Bringing an AI Ecosystem to the Domain Expert and Enterprise AI Developer

Fred Reiss

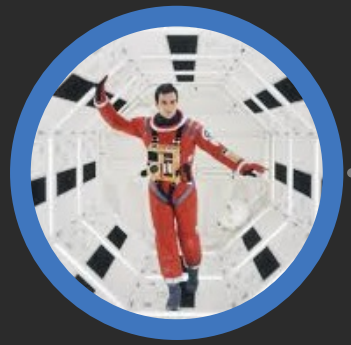
Vijay Bommireddipalli

IBM Center for Open-Source Data & AI Technologies
(http://codait.org)

IBM’s history of strong AI leadership

1968, 2001: A Space Odyssey

- IBM was a technical advisor
- HAL is “the latest in machine intelligence”



1997: Deep Blue

- Deep Blue became the first machine to beat a world chess champion in tournament play

2011: Jeopardy!

- Watson beat two top Jeopardy! champions

2018: Open Tech, AI & emerging standards

- New IBM centers of gravity for AI
- OS projects increasing exponentially
- Emerging global standards in AI

Center for Open Source Data and AI Technologies

- CODAIT aims to make AI solutions dramatically easier to create, deploy, and manage in the enterprise

- Relaunch of the Spark Technology Center (STC) to reflect expanded mission

- CODAIT
- codait.org

- codait (French)
- = coder/coded

- https://m.interglot.com/fr/en/codait

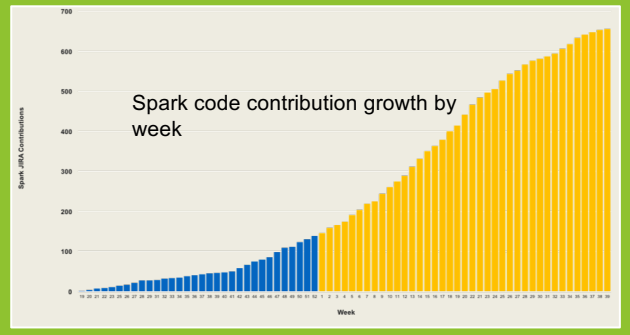
CODAIT by the numb3rs

- The team contributes to over 10 open source projects. These projects include - Spark, Tensorflow, Keras, SystemML, Arrow, Bahir, Toree, Livy, Zeppelin, R4ML, Stocator, Jupyter Enterprise Gateway
- 17 committers and many contributors in Apache projects- Spark, Arrow, systemML, Bahir, Toree, Livy
- Over 980 JIRAs and 50,000 lines of code committed to Apache Spark itself, and Over 65,000 LoC into SystemML
  - Established IBM as the number 1 contributor to Spark Machine Learning in Spark 2.0 release
- Over 25 product lines within IBM leveraging Apache Spark in some form or another. CODAIT engineers have interacted and interlocked with many of them.
- Speakers at over 100 conferences, MeetUps, un-conferences etc.

- CODAIT
- codait.org

- codait (French)
- = coder/coded

- https://m.interglot.com/fr/en/codait



Center for Open Source Data and AI Technologies

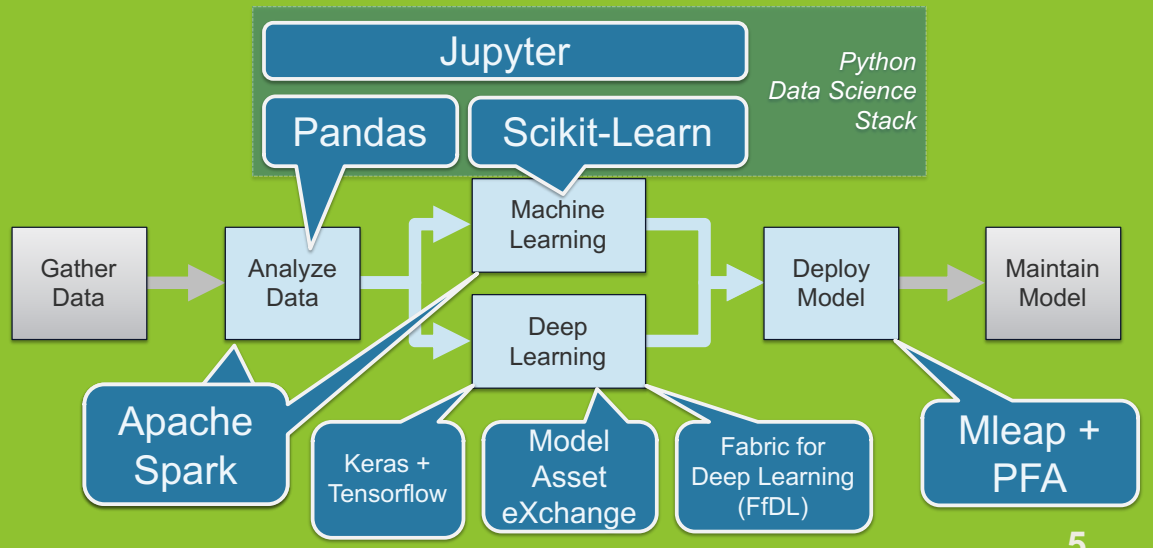
- **Code** - Build and improve practical frameworks to enable more developers to realize immediate value (e.g. FfDL, Tensorflow Jupyter, Spark)
- **Content** – Showcase solutions to complex and real world AI problems
- **Community** – Bring developers and data scientists to engage with IBM (e.g. MAX)

CODAIT

codait.org

- codait (French)
- = coder/coded
- https://m.interglot.com/fr/en/codait

Improving the Enterprise AI lifecycle in Open Source



Fabric for Deep Learning

https://github.com/IBM/FfDL

https://www.youtube.com/watch?v=nQsYWmkfLP4

- FfDL provides a scalable, resilient, and fault tolerant deep-learning framework
- Fabric for Deep Learning or FfDL (pronounced as ‘fiddle’) is an open source project which aims at making Deep Learning easily accessible to the people it matters the most i.e. Data Scientists, and AI developers.
- FfDL Provides a consistent way to deploy, train and visualize Deep Learning jobs across multiple frameworks like TensorFlow, Caffe, PyTorch, Keras etc.
- FfDL is being developed in close collaboration with IBM Research and IBM Watson. It forms the core of Watson`s Deep Learning service in open source.

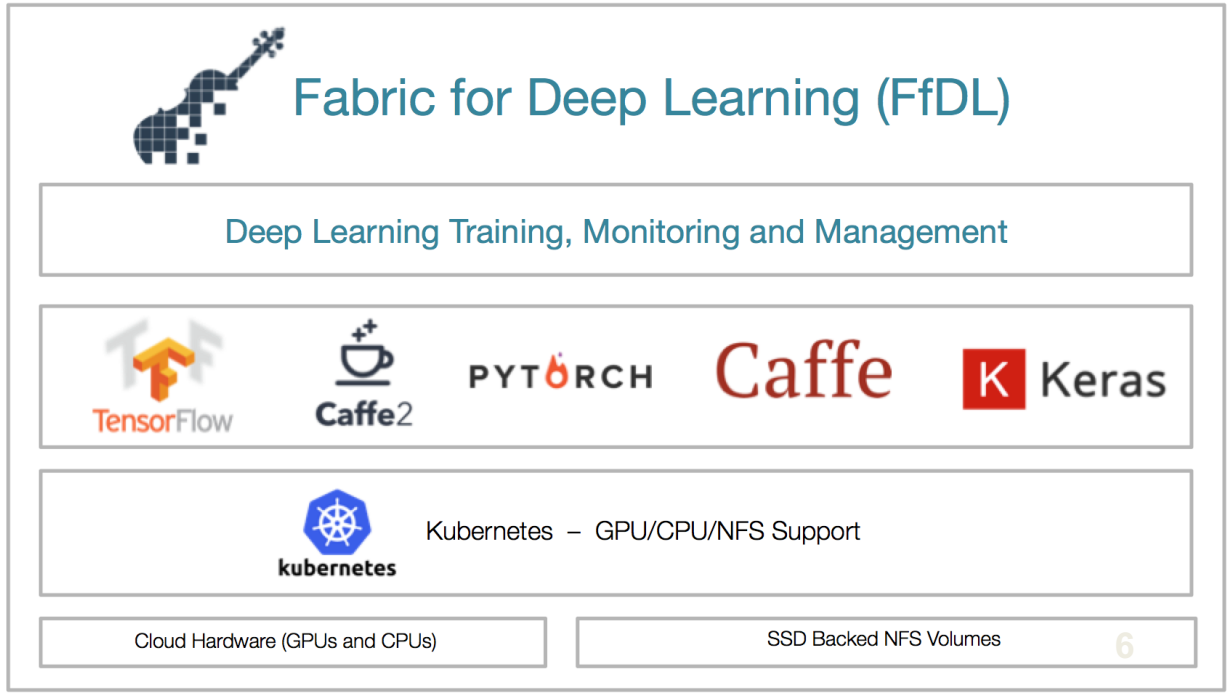
- FfDL Github Page
  https://github.com/IBM/FfDL

  FfDL dwOpen Page
  https://developer.ibm.com/code/open/projects/fabric-for-deep-learning-ffdl/

  FfDL Announcement Blog
  http://developer.ibm.com/code/2018/03/20/fabric-for-deep-learning

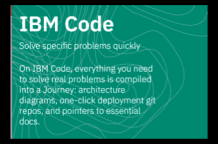
  FfDL Technical Architecture Blog
  http://developer.ibm.com/code/2018/03/20/democratize-ai-with-fabric-for-deep-learning

  Deep Learning as a Service within Watson Studio
  https://www.ibm.com/cloud/deep-learning
- Research paper: “Scalable Multi-Framework Management of Deep Learning Training Jobs”
  http://learningsys.org/nips17/assets/papers/paper_29.pdf



Jupyter Enterprise Gateway

- A lightweight, multi-tenant, scalable and secure gateway that enables Jupyter Notebooks to share resources across an Apache Spark or Kubernetes cluster for Enterprise/Cloud use cases



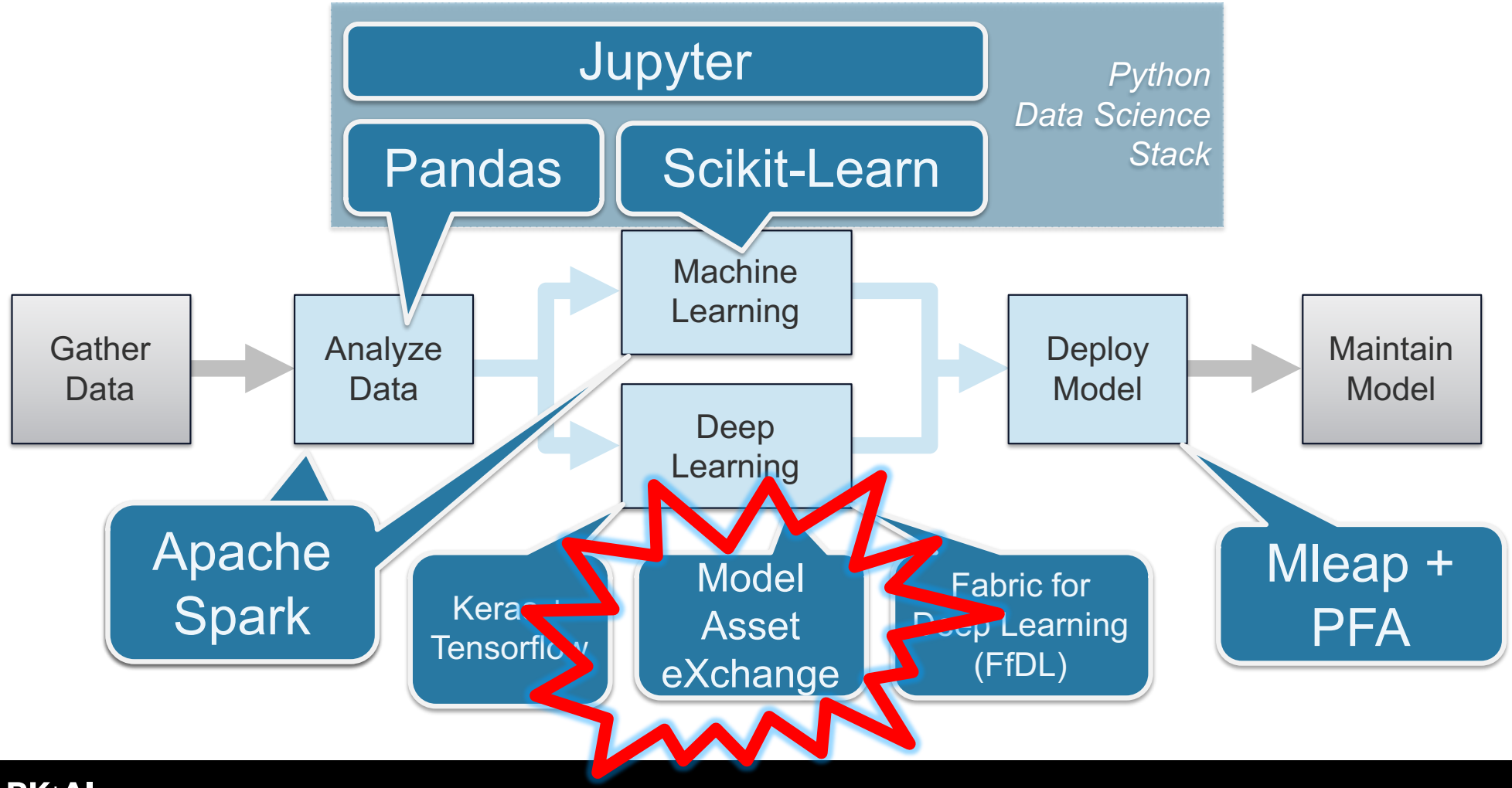
- Jupyter Enterprise Gateway at IBM Code
- https://developer.ibm.com/code/openprojects/jupyter-enterprise-gateway/

- Jupyter Enterprise Gateway source code at GitHub
- https://github.com/jupyter-incubator/enterprise_gateway
- Jupyter Enterprise Gateway Documentation
- http://jupyter-enterprise-gateway.readthedocs.io/en/latest/

Road Map

- Background: Deep Learning Models
- The IBM Code Model Asset Exchange
- Demo
- What's Next

CODAIT: Enabling End-to-End AI in the Enterprise



Making AI as Ubiquitous as the Telephone

This talk is about enabling **domain experts** to use **deep learning** in the enterprise.

Q: What is deep learning?

A: Machine learning using **deep neural networks.**

Q: What is a deep neural network?

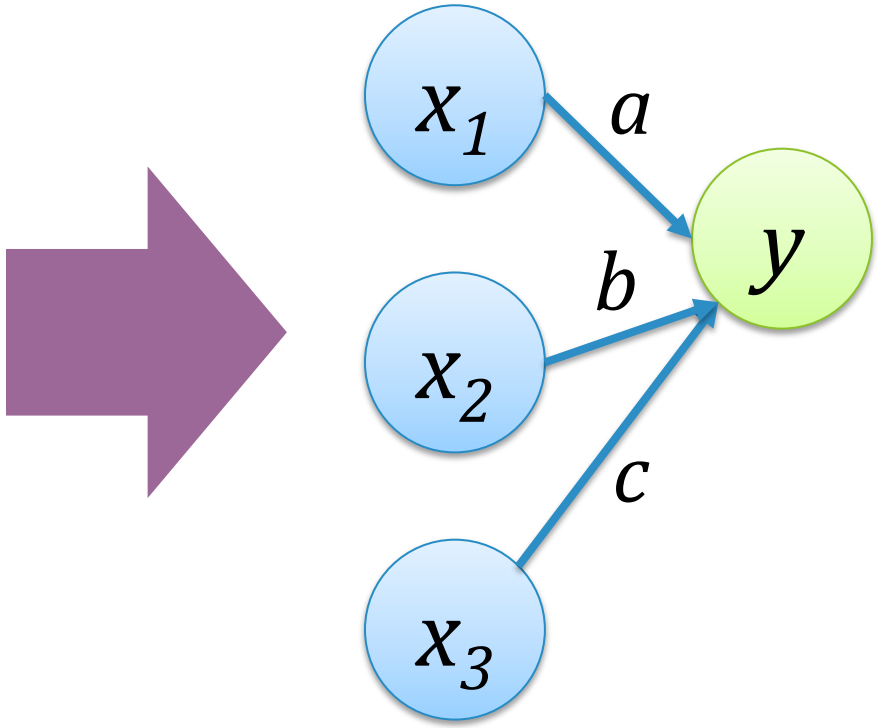
A: A neural network with **multiple hidden layers.**

Q: What is a **neural network**?

What is a neural network?

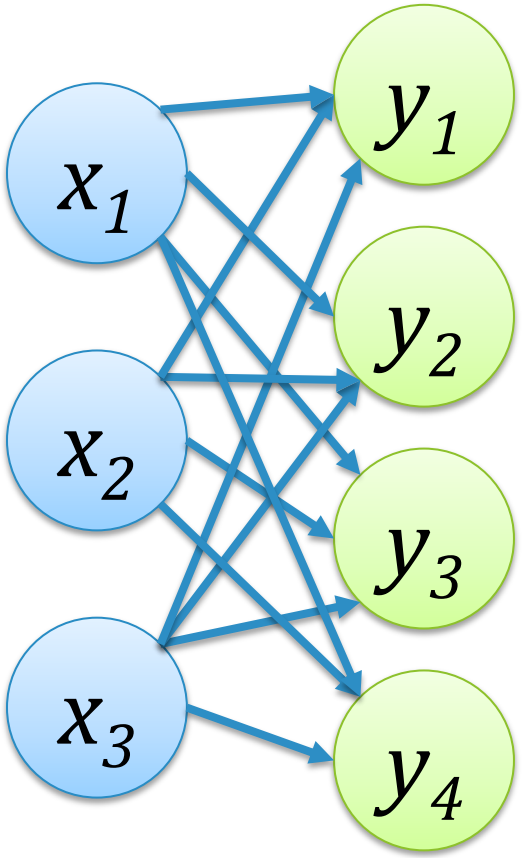
$$y = ax_1 + bx_2 + bx_3$$

Linear regression

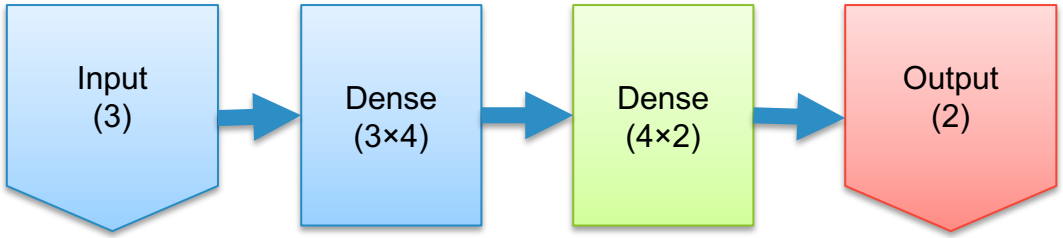


What is a neural network?

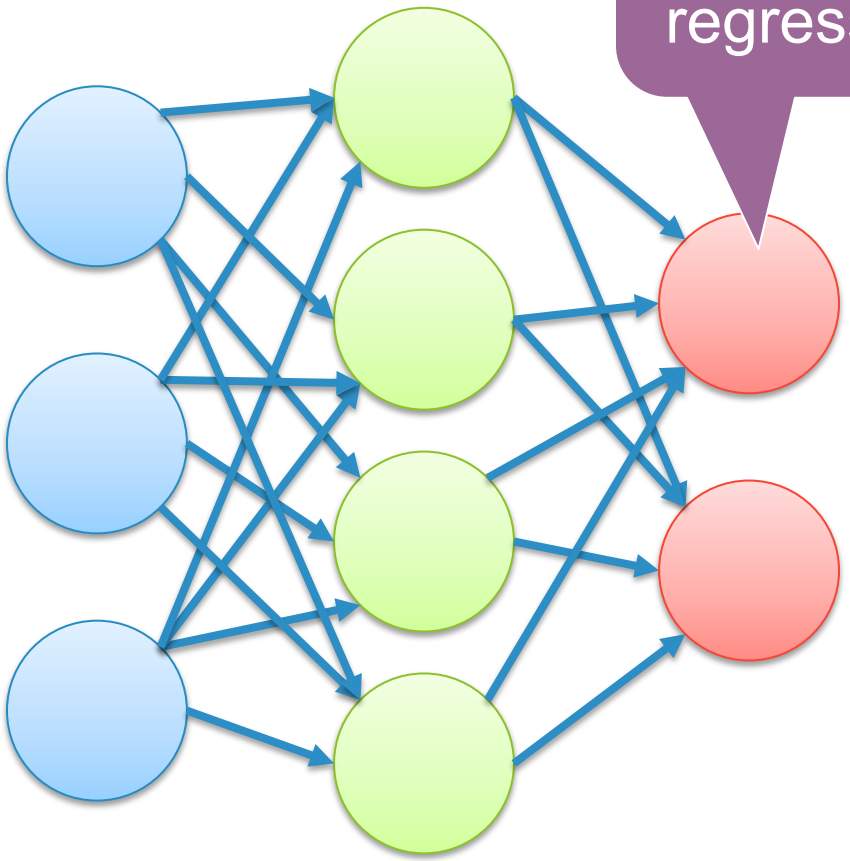
Multiple linear regressions at the same time



What is a neural network?

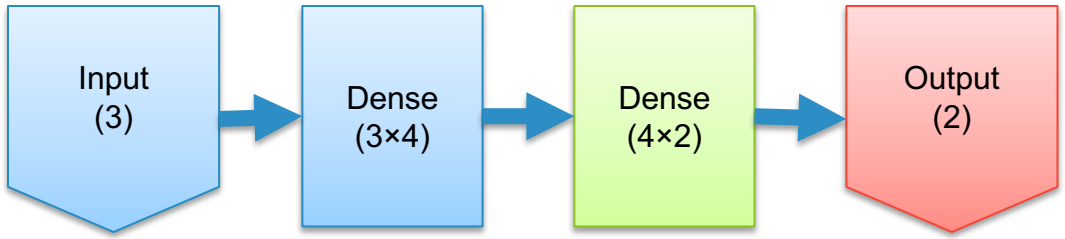


Same network in a more compact notation

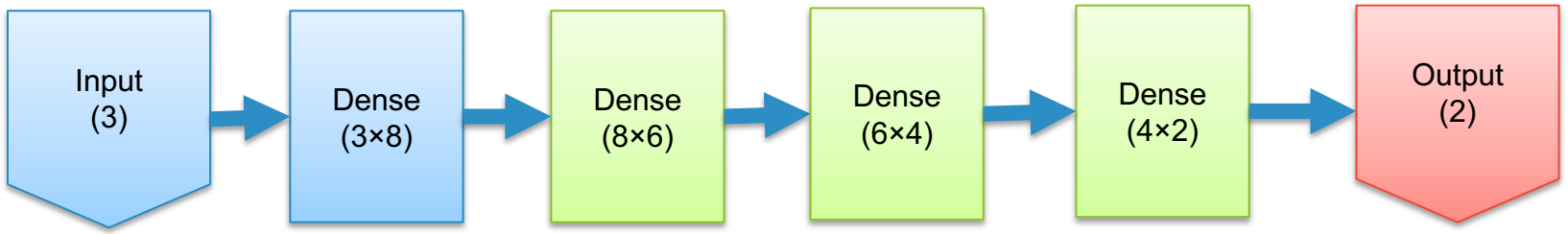


Multilayer Perceptron
Neural Network

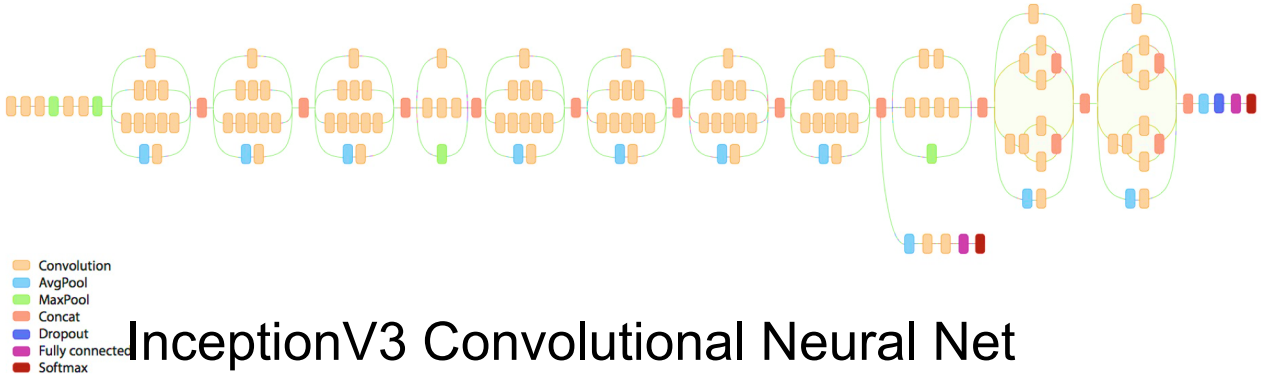
Q: What is a deep neural network?
A: A neural network with multiple hidden layers.



Q: What is a deep neural network?
A: A neural network with multiple hidden layers.



Q: What is deep learning?
A: Machine learning using deep neural networks.

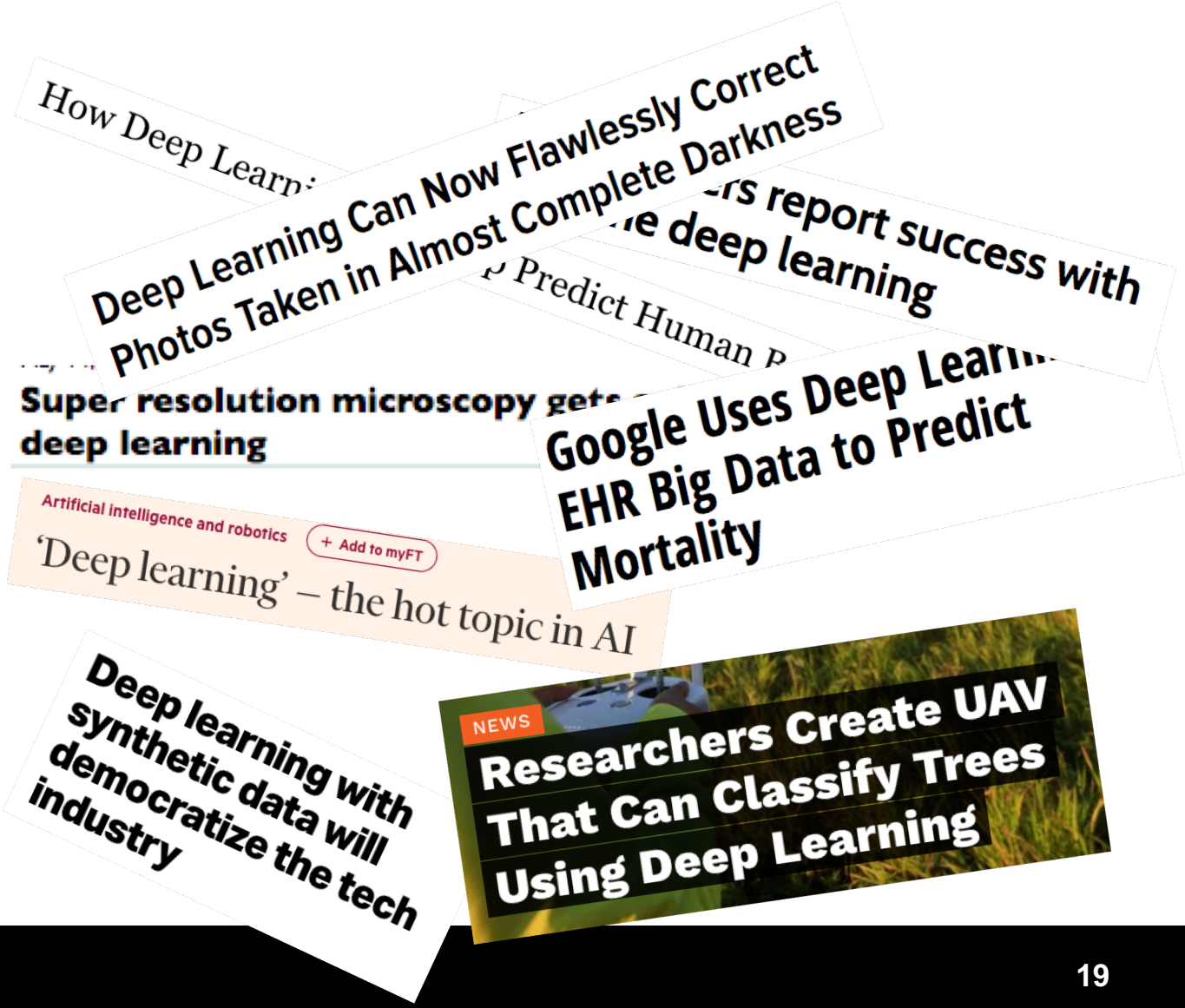


InceptionV3 Convolutional Neural Net
(A "medium-sized" deep learning model)

Image Source:
https://github.com/tensorflow/models/blob/master/research/inception/g3doc/inception_v3_architecture.png

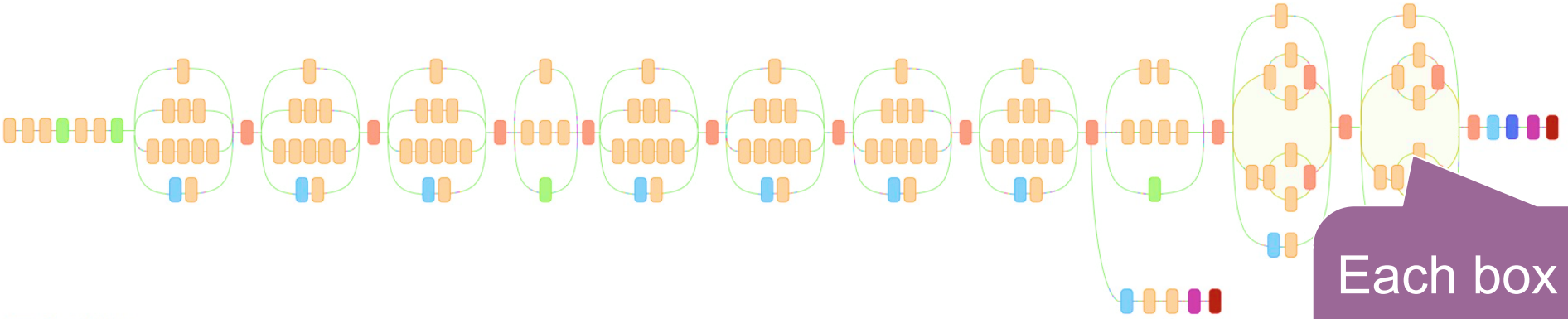
Characteristics of Deep Learning (1)

- State-of-the-Art prediction quality in many domains
  - Image classification
  - Machine translation
  - Facial recognition
  - Time series prediction
  - Many more



Characteristics of Deep Learning (2)

- Large, complex models
  - Model size generally determined by “how big a model can you fit on your device?”



InceptionV3 Convolutional Neural Net
(A “medium-sized” deep learning model)

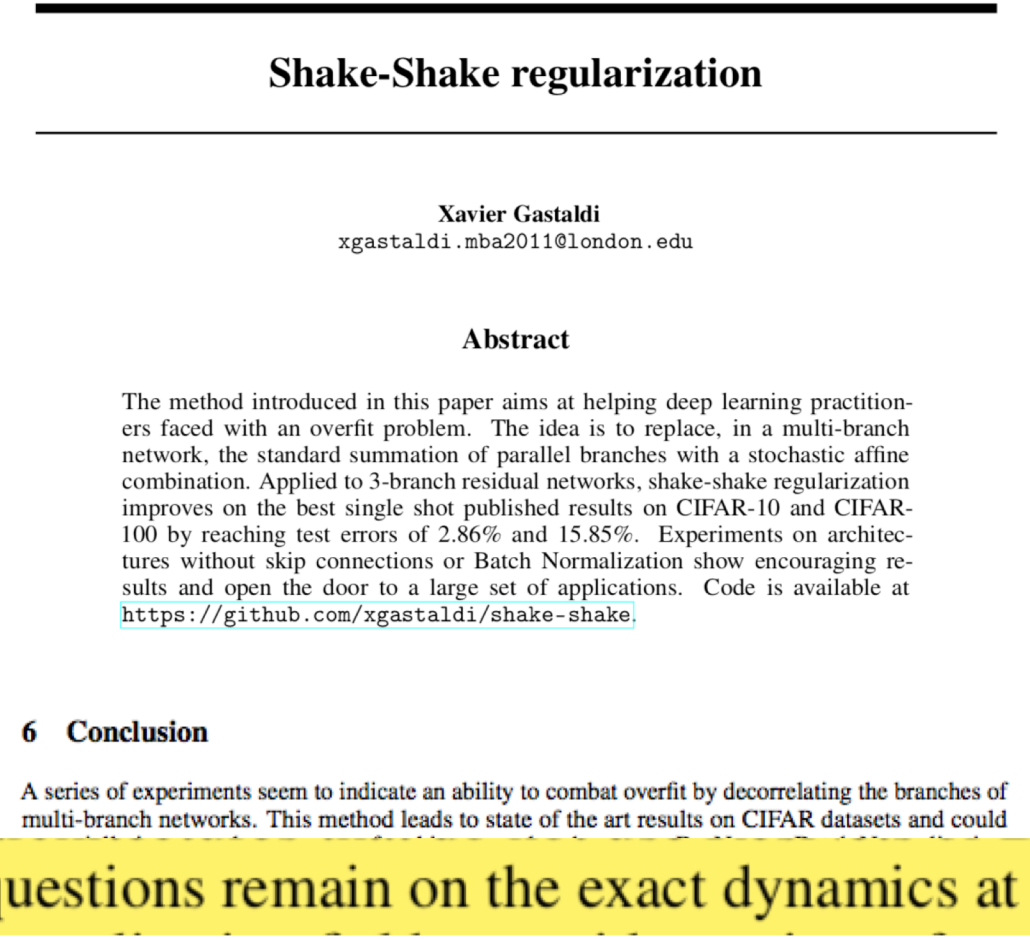
Each box ≈ between 32 and 768 linear regression models

Characteristics of Deep Learning (3)

Poorly understood today

…even by experts

- Why do the models converge?
- Why do the models converge with low loss?
- Why do the models generalize?



Shake-Shake regularization

Xavier Gastaldi
xgastaldi.mba2011@london.edu

Abstract

The method introduced in this paper aims at helping deep learning practitioners faced with an overfit problem. The idea is to replace, in a multi-branch network, the standard summation of parallel branches with a stochastic affine combination. Applied to 3-branch residual networks, shake-shake regularization improves on the best single shot published results on CIFAR-10 and CIFAR-100 by reaching test errors of 2.86% and 15.85%. Experiments on architectures without skip connections or Batch Normalization show encouraging results and open the door to a large set of applications. Code is available at https://github.com/xgastaldi/shake-shake

6 Conclusion

A series of experiments seem to indicate an ability to combat overfit by decorrelating the branches of multi-branch networks. This method leads to state of the art results on CIFAR datasets and could

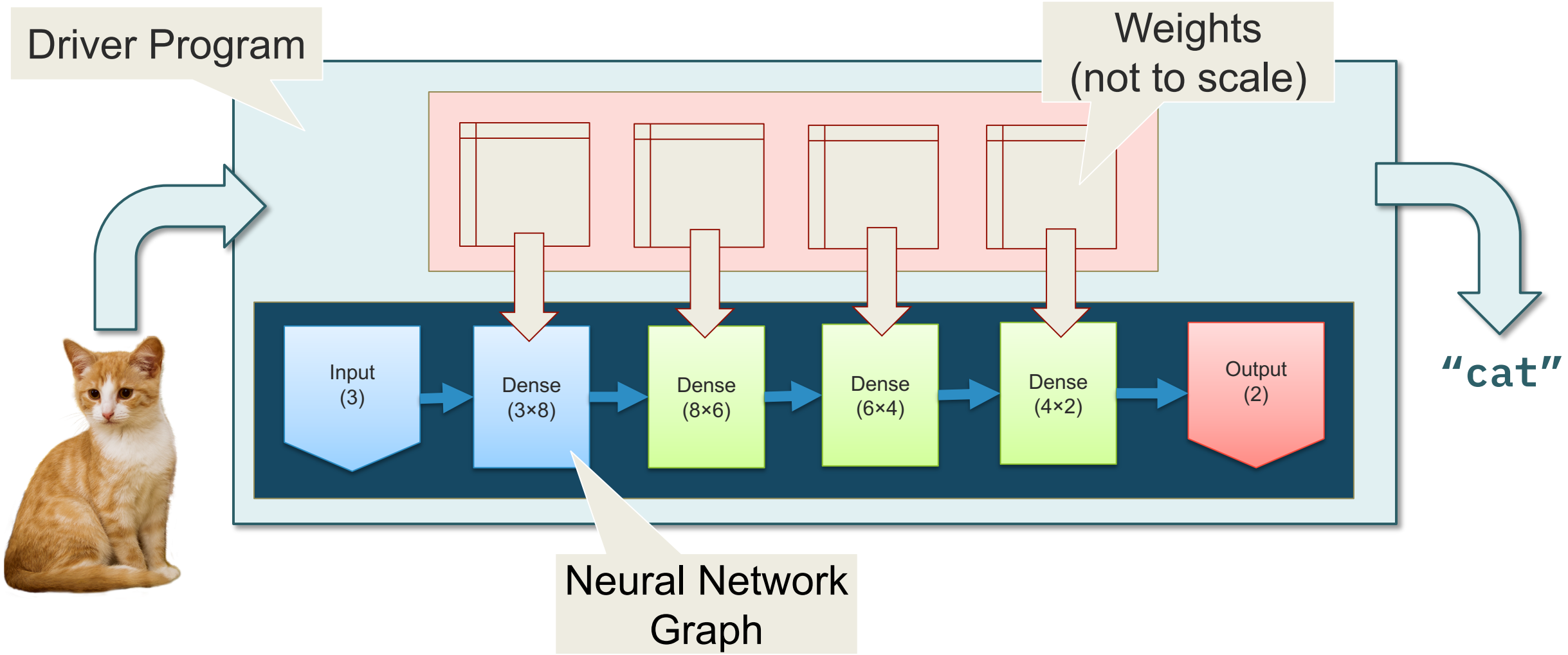
While these results are encouraging, questions remain on the exact dynamics at play.

Focus of this Talk

Incorporating **well-understood** deep learning models into enterprise applications.

Sounds easy!

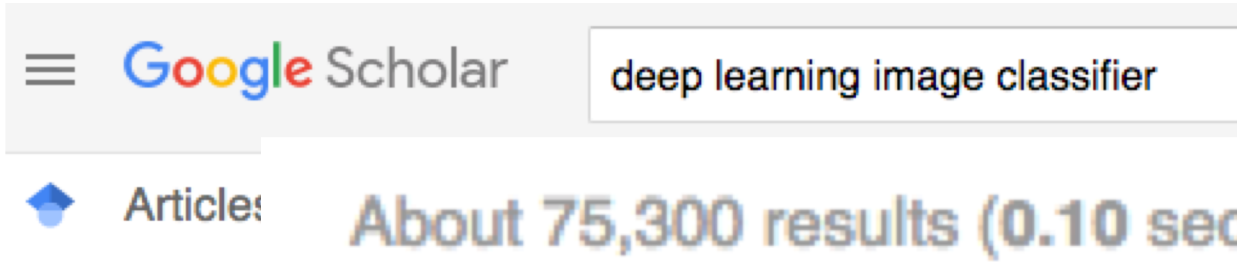
The Parts of a Deep Learning Model

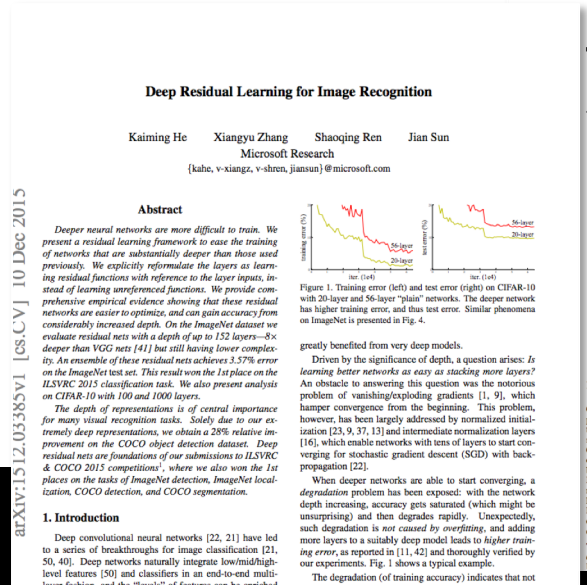


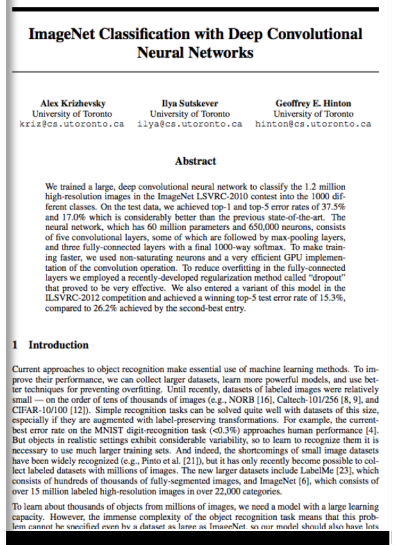
Example: Get an Image Classifier

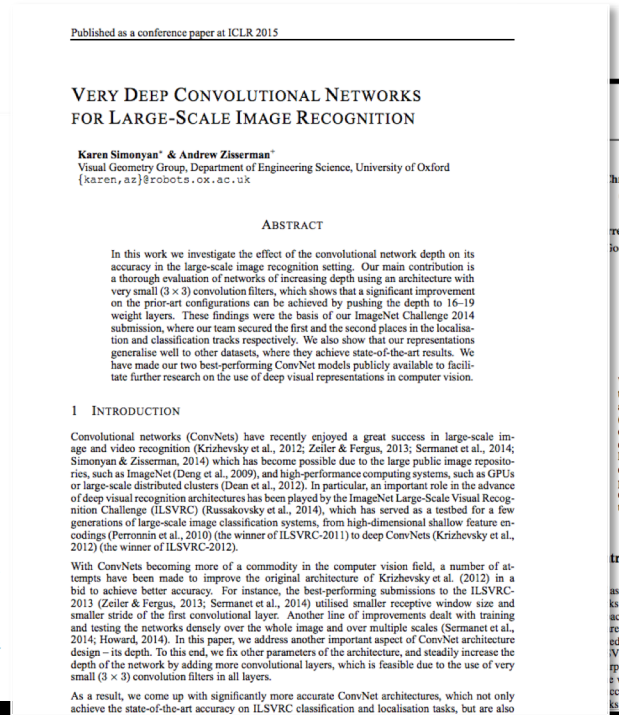
Step 1: Find a suitable neural network graph.

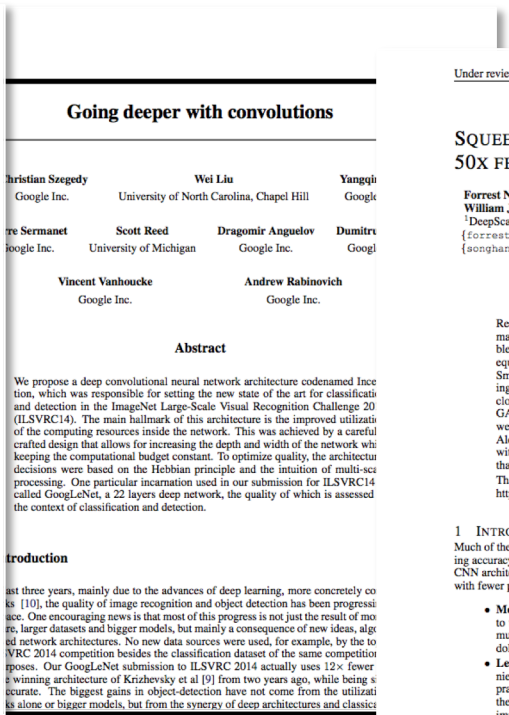
- Need to read some papers

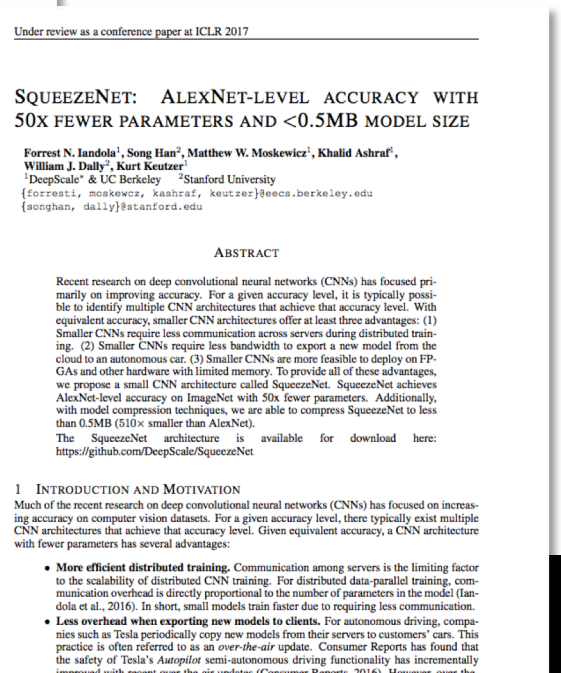












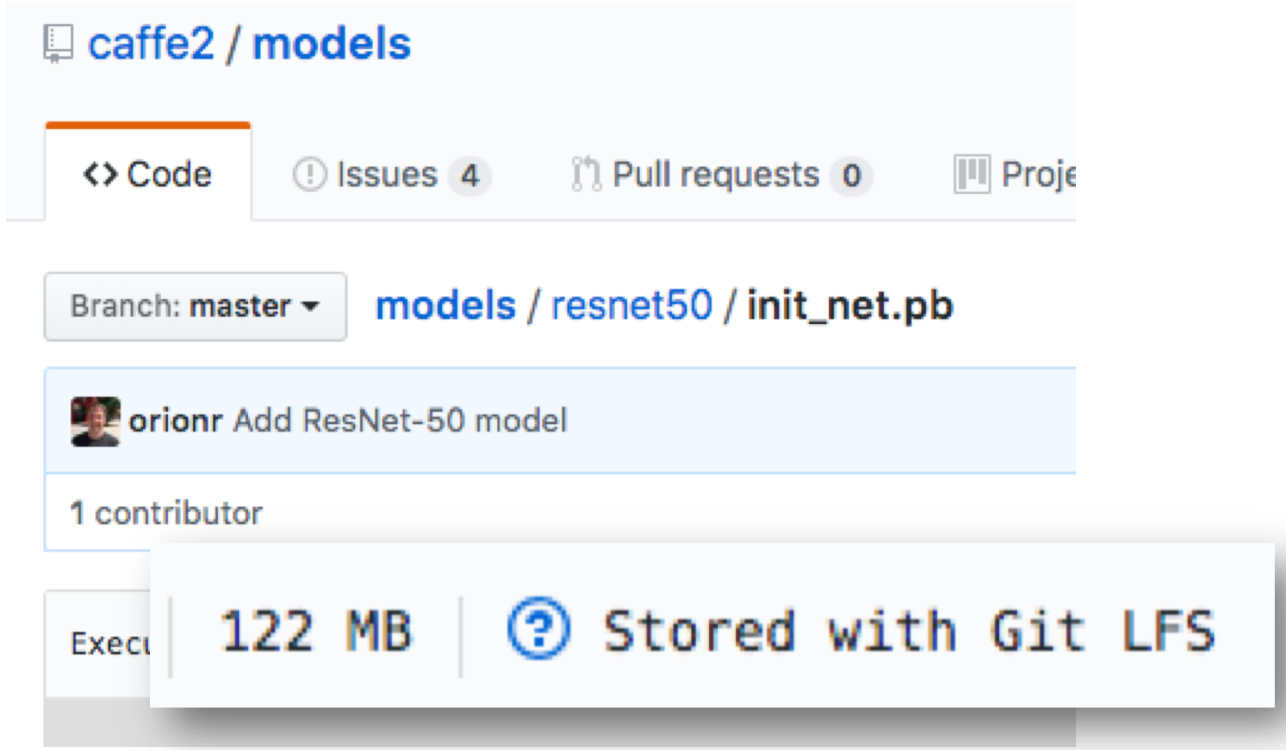
Example: Get an Image Classifier

Step 2: Find code to generate the neural network graph

TensorFlow code to build ResNet50 neural network graph

Example: Get an Image Classifier

Step 3: Find some pre-trained weights for your graph

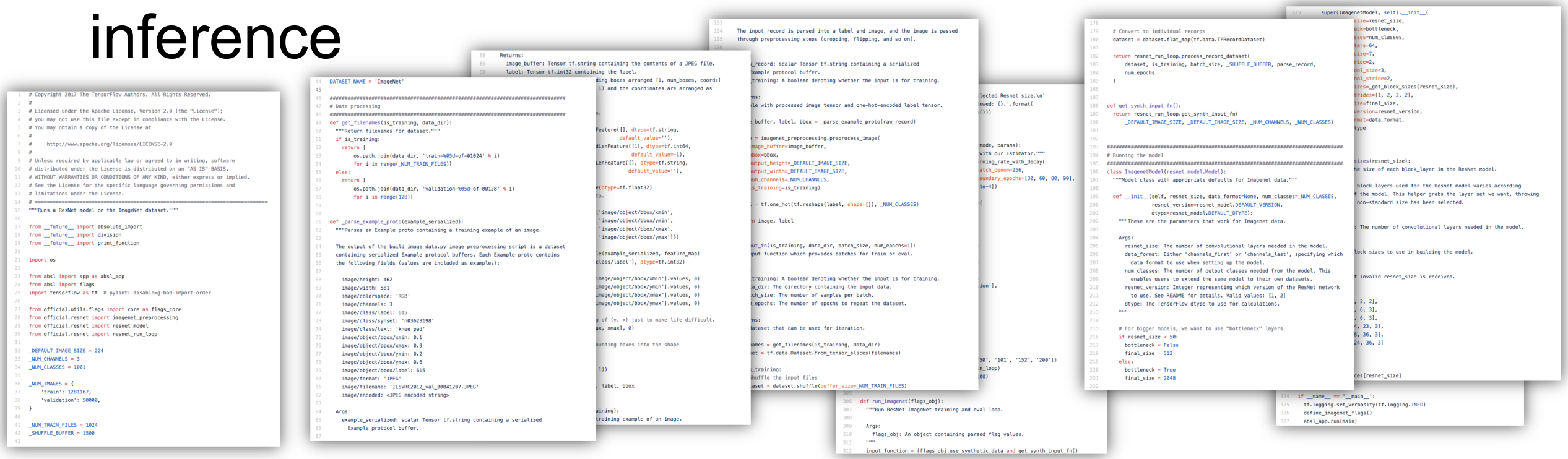


*Caffe2 ResNet50 model weights**

* Caffe2 only. Find a different binary file if your framework is not Caffe2.

Example: Get an Image Classifier

Step 4: Find example code that performs model inference



*TensorFlow code for **training and batch inference*** on ResNet50*

* Single-crop inference only. Additional code required to use multiple crops.

Example: Get an Image Classifier

Step 5: Write your own code to perform model inference on one image at a time

Step 6: Package your inference code, graph creation code, and pre-trained weights together

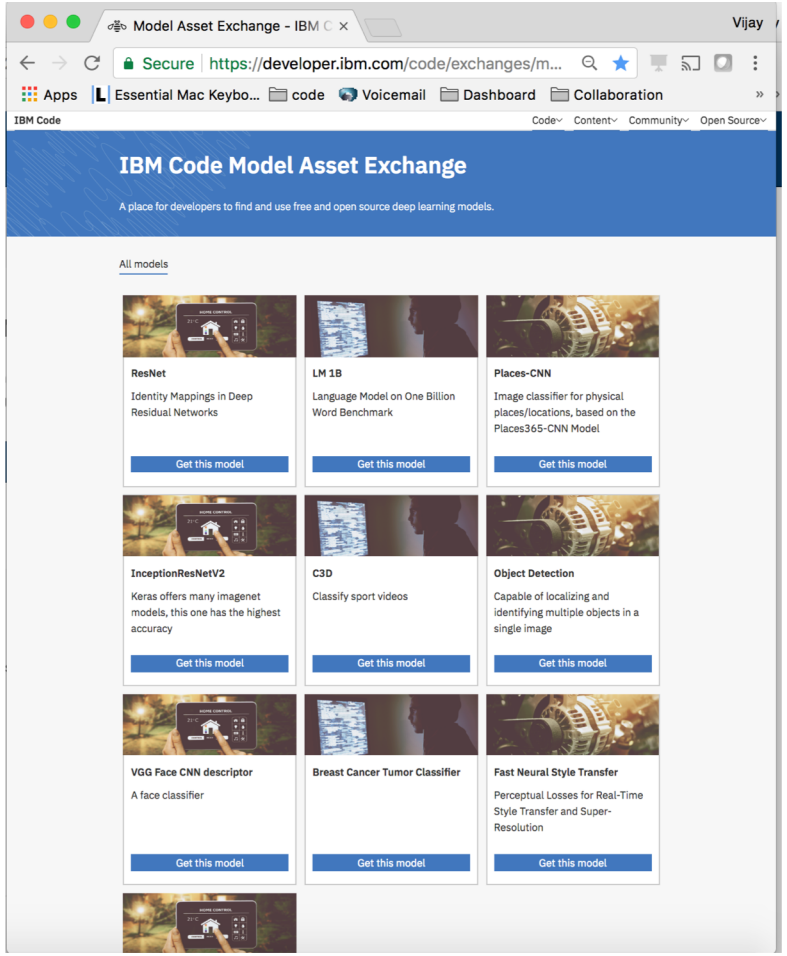
Step 7: Deploy your package

Model Marketplaces

- Collections of well-understood deep learning models
- Provide a central place to find known-good implementations of these models

The IBM Code Model Asset eXchange

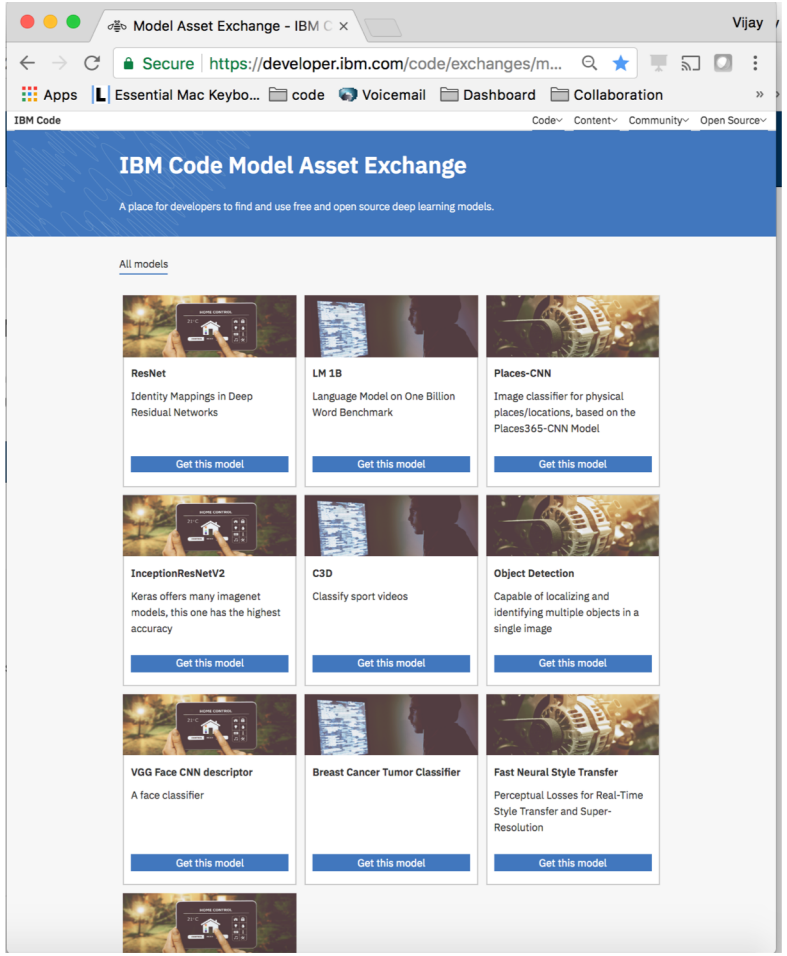
- Free, open-source models.
- Wide variety of domains.
- Multiple deep learning frameworks.
- Vetted and tested code and IP.
- Build and deploy a web service in 30 seconds.
- Start training on **Watson Studio** in minutes.



- Demo!

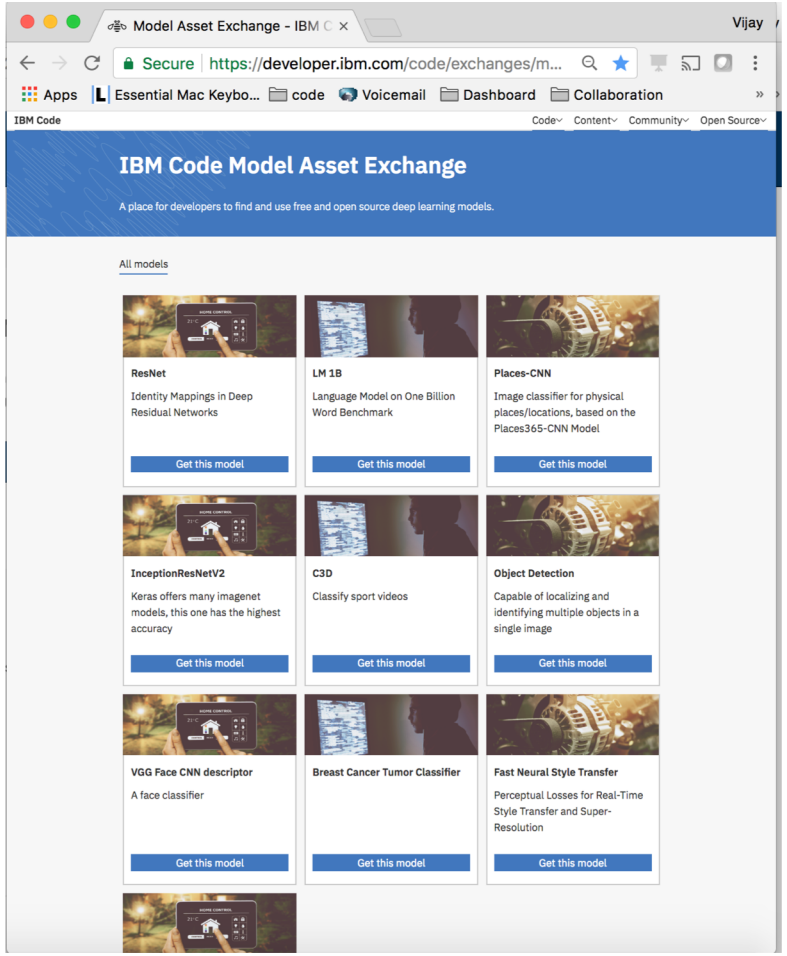
Model Asset eXchange: Summary

- Free, open-source models.
- Wide variety of domains.
- Multiple deep learning frameworks.
- Vetted and tested code and IP.
- Build and deploy a web service in 30 seconds.
- Start training on **Watson Studio** in minutes.



Model Asset eXchange: What's Next

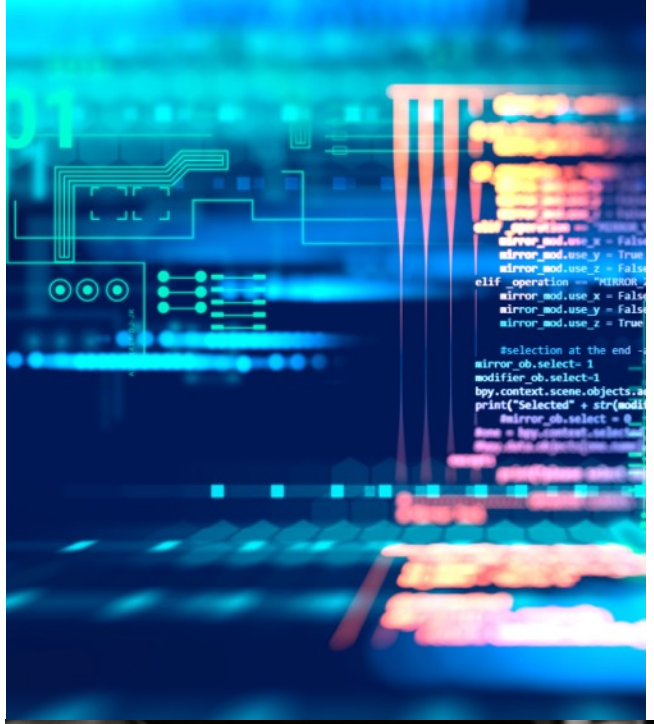
- More models
- More deployment options
- Code Patterns showing how to use the models (including today's demo!)

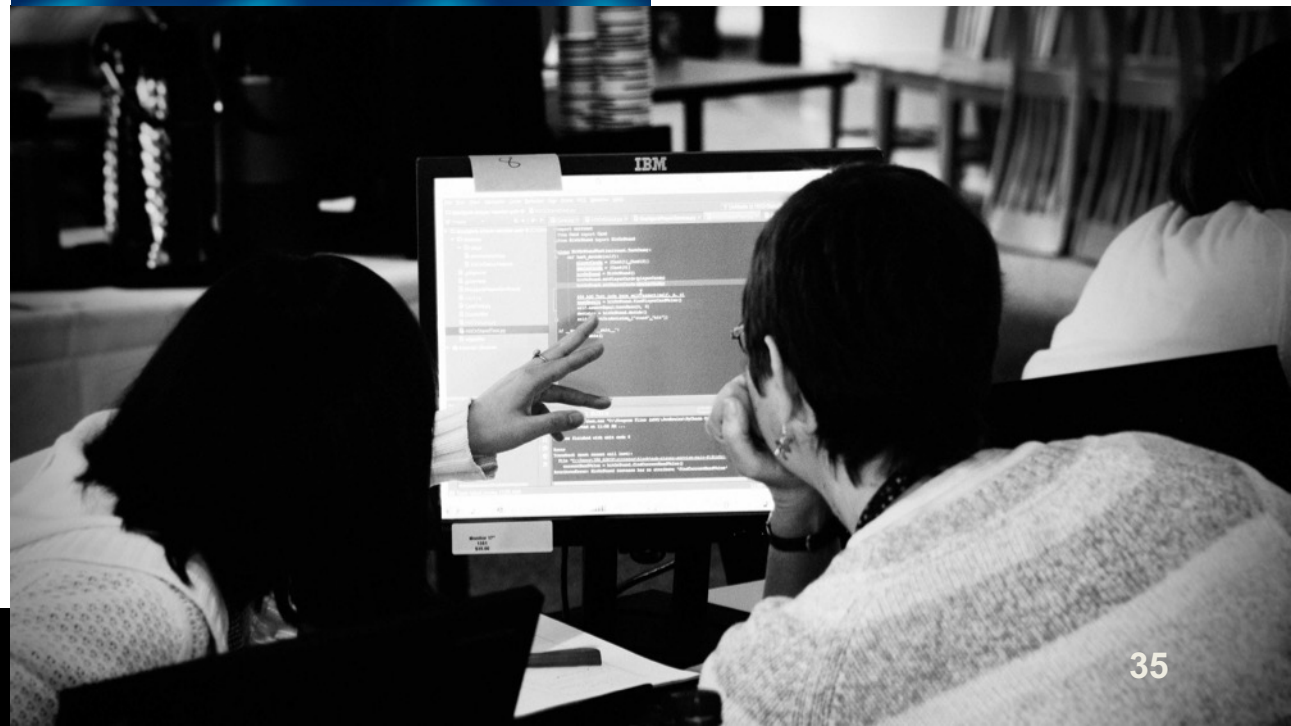


Thank you!

- http://codait.org
- https://developer.ibm.com/code/exchanges/models/

- github.com/codait
- developer.ibm.com/code





CALL FOR CODE
GLOBAL INITIATIVE 2018

Commit to the cause. Push for change.

Call for Code inspires developers to solve **pressing global problems** with **sustainable software solutions**, delivering on their vast potential to do good.

Bringing together NGOs, academic institutions, enterprises, and startup developers to compete build effective **disaster mitigation solutions**, with a focus on health and well-being.

International Federation of Red Cross/Red Crescent, **The American Red Cross**, and the **United Nations Office of Human Rights** combine for the ***Call for Code Award*** *to* elevate the profile of developers.

Award winners will receive **long-term support** through **open source foundations**, **financial prizes**, the **opportunity to present their solution to leading VCs**, and will deploy their solution through **IBM's Corporate Service Corps**.

Developers will jump-start their project with dedicated **IBM Code Patterns**, combined with **optional enterprise technology** to build projects over the course of three months.

Judged by the world's most **renowned technologists**, the **grand prize** will be presented in **October** at an Award Event.

developer.ibm.com/callforcode

IBM Sessions at Spark+AI Summit 2018 (Tuesday, June 5)

Date, Time, Location & Duration	Session title and Speaker
Tue, June 5 | 11 AM 2010-2012, 30 mins	Productionizing Spark ML Pipelines with the Portable Format for Analytics Nick Pentreath (IBM)
Tue, June 5 | 2 PM 2018, 30 mins	Making PySpark Amazing—From Faster UDFs to Dependency Management and Graphing! Holden Karau (Google) Bryan Cutler (IBM)
Tue, June 5 | 2 PM Nook by 2001, 30 mins	Making Data and AI Accessible for All Armand Ruiz Gabernet (IBM)
Tue, June 5 | 2:40 PM 2002-2004, 30 mins	Cognitive Database: An Apache Spark-Based AI-Enabled Relational Database System Rajesh Bordawekar (IBM T.J. Watson Research Center)
Tue, June 5 | 3:20 PM 3016-3022, 30 mins	Dynamic Priorities for Apache Spark Application's Resource Allocations Michael Feiman (IBM Spectrum Computing) Shinnosuke Okada (IBM Canada Ltd.)
Tue, June 5 | 3:20 PM 2001-2005, 30 mins	Model Parallelism in Spark ML Cross-Validation Nick Pentreath (IBM) Bryan Cutler (IBM)
Tue, June 5 | 3:20 PM 2007, 30 mins	Serverless Machine Learning on Modern Hardware Using Apache Spark Patrick Stuedi (IBM)
Tue, June 5 | 5:40 PM 2002-2004, 30 mins	Create a Loyal Customer Base by Knowing Their Personality Using AI-Based Personality Recommendation Engine; Sourav Mazumder (IBM Analytics) Aradhna Tiwari (University of South Florida)
Tue, June 5 | 5:40 PM 2007, 30 mins	Transparent GPU Exploitation on Apache Spark Dr. Kazuaki Ishizaki (IBM) Madhusudanan Kandasamy (IBM)
Tue, June 5 | 5:40 PM 2009-2011, 30 mins	Apache Spark Based Hyper-Parameter Selection and Adaptive Model Tuning for Deep Neural Networks Yonggang Hu (IBM) Chao Xue (IBM)

IBM Sessions at Spark+AI Summit 2018 (Wednesday, June 6)

Date, Time, Location & Duration	Session title and Speaker
Wed, June 6 | 12:50 PM	Birds of a Feather: Apache Arrow in Spark and More Bryan Cutler (IBM) Li Jin (Two Sigma Investments, LP)
Wed, June 6 | 2 PM 2002-2004, 30 mins	Deep Learning for Recommender Systems Nick Pentreath (IBM))
Wed, June 6 | 3:20 PM 2018, 30 mins	**Bringing an AI Ecosystem to the Domain Expert and Enterprise AI Developer Frederick Reiss (IBM) Vijay Bommireddipalli (IBM Center for Open-Source Data & AI Technologies)**

Meet us at IBM booth in the Expo area.

Thank you!

- http://codait.org
- https://developer.ibm.com/code/exchanges/models/
- https://www.linkedin.com/in/fred-reiss/
- https://www.linkedin.com/in/vijayrb

- github.com/codait
- developer.ibm.com/code