The Ground Truth Grind

Strategies for evaluating the quality of labeled data

Teresa O'Neill
Solutions Architect, Natural Language Services

iMerit
human-empowered computing

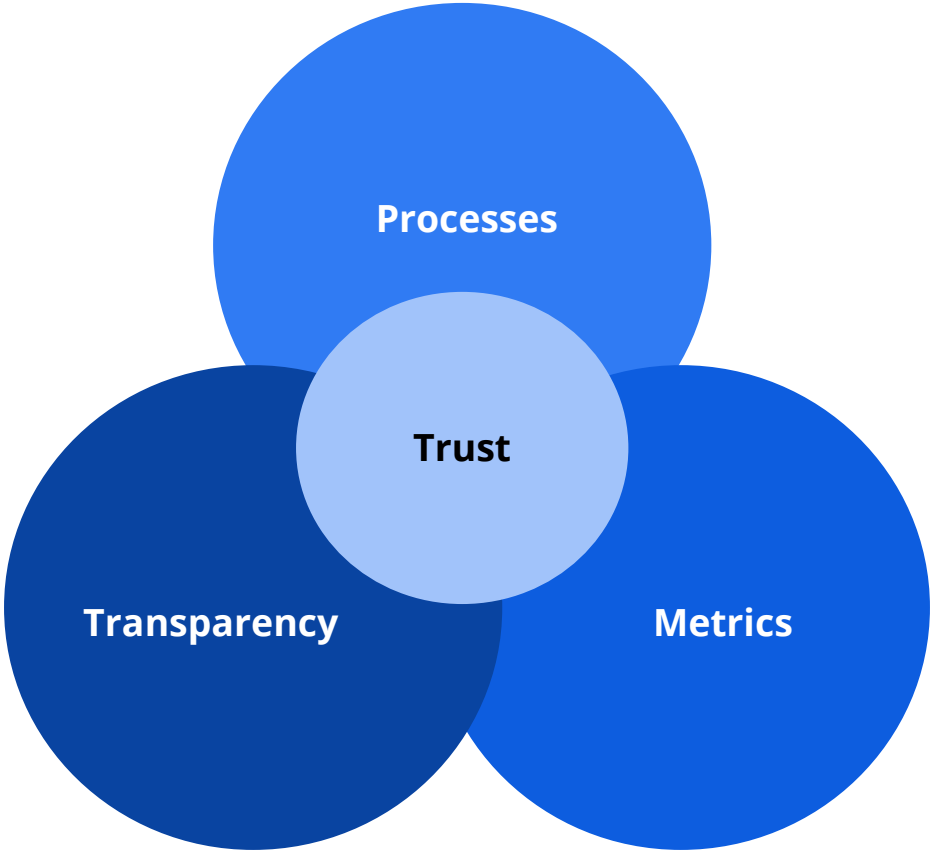
https://xkcd.com/1838/

iMerit

- What else can we do besides stir the pot?
- Pour better data in!
  - Training data for supervised learning
  - Evaluation / ground truth datasets
- How can you trust your labeled data?

Building Trust in Data Quality

iMerit



We Are iMerit

iMerit

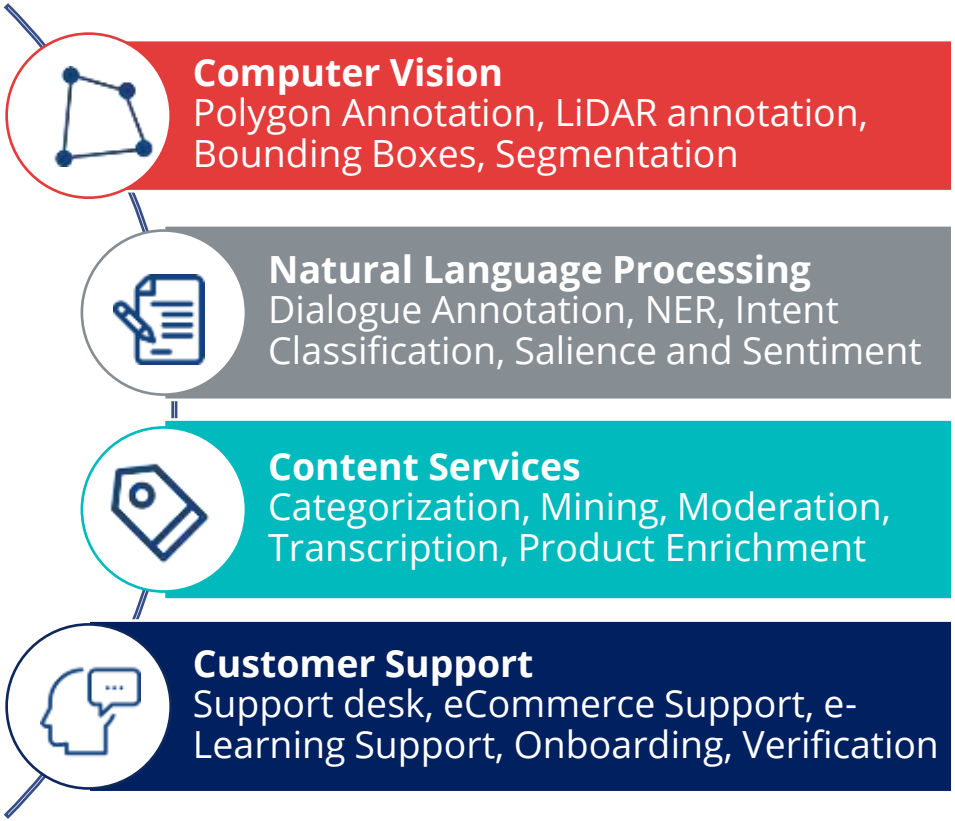
iMerit leverages human intelligence to label and enrich data.

We power algorithms in machine learning and computer vision.

We effect **positive social and economic change.** We tap into a talent pool which was under-resourced and digitally excluded.

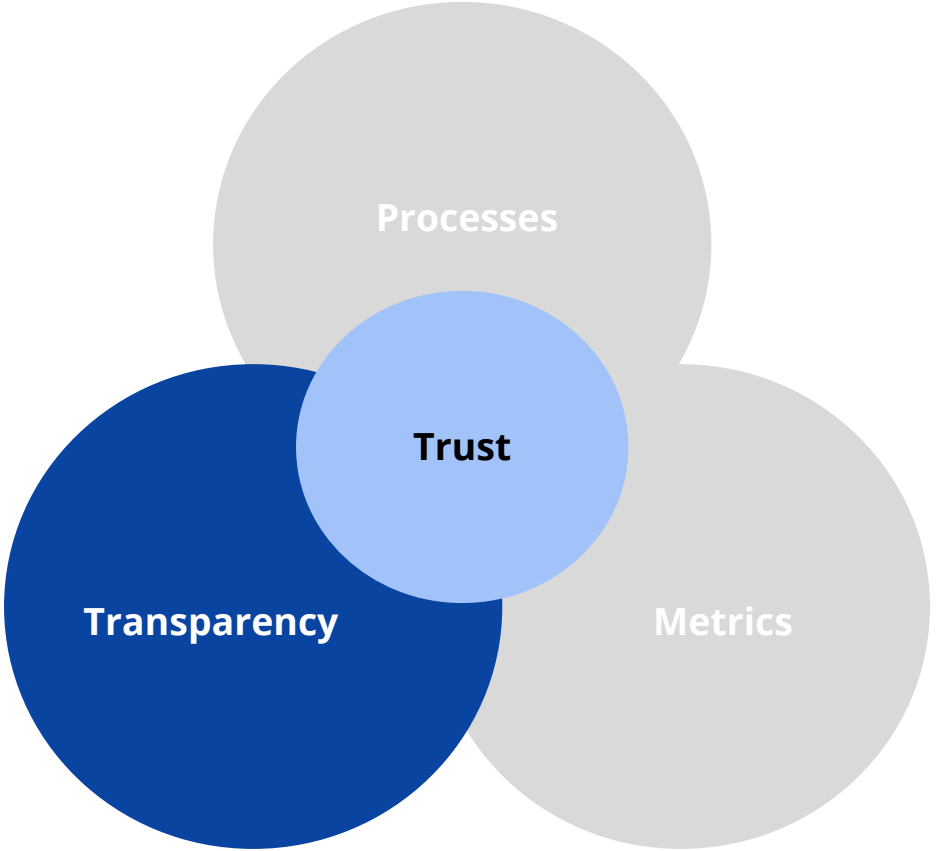
2,700+	**100+**	**9**	**< 5%**	**24x7**
Employees	Clients	Centers	Attrition	Operations

iMerit Expertise



Transparency

- What does “high quality” mean?
- What assumptions do we start with?
- What expectations do we and our labelers have of each other?



What does “high quality” mean?

High-quality labeled data is...

- Complete
- Relevant
- Unbiased
- Consistent
- Accurate

What does “high quality” mean?

High-quality labeled data is… use-case dependent

- Complete
- Relevant
- Unbiased
- Consistent
- Accurate

what’s good for the goose…
may not be good for the gander

Wikipedia commons

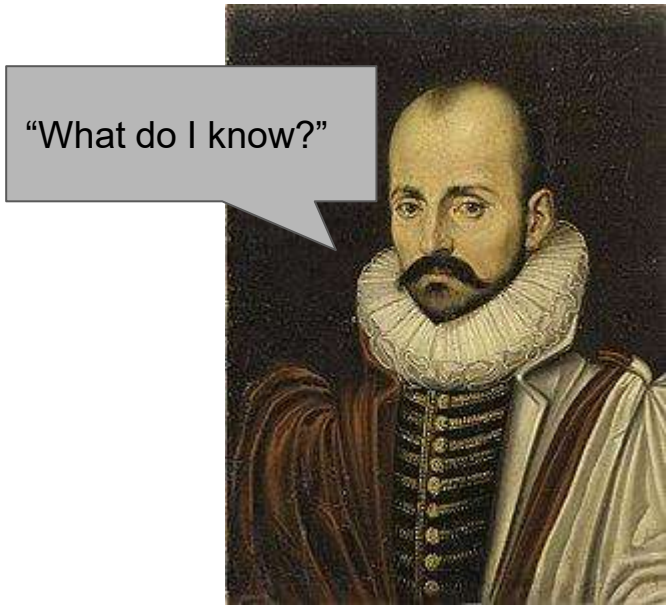
What assumptions do we start with?

iMerit

- Quality is measurable
- Ground truth exists
- Ground truth is knowable
- Consistency indicates accuracy
- Consistency is desirable

What assumptions do we start with?

- Quality is measurable
- Ground truth exists
- Ground truth is knowable
- Consistency indicates accuracy
- Consistency is desirable



What assumptions do we start with?

iMerit

Quality is measurable	… but not for all data points
Ground truth exists	… but not if the task is subjective or unbounded
Ground truth is knowable	… but not if the datapoint is ambiguous
Consistency indicates accuracy	… but not if it indicates bias
Consistency is desirable	… but not if it eliminates insights

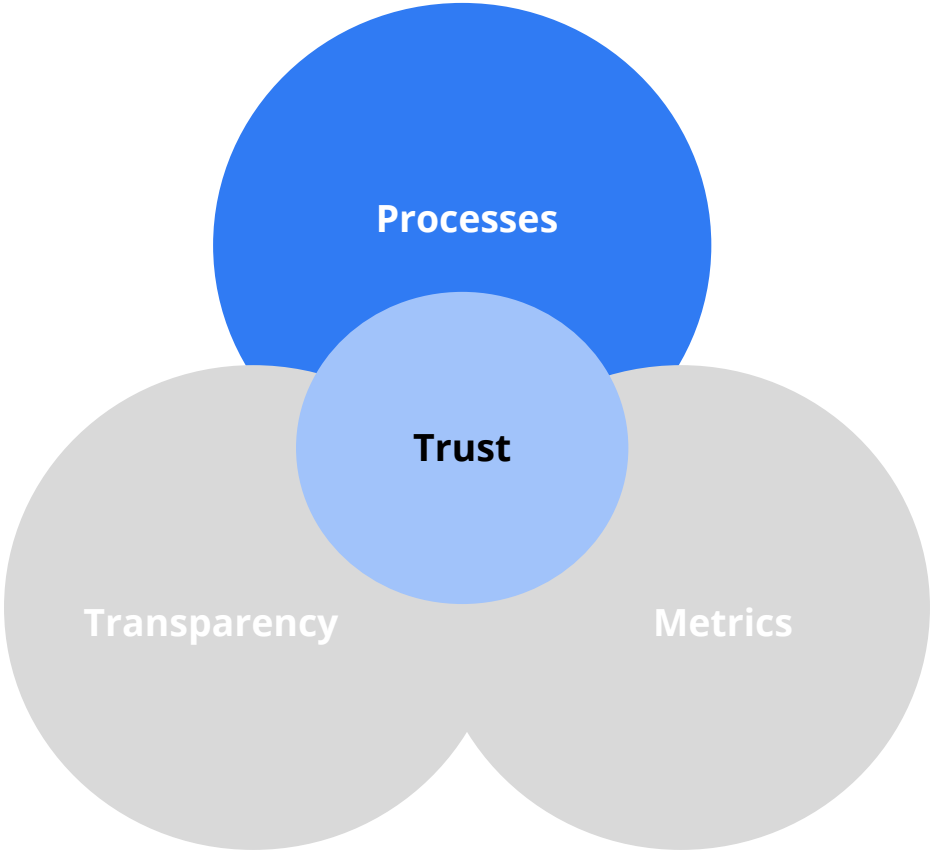
What shared expectations do we have?

Success requires mutual investment in

- Guidelines
- Definitions of quality parameters
- Development of metrics appropriate for use case
- Feedback cycle
- Transparent reporting

Processes

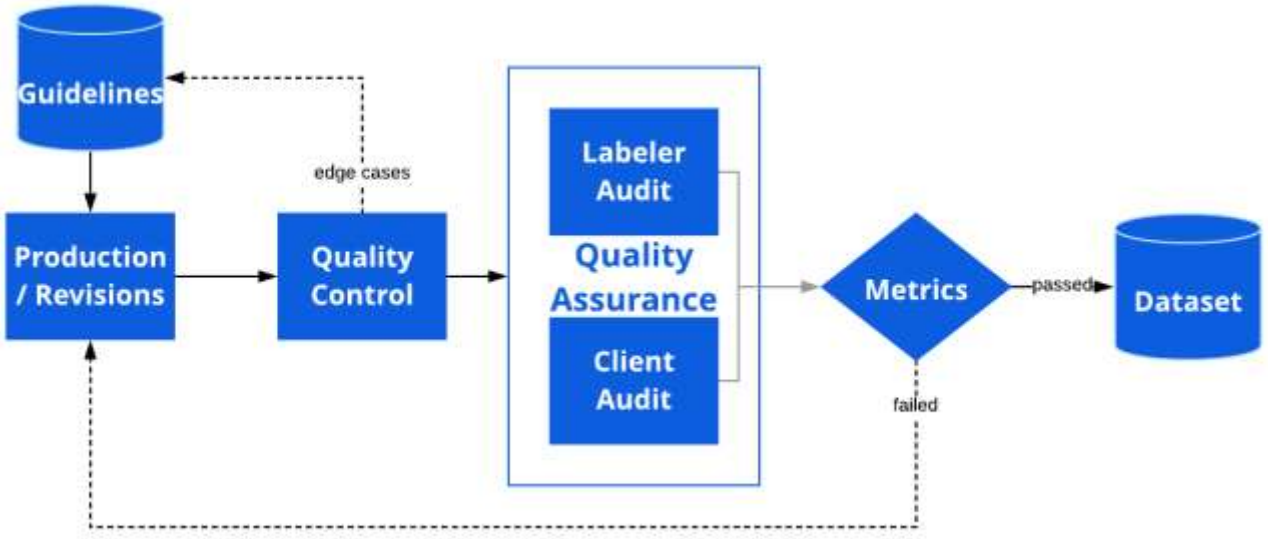
- Guidelines & training
- Workflow customization
- **Quality control**
- **Quality assurance**



Quality Control vs. Quality Assurance

QC: Workflow designed to detect and correct defects

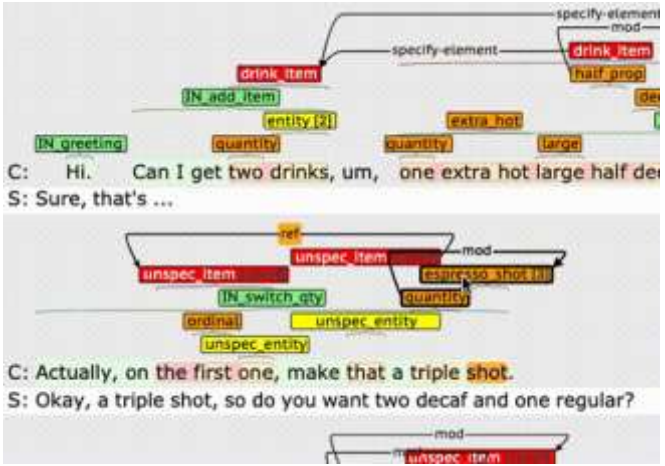
QA: Post-hoc audit designed to measure quality of a dataset



Quality Control

QC: Workflow designed to detect and correct defects

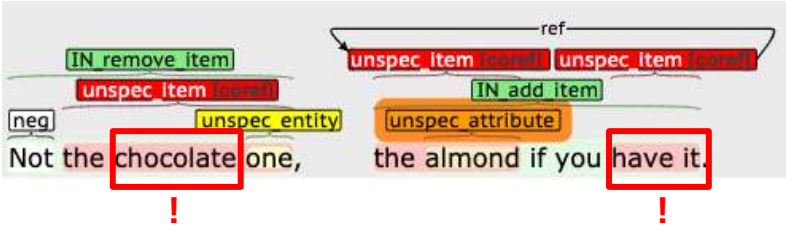
- **Optimal method depends on use case**



Quality Control

QC: Workflow designed to detect and correct defects

- Optimal methods depend on use case
- **Manual vs. automatic processes**



Quality Control

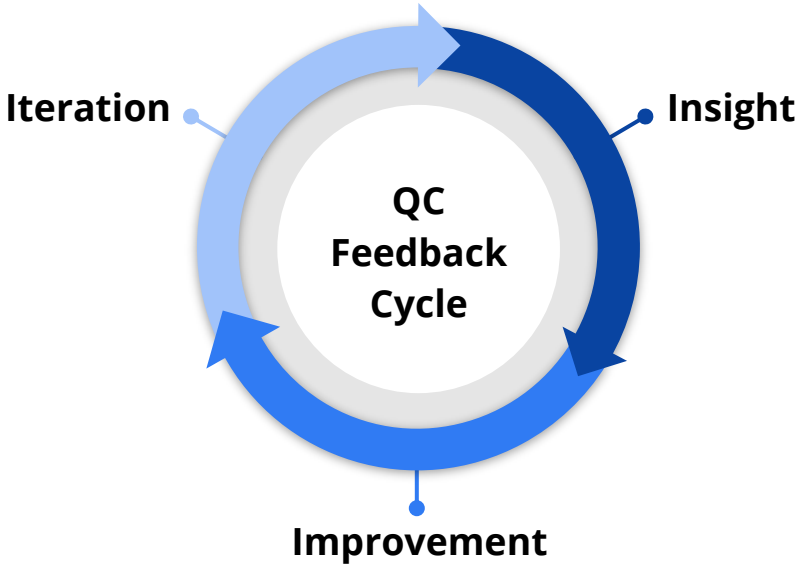
QC: Workflow designed to detect and correct defects

- Optimal methods depend on use case
- Manual vs. automatic processes
- **Process structure**
    - Multiple annotation
    - Multiple pass (expert reviewer)
    - Multiple annotation with adjudicator

Quality Control

QC: Workflow designed to detect and correct defects

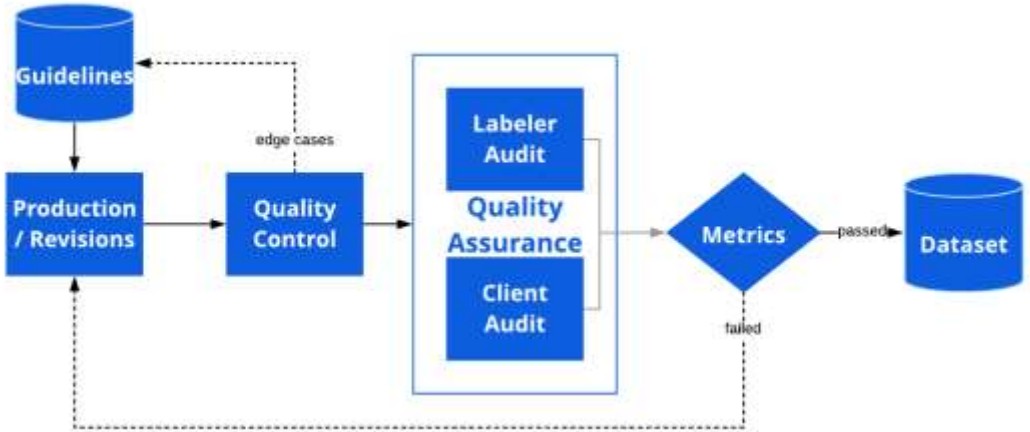
- Optimal methods depend on use case
- Manual vs. automatic processes
- Process structure
- **Interpretation of results**



Quality Assurance

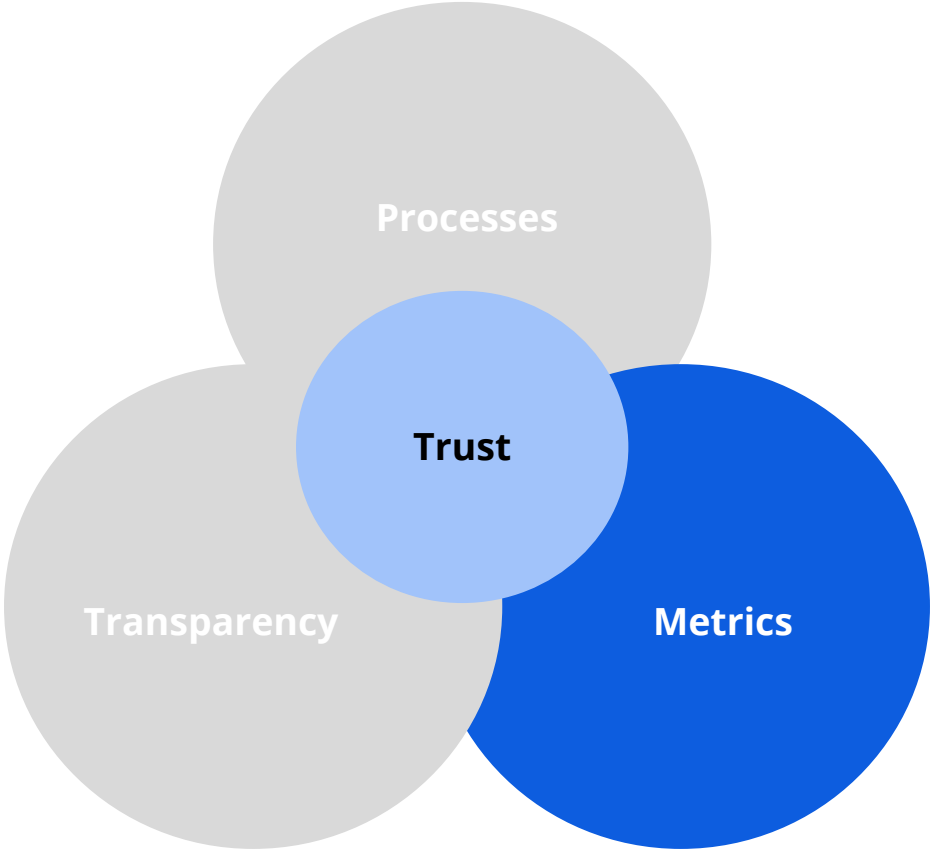
Owner: Client/internal team
Expert labelers or domain experts
Shared responsibility, higher cost burden on internal team during initial stage

Method: Double-annotation w/ consistency measures
Post-hoc comparison w/ random gold sample
Benchmarking



Metrics: More of an art than a science

- Consistency vs. accuracy
- Priority alignment
- Transparent, actionable reports



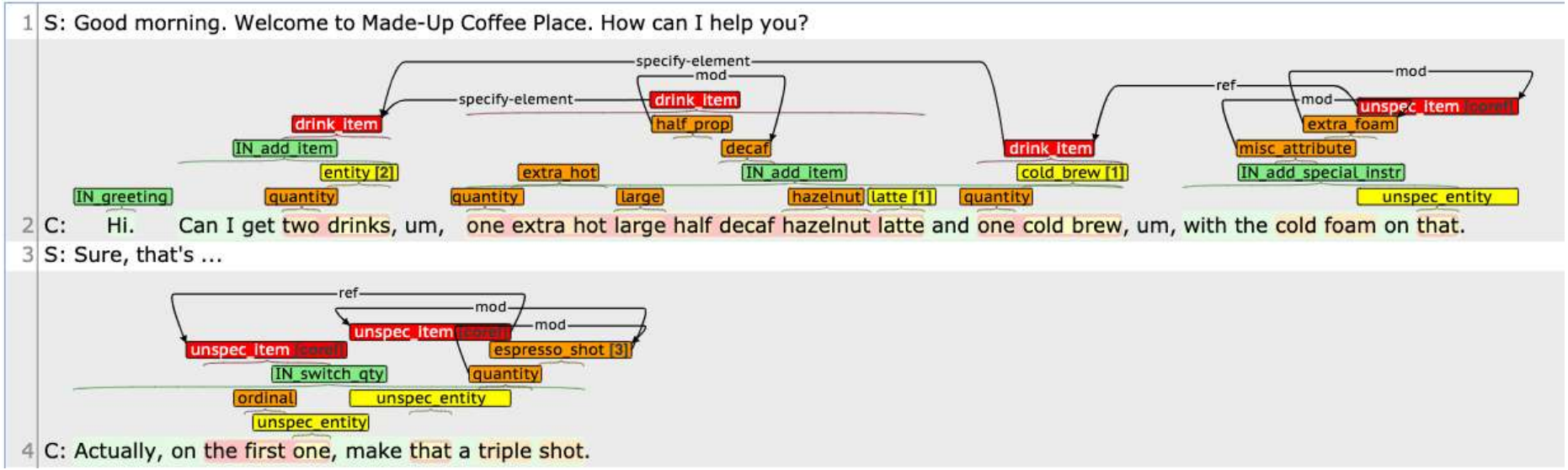
Sentiment Analysis: Consistency >>

iMerit

The best things about NOLA are the waitstaff and the atmosphere. It's a lovely space, and whoever trains the staff is a pro: they are friendly and very efficient. The food is just average. I ordered a crab cake, it was virtually flavorless and the cornbread was dry as dust. A standout for us was desert: delicious bread pudding and turtle pie. We will be coming back for the ambiance and deserts.

Positive Neutral Negative

Intent, Entity, Relation Labeling: Accuracy >>



Metrics: More of an art than a science

Set priorities

- Macro level vs. micro level
- Expected vs. actual distribution of classes
- Objective vs. subjective categories
- Major vs. minor error categories
- Primary vs. dependent labels

{We|PRON} attended {Conversational Interaction|CONFERENCE}...

{We|PRON} attended {Conversational Interaction|COMPANY}...

{We|PRON} attended Conversational Interaction...

Metrics: More of an art than a science

Set priorities

- Macro level vs. micro level
- Expected vs. actual distribution of classes
- Objective vs. subjective categories
- Major vs. minor error categories
- Primary vs. dependent labels

{We|PRON$_k$} attended {Conversational Interaction|CONFERENCE$_j$}... {it|PRON$_j$} was awesome

{We|PRON$_k$} attended {Conversational Interaction|COMPANY$_j$}... {it|PRON$_j$} was awesome

Metrics: More of an art than a science

Set priorities

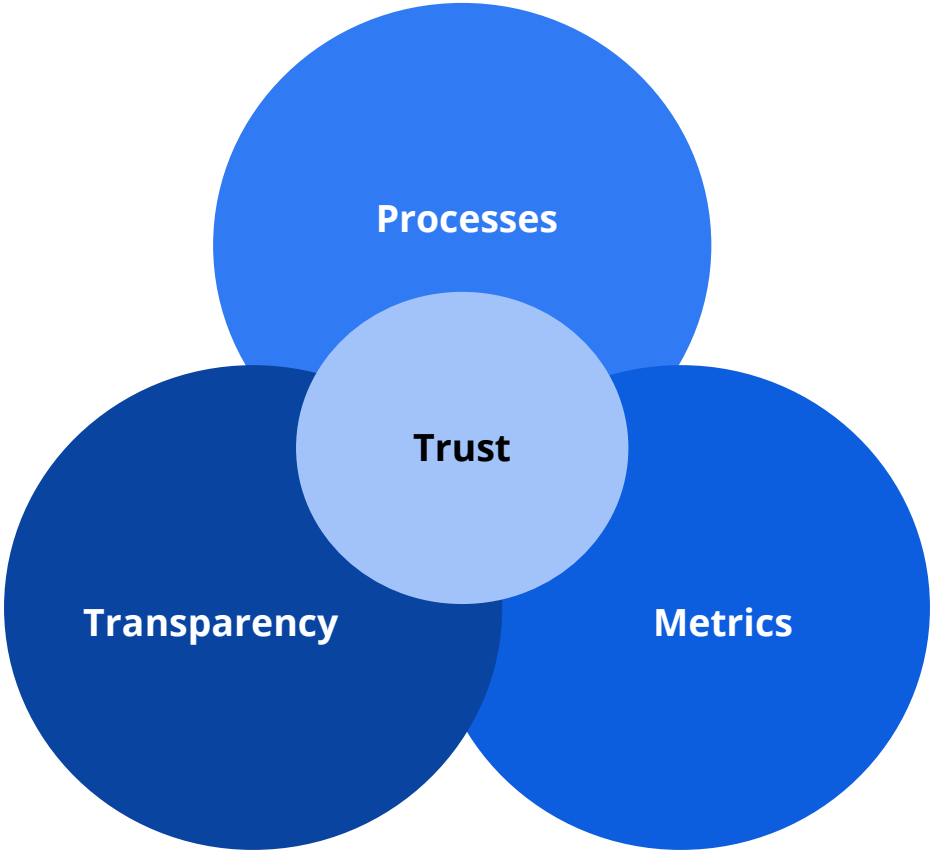
Build transparent, actionable reports

- Insights drive improvement

Entity Label Error Categories	Score
Perfect matches	90%
Misses	2%
Additions	3%
Pure label errors--diff maj. class	1%
Pure label errors--diff min. class	2%
Pure span error--nesting	0.2%
Pure span error--shifting	0.4%
Label+span error…	…

Building Trust in Data Quality

iMerit



Thank You!

teresa@imerit.net

iMerit
human-empowered computing