

Dave Beulke
Principal Consultant Pragmatic Solutions, Inc.

 @DBeulke

 davebeulke

Best Performance and Design Practices for Analytic, ML and AI Applications

I am honored to have been a presenter at all 30 years of IDUG

think 2019

2018 – Malta/Philadelphia/San Francisco

-Best Design and Performance Practices for Analytics

2017 – Anaheim -Understand IDAA Performance and Justify an IDAA Appliance

2016 – Austin Performance Enterprise Architectures for Analytic Design Patterns
How to do your own Db2 Security Audit

2015 - Valley Forge Db2 Security Practices

Big Data Performance Analytics Insights

2014 – Phoenix Big Data SQL Considerations

2013 – Orlando Big Data Disaster Recovery Performance

2012 – Denver Agile Big Data Analytics

2011 – Anaheim Db2 Temporal Tables Performance Designs

2010 - Tampa - Java DB2 Developer Performance Best Practices

2009 – Denver -Java Db2 Perf with pureQuery and Data Studio
Improve Performance with Db2 Version 9 for z/OS

2008 – Dallas - Java pureQuery and Data Studio Performance

2007 - San Jose - Developing High Performance SOA Java Db2 Apps
Why I want Db2 Version 9

2006 - Tampa - Class - How to do a Db2 Performance Review
Db2 Data Sharing

Data Warehouse Designs for Performance

2005 – Denver - High Performance Data Warehousing

2004 – Orlando – Db2 V8 Performance

President of IDUG

2003 - Las Vegas - Db2 UDB Server for z/OS V8 Breaking all the Limits
Co-author IBM Business Intelligence Certification Exam

2002 - San Diego - Db2 UDB for LUW 8 - What is new in Db2 Version 8
Data Warehouse Performance

2001 – Orlando -Data Sharing Recovery Cookbook
Designing a Data Warehouse for High Performance
Co-authored the first IBM Db2 z/OS Certification Exam

2000 – Dallas - Db2 Data Warehouse Performance Part II

1999 – Orlando - Store Procedures & Multi-Tier Performance
Developing your Business Intelligence Strategy
Evaluating OLAP Tools

1998 - San Francisco - Db2 Version 6 Universal Solutions
Db2 Data Warehouse Performance
Db2 & the Internet Part II

1997 – Chicago - Db2 & the Internet

1996 – Dallas- Sysplex & Db2 Data Sharing
Best Speaker Award at CMG Conference Mullen Award

1995 – Orlando/Sydney - Practical Performance Tips
Improving Application Development Efficiency

1994 - San Diego - Database Design for Time Sensitive Data &
Guidelines for Db2 Column Function Usage

1993 – Dallas - High Availability Systems: A Case Study &
Db2 V3: A First-Cut Analysis

1992 - New York -Db2 –CICS Interface Tuning

1991 - San Francisco - Pragmatic Db2 Capacity Planning for DBAs

1990 – Chicago - Performance Implication of Db2 Design Decisions

1989 – Chicago - Db2 Performance Considerations



Dave@davebeulke.com

- Member of the inaugural IBM Db2 Information Champions
- One of 40 IBM Db2 Gold Consultant Worldwide
- President of DAMA-NCR
- Past President of International Db2 Users Group - IDUG
- Best speaker at CMG conference & former TDWI instructor
- Former Co-Author of certification tests
 - Db2 DBA Certification tests
 - IBM Business Intelligence certification test
- Former Columnist for IBM Data Management Magazine
- Extensive experience in Big Data systems, DW design and performance
 - Working with Db2 on z/OS since V1.2
 - Working with Db2 on LUW since OS/2 Extended Edition
 - Designed/implemented first data warehouse in 1988 for E.F. Hutton
 - **Syspedia** for data lineage and data dependencies since 2001 –
 - Find, understand and integrate your data faster!

Proven Performance Tips:
www.DaveBeulke.com

➤ Consulting

- **Security Audit & Compliance**
- **Db2 Performance Review**
- **CPU MLC Demand Reduction**
- **Analytics & Database Design Review**
- **Db2 12 Migration Assistance**
- **Java Application Performance Tuning**

➤ Educational Seminars

- **Java Security for Application Developers**
- **Db2 Version 12 Transition**
- **Db2 Performance for Java Developers**
- **Data Warehousing Designs for Performance**
- **How to Do a Java Performance Review**

World hasn't changed-the same w/new names ML-AI **think** 2019

- DW analysis
- Software Releases
- Production release
- Programs
- Files/Databases
- Documentation
- Project Plan
- Report
- Data Stewards
- Maintenance

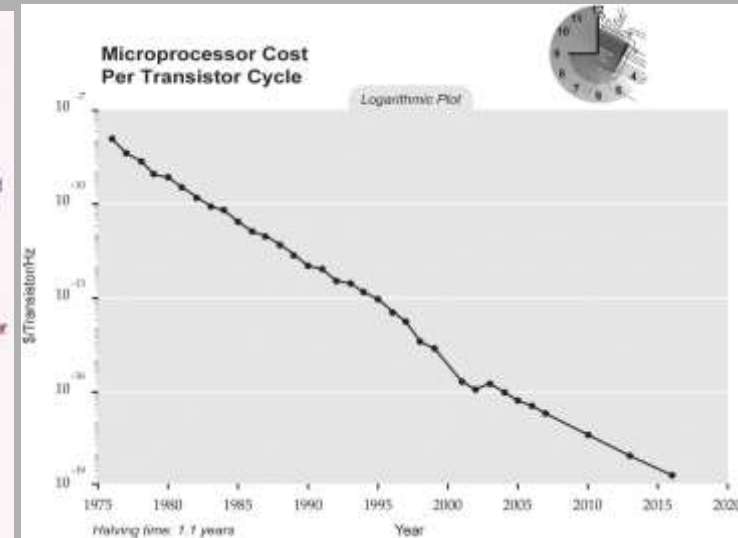
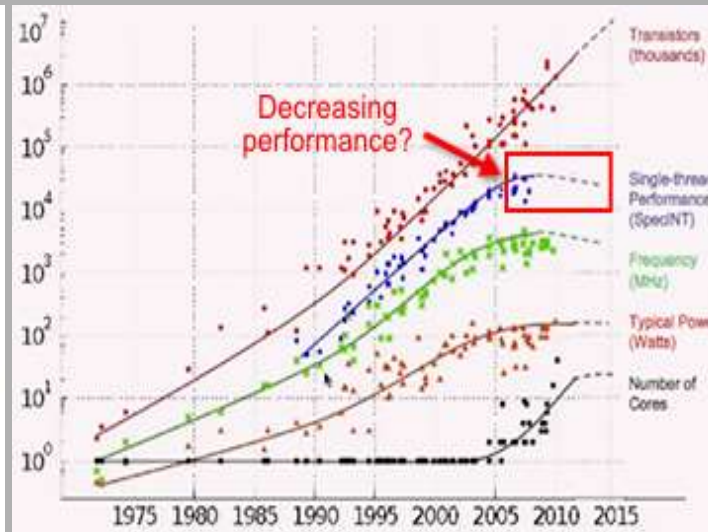
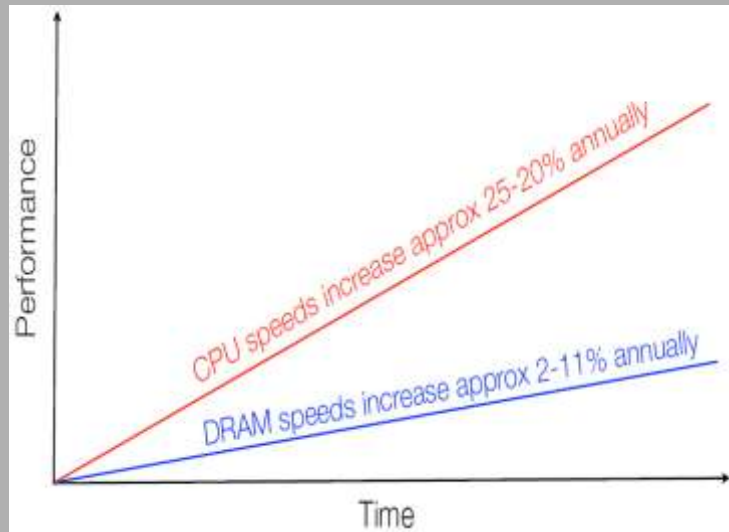


- Overlapping experiments
- Continuous builds/integration
- Confidence interval release
- Automated services/APIs
- Unstructured/data lake/fluid
- Scoring model wiki
- Infrastructure updates
- Output as Input
- String Indexer meetings
- Github fork

Cloud, ML and AI changed the names-still depend on good data management practices!

Moore's Law, commodity everything and Cloud

- Cost of the hardware/software is non issue
 - Open source, elastic pricing pay for only what is used





- **Harvard's Data Science Course**
- End-to-end data science course. While there's less emphasis on ML than in Andrew Ng's course, you'll get more practice with the entire data science workflow from data collection to analysis
 - [Course Homepage](#)
[Lecture Videos and Slides](#)
[Homework Assignments](#)



- **Stanford's Machine Learning Course**
- This is the famous course taught by Andrew Ng, and it's the gold standard when it comes to learning machine learning theory. These videos really clear up the core concepts behind ML.
If you only have time for 1 course, I recommend this one.
 - [Course Videos](#)

Analysis of every situation - track everything

Scene from the book '*Snow Crash*' by Neal Stephenson (required reading at Facebook)

- Novel shows a world where the internet is replaced with the *Metaverse*, a shared virtual reality
- Y.T.'s mom reads a memo on a cost-saving program and is tracked by her employer

How long did it take her to read that memo? If.....

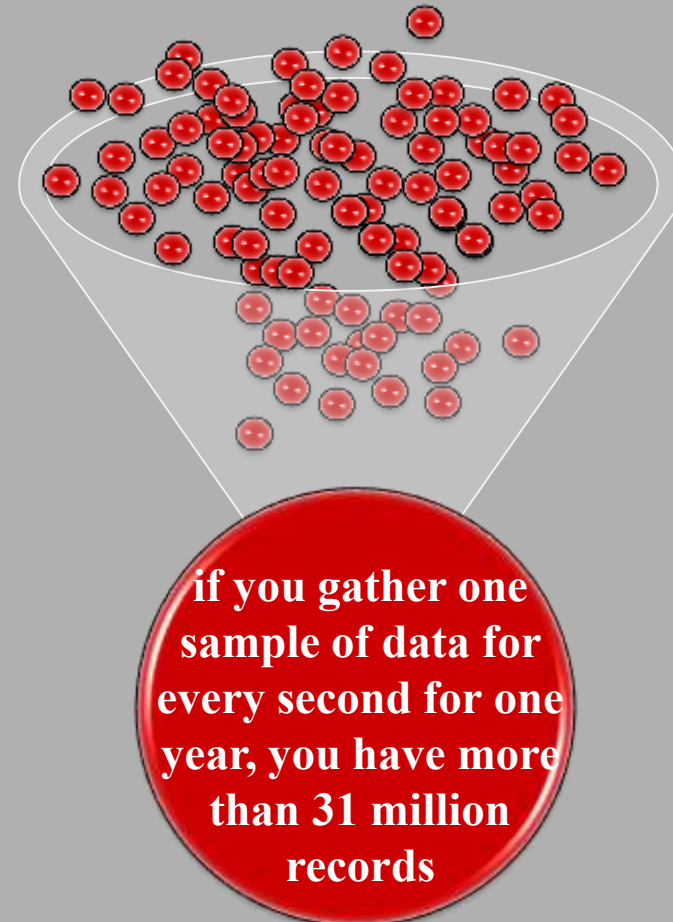
- **Less than 10 min.:** Time for an employee conference and possible attitude counseling.
- **10-14 min.:** Keep an eye on this employee; may be developing slipshod attitude.
- **14-15.61 min.:** Employee is an efficient worker, may sometimes miss important details.
- **Exactly 15.62 min.:** Smartass. Needs attitude counseling because that is exactly correct.
- **15.63-16 min.:** Not to be trusted.
- **16-18 min.:** Employee is a methodical worker, may sometimes get hung up on minor details.
- **More than 18 min.:** Check the security videotape, see just what this employee was up to (e.g., possible unauthorized restroom break).



Cloud Security

- Technology bandwidth –
 - Regulatory drivers
 - Use it for DR and
 - Drive business growth competition
- Security impacts
 - PII, HIPPA, Masking, Encryption etc....
- Framework for business continuity
 - Physical to VM
 - VM to Physical
 - VM to VM
 - Logical sync point - Local or remote
 - File or transaction
 - Requirements/Money/Technology

IoT - Sensor Data



Best Practice for analytical database design

- Delineate physical objects I/O bound operations

- Partition the database tables to minimize the data required for daily SQL
- Use the thousands of partitions available for a design
 - How many parallel processes are your applications running today?

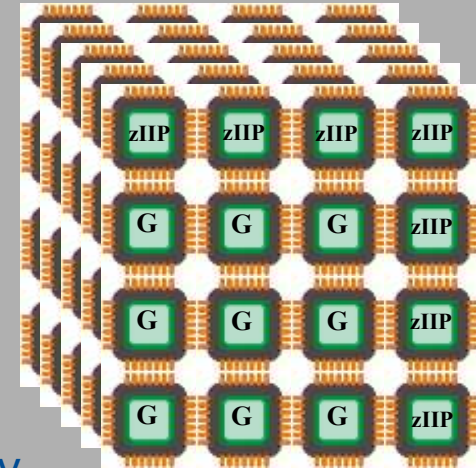
- Separate old data from new data

- Current Year, Quarter, Week, Day
- Temporal tables with the HISTORY tables
 - Complicates the SQL also can make a lot of data quickly
- Materialized Query Table – YTD sales figures
 - Or composite tables to separate via TIME axis Year, AP, Quarter, Week, Day
 - Or composite tables to separate via Sales territory axis Country, Region, State, City, Zip code



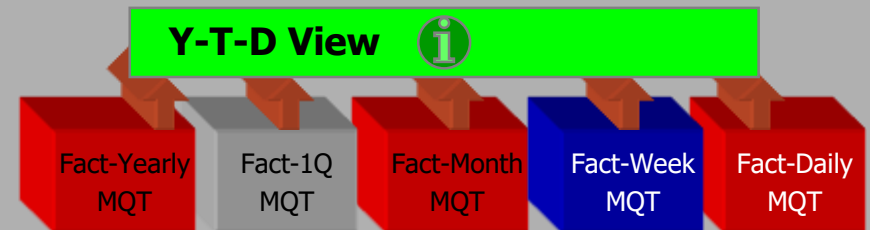
Leverage Free zIIP CPU thru parallelism

- Partitioning to leverage **free zIIP** more parallel processes
 - Same partitioning limit keys across multiple table spaces
 - Via Customer number across those related tables
 - Via Product SKU number across all the product related data
- Partitioning design leverages customer, product or time properly
 - The active partitions are only a segment of the entire table
 - Concentrates the I/O into the right sized portion of the database
 - Current history available - Ancient history is in database as AOT/archived easily
- Indexes (NPIs or DPSIs) are appropriately designed
 - Partitioned for parallelism and recovery time objectives (RTO)
 - Table clustered for SQL efficiencies



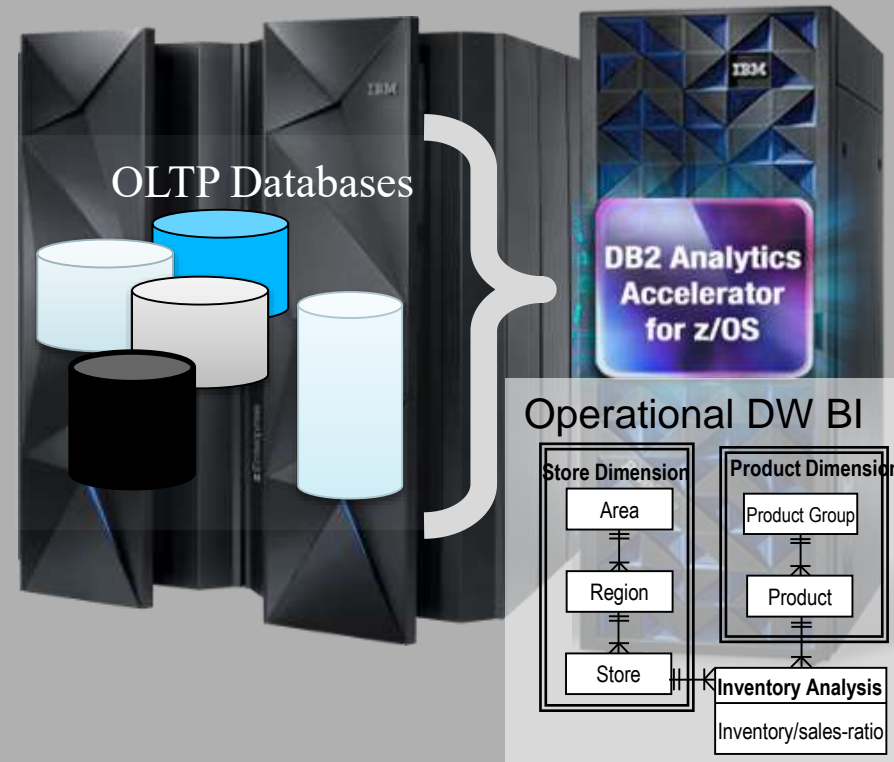
Best practice use MQT – 10 to 1000 times improvement!

- 5B rows per year–10 per 4k page= ½B pages
- MQT aggregates save large amounts of everything
 - Create aggregates for every possibility
 - “On Demand” information
 - Sales by department
 - Sales by zip code
 - Sales by time period – day/week/month/quarter/AP
 - All reporting and analysis areas
 - Trace usage to create/eliminate aggregates
- Total by month ½B I/Os versus 12 I/Os

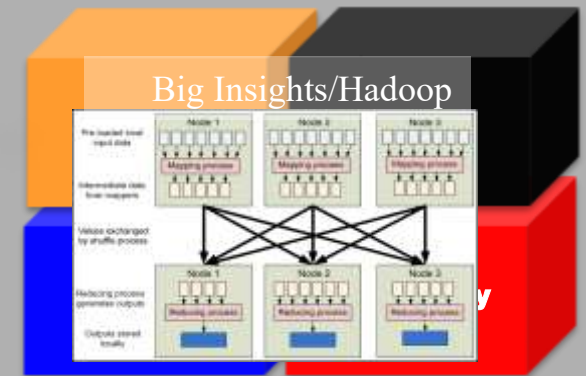


Big Data HTAP Architecture

- Capacities- I/O & CPU
- Latency/Integrity
- All APIs & interfaces
 - JSON
- Security
- Disaster Recovery

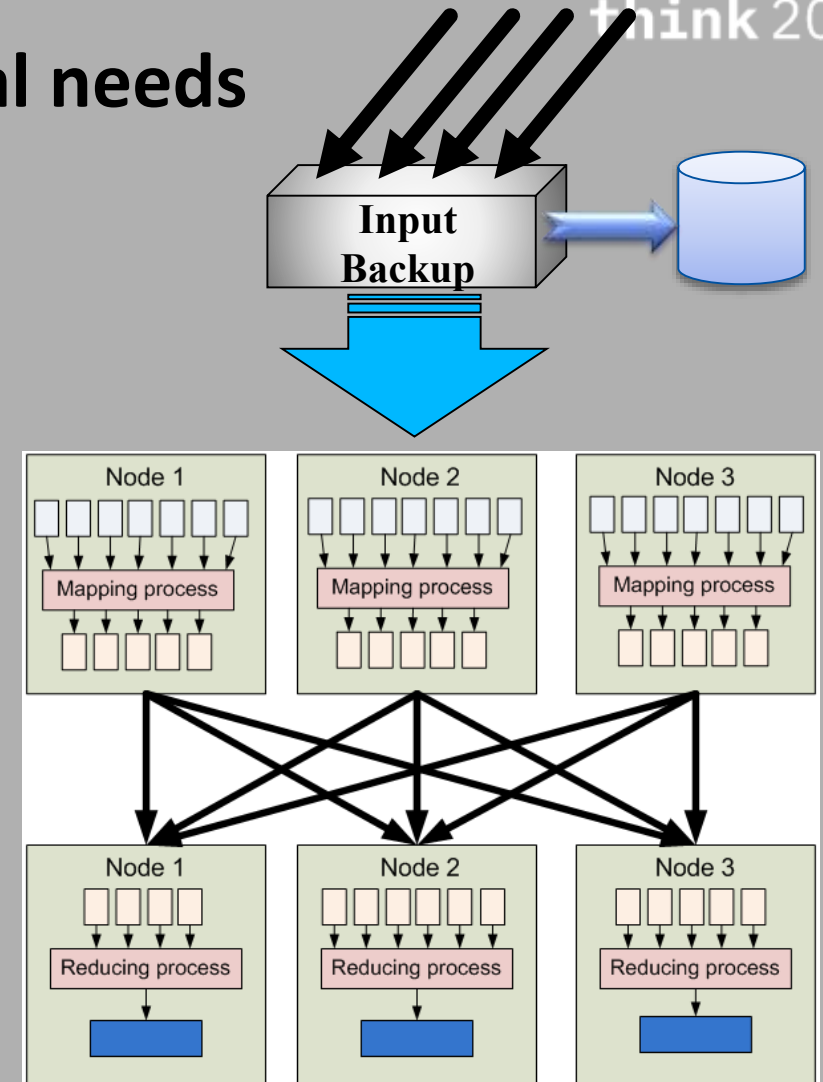


Cloud!



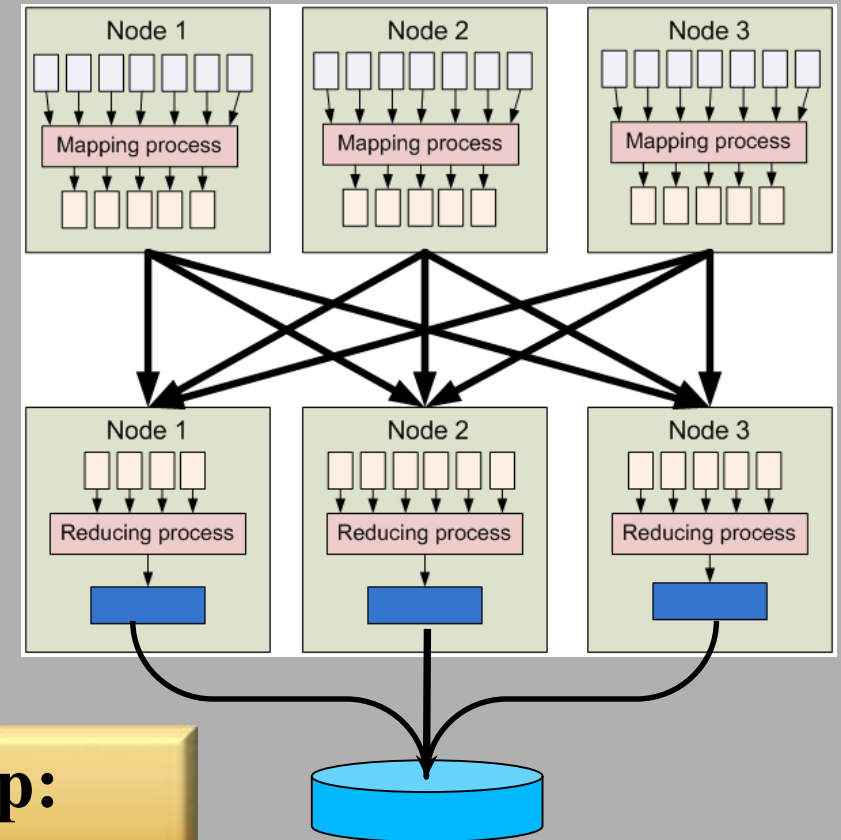
Hadoop's Disaster Recovery - special needs

- Hadoop Three Sections
 - Data, System & Configuration
- Biggest is input data
 - **Data is REPLICATED to 2 or 3 nodes**
 - Compression considerations
 - Best: backup as it comes in
 - Disk, Node, Rack, Site failures
 - Standard DR
- System/Application(s)
 - Frequent regular backups
- Configuration
 - Frequent regular backups



Hadoop's Disaster Recovery - special needs

- Community is working on HDFS Snapshot capabilities
 - Maprfs – Amazon
 - Provides snapshots
 - Namenode single point
 - Dual nodes heartbeat sync
- Map Reduce Output(s)
 - Standard DR Backups
 - Transform to Archive
 - Standard reports/files

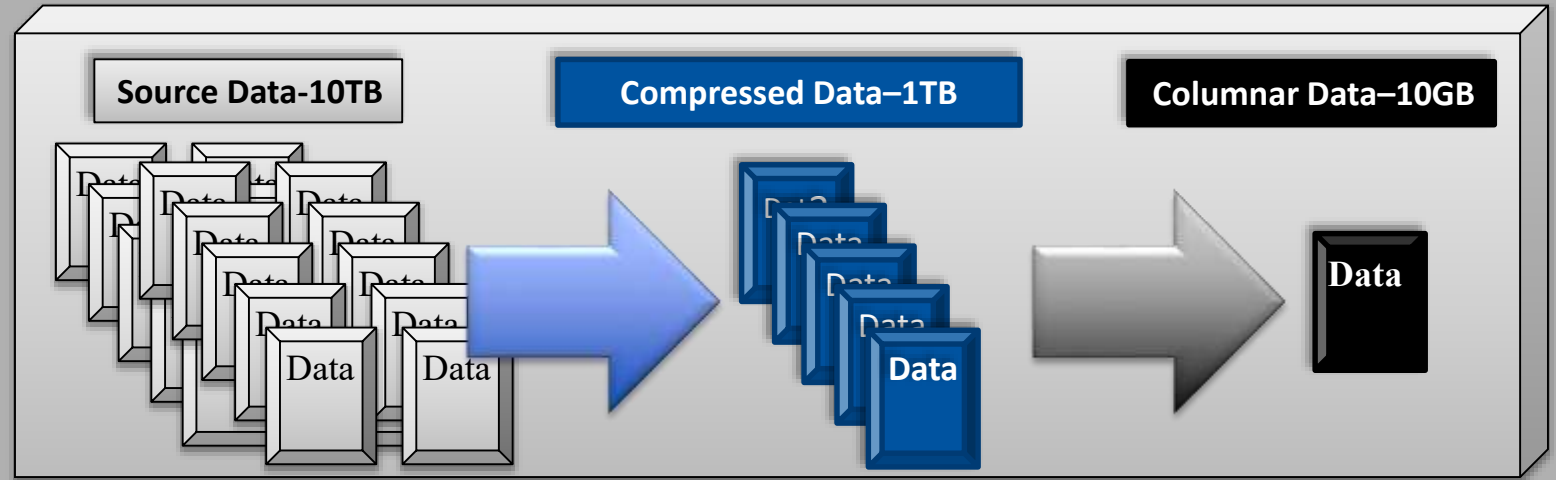


**Best backup Hadoop:
Do only Config, Inputs, Outputs**

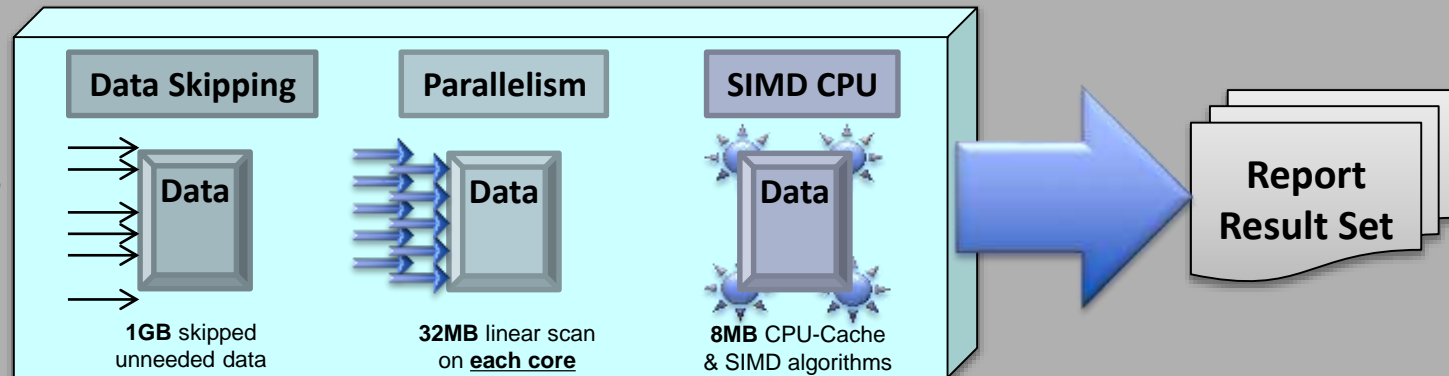
Db2 BLU – Columnar Data Store

think 2019

- 10TB to 10GB for data at rest
 - Extreme Compression
 - 1/1000th of the storage space

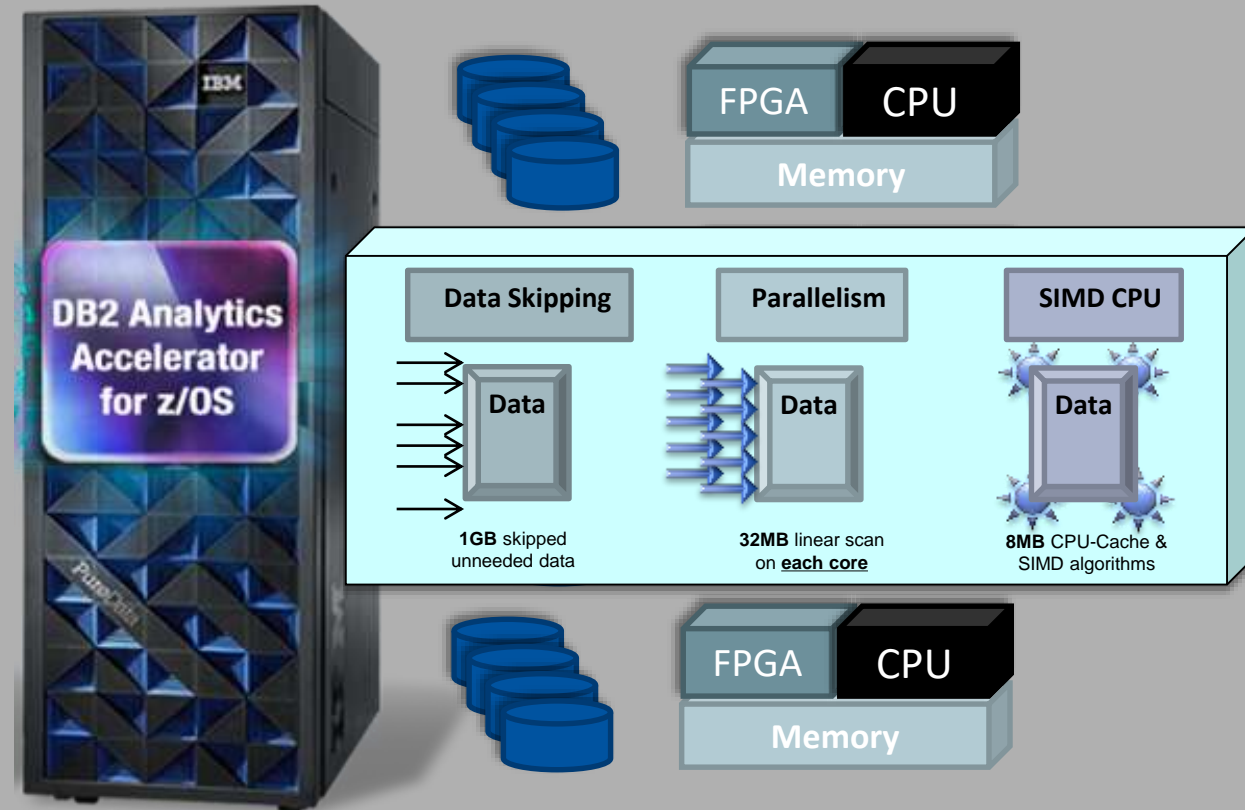


- Processing Data Set
 - Data Skip reduces I/O
 - Parallelism of 32MB linear scans
 - SIMD Cache operates on data



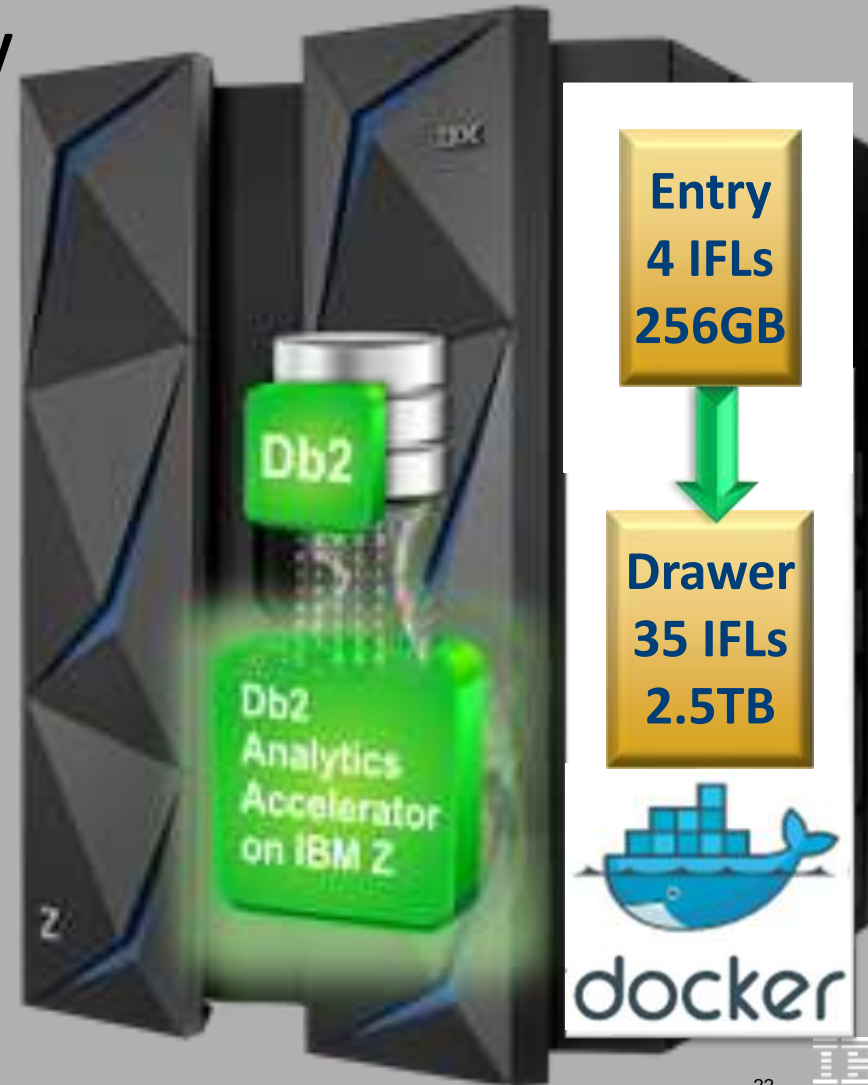
Db2 combined with IDAA

- Db2 routing SQL to IDAA
- Hardware Processing Speed
- Processing minimizes data
- Table level customization
- Great AOT options also
- DR is fast & easily resolved



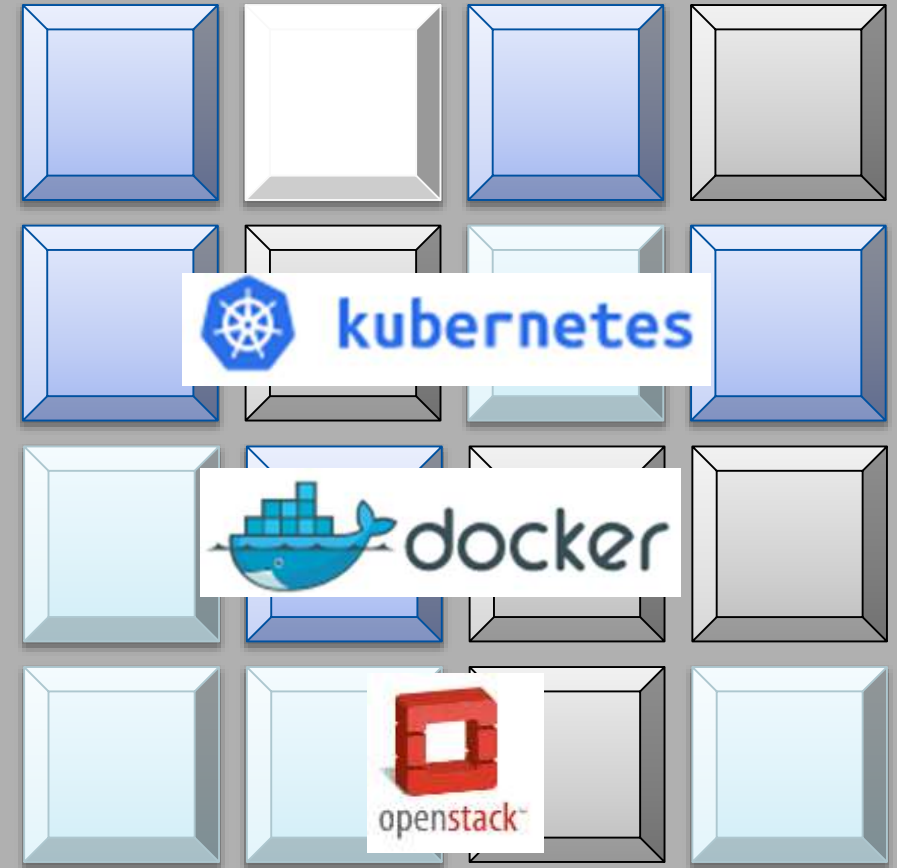
Best practices leverage the latest technology

- Enhancement for IDAA
 - Further integration of HTAP features for Db2
 - Better within the large memory space of SystemZ
- Can be configured just like any IDAA
- Direction of IDAA
 - Same private network connections
- Provides growth path for Accelerator
- Make sure to order z14 with IDAA
 - Need to make sure that “*Special Bid Pricing*” can include an Accelerator in your new z14



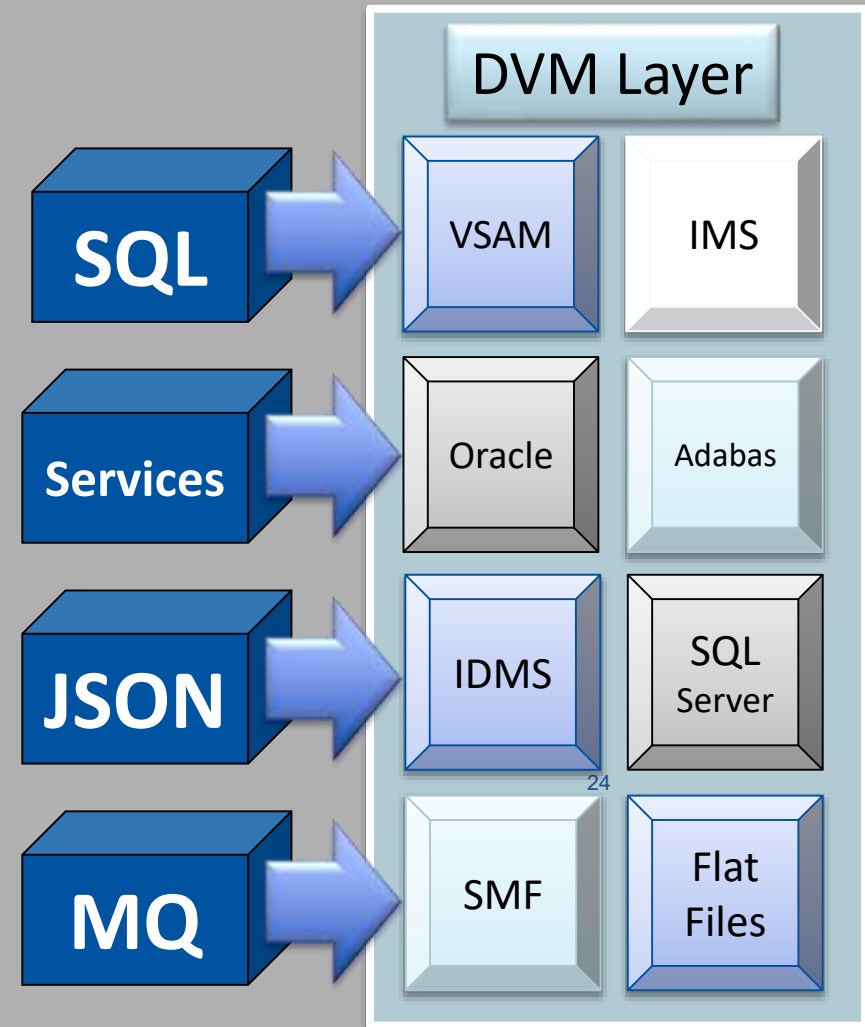
Data Grid, In memory or containers

- Cache Size
 - Cache expiration
 - Independent updates to the underlying data store
 - Synchronous or asynchronous updates
- Consistent Client view of your data
 - How to do scale up, replication and failover?
 - Which container provides best security/performance?
 - New OpenStack Kata, Docker or kubernetes?
- There is a huge complexity cost of cache/container management!



Data Virtual Manager - DVM

- Relational SQL access for any data store
 - VSAM
 - IMS
 - IDMS
 - SMF or flat files
 - SQL Server
 - Oracle
 - Adabas
- Provides SQL, JSON & RESTful services interface
- MQGet/MQPut messages access virtual layer
- SQL JOINS with non-relational data stores!



Best practice leverage Watson ML AI partner resources

- Automatic data relationship discovery
 - Automatically reviews and develops the relationships within the data
 - If you are subscribed to the Professional edition or the Plus edition of Watson
- Analytics, you have access to more types of data:
 - Cognos® BI reports
 - Databases such as IBM Db2®, IBM dashDB®, IBM SQL Database for Bluemix®,
 - Microsoft SQL Server, MySQL, Oracle, PostgreSQL
- Best practice: Determine the profits from questions and answers
- Free Db2 on cloud trial: <https://www.ibm.com/cloud/db2-on-cloud>

Any ML and AI types already used?

- What analysis type is best for your business problem/optimization?

Supervised Learning

- Using a given set of variables a function is generated that maps the inputs to the desired outputs executed until model achieves a desired level of accuracy

Unsupervised Learning

- Used for clustering population in different data groups which is widely used for segmenting customer in different groups for specific intervention
- No target or outcome variable to predict or estimate

Reinforced Learning

- Tries to capture the best possible knowledge acquired from past experience to make accurate business decisions
- The machine is trained to make specific decisions where it is exposed to an environment and makes the best possible decisions through trial and error.

- ML on Watson versus other data stores

- Performance matters & Data Refresh time matters

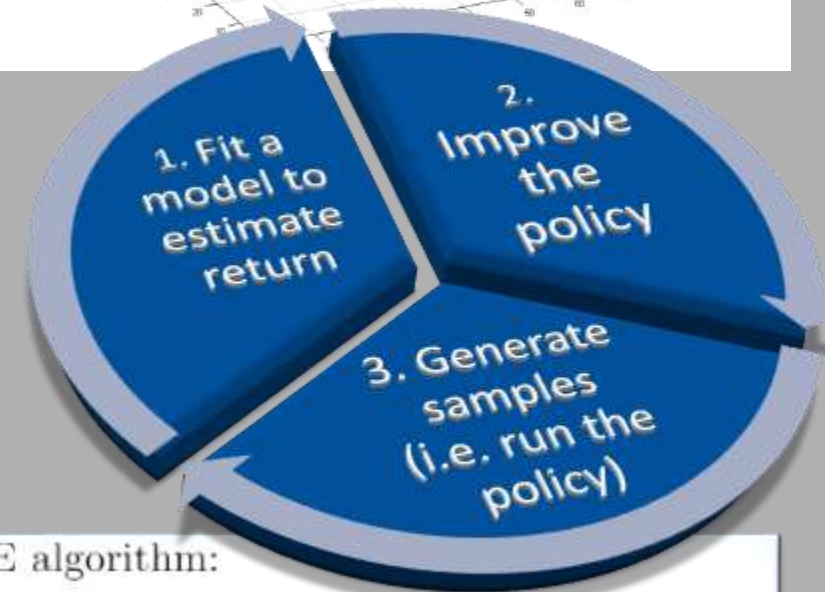
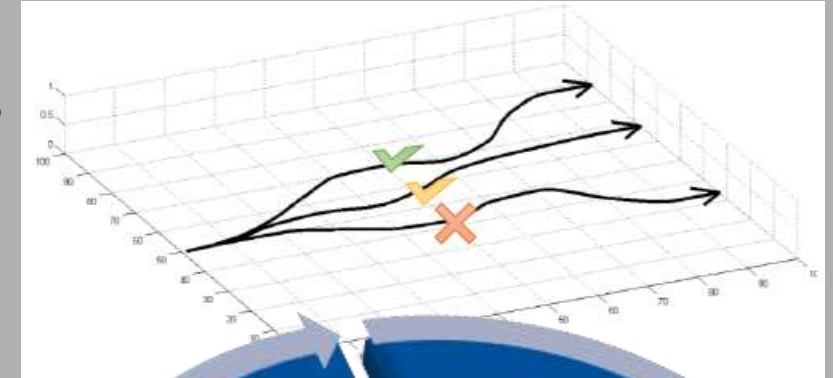
ML initiatives mean different ideas within the **same** meeting

- Encounter skepticism and resistance to change when implementing AI and machine learning
- Which type of algorithm?
 - Iterative Algorithms
 - What algorithm type & formula best?
 - What algorithms is the business use today?
- Identify patterns in systems
 - feedback loop so that previous recommendations are input to improve the next recommendations
- SQL **WHERE** Filters → train ML algorithms for desired outcome
 - **How big is your required sample size?**



ML and AI Challenges

- Redesign the accountabilities and verify business is prepared to consume ML/AI conclusions
- Applying machine learning technologies, choose one with measurable results and economic effect
- You **can't** do analytics...
 - Without a good understanding of the data requirements
 - Without the best algorithm for the business situation
 - Without a solid high performance data infrastructure
 - AI without machine learning
- Machine learning systems can predict fraud and have gotten sophisticated enough to detect when behavior deviates

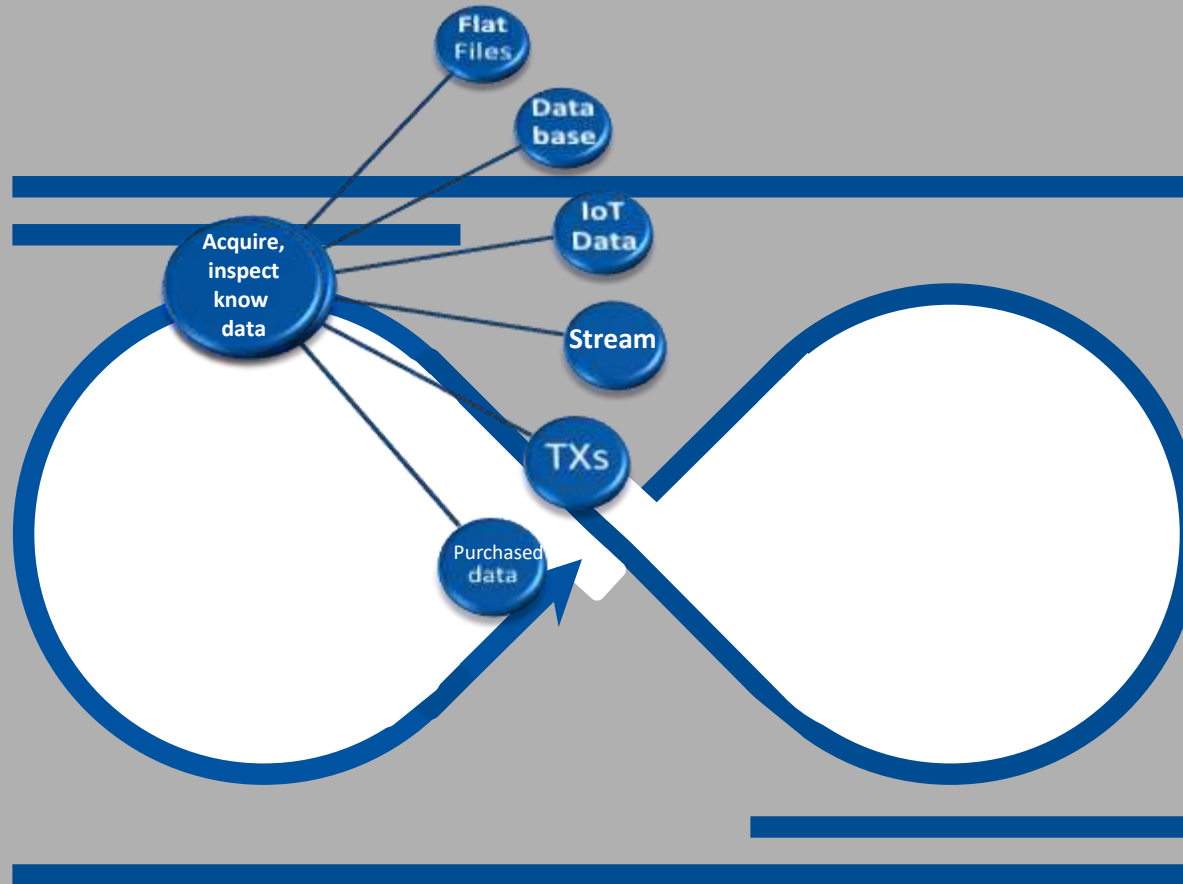


REINFORCE algorithm:

1. sample $\{\tau^i\}$ from $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ (run the policy)
2. $\nabla_\theta J(\theta) \approx \sum_i (\sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i|\mathbf{s}_t^i)) (\sum_t r(\mathbf{s}_t^i, \mathbf{a}_t^i))$
3. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

ML & AI Complex Ecosystems

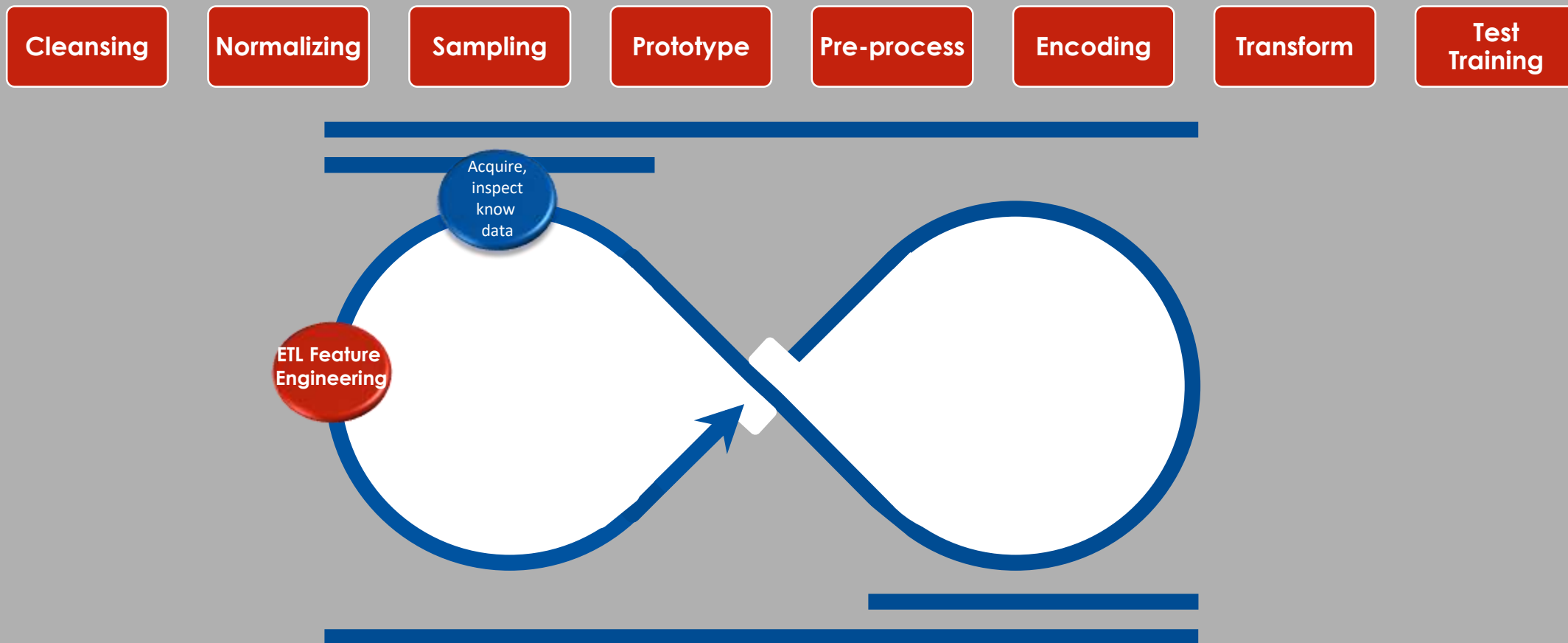
- Know the data context – time grain & domains



ML & AI Complex Ecosystems

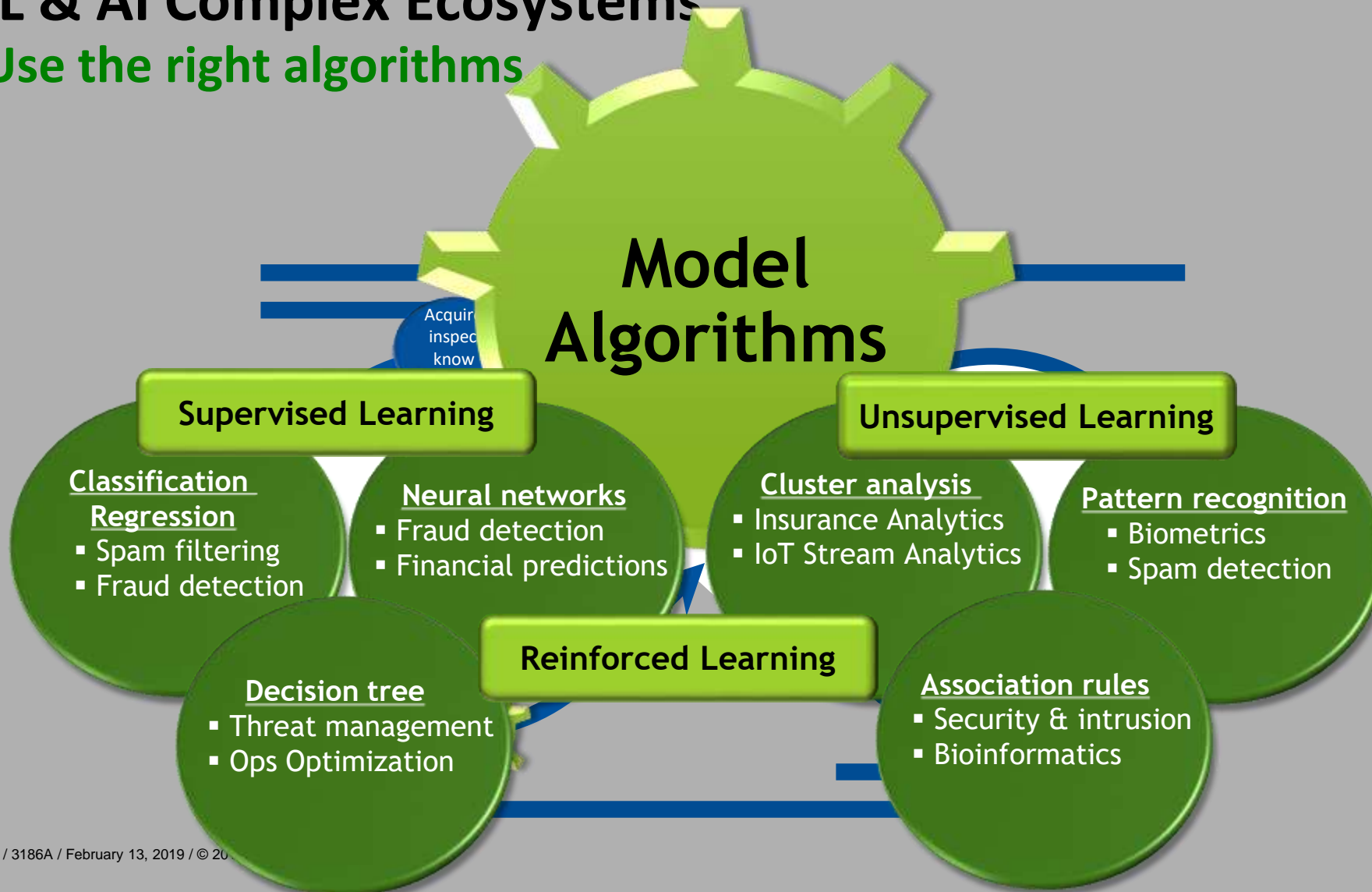
think 2019

- Preparing the data



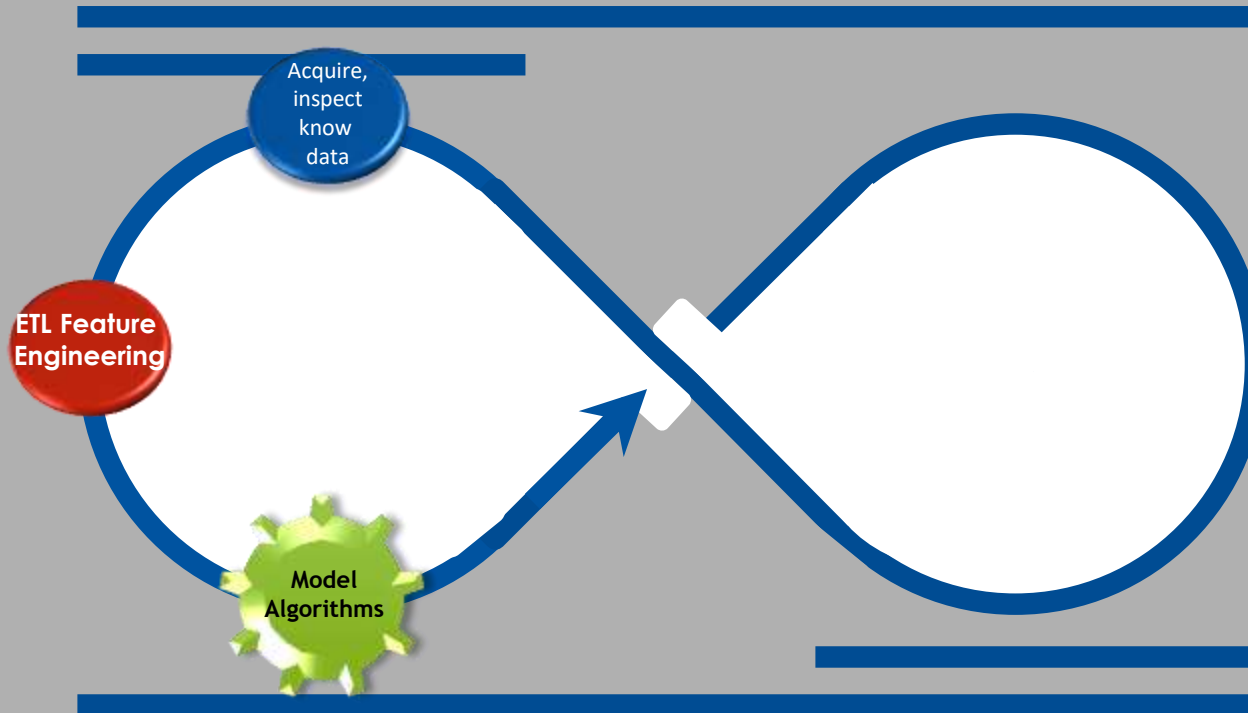
ML & AI Complex Ecosystems

- Use the right algorithms



ML & AI Complex Ecosystems

- Use the right algorithms for the business problem



ML & AI Complex Ecosystems

- Which algorithms is best for the business issue?

NEURAL NETWORKS

$$f(x) = o = w_0 + \sum_{i=1}^n w_i x_i$$

SIMILARITY

$$w_{ij} = \frac{\sum_k (R_{ix} - \bar{R}_i) \cdot (R_{jk} - \bar{R}_j)}{\sqrt{\sum_k (R_{ix} - \bar{R}_i)^2 \cdot (R_{jk} - \bar{R}_j)^2}}$$

DECISION TREE

$$Entropy = \sum_{v=0}^1 -P \cdot \log$$

$$InfoGain = P_{+} \cdot [-P_{+t} \cdot \log(P_{+t}) - P_{+(t-1)} \cdot \log(P_{+(t-1)})]$$

Which is the best algorithm for the business situation?

- Linear Regression
- Logistic Regression
- Decision Tree
- SVM
- Naive Bayes
- kNN
- K-Means
- Random Forest
- Dimensionality Reduction
- Gradient Boosting algorithms
-and hundreds more

MAXIMUM LIKELIHOOD

$$L = \arg \max P(c|a)$$

TOTAL PROBABILITY

$$TotalP(B) = P(B|A) \cdot P(A)$$

VARIANCE

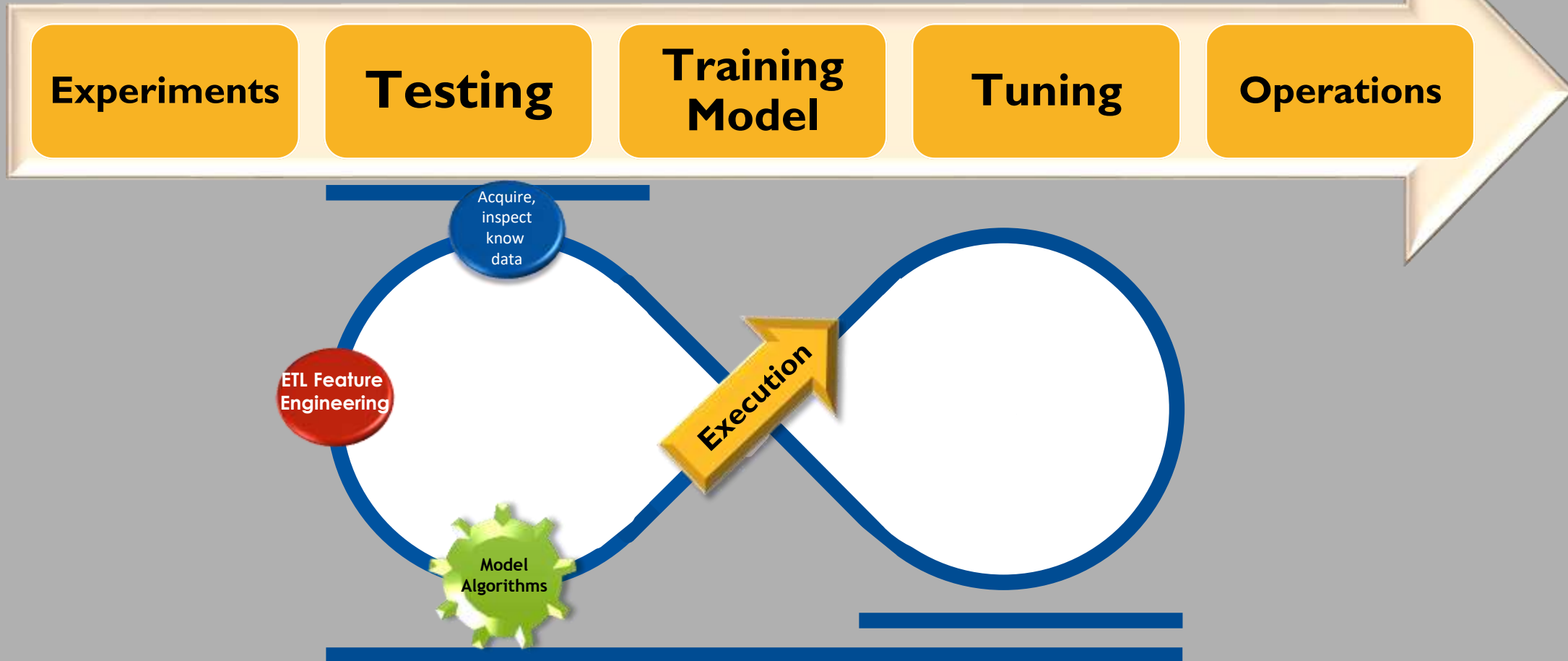
$$= \frac{\sum (x - \bar{x})^2}{n - 1}$$

REGRESSION

$$m_1 = \frac{\sum x_2^2 \sum x_1 y - \sum x_1 x_2 \sum x_2 y}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

ML & AI Complex Ecosystems

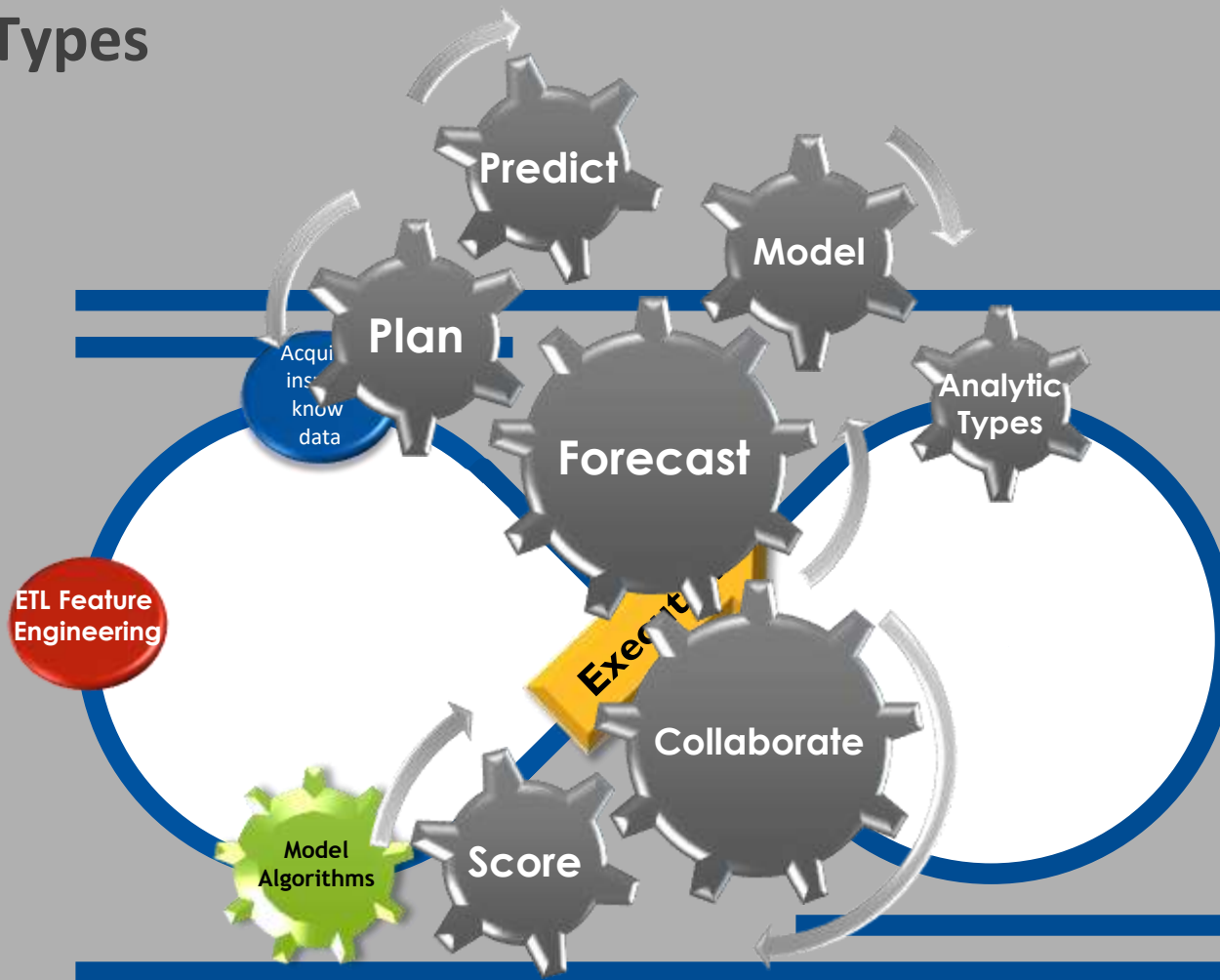
- Execution & Operations



ML & AI Complex Ecosystems

think 2019

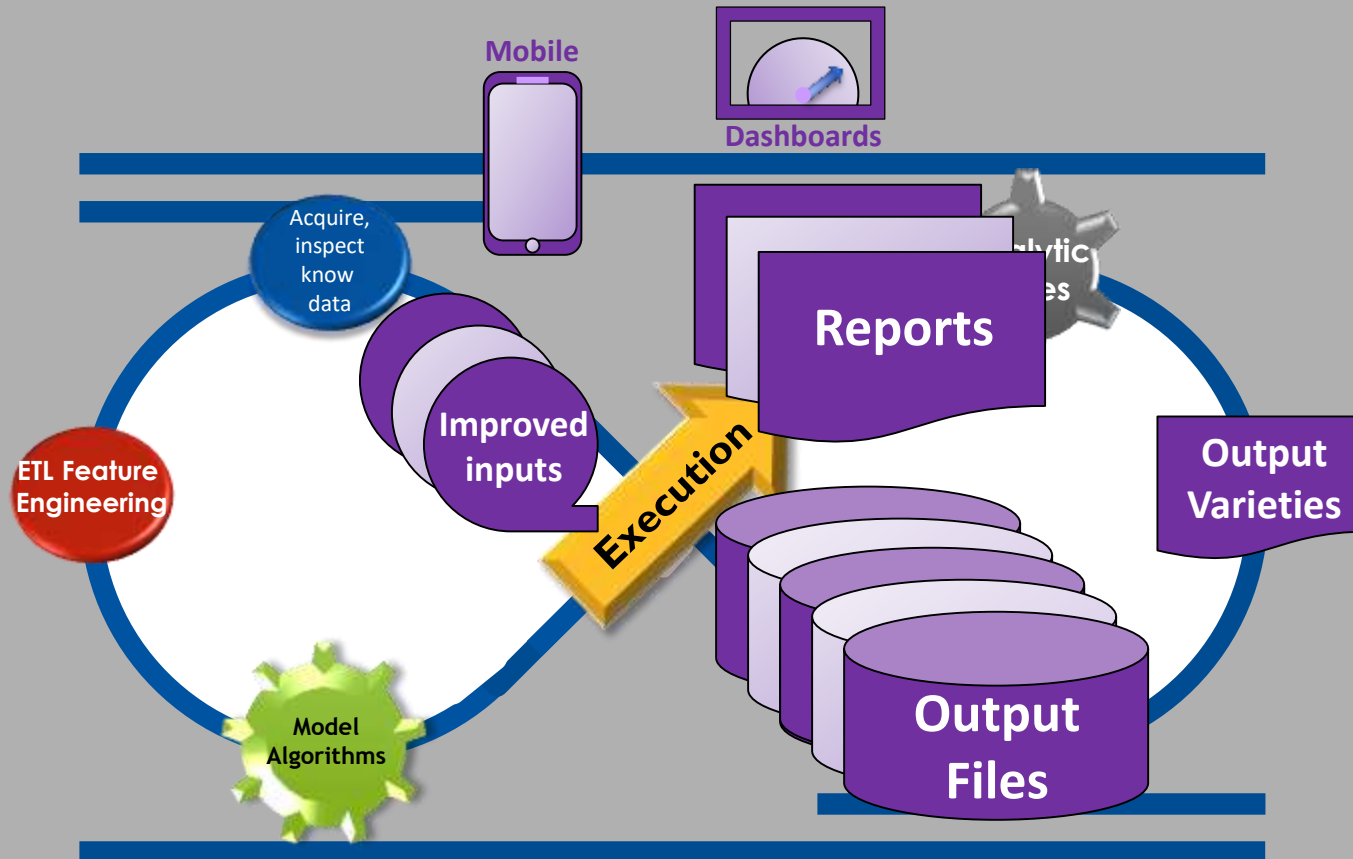
- Analytic Types



ML & AI Complex Ecosystems

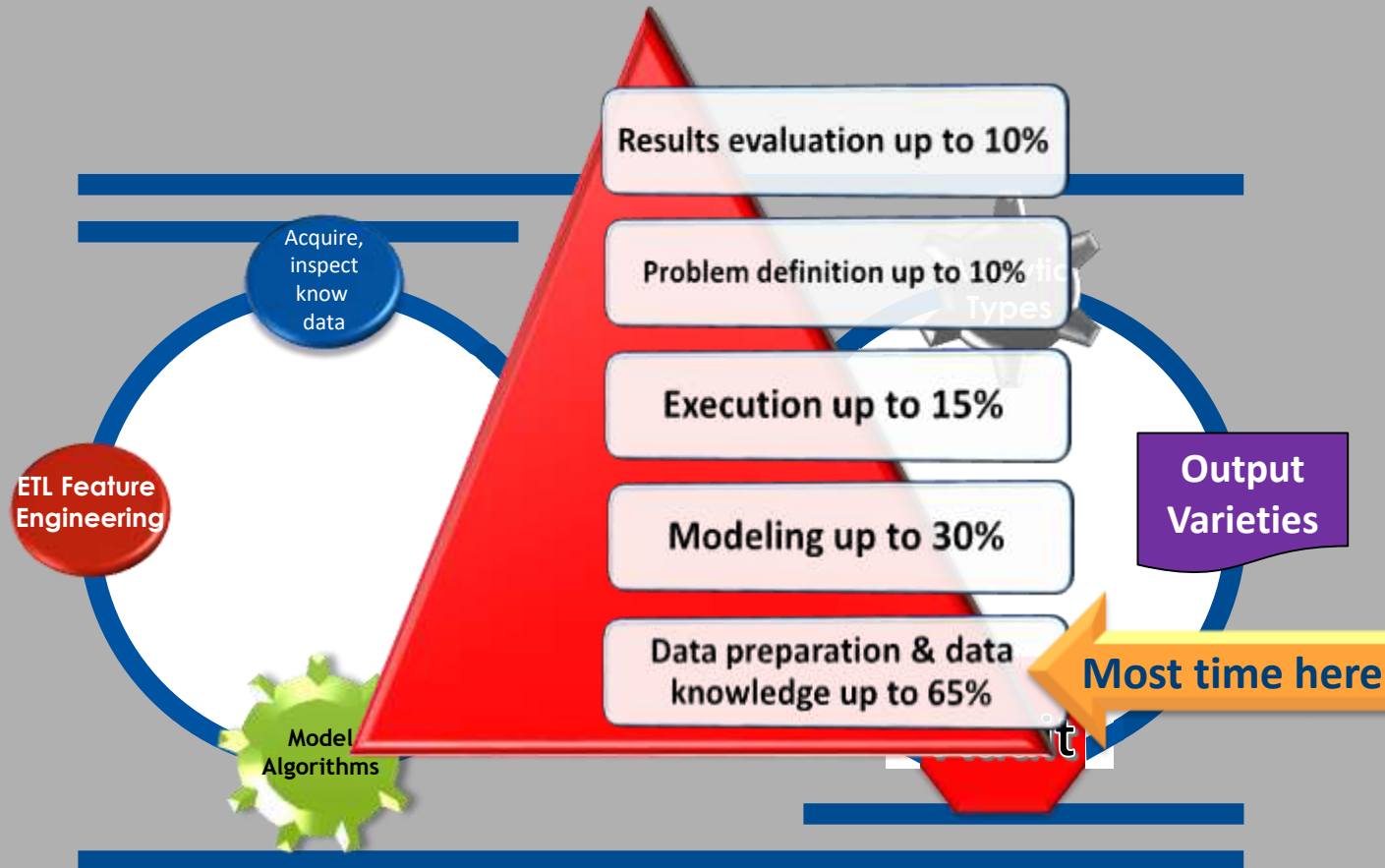
think 2019

- Output Varieties



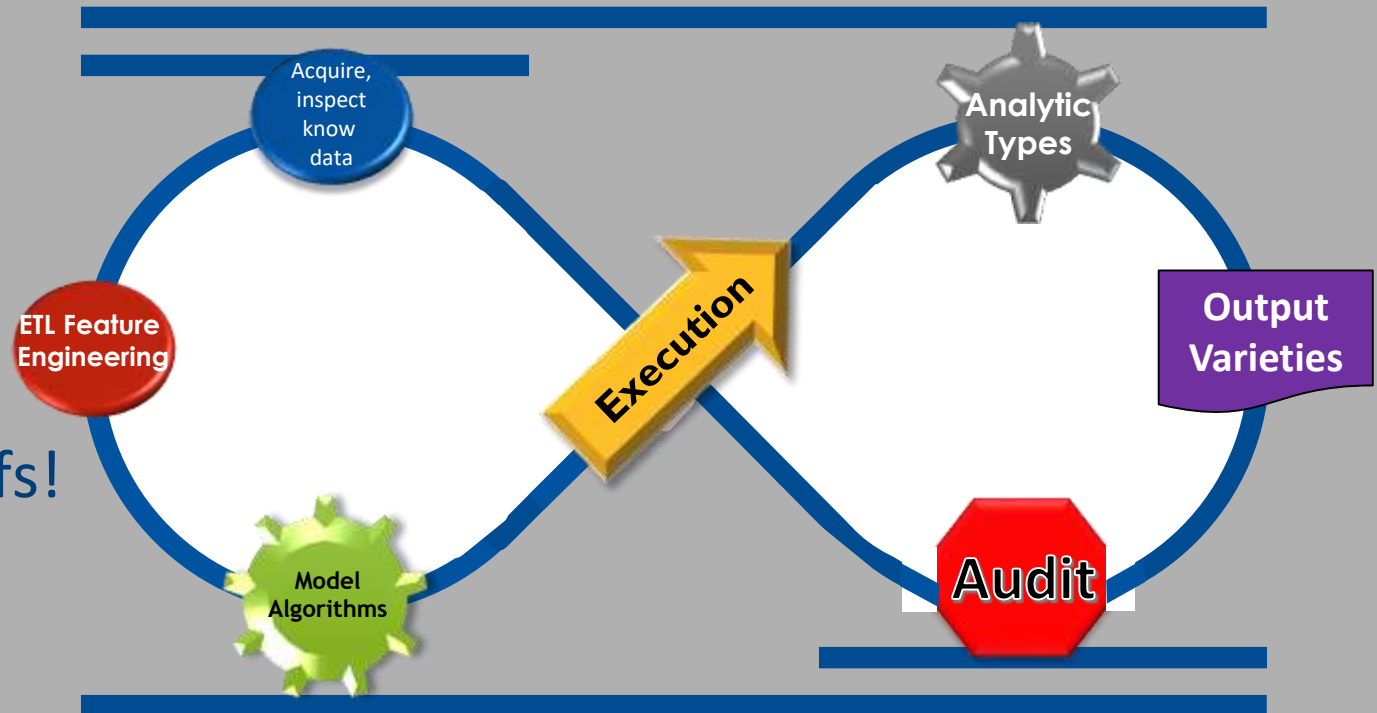
ML & AI Complex Ecosystems

- Audit ML AI Processing for improvements



ML & AI Complex Ecosystems

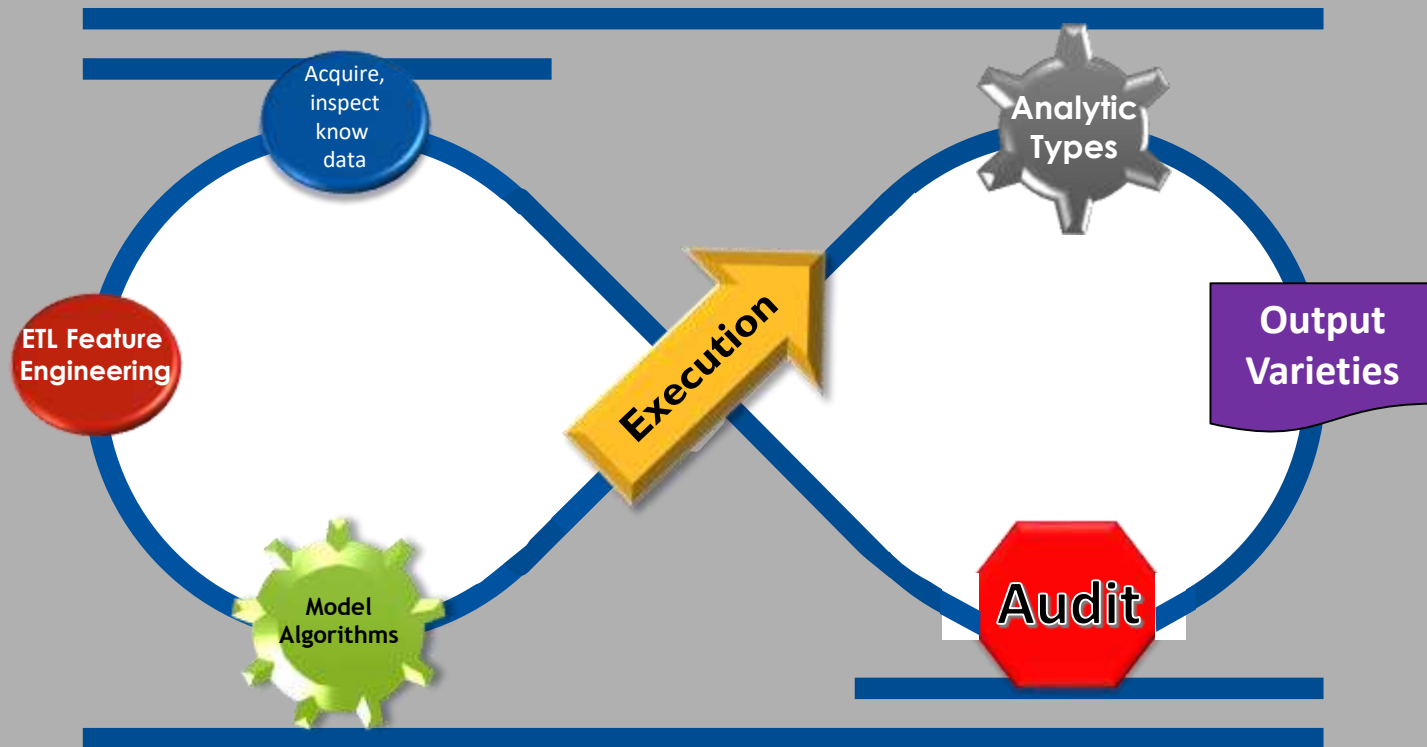
- Research, run and repeat
 - *Add ML or AI to any project proposal and it will be approved*
- Inflection point
 - Market share competition
 - Operational optimization
 - Grow market differentiators
 - Eliminate fraud
- Long term, big payoffs!



Best practice simplify complexities

- **More computing power needed for every aspect**

- Any single phase can impact all the other components within your ML or AI project
- Each phase should have its SME to optimize its planning, testing and performance
- Understand self improving formulas in Java, C#, Python or R
- Develop prototype and test your **formulas** that demonstrate **ROI**



$$Q(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} - \underbrace{Q(s_t, a_t)}_{\text{old value}} \right)$$

Continuous experiments leveraging DM best practices

Classify the
issue and
business case
ROI

Acquire, inspect
and know your
data

Model learning
algorithms path
and their outputs

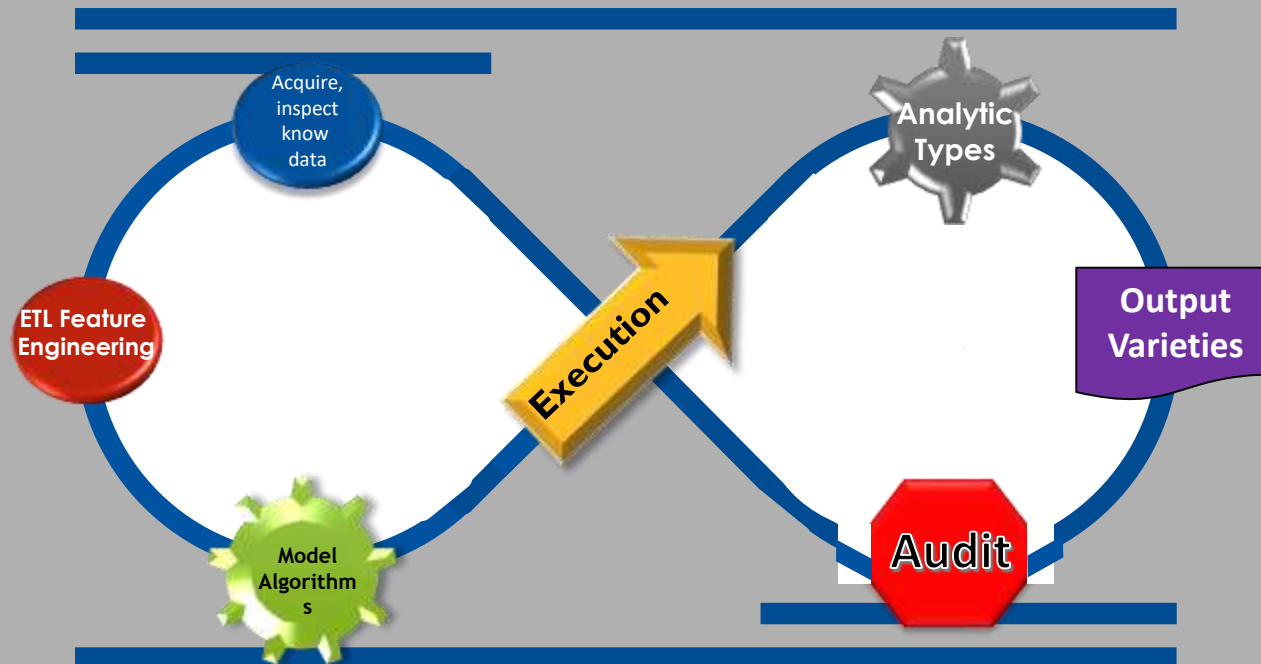
Evaluate score
your algorithm
effectiveness

Optimize
workflow,
enhance
algorithm and
refine data

**Reiterate
improving
algorithm and
data inputs**

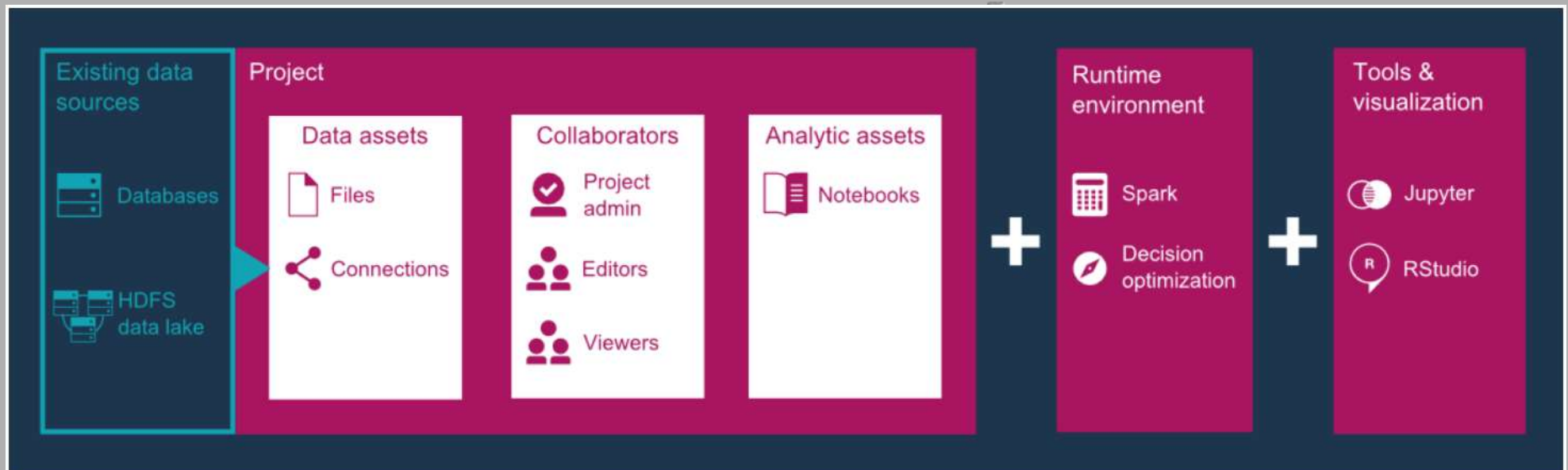
IBM Cloud opportunities

- Three types of computing – Local, Cloud or Hybrid
- IBM Watson® Studio
 - Platform agnostic product interface for implementing your DS model
 - Runs on any platform with any type of server, data or languages



IBM Cloud opportunities

- Three types of computing – Local, Cloud or Hybrid
- IBM Watson® Studio
 - Platform agnostic product interface for implementing your DS model
 - Runs on any platform with any type of server, data or languages



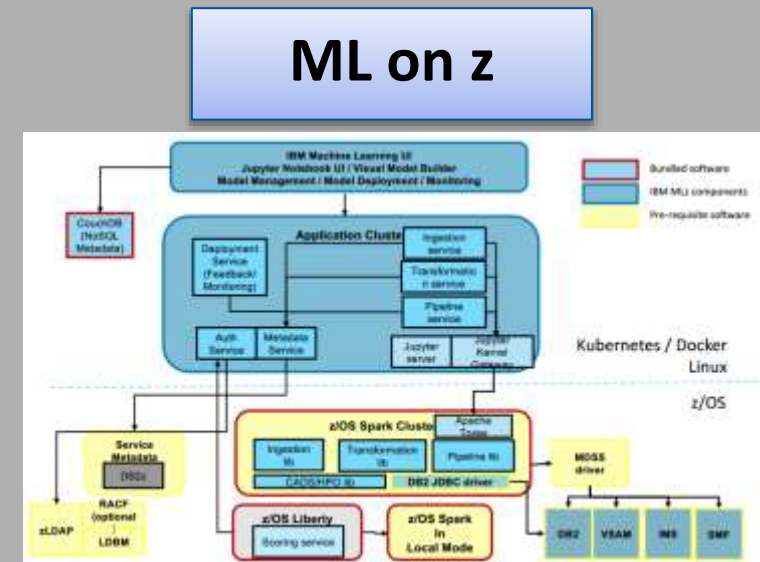
IBM Watson® Studio Family

think 2019

 Compute	 Network	 Storage	 Management
 Security	 Databases	 Analytics	 AI
 IoT	 Mobile	 Developer Tools	 Blockchain
 Integration	 Migration	 Private Cloud	 VMware

- Each category has wide variety of tools, products and options
- Immediate infrastructure for collaborating
- All types of data science developers
 - Leverage the latest open source tools such as Jupyter Notebooks, R Studio, Python and other
 - Developers can share open source tools
 - Provides ability to use the new TensorFlow, Keras, Pytorch, Caffe opens source tools
 - More tools and products are added all the time to this IBM !
- Watson Studio helps you deploy your data science solution to any cloud configuration

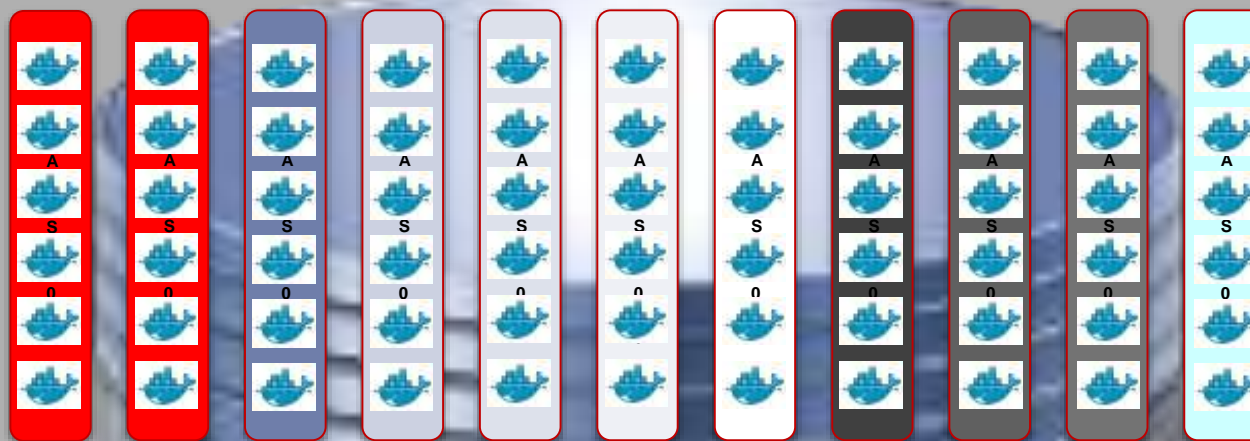
- # Cloud
- <https://www.ibm.com/cloud/get-started>
- 
- The diagram illustrates the IBM DSX Data Science Platform architecture. It is divided into four main functional areas, each represented by a red box with a white header and a list of components:
- Project**: This area is further divided into three sub-sections:
 - Data assets**: Includes 'File' and 'Externalities'.
 - Collaboration**: Includes 'Project center', 'Collaborate', and 'Share'.
 - Analytics assets**: Includes 'Modeling'.
 - Runtime environment**: Includes 'Start' and 'Deploy applications'.
 - Tools & visualization**: Includes 'Applet' and 'Attitude'.
- Arrows indicate the flow of data and processes between these components. A large plus sign (+) is placed between the 'Project' and 'Runtime environment' boxes, and another plus sign is between the 'Runtime environment' and 'Tools & visualization' boxes, suggesting a sequential or integrated workflow.
- Data Science made simple with IBM DSX |**
by Krishna Chaitanya



IBM Watson® Studio

think 2019

- Watson Studio Cloud, Watson Studio Local or Watson Studio Desktop
- Platform agnostic product interface for implementing your DS model
 - Runs on any platform with almost any type of server, un/structured data or programming languages



Many factors for Analytic Performance

- Storage efficiency/data amount vs. cost

- Private to Public info

-

- S

- Pro ... ing CPU

- Reco ... y

All the architectures work!
IBM Watson, Db2 Family and IDAA have
the reliability, availability and performance
scalability for any analytics situation!

By Dave Beulke
Dave Beulke and Associates

Dave @ davebeulke.com

think 2019

 @DBeulke

 davebeulke

Best Performance and Design Practices for Analytic, ML and AI Applications

E01

Proven Performance Tips:
www.DaveBeulke.com

Thank you!

More ML and AI information

- **The best Machine Learning Resources**
<https://medium.com/machine-learning-for-humans/how-to-learn-machine-learning-24d53bb64aa1>
- **Preparing and Architecting for Machine Learning - Gartner Inc.**
https://www.gartner.com/binaries/content/assets/events/keywords/catalyst/catus8/preparing_and_architecting_for_machine_learning.pdf
- **Three Real Use-Cases of Machine Learning in Business Applications**
https://www.huffingtonpost.com/entry/three-real-use-cases-of-machine-learning-in-business_us_593a0e91e4b014ae8c69df37
- **Smart Implementation of Machine Learning and AI in Data Analysis**
<https://callminer.com/blog/smart-implementation-machine-learning-ai-data-analysis-50-examples-use-cases-insights-leveraging-ai-ml-data-analytics/>
- **140 Machine Learning Formulas**
<https://www.datasciencecentral.com/profiles/blogs/140-machine-learning-formulas>
- **10 Algorithms Machine Learning Engineers Need to Know**
<https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article>
- **Hybrid Cloud with IBM Cloud Manager with OpenStack on z Systems**
[https://www-01.ibm.com/events/wwc/grp/grp019.nsf/vLookupPDFs/2_2_2_Heimes/\\$file/2_2_2_Heimes.pdf](https://www-01.ibm.com/events/wwc/grp/grp019.nsf/vLookupPDFs/2_2_2_Heimes/$file/2_2_2_Heimes.pdf)
- **NIST Definition of Cloud Computing**
<https://csrc.nist.gov/publications/detail/sp/800-145/final>
- **IBM Integration Bus – MQ Version 9.04 download**
http://ibm.biz/MQ_V9_FAQ & <https://www-01.ibm.com/support/docview.wss?uid=swg24043348>
- **Machine learning algorithm cheat sheet**
<https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-cheat-sheet>

More ML and AI information

think 2019

- **Azure Machine Learning first impressions**
https://medium.com/@markryan_69718/azure-machine-learning-first-impressions-f7c8366b4971
- **Machine Learning for Humans**
<https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12>
- **Three Real Use-Cases of Machine Learning in Business Applications**
https://www.huffingtonpost.com/entry/three-real-use-cases-of-machine-learning-in-business_us_593a0e91e4b014ae8c69df37