Trust in the AI world

Trust: Eye Networks Shared Insights 2019

Ted Tøraasen, Microsoft

?

- What is Artificial intelligence (AI)
- Can we trust AI to solve our problems
- Do we need to give up all privacy in the new brave AI world

What is Artificial intelligence

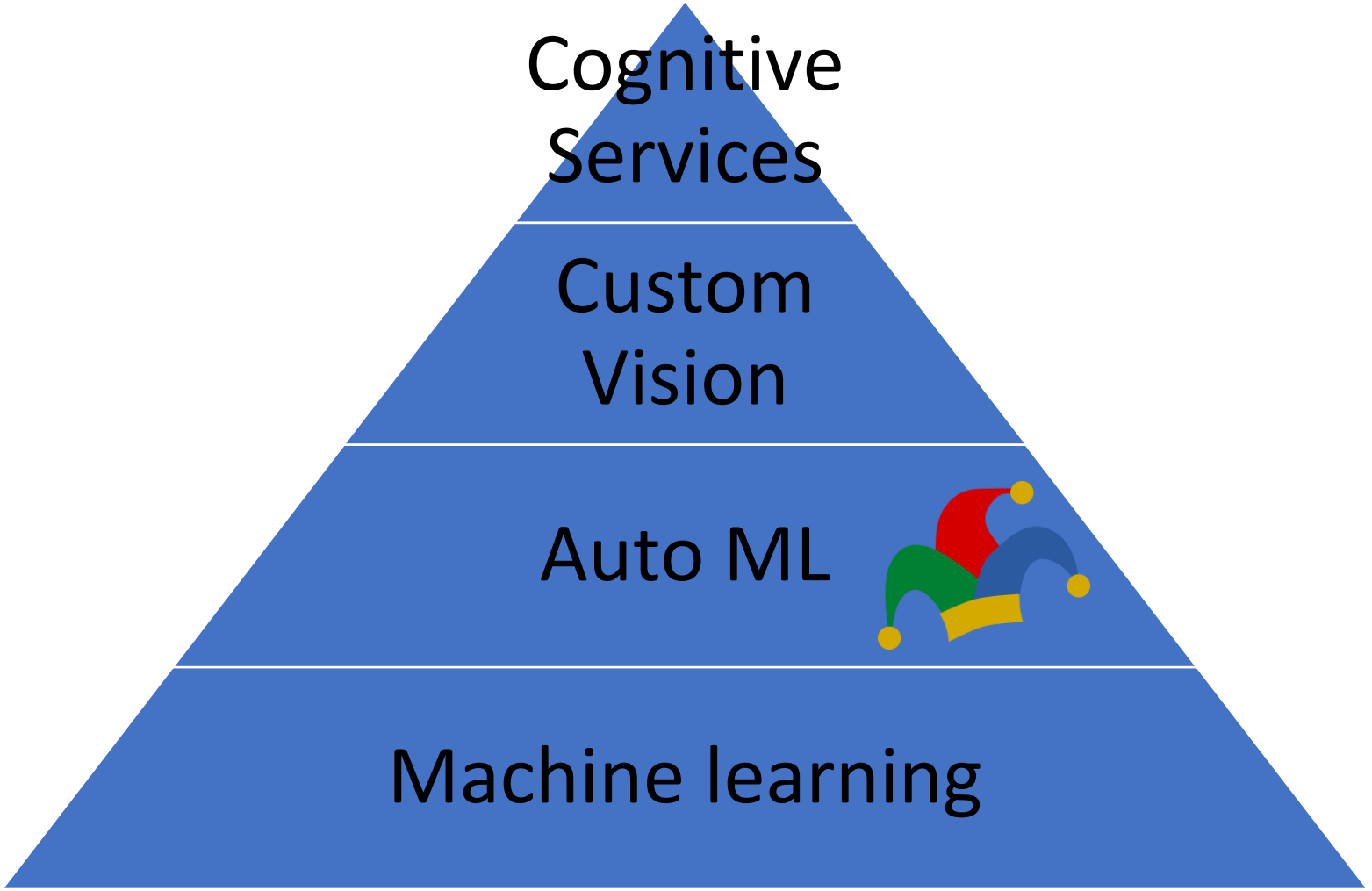
Artificial intelligence (AI) refers to systems that show intelligent behaviour: by analysing their environment they can perform various tasks with some degree of autonomy to achieve specific goals.

European Commission - Factsheet: Artificial Intelligence for Europe

- Not Terminator or West World, domain specific
- Most of us use it every day
  - Personal assistants (Alexa, Cortana, Siri)
  - Language translation services
  - Image recognition
  - Recommendation services (Netflix, Amazon)
- Mostly based on machine learning
- Tend to need a **lot** of data
- Ideal: like us, only better

```
# ##############################################################################
# Train a SVM classification model

print("Fitting the classifier to the training set")
t0 = time()
param_grid = {'C': [1e3, 5e3, 1e4, 5e4, 1e5],
              'gamma': [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1], }
clf = GridSearchCV(SVC(kernel='rbf', class_weight='balanced'),
                   param_grid, cv=5)
clf = clf.fit(X_train_pca, y_train)
print("done in %0.3fs" % (time() - t0))
print("Best estimator found by grid search:")
print(clf.best_estimator_)


# ##############################################################################
# Quantitative evaluation of the model quality on the test set

print("Predicting people's names on the test set")
t0 = time()
y_pred = clf.predict(X_test_pca)
print("done in %0.3fs" % (time() - t0))

print(classification_report(y_test, y_pred, target_names=target_names))
print(confusion_matrix(y_test, y_pred, labels=range(n_classes)))


# ##############################################################################
# Qualitative evaluation of the predictions using matplotlib

def plot_gallery(images, titles, h, w, n_row=3, n_col=4):
    """Helper function to plot a gallery of portraits"""
    plt.figure(figsize=(1.8 * n_col, 2.4 * n_row))
    plt.subplots_adjust(bottom=0, left=.01, right=.99, top=.90, hspace=.35)
    for i in range(n_row * n_col):
        plt.subplot(n_row, n_col, i + 1)
        plt.imshow(images[i].reshape((h, w)), cmap=plt.cm.gray)
        plt.title(titles[i], size=12)
        plt.xticks(())
        plt.yticks(())


# plot the result of the prediction on a portion of the test set

def title(y_pred, y_test, target_names, i):
    pred_name = target_names[y_pred[i]].rsplit(' ', 1)[-1]
```

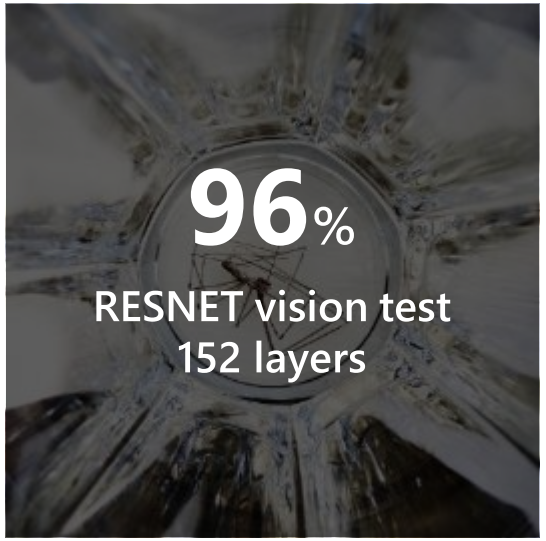
Different levels of Machine learning/AI



Examples in Azure context

Microsoft AI breakthroughs

Vision

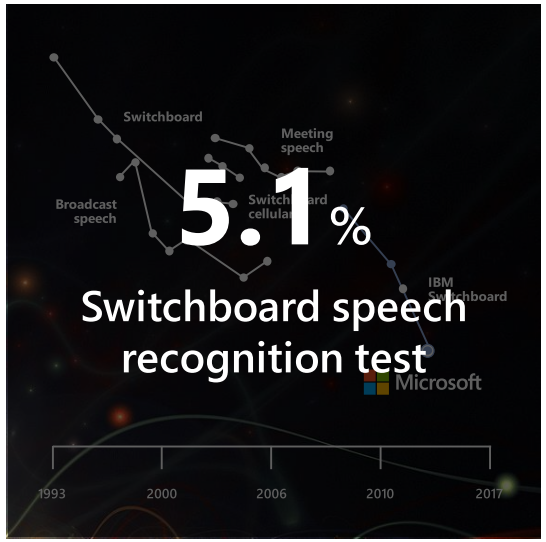


2016

First to achieve

Object recognition
Human parity

Speech



2017

First to achieve

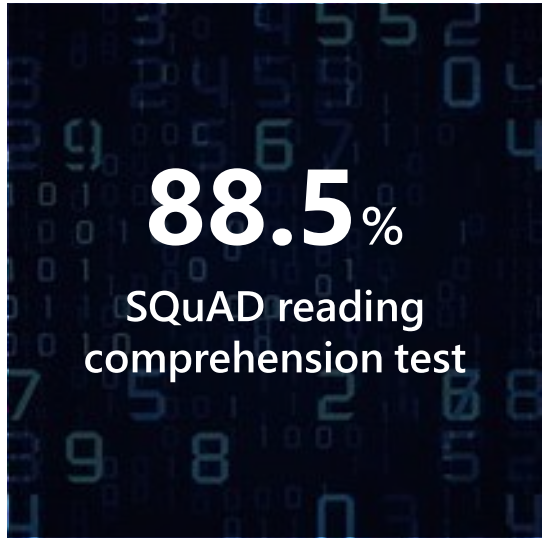
Speech recognition
Human parity

March 2018

First to achieve

Machine translation
Human parity

Language



January 2018

First to achieve

Machine reading comprehension
Human parity

Can we trust AI ?

Human parity, right?

What could possibly go wrong?

What did we learn?

What did Tay teach us?

What the world thinks:

Tay was an experiment gone wrong.

Example of bad design.

We did not do our due diligence.

"It's 2016. If you're not asking yourself 'how could this be used to hurt someone' in your design/engineering process, you've failed."

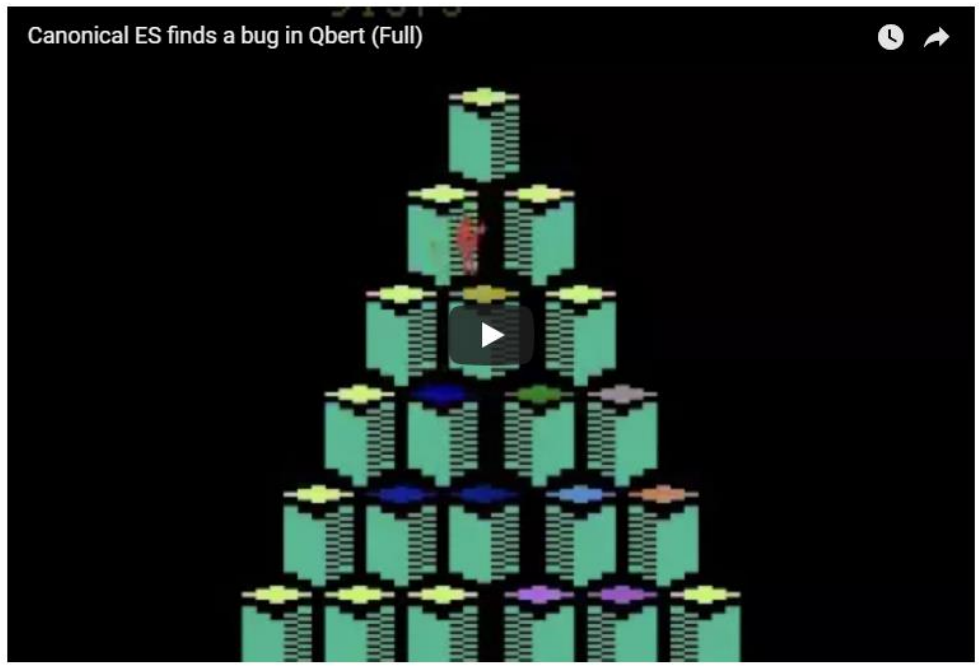
Zoe Quinn
Co-Founder
Crash Override Network
A crisis helpline, advocacy group and resource center for people who are experiencing online abuse.

"Thinking About You" (TAY)

Gaming AI beats human top scores by cheating

The artificial intelligence system had no qualms about exploiting ancient bugs to win.

By Charlie Osborne for Between the Lines | March 2, 2018 -- 11:12 GMT (03:12 PST) | Topic: Innovation



RELATED STORIES

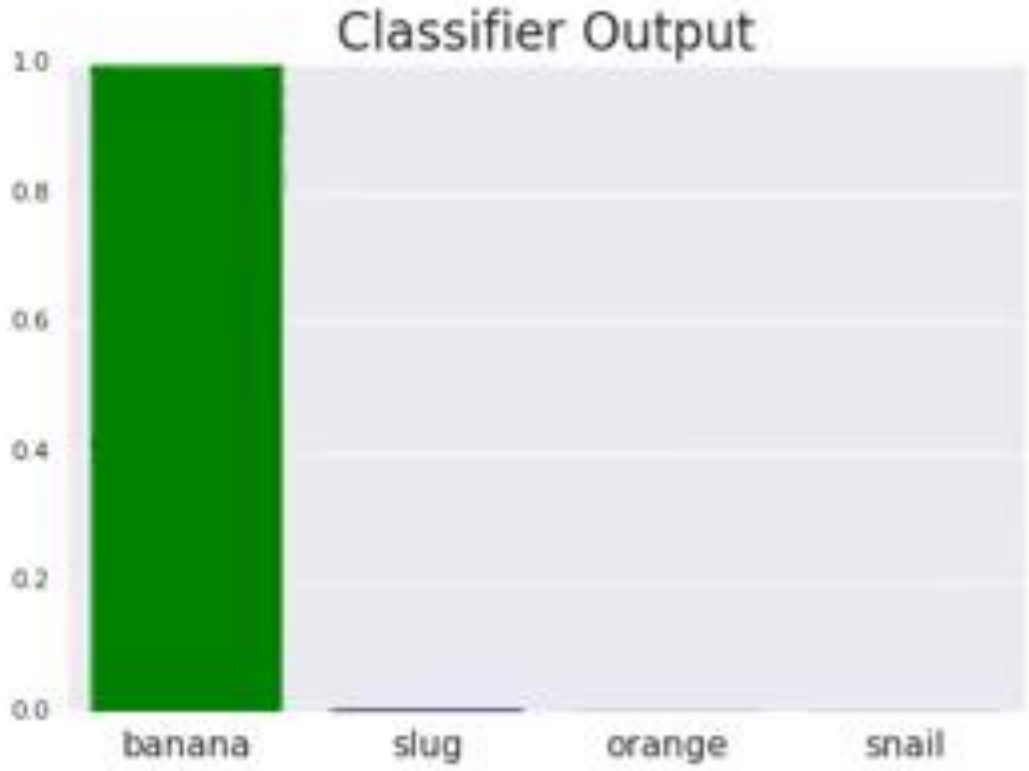

Robotics
Artificial empathy: Call center employees are using voice analytics to predict how you feel


Innovation
Gatwick Airport to trial British self-driving car system from Oxbotica


Mobility
Senate committee recommends expanding copyright safe harbours


Innovation
MicroLED patents nearly double in a year in South Korea

Adversarial Patch vs VGG16

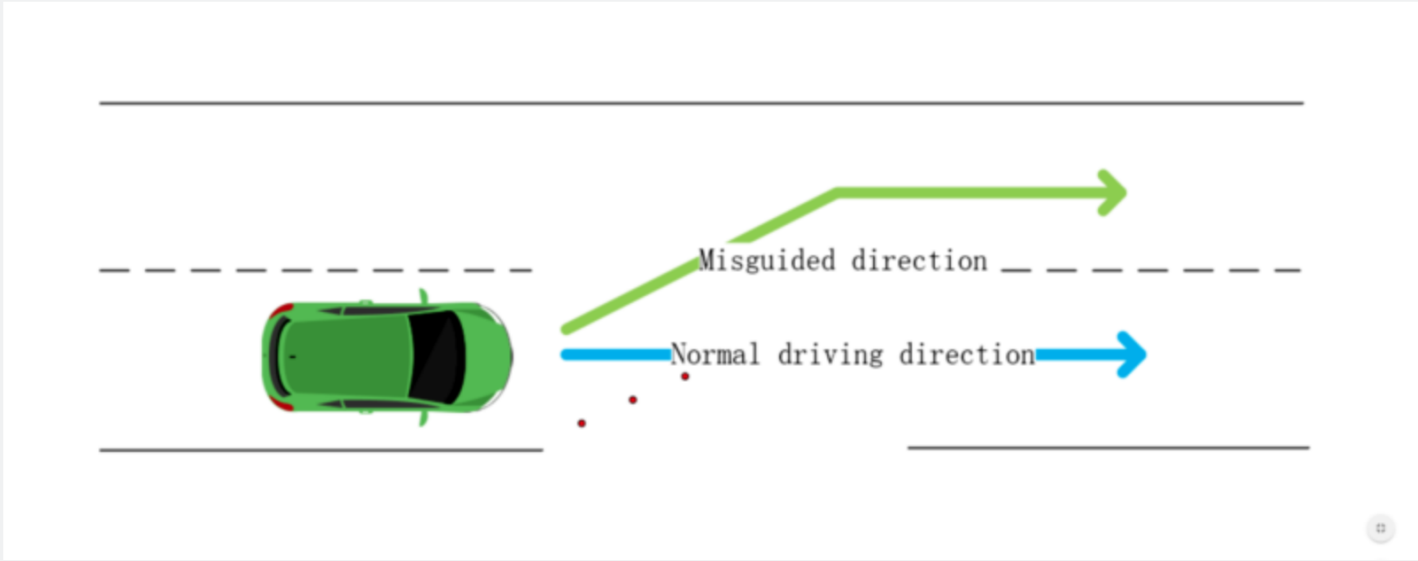


TESLA AUTOPILOT —

Researchers trick Tesla Autopilot into steering into oncoming traffic

Stickers that are invisible to drivers and fool autopilot.

DAN GOODIN - 4/2/2019, 2:50 AM



Keen Security Lab

Enlarge

Do we need to give up all our data?

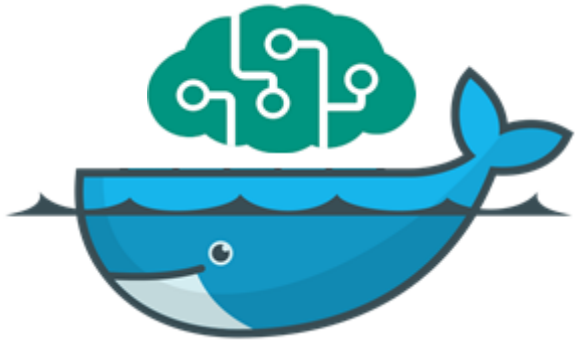
More data is always better

Two sides

- Service creator (Company)
- Service user (end user)

- Some ways to alleviate the pain
  - Transferred learning
  - Models in containers
  - Homomorphic encryption

Container Support for Azure Cognitive Services



Why Containerization?

Enterprises interested in Azure AI, but:

- Unable to load all their data into the cloud.
- Regulatory requirements on handling customer data.
- Security & Privacy concerns
- Low Bandwidth or intermittently connected environments

Container enables new possibilities:

- Run AI locally in their own context, and in their own network.
- Easier to deploy and manage software.
- It opens the doors and democratizes these AI techniques to a variety of people who may not have been able to use them before.
- Makes AI portable so that it could go to a variety of different environments.
- Provides a high-throughput scenario for faster operations

Customer Benefits

- **Control over data:** It is essential for customers that cannot send data to the cloud but need access to Cognitive Services technology.

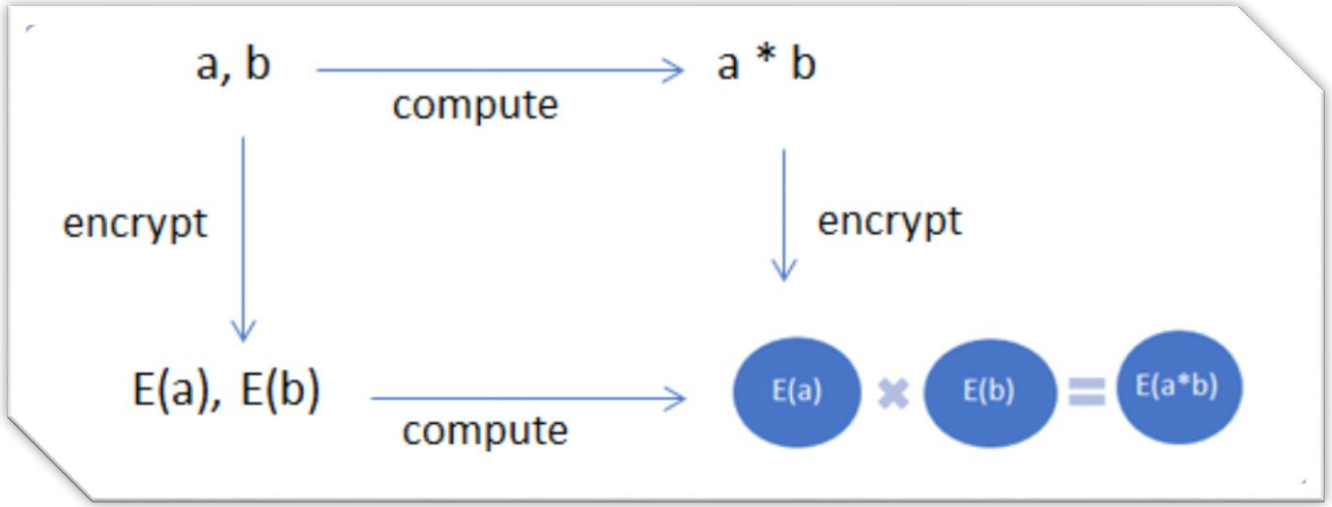
- **Control over model updates**: Provide customers flexibility in versioning and updating of models deployed in their solutions.

- **Portable architecture**: Enable the creation of a portable application architecture that can be deployed in the cloud, on-premises and the edge.

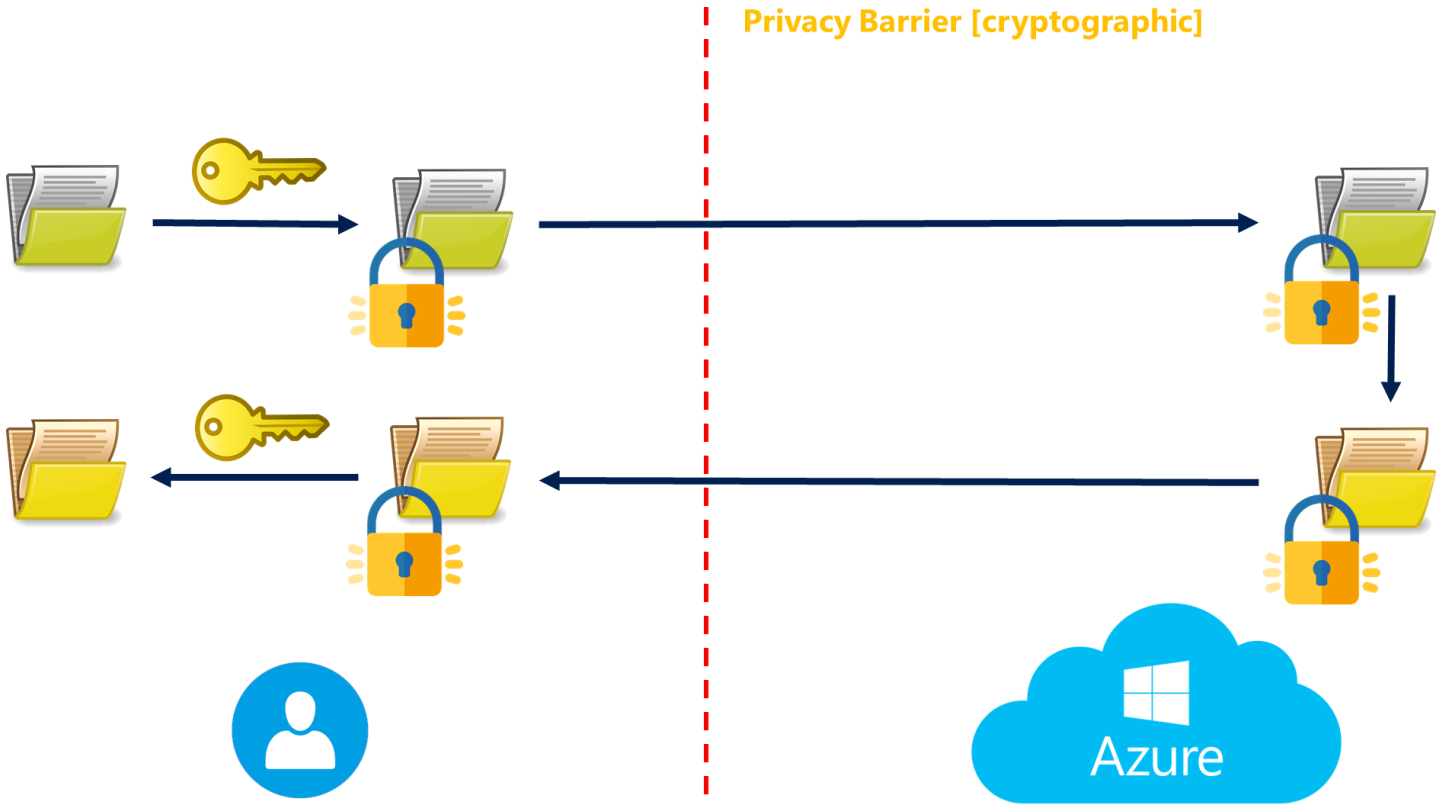
- **High throughput / low latency**: Provide customers the ability to scale for high throughput, low latency, requirements by enabling to run with more resources.

Homomorphic Encryption (HE)

- Computation on encrypted data without decrypting it!
- 2009: Considered impractical
- 2011: Surprise breakthrough at Microsoft Research
- Widespread enthusiasm about results
- 2016 breakthrough: neural nets on encrypted data!



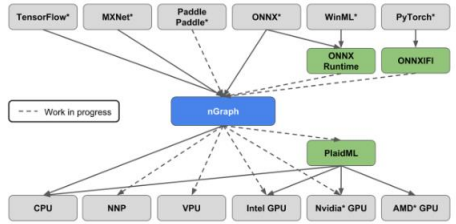
FHE Promise



HE-TRANSFORMER FOR NGRAPH: ENABLING DEEP LEARNING ON ENCRYPTED DATA

We are pleased to announce the open source release of HE-Transformer, a homomorphic encryption (HE) backend to nGraph, Intel's neural network compiler. HE allows computation on encrypted data. This capability, when applied to machine learning, allows data owners to gain valuable insights without exposing the underlying data; alternatively, it can enable model owners to protect their models by deploying them in encrypted form. HE-transformer is a research tool that enables data scientists to develop neural networks on popular open-source frameworks, such as TensorFlow*, then easily deploy them to operate on encrypted data.

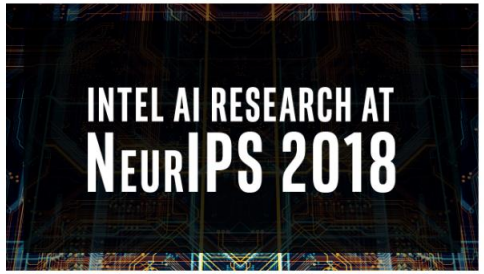
We are also pleased to announce that HE-Transformer uses the Simple Encrypted Arithmetic Library (SEAL) from Microsoft Research to implement the underlying cryptography functions. Microsoft has just released SEAL as open source, a significant contribution to the community. "We are excited to work



NGRAPH COMPILER STACK-BETA RELEASE

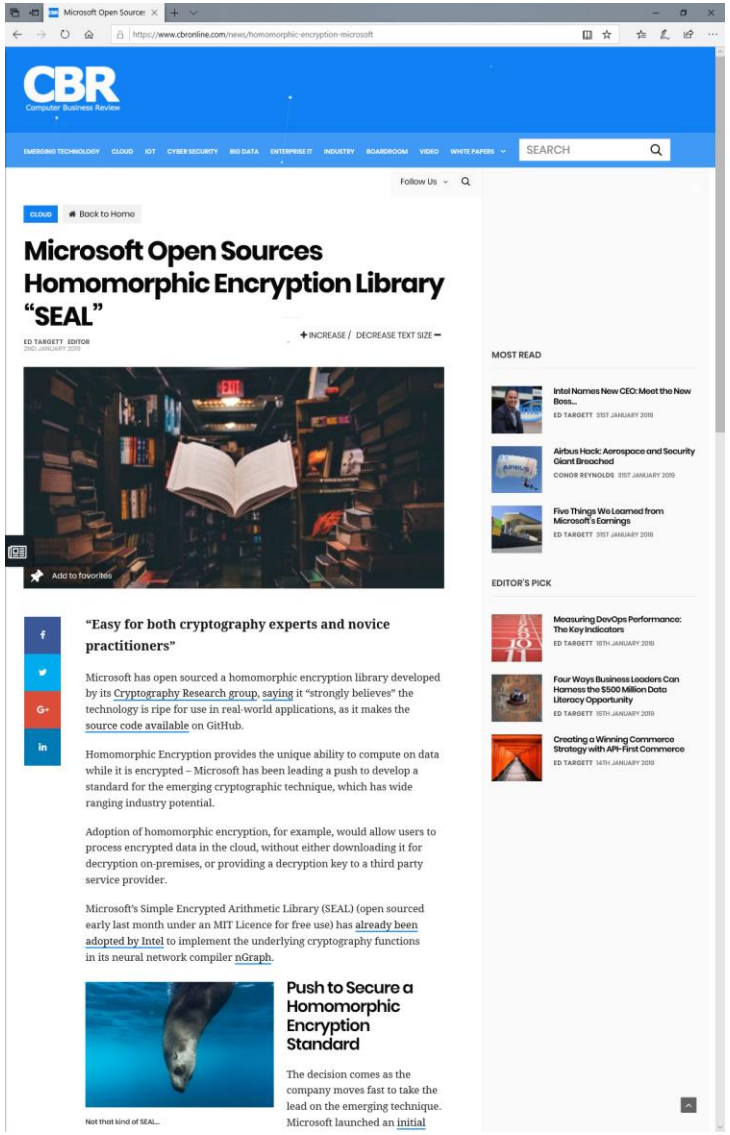
Deep Learning (DL) computational performance is critical for scientists and engineers applying deep learning techniques to many challenges in healthcare,...

Read More



INTEL AI RESEARCH AT NEURIPS 2018

Now that the excitement of the 32nd



Microsoft SEAL

- Homomorphic Encryption library by MSR
- First release in 2015; actively developed today
- Open source (MIT license)
- https://github.com/Microsoft/SEAL
- Developed in C++17
- .NET wrappers n SEAL 3.2, supporting Azure Functions
  - **Serverless Secure Compute**
- By far the most sophisticated HE library available
- Post-quantum secure
- GPU/FPGA acceleration

Key takeaways

- AI/ML don’t see the world as humans
- Behavior may change over time
- Like any other systems they will get attacked

- Inference can be done without revealing data

AI Threat Modeling Q&A

- What is your AI's specific use scenarios? Is it voice and/or gesture driven?
- Document your assumptions about expected behavior and failure scenarios:
  - AI interactions:
    - How is a user expected to interact with the AI?
    - What is considered an accepted reasonable outcome?
    - what is upper and lower bounds to acceptable behavior?
    - What should be the response to a negative behavior?
  - User interactions
    - What kind of assumptions does your AI have about the intended user?
    - How does your AI handle misdirected requests or commands?
    - Have these assumptions been defined and described by an individual or have they been peer-reviewed?
    - Do these assumptions align to the company's core values and the AI principles?
  - Attacker interactions
    - How could I make your AI look bad to others?
    - How could I prevent your AI from servicing others?
    - How can I make your AI give up information about itself or others?
    - How can I make your AI offend, discriminate or incriminate others?