



# Privacy for Data Scientists

Utilizing Privacy-Aware Methods for Data Science

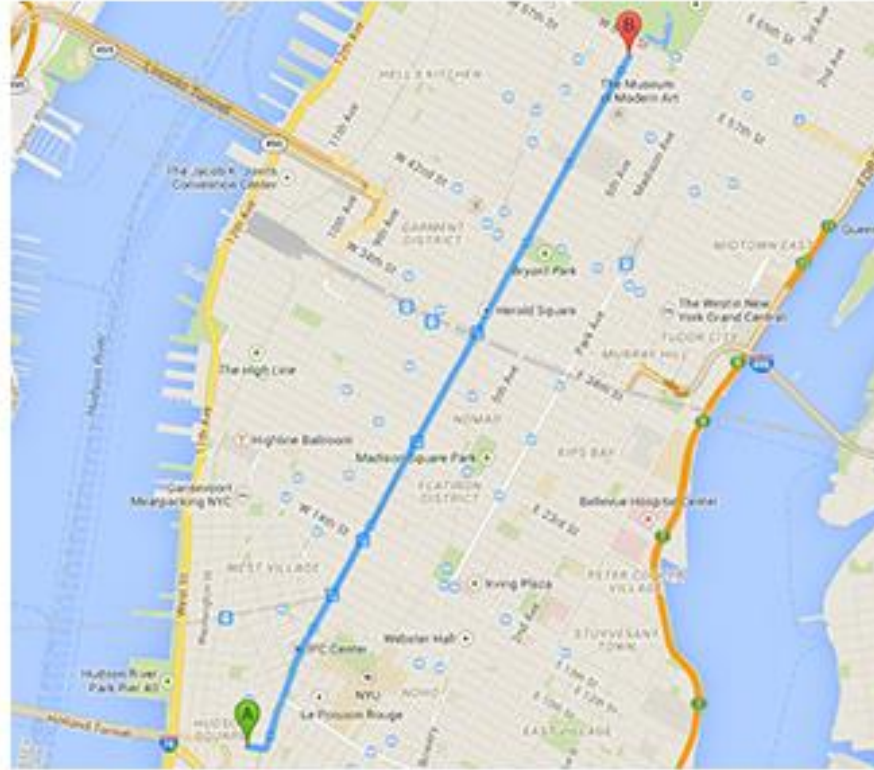
**KIProtect**

EuroPython 2018

# Data Privacy: For Whom?



KOURTNEY KARDASHIAN  
SCOTT DISICK

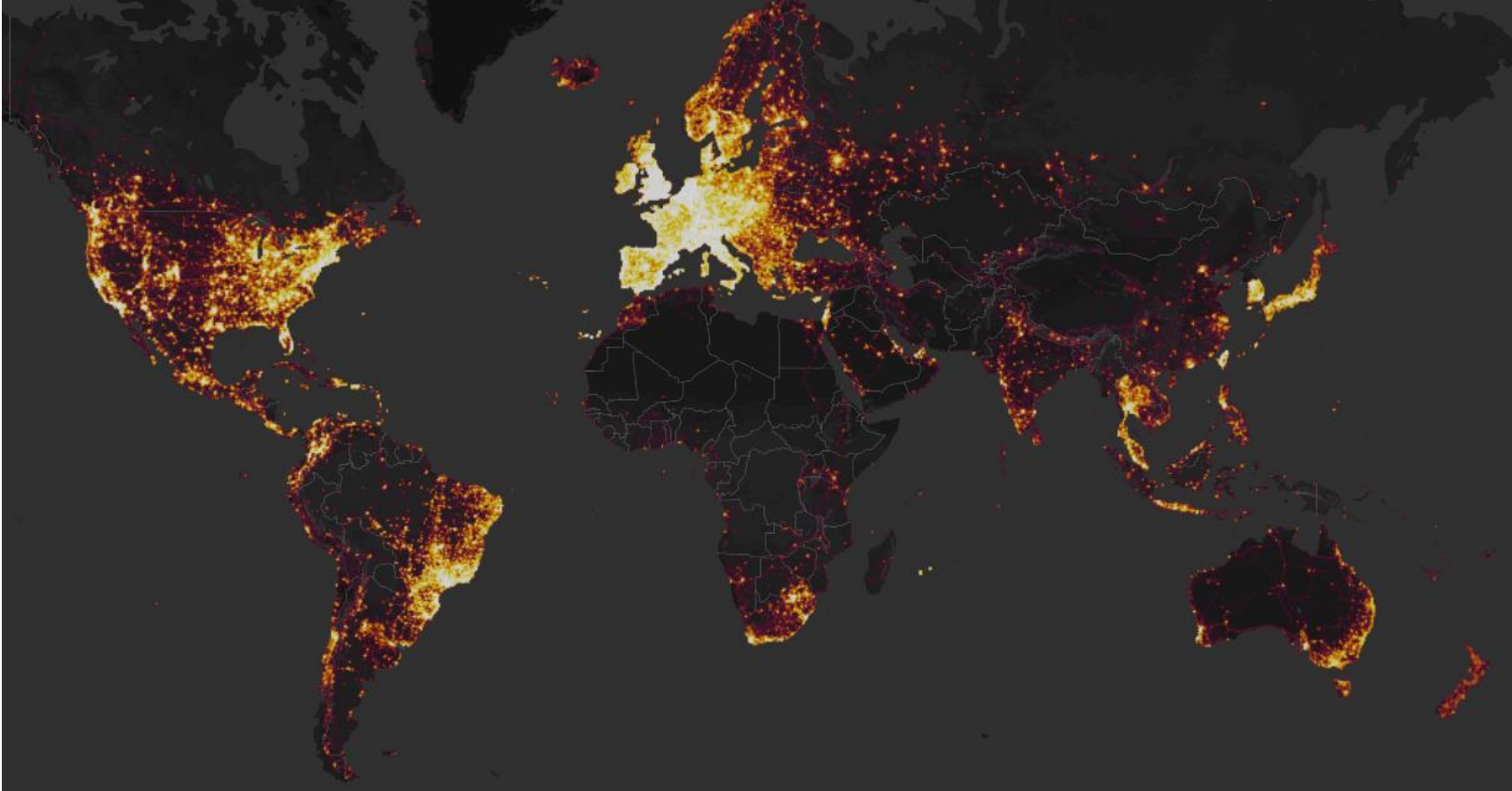


NOVEMBER 4, 2013 • 12:11 PM - 12:36 PM  
246 SPRING ST. TO 1412 6TH AVE  
\$16.50 FARE • \$3.40 TIP • ©SPLASH

Poorly  
“Anonymized”  
Data: NYC Taxi  
Rides

(released under  
FOIA and used a  
md5 hash)

# Data Privacy: From Whom?



Aggregated  
Anonymized Data  
Leaks Secrets

(secret military  
bases revealed  
from aggregated  
sport data)

# About the Instructors



**Dr. Andreas Dewes**  
Dr. rer. nat., Dipl. Phys., Dipl. Kfm.



**Katharine Jarmul**



# What even is Privacy?

## OPINIONS ON INTERNET PRIVACY

### THE PHILOSOPHER:

"PRIVACY" IS AN IMPRACTICAL WAY TO THINK ABOUT DATA IN A DIGITAL WORLD SO UNLIKE THE ONE IN WHICH OUR SOCI-

SO BORED.



### THE CRYPTO NUT:

MY DATA IS SAFE BEHIND SIX LAYERS OF SYMMETRIC AND PUBLIC-KEY ALGORITHMS.

WHAT DATA IS IT?

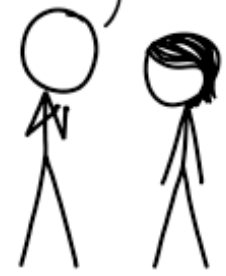
MOSTLY ME EMAILING WITH PEOPLE ABOUT CRYPTOGRAPHY.



### THE CONSPIRACIST:

THESE LEAKS ARE JUST THE TIP OF THE ICEBERG. THERE'S A WAREHOUSE IN UTAH WHERE THE NSA HAS THE ENTIRE ICEBERG.

I DON'T KNOW HOW THEY GOT IT THERE.



### THE NIHILIST:

JOKE'S ON THEM, GATHERING ALL THIS DATA ON ME AS IF ANYTHING I DO MEANS ANYTHING.



### THE EXHIBITIONIST:

MMMM? I SURE HOPE THE NSA ISN'T WATCHING ME BITE INTO THESE JUICY STRAWBERRIES!!

OOOPS, I DRIPPED SOME ON MY SHIRT! BETTER TAKE IT OFF.

GOOGLE, ARE YOU THERE?

GOOGLE, THIS LOTION FEELS SOOOO GOOD.



### THE SAGE:

I DON'T KNOW OR CARE WHAT DATA ANYONE HAS ABOUT ME.

DATA IS IMAGINARY. THIS BURRITO IS REAL.



# Privacy Definitions (which we will cover today)

- Pseudonymization: Personal information not directly disclosed, still vulnerable to statistical / informed attacks.
- K-Anonymity: Some anonymity guarantees (identifiable with outside information).
- Differential Privacy: “Gold Standard” of anonymity, may still leak group information.

# Pseudonymization

- Originating from Greek (pseudonymos) "having a false name, under a false name"
- Under GDPR, it allows for processing of data in conditions not expressly defined in the initial collection. It is also recommended for Privacy by Design.
- Not safe for public release or release to unsecured partners

# k-Anonymity

- Attack Model: Identify individuals via attribute combinations (e.g. age, gender, zip code and weight)
- If enough attributes (quasi identifiers) are available, a unique identifier can be constructed even in large data sets.
- Protect individuals by ensuring that for each possible identifier combination there are at least  $k$  individuals in the data set.
- Naive method has several drawbacks that can be partially fixed.



# Differential Privacy

- Differential Privacy (DP) is not an anonymization method but a way to quantify information leakage in probabilistic querying or data transformation algorithms.
- It is more general than k-anonymity and has a much stricter risk model (as it does not distinguish between sensitive and non-sensitive attributes)
- For simple data types (e.g. binary data) we can easily implement differentially private querying / release mechanism.

# Privacy-Preserving ML:

## PATE

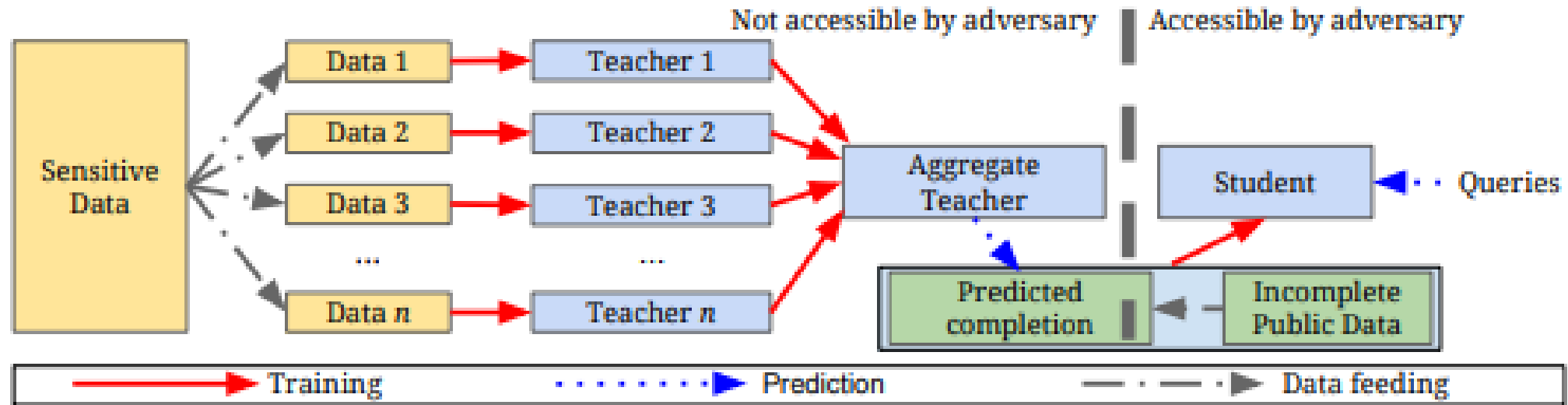
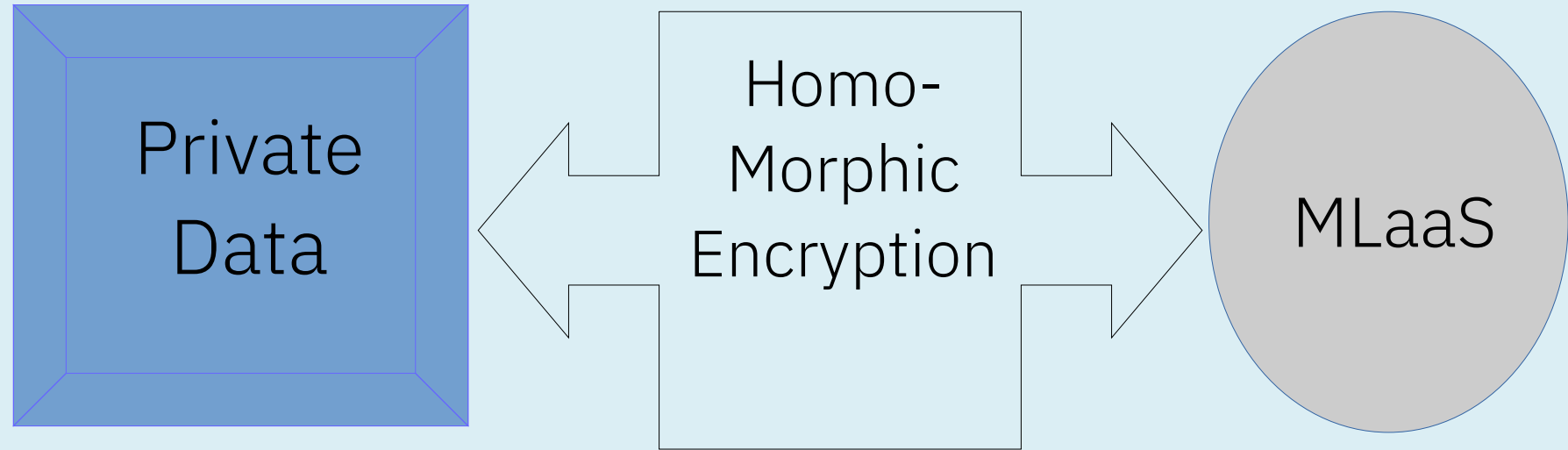


Figure 1: Overview of the approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble.

Papernot et al (2016):

<https://arxiv.org/abs/1610.05755>

Privacy-  
Preserving  
ML:  
CryptoNets



Dowlin et al (2016):

<http://proceedings.mlr.press/v48/gilad-bachrach16.pdf>

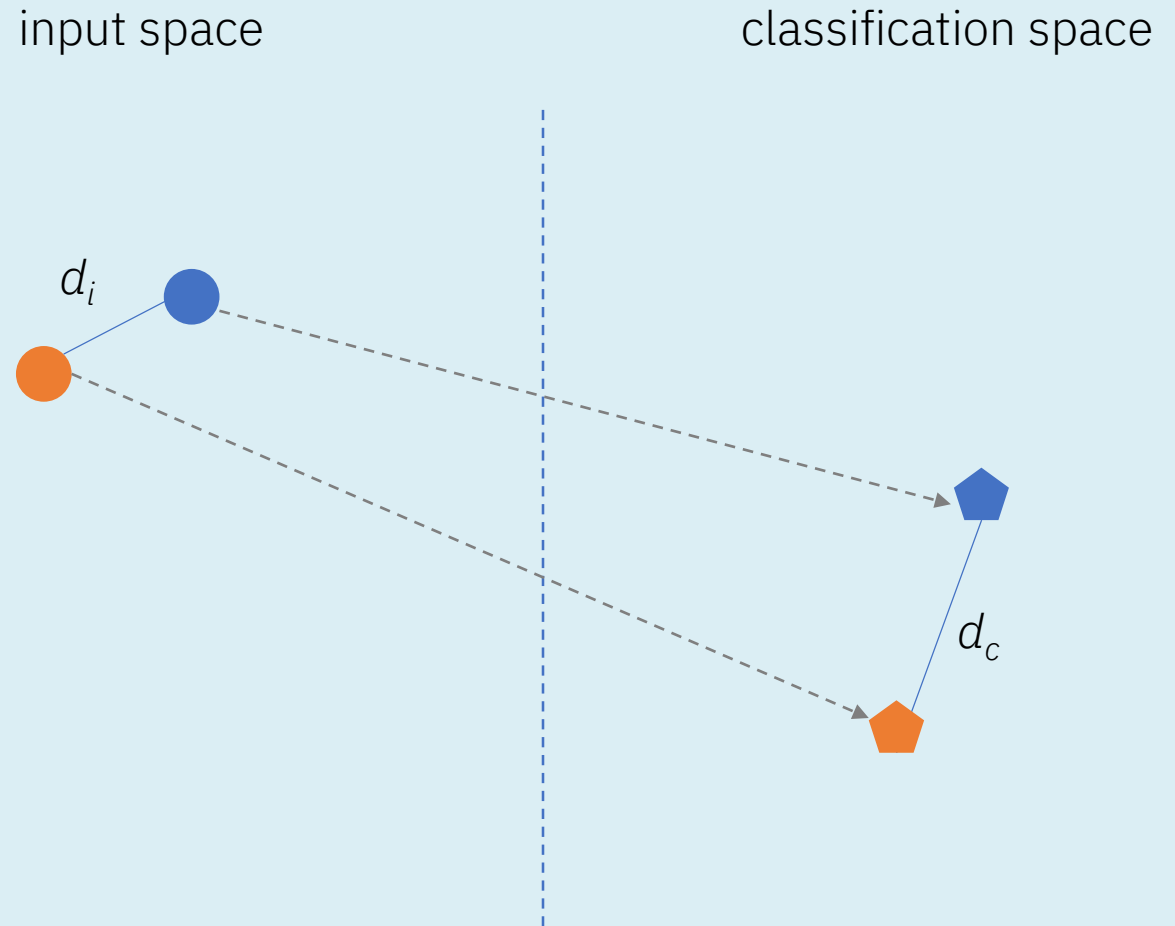
# Privacy- Preserving ML

Individual-based fairness is enforced through a „Lipschitz property“: The statistical distance of classifications for two individuals should be bounded by their distance in the input space. Interesting takeaway: **Fairness (usually) implies privacy.**

The input space distance is calculated based on a fair metric (that we need to define).

Note: Implied „affirmative action“ not always legal!

## C. Dwork et. al.: Fairness Through Awareness



If all  
else fails...



## YOUR PERSONAL INFORMATION

PLEASE DON'T SEND US YOUR PERSONAL INFORMATION. WE DO NOT WANT YOUR PERSONAL INFORMATION. WE HAVE A HARD ENOUGH TIME KEEPING TRACK OF OUR OWN PERSONAL INFORMATION, LET ALONE YOURS.

IF YOU TELL US YOUR NAME, OR ANY IDENTIFYING INFORMATION, WE WILL FORGET IT IMMEDIATELY. THE NEXT TIME WE SEE YOU, WE'LL STRUGGLE TO REMEMBER WHO YOU ARE, AND TRY DESPERATELY TO GET THROUGH THE CONVERSATION SO WE CAN GO ONLINE AND HOPEFULLY FIGURE IT OUT.

# Thank you for attending!

Questions? We'd Love to hear them!

Or reach out anytime:

[info@kiprotect.com](mailto:info@kiprotect.com)

@KIProtect (Twitter)

<https://github.com/kiprotect>

Andreas Dewes

[andreas@kiprotect.com](mailto:andreas@kiprotect.com)

@japh44 (Twitter)

Katharine Jarmul

[katharine@kiprotect.com](mailto:katharine@kiprotect.com)

@kjam (Twitter)

7scientists GmbH

KIProtect

Bismarckstr. 10-12

10625 Berlin

