Less rules, more engineering with encrypted data-in-use

Alexander C. Mueller

July 2022

Agenda

Goal: awareness of future-proof new privacy technology

Trans-Atlantic data pipelines in need of future proofing and why

- (In-) Adequacy Frameworks
- Privacy Shield, etc. vs. Schrems I and II
- Notable EDPB Recommendations

Exciting new approaches

- Something for you to play with
- Qualitatively different encryption
- Some nomenclature
- Encrypted *data-in-use* and private pipelines

Who is the speaker?

30 second resume:

- an ancient metro Saint Louis townie
- Ph.D. University of Michigan 2013 (math)
- a few years of data science
- founded data privacy tech company Capnion, Inc.

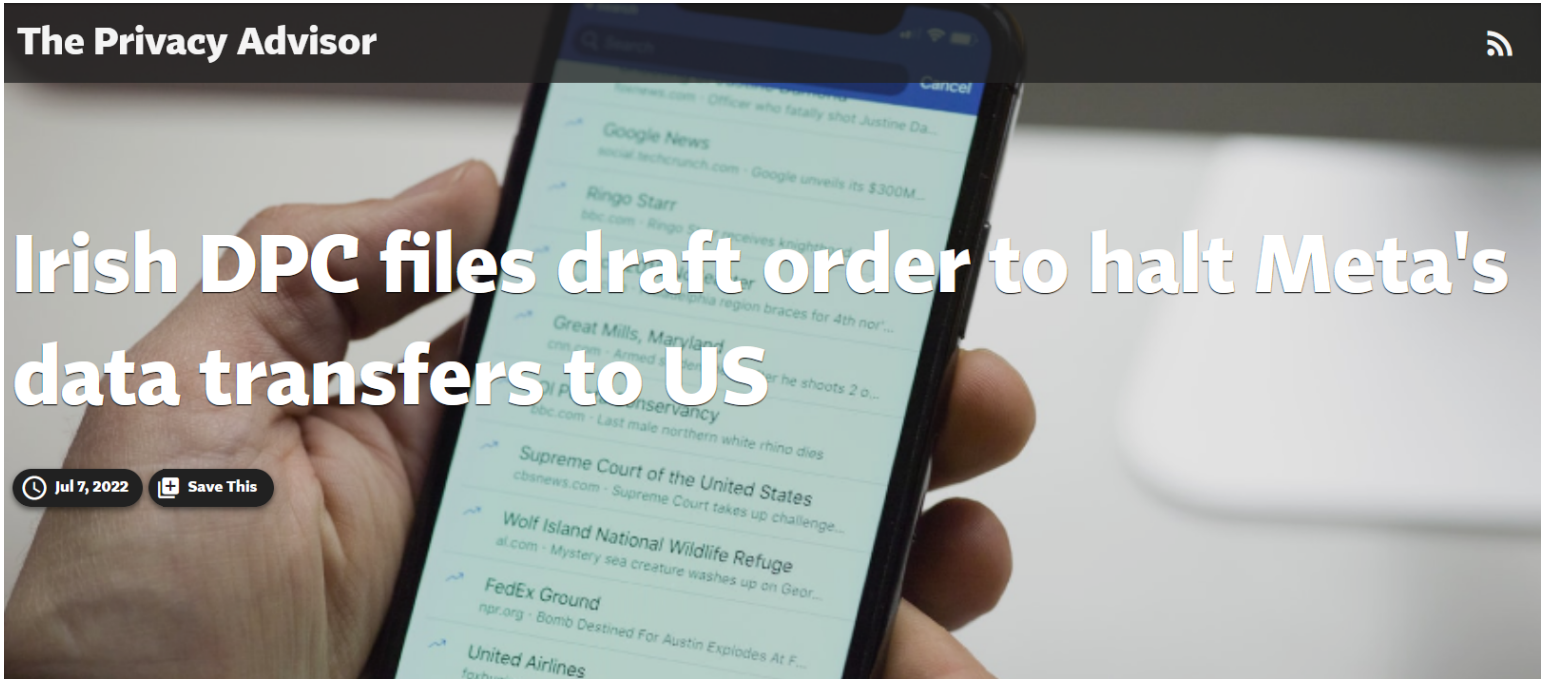
You might find it worthwhile to download the deck...

`https://github.com/capnion/random/blob/master/acm_capnion_he_compliance_utah_jul22.pdf`

...to use these links to my Email, LinkedIn, or GitHub.

GHOST PII

Very Current Events



View the article via this link.

Rewind: Adequacy and Not

Twice now roadmaps for Americans to comply with EU privacy law have been developed and then later struck down.

- Safe Harbor 1998 - October 2015
- Privacy Shield February 2016 - July 2020

The earliest round of complaints were filed with the Irish Data Protection Commissioner in 2011 and later moved to the courts.

Today we might discuss relevant issues with reference to GDPR, but the underlying problems are older and enduring.

Future Proof This: Instability Continues

US and EU reach breakthrough in data protection dispute

The powers announced an "agreement in principle" during a visit by President Biden to Europe to build transatlantic unity over the Russian invasion of Ukraine.

Alexander Martin
Technology reporter @AlexMartin

Friday 25 March 2022 11:31, UK

This *adequacy framework*, which does not yet even exist in detail, is already being discussed as legally inadequate.

GHOST PII

The Issue in the United States

The problem is the level of government digital surveillance in the United States (notably under PRISM), which is regarded as illegal under EU law when the subjects are EU citizens.

Schrems I and II

Maximilian Schrems is an activist and privacy campaigner. In 2013, he filed a long-running and impactful complaint against Facebook that eventually reached the CJEU.

Beware the “round-trip to Luxembourg” as Schrems famously called Privacy Shield early on. Was it a solution or a band-aid?

Tangible Examples

It is easy to run into problematic data transfer if you are a consumer-facing US business operating in the EU.

- basic collection of customer data
- information on EU citizens in cloud applications
- analytics involving EU citizens data

Many businesses have many of these practices integrated with one another in some way, and the more you have the more likely you were to be leaning on an adequacy framework and not a limited standard contractual clause.

The State of Affairs

We are between privacy frameworks right now, and some companies are in a bit of limbo. Common approaches...

- avoid data in the clear
- standard contractual clauses
- hope for a new framework soon
- gaslight or intimidate the Irish government

It should be noted that the privacy frameworks were really for managing a data “firehose” where it is difficult to explicitly enumerate how and why the data will be used.

EDPB Guidance: Overview

Recommendations 01/2020 on measures that supplement transfer tools to ensure compliance with the EU level of protection of personal data

Adopted on 10 November 2020

The point of the next few slides is you can probably stay out of trouble long-term if you can keep data encrypted. This used to be a severe restriction but with the advent of *encrypted data-in-use* it is not so much anymore.

EDPB Guidance: Technical Fixes 1

If you can avoid data in the clear...

> Use Case 1: Data storage for backup and other purposes that do not require access to data in the clear

Your transfer is permitted if you follow some guardrails about doing a good job.

> then the EDPB considers that the encryption performed provides an effective supplementary measure.

The notable new trick (and topic for Part 3 of this presentation) is that you can do an awful lot without data in the clear, and without otherwise learning too much about what is in the data.

EDPB Guidance: Technical Fixes 2

For some of the examples we have discussed and others...

> Use Case 6: Transfer to cloud services providers or other processors which require access to data in the clear

you are probably quite a ways from being right by the EDPB if you need data in the clear...

> then the EDPB is, considering the current state of the art, incapable of envisioning an effective technical measure to prevent that access from infringing on data subject rights. The EDPB does not

Data in the clear and not

There are many approaches out there under a million different names, sometimes overlapping.

- tokenization
- masking
- blurring
- ...

Encryption is one way, but this word actually covers much ground. The latter part of this presentation is about new methods of encryption that lets you get more done with the encrypted data.

Terminology

There are a variety of new techniques that allow you to do useful work on encrypted data without gaining access to it. In context of some older terms, you might say that you have...

- encrypted data at rest
- encrypted data in motion
- and now *encrypted data in use*

This is perhaps a little counter-intuitive and thus we will examine two encrypted data in use techniques, homomorphic encryption and zero-knowledge proof, in more detail.

For those who learn by doing...

We just (last week) open-sourced our Python encrypted data in use tool. You can tinker with it in **Jupyter notebooks via the Binder link below** if you like. Please avoid the notebooks about distance calculations as Binder uses an older version of Python that gives them trouble.

`https://mybinder.org/v2/gh/capnion/ghostpii_demos/main`

The rest of the presentation will be about what you are looking at and why it's important (although presented in much more general terms).

Big Idea: Computation without Decryption

Does “**as5ga4dsg**” decrypt the same as “**p44hdfj3jdk**”? Y/N?

The big idea is that you might be interested in

- a high-level insight which is not sensitive, or
- preparing something sensitive for someone else that you yourself do not need to see, and

These techniques allow you to get that insight without learning more about the underlying data.

Back to Basics: Simple Arithmetic

Would you be willing to do some simple arithmetic for me?

Assignment: Basic Bookkeeping

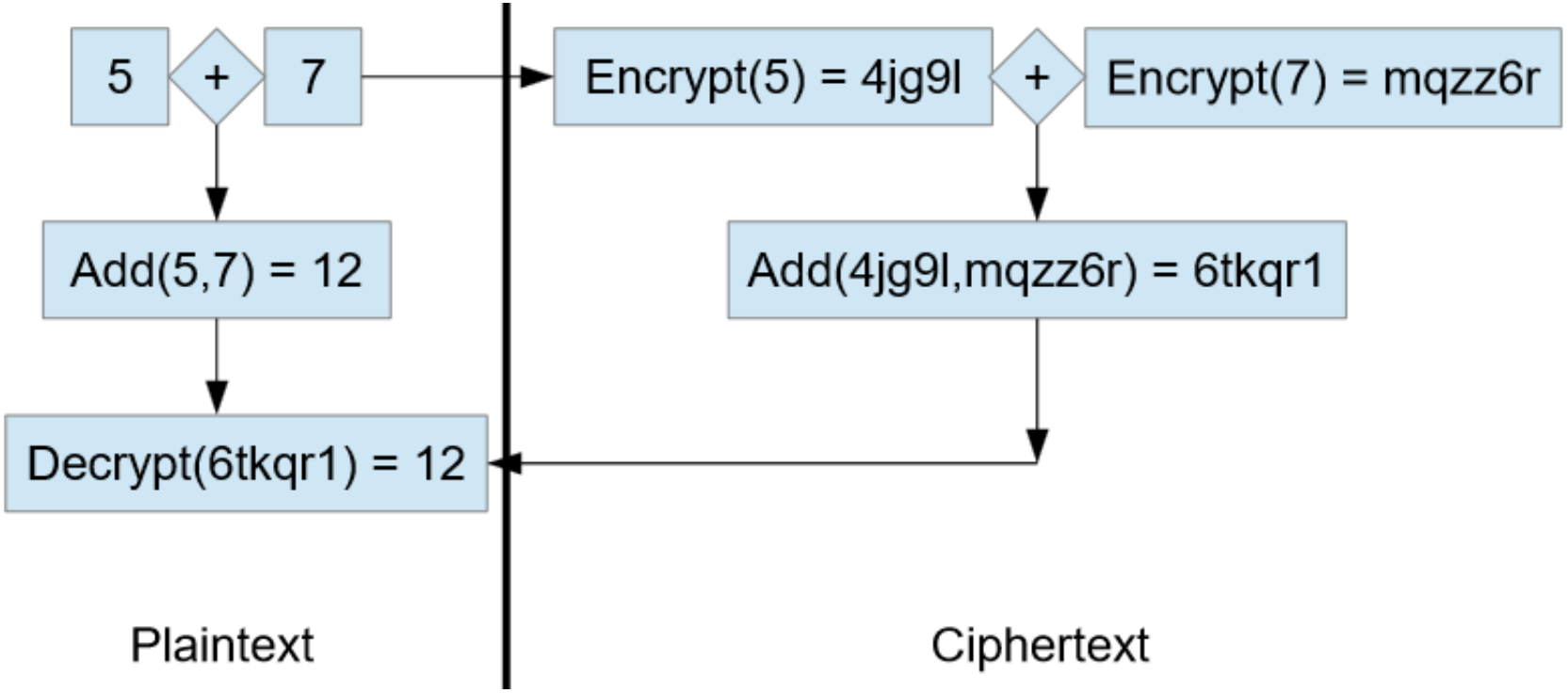
Please go ahead and...

- verify that the ending funds figure is correct, but
- maintain secrecy of expense amounts and starting funds

Expense	Amount	Starting Funds
fancy pens	XXXXXXXXX	XXXXXXXXX
private jet	XXXXXXXXX	**Ending Funds**
"business entertainment"	XXXXXXXXX	$30173
attorneys	XXXXXXXXX	

Thanks! Easy as pie, right? RIGHT?!

Homomorphic Encryption



Homomorphic encryption allows one to do computations on encrypted data and get the “correct” answer after decryption.

What is a zero knowledge proof?

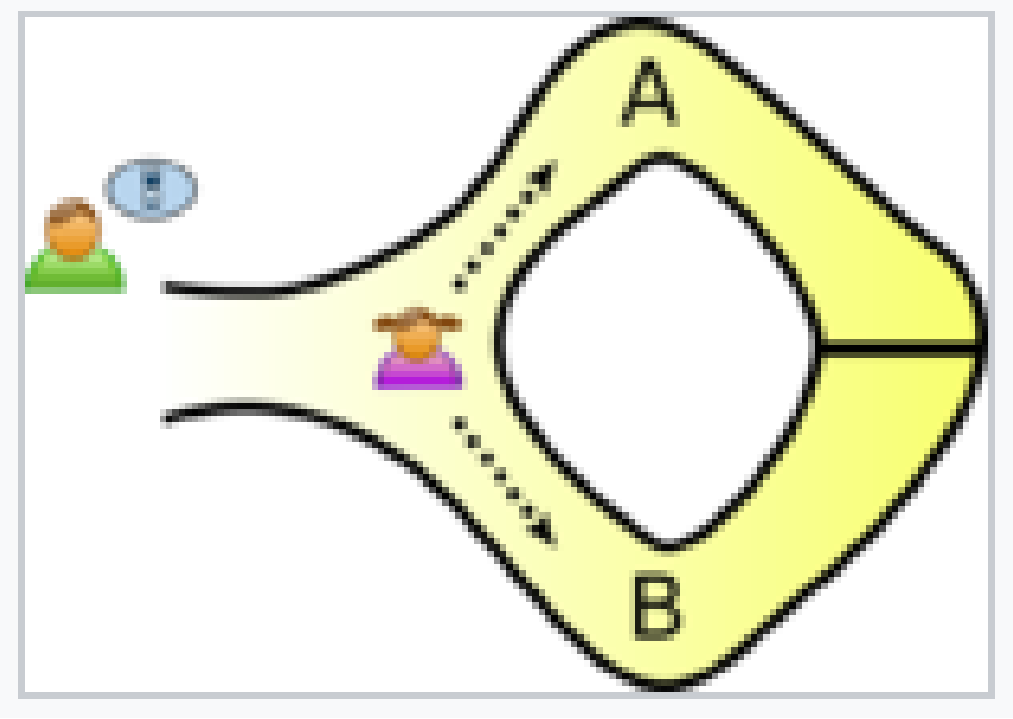
A **zero-knowledge proof** is a method by which a *prover*, traditionally called Peggy, proves to a *verifier*, traditionally called Victor, that

- Peggy possess a particular piece of information, and...
- Peggy does not reveal that information to Victor.

Example: Digital Signatures

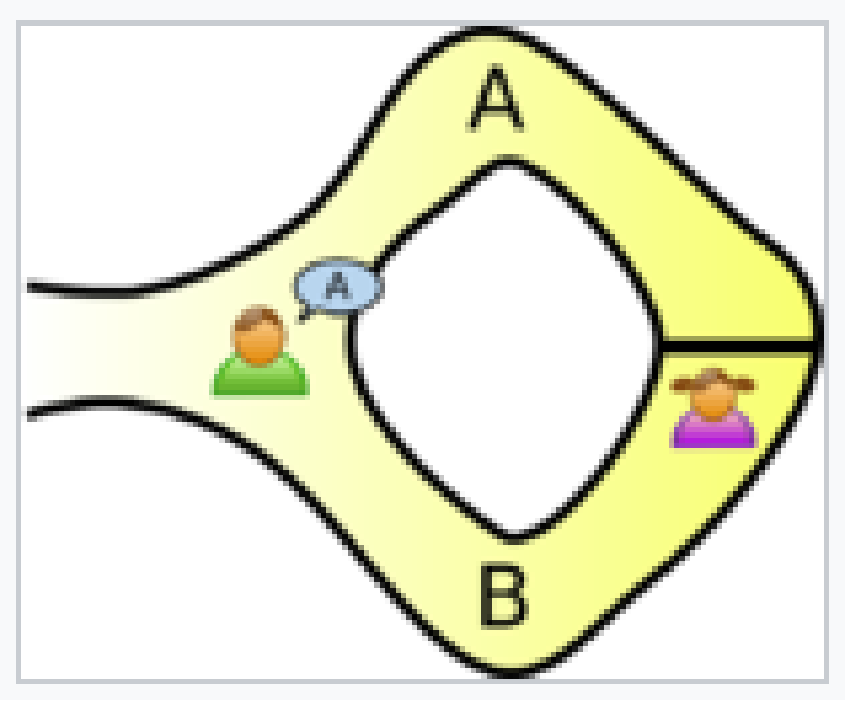
Peggy proves to Victor that she holds the private key corresponding to a particular public key but Victor does not receive any information about the private key itself.

Classic Cave Example 1



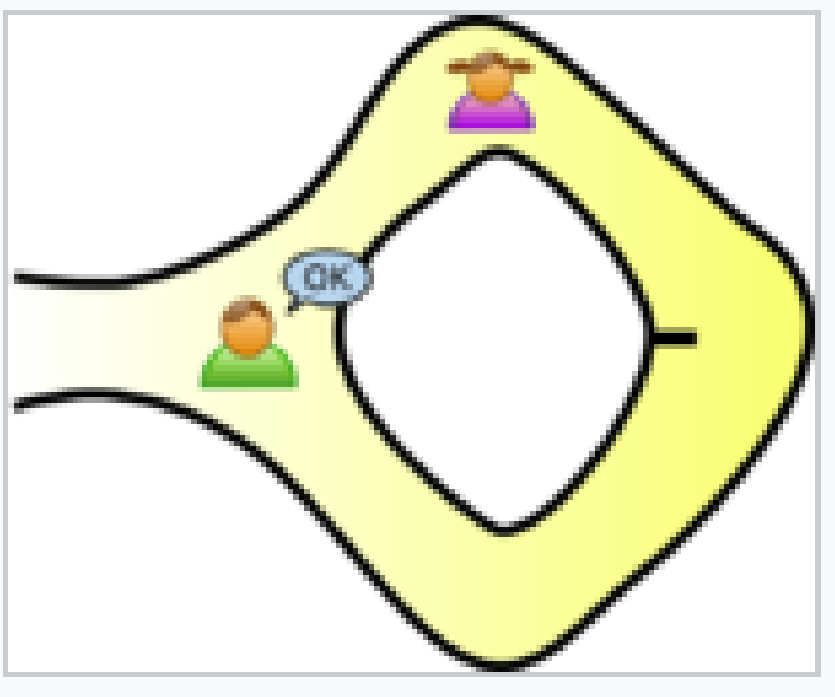
Peggy wishes to prove to Victor that she can **open the magic door** in the back of a donut-shaped cave, yet she wants to **keep her method for opening the door** a secret from Victor.

Classic Cave Example 2



After Peggy has entered the cave by one path or the other, Victor shouts into the cave whether Peggy should return by Path A or Path B.

Classic Cave Example 3



If Peggy can consistently return by the path Victor requests, she has demonstrated that she has a method for passing through the magic door, yet Victor has not seen Peggy open the door.

Basic Bookkeeping Revisited

Expense	Amount	Starting Funds
fancy pens	XXXXXXXXX	XXXXXXXXX
private jet	XXXXXXXXX	**Ending Funds**
"business entertainment"	XXXXXXXXX	$30173
attorneys	XXXXXXXXX	

Adding up the unknown, encrypted amounts to get an unknown and correct but still encrypted amount is **homomorphic encryption**.

Verifying that this encrypted amount is $30,173 without decrypting anything is a **zero-knowledge proof**.

Related Technologies

Use cases in two categories

- Federated learning: "code to data not data to code"
- Homomorphic encryption: blind data processing

Know just what you need...

Encrypt data like this...

#'F&?@!&<&)q"!%"?y#L%%[!"ei	%*1#&e#zt"m-"bQ"sq&%L!15!yV"p4	!A6"u_#d6"jR	"jZ!GU	"na!?p	!-+"Kj#-C
"^H"z7%at#iy"=`#&5%4p#RP#-t	%4?"@n"0!#>q"vf!ty!zi%E?%sE%DZ	"Oh!]2&*5#7;	%4I%dF	&90%Bx	!`S!4+%FM
#;O!.T#V2"FH%,.!sL"j!%y,#q0	":C"T&#99#@t"Ts"#N!'&"pp%z]%NP	#QL!@O!K<!VM	#H#"dt	&#F#I%	&=(%+b#jQ
"eH%e3"qB%X+%@t!2["ew%9w!ly	!-T!a%&-R"1;!M6#gX!,6&:J!S(#X(	&.H"9M&=C#,?	!*8#H#	%#6!>0	#Bj![`%+A
!zf##Z#r<%%e#@E!2c#LH!Q"#B<	#.""r"#%K#>i"X!#AW#LI#a3"MY"Wt	!.@!Jb!zh#]s	#Y5#(R	#Qb"nX	!Ou&&c"mF

...but answer questions like this...

```
#in this example we are comparing the first three letters
print(myCipherFrame[1][2][0:3]==myCipherFrame[1][1][0:3])
print(myCipherFrame[1][1][0:3]==myCipherFrame[1][0][0:3])
```

```
https://ghostpii.com/recordlink/?lowerOne=6943865&upperOne
other endpoint
False
https://ghostpii.com/recordlink/?lowerOne=6943855&upperOne
other endpoint
False
```

GHOST PII

Solution Detail

Ghost PII is a system for working with encrypted data-in-use that consists of...

- a private computation API hosting special keys in the cloud
- a Python client automating contact with this API

You can perform a computation locally on encrypted data and download an **answer key** that can **decrypt only the answer but not the underlying data**.

This client is designed to integrate easily with **common Python tools** for data engineering and analytics and require minimal retraining for those that are familiar with Python.

Federated Learning Model

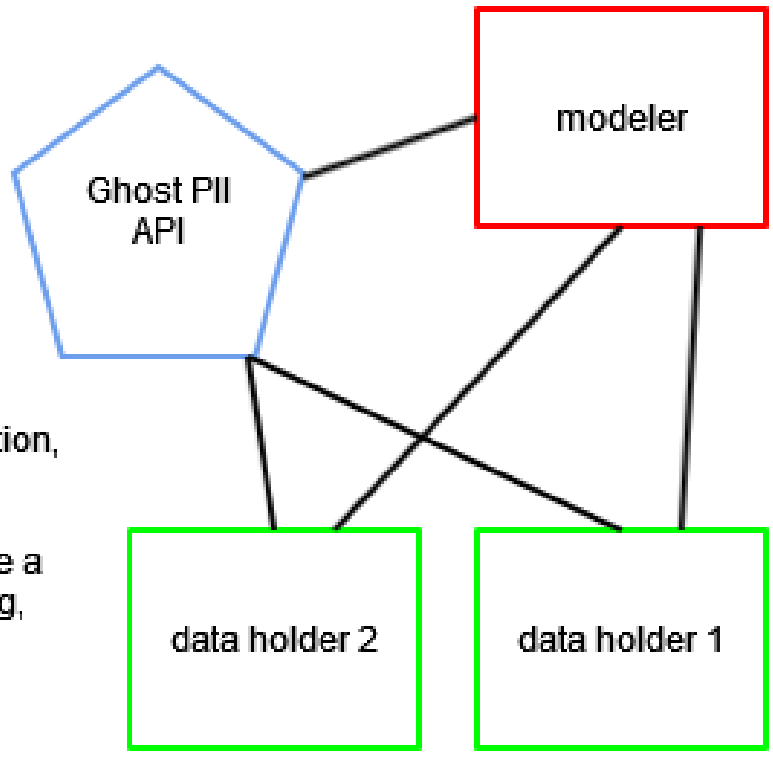
federated learning

"code to data not data to code"

Encrypted computations
allow the modeler to train as
if on the pooled data of holders
1 and 2 but only desired data,
or none, is revealed in plaintext.

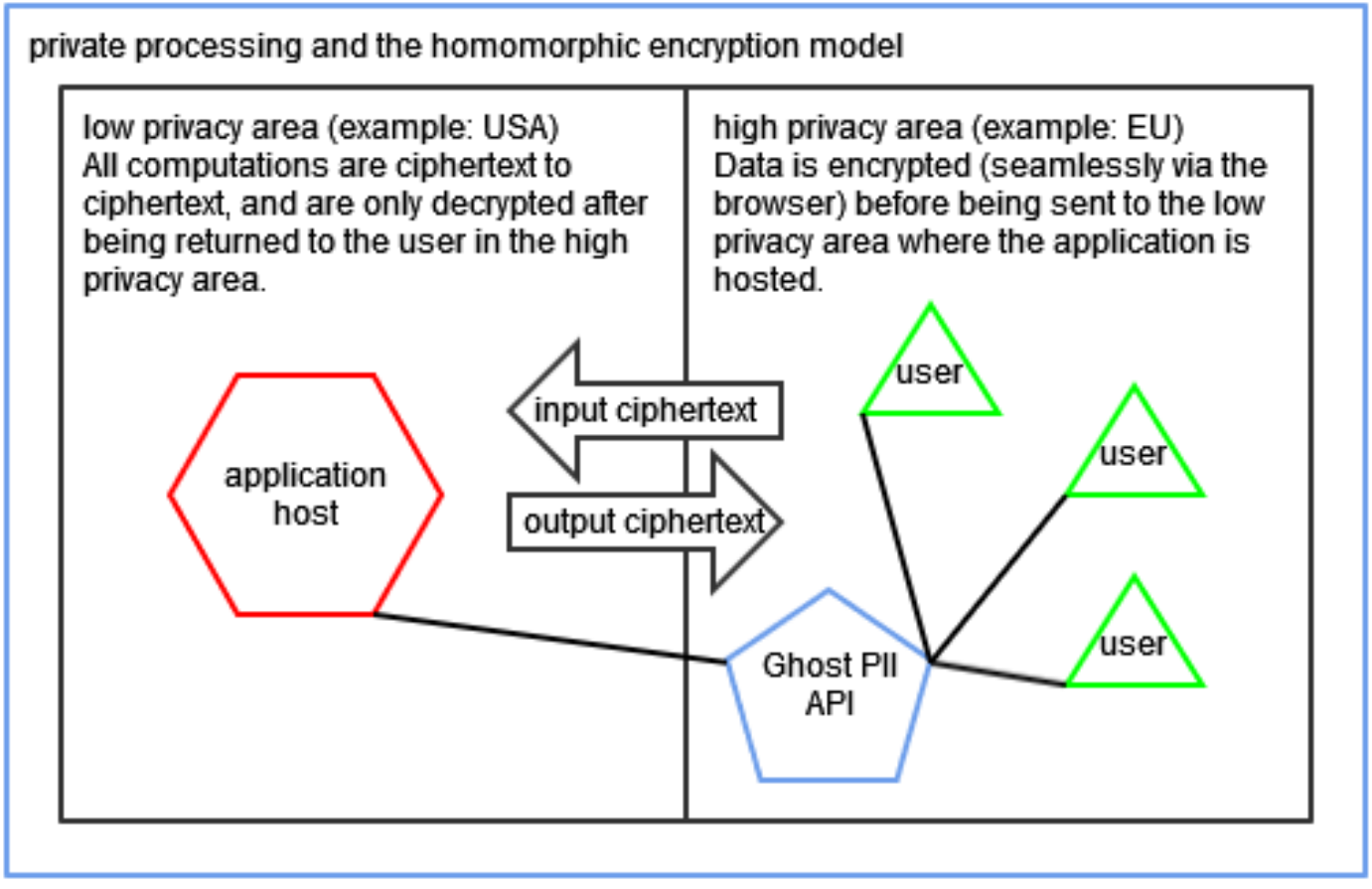
Examples:
inter hospital network collaboration,
advertising measurement.

Data holders 1 and 2 might have a
variety of concerns about pooling,
sharing, or mingling their data.



GHOST PII

Homomorphic Encryption Model



Various Tools and Organizations

Federated Learning

- OpenMined
- FATE AI

Secure multi-party computation

- CipherCore

Fully homomorphic encryption

- Palisade

I can't make too many guarantees about these, especially not about how easy (or hard) you will find them to implement.

Questions and Conversation

Any questions?

acmueller@capnion.com

https://www.linkedin.com/in/
alexander-c-mueller-phd-0272a6108/

https://github.com/capnion/ghostpii_client

GHOST PII