data: PPF Lab10a lecture on RL

chris.wiggins@columbia.edu , 2020-04-09

machine learnings:
The 3 types of learning

learning (a la Cisco)

Figure 1: actual Cisco.com image called ‘get to know machine learning’

2 types of learning (a la MATLAB)

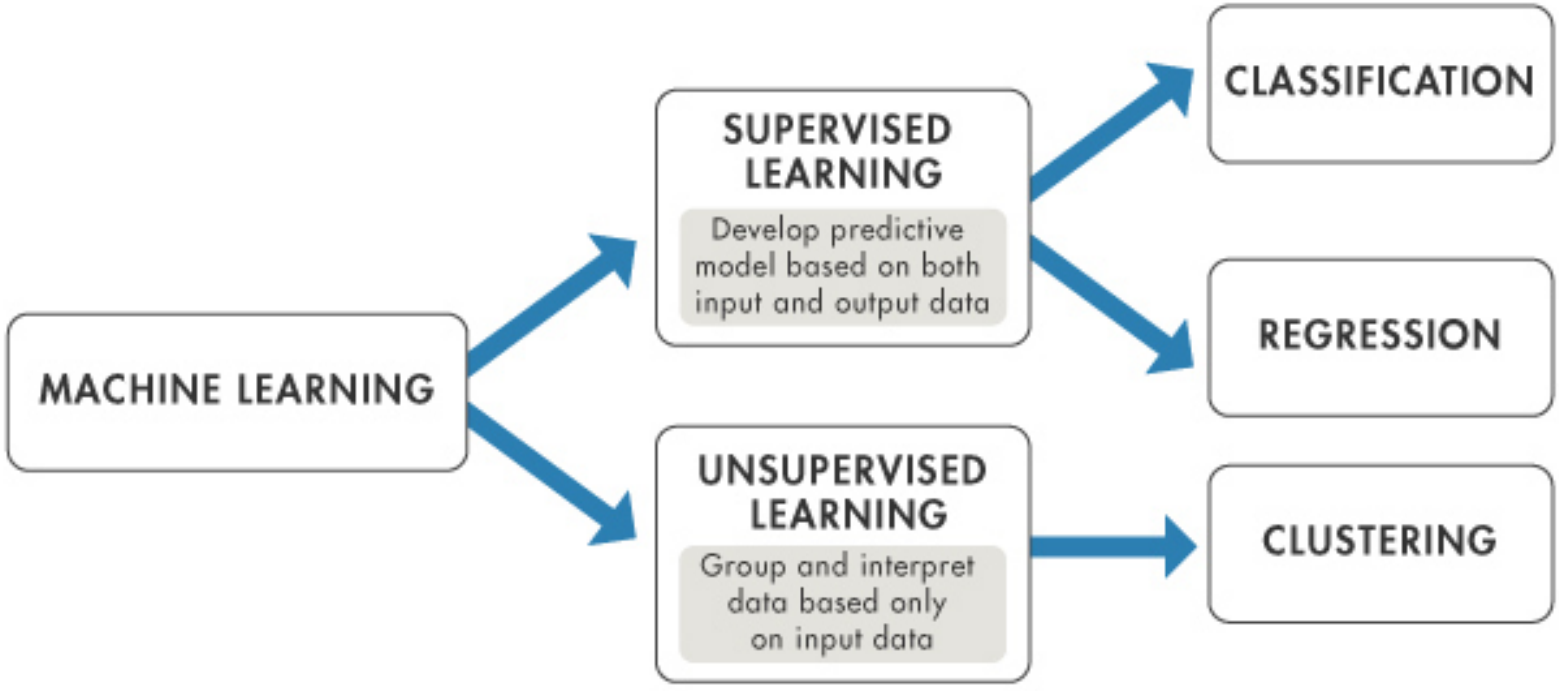


Figure 2: mathworks.com

2 types of learning (a la MATLAB)

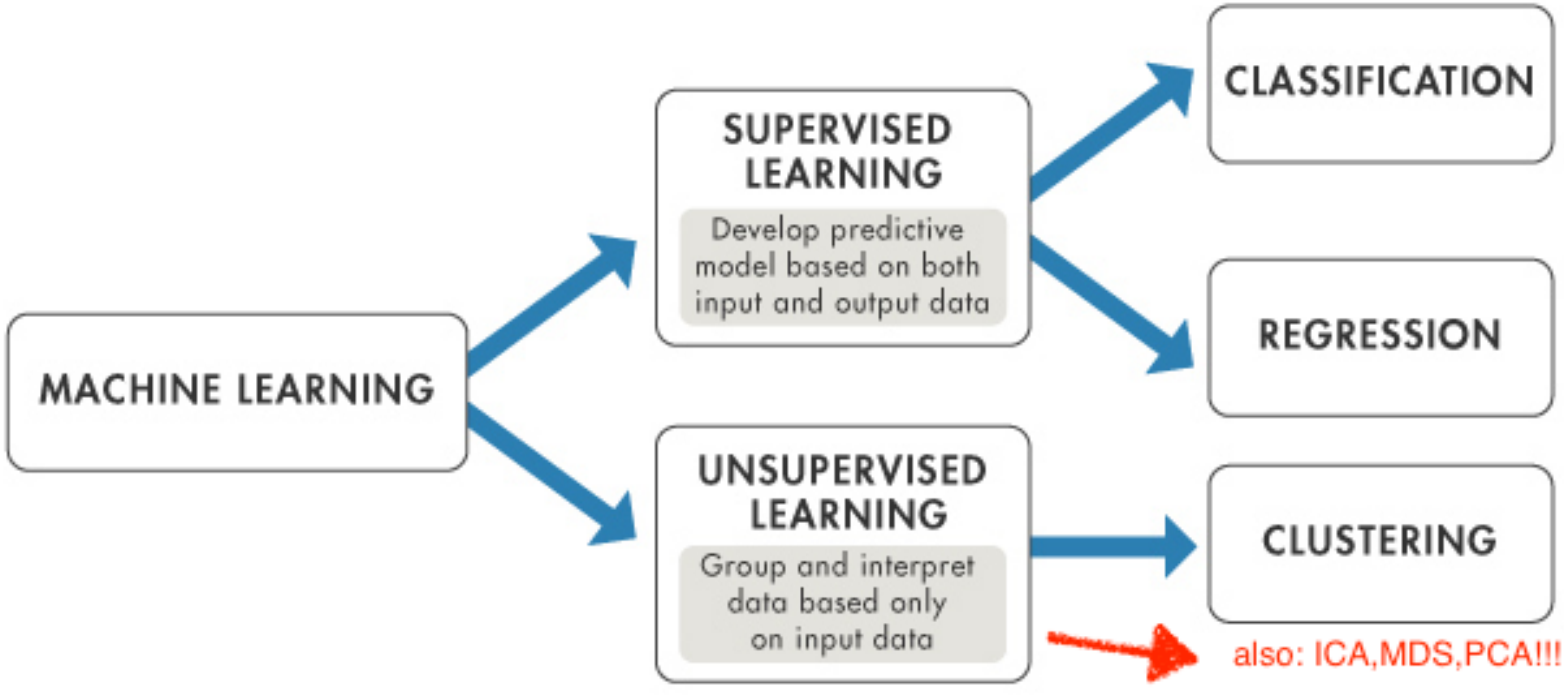


Figure 3: mathworks.com

2 types of learning (a la Python/scikit)

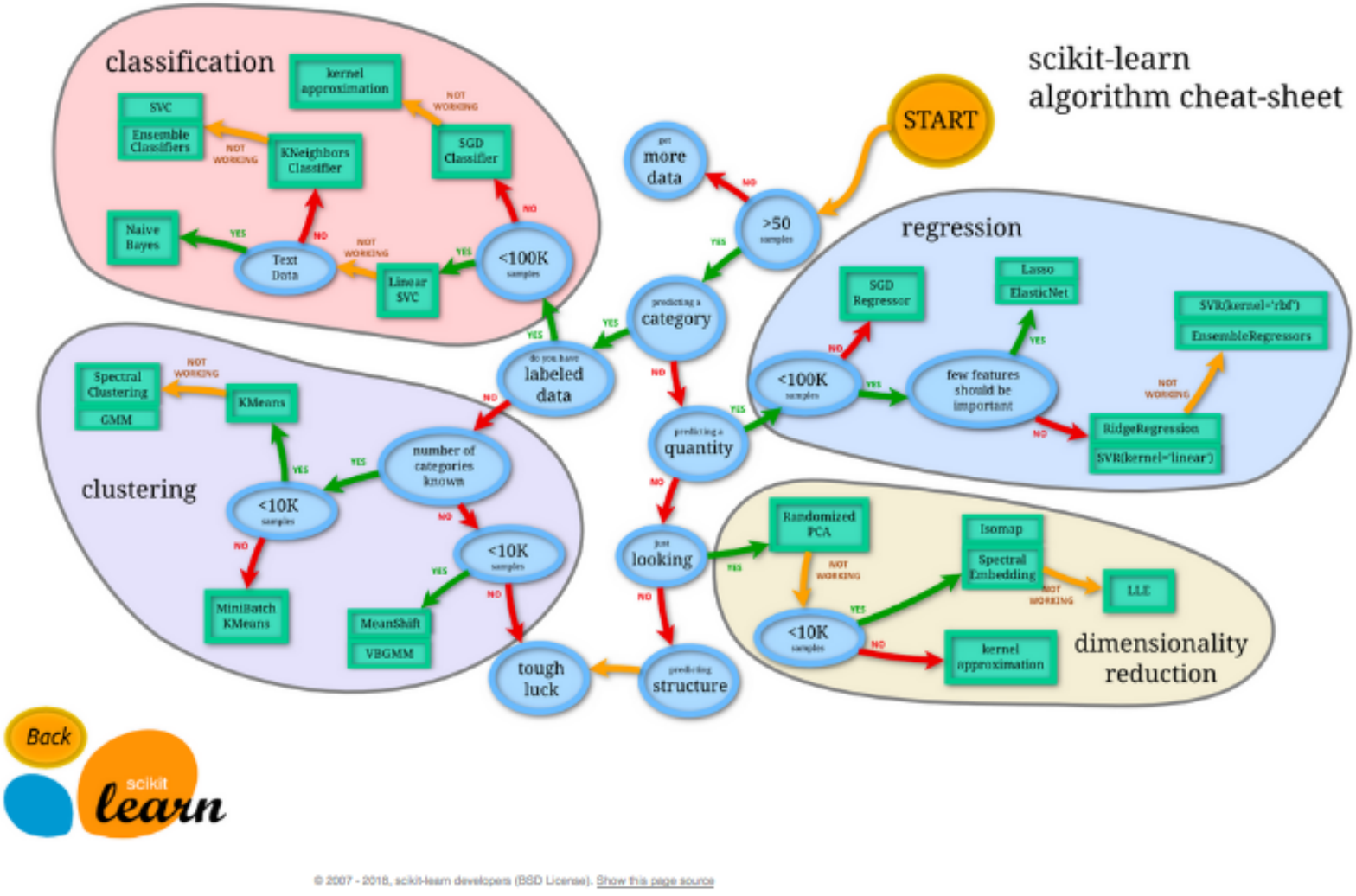


Figure 4: scikit documentation. Note: familiar OLS, PCA

3 types of learning (my view)

descriptive:	specify x; learn $z(x)$ or $p(z\|x)$ where z is "simpler" than x
predictive:	specify x and y; learn to predict y from x
prescriptive:	specify x, y, and a; learn to prescribe a given x to maximize y

Figure 5: desc/pred/pres, a human (also database) view

RL, control theory, cybernetics

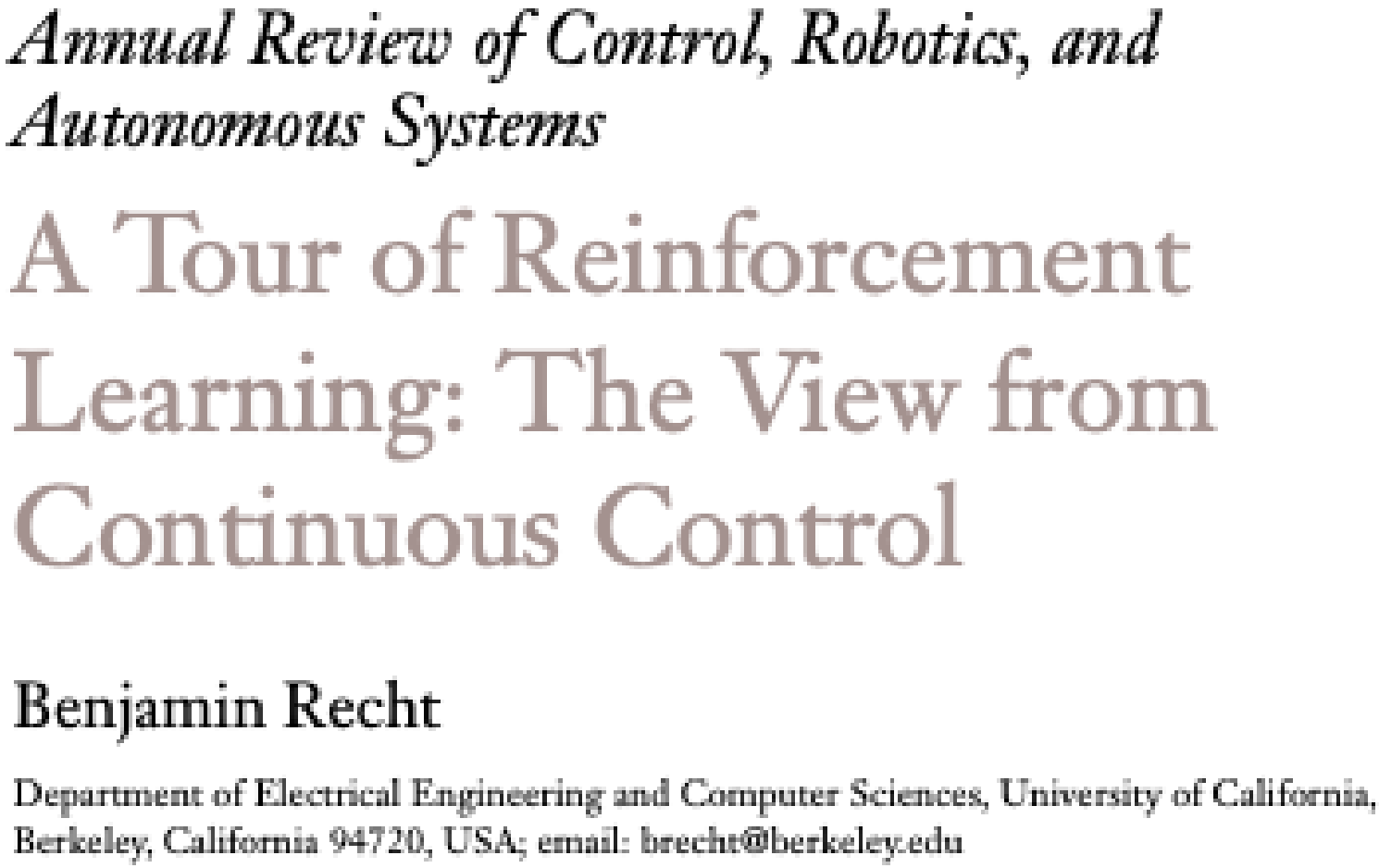
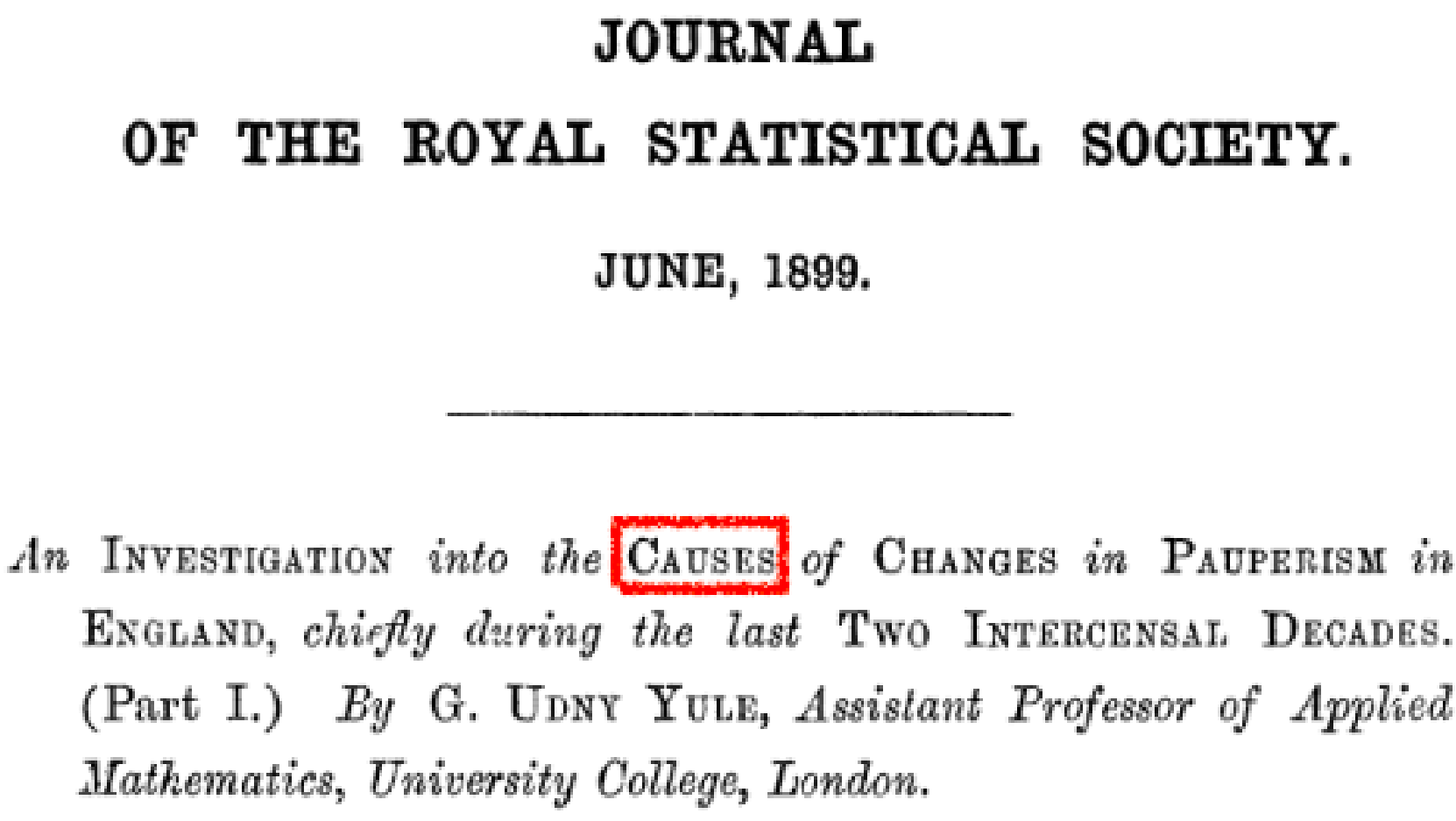


Figure 6: "view from control"

warning: don't do this



JOURNAL

OF THE ROYAL STATISTICAL SOCIETY.

JUNE, 1899.

An INVESTIGATION *into the* CAUSES *of* CHANGES *in* PAUPERISM *in* ENGLAND, *chiefly during the last* TWO INTERCENSAL DECADES. (Part I.) *By* G. UDNY YULE, *Assistant Professor of Applied Mathematics, University College, London.*

Figure 7: Yule 1889; $p(y|a,x)p(a|x)p(x) \neq p(y|a,x)p(a)p(x)$

3 types (a la Gartner)

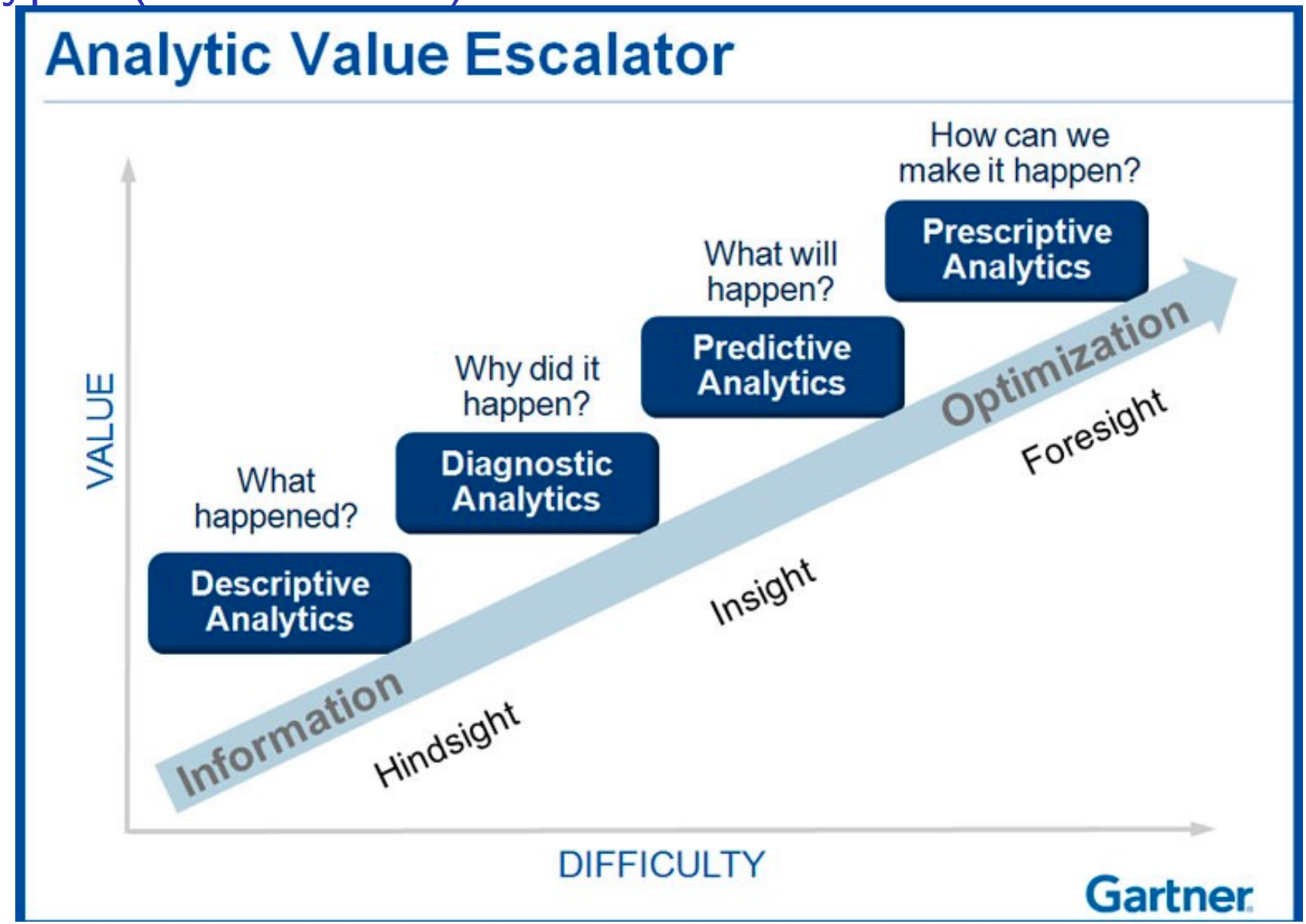


Figure 8: Gartner

3 types (a la Gartner, mod)

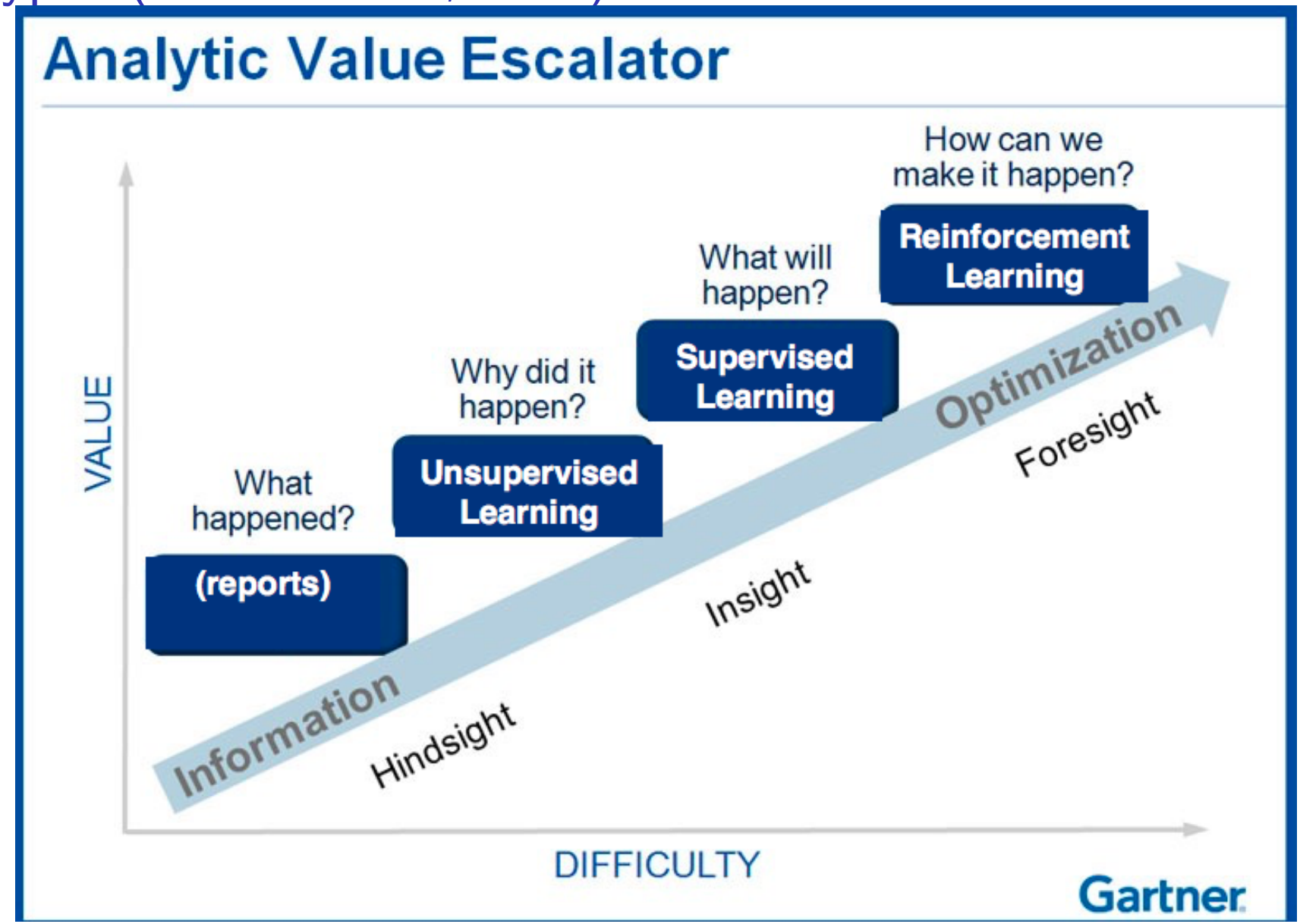


Figure 9: Gartner, translated

3 types: RL

Figure 10: atari

3 types: RL

DQN

Better than human-level control of classic Atari games through Deep

3 types: RL modern impact in games

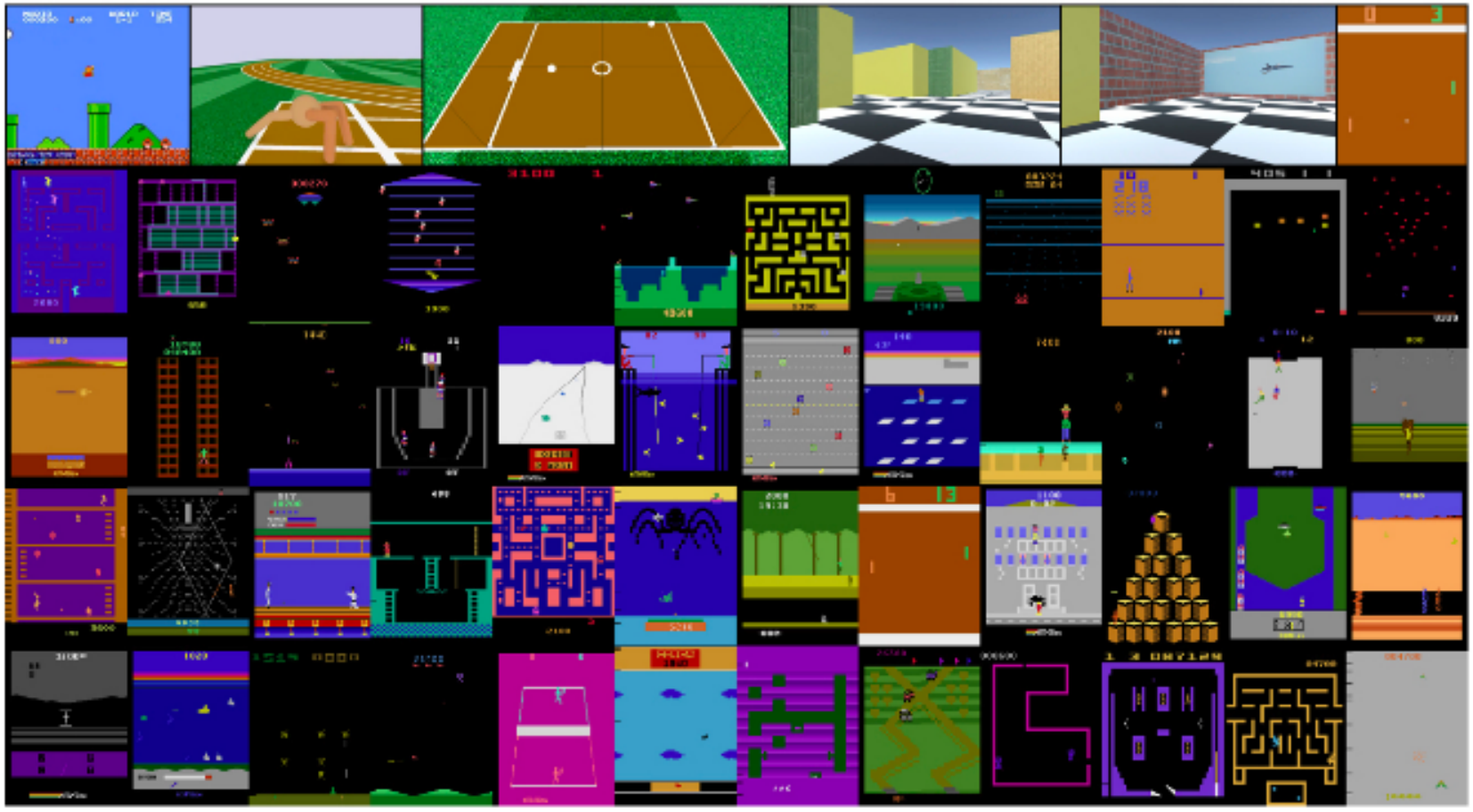


Figure 1: A snapshot of the 54 environments investigated in the paper. We show that agents are able to make progress using no extrinsic reward, or end-of-episode signal, and only using curiosity. Video results, code and models at https://pathak22.github.io/large-scale-curiosity/.

Figure 12: Large-Scale Study of Curiosity-Driven Learning 2018

3 types: RL (history note: AI+ML born of games)

Some Studies in Machine Learning Using the Game of Checkers

Arthur L. Samuel

Abstract: Two machine-learning procedures have been investigated in some detail using the game of checkers. Enough work has been done to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program. Furthermore, it can learn to do this in a remarkably short period of time (8 or 10 hours of machine-playing time) when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters which are thought to have something to do with the game, but whose correct signs and relative weights are unknown and unspecified. The principles of machine learning verified by these experiments are, of course, applicable to many other situations.

Figure 13: 1959: "ML" born to play checkers

3 types: RL modern impact in games + commerce

Figure 14: Stitch Fix: Stitch Fix for $, 2018

3 types: RL modern impact in games + commerce

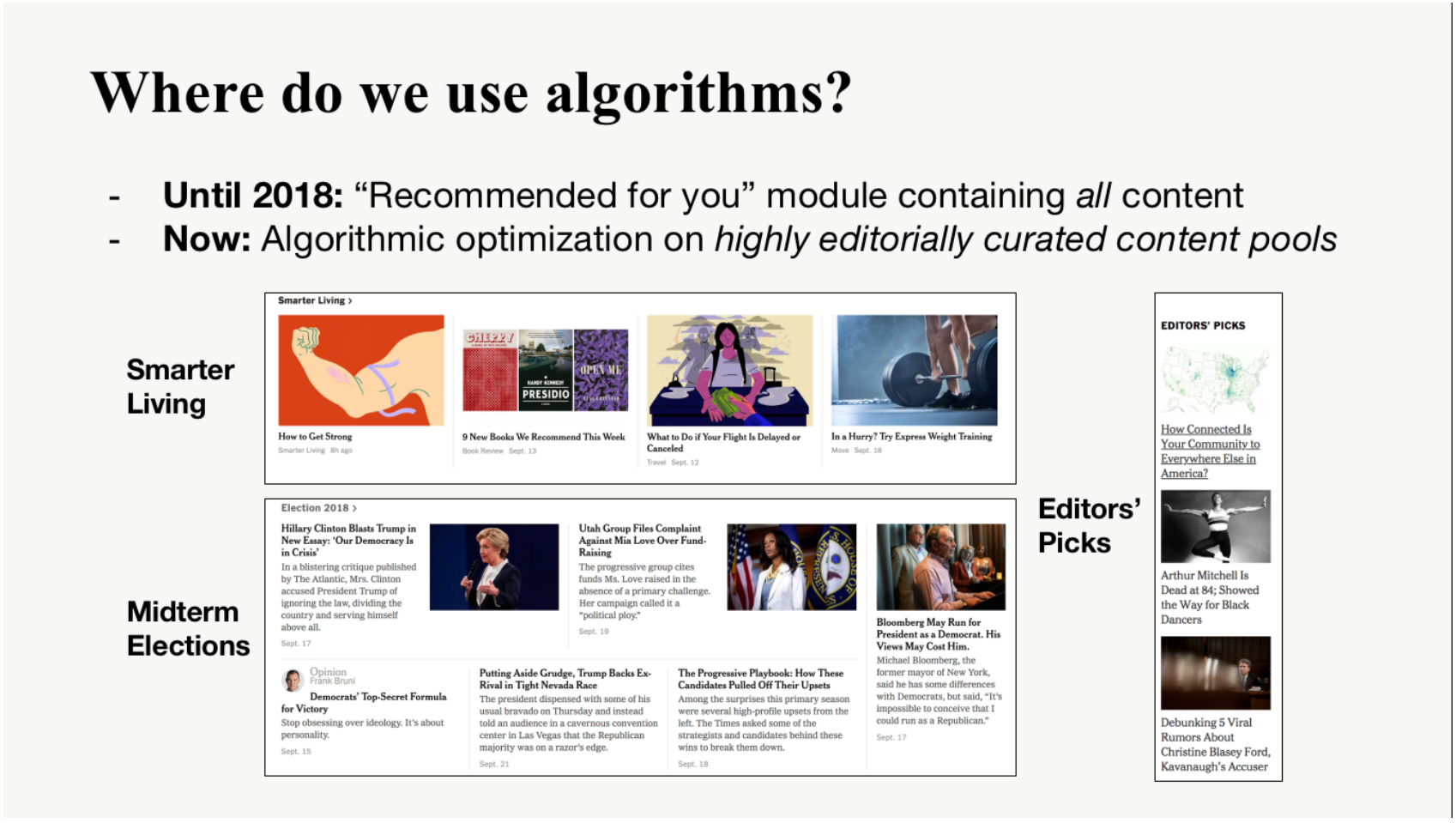
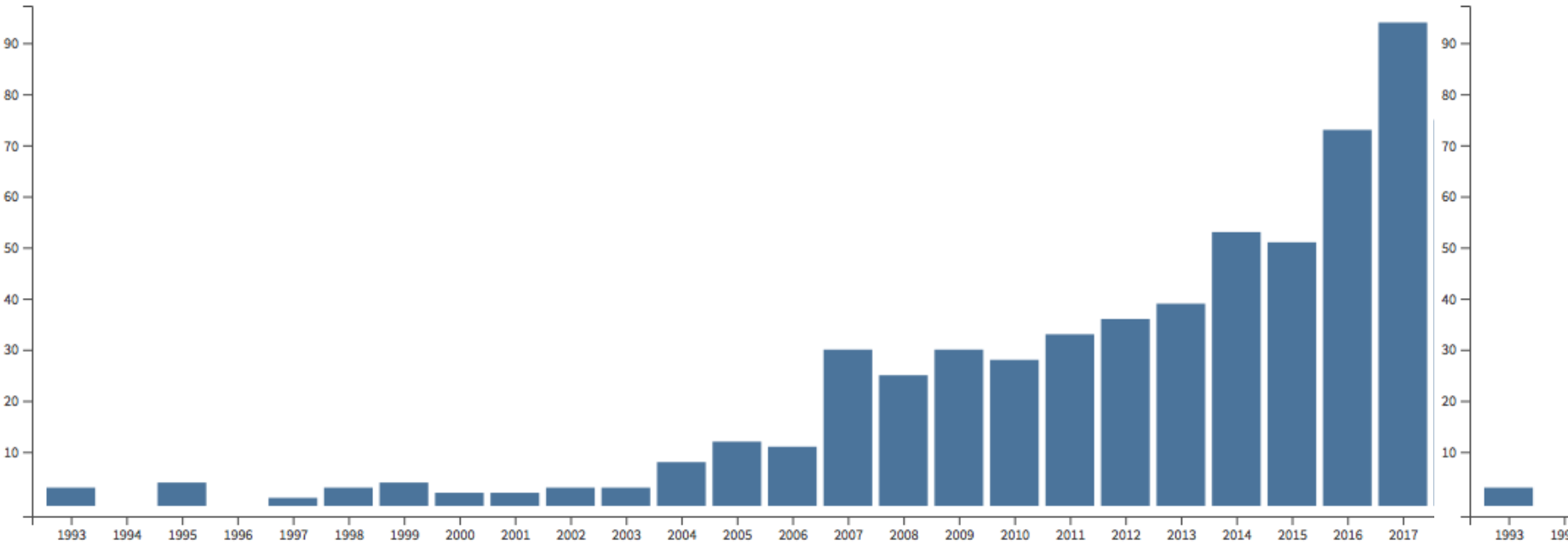


Figure 15: NYT: RL for engagement, 2018

3 types: RL impact in games + commerce + health



3 types: “bandits”=stateless, simplest RL

how to optimize from data: slow

- explore then exploit
  - R.A. Fisher, 1925
  - in industry: "A/B test" (aka MVT, RCT...)
- downsides: requires a meeting, time, regret (unnecessary patients w/worse treatment, etc.)
  - "go with winner"...
  - ...but not until p-value met,
  - (warning: Wasserstein, Ronald L., and Nicole A. Lazar. "The ASA's statement on p-values: context, process, and purpose." The American Statistician 70, no. 2 (2016): 129-133.)

how to optimize from data: fast

3 common approaches:

- iterate the above: epsilon greedy bandit:
  - $p(a) = \epsilon/K + (1-\epsilon)[a = \text{argmax}_{a'} \hat{\mu}_{a'}]$
- OR cautious optimism: UCB (2002) bandit, e.g.,
  - $p(a) = [a = \text{argmax}_{a'} \left(\mu_{a'} + \sqrt{\frac{2 \ln n}{n_{a'}}}\right)]$
- OR just do what you think is best, and learn (1933/TS)
  - $p(a) = p(a = \text{argmax}_{a'} \mu_{a'} | D)$
  - requires 2 models: $p(y|a, \vartheta) = R(y|a, \vartheta)$ and $p(\vartheta)$
  - Note: $\mu_a := \sum_y y R(y|a, \vartheta)$, not empirical $\hat{\mu}_a$

how to optimize: do what you think is best (1933)

ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES.

By WILLIAM R. THOMPSON. From the Department of Pathology, Yale University.

Section 1.

In elaborating the relations of the present communication interest was not centred upon the interpretation of particular data, but grew out of a general interest in problems of research planning. From this point of view there can be no objection to the use of data, however meagre, as a guide to action required before more can be collected; although serious objection can otherwise be raised to argument based upon a small number of observations. Indeed, the fact that such objection can never be eliminated entirely—no matter how great the number of observations—suggested the possible value of seeking other modes of operation than that of taking a large number of observations before analysis or any attempt to direct our course. This problem is more general than that treated in *Section* 2, and is directly concerned with any case where probability criteria may be established by means of which we judge whether one mode of operation is *better* than another in some given sense or not.

how to optimize: do what you think is best (1933)

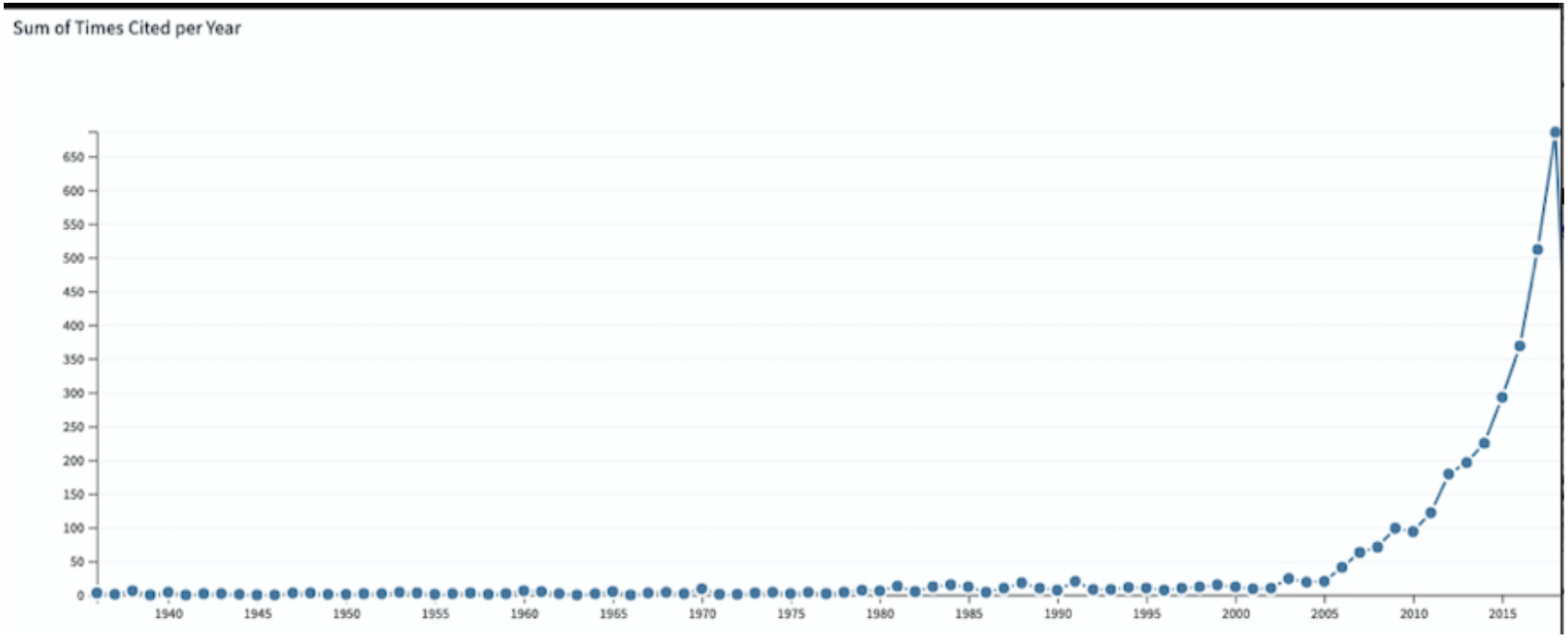


Figure 18: citations

how to optimize: do what you think is best (1933)

given $p(\vartheta)$ and $R(y|a,\vartheta)$; denote $D := \{a_{1:t}, y_{1:t}\}$

$$
\begin{aligned}
p(a) &= p(a = \operatorname{argmax}_{a'} \mu_{a'} | D) && (1)\\
&= \int d\vartheta [a = \operatorname{argmax}_{a'} \mu_{a'}(\vartheta)] p(\vartheta|D) && (2)
\end{aligned}
$$

equivalently:

1. draw $\vartheta \sim p(\vartheta|D) \propto p(D|\vartheta)p(\vartheta)$
2. set $a = \operatorname{argmax}_{a'} \mu_{a'}(\vartheta) := \operatorname{argmax}_{a'} \sum_y yR(y|a',\vartheta)$

(and repeat)

how to optimize: do what you think is best, example

"Bernoulli" case, with data=counts of successes S_a & failures F_a for each arm

- $D = \{S_a, F_a\}$
- $y \in \{0, 1\}$
- $R(y|a, \vartheta_{a=1:K}) \propto \vartheta_a^y (1 - \vartheta_a)^{(1-y)}$
- Prior $p(\vartheta_a) \propto \vartheta_a^{\alpha-1}(1 - \vartheta_a)^{\beta-1}$

Repeat:

1. sample $\vartheta_a \sim p(D, \vartheta_a) \propto \vartheta_a^{S_a+\alpha-1}(1 - \vartheta_a)^{F_a+\beta-1}$.
2. $a = \mathrm{argmax}_{a'} \mu_{a'} = \mathrm{argmax}_{a'} \vartheta_{a'}$
3. world increments either $S_a = S_a + 1$ or $F_a = F_a + 1$