Superintelligence

Swiss Study Foundation

Kaspar Etter, kaspar.etter@gbs-schweiz.org
Adrian Hutter, adrian.hutter@gbs-schweiz.org

Robin Li, Bill Gates, Elon Musk on AI

The Huge Challenge of Artificial Intelligence
www.youtube.com/watch?v=vHzJ_AJ34uQ

More Money for Entertainment

… than ensuring a good outcome!

Ex Machina Movie
exmachina-movie.com

Intelligence and Rationality

What are we talking about?

What is Intelligence?

«Innumerable tests are available for measuring intelligence, yet no one is quite certain of what intelligence is, or even just what it is that the available tests are measuring.» – R. L. Gregor

«The capacity to learn or to profit by experience.» – W. F. Dearborn

Intelligence
en.wikipedia.org/wiki/Intelligence

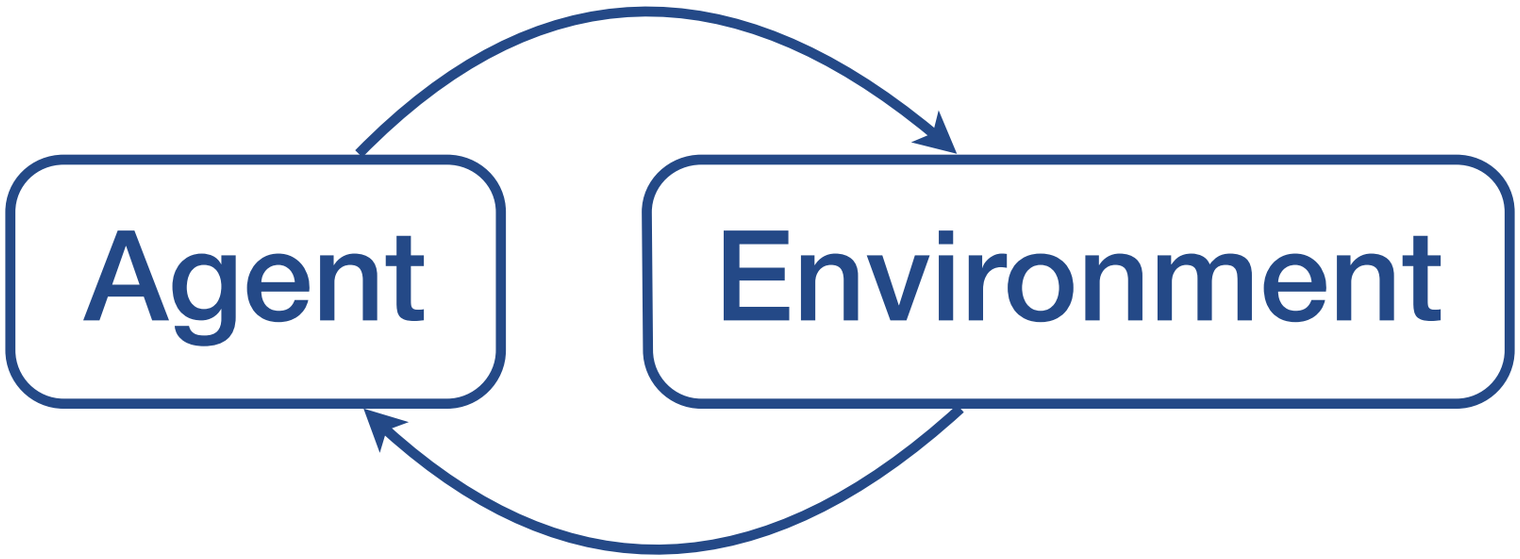
A Useful Definition

«Intelligence measures an agent's ability to achieve its goals in a wide range of unknown environments.»

$$\text{Intelligence} = \frac{\text{Optimization Power}}{\text{Comp. Resources}}$$

Universal Intelligence
arxiv.org/pdf/0712.3329.pdf

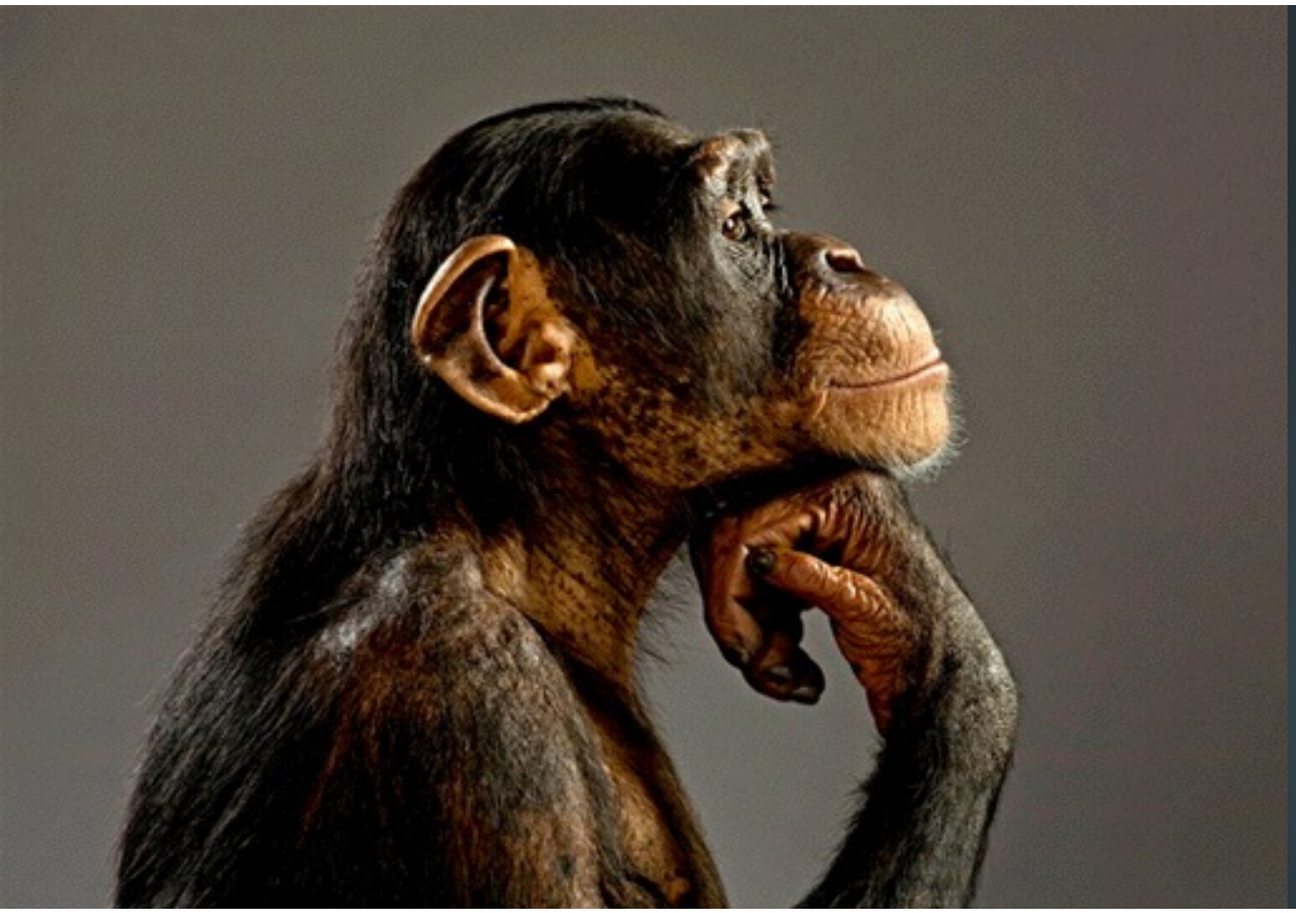
Required Ingredients



Learn, predict, rate and plan!

Marcus Hutter: Universal Artificial Intelligence
www.youtube.com/watch?v=I-vx5zbOOXI

Intelligence is a Big Deal



6 million years ago, 96% common DNA

Fast-Evolving Human DNA Leads to Bigger-Brained Mice
phenomena.nationalgeographic.com/2015/[…]

Technology = (Neutral) Lever

Greed and laziness drive us to

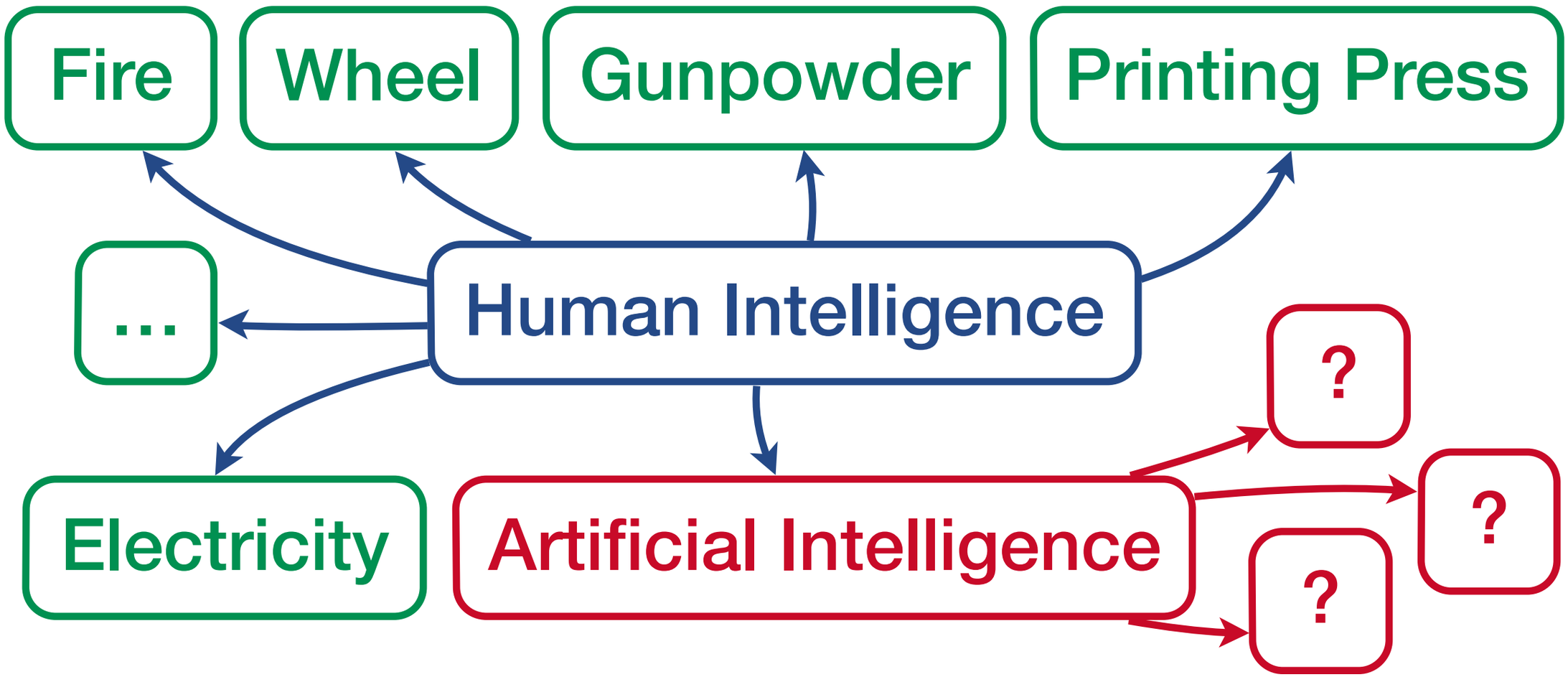
- increase our **productivity**
- make our **lives easier**

… which is fine except:

Our technological progress far outperforms our moral progress!

Dual-Use Technology
en.wikipedia.org/wiki/Dual-use_technology

Artificial Intelligence



Intelligence is a technology like no other!

Rationality

It's about **achieving** your goals!

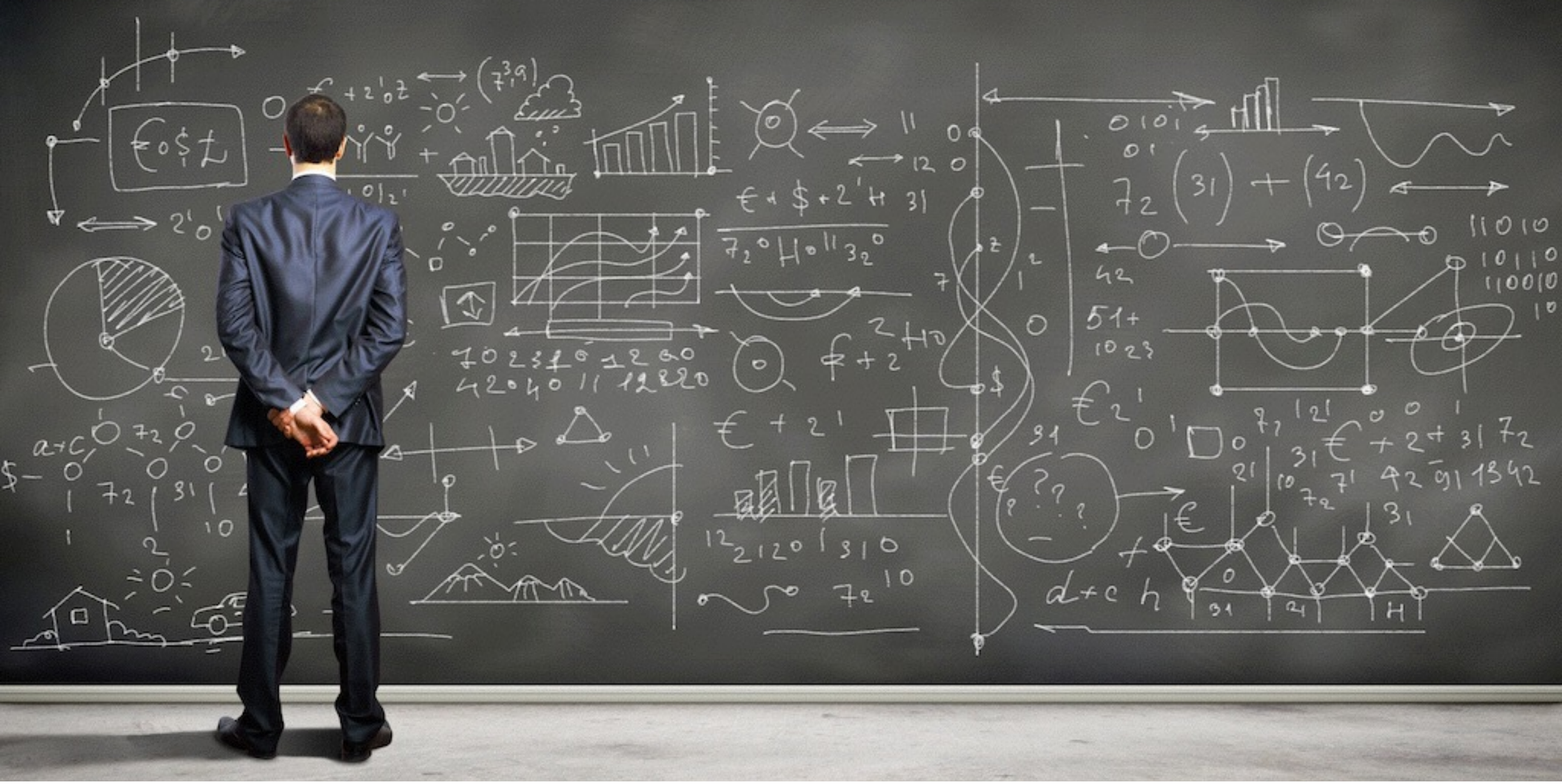
Rationality is the science of winning:

- **Epistemic rationality**: accurate beliefs
- **Instrumental rationality**: good strategy

What do we mean by "rationality"?
lesswrong.com/lw/31/what_do_we_mean_by_[…]

Rationality Continued

- **Normative Rationality**: probability theory, decision theory, game theory
- **Descriptive Rationality**: cognitive biases, system 1 and system 2
- **Prescriptive Rationality**: applied rationality; train:



Probability Theory

What can we know about the world?

What are Probabilities?

Probabilities can be interpreted as ...

- **the frequency** with which an event occurs if an experiment is repeated a large number of times (**Frequentist**)
- or as **the credence** in a statement (**Bayesian** or **subjective** probability).

What is the P that the 9^{th} digit of pi is 7?

Probability Interpretations
en.wikipedia.org/wiki/Probability_interpretations

Relevance of Probabilities

What bets should we take, given our uncertainty about the world we live in?

Living is acting, acting is betting!

Prediction Market
en.wikipedia.org/wiki/Prediction_market

Properties of Probabilities

- For all propositions p: $0 \leq P(p) \leq 1$
- If p is certainly true: $P(p) = 1$
- If p and q are mutually exclusive: $P(p \cup q) = P(p) + P(q)$
- Consequences: $P(\neg p) = 1 - P(p)$
- And: $P(p \cup \neg p) = P(p) + P(\neg p) = 1$

Probability Axioms
en.wikipedia.org/wiki/Probability_axioms

Conditional Probabilities

P(A|B) is the probability of A **given** B.

Example:

Alice has two children. You learn either

a) that the **older child** is a girl – or

b) that at least **one of them** is a girl.

What is the P of **two girls** in a) and b)?

Conditional Probability
en.wikipedia.org/wiki/Conditional_probability

Joint Probabilities

P(A, B): probability that A **and** B occur

P(A, B) **not determined** by P(A) & P(B):

A = “she voted for SP”, B = “… for SVP”

P(A) = 27%, P(B) = 19%, P(A, B) = 0%

A = “older than 55”, B = “older than 65”

P(A) = 27%, P(B) = 19%, P(A, B) = 19%

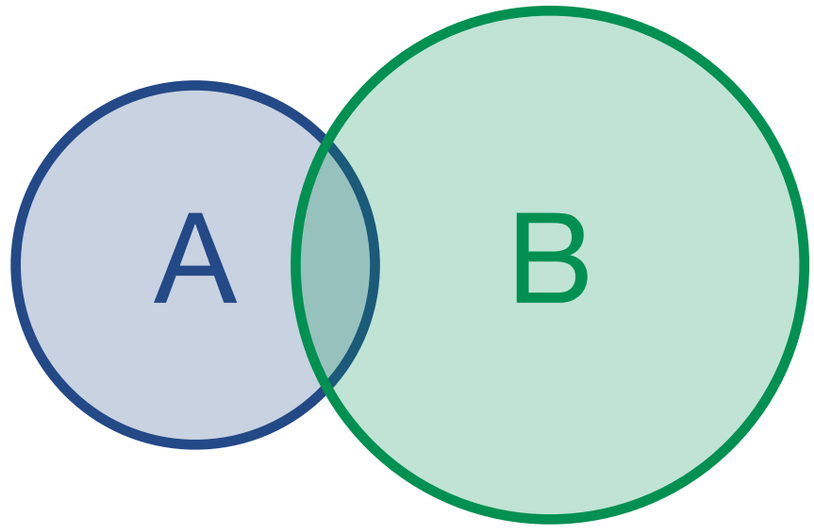
Joint Probability Distribution
en.wikipedia.org/wiki/Joint_probability_distribution

Joint and Conditional Probabil.

$$P(A, B) = P(A \mid B) * P(B)$$

$$P(B, A) = P(B \mid A) * P(A)$$

$$P(A, B) = P(B, A)$$



$$\boxed{P(A \mid B) * P(B) = P(B \mid A) * P(A)}$$

Euler Diagram
en.wikipedia.org/wiki/Euler_diagram

Bayes' Theorem

- Thomas Bayes
- 1701 – 1761
- English Statistician

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Bayes' Theorem
en.wikipedia.org/wiki/Bayes%27_theorem

Testing for a Disease

- 1% of people have a certain disease
- 80% with disease are tested positively
- 9.6% w/o disease are tested positively
- Given a positive test, how likely is “it”?
- Out of 100 people, 1 are ill, 99 aren’t
- Given positive test, only with 7.8% ill

Bayes’ Rule
en.wikipedia.org/wiki/Bayes%27_rule

Bayesian Inference

How to update your beliefs:

$$P_{new}(h) = P_{old}(h) \cdot \frac{P_{old}(e|h)}{P_{old}(e)}$$

The confidence in a **hypothesis h** increases if the **evidence e** is more likely to happen given h than without.

Bayesian Inference
en.wikipedia.org/wiki/Bayesian_inference

Base Rate Fallacy

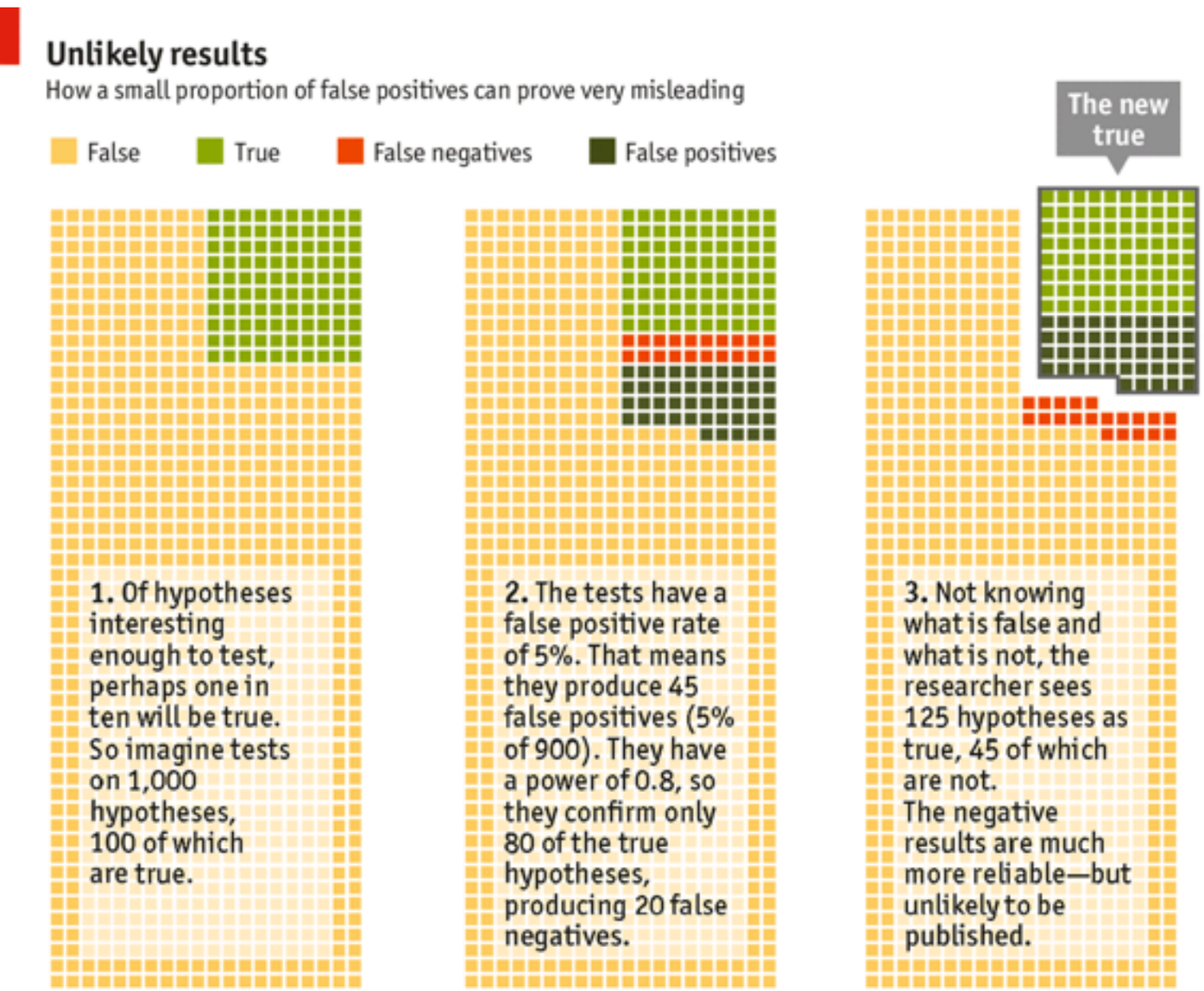
We tend to replace P(h|e) with P(e|h) instead of updating the **prior belief!**

$$P_{new}(h) = P_{old}(h) \cdot \frac{P_{old}(e|h)}{P_{old}(e)}$$

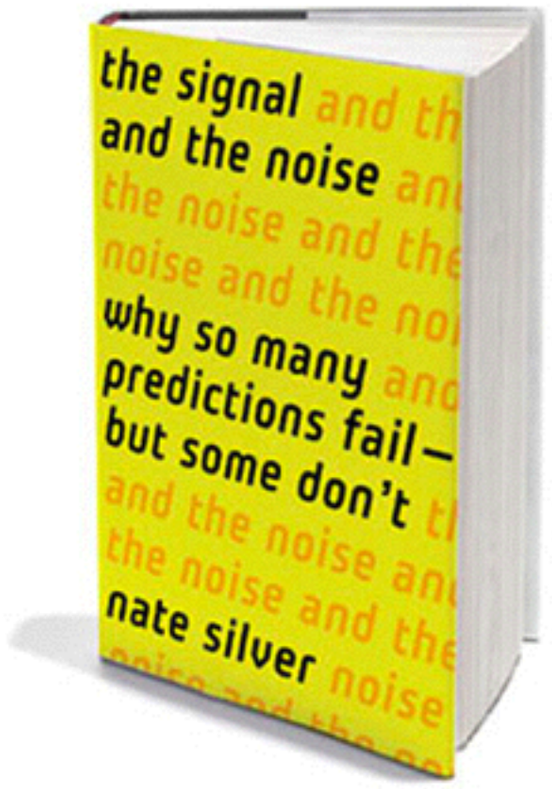
Extraordinary claims require extraordinary evidence!

Base Rate Fallacy
en.wikipedia.org/wiki/Base_rate_fallacy

Broken Science



Source: *The Economist*



Null-Hypothesis Testing: P(e|h) instead of P(h|e)!

Trouble at the Lab
econ.st/19QXzFe

Conservation of Exp. Evidence

$$P(h) = P(h \cap e) + P(h \cap \overline{e})$$

$$P(h) = P(h|e) \cdot P(e) + P(h|\overline{e}) \cdot P(\overline{e})$$

$$P(e) + P(\overline{e}) = 1$$

$$\Rightarrow P(h|\overline{e}) \leq P(h) \leq P(h|e)$$

Absence of evidence is evidence of absence!

Prior Probability

Where do we get the priors from?

- **Irrelevant** given enough evidence
- **Knowledge** about the statistics
- **Symmetries** (even distribution)

But:

- What is the prior that Zeus exists?
- The prior that the universe is finite?

Solomonoff Induction

Ray Solomonoff, 1960: **Universal Prior**

Combines and formalizes ideas by

- **Epicurus**: “Keep all hypotheses that are consistent with the data.”
- **Ockham**: “Among all hypotheses consistent with the observations, choose the simplest.” (O.’s Razor)

An Intuitive Explanation of Solomonoff Induction
lesswrong.com/lw/dhg/an_intuitive_explanation_of_[…]

Kolmogorov Complexity

Simplicity: Not how easy something is to understand for humans – but rather the length of the shortest program that is able to reproduce the observations.

The **Solomonoff prior** is exponentially small in this length & not computable. (Formal def. involves *Turing machines*.)

Kolmogorov Complexity
en.wikipedia.org/wiki/Kolmogorov_complexity

Decision Theory (Part 1)

How shall we decide in the face of uncertainty?

Utility Function

A utility function describes an agent's **preferences** over different outcomes.

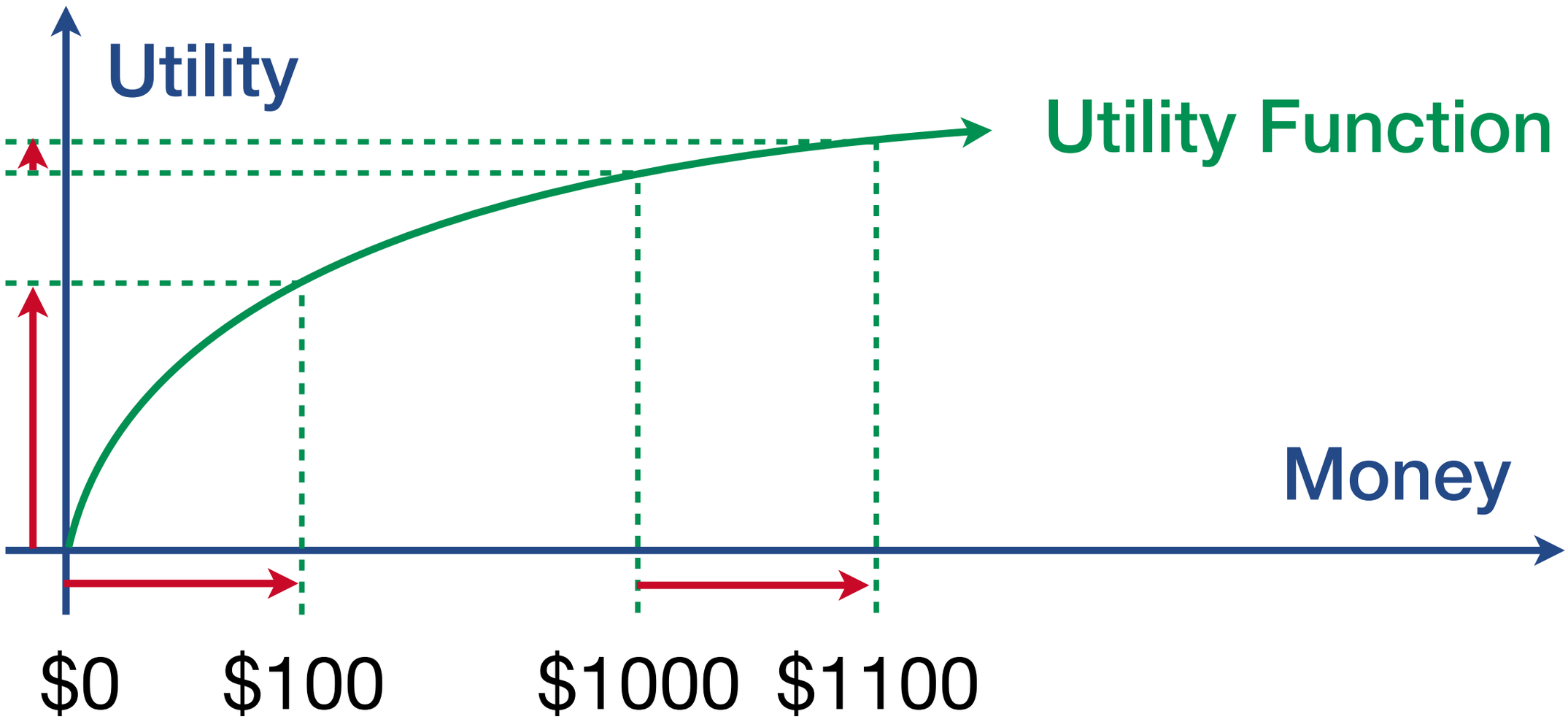
Being nourished is a state of **higher utility** for animals than being starving.

A utility function doesn't have to be an explicit or **conscious goal** of an agent.

Utility
en.wikipedia.org/wiki/Utility

Marginal Utility

Money has diminishing marginal utility:



Marginal Utility
en.wikipedia.org/wiki/Marginal_utility

Expected Utility

Under **uncertainty** (i.e. in reality) we can only maximize expected utilities.

If there are 3 outcomes with **utilities** u_1, u_2, u_3 and **probabilities** p_1, p_2, p_3, the expected utility is calculated as $p_1 * u_1 + p_2 * u_2 + p_3 * u_3$.

Expected Utility Hypothesis
en.wikipedia.org/wiki/Expected_utility_hypothesis

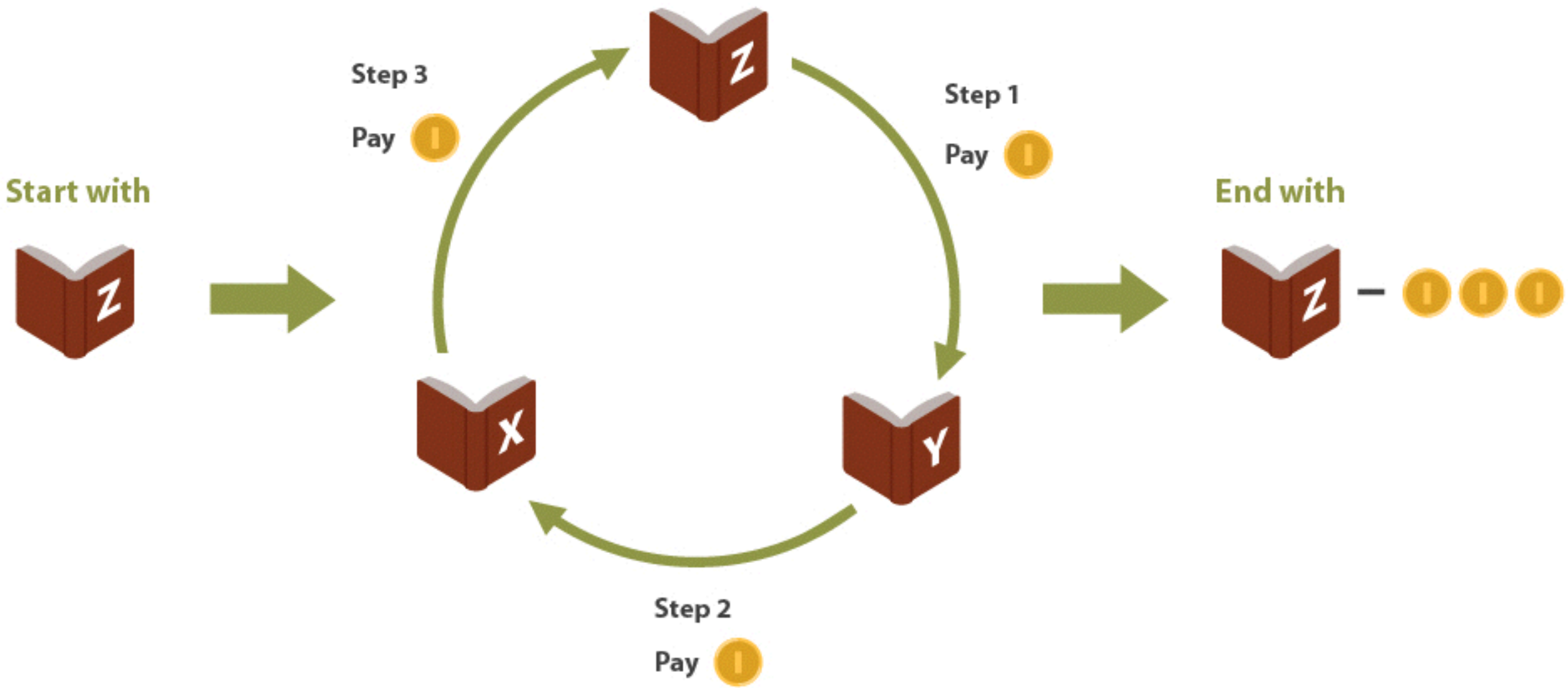
Von Neumann–Morgenstern

Von Neumann and Morgenstern, 1947: Every *rational* agent (i.e. one satisfying four natural axioms) behaves *as if* it tries to maximize the expected utility of some utility function.

Axioms: **Completeness**, **Transitivity**, **Continuity** and **Independence** (VNM).

Von Neumann-Morgernstern Utility Theorem
en.wikipedia.org/wiki/Von_Neumann–Morgenstern_[…]

Money-Pump Argument



Decision Theory FAQ
lesswrong.com/lw/gu1/decision_theory_faq/

Loss Aversion

You have been given $1000. Choose:

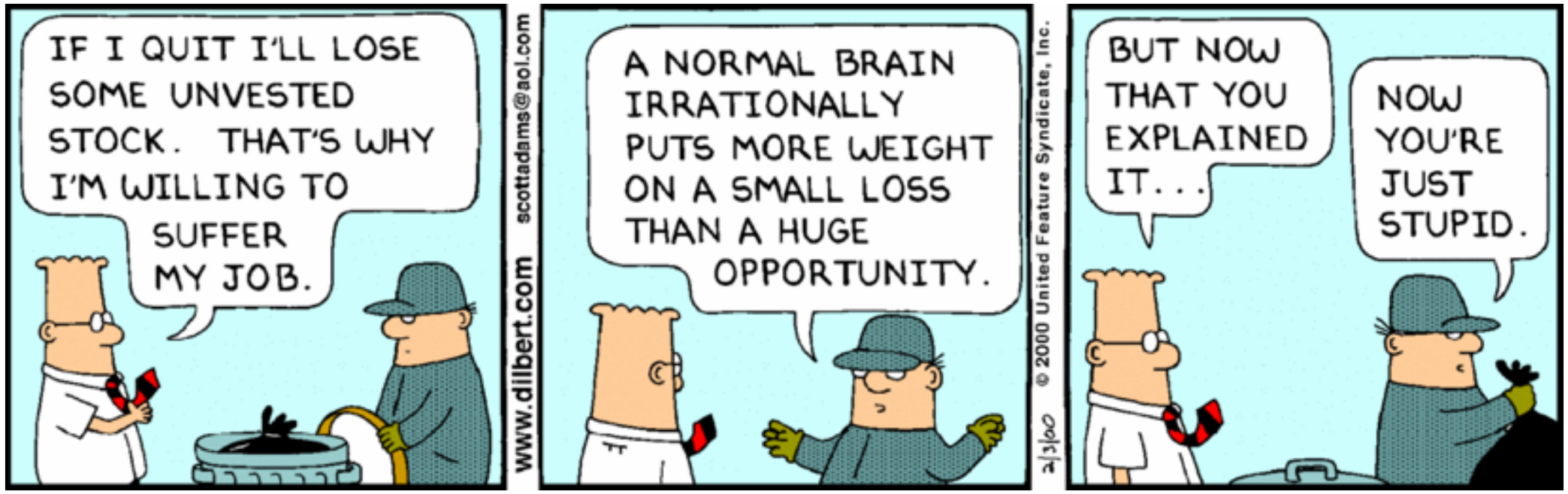
- Win $1000 with 50% (risky)
- Win $500 with certainty (safe)

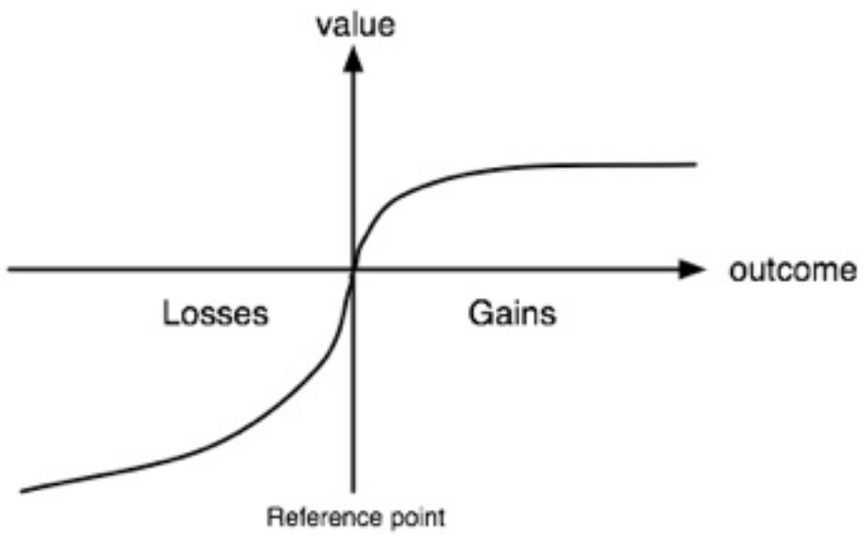
You have been given $2000. Choose:

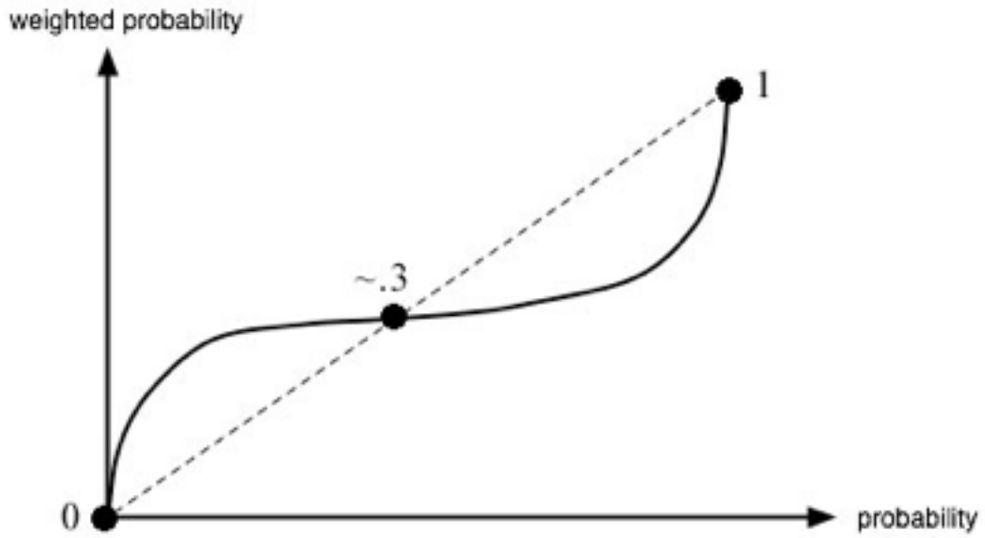
- Lose $1000 with 50% (risky)
- Lose $500 with certainty (safe)

TED Talk by Laurie Santos
www.ted.com/talks/laurie_santos.html

Prospect Theory







Prospect Theory
en.wikipedia.org/wiki/Prospect_theory

Framing Effect

Tversky/Kahneman:
A: 72%, B: 28%
C: 22%, D: 78%

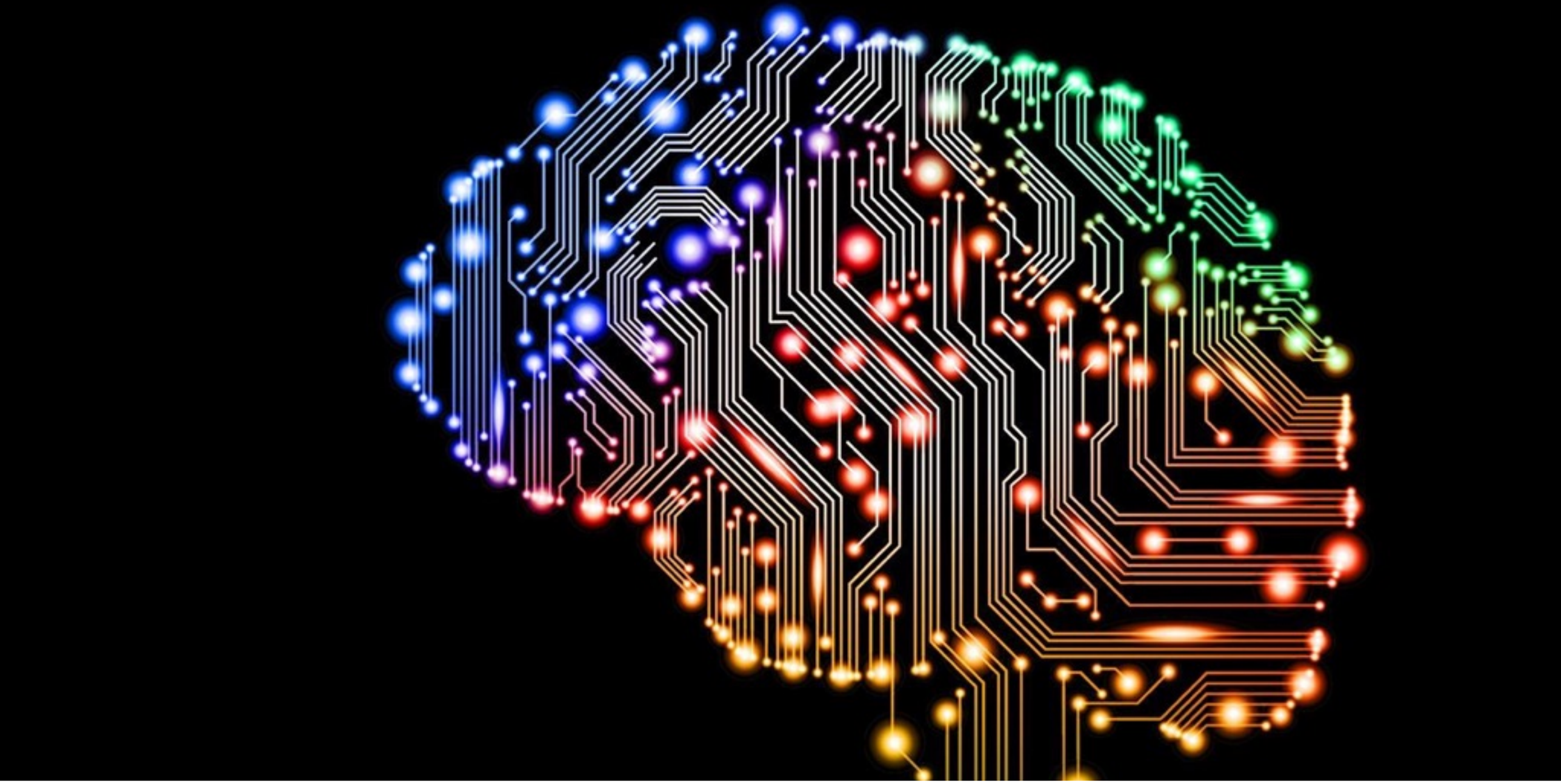
Outbreak of disease, choose between:

- A: 200 out of 600 people will be saved
- B: ⅓ pr. of saving everyone, ⅔ no one

- C: 400 out of 600 people will die
- D: ⅓ prob. nobody will die, ⅔ 600 die

Don’t trust your (moral) intuitions!

Framing (Social Sciences)
en.wikipedia.org/wiki/Framing_(social_sciences)



Universal Intelligence

Can we define an optimally intelligent agent?

Swiss AI Lab in Lugano

World-leading in **pattern recognition** with Artificial Neural Networks (ANNs).

Theory of **optimally intelligent agents:**

- AIXI (Marcus Hutter)
- Gödel Machine (Jürgen Schmidhuber)

Former PhD student co-founded Deep-Mind (sold to Google for $500m, 2014).

An Optimal Agent

Active agents in a **known** environment maximize **expected utility**.

Passive agents in an **unknown** environment use **Solomonoff induction** to get probability distribution over possible environments.

What about **active** agents in **unknown** environments (like reality)?

Intelligent Agent
en.wikipedia.org/wiki/Intelligent_agent

Universal Intelligence: AIXI

Developed by Marcus Hutter in 2000

$$a_k := \arg\max_{a_k} \sum_{o_k r_k} \ldots \max_{a_m} \sum_{o_m r_m} [r_k + \ldots + r_m] \sum_{q : U(q, a_1 .. a_m) = o_1 r_1 .. o_m r_m} 2^{-\ell(q)}$$

At each step, update your probability distribution over **all possible worlds** (Bayes/Solomonoff) and choose the action which maximizes expected utility over **all remaining steps.**

Universal Algorithmic Intelligence
www.hutter1.net/ai/aixigentle.htm

Monte Carlo AIXI

It is the most intelligent agent possible.

Like the Solomonoff prior, AIXI **cannot be computed** but only approximated.

AIXI is *not* intended as a proposal for building an AI, but as an **upper bound** on how intelligent an agent can be.

A Monte Carlo AIXI Approximation
arxiv.org/abs/0909.0801

Shortcomings of AIXI

Cartesian model: The agent and the environment are modelled as separate Turing machines. In reality, the agent is part of the world in which it lives.

Wireheading: The most intelligent thing to do for a reinforcement learner is to get control over its reward-channel.

Universal Algorithmic Intelligence
arxiv.org/abs/cs/0701125

Gödel Machine

Named after Kurt Gödel (1906–1978), developed by J. Schmidhuber (2003).

Mathematically rigorous, general, fully self-referential, self-improving, optimally efficient **problem solver.**

The GM solves all large enough problems almost as quickly as if it already knew the best (unknown) algorithm for solving them.

Gödel Machine by Jürgen Schmidhuber
people.idsia.ch/~juergen/goedelmachine.html

Decision Theory (Part 2)

What kind of problems can we run into?

Complications

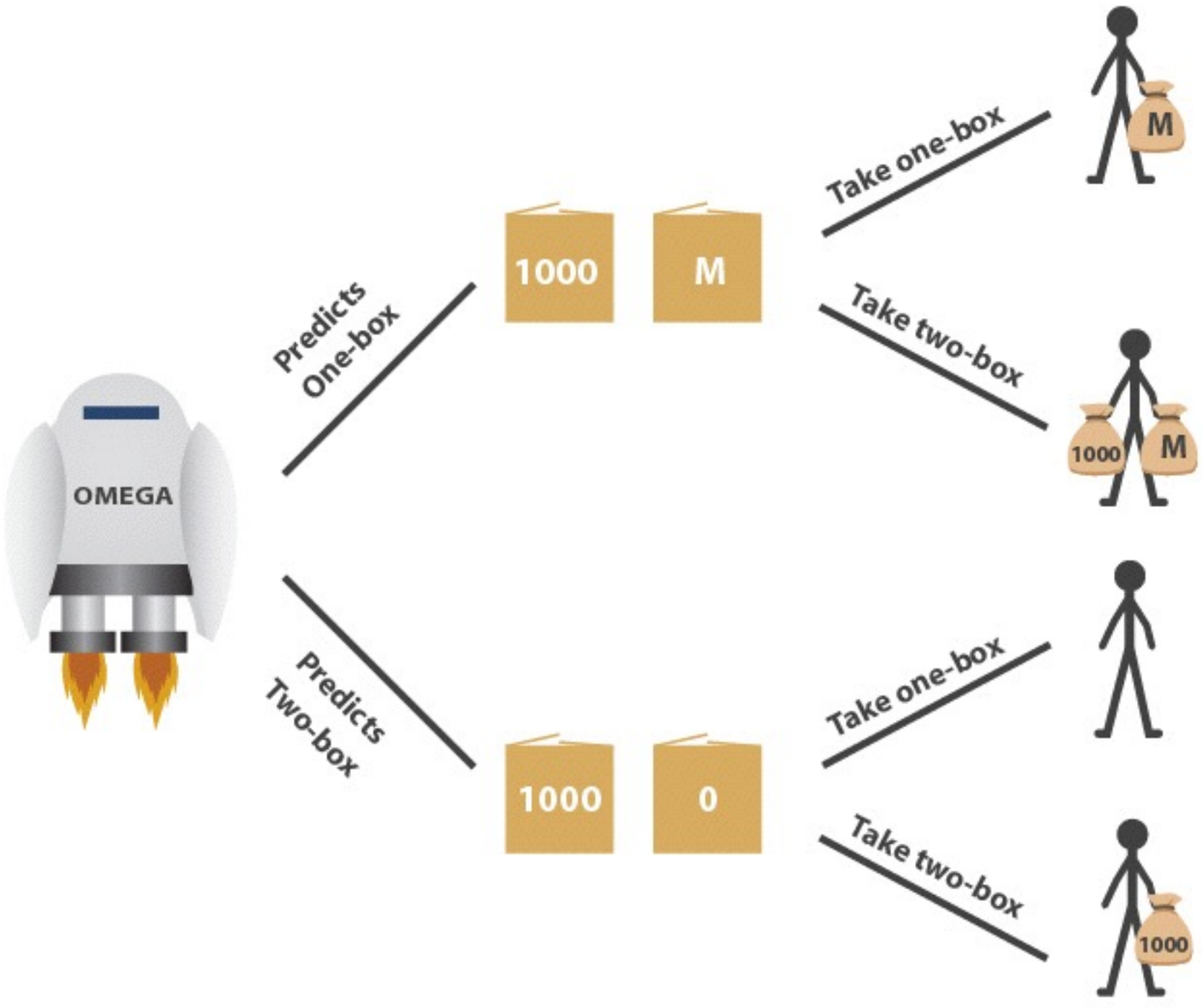
What if the payout-structure itself depends on the action you take?

Causal Decision Theory: Take the action that causes the best expected outcome.

Evidential Decision Theory: Choose the action which, *conditional on you having chosen it*, gives you the best e. outcome.

Evidential Decision Theory
en.wikipedia.org/wiki/Evidential_decision_theory

Newcomb’s Paradox



Newcomb’s Paradox
en.wikipedia.org/wiki/Newcomb%27s_paradox

Possible Answers

EDT > CDT?

Causal reasoning: Content of boxes is fixed and not affected by my decision. Taking both gives me more than one.

Evidential reasoning: If I take only one box, the predictor had predicted this and filled it with $1m. If I take both boxes, I will walk away with only $1k.

Newcomb's Problem
wiki.lesswrong.com/wiki/Newcomb's_problem

Where CDT (Arguably) > EDT

«Solomon is an ancient monarch vaguely reminiscent of the Israelite King. (Every part of this story is Biblically inaccurate.) He is pondering whether to summon Bathsheba, another man's wife. But Solomon is also fully informed as to the peculiar connection between his choice in this matter and the likelihood of his eventually suffering a successful revolt: "Kings have two basic personality types, charismatic and uncharismatic. A king's degree of charisma depends on his genetic make-up and early childhood experiences, and cannot be changed in adulthood. Now charismatic kings tend to act justly and uncharismatic kings unjustly. Successful revolts against charismatic kings are rare, whereas successful revolts against uncharismatic kings are frequent. Unjust acts themselves, though, do not cause successful revolts… Solomon does not know whether or not he is charismatic; he does know that it is unjust to send for another man's wife."»

Paradoxes in Probability Theory by William Eckhardt
www.springer.com/us/book/9789400751392

Where Both (Arguably) Fail

- Blackmailing
- Parfit's Hitchhiker
- Counterfactual Mugging

Possible solution: **Updateless DT** ("Do what you would have precommitted.")

EDT & CDT not reflectively consistent.

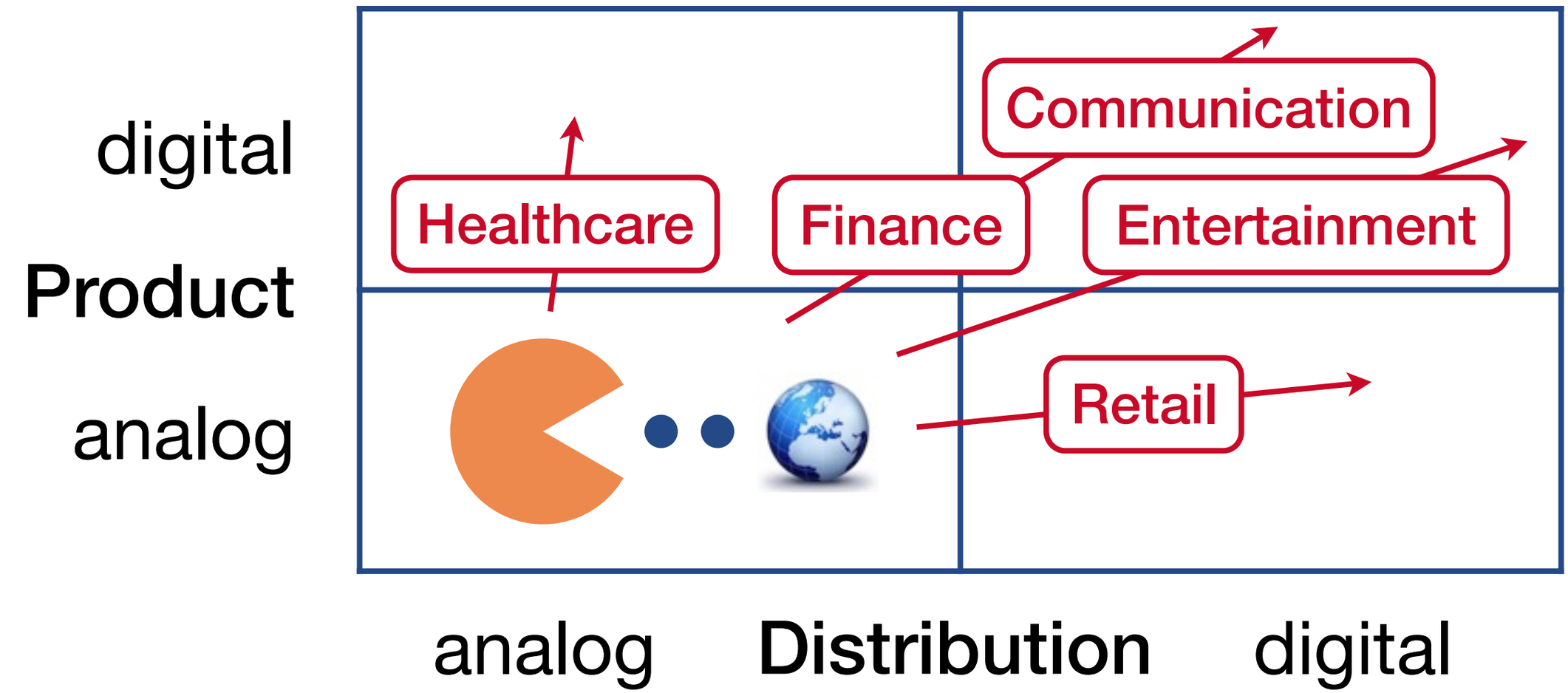
UDT with Known Search Order by Tsvi Benson-Tilsen
intelligence.org/files/UDTSearchOrder.pdf

Machine Learning

How can algorithms separate the signal from the noise?

Software is eating the world

… because it is more productive!



Marc Andreessen
on.wsj.com/1w2FbVs

Machine Intelligence LANDSCAPE

CORE TECHNOLOGIES

ARTIFICIAL INTELLIGENCE
IBM WATSON · MetaMind · Numenta · ai-one · Cycorp · Microsoft Research · nara · Reactor · SI Scaled Inference

DEEP LEARNING
vicarious · Vision Factory · facebook · Google · Baidu IDL · ersatz · SKYMIND · SignalSense

MACHINE LEARNING
rapidminer · context relevant · 0xdata H2O · DATARPM · LiftSigniter · SPARKBEYOND · Azure ML · yhat · wise.io · Sense · GraphLab · Alpine · Ersonian

NLP PLATFORMS
cortical.io · idibon · LUMINOSO · wit.ai · Maluuba

PREDICTIVE APIS
AlchemyAPI · MINDOPS · Google · bigml · indico · ALGORITHMIA · Expect Labs · PredictionIO

IMAGE RECOGNITION
clarifai · MADBITS · DNNresearch · DEXTRO · ViSENZE · lookflow

SPEECH RECOGNITION
GRIDSPACE · popup archive · NUANCE

RETHINKING ENTERPRISE

SALES
Preact · AVISO · RelateIQ · NGDATA · CLARABRIDGE · FRAMED · infer · ATTENSITY · causata

SECURITY / AUTHENTICATION
CROSSMATCH · conjur · EYEVERIFY · BITSIGHT · CYLANCE · AREA 1 SECURITY · bionym

FRAUD DETECTION
sift science · SOCURE · ThreatMetrix · feedzai · BrighterionTM · VERAFIN

HR / RECRUITING
TalentBin · entelo · predikt · Connectifier · gild · hiQ · CONCEPTNODE

MARKETING
brightfunnel · bloomreach · CommandIQ · AIRPR · RADIUS · TellApart · people pattern · Freshplum

PERSONAL ASSISTANT
Siri · Google now · Cortana · cleversense · tempo · Robinlabs · KASISTO · fuse|machines · VIV · CLARA LABS

INTELLIGENCE TOOLS
ADATAD · Palantir · Quid · Digital Reasoning · FirstRain

RETHINKING INDUSTRIES

ADTECH
METAMARKETS · dstillery · rocketfuel · YieldMo · ADBRAIN

AGRICULTURE
BLUE RIVER · TerrAvion · ceresimaging · HONEYCOMB · THE CLIMATE CORPORATION · tule

EDUCATION
declara · coursera · KNEWTON · kidaptive

FINANCE
Bloomberg · FinGenius · alphasense · KENSHO · Dataminr · minettabrook · BINATIX

LEGAL
Lex Machina · brightleaf · COUNSELYTICS · RAVEL · JUDICATA · Brevia · DiligenceEngine

MANUFACTURING
SIGHT MACHINE · MICROSCAN · IVISYS · BOULDER IMAGING

MEDICAL
Parzival · transcriptic · ZEPHYR HEALTH · Genescient · grand round table · bina · TUTE GENOMICS

OIL AND GAS
kaggle · AYASDI · TACHYUS · biota · Flutura

MEDIA / CONTENT
Outbrain · newsle · ARRIA · SAILTHRU · wavii · Owlin · NarrativeScience · yscap · Summly · Prismatic · ai

CONSUMER FINANCE
Affirm · inVenture · zesty.finance · BILL GUARD · LendUp · LendingClub · Kabbage

PHILANTHROPIES
DataKind · thorn · THE DATA GUILD

AUTOMOTIVE
Google · Continental · Tesla · MobileEye · CRUISE

DIAGNOSTICS
enlitic · 3SCAN · lumiata · ENTOPSIS

RETAIL
BAY SENSORS · PRISM SKYLABS · celect · euclid

RETHINKING HUMANS / HCI

AUGMENTED REALITY
wearable intelligence · APX · blippAR · META · layar

GESTURAL COMPUTING
THALMICLABS · omek · FLUTTER · LEAP MOTION · eyeSight · 3Gear · GestureTek · nod

ROBOTICS
intel · LIQUID ROBOTICS · SoftBank · iRobot · Boston Dynamics · jibo · ANKI · evolution robotics

EMOTIONAL RECOGNITION
affectiva · BEYONDVERBAL · EMOTIENT · cogito

SUPPORTING TECHNOLOGIES

HARDWARE
NVIDIA · XILINX · QUALCOMM · NERVANA SYSTEMS · TERADEEP · Artificial Learning · rigetti

DATA PREP
TRIFACTA · Paxata · tamr · Alation

DATA COLLECTION
diffbot · kimono · CrowdFlower · Connotate · WorkFusion · import.io

www.shivonzilis.com/machineintelligence

Bloomberg BETA

Computers have just learned …

… how to see, read, write and classify.

Jeremy Howard
go.ted.com/bbZC

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

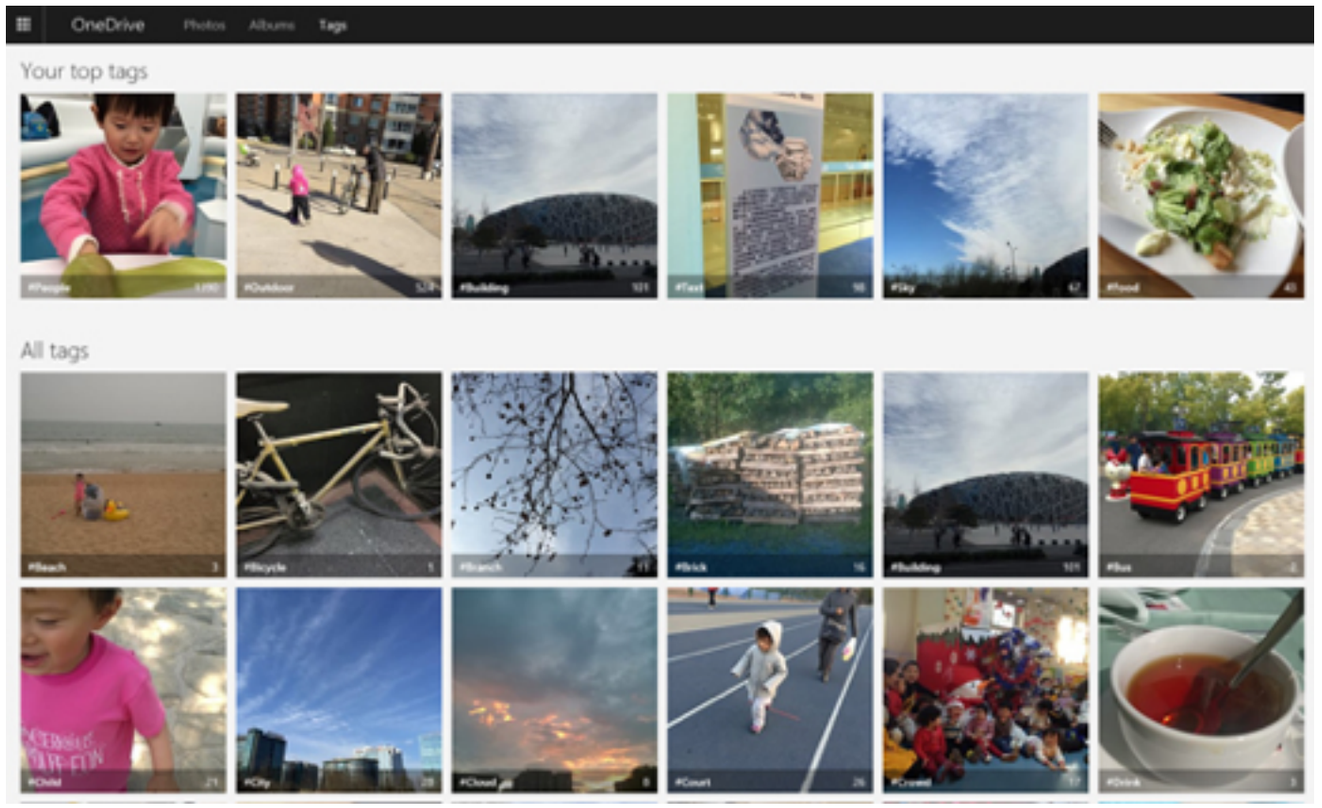
A close up of a cat laying on a couch.

A red motorcorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.

Superhuman Image Recognition

- With convolutional neural networks
- 1.2 m training images, ~ 30 layers



Inside Microsoft Research, February 2015
blogs.technet.com/b/inside_microsoft_research/[…]

Machine

A **machine** runs an **algorithm** which is

- a procedure for
- solving a specified problem
- in a finite number of steps
- (i.e. eventually producing an output).

Turing Machine
en.wikipedia.org/wiki/Turing_machine

Learning

- Building **predictive** models from data
- Algorithms improve with “experience”
- Useful if you don’t know the solution

- **Machine Learning**: Learn and predict
- **Data Mining**: Discover new properties

Machine Learning
en.wikipedia.org/wiki/Machine_learning

Feedback Mechanism

Categorization based on feedback:

- **Supervised Learning**: Given a set of inputs & outputs, learn a general rule
- **Unsupervised Learning**: Learn the structure in input without training set
- **Reinforcement Learning**: Achieve a certain goal in dynamic environment

Supervised Learning
en.wikipedia.org/wiki/Supervised_learning

Tasks and Models

- **Classification**: Assign unseen input to descrete categories (supervised)
- **Regression**: Map the unseen input to continuous values (supervised)
- **Clustering**: Divide set of inputs into a number of groups (unsupervised)
- **Dimensionality Reduction**: Simplify

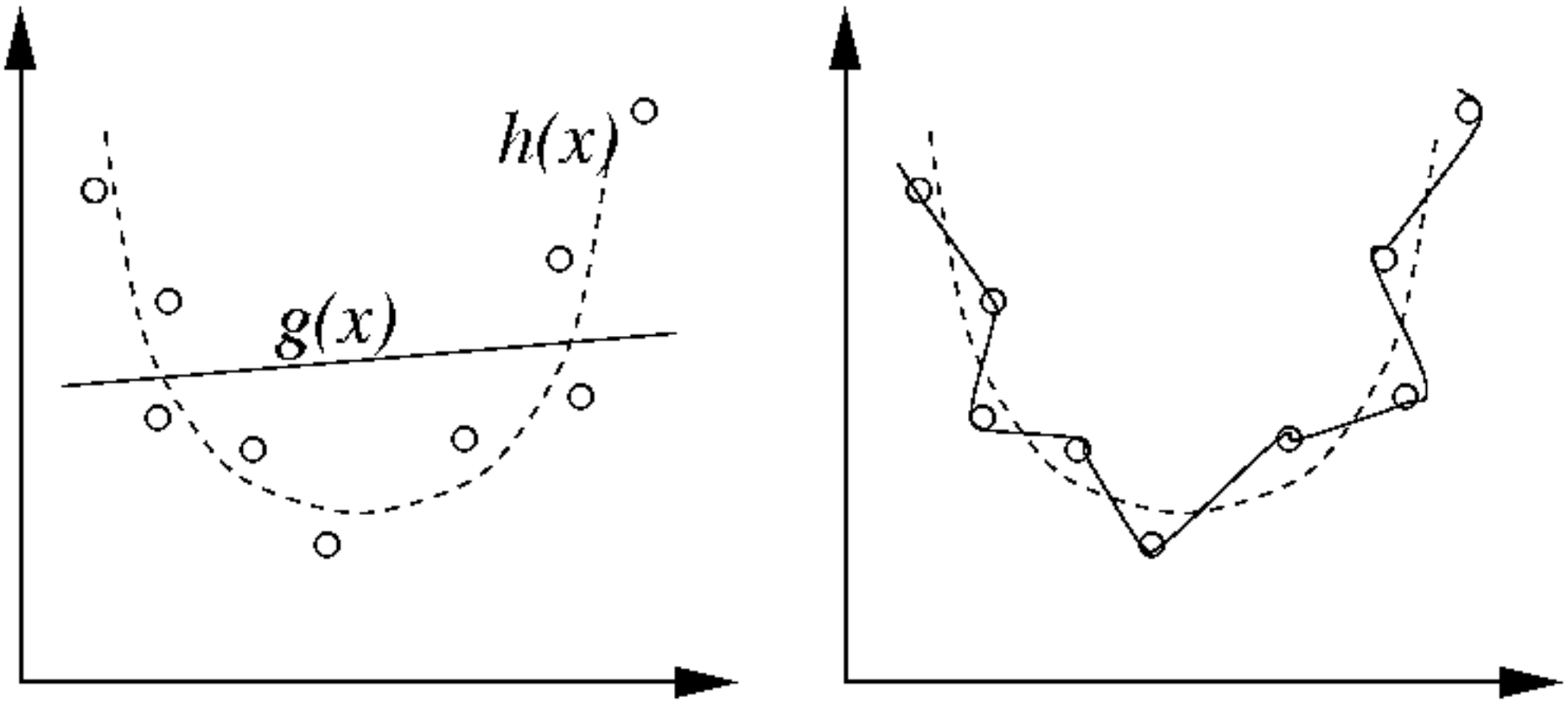
Statistical Classification
en.wikipedia.org/wiki/Statistical_classification

Applications

- Spam filtering
- Search engines
- Computer vision
- Medical diagnosis
- Stock market prediction
- Optical character recognition
- Recognizing credit card fraud
- Speech-recognition (e.g. SIRI)
- Autonomous systems (e.g. self-driving cars)

Six Novel Machine Learning Applications
www.forbes.com/sites/85broads/2014/01/06/[…]

Problem: Overfitting



Overfitting
en.wikipedia.org/wiki/Overfitting

Cross-Validation

- Helps to prevent overfitting (i.e. find the **optimal number of parameters**).
- Helps to estimate how good a model will **generalize to unseen data.**

Partition data into a "training set" and a "validation set".

Cross-Validation (Statistics)
en.wikipedia.org/wiki/Cross-validation_(statistics)

Exploration-Eploitation Trade-Off

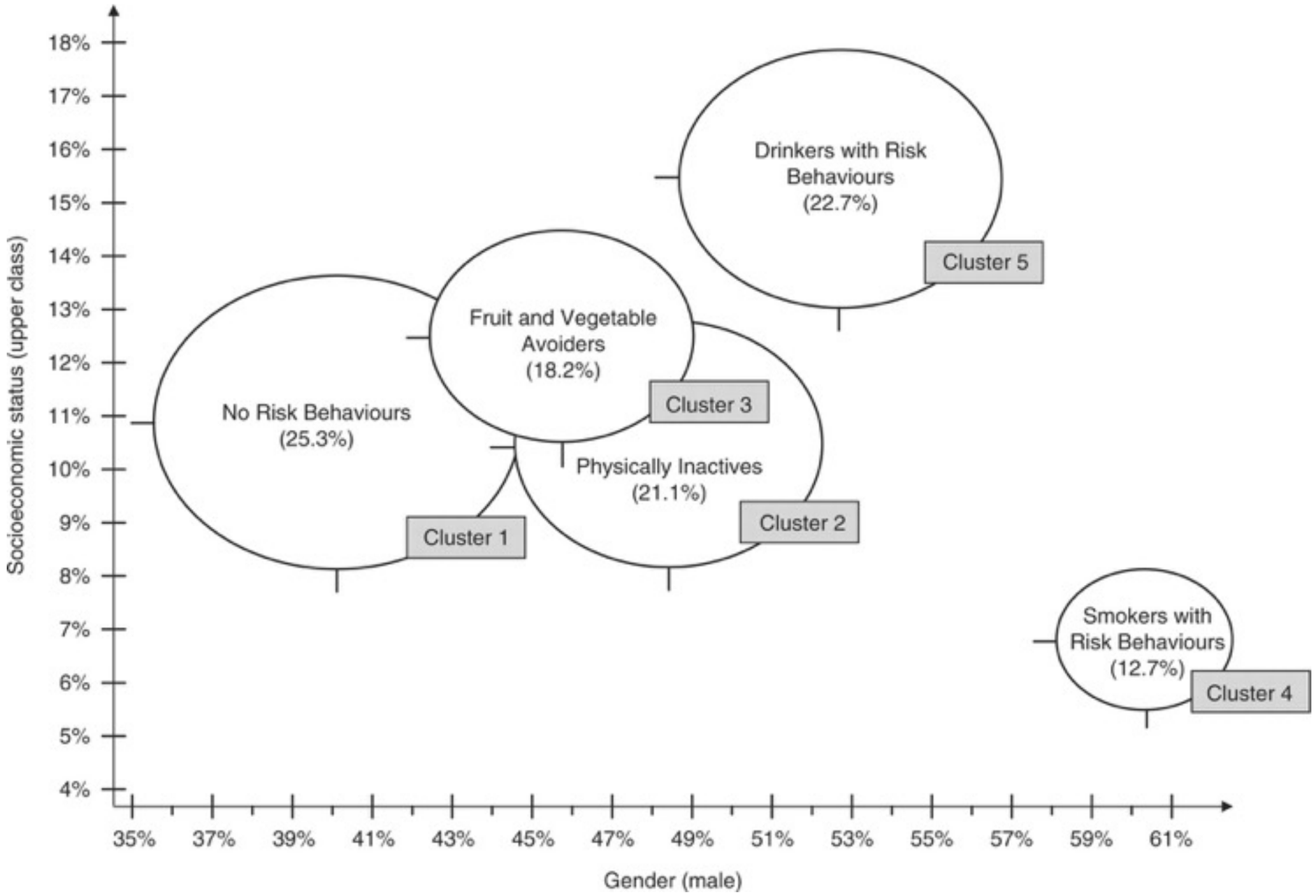
You have 1000 gambling machines in front of you, which give different pay-outs with different prob. distributions.

You have 1000 trials for all machines.

How many of your trials do you spend finding a good machine, how many to exploit the best machine found so far?

Reinforcement Learning
en.wikipedia.org/wiki/Reinforcement_learning

Clustering: Example



Health Behaviour Patterns in a German 50+ survey
www.ncbi.nlm.nih.gov/pubmed/19164433

Clustering: k-means Algorithm

How to find **k clusters** in N data points?

Goal: Minimize the sum of the **distances** of each point to the **center** of its cluster.

Solving this problem exactly is hard (NP).

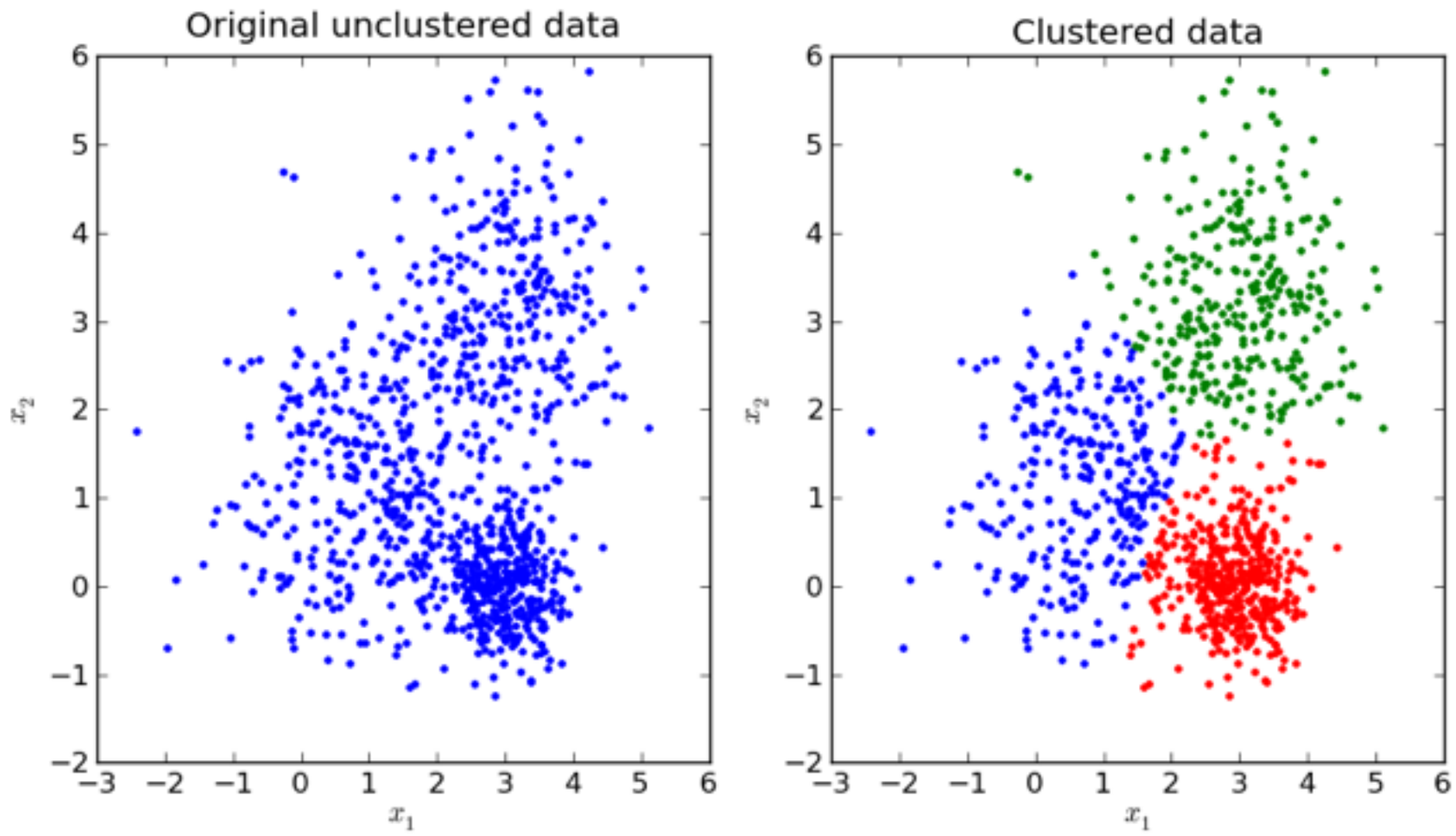
Choose k of the N points randomly.

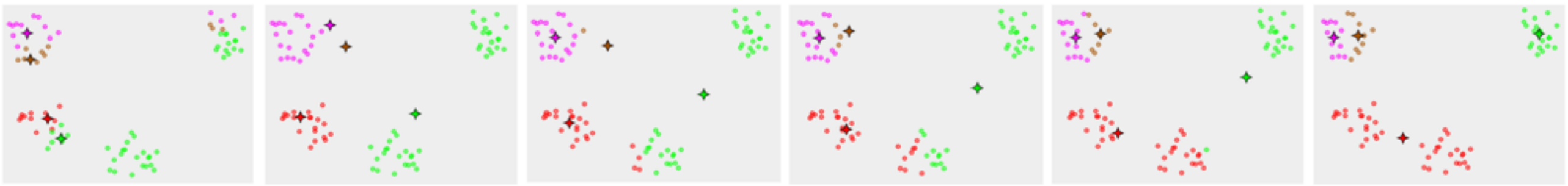
Assign each point to the nearest mean.

Calculate the new means of each cluster.

k-means Clustering
en.wikipedia.org/wiki/K-means_clustering

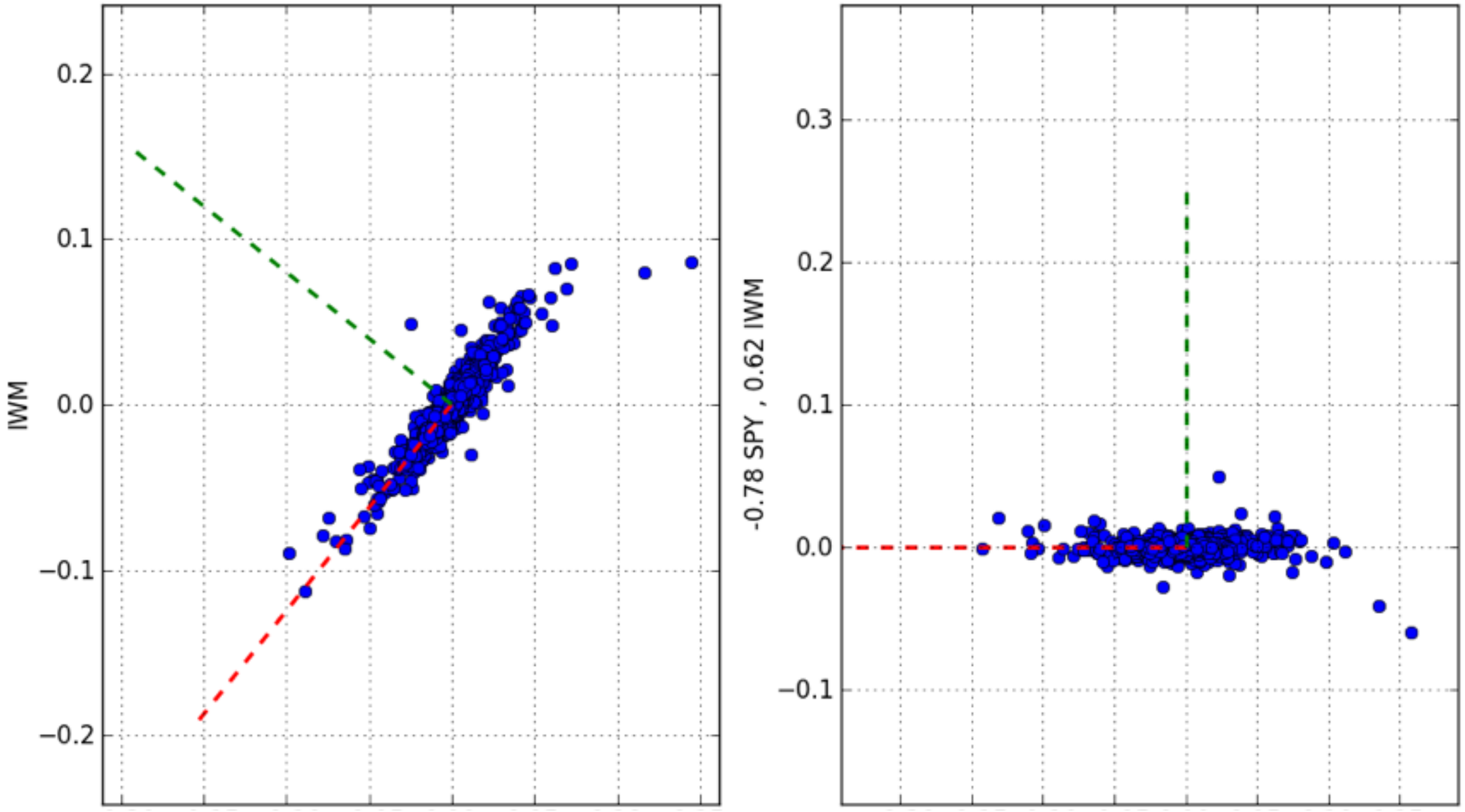
k-means Clustering Outcomes





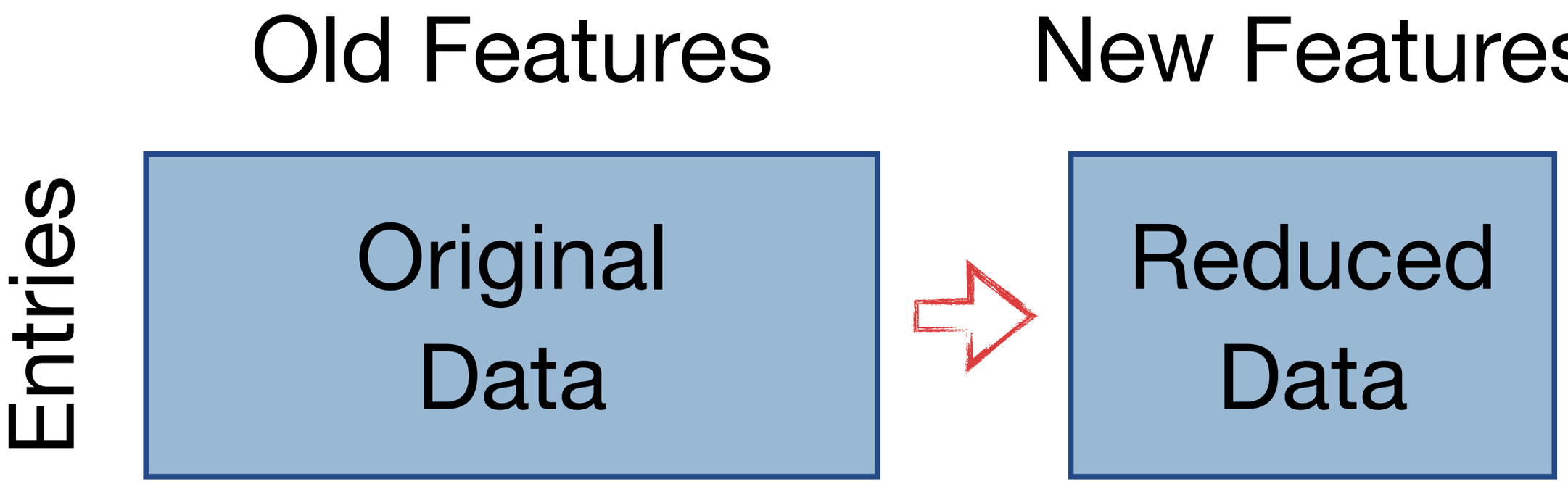
k-means Clustering
en.wikipedia.org/wiki/K-means_clustering

Dimensionality Reduction



Dimensionality Reduction
en.wikipedia.org/wiki/Dimensionality_reduction

Goal: Reduce Dimensionality



Retain as much information as possible with as little information as possible!

Principal Component Analysis
en.wikipedia.org/wiki/Principal_component_analysis

Recommendation Systems

Example: If each song is a dimension, a user's taste can be viewed as vector.

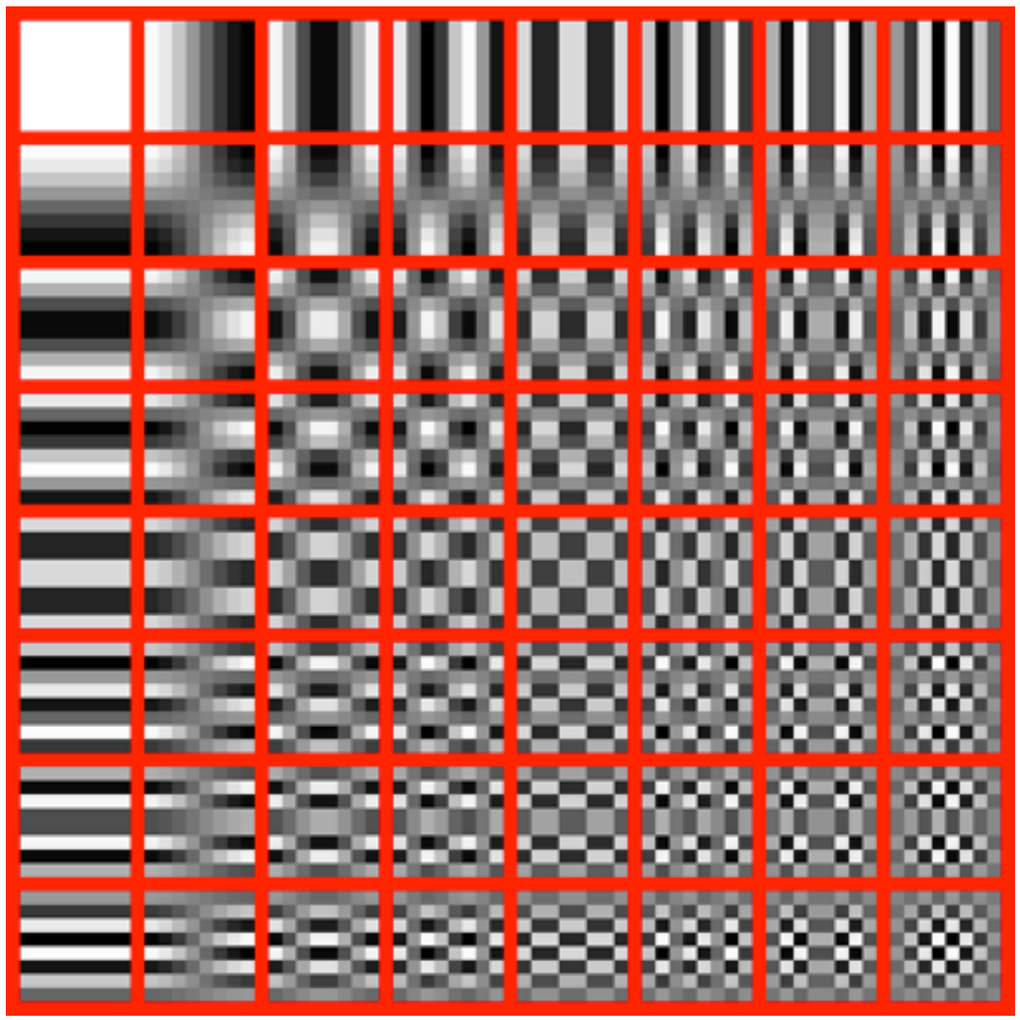
Observation: As groups of users rate the same kind of songs high and low, try to distill the tastes as base vectors.

Apply **Principal Component Analysis!**

Recommender System
en.wikipedia.org/wiki/Recommender_system

Lossy Compression

- Downside: New base vectors need to be stored to recover the data.
- Solution: Choose appropriate base vectors (in case of JPEG: DCT)!

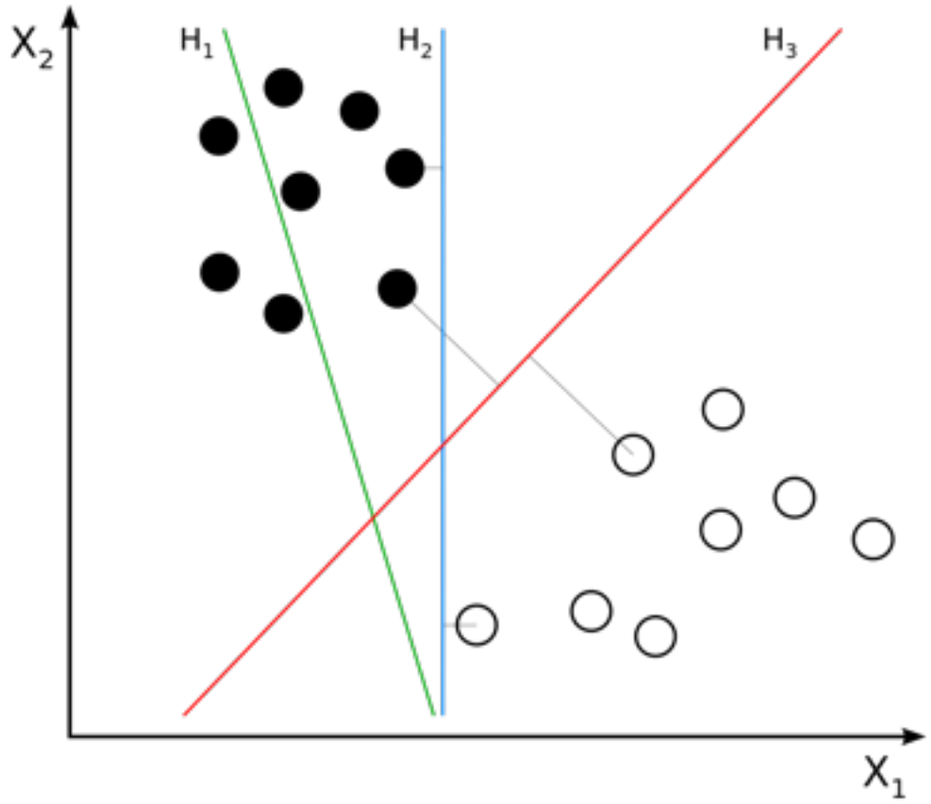


Data Compression
en.wikipedia.org/wiki/Data_compression

Support Vector Machines

State-of-the-art supervised learning models used for **classification**.

Binary, linear classifier:
Find a plane that max. distance to the closest data points from both classes.

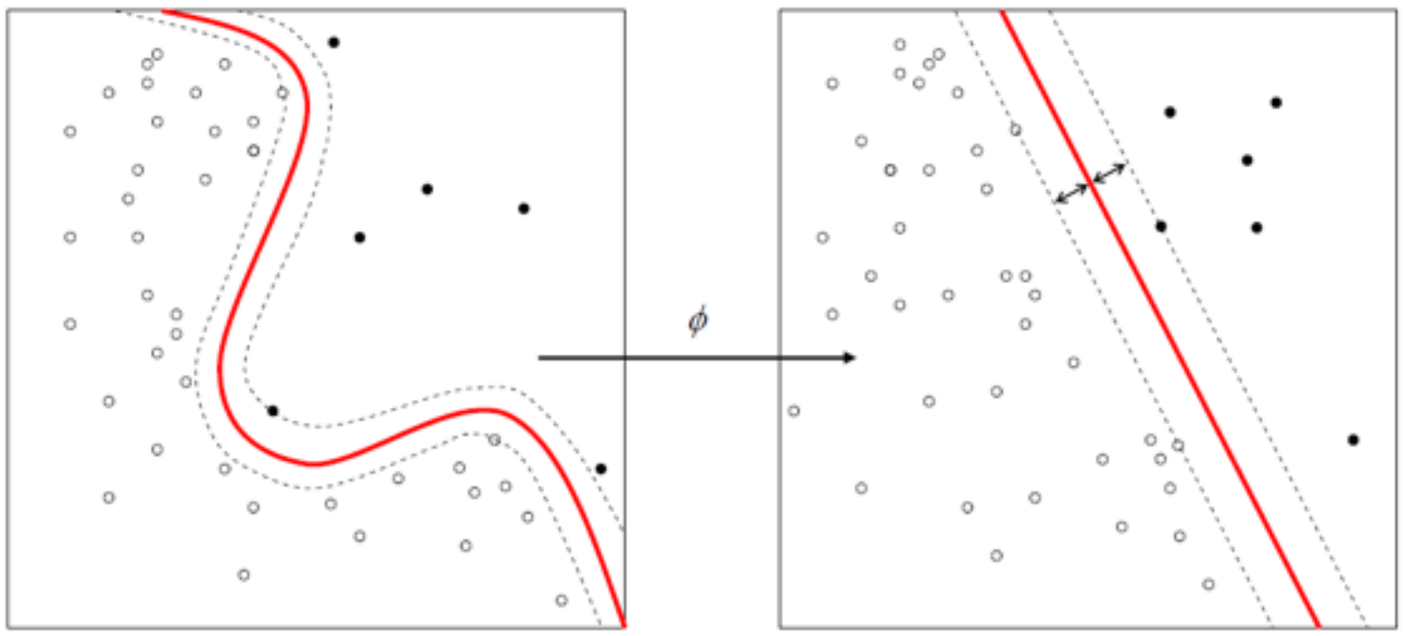


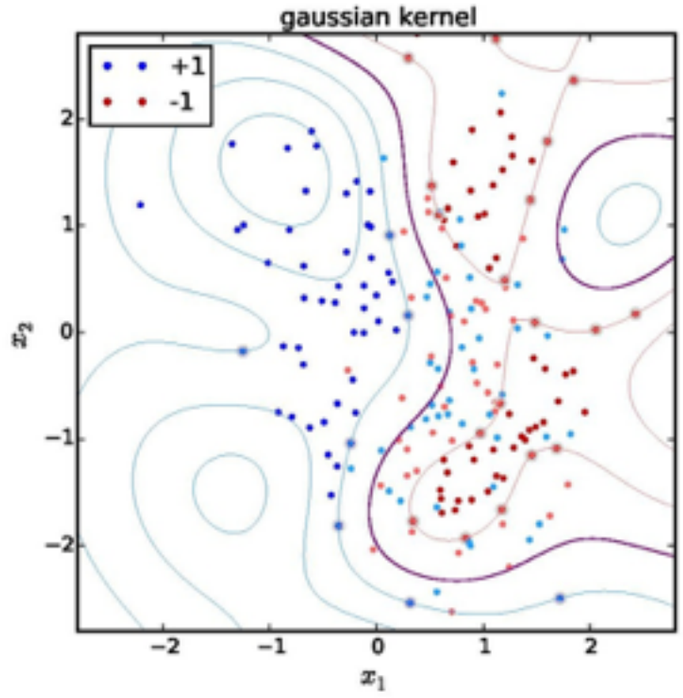
Support Vector Machine
en.wikipedia.org/wiki/Support_vector_machine

SVM's Extensions

In practice, there will often be no plane that perfectly separates the two classes.

Soft margin: Allow violations w. penalty.





Kernel Method
en.wikipedia.org/wiki/Kernel_method

Reinforcement Learning (CNN)

Google develops self-learning computer program
www.theguardian.com/technology/2015/feb/25/[…]

Outlook: Huge Responsibility

… in the near-term. For the long-term, see www.superintelligence.ch.

Train CNN

Deploy CNN

Campaign to Stop Killer Robots
www.stopkillerrobots.org