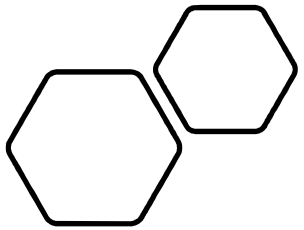
2021: extracting knowledge from data

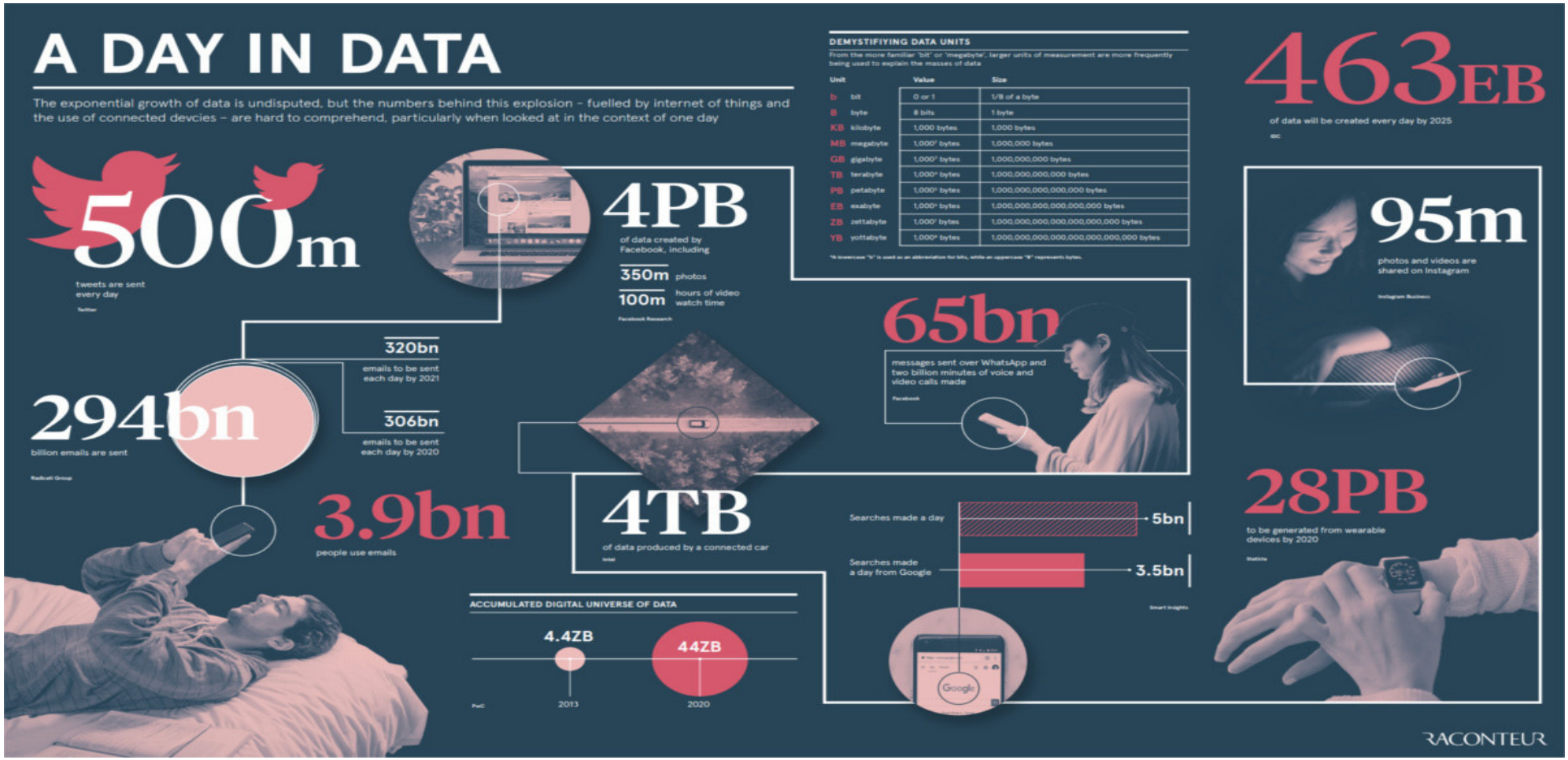


- Definitions
- Data flows
- Measuring, quantifying
- At scale
- Delivery
- Actions speaks
- Thoughts

What is data science?

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Are we “drowning in data”? extracting knowledge from data



data flows : Signs you need a data scientist

- Too much data, too little information
- Overwhelmed by operations
- Data having low business impact
- Lack of collaboration in the business
- Analysis and models not scalable, repeatable
- Poor experience with unstructured data or statistical analysis
- Cognitive biases in your business is impacting your customer relationships

data flows : Job description

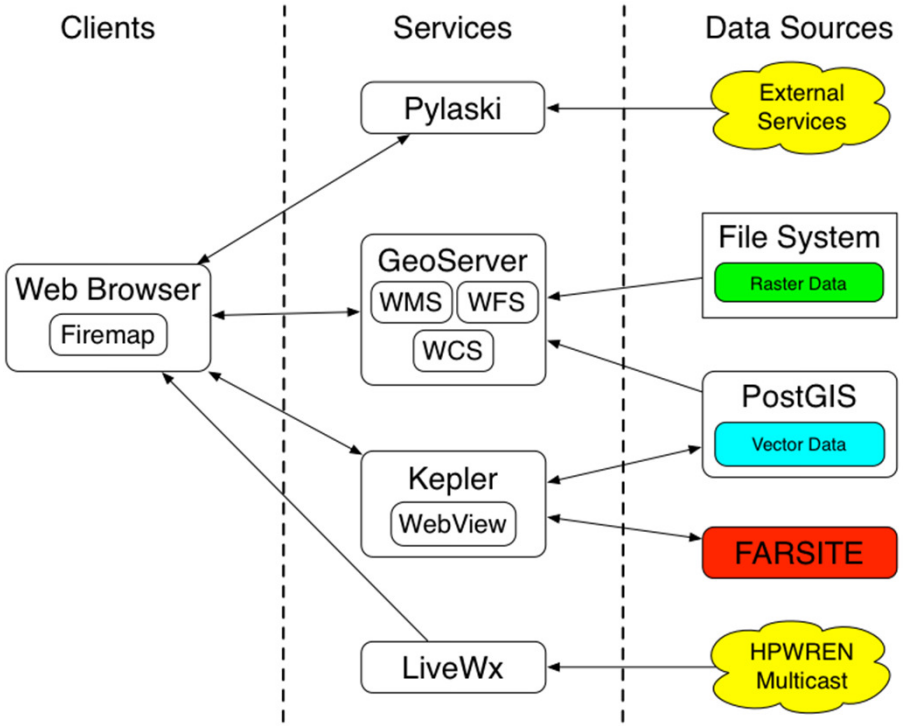
- Will help discover the information hidden in vast amounts of data.
- Will help make smarter decisions to deliver even better products.
- Primary focus will be in applying data mining techniques, doing statistical analysis, and building high quality prediction systems integrated with products.

data flows : Job Skills and Qualifications

- Excellent understanding of machine learning techniques and algorithms, such as k-NN, Naive Bayes, SVM, Decision Forests, etc.
- Experience with common data science toolkits, such as R, Weka, NumPy, MatLab, etc.
- Experience with data visualisation tools, such as D3.js, GGplot, etc.
- Proficiency in using query languages such as SQL, Hive, Pig
- Experience with NoSQL databases, such as MongoDB, Cassandra, Hbase
- Good applied statistics skills, such as distributions, statistical testing, regression, etc.
- Good scripting & programming skill
- Data-oriented personality
- Great communication skills

data flows : WIFIRE

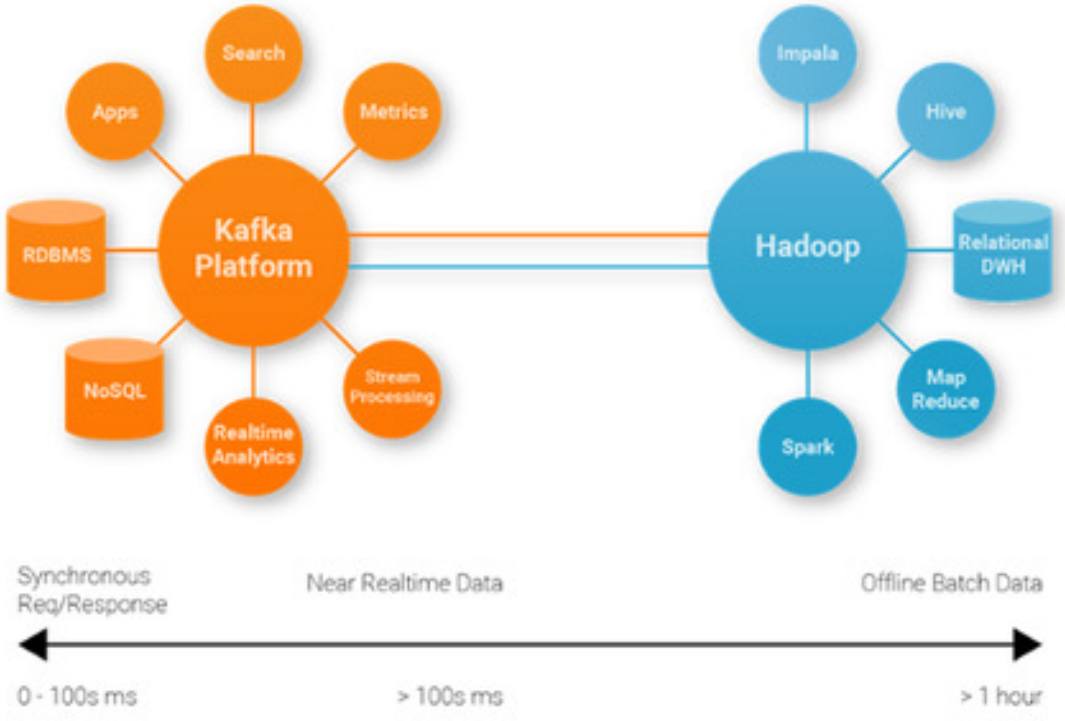
Firemap is a tool developed by WIFIRE researchers to perform data-driven predictive modeling and analysis of fires that have a high potential for rapid spread; and enables what-if analysis of fire scenarios ahead of the time as well as real-time fire forecasting.



- Fire modeling: FARSITE
- Weather stations: HPWREN, SDG&E, and MesoWest & SynopticsLabs
- Weather forecast: NOAA HRRRX and NWS National Digital Forecast Database
- Cameras: HPWREN, SDG&E, UNR Seismological Laboratory, and NV BLM
- Historical fire perimeters: CAL FIRE FRAP Program and USGS GeoMAC
- Fuels: USGS LANDFIRE Program
- Satellite fire detections: NASA FIRMS
- Air quality: OpenAQ

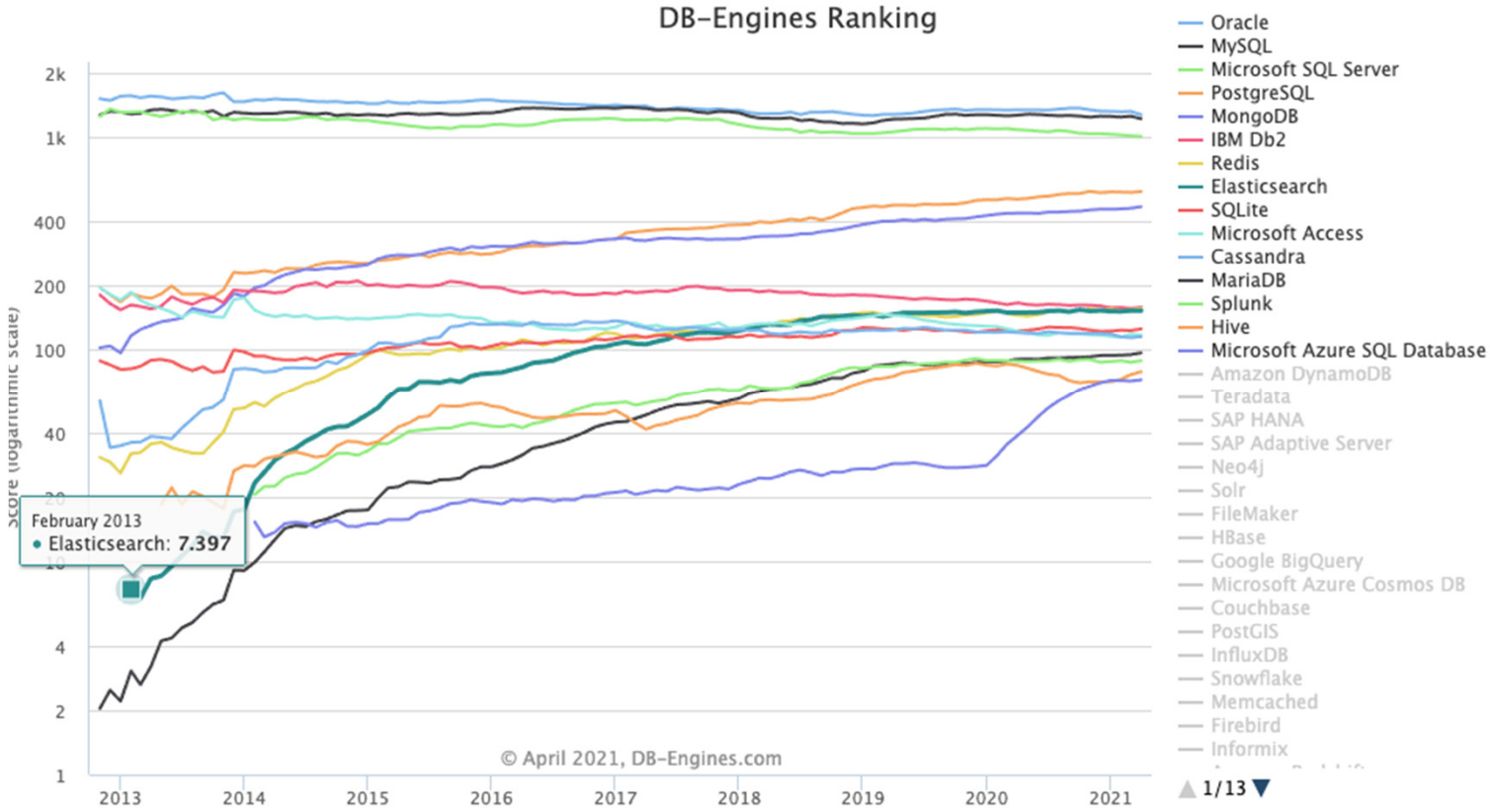
data flows: KAFKA

Apache Kafka is an open-source distributed event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, data integration, and mission-critical applications



- OPC much?
- Batch vs Event Driven
- ETL vs ELT
- Shards vs SQL

data flows: ElasticSearch



- SPEED vs SCHEMA

- SPEED vs ATOMIC

- SHARDS vs NORMALIZATION

- API vs SQL (vs SOAP)

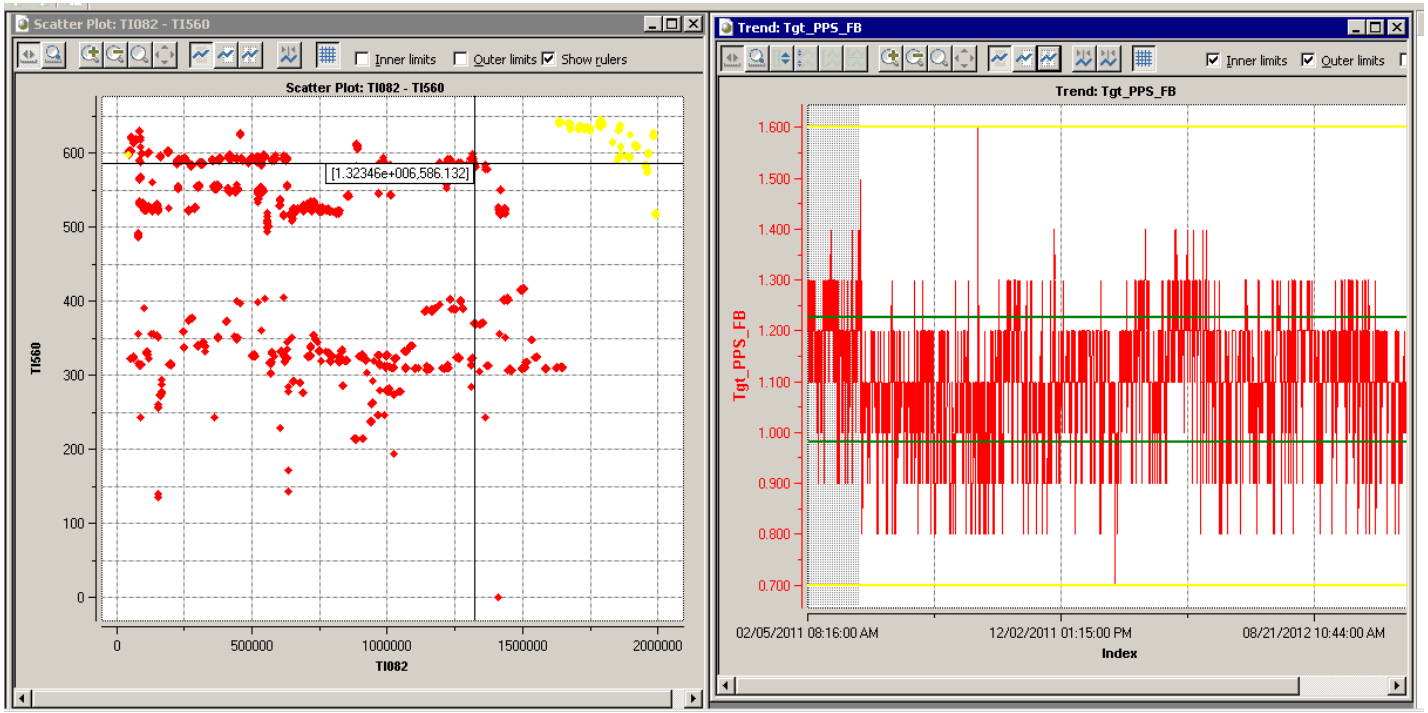
- JSON vs XML

- S3 vs AZURE BLOB vs GOOGLE DRIVE

measuring: pps smoothness C1

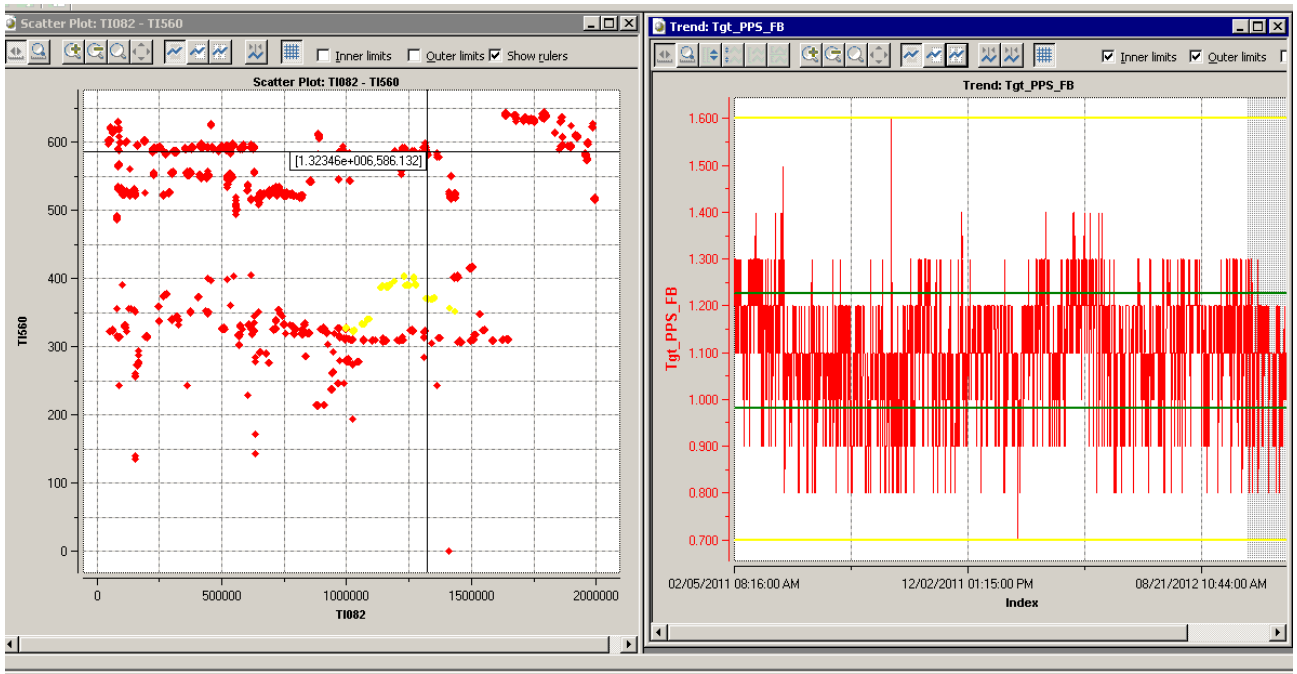
- C1 Paper Machine, MeadWestVaco, West Virginia
- Makes carton in a semi batch process, Parker Print-Surf (PPS) is a KPI.
- Smoother carton, prints better, makes status symbols out of cigarette boxes
- Investigate the driving factors for PPS smoothness variation on the C1 paper machine
- Propose additions to the PPS C1 paper machine control system to avoid higher than expected PPS scores
- CSense Advanced Analytics Platform
- Dumped 1000 tags of a year's data on me at 1min sampling rate, no names, no units, no process diagram. I.P. protection!

measuring: PPS Machine C1 operation variation



On the right: PPS as a function of time. On the Left TI560 vs. TI082. The yellow/grey lighted area correspond to the high PPS scores at the beginning of 2011. The same plot will highlight the change in operation from the begin of 2011 to the end of 2012.

measuring: Change in operation process tags



On the right: PPS as a function of time. On the left TI560 vs. TI082. The yellow/grey lighted area correspond to the lower PPS scores at the end of 2012. The biggest change in operation from 2011 to the 2012, from the selected tag list, is a downward shift in both TI082 and in TI560. Overall lower values for TI560 appears to be good for PPS smoothness and blade usage.

measuring: KAGGLE

Inside Kaggle you'll find all the code & data you need to do your data science work. Use over 50,000 public datasets and 400,000 public notebooks to conquer any analysis in no time.

Discussion topic

Electricity Price or Electricity source datasets

by Matthew P. McAteer

Datasets

Electricity Price or Electricity source datasets

Discussion topic

Electricity Consumption by Month, Year (new datasets)

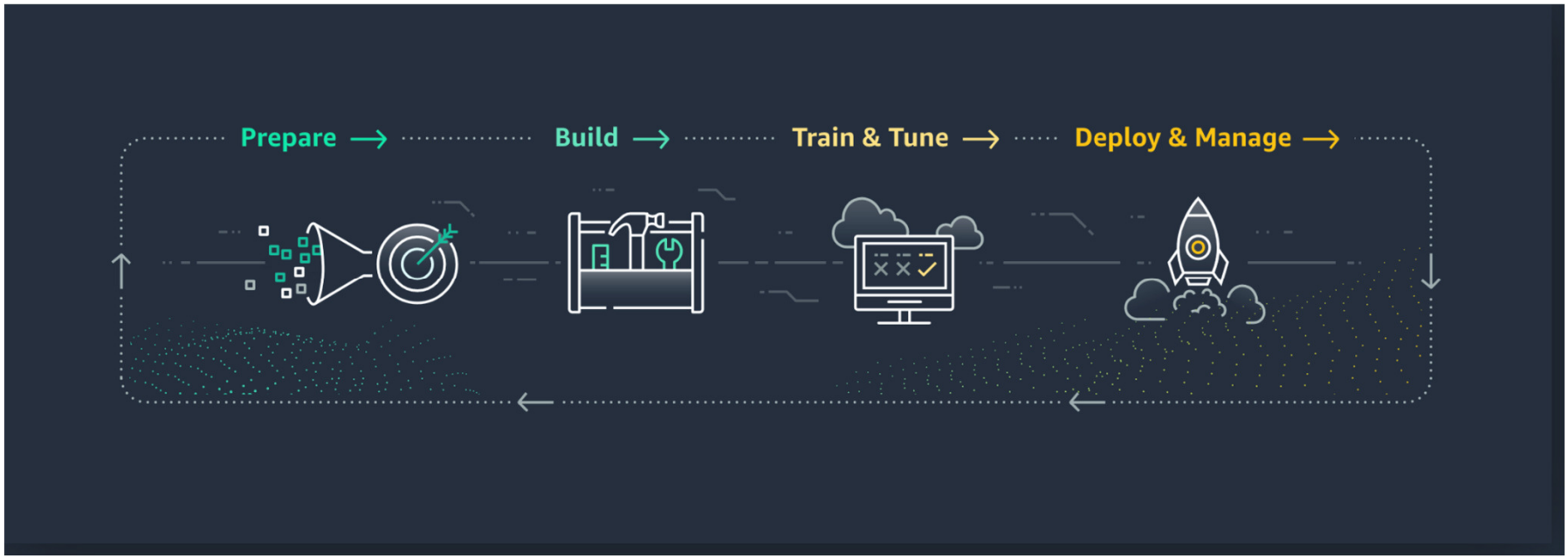
by V. Gates

DS4G - Environmental Insights Explorer

Electricity Consumption by Month, Year (new datasets)

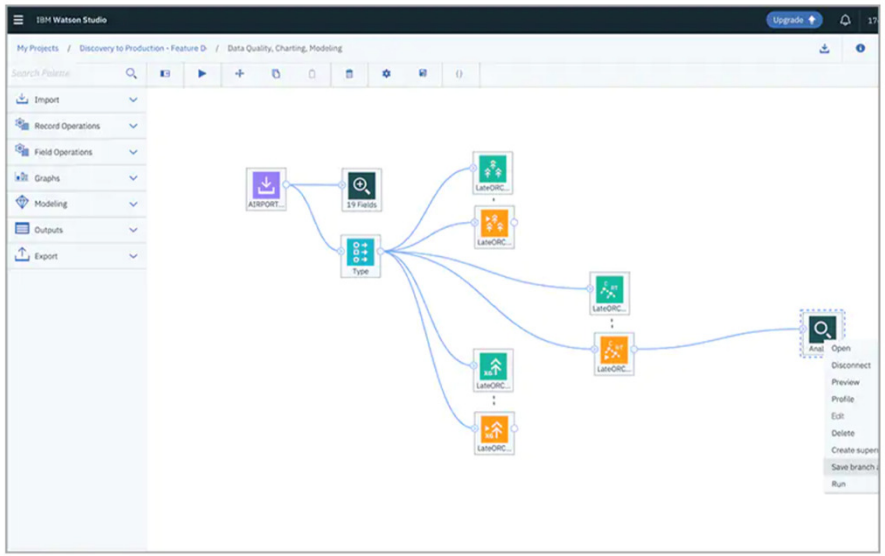
at scale: AWS SAGEMAKER

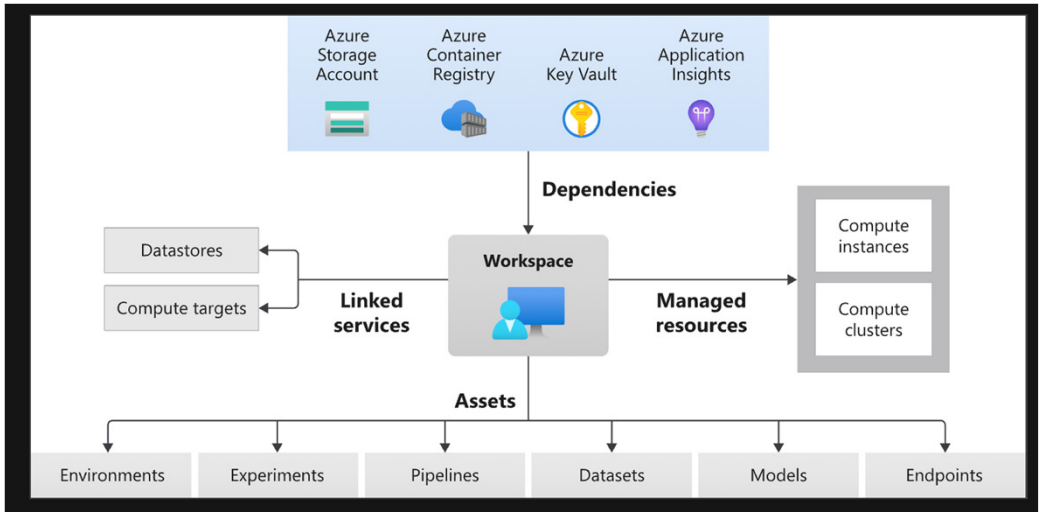
Amazon SageMaker helps data scientists and developers to prepare, build, train, and deploy high-quality machine learning (ML) models quickly by bringing together a broad set of capabilities purpose-built for ML.



at scale: AZURE, WATSON

Azure Machine Learning, a cloud-based environment you can use to train, deploy, automate, manage, and track ML models.



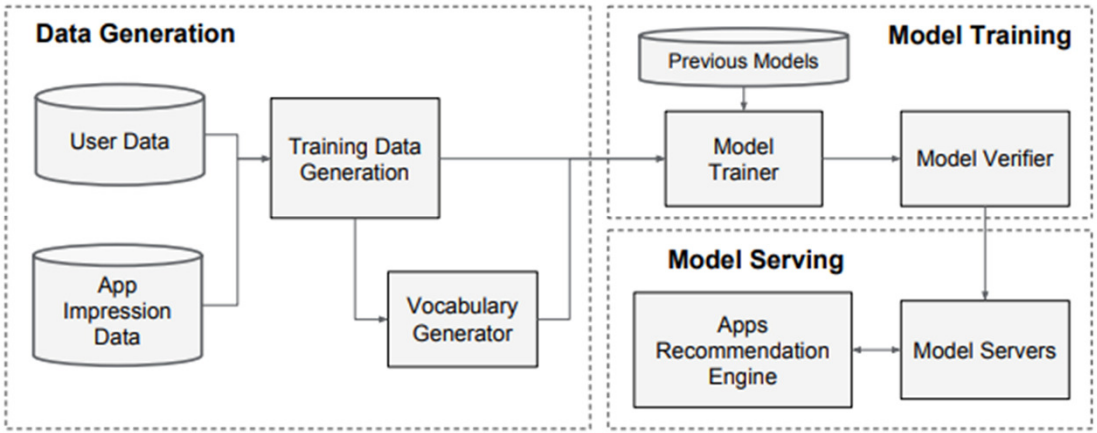


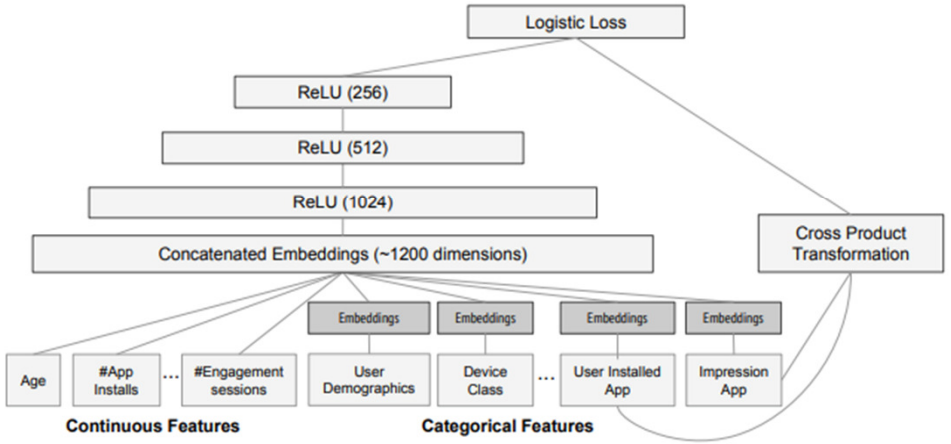
As part of IBM Watson® Studio, IBM Watson Machine Learning helps data scientists and developers accelerate AI and machine learning deployment on IBM Cloud Pak® for Data. Deploy AI models at scale across any cloud on an open, extensible architecture.

at scale: RECOMMENDERS

The birth of Tensorflow: https://arxiv.org/pdf/1606.07792.pdf

At peak traffic, our recommender servers score over 10 million apps per second. With single threading, scoring all candidates in a single batch takes 31 ms. We implemented multithreading and split each batch into smaller sizes, which significantly reduced the client-side latency to 14 ms (including serving overhead)

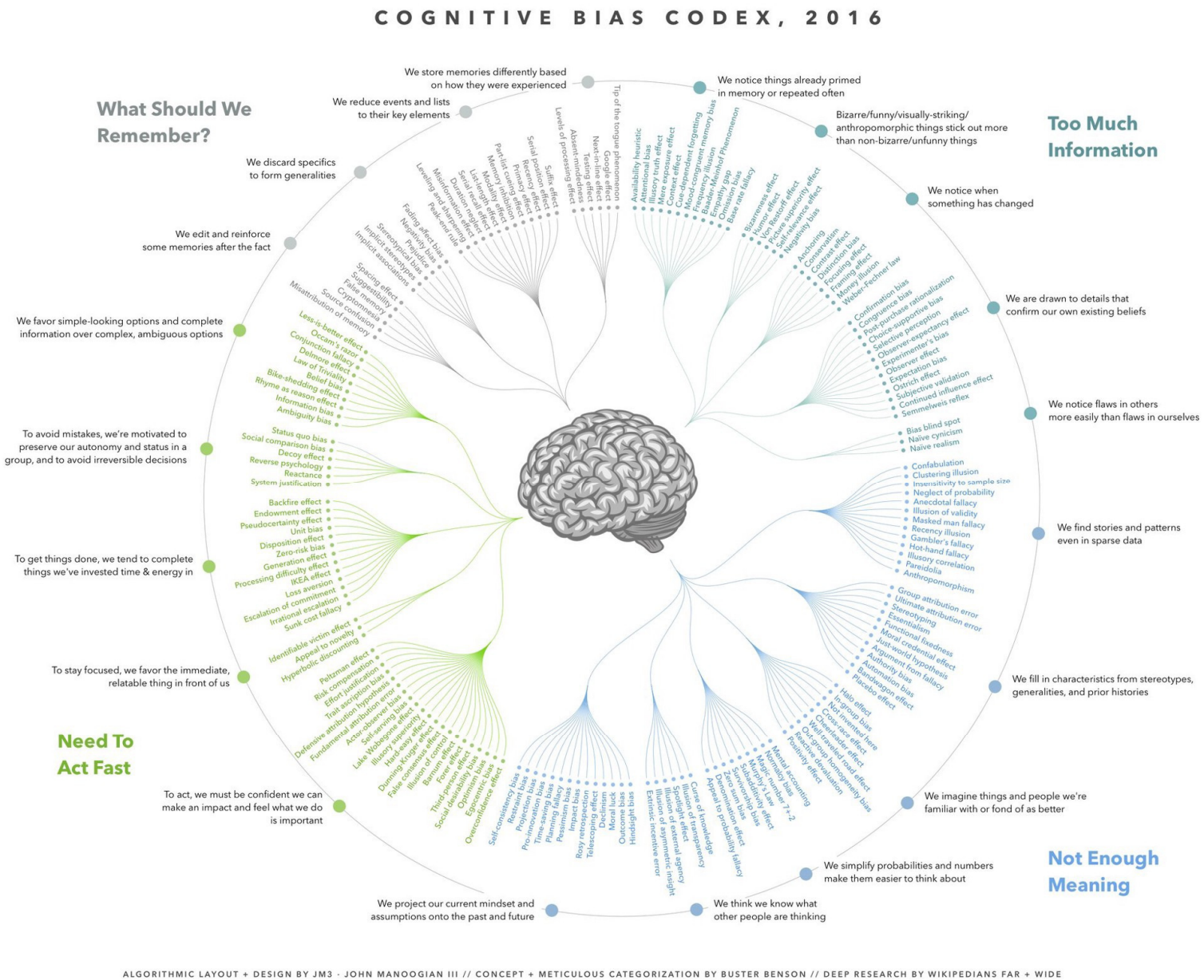




delivery: change minds

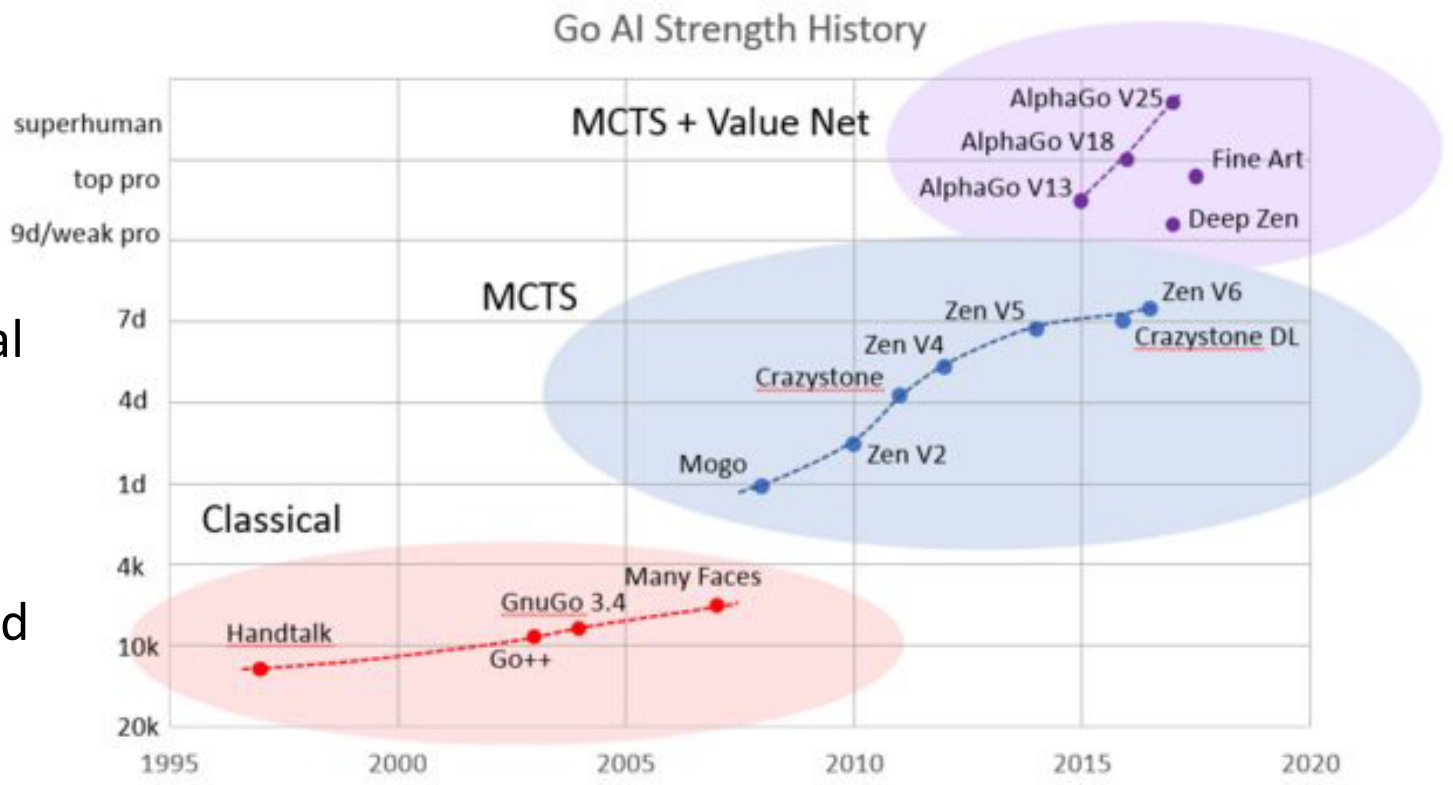
Cognitive biases are systematic patterns of deviation from norm and/or rationality in judgment. They are often studied in psychology and behavioral economics.
Although the reality of most of these biases is confirmed by reproducible research, there are often controversies about how to classify these biases or how to explain them.

TELL STORIES!
GPS NARATIVE
DIGITAL TWIN



actions speaks: AlphaGo Zero

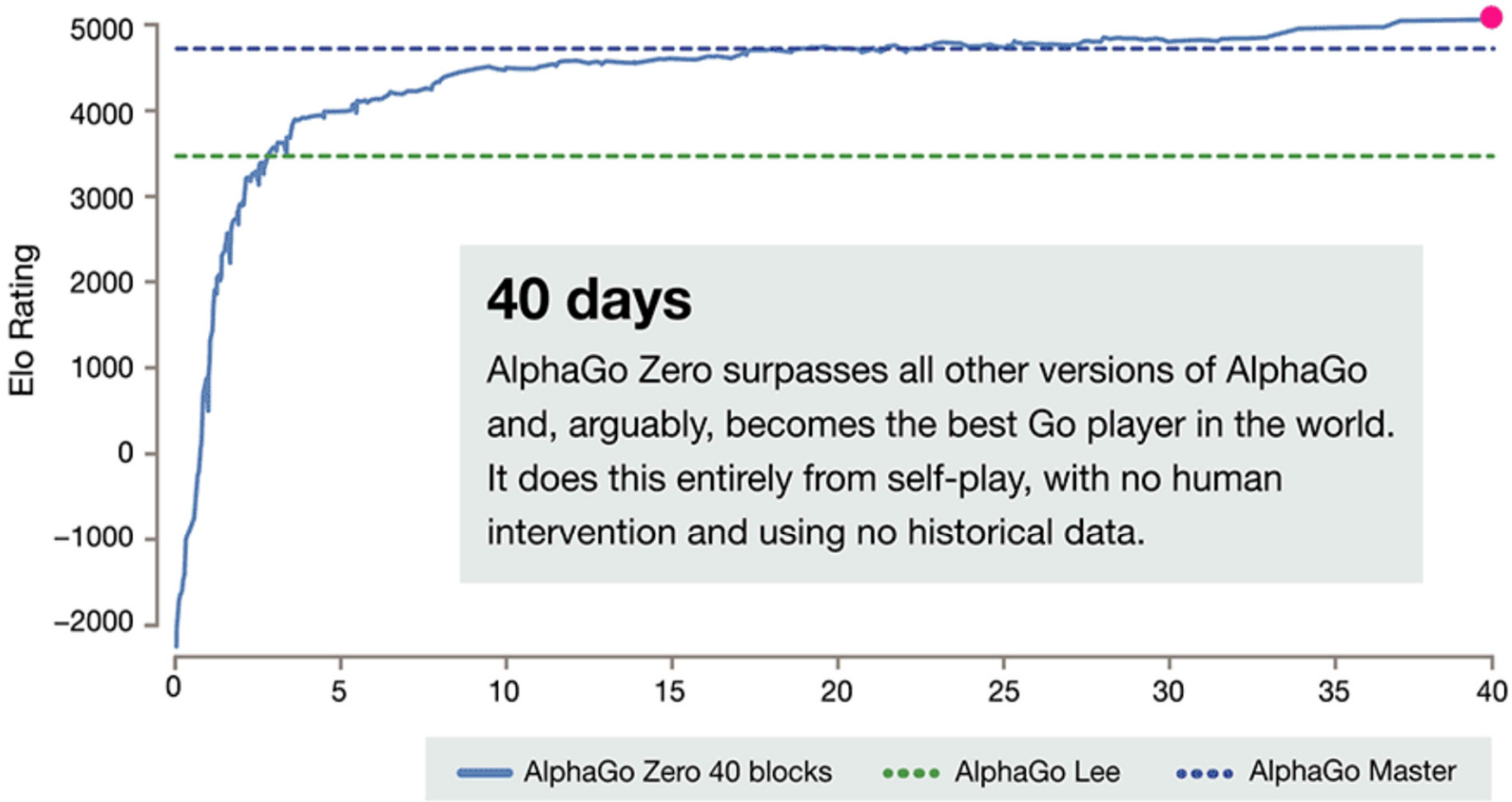
Deep reinforcement learning (**deep RL**) is a subfield of machine learning that combines reinforcement learning (RL) and deep learning. RL considers the problem of a computational agent learning to make decisions by trial and error. Deep RL incorporates deep learning into the solution, allowing agents to make decisions from unstructured input data without manual engineering of the state space.



actions speaks: AlphaGo Zero

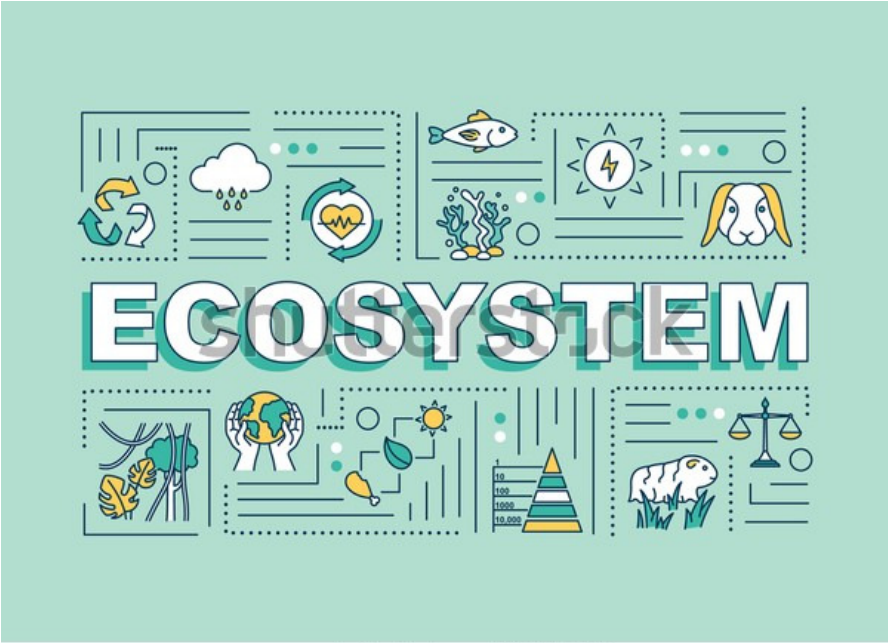
The system starts off with a neural network that knows nothing about the game of Go. It then plays games against itself, by combining this neural network with a powerful search algorithm. As it plays, the neural network is tuned and updated to predict moves, as well as the eventual winner of the games.

This updated neural network is then recombined with the search algorithm to create a new, stronger version of AlphaGo Zero, and the process begins again.



thoughts: does Go relate to electricity use?

The **competitive exclusion principle** tells us that two species can't have exactly the same niche in a habitat and stably coexist. That's because species with identical niches also have identical needs, which means they would compete for precisely the same resources.



On a Go board two species compete for a resource in the same niche. The species that controls more of the board, wins.

Boundary conditions play an important role on a board.

Can we create an Ecosystem game?

thoughts: human knowledge?

A 2016 poll of 1,500 scientists reported that 70% of them had failed to reproduce at least one other scientist's experiment (50% had failed to reproduce one of their own experiments).[9] In 2009, 2% of scientists admitted to falsifying studies at least once and 14% admitted to personally knowing someone who did.

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may

Top languages for data science (2020)

	Bultin	Flatiron	Simplilearn	bigdata-madesimple
1	Python	Python	Python	Python
2	R	Javascript	R	Java
3	Julia	Java	Java	R
4	C/C++	R	Javascript	SQL
5	Java	C/C++	SAS	Julia

It’s clear Python, R, Java, SQL are key resources

Database choices (Task drives selection)

- Open source or Licenced
- SQL or NOSQL
- Relational or not
- Storage of Structured or unstructured data (Data,objects,images,maps,text..)
- Logging & recovery
- Capacity
- Scalability
- Realtime or eventual Consistency

thoughts: Data science

- Data science does not exist for itself, it creates a digital twin of something, its just a mirror
- Data science is a team sport, Data Scientists tend to be anti social
- The role of automation
- Code whatever you are comfortable in, but, R, Python, Scala, REST, SQL, Javascript has the best libraries for data.
- Some data scientist despise coding; Some data scientists despise statisticians.
- Tensorflow, Spark, SQL, Kafka, etc are clusters/cluster expression trees. Think DOM construction, setup for execution.
- A data scientist represents years of experience in a range of disciplines, worked by a brilliant mind. Payment must reward the effort. Or else you will pay for a team. Good luck: do you have line of sight on a solution, with a siloed team?

THANK YOU

thoughts: human security?

5 October 2020 **OAuth 2.0 Security Best Current Practice**

draft-ietf-oauth-security-topics-16

Lemma 10 (Third parties do not learn state). Let ρ be a run of an OAuth web system with web attackers $\mathcal{OWS}^w$, (S^j, E^j, N^j) be a state of ρ, $r \in \mathsf{RP}$ be an RP that is honest in S_j, $i \in \mathsf{IDP}$ be an IdP that is honest in S_j, b be a browser that is honest in S_j.

Then there exists no $l \leq j$, with (S^l, E^l, N^l) being a state in ρ, a nonce $\mathit{loginSessionId} \in \mathcal{N}$, a nonce $\mathit{state} \in \mathcal{N}$, a domain $h \in \mathsf{dom}(r)$ of r, terms x, y, x', y', $z \in \mathcal{T}_{\mathcal{N}}$, cookie $c := \langle \texttt{loginSessionId}, \langle \mathit{loginSessionId}, x', y', z \rangle \rangle$, an atomic DY process $p \in \mathcal{W} \setminus \{b, i, r\}$ such that $\mathit{state} \in d_\emptyset(S^l(p))$, $\langle \mathit{loginSessionId}, \langle g, \mathit{state}, x, y \rangle \rangle \in^{\langle\rangle} S^l(r).\texttt{loginSessions}$ and $\langle h, c \rangle \in^{\langle\rangle} S^l(b).\texttt{cookies}$.

PROOF. To prove Lemma 10, we track where the login session identified by *loginSessionId* is created and used.

We have that $\langle h, c \rangle \in^{\langle\rangle} S^l(b).\texttt{cookies}$. Login sessions are only created in Line 100 of Algorithm 10

https://tools.ietf.org/html/draft-ietf-oauth-security-topics-16