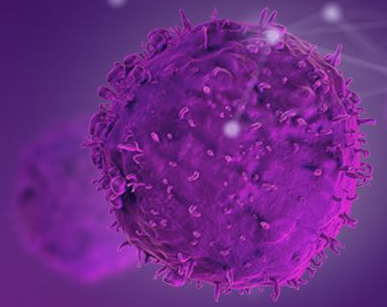
ardigen

Artificial Intelligence
& Bioinformatics
for Precision Medicine



Hand in hand with weak supervision using snorkel

Szymon Wojciechowski

snorkel?

My team

ardigen

Kaja Milanowska, PhD
Executive VP

Jan Majta
Bioinformatician

Sonia Wróbel
Bioinformatician

Vladyslav Hubar
Bioinformatics Intern

Krzysztof Odrzywołek
Data Scientist II

Szymon Wojciechowski
Data Scientist II

Biology & Bioinformatics

Artificial Intelligence & Data Science

Monika Brzychczy-Włoch, PhD
Senior Microbiologist

Emilia Strycharz-Angrecka
Laboratory QA Manager

Karol Horosin
Software Engineer

Laboratory

Software Engineering

Outline of the problem

Getting **useful** information

ardigen

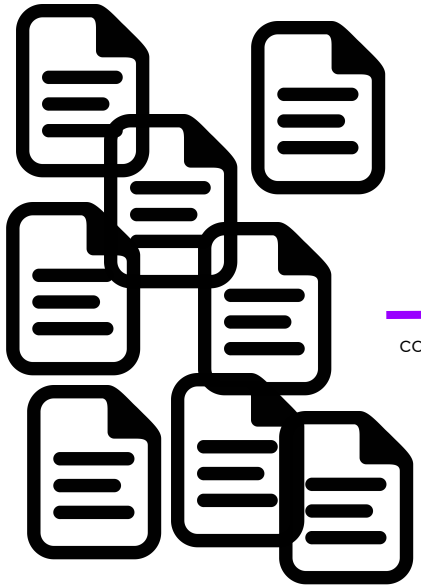
comprises information on

resistant / susceptible to

Document **Bacterium** **Antibiotic**

Outline of the problem

Getting **useful** information



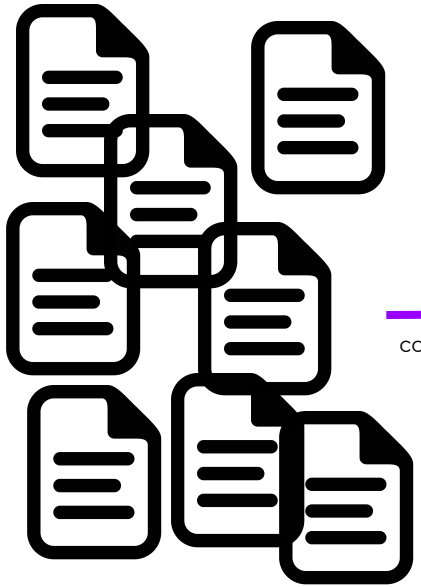
comprise information on

resistant / susceptible to

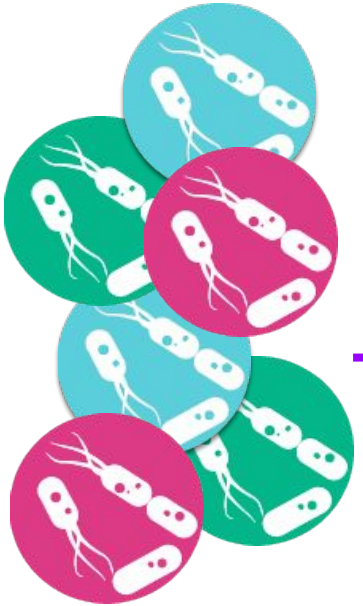
Documents **Bacterium** **Antibiotic**

Outline of the problem

Getting **useful** information



comprise information on



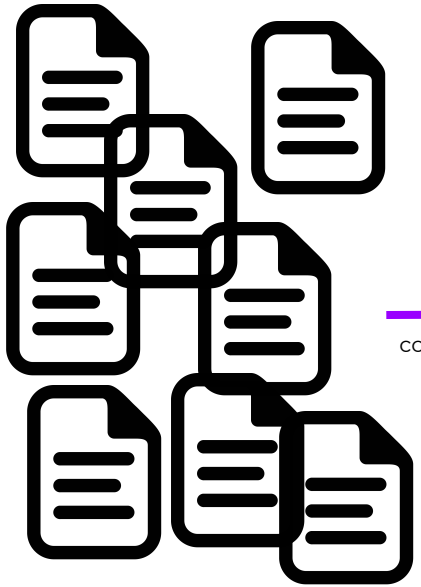
resistant / susceptible to

Documents **Bacteria** **Antibiotic**

Outline of the problem

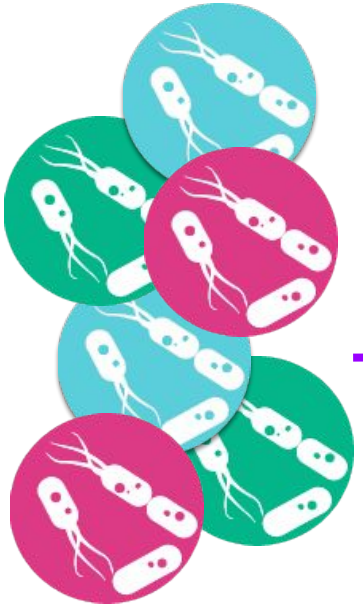
Getting **useful** information

ardigen



Documents

comprise information on



Bacteria

resistant / susceptible to

Antibiotics

Putting it into perspective...

ardigen

~150’000 x

Document

~10’000 x

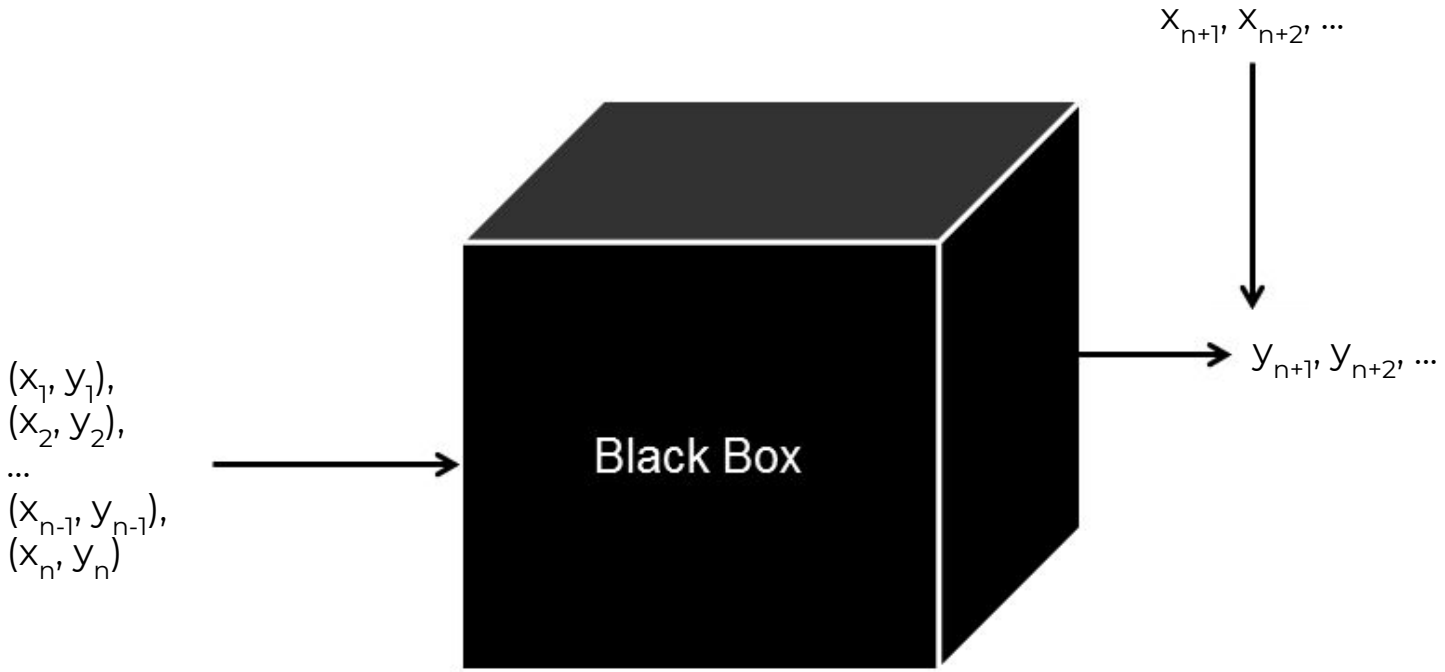
Bacterium

~120 x

Antibiotic

Let’s use deep / machine learning / AI!

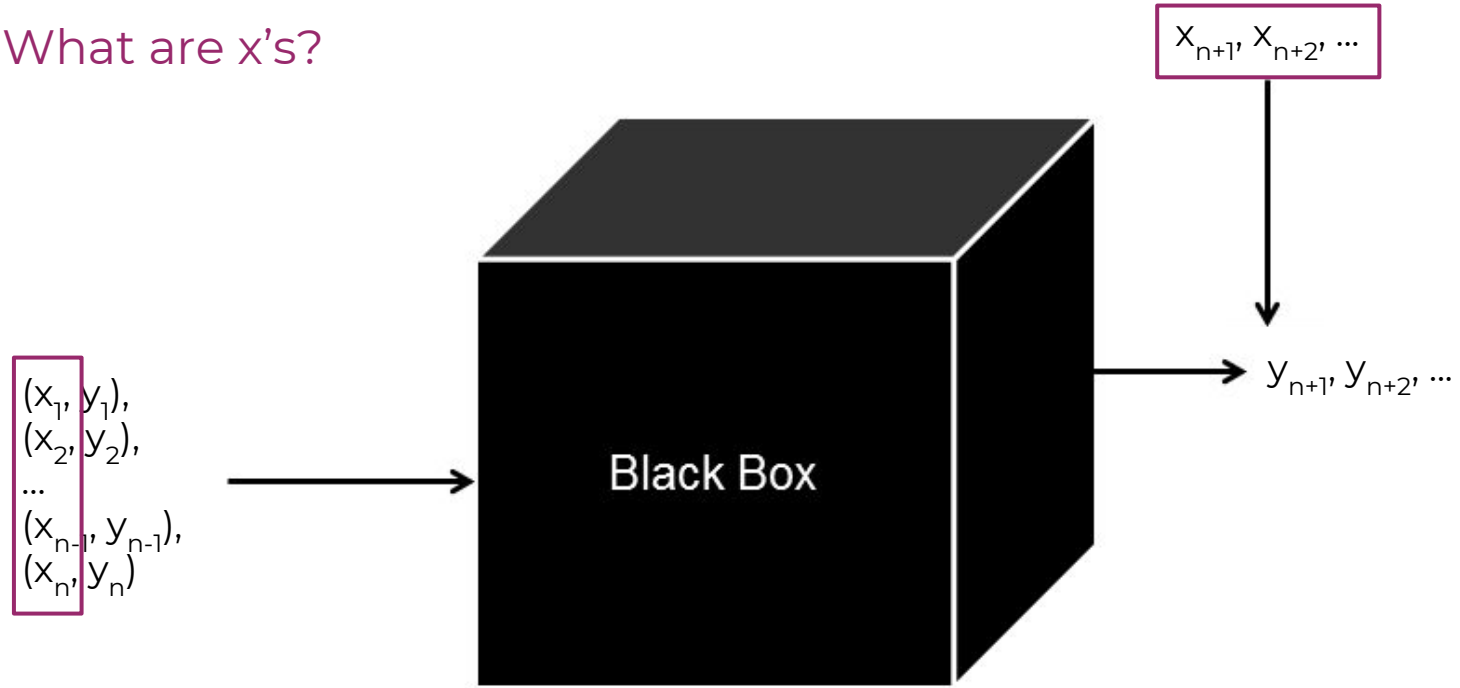
But…



Let's use deep / machine learning / AI!

But...

What are x's?

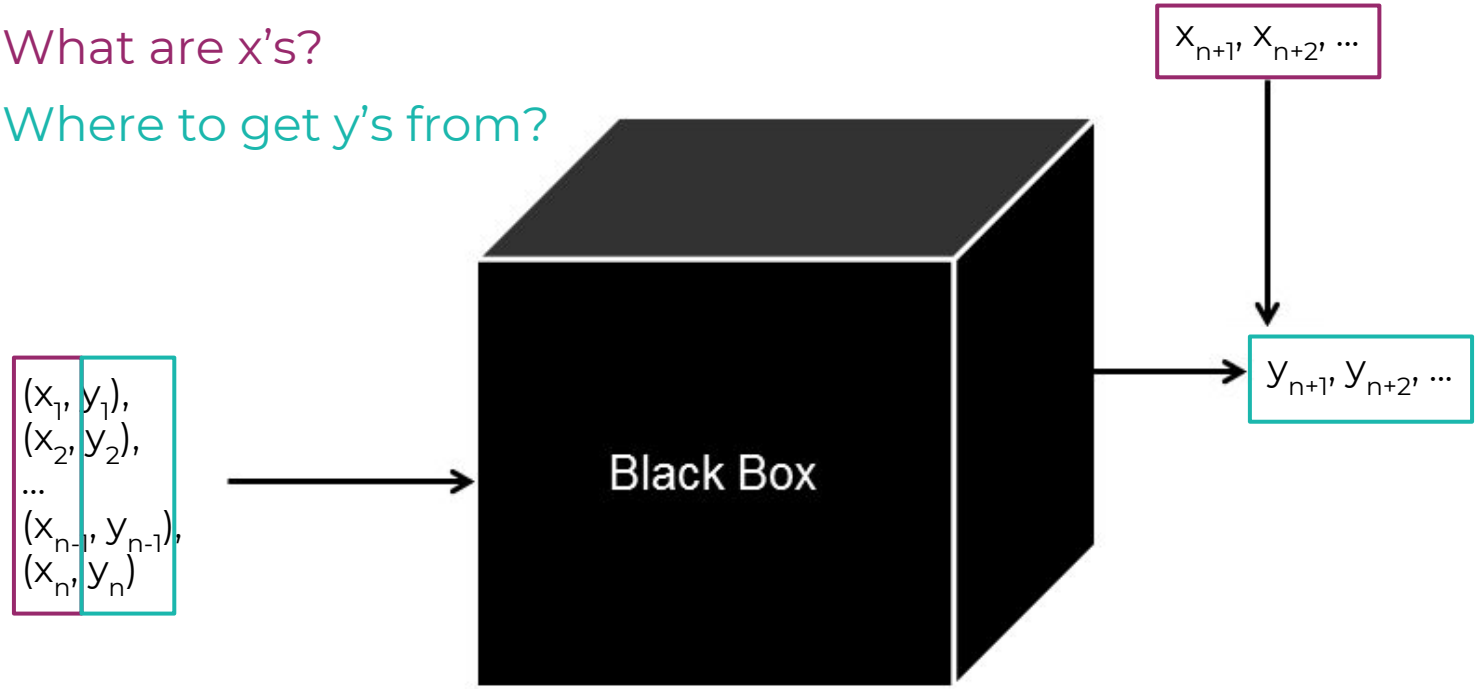


Let's use deep / machine learning / AI!

But...

What are x's?

Where to get y's from?



In the search for x's

Our aim was to evaluate the behavior of *Lactobacillus rhamnosus* GGATCC 53103, a well-known probiotic microorganism, during exposure to erythromycin, tetracycline, amoxicillin/clavulanate and ciprofloxacin. *

This is particularly worrisome as ceftriaxone is the last remaining option for empirical first-line treatment of gonorrhea. *N. gonorrhoeae* now seems to be evolving into a true superbug and, in the near future, gonorrhea may become untreatable in certain circumstances. **

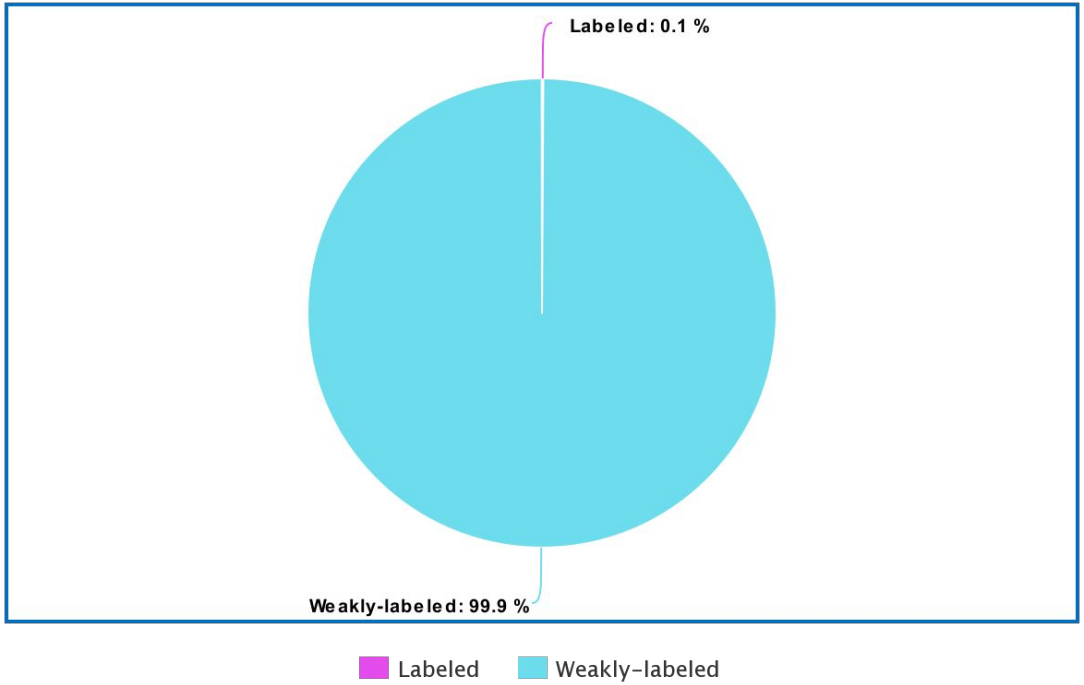
Among strains of *S. xylosus*, the incidence of resistance ranged from 22% for tetracycline up to 69% for penicillin. ***

* Drago L., Rodighiero V., Mattina R., Toscano M. & De Vecchi E. (2011) In Vitro Selection of Antibiotic Resistance in the Probiotic Strain Lactobacillus rhamnosus GG ATCC 53103, Journal of Chemotherapy, 23:4, 211-215, DOI: 10.1179/joc.2011.23.4.211

** Unemo M., Shafer WM. (2011) Antibiotic resistance in Neisseria gonorrhoeae: origin, evolution, and lessons learned for the future, Ann N Y Acad Sci, 2011;1230:E19–28.

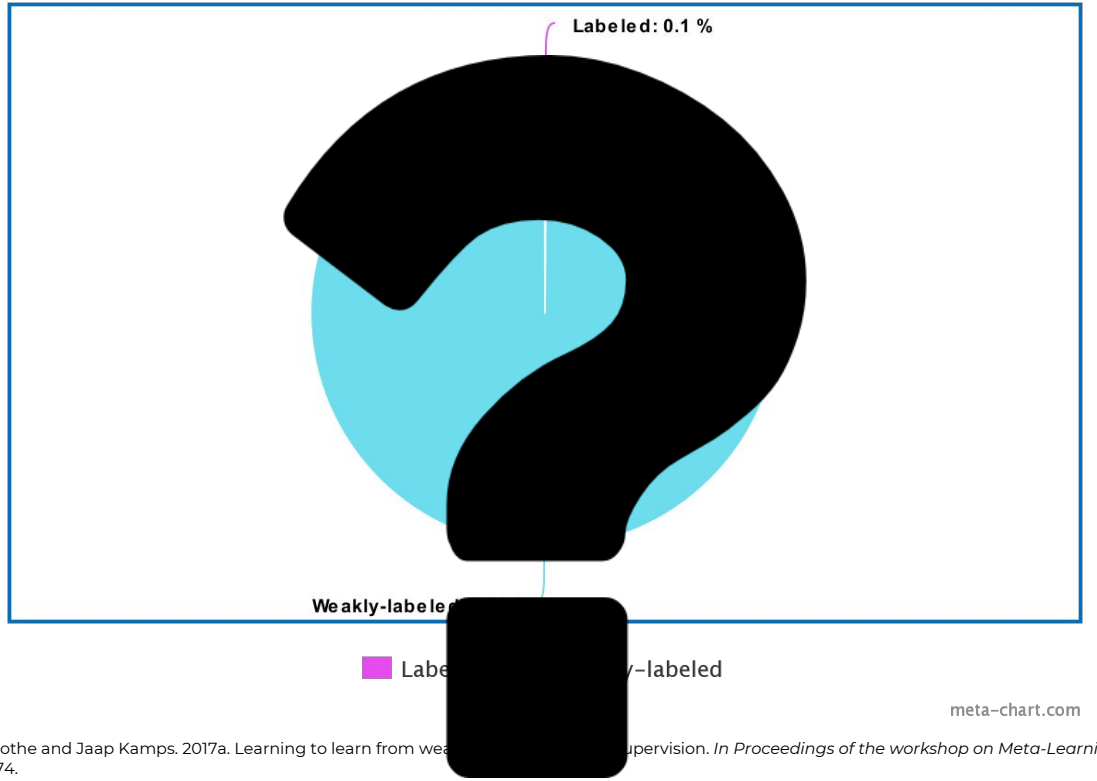
*** Resch M., Nagel V. & Hertel C. (2008) Antibiotic resistance of coagulase-negative staphylococci associated with food and used in starter cultures, *International Journal of Food Microbiology*, 127:1-2, 99-104, DOI: https://doi.org/10.1016/j.ijfoodmicro.2008.06.013

y's - a creative mixture



Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe and Jaap Kamps. 2017a. Learning to learn from weak supervision by full supervision. *In Proceedings of the workshop on Meta-Learning at Advances in Neural Information Processing Systems 31* (NIPS 2017), pages 65–74.
Shnarch, E., Alzate, C., Dankin, L., Gleize, M., Hou, Y., Choshen, L., ... & Slonim, N. (2018). Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (Vol. 2, pp. 599-605).

y's - a creative mixture



Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe and Jaap Kamps. 2017a. Learning to learn from wea upervision. *In Proceedings of the workshop on Meta-Learning at Advances in Neural Information Processing Systems 31* (NIPS 2017), pages 65–74.
Shnarch, E., Alzate, C., Dankin, L., Gleize, M., Hou, Y., Choshen, L., ... & Slonim, N. (2018). Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) (Vol. 2, pp. 599-605).

snorkel - x's and y's in one go

What is snorkel?

ardigen

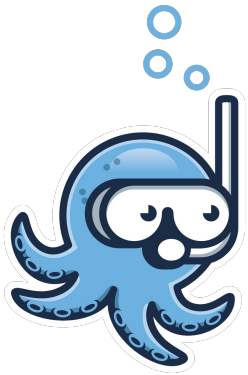
The authors - a research group led by Prof. Chris Ré at the Department of Computer Science at **Stanford University:**

*"A system for quickly generating training data with weak supervision."**

*"Snorkel is a system for rapidly creating, modeling, and managing training data."***

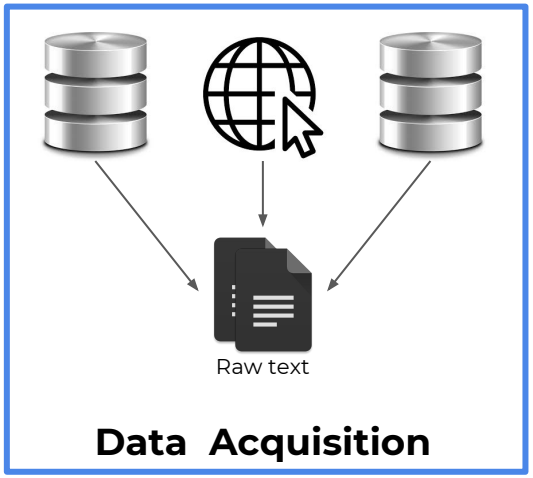
Me:

"A Python module."

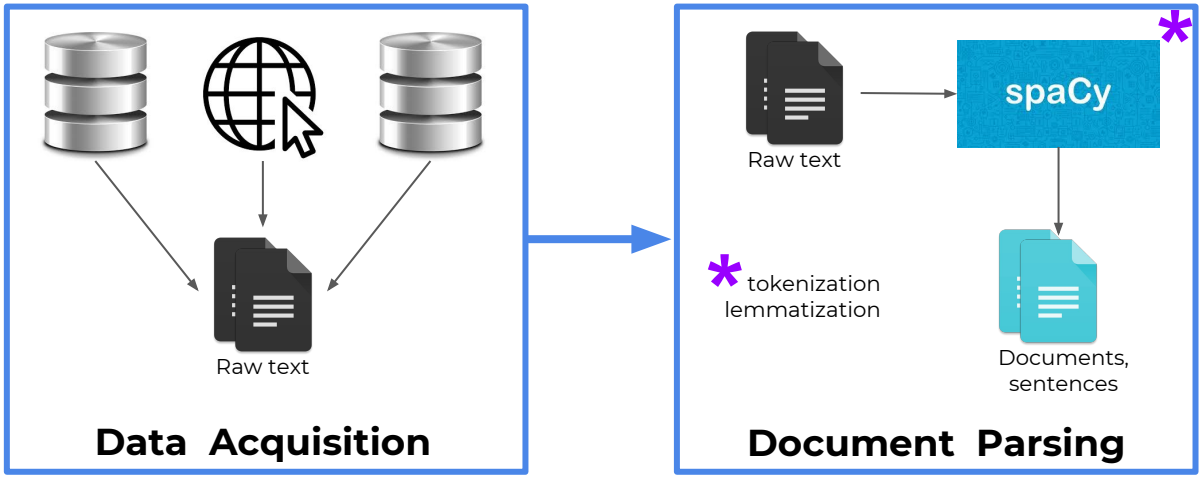


snorkel

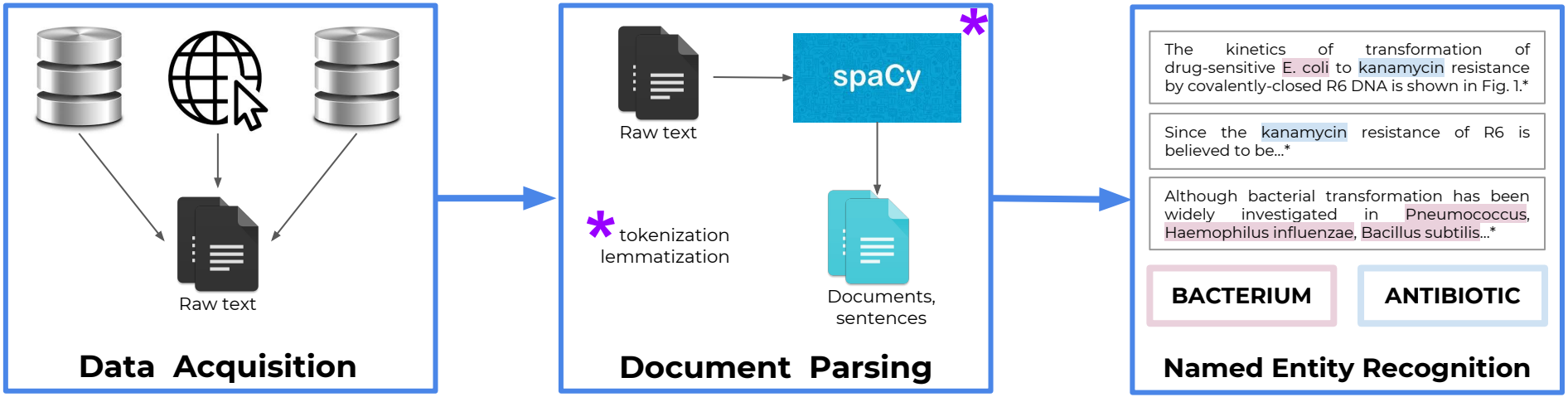
* https://github.com/HazyResearch/snorkel
** https://hazyresearch.github.io/snorkel/



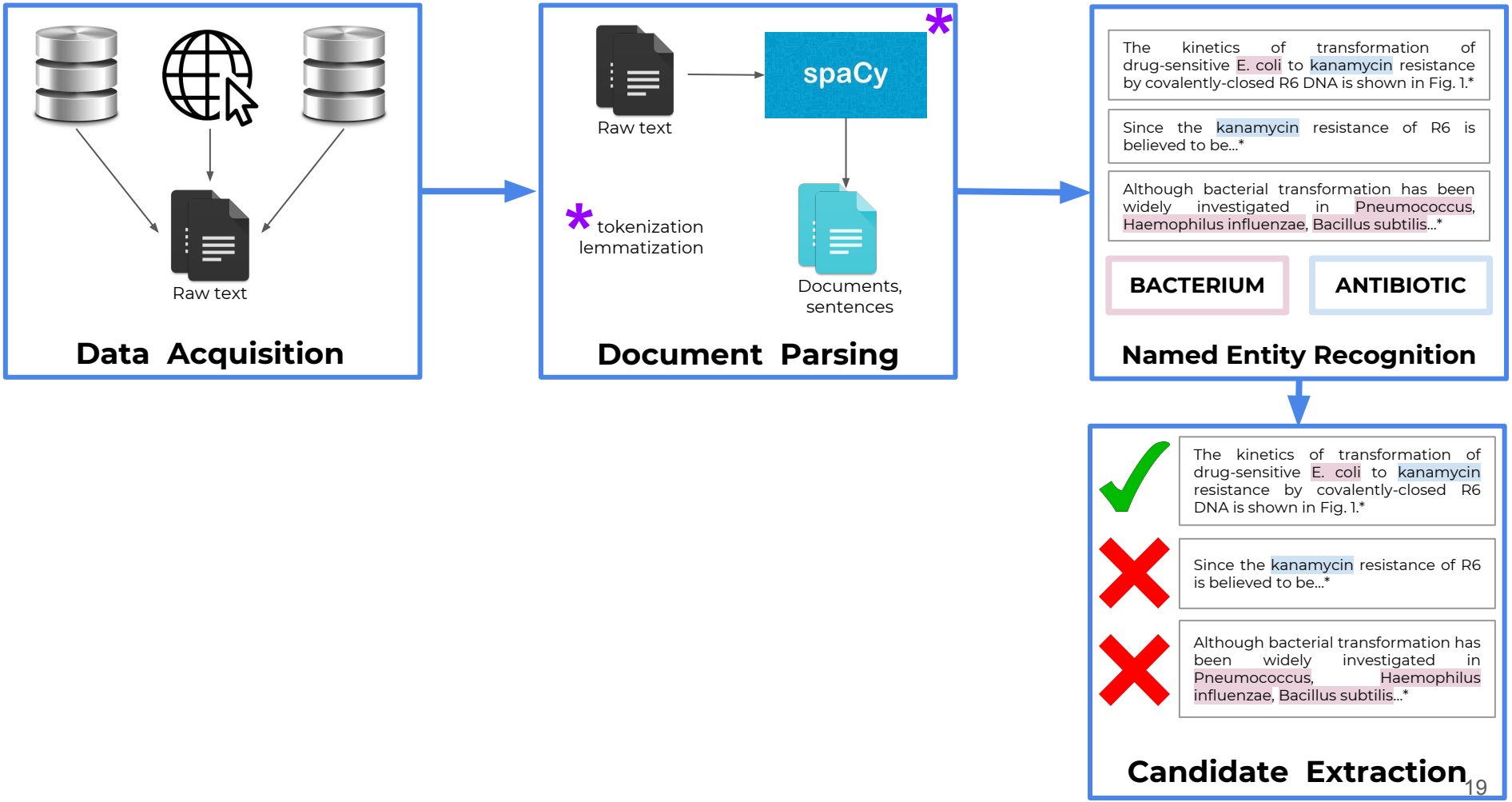
* Cohen SN, Chang AC, Hsu L. Nonchromosomal antibiotic resistance in bacteria: Genetic transformation of Escherichia coli by R-factor DNA. Proc Natl Acad Sci USA. 1972;69:2110–2114.



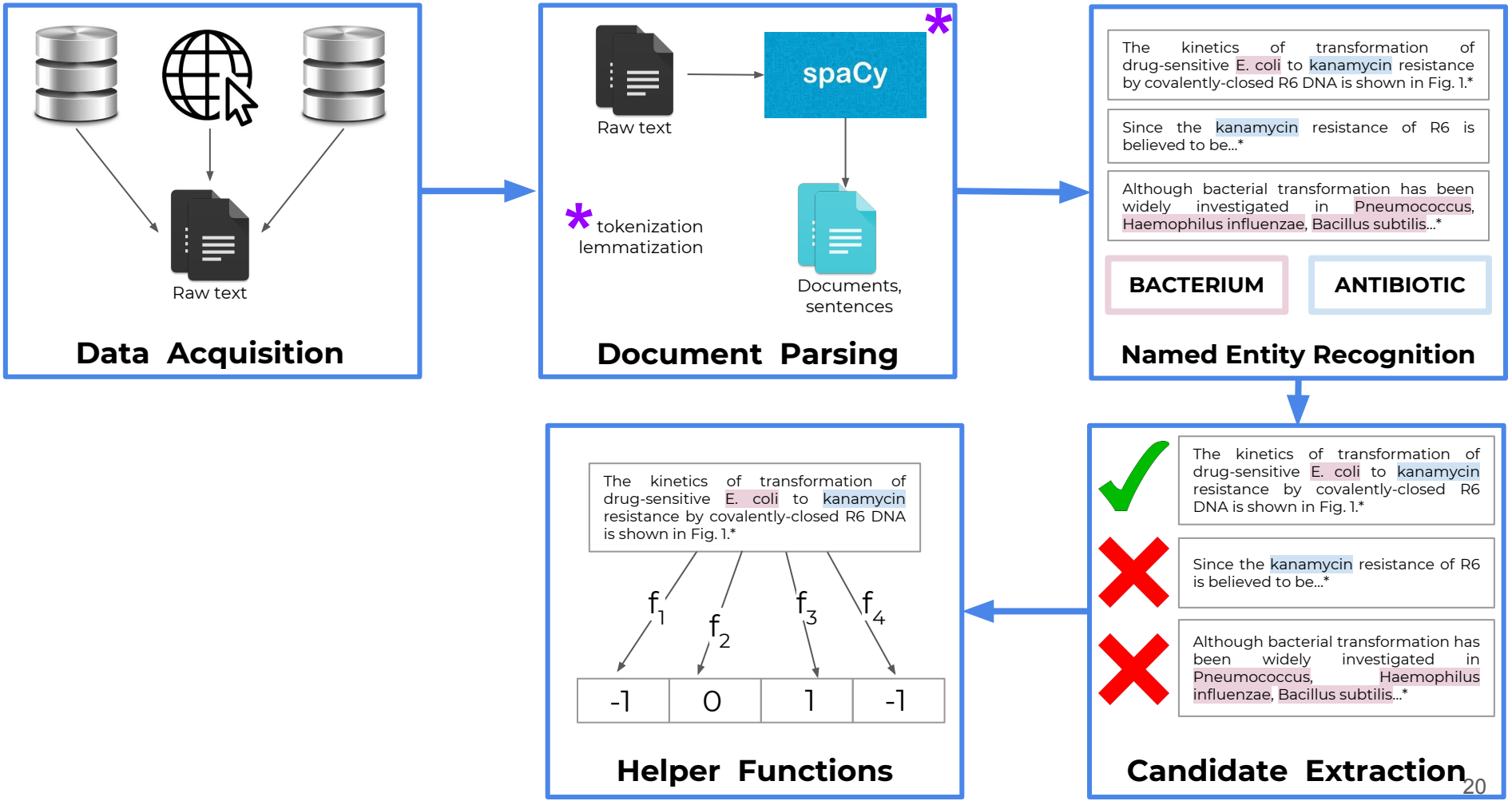
* Cohen SN, Chang AC, Hsu L. Nonchromosomal antibiotic resistance in bacteria: Genetic transformation of Escherichia coli by R-factor DNA. Proc Natl Acad Sci USA. 1972;69:2110–2114.



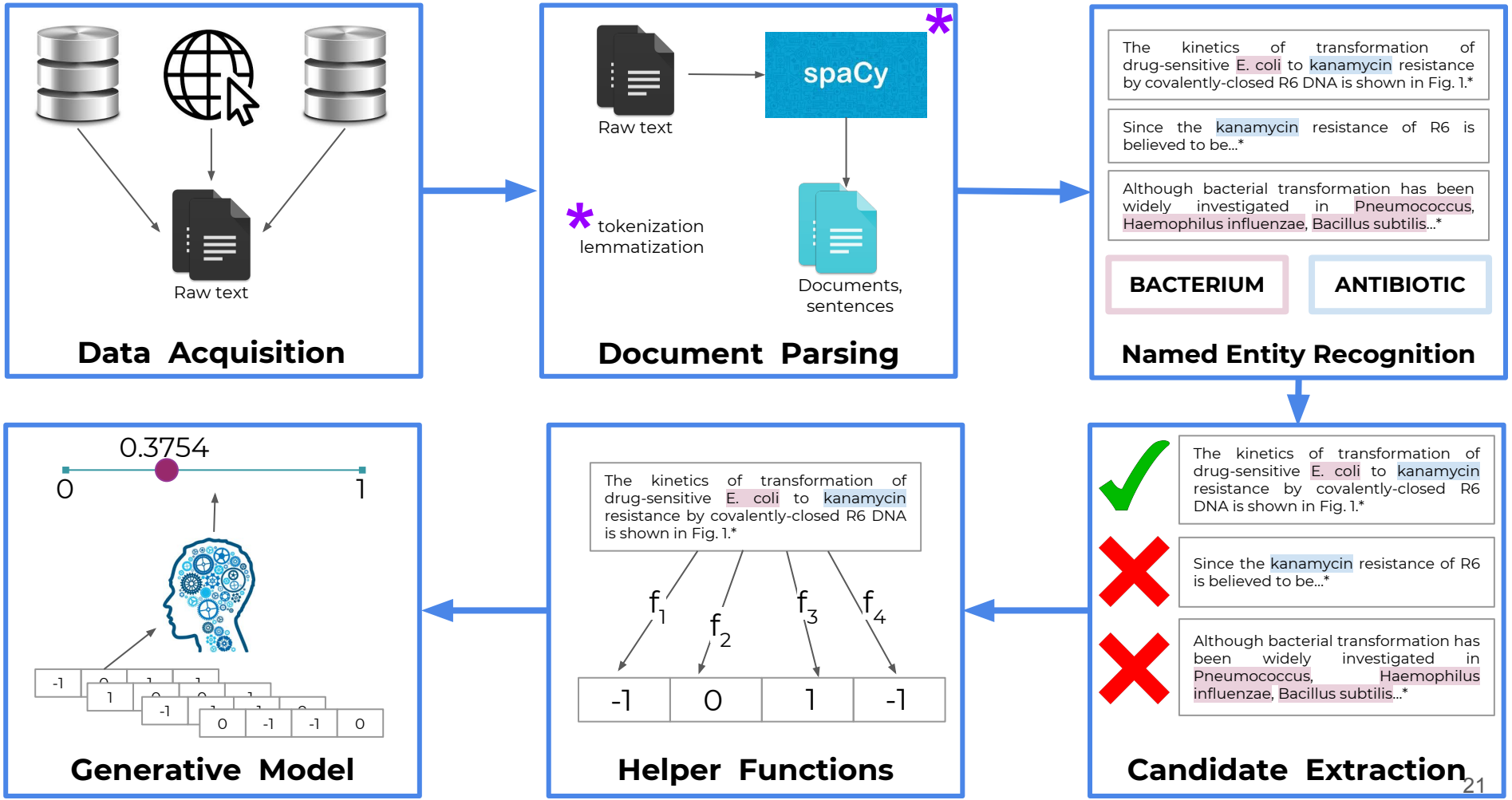
* Cohen SN, Chang AC, Hsu L. Nonchromosomal antibiotic resistance in bacteria: Genetic transformation of Escherichia coli by R-factor DNA. Proc Natl Acad Sci USA. 1972;69:2110–2114.



* Cohen SN, Chang AC, Hsu L. Nonchromosomal antibiotic resistance in bacteria: Genetic transformation of Escherichia coli by R-factor DNA. Proc Natl Acad Sci USA. 1972;69:2110–2114.



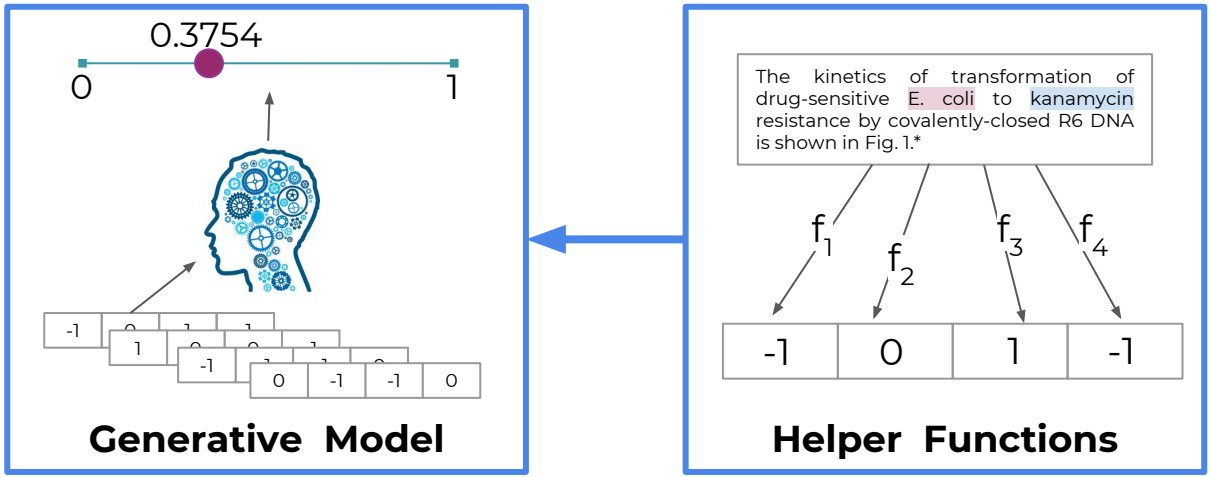
* Cohen SN, Chang AC, Hsu L. Nonchromosomal antibiotic resistance in bacteria: Genetic transformation of Escherichia coli by R-factor DNA. Proc Natl Acad Sci USA. 1972;69:2110–2114.



* Cohen SN, Chang AC, Hsu L. Nonchromosomal antibiotic resistance in bacteria: Genetic transformation of Escherichia coli by R-factor DNA. Proc Natl Acad Sci USA. 1972;69:2110–2114.

Snorkel is a general framework

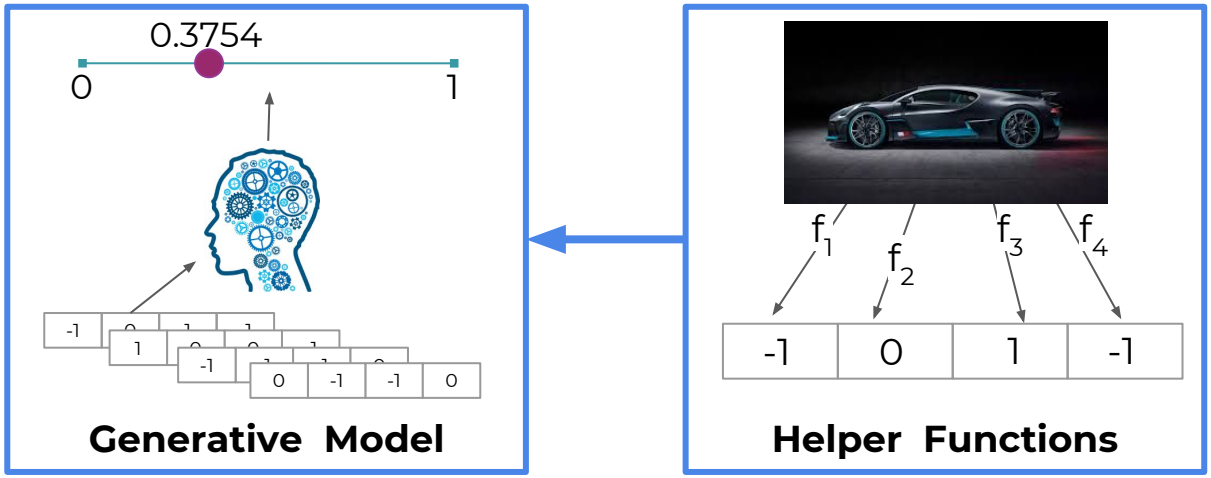
It goes beyond text data



* Cohen SN, Chang AC, Hsu L. Nonchromosomal antibiotic resistance in bacteria: Genetic transformation of Escherichia coli by R-factor DNA. Proc Natl Acad Sci USA. 1972;69:2110–2114.

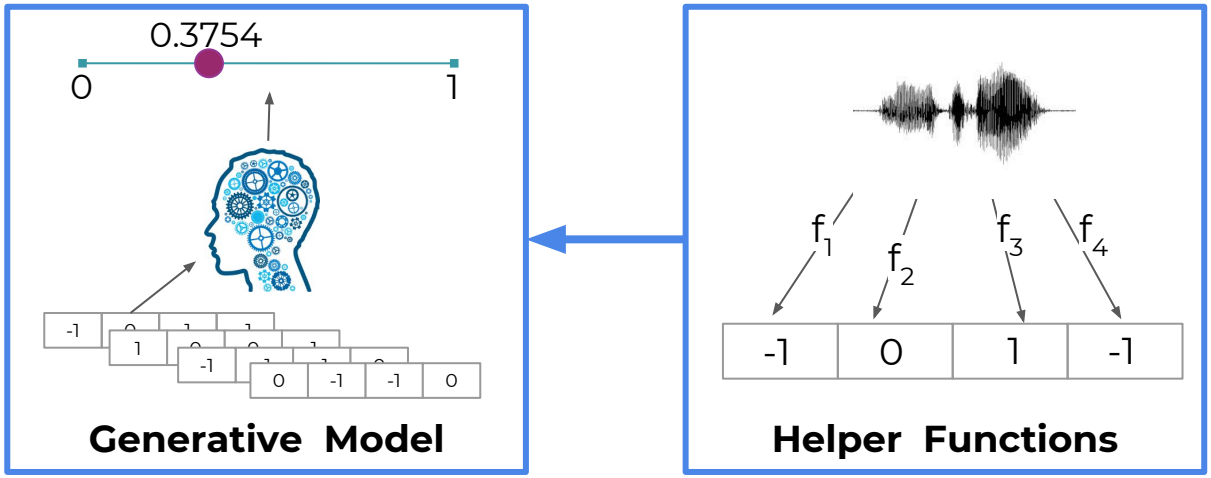
Snorkel is a general framework

It goes beyond text data



Snorkel is a general framework

It goes beyond text data



What are the helper functions?

Mediums of any heuristics

ardigen

The kinetics of transformation of drug-sensitive E. coli to kanamycin resistance by covalently-closed R6 DNA is shown in Fig. 1.*

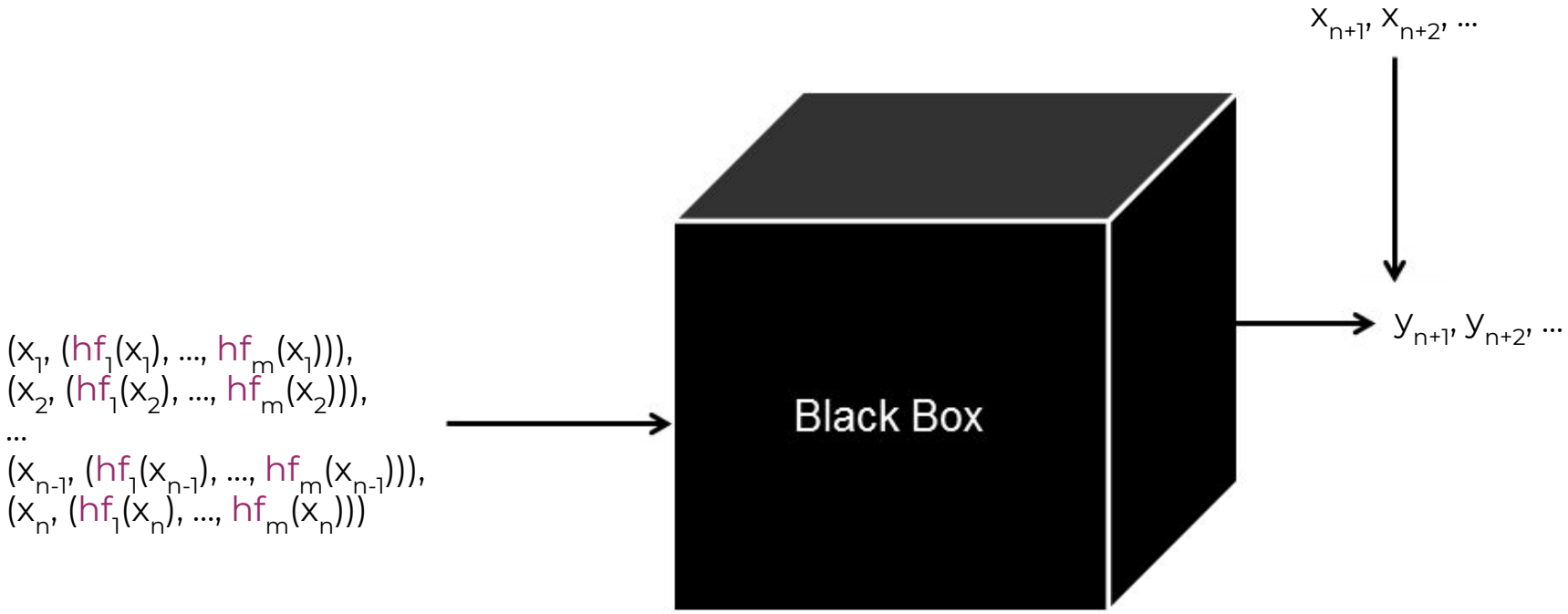
Does the sentence comprise word “***resistant***”?

Is there a ***reddish*** patch of color somewhere in the picture?

Does the wave consist mostly of ***high pitches***?

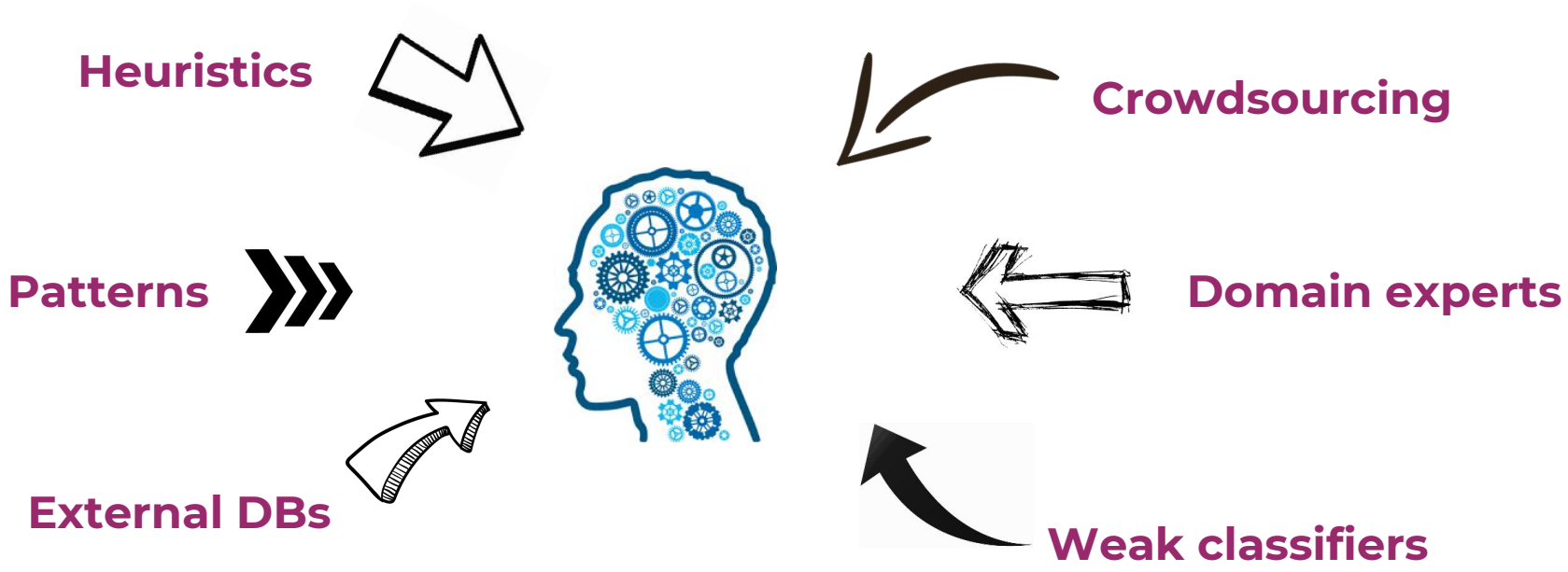
Helper functions

Big picture



Generative model

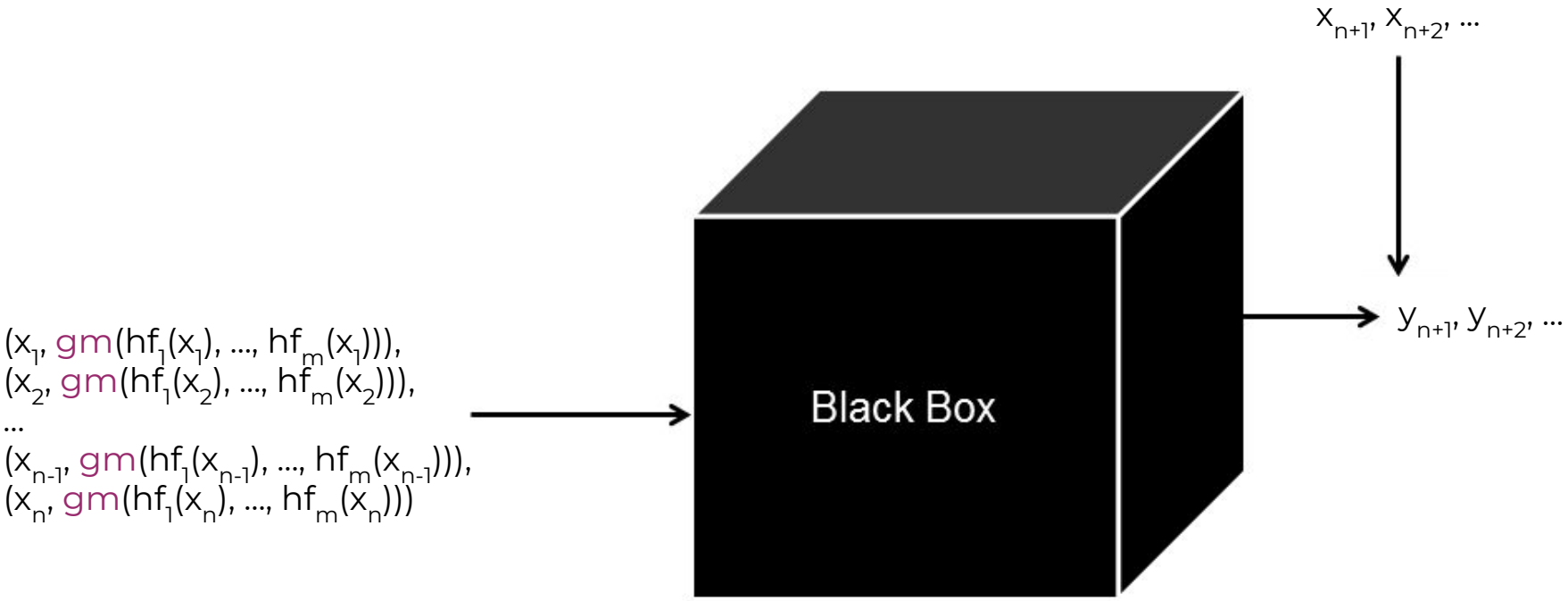
The heart of snorkel



Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, *11*(3), 269-282.

Generative model

Big picture



Snorkel on technical side

ardigen

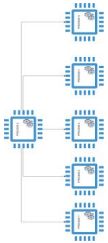
ORM software with support for most of the popular DB providers through **SQLAlchemy**.

Embedded extraction model is written in **Pytorch** (previously in tensorflow).

Devised as a **jupyter notebook**-based solution.

Built-in support for **Apache Spark** clusters.

NLP tasks can be performed out-of-the-box in **spaCy** or **NLTK**. Additionally, custom pipelines are also possible.



Multiprocessing is ensured through **processes** and **queues** (*multiprocessing* module).

Downsides *

- Snorkel is relatively **inflexible**, which makes changes to its very core **complicated** beyond measure.

* All the remarks apply to **v0.6**. They might have been eliminated / mitigated in the **v0.7**.

Downsides *

ardigen

- Snorkel is relatively **inflexible**, which makes changes to its very core **complicated** beyond measure.
- There exists a performance problem when huge data is loaded into the underlying database - **load speed** can be greatly enhanced. On top of that, by default, there is **no way of appending** documents to an existing database, everything needs to be calculated from scratch whenever such a need arises.

* All the remarks apply to **v0.6**. They might have been eliminated / mitigated in the **v0.7**.

Downsides *

ardigen

- Snorkel is relatively **inflexible**, which makes changes to its very core **complicated** beyond measure.
- There exists a performance problem when huge data is loaded into the underlying database - **load speed** can be greatly enhanced. On top of that, by default, there is **no way of appending** documents to an existing database, everything needs to be calculated from scratch when when such a need arises.
- All the tasks (NLP and learning) are performed **in-memory**, which is unsuitable for (relatively) big data sets, as the system quickly runs out of RAM.

* All the remarks apply to **v0.6**. They might have been eliminated / mitigated in the **v0.7**.

Snorkel enables:

- end-to-end processing in NLP tasks;
- generation of probabilistic labels to be used as weak supervision indications without the access to ground truth;
- combination of multiple, possibly noisy, labels to enhance the ultimate scores.

ardigen

Artificial Intelligence & Bioinformatics
for Precision Medicine

ardigen.com

Hand in hand with weak supervision using snorkel

Szymon Wojciechowski szymon.wojciechowski@ardigen.com