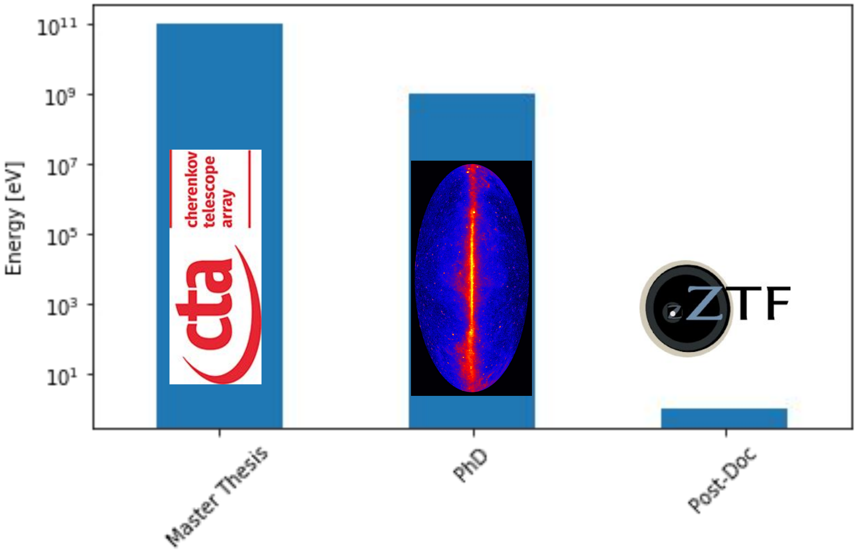
# From Astrophysics to Differential Privacy

Data science seminar
DESY Zeuthen, April 2020
Matteo Giomi

# About me

Science: time-domain astrophysics.




Now: privacy researcher.

Statice

Startup developing software for **privacy-preserving** data sharing.

# Privacy

“Privacy is the ability of an individual to seclude themselves or information about themselves, and thereby express themselves selectively.” (wikipedia)

Lack of privacy → behavioural change.

# Privacy in the digital era

Every interaction with technology creates data about the user.

- In the wrong hands data can be used for blackmail, social engineering, mass surveillance and the like.
- If used correctly, data can also lead to collective benefits.

Complete non-disclosure is not the best option.  Also,

- We do share sensitive data with strangers (i.e. doctors)

How to share data in a privacy preserving way?

# Example dataset

| phone | race | birth year | sex | zip code | medical condition |
|---|---|---|---|---|---|
| 015940192 | white | 1964 | f | 1203002 | chest_pain |
| 010405919 | white | 1964 | f | 1203505 | obesity |
| 011500159 | white | 1964 | f | 1203106 | short_breath |
| 010192042 | black | 1965 | m | 5403221 | heart_disease |
| 015909191 | black | 1965 | m | 5403221 | heart_disease |
| 015553436 | black | 1965 | m | 5403221 | heart_disease |
| 016901095 | white | 1960 | f | 3003202 | ovarian cancer |
| 017497297 | white | 1960 | f | 3003555 | ovarian cancer |
| 018206810 | white | 1960 | m | 3003890 | prostate cancer |

# Identifiers and quasi-identifiers

| phone | race | birth year | sex | zip code | medical condition |
|---|---|---|---|---|---|
| 015940192 | white | 1964 | f | 1203002 | chest_pain |
| 010405919 | white | 1964 | f | 1203505 | obesity |
| 011500159 | white | 1964 | f | 1203106 | short_breath |
| 2 | black | [illegible] | [illegible] | 3221 | he |
| 91 | black | [illegible] | [illegible] | 3221 | he |
| 015553436 | black | 1965 | m | 5403221 | heart_disease |
| 016901095 | white | 1960 | f | 3003202 | ovarian cancer |
| 017497297 | white | 1960 | f | 3003555 | ovarian cancer |
| 018206810 | white | 1960 | m | 3003890 | prostate cancer |

Personally identifying information (PII)

“Quasi” identifiers

Sensitive information

# “Sanitizing” a dataset

| phone | race | birth year | sex | zip code | medical condition |
|---|---|---|---|---|---|
| [illegible]1594019[illegible] | white | 1964 | f | 1203002 | chest_pain |
| 0[illegible]04059[illegible]19 | white | 1964 | f | 1203505 | obesity |
| 01[illegible]500[illegible]59 | white | 1964 | f | 1203106 | short_breath |
| [illegible]2 | black | [illegible] | [illegible] | [illegible]3221 | he[illegible] |
| [illegible]91 | black | [illegible] | [illegible] | [illegible]3221 | he[illegible] |
| 015[illegible]53[illegible]36 | black | 1965 | m | 5403221 | heart_disease |
| 01[illegible]3901[illegible]95 | white | 1960 | f | 3003202 | ovarian cancer |
| [illegible]1749729[illegible] | white | 1960 | f | 3003555 | ovarian cancer |
| 018206810 | white | 1960 | m | 3003890 | prostate cancer |

Personally identifying information (PII)

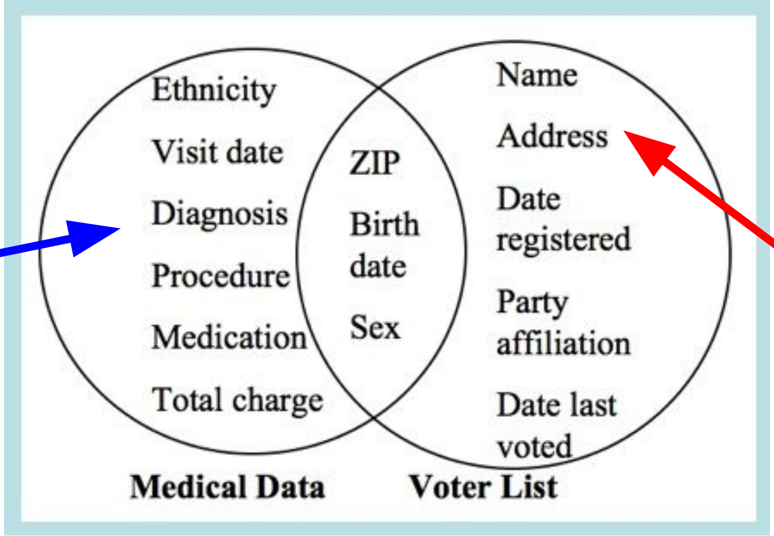
“Quasi” identifiers

Sensitive information

# Re-identification via linkage

Even without PII individuals can be re-identified by linking with external information.

‘Sanitized’ record, with sensitive data

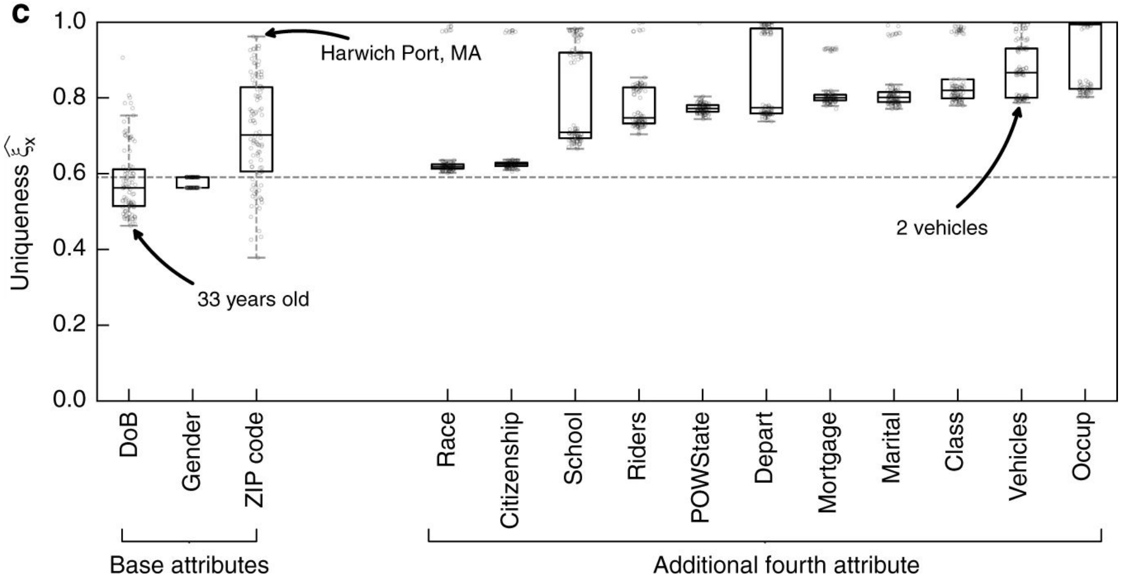



Identifiable, no sensitive data

[Sweeney, Latanya. Weaving Technology and Policy Together to Maintain Confidentiality. Journal of Law, Medicine and Ethics, Vol. 25 1997, p. 98-110]

# “We are all special”

Given enough quasi-identifiers everyone is unique → can be re-identified with certainty.




Rocher, L., et al. Estimating the success of re-identifications in incomplete datasets using generative models.

# Removing PII is not enough

The notion of PII has no technical meaning: everything is PII!

However:

*"We do not share information of data in any personally identifiable form.."*

[Arvind Narayanan and Vitaly Shmatikov, Myths and and fallacies of "Personally identifiable information"](#)

# Old solution: k-anonymity

Avoid unique joints: "any combination of quasi-identifiers must appear at least *k* times"

| race | birth year | sex | zip code | medical condition |
|---|---|---|---|---|
| white | 1964 | * | 1203* | chest_pain |
| white | 1964 | * | 1203* | obesity |
| white | 1964 | * | 1203* | short_breath |
| black | 1965 | * | 5403* | heart_disease |
| black | 1965 | * | 5403* | heart_disease |
| black | 1965 | * | 5403* | heart_disease |
| white | 1960 | * | 3003* | ovarian cancer |
| white | 1960 | * | 3003* | ovarian cancer |
| white | 1960 | * | 3003* | prostate cancer |

P. Samarati and L. Sweeney, Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression

# No solution: k-anonymity

Problems with k-anonymity: lack of diversity, background knowledge

| phone | race | birth year | sex | zip code |
| --- | --- | --- | --- | --- |
| 015940192 | white | 1964 | f | 1203002 |

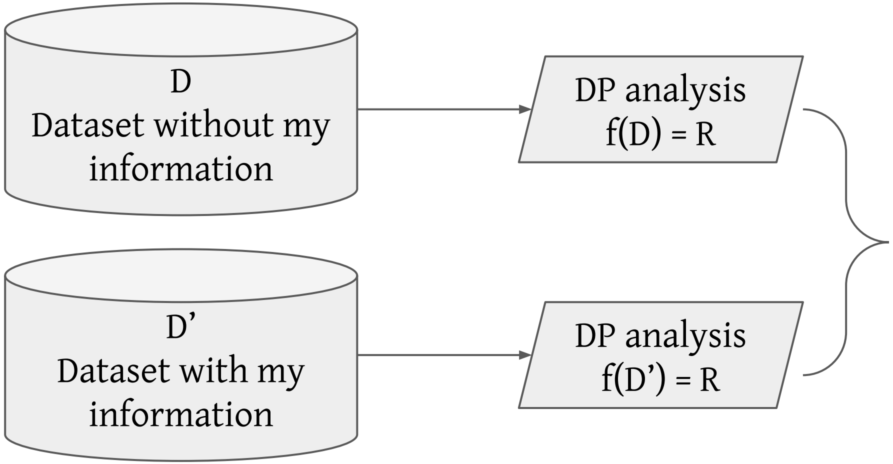
| phone | race | birth year | sex | zip code |
| --- | --- | --- | --- | --- |
| 015909191 | black | 1965 | f | 5403014 |
| 018206810 | white | 1960 | m | 3003890 |

| race | birth year | sex | zip code | medical condition |
| --- | --- | --- | --- | --- |
| white | 1964 | * | 1203* | chest_pain |
| white | 1964 | * | 1203* | obesity |
| white | 1964 | * | 1203* | short_breath |
| black | 1965 | * | 5403* | heart_disease |
| black | 1965 | * | 5403* | heart_disease |
| black | 1965 | * | 5403* | heart_disease |
| white | 1960 | * | 3003* | ovarian cancer |
| white | 1960 | * | 3003* | ovarian cancer |
| white | 1960 | * | 3003* | prostate cancer |

# Differential privacy

Suppose I have a secret and I'm considering whether to share my data. I wish that

Pr(someone guess my secret | data) ~ Pr(someone guess my secret)




$$\frac{Pr[f(D) = R]}{Pr[f(D\prime) = R]} \leq e^{\varepsilon}$$

[Dwork C., et al.. (2006) Calibrating Noise to Sensitivity in Private Data Analysis]

# Differential privacy

Suppose I have a secret and I'm considering whether to share my data. I wish that

Pr(someone guess my secret | data) ~ Pr(someone guess my secret)

The ε parameter is the 'privacy guarantee':

- The smaller the ε the stronger the privacy.

$$\frac{Pr[f(D) = R]}{Pr[f(D\prime) = R]} \leq e^{\varepsilon}$$

[Dwork C., et al.. (2006) Calibrating Noise to Sensitivity in Private Data Analysis]

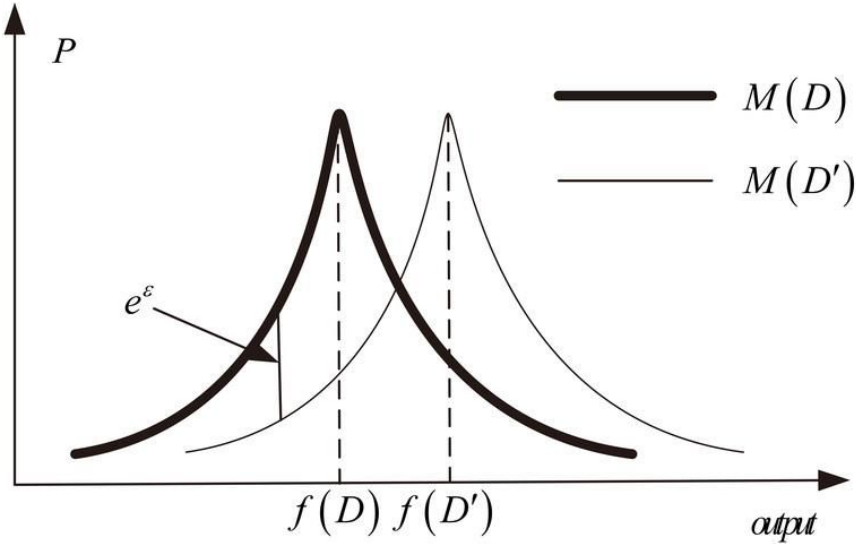
# DP example: laplacian mechanism

Adds noise draws from laplacian distribution:

$$M(f, \varepsilon, D) = f(D) + \text{Laplace}(\Delta f/\varepsilon)$$

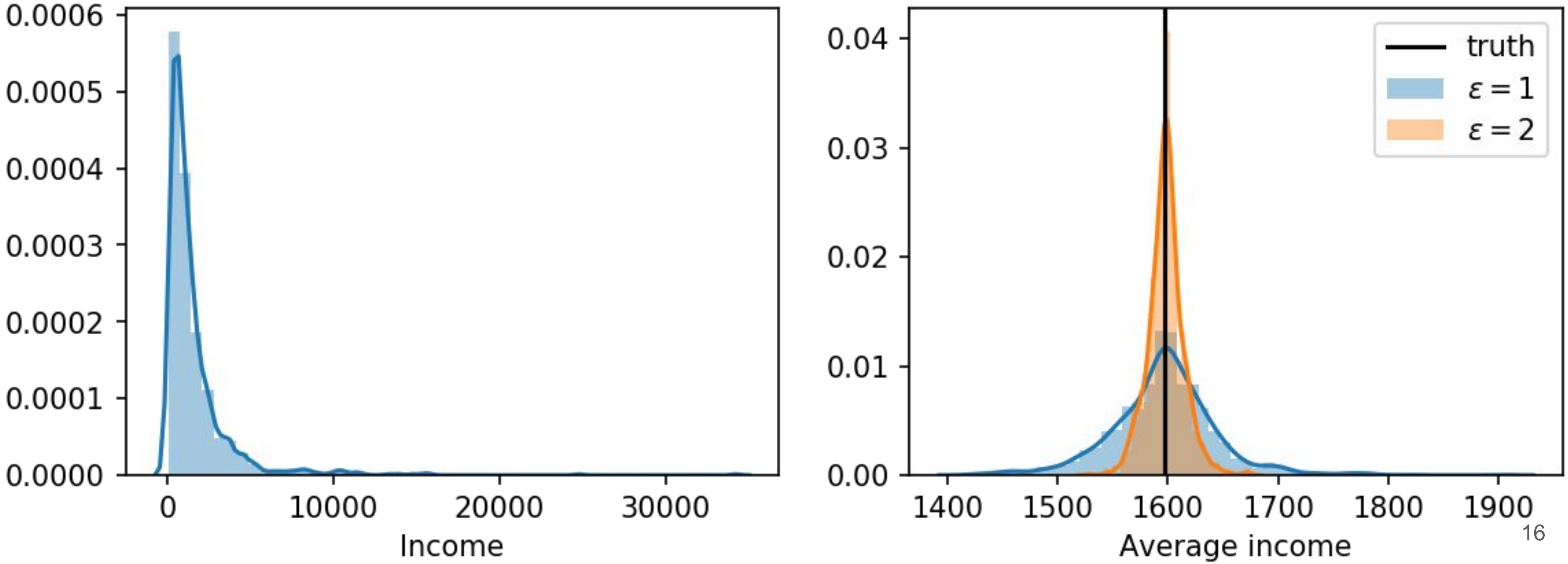
- $\Delta$f is the <u>sensitivity</u> of the query f: maximum change in its output due to single records in the dataset.

- For a given epsilon, high-sensitivity queries requires more noise.



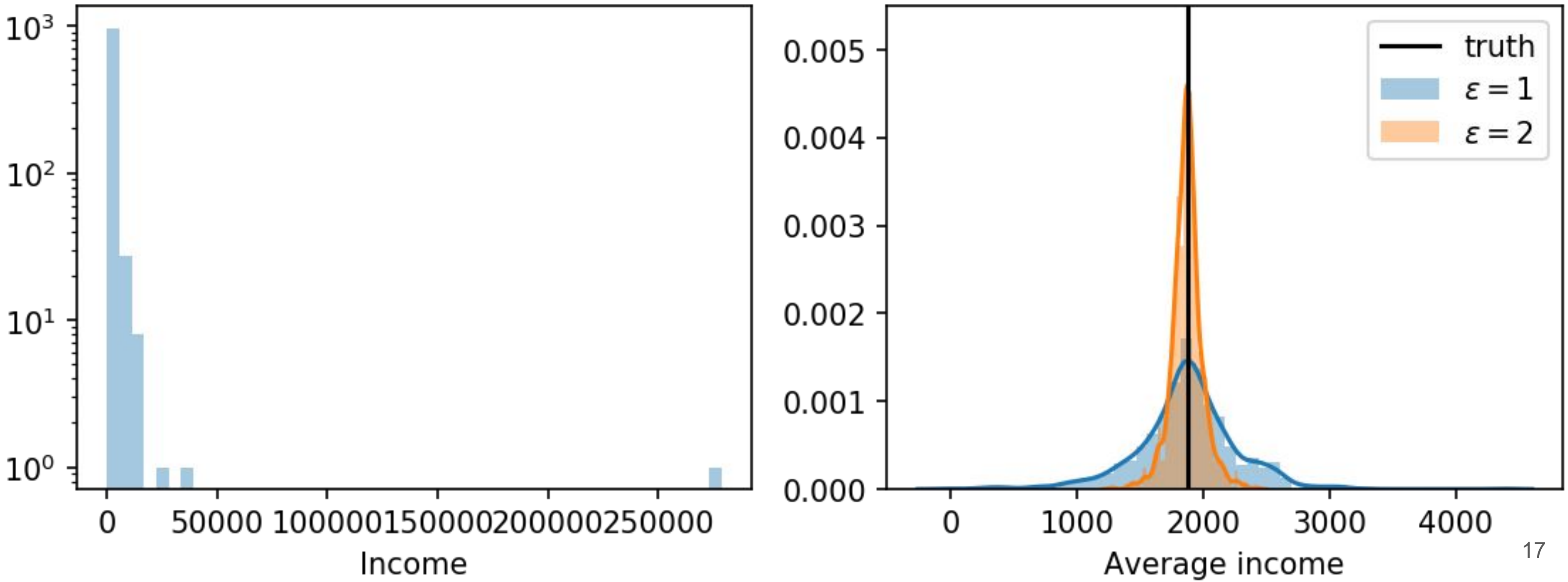
# Toy example

Use laplacian mechanism to estimate the average income of a population.



# Problems: outliers

To mask an outlier much more noise have to be added.



# Differential privacy

DP is a property of the analysis and is based on the addition of ‘just enough’ noise.

Pros:

- Makes no assumption on the attacker.
- Robust against post-processing.
- Different DP analysis compose.

Cons:

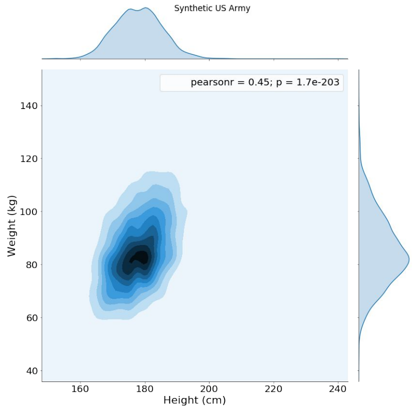
- Outliers requires large amount of noise.
- Noise can be averaged out via multiple queries.

# Privacy-preserving synthetic data

- Use deep generative models to learn the data-generating distribution.
- Sample from this distribution to obtain synthetic data.

| | Height (cm) | Weight (kg) |
|---|---|---|
| 0 | 165.10 | 68.03880 |
| 1 | 162.56 | 88.45044 |
| 2 | 170.18 | 61.23492 |
| 3 | 172.72 | 72.57472 |
| 4 | 177.80 | 88.45044 |

Original data




| | Height (cm) | Weight (kg) |
|---|---|---|
| 0 | 165.059741 | 67.628771 |
| 1 | 162.493573 | 89.085417 |
| 2 | 169.926607 | 50.631068 |
| 3 | 172.567460 | 72.163759 |
| 4 | 177.599143 | 88.280858 |

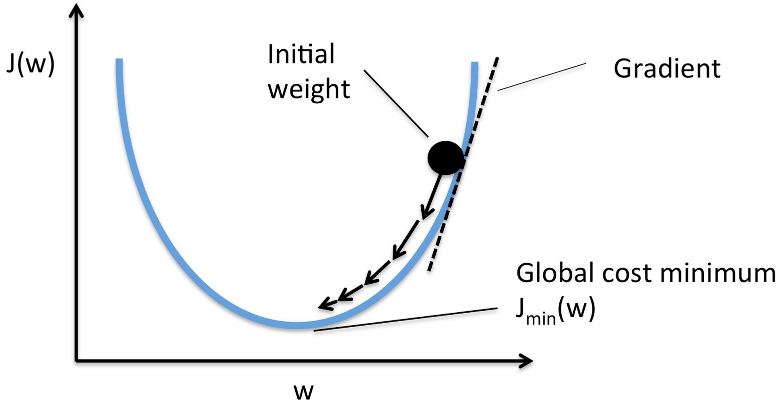
Synthetic data

# Deep learning with differential privacy

Models learn by minimizing a loss function.

Minimization via stochastic gradient descent (SGD).

The model interacts with the data exclusively via the gradients




Fitting a model with DP:

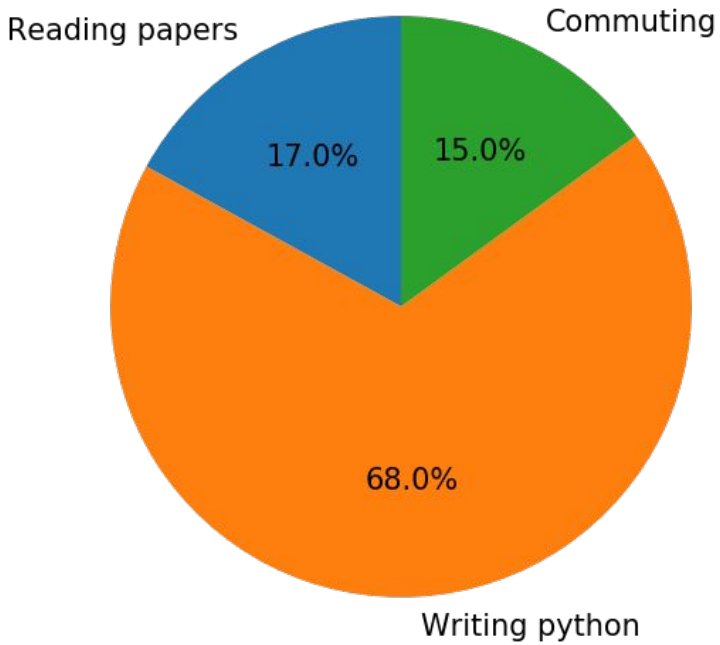
- Each gradient is clipped to a maximum length (fix the sensitivity).
- Noise is added to each gradient component.

This makes sure that the model won't learn from individual training examples.

M. Abadi et al, Deep Learning with Differential Privacy (2016)

# What am I doing every day

Implement DP algorithms and develop privacy evaluations (attack is the best defence)



# Transition to industry: applying for a job

Before you start:

- Time (3 to 6 months) and patience.
- Sleek 1-page CV, honest cover letter (sometimes not asked).
- Profile on linkedin, glassdoors and the like.

Typical job interview:

- Initial call.
- Code challenge / test task.
- Final interview.

Job interview is a skill to be learned: don’t be afraid to apply generously.

# Transition to industry: tips and tricks

This is what I gathered

- For position you really like, get in contact through company website.
- Sometimes it’s hard to convince people that you can “deliver”.

Good to have:

- Familiarity with usual data-science stack: numpy, pandas, seaborn, sklearn.
- Plenty of material on machine learning (Andrew Ng on yt, for example).
- Plenty of datasets and tasks on kaggle, start playing around.
- Luck.

Thanks for the attention