SAS® GLOBAL FORUM 2018

USERS PROGRAM

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF

Data Science And SAS®™ - A Data Scientist Perspective

Richard La Valley & Jay Revere
OGSystems, Inc. SAS® Institute

Presenter

RICH LA VALLEY, OGSYSTEMS, INC, DATA SCIENTIST/STATISTICIAN

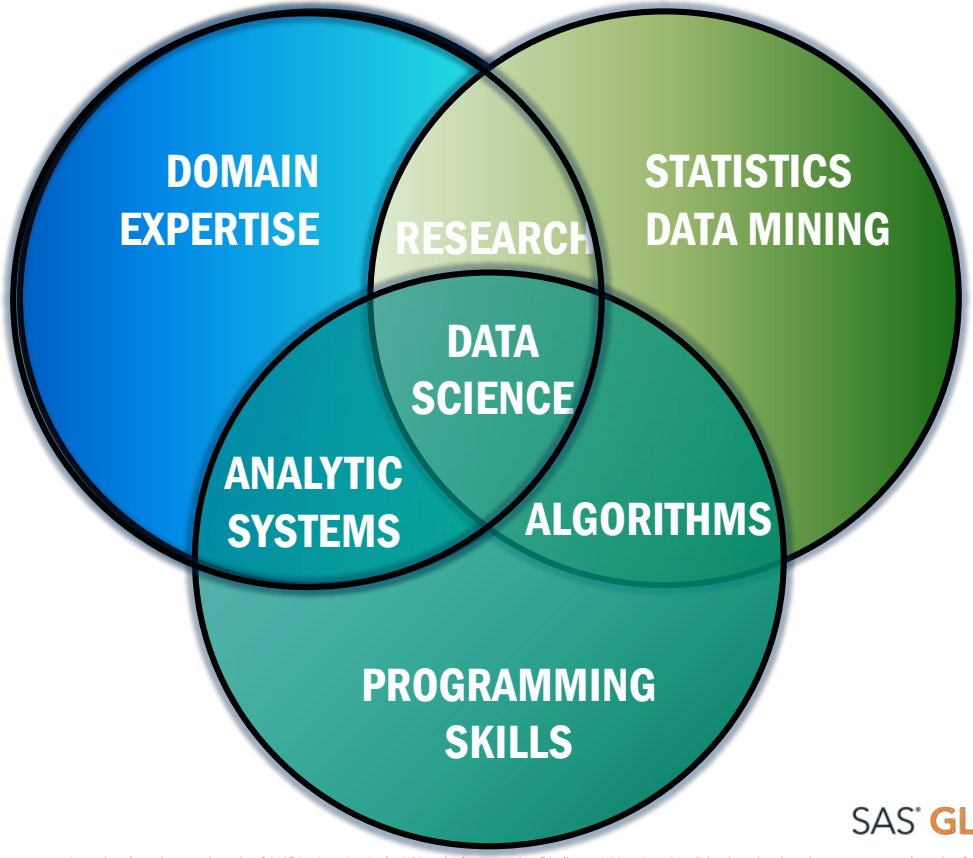
Rich La Valley is a data scientist and statistician at OGSYSTEMS, INC. where he has been since 2017. Rich uses SAS®® in his data science and data analysis responsibilities as a data scientist for his customers. He has used SAS® for more than 39 years.

Presenter

JAY REVERE, SAS® INSTITUTE

Jay Revere is a Principal Technical Consultant at the SAS® Institute where he has been since 2009. Jay uses SAS® in his daily responsibilities, helping clients leveraging SAS® and the Power To Know and derive actionable intelligence. He has used SAS® for more than 27 years.

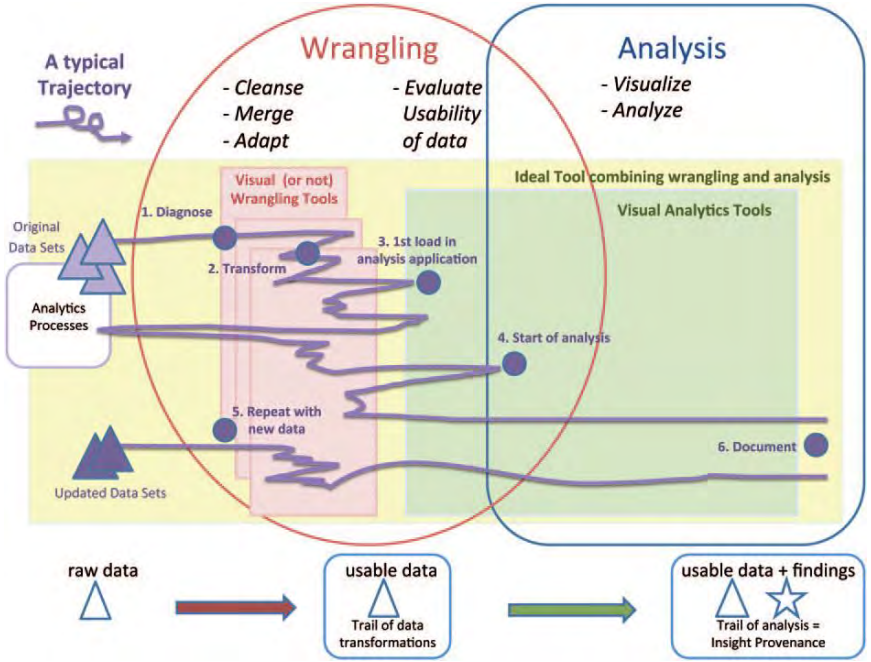
What is Data Science?



What is the biggest problems that Data Scientists face?

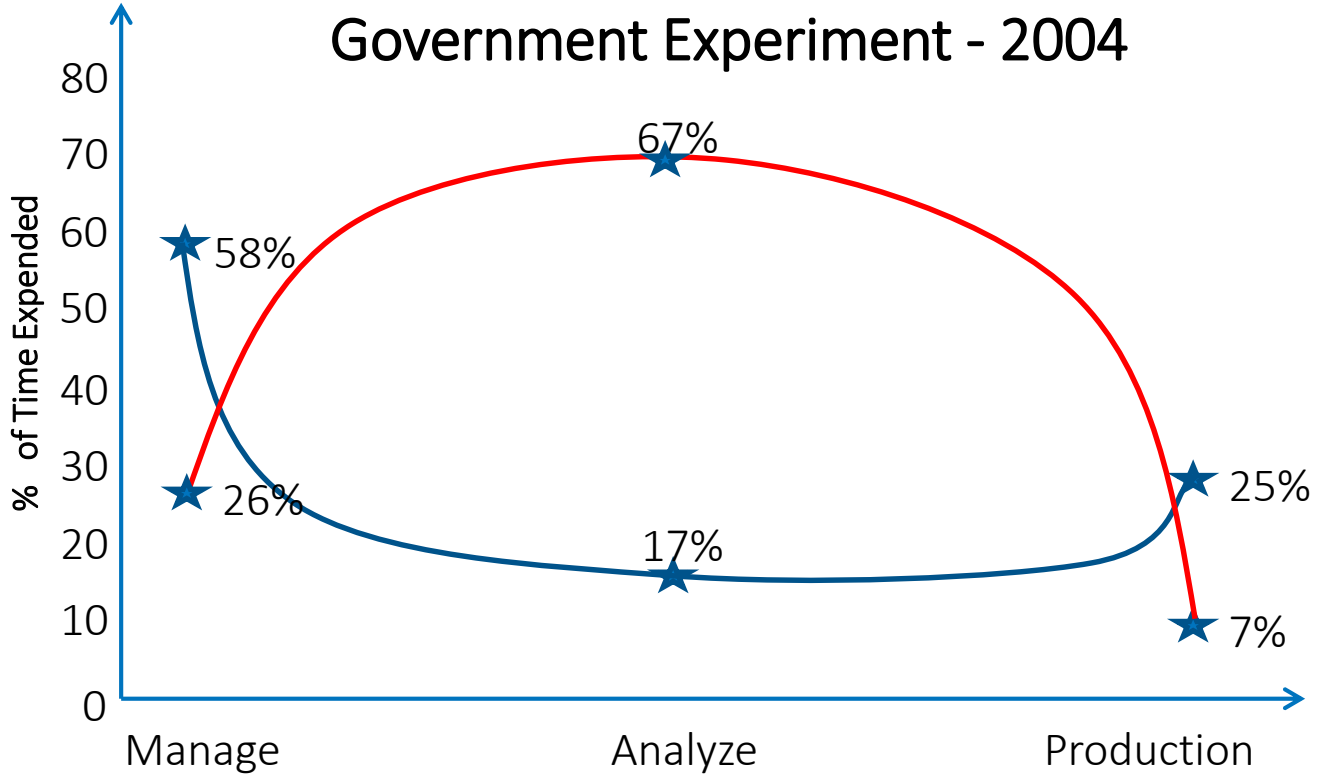
Data Wrangling or Data Munging

- Process of iterative data exploration and transformation that enables analysis
- Process of making data useful



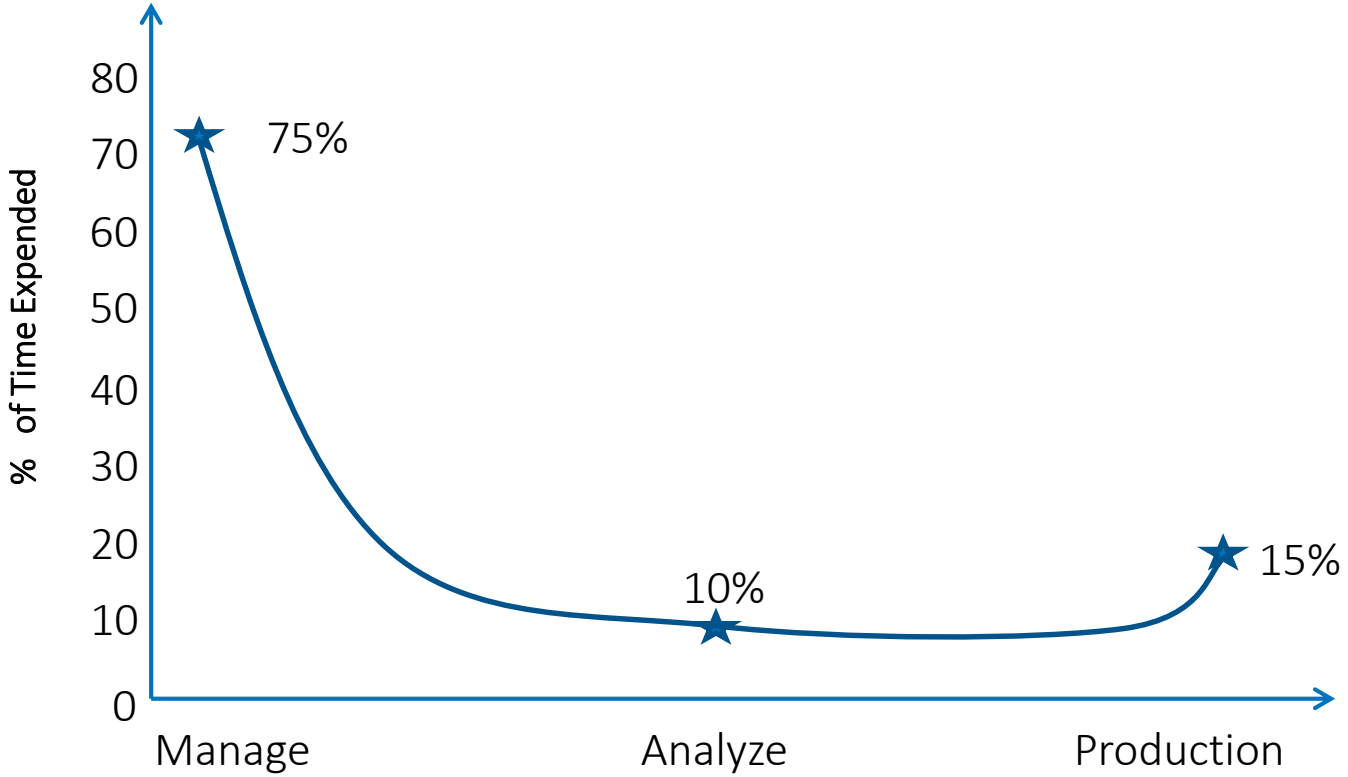
What is the Point?

Inverting the Bath Tub Curve

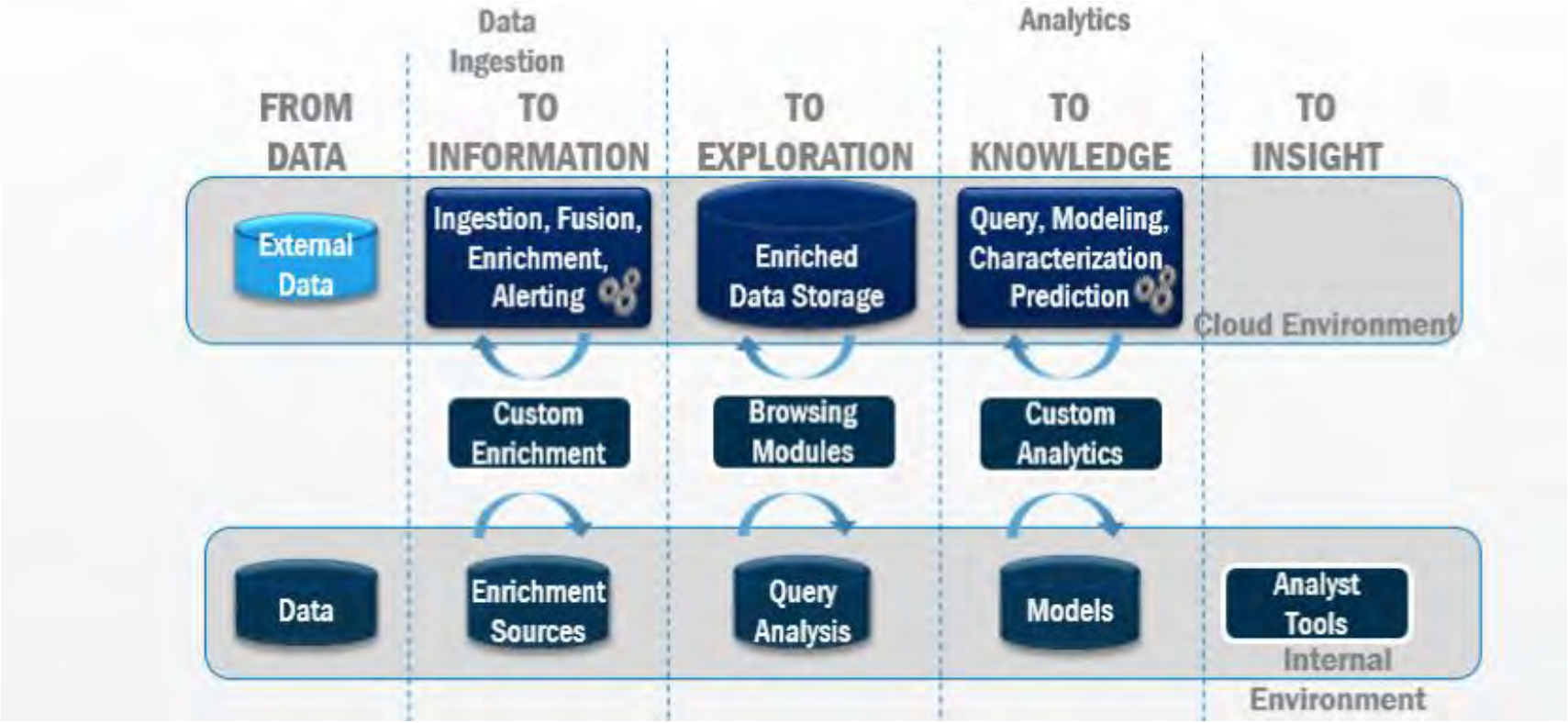


Inverting the Bath Tub Curve

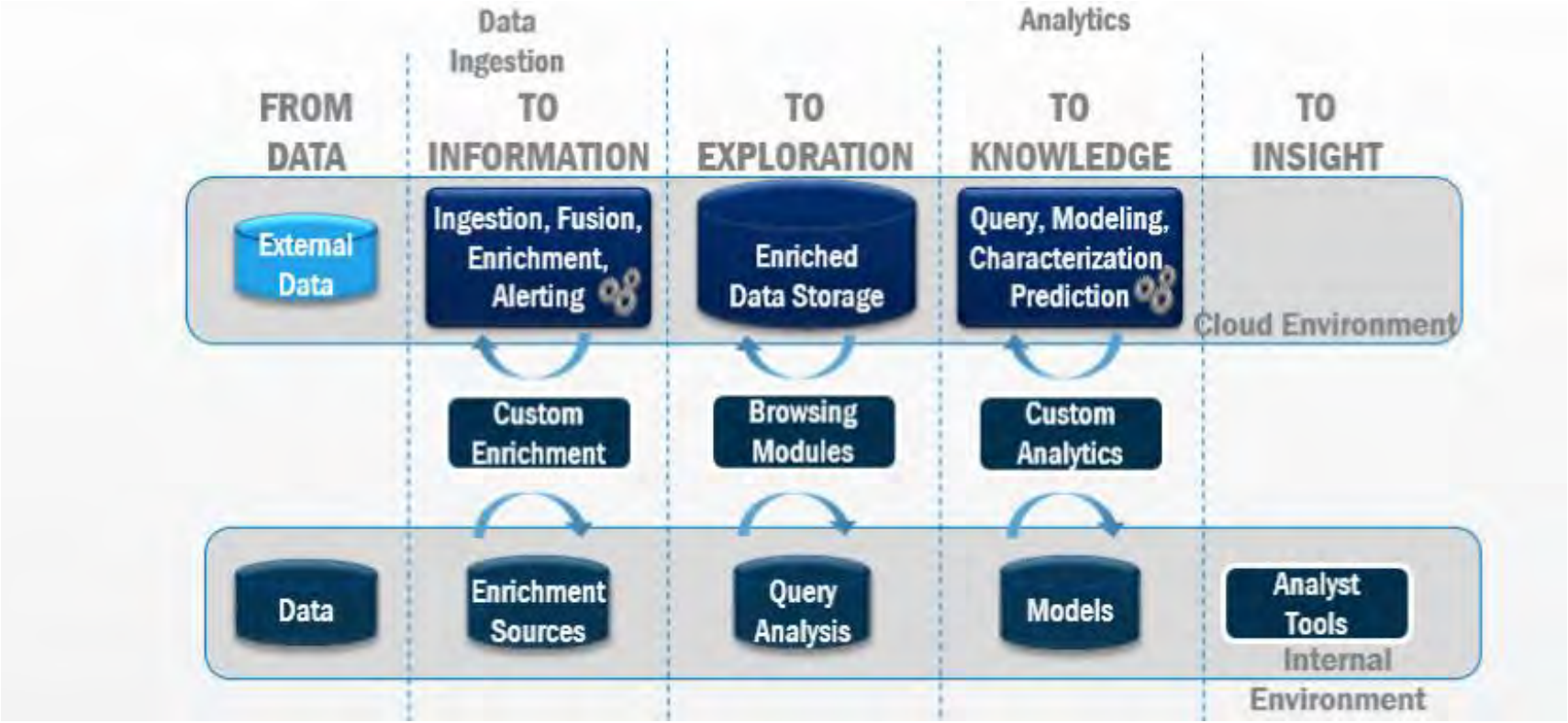
BIG DATA - 2014



Data Science Process – What Data Scientists do?

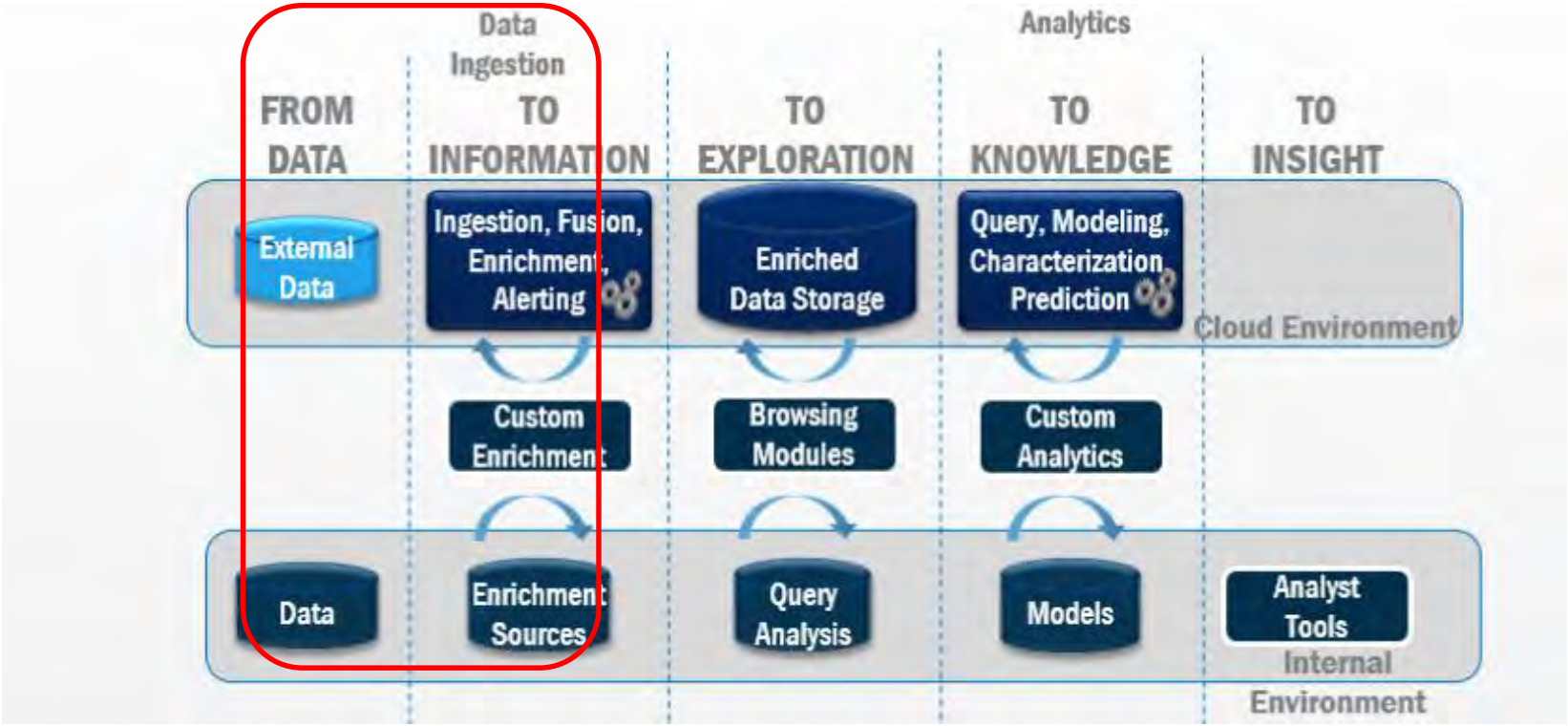


Where does SAS® fit in with this process?



SAS® & INGESTING DATA

SAS® ACCESS & API CAPABILITIES



DATA SCIENCE & SAS®® DATA INGESTION

SAS®® Data Ingestion with SAS®® \APIs and SAS®\ACCESS

Most SAS®® APIs provide data ingestion capabilities.

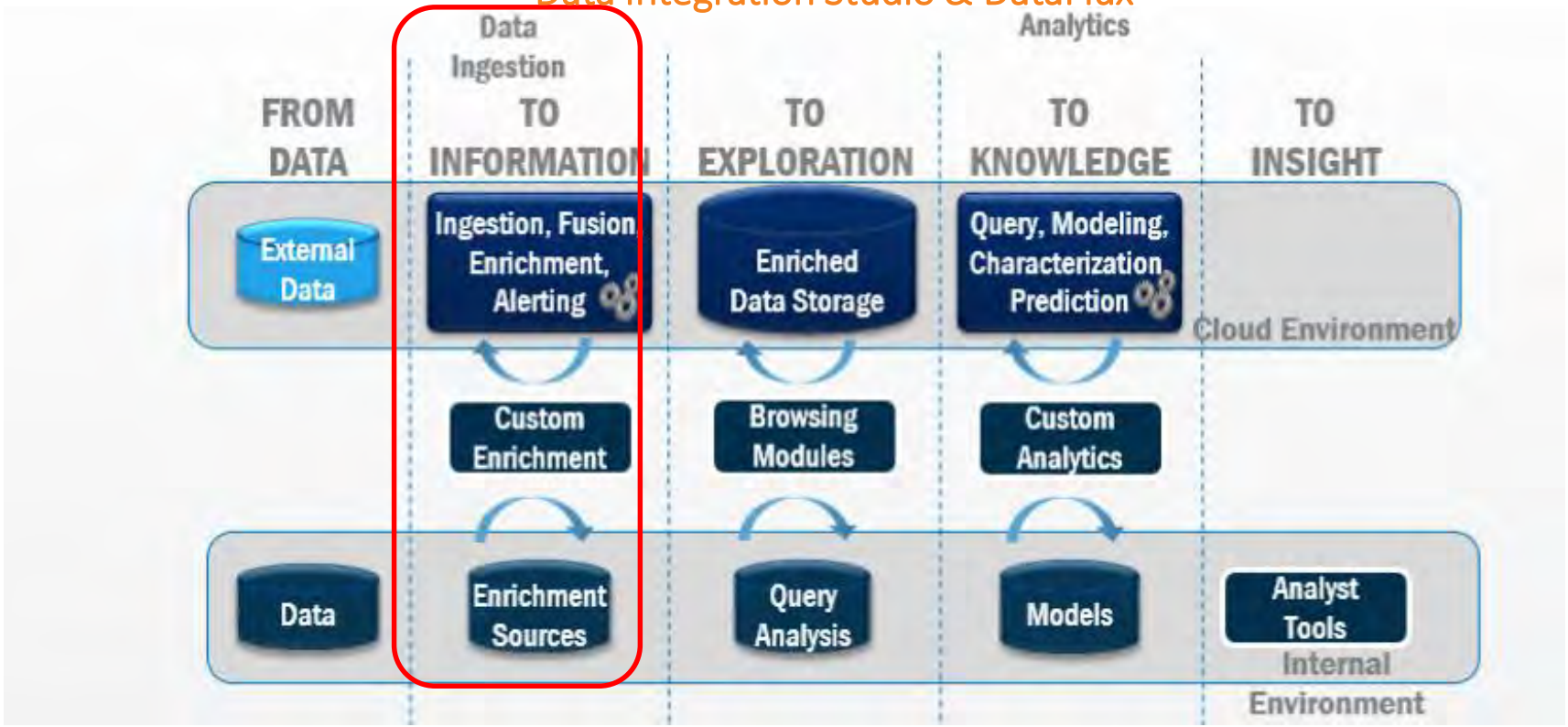
- SAS® Visual Analytics/Visual Statistics allows the user to simply browse out to their data, import it into SAS® and work with that ingested data in SAS® format for visualizations.
- SAS® Enterprise Guide and SAS® Data integration Studio also allow the user to browse to their data and import it using API controls and wizards.

SAS®\Access Engines

- Using SAS® Management Console, creating library references to DBMS allows users to click and drag their data into the applications to work with.
- SAS® programs using the SAS®\Access engines can ingest data into a SAS® data warehouse that all users can access

DATA SCIENCE & SAS® DATA MANAGEMENT

Data Integration Studio & DataFlux



DATA SCIENCE & SAS® Data Management

SAS® Data Integration Studio / Dataflux

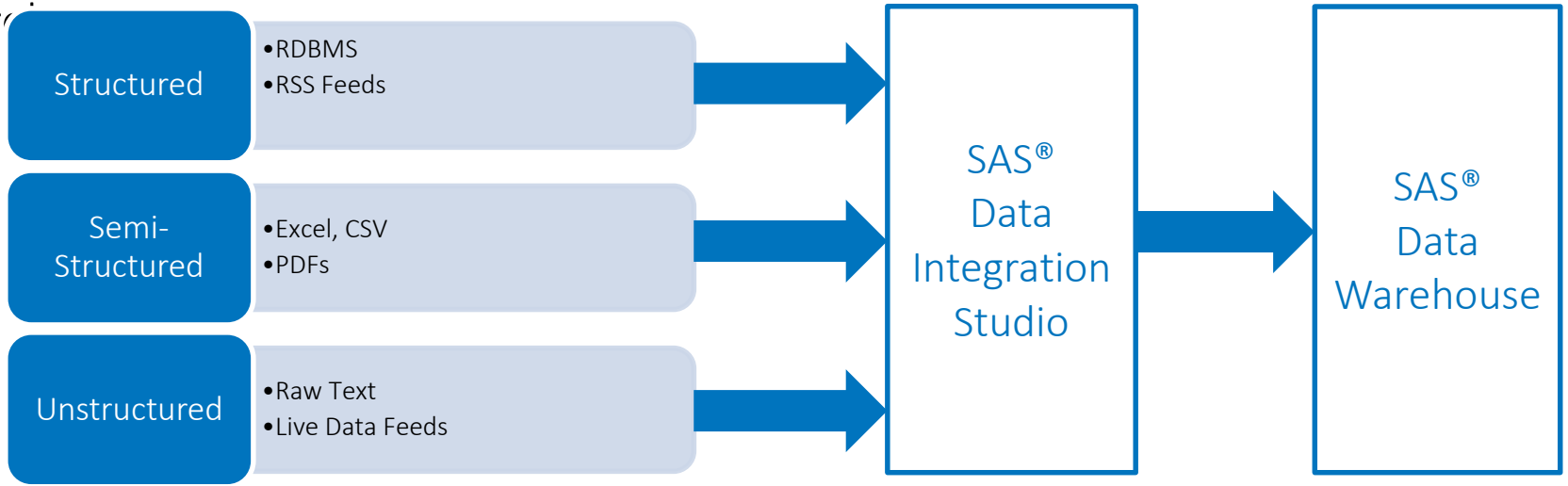
- SAS® Data Integration Studio and Dataflux contain the key core Data Science capabilities to extract raw (structured, semi-structured and unstructured) data sources, transform them into SAS® datasets and load them into a SAS® data warehouse.

SAS® Data Integration Studio	SAS® Dataflux
• Parameterized, Iterative visual workflows	• Data Explorations and Profiling
• Multithreading and scheduling	• Multithreading and Scheduling
• Forecasting and Modeling Integration	• Data Enrichment and Standardization
• Cleansing and standardization integration	• Business rules and entity resolution
• Cross Framework integration with other SAS® components and processes	

DATA SCIENCE & SAS® Data Management

SAS® Data Integration Studio

SAS® Data Integration Studio has the capability to extract data from structured, semi-structured and unstructured data sources and load them into a SAS® data ware



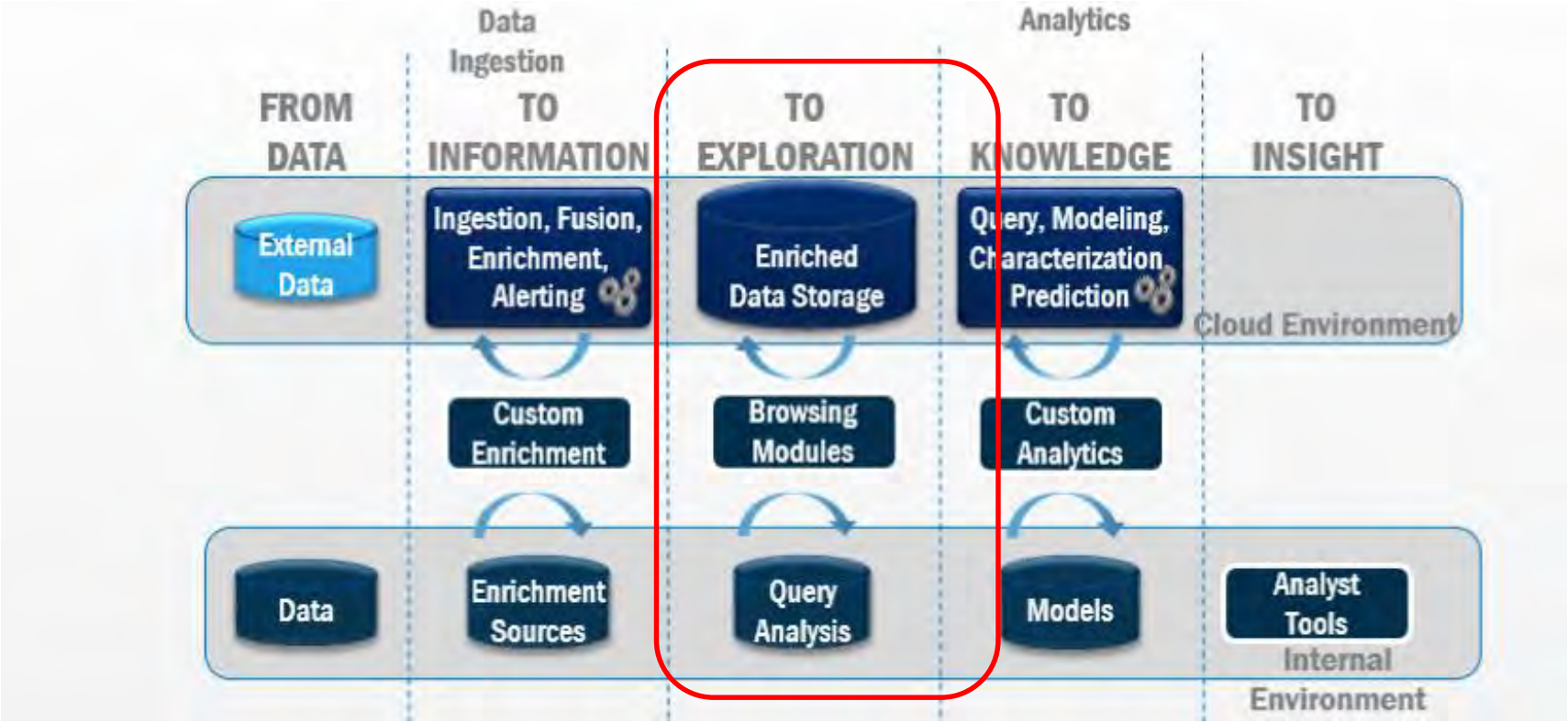
The core Data Science functions of SAS® Data Integration Studio:

- **ETL** – SAS® Data integration studio provides out of the box transformations for the ingestion of structured, semi-structured and unstructured data, both at rest and in motion(live stream ingestion).
- **Data cleansing** – both SAS® Data Integration Studio and Dataflux have out of the box data cleansing transformations that make cleansing the data a snap, even when ingesting huge volumes of data
- **Multithreading** – the loop control transformation allows the seamless multithreading across multiple cores, and is tunable to distribute the load appropriately.
- **Scheduling** – with the integration of the OS scheduler, all jobs can be deployed to be run automatically
- **Real time scoring** – SAS® Data integration studio allows the seamless application of modeling and forecasting allowing real time scoring and alerting during the ETL process. The SAS® infrastructure allows this to be applied to data at rest, in the cloud and live streaming data.

KEY DATA SCIENCE CAPABILITIES of SAS® Dataflux

- **Insuring Quality** – Dataflux can quickly and easily identify data integrity issues and suggest solutions
- **Standardization of Data**– Dataflux offers out of the box, and customizable data standardization algorithms. Comparative analysis can easily demonstrate the improved accuracy of statistics generated between the raw and standardized data.
- **Profiling** – Dataflux can quickly and easily offer insight into the data via data profiling allowing the discovery of questions not originally known to be asked from the data. Data profiling can also support modeling and forecasting initiatives.
- **Standardization of Models** – Dataflux leverages its standardization capabilities when applied to the data model ensuring data model standardization across collections of data included in the SAS® data warehouse
- **Data Enrichment** – Dataflux has out of the box data enrichment capabilities to ensure categorical uniformity across analysis from disparate data sources.

DATA SCIENCE & SAS® ANALYZING DATA



DATA SCIENCE & SAS® DATA ANALYSIS

The Key Data Science tools in SAS® Software Suite offering data analysis capabilities are:

a) **SAS® Visual Analytics** – for targeted data analysis

b) **SAS® Visual Statistics** – for targeted modeling analysis

c) **SAS® Visual Forecasting** – for building forecasting models

d) **Enterprise Guide** – can do modeling, forecasting and analysis

e) **Data Integration Studio** – can do forecasting, modeling and analysis

f) **SAS® Enterprise Miner** – for building predictive modeling analysis

g) **SAS® Forecast Studio** – for building forecasting analysis

h) JMP

DATA SCIENCE & SAS® DATA VISUALIZATION

SAS® VISUAL ANALYTICS

The Key Data Science data visualization capabilities in SAS® come from:

1.) SAS® Visual Analytics – for generating targeted visualizations on specific data

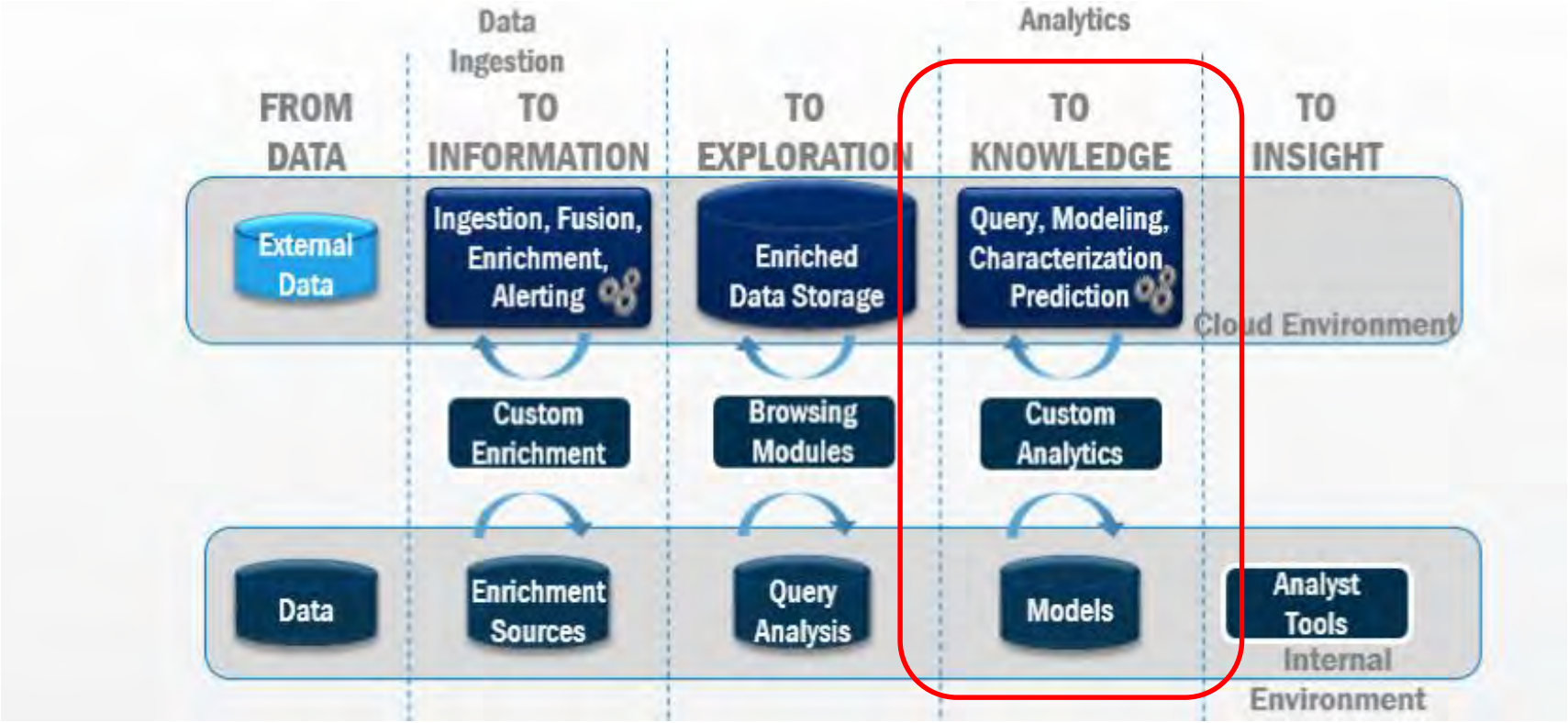
2.) Enterprise Guide – for generating visualizations on collections of data

3.) Data Integration Studio – for generating visualizations on large collections

4.) SAS® Enterprise Miner – for building predictive modeling visualizations

5.) SAS® Forecast Studio – for building forecasting visualizations

SAS® & Modeling

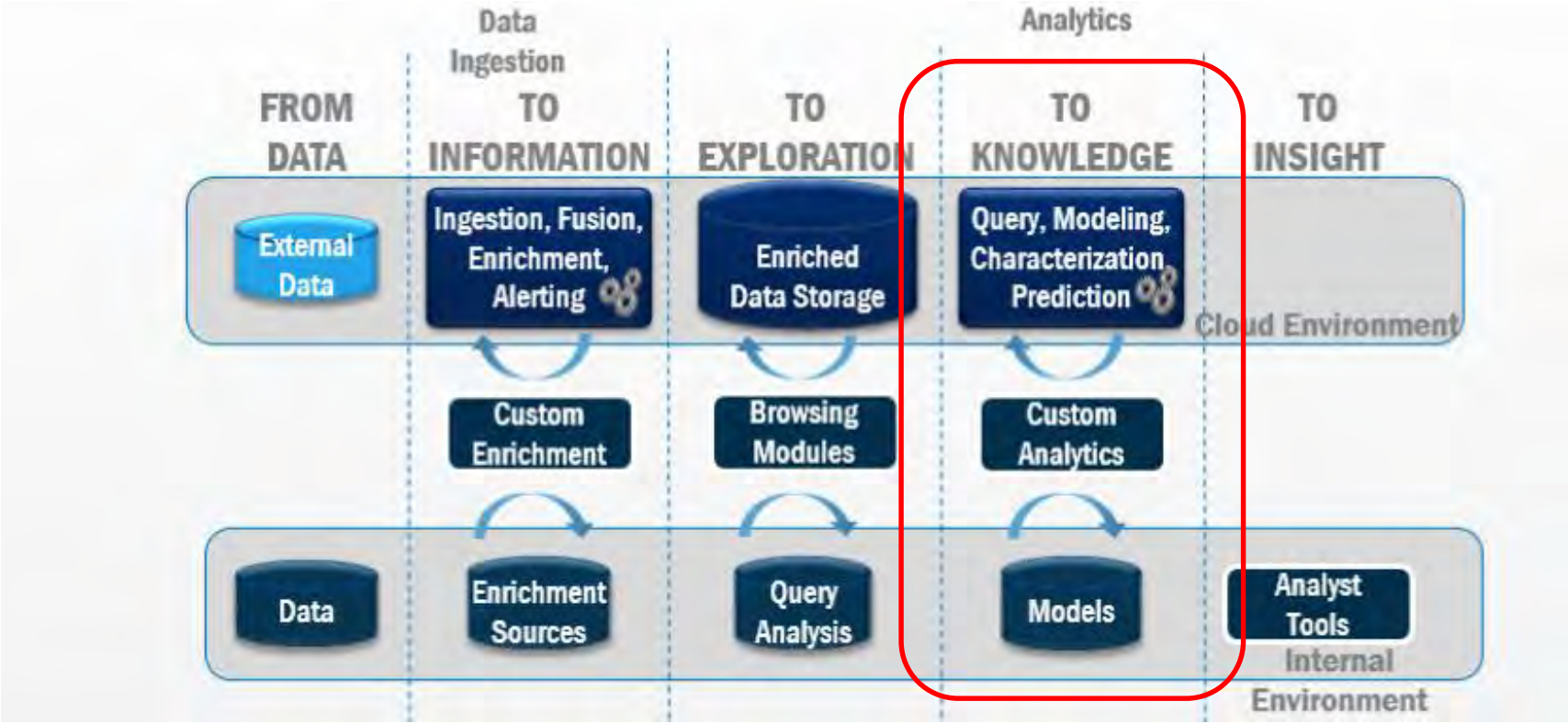


DATA SCIENCE & MODELING IN SAS®

SAS® Modeling Capabilities

- Data Scientists primarily choose **SAS® Enterprise Miner** as the primary tool for building predictive models
- **SAS® Visual Statistics** offers data scientists model building capabilities as an alternative to SAS® Enterprise Miner
- Data scientists can leverage the rapid modeling building capability within **SAS® Enterprise Guide, t**hese results can later be imported and augmented within SAS® Enterprise Miner

SAS® & FORECASTING



DATA SCIENCE & FORECASTING IN SAS®

SAS® Forecasting Capabilities

- Data scientists use **SAS® Forecasting Studio** as the primary tool for building forecasting models
- Data scientists can also leverage **SAS® Visual Forecasting** that offers forecasting model building capabilities with the expanded tools with **SAS® Viya**
- Data scientists and data managers conducting big data processing will leverage **SAS® Data Integration Studio** that has its own Forecasting module and can integrate forecasting models built within **SAS® Forecasting Studio** into ETL flows for generating forecasts at the time of data ingestion

Data Science & BIG Data

How does a Data Scientist look at Big Data?

Engineering View

- ✓ Volume
- ✓ Velocity
- ✓ Variety

Business View

- ✓ Value

Science

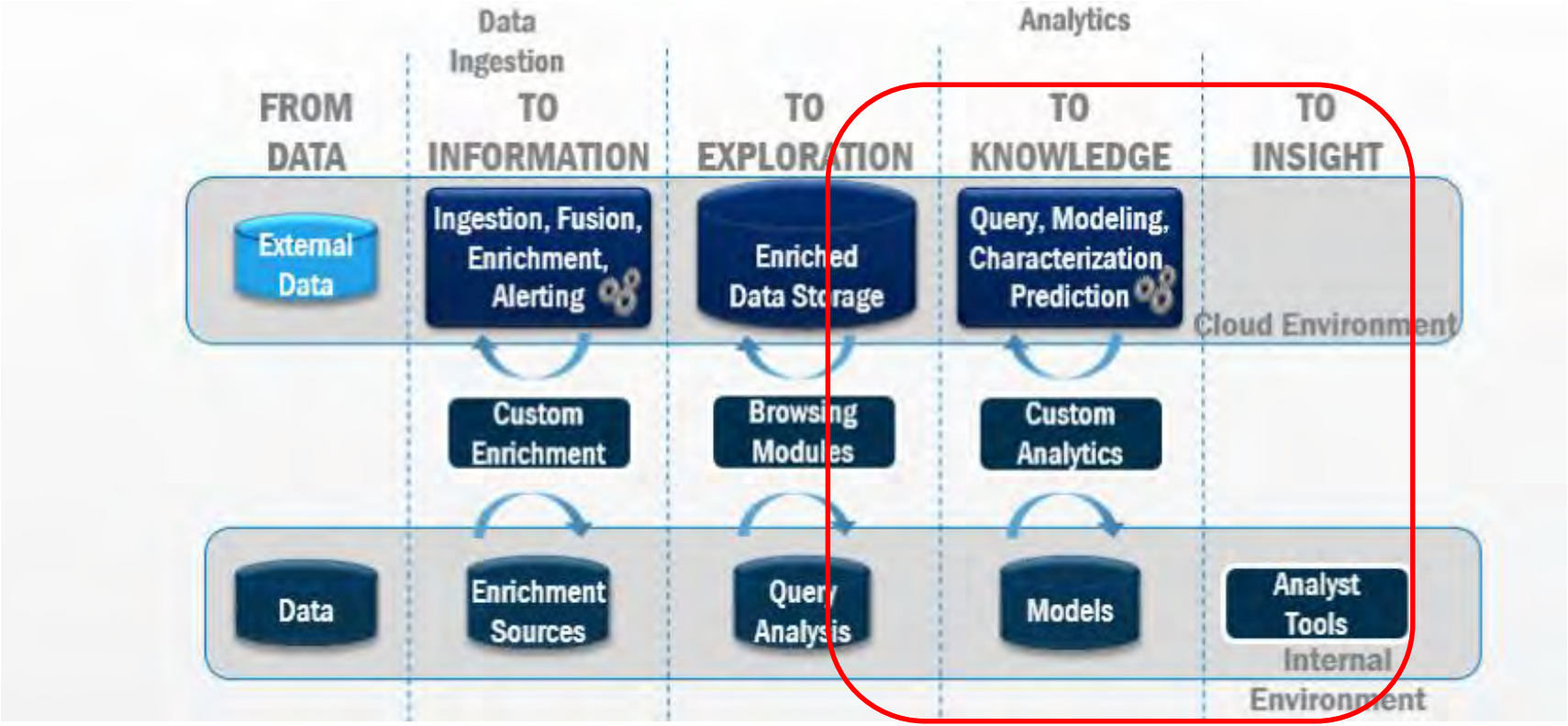
- ✓ Veracity
- ✓ Complexity
- ✓ Cleanliness
- ✓ Completeness
- ✓ Consistency
- ✓ Latency
- ✓ Provenance

Data Types

- ✓ Structured
- ✓ Semi-structured
- ✓ Unstructured

SAS® & BIG DATA

High Performance & Cloud Implementations



SAS® Suite of High Performance Procedures

- Data scientists and big data analysts benefit from a collection of high performance procedures that leverage multi-node\grid\cloud infrastructure with emphasis on data mining, econometrics, forecasting, optimization, statistics and text mining, empowering data scientists with the tools they need to accomplish any task on any scale of data.

SAS® Cloud Computing Capabilities

Data scientists and big data analyst extend their capabilities by leveraging a collection of SAS® tools that possess capability of running in a cloud environment.

- SAS® Analytics for Containers
- SAS® Enterprise Miner
- SAS® Forecast Studio
- SAS® Data Integration Server
- SAS® Visual Analytics
- SAS® Visual Statistics

DATA SCIENCE & SAS® BIG DATA

VIRTUAL MACHINE

To meet the challenge of big data processing, the SAS® software can be deployed in four logical components or tiers and each tier can span multiple virtual machines:

Metadata Tier	Compute (Application) Tier
MDT_VM1, MDT_VM2 ……. MDT_VM(N)	COT_VM1, COT_VM2 ……. COT_VM(N)
Web (middle) Tier	**Client Tier**
WMT_VM1, WMT_VM2 ……. WMT_VM(N)	CLT_VM1, CLT_VM2 ……. CLT_VM(N)

QUESTIONS?

Your feedback counts!

Don't forget to complete the session survey
in your conference mobile app.

1. Go to the Agenda icon in the conference app.
2. Find this session title and select it.
3. On the sessions page, scroll down to Surveys and select the name of the survey.
4. Complete the survey and click Finish.