The Deep Learning Revolution

Alex Acero

Agenda

The Deep Learning Revolution

Fundamentals of Deep Learning

Why now? A brief history

Transforming our Digital Lives

Acknowledgments

John Bridle and the Siri team

Josh Suskind, Sofien Bouaziz, and Apple’s Video Team

The Deep Learning Revolution

Technology Disruptions

Content Creation:

- Text
- Photography

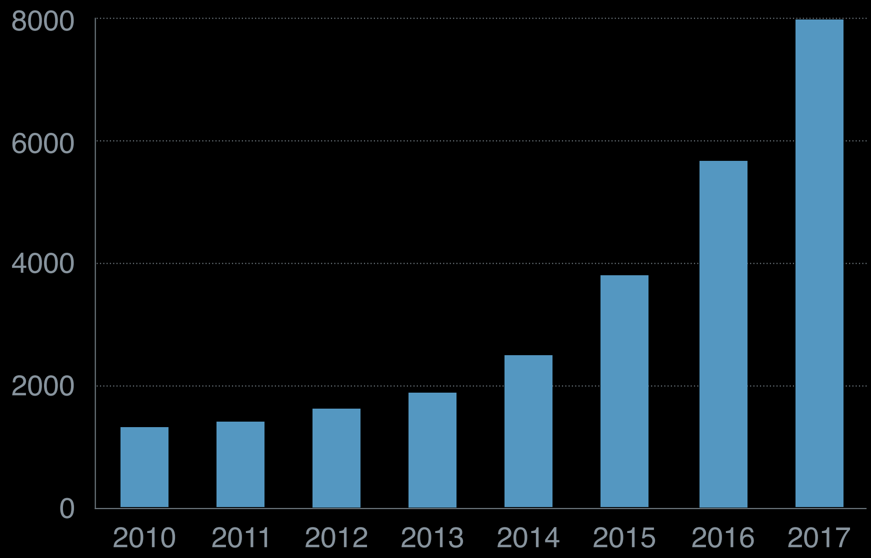
Content Consumption:

- Text
- Photography
- Music
- Video

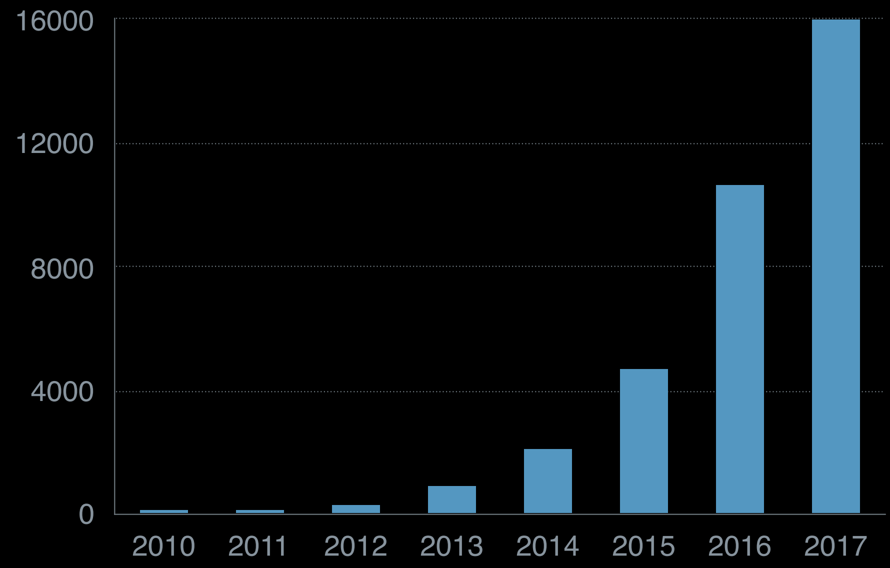
Our daily lives:

- Transportation
- Communication
- Shopping
- Travel

The Deep Learning Revolution



Neural Information Processing Systems (NIPS) Attendees

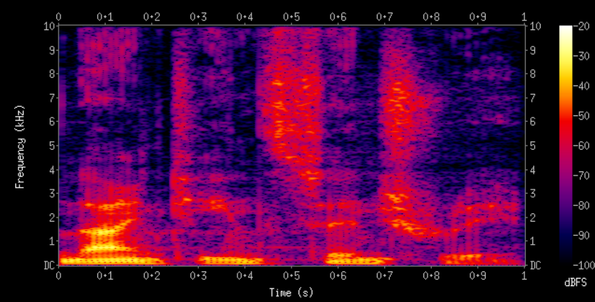


Papers with “Deep Neural Networks”

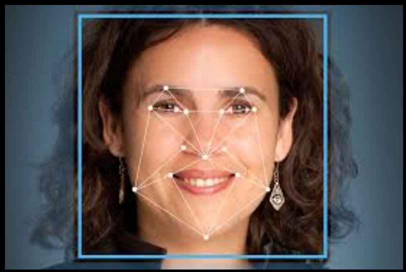
Fundamentals of Deep Learning

Binary Classification

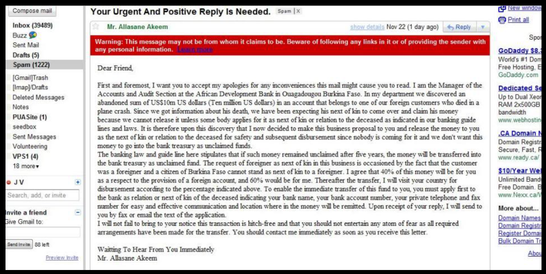
TouchID



Speaker Verification



Face ID



Email Spam

Motion Detection

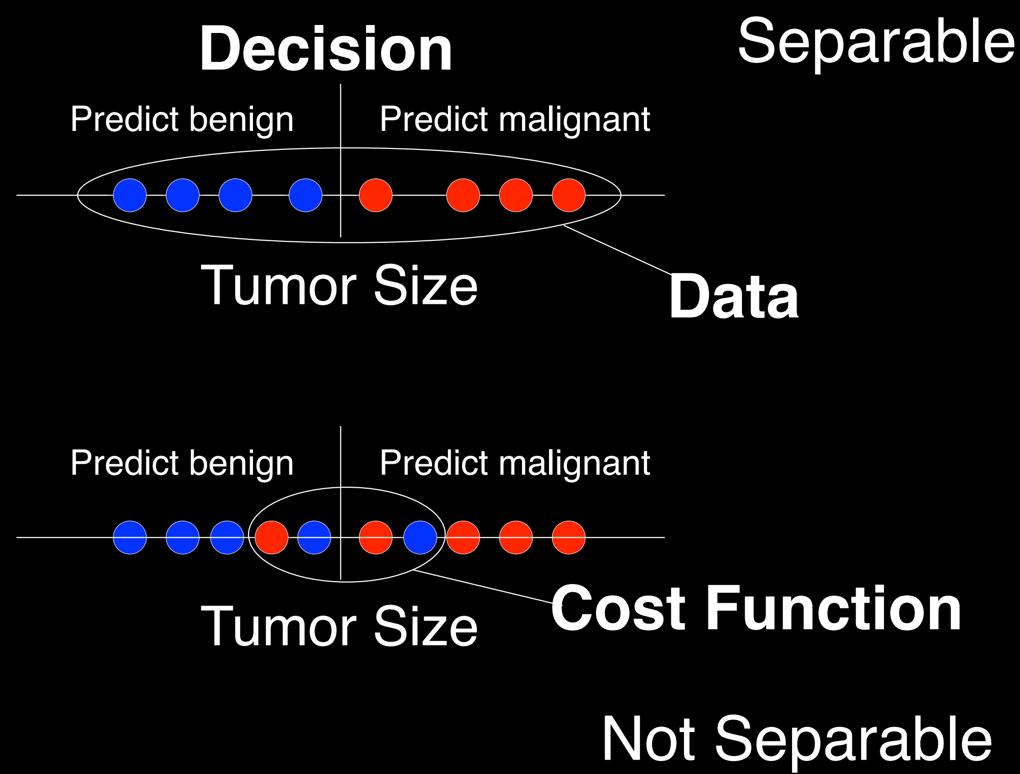
Credit Card Fraud

Binary Classification

Output Labels

Breast Cancer

- Benign
- Malignant



Input Features

Tumor Size

Binary Classification

Output Labels

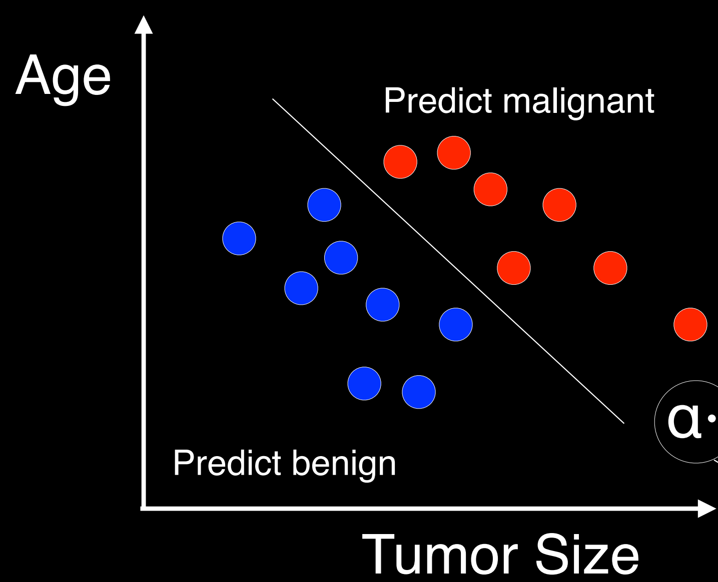
Breast Cancer

- Benign
- Malignant

Input Features

Tumor Size

Age



Linearly Separable

More features:

Clump thickness

Uniformity of cell size

…

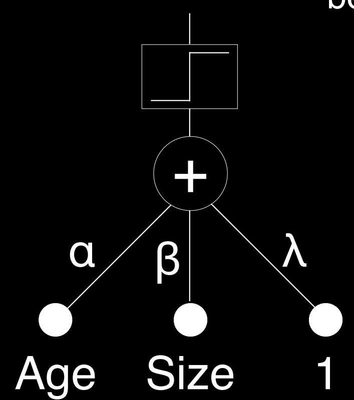
$$\alpha \cdot \text{Age} + \beta \cdot \text{Size} + \lambda > 0$$

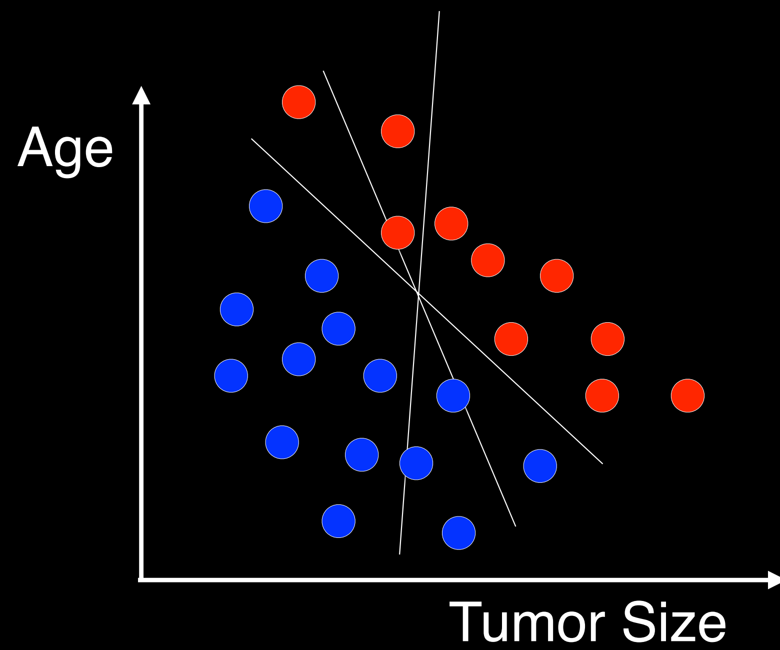
Learning

Perceptron Learning

Rosenblatt, 1958

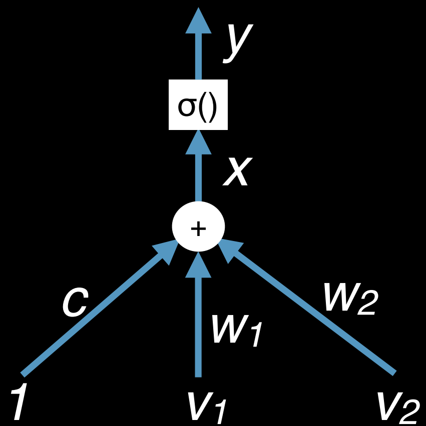
$$\alpha \cdot \text{Age} + \beta \cdot \text{Size} + \lambda \underset{\text{benign}}{\overset{\text{malignant}}{\gtrless}} 0$$





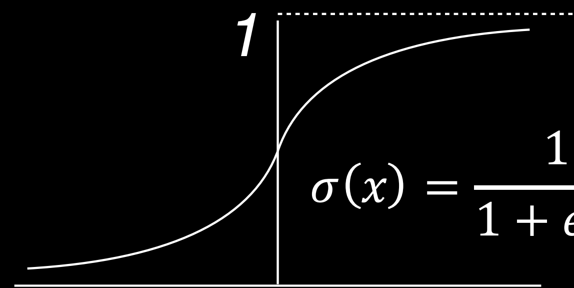
$$\alpha(i) = \alpha(i - 1) + \eta \cdot \{\text{Target}(i) - \text{Output}(i)\} \cdot \text{Age}(i)$$
$$\beta(i) = \beta(i - 1) + \eta \cdot \{\text{Target}(i) - \text{Output}(i)\} \cdot \text{Size}(i)$$
$$\lambda(i) = \lambda(i - 1) + \eta \cdot \{\text{Target}(i) - \text{Output}(i)\}$$

Stochastic Gradient Descent (SGD)



$$y = \sigma(x)$$

$$x = c + \mathbf{v}^T \mathbf{w}$$



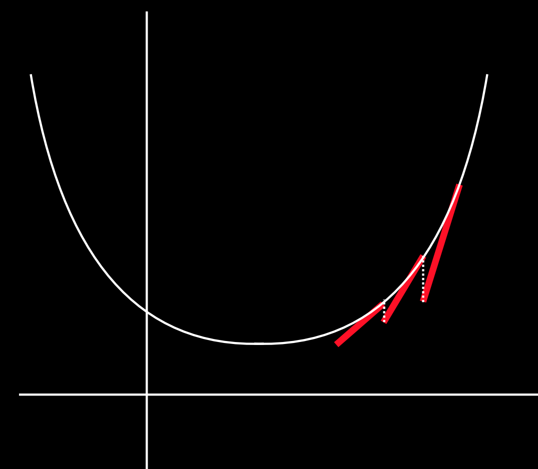
$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$p(t|\mathbf{v}) = y^t (1-y)^{1-t}$$

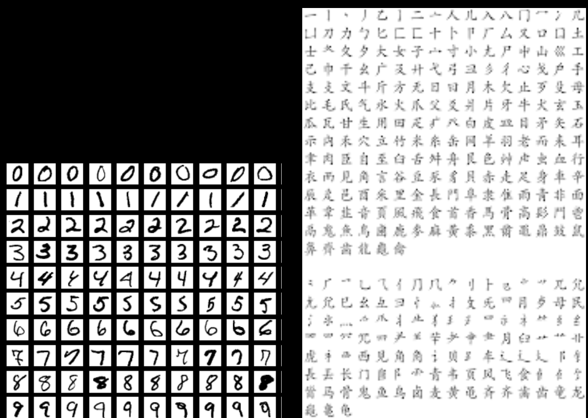
$$L = \ln p(t|\mathbf{v}) = t \ln y + (1-t) \ln(1-y)$$

$$\frac{\partial L}{\partial w_1} = \left(\frac{\partial L}{\partial y}\right)\left(\frac{\partial y}{\partial x}\right)\left(\frac{\partial x}{\partial w_1}\right) = (y-t)v_1$$

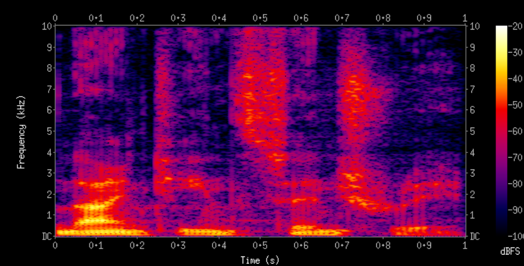
$$w_j^{(i)} = w_j^{(i-1)} - \eta \frac{\partial L}{\partial w_j} = w_j^{(i-1)} + \eta v_j \left(t - y^{(i-1)}\right)$$



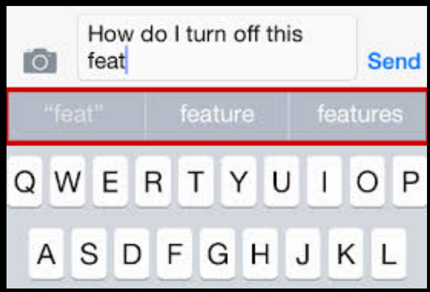
N-ary Classification



Handwriting Recognition

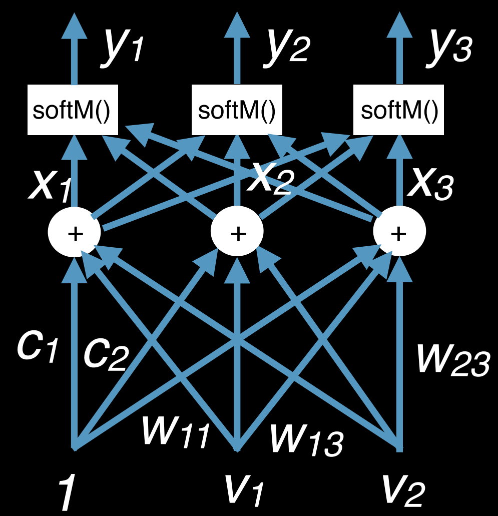


Speaker Identification



Word prediction

N-ary Classification



$$y_i = p(i|\mathbf{v}) = \frac{e^{x_i}}{\sum_{l=1}^{N} e^{x_l}}$$

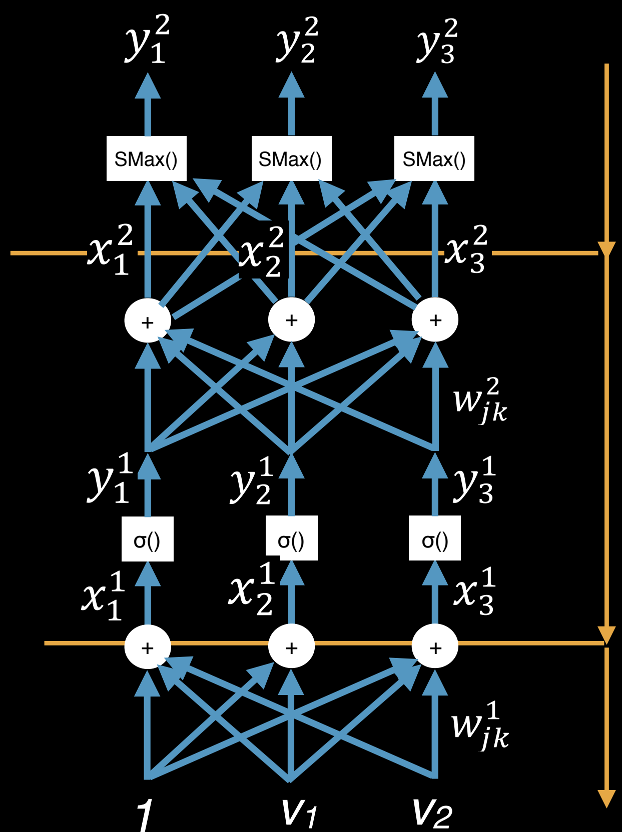
Softmax

$$L = \sum_{i=1}^{N} t_i \ln y_i$$

$$w_{nj}^{(i)} = w_{nj}^{(i-1)} + \eta v_n^{(i-1)}\left(t_j - \mathrm{y}_j^{(i-1)}\right)$$

Perceptron Learning

Werbos, 1974; Rumelhart, Hinton, Williams 1986



Two-layers

2 input features

3 output labels

$$\nabla_n^2(m) = y_n^2(m) - t_n(m)$$

$$\left[w_{jn}^2\right]^{(i)} = \left[w_{jn}^2\right]^{(i-1)} - \eta \frac{1}{M} \sum_{m=1}^{M} y_n^1(m) \nabla_n^2(m)$$

$$\nabla_n^1(m) = y_n^1(m)\left(1 - y_n^1(m)\right) \sum_{k=1}^{N} w_{nk}^2 \nabla_k^2(m)$$

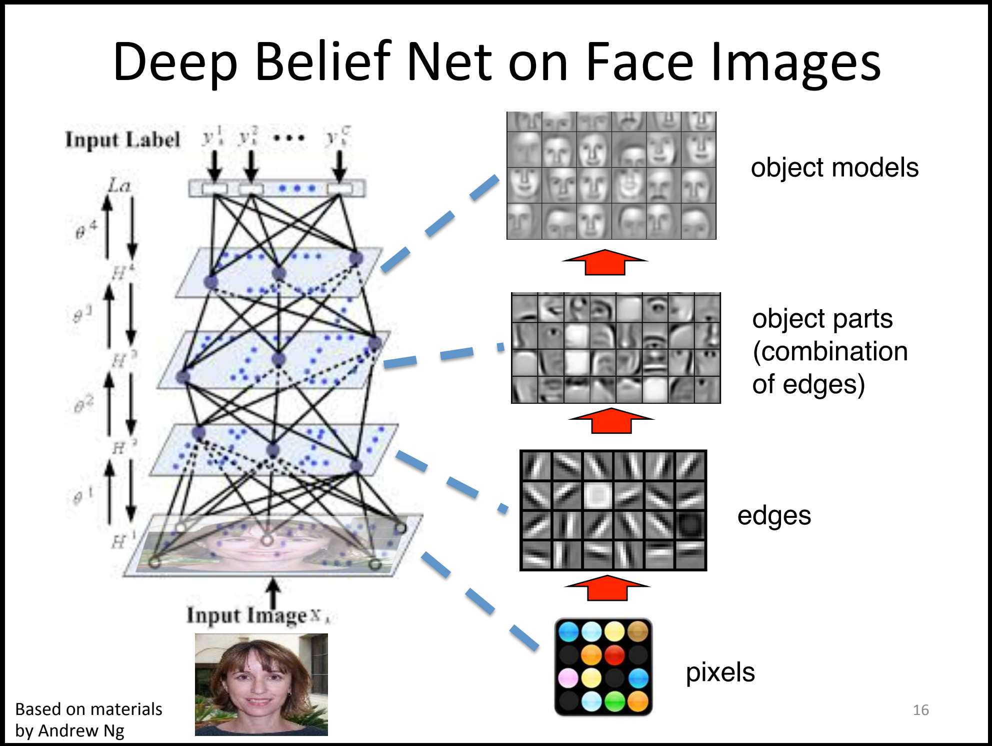
$$\left[w_{jn}^1\right]^{(i)} = \left[w_{jn}^1\right]^{(i-1)} - \eta \frac{1}{M} \sum_{m=1}^{M} v_j(m) \nabla_n^1(m)$$

backpropagation

Mini-batch

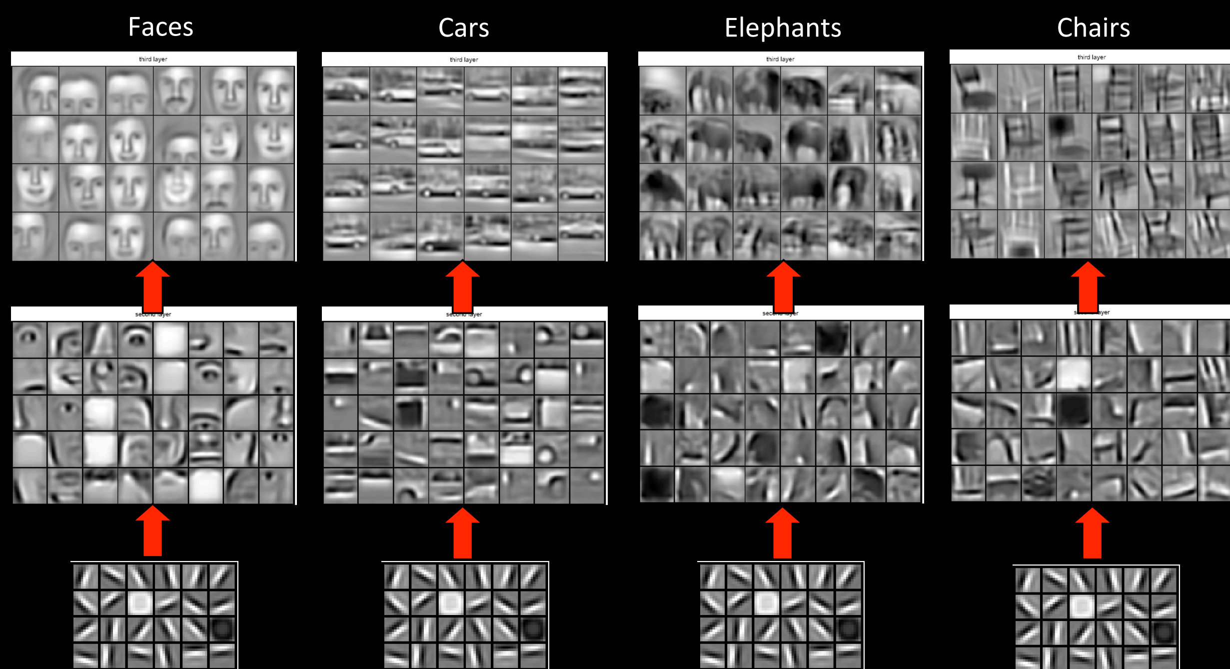
CNN on Face Images

2012



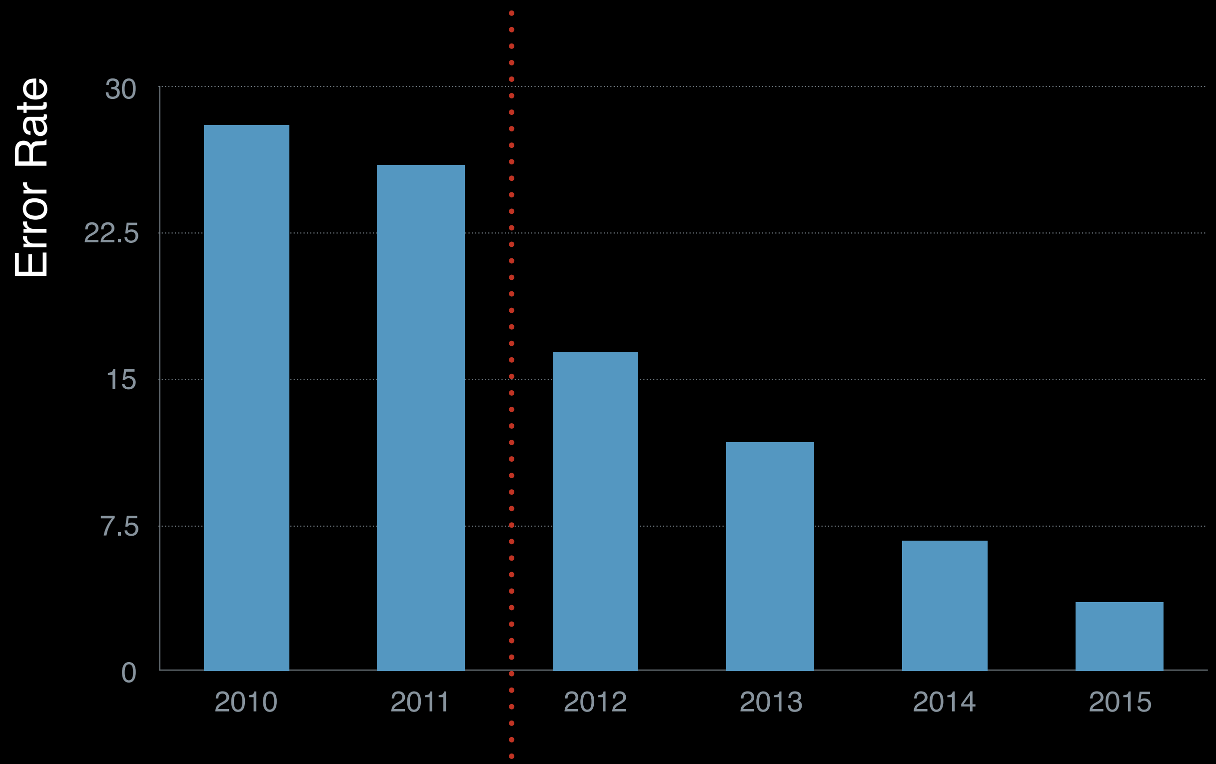
ImageNet Large Scale
Visual Recognition Challenge, 2012

Examples of learned object parts from object categories



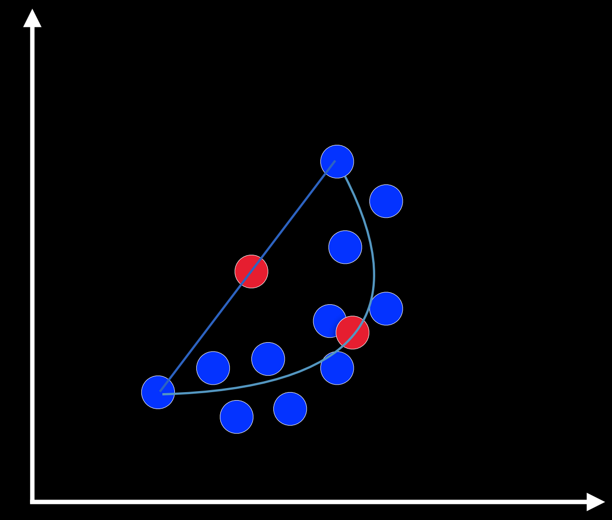
Credit: Andrew Ng

ImageNet Task Progress



Non-Linear Manifolds

Such non-linearity requires multiple layers

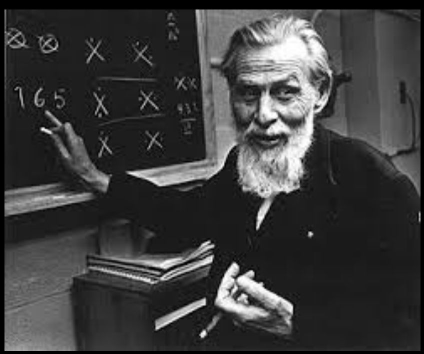


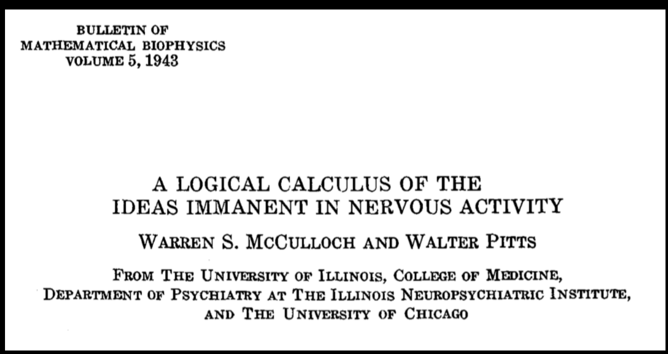
Why Now?

A brief history

McCulloch-Pitts Neurons

1943



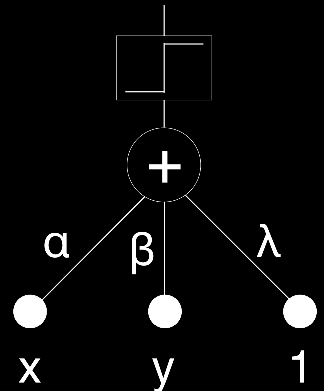


BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943

A LOGICAL CALCULUS OF THE
IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,
AND THE UNIVERSITY OF CHICAGO



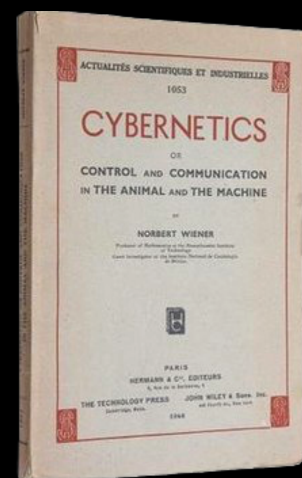
Norbert Wiener

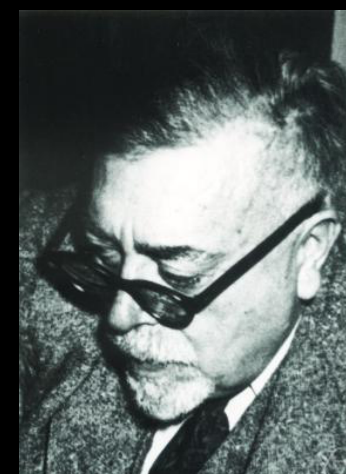
Wiener–Khinchin Theorem (1930)
Wiener Filter (1949)

McCulloch & Pitts joined Wiener at MIT (1943)

Cybernetics (1948)
  5. Computing Machines and the Nervous System
  10. Brain Waves and Self-Organising Systems

Suggested chess playing programs





Checkers

Arthur Samuel, 1956

ELIZA

Weizenbaum,1966

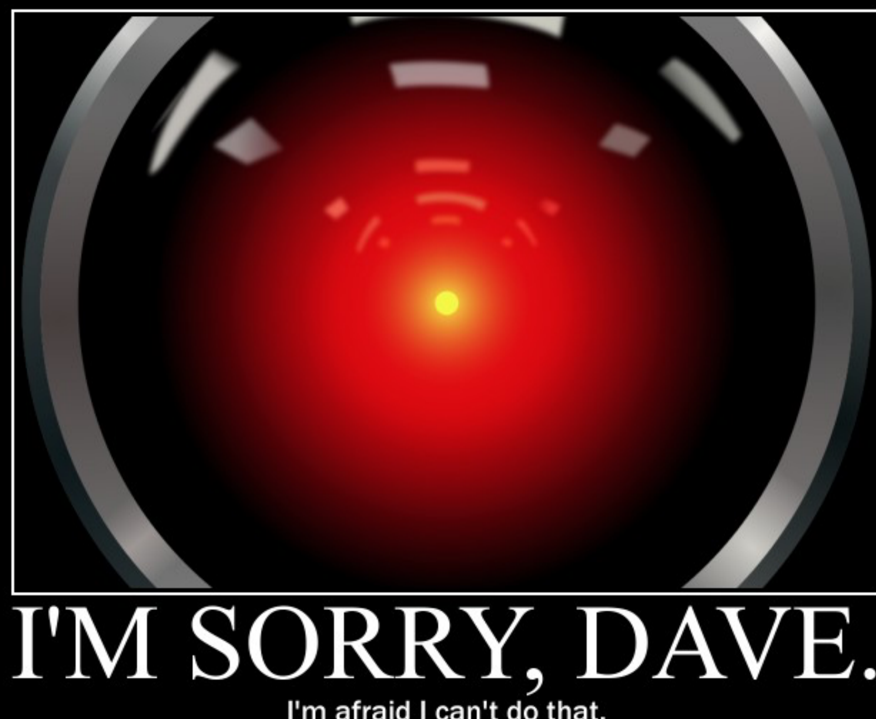
```
===============================================================================
           EEEEEEEE  L         IIIIIII  ZZZZZZZ      AAA
           E         L            I         Z      A   A
           E         L            I        Z      A     A
           EEEEE     L            I       Z       A     A
           E         L            I      Z        AAAAAAA
           E         L            I     Z         A     A
           EEEEEEEE  LLLLLLLL  IIIIIII  ZZZZZZ    A     A

===============================================================================
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...

===============================================================================
```

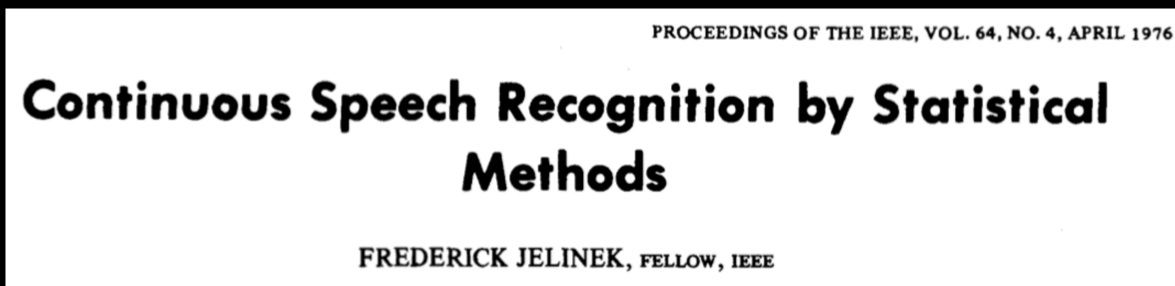
2001 Space Odyssey

Stanley Kubrick, 1968



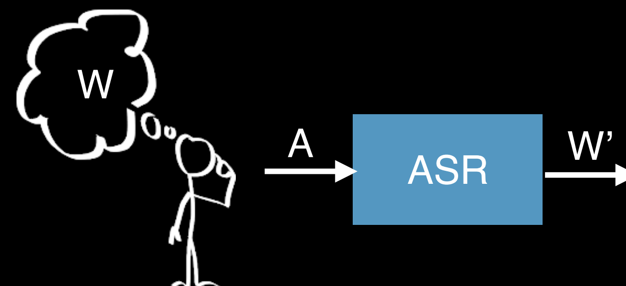
AI Winter

Fundamental Equation of Speech Recognition


PROCEEDINGS OF THE IEEE, VOL. 64, NO. 4, APRIL 1976

Continuous Speech Recognition by Statistical Methods

FREDERICK JELINEK, FELLOW, IEEE



$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p(W|A) = \underset{W}{\operatorname{argmax}}\, p(A|W)p(W) = \underset{W}{\operatorname{argmax}}\{\ln p(A|W) + \ln p(W)\}$$

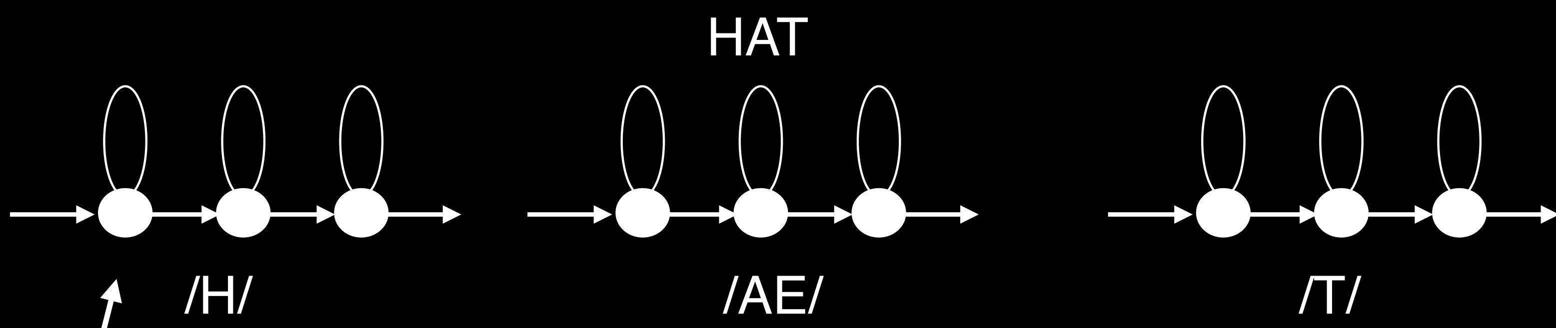
$$\hat{W} = \underset{W}{\operatorname{argmax}}\{\lambda \ln p(A|W) + \ln p(W)\}$$

Acoustic Model

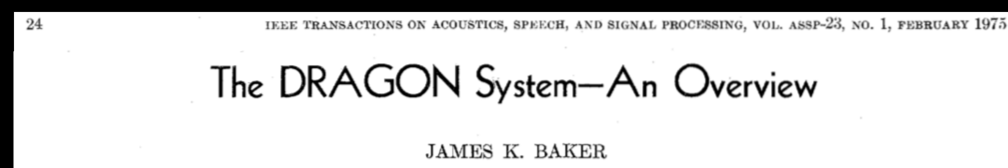
Language Model

Acoustic Model

Hidden Markov Models



$$p(a_t|s_j) = \sum_{i=1}^{I} \alpha_i \mathcal{N}\left(a_t, \mu_{ij}, \Sigma_{ij}\right)$$



24 IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-23, NO. 1, FEBRUARY 1975

The DRAGON System—An Overview

JAMES K. BAKER

PROCEEDINGS OF THE IEEE, VOL. 64, NO. 4, APRIL 1976

Continuous Speech Recognition by Statistical Methods

FREDERICK JELINEK, FELLOW, IEEE

Neural Networks for Speech Recognition in the 1990’s

328 IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

Merging Multilayer Perceptrons and Hidden Markov Models: Some Experiments in Continuous Speech Recognition

H. Bourlard[1], N. Morgan[2]

TR-89-033

July 1989

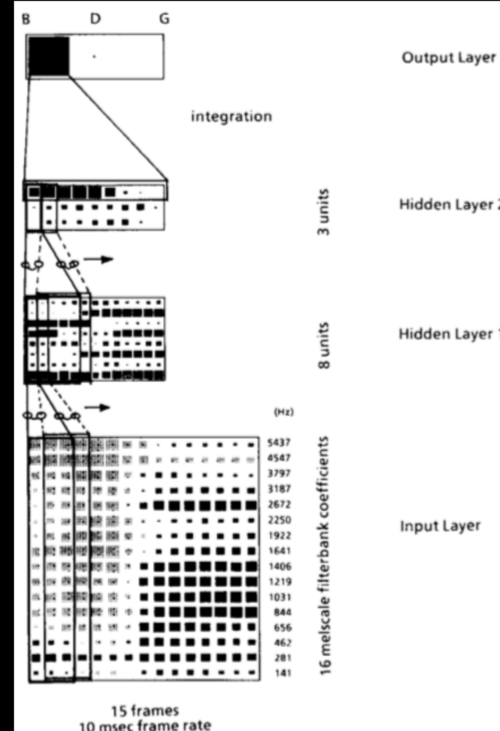
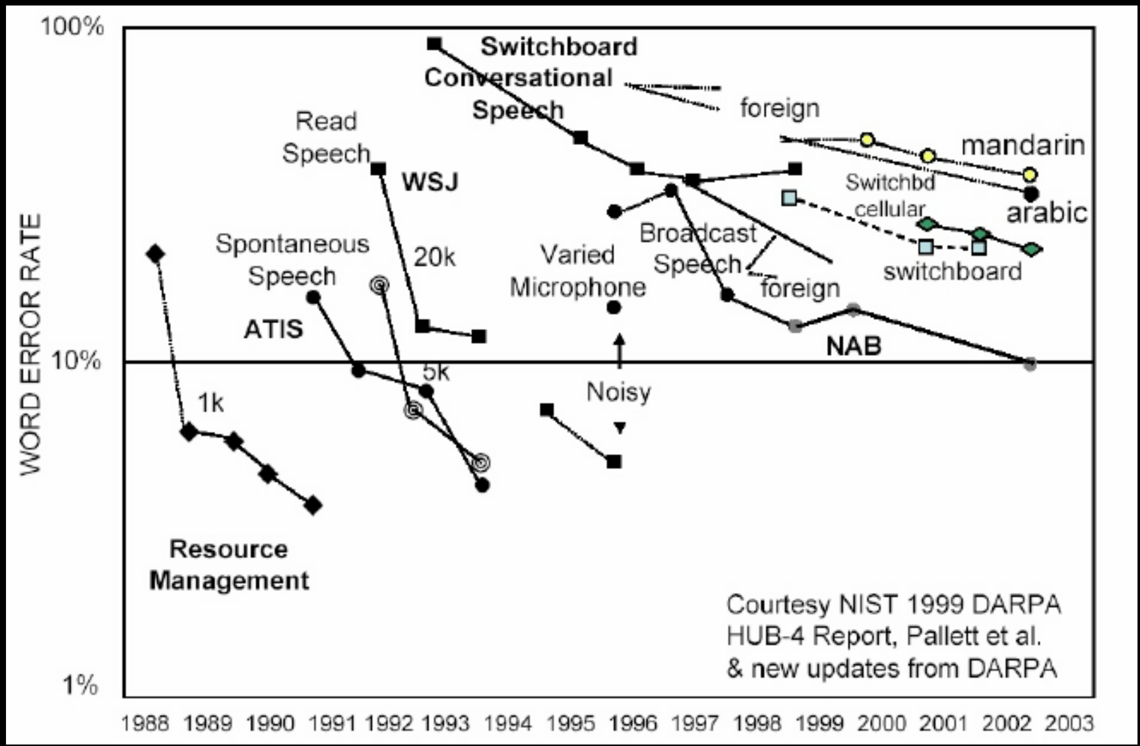


Fig. 2. The architecture of the TDNN.

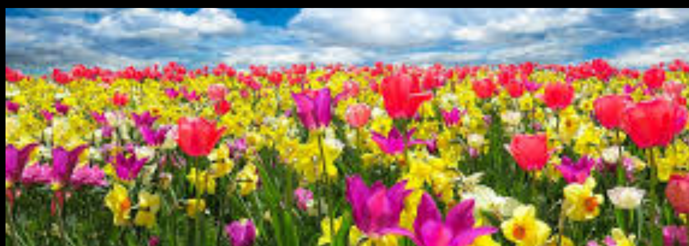
Neural Network Winter for Speech Recognition

Open Challenge Tasks

DARPA



Deep Learning



Reducing the Dimensionality of Data with Neural Networks

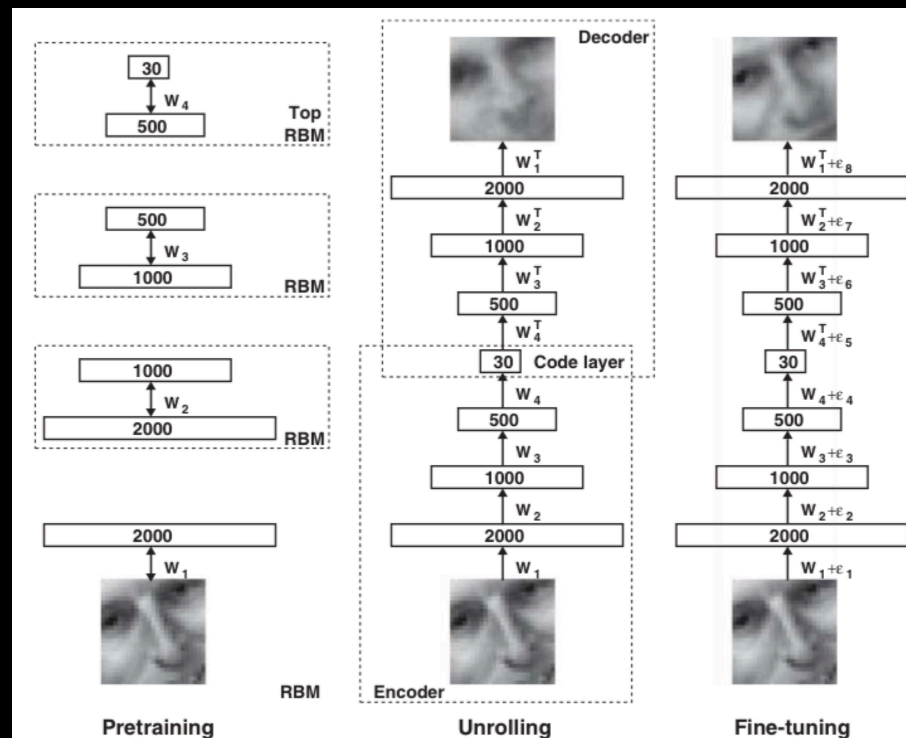
G. E. Hinton* and R. R. Salakhutdinov

28 JULY 2006 VOL 313 **SCIENCE** www.sciencemag.org

Deep Boltzmann Machines

Ruslan Salakhutdinov
Department of Computer Science
University of Toronto
rsalakhu@cs.toronto.edu

Geoffrey Hinton
Department of Computer Science
University of Toronto
hinton@cs.toronto.edu



Deep Belief Networks ➡ Deep Neural Networks

INTERSPEECH 2010

Investigation of Full-Sequence Training of Deep Belief Networks for Speech Recognition

Abdel-rahman Mohamed [1], Dong Yu[2], Li Deng[2]*

[1] Department of Computer Science, University of Toronto, Toronto, ON Canada
[2] Microsoft Research, Redmond, WA USA

INTERSPEECH 2011

Conversational Speech Transcription Using Context-Dependent Deep Neural Networks

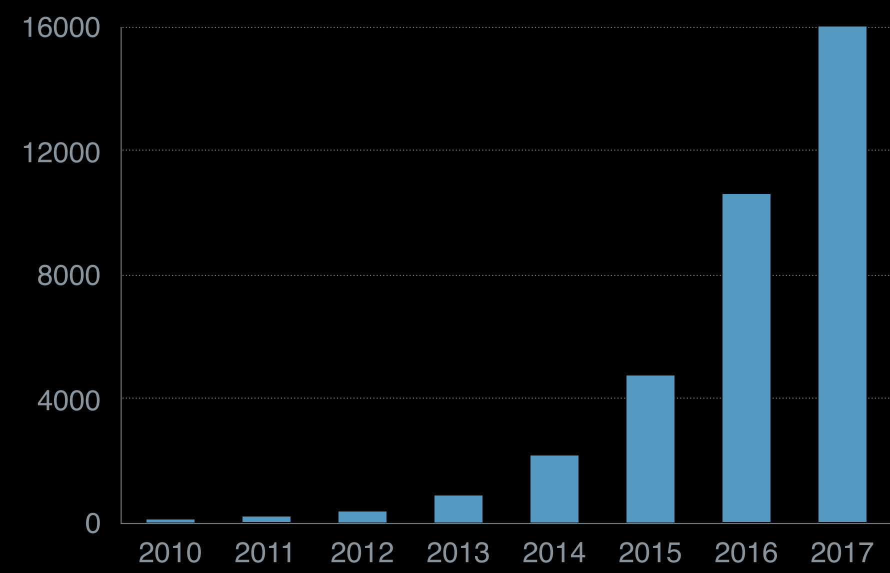
Frank Seide[1], Gang Li,[1] and Dong Yu[2]

[1]Microsoft Research Asia, Beijing, P.R.C.
[2]Microsoft Research, Redmond, USA

30 IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 1, JANUARY 2012

Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition

George E. Dahl, Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, and Alex Acero, *Fellow, IEEE*



Papers with “Deep Neural Networks”

Deep Learning for Speech

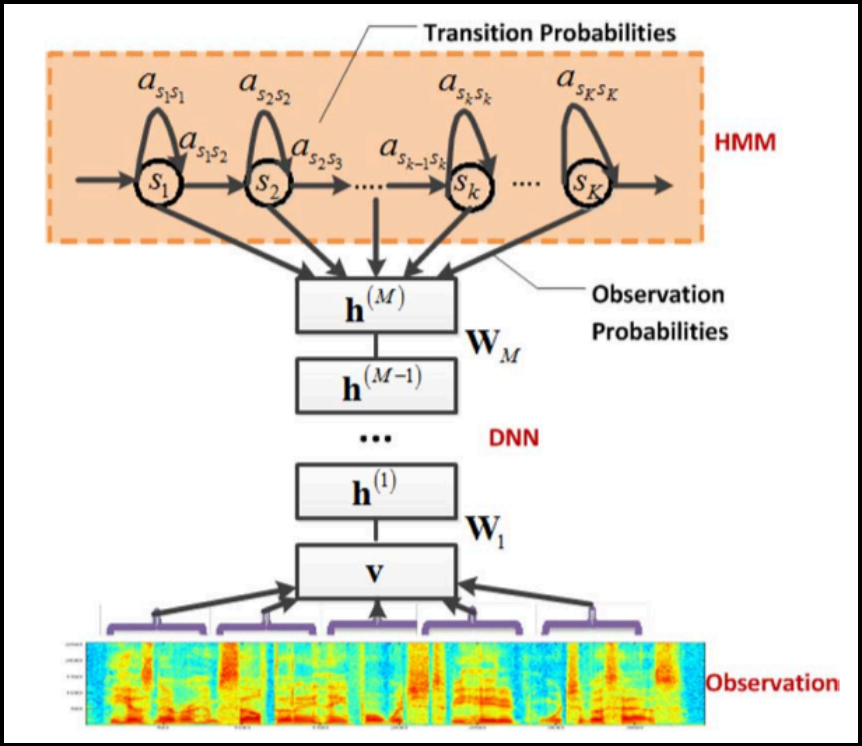
Deng et al., 2010

DNNs for large vocabulary

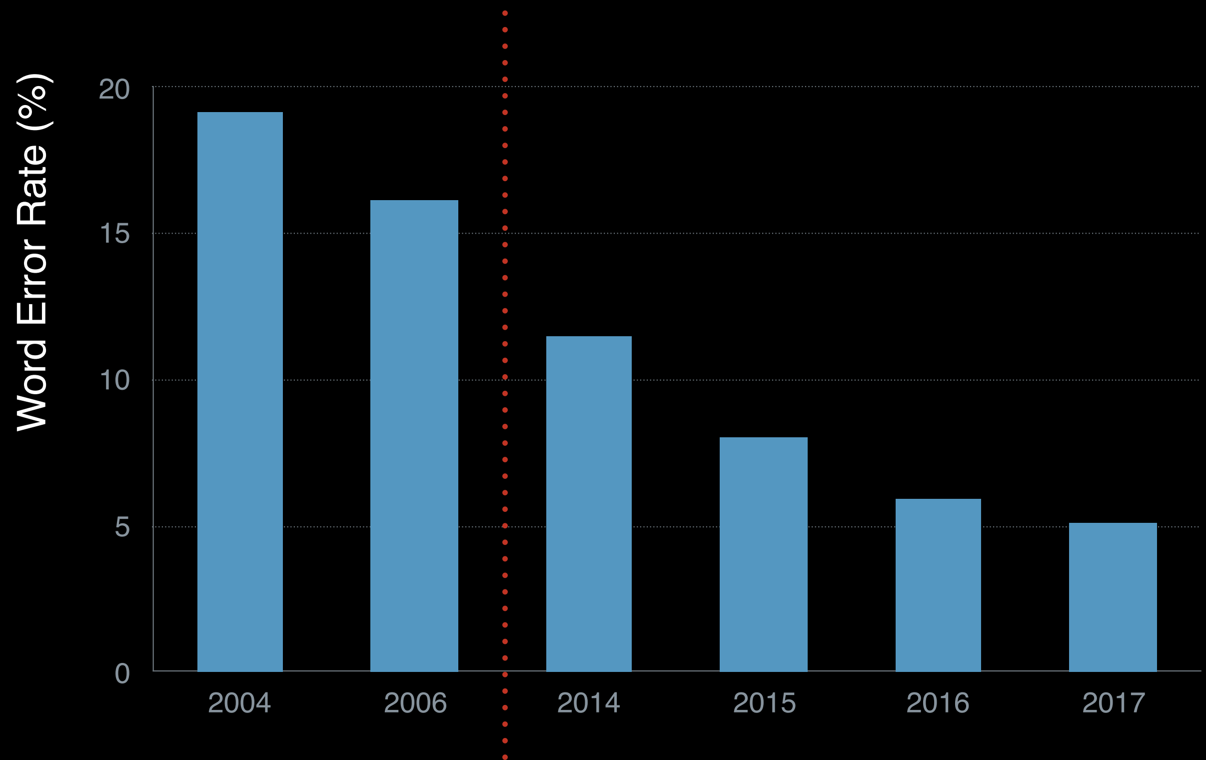
- 800 input features
- 5 layer network
- 1000 neurons per layer
- 8000 output labels
- 12 Million weights

Training

- 300 hours of speech with transcripts
- 1 week training time on a GPU

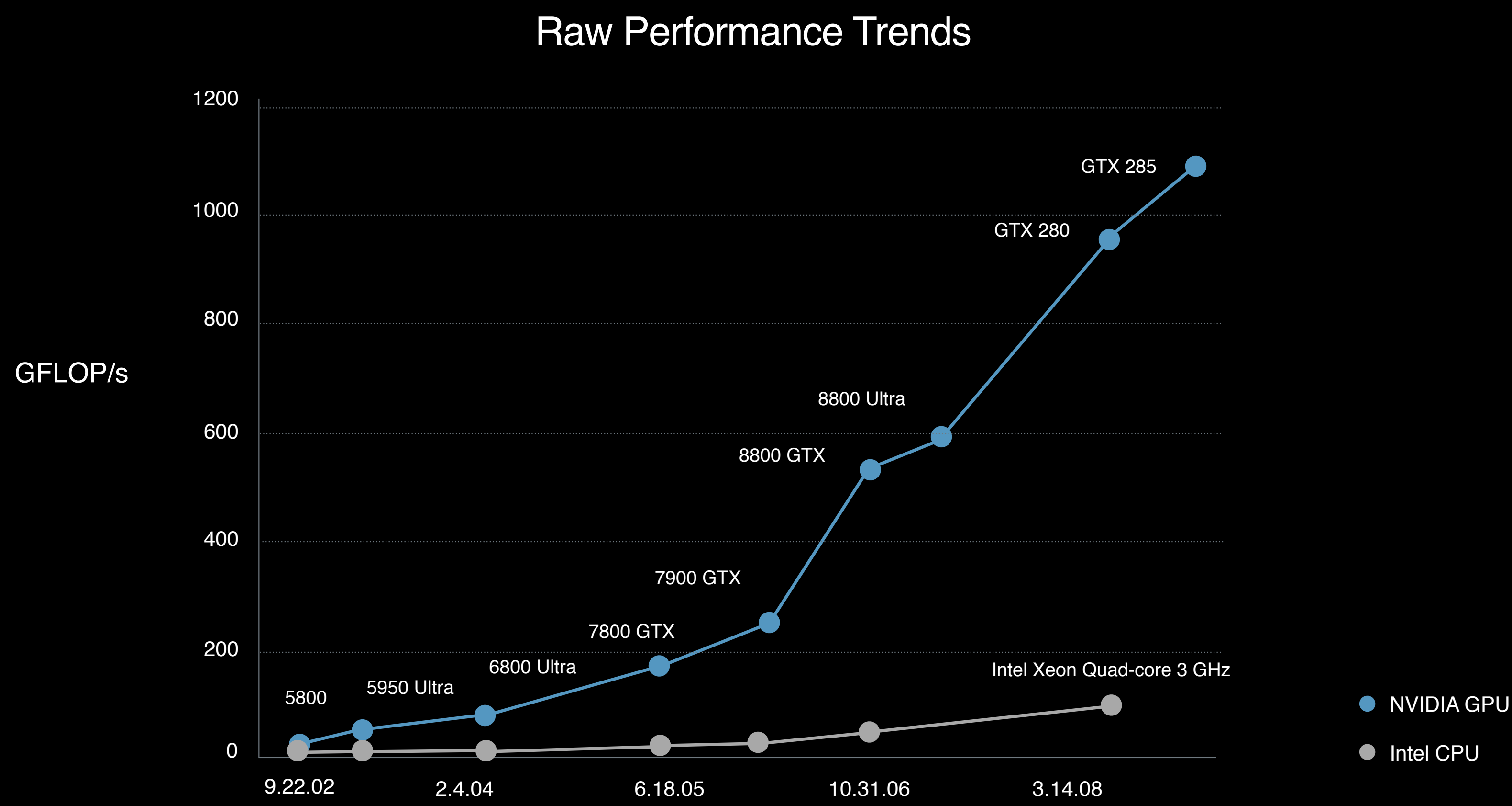


Switchboard



Why Now?

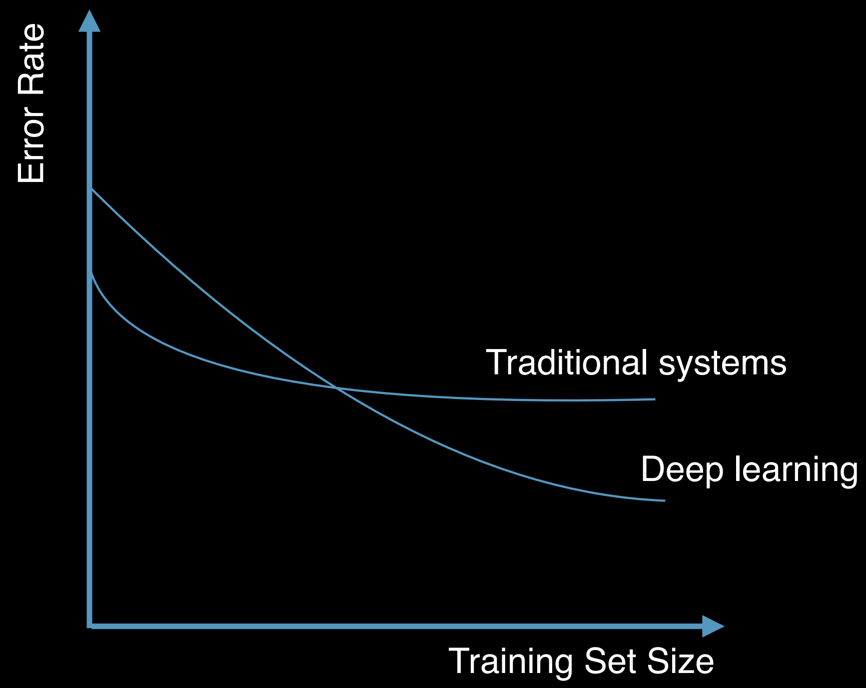
GPUs



Graph is courtesy of NVIDIA, 7/26/2009

Why Now?

Large Amounts of Data



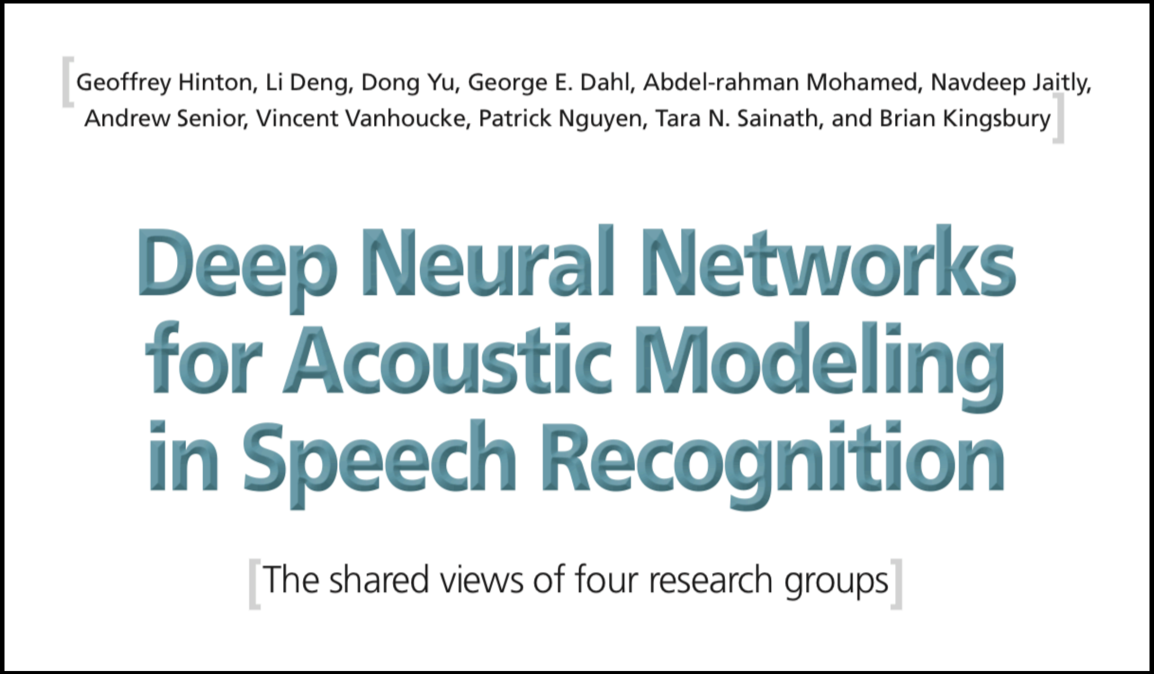
Why Now?

Algorithms

- Direct modeling of context-dependent (tied triphone states) through the DNN
- ~~Unsupervised Pre-training~~
- Deeper networks

Why Now?
Open sharing

U. Toronto
Microsoft
Google
IBM


[Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury]

Deep Neural Networks for Acoustic Modeling in Speech Recognition

[The shared views of four research groups]

Why Now?

Tools

theano

torch

mxnet

TensorFlow

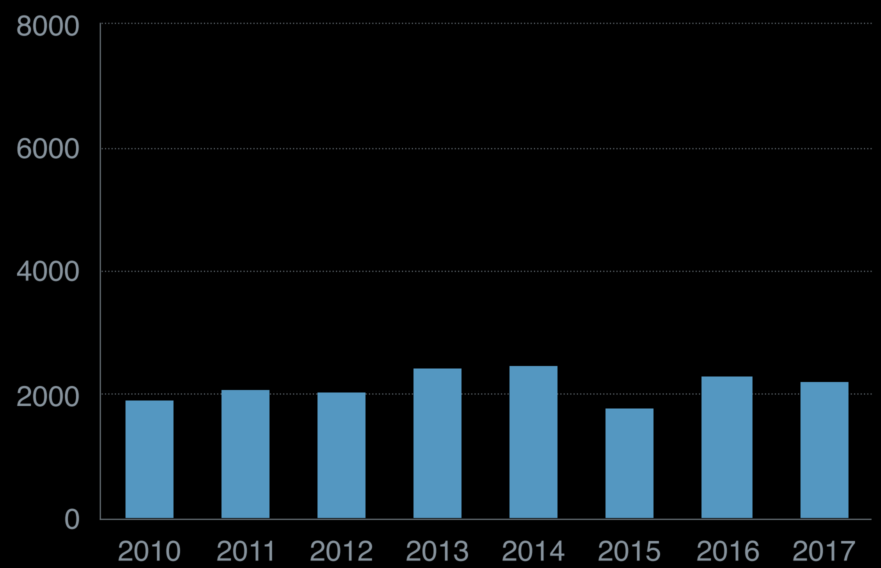
PYTORCH

Caffe2

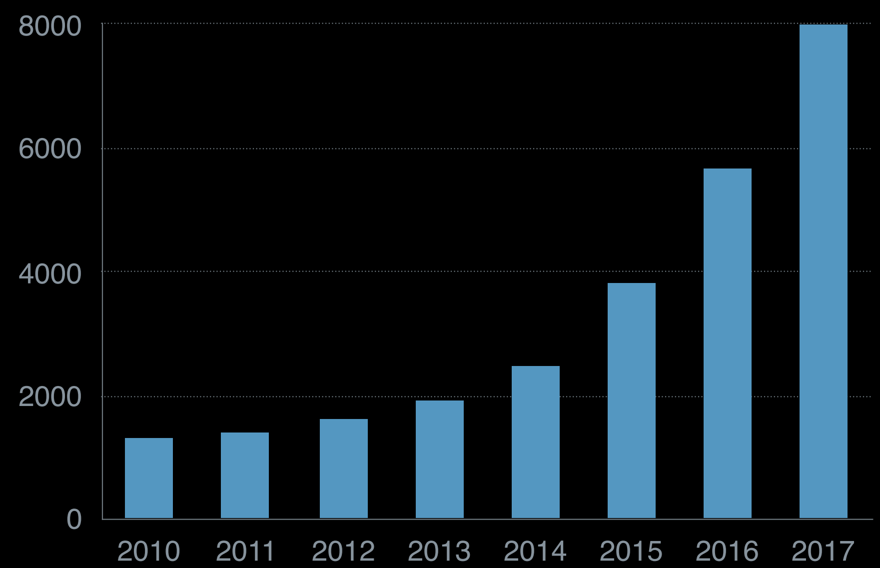
Keras

Microsoft CNTK

Deep Learning Has Roots in Signal Processing



ICASSP Attendees



NIPS Attendees

Transforming Our Digital Lives

ML Becomes Mainstream

On-device scene recognition

Portrait Mode

Language modeling

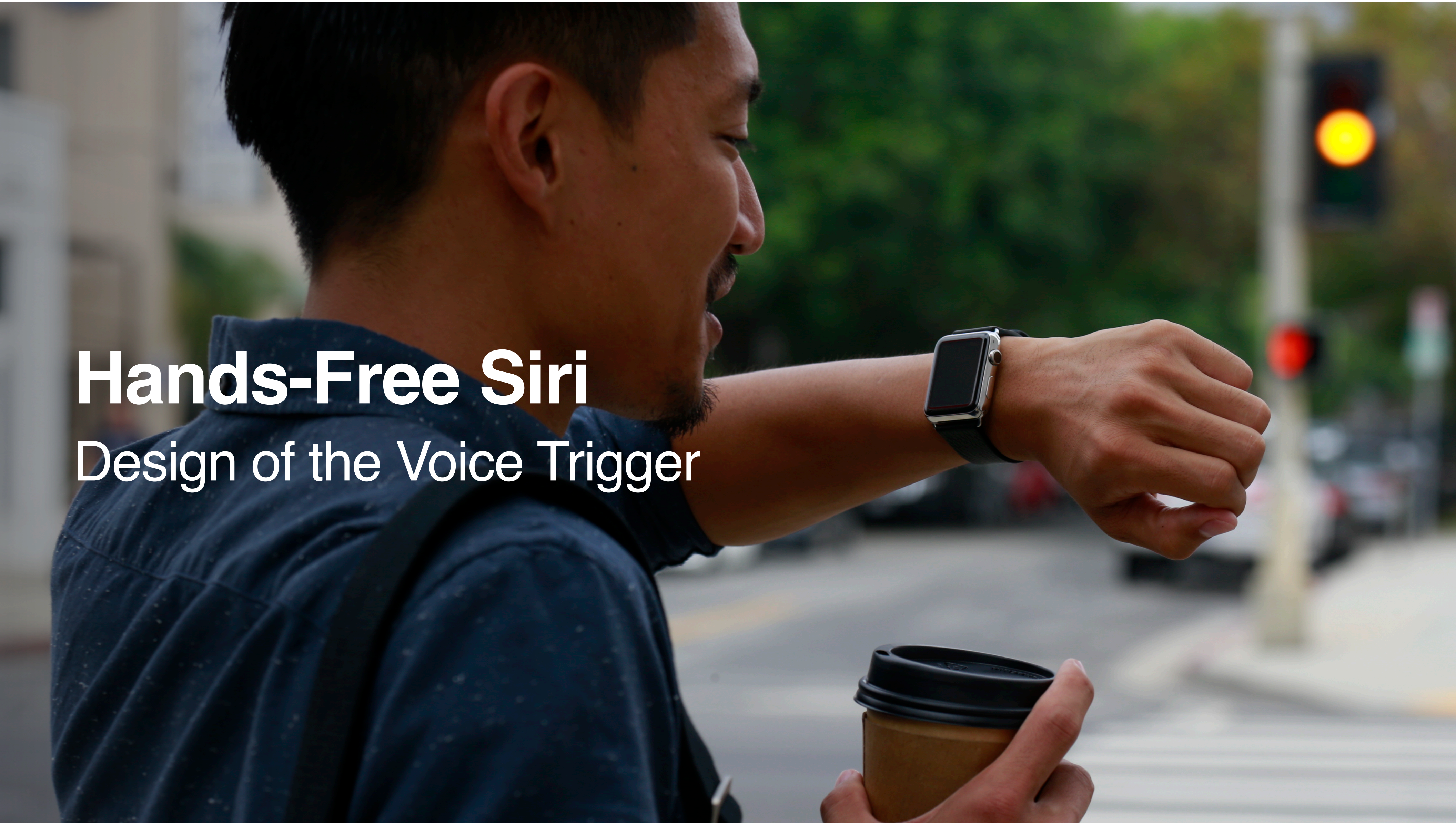
Handwriting recognition

News recommendation

Intelligent assistant

Siri

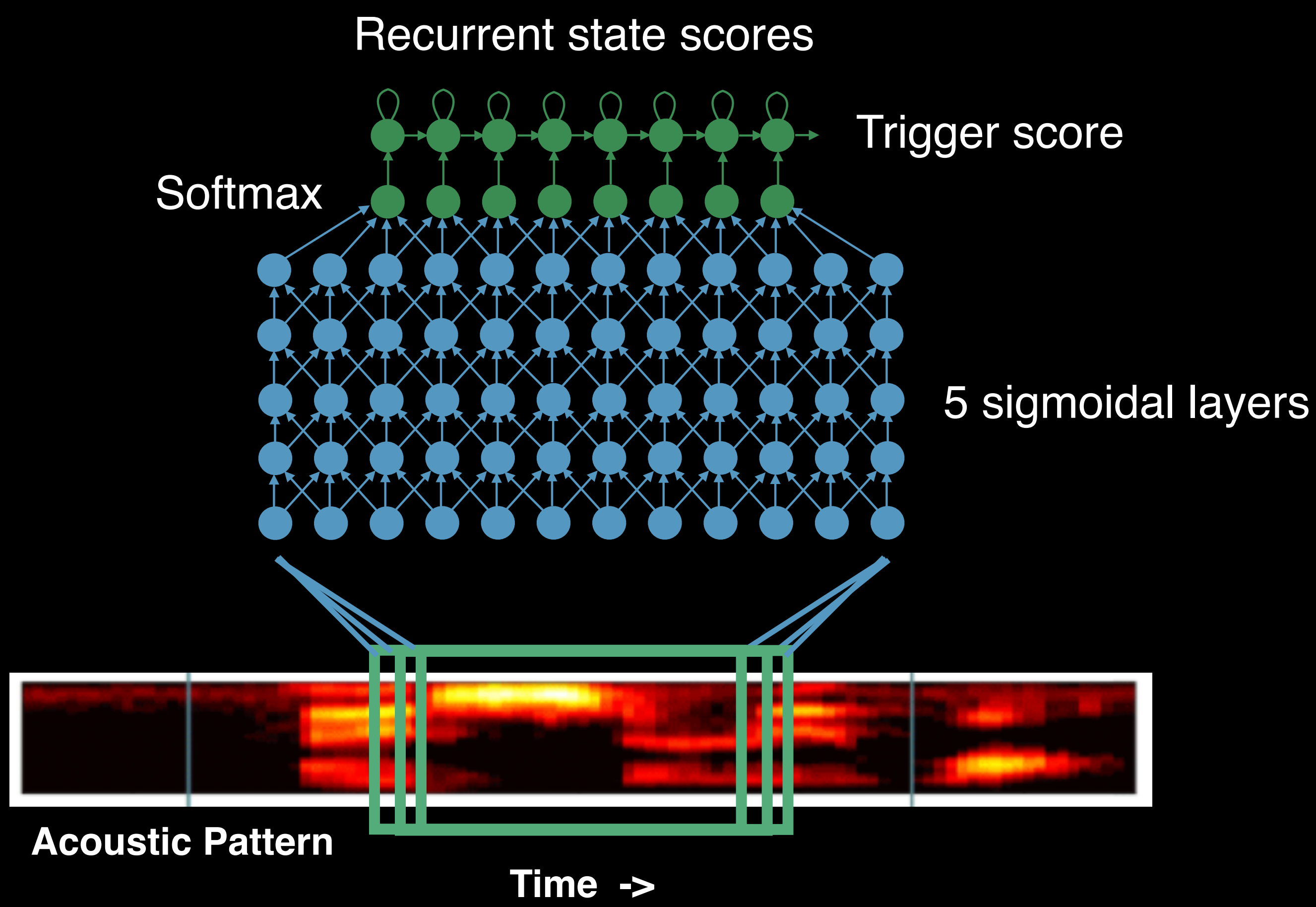
Apple, 2011



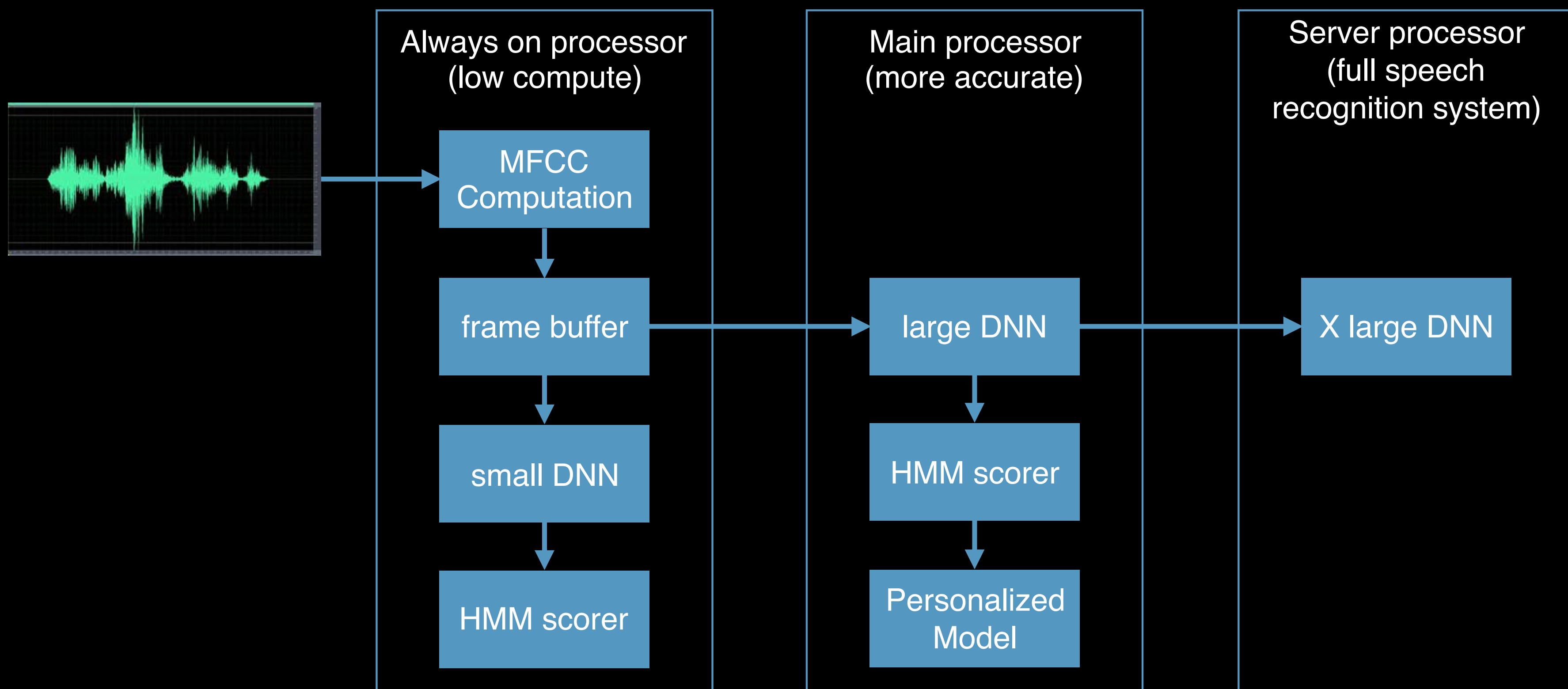
Hands-Free Siri

Design of the Voice Trigger

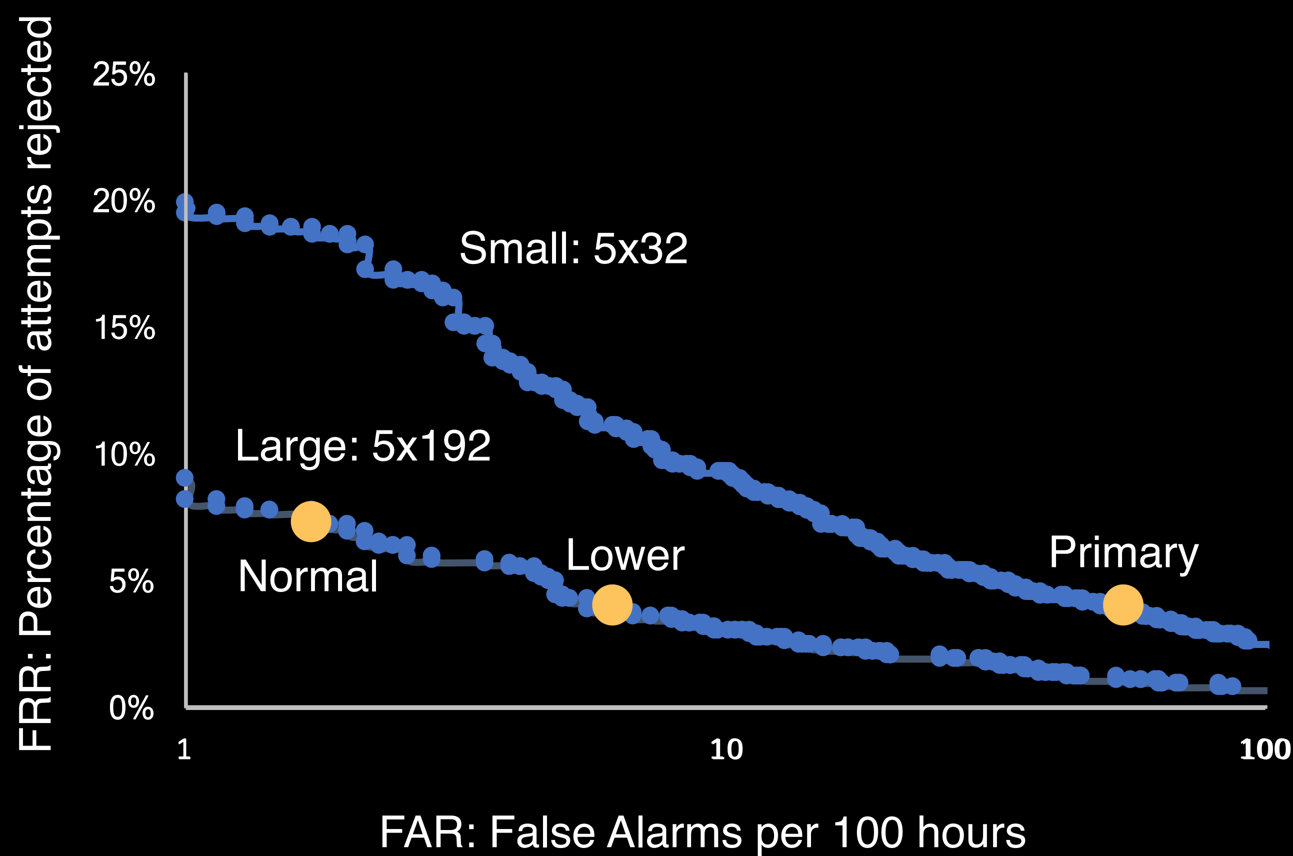
Hey Siri DNN


Recurrent state scores
Trigger score
Softmax
5 sigmoidal layers
Acoustic Pattern
Time ->

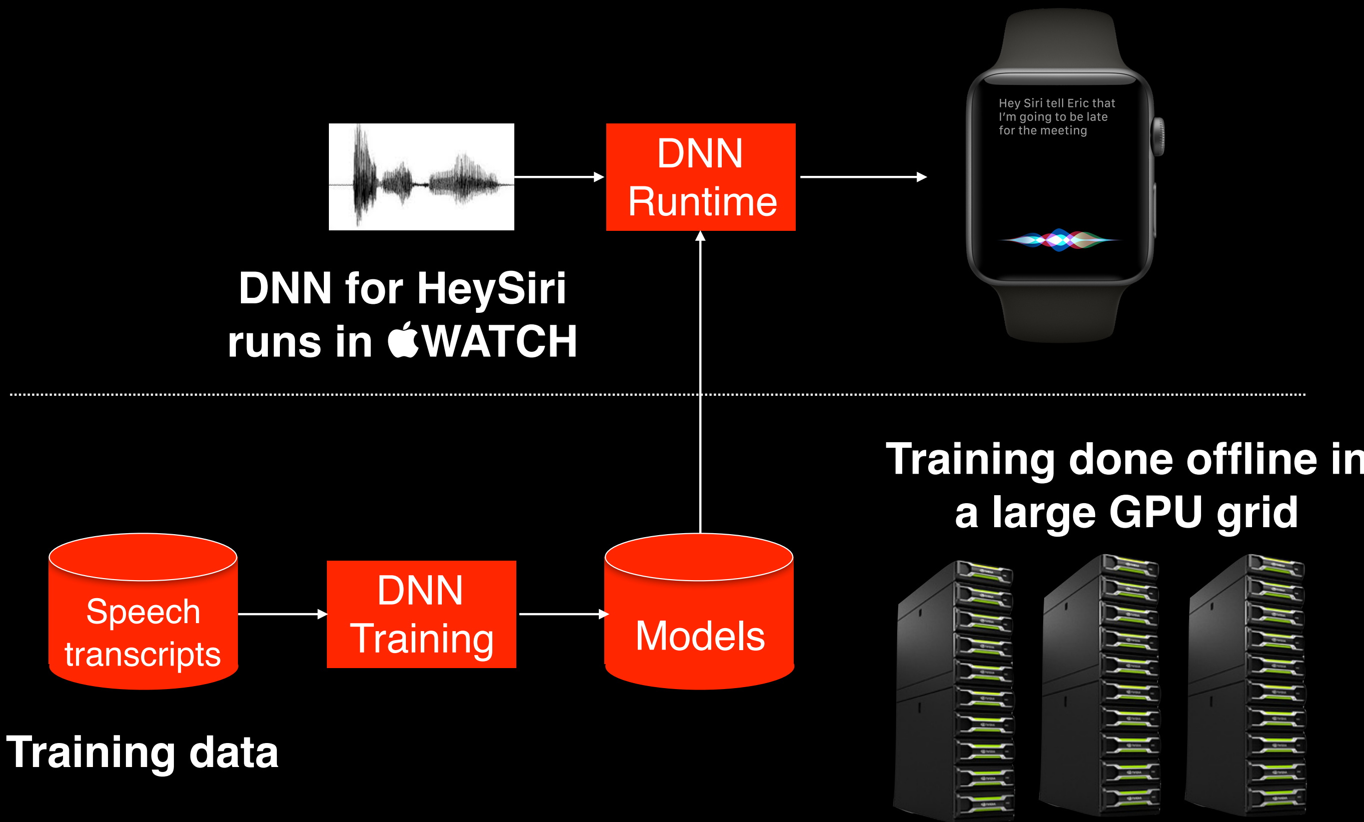
Multi-Pass Detection



Two-Pass Detection

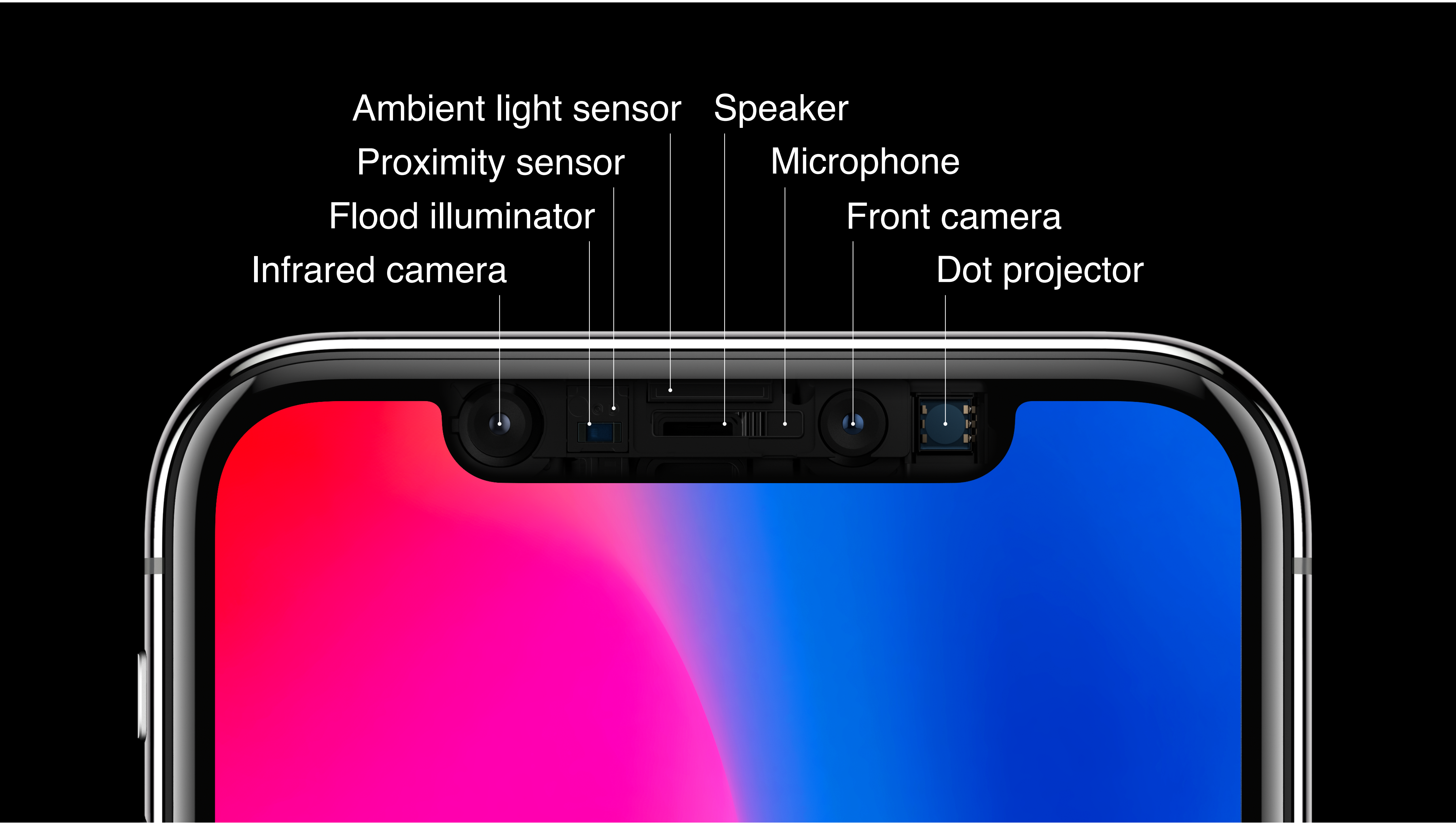


Computing for Deep Learning



Face ID

Apple, 2017

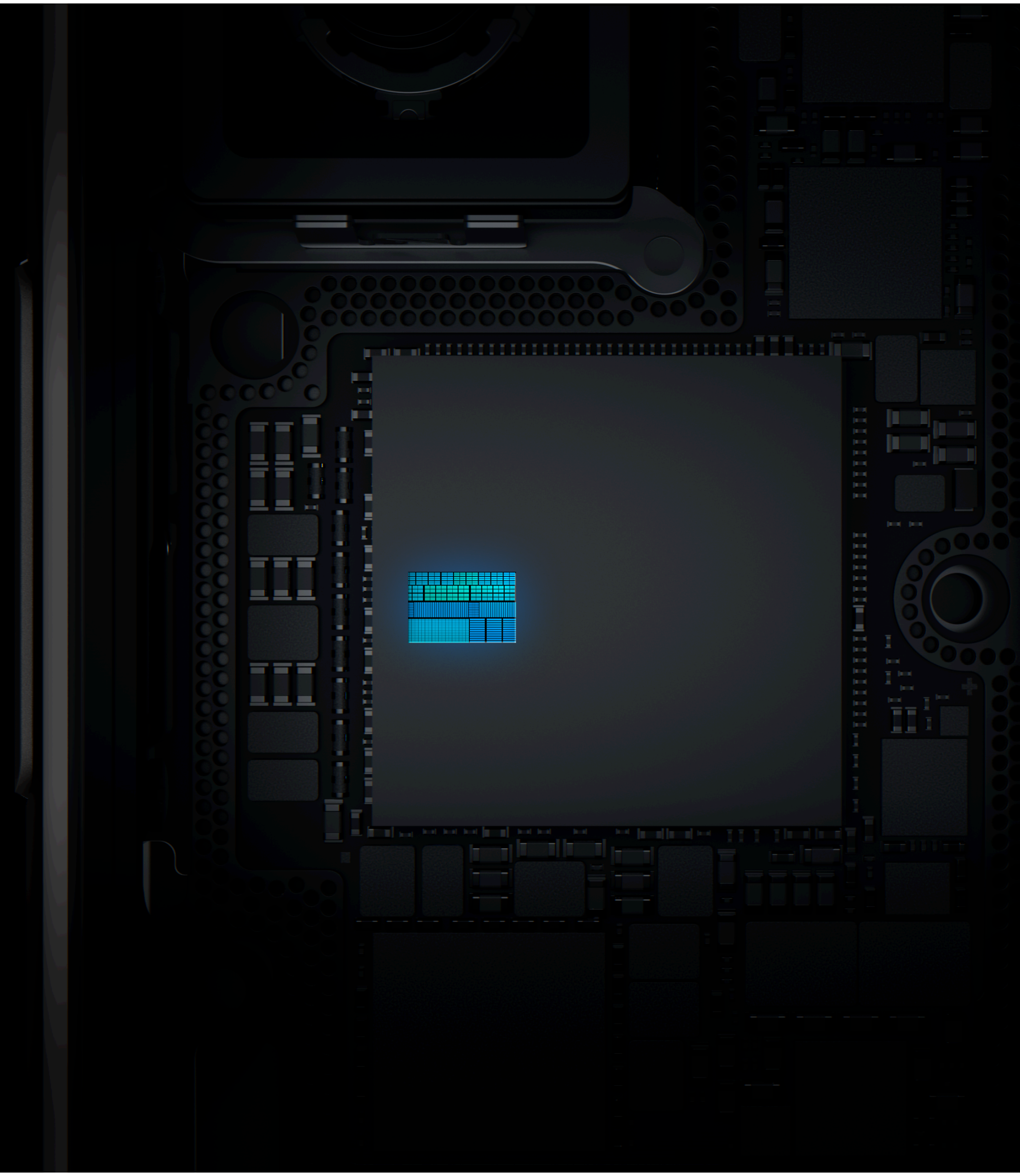

Ambient light sensor
Speaker
Proximity sensor
Microphone
Flood illuminator
Front camera
Infrared camera
Dot projector

Neural engine

Dual-core design

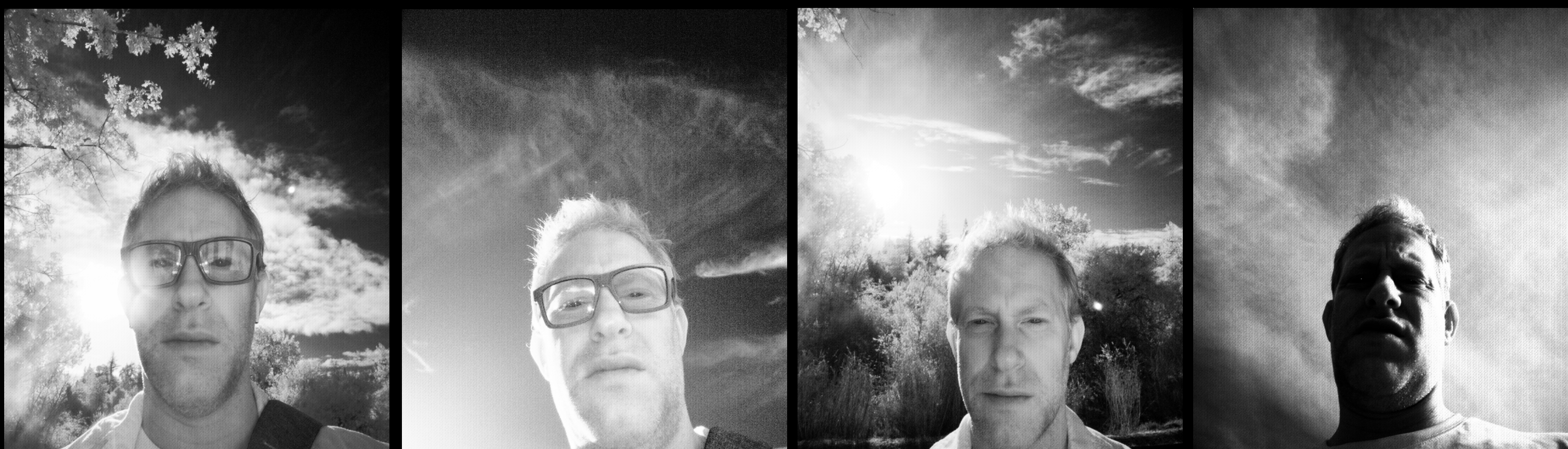
600 billion operations per second

Real-time processing

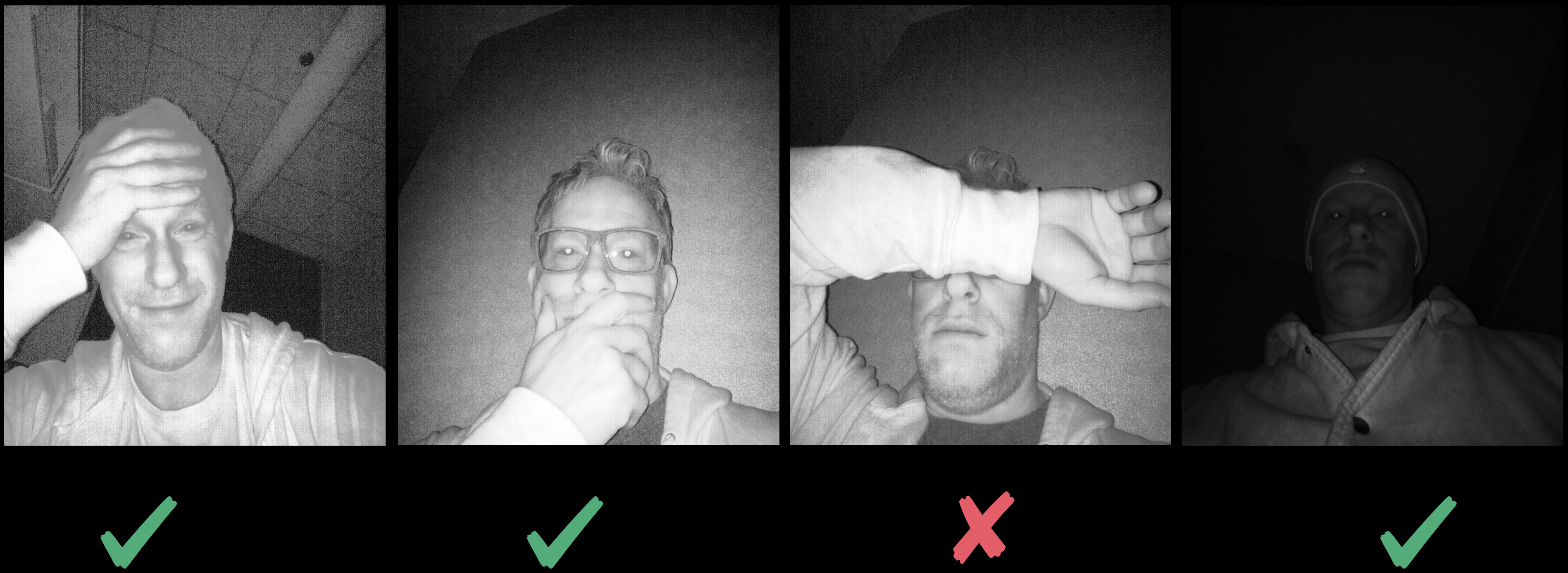


Unconstrained Face Matching

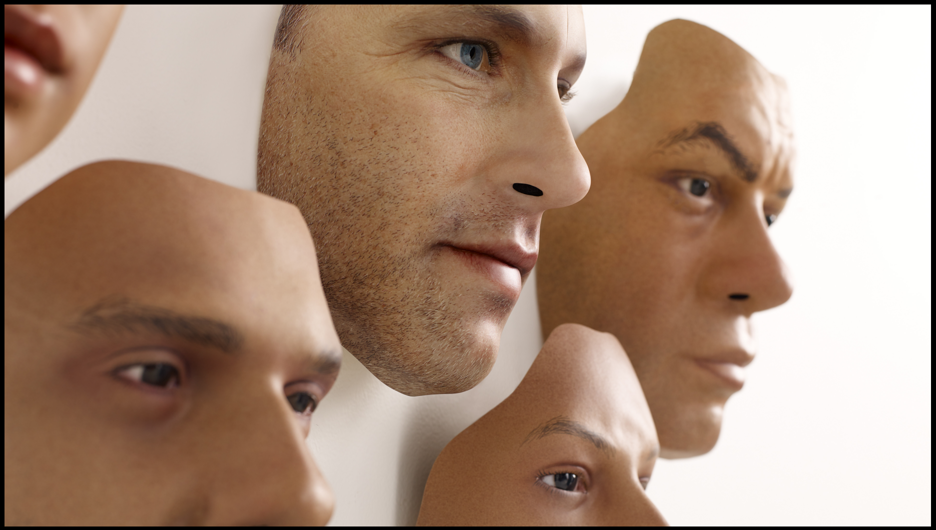
Works in Bright Sunlight and Shadows



Robust to Occlusions

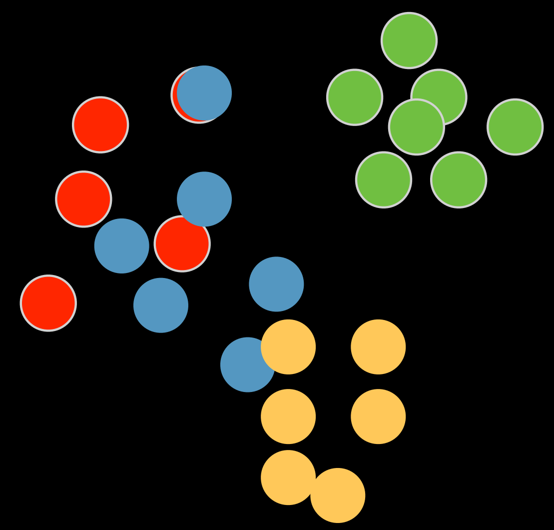


Anti-Spoofing



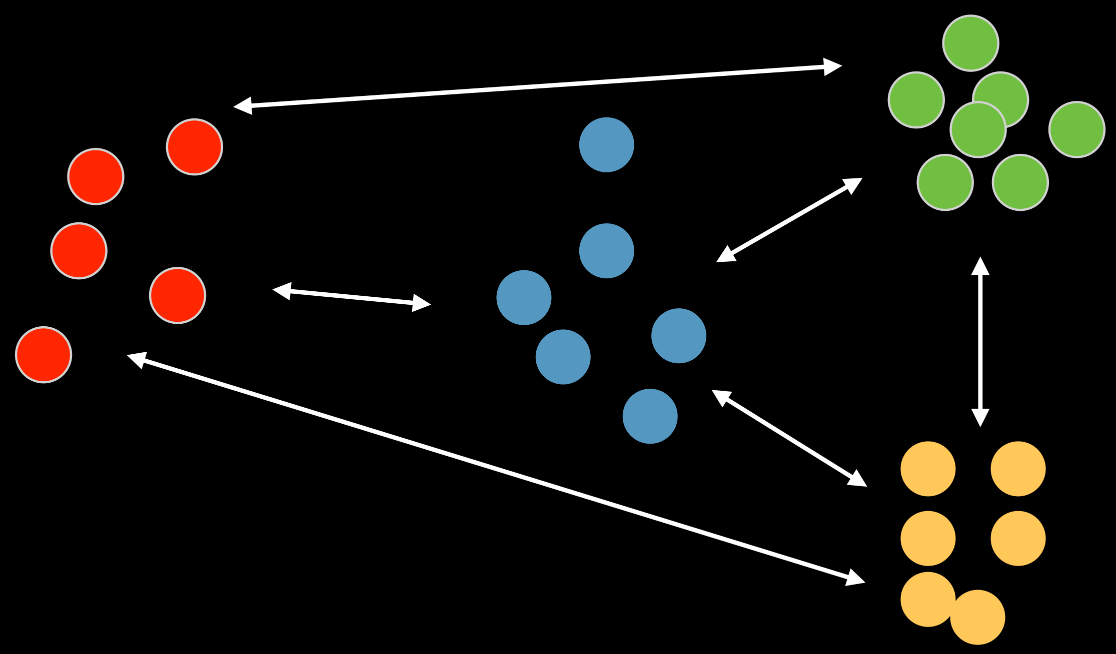
Face ID is a Machine Learning Problem

Goal is to pull same identity pairs together and push different identity pairs apart



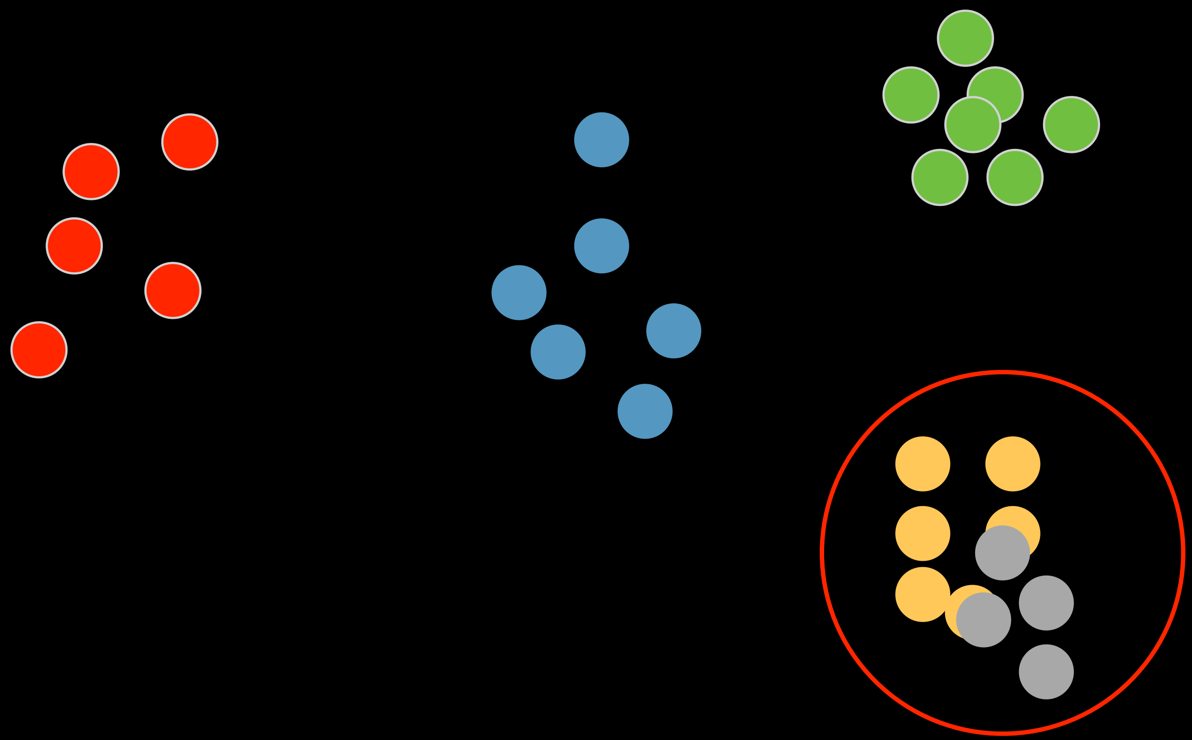
Face ID is a Machine Learning Problem

Goal is to pull same identity pairs together and push different identity pairs apart



Face ID is a Machine Learning Problem

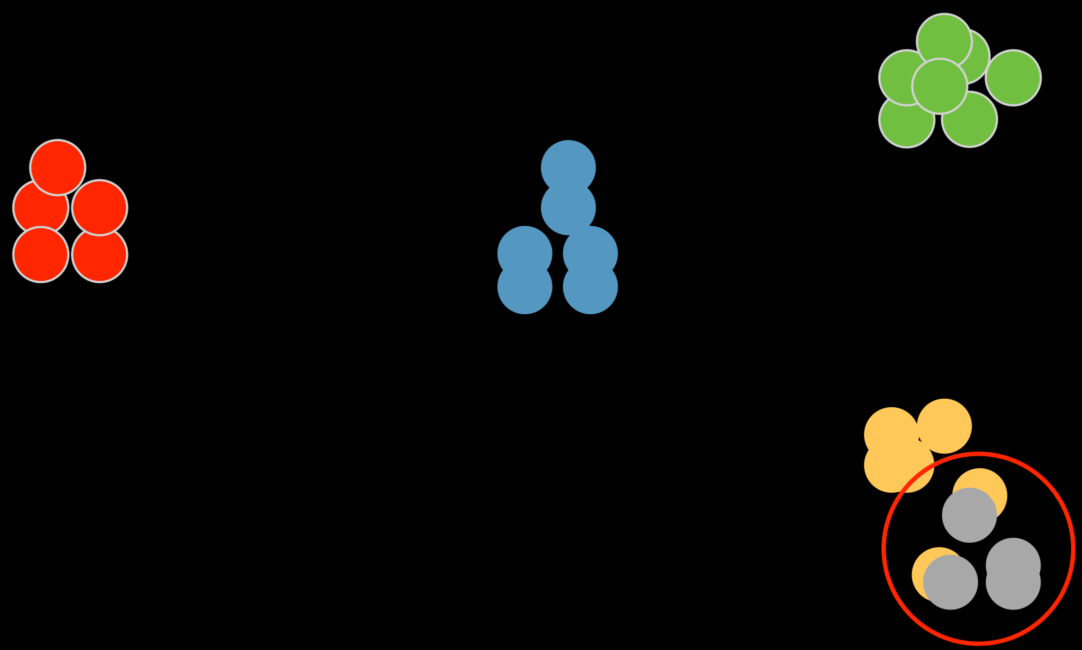
Most faces are not similar—
needles in a haystack



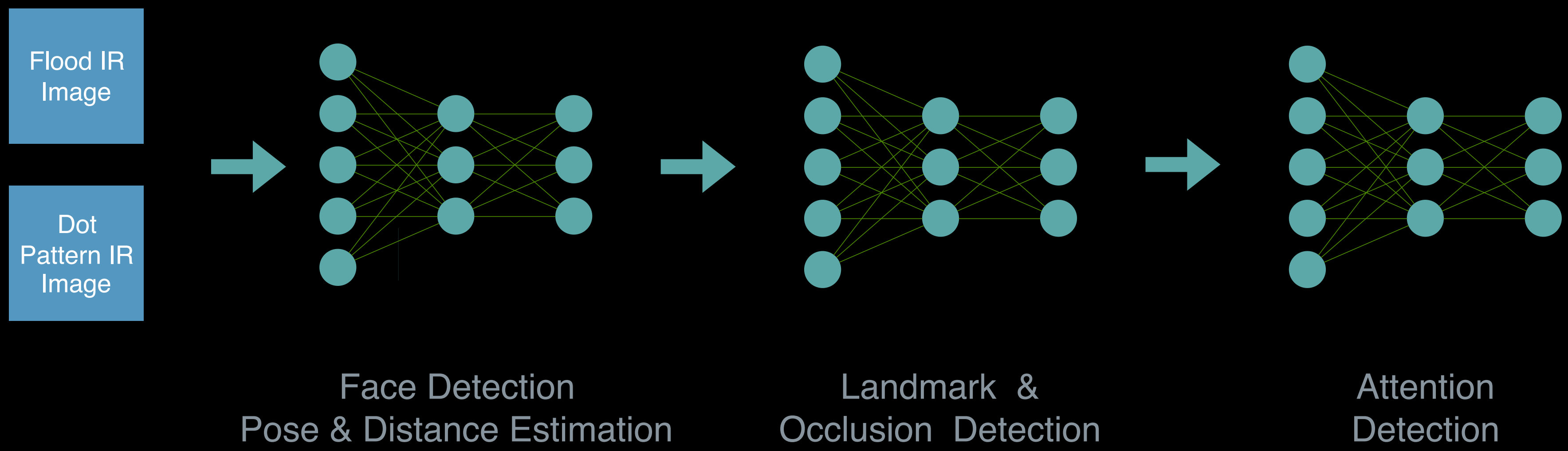
Face ID is a Machine Learning Problem

Sometimes it is very hard to find patterns that separate people that are not spurious

A model that is better at all the easy cases is not necessarily better at solving the hard cases



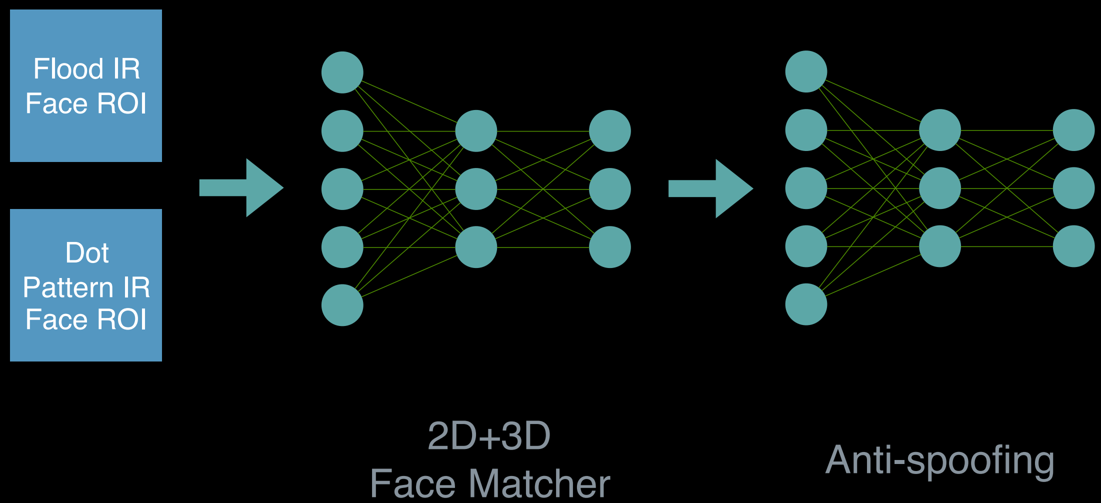
Neural Network Face Matching Pipeline: Detection



Makes decision at any point (no face, out of spec, inattention)

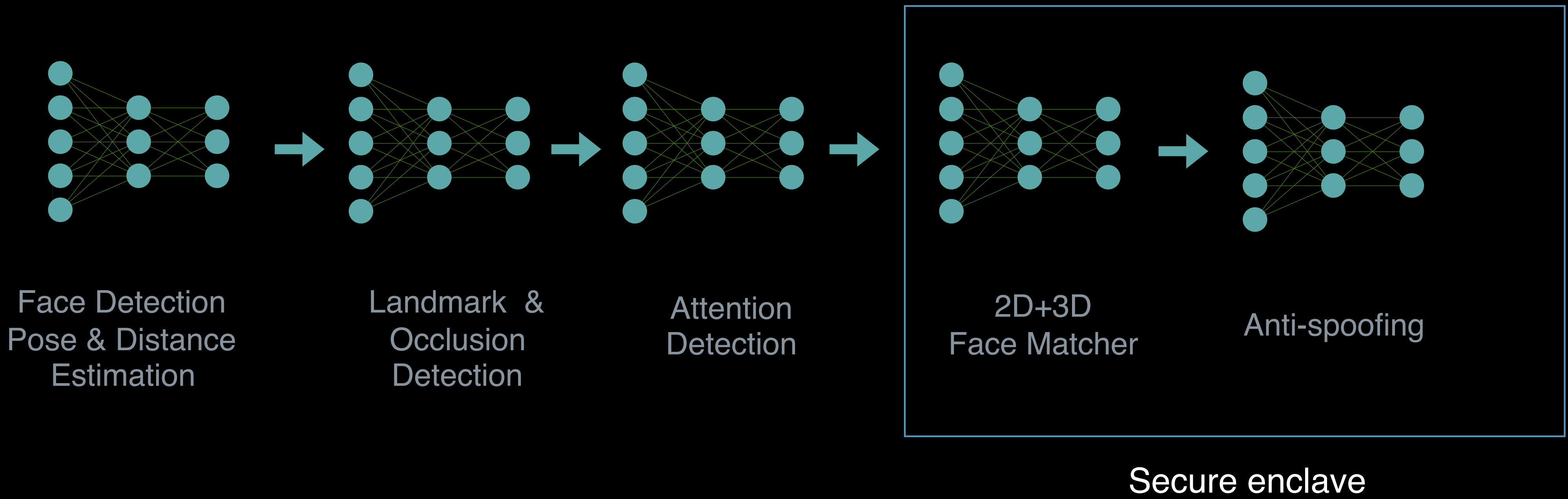
Localizes faces for matching

Neural Network Face Matching Pipeline: Verification

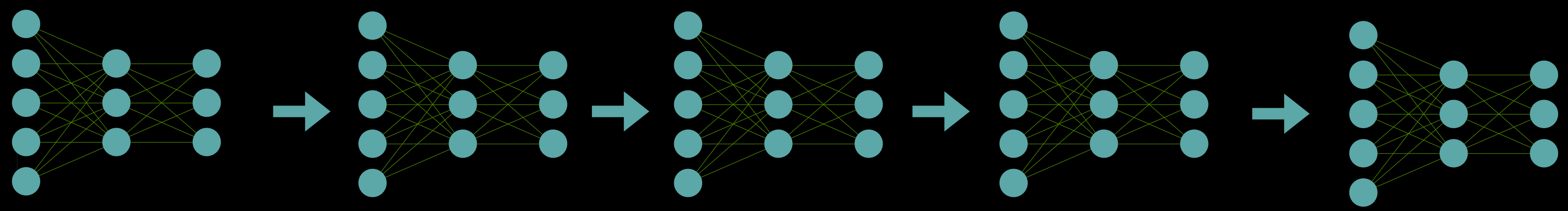


Multimodal learning problem (how to fuse 2D and 3D representations)

Neural Network Face Matching Pipeline: End-To-End



Neural Network Face Matching Pipeline: End-To-End



Has to be really fast

Small memory footprint

Limited power impact

Optimized for full system performance

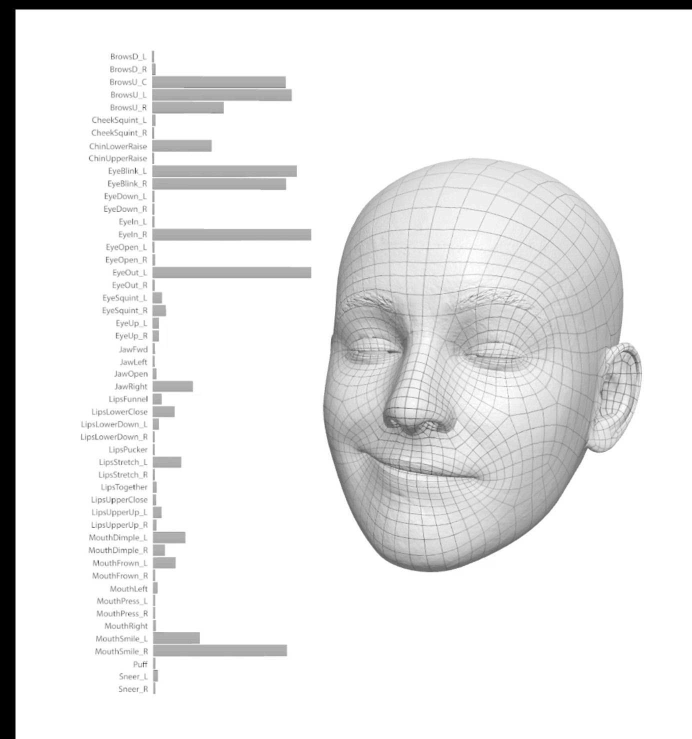
Animoji

Apple, 2017

Realtime Facial Animation

Blendshape Model

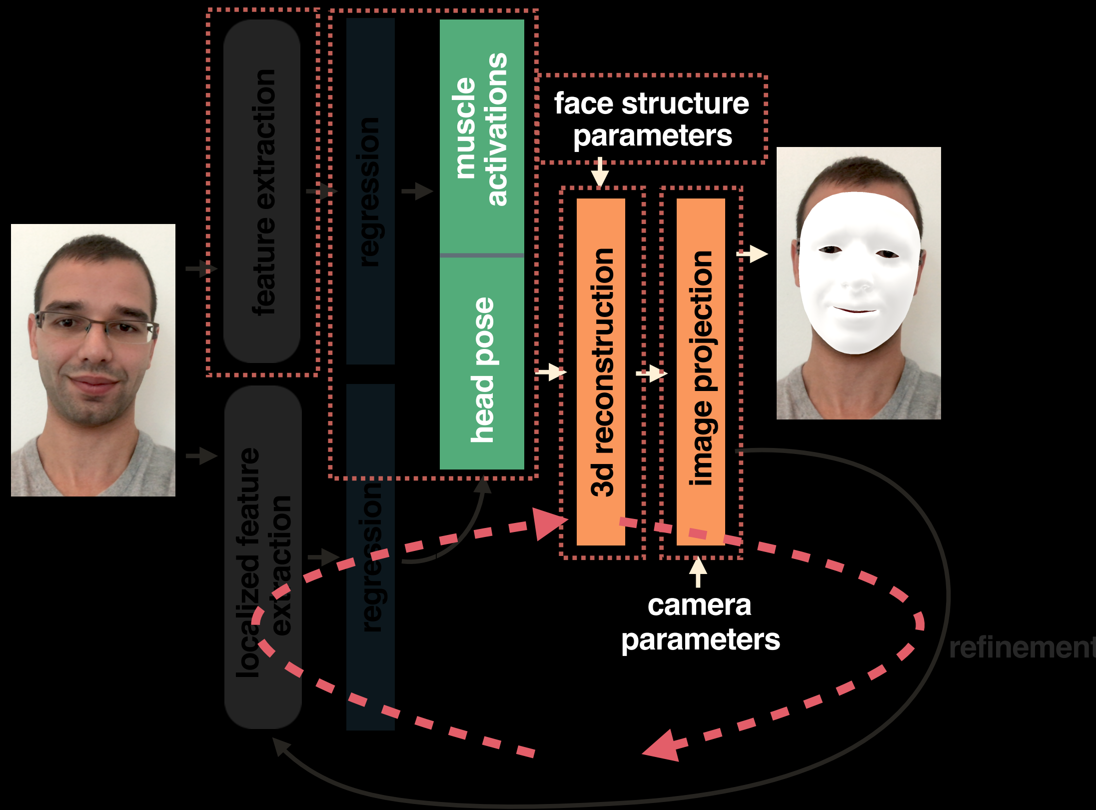
51 blendshapes ("muscles") driving more than 100 shapes



Animojis Driven by Blendshape Model

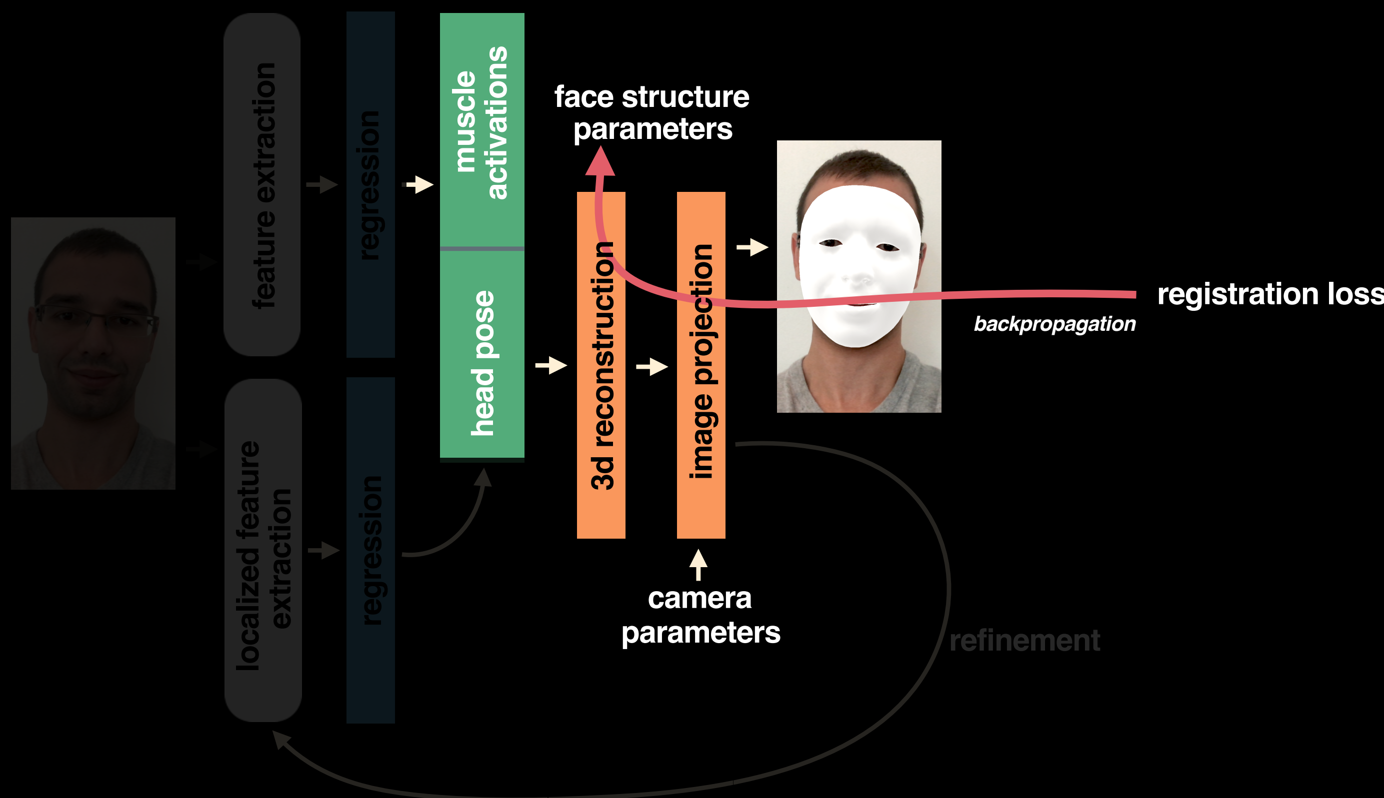
Realtime Facial Animation

Model-based RNN



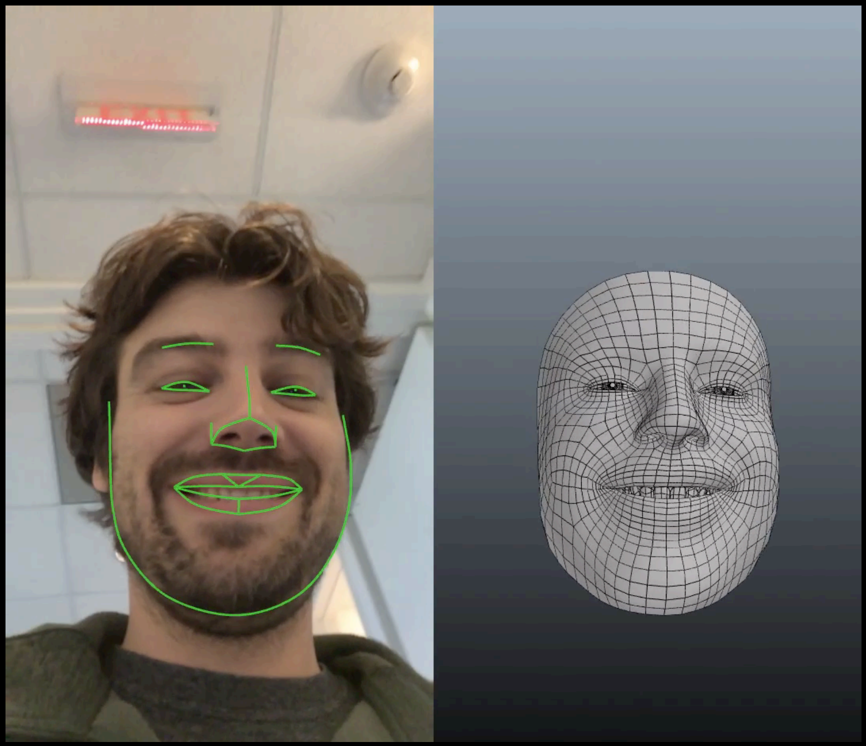
Online Identity Adaptation

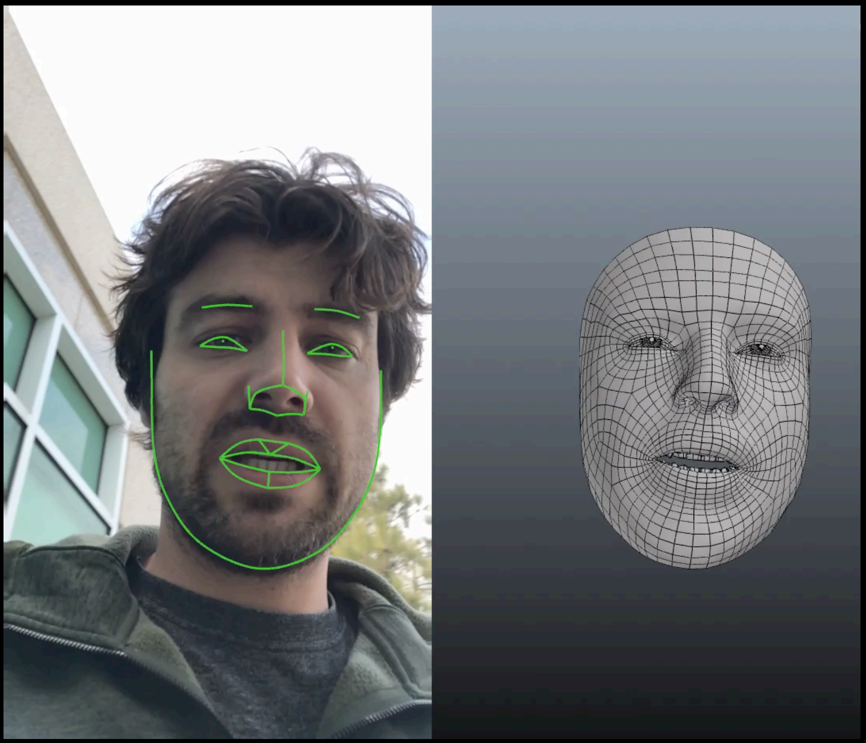
Geometric backpropagation



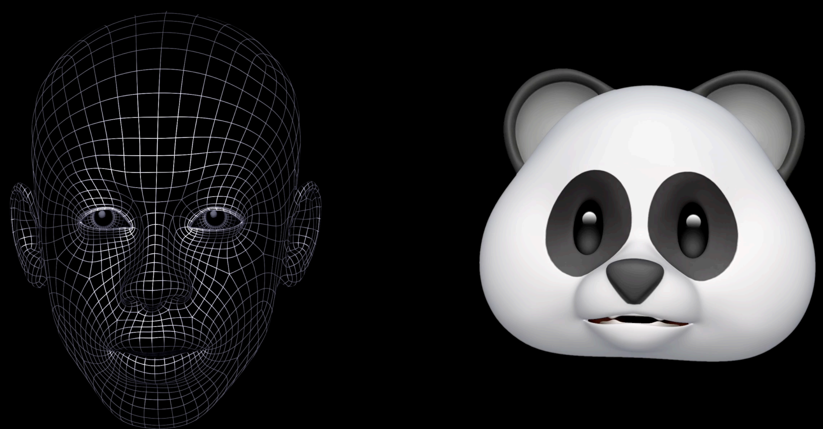
Results

Indoor and outdoor





Animoji
Performance



The animation runs sustainably at 60fps

And of course…

Animoji Karaoke