**Motivation**

**Goal:** compute best-fit curve for a large number of data points revealing no other information about the input.

**Recommendation system**
- Users submit encrypted ratings
- Users get recommendations

**Medical research**
- Users submit encrypted surveys
- Researcher learns a statistical model

**Summary**

- **Contributions:**
  - Built a system to compute ridge regression preserving privacy (find best-fit curve for a collection of encrypted data points).
  - Achieved excellent scalability in the number of users (1,000,000 users).
  - Showed good performance on real datasets.

- **Challenges:**
  - Method should be efficiently represented with a boolean circuit.
  - All operations should be data oblivious.

- **Tools:**
  - Yao’s garbled circuits.
  - Homomorphic encryption.

**Ridge Regression**

Given \( n \) pairs of variables \((x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\), learn a linear function \( f : \mathbb{R}^d \rightarrow \mathbb{R} \), s.t. \( y_i = f(x_i) \), where \( x_i, \beta \in \mathbb{R}^d \).

\( \beta \) is fit to the data by minimizing the following quadratic form:

\[
F(\beta) = \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \| \beta \|^2.
\]

**Cholesky Decomposition**

The minimizer, \( \beta \), can be computed by solving the linear system \( A \beta = b \), where \( A = X^T X + \lambda I \in \mathbb{R}^{d \times d} \) and \( b = X^T y \in \mathbb{R}^d \).

\( A \) is Symmetric Positive Definite and efficient solution can be found through Cholesky decomposition.

**Cholesky Decomposition**

Computes lower triangular matrix \( L \), s.t. \( A = L^T L \).

- \( \Theta(d^2) \) additions
- \( \Theta(d^2) \) multiplications
- \( \Theta(d^2) \) divisions
- \( \Theta(d) \) square roots

**Solving Linear System**

Input: \( L, b \), s.t. \( A = L^T L \).

Output: \( x, s.t. \ Ax = b \).

Solve triangular system: \( L^T y = b \) Solve triangular system: \( L x = y \)

**Cryptographic primitives**

**Garbled circuits:**

Oblivious transfer (OT):

**Homomorphic Encryption:**

\( HE(a) \oplus HE(b) = HE(a + b) \)

Allows to carry one type of operation on the ciphertexts.

**System design**

**Crypto Service Provider**

1. Create and send (\( \xi \))
2. Generate and broadcast keys \( (pk, sk) = HE(\xi) \)
3. Receive and decrypt \( a \) and \( b \) (assuming \( HE(a+b) \) is received)
4. Send \( G(A + mask), G(b + mask) \)
5. OT on mask

**Evaluation**

1. Receive circuit \( C \)
2. Compute \( \omega_i = \sum_{i=1}^{n} y_i \), \( \omega_f = \sum_{i=1}^{n} x_i y_i \), \( \omega_{i,i} = \sum_{i=1}^{n} x_i^2 \)\( \omega_{j,j} = \sum_{i=1}^{n} y_i^2 \), \( \omega_{i,j} = \sum_{i=1}^{n} x_i y_i \), \( \omega_{j,i} = \sum_{i=1}^{n} x_j y_i \)
3. Send these values to CSP
4. Receive inputs:
   - \( (\alpha, \beta) \)
   - \( y \in \mathbb{R} \)
5. OT on mask: receive \( G(mask) \)
6. Compute circuit \( \omega_i + \frac{G(mask)}{G(b + mask)} \)

**Performance Evaluation**

- **Experimental results using UCI datasets, target error \( 10^{-5} \)**

<table>
<thead>
<tr>
<th>Name</th>
<th>n</th>
<th>d</th>
<th>Bits</th>
<th>Comm. (MB)</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>automobile</td>
<td>205</td>
<td>14</td>
<td>3189</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>communities</td>
<td>204</td>
<td>20</td>
<td>2134</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>concrete</td>
<td>1000</td>
<td>19</td>
<td>27</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>forest fires</td>
<td>517</td>
<td>12</td>
<td>23</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>insurance</td>
<td>512</td>
<td>14</td>
<td>21</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>flare</td>
<td>243</td>
<td>20</td>
<td>1760</td>
<td>92</td>
<td></td>
</tr>
</tbody>
</table>

**Conclusion**

- On commodity server a regression model for 100 million user records, 20 features, runs in 8.75 hours.
- Further research: matrix factorization, support vector machines, logistic regression, etc.