AWS Ground Truth
□ *is*
□ *is not*
A silver bullet

What AWS has done for you
and what you must be prepared to do

PAVEL DERBAN

Lead BI Business Analyst

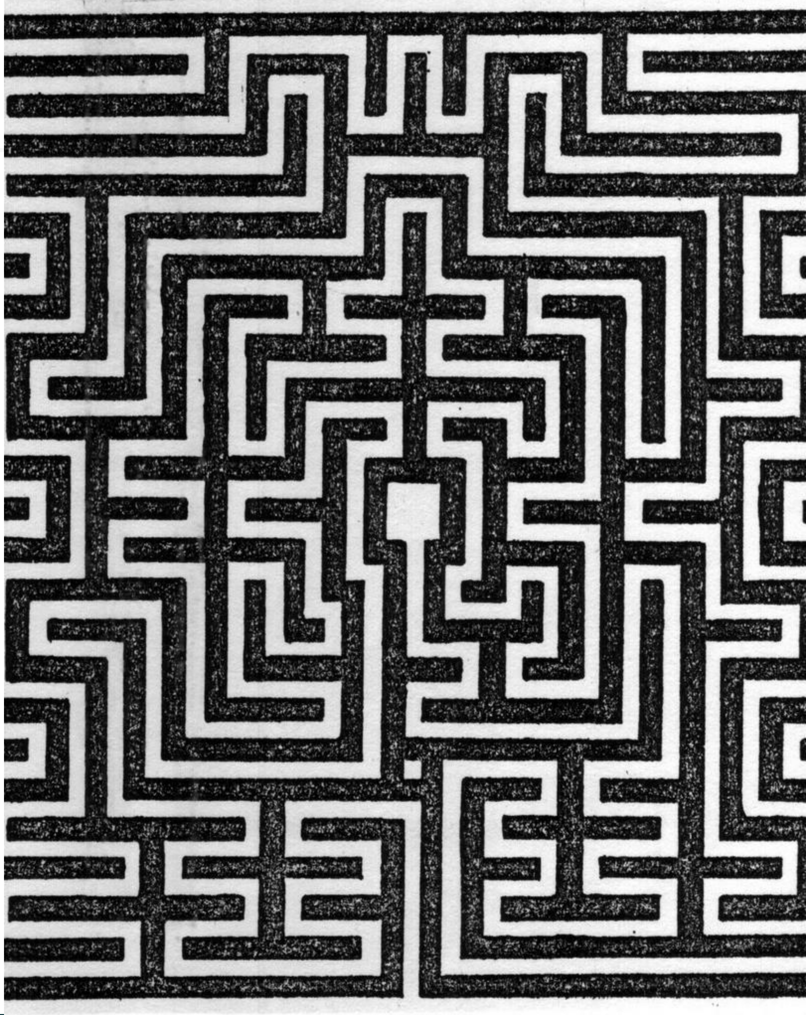
Fascinated by ML

AGENDA

1. USE CASE: COMPLEX TEXT ANNOTATION
2. ANNOTATION SOLUTION REQUIREMENTS
3. WHEN CHOOSE AWS GT: GT VS OPEN SOURCE
4. GT CUSTOMIZATION: UI, LAMBDAS, AURORA
5. PROJECT TAKEAWAYS
6. Q&A SESSION

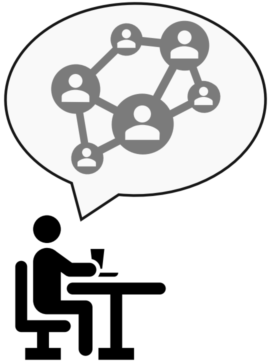
TERMINOLOGY INTRO

- **Supervised Machine Learning** is a learning from labeled training data
- **Training Data Set** is labeled data used to train machine learning algorithms
- **Named Entity (NE)** is a proper name referring to important business object (company name, person name, security name)
- **Coreference/Anaphora** is an expression referencing to the mentioned named entity (he, the suspect, the president)
- **Taxonomy** is treelike structure of classes of objects within a domain
- **Knowledge Graph** is a brain-like, structured database that stores facts with flexible, bi-directional relationships
- **Knowledge Graph triple feed (RDF triple)** is a set of three entities that codifies a statement about semantic data in the form of subject–predicate–object expressions



BUSINESS CASE

• **EPAM client’s business:**	news aggregation, information extraction and provision
• **Final product:**	access to knowledge graph containing information
• **Client’s customer:**	investment banks
• **Business Value:**	investment risk assessment
• **Enabler product:**	person-risk relation automated discovery
• **Enabler of enabler product:**	training data set



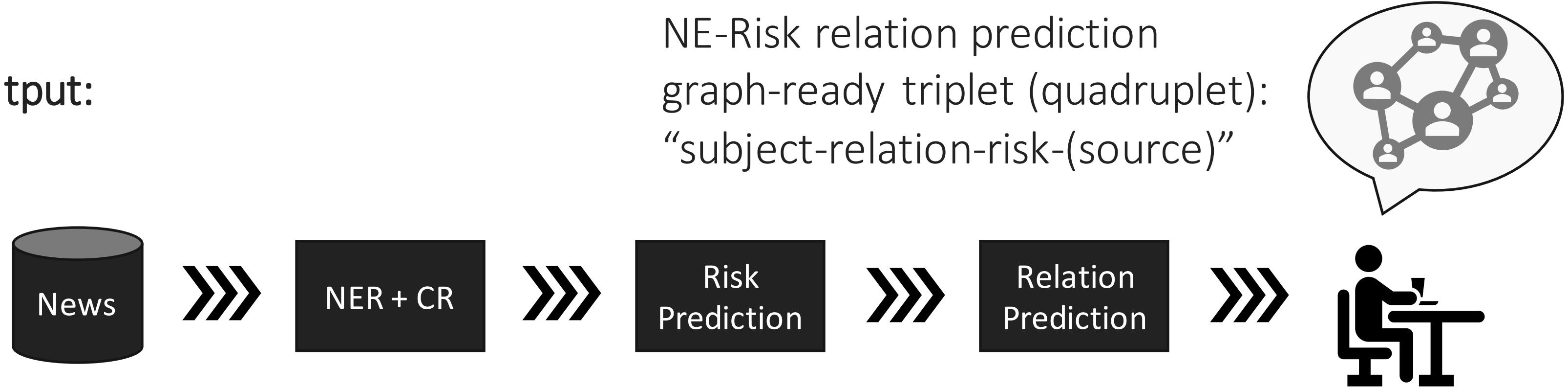
Where is ML?

Where is Ground Truth?

Where is Cloud?

ML TASK FORMULATION

- Business Value: automated news classification, human effort reduction
- Input: news feed, xml files
- Preprocessed (3rd Party): named entity recognition (NER), co-reference resolution (CR)
- Processing: taxonomy class (Risk) prediction, NE-Risk relation prediction
- Output: graph-ready triplet (quadruplet): “subject-relation-risk-(source)”



TRAINING SET TASK

- Business Value: provide human labeled “ground truth” to train model, reduce cost of training set through provision of convenient annotation tool
- Input: batch of xml files, custom schema
- NE/Coreference resolved NO
- Processing: manually annotate 3 times by different persons

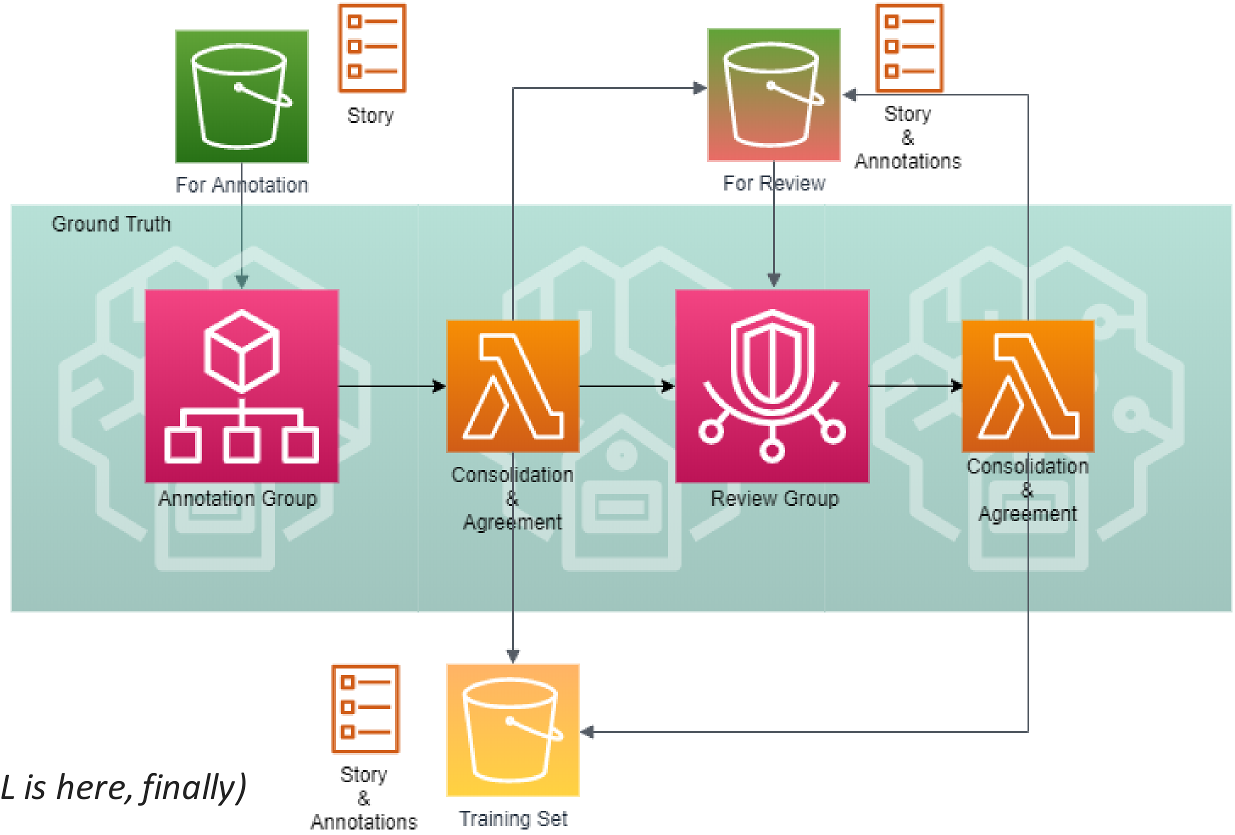
  (4 types of labels),

  automatically consolidate,

  automatically and manually assure Annotation Quality
- Output: Json file, containing consolidated annotation and source text url

Where is ML, still?

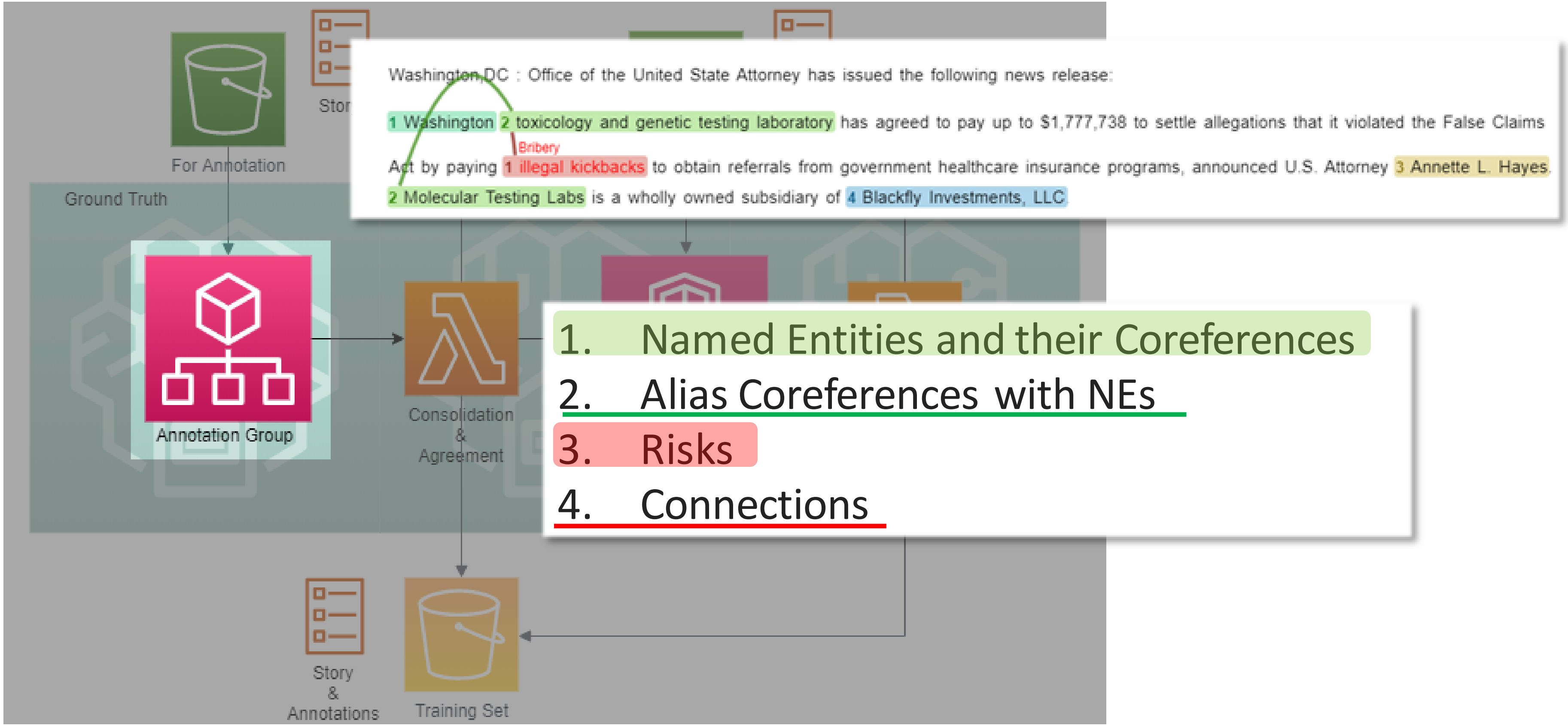
ANNOTATION WORKFLOW DIAGRAM



- Workflow is cyclic in order to resolve mis-agreement issues
- Lambda functions implement the logic and do powerlifting
- S3 buckets provide structured storage (folders)

ML is here, finally)

ANNOTATION TASK



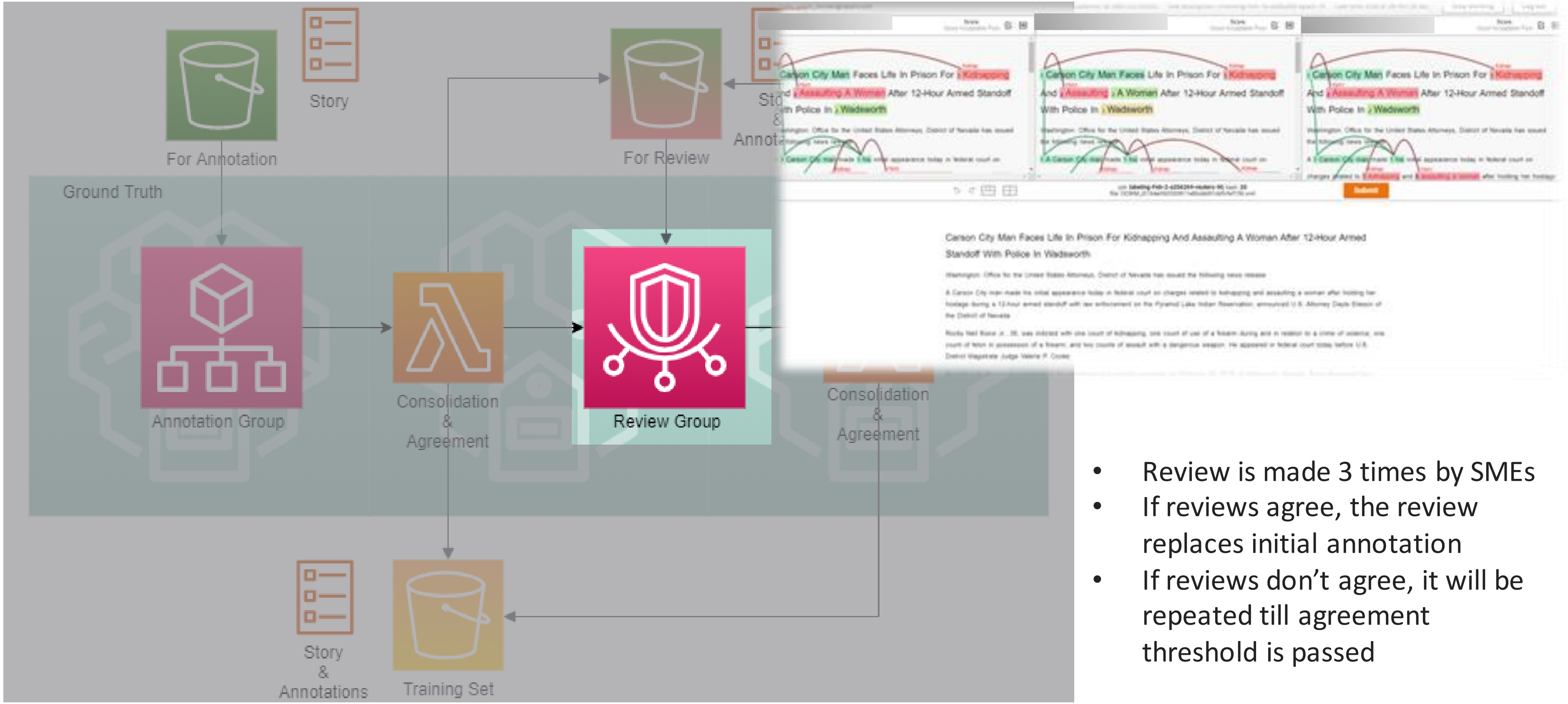
Washington DC : Office of the United State Attorney has issued the following news release:

1 Washington 2 toxicology and genetic testing laboratory has agreed to pay up to $1,777,738 to settle allegations that it violated the False Claims Act by paying 1 illegal kickbacks to obtain referrals from government healthcare insurance programs, announced U.S. Attorney 3 Annette L. Hayes.

2 Molecular Testing Labs is a wholly owned subsidiary of 4 Blackfly Investments, LLC.

1. Named Entities and their Coreferences
2. Alias Coreferences with NEs
3. Risks
4. Connections

REVIEW TASK



Carson City Man Faces Life In Prison For Kidnapping And Assaulting A Woman After 12-Hour Armed Standoff With Police In Wadsworth

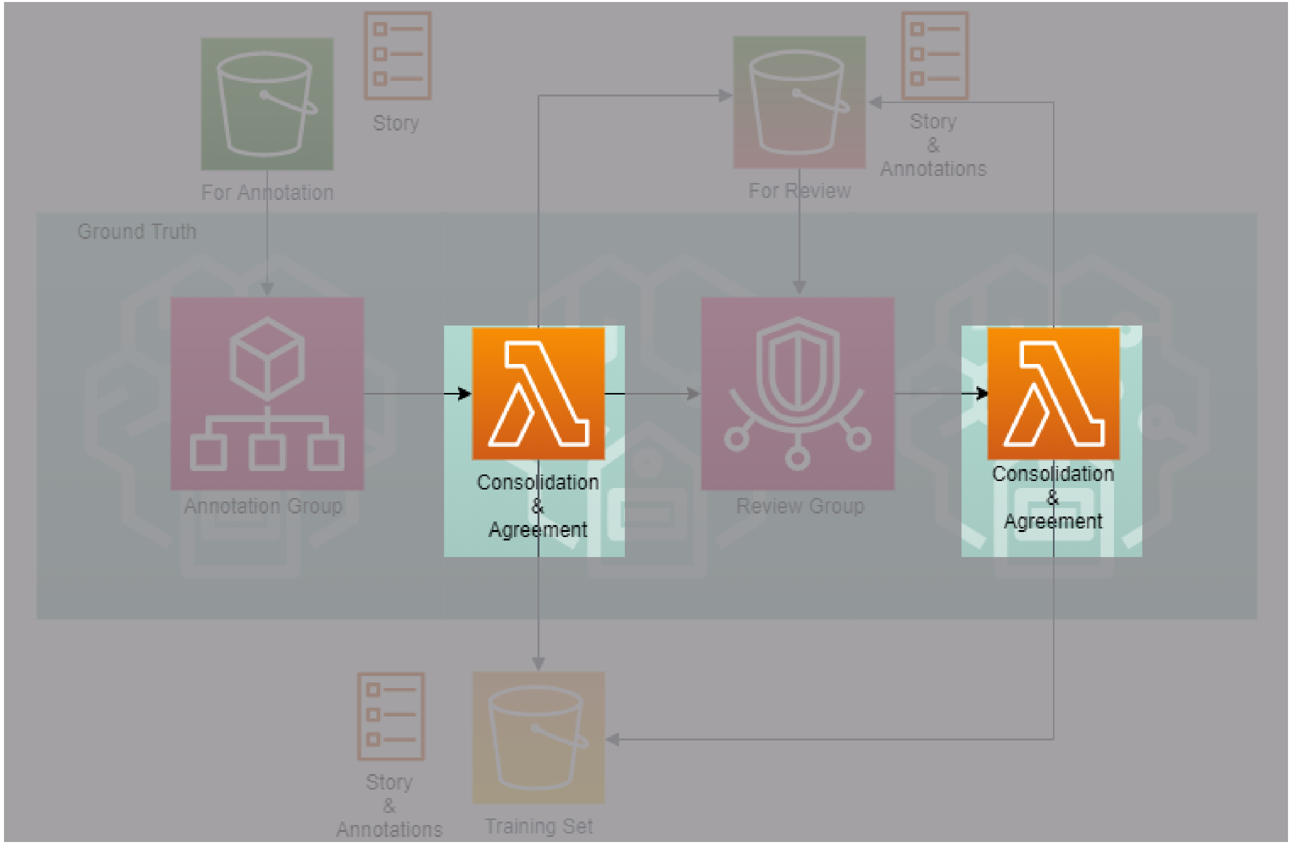
Washington: Office for the United States Attorneys, District of Nevada has issued the following news release:

A Carson City man made his initial appearance today in federal court on charges related to kidnapping and assaulting a woman after holding her hostage during a 12-hour armed standoff with law enforcement on the Pyramid Lake Indian Reservation, announced U.S. Attorney Dayle Elieson of the District of Nevada.

Rocky Neil Boice Jr., 38, was indicted with one count of kidnapping, one count of use of a firearm during and in relation to a crime of violence, one count of felon in possession of a firearm, and two counts of assault with a dangerous weapon. He appeared in federal court today before U.S. District Magistrate Judge Valerie P. Cooke.

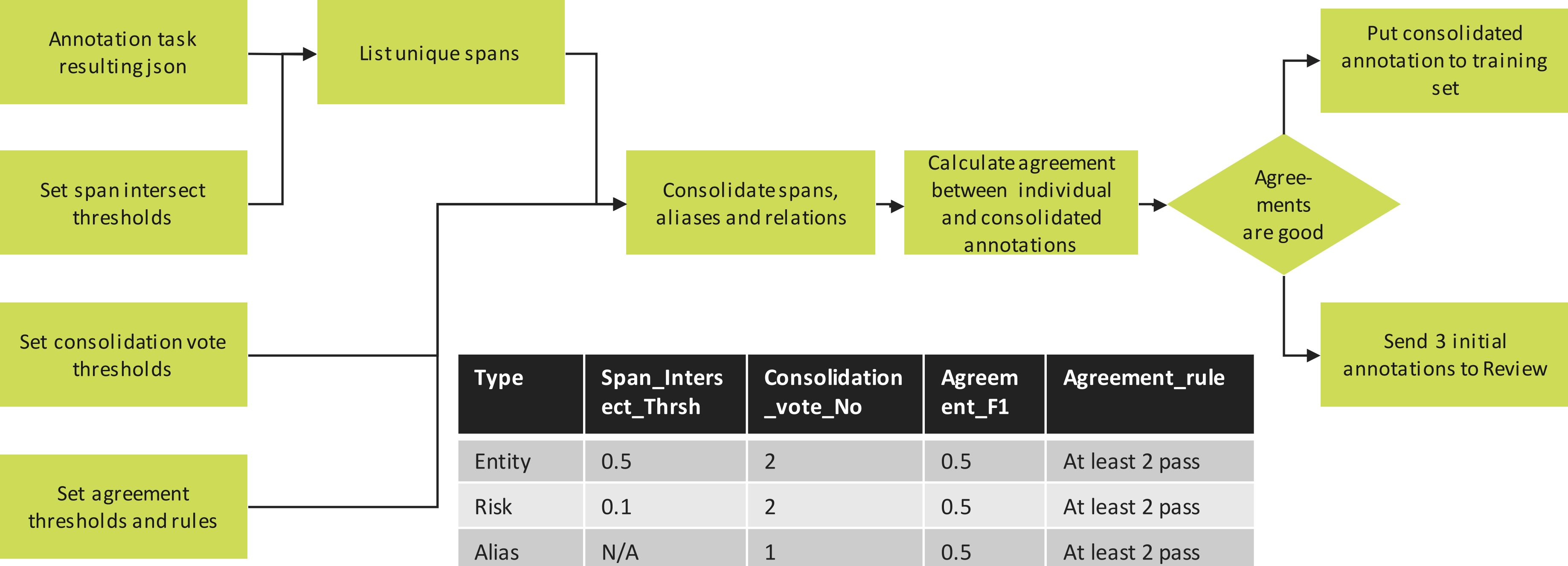
- Review is made 3 times by SMEs
- If reviews agree, the review replaces initial annotation
- If reviews don’t agree, it will be repeated till agreement threshold is passed

ANNOTATION CONSOLIDATION AND AGREEMENT



- Lambda functions perform the annotation consolidation logic
- They are the gauges of control and adjustment
- End-to-end algorithm performance is the KPI to use when tuning consolidation

CONSOLIDATION AND AGREEMENT WORKFLOW



Annotation task resulting json

Set span intersect thresholds

List unique spans

Set consolidation vote thresholds

Set agreement thresholds and rules

Consolidate spans, aliases and relations

Calculate agreement between individual and consolidated annotations

Agree- ments are good

Put consolidated annotation to training set

Send 3 initial annotations to Review

Type	Span_Intersect_Thrsh	Consolidation_vote_No	Agreement_F1	Agreement_rule
Entity	0.5	2	0.5	At least 2 pass
Risk	0.1	2	0.5	At least 2 pass
Alias	N/A	1	0.5	At least 2 pass
Relation	N/A	2	0.5	At least 2 pass

ANNOTATION ADMINISTRATION AND ANALYTICS

ADMINISTRATION TASKS

- Create users and user groups
- Allocate jobs to specific groups
- Stop and resume annotation tasks
- Find annotation by its attributes
- Show/compare annotations
- Trace consolidated/reviewed annotations to initial ones
- Log annotation sessions (security and time tracking)

ANALYTICS TASKS

- Annotation velocity: stories, lines, spans per period
- Annotation quality: agreement scores
- Risk topic coverage: snapshot & dynamics
- Annotator’s performance: quantity & quality
- Reviewer’s performance: quantity & quality
- Training set funnel rates: selected-annotated-consolidated-reviewed-delivered

THIS IS ALL ABOUT USE CASE AND SOLUTION REQUIRENMENTS

QUESTIONS SO FAR?

WHEN CHOOSE AWS GT: OPEN-SOURCE VS GT

BRAT

- UI ready for complex text annotation
- Annotation browser + Stop/resume available
- Pairwise comparison of annotated texts
- Security: login, password, IP-address
- Manual user management
- Manual workflow administration
- No annotation analytics

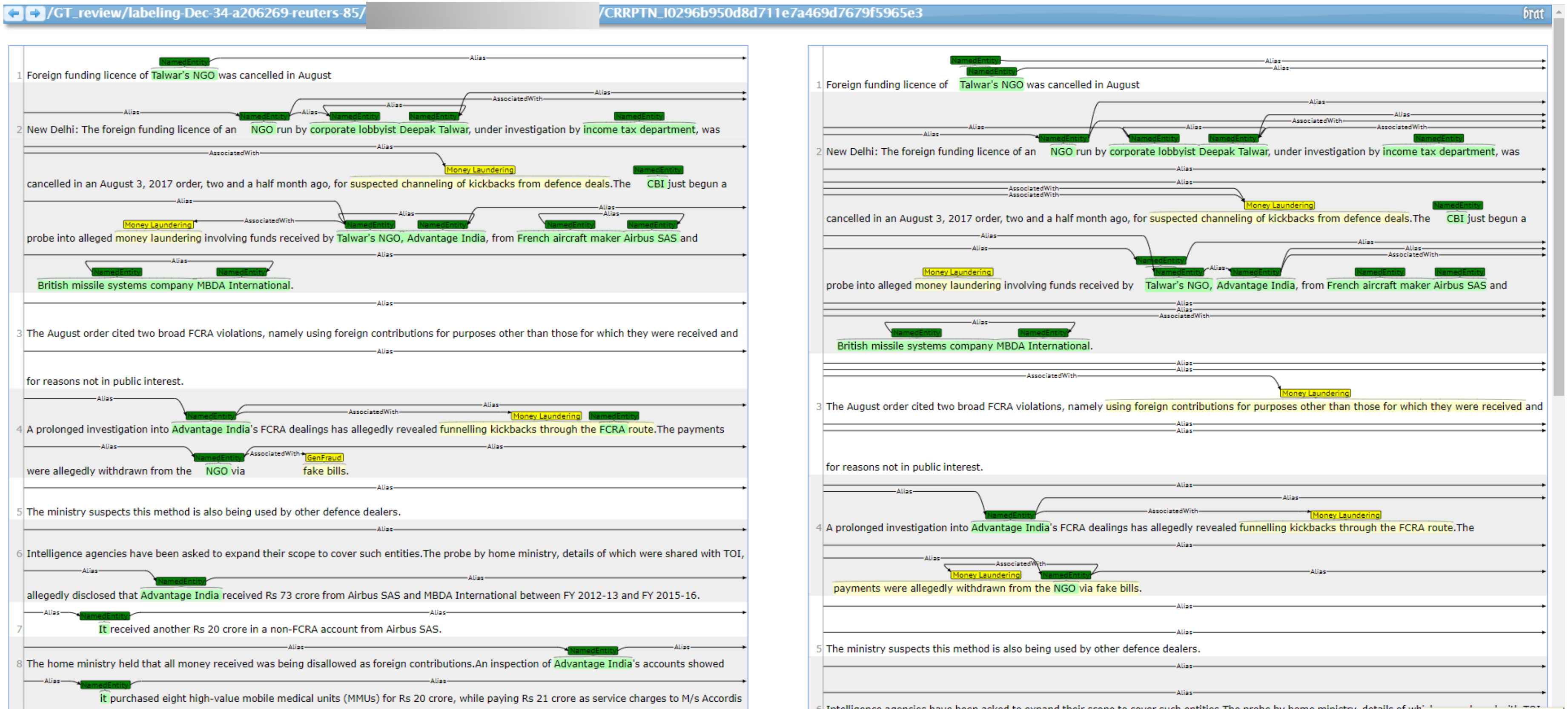
GOOD FIT FOR SMALL-SCALE LEAN TEXT ANNOTATION

GT

- Automated user management (AWS Cognito)
- Automated workflow (AWS Lambdas, Step functions)
- Automated annotation task fulfillment(count) control
- Security: AWS managed
- UI ready for simple annotation
- No annotation browser, no stop/resume available
- No review/comparison functionality
- No annotation analytics
- Ready for crowed-sourced work force
- Highly scalable

GOOD FIT FOR LARGE SCALE VISUAL ANNOTATION FOR ANN TRAINING

BRAT UI EXAMPLE: KIND OF OLD-FASHIONED



/GT_review/labeling-Dec-34-a206269-reuters-85/[redacted]/CRRPTN_I0296b950d8d711e7a469d7679f5965e3

1 Foreign funding licence of Talwar's NGO was cancelled in August

2 New Delhi: The foreign funding licence of an NGO run by corporate lobbyist Deepak Talwar, under investigation by income tax department, was cancelled in an August 3, 2017 order, two and a half month ago, for suspected channeling of kickbacks from defence deals.The CBI just begun a probe into alleged money laundering involving funds received by Talwar's NGO, Advantage India, from French aircraft maker Airbus SAS and British missile systems company MBDA International.

3 The August order cited two broad FCRA violations, namely using foreign contributions for purposes other than those for which they were received and for reasons not in public interest.

4 A prolonged investigation into Advantage India's FCRA dealings has allegedly revealed funnelling kickbacks through the FCRA route.The payments were allegedly withdrawn from the NGO via fake bills.

5 The ministry suspects this method is also being used by other defence dealers.

6 Intelligence agencies have been asked to expand their scope to cover such entities.The probe by home ministry, details of which were shared with TOI, allegedly disclosed that Advantage India received Rs 73 crore from Airbus SAS and MBDA International between FY 2012-13 and FY 2015-16.

7 It received another Rs 20 crore in a non-FCRA account from Airbus SAS.

8 The home ministry held that all money received was being disallowed as foreign contributions.An inspection of Advantage India's accounts showed it purchased eight high-value mobile medical units (MMUs) for Rs 20 crore, while paying Rs 21 crore as service charges to M/s Accordis

1 Foreign funding licence of Talwar's NGO was cancelled in August

2 New Delhi: The foreign funding licence of an NGO run by corporate lobbyist Deepak Talwar, under investigation by income tax department, was cancelled in an August 3, 2017 order, two and a half month ago, for suspected channeling of kickbacks from defence deals.The CBI just begun a probe into alleged money laundering involving funds received by Talwar's NGO, Advantage India, from French aircraft maker Airbus SAS and British missile systems company MBDA International.

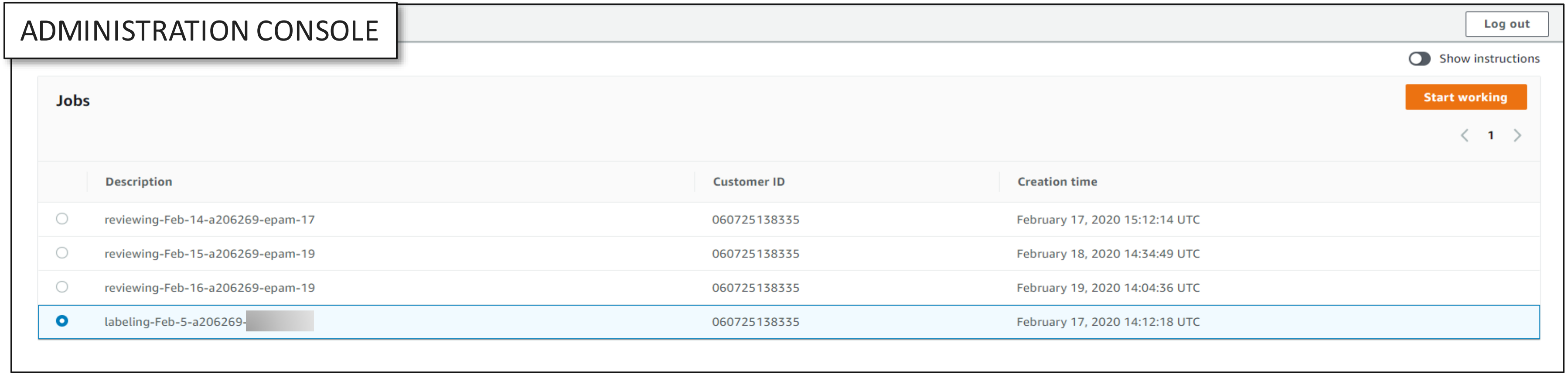
3 The August order cited two broad FCRA violations, namely using foreign contributions for purposes other than those for which they were received and for reasons not in public interest.

4 A prolonged investigation into Advantage India's FCRA dealings has allegedly revealed funnelling kickbacks through the FCRA route.The payments were allegedly withdrawn from the NGO via fake bills.

5 The ministry suspects this method is also being used by other defence dealers.

6 Intelligence agencies have been asked to expand their scope to cover such entities.The probe by home ministry, details of which were shared with TOI

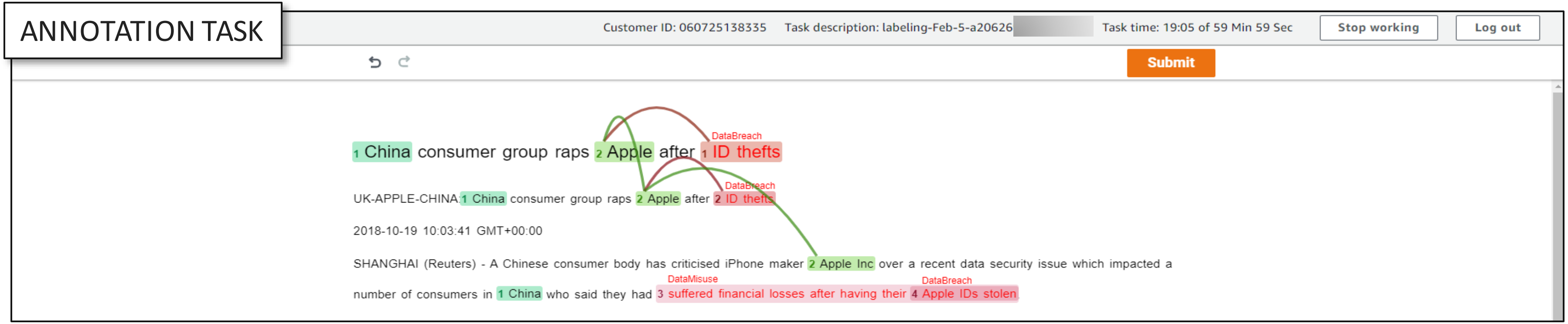
GT UI EXAMPLES: CLEAN AND MINIMALISTIC



ADMINISTRATION CONSOLE

Jobs

	Description	Customer ID	Creation time
○	reviewing-Feb-14-a206269-epam-17	060725138335	February 17, 2020 15:12:14 UTC
○	reviewing-Feb-15-a206269-epam-19	060725138335	February 18, 2020 14:34:49 UTC
○	reviewing-Feb-16-a206269-epam-19	060725138335	February 19, 2020 14:04:36 UTC
●	labeling-Feb-5-a206269-[redacted]	060725138335	February 17, 2020 14:12:18 UTC



ANNOTATION TASK

Customer ID: 060725138335 Task description: labeling-Feb-5-a20626[redacted] Task time: 19:05 of 59 Min 59 Sec

DataBreach

1 China consumer group raps 2 Apple after 1 ID thefts

DataBreach

UK-APPLE-CHINA:1 China consumer group raps 2 Apple after 2 ID thefts

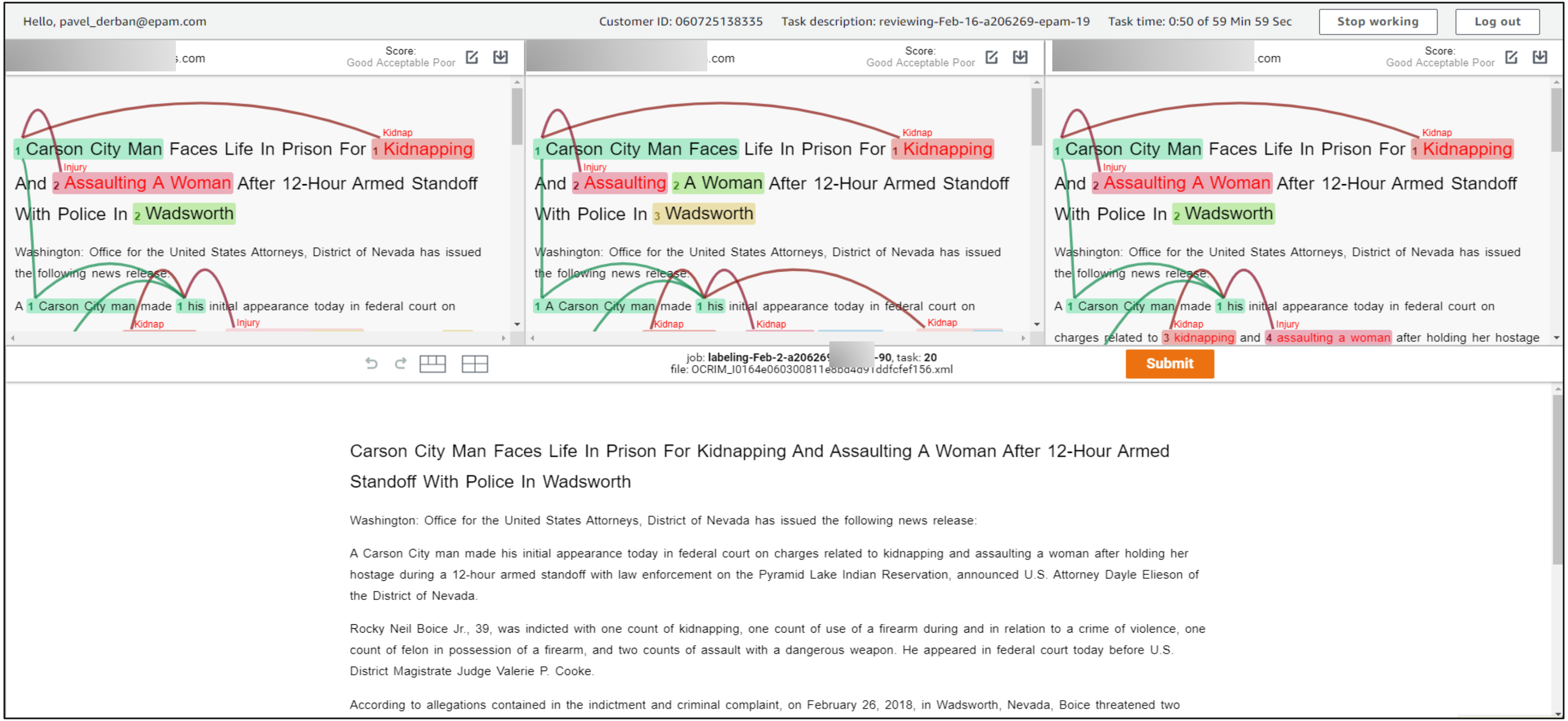
2018-10-19 10:03:41 GMT+00:00

SHANGHAI (Reuters) - A Chinese consumer body has criticised iPhone maker 2 Apple Inc over a recent data security issue which impacted a

DataMisuse DataBreach

number of consumers in 1 China who said they had 3 suffered financial losses after having their 4 Apple IDs stolen

REVIEW TASK



Hello, pavel_derban@epam.com

Customer ID: 060725138335 Task description: reviewing-Feb-16-a206269-epam-19 Task time: 0:50 of 59 Min 59 Sec

[redacted].com Score: Good Acceptable Poor

1 Carson City Man Faces Life In Prison For 1 Kidnapping (Kidnap) And 2 Assaulting A Woman (Injury) After 12-Hour Armed Standoff With Police In 2 Wadsworth

Washington: Office for the United States Attorneys, District of Nevada has issued the following news release:

A 1 Carson City man made 1 his initial appearance today in federal court on

[redacted].com Score: Good Acceptable Poor

1 Carson City Man Faces Life In Prison For 1 Kidnapping (Kidnap) And 2 Assaulting (Injury) 2 A Woman After 12-Hour Armed Standoff With Police In 3 Wadsworth

Washington: Office for the United States Attorneys, District of Nevada has issued the following news release:

1 A Carson City man made 1 his initial appearance today in federal court on

[redacted].com Score: Good Acceptable Poor

1 Carson City Man Faces Life In Prison For 1 Kidnapping (Kidnap) And 2 Assaulting A Woman (Injury) After 12-Hour Armed Standoff With Police In 2 Wadsworth

Washington: Office for the United States Attorneys, District of Nevada has issued the following news release:

A 1 Carson City man made 1 his initial appearance today in federal court on charges related to 3 kidnapping and 4 assaulting a woman after holding her hostage

job: labeling-Feb-2-a20626[redacted]-90, task: 20
file: OCRIM_I0164e060300811e8bd4d91ddfcfef156.xml

Carson City Man Faces Life In Prison For Kidnapping And Assaulting A Woman After 12-Hour Armed Standoff With Police In Wadsworth

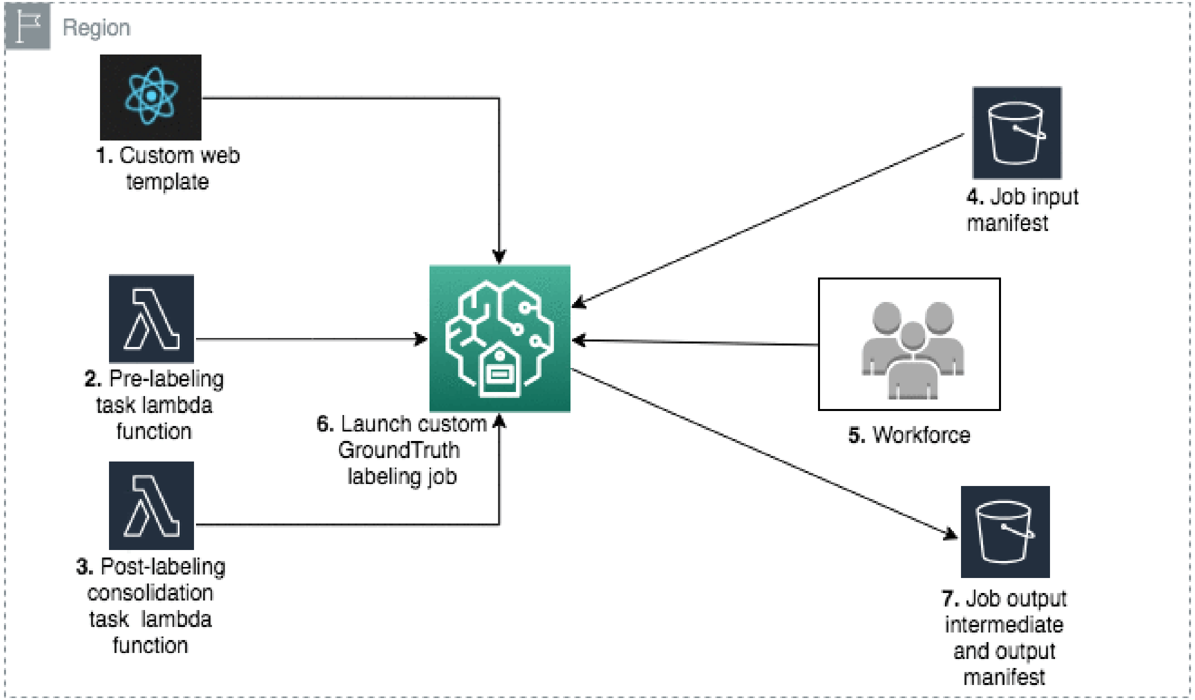
Washington: Office for the United States Attorneys, District of Nevada has issued the following news release:

A Carson City man made his initial appearance today in federal court on charges related to kidnapping and assaulting a woman after holding her hostage during a 12-hour armed standoff with law enforcement on the Pyramid Lake Indian Reservation, announced U.S. Attorney Dayle Elieson of the District of Nevada.

Rocky Neil Boice Jr., 39, was indicted with one count of kidnapping, one count of use of a firearm during and in relation to a crime of violence, one count of felon in possession of a firearm, and two counts of assault with a dangerous weapon. He appeared in federal court today before U.S. District Magistrate Judge Valerie P. Cooke.

According to allegations contained in the indictment and criminal complaint, on February 26, 2018, in Wadsworth, Nevada, Boice threatened two

GT CUSTOMIZATION STANDARD OFFERING


Region
1. Custom web template
2. Pre-labeling task lambda function
3. Post-labeling consolidation task lambda function
6. Launch custom GroundTruth labeling job
4. Job input manifest
5. Workforce
7. Job output intermediate and output manifest

- Custom UI template(s)
- Pre-lambda: Task creation
- Post-lambda: Task consolidation

PROJECT TEAM

• Senior frontend engineer	1 FTE
• Senior .NET/python engineer	1.5 FTE
• Senior cloud devops engineer	1 FTE
• Senior business analyst	1 FTE
• Data scientist	0.5 FTE

Project duration – Six month

UI CUSTOMIZATION

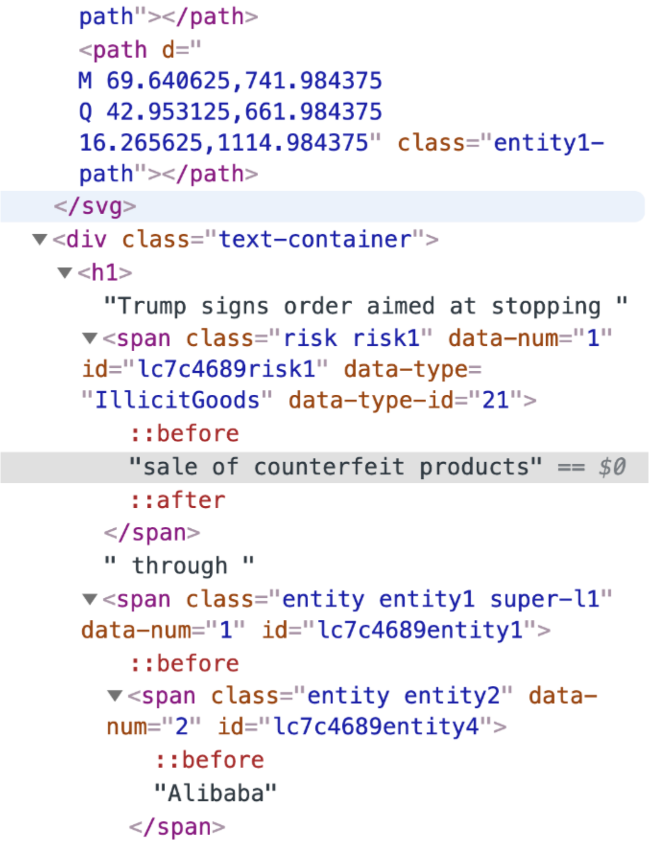
REQUIREMENTS:

- No external libs/dependencies
- Unicode-compatible
- Works in latest Chrome and Firefox

TECHNOLOGY:

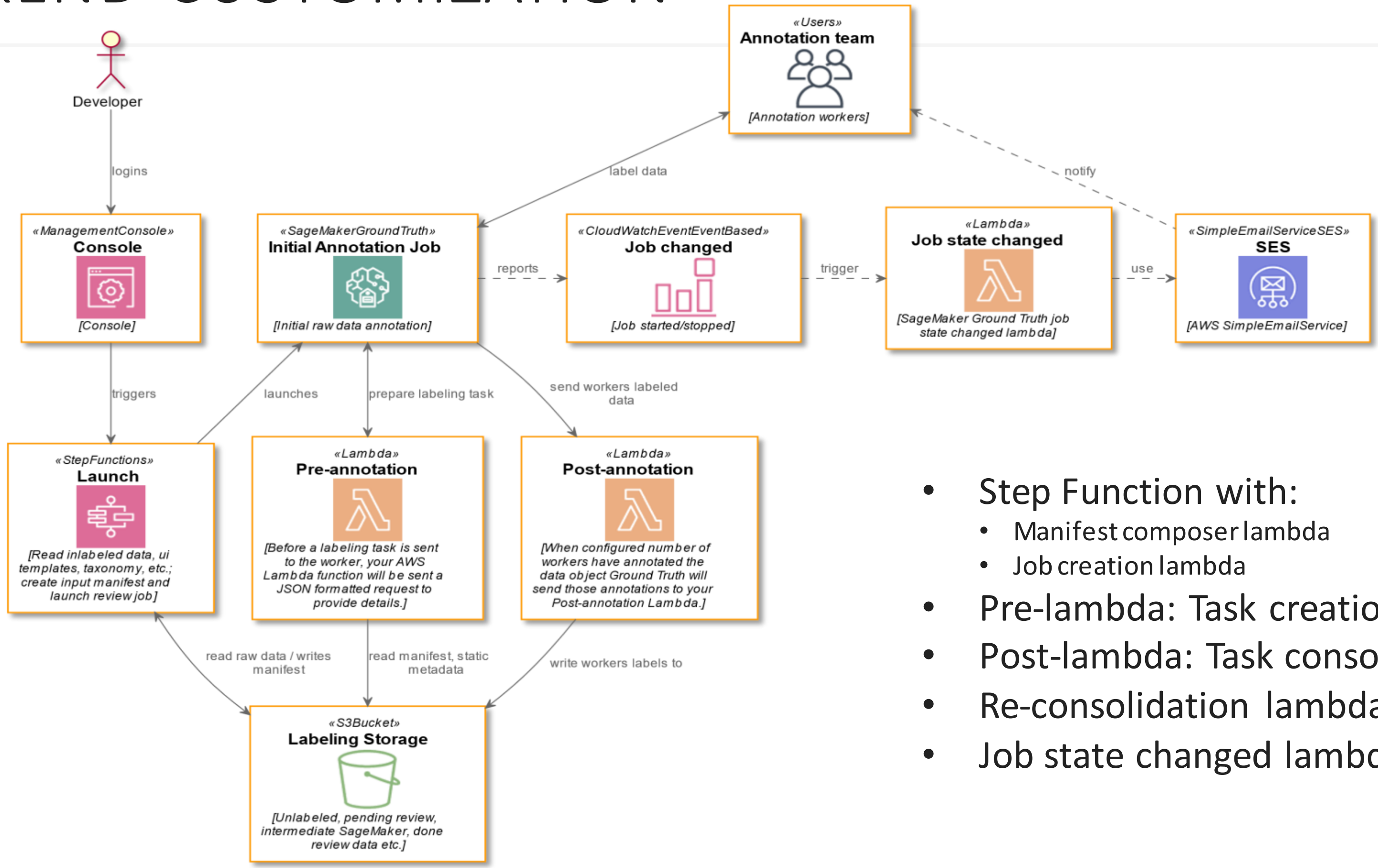
- HTML
- CSS Grid Layout
- SVG
- JS (ES6)
- HTML DOM selection API

CHALLENGE	SOLUTION
JS string functions, DOM selection API and Python backend count characters differently	Custom character counting function based on ES6 features; Unicode friendly
Visualize relations between spans as arches in html	Dynamically generated in-line SVG
DOM selection API is quite stubborn one	Patience)



```
  path"></path>
  <path d="
  M 69.640625,741.984375
  Q 42.953125,661.984375
  16.265625,1114.984375" class="entity1-
  path"></path>
</svg>
▼<div class="text-container">
  ▼<h1>
     "Trump signs order aimed at stopping "
    ▼<span class="risk risk1" data-num="1"
    id="lc7c4689risk1" data-type=
    "IllicitGoods" data-type-id="21">
       ::before
       "sale of counterfeit products" == $0
       ::after
     </span>
     " through "
    ▼<span class="entity entity1 super-l1"
    data-num="1" id="lc7c4689entity1">
       ::before
      ▼<span class="entity entity2" data-
      num="2" id="lc7c4689entity4">
         ::before
         "Alibaba"
       </span>
```

BACKEND CUSTOMIZATION



Developer

logins

«ManagementConsole»
Console
[Console]

triggers

«StepFunctions»
Launch
[Read inlabeled data, ui templates, taxonomy, etc.; create input manifest and launch review job]

launches

«SageMakerGroundTruth»
Initial Annotation Job
[Initial raw data annotation]

prepare labeling task

«Lambda»
Pre-annotation
[Before a labeling task is sent to the worker, your AWS Lambda function will be sent a JSON formatted request to provide details.]

send workers labeled data

«Lambda»
Post-annotation
[When configured number of workers have annotated the data object Ground Truth will send those annotations to your Post-annotation Lambda.]

read raw data / writes manifest

read manifest, static metadata

write workers labels to

«S3Bucket»
Labeling Storage
[Unlabeled, pending review, intermediate SageMaker, done review data etc.]

label data

«Users»
Annotation team
[Annotation workers]

reports

«CloudWatchEventEventBased»
Job changed
[Job started/stopped]

trigger

«Lambda»
Job state changed
[SageMaker Ground Truth job state changed lambda]

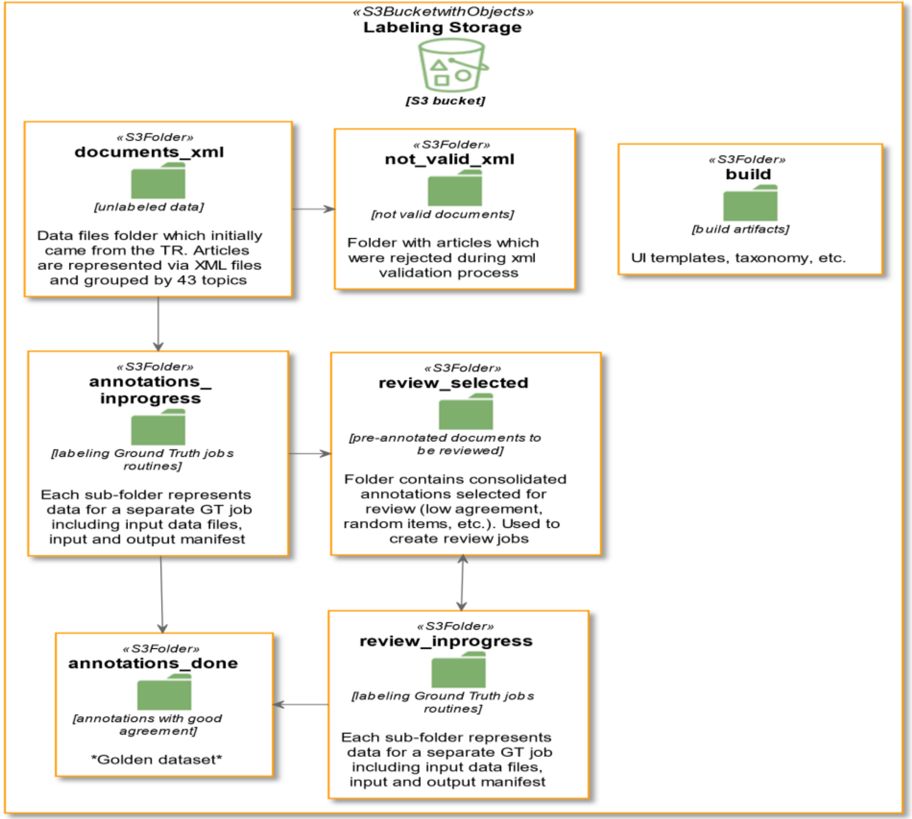
use

«SimpleEmailServiceSES»
SES
[AWS SimpleEmailService]

notify

- Step Function with:
  - Manifest composer lambda
  - Job creation lambda
- Pre-lambda: Task creation
- Post-lambda: Task consolidation
- Re-consolidation lambda
- Job state changed lambda

S3 OPERATIONAL STORAGE



«S3BucketwithObjects»
Labeling Storage
[S3 bucket]

«S3Folder»
documents_xml
[unlabeled data]
Data files folder which initially came from the TR. Articles are represented via XML files and grouped by 43 topics

«S3Folder»
not_valid_xml
[not valid documents]
Folder with articles which were rejected during xml validation process

«S3Folder»
build
[build artifacts]
UI templates, taxonomy, etc.

«S3Folder»
annotations_ inprogress
[labeling Ground Truth jobs routines]
Each sub-folder represents data for a separate GT job including input data files, input and output manifest

«S3Folder»
review_selected
[pre-annotated documents to be reviewed]
Folder contains consolidated annotations selected for review (low agreement, random items, etc.). Used to create review jobs

«S3Folder»
annotations_done
[annotations with good agreement]
Golden dataset

«S3Folder»
review_inprogress
[labeling Ground Truth jobs routines]
Each sub-folder represents data for a separate GT job including input data files, input and output manifest

STORY FILES ARE MOVING FROM FOLDER TO FOLDER AS THEY PROGRES ALONG PIPELINE TO “GOLDEN DATA SET”

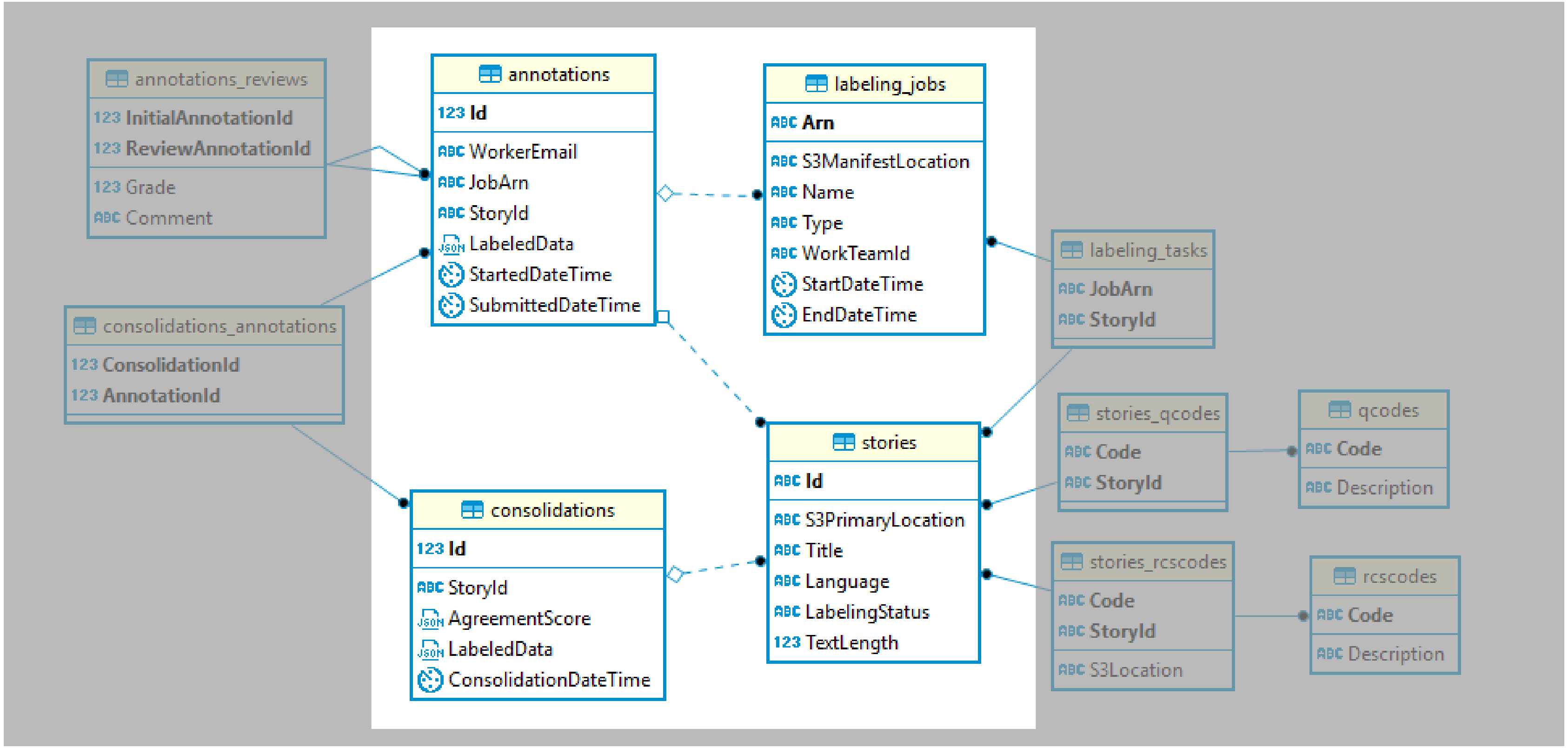
Folders:

- Documents_xml
- Annotations_inprogress
- Review_selected
- Review_inprogress
- Annotations_done

ANALYTICAL DWH: DYNAMO vs ATHENA vs AURORA

Analytics fitness	☹	☺	☺
Implementation complexity	☺	☹	☺
Cost of ownership	☺	☹	☺

AURORA SERVERLESS: DWH ERD (simplified)



annotations_reviews
123 InitialAnnotationId
123 ReviewAnnotationId
123 Grade
ABC Comment

annotations
123 Id
ABC WorkerEmail
ABC JobArn
ABC StoryId
JSON LabeledData
StartedDateTime
SubmittedDateTime

labeling_jobs
ABC Arn
ABC S3ManifestLocation
ABC Name
ABC Type
ABC WorkTeamId
StartDateTime
EndDateTime

labeling_tasks
ABC JobArn
ABC StoryId

consolidations_annotations
123 ConsolidationId
123 AnnotationId

stories
ABC Id
ABC S3PrimaryLocation
ABC Title
ABC Language
ABC LabelingStatus
123 TextLength

stories_qcodes
ABC Code
ABC StoryId

qcodes
ABC Code
ABC Description

consolidations
123 Id
ABC StoryId
JSON AgreementScore
JSON LabeledData
ConsolidationDateTime

stories_rcscodes
ABC Code
ABC StoryId
ABC S3Location

rcscodes
ABC Code
ABC Description

ANOTATION KNOWLEDGE MANAGEMENT

- Annotation rules
- Annotation FAQ
- Crime taxonomy 2.3
- Corporate crime : Guidance on how tagging should be applied for this part of the taxonomy
- Non-corporate crime: Guidance on how tagging should be applied to this part of the taxonomy
- Cases for annotation quality improvement

CORRUPTION VS BRIBERY
DRUG TRAFFICKING VS DRUG ABUSE
SEXUAL EXPLOITATION VS SEX OFFENCE
THEFT VS EMBEZZLEMENT VS ROBBERY VS BURGLARY
GENERAL FRAUD VS SECURITIES FRAUD

"DUMMY" NAMED ENTITIES
SANCTIONED ENTITIES
TITLES & OCCUPATIONS
SUPER- AND SUB-ENTITIES
GENERIC CRIMES

THIS IS ALL ABOUT GT RESHAPING FOR TEXT ANNOTATION

QUESTIONS?

ANOTATION OUTCOME STATS

Annotation team size:	5 persons
Annotation period:	6 month
Stories annotated:	1200
Risk coverage min-avg-max :	1-50-420
Stories reviewed:	~ 50
Median NE agreement:	0.92
Median alias agreement:	0.85
Median risk agreement:	0.98
Median relation agreement:	0.86

LESSONS LEARNED

- Comprehensive annotation tool is complex and expensive
- Ground truth is difficult to adopt to complex text annotation
- **Annotation tool is only a tool**
- Good taxonomy is very important
- Annotation rules body of knowledge is a must
- Expect lot of investments into training, review and re-training

FINAL JUDGEMENT

- Tool is used by several teams within the customer organization
- Nice corporate toy to brag about
- **Proof of investment is difficult, maybe, better spend this money for more qualified and motivated annotators?!**

Whoops! Demolition company accidentally tears down wrong house.

Burglary OR Theft?? 😊