

IDAR virtual AI assistant to automate statistical report generation

Janssen R&D



Presenters



Shobha Rani

Lead, Intelligent Automation,
IT, Janssen R&D, New Jersey, US



Harry Chen

Analytics Innovation Technical Lead,
SP&A, Janssen R&D, Shanghai, China



Sumesh Kalappurakal

Head, Analytics Innovation, SP&A,
Janssen R&D, New Jersey US

Acknowledgement

Business

- ❖ Darren Weston
- ❖ Olivier Leconte

IT

- ❖ Jitendra Sinha
- ❖ Richard Franckowiak Jr.
- ❖ Srikanth Ramakrishnan
- ❖ Akash Srivastava

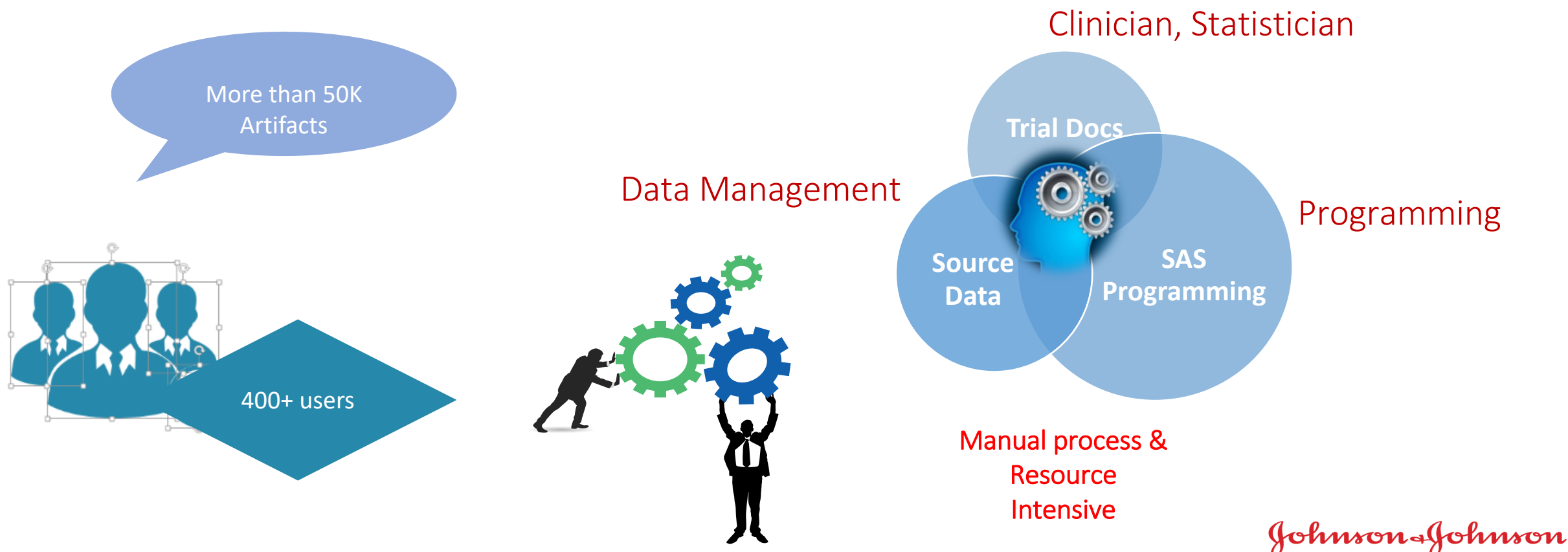
Special thanks to PhUSE Organization/Stream chairs for providing an opportunity to participate and present in this event

Problem Statement



Problem Statement

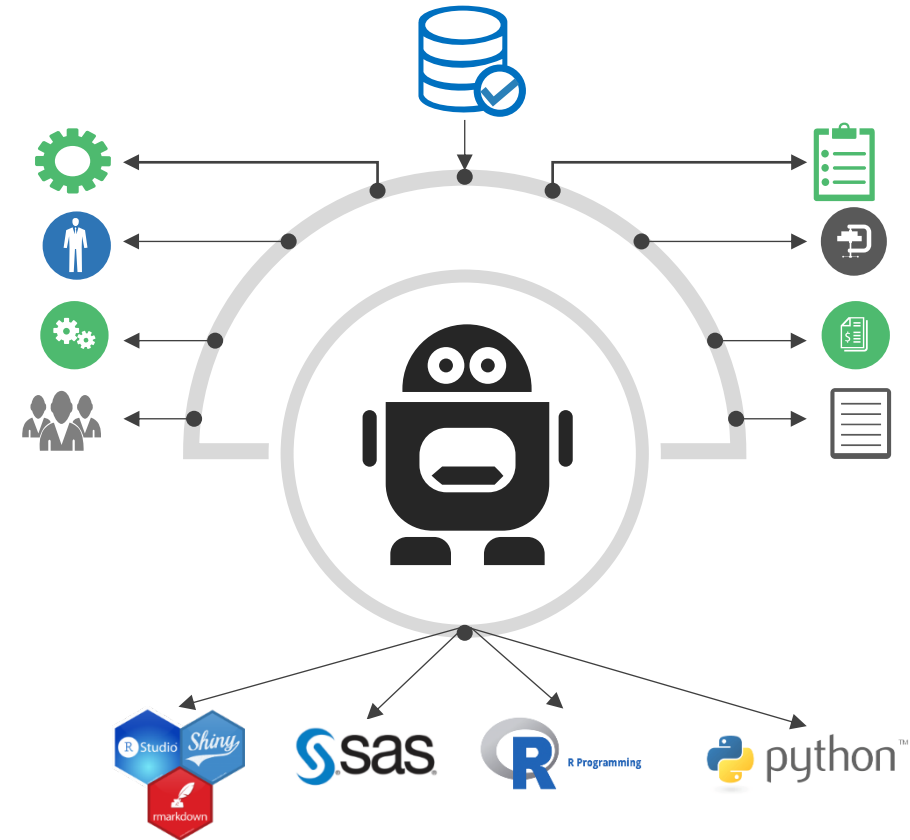
Objective of this project is to develop a process using cutting edge technology and methodology to **eliminate repetitive work, increase automation, enhance operational process and reduce cycle time**. Also engage highly qualified resources to focus on highly valued areas to **meet the unmet needs**.



Advance Automation



Current manual process



Future Automation

Journey toward Automation



Journey towards Automation

Part 1

Part 1 of the Autocode project, and we focus on macro call statement and IF-THEN/ELSE logic generation from tabular metadata by rule-based system

Macro call

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851	852	853	854	855	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883	884	885	886	887	888	889	890	891	892	893	894	895	896	897	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912	913	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930	931	932	933	934	935	936	937	938	939	940	941	942	943	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960	961	962	963	964	965	966	967	968	969	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	988	989	990	991	992	993	994	995	996	997	998	999	1000
DATASET	VARIABLE	VARVER	VARIABLE	REASON	COMPLT	CRISCOM	REVIEWP	SAMMR																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															

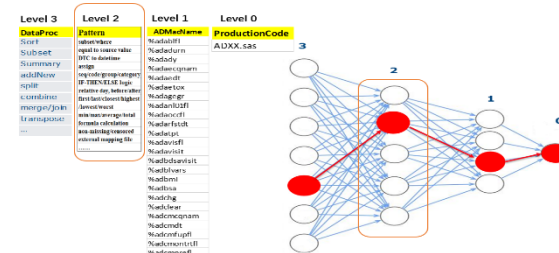
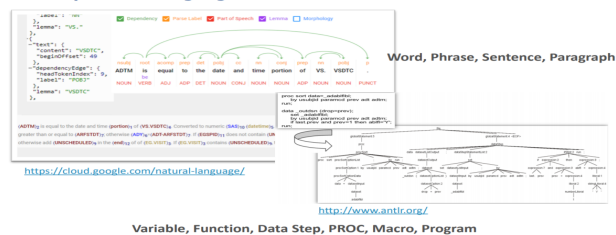
IF-THEN/ELSE logic in Data Step

Overall response at 1st time point (X ₁)	Overall response at 2nd time point (X ₂)	Overall response at 3rd time point (X ₃)	Best overall Response
CR	Unknown	Unknown	Final: SD if X ₁ = XX otherwise, NE Interim: If additional disease assessments are expected* then SD if X ₁ = XX otherwise, NE else UCR
CR with X ₁ = X ₂ = YV	Any	CR	
CR with X ₁ = YV	CR	CR	
1 CR	R01 R02	X1 X1	R03 R04 R05 R06 R07 R08 R09
2 CR	Unknown	Unknown	R1+XX SD NE
3 CR	R2+YY	Unknown	CR CR
4 CR	R2+YY PR	Unknown	R2+XX SD NE
5 CR	R2+YY PR	R2+XX	SD NE
6 CR	R2+YY SD	R2+XX	PD
7 CR	R2+YY SD	R2+XX	PD
8 CR	R2+YY SD	R2+XX	PD
9 CR	R2+YY SD	R2+XX	PD
10 CR	R2+YY SD	R2+XX	PD
11 CR	R2+YY SD	R2+XX	PD
12 CR	R2+YY SD	R2+XX	PD
13 PR	R2+YY	R2+XX	PD
14 CR	R2+YY	R2+XX	PD
15 PR	R2+YY	R2+XX	PD
16 CR	R2+YY	R2+XX	PD
17 CR	R2+YY	R2+XX	PD
18 CR	R2+YY	R2+XX	PD
19 CR	R2+YY	R2+XX	PD
20 CR	R2+YY	R2+XX	PD
21 CR	R2+YY	R2+XX	PD
22 CR	R2+YY	R2+XX	PD
23 CR	R2+YY	R2+XX	PD
24 CR	R2+YY	R2+XX	PD
25 CR	R2+YY	R2+XX	PD
26 CR	R2+YY	R2+XX	PD
27 CR	R2+YY	R2+XX	PD
28 CR	R2+YY	R2+XX	PD
29 CR	R2+YY	R2+XX	PD
30 CR	R2+YY	R2+XX	PD
31 CR	R2+YY	R2+XX	PD
32 CR	R2+YY	R2+XX	PD
33 CR	R2+YY	R2+XX	PD
34 CR	R2+YY	R2+XX	PD
35 CR	R2+YY	R2+XX	PD
36 CR	R2+YY	R2+XX	PD
37 CR	R2+YY	R2+XX	PD
38 CR	R2+YY	R2+XX	PD
39 CR	R2+YY	R2+XX	PD
40 CR	R2+YY	R2+XX	PD
41 CR	R2+YY	R2+XX	PD
42 CR	R2+YY	R2+XX	PD
43 CR	R2+YY	R2+XX	PD
44 CR	R2+YY	R2+XX	PD
45 CR	R2+YY	R2+XX	PD
46 CR	R2+YY	R2+XX	PD
47 CR	R2+YY	R2+XX	PD
48 CR	R2+YY	R2+XX	PD
49 CR	R2+YY	R2+XX	PD
50 CR	R2+YY	R2+XX	PD
51 CR	R2+YY	R2+XX	PD
52 CR	R2+YY	R2+XX	PD
53 CR	R2+YY	R2+XX	PD
54 CR	R2+YY	R2+XX	PD
55 CR	R2+YY	R2+XX	PD
56 CR	R2+YY	R2+XX	PD
57 CR	R2+YY	R2+XX	PD
58 CR	R2+YY	R2+XX	PD
59 CR	R2+YY	R2+XX	PD
60 CR	R2+YY	R2+XX	PD
61 CR	R2+YY	R2+XX	PD
62 CR	R2+YY	R2+XX	PD
63 CR	R2+YY	R2+XX	PD
64 CR	R2+YY	R2+XX	PD
65 CR	R2+YY	R2+XX	PD
66 CR	R2+YY	R2+XX	PD
67 CR	R2+YY	R2+XX	PD
68 CR	R2+YY	R2+XX	PD
69 CR	R2+YY	R2+XX	PD
70 CR	R2+YY	R2+XX	PD
71 CR	R2+YY	R2+XX	PD
72 CR	R2+YY	R2+XX	PD
73 CR	R2+YY	R2+XX	PD
74 CR	R2+YY	R2+XX	PD
75 CR	R2+YY	R2+XX	PD
76 CR	R2+YY	R2+XX	PD
77 CR	R2+YY	R2+XX	PD
78 CR	R2+YY	R2+XX	PD
79 CR	R2+YY	R2+XX	PD
80 CR	R2+YY	R2+XX	PD
81 CR	R2+YY	R2+XX	PD
82 CR	R2+YY	R2+XX	PD
83 CR	R2+YY	R2+XX	PD
84 CR	R2+YY	R2+XX	PD
85 CR	R2+YY	R2+XX	PD
86 CR	R2+YY	R2+XX	PD
87 CR	R2+YY	R2+XX	PD
88 CR	R2+YY	R2+XX	PD
89 CR	R2+YY	R2+XX	PD
90 CR	R2+YY	R2+XX	PD
91 CR	R2+YY	R2+XX	PD
92 CR	R2+YY	R2+XX	PD
93 CR	R2+YY	R2+XX	PD
94 CR	R2+YY	R2+XX	PD
95 CR	R2+YY	R2+XX	PD
96 CR	R2+YY	R2+XX	PD
97 CR	R2+YY	R2+XX	PD
98 CR	R2+YY	R2+XX	PD
99 CR	R2+YY	R2+XX	PD
100 CR	R2+YY	R2+XX	PD
101 CR	R2+YY	R2+XX	PD
102 CR	R2+YY	R2+XX	PD
103 CR	R2+YY	R2+XX	PD
104			

Part 2

Part 2 of the Autocode project, our exploration was using NLP – both supervised and un-supervised methods

NLP (NATURAL LANGUAGE PROCESSING)
- Map human language to Code

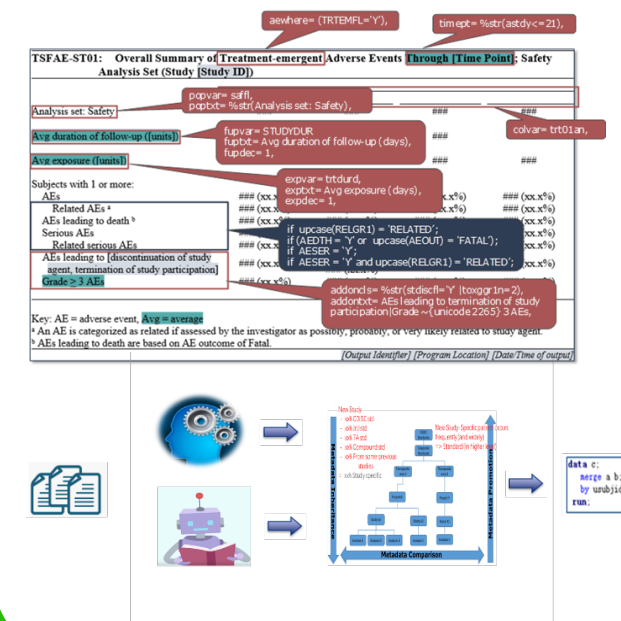


Assign to Y on last non-missing record within PARAMCD where ADT <= ARFSTD.

```
%Assign(indsn=xxx,value='Y',by=PARAMCD,select=last non-missing,subset=%str(<ADT<=ARFSTDT) )
```

Part 3

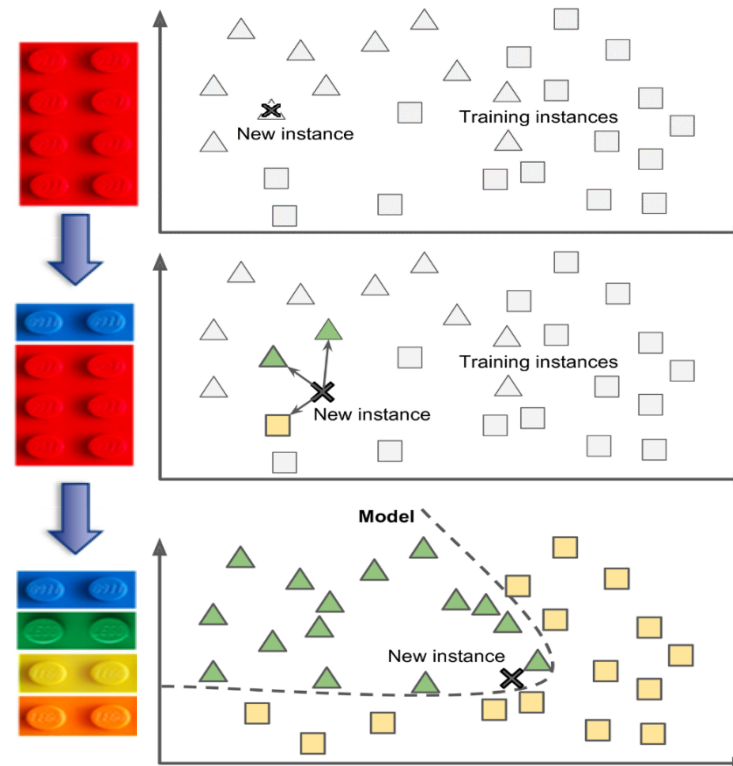
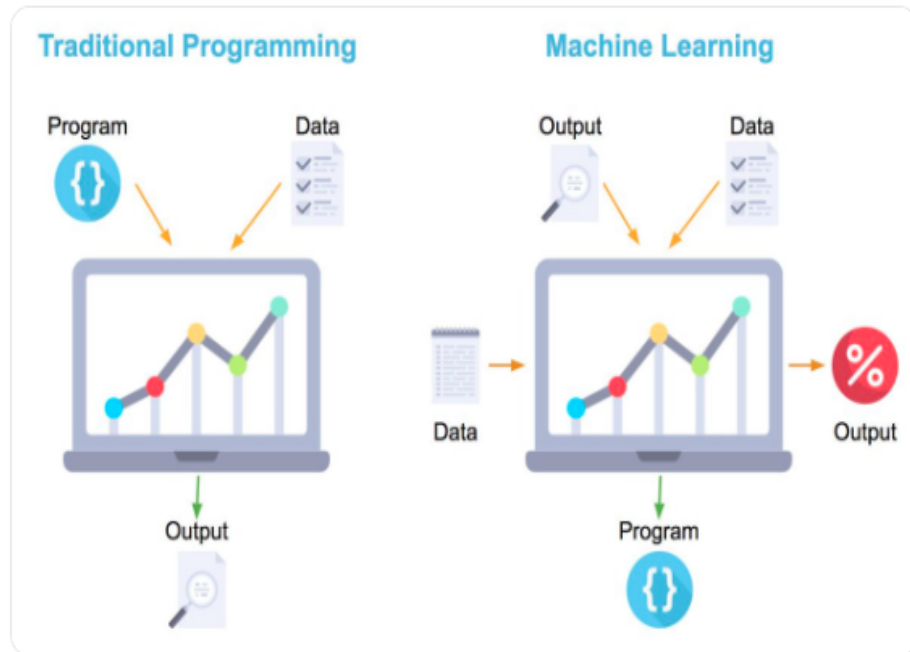
Part 3 of the Autocode project, we explored with NLU and NER engine to generate TLG using a Virtual AI assistant to automate this task with minimum inputs and manual effort



Approach



Approach



- ✓ **Exact match**
(rule-based system)
- ✓ **Fuzzy match**
(instance-based learning)
- ✓ **Smart match**
(model-based learning)

Hands-On Machine Learning with Scikit-Learn and TensorFlow, Aurélien Géron Published by O'Reilly Media, Inc.

Approach

TLG title NLU - title **auto** analyzed to **structured** data – map to code

Sentence

Treatment-emergent Serious Adverse Events with Grade 3 or Higher by System Organ Class and Preferred Term (at Least 1 Percent in Any Treatment Group)

Intent

aesummary

Entities

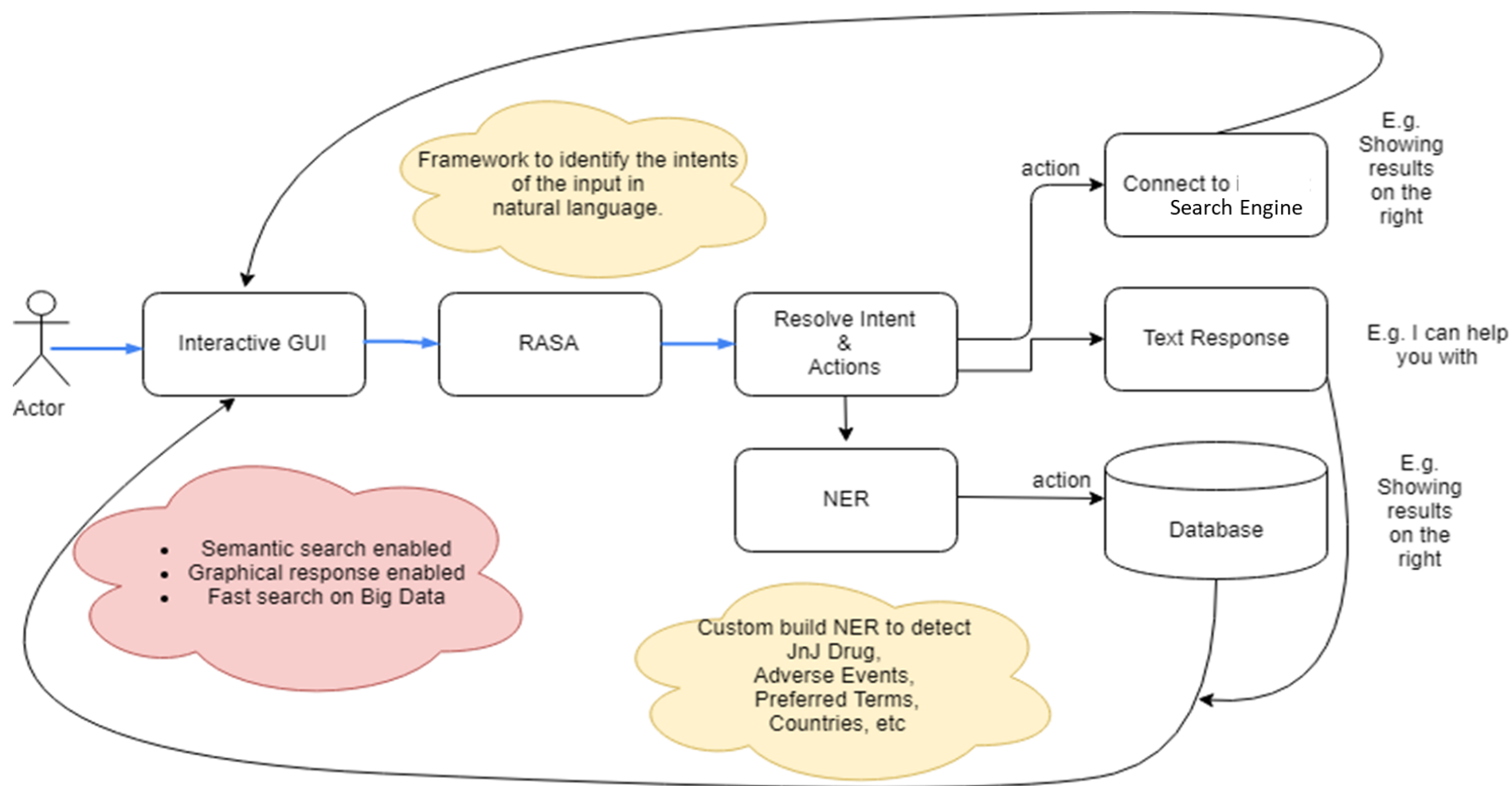
Synonym	Entity
Treatment-emergent	teae
Serious	seriousAE
Grade 3 or Higher	aeGrade
System Organ Class	bygroup
Preferred Term	bygroup2
Any Treatment Group	inaegroup
at Least 1 Percent	aePercent

```
%astrefst001(tbltype = CORE,  
  colvar = trt01an,  
  inbdsdsn = a_in.adae,  
  insldsn = a_in.adsl,  
  popvar = saffl,  
  poptxt = %str(Analysis set: Treated),  
  ...  
  cutoff = pct1>=1 or pct2>=1,  
  bdswhere = %str(TRITEMFL='Y' and aetoxgrn>=3 and aeser = 'Y'),  
  ...  
  subj1txt = AEs,  
  lvlvars = aebodsys*aedecod,  
  lvltxt = %str(System organ class|Preferred term),  
  ...);
```

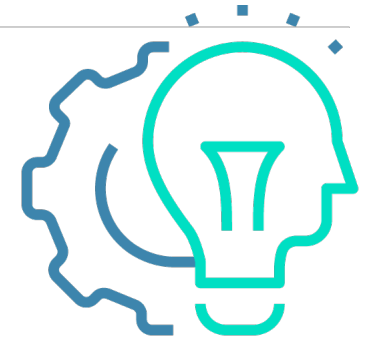
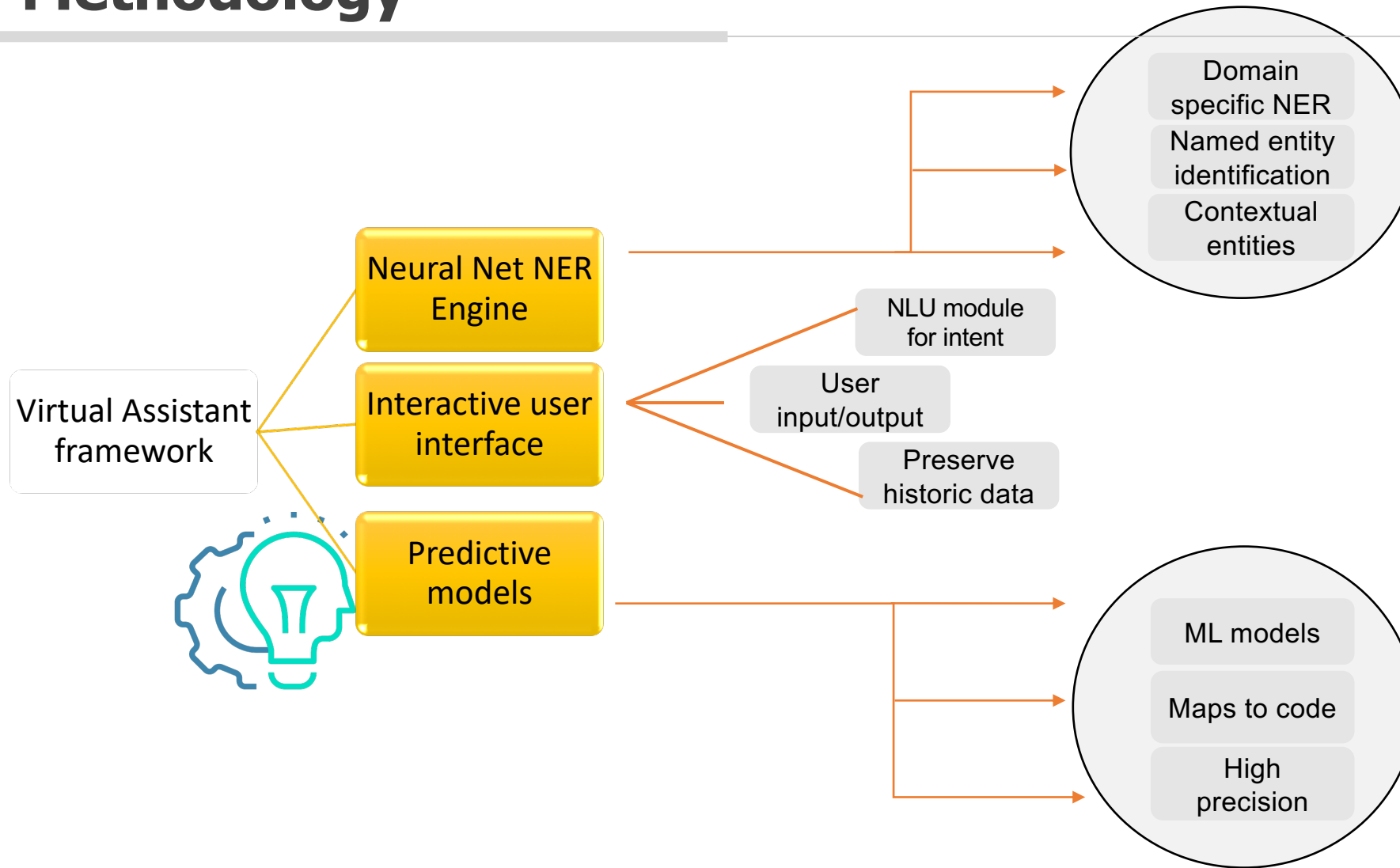
Methodology



Methodology



Methodology

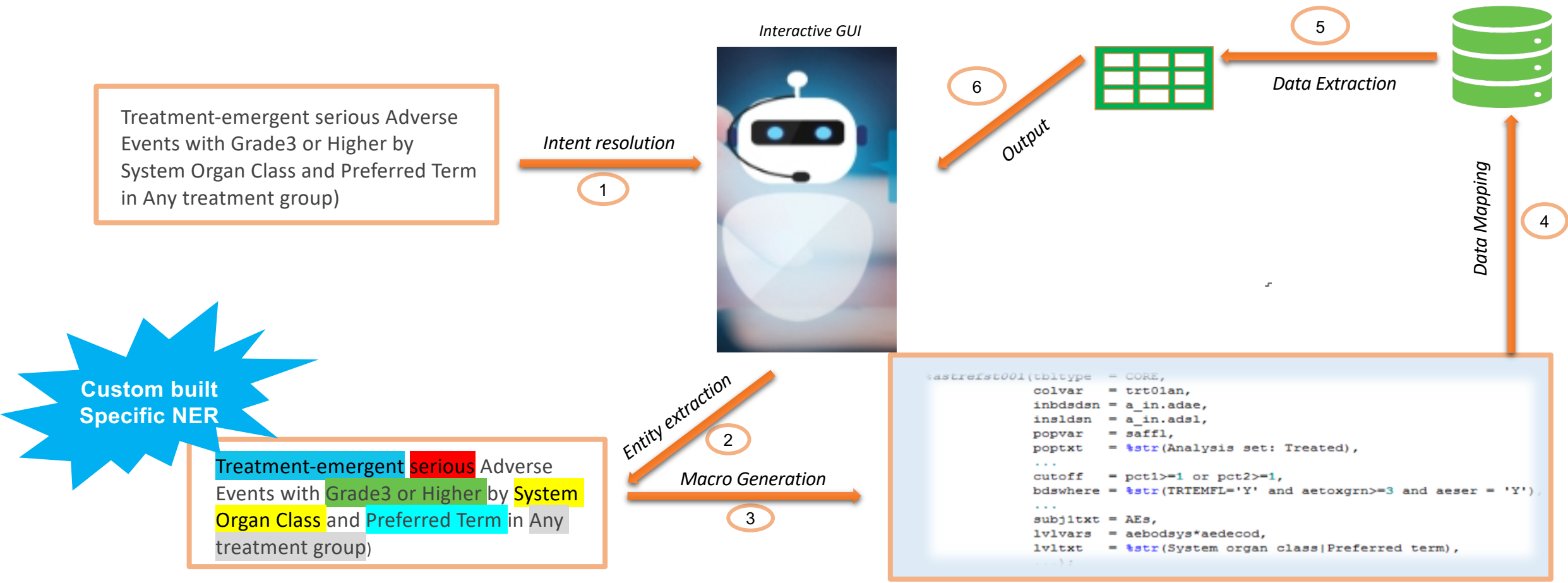


- ❖ *Named Entity Extraction*
- ❖ *Contextual entity extraction*
- ❖ *Macros/functions mapping.*
- ❖ *Active feedback.*
- ❖ *Models optimized for Precision.*

Process flow



Process flow



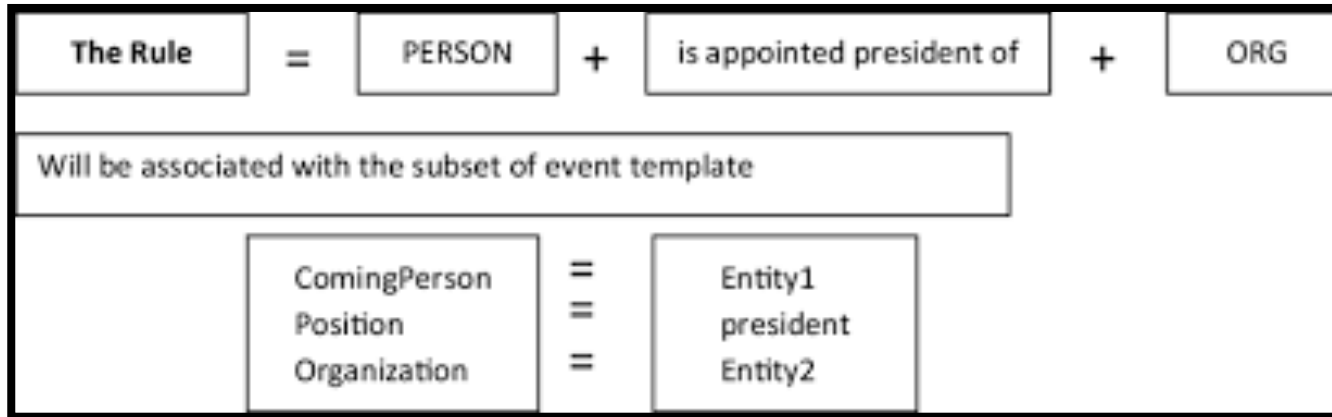
ENVISIONED DATA PROCESS

NER (Named Entity Recognition)



NER (Named Entity Recognition)

1) TEXT RULE-BASED APPROACH



Find what : `\d+\s*(percent|pct|%)`

stem Organ Class and Preferred Term (at Least 1 Percent in Any Treatment Group)
at Least 5% by System Organ Class and Preferred Term
rib With Frequency of at Least 5 pct by System Organ Class and Preferred Term

NER (Named Entity Recognition)

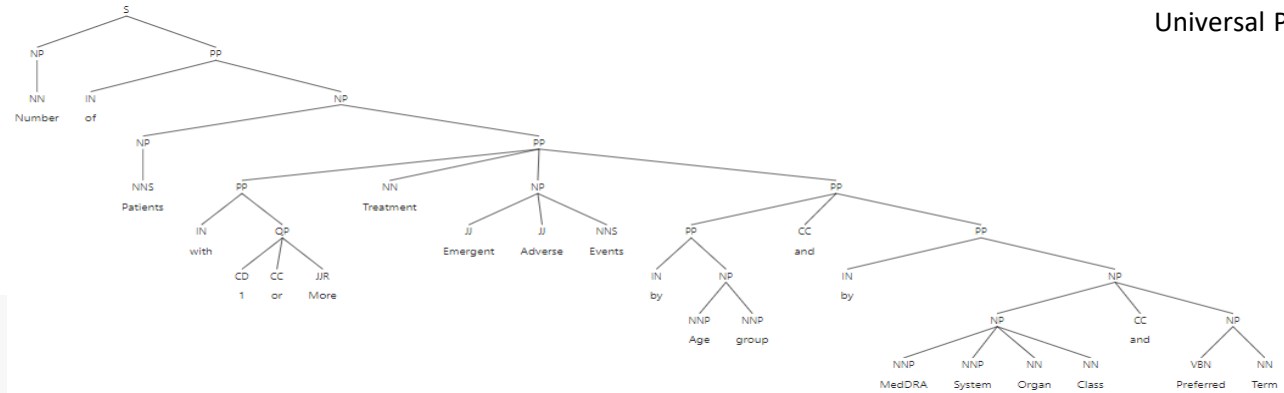
2) LINGUISTIC RULE-BASED APPROACH

Number of Patients with 1 or More Treatment-Emergent Adverse Events by Age group BYGROUP and by MedDRA System Organ Class and Preferred Term

Preferred Term BYGROUP

```
import spacy
from spacy.pipeline import EntityRuler
nlp = spacy.load("en_core_sci_md", disable = ['ner'])
nlp.add_pipe(merge_noun_chunks)
nlp.add_pipe(merge_entities)

bygroup_pattern = [
    {"LOWER": "by"},
    {"POS": "{", "OP": "*"},
    {"POS": {"NOT_IN": ['ADP']}, "OP": "*"},
    {"POS": {"NOT_IN": ['ADP', 'VERB', 'PART', 'CCONJ']}]
]
```



by Age group NOUNPHRASE

by MedDRA System Organ Class NOUNPHRASE and Preferred Term NOUNPHRASE

```
np_pattern = [{"TAG": {"IN": ['NN', 'NNP', 'NNS', 'NNPS']}, "OP": "+"}]
```

NN Noun, singular or mass
NNS Noun, plural
NNP Proper noun, singular
NNPS Proper noun, plural

NER (Named Entity Recognition)

3) MACHINE LEARNING APPROACH

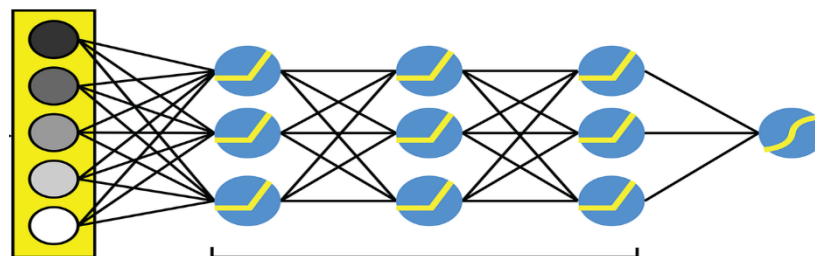


Test Data
(text corpus to be labeled)



Documents

Deep Learning-based NLP



Annotate Entities

Selected word/phrase	Entity type
Drug-related	aerel
Serious	seriousAE
Grade 3 or Higher	aeGrade
In Any Treatment Gro	inaeGroup
1 Percent	aePercent
Treatment-emergent	TEAE
System Organ Class	bygroup
Preferred Term	bygroup

Demo



Challenges



Challenges

- ❖ High-quality training data is the key factors for the success and more training data need to be collected and labeled.
- ❖ Limited to **standard macros/functions** – Open code is yet explored.
- ❖ Current exploration is limited to analysis data should meet the **one-proc-away** requirement for output generation.

Any Questions



THANK YOU



ADDITIONAL SLIDES



Data Flow and solution

1. Intent prediction → Macro Name

aesummary (0.95)

2. Entity(role) recognition → Macro parameter

Treatment-emergent Serious Adverse Events with Grade 3 or Higher by System Organ Class and Preferred Term (at Least 1 Percent in Any Treatment Group)

3. Fuzzy match to standard label(variable) → Macro parameter value

```
input by variable label  
{'bygroup': 'Organ Class', 'bygroup': 'Outcome'}
```

```
mapped variable and label  
{'bygroup': ['AEBODSYS', 'System Organ Class'], 'bygroup': ['AEOUT', 'Outcome']}
```

4. Rule-based mapping → Macro parameter value (e.g., "Serious" → AESER = "Y")

5. SAS Code generation (header, **body**, end)

6. Command file (.bat) generation to trigger xxx.sas/xxx.R and generate xxx.rtf