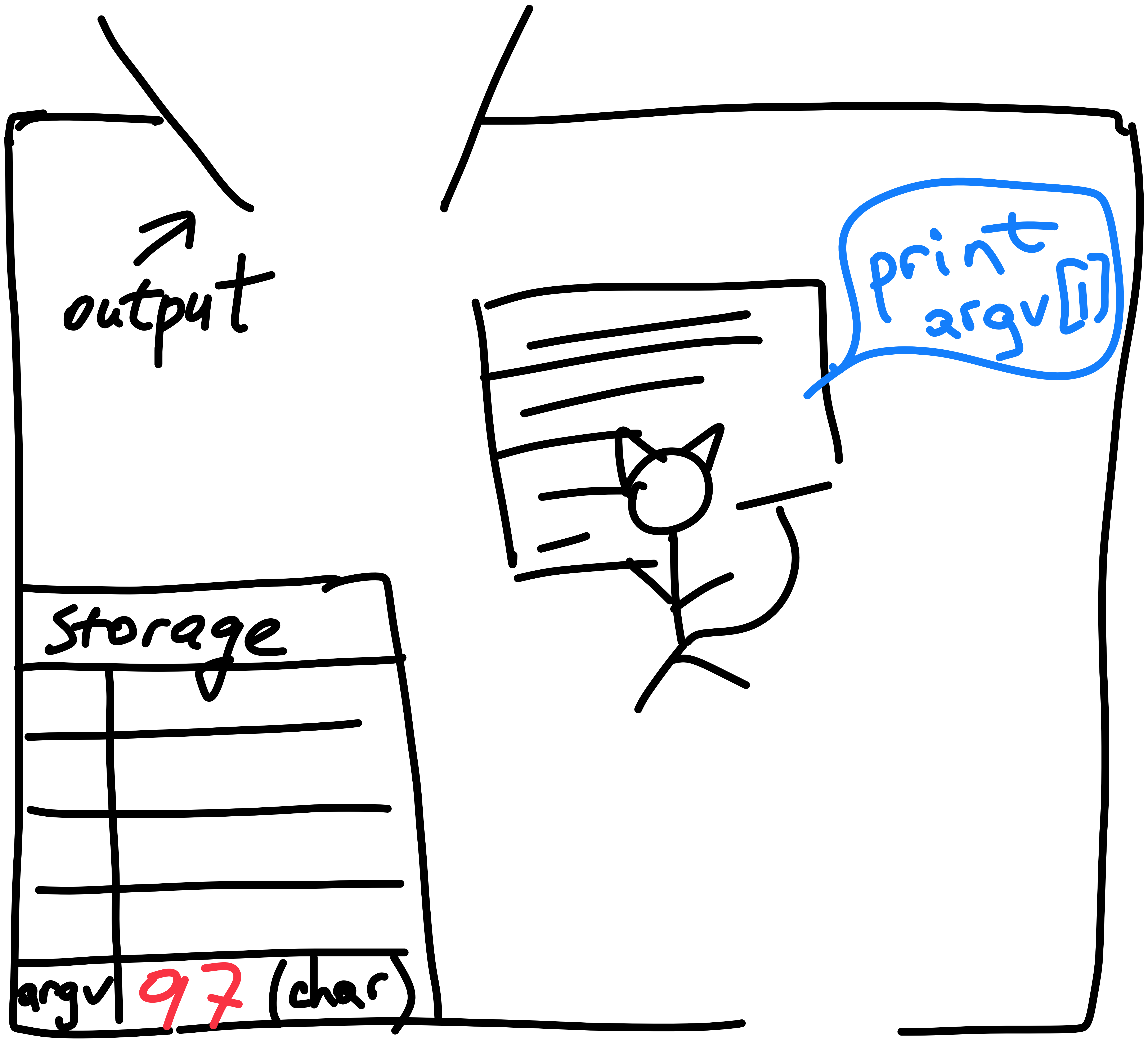
why they matter

- The computer stores everything as numbers
  - including characters



Encodings
why they matter

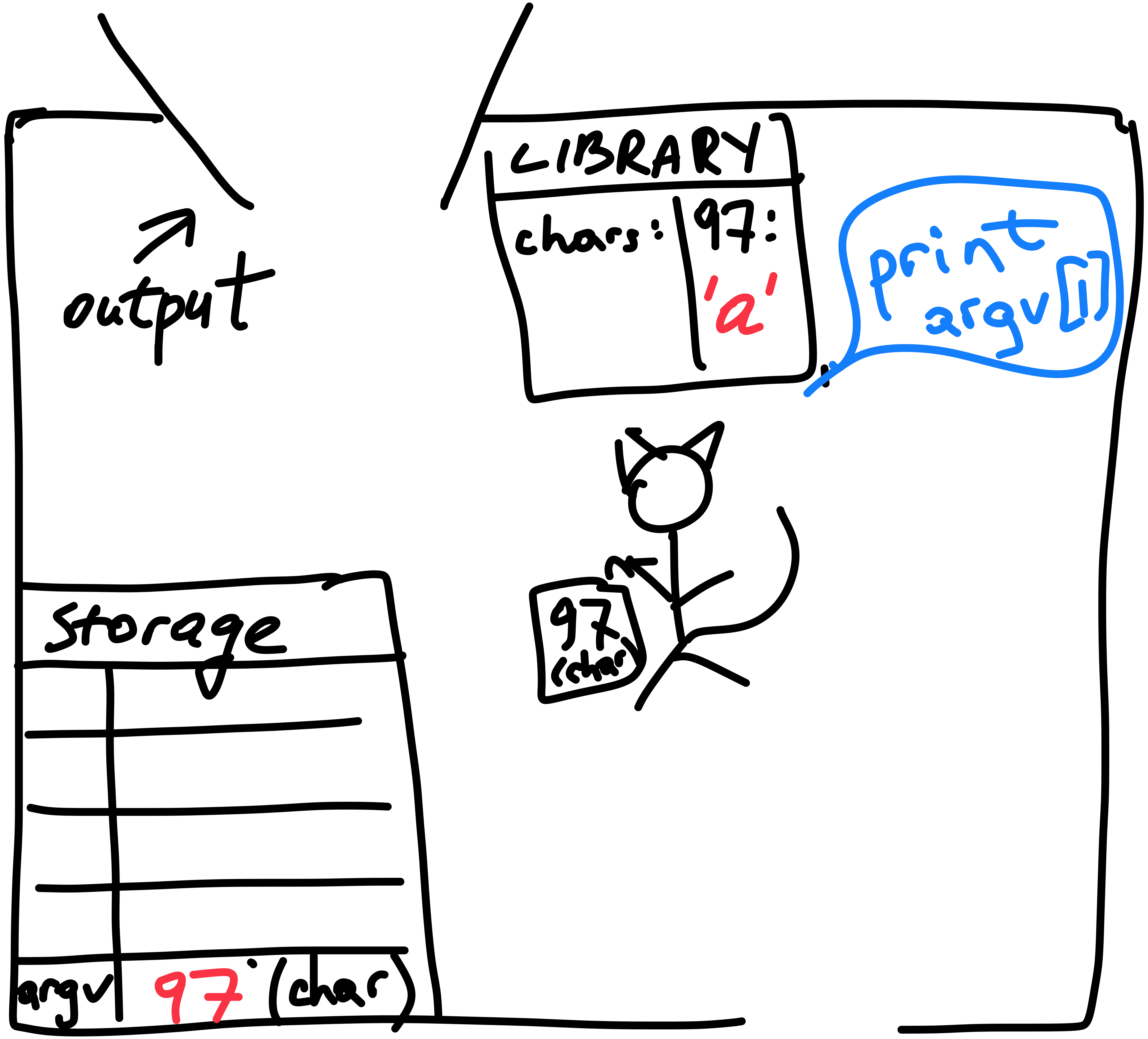
- The computer stores everything as numbers
  - including characters



Encodings

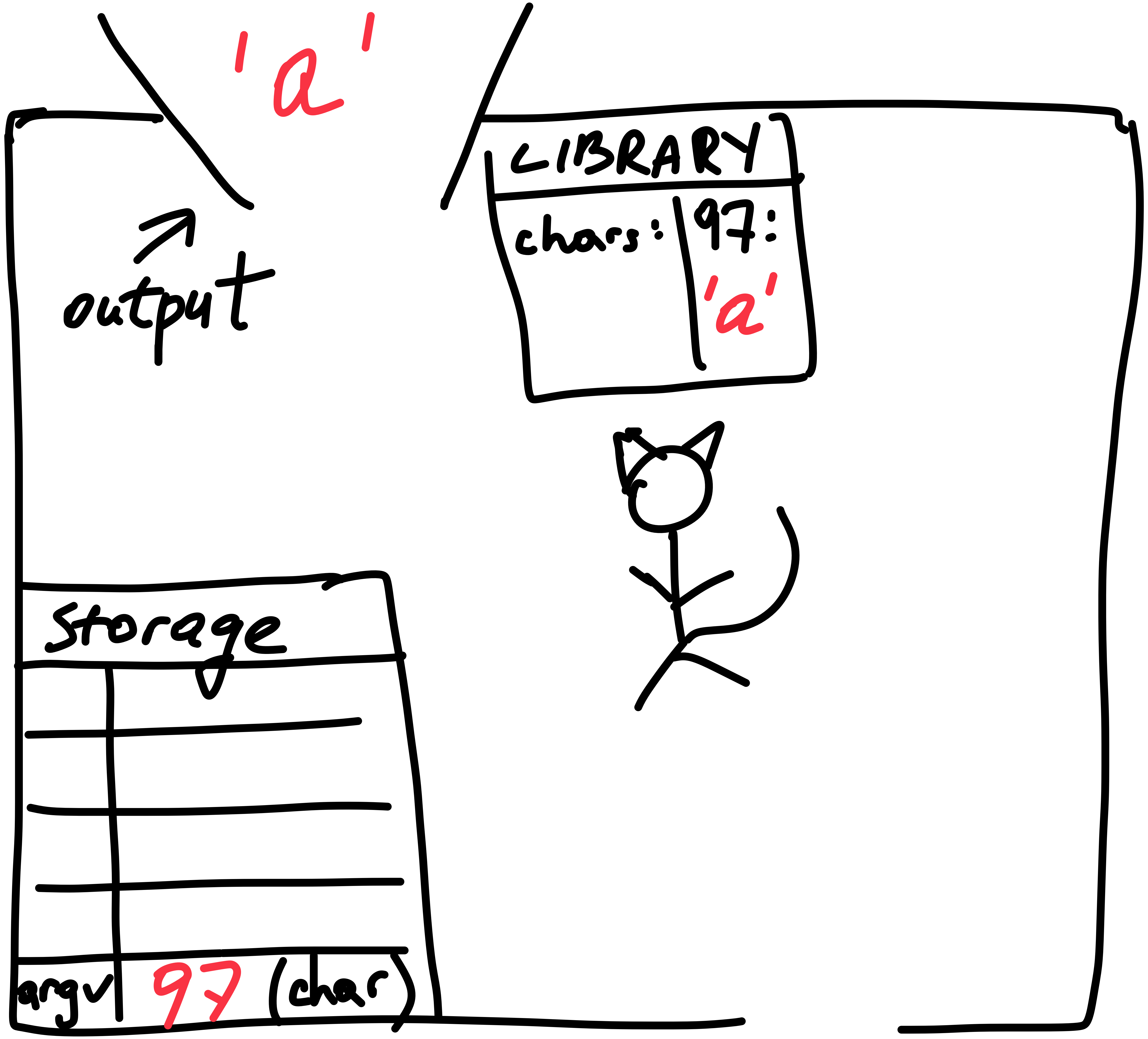
why they matter

- The computer stores everything as numbers
  - including characters



Encodings
why they matter

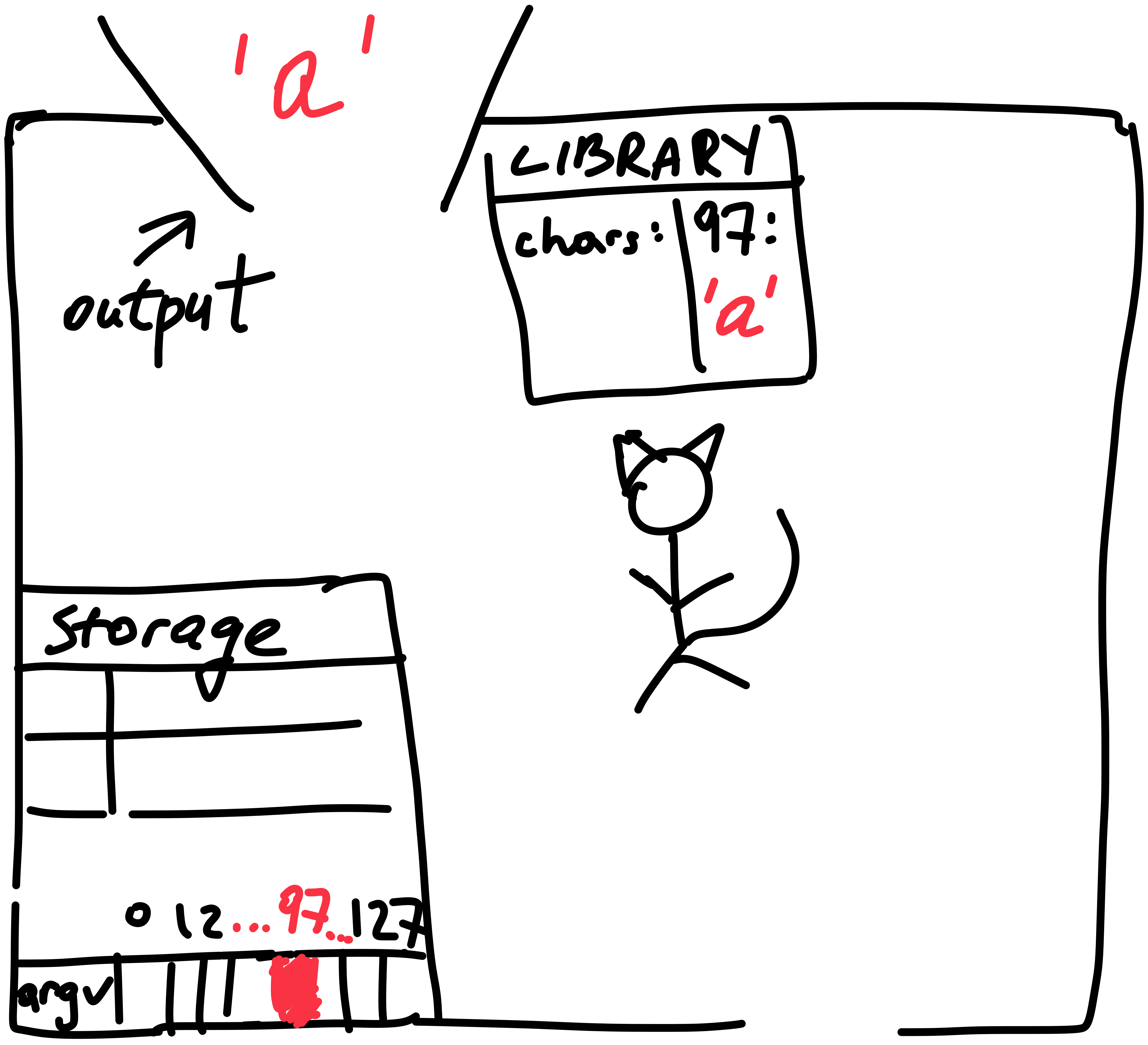
- The computer stores everything as numbers
  - including characters
- What's output is a **picture**
  - pictures take **a lot** of space
  - are difficult to **compare**
  - so you invoke the actual picture as little as possible



Encodings

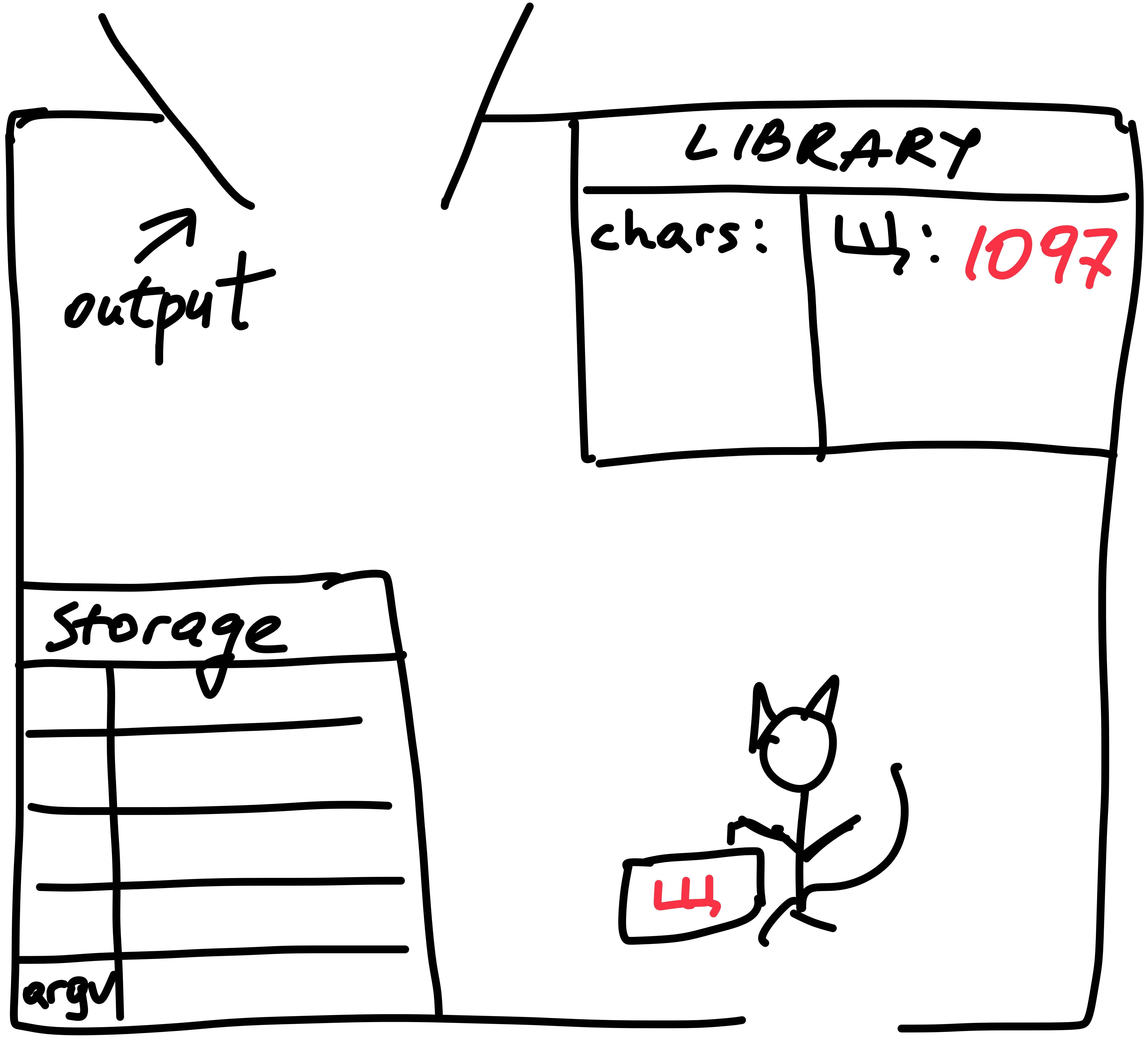
why they matter

- Needed a **convention** for operating systems, graphical adapters, etc.
  - **which** number to map **each** character to
- **ASCII:**
  - American Standard Code for Information Interchange
  - Widely used until recently
    - python2
  - Allows only for 127 characters on most operating systems
    - whatever doesn’t fit is rendered as **<?>**
    - Why? To save **space** (prioritizing English)



Unicode
consortium

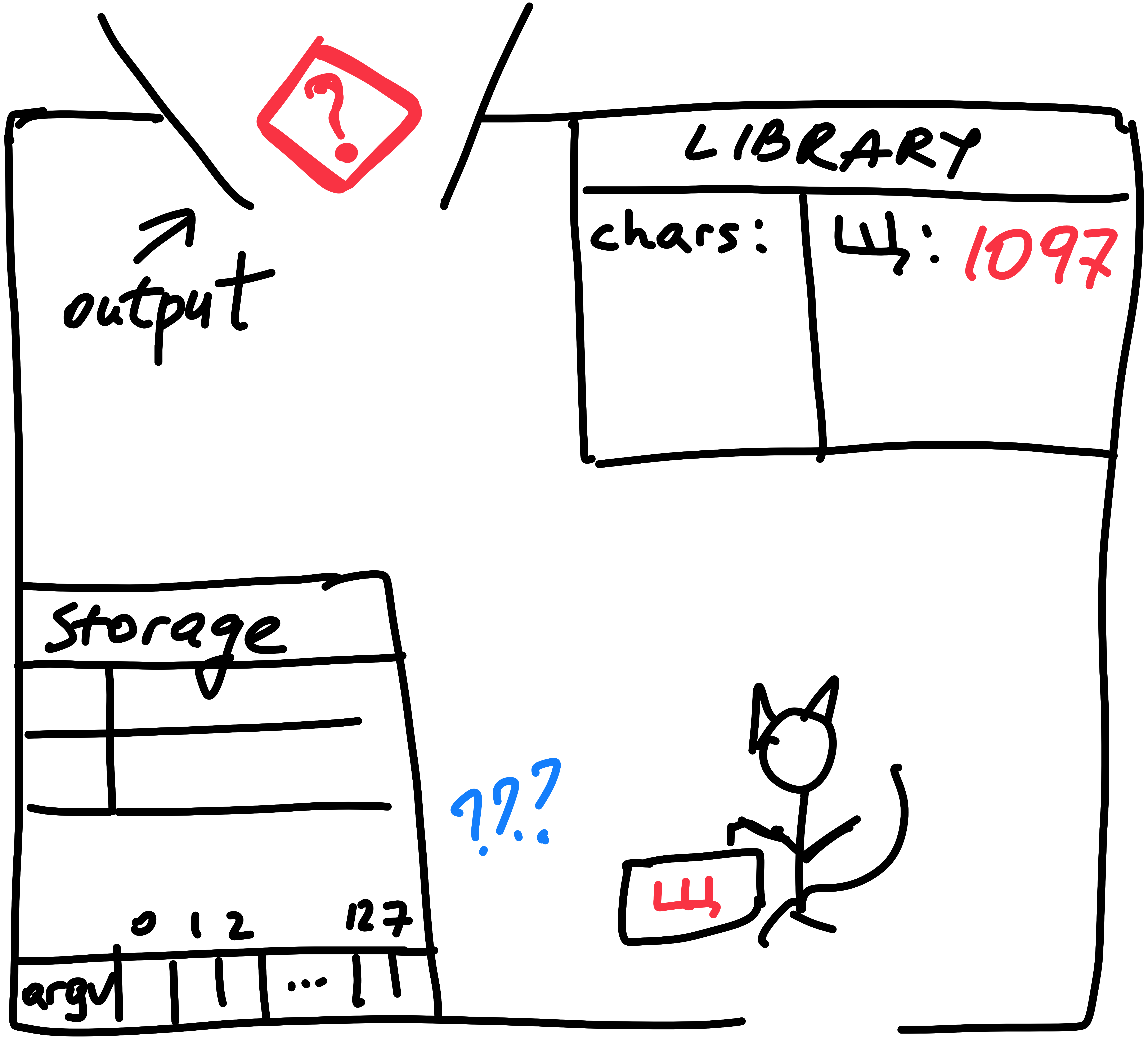
- Catalogue all known characters and assign numbers to them
  - Obviously, will need a lot more numbers than 127
  - e.g. **'a'**: 0061
    - 97 in decimal notation
  - e.g. 😀: 1F600
    - 128512 in decimal notation
- The catalogue **keeps growing**!



Non-ASCII characters

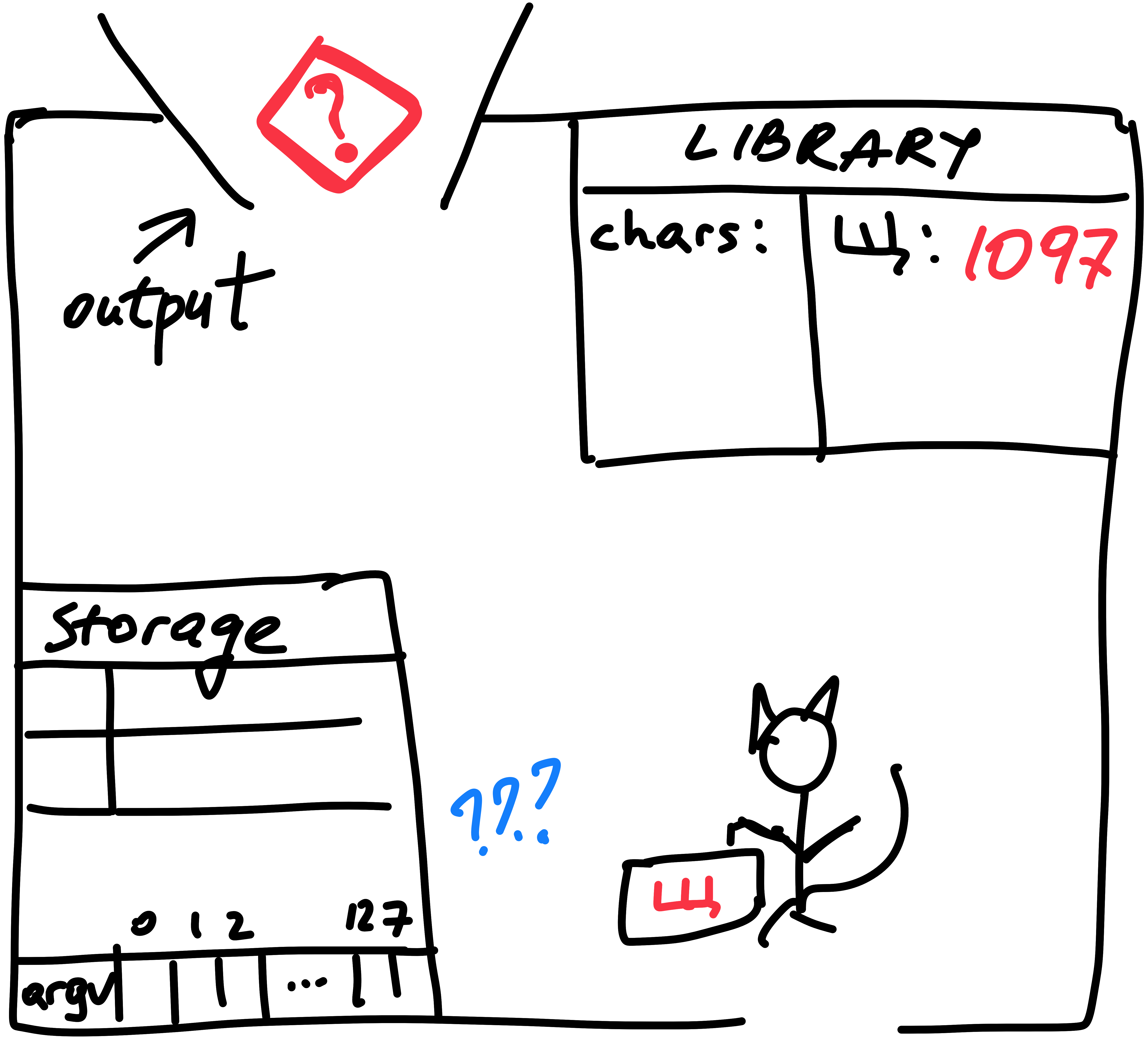
in ASCII systems

- There is only 127 possible chars
  - Everything else is simply output as a special char <?>
  - e.g. in **python2**
    - need to explicitly **change** encoding
    - open(filename,'r',**encoding**='utf8')



Unicode support
in python3

- **Enough** space reserved for **all** characters
  - ...at least the ones currently catalogued!
- **No need** to worry to much about different encodings
  - ...but only because most files are currently **saved as unicode**!
  - may **still** need to be **aware** of encodings, particularly ascii
  - ...to open files **correctly**



Evaluation in data science and NLP

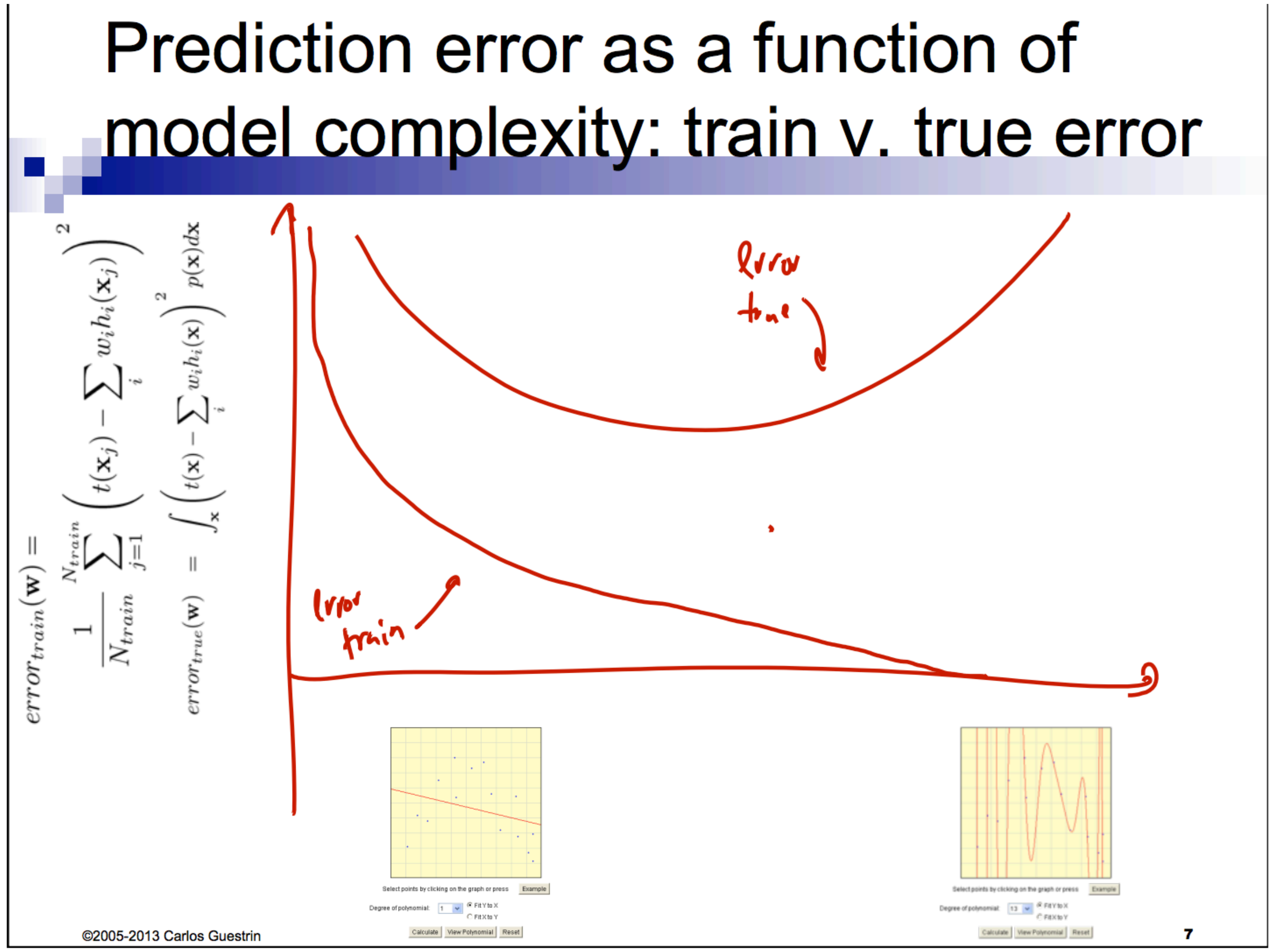
Evaluation
in computational fields

https://www.rosette.com/blog/make-your-choice-its-more-than-a-score-for-evaluating-nlp/

- Computational approaches:
  - Allow for **numeric** evaluation
  - ...and for system **comparison**
  - ...and for feedback on **system changes**
- Computational fields are often **defined** by evaluation
  - what they do is **driven** by evaluation scores on **concrete** datasets

Evaluation
in machine learning

- Machine learning:
  - Algorithm **trains** on labeled data points
  - To evaluate:
    - need **unseen** data points
- **Train**/Dev/**Test split** in datasets
  - Dev: to **tune** various parameters
- Does it ever make sense to evaluate on Train?
  - Yes! But very carefully :)



A picture from Carlos Guestrin's lecture on ML

Evaluating
without a train/test split

5-fold CV

DATASET

Estimation 1	Test	Train	Train	Train	Train
Estimation 2	Train	Test	Train	Train	Train
Estimation 3	Train	Train	Test	Train	Train
Estimation 4	Train	Train	Train	Test	Train
Estimation 5	Train	Train	Train	Train	Test

https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789617740/2/ch02lvl1sec14/k-fold-cross-validation

- Sometimes there isn’t enough data
- Cross-validation:
  - reserve 1 (or a few) data points in each iteration of training
  - at every iteration, the evaluation is then done on small held-out data
- Also:
  - Sometimes you are not really training!
    - e.g. Assignment 2—3, “simplistic prediction”
    - Any training happening there?
      - Does the next prediction depend on the previous ones?)
      - ?

Evaluating
without a train/test split

5-fold CV **DATASET**

Estimation 1	Test	Train	Train	Train	Train
Estimation 2	Train	Test	Train	Train	Train
Estimation 3	Train	Train	Test	Train	Train
Estimation 4	Train	Train	Train	Test	Train
Estimation 5	Train	Train	Train	Train	Test

https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789617740/2/ch02lvl1sec14/k-fold-cross-validation

- Sometimes there isn't enough data
- Cross-validation:
  - reserve 1 (or a few) data points in each iteration of training
  - at every iteration, the evaluation is then done on small held-out data
- Also:
  - Sometimes you are not really training!
    - e.g. Assignment 2—3, "simplistic prediction"
    - Any training happening there?
      - Does the next prediction depend on the previous ones?)
      - No! "Simplistic prediction" is a **symbolic** method (logic)

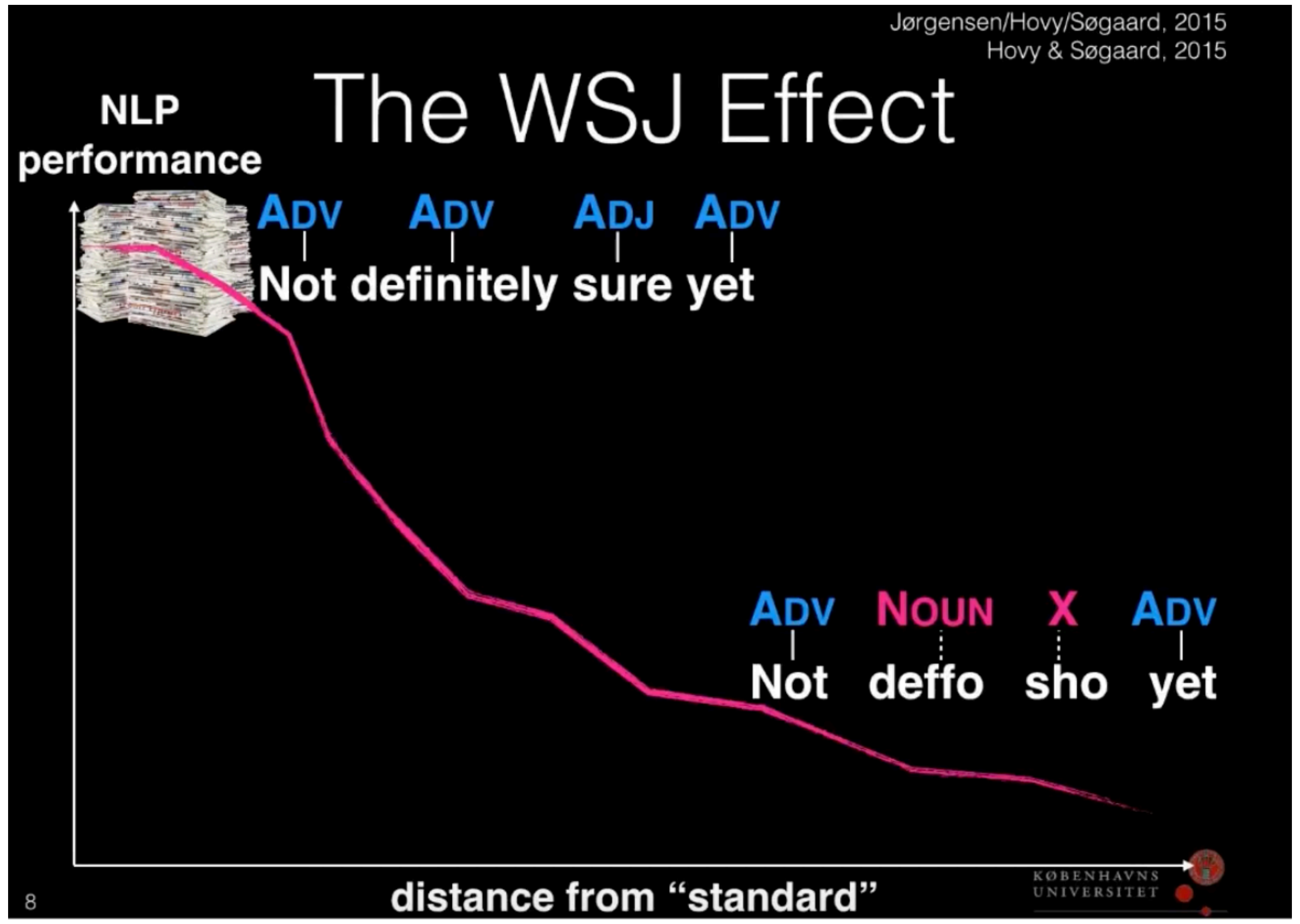
Evaluation
in NLP/data science

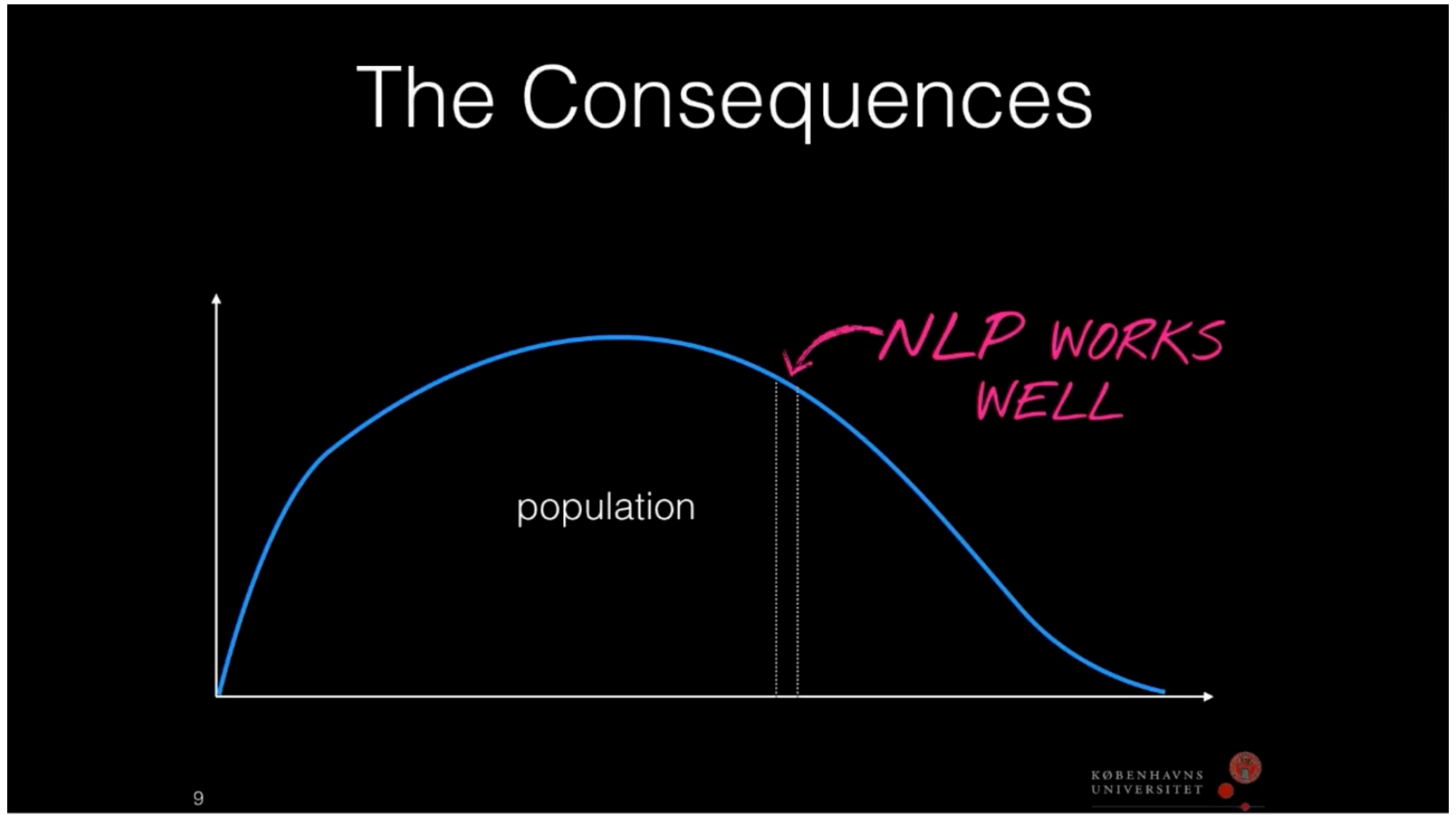
https://www.rosette.com/blog/make-your-choice-its-more-than-a-score-for-evaluating-nlp/

- NLP is defined/driven by evaluation
- Benchmarks:
  - Classic datasets
    - e.g. the Wall Street Journal
  - Systems are compared based on how well they do on the same dataset(s)
- Makes sense?

The WSJ effect

- Years of retraining systems on WSJ
  - enshrined certain **biases** in NLP
  - ...but also, led to systems **adapting** to the **test** portion of WSJ
    - **even** though the **train/test** division in the dataset was observed!
    - So, not only we are biased towards WSJ, but we **aren't even sure** what our **numbers** mean

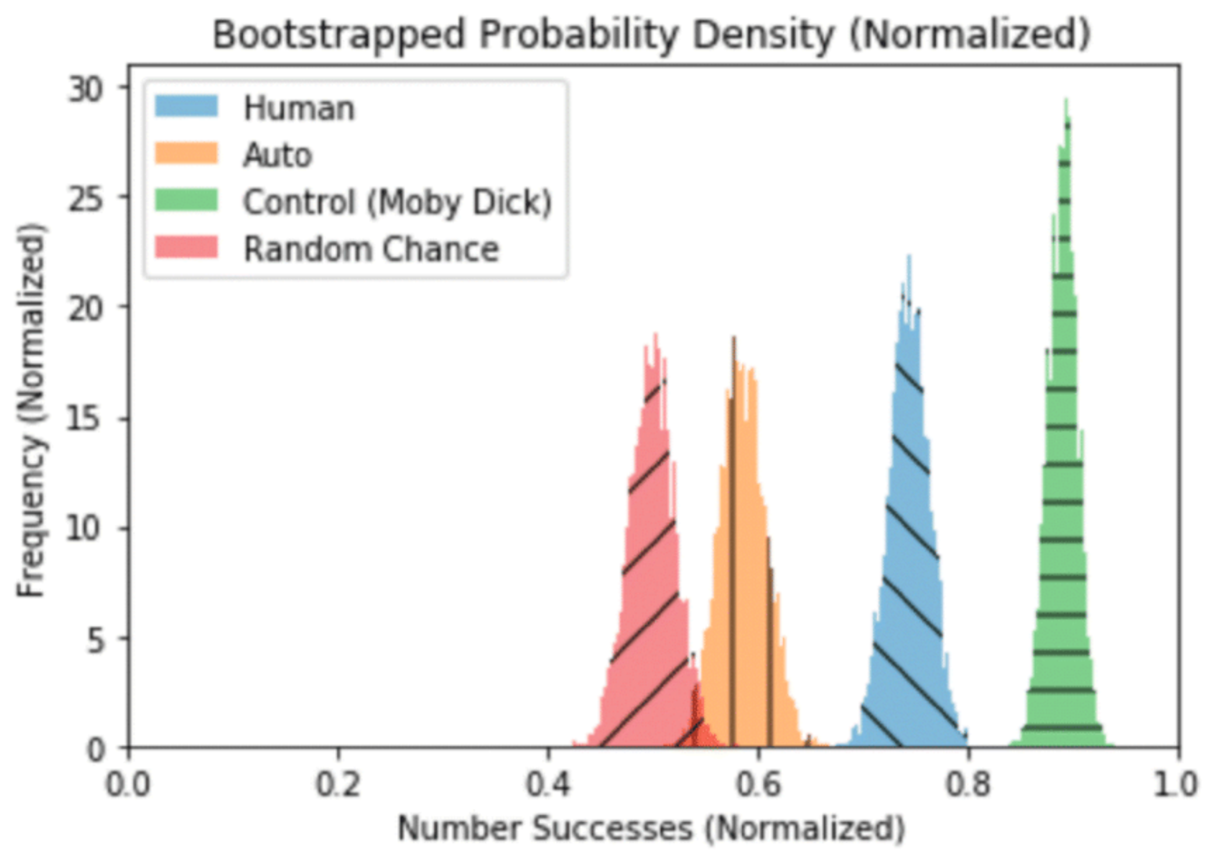




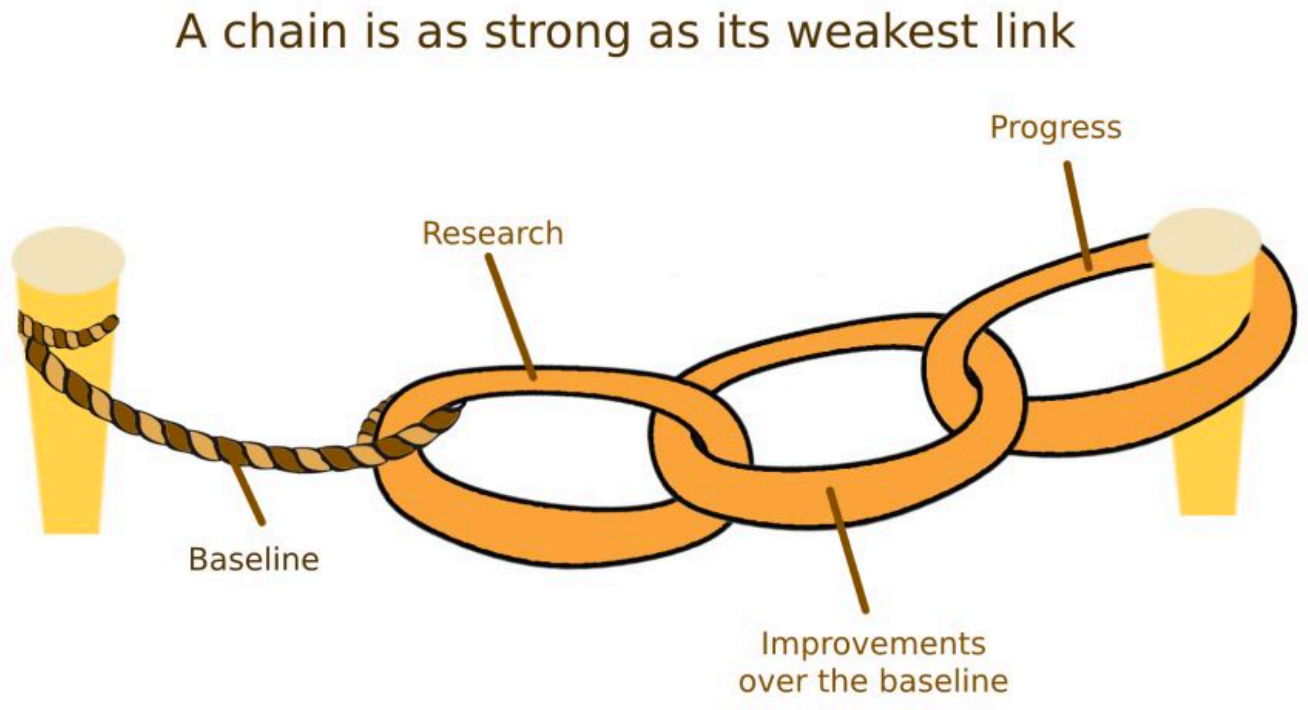
https://hch19.cl.uni-heidelberg.de/program/slides/l/HCH19_lecture_Dirk_Hovy.pdf

Evaluation
Baseline

- A “starting point” for **comparison**
  - What you want to “beat”, in your experiment
  - e.g. 0
  - e.g. random/**chance** performance
  - e.g. most **common** value
    - e.g., predict word order in an unknown language is SOV :).
  - e.g. **least restrictive** value
    - free word order in grammar inference setting
  - e.g. an **older** system/algorithm/model
  - e.g. a **basic** pipeline/architecture
    - then **add** a module to it, see if performance **changes**



https://www.researchgate.net/figure/Bootstrapped-distribution-of-performance-for-cluster-pairings-Human-auto-random_fig4_341893966

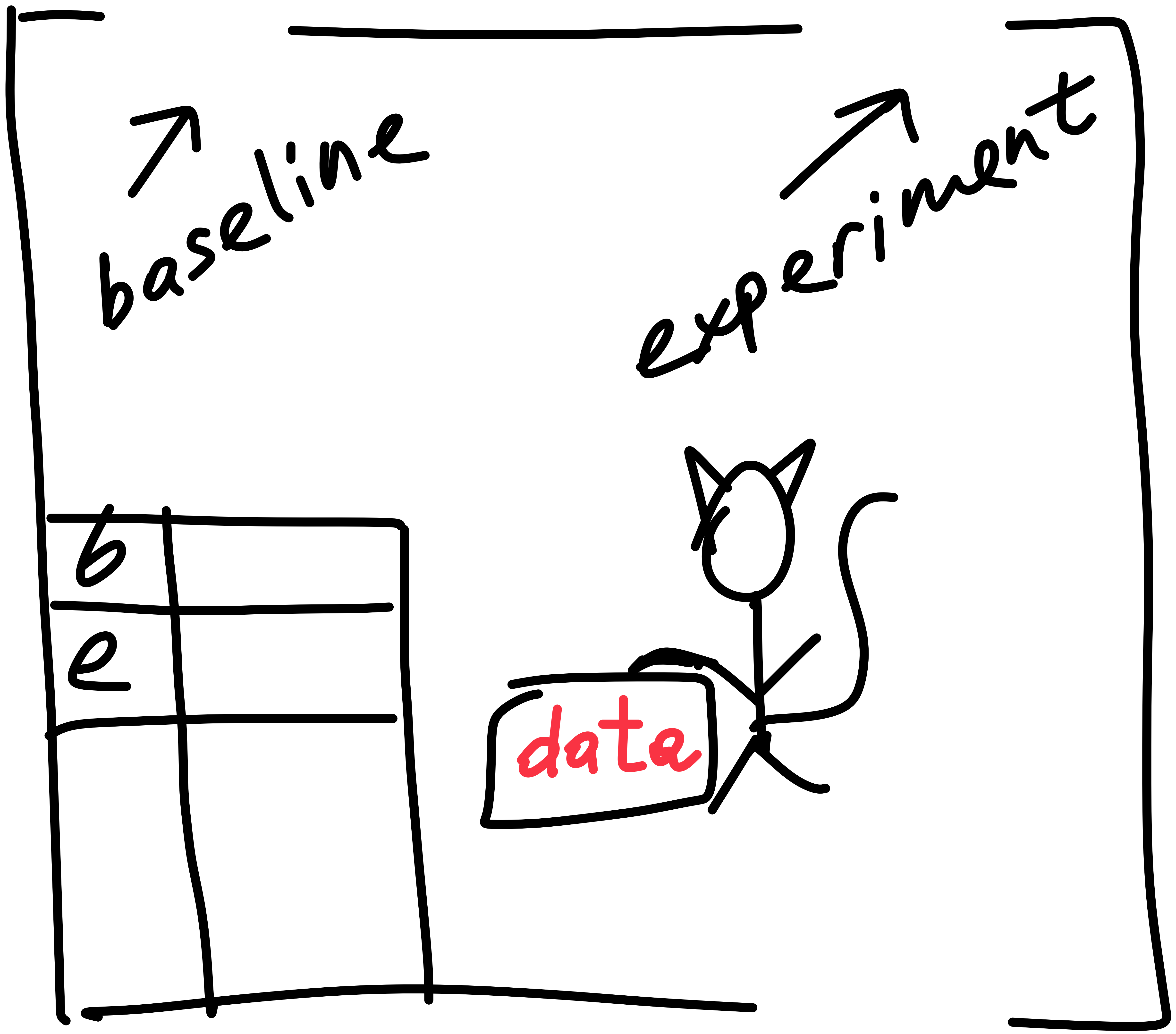


https://blog.ml.cmu.edu/2020/08/31/3-baselines/

Evaluation

Baseline

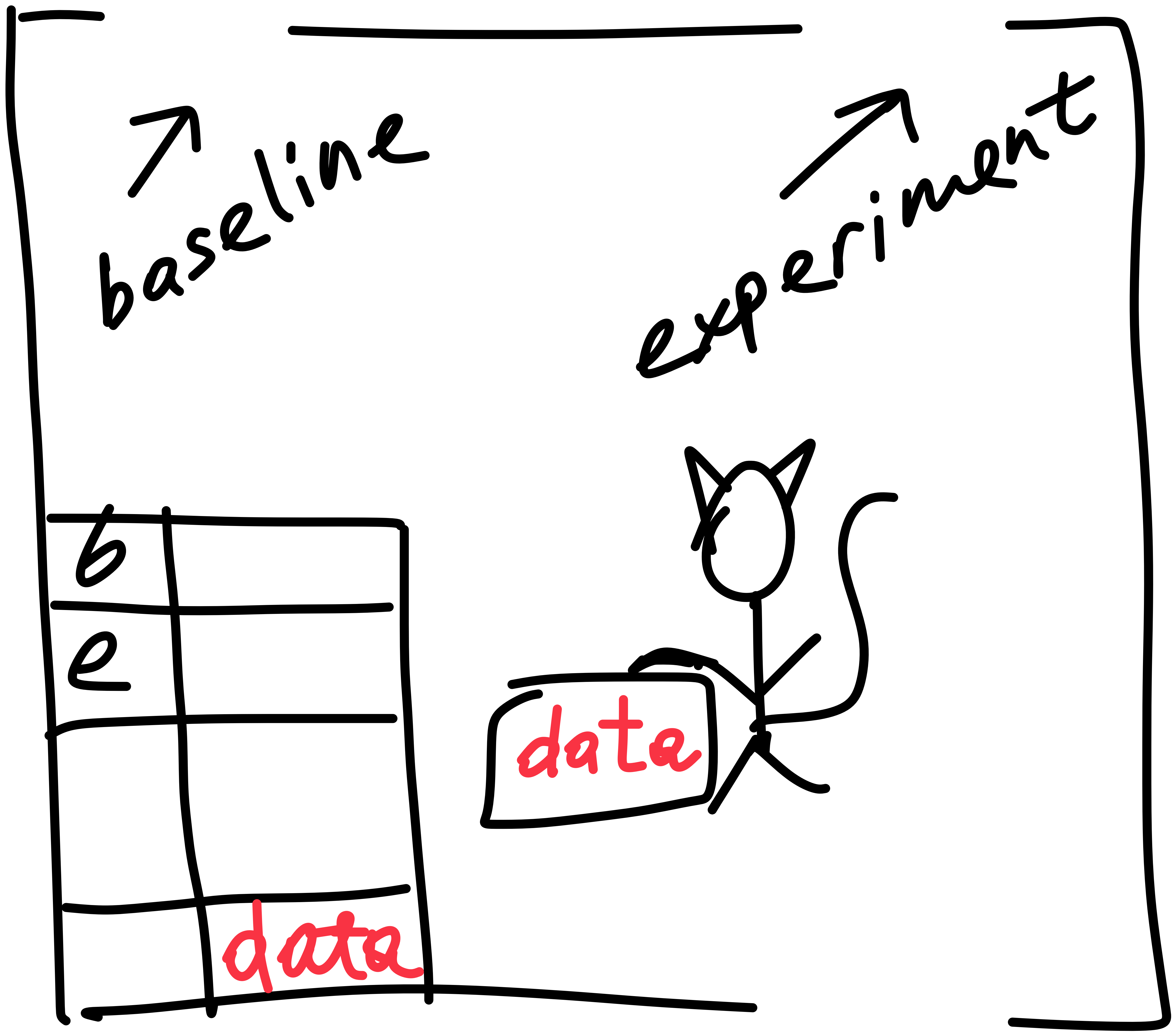
- A "starting point" for comparison
  - What you want to "beat", in your experiment
  - e.g. 0
  - e.g. random/chance performance
  - e.g. most common value
    - e.g., predict word order in an unknown language is SOV :).
  - e.g. an older system/algorithm/model
  - e.g. a basic pipeline/architecture
    - then add a module to it, see if performance changes



Evaluation

Baseline

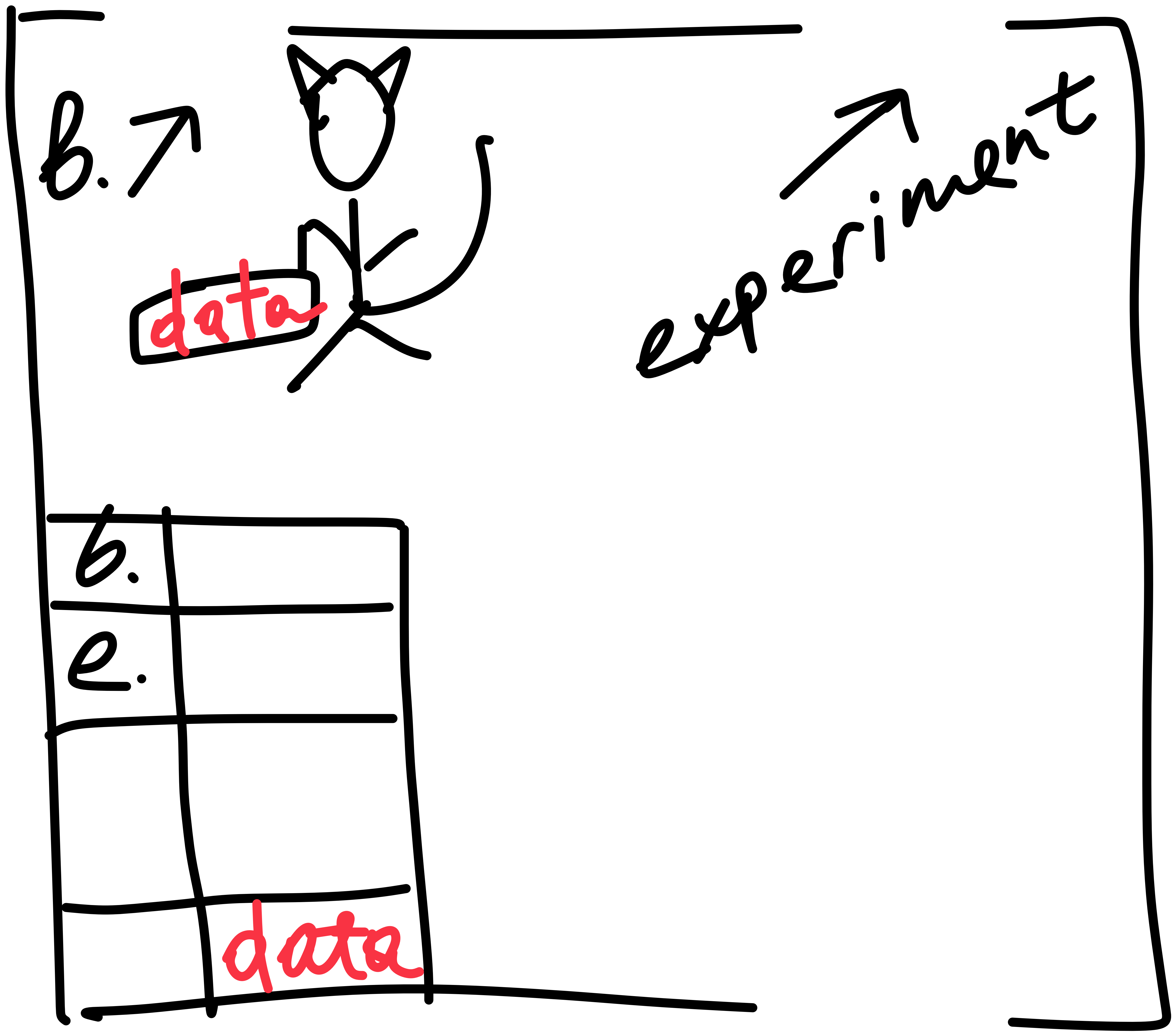
- A "starting point" for comparison
  - What you want to "beat", in your experiment
  - e.g. 0
  - e.g. random/chance performance
  - e.g. most common value
    - e.g., predict word order in an unknown language is SOV :).
  - e.g. an older system/algorithm/model
  - e.g. a basic pipeline/architecture
    - then add a module to it, see if performance changes



Evaluation

Baseline

- A “starting point” for comparison
  - What you want to “beat”, in your experiment
  - e.g. 0
  - e.g. random/chance performance
  - e.g. most common value
    - e.g., predict word order in an unknown language is SOV :).
  - e.g. an older system/algorithm/model
  - e.g. a basic pipeline/architecture
    - then add a module to it, see if performance changes



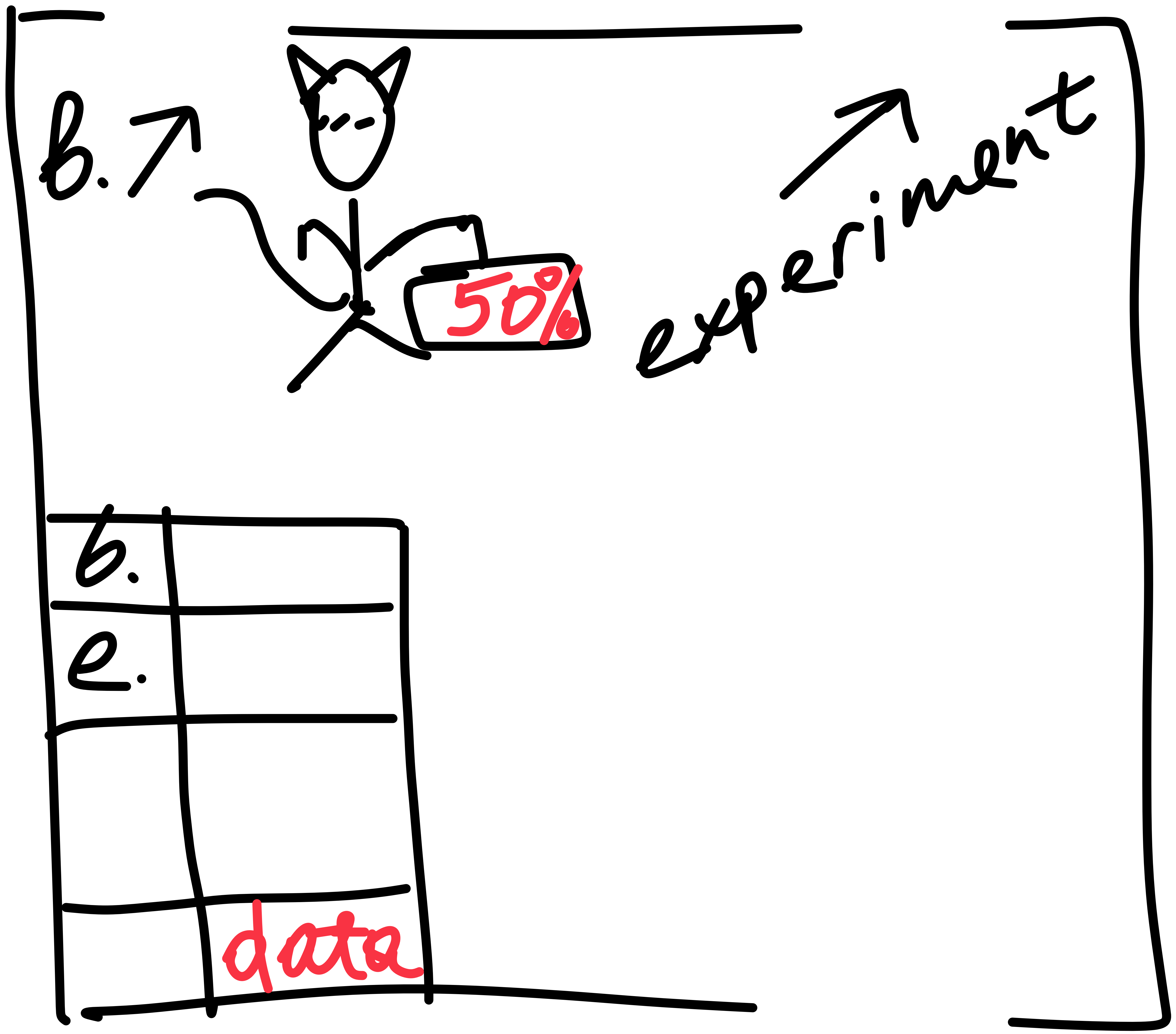
Evaluation
Baseline

- A “starting point” for comparison
  - What you want to “beat”, in your experiment
  - e.g. 0
  - e.g. random/chance performance
  - e.g. most common value
    - e.g., predict word order in an unknown language is SOV :).
  - e.g. an older system/algorithm/model
  - e.g. a basic pipeline/architecture
    - then add a module to it, see if performance changes

Evaluation

Baseline

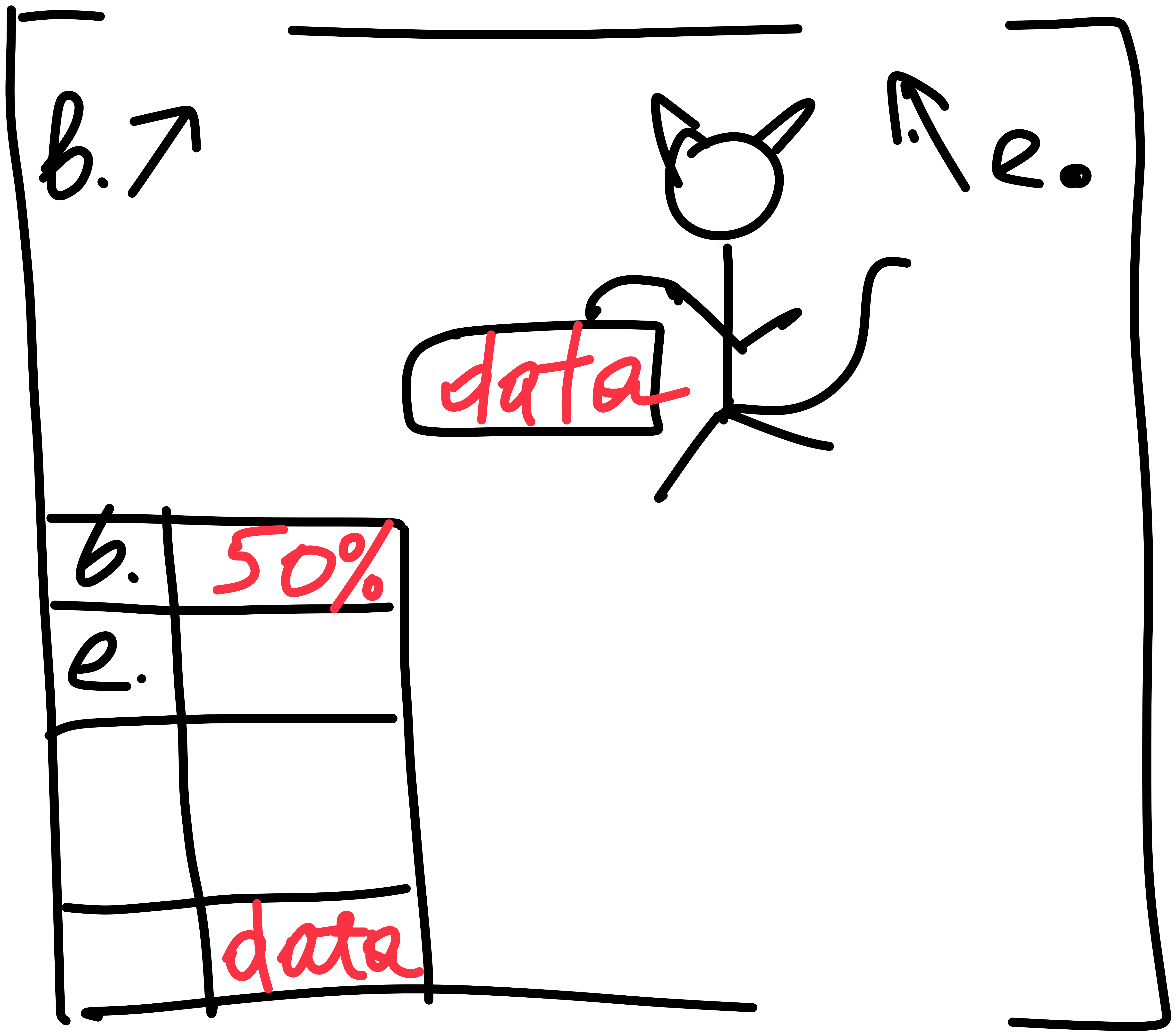
- A "starting point" for comparison
  - What you want to "beat", in your experiment
  - e.g. 0
  - e.g. random/chance performance
  - e.g. most common value
    - e.g., predict word order in an unknown language is SOV :).
  - e.g. an older system/algorithm/model
  - e.g. a basic pipeline/architecture
    - then add a module to it, see if performance changes



Evaluation

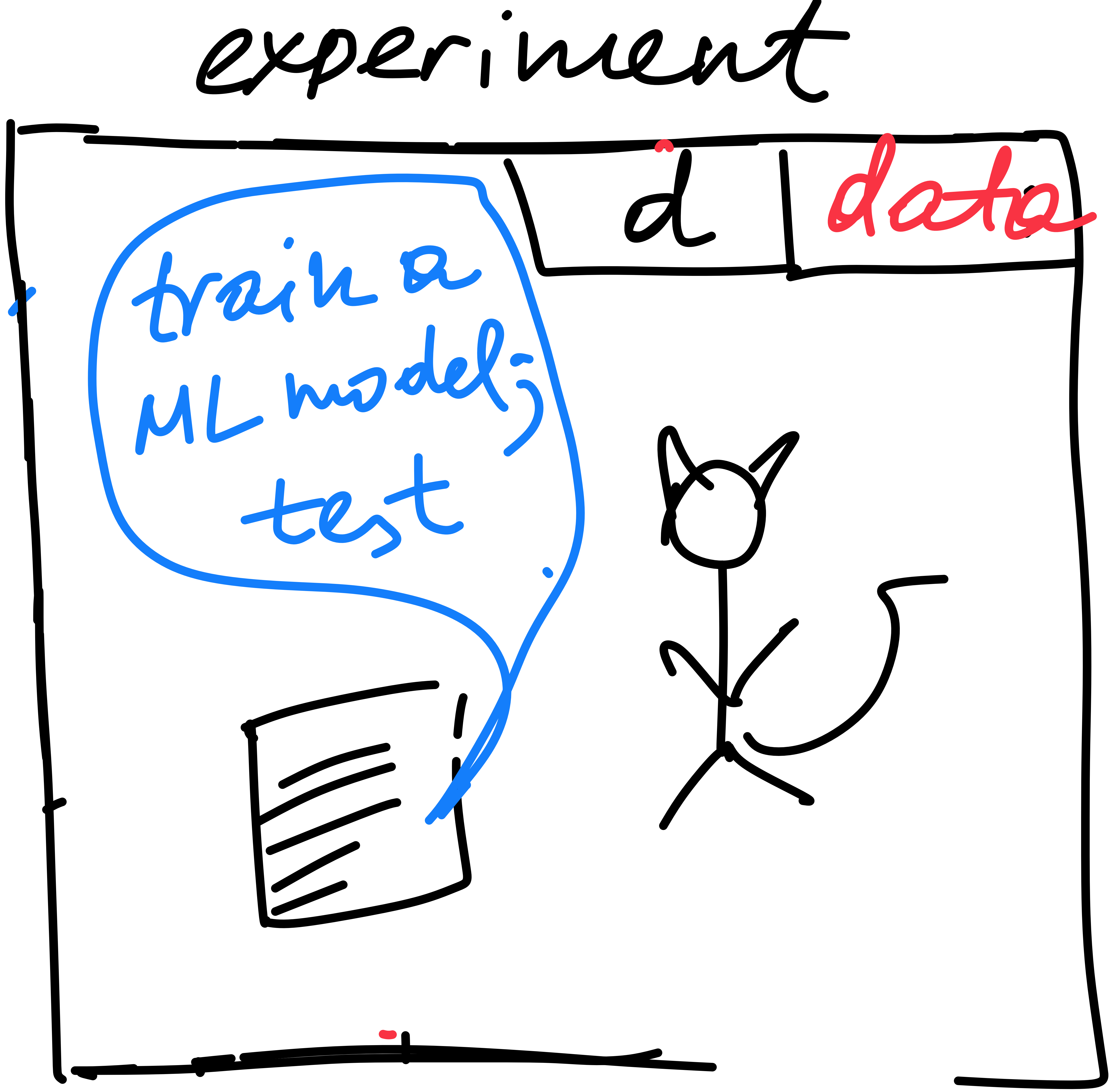
Baseline

- A "starting point" for comparison
  - What you want to "beat", in your experiment
  - e.g. 0
  - e.g. random/chance performance
  - e.g. most common value
    - e.g., predict word order in an unknown language is SOV :).
  - e.g. an older system/algorithm/model
  - e.g. a basic pipeline/architecture
    - then add a module to it, see if performance changes



Evaluation
Baseline

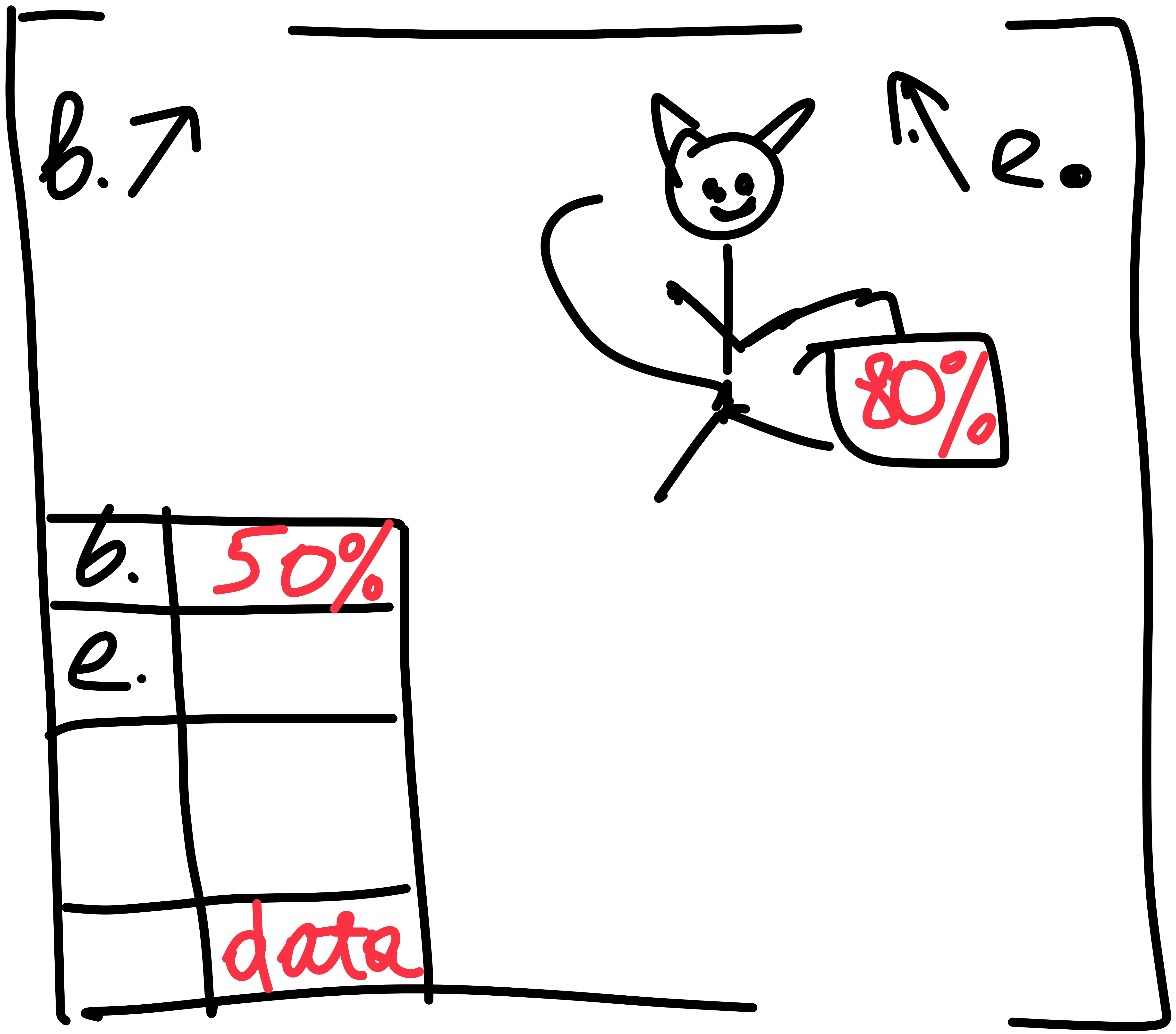
- A "starting point" for comparison
  - What you want to "beat", in your experiment
  - e.g. 0
  - e.g. random/chance performance
  - e.g. most common value
    - e.g., predict word order in an unknown language is SOV :).
  - e.g. an older system/algorithm/model
  - e.g. a basic pipeline/architecture
    - then add a module to it, see if performance changes



Evaluation

Baseline

- A “starting point” for comparison
  - What you want to “beat”, in your experiment
  - e.g. 0
  - e.g. random/chance performance
  - e.g. most common value
    - e.g., predict word order in an unknown language is SOV :).
  - e.g. an older system/algorithm/model
  - e.g. a basic pipeline/architecture
    - then add a module to it, see if performance changes



Gold Standard

“ground truth”

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

https://medium.com/@metalscom/the-history-of-the-gold-standard-in-the-united-states-6556229954e2

Gold Standard

"ground truth"

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

“ground truth”

- Labels treated as correct in evaluation
    - E.g. labels provided with the IMDB dataset

Gold Standard

“ground truth”

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

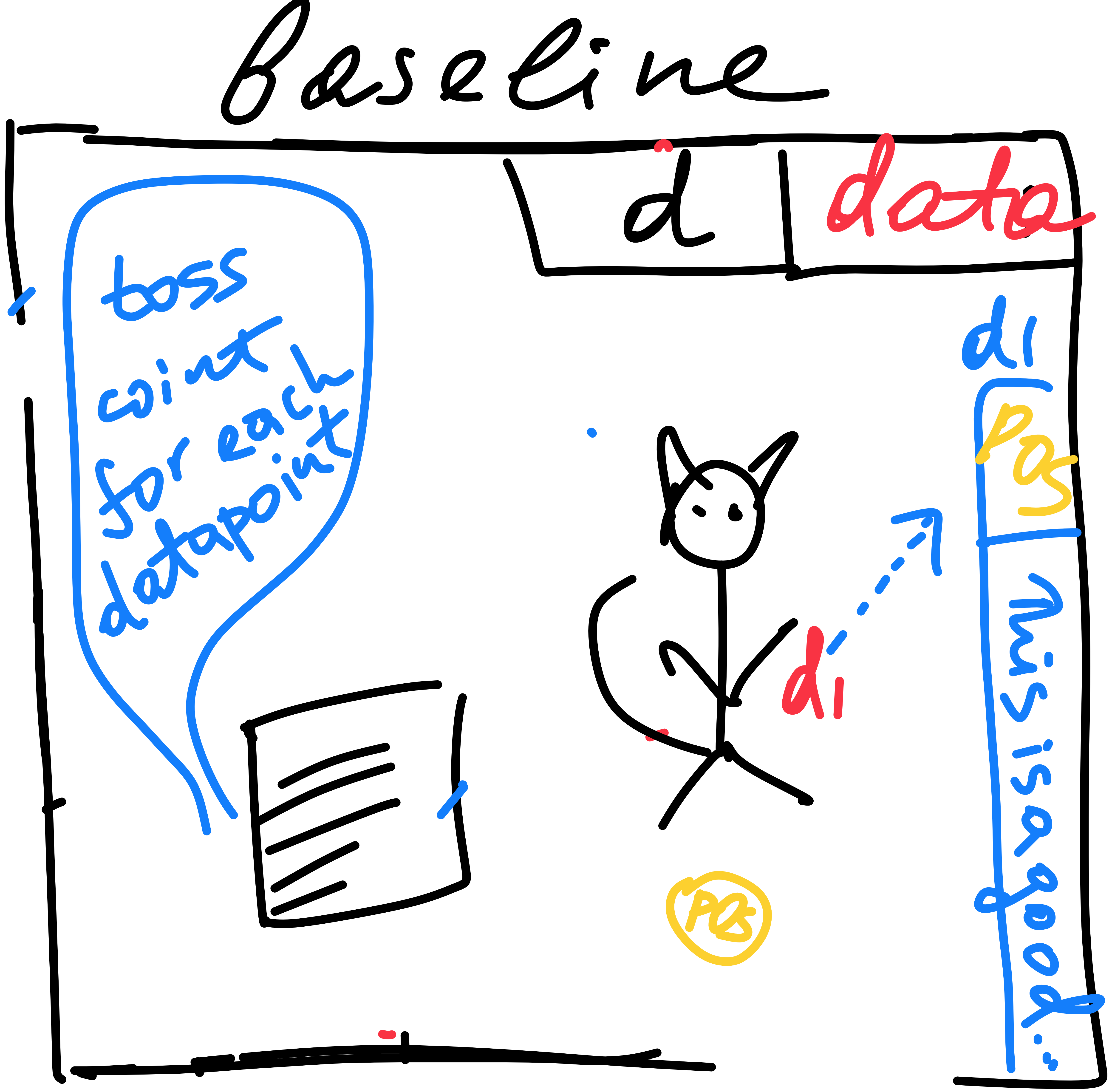
"ground truth"

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

"ground truth"

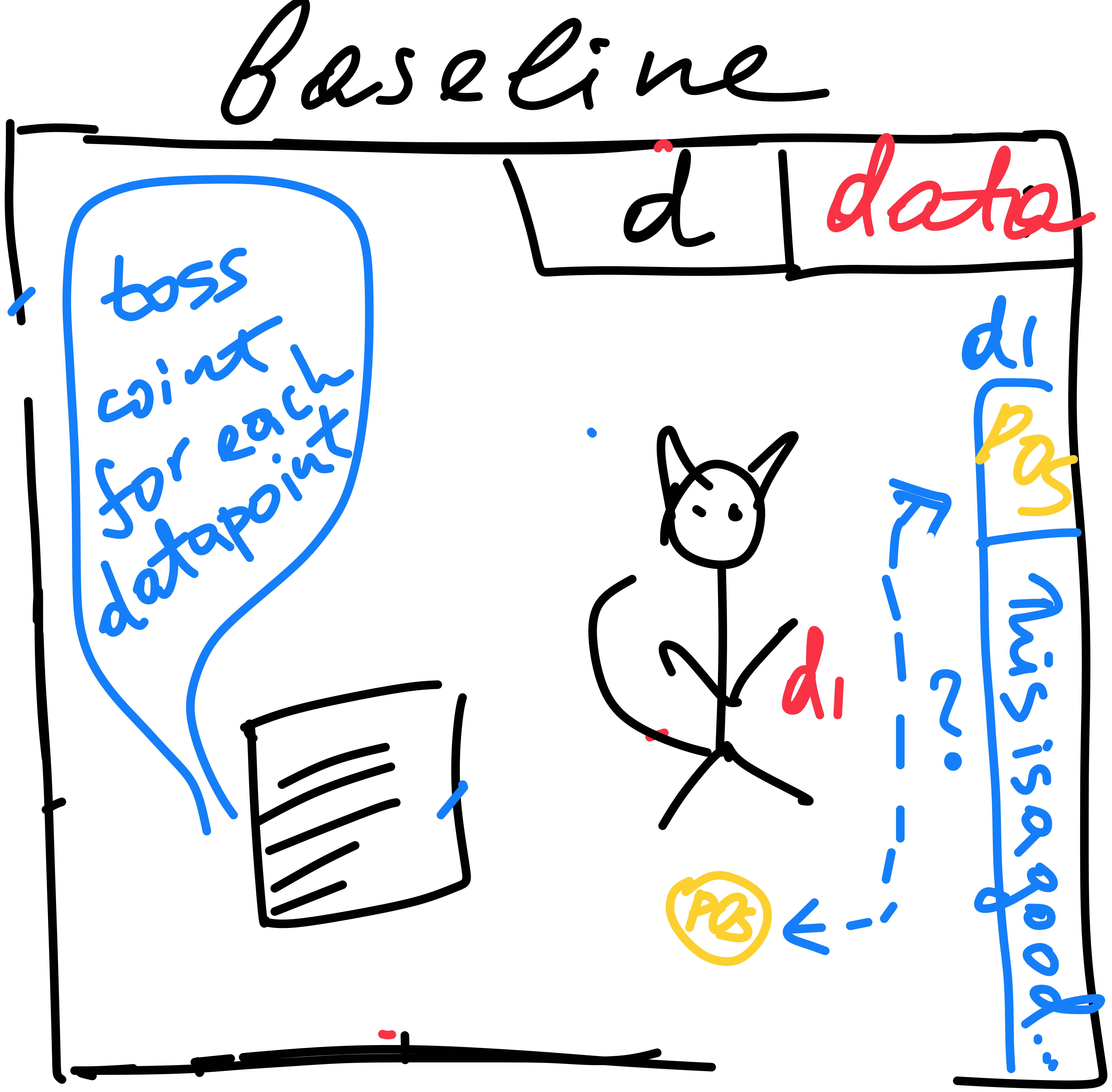
- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset



Gold Standard

“ground truth”

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset



Gold Standard

“ground truth”

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

"ground truth"

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

"ground truth"

- Labels treated as correct in evaluation
    - E.g. labels provided with the IMDB dataset

Gold Standard

"ground truth"

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

"ground truth"

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

"ground truth"

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

“ground truth”

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

“ground truth”

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Gold Standard

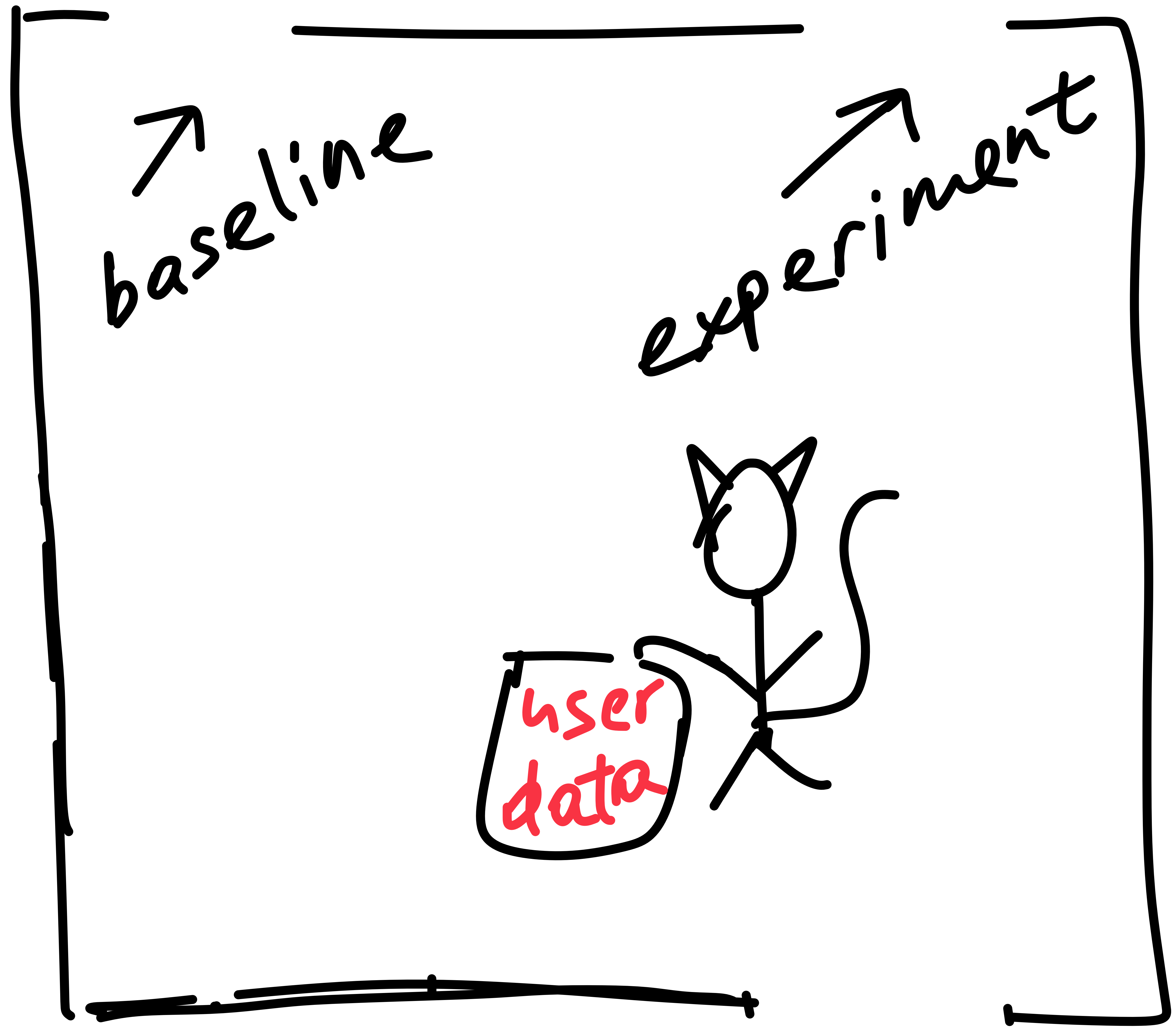
“ground truth”

- Labels treated as correct in evaluation
  - E.g. labels provided with the IMDB dataset

Questions?

Evaluation
Extrinsic v. Intrinsic

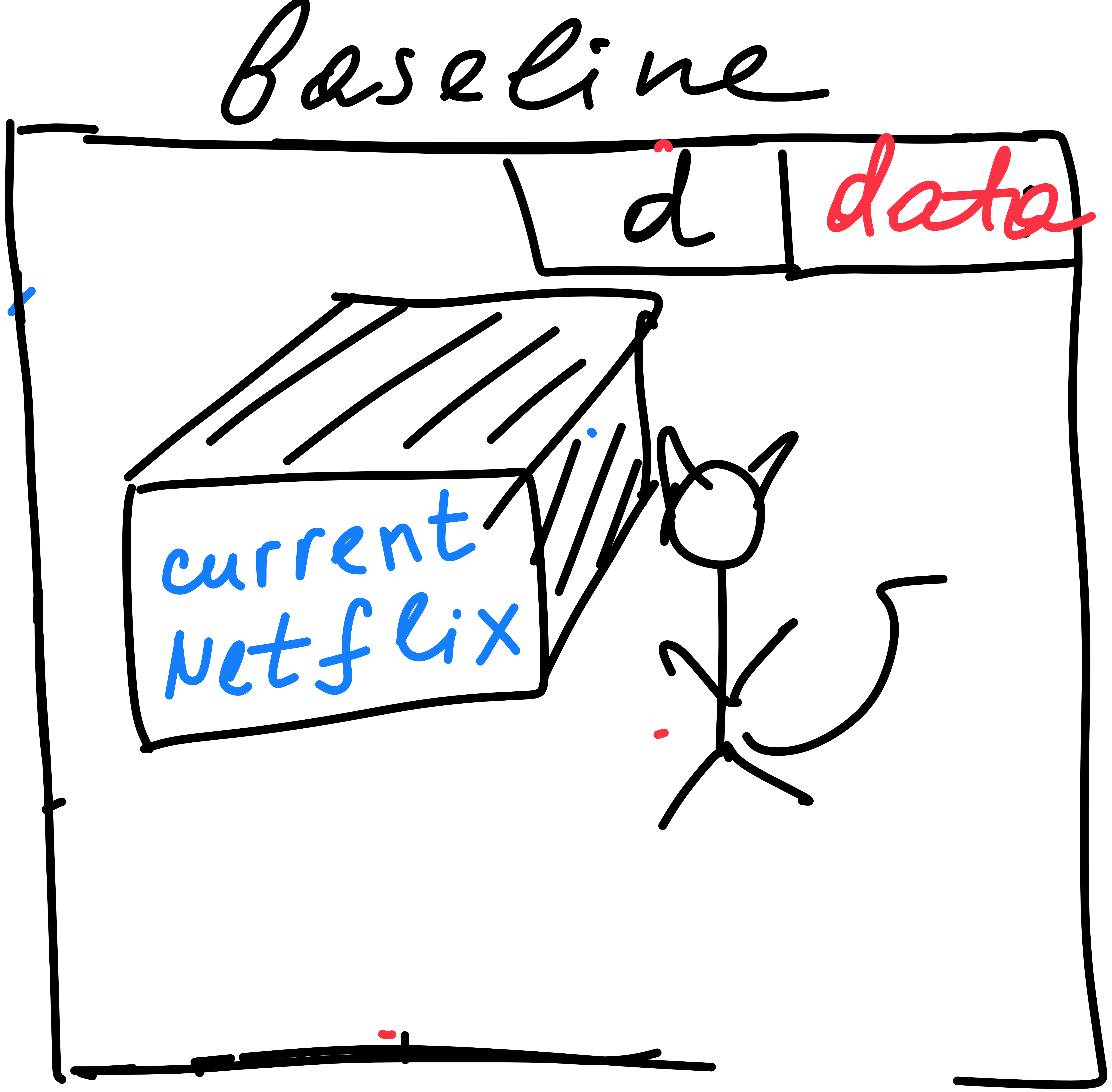
- Intrincic:
  - How well the system does based on its own criteria
    - e.g. How well does our system predict movie review labels?
- Extrinsic:
  - Does the system improve the performance of some other system down the pipeline?
    - e.g.: With our system added, does another system which makes movie suggestions lead to more users clicking on/watching the suggestions?



Evaluation

Extrinsic v. Intrinsic

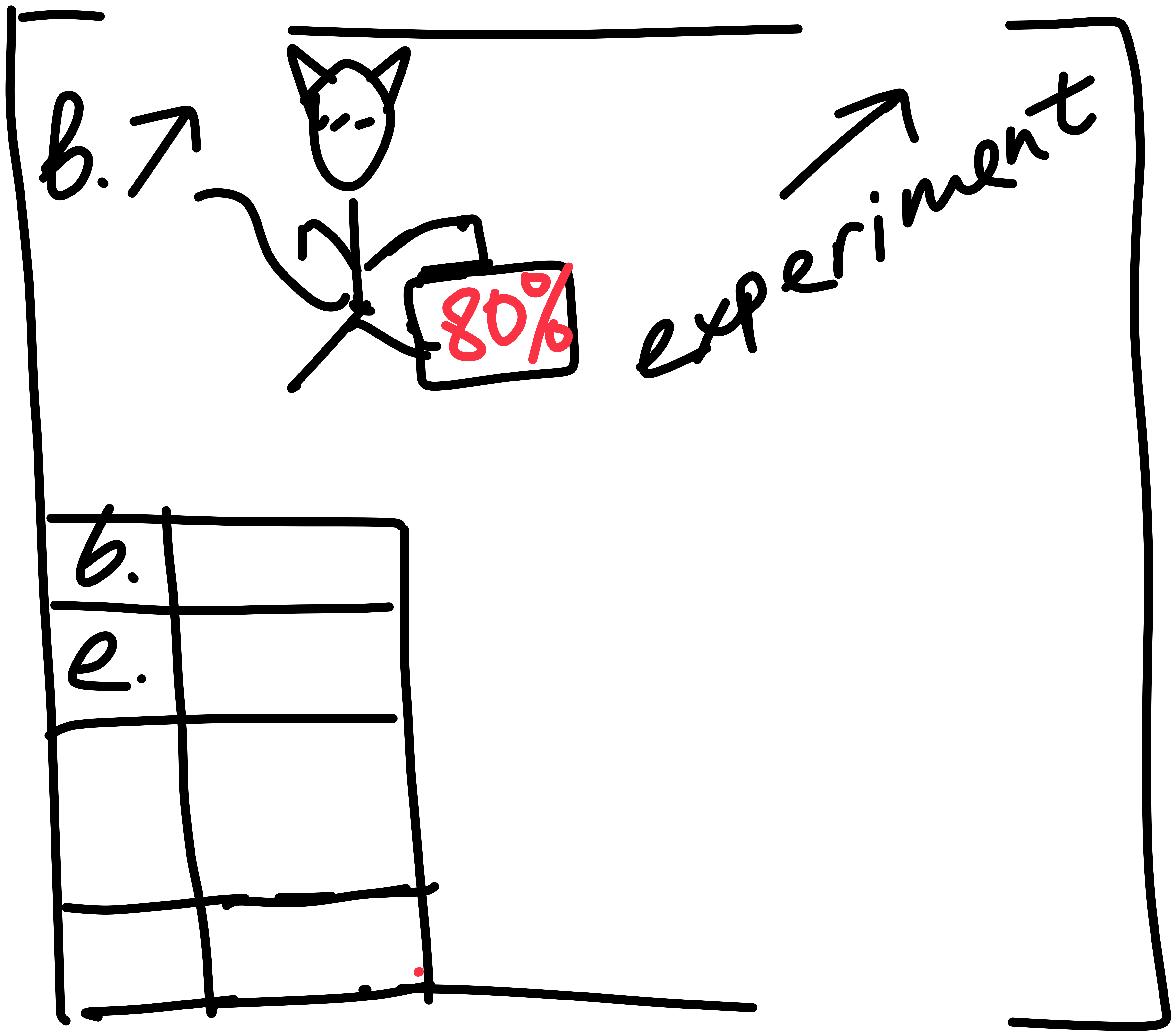
- Intrincic:
  - How well the system does based on its own criteria
    - e.g. How well does our system predict movie review labels?
- Extrinsic:
  - Does the system improve the performance of some other system down the pipeline?
    - e.g.: With our system added, does another system which makes movie suggestions lead to more users clicking on/watching the suggestions?



Evaluation

Extrinsic v. Intrinsic

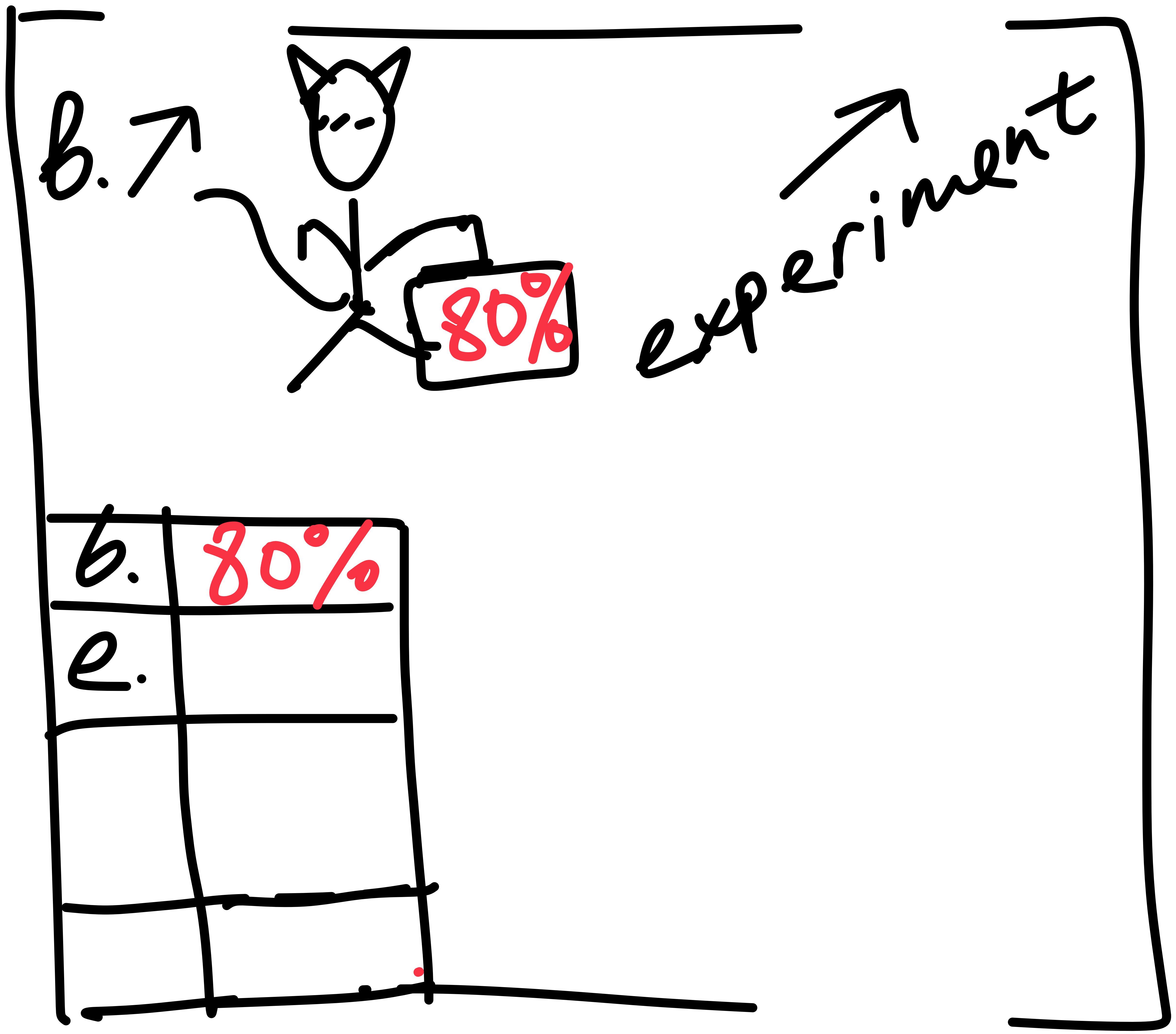
- Intrincic:
  - How well the system does based on its own criteria
    - e.g. How well does our system predict movie review labels?
- Extrinsic:
  - Does the system improve the performance of some other system down the pipeline?
    - e.g.: With our system added, does another system which makes movie suggestions lead to more users clicking on/watching the suggestions?



Evaluation

Extrinsic v. Intrinsic

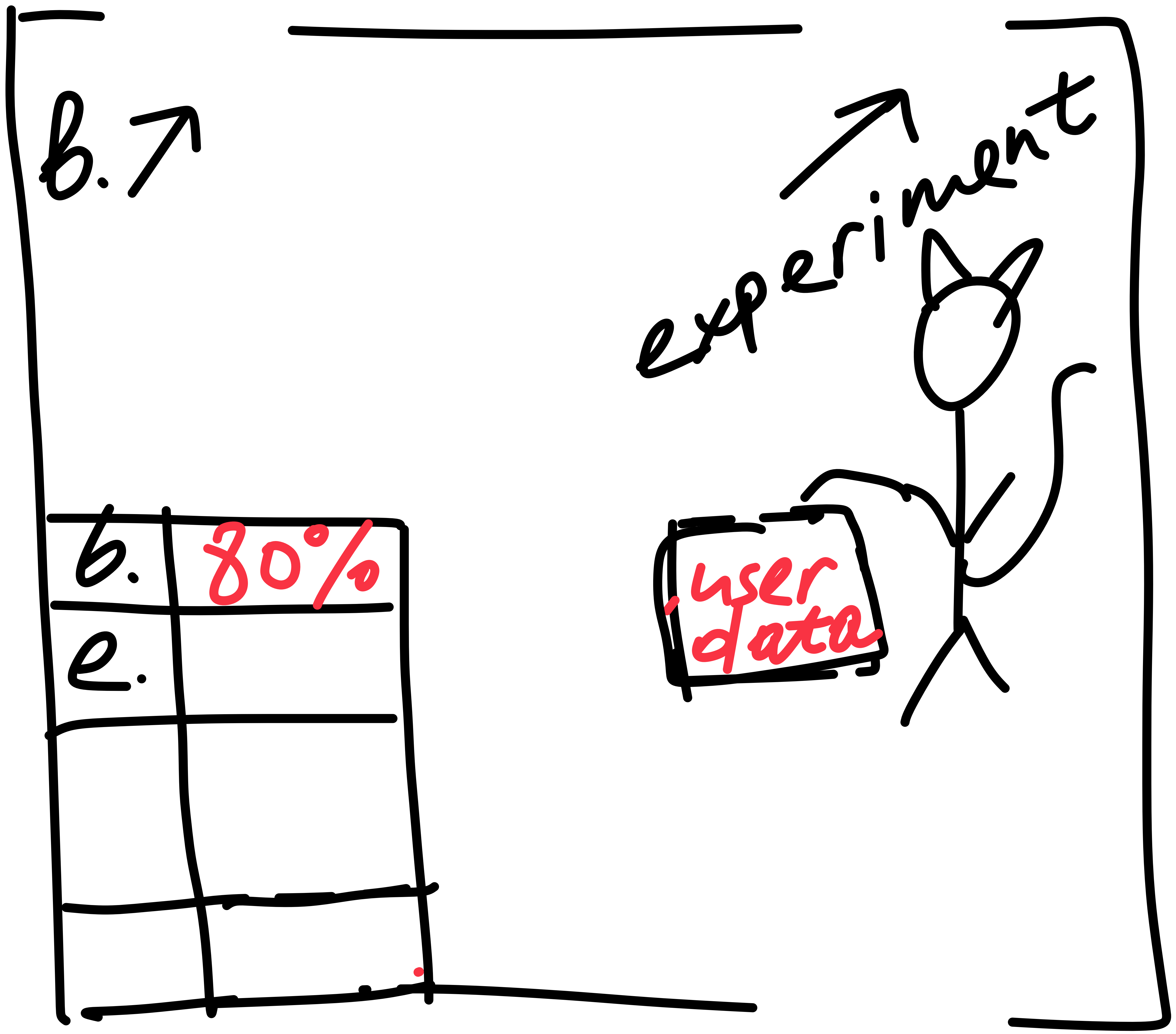
- Intrincic:
  - How well the system does based on its own criteria
    - e.g. How well does our system predict movie review labels?
- Extrinsic:
  - Does the system improve the performance of some other system down the pipeline?
    - e.g.: With our system added, does another system which makes movie suggestions lead to more users clicking on/watching the suggestions?



Evaluation

Extrinsic v. Intrinsic

- Intrincic:
  - How well the system does based on its own criteria
    - e.g. How well does our system predict movie review labels?
- Extrinsic:
  - Does the system improve the performance of some other system down the pipeline?
    - e.g.: With our system added, does another system which makes movie suggestions lead to more users clicking on/watching the suggestions?



Evaluation

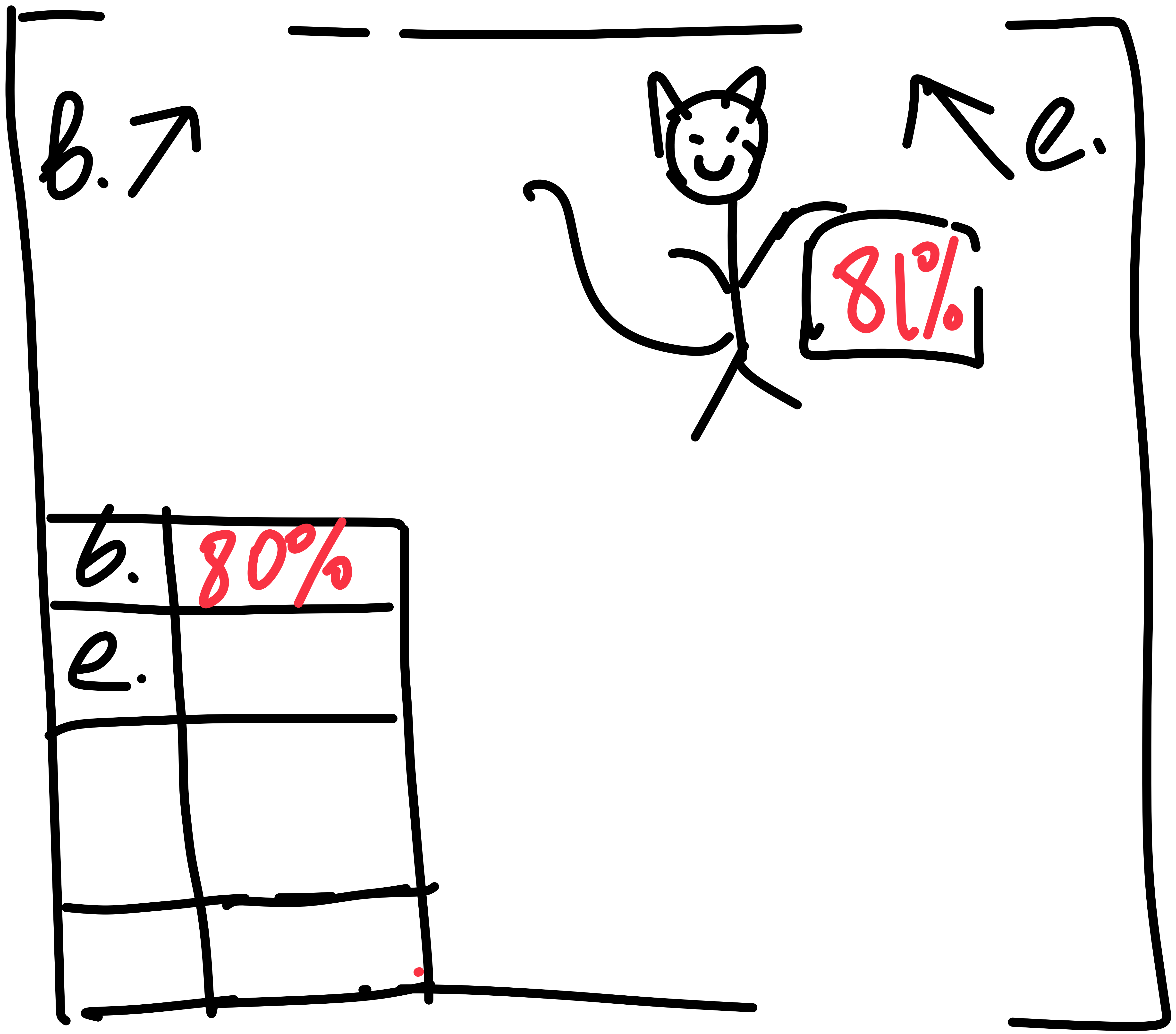
Extrinsic v. Intrinsic

- Intrincic:
  - How well the system does based on its own criteria
    - e.g. How well does our system predict movie review labels?
- Extrinsic:
  - Does the system improve the performance of some other system down the pipeline?
    - e.g.: With our system added, does another system which makes movie suggestions lead to more users clicking on/watching the suggestions?

Evaluation

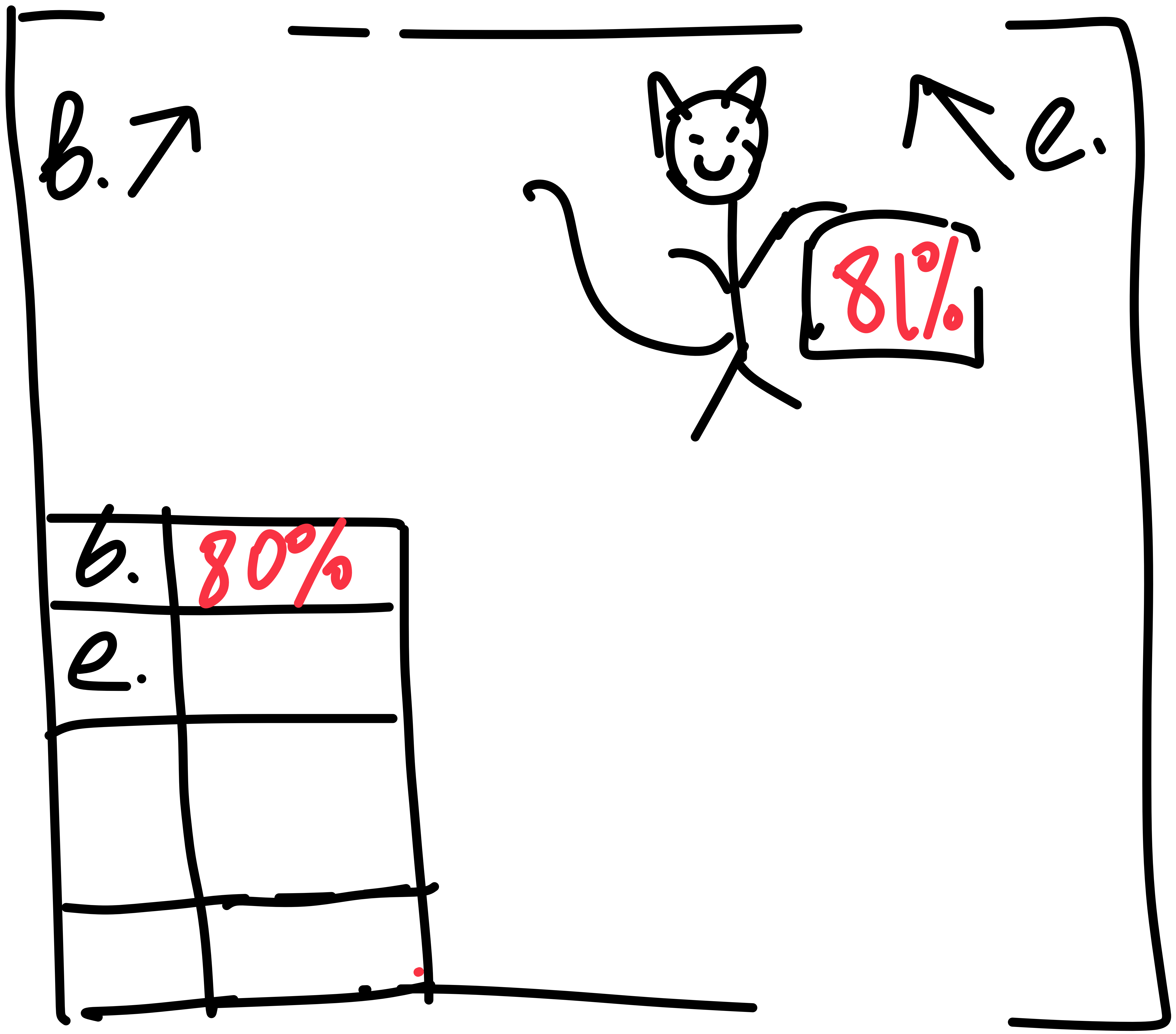
Extrinsic v. Intrinsic

- Intrincic:
  - How well the system does based on its own criteria
    - e.g. How well does our system predict movie review labels?
- Extrinsic:
  - Does the system improve the performance of some other system down the pipeline?
    - e.g.: With our system added, does another system which makes movie suggestions lead to more users clicking on/watching the suggestions?



Evaluation
drives NLP

- ...is to say:
  - people are happy about **incremental** improvements
  - ...and they **design** experiments so as to obtain those
  - ...and they sometimes worry **less** about whether the numbers are **meaningful**
- **Data science** tries to **make sense** of the numbers



Please consider filling out the survey:
https://canvas.uw.edu/courses/1465777/quizzes/1435948