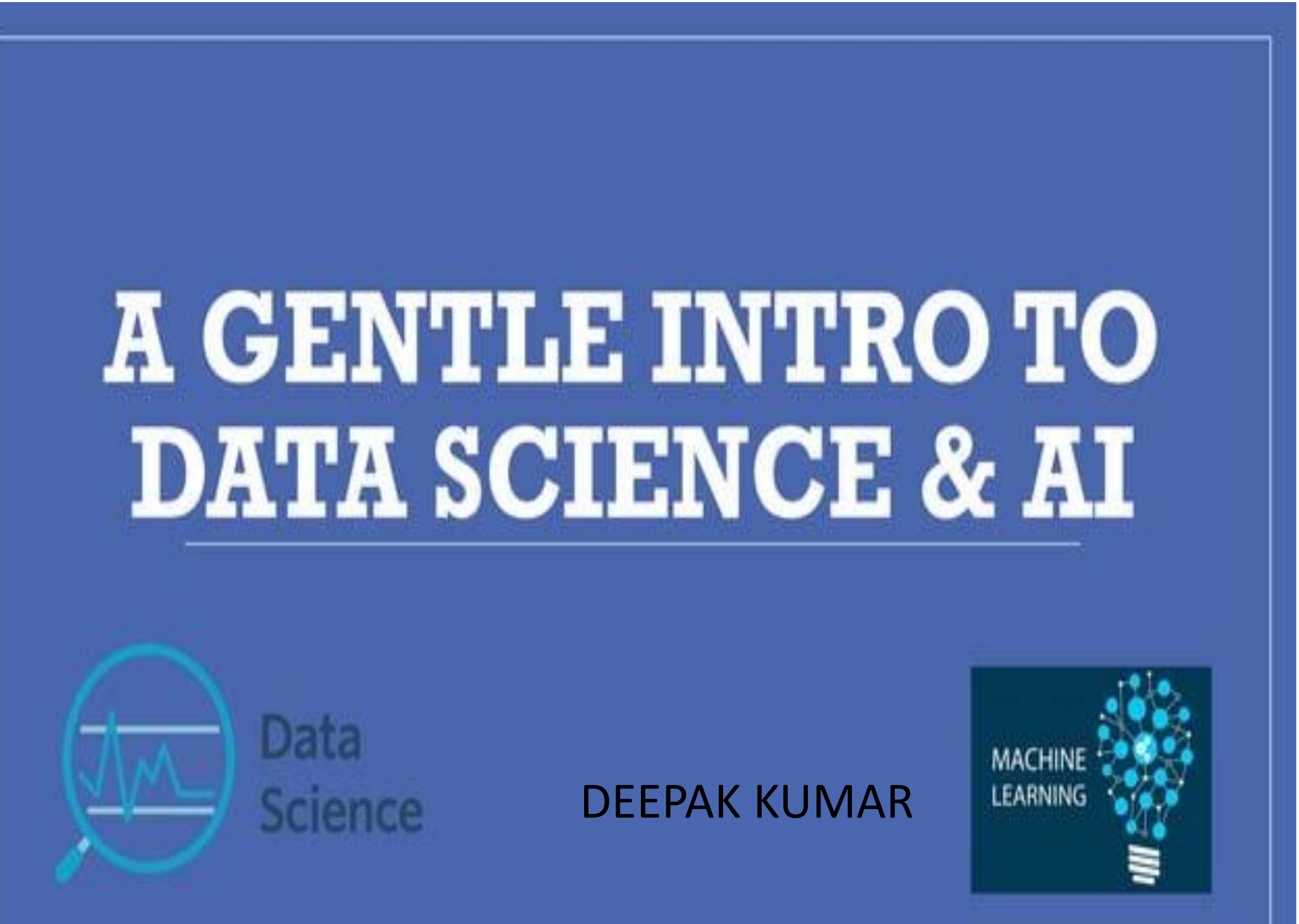


A GENTLE INTRO TO DATA SCIENCE & AI

DEEPAK KUMAR

Computer Science Hot Topics

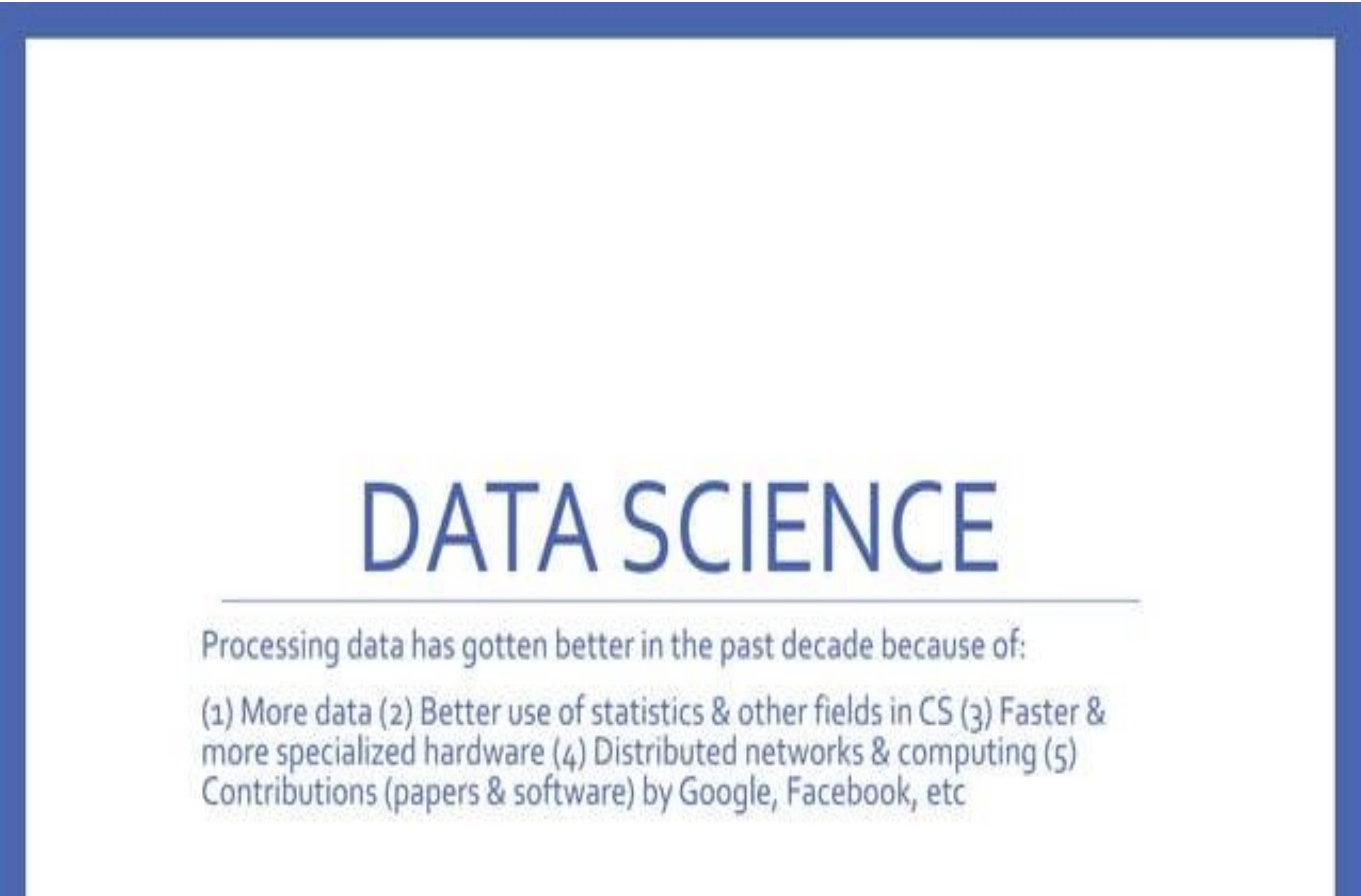
- Mid 1980s – 1990s: desktop applications
  - Networking, graphics & graphical user interfaces (GUIs), some AI / ML

- Mid 1990s – 2006: websites & web applications
- 2007 – 2014: mobile apps

```
starting_date = '2016070
sample_numpy_data = np.a
dates_index = pd.date_ra
sample_df = pd.DataFrame

sample_df_2 = sample_df.c
sample_df_2['Fruits'] = 
```

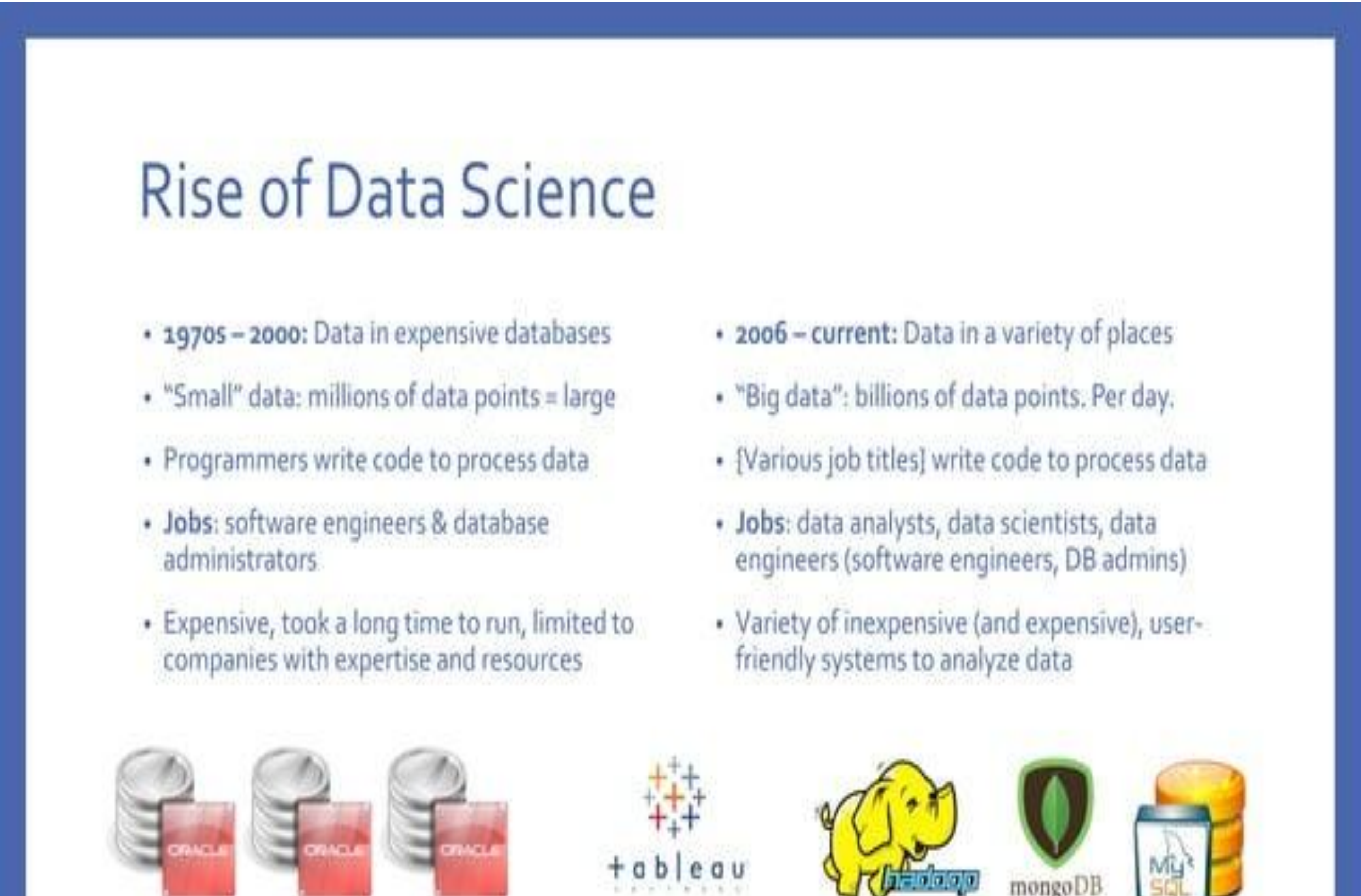
- 2012 – 2017: data science
  - Maybe some virtual reality (VR) and augmented reality (AR)
- 2016 – current: artificial intelligence (AI) & machine learning (ML)
- 2017 – early 2018: Bitcoin! (Crypto-currencies)
  - IMHO, passing fad & pure speculation



DATA SCIENCE

Processing data has gotten better in the past decade because of:

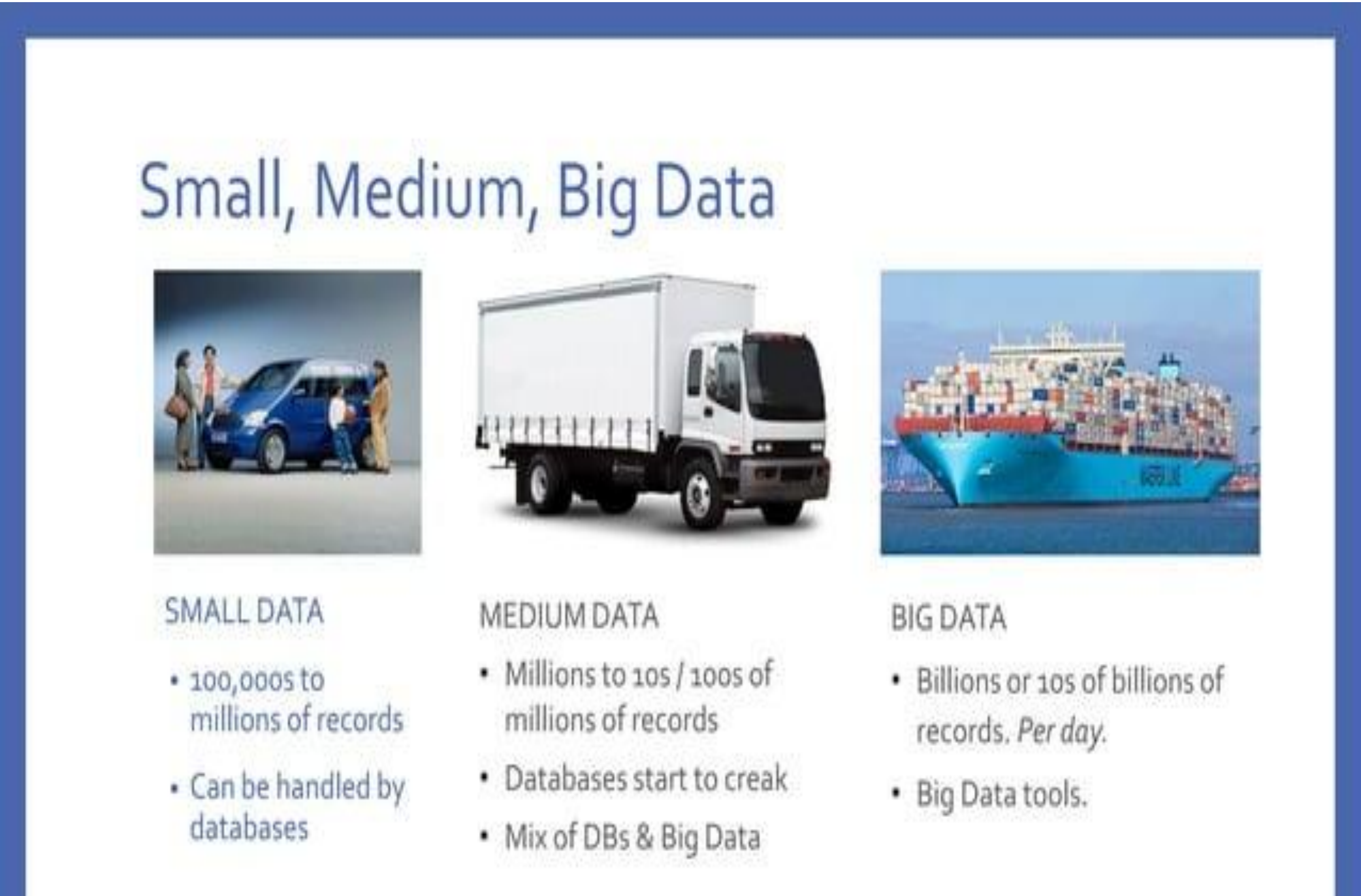
(1) More data (2) Better use of statistics & other fields in CS (3) Faster & more specialized hardware (4) Distributed networks & computing (5) Contributions (papers & software) by Google, Facebook, etc



Rise of Data Science

- **1970s – 2000:** Data in expensive databases
- "Small" data: millions of data points = large
- Programmers write code to process data
- **Jobs**: software engineers & database administrators
- Expensive, took a long time to run, limited to companies with expertise and resources

- **2006 – current:** Data in a variety of places
- "Big data": billions of data points. Per day.
- [Various job titles] write code to process data
- **Jobs**: data analysts, data scientists, data engineers (software engineers, DB admins)
- Variety of inexpensive (and expensive), user-friendly systems to analyze data



Small, Medium, Big Data

SMALL DATA

- 100,000s to millions of records
- Can be handled by databases

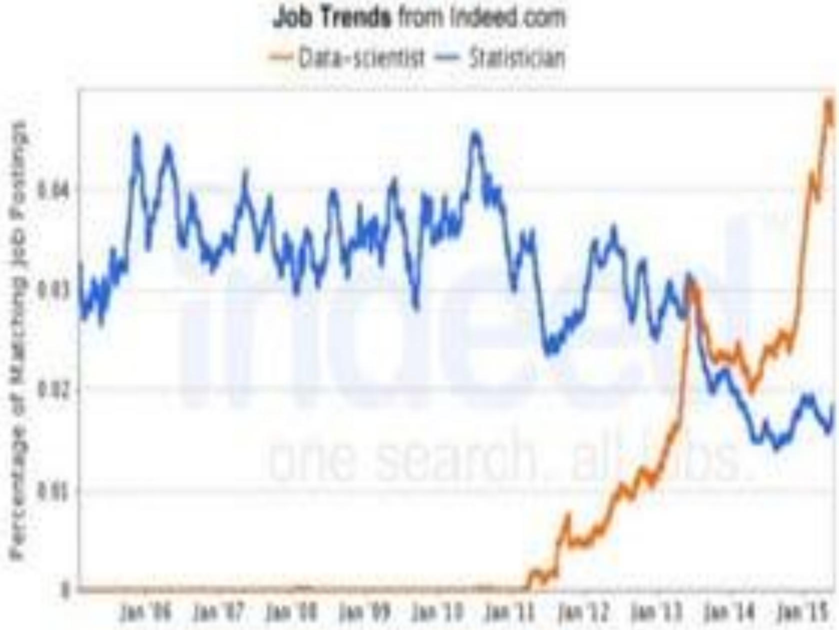
MEDIUM DATA

- Millions to 10s / 100s of millions of records
- Databases start to creak
- Mix of DBs & Big Data

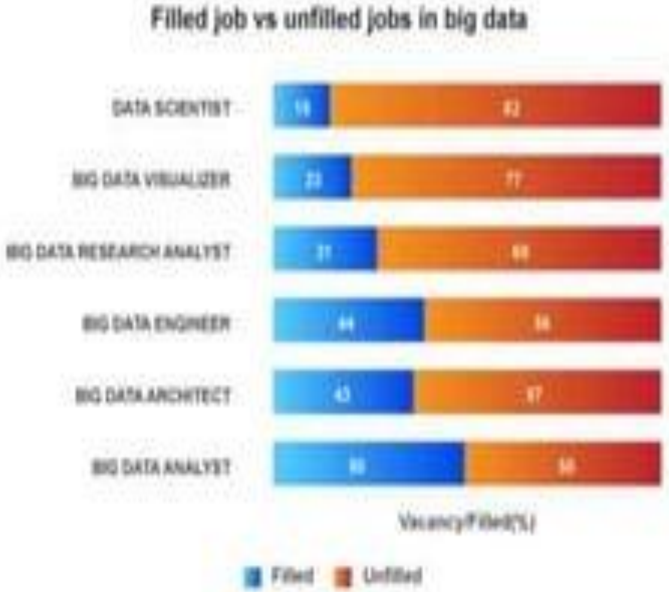
BIG DATA

- Billions or 10s of billions of records. *Per day.*
- Big Data tools.

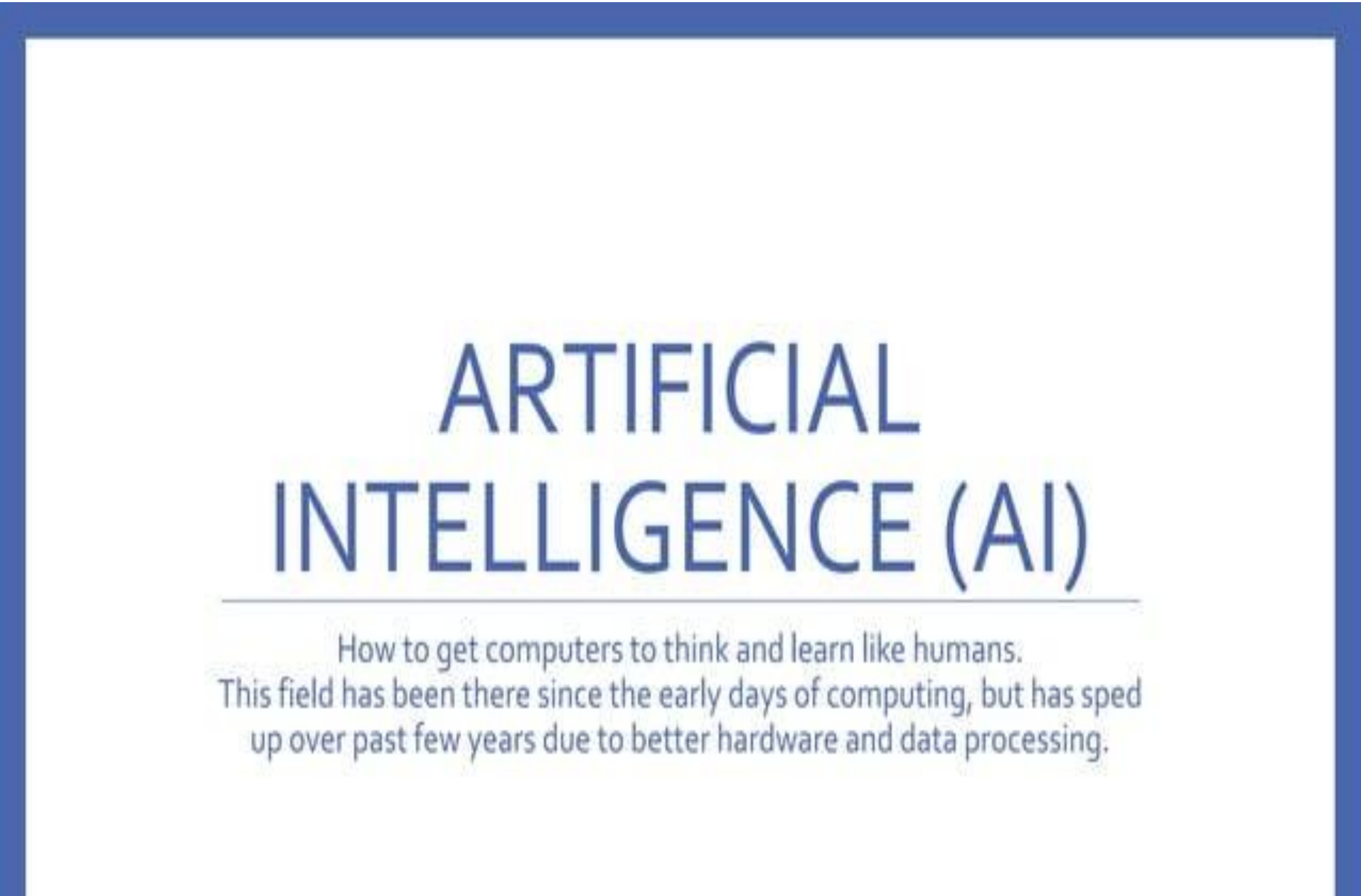
Job trends in Data Science



Indeed.com postings, Sep 2017

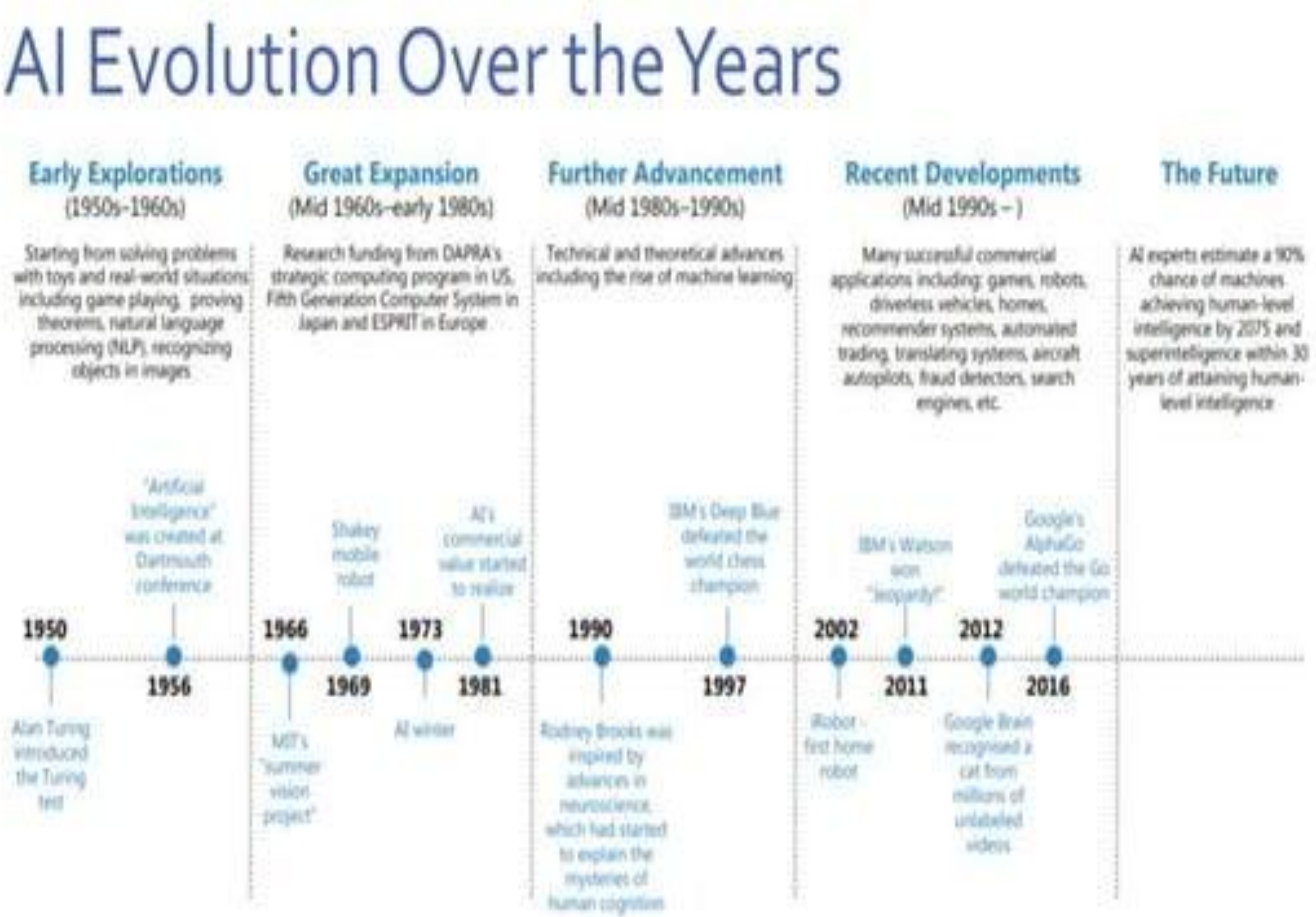


Quora, Dec 2017



ARTIFICIAL INTELLIGENCE (AI)

How to get computers to think and learn like humans.
This field has been there since the early days of computing, but has sped up over past few years due to better hardware and data processing.



AI Evolution Over the Years

Early Explorations

(1950s–1960s)

Starting from solving problems with toys and real-world situations including game playing, proving theorems, natural language processing (NLP), recognizing objects in images

Great Expansion

(Mid 1960s–early 1980s)

Research funding from DAPRA's strategic computing program in US, Fifth Generation Computer System in Japan and ESPRIT in Europe

Further Advancement

(Mid 1980s–1990s)

Technical and theoretical advances including the rise of machine learning

Recent Developments

(Mid 1990s –)

Many successful commercial applications including: games, robots, driverless vehicles, homes, recommender systems, automated trading, translating systems, aircraft autopilots, fraud detectors, search engines, etc.

The Future

AI experts estimate a 90% chance of machines achieving human-level intelligence by 2075 and superintelligence within 30 years of attaining human-level intelligence

1950 Alan Turing introduced the Turing test

1956 "Artificial Intelligence" was created at Dartmouth conference

1966 MIT's "summer vision project"

1969 Shakey mobile robot

1973 AI winter

1981 AI's commercial value started to realize

1990 Rodney Brooks was inspired by advances in neuroscience, which had started to explain the mysteries of human cognition

1997 IBM's Deep Blue defeated the world chess champion

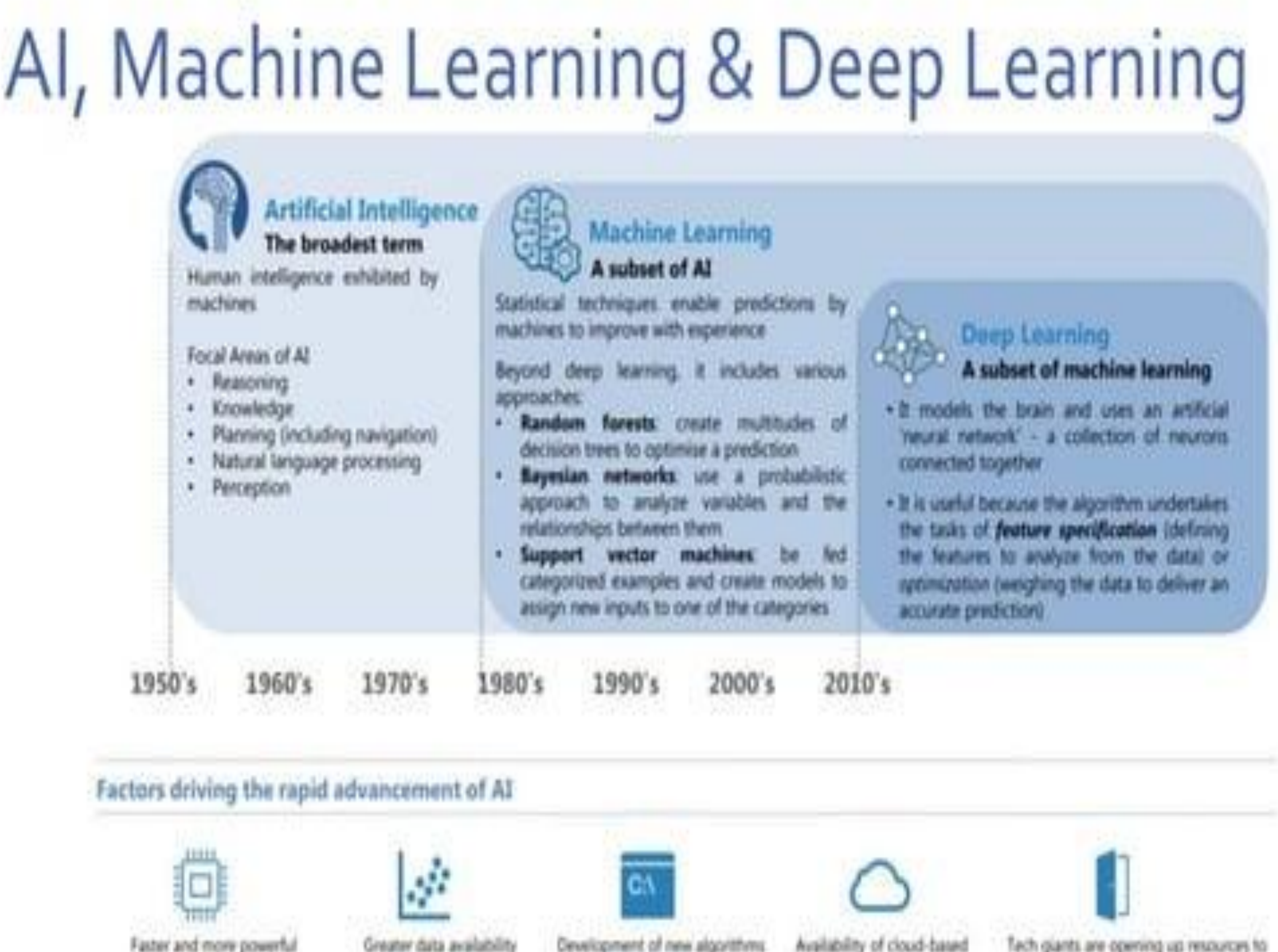
2002 iRobot – first home robot

2011 IBM's Watson won "Jeopardy!"

2012 Google Brain recognised a cat from millions of unlabeled videos

2016 Google's AlphaGo defeated the Go world champion

AI, Machine Learning & Deep Learning



Artificial Intelligence

The broadest term

Human intelligence exhibited by machines

Focal Areas of AI

- Reasoning
- Knowledge
- Planning (including navigation)
- Natural language processing
- Perception

Machine Learning

A subset of AI

Statistical techniques enable predictions by machines to improve with experience

Beyond deep learning, it includes various approaches:

- **Random forests**: create multitudes of decision trees to optimise a prediction
- **Bayesian networks**: use a probabilistic approach to analyze variables and the relationships between them
- **Support vector machines**: be fed categorized examples and create models to assign new inputs to one of the categories

Deep Learning

A subset of machine learning

- It models the brain and uses an artificial 'neural network' - a collection of neurons connected together
- It is useful because the algorithm undertakes the tasks of ***feature specification*** (defining the features to analyze from the data) or *optimization* (weighing the data to deliver an accurate prediction)

Factors driving the rapid advancement of AI

Faster and more powerful

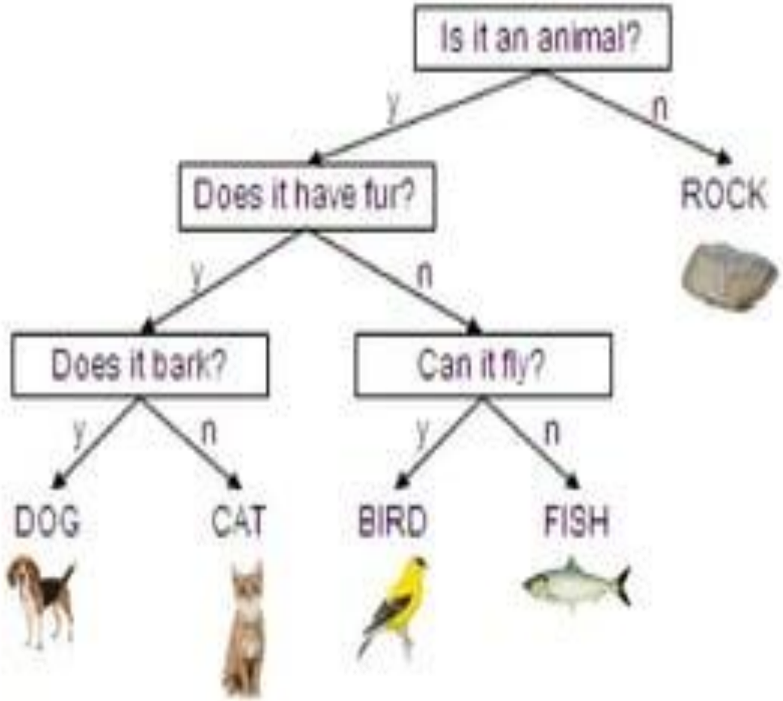
Greater data availability

Development of new algorithms

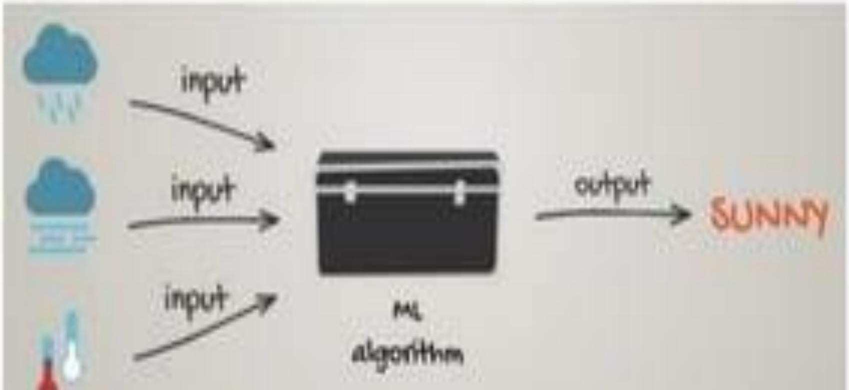
Availability of cloud-based

Tech giants are opening up resources to

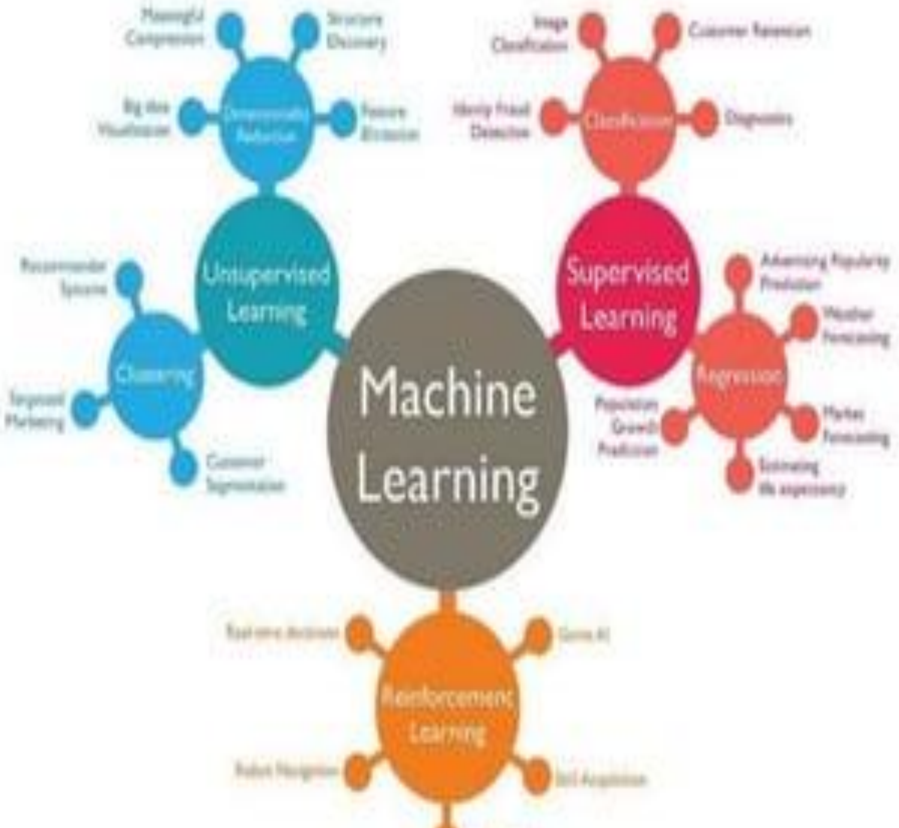
Machine Learning (ML)



- Limited AI; machine learns from existing data to give (hopefully) correct response
- Build model that outputs correct information given training on input data
  - Model often built by a human and computer is "trained" using existing data

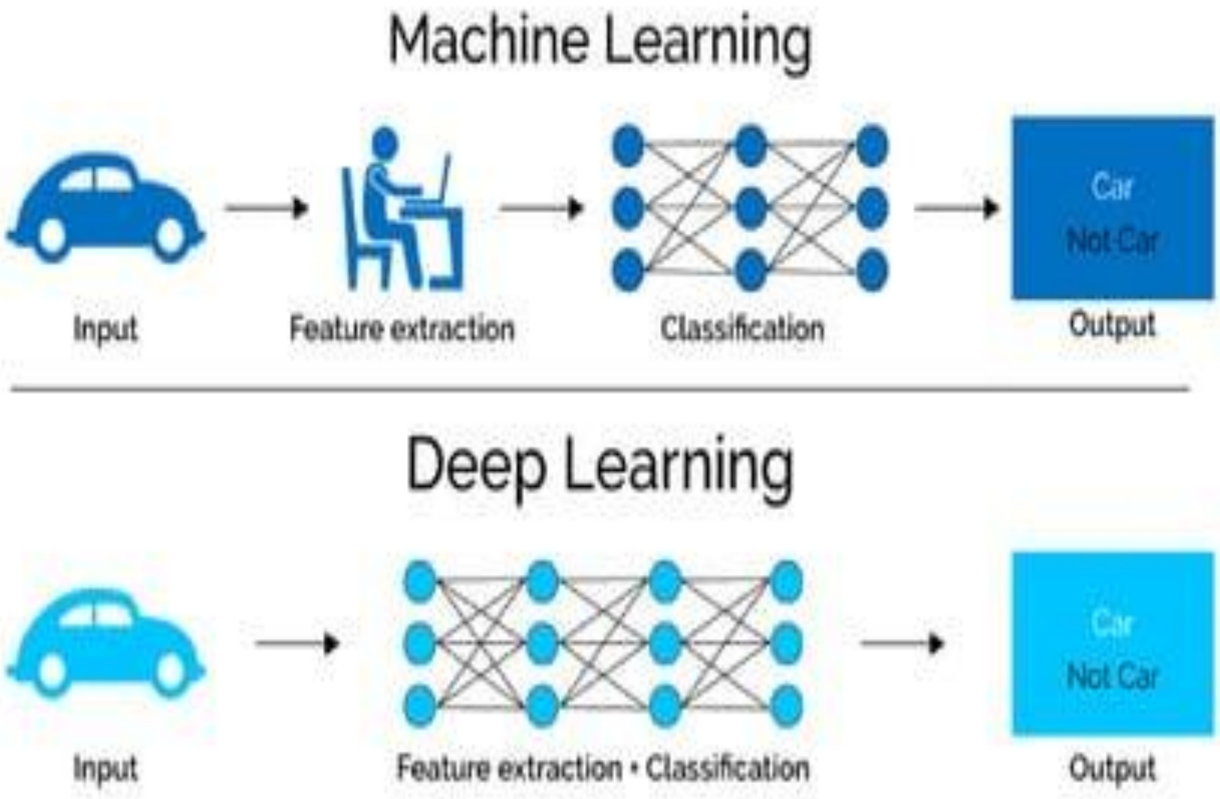


ML Fields: Supervised, Unsupervised & Reinforcement Learning

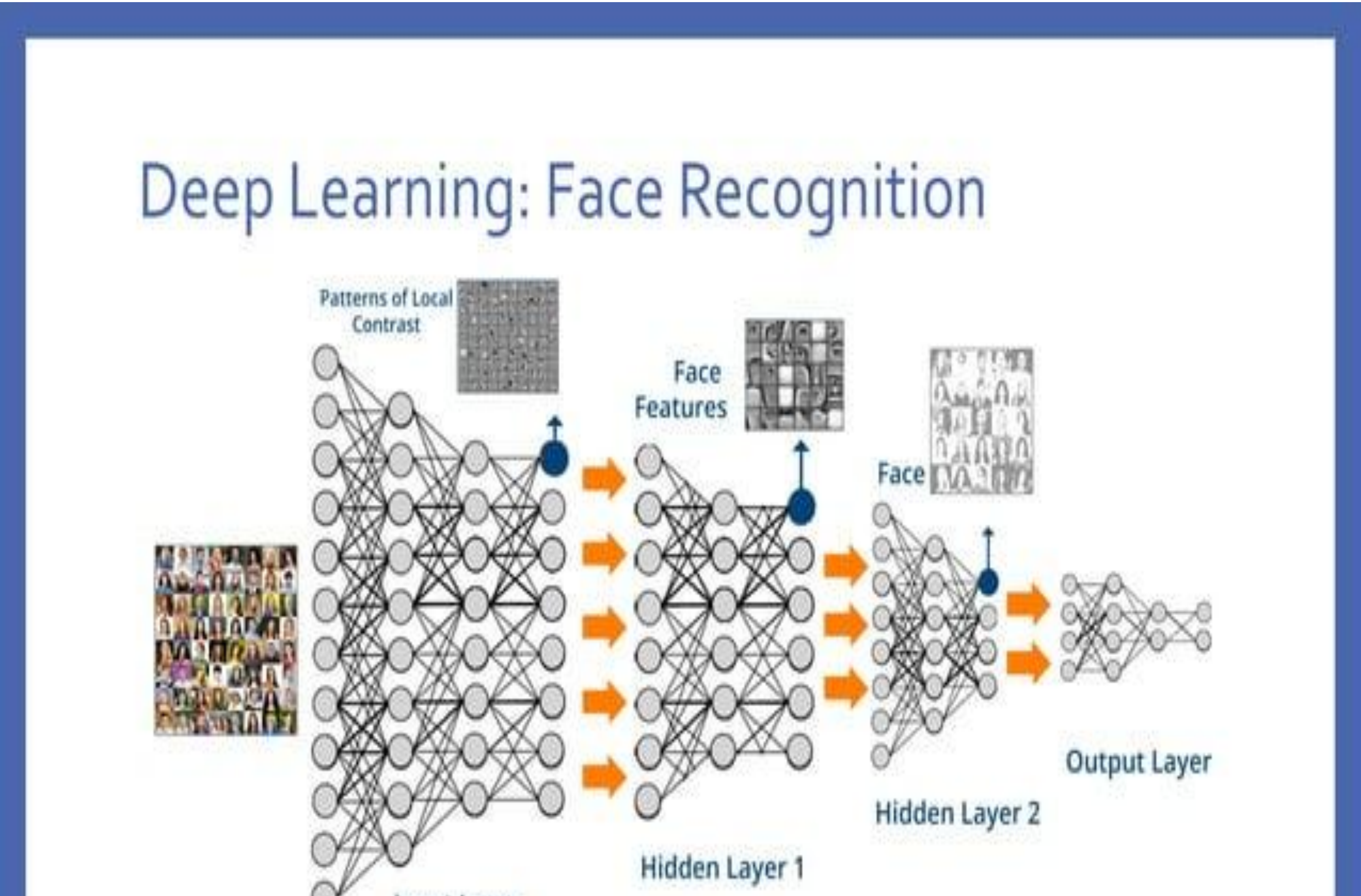


- **Supervised Learning:** develop model to *predict* output based on *existing* input-output (done by human)
  - **Classification**: Is this a cat or not?
  - **Regression**: Will user click on this ad?
- **Unsupervised Learning:** group & interpret data based on input only
  - **Clustering**: Identify patterns that are not obviously visible.
- **Reinforcement Learning:** actions to maximize "rewards"
  - Recommendations on shopping websites (Amazon), videos (Netflix)
  - Computer vs human gaming. Chess (IBM Watson). Go (Google's AlphaGo).

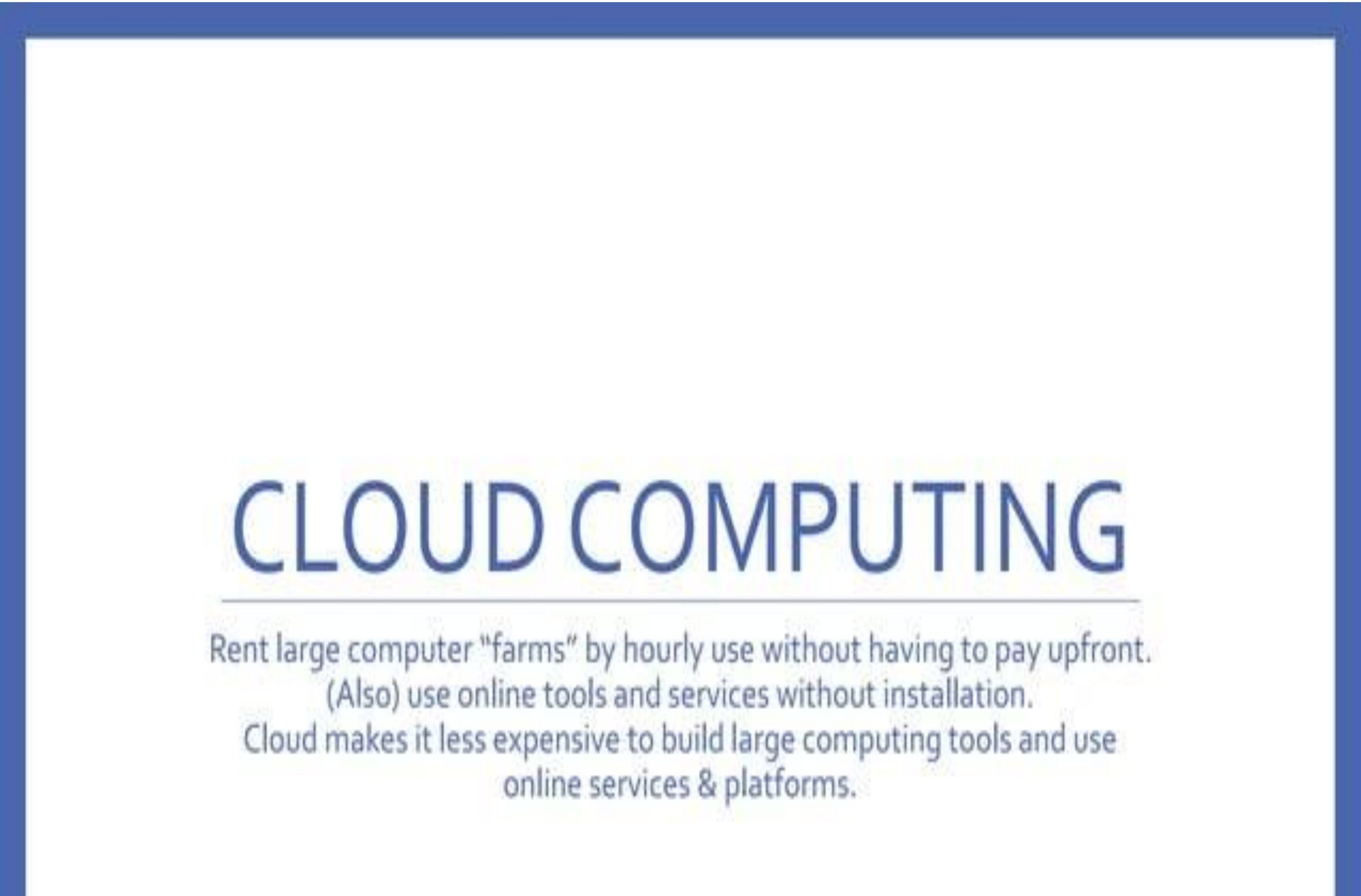
Deep Learning



- ML requires human input to train and classify
- Deep Learning uses multiple levels of CNNs (Convolutional Neural Networks) to learn by itself
- But: setting up & programming deep learning neural network is hard.



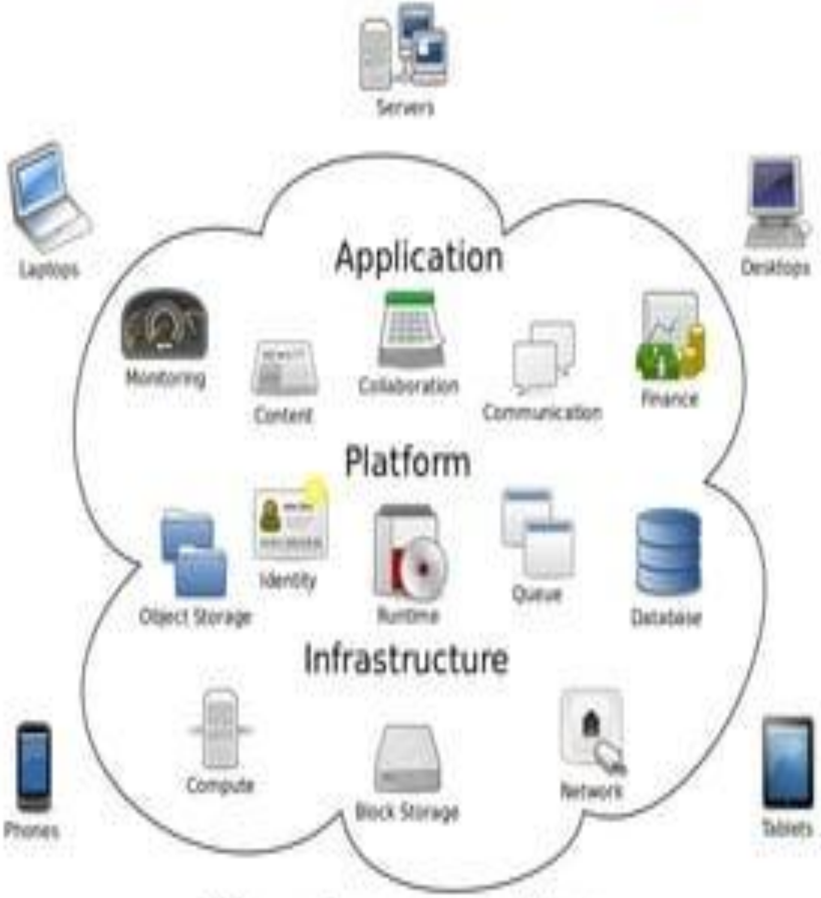
Deep Learning: Face Recognition



CLOUD COMPUTING

Rent large computer "farms" by hourly use without having to pay upfront.
(Also) use online tools and services without installation.
Cloud makes it less expensive to build large computing tools and use online services & platforms.

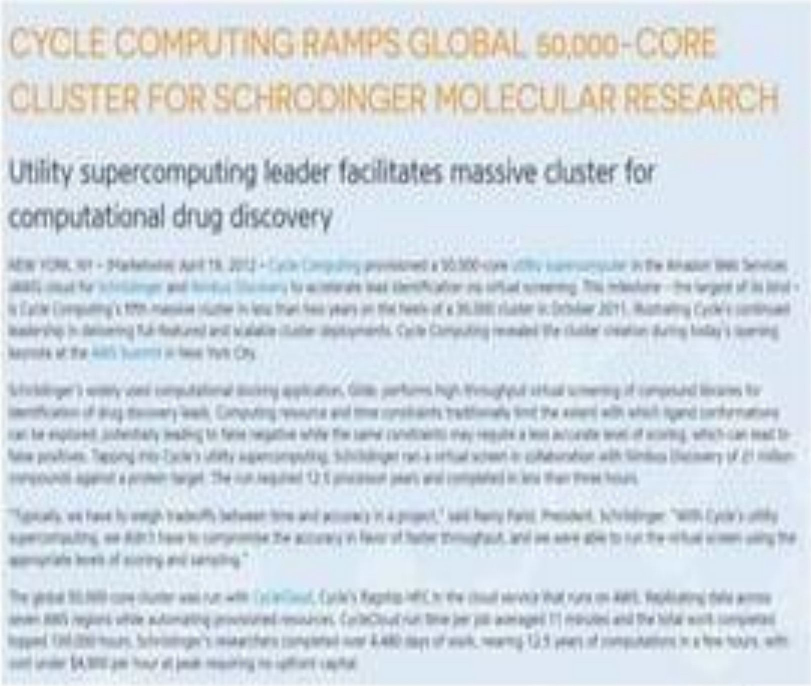
Cloud Computing



- Run computer services on a "cloud"
  - Remote location run by service provider
- **IaaS = Infrastructure** as a service
  - "Rent" computers, storage, networking. Install your own software.
- **PaaS = Platform** as a service
  - Higher level services, such as databases, web servers, etc. Web hosting.
- **SaaS = Software** as a service
  - Run software from the cloud. Websites, online applications, e-mail, IM / messaging,

Cloud Computing -> Data Science

- "Rent" computing servers instead of buying outright
  - Zero setup time; no setting up hardware
- "… building a 50,000 core cluster could easily cost $20 million to $30 million, he said. The Schrödinger project, by contrast, cost about $4,850 per hour to run."
  - GigaOm: Cycle Computing spins up 50K core Amazon cluster
- "By leveraging Cycle Computing software and AWS Cloud infrastructure, Novartis was able to accomplish the same work faster, and for far less money.
  - $44 Million in infrastructure; 10 Million compounds screened; 39 drug design years in 11 hours for a cost of $4,232; 3 compounds identified for future work"
  - Chef: Novartis Conducts 39 Yrs of Computing in 11 Hours w/Cycle Computing and Chef



CYCLE COMPUTING RAMPS GLOBAL 50,000-CORE CLUSTER FOR SCHRODINGER MOLECULAR RESEARCH

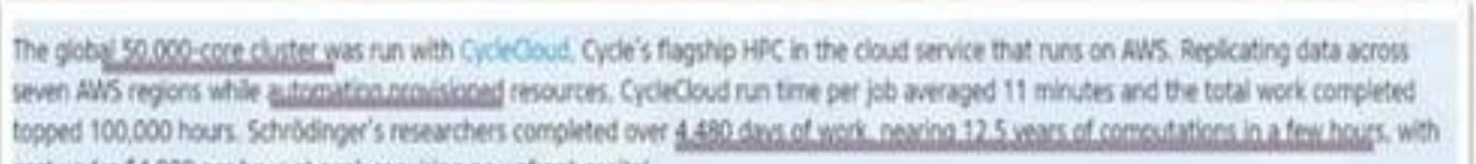
Utility supercomputing leader facilitates massive cluster for computational drug discovery

NEW YORK, NY – (Marketwire) April 19, 2012 – Cycle Computing provisioned a 50,000-core utility supercomputer in the Amazon Web Services (AWS) cloud for Schrödinger and Nimbus Discovery to accelerate lead identification via virtual screening. This milestone – the largest of its kind – is Cycle Computing's fifth massive cluster in less than two years on the heels of a 30,000 cluster in October 2011, illustrating Cycle's continued leadership in delivering full-featured and scalable cluster deployments. Cycle Computing revealed the cluster creation during today's opening keynote at the AWS Summit in New York City.

Schrödinger's widely used computational docking application, Glide, performs high throughput virtual screening of compound libraries for identification of drug discovery leads. Computing resource and time constraints traditionally limit the extent with which ligand conformations can be explored, potentially leading to false negative while the same constraints may require a less accurate level of scoring, which can lead to false positives. Tapping into Cycle's utility supercomputing, Schrödinger ran a virtual screen in collaboration with Nimbus Discovery of 21 million compounds against a protein target. The run required 12.5 processor years and completed in less than three hours.

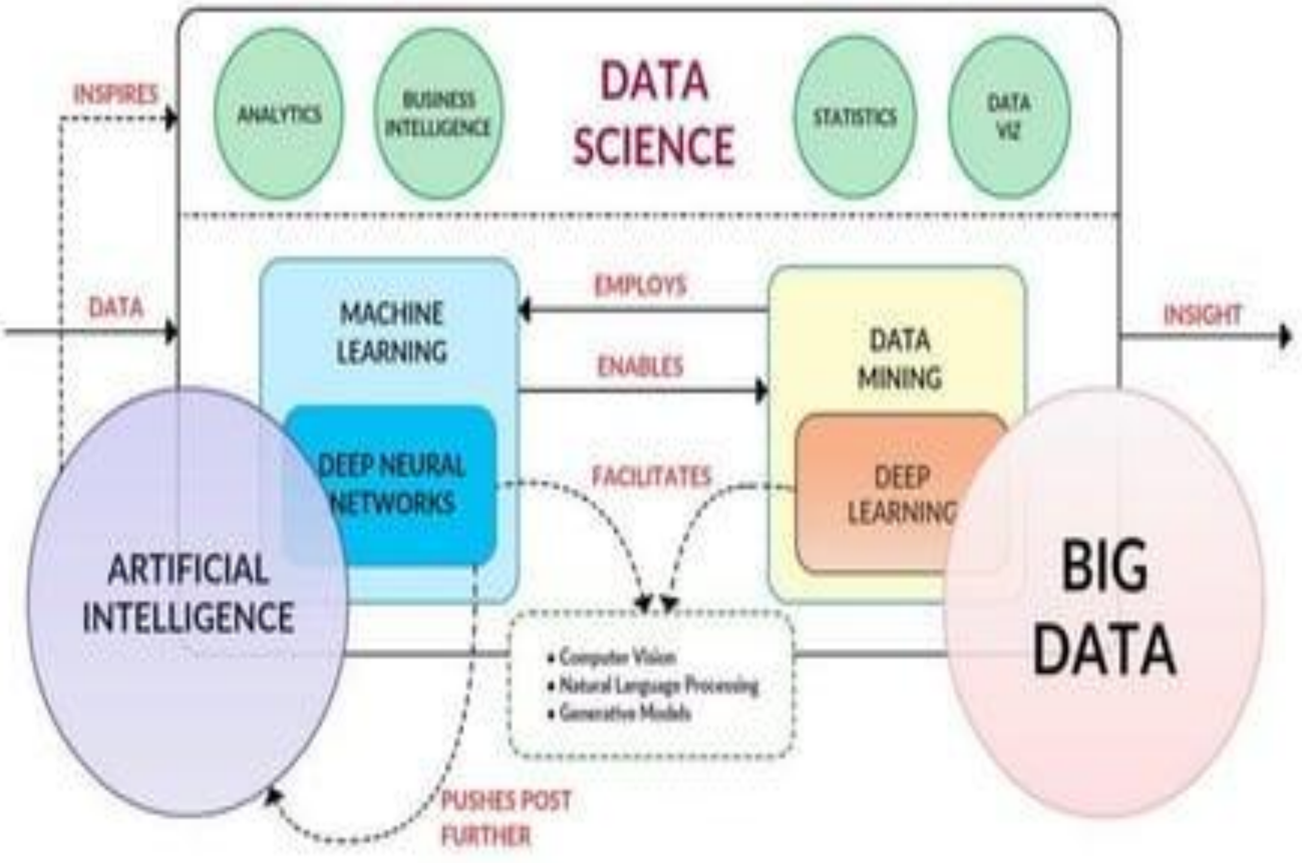
"Typically, we have to weigh tradeoffs between time and accuracy in a project," said Ramy Farid, President, Schrödinger. "With Cycle's utility supercomputing, we didn't have to compromise the accuracy in favor of faster throughput, and we were able to run the virtual screen using the appropriate levels of scoring and sampling."

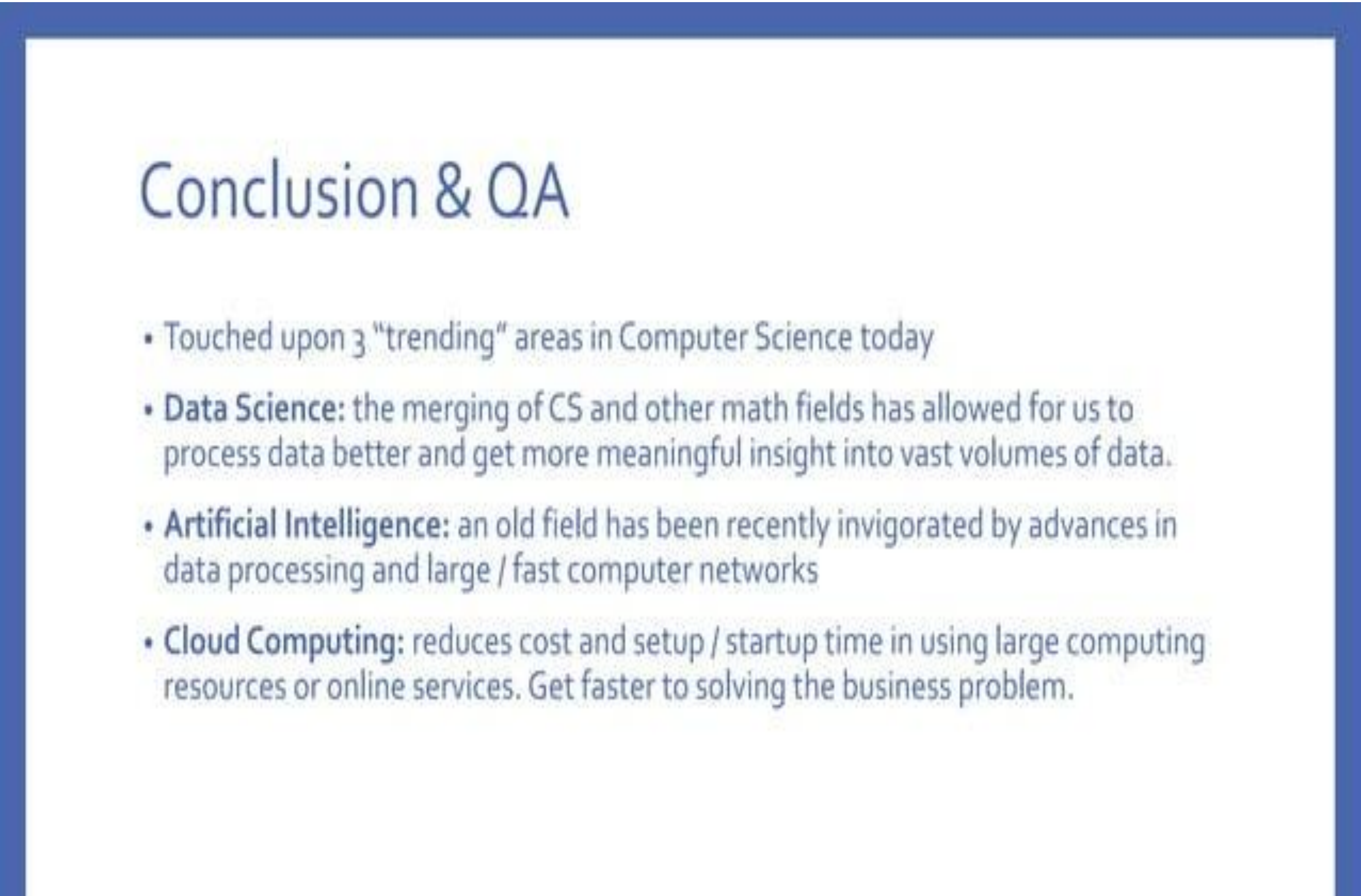
The global 50,000-core cluster was run with CycleCloud, Cycle's flagship HPC in the cloud service that runs on AWS. Replicating data across seven AWS regions while automating provisioned resources. CycleCloud run time per job averaged 11 minutes and the total work completed topped 100,000 hours. Schrödinger's researchers completed over 4,480 days of work, nearing 12.5 years of computations in a few hours, with cost under $4,900 per hour at peak requiring no upfront capital.



The global 50,000-core cluster was run with CycleCloud, Cycle's flagship HPC in the cloud service that runs on AWS. Replicating data across seven AWS regions while automation provisioned resources. CycleCloud run time per job averaged 11 minutes and the total work completed topped 100,000 hours. Schrödinger's researchers completed over 4,480 days of work, nearing 12.5 years of computations in a few hours, with

Data Science <-> AI / ML





Conclusion & QA

- Touched upon 3 "trending" areas in Computer Science today
- **Data Science:** the merging of CS and other math fields has allowed for us to process data better and get more meaningful insight into vast volumes of data.
- **Artificial Intelligence:** an old field has been recently invigorated by advances in data processing and large / fast computer networks
- **Cloud Computing:** reduces cost and setup / startup time in using large computing resources or online services. Get faster to solving the business problem.