
AWS
re:Invent

SEC401

Privacy-preserving machine learning

Joan Feigenbaum

Amazon Scholar, AWS Cryptographic Algorithms Group

Grace Murray Hopper Professor of Computer Science, Yale University

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Motivation and vision

Our approach: Computing on encrypted data

Two prototypes

- Logistic-regression training
- Extreme Gradient Boosting prediction

Summary



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Motivating examples



A medical data-analysis company offers disease predictions to customers based on their medical records; it wants to use state-of-the-art models and must preserve customers' privacy

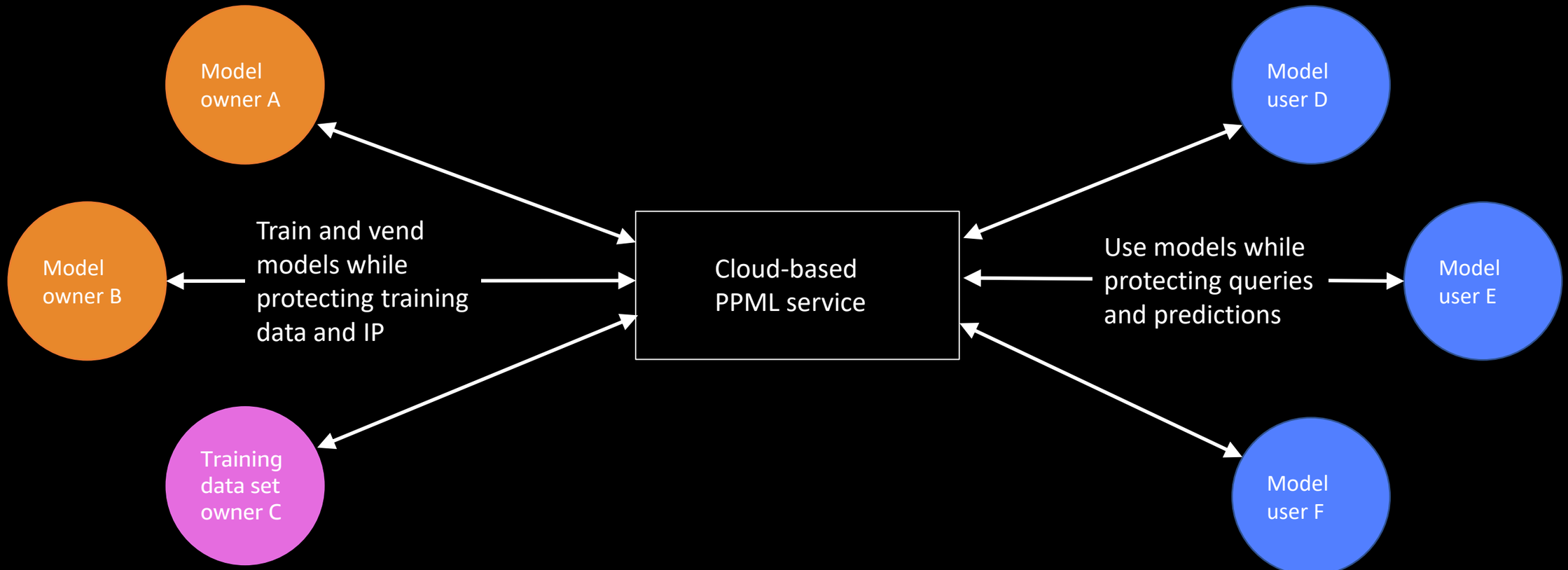


A machine-learning company has a proprietary disease-prediction model trained on data sets compiled at teaching hospitals; it wants to sell access to this model while protecting its intellectual property



A teaching hospital develops new treatments and diagnosis procedures, and it compiles patient-data sets; it wants to make these data sets available for model training but must preserve patients' privacy

Vision: A privacy-preserving machine learning (PPML) marketplace



Approach: Perform training and prediction on **encrypted data**

Our approach: computing on encrypted data



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

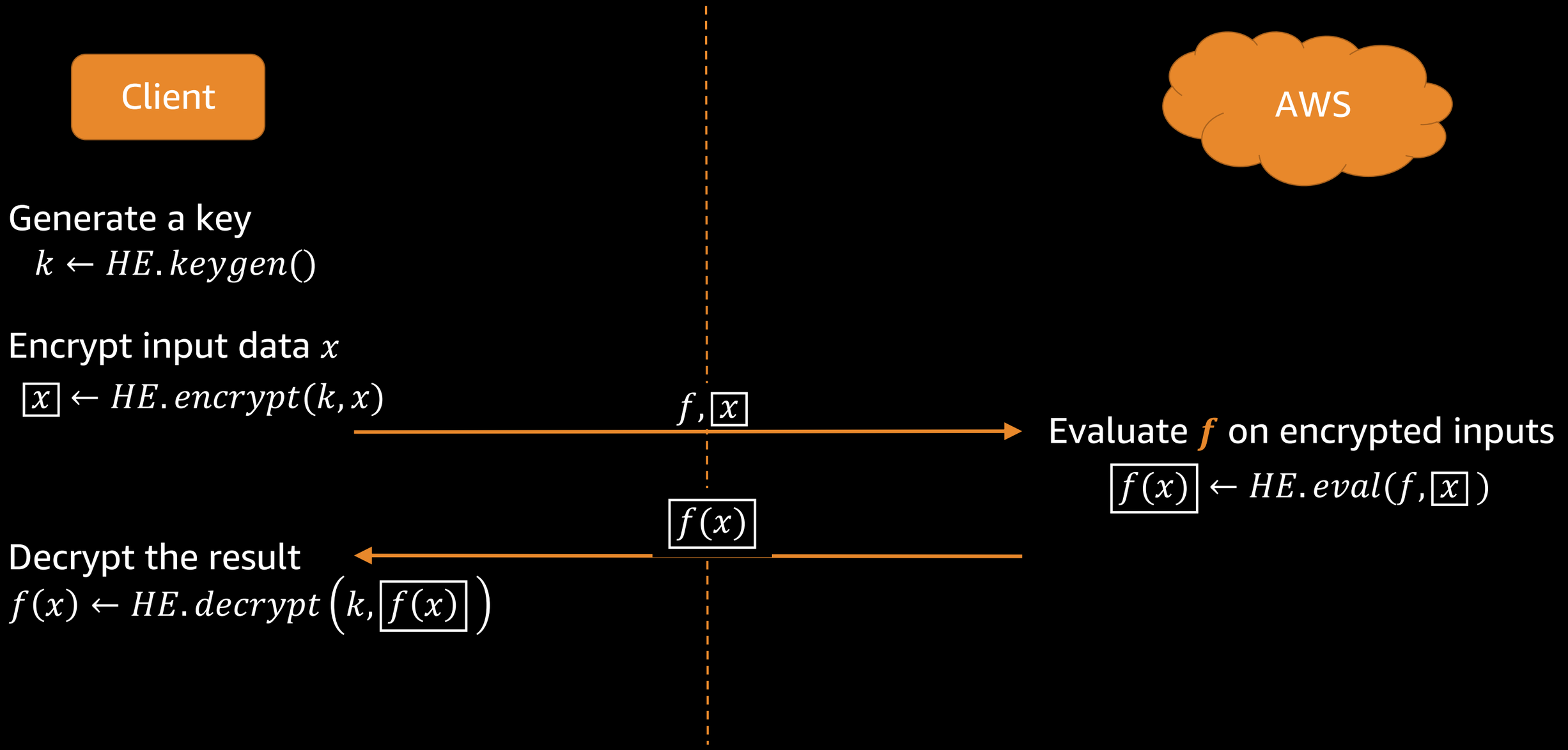
Homomorphic encryption

- A fully homomorphic encryption scheme supports **addition** and **multiplication** on ciphertexts
- Craig Gentry exhibited an FHE encryption scheme in 2009

```
C1 <— FHE.Enc(m1)
C2 <— FHE.Enc(m2)
C’  <—  C1 + C2        [C’  = FHE.Enc(m1 + m2)]
C” <—  C1 * C2         [C” = FHE.Enc(m1 * m2)]
```

- Recall that + and * are a **complete** set of operations [arithmetic circuits]
- For **any** function f, Y <— FHE.Dec(f(FHE.Enc(X))) [Y = f(X)]
- FHE is not (yet!) fast enough to be used in all applications, but there are fast SHE schemes for some f that are useful in PPML [S for somewhat]

Generic deployment model



Two prototypes



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Logistic regression on encrypted training data

(ERIC CROCKETT, APPLIED SCIENTIST, AWS CRYPTOGRAPHIC ALGORITHMS GROUP)

- In statistics, the **logistic model** (or **logit model**) is used to model the probability of a certain class or event existing: pass/fail, win/lose, **likely/not likely to get a certain disease**
- Training process
  - Input: A set of labeled (1/0) instances
  - Output: A model, i.e., function that maps an (as-yet unlabeled) instance to 1 or 0
- Previous work
  - Use FHE (based on Gentry) to encrypt training-data set
  - Use an **arithmetic circuit** to produce an encrypted model of the required form
- Goal: Construct arithmetic circuits that are more efficient

HELR results

- How can an **arithmetic circuit** perform LR training?

  Approximate a **sigmoid function** $\sigma(x) = \frac{1}{1+e^{-x}}$

  with a **polynomial** such as $c_3 \cdot x^3 + c_1 \cdot x + c_0$

- Crockett's faster training method **halves circuits' multiplicative depth**

	Multiplicative depth	Multiplications	Multiplicative width	Training time (s)
Prev. best '18	30	207	6.9	77
Crockett '20	15	507	33.8	16

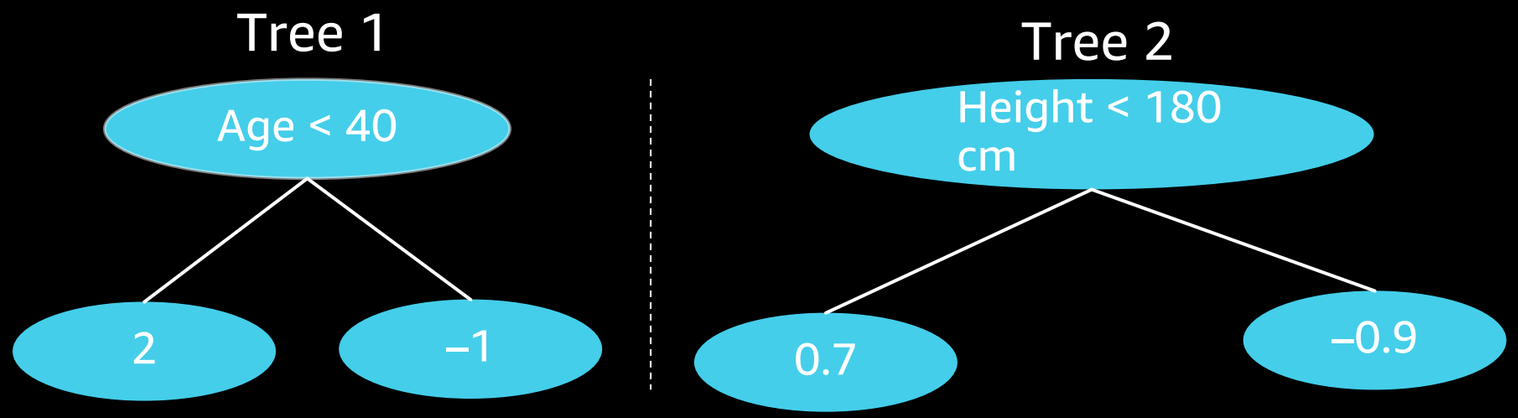
- Nonstandard threat model in our use case: Compromise of data scientists' computers may lead to exfiltration of customers' data lake

XGBoost prediction on encrypted queries

(XIANRUI MENG [APPLIED SCIENTIST, AWS CRYPTOGRAPHIC ALGORITHMS GROUP] AND JF)

Model produced by XGBoost training is **a set of classification and regression trees** (CARTs)

- Each CART assigns to each input query a (real) numerical score
- The prediction algorithm **aggregates (sum or softmax) the set of scores** output by the CARTs



Query: [age = 36, height = 175]

XGBoost prediction = 2 + 0.7 = 2.7

PPXGBoost techniques and results

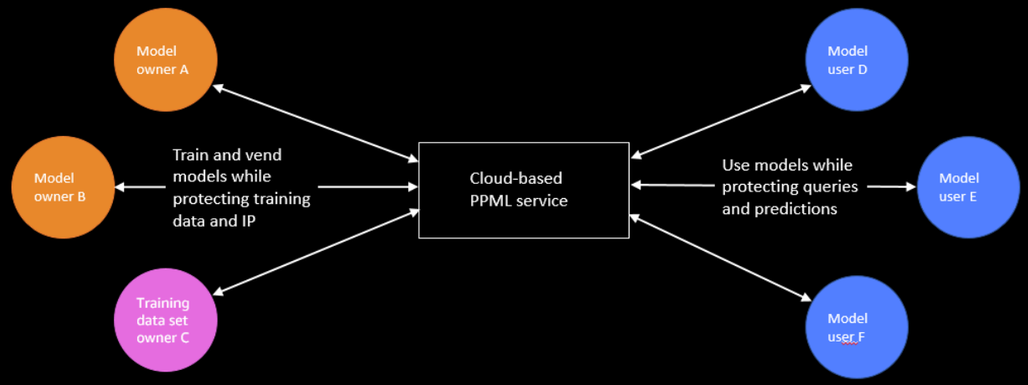
- A personalized **encrypted model** is created for each user
- Model encryption and query processing both use
  - **Order-preserving encryption** (for all *k*, $x < y$ if and only if $Enc(k,x) < Enc(k,y)$)
  - **An SHE scheme that is additive and asymmetric**
- Performance

Dataset	XGBoost query time	PPXGBoost query time	XGBoost model size	PPXGBoost model size
Amazon	1 ms	0.43 s	506 KB	4.2 MB
Titanic	<1 ms	0.32 s	3 KB	12 KB
US Census	1 ms	0.49 s	210 KB	2.5 MB

- Efficient enough for many use cases, e.g., phone-based apps that submit queries to servers and our disease-prediction example

Summary

- Vision: Cloud-based PPML marketplace



- Promising prototypes of HELR training and PPXGBoost prediction
- Short-term plan: Release improved versions of these prototypes
- Ongoing work: Applications in which HE might be used include
  - Federated learning
  - Advertising metrics
  - Fraud detection

Thank you!

Joan Feigenbaum

jfbaum@amazon.com

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.