Real World Evidence using AI/ML on SAS® Viya

SAS Institute Japan

Ryosuke Horiuchi – Senior Consultant, Advanced Analytics CoE

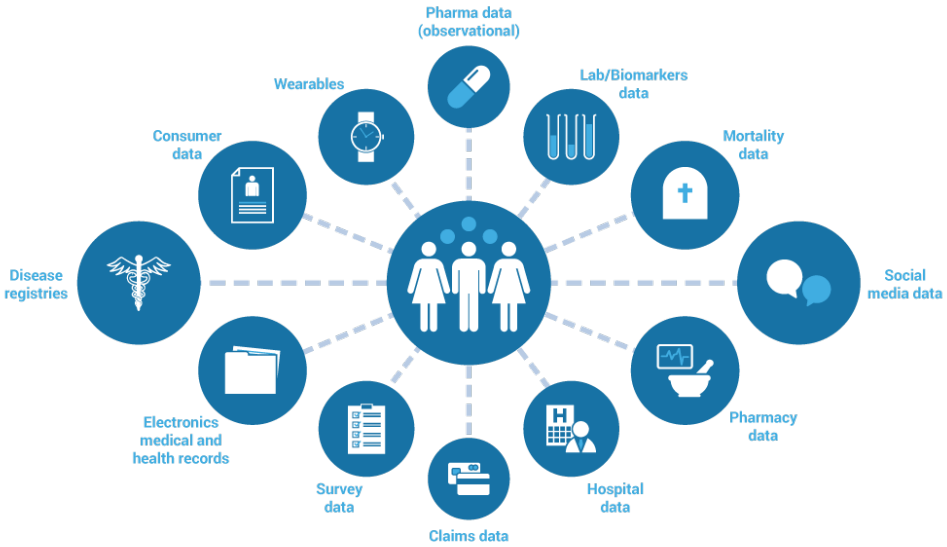
PharmaSUG Tokyo 2019
October 24th

Agenda

- Introduction
- Problem definition
- Technique & Mechanism
- The use case of Text Mining with Viya
- The demo for implementing Factorization Machines in Viya
- Conclusions

Introduction

Real World Evidence (RWE) means evidence obtained from Real World Data (RWD), which are observational data obtained outside the context of randomized controlled trials (RCTs) and generated during routine clinical practice.



- ✓ The characteristics of RWD, the game changer
- ✓ Various formats of data means various analytics tools & solutions are needed
- ✓ Unconventional techniques are needed

Source: "GLOBAL REAL WORLD EVIDENCE SOLUTIONS MARKET RESEARCH REPORT FORECAST 2018–2025"

Problem Definition

The characteristics of RWD compared to Clinical Trial Data

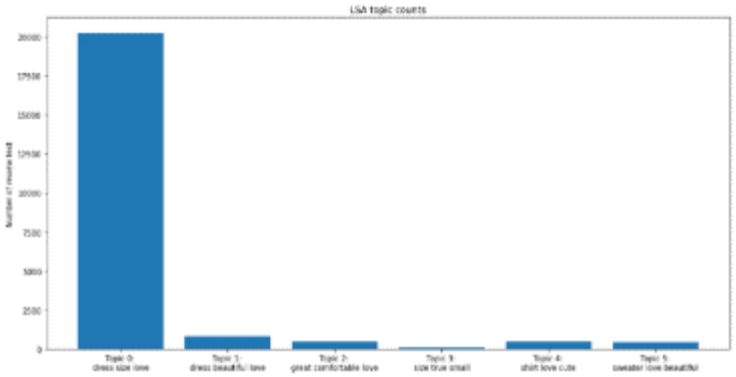
Real World Data	Clinical Trial Data
• Big data • Unstructured data • Longitudinal data • Sparse data • Not randomized	• Not so big • Structured data • Cross sectional data • Dense data • Randomized

The sources of RWD:
Electronic health records, medical claims, wearable devices, patient-reported outcomes such as forums, social media, etc.

Unstructured data (example)

Many sources of RWD contain large amounts of texts (e.g., EHRs, patient-reported outcomes such as forums, social media). We need to parse and vectorize it in order to proceed analytics using AI. Topic extraction, sentimental analysis, etc., can be applied. These techniques are crucial especially in the field of adverse event detection and pharmacovigilance activities as well as regulatory compliance activities.

	Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating
0	0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4
1	1	1080	34	NaN	Love this dress! it's sooo pretty. I happene...	5
2	2	1077	60	Some major design flaws	had such high hopes for this dress and reall...	3
3	3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5
4	4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5



Topic segmentation based on LSA

Source: “Women’s E-Commerce Clothing Reviews” (Kaggle dataset)

Sparse data (example)

Alice	Bob	Charlie	drug1	drug2	drug3	drug4	drug5	drug6	drug7	drug8	Outcome
1	0	0	1	0	0	0	0	0	0	0	1
1	0	0	0	0	1	0	0	0	0	0	5
0	1	0	0	1	0	0	0	0	0	0	3
0	1	0	0	0	0	0	0	0	1	0	11
0	0	1	0	0	0	0	1	0	0	0	9

Patient	Treatment	Outcome
Alice	drug 1	1
Alice	drug 2	
Alice	drug 3	5
Alice	drug 4	
Alice	drug 5	
Alice	drug 6	
Alice	drug 7	
Alice	drug 8	

Patient	Treatment	Outcome
Bob	drug 1	
Bob	drug 2	3
Bob	drug 3	
Bob	drug 4	
Bob	drug 5	
Bob	drug 6	
Bob	drug 7	11
Bob	drug 8	

Patient	Treatment	Outcome
Charlie	drug 1	
Charlie	drug 2	
Charlie	drug 3	
Charlie	drug 4	
Charlie	drug 5	9
Charlie	drug 6	
Charlie	drug 7	
Charlie	drug 8	

The reason why RWD tends to be sparse is that many different kinds of treatments exist.

However, the patients of interest do not take all. Here emerge the missing values that makes the matrix sparse when one-hot encoding is applied.

The drawback of sparse data:
In general, it's difficult to learn compared to with dense data, e.g., SVMs fail to estimate parameters during training.

Not randomized data (example)

Patient	Untreated	Treated	Outcome (y_0)	Outcome (y_1)
Alice	0	1		3
Bob	0	1		7
Charlie	0	1		9
Dave	1	0	6	
Eve	1	0	7	
Frank	1	0	6	

RWD is not designed as a randomized control trial. You cannot compare the outcomes between treated and untreated.

$$\mathrm{ATE} = \frac{1}{N}\sum_i y_1(i) - y_0(i)$$

In order to calculate the Average Treatment Effect (ATE) under this restriction, we ideally need both Outcome y_0 and Outcome y_1 for each patient. However, it is not realistic to have both in real world.

Traditionally, Propensity Score Matching, Inverse Probability of Treatment Weights (IPTWs) have been used to calculate ATE. However, those methods are vulnerable to noise which increases as the number of covariates increases.

Here counterfactual prediction methods such as G Computation, Targeted Maximum Likelihood Estimation (TMLE), Variational Autoencoder (VAE) , etc., come into play.

AI IN HEALTH CARE AND LIFE SCIENCES

INFERENCE

- Two-sample Hypothesis Testing
- ANOVA
- ANCOVA
- Linear Regression Analysis
- Generalized Linear Models

Randomized Controlled Trial Data

PREDICTION

- Generalized Linear Models
- Support Vector Machine
- Decision Tree
- Deep Neural Network
- Factorization Machines
- Generative Adversarial Networks

Longitudinal Data (Repeated Cross-sectional Data)

CAUSAL INFERENCE

- Propensity Score Matching
- Invers Probability of Treatment Weights
- G Computation
- Targeted Maximum Likelihood Estimation
- Variational Autoencoder

Not-randomized Data

sas

Technique & Mechanism

SAS® Viya™

Source-based Engines

In-Stream

In-Hadoop

In-Database

Parallel & Serial, Pub / Sub, Web Services, MQs

In-Memory Runtime Engine

Microservices

- UAA
- Data Source Mgmt
- Folders
- etc...
- BI GUIs
- CAS Mgmt
- Log
- Analytics GUIs
- Data Mgmt GUIs
- Query Gen
- Env Mgr
- Model Mgmt
- Audit

Cloud Analytics Services (CAS)

Solutions

- Customer Intelligence
- Analytics
- Risk Management
- Business Visualization
- Fraud and Security Intelligence
- Data Management

APIs

sas, REST API, python, Java, R, Lua

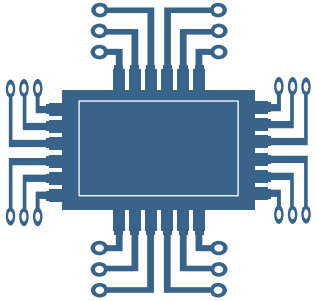
Platform

CLOUD FOUNDRY, docker, CI, ANSIBLE, rpm

amazon web services, Azure, Google Cloud Platform, openstack, vmware, ORACLE CLOUD

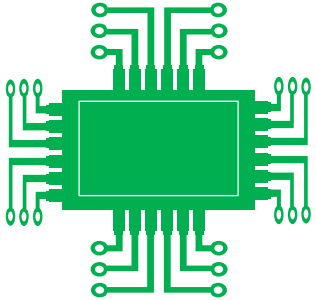
CAS & SPRE

Viya has two compute engines



CAS: Cloud Analytic Services

- The in-memory engine
- Data is distributed across all CAS worker nodes and threads
- Processing is done in parallel



SPRE: SAS Programming Runtime Environment

- Also referred to as Viya's Compute Server
- Also referred to as Viya's Workspace Server

The use case of Text Mining with Viya

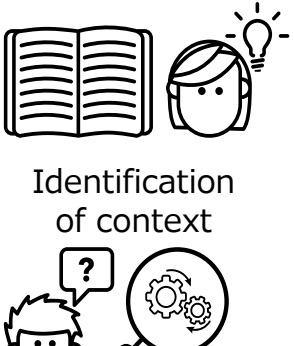
SAS Visual Text Analytics

NLP by combination of ML & Rules-based methods

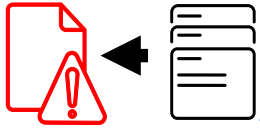
SAS Visual Text Analytics (VTA) provides integrated functions for text analytics through GUI. VTA automatically extracts information and enables to generate valuable insights with combinations of other analytical & machine learning methods from text.

Major use cases

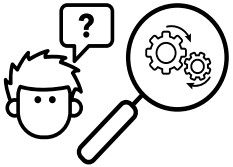
Automatic tagging (categorization)



Identification of context



Early detection of problem

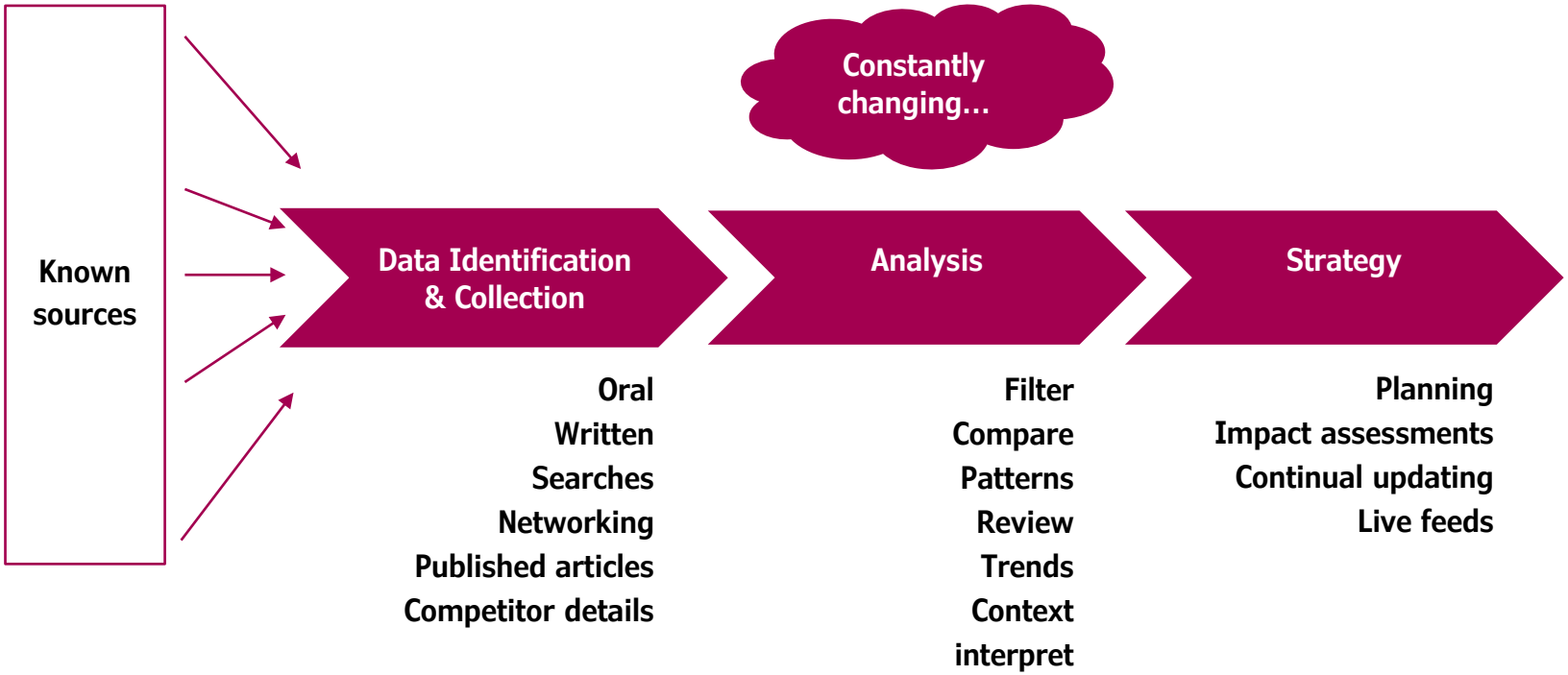


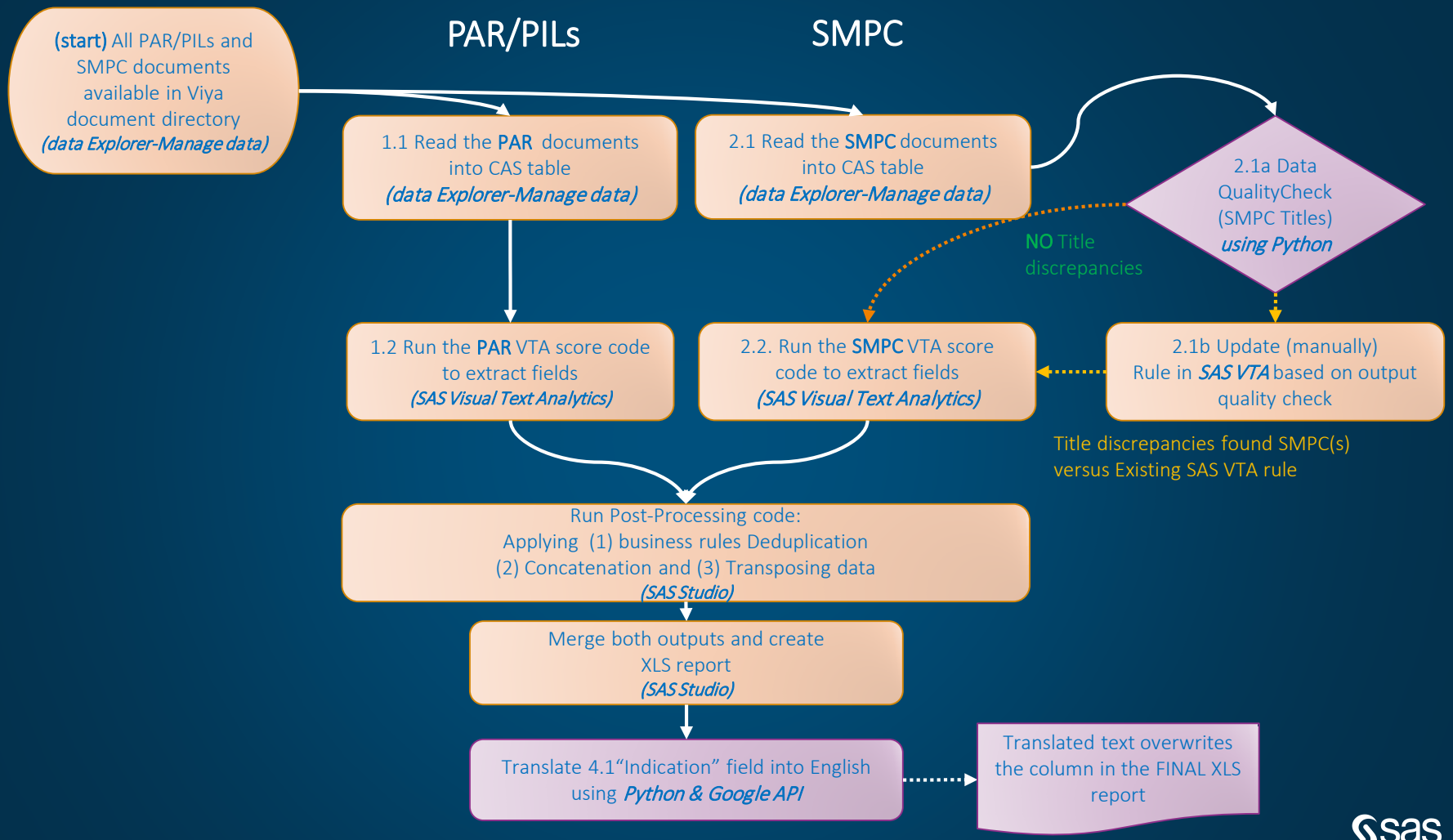
Cause analysis of problem

Major functions

- Enables NLP with GUI on web browser
- Creates visual text analytics model through only a few clicks
- Realizes not only automatic creation but also fine tuning of text analytics model
- 32 languages including English & Japanese are supported.
- The following functions are provided.
  - Extraction of concept
  - Text mining
  - Sentiment analysis
  - Automatic extraction of topics by machine learning and rule creation for categorization
  - Tagging
  - Deployment of model for categorizing and scoring (Real time scoring is also available)
  - Structures text and integrates with predictive model

Use case of demo for NLP: Regulatory Intelligence process





Text Analytics for automated report generation (example)

Paracetamol and phenylephrine
Spain

Product name	Marketing Authorisation Holder	MAH Number	Composition (API Info, Strength)	Procedure	CMS	Legal basis	Prescription Status	Indication (English translation)
Paracetamol /Fenilefrina Sandoz 1.000 mg/12,2 mg polvo para solución oral EFG	Sandoz	79272	1000mg paracetamol and 12,2 mg phenylephrine HCl	UK/H/5441/02/DC	BG, CZ, EE, EL, ES, HU, IE, LV, LT, RO	Art. 10.1 generic application	Rx, not marketed	For the relief of the symptoms of colds and flu-like infections, including the relief of pain, headache, fever and nasal congestion in adults and children over 16 years of age
Cinfacold 1000 mg/10 mg oral solution	Laboratorios Cinfa	69931	1000mg paracetamol and 12,2 mg phenylephrine HCl	National	N/A	Art. 10.1 generic application	Rx	Symptomatic relief of catarrh or flu-like processes with mild or moderate pain, fever and nasal congestion. in adults and children over 18 years of age
pharmagrip dúo 650 mg/8,2 mg oral solution	Sandoz	69932	650mg paracetamol and 10 mg phenylephrine HCl	National	N/A	N/A	OTC	For the relief of the symptoms of colds and flu-like infections, including the relief of pain, headache, fever and nasal congestion in adults and children over 14 years of age

AT BE BG CZ DE DK EE EL ES FI FR HU HR IE IT LT LV MT NL NO PL PT RO SE SI SK

The Demo for Implementing Factorization Machines with Viya

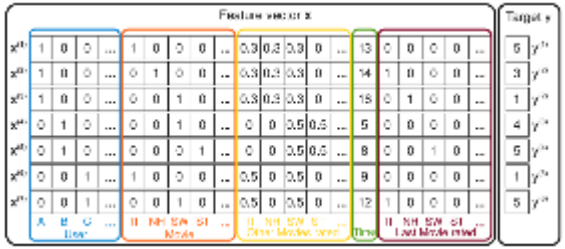
Factorization Machines
The overview and its basic algorithm

Factorization Machines are proposed by Steffen Rendle when he was studying in Osaka Universty, Japan in 2010.

Users = {Alice (A), Bob (B), Charlie (C), …}
Items = {Titanic (TI), Notting Hill (NH), Star Wars (SW), Star Trek (ST), …}

The observed data =

{(A, TI, 2010-1, 5), (A, NH, 2010-2, 3), (A, SW, 2010-4, 1), (B, SW, 2009-5, 4), (B, ST, 2009-8, 5), (C, TI, 2009-9, 1), (C, SW, 2009-12, 5)}



$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i\, x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle\, x_i\, x_j$$

w_0:
global bias (the average rating over all users and movies)

w_i, <v_i, v_j>:
per-user bias (the average of the ratings given by the user)
per-item bias (the average of the ratings given to that movie)
pairwise interaction term between the user and that particular movie

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle := \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f}$$

PROC FACTMAC

The following statements show how to use PROC FACTMAC to predict movie ratings:

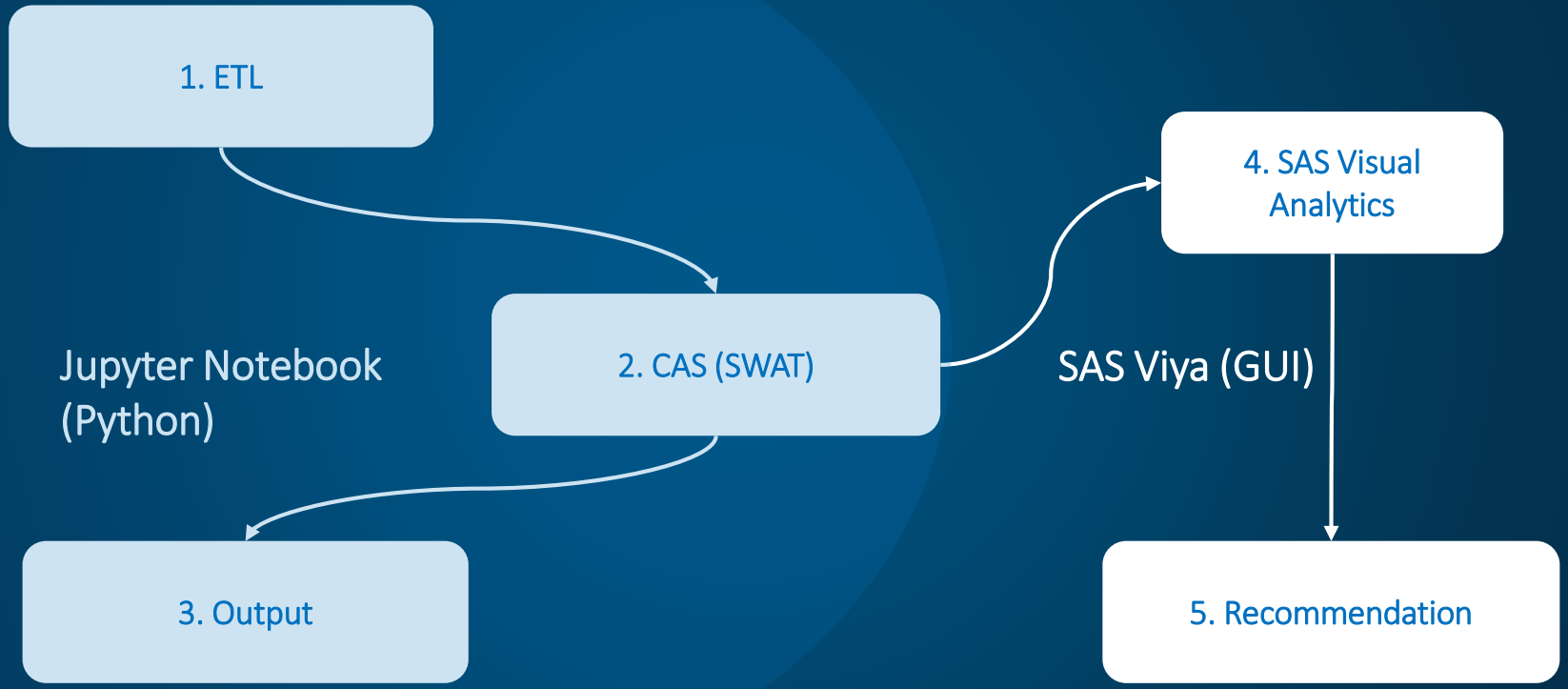
```
-------------------------------------------
proc factmac data=mycas.movlens nfactors=10 learnstep=0.15
          maxiter=20  outmodel=mycas.factors;

  input userid itemid /level=nominal;
  target rating /level=interval;
  output out=mycas.out1 copyvars=(userid itemid rating);
run;
```

Obs	userid	itemid	rating	P_rating
1	196	242	3	4.09834
2	186	302	3	3.78284
3	22	377	1	1.42463
4	244	51	2	3.20907
5	166	346	1	2.58391
6	298	474	4	4.60470
7	115	265	2	3.43183
8	253	465	5	4.57718
9	305	451	3	2.96174
10	6	86	3	4.44866
11	62	257	2	3.08217

Generalized matrix factorizations, among other techniques read and write data in distributed form, and perform factorization in parallel by making full use of multicore computers or distributed computing environments.

Source: "SAS Visual Data Mining and Machine Learning 8.1: Data Mining and Machine Learning Procedures" (SAS Institute Inc.)

Process Flow



DEMO

Stay tuned…

PROC CAUSALTRT

Estimates the average causal effect of a binary treatment, T, on a continuous or discrete outcome, Y.
Good for data from nonrandomized trials or observational studies.

The following statements invoke PROC CAUSALTRT to estimate the ATE of attending a Catholic high school on math scores:

```
-----------------------------------------
ods graphics on;
proc causaltrt data=school covdiffps poutcomemod nthreads=2;
        class Income FatherEd MotherEd;
        psmodel Catholic(ref='No') = Income FatherEd MotherEd;
        model Math = BaseMath Income FatherEd MotherEd;
        bootstrap seed=1234 plots=hist(effect);
run;
```

Source: "Estimating Causal Effects from Observational Data with the CAUSALTRT Procedure" (Michael Lamm and Yiu-Fai Yung, SAS Institute Inc.)
"ENHANCING THE APPLICATION OF REAL WORLD EVIDENCE WITH EPISODE-OF-CARE ANALYTICS" (David Olaleye and Lina Clover, SAS, Cary, NC)

Conclusion

- ✓ For sparse data, SAS Viya can be used for handling sparse matrix and missing values
- ✓ For unstructured data, SAS Viya can be used for handling text data and further NLP
- ✓ For causal inference, SAS provides special procedures to calculate ATE
- ✓ SAS Viya provides various types of capabilities to handle RWD depending on its forms

Thank you !

sas.com

THE POWER TO KNOW.