Continued Improvements in Patient Deduplication:

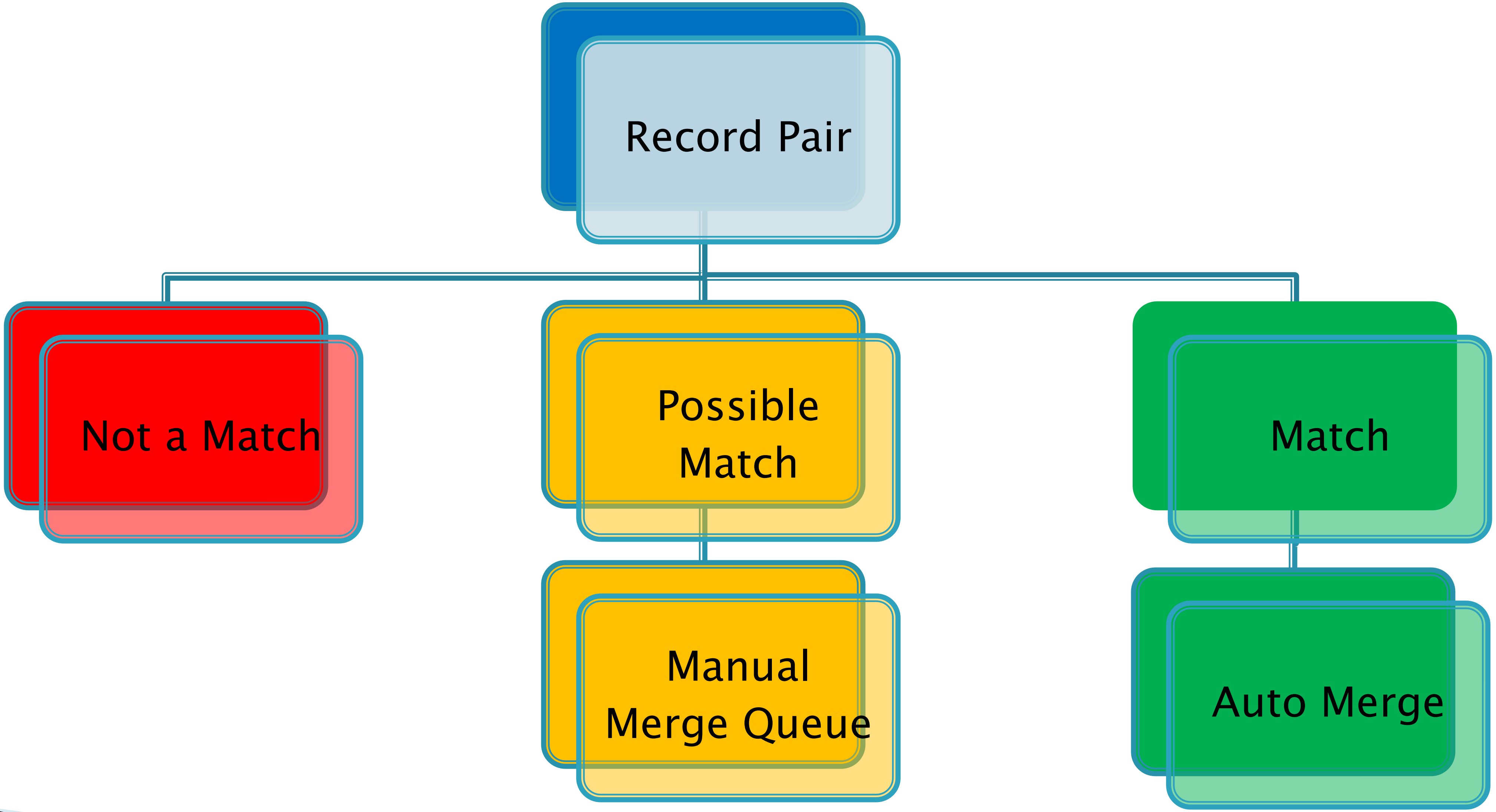
Improving patient deduplication and overall data quality using artificial intelligence

Sam Belton
Strategic Solutions Group
sbelton@ssg-llc.com
MIIS IT Project Team
Production Support

Agenda

1. MIIS deduplication methods
2. Methodology and tools for analysis
3. High level goals of analysis
4. Comparing the 2018 data to 2017
5. Detailed review of individual match cases
6. Outcomes and updates

Patient Deduplication System



Massachusetts IIS Patient Deduplication

- The Massachusetts IIS (MIIS) currently uses a weight based algorithm that has been trained using both test data as well as production data.
- Prior to using an AI tool to evaluate the system we had to rely on users to report issues with deduplication.

- Last year we were able to evaluate and re-train the MIIS using Lantern (our AI tool).
- This was the largest retraining effort for our deduplication system since we began training with production data.

SSG

Timeline of Patient Deduplication in the MIIS

06/2013	04/2014	10/2015	08/2017
MIIS v3.0	MIIS v3.4.3	MIIS v3.5	MIIS v5.0

2013 – The deduplication engine at this time had been trained only using test data.

2014 - The deduplication algorithm was re-trained with a large subset of production data.

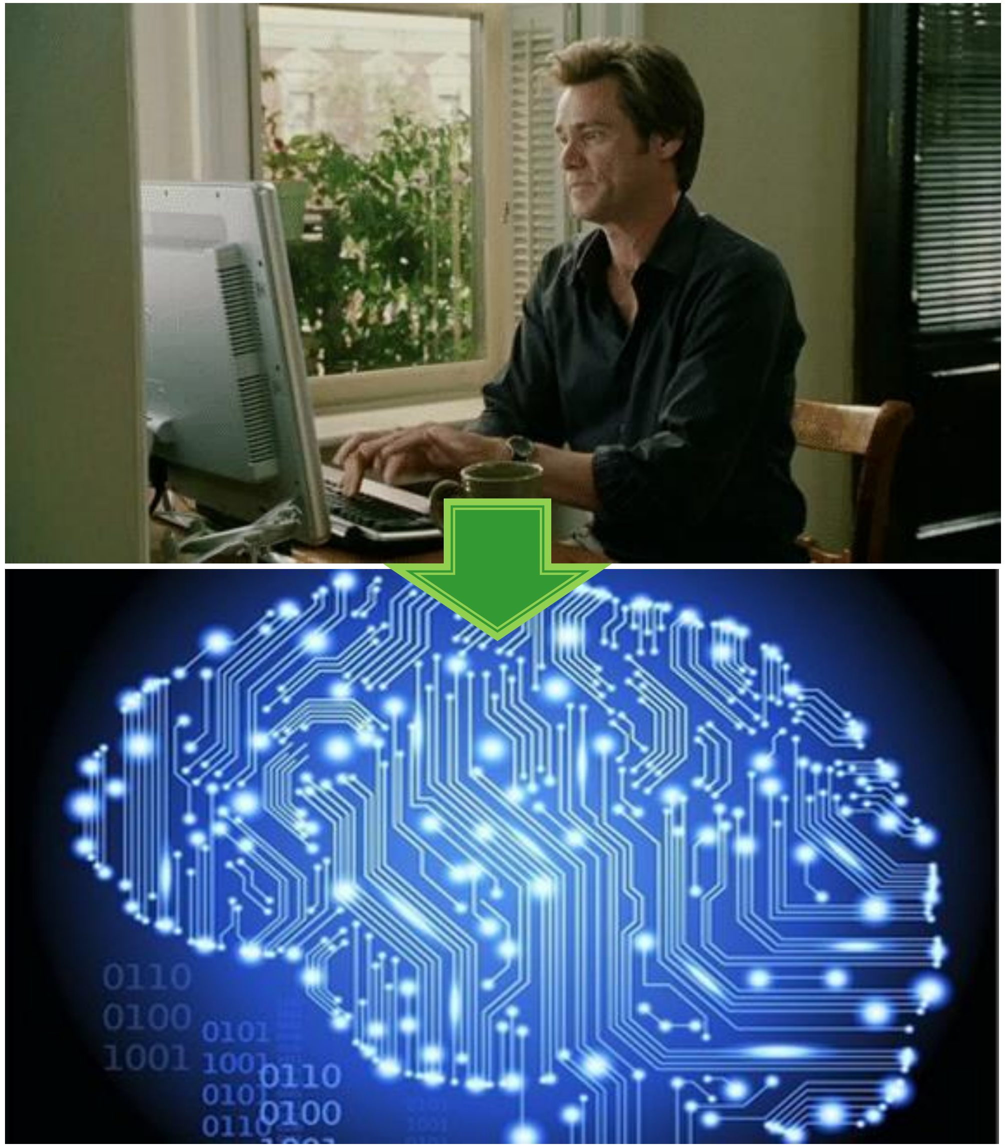
2015 - The MIIS was updated to send a multiple birth indicator into the deduplication algorithm.

2017 – The Lantern tool was used to identify both code fixes and necessary re-training. Most notably training for sparsely populated records with null values in key fields.

SSG

Concept for Lantern Person Matching Tool

- Develop a tool to do this analysis work automatically.
- The tool can sit outside of the IIS application, accessing the data via a DB connection.
- Configurable to be used with any record system.
- Lantern, would use deep learning technology (a form of AI) to learn patterns and accurately determine the probability that any two pairs are a match.
- Trained with Massachusetts production data.
- For any pair of records analyzed, a predicted matching probability would be produced.



SSG

Concept for Lantern Person Matching Tool

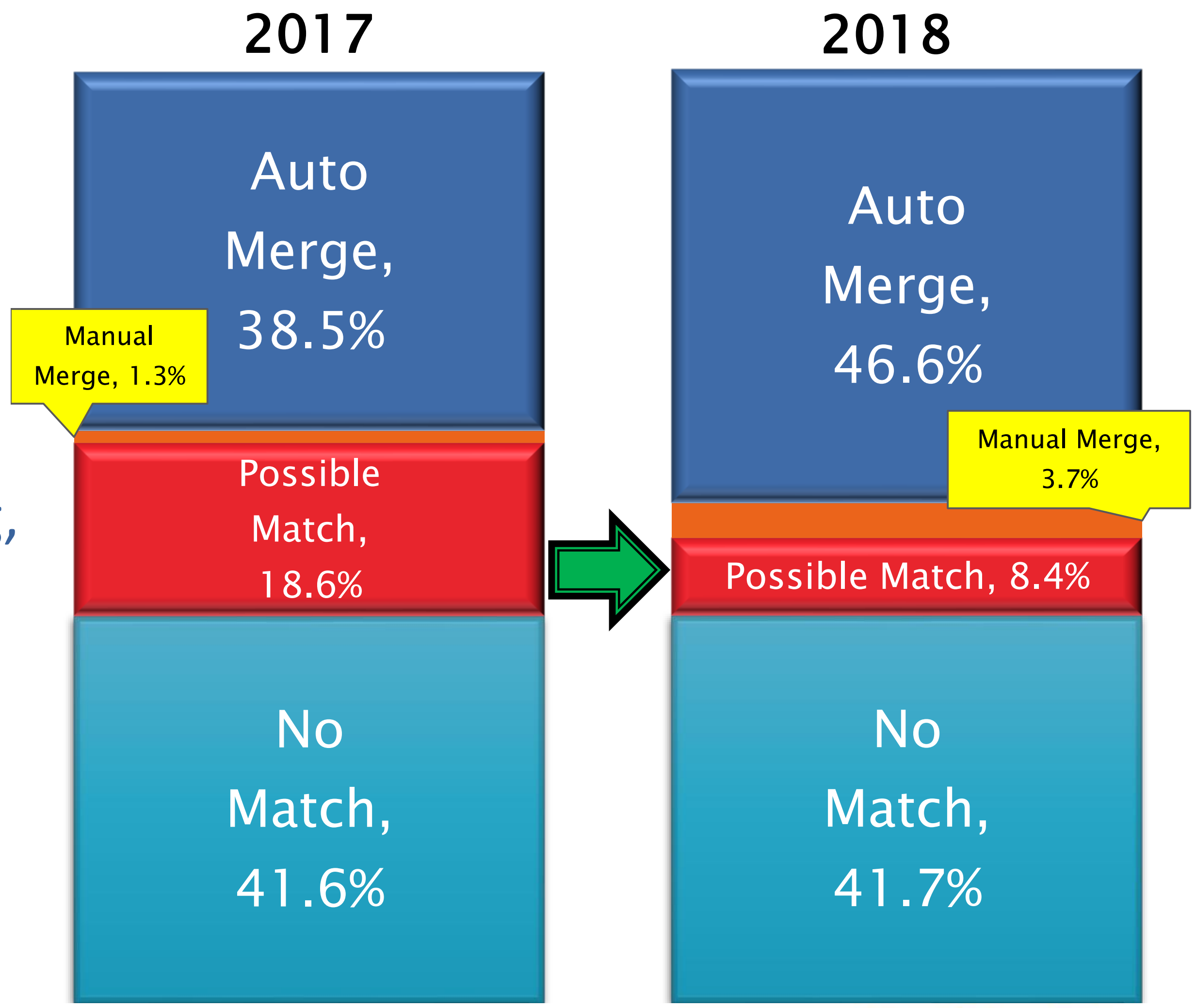
- Develop a tool to do this analysis work automatically.
- The tool can sit outside of the IIS application, accessing the data via a DB connection.
- Configurable to be used with any record system.
- Lantern, would use deep learning technology (a form of AI) to learn patterns and accurately determine the probability that any two pairs are a match.
- Trained with Massachusetts production data.
- For any pair of records analyzed, a predicted matching probability would be produced.

SSG

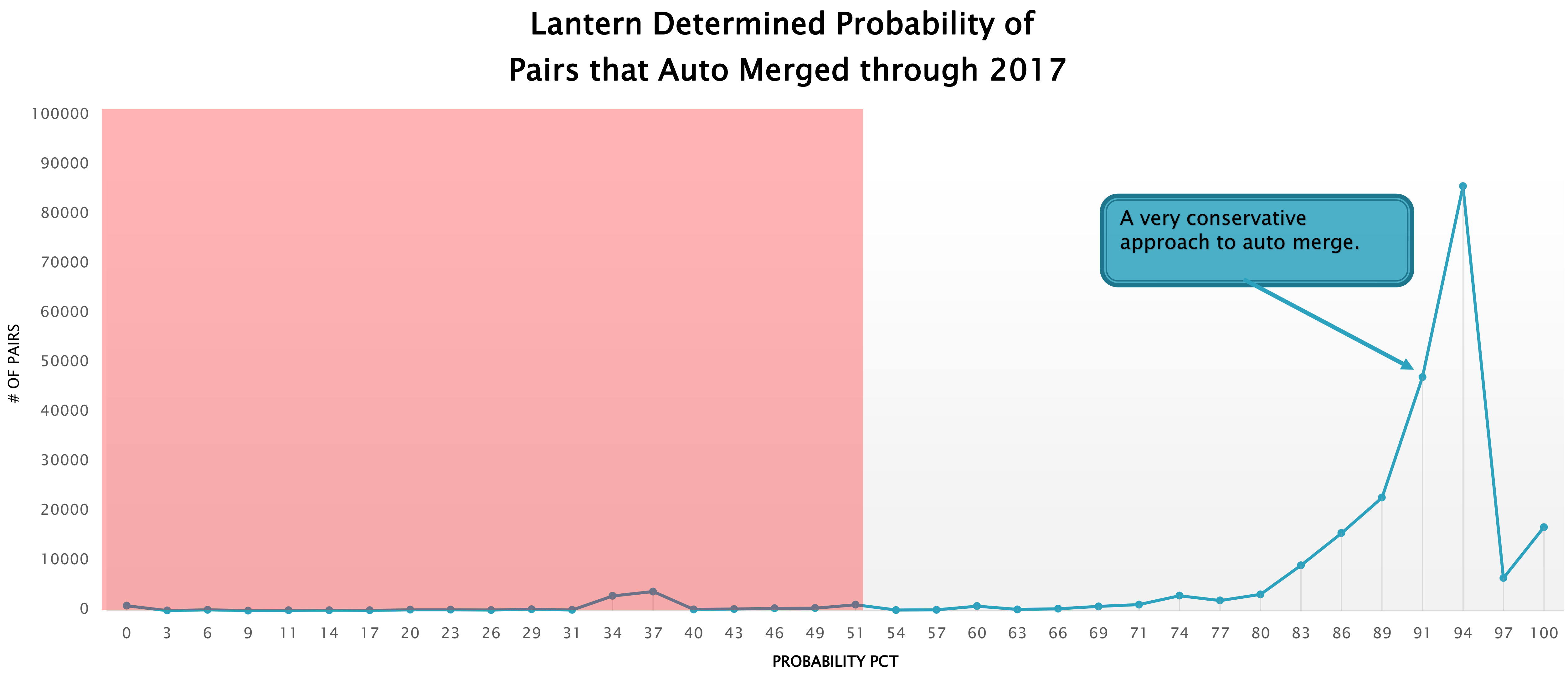
Reducing the burden of Manual Matching

- In 2017 we had two goals for deduplication analysis:
  1) Ensure that there are no incorrect auto merges occurring.
  2) Reduce the size of the manual merge queue, the queue was too large for providers to effectively use.
- We were able to reduce the manual merge queue from over 660k pairs prior to retraining, down to 270k pairs on the queue.
- How did we do for accuracy?

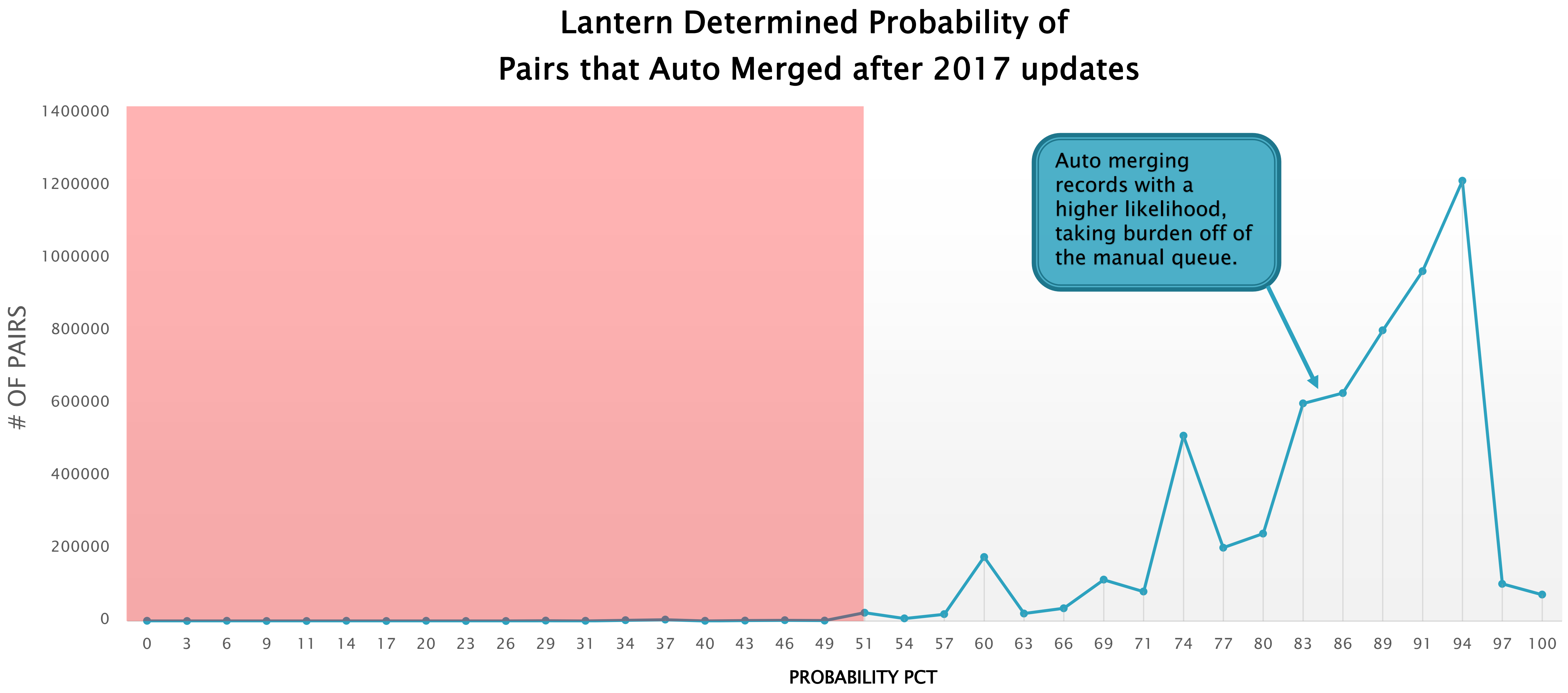
Patients in MIIS by Category



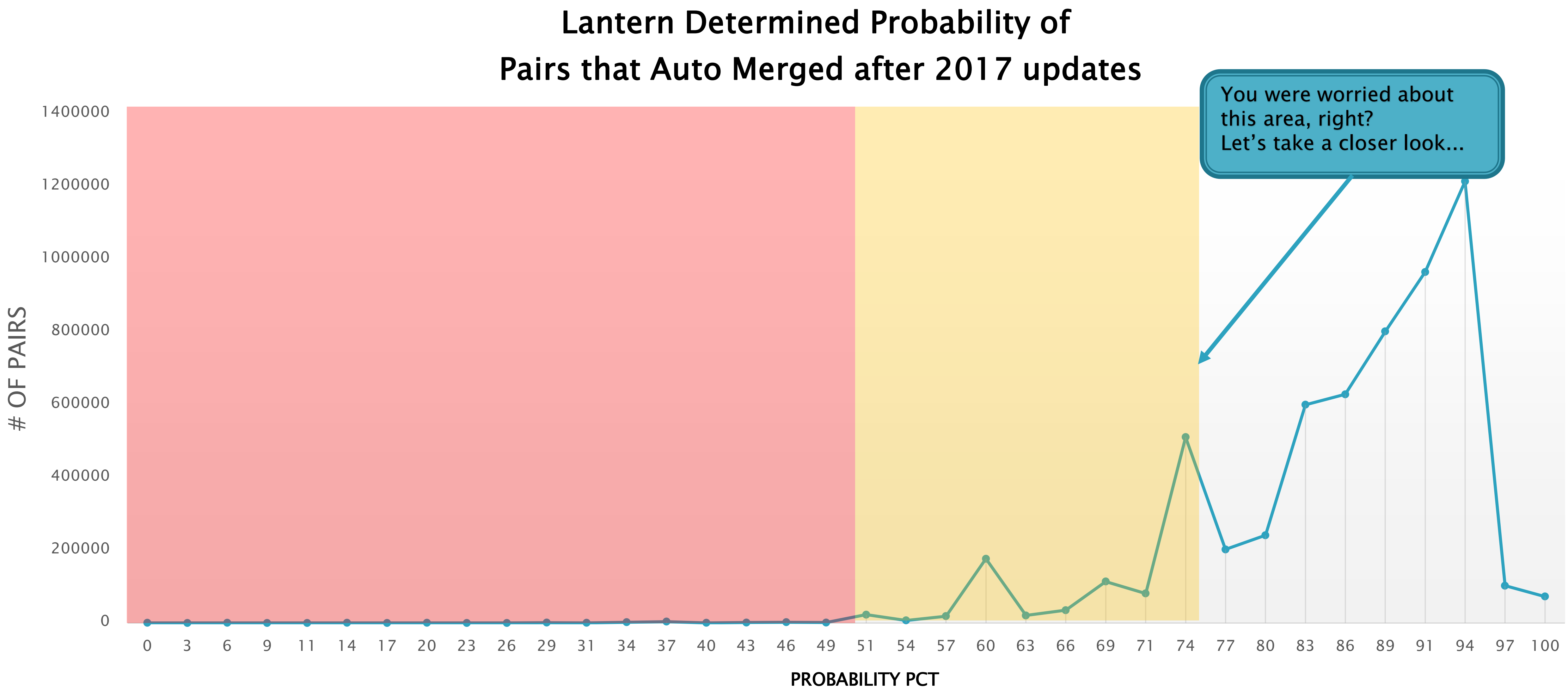
Probabilistic Analysis of Auto Merges



Probabilistic Analysis of Auto Merges



Probabilistic Analysis of Auto Merges



Auto Merge pairs with 50% – 75% Confidence

Last Match	First Match	DOB Match	Middle Match	Address Match	Total Pairs	Pct Pairs
FALSE	FALSE	TRUE	TRUE	TRUE	1	0.12%
	TRUE	FALSE		FALSE	1	0.12%
		TRUE	FALSE	FALSE	71	8.73%
				TRUE	2	0.25%
			TRUE	FALSE	2	0.25%
TRUE	FALSE		FALSE	FALSE	11	1.35%
				TRUE	1	0.12%
	TRUE	FALSE		FALSE	160	19.68%
				TRUE	1	0.12%
			TRUE	FALSE	29	3.57%
				TRUE	1	0.12%
		TRUE	FALSE	FALSE	533	65.56%
Grand Total					813	

Auto Merge pairs with 50% – 75% Confidence

Last Match	First Match	DOB Match	Middle Match	Address Match	Total Pairs	Pct Pairs
FALSE	FALSE	TRUE	TRUE	TRUE	1	0.12%
	TRUE	FALSE		FALSE	1	0.12%
		TRUE	FALSE	FALSE	71	8.73%
				TRUE	2	0.25%
			TRUE	FALSE	2	0.25%
TRUE	FALSE		FALSE	FALSE	11	1.35%
				TRUE	1	0.12%
	TRUE	FALSE		FALSE	160	19.68%
				TRUE	1	0.12%
			TRUE	FALSE	29	3.57%
				TRUE	1	0.12%
		TRUE	FALSE	FALSE	533	65.56%
Grand Total					813	

Example Patient Pair with 65% confidence

- Here is an example where Last, Middle, Address are not matching.
- We can see that we flagged the last name as not matching due to a space in one name. The Middle and Address do not match due to missing data, which we specifically trained for last year.

Field	Patient A	Patient B
Last Name	De Jesus	DeJesus
Middle Name	Mark	
First Name	Samuel	Samuel
Gender	Male	Male
Date of Birth	01/01/2001	01/01/2001
Street Address	4 Oak Street	
City	Boston	
State	MA	

- Of this subset, where the last name, middle, and address did not match; a hyphen, space, or apostrophe in one of the last names accounted for 100% of the 71 merges that occurred.

SSG

Auto Merge pairs with 50% – 75% Confidence

Last Match	First Match	DOB Match	Middle Match	Address Match	Total Pairs	Pct Pairs
FALSE	FALSE	TRUE	TRUE	TRUE	1	0.12%
	TRUE	FALSE		FALSE	1	0.12%
		TRUE	FALSE	FALSE	71	8.73%
				TRUE	2	0.25%
			TRUE	FALSE	2	0.25%
TRUE	FALSE		FALSE	FALSE	11	1.35%
				TRUE	1	0.12%
	TRUE	FALSE		FALSE	160	19.68%
				TRUE	1	0.12%
			TRUE	FALSE	29	3.57%
				TRUE	1	0.12%
		TRUE	FALSE	FALSE	533	65.56%
Grand Total					813	

Pairs with last name that differ due to space, hyphen, or apostrophe.

Example Patient Pair with 55% confidence

- Here is an example of First and Last Match and DOB, Middle, Address Not matching.
- We can see that training around swapped month and year was seen as just as confidant as matching DOB in the MIIS.

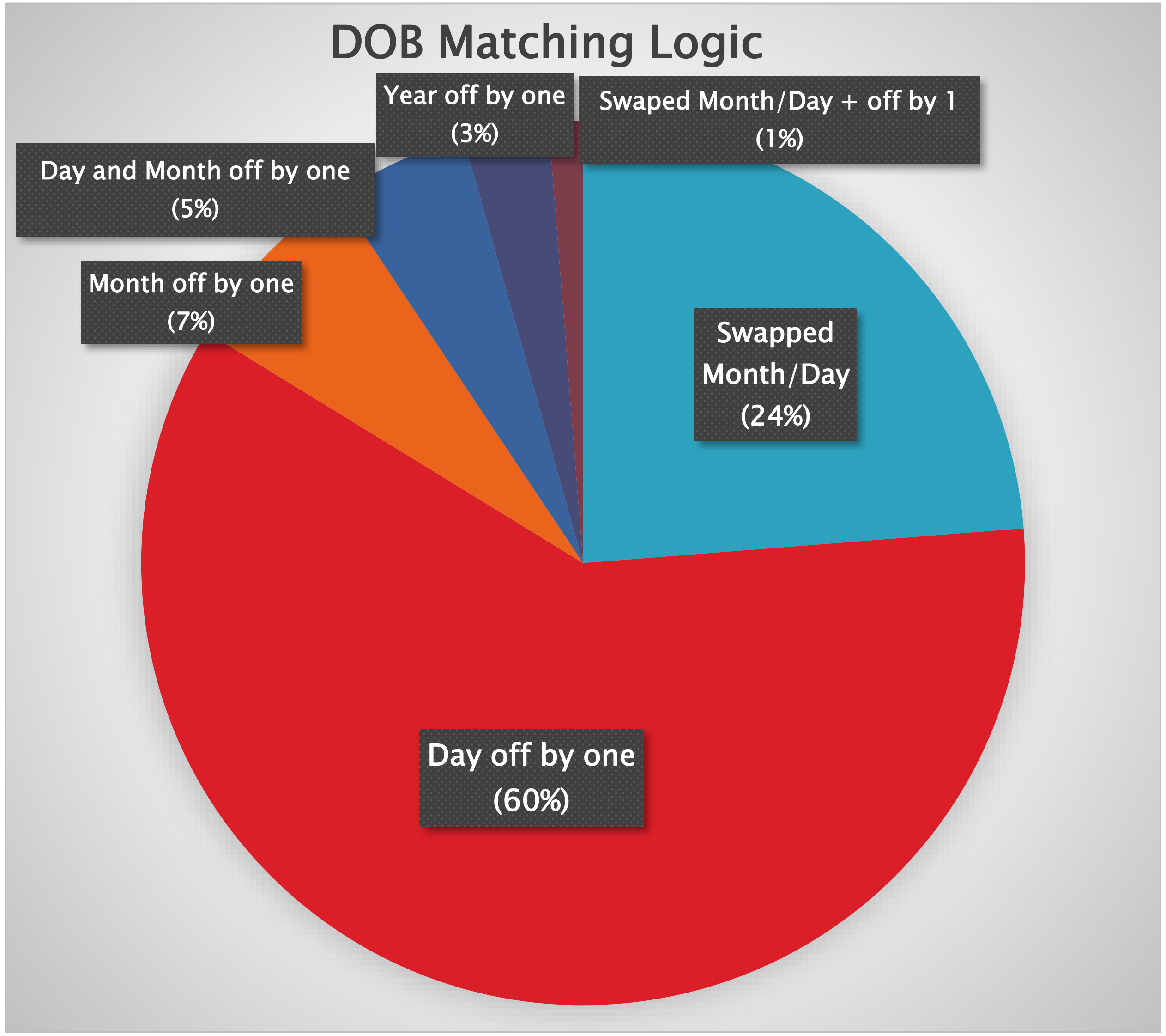
Field	Patient A	Patient B
Last Name	Greenwood	Greenwood
Middle Name	M	
First Name	Samuel	Samuel
Gender	Male	Male
Date of Birth	08/07/2001	07/08/2001
Street Address		4 Oak Street
City		Boston
State		MA

- These might be a better candidate for the manual merge queue, especially since we became less restrictive on matching middle and address we shouldn't be auto merging as much with swapped month and day.

SSG

DOB Analysis Breakdown

- We evaluated the160 pairs in this group that differ on DOB.
- When EHR and IIS systems first launched there were many data quality issues, this required adjustments to compensate. Over time that data has become much cleaner.
- Here “fuzzy” logic designed to correct for typos in the DOB was pushing these pairs into the Auto Merge category.



SSG

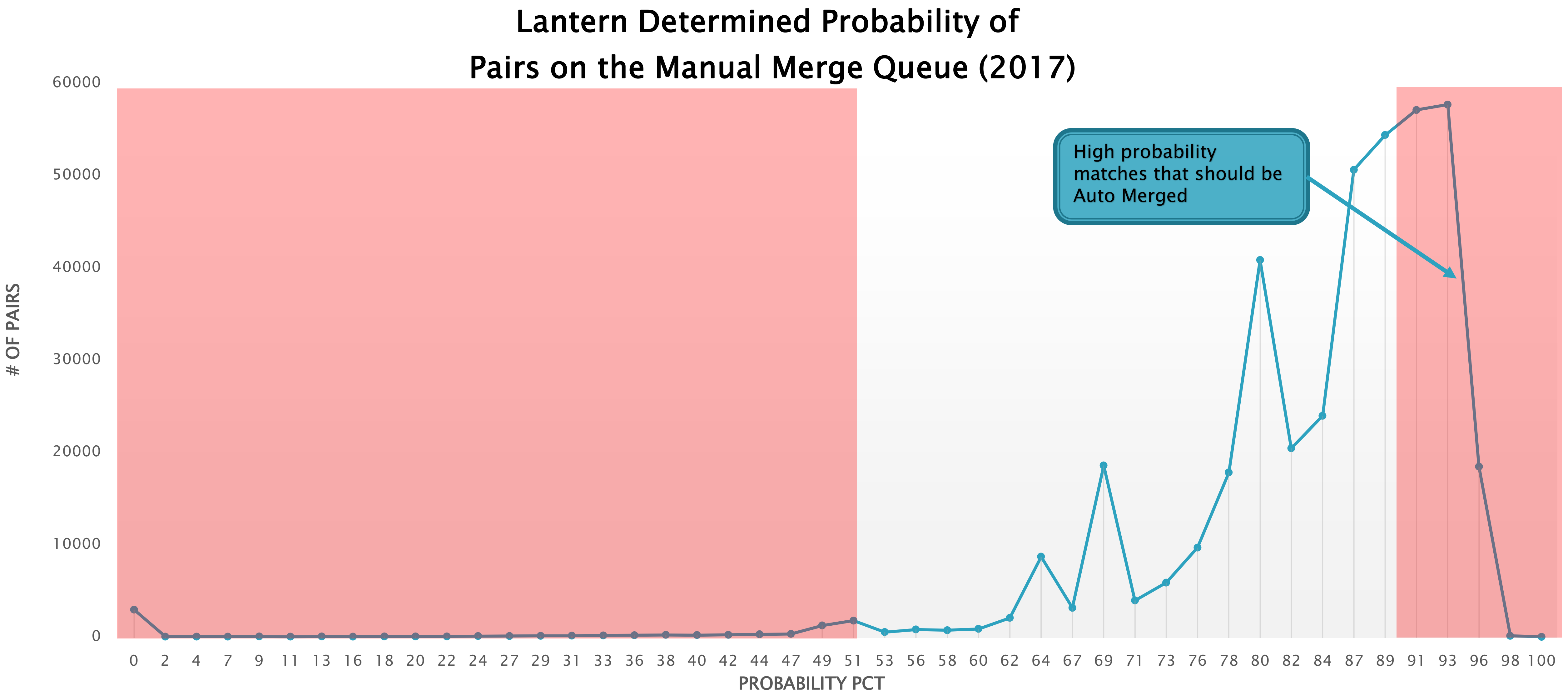
Auto Merge pairs with 50% – 75% Confidence

Last Match	First Match	DOB Match	Middle Match	Address Match	Total Pairs	Pct Pairs
FALSE	FALSE	TRUE	TRUE	TRUE	1	0.12%
	TRUE	FALSE		FALSE	1	0.12%
		TRUE	FALSE	FALSE	71	8.73%
				TRUE	2	0.25%
			TRUE	FALSE	2	0.25%
TRUE	FALSE		FALSE	FALSE	11	1.35%
				TRUE	1	0.12%
	TRUE	FALSE		FALSE	160	19.68%
				TRUE	1	0.12%
			TRUE	FALSE	29	3.57%
				TRUE	1	0.12%
		TRUE	FALSE	FALSE	533	65.56%
Grand Total					813	

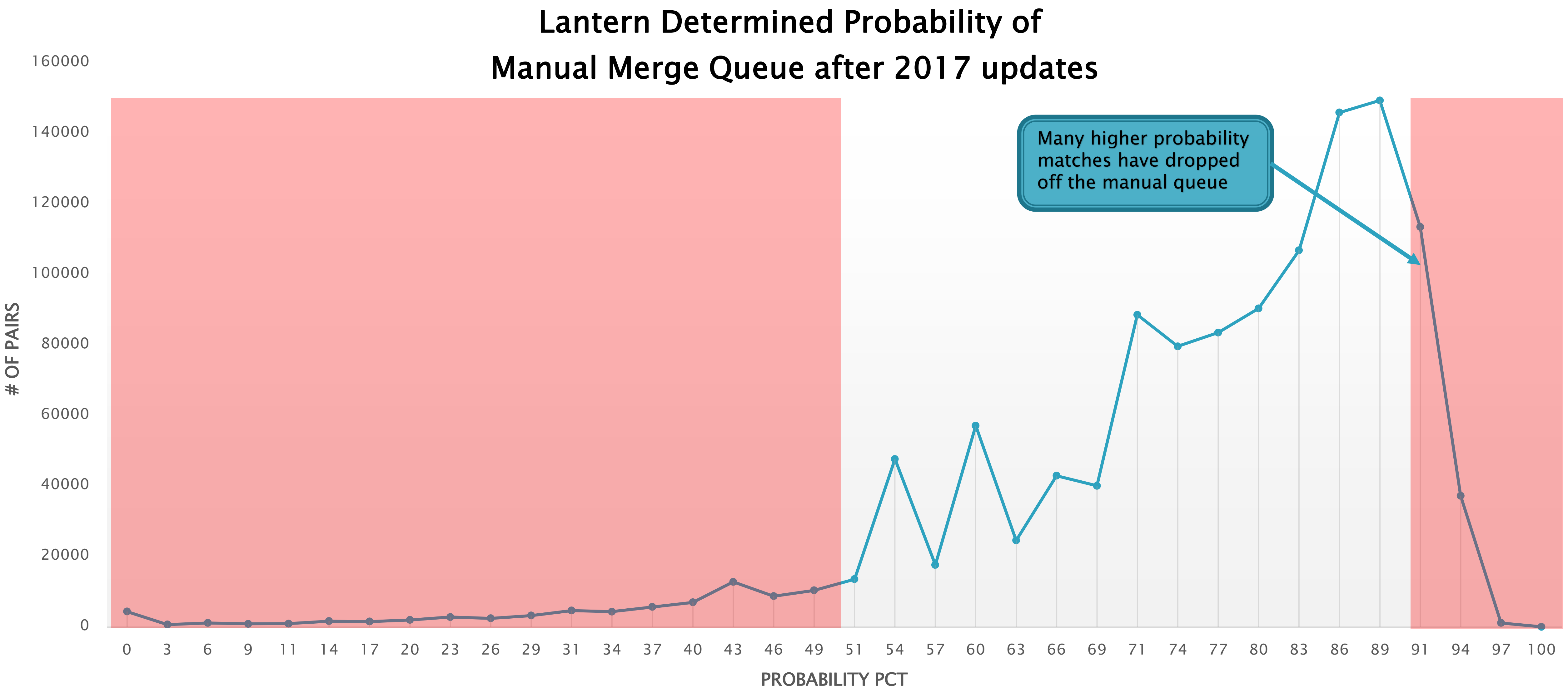
Pairs with last name that differ due to space, hyphen, or apostrophe.

Pairs used “fuzzy logic” to correct for DOB typos. May be good candidates for retraining.

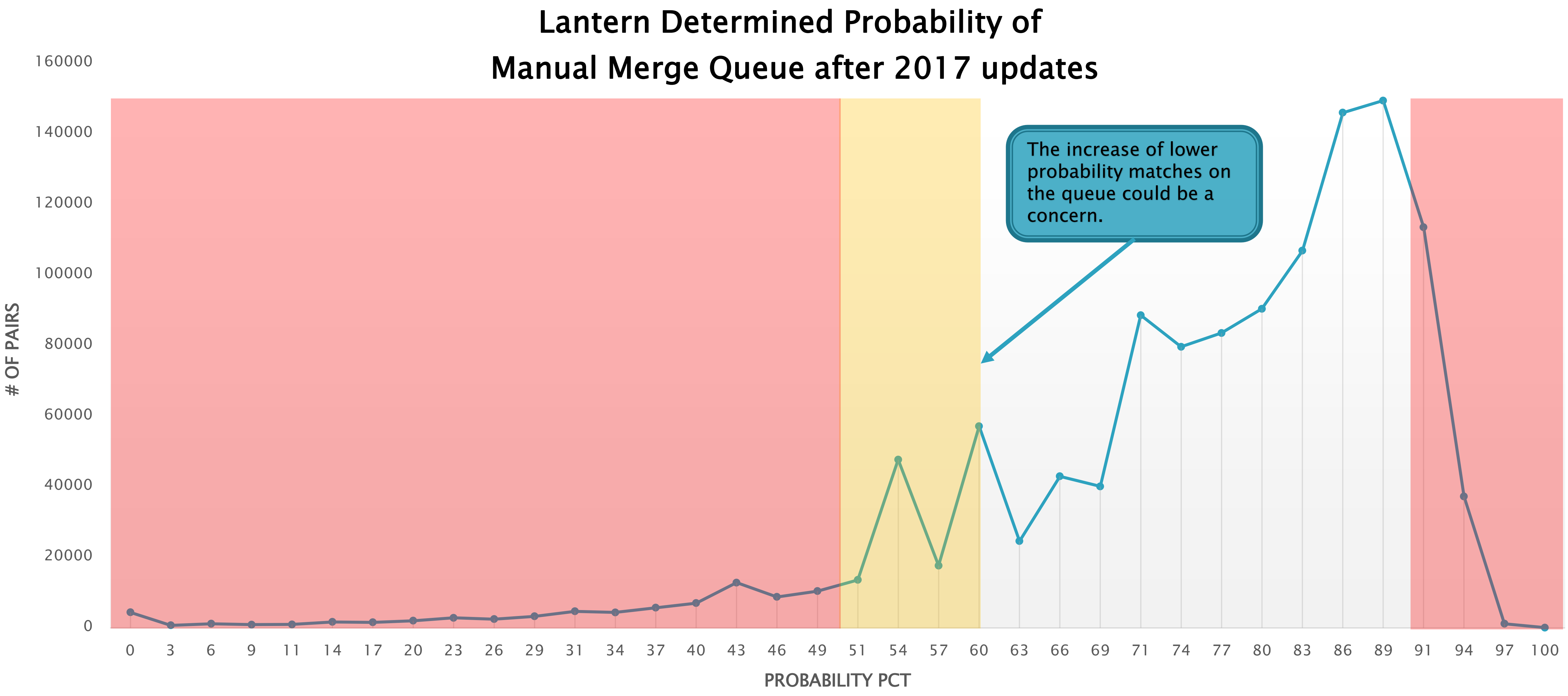
Probabilistic Analysis of Manual Merge Queue



Probabilistic Analysis of Manual Merge Queue



Probabilistic Analysis of Manual Merge Queue



Pairs on the Manual Queue with <60% Confidence

Last Match	First Match	DOB Match	Middle Match	Total Pairs	Pct of Pairs
FALSE	FALSE	FALSE	FALSE	**20**	**4.80%**
			TRUE	**8**	**1.92%**
		TRUE	FALSE	80	19.18%
			TRUE	11	2.64%
	TRUE	FALSE	FALSE	49	11.75%
			TRUE	12	2.88%
		TRUE	FALSE	180	43.17%
			TRUE	3	0.72%
TRUE	FALSE	FALSE	FALSE	5	1.20%
			TRUE	1	0.24%
		TRUE	FALSE	43	10.31%
	TRUE	FALSE	FALSE	5	1.20%
Grand Total				**417**	

SSG

Pairs on the Manual Queue with <60% Confidence

Last Match	First Match	DOB Match	Middle Match	Total Pairs	Pct of Pairs
FALSE	FALSE	FALSE	FALSE	**20**	**4.80%**
			TRUE	**8**	**1.92%**
		TRUE	FALSE	80	19.18%
			TRUE	11	2.64%
	TRUE	FALSE	FALSE	49	11.75%
			TRUE	12	2.88%
		TRUE	FALSE	180	43.17%
			TRUE	3	0.72%
TRUE	FALSE	FALSE	FALSE	5	1.20%
			TRUE	1	0.24%
		TRUE	FALSE	43	10.31%
	TRUE	FALSE	FALSE	5	1.20%
Grand Total				**417**	

SSG

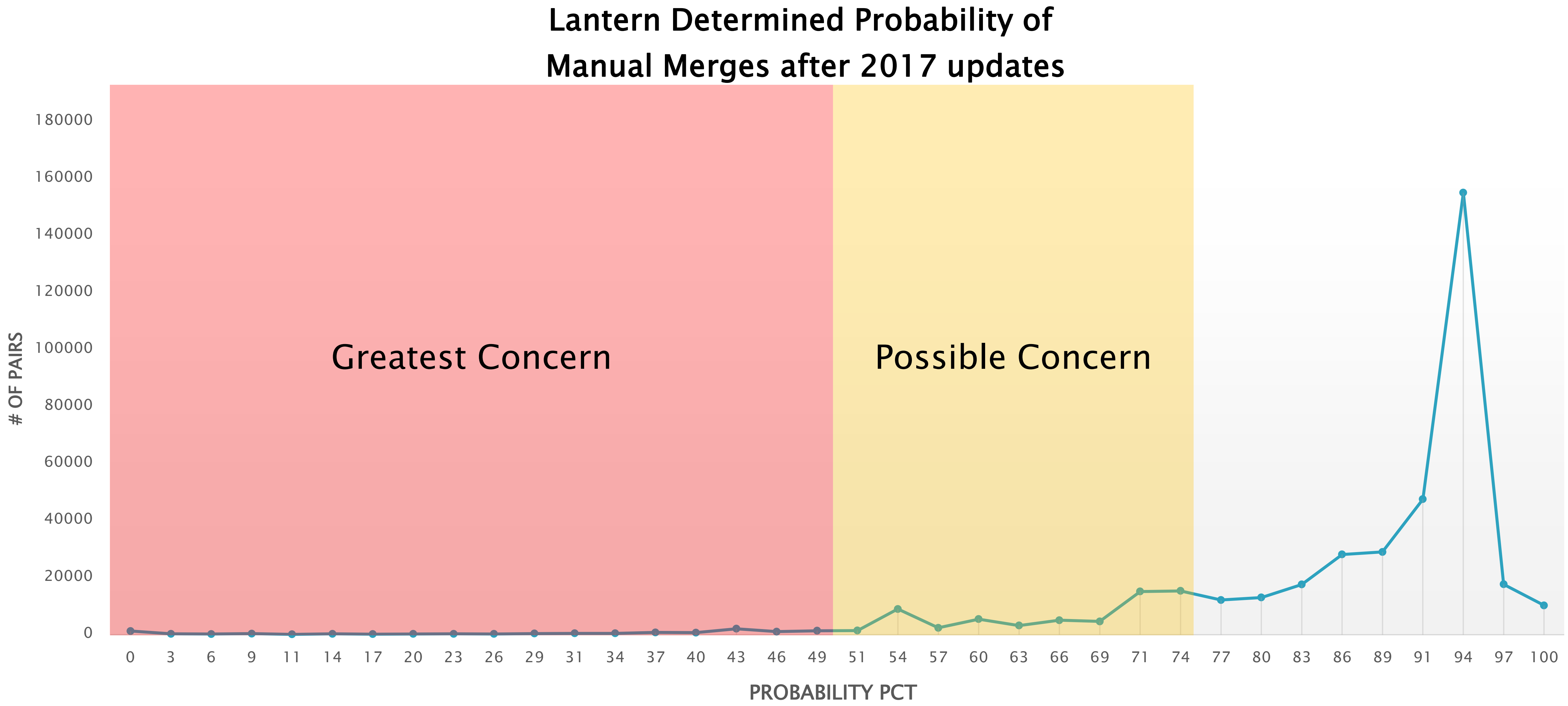
Example Patient Pair on Manual Queue with low confidence

- This is an example of one of the lowest confidence pairs on the queue, at just 14% confidence from Lantern.
- We can see enough similarities to understand why the system picked these records, but we may not want to allow users the opportunity to merge examples like this.
- This follows a similar pattern with the DOB still ranking at high confidence. If we do retraining for those DOB issues from the auto merge, the confidence level on this patient would drop too.

Field	Patient A	Patient B
Last Name	Sample-One	Sample-Two
Middle Name	M	
First Name	Julien	Jariel
Gender	Male	Male
Date of Birth	10/02/2001	02/10/2001
Street Address		2 Maple Street
City		Boston
State		MA

SSG

Probabilistic Analysis of Manual Merges by Users



Manual Merge pairs with 50% – 75% Confidence

Last Match	First Match	DOB Match	Middle Match	Address Match	Total Pairs	Pct Pairs
FALSE	FALSE	TRUE	FALSE	FALSE	19	2.64%
FALSE	FALSE	TRUE	FALSE	TRUE	4	0.56%
FALSE	FALSE	TRUE	TRUE	FALSE	2	0.28%
FALSE	TRUE	FALSE	FALSE	FALSE	4	0.56%
FALSE	TRUE	FALSE	FALSE	TRUE	1	0.14%
FALSE	TRUE	FALSE	TRUE	FALSE	1	0.14%
FALSE	TRUE	TRUE	FALSE	FALSE	392	54.44%
FALSE	TRUE	TRUE	FALSE	TRUE	3	0.42%
FALSE	TRUE	TRUE	TRUE	FALSE	49	6.81%
TRUE	FALSE	FALSE	TRUE	FALSE	1	0.14%
TRUE	FALSE	TRUE	FALSE	FALSE	84	11.67%
TRUE	FALSE	TRUE	FALSE	TRUE	2	0.28%
TRUE	FALSE	TRUE	TRUE	FALSE	3	0.42%
TRUE	TRUE	FALSE	FALSE	FALSE	10	1.39%
TRUE	TRUE	FALSE	FALSE	TRUE	2	0.28%
TRUE	TRUE	FALSE	TRUE	FALSE	3	0.42%
TRUE	TRUE	TRUE	FALSE	FALSE	140	19.44%
Grand Total					**720**	

SSG

Example Manually Merged Pair with low confidence

- While in this case the user may know more about the patient as they have direct patient contact, this match seems unlikely.
- This was either a result of bad data collected by one of the providers, or an incorrect merge.
- This is an opportunity to not just create a better algorithm, but improve data quality and/or train end users.

Field	Patient A	Patient B
Last Name	Sample	Sampson
Middle Name		James
First Name	Samuel	Samuel
Gender	M	M
Date of Birth	01/17/2000	01/18/2000
Street Address		15 Oak Street
City		Boston
State		MA

SSG

What did we learn?

- Updates from 2017 have significantly helped to alleviate the burden of the manual queue off of users. We accomplished this while staying very accurate, in most cases.
- We were able to identify marginal auto merges and unlikely possible matches, which are an opportunity for the MIIS to re-train.
- Running the analysis for a second year in a row, we were able to use an efficient methodological approach, along with an external tool, to quickly and effectively identify issues (and successes).

SSG

Keys to successful Deduplication Analysis & Improvements

1. Always follow the **DATA**!
   a) Start with high level trends and narrow down on specific cases. This will translate into the most effective updates.
   b) Don't immediately focus on one type of case just because a provider reported it. This may be a valid area to work on, but let the data lead you there.
2. Don't forget the **BIG** picture!
   a) It's easy to get caught up in the details of an individual case. Be careful not to make changes specifically for rare cases that could have a negative effect on the most common matches in the system.

SSG

How can this technology help you?

1) **Stand alone tool can be plugged into any database for analysis**

- More advanced training: incorporating data from other IIS registries and Public Health projects to create a larger training set will make the tool more detailed and accurate.
- AI Technology likes larger data sets, the more data the better!!

2) **Full function deduplication engine - used for real time matching**

1) We are now using the Lantern engine built into SSG's case management system, Casetivity, for the following organizations:

   a) MA Childhood Lead Poisoning Prevention Program
      - 2 million client records

   b) MA Early Intervention Program
      - System launching 2019

Thank You AIRA Attendees!

Sam Belton
Strategic Solutions Group
sbelton@ssg-llc.com
MIIS IT Project Team
Production Support