The Future of the Future:

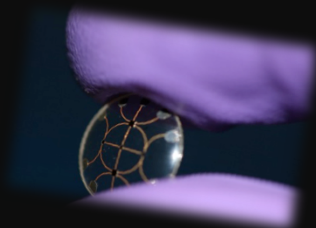
reasoning on the realities and the hypes of technology

Alberto Sangiovanni-Vincentelli,
The Edgar L. and Harold H. Buttner Chair
Department of EECS,
University of California, Berkeley

Co-founder, Member of the Board, Chief Technology Advisor
Cadence Design Systems





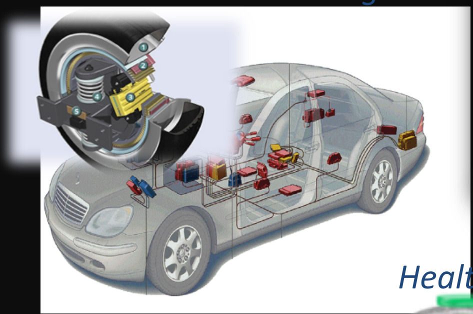

Outline

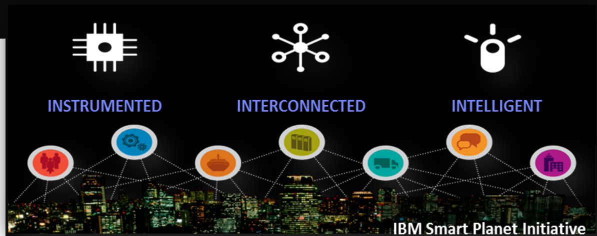
- Setting the stage
- Artificial Intelligence
  - Economics
  - Neural networks and learning
  - Hype vs. reality

The Emerging Technology Scene: IoT, Mobile Devices, Cloud

Cyber-Physical Systems (CPS) Interconnect the World Around Us and Make It "Smarter"

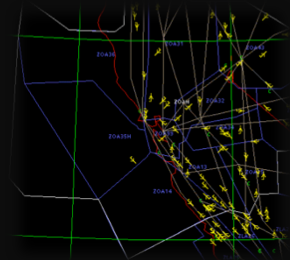
Autonomous Driving





Avionics

Transportation (Air traffic control)



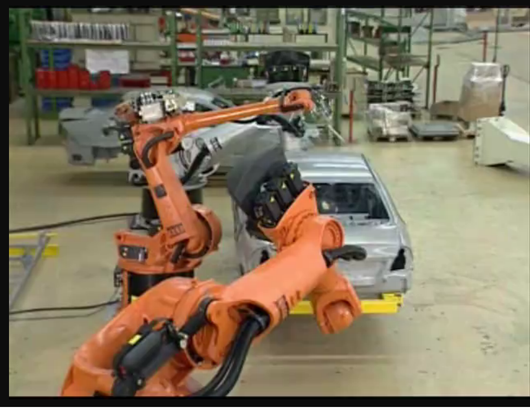
Telecommunications

Health care

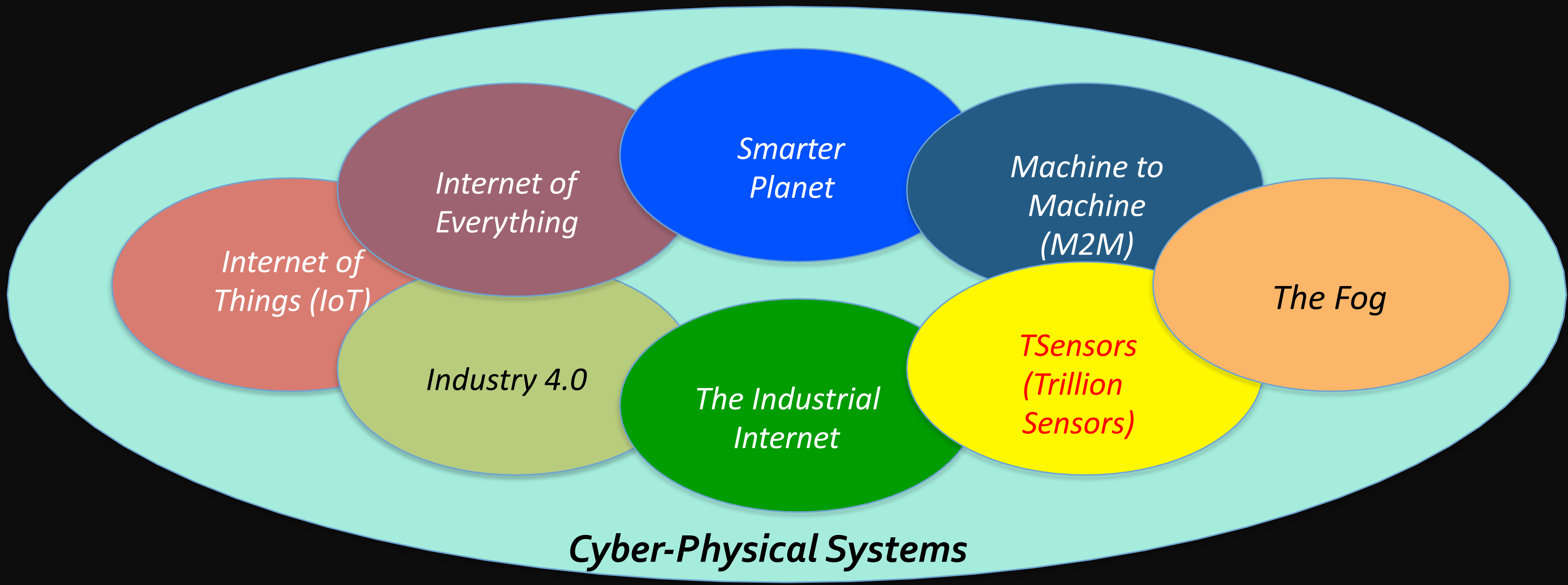
Buildings

Power generation and distribution

Factory automation



How Buzzwords Relate to CPS



Economic Potential

The Internet of Things	**300%** Increase in connected machine-to-machine devices over past 5 years **80–90%** Price decline in MEMS (microelectromechanical systems) sensors in past 5 years	**1 trillion** Things that could be connected to the Internet across industries such as manufacturing, health care, and mining **100 million** Global machine to machine (M2M) device connections across sectors like transportation, security, health care, and utilities	**$36 trillion** Operating costs of key affected industries (manufacturing, health care, and mining)
Cloud technology	**18 months** Time to double server performance per dollar **3x** Monthly cost of owning a server vs. renting in the cloud	**2 billion** Global users of cloud-based email services like Gmail, Yahoo, and Hotmail **80%** North American institutions hosting or planning to host critical applications on the cloud	**$1.7 trillion** GDP related to the Internet **$3 trillion** Enterprise IT spend
Advanced robotics	**75–85%** Lower price for Baxter[3] than a typical industrial robot **170%** Growth in sales of industrial robots, 2009–11	**320 million** Manufacturing workers, 12% of global workforce **250 million** Annual major surgeries	**$6 trillion** Manufacturing worker employment costs, 19% of global employment costs **$2–3 trillion** Cost of major surgeries

Computers and mobiles to disappear

7 trillions devices **for** **7 billions** persons

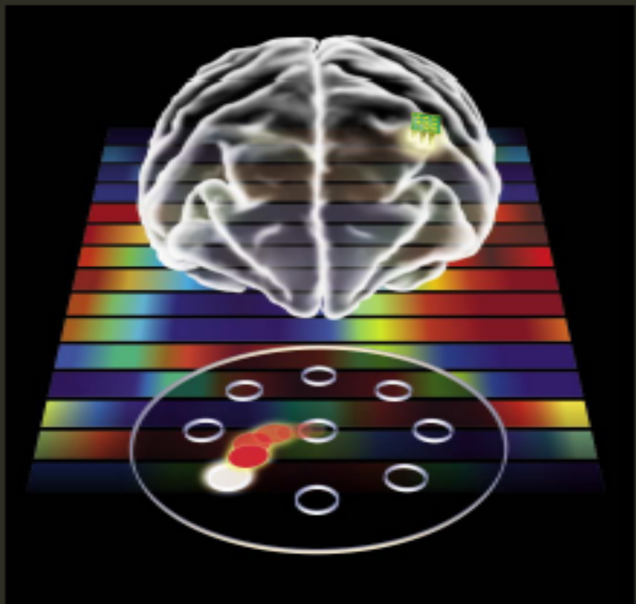
1,000 Devices per person by 2025

The Immersed Human

Real-life interaction between humans and cyberspace, enabled by enriched input and output devices on and in the body and in the surrounding environment

Another One: BioCyber (?) Systems

Linking the Cyber and Biological Worlds



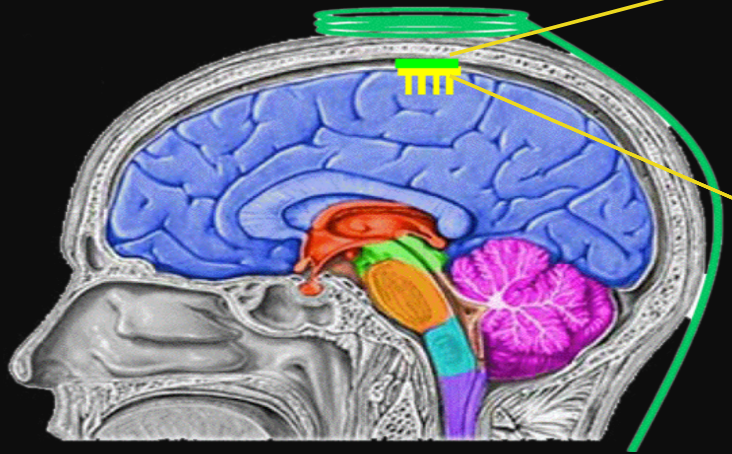


Examples: Brain-machine interfaces and body-area networks

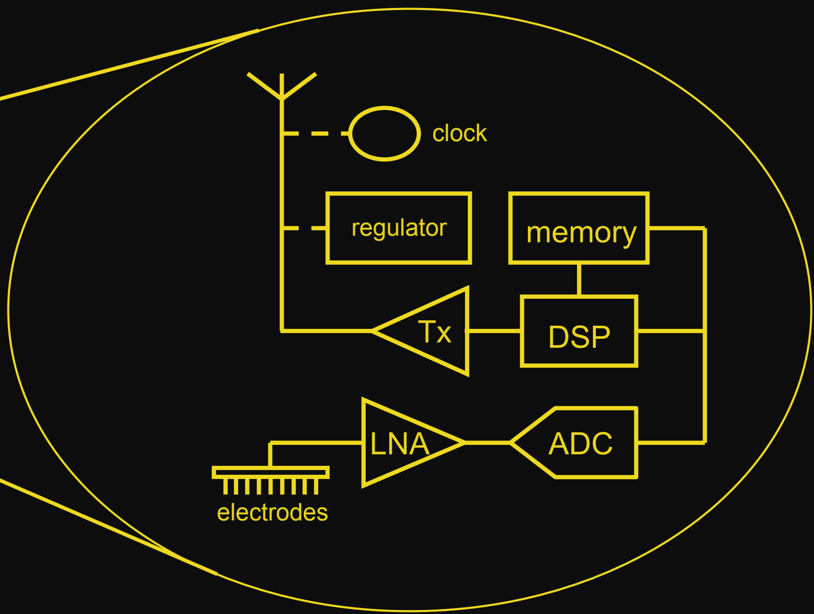
Courtesy: J. Rabaey

Towards Integrated Wireless Implanted Interfaces

Moving the state-of-the-art in wireless sensing

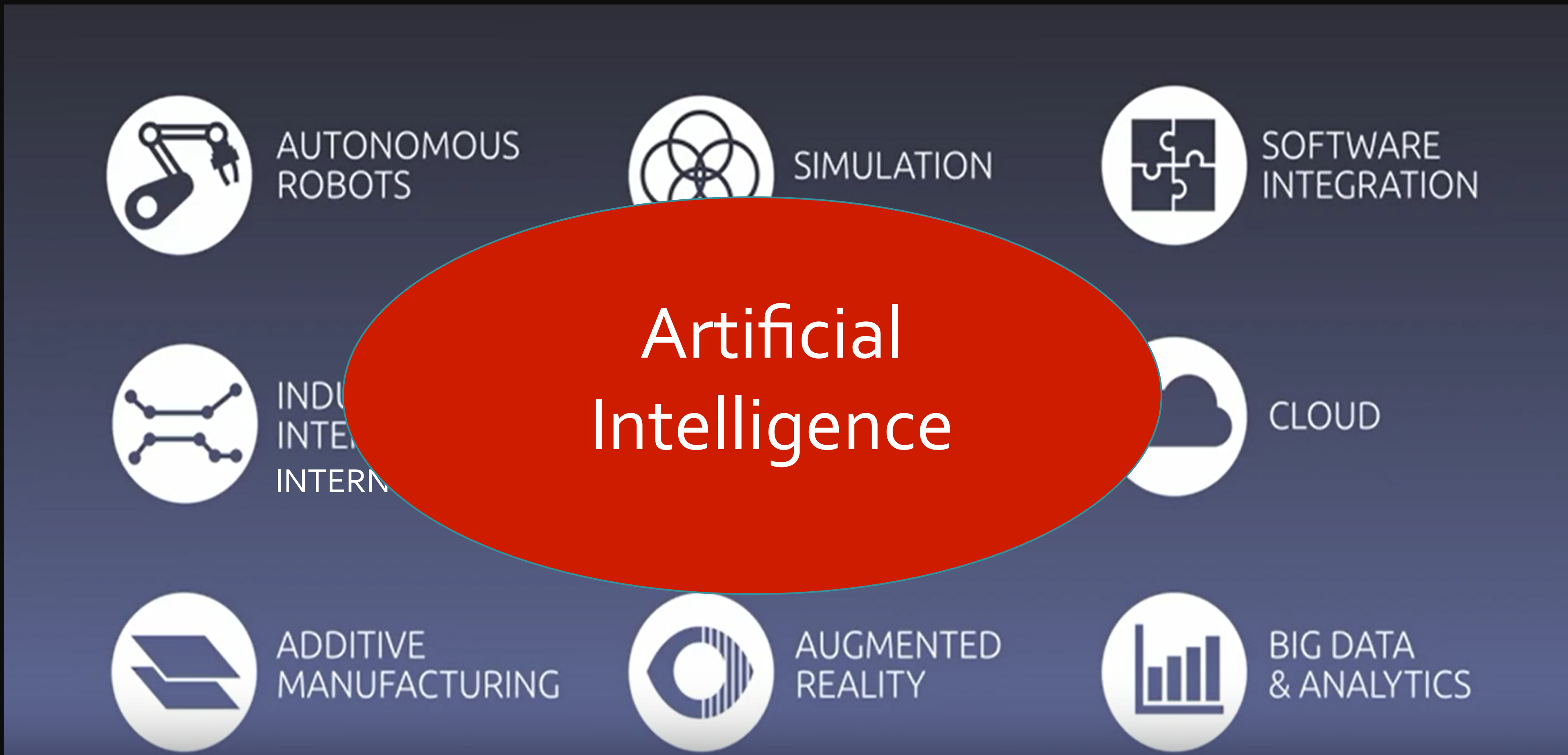


[Illustration art: Subbu Venkatraman]



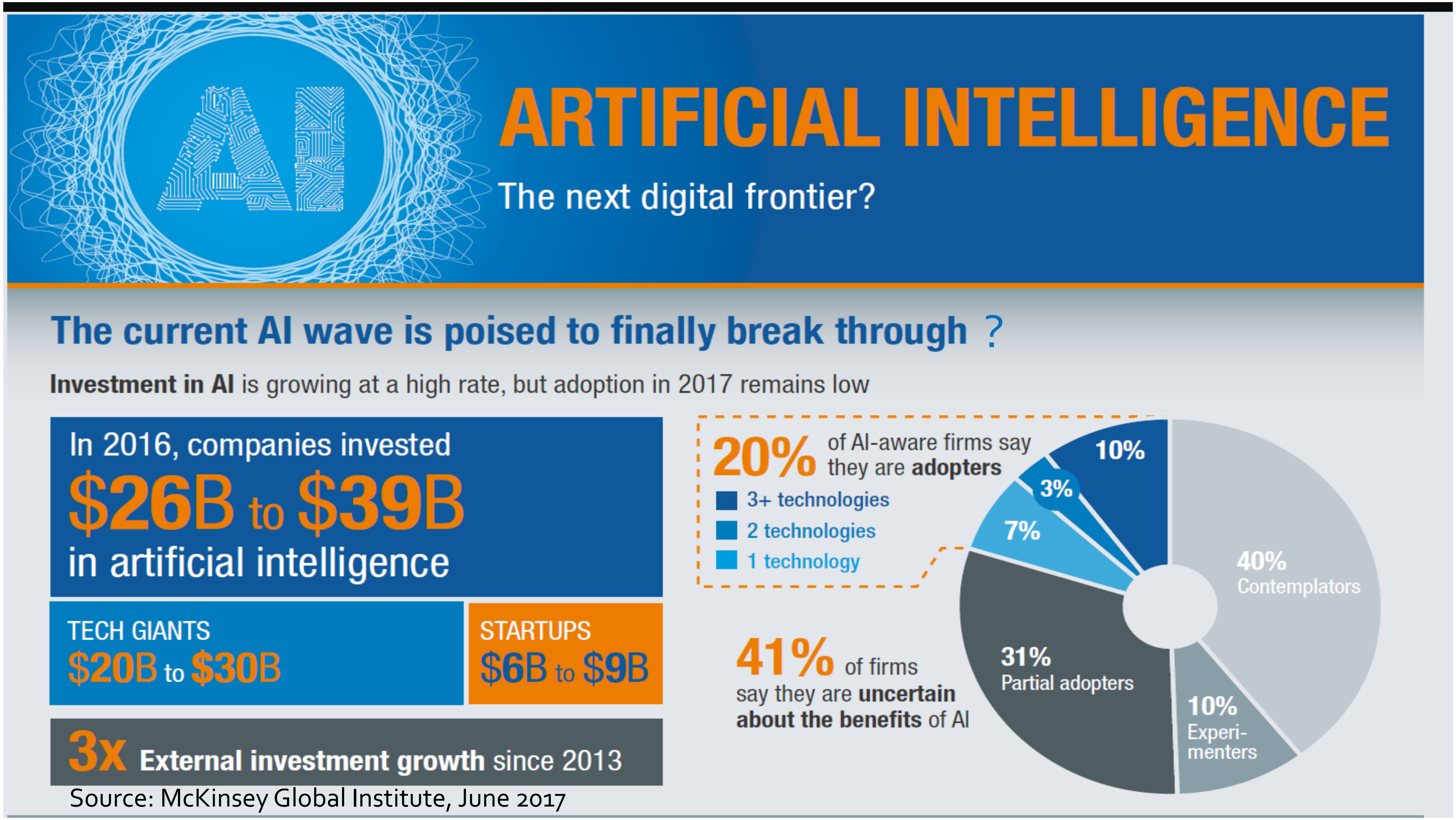
Power budget: mWs to 1 mW

Digitalization: Enabling Technologies



Outline

- Setting the stage
- Artificial Intelligence
  - Economics
  - Neural networks and learning
  - Hype vs. reality



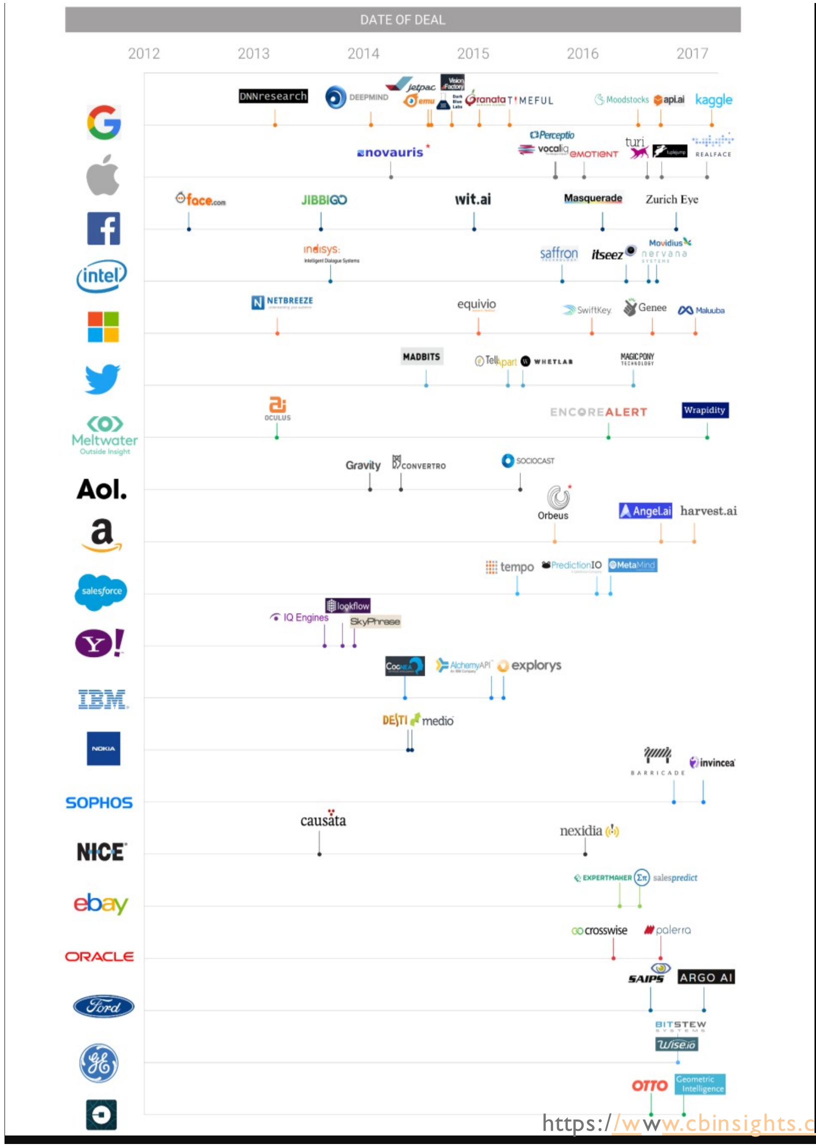
Source: McKinsey Global Institute, June 2017

AI Equity Funding Since 2012:

$14.9 BILLION

across

2250 DEALS



RACE FOR INTELLIGENCE

Google, Facebook, Apple, Intel and other big corporations acquiring AI startups

200+
Acquisitions since 2012

30+
M&A deals in Q1'17

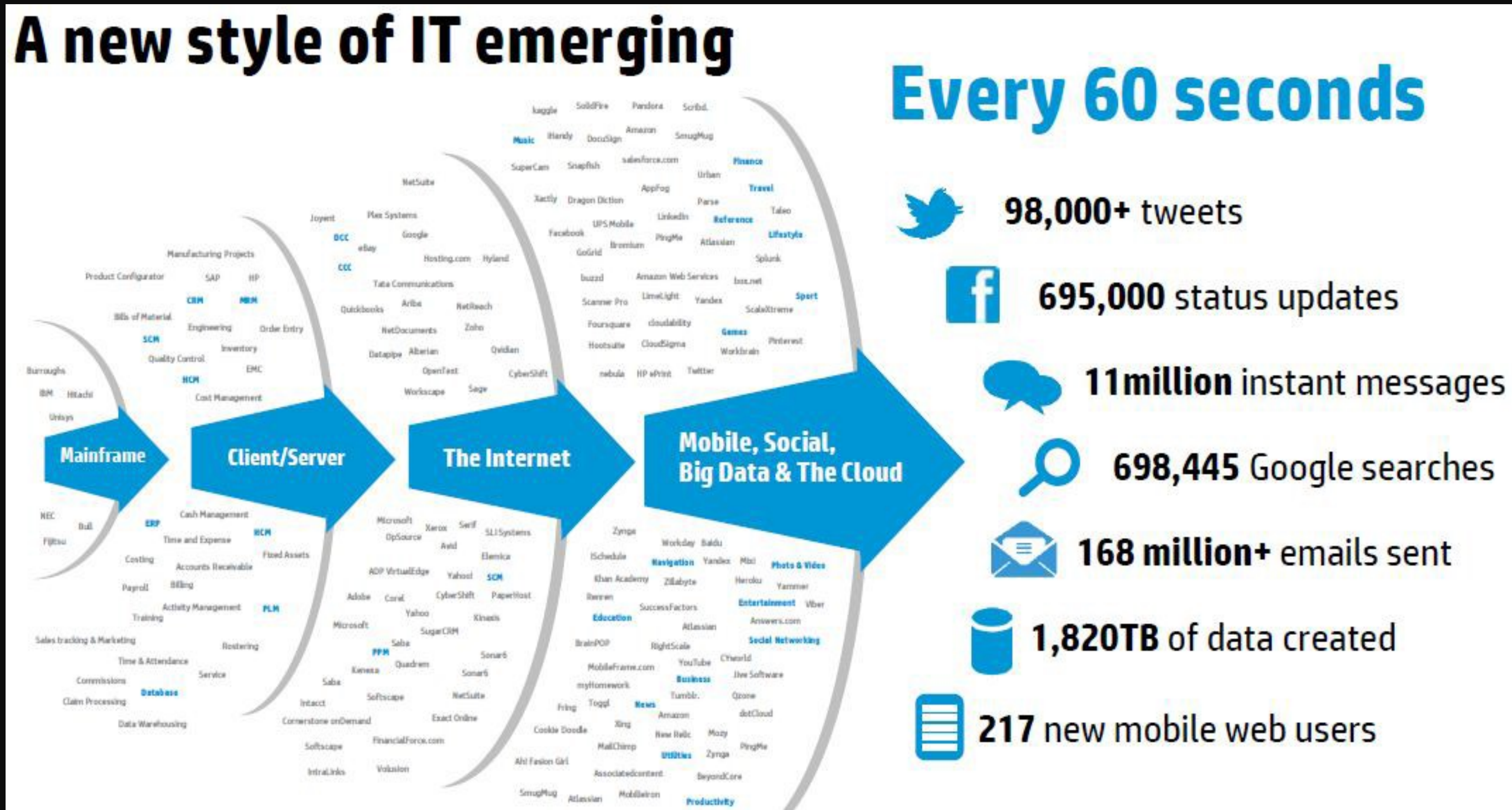
11
Acquisitions by Google

Google is the most active acquirer of AI startups, having acquired 11 startups since 2012. Apple, which has been ramping up its M&A efforts, ranked second with 7 acquisitions under its belt. Newer entrants in the race include Ford, which acquired Argo AI for $1B in Q1'17, cybersecurity company Sophos, and Amazon.

https://www.cbinsights.com/blog/top-acquirers-ai-startups-ma-timeline/

But WHY?

Mobile Creates 24/7 Data Deluge



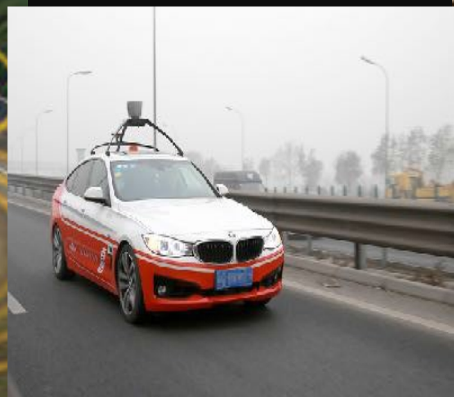
Big Data + Processing Power =

New Age for Artificial Intelligence

A few AI applications today

A LOT OF NUMBER CRUNCHING	VISION	LANGUAGE PROCESSING
BUSINESS INTELLIGENCE	AUTO TECH AND DRONE COLLISION AVOIDANCE	CHATBOTS
IOT PREDICTIVE MAINTENANCE	E-COMMERCE SEARCH	NEWS & MEDIA CONTENT CREATION
SEARCH RECOMMENDATIONS	PICK AND PLACE ROBOTS	SMART HOME VOICE INTERFACES
FORECASTING MODELS	HEALTHCARE DIAGNOSTICS	TEXT ANALYTICS

Disruption in Consolidated Industrial Segments: Autonomous Vehicles



U.S. Department of Transportation (DOT) federal policy framework for autonomous vehicles: *Automated Driving Systems 2.0: A Vision for Safety*.

October 13, 2017

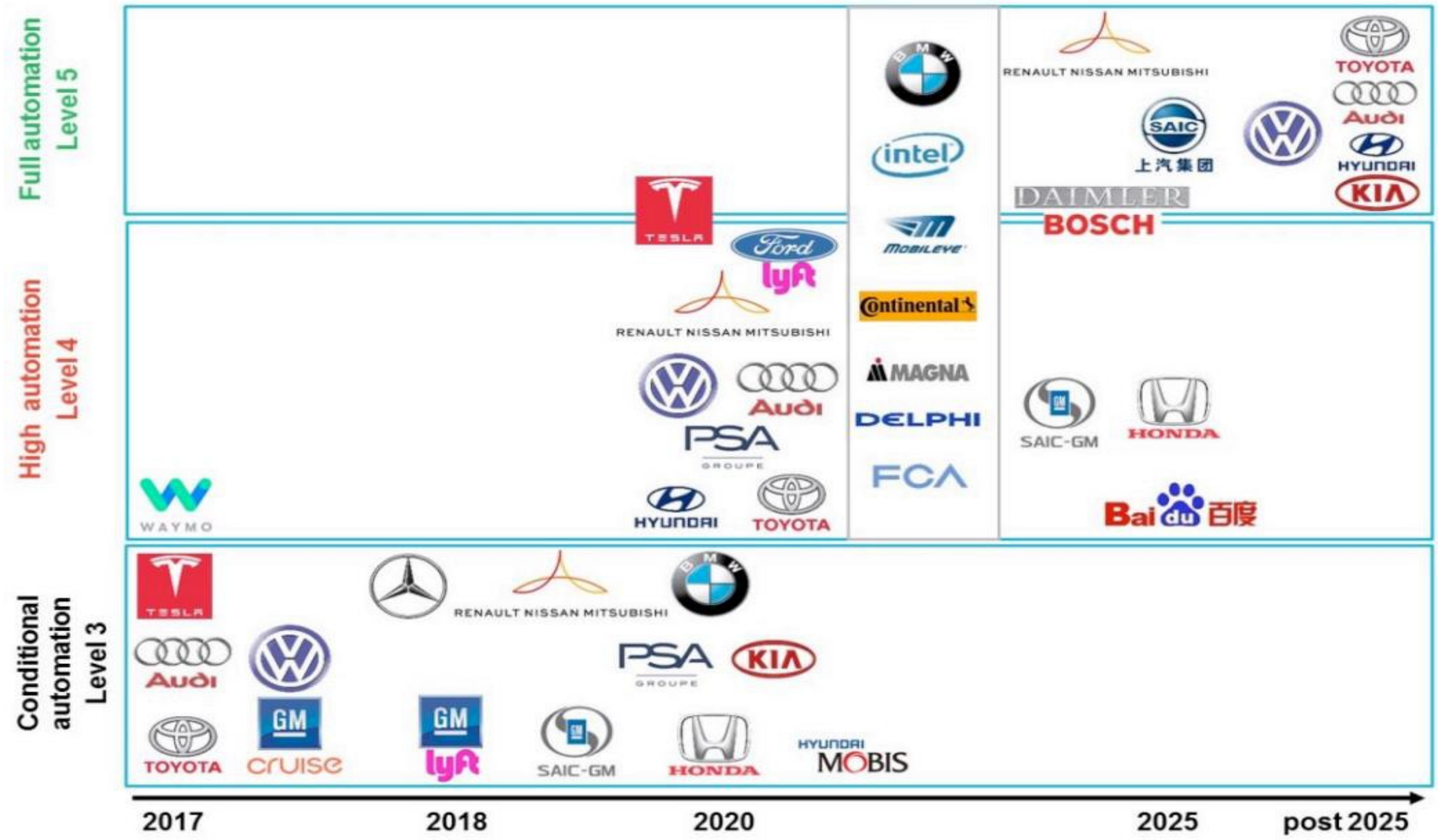
Every year, **1.2 million lives** are lost to traffic crashes around the world, and in the U.S. the number of tragedies is growing. A common element of these crashes is that **94% involve human error**. Driving is not as safe or as easy as it should be, while distracted driving is on the rise. We believe our technology could save thousands of lives now lost to traffic crashes every year.

MARCH 21, 2017

21 Industries Other Than Auto That Driverless Cars Could Turn Upside Down

Launch timelines all over the board

Figure 16: Autonomous vehicle launch timelines based on public announcements

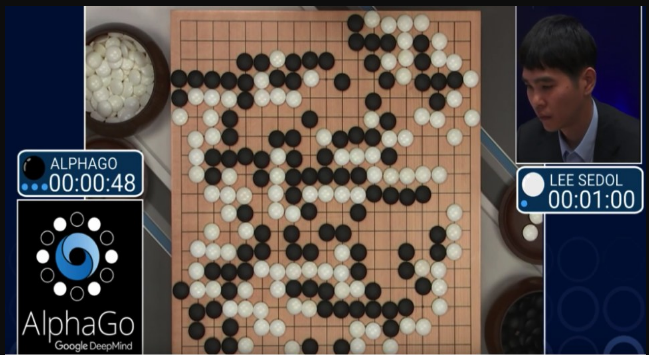


Four Fundamental Questions

1. **Where Am I?**
   - Sensing technology: GPS, Inertial,… (mapping technology)
2. **What's Around Me?**
   - «Vision» systems: Radars, Lidars, Camera systems (neural networks for image recognition)
3. **What Will Happen Next?**
   - Predictive systems: software and algorithms (dynamical systems)
4. **What Should I Do?**
   - Decision systems (neural networks for decision making, connected cars, trip planning)

SENSOR FUSION AND BIG DATA

Reaching Human Level Performance

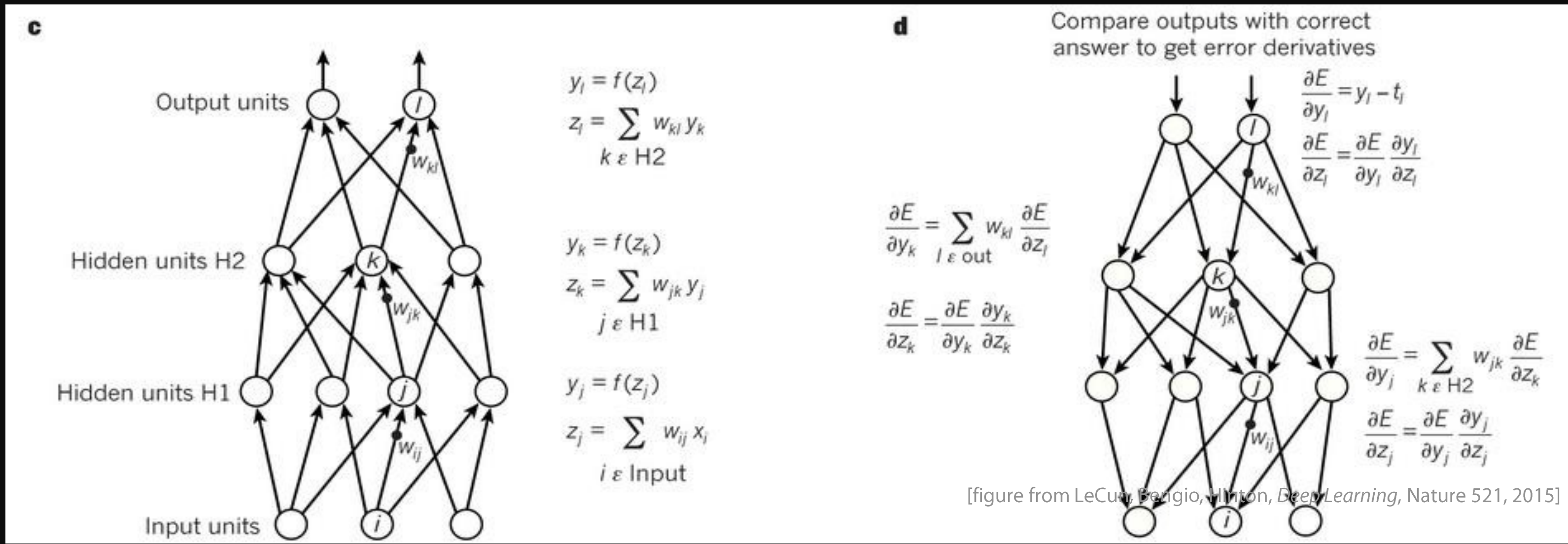


2017

Outline

- Setting the stage
- Artificial Intelligence
  - Economics
  - Neural networks and learning
  - Hype vs. reality

Artificial Neural Network



[figure from LeCun, Bengio, Hinton, *Deep Learning*, Nature 521, 2015]

Function approximation

Basically, this is what an artificial neural network does

Supervised learning

The parameters (i.e. *weights*) are "learnt" from a dataset of inputs and expected outputs pairs

Incremental optimization — a.k.a. "*backward propagation*"

Weights are progressively corrected to reduce *the difference* between actual and expected outputs

An OLD problem

The field of **system identification** uses statistical methods to build mathematical models of dynamical systems from measured data. System identification also includes the optimal design of experiments for efficiently generating informative data for fitting such models as well as model reduction.

The origin of the field can be traced to the work in astronomy by Kepler and others (**1772**)

Goodwin, Graham C. & Payne, Robert L. *Dynamic System Identification: Experiment Design and Data Analysis*. Academic Press, 1977.

M. Deistler, *System Identification and Time Series Analysis: Past, Present, and Future*, B. Pasik-Duncan (Ed.): Stochastic Theory and Control, LNCIS 280, pp. 97–109, 2002.Springer-Verlag Berlin Heidelberg .

Feed-Forward Neural Network

- Approximating a target function

$$y = f^*(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d$$

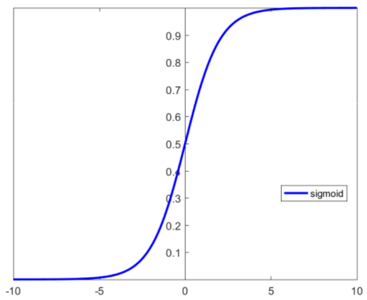
(shallow) feed-forward neural network

$$\tilde{y} = \boldsymbol{w} \cdot g(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) + b, \quad \boldsymbol{W} \in \mathbb{R}^{h \times d}, \ \boldsymbol{w}, \boldsymbol{b} \in \mathbb{R}^h, b \in \mathbb{R}$$

Popular choices for the non-linear function:

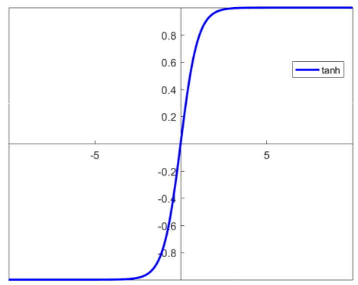
this is somewhat special...

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



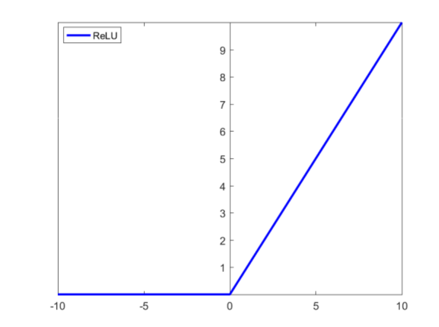
Sigmoid

$$g(x) = \tanh(x)$$



Hyperbolic Tangent

$$g(x) = \max(0, x)$$



ReLU

See http://vision.unipv.it/AI/AIRG.html

Universal approximation theorem
(Cybenko, 1989, Hornik, 1991)

For any target function

$$y = f^*(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d$$

(which is continuous and Borel measurable)

and any $\varepsilon > 0$ there exists parameters

$$h \in \mathbb{Z}^+, \boldsymbol{W} \in \mathbb{R}^{h \times d}, \ \boldsymbol{w}, \boldsymbol{c} \in \mathbb{R}^h, c \in \mathbb{R}$$

this is the dimension of the hidden layer: it is a parameter in the theorem

such that the ***(shallow) feed-forward neural network***

$$\tilde{y} = \boldsymbol{w} \cdot g(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{c}) + c$$

approximates the target function by less than ε

$$|\, f^*(\boldsymbol{x}) - \boldsymbol{w} \cdot g(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{c}) + c \,| < \varepsilon$$

(on a compact subset of $\mathbb{R}^d$)

http://vision.unipv.it/AI/AIRG.html

The Mother of all DCNNs

Deep Convolutional Neural Network (DCNN)

- ***AlexNet*** [Krizhevsky, Sutskever & Hinton, 2012]

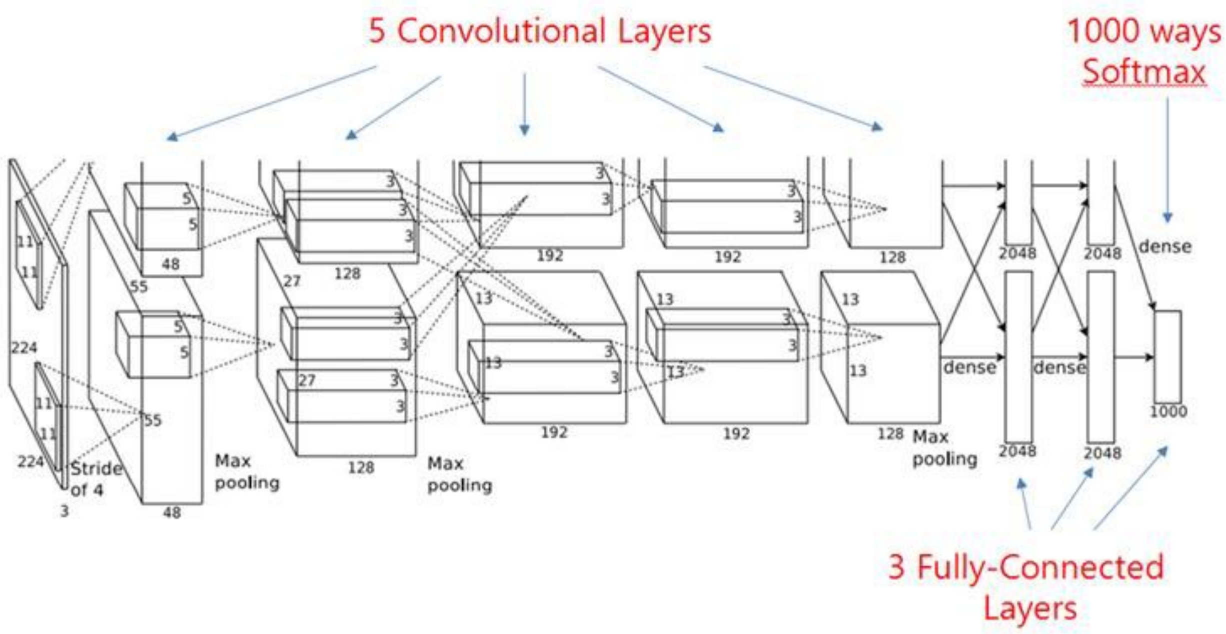


Image from [Krizhevsky, Sutskever & Hinton, 2012]

http://vision.unipv.it/AI/AIRG.html

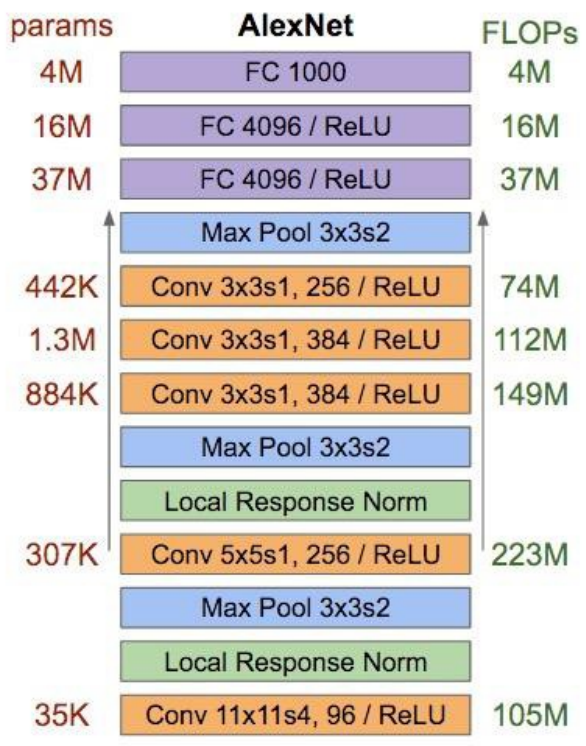
AlexNet Architecture

- ***AlexNet*** [Krizhevsky, Sutskever & Hinton, 2012]
  - number of parameters, per layer
    in red on the left
  - number of floating point operations, (FLOP) per layer in single forward pass
    in green on the right

Higher layers have more parameters but the bulk of the computation takes place at lower layers

Totals:

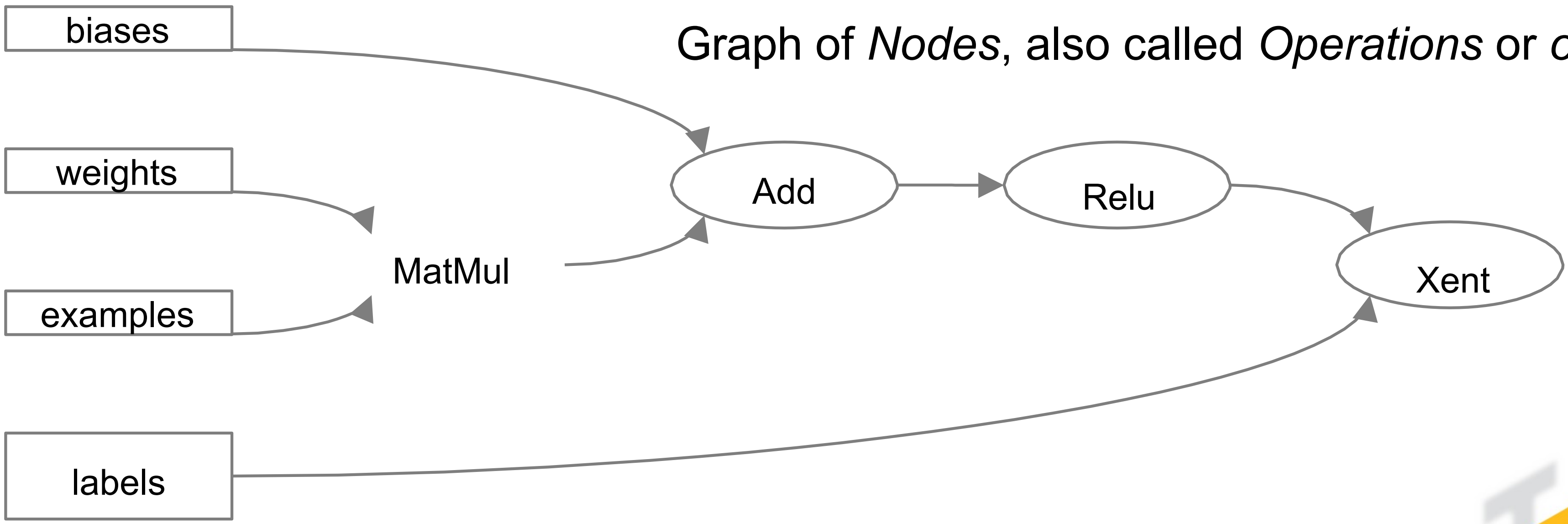
- around 60M parameters
- around 837M FLOPs for a single pass



http://vision.unipv.it/AI/AIRG.html

Computation is a dataflow graph

Graph of *Nodes*, also called *Operations* or *ops.*

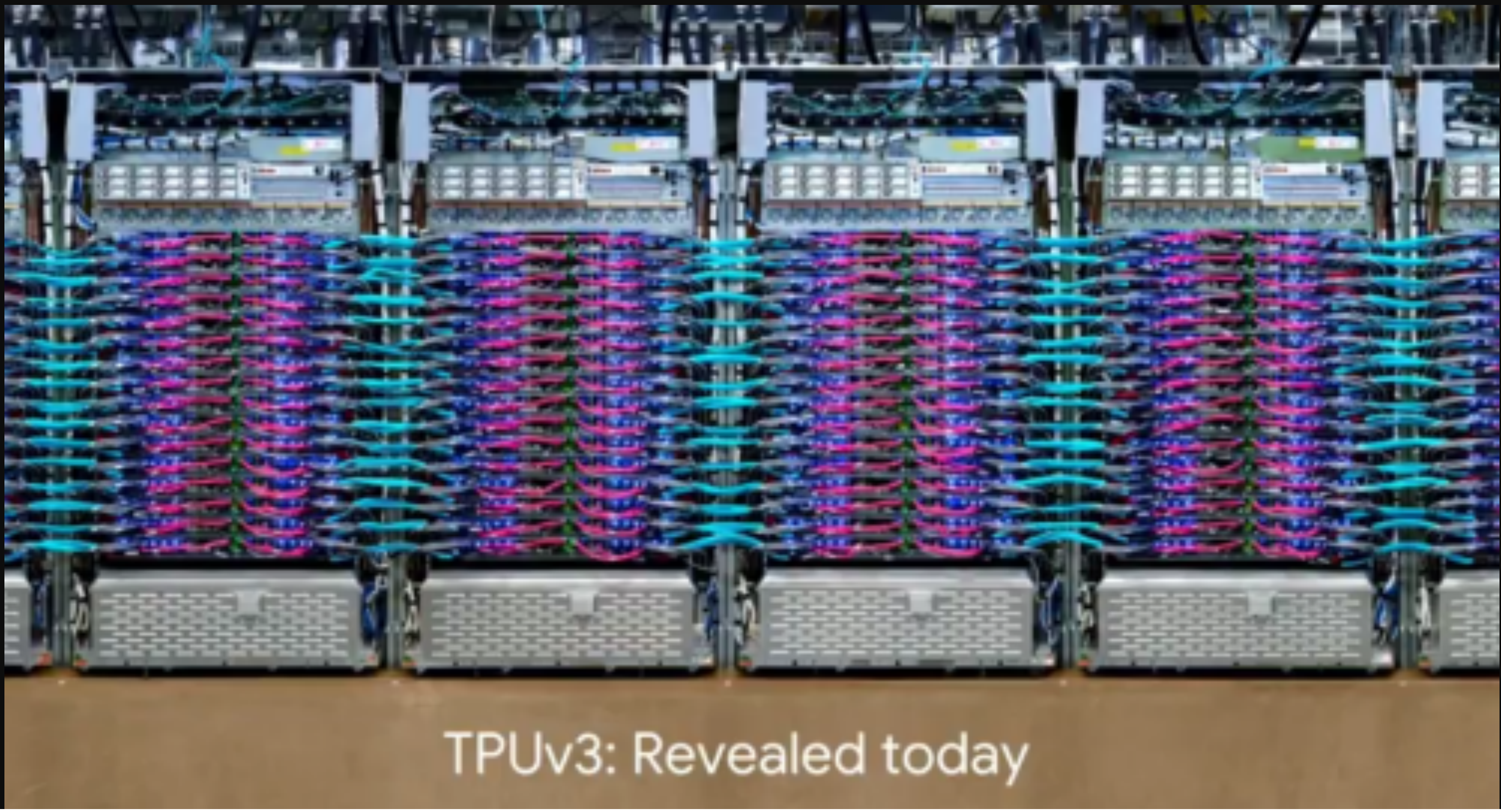


AI Chips: the race is ON!

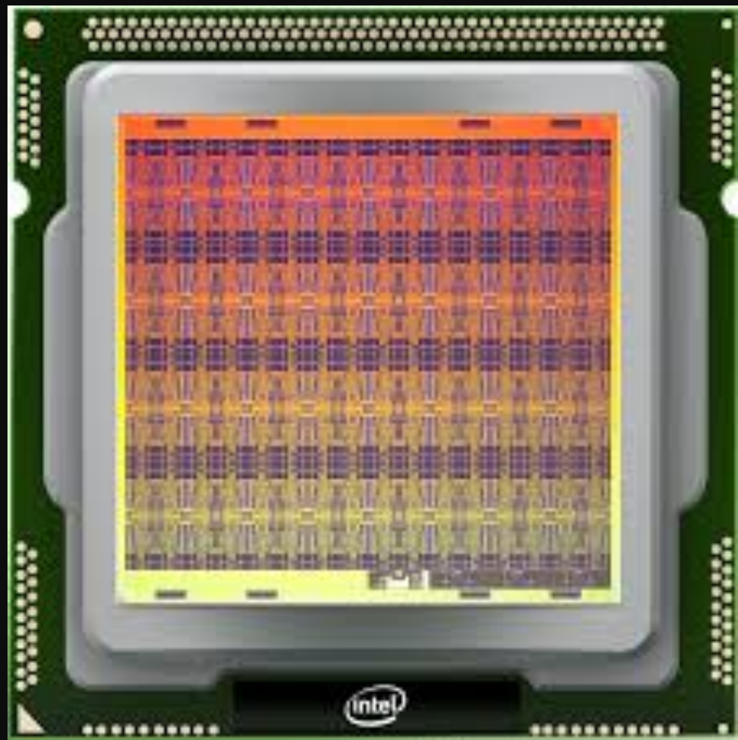
IC Vendors	Intel, Qualcomm, Nvidia, Samsung, AMD, Xilinx, IBM, STMicroelectronics, NXP, MediaTek, HiSilicon, Rockchip
Tech Giants & HPC Vendors	Google, Amazon_AWS, Microsoft, Apple, Aliyun, Alibaba Group, Tencent Cloud, Baidu, Baidu Cloud, HUAWEI Cloud, Fujitsu, Nokia, Facebook
IP Vendors	ARM, Synopsys, Imagination, CEVA, Cadence, VeriSilicon, Videantis
Startups in China	Cambricon, Horizon Robotics, DeePhi, Bitmain, Chipintelli, Thinkforce
Startups Worldwide	Cerebras, Wave Computing, Graphcore, PEZY, KnuEdge, Tenstorrent, ThinCI, Koniku, Adapteva, Knowm, Mythic, Kalray, BrainChip, AImotive, DeepScale, Leepmind, Krtkl, NovuMind, REM, TERADEEP, DEEP VISION, Groq, KAIST DNPU, Kneron, Esperanto Technologies, Gyrfalcon Technology, SambaNova Systems, GreenWaves Technology

Nvidia Volta

Google TPUv3



Intel Loihi Neuromorphic



Outline

- Setting the stage
- Artificial Intelligence
  - Economics
  - Neural networks and learning
  - Hype vs. reality

Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts

Big-data boondoggles and brain-inspired chips are just two of the things we're really getting wrong

By Lee Gomes
Posted 20 Oct 2014 | 19:37 GMT

Photo-illustration: Randi Klett

Michael Jordan: There are no spikes in deep-learning systems. There are no dendrites. And they have bidirectional signals that the brain doesn't have.

We don't know how neurons learn. Is it actually just a small change in the synaptic weight that's responsible for learning? That's what these artificial neural networks are doing. In the brain, we have precious little idea how learning is actually taking place.

Spectrum**:** I read all the time about engineers describing their new chip designs in what seems to me to be an incredible abuse of language. They talk about the "neurons" or the "synapses" on their chips. But that can't possibly be the case; a neuron is a living, breathing cell of unbelievable complexity. Aren't engineers

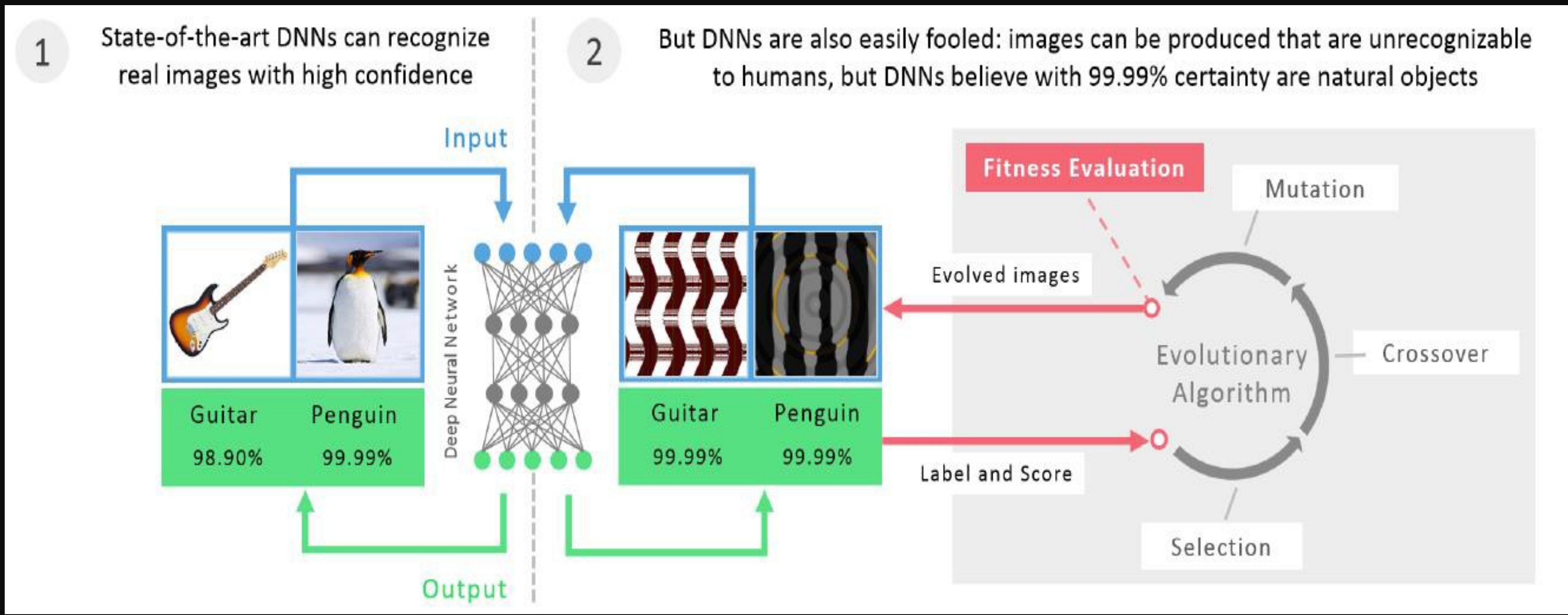
appropriating the language of biology to describe structures that have nothing remotely close to the complexity of biological systems?

Michael Jordan: Well, I want to be a little careful here. I think it's important to distinguish two areas where the word *neural* is currently being used.

One of them is in deep learning. And there, each "neuron" is really a cartoon. It's a linear-weighted sum that's passed through a nonlinearity. Anyone in electrical engineering would recognize those kinds of nonlinear systems. Calling that a neuron is clearly, at best, a shorthand. It's really a cartoon. There is a procedure called logistic regression in statistics that dates from the 1950s, which had nothing to do with neurons but which is exactly the same little piece of architecture.

http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts

A DCNN can be fooled…



[images from https://arxiv.org/pdf/1412.1897v4.pdf]

McKinsey Report on AI, 2018

- Two-thirds of the opportunities to use AI are in improving the performance of existing analytics use cases
- The biggest value opportunities for AI are in marketing and sales and in supply-chain management and manufacturing
- Limitations include the need for massive data sets, difficulties in explaining results, generalizing learning, and potential bias in data and algorithms
  - Data requirements for deep learning are substantially greater than for other analytics, in terms of both volume and variety
  - Ongoing data acquisition for retraining AI systems is necessary; one out of three use cases requires model refreshes at least monthly and sometimes daily

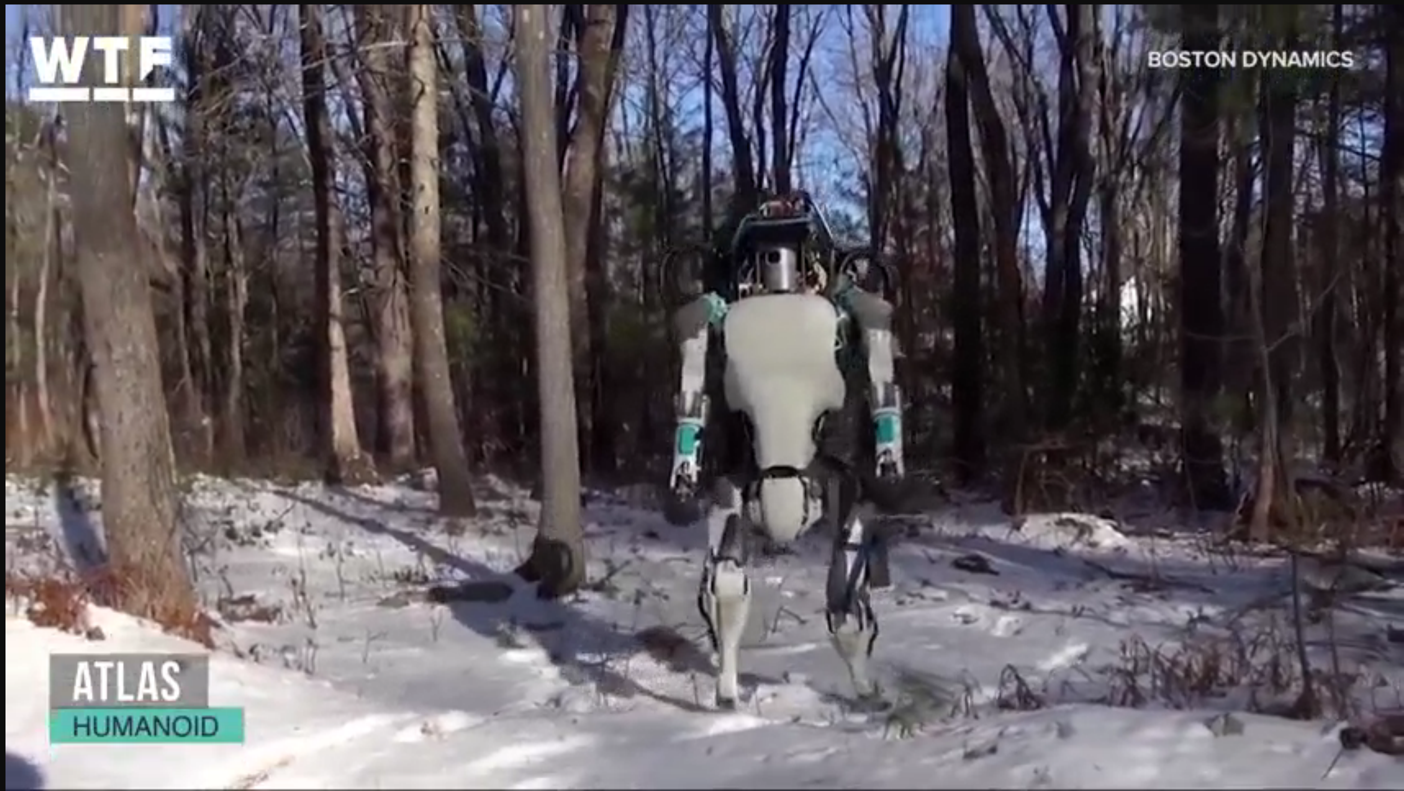
Final Words of Wisdom

- AI can seem an elusive business case for now
  - In some areas, the techniques today may be mature and the data available, but the **cost and complexity of deploying AI may simply not be worthwhile, given the value that could be generated**
- Societal concerns and regulations can constrain AI use

Last Fatal Accidents

In the **Uber crash** of March 19th, the ride services company was testing a fully driverless system intended for commercial use when the prototype vehicle struck and killed a woman walking across an Arizona road. Video of the crash, taken from inside the vehicle, shows the driver at the wheel, who appears to be looking down and not at the road. Just before the video stops, the driver looks upwards toward the road and suddenly looks shocked.

In the **Tesla incident** last month, which involved a car that any consumer can buy, a Model X vehicle was in semi-autonomous Autopilot mode when it crashed, killing its driver. The driver had received earlier warnings to put his hands on the wheel, Tesla said.


WTF
BOSTON DYNAMICS
ATLAS
HUMANOID

- Artificial Intelligence touted as the root of all future revolutions
- Overblown claims about capabilities
- Applicable to subsets of problems, not to ALL problems
- Andrew Ng: “There's a big difference between intelligence and sentience”.
- Data Analytics using ML can go only so far… If we know the physics of the problem, we are better off using our knowledge coded in physics-based mathematical models
- Important education on LIMITATIONS