SEQUOIA

Robust algorithms for learning from modern data

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

ERC, Brussels - October 2018

Machine learning
Scientific context

- **Proliferation of digital data** ($+10^{19}$ bytes per day)
  - Personal data
  - Industry
  - Scientific: from bioinformatics to humanities

- **Need for automated processing of massive data**

Machine learning
Scientific context

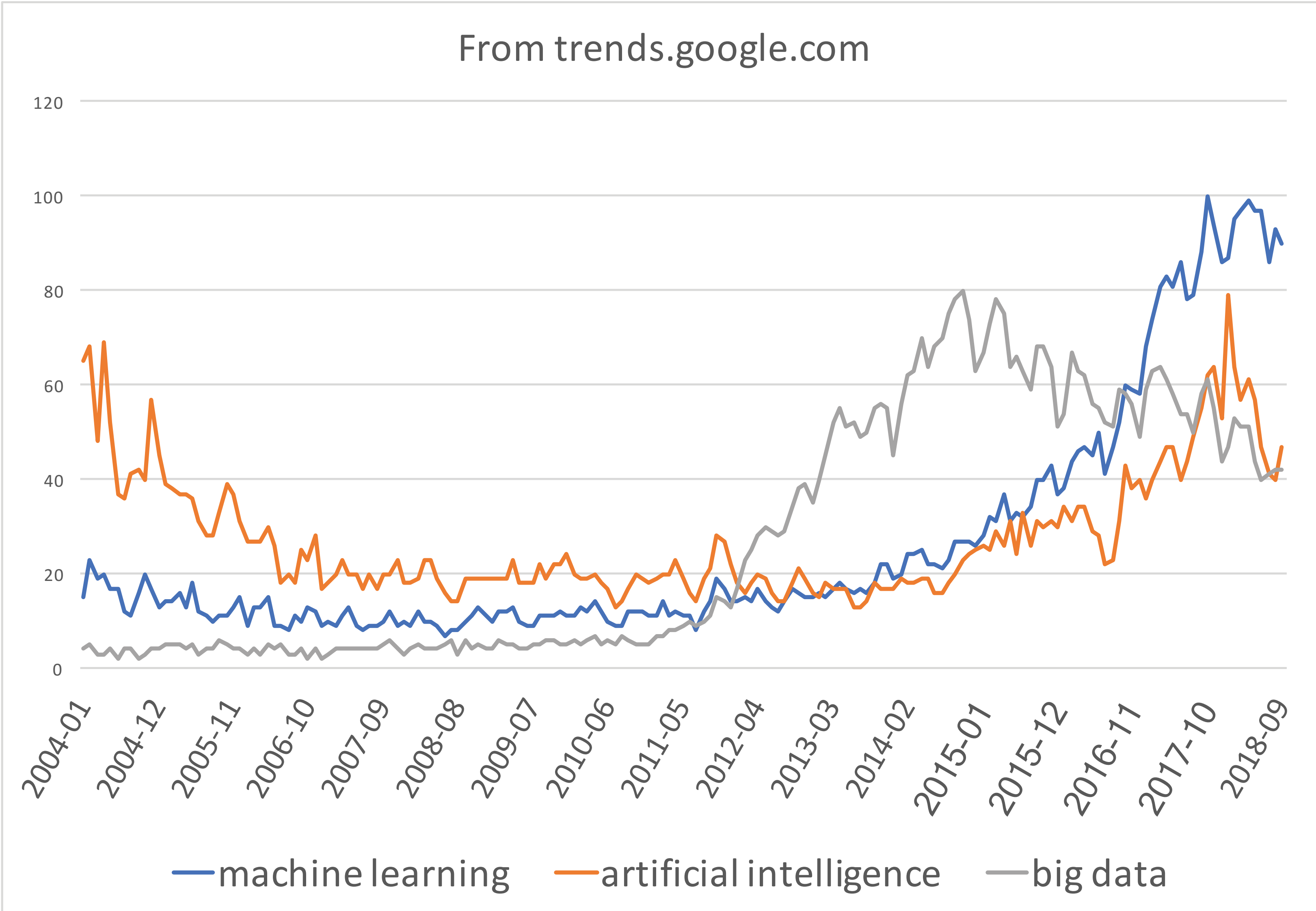
- **Proliferation of digital data** ($+10^{19}$ bytes per day)
  - Personal data
  - Industry
  - Scientific: from bioinformatics to humanities

- **Need for automated processing of massive data**

- **Series of "hypes"**

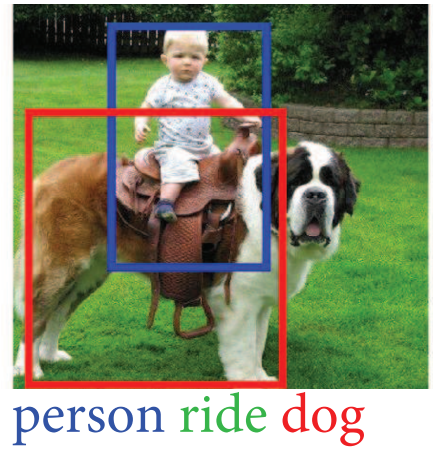
  Big data $\rightarrow$ Data science $\rightarrow$ Machine Learning
  $\rightarrow$ Deep Learning $\rightarrow$ Artificial Intelligence

An AI revolution?



Recent progress in perception (vision, audio, text)

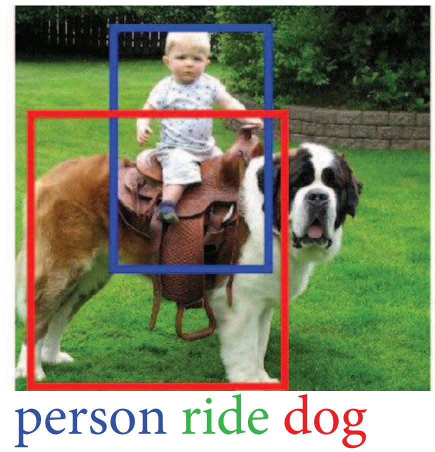
From translate.google.fr



From Peyré et al. (2017)

Recent progress in perception (vision, audio, text)

From translate.google.fr



From Peyré et al. (2017)

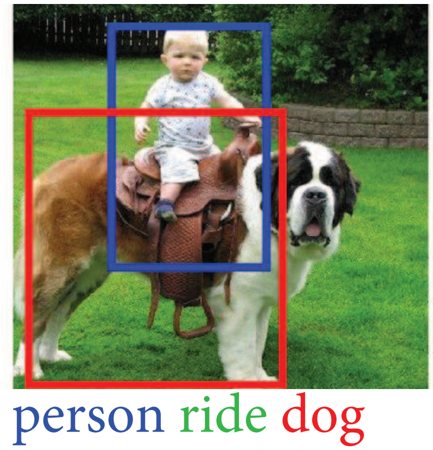
(1) **Massive data**

(2) **Computing power**

(3) **Methodological and scientific progress**

Recent progress in perception (vision, audio, text)

From translate.google.fr



From Peyré et al. (2017)

(1) **Massive data**

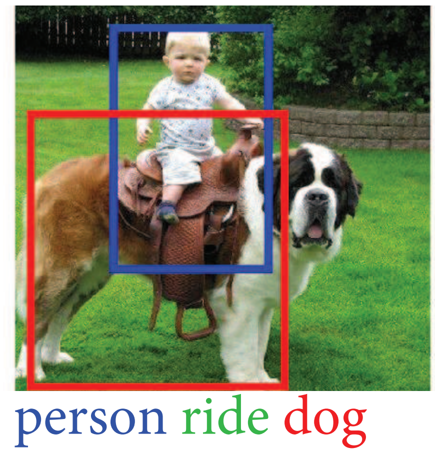
(2) **Computing power**

(3) **Methodological and scientific progress**

"Intelligence" = models + algorithms + data + computing power

Recent progress in perception (vision, audio, text)

From translate.google.fr



From Peyré et al. (2017)

(1) **Massive data**

(2) **Computing power**

(3) **Methodological and scientific progress**

"Intelligence" = models + algorithms + data + computing power

Machine learning

- **Scientific domain for 30+ years** $\neq$ **AI**
  - Building predictions from examples
  - Conferences NIPS, COLT and ICML + Journal JMLR
- **Theory, algorithms and applications**

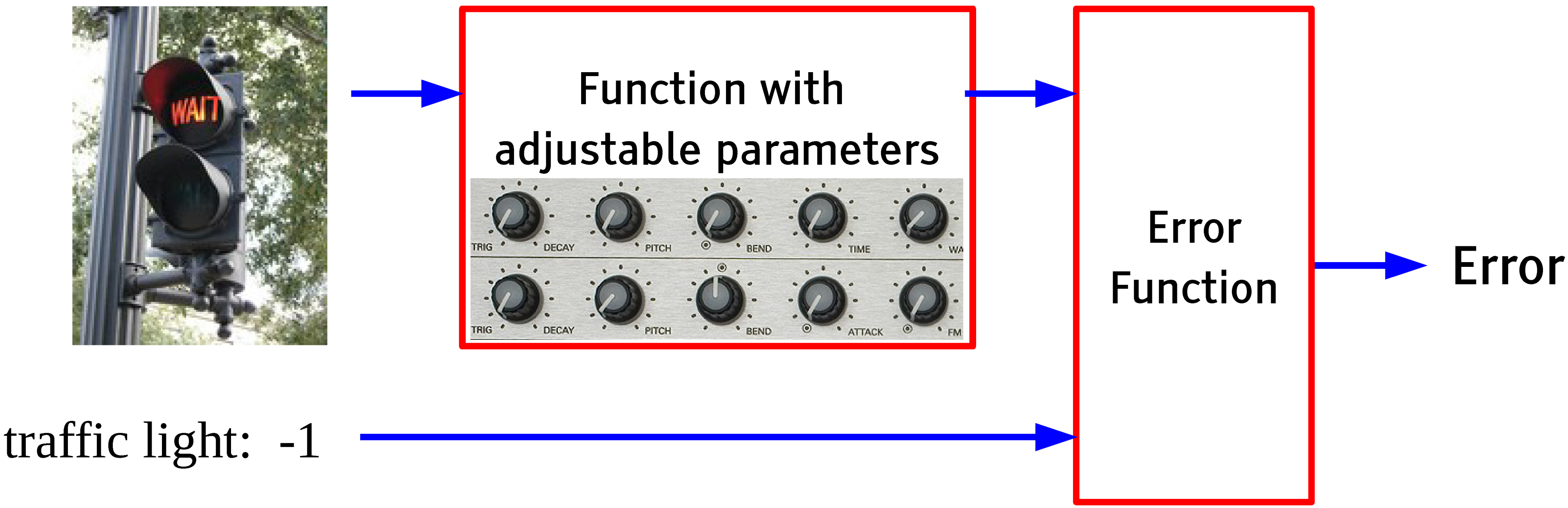
Machine learning

- **Scientific domain for 30+ years** $\neq$ **AI**
  - Building predictions from examples
  - Conferences NIPS, COLT and ICML + Journal JMLR

- **Theory, algorithms and applications**

- **Growth from 2000 to 2018**
  - NIPS: from 150 to 1000 articles, from 300+ to 8000 attendees
  - Impact from/on industry: between users and contributors

Supervised machine learning
A simplified view



From Yann Le Cun's lecture

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$
- Prediction as linear functions $\langle \theta, \Phi(x) \rangle = \sum_{j=1}^{d} \theta_j \Phi_j(x)$ of features $\Phi(x) \in \mathbb{R}^d$
- **Empirical risk minimization**:

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, \langle \theta, \Phi(x_i) \rangle\big) \quad + \quad \mu \, \Omega(\theta)$$

Data fitting term $+$ regularization

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$
- Prediction as linear functions $\langle \theta, \Phi(x) \rangle = \sum_{j=1}^{d} \theta_j \Phi_j(x)$ of features $\Phi(x) \in \mathbb{R}^d$
- **Empirical risk minimization**:

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, \langle \theta, \Phi(x_i) \rangle\big) \quad + \quad \mu \, \Omega(\theta)$$

Data fitting term + regularization

- Applications to any data-oriented field
  - Computer vision, bioinformatics
  - Natural language processing, etc.

Supervised machine learning

- **Data**: n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$

- Prediction as linear functions $\langle \theta, \Phi(x) \rangle = \sum_{j=1}^{d} \theta_j \Phi_j(x)$ of features $\Phi(x) \in \mathbb{R}^d$

- **Empirical risk minimization**:

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, \langle \theta, \Phi(x_i) \rangle\big) \quad + \quad \mu \, \Omega(\theta)$$

Data fitting term $\quad + \quad$ regularization

- Main practical challenges
  - Designing/learning good features $\Phi(x)$
  - Efficiently solving the optimization problem

New scientific challenges in machine learning

- **Supervised machine learning well understood**
  - Running at scale with optimization methods (single machine)
  - Dealing with high dimension through sparsity
  - Neural networks

Deep learning

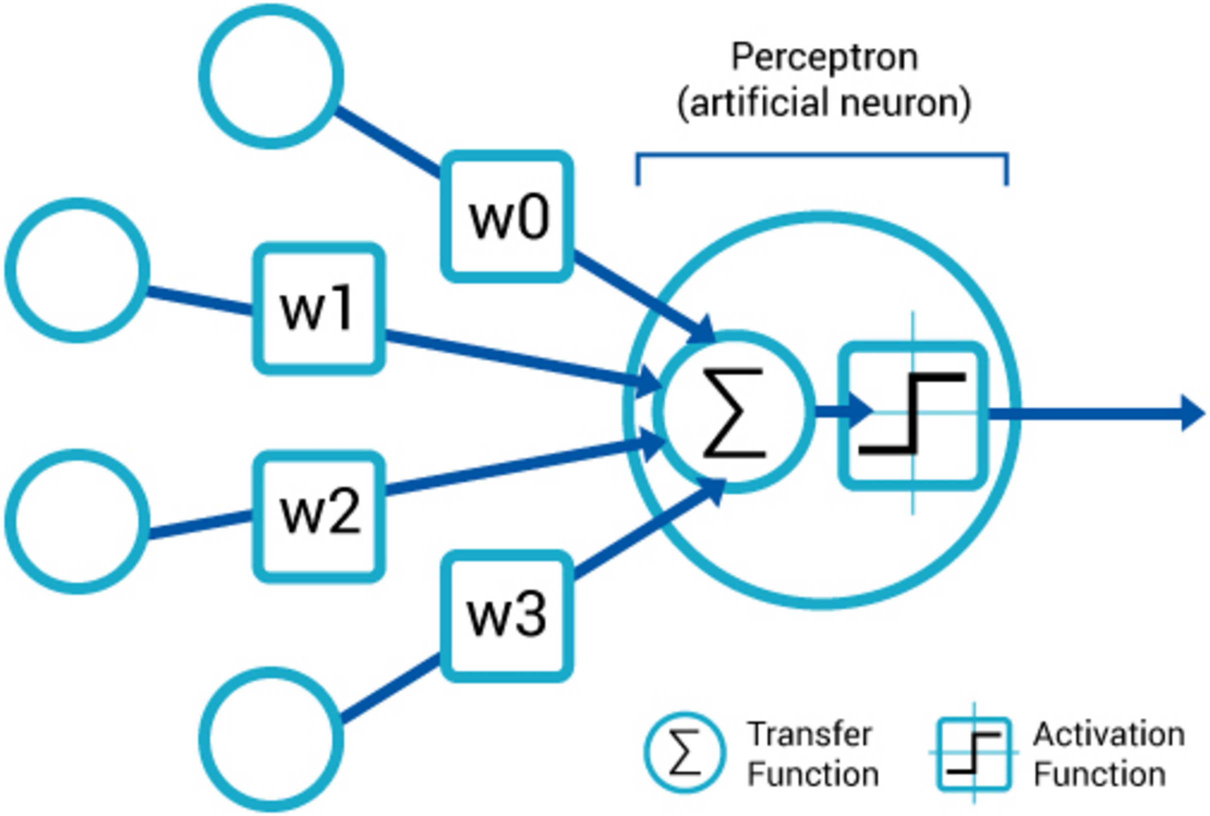
- **Shallow / non-deep learning**
  - Prediction as linear function $\langle \theta, \Phi(x) \rangle = \sum_{j=1}^{d} \theta_j \Phi_j(x)$ of known features $\Phi(x) \in \mathbb{R}^d$
  - Optimization (single machine) and theory well understood
  - Widespread use in industry (e.g., marketing and advertising)

Deep learning

- **Shallow / non-deep learning**
  - Prediction as linear function $\langle \theta, \Phi(x) \rangle = \sum_{j=1}^{d} \theta_j \Phi_j(x)$ of known features $\Phi(x) \in \mathbb{R}^d$
  - Optimization (single machine) and theory well understood
  - Widespread use in industry (e.g., marketing and advertising)

- **Deep neural networks**
  - Learning of features from data
  - Parametrization by combination of simple operations (+GPU)
  - Optimization and theory not totally understood
  - Works very well in vision / NLP with lots of training examples

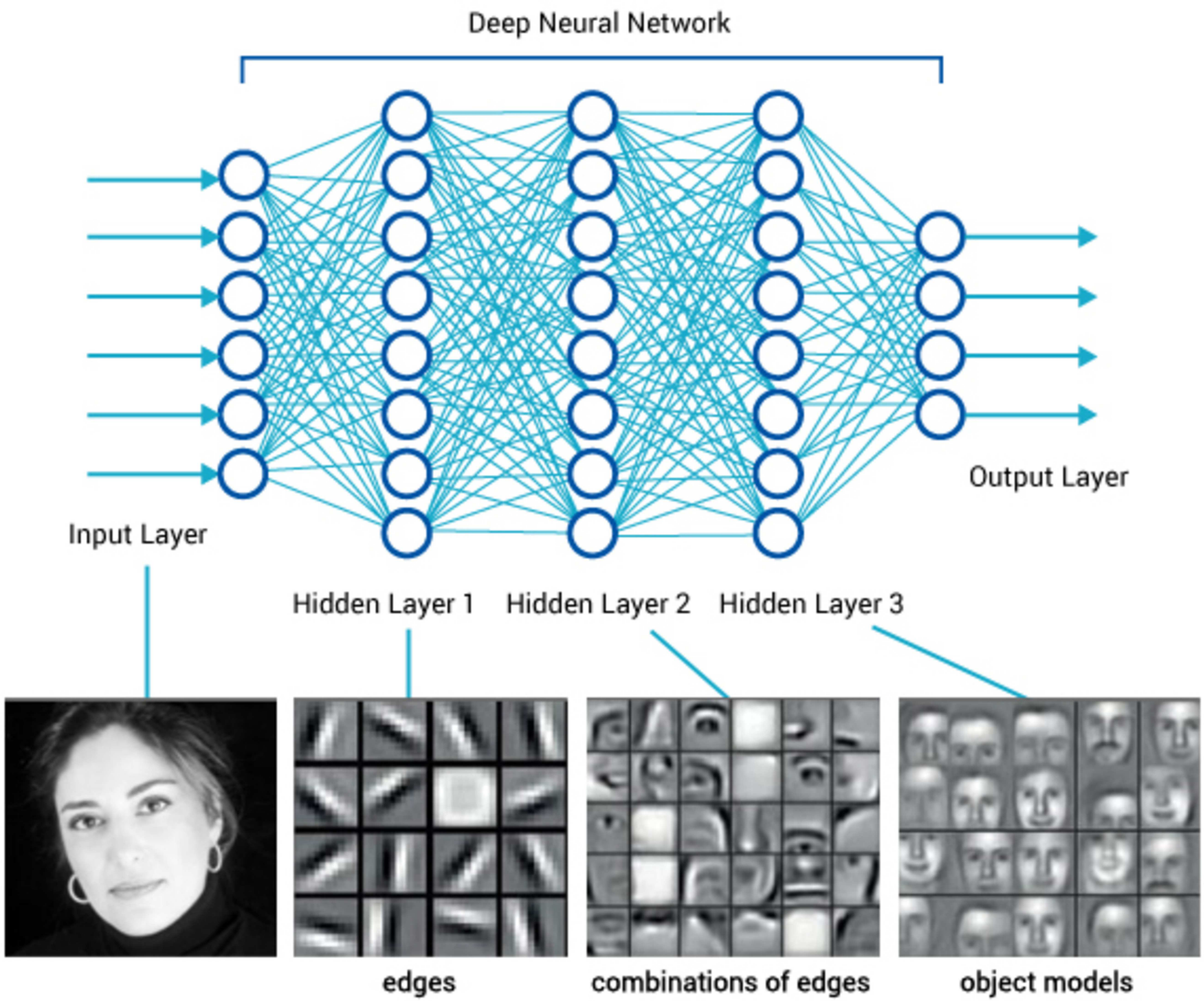
Neural networks
A single neuron



Figures from Goodfellow et al. (2016)

Linear prediction: $\sigma(w_0x_0 + w_1x_1 + w_2x_2 + w_3x_3)$

Deep neural networks



Non-linear prediction: $\theta_m^\top \sigma(\theta_{m-1}^\top \sigma(\cdots \theta_2^\top \sigma(\theta_1^\top x))$

New scientific challenges in machine learning

- **Supervised machine learning well understood**
  - Running at scale with optimization (single machine)
  - Dealing with high dimension through sparsity
  - Neural networks

- **Structured prediction**: beyond binary or real-valued outputs

- **Unsupervised learning**: weak supervision and relevance of results

- **Reinforcement learning**: mixing actions and predictions

- **Distributed optimization**: GPU / multi-cores / cloud

- **Non-convex optimization**: neural networks

- **Robust optimization**: beyond i.i.d. assumption

SEQUOIA : Robust algorithms for learning from modern data

- **Consolidator grant started in September 2017**
  - Between theory, algorithms and applications
- **Ambition**
  - Provable robustness and adaptivity to modern hardware and learning problems
- **Main focus**
  - Optimization algorithms
  - Theoretical guarantees and good empirical performance

Hot from the press
(Pillaud-Vivien, Rudi and Bach, NIPS 2018)

- **Stochastic gradient descent for large-scale machine learning**
  - Processes observations one by one

Hot from the press
(Pillaud-Vivien, Rudi and Bach, NIPS 2018)

- **Stochastic gradient descent for large-scale machine learning**
  - Processes observations one by one

- **Theory**: Single pass SGD is optimal

- **Practice**: Multiple pass SGD always works better

Hot from the press (Pillaud-Vivien, Rudi and Bach, NIPS 2018)

- **Stochastic gradient descent for large-scale machine learning**
  - Processes observations one by one
- **Theory**: Single pass SGD is optimal
  - Only for "easy" problems
- **Practice**: Multiple pass SGD always works better
  - Provable for "hard" problems
  - Quantification of required number of passes
  - Optimal statistical performance