Data Science on TAP

Kyle H. Ambert, PhD

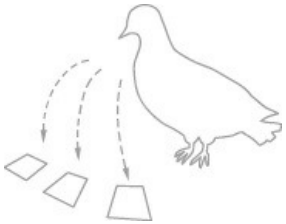
Intel Big Data Solutions, Datacenter Group

About me.

About me.

About me.

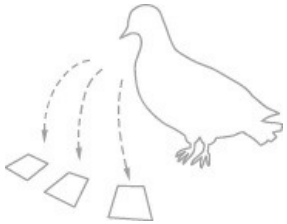
About me.



About me.


WHEATON COLLEGE
CHRISTO
ET REGNO
EJUS
ILLINOIS

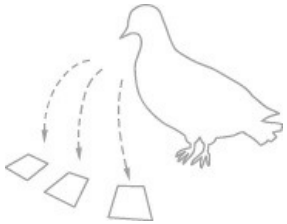

10001001010
0111000100111001
0101001011010010 1001
1011010011101001001010
0110100100101010011010 1
1001001010010111 1000100
1011011001010010 00101
0110100100100100 10
10010010010
011010101
11 0010
0



About me.


WHEATON COLLEGE
CHRISTO
ET REGNO
EJUS
ILLINOIS


10001001010
0111000100111001
0101001011010010 1001
1011010011101001001010
0110100100101010011010 1
1001001010010111 1000100
1011011001010010 00101
0110100100100100 10010
10010010010
011010101
11 0010
0

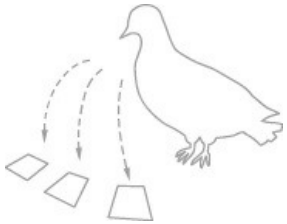


About me.


WHEATON COLLEGE
CHRISTO
ET REGNO
EJUS
ILLINOIS


10001001010
0111000100111001
0101001011010010 1001
1011010011101001001010
0110100100101010011010 1
1001001010010111 1000100
1011011001010010 00101
0110100100100100 10010
10010010010
011010101
11 0010
0


OREGON
HEALTH
&SCIENCE
UNIVERSITY
OHSU

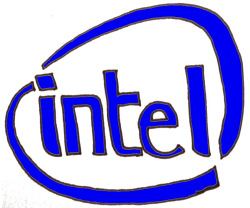


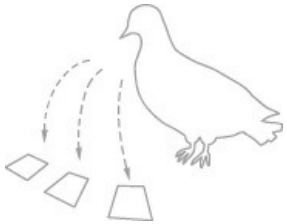
About me.

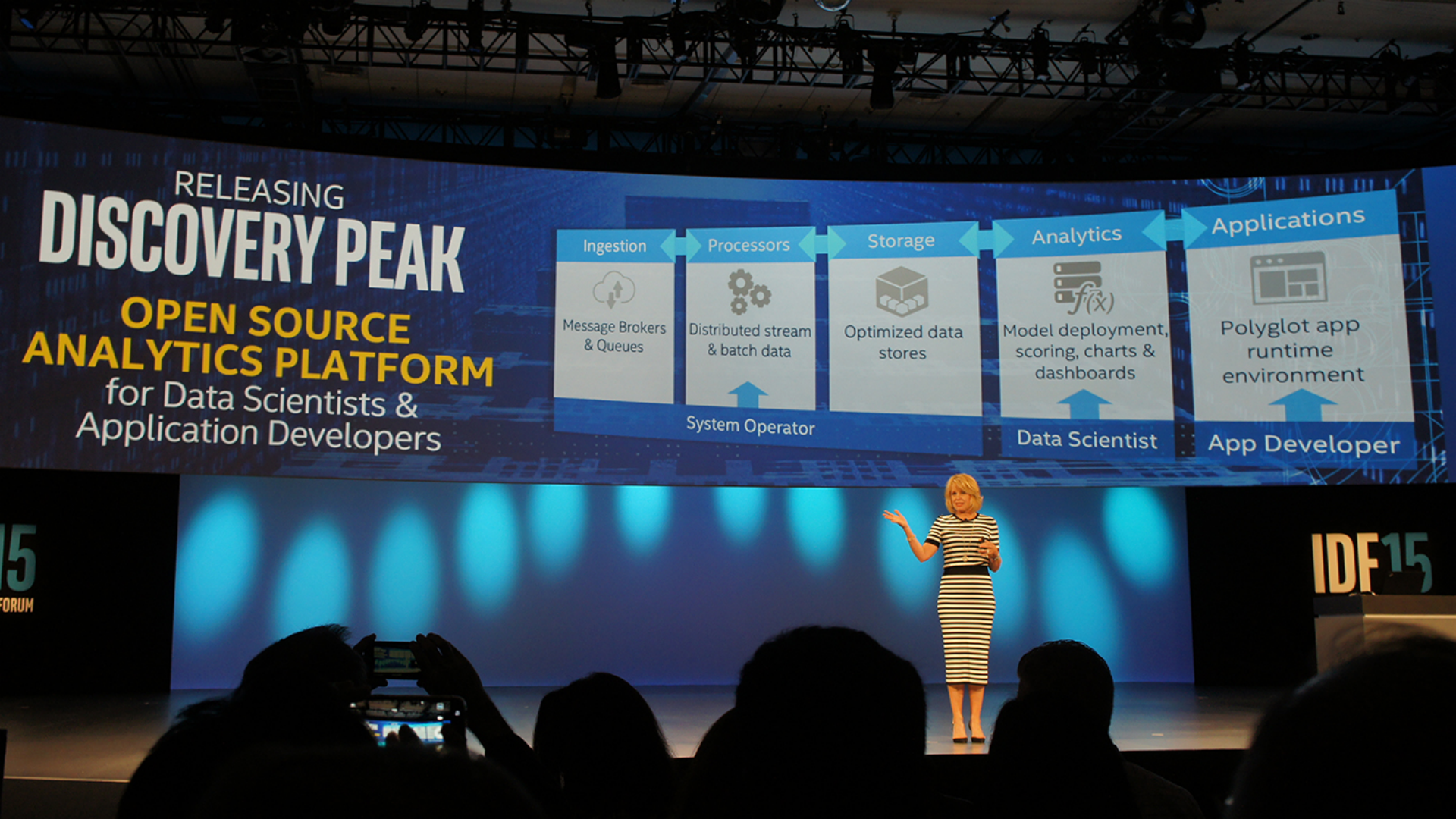

WHEATON COLLEGE
CHRISTO
ET REGNO
EJUS
ILLINOIS


10001001010
0111000100111001
0101001011010010 1001
1011010011101001001010
0110100100101010011010 1
1001001010010111 1000100
1011011001010010 00101
0110100100100100 10010
10010010010
011010101
11 0010
0


OREGON
HEALTH
&SCIENCE
UNIVERSITY
OHSU


intel




RELEASING
DISCOVERY PEAK
OPEN SOURCE ANALYTICS PLATFORM
for Data Scientists & Application Developers
Ingestion
Message Brokers & Queues
Processors
Distributed stream & batch data
Storage
Optimized data stores
Analytics
Model deployment, scoring, charts & dashboards
Applications
Polyglot app runtime environment
System Operator
Data Scientist
App Developer
IDF15

Problem Statement

From data science to big data analytics: Less alchemy, more chemistry

Problem Statement

Data Science:

From data science to big data analytics: Less alchemy, more chemistry

Problem Statement

Data Science:

- Iterative error-prone drudgery

From data science to big data analytics: Less alchemy, more chemistry

Problem Statement

Data Science:

- Iterative error-prone drudgery
- One-off, ad hoc models in isolation

From data science to big data analytics: Less alchemy, more chemistry

Problem Statement

Data Science:

- Iterative error-prone drudgery
- One-off, ad hoc models in isolation

Analytics Processing:

From data science to big data analytics: Less alchemy, more chemistry

Problem Statement

Data Science:

- Iterative error-prone drudgery
- One-off, ad hoc models in isolation

Analytics Processing:

- Single-threaded, single-node processing

From data science to big data analytics: Less alchemy, more chemistry

Problem Statement

Data Science:

- Iterative error-prone drudgery
- One-off, ad hoc models in isolation

Analytics Processing:

- Single-threaded, single-node processing
- Proprietary, fixed-function solutions

From data science to big data analytics: Less alchemy, more chemistry

Problem Statement

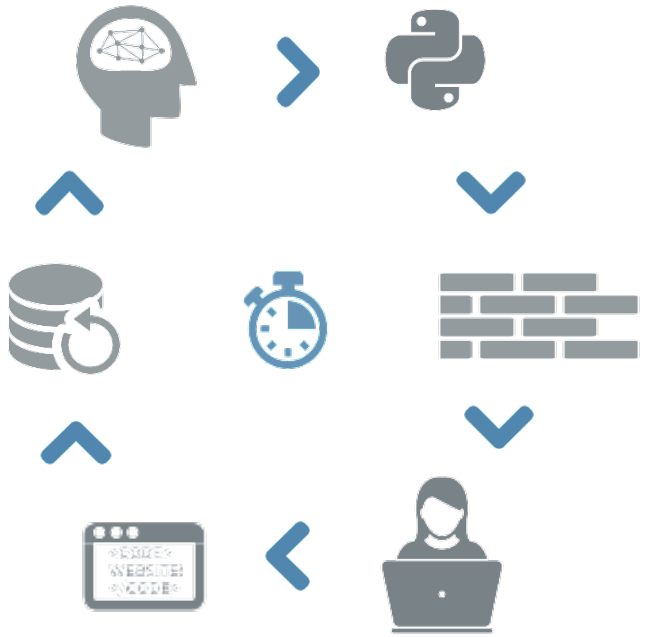
Data Science:

- Iterative error-prone drudgery
- One-off, ad hoc models in isolation

Analytics Processing:

- Single-threaded, single-node processing
- Proprietary, fixed-function solutions

Application Code:



From data science to big data analytics: Less alchemy, more chemistry

Problem Statement

Data Science:

- Iterative error-prone drudgery
- One-off, ad hoc models in isolation

Analytics Processing:

- Single-threaded, single-node processing
- Proprietary, fixed-function solutions

Application Code:

- Monolithic architecture

From data science to big data analytics: Less alchemy, more chemistry

Problem Statement

Data Science:

- Iterative error-prone drudgery
- One-off, ad hoc models in isolation

Analytics Processing:

- Single-threaded, single-node processing
- Proprietary, fixed-function solutions

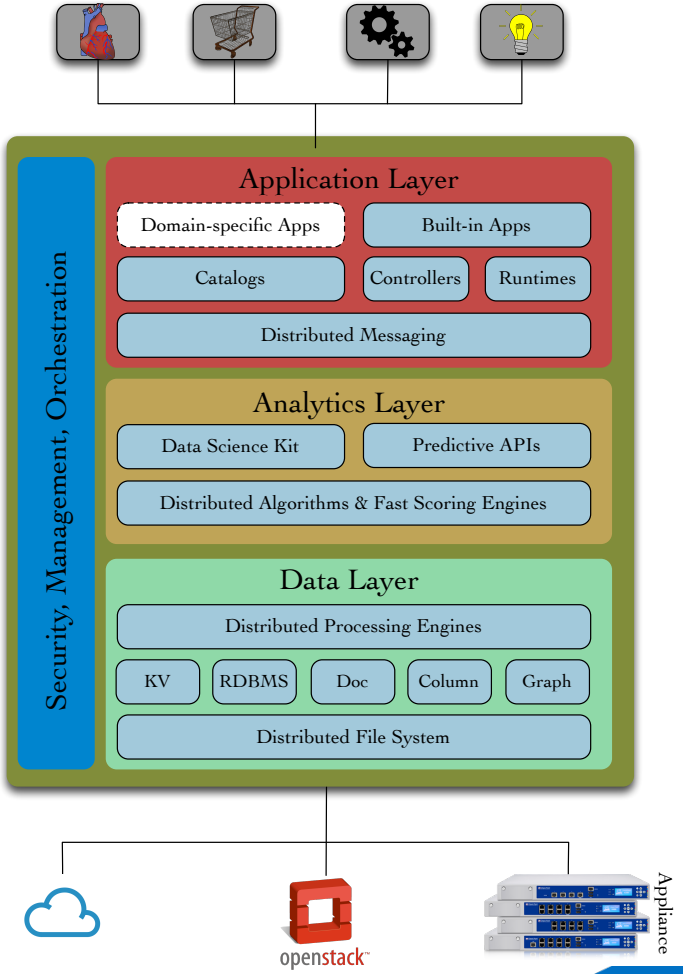
Application Code:

- Monolithic architecture
- Legacy components

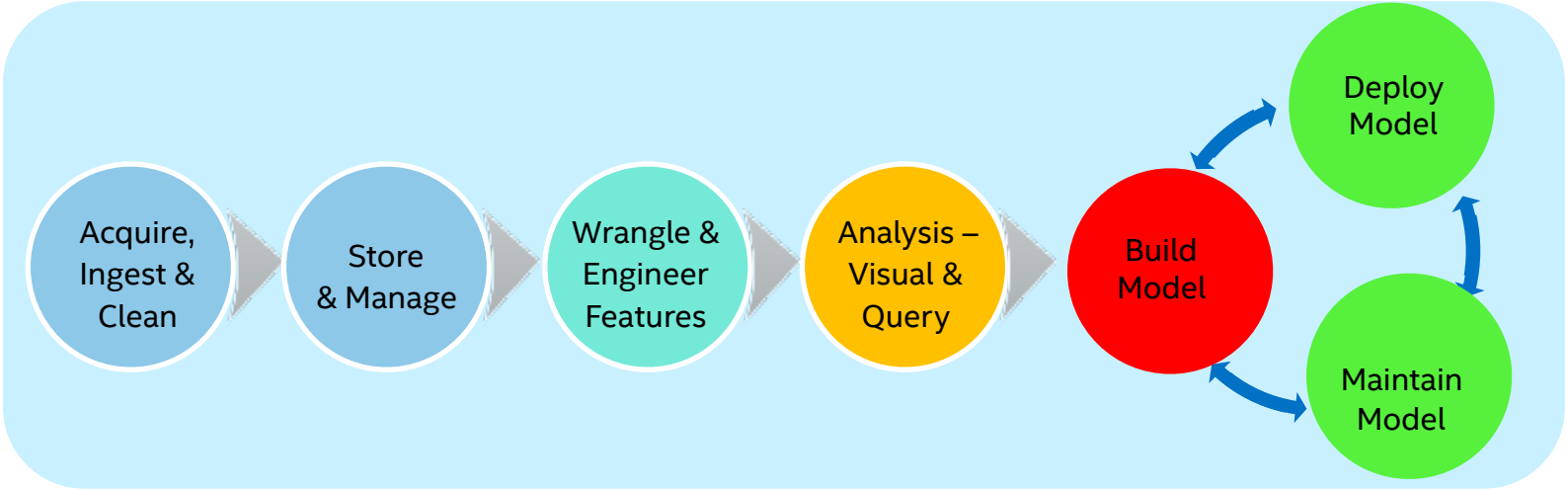
From data science to big data analytics: Less alchemy, more chemistry

Trusted Analytics Platform (TAP)

Open source software project that accelerates creation of Cloud-native apps driven by big data analytics. TAP makes it easier for developers to collaborate with data scientists by providing a shared environment for advanced analytics on big data.



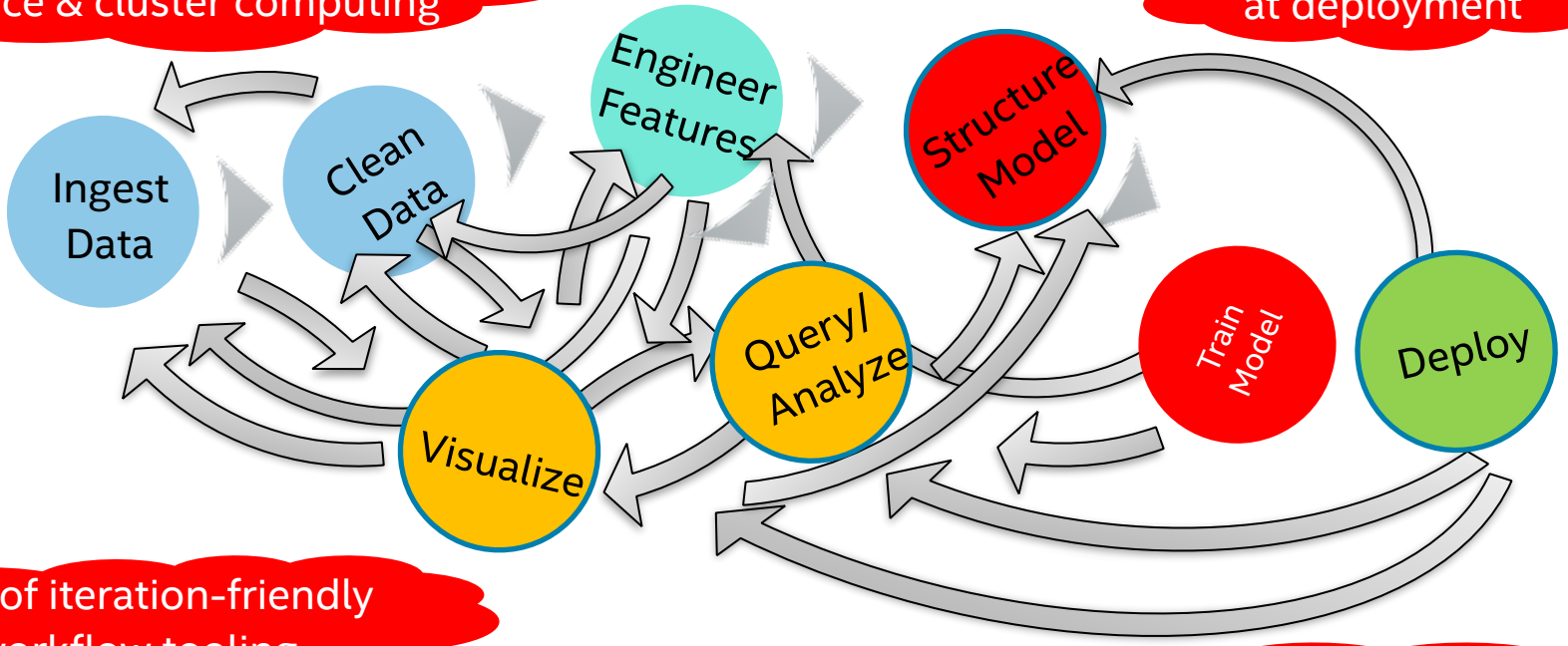
The Data Science Workflow



Challenges Across The Workflow

Skill shortages: data science & cluster computing

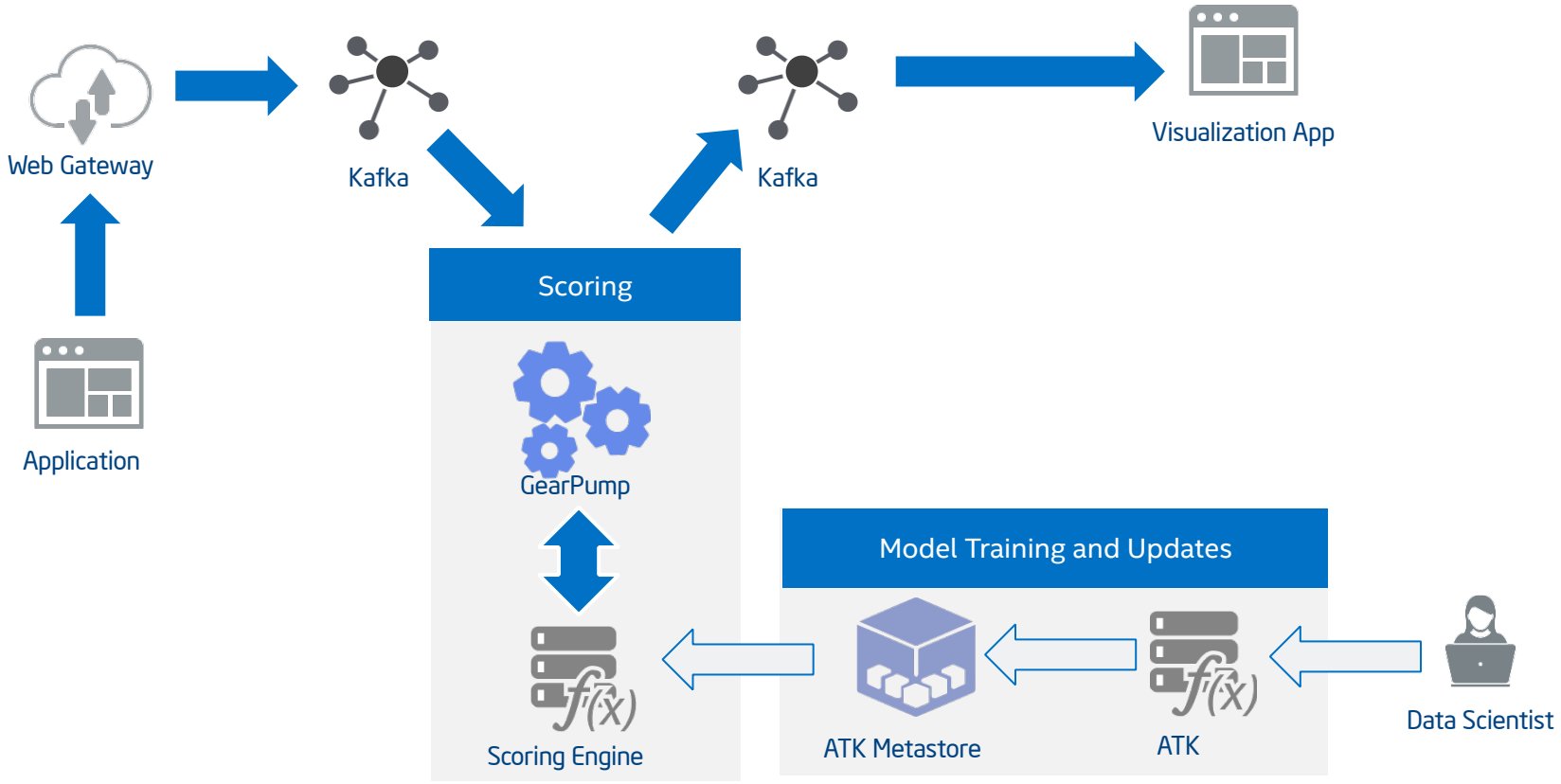
Process starts over at deployment



Lack of iteration-friendly workflow tooling

Limited methods to explore hidden patterns in big data

Example: App With End to End Scoring

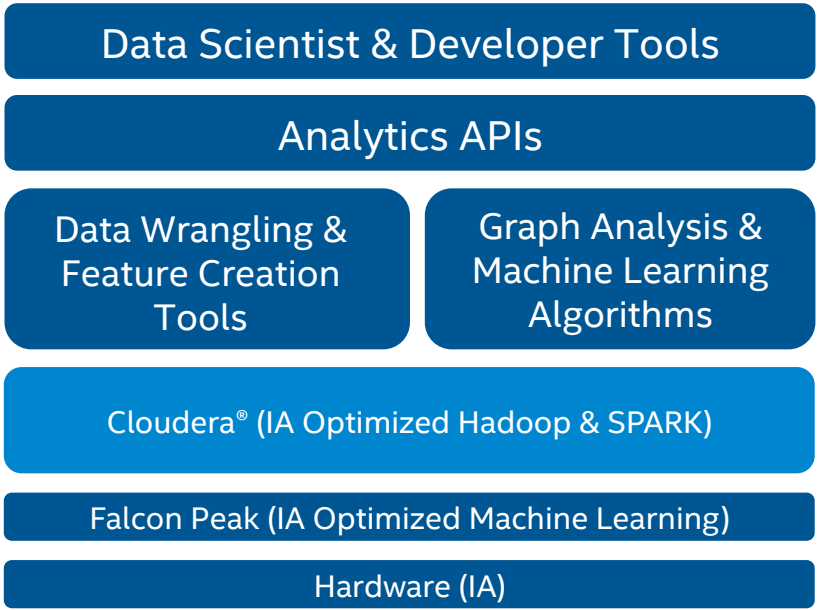


Our Analytics Toolkit

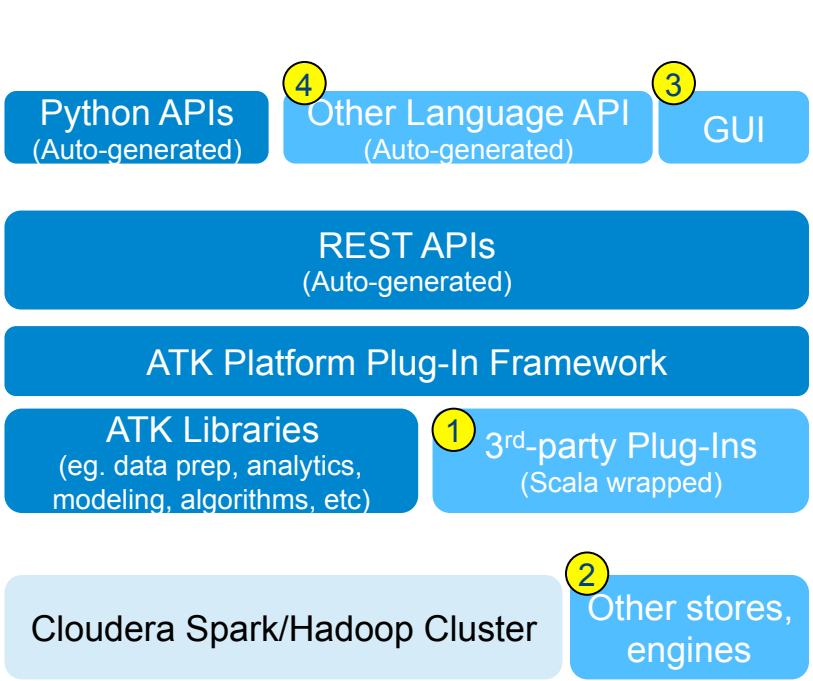
A unifying framework for analytics pipelines

An extensible platform for data science

- Development
  - Interactive
  - Language integration
  - Familiar data structures
  - Extensible data science pipeline (methods, functions, engines)
  - Separates the role of Data Scientist from Hadoop Expert
- Production
  - Interactive solution deployed to production as prediction service / scoring engine



Our Analytics Toolkit: Extensibility



1. Add algorithms & routines that can consume frame & graph objects
   - Isolates libraries from conflicts
   - API presentation auto-generated
   - ATK source not needed; unify GPL codes
2. Optionally unify workflow with other (off Hadoop) data stores & engines
3. GUI presentation easily supported, enabling UIs to expose capabilities to other data science & analyst preferences
4. API auto-generation extensible to other languages – eg “R” or Javascript

Machine Learning for Predicting Risk for Emergency Room Readmittance

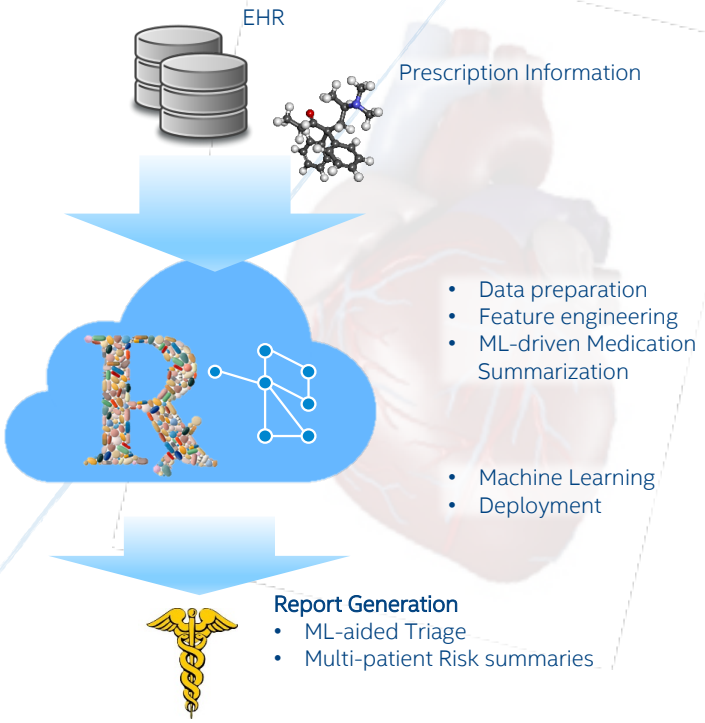
Patient Outcomes

- At-risk patients can be identified on admit
- Enables more personalized treatment

Hospital Outcomes

- Decreased costs
- More efficient use of resources

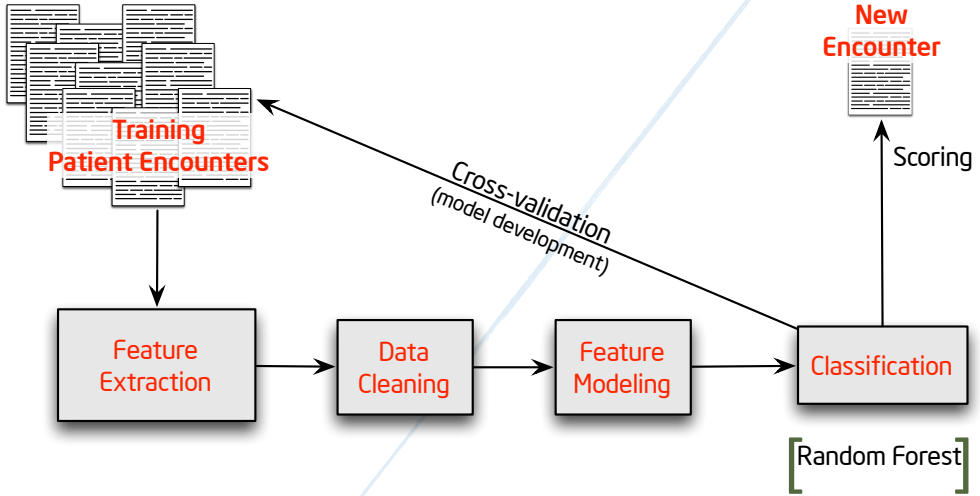
By leveraging big data in health analytics, patients at risk for negative outcomes can be identified early on, enabling better care



Iterative Model Building in the Trusted Analytics Platform (TAP)

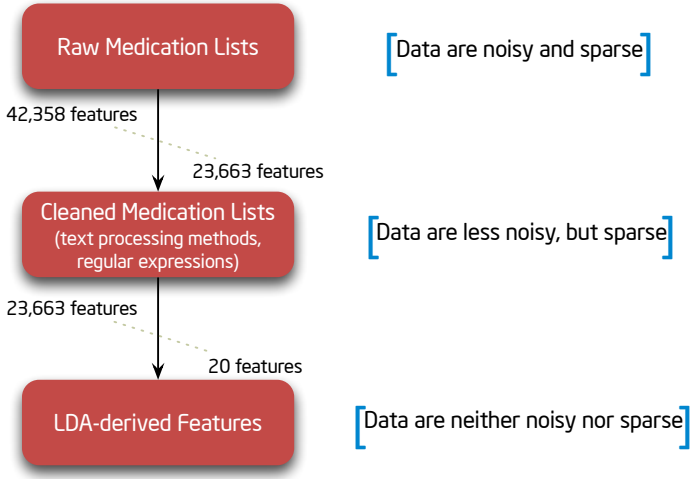
- De-identified electronic health record data was obtained for one year of patient encounters at Penn Med
- A cohort of emergency-admitted cardiac-related patients was identified (≈140K)

- We conducted feature selection and modeling experiments to iteratively construct a model useful for scoring previously-unseen encounters
- We built a classifier to identify which patients would be readmitted within 30 and 90 days of being discharged



TAP enabled us to efficiently conduct iterative experiments in a distributed environment!

Contribution of Medication Modeling with Topic Modeling



[Data are noisy and sparse]

[Data are less noisy, but sparse]

[Data are neither noisy nor sparse]

Minimal:	Age, Gender, Dead (y/n), Illness Severity
Baseline:	(*minimal model*), average length of stay, weight, height, race, marital status, mortality risk, hospice discharge, readmit, days since previous readmit,
Baseline & Discharge Medications	(*baseline model*), medications (LDA)
Baseline & Admitting Features	(*baseline model*), Hx medications (LDA), admitting reasons (LDA), Hx comborbid diagnoses (LDA), Hx order names (LDA), Hx discharge diagnoses (LDA)

Including LDA-derived medication features resulted in a 15% improvement!