Opening the Black Box

Visualising, Explaining, Interpreting

Jules Françoise

jules.francoise@limsi.fr
LIMSI-CNRS
Orsay, France

HCID - IML Lecture 3 — 13/01/2020

Example #1: Wekinator

Wekinator

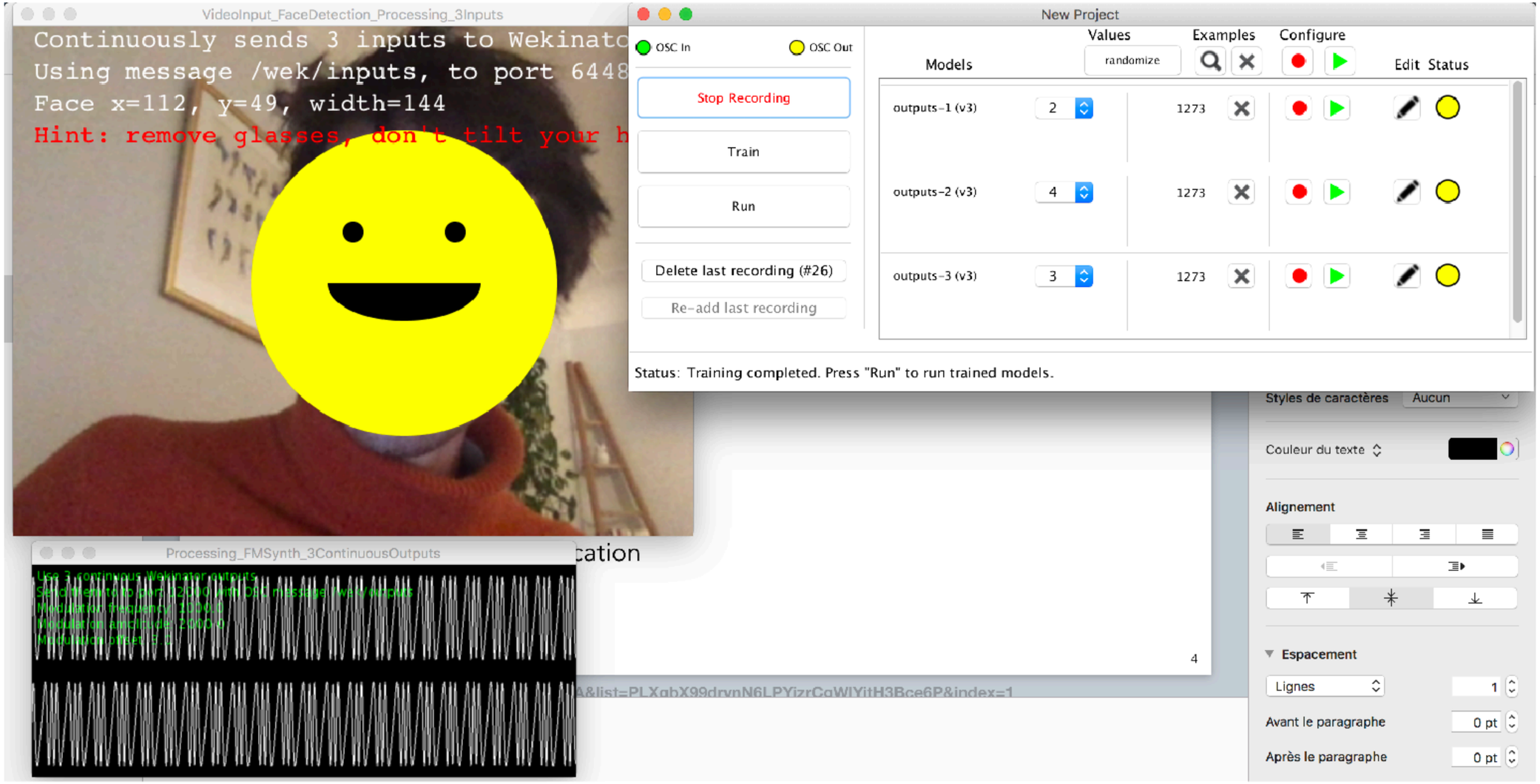
- **Task:** Music performance
- Goal: create novel gesture-based instruments
- Regression to learn movement-sound mapping

http://www.wekinator.org/example-projects/

=> Users learn how to give examples and update their expectations along the process

Fiebrink, R., Cook, P. R., & Trueman, D. (2011, May). Human model evaluation in interactive supervised learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 147-156). ACM.

Wekinator: Demo



http://www.wekinator.org/

Wekinator: User Study

3 studies:

- 7 composers => weekly meetings for 10 weeks
- 21 students => assignment: create a gesturally composed performance
- professional cellist => build a gesture recognition system for a sensor-equipped cello bow (“K-Bow”)

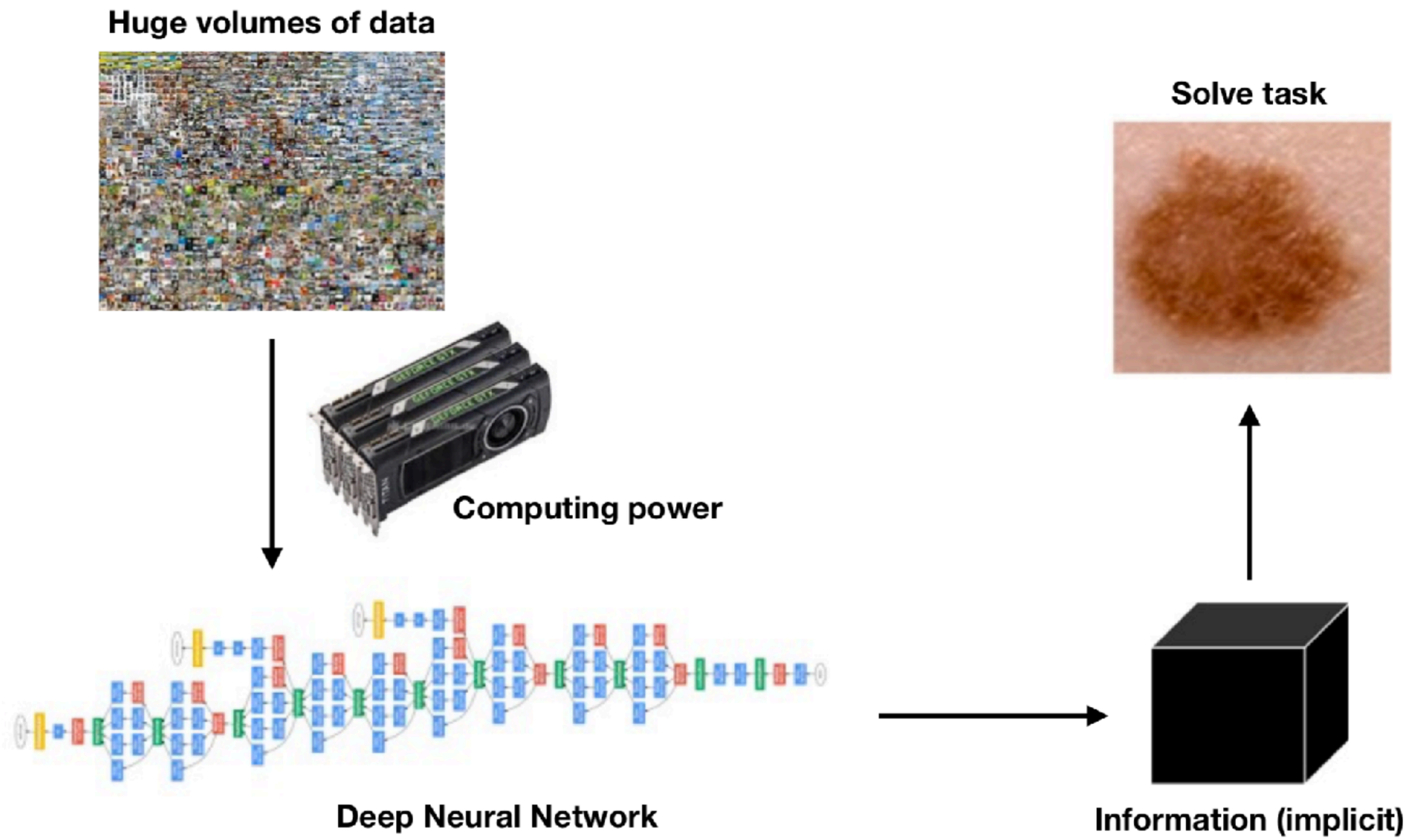
Findings

- People prefer Direct Evaluation to cross-validation
- Criteria: Correctness, Decision Boundary Shape, Complexity and Unexpectedness, …

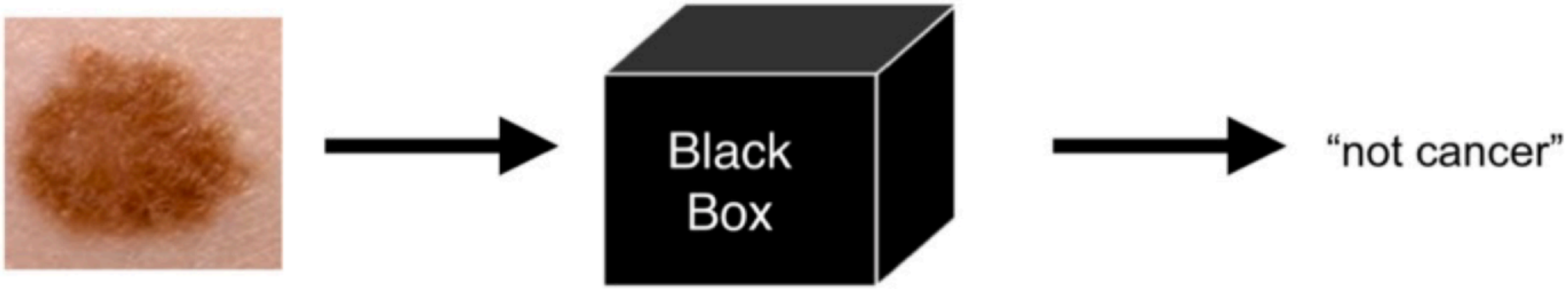
Fiebrink, R., Cook, P. R., & Trueman, D. (2011, May). Human model evaluation in interactive supervised learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 147-156). ACM.

Opening the black box

Black box models



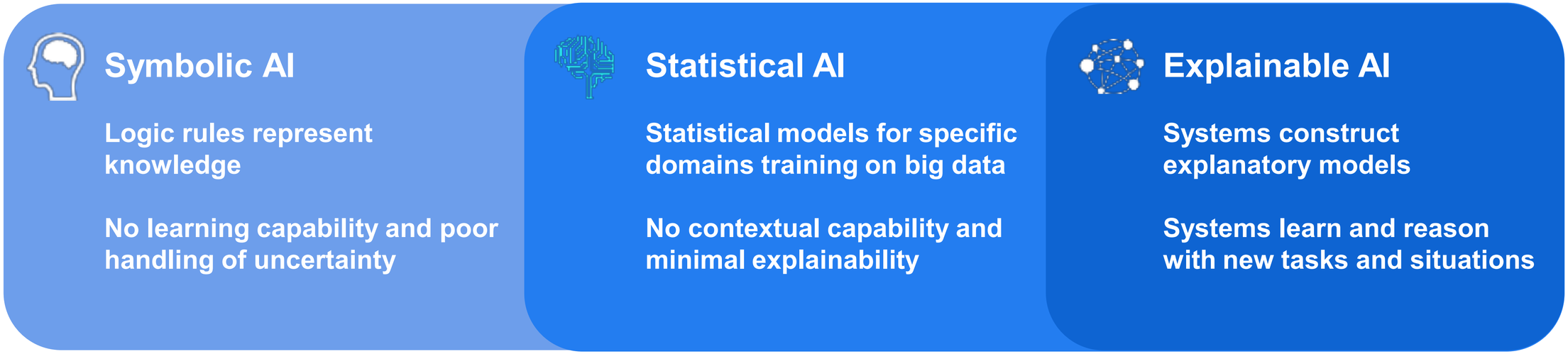
Black box models



Is minimizing the error a guarantee for the model to work well in practice?

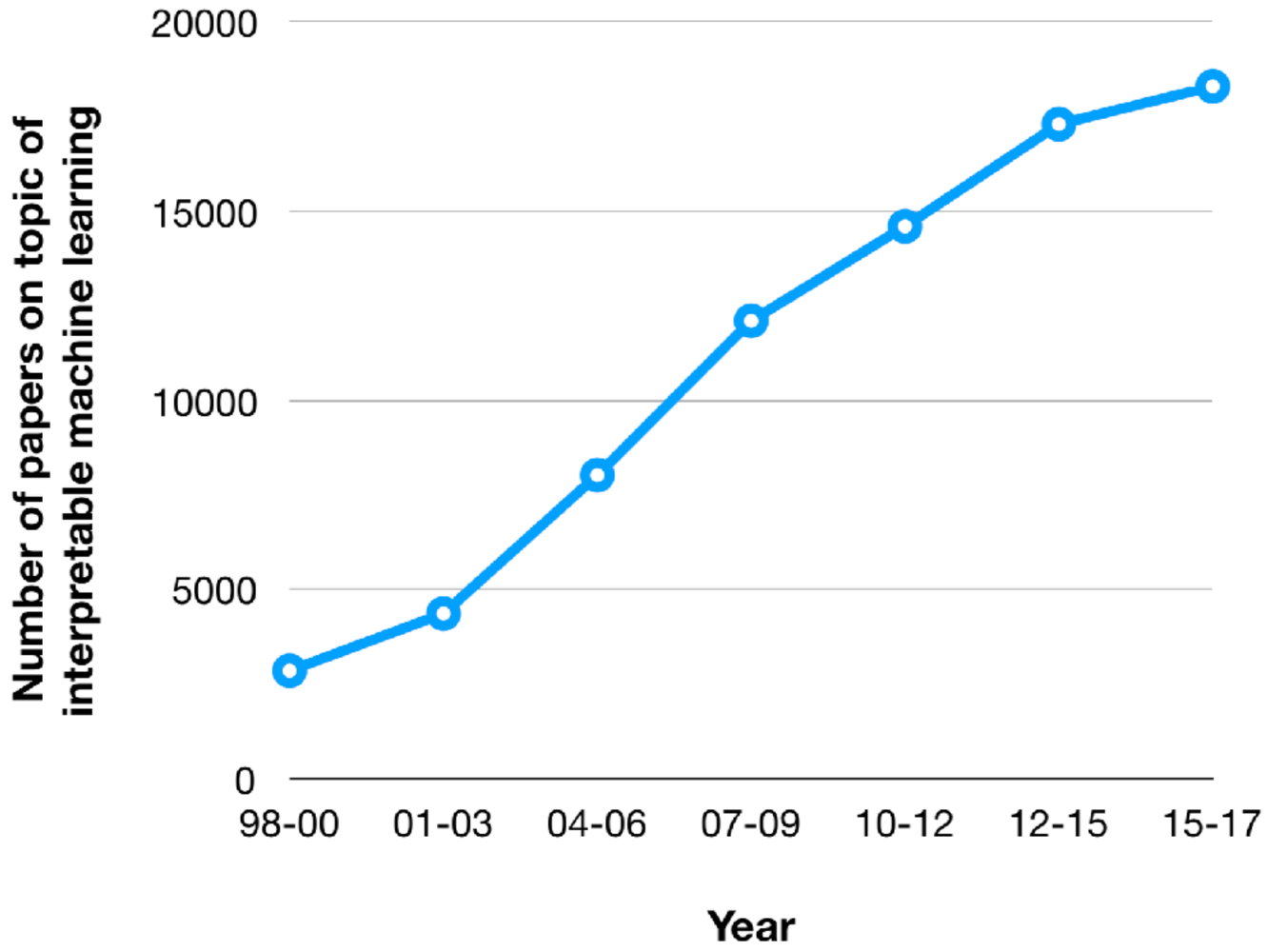
Trends

=>Third Wave of AI



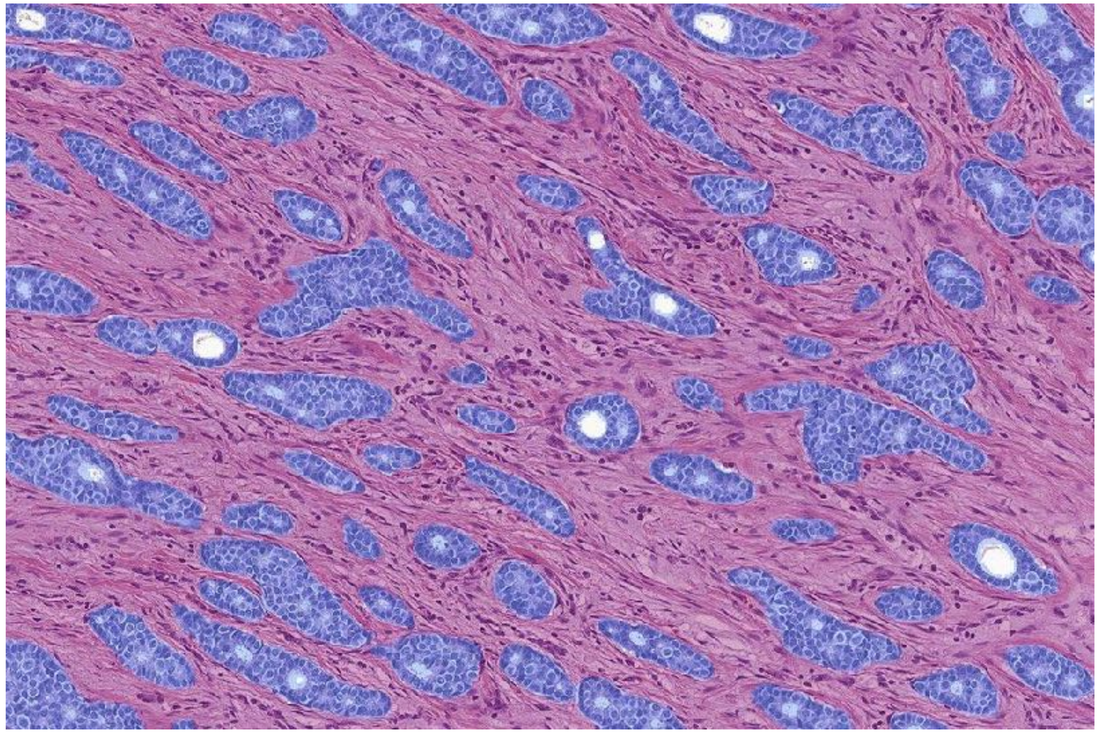
From Sahin Cem Geyik, Krishnaram Kenthapadi & Varun Mithal, ***Explainable AI in Industry*** *(KDD 2019 Tutorial)*

Trends


20000
15000
10000
5000
0
Number of papers on topic of interpretable machine learning
98-00
01-03
04-06
07-09
10-12
12-15
15-17
Year

Why Interpretability?

=> Wrong decisions can be costly and dangerous



Why Interpretability?

Critical systems

User Acceptance & Trust

Legal requirements

- Conformance to ethical standards, fairness
- Right to be informed
- Contestable decisions

Increase Insightfulness

- Informativeness
- Uncovering causality

Explanations & Interpretability

explanation | ɛksplə'neɪʃ(ə)n |

noun

a statement or account that makes something clear: *the birth rate is central to any explanation of population trends*.

interpret | ɪn'tə:prɪt |

verb **(interprets, interpreting, interpreted)** *[with object]*

1 explain the meaning of (information or actions): *the evidence is difficult to interpret*.

Oxford Dictionary of English

Role-based Interpretability

~~"Is the explanation interpretable?"~~ => "**To whom** is the explanation interpretable?"
No Universally Interpretable Explanations!

- **End users** "Am I being treated fairly?" | "Can I contest the decision?" | "What could I do differently to get a positive outcome?"
- **Engineers, data scientists**: "Is my system working as designed?"
- **Regulators** " Is it compliant?"

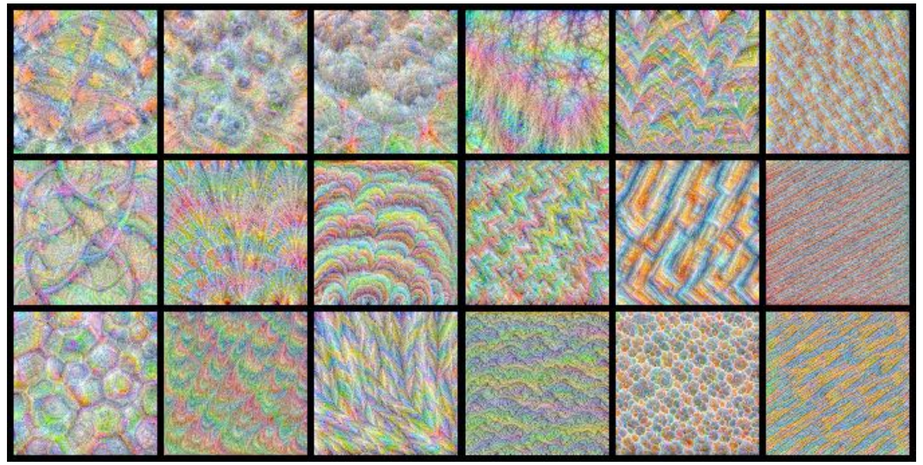
=> An ideal explainer should model the user background.

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. arXiv preprint arXiv:1806.07552.

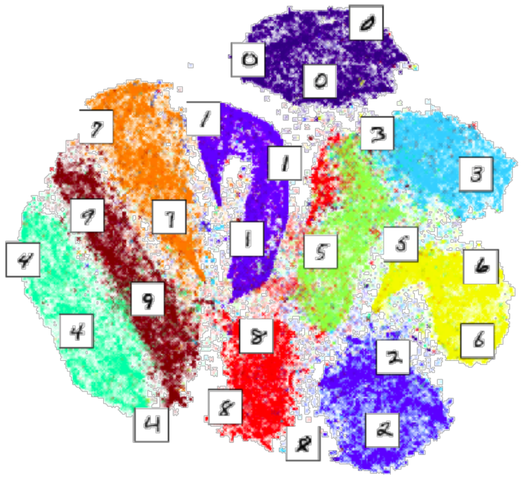
Dimensions of Interpretability

Prediction

Model

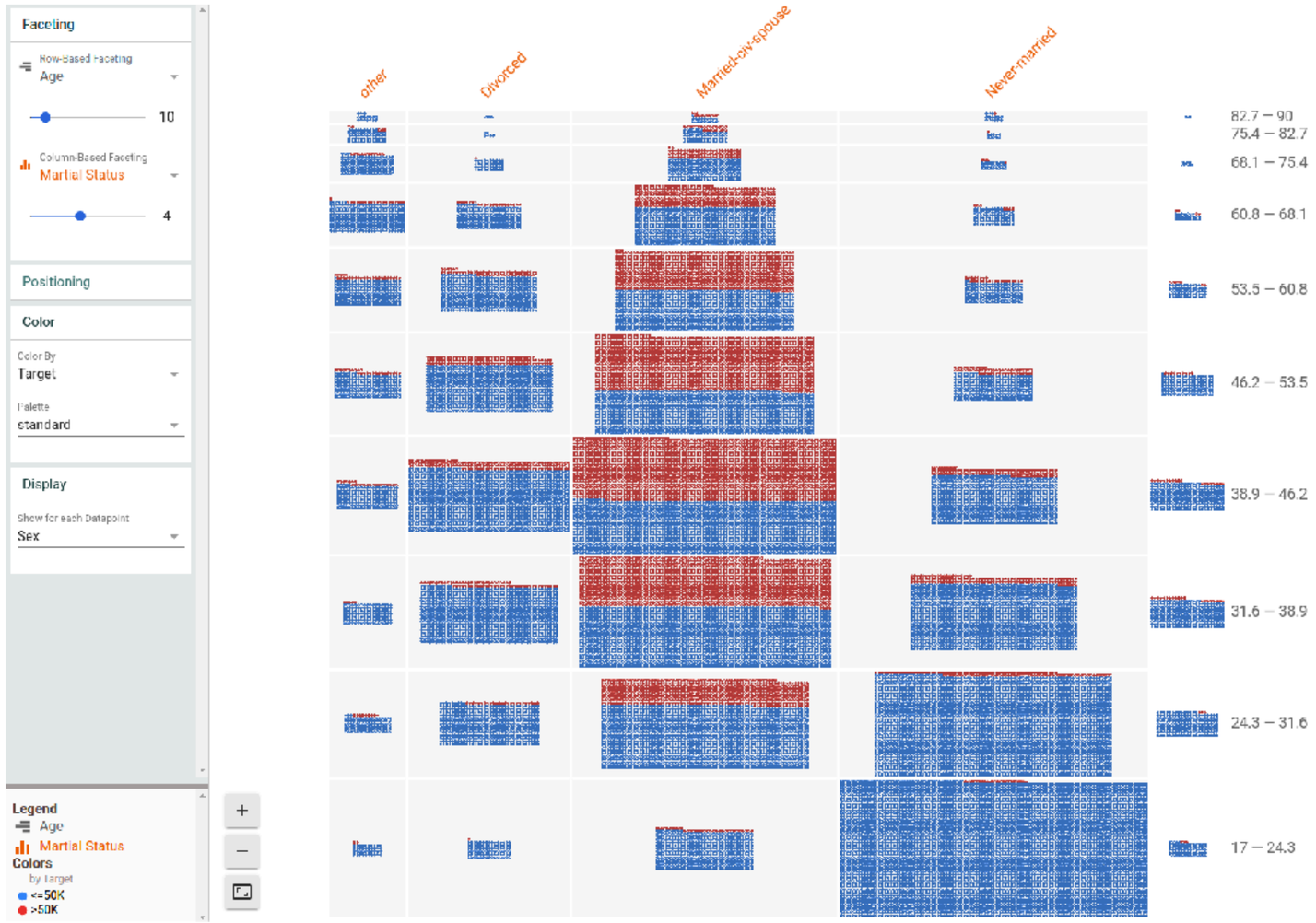


Data



Visualizing training data

Facets (Google PAIR)

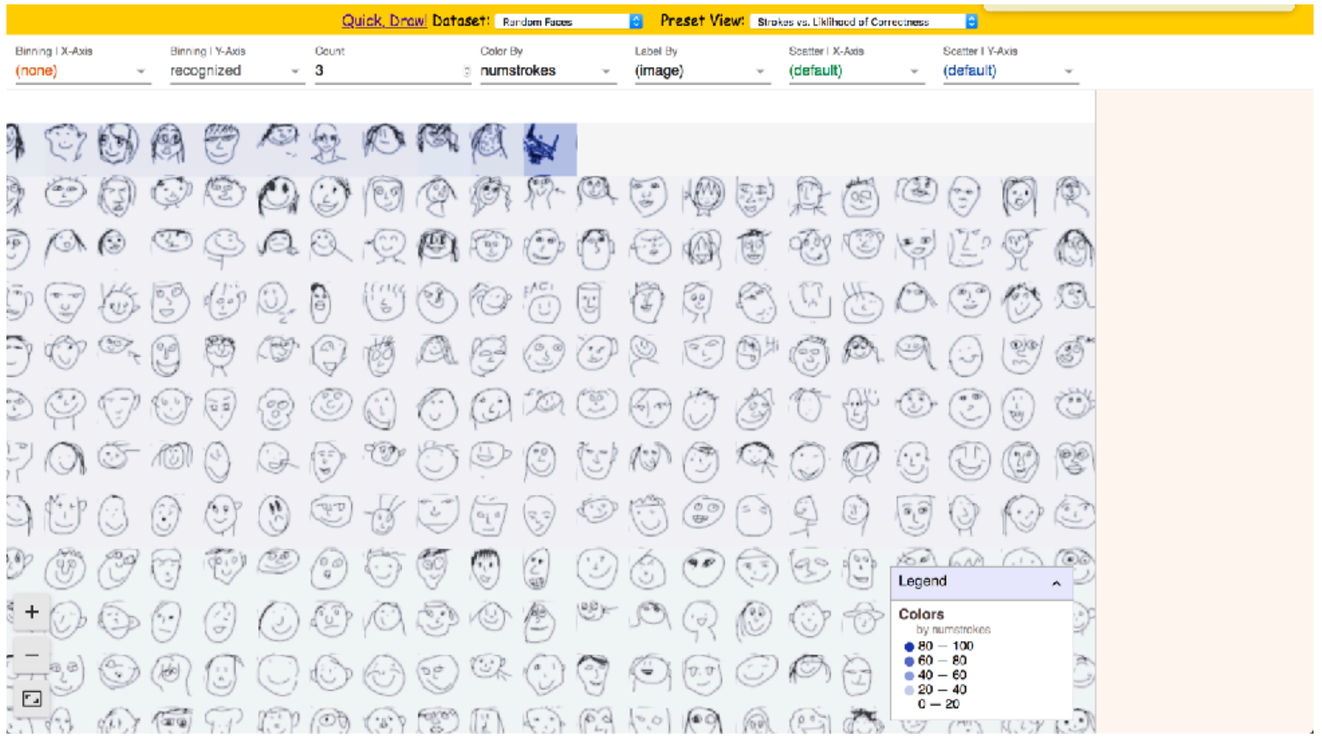


https://pair-code.github.io/facets/

Example : Quickdraw

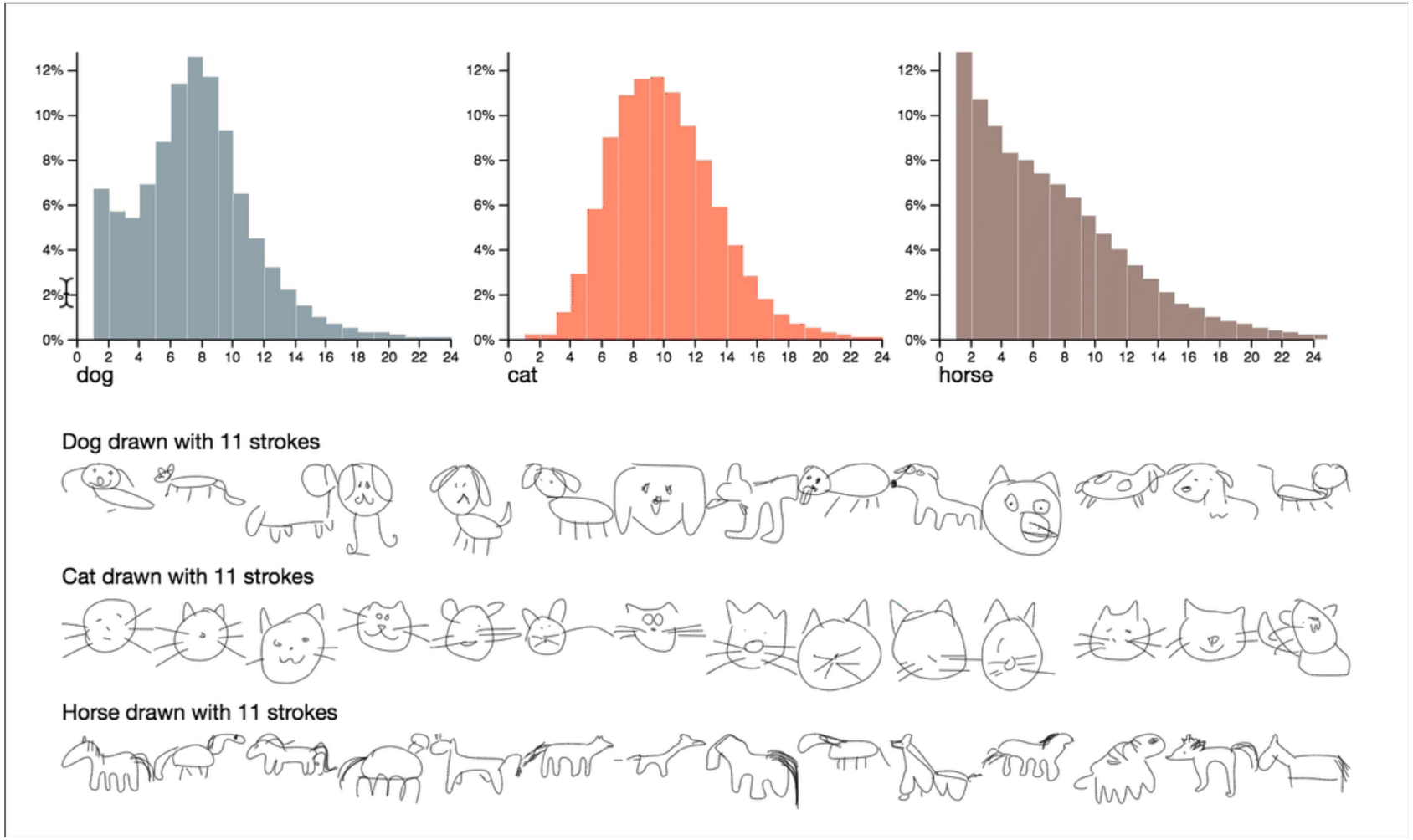
Example with the QuickDraw Data:

=> https://pair-code.github.io/facets/quickdraw.html



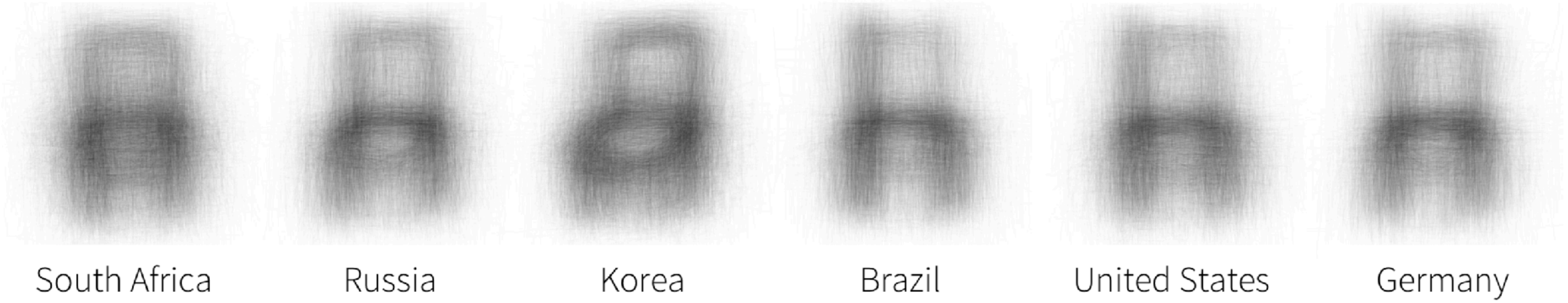
https://quickdraw.withgoogle.com/data/

Example : Quickdraw



https://medium.com/@enjalot/machine-learning-for-visualization-927a9dff1cab

Example : Quickdraw



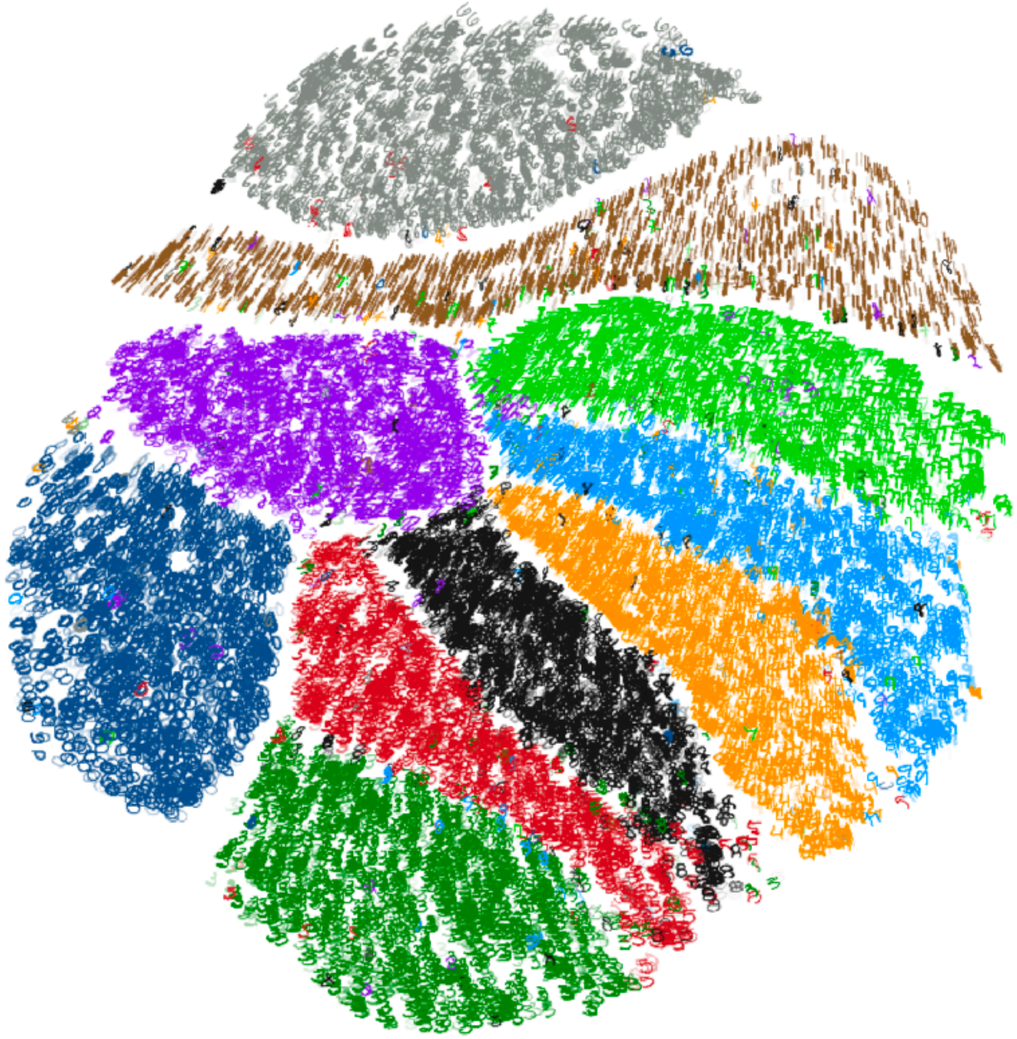
Fernanda Viégas, Martin Wattenberg, Visualization for Machine Learning (Google Brain)

Visualizing High-dimensional Data

Main approaches:

- Linear
  - Principal Component Analysis
  - Visualization of Labeled Data Using Linear Transformations (Koren & Carmel)
- Non-linear
  - Multidimensional scaling
  - Isomap
  - t-SNE

t-SNE



t-SNE

- Complex non-linear technique
- Goal: take a set of points in a high-dimensional space and find a faithful representation of those points in a lower-dimensional space (2D)
- The algorithm is non-linear and adapts to the underlying data, performing different transformations on different regions, using an adaptive sense of « distance. »
- Example: MNIST Visualization: https://nicola17.github.io/tfjs-tsne-demo/

t-SNE is Tricky!

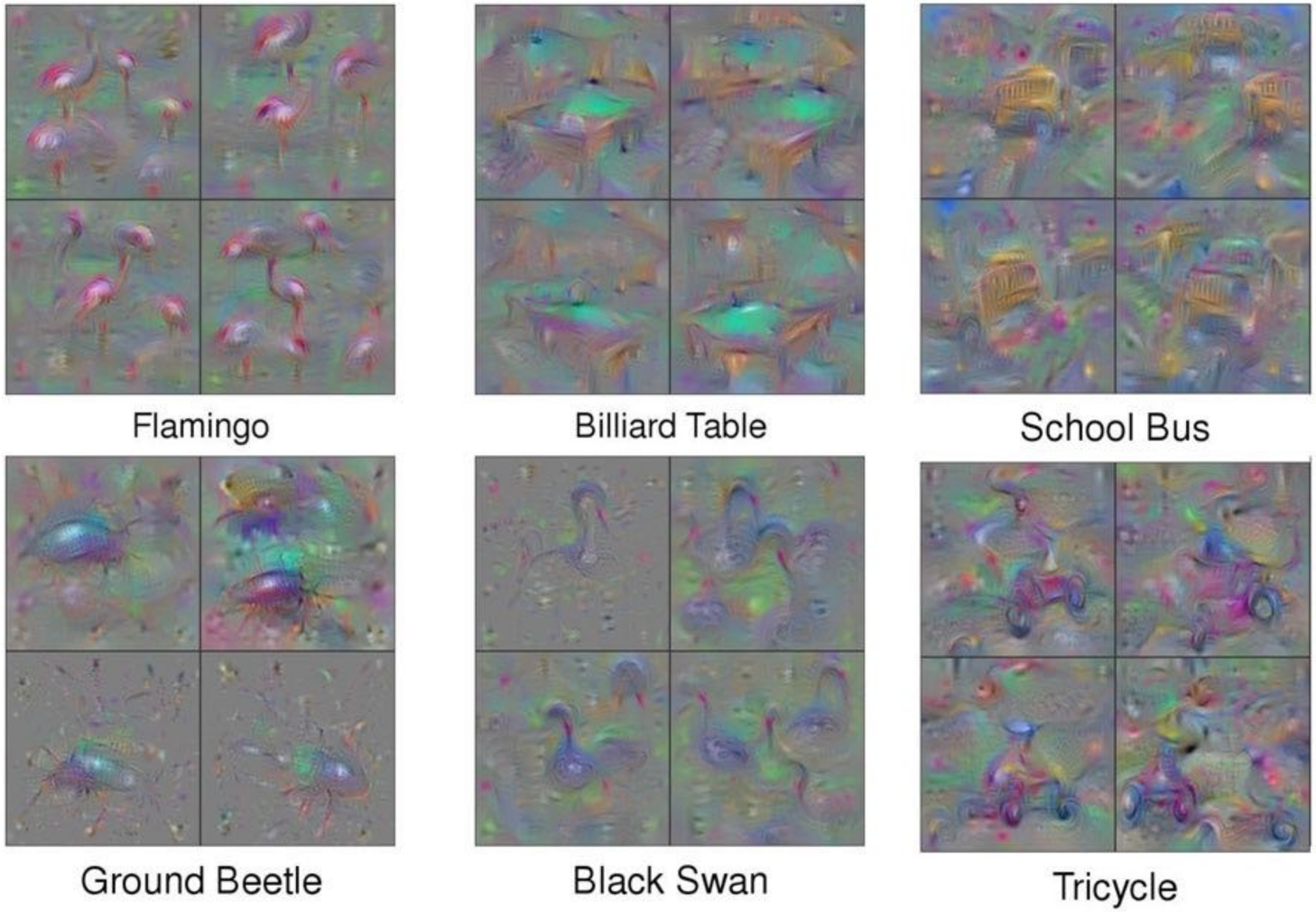
https://distill.pub/2016/misread-tsne/

Wayne McGregor - Living Archive

https://artsexperiments.withgoogle.com/living-archive

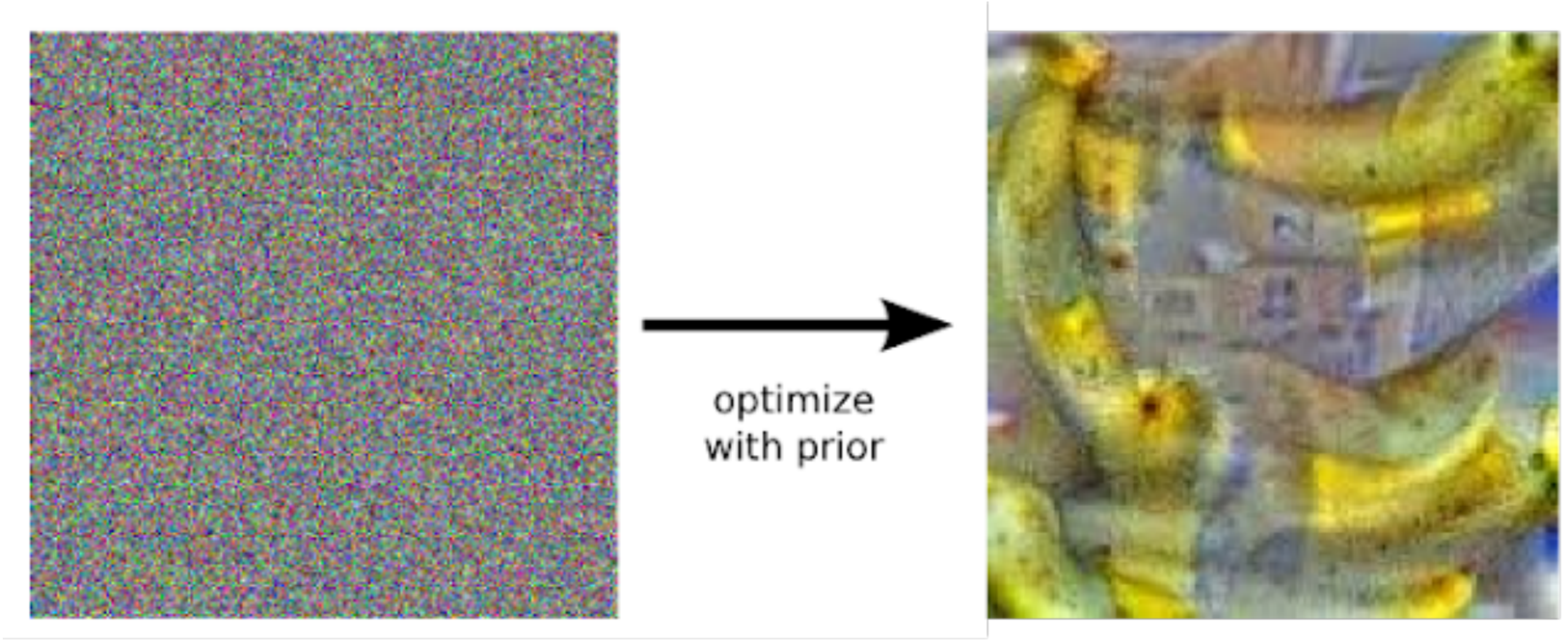
Visualizing Models

Deepvis



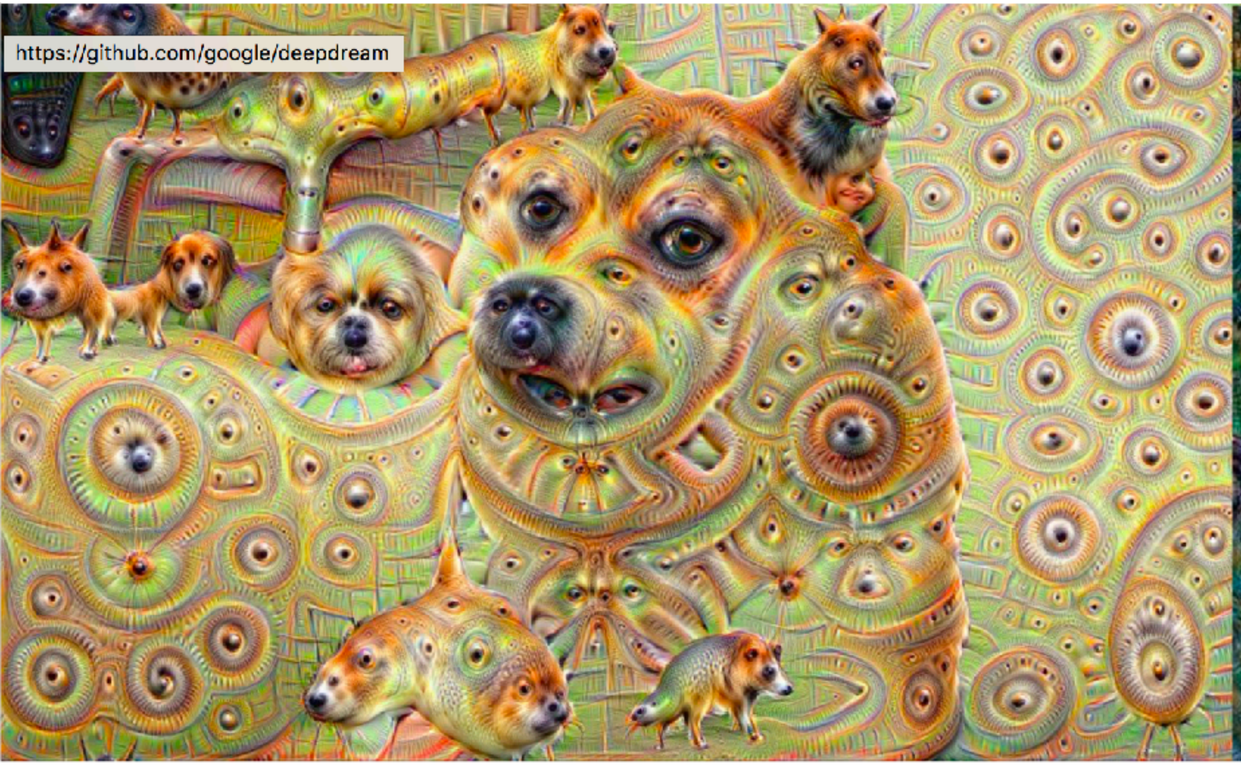
Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.

DeepDream



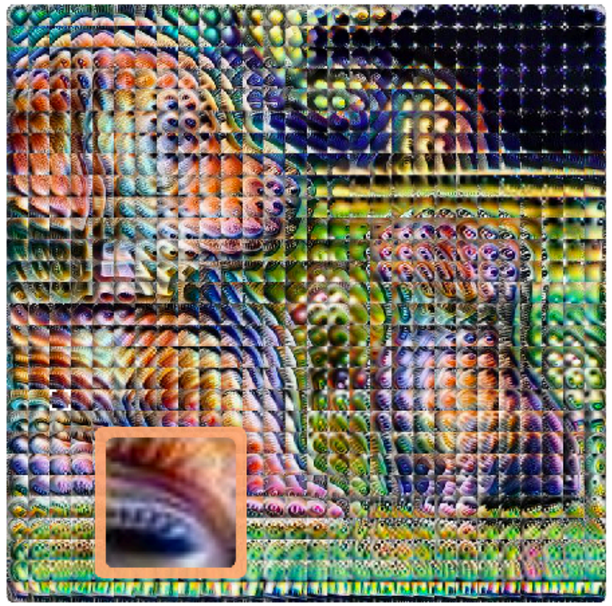
https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

DeepDream

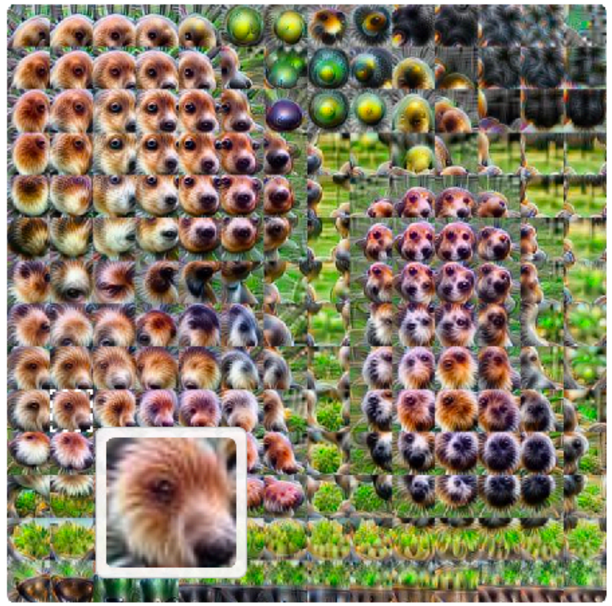


https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

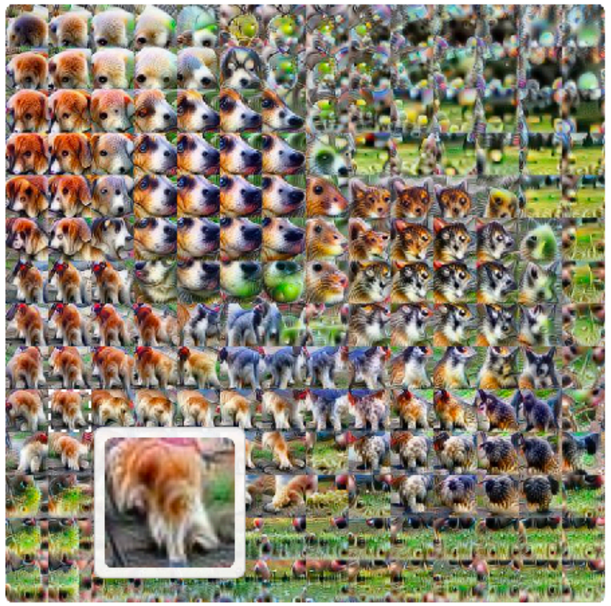
Combining techniques



MIXED3A



MIXED4A



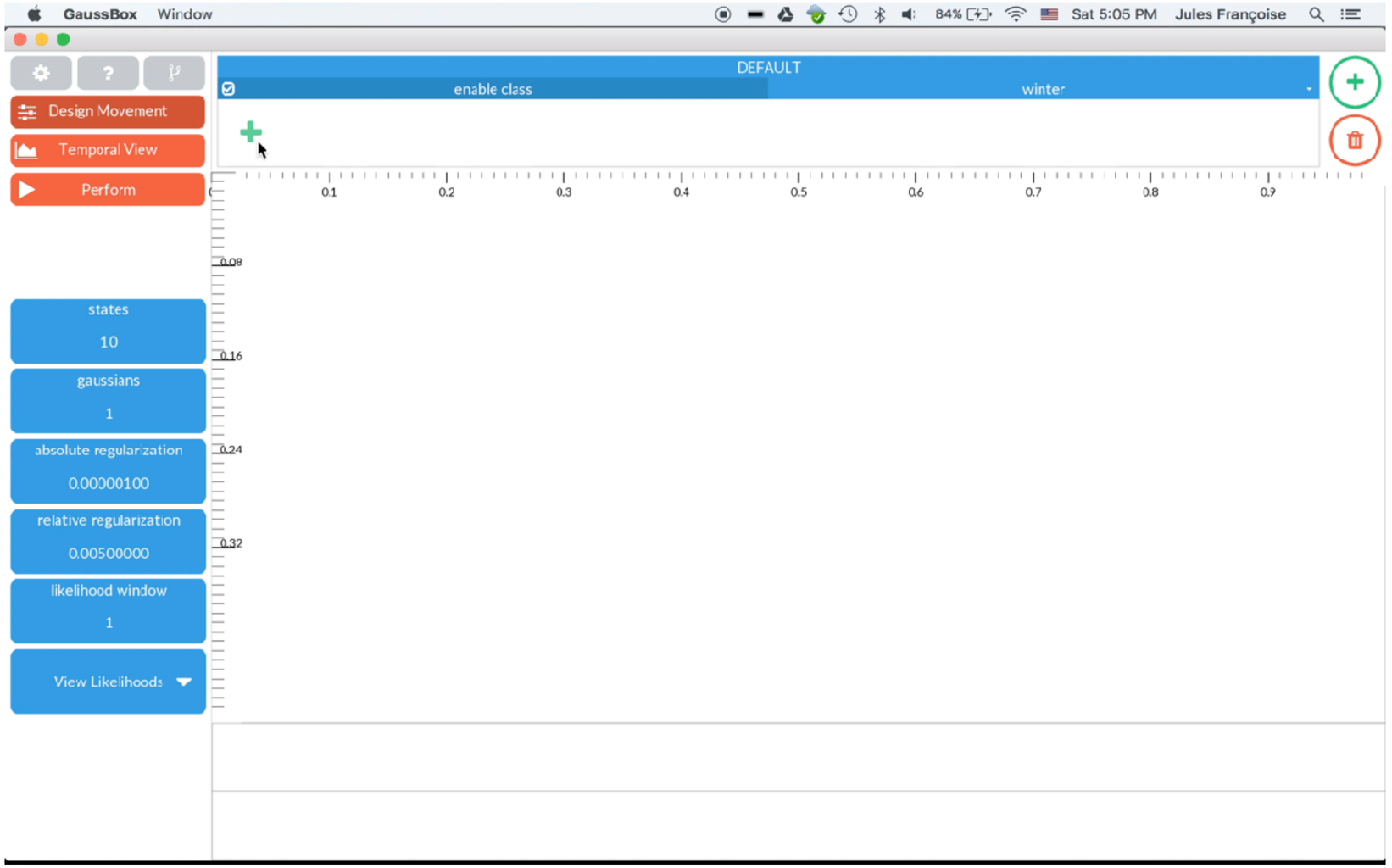
MIXED4D

MIXED5A

https://distill.pub/2018/building-blocks/

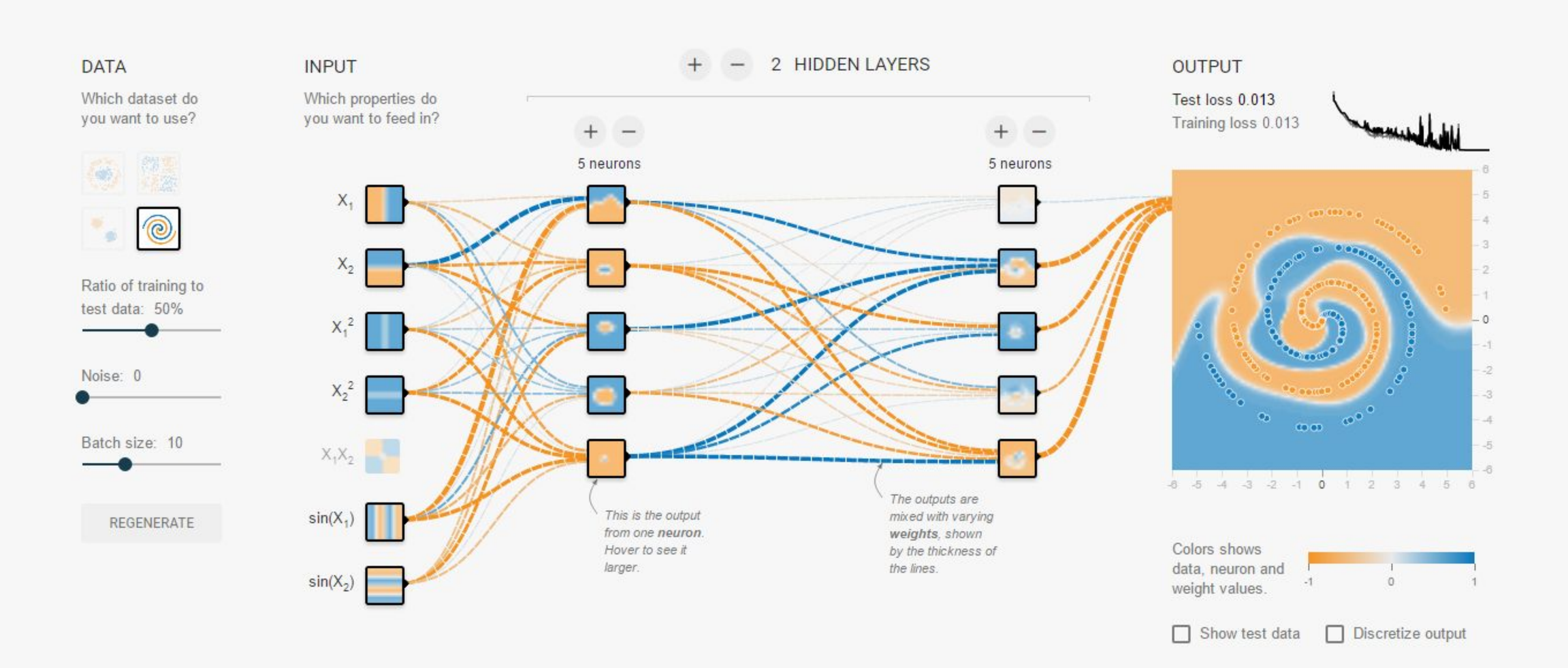
Visualizations for Education

Gaussbox



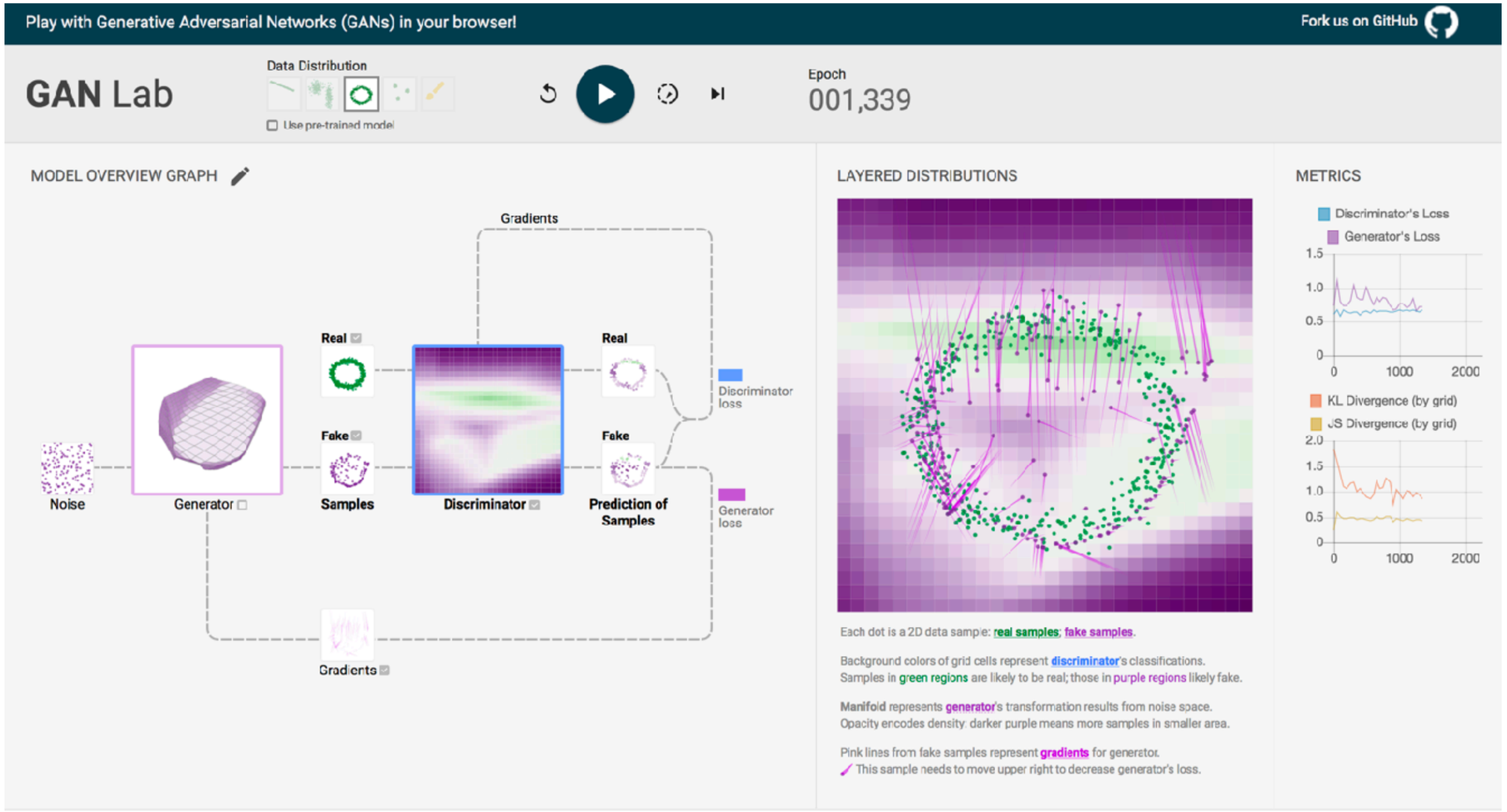
Françoise, J., Bevilacqua, F., & Schiphorst, T. (2016, May). GaussBox: Prototyping movement interaction with interactive visualizations of machine learning. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 3667-3670).

Tensorflow Playground



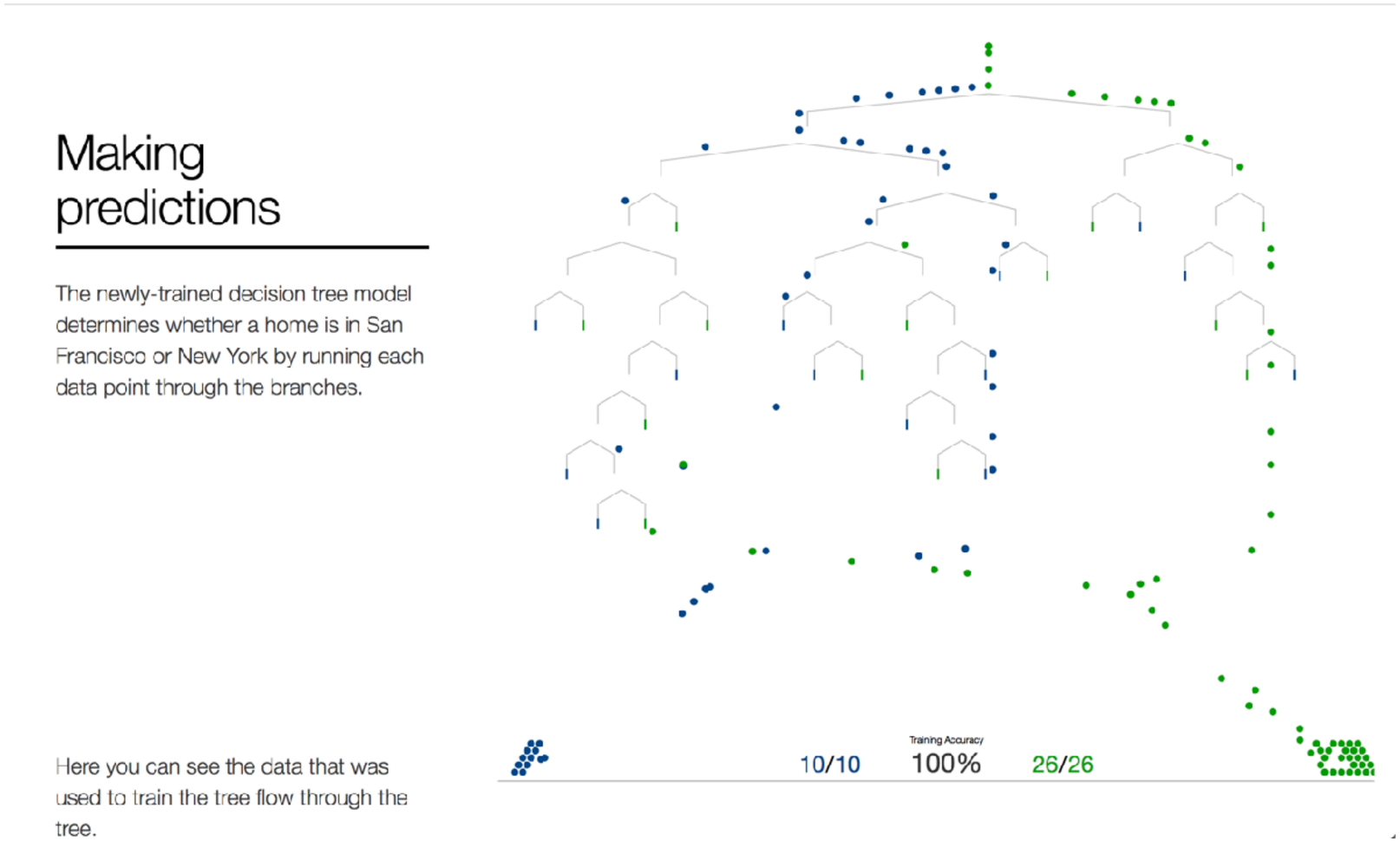
http://playground.tensorflow.org/

GAN Lab



https://poloclub.github.io/ganlab/

Visual intro to ML

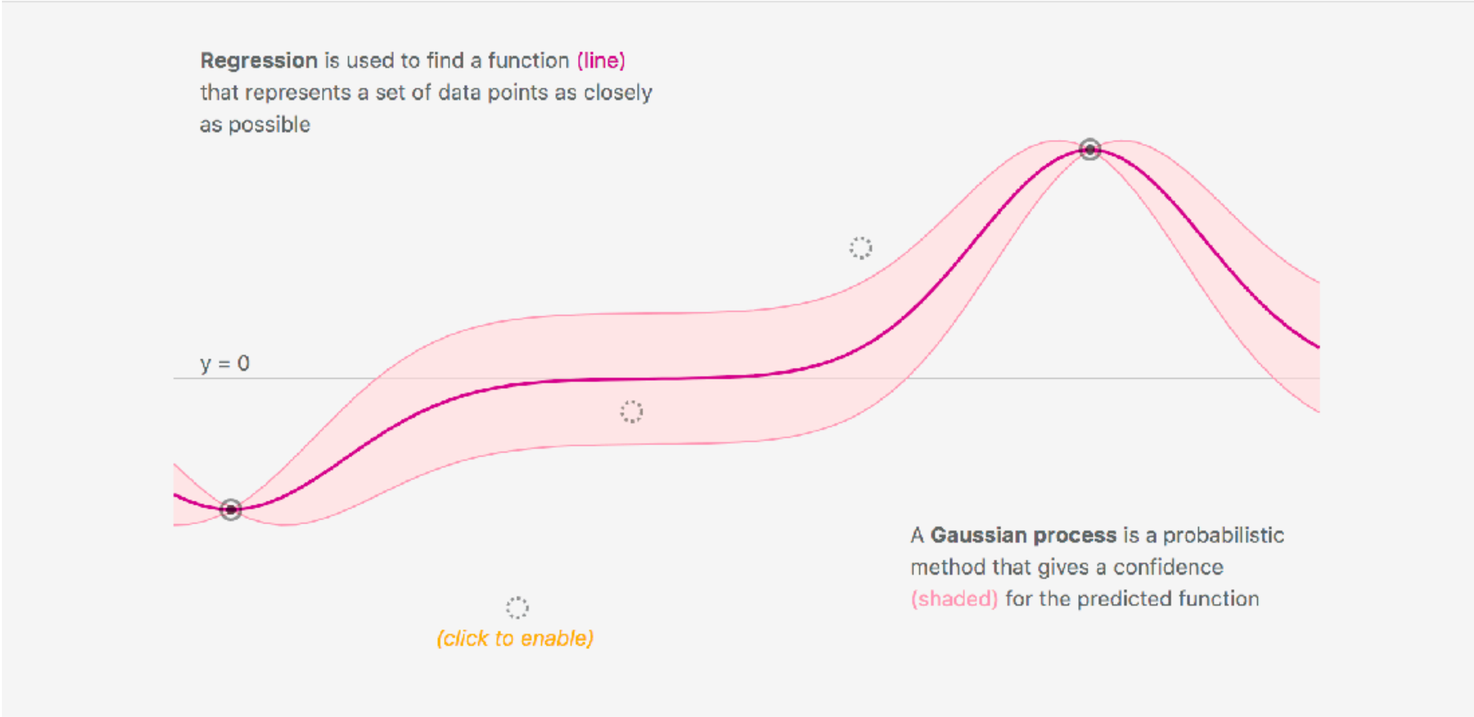


http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Distill.pub

A Visual Exploration of Gaussian Processes

How to turn a collection of small building blocks into a versatile tool for solving regression problems.



https://distill.pub/

Explanations

Explainable AI

Explainable-AI explores and investigates methods to **produce** or **complement** AI models to make accessible and interpretable the internal logic and the outcome of the algorithms, making such process **understandable by humans**.

Intrinsic or post hoc?

Interpretability can be achieved:

=> by restricting the complexity of the machine learning model (**intrinsic**)

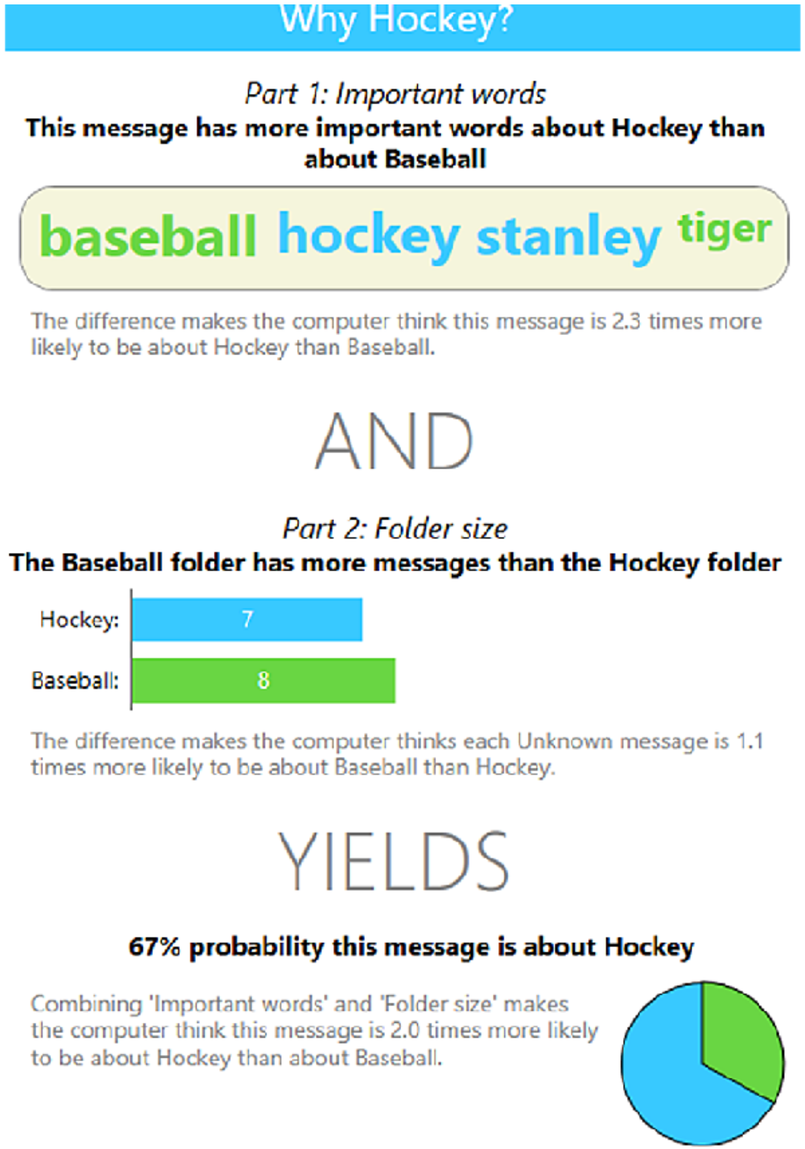
Some models are considered interpretable due to their simple structure (short decision trees, sparse linear models, …)

=> by applying methods that analyze the model after training (**post hoc**).

Application of (model-agnostic) interpretation methods after model training

Elucidebug

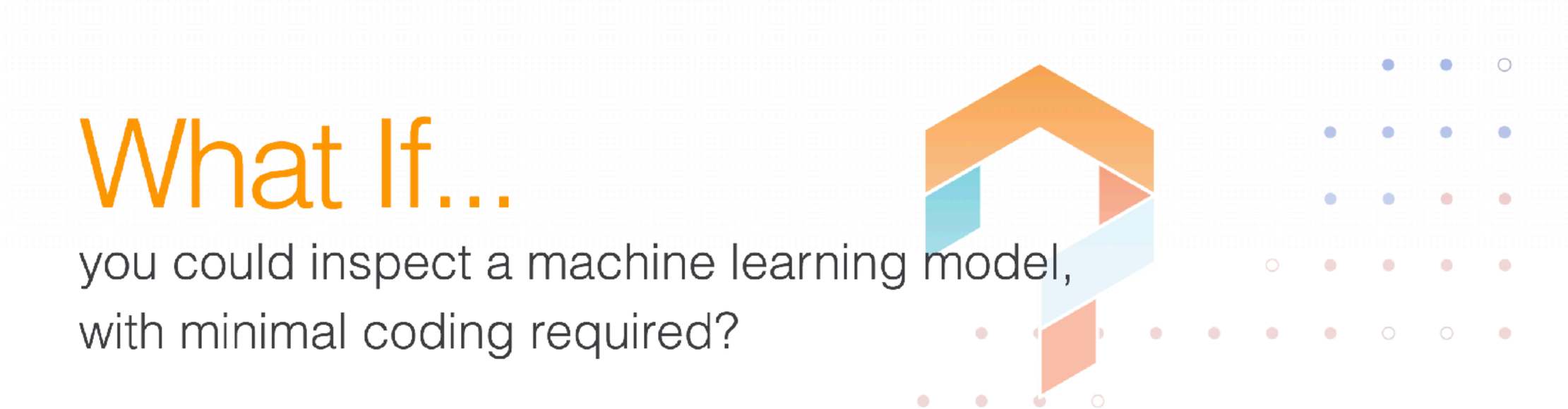
- **Task:** text analysis (spam filtering & classification)
- Goal: give users explanations about the algorithm's decisions
- Allow users to query the system about particular decisions
- Provides feedback about uncertainty



Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. Principles of explanatory debugging to personalize interactive machine learning. ACM IUI 2015

Example #2 : What-If ?

What-If



https://pair-code.github.io/what-if-tool/

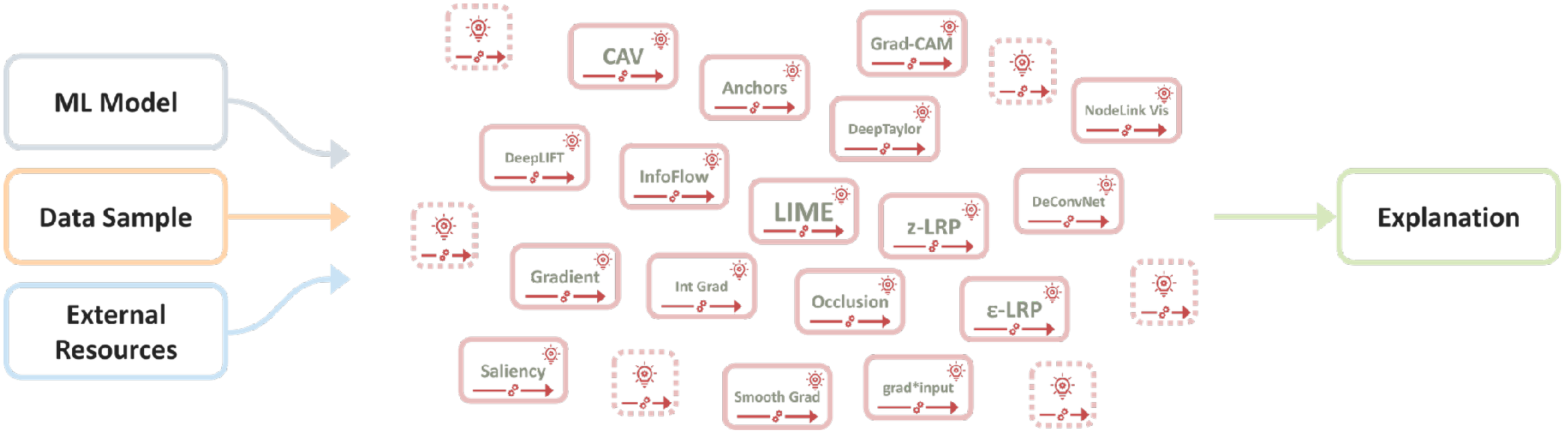
Example #3 : ExplAIner

ExplAIner

https://explainer.ai/

Spinner, T., Schlegel, U., Schäfer, H., & El-Assady, M. (2019). explAIner: A visual analytics framework for interactive and explainable machine learning. IEEE transactions on visualization and computer graphics, 26(1), 1064-1074.

ExplAIner



https://explainer.ai/

Spinner, T., Schlegel, U., Schäfer, H., & El-Assady, M. (2019). explAIner: A visual analytics framework for interactive and explainable machine learning. IEEE transactions on visualization and computer graphics, 26(1), 1064-1074.

Sources

- Sahin Cem Geyik, Krishnaram Kenthapadi & Varun Mithal, **Explainable AI in Industry** (KDD 2019 Tutorial)
- Wojciech Samek & Alexander Binder, **Tutorial on Interpretable Machine Learning** (MICCAI 2018).
- Fernanda Viégas, Martin Wattenberg, **Visualization for Machine Learning** (Google Brain)