Administration for Children's Services (ACS)

Accelerated Safety Analysis Protocol (ASAP) Tool

First Used: May 2018

Computation Type: Ranking	**Purpose Type**: Performance evaluation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

Predictions of Severe Harm (identifying likelihood of substantiated allegations of physical or sex abuse within the next 24 months) are based on a machine learning methodology and are calculated for all children involved in active investigations early in the investigation (day 10). An investigation is assigned a numeric likelihood of this outcome based on the child in the case with the highest likelihood. The ACS Quality Assurance unit in the Division of Child Protection reviews about 3,000 active investigations annually, selecting those with the highest likelihood of severe harm.

Tool Purpose

The Quality Assurance Unit in the Division of Child Protection at ACS has the capacity to review about 3,000 investigation cases out of about 50,000 investigations annually. ACS developed this predictive model to support the selection of cases for Quality Assurance review. Open investigations involving children with the greatest likelihood to experience future severe harm – substantiated allegations of physical or sex abuse in the following 24 months – are selected for review. The tool does not support decisions about services or interventions for individuals or families involved with ACS, beyond the selection of the case for this additional Quality Assurance review.

If the Quality Assurance review team identifies gaps in routine, required documentation or practice, the team speaks with the field office conducting the investigation and follows up to make certain these gaps have been addressed. Scores are not shared with staff in the Quality Assurance unit or the investigative unit. The model only supports the decision about which investigations are prioritized for review by the Quality Assurance (QA) unit.

Populations Impacted

Individuals

Data Analyzed

Training Data	ACS trained the model on ACS historic administrative data about closed investigations from April 2014 to April 2016. The training set included about 142,026 observations. The model was tested on closed investigations from April 2016 to April 2017 with 53,477 observations.
Input Data	Predictions are based on administrative data about prior and current child welfare involvement including investigations triggered by a New York State Central Register (SCR) call and time spent in foster care. Only ACS administrative data are used in the model.

Output Data	Rank ordered list of open investigation cases involving children with the highest likelihood to experience future severe harm, defined as substantiated allegations of physical or sex abuse in the following 24 months to be reviewed by a special QA Review Team.

Caseloads Projection

First Used: Jul 2024

Computation Type: Forecasting	**Purpose Type**: Resource allocation
Identifying Information: No	**Updated in 2024**: Tool was created in CY2024

Tool Description

The model estimates weekly caseloads by area for the next 12 months from the run date (day 14 of the prevention case).

The model is a time series model that forecasts caseloads Citywide. The forecast analysis was performed at a more granular geography including Borough and Zone levels. Borough level analysis worked well but the data at Zone level is noisy for data modeling purposes. The model accounts for seasonality, trend, and variation in data to accurately project future caseloads.

Staffing decisions are made based on projected caseloads, historical attrition rate, and current team size.

Tool Purpose

Caseloads projections help identify how busy certain neighborhoods are likely to be in the next 12 months to optimize workforce distribution across different Boroughs in the City. As social workers complete their training program, they are assigned to field offices. The right number of Child Protective Service (CPS) Social Workers needs to be assigned to each Borough based on the workload.

Populations Impacted

Geographic space

Data Analyzed

Training Data	The model was trained on caseloads from January 2021 to July 2023 and tested on caseloads from July 2023 to July 2024.
Input Data	Predictions are based on administrative data on investigations and FSU involvement were extracted from Connections (CNNX). A snapshot of all open investigations was taken at the start of the week (every Monday). This value was treated as the average caseload for the week. Only ACS administrative data are used in the model.
Output Data	List of Boroughs with open investigation cases involving children.

Housing Prioritization

First Used: Apr 2023

Computation Type: Ranking	**Purpose Type**: Resource allocation
Identifying Information: No	**Updated in 2024**: No

Tool Description

The model estimates a risk score for a child receiving prevention services whose family will apply and be eligible for a homeless shelter within 12 months from the start of service (day 14 of the prevention case).

The model helps predict the risk of application for homeless shelters among families receiving prevention services. With a limited number of vouchers available, the risk model helps ACS prioritize housing assistance for those families at greatest risk of becoming homeless.

The Service Provider meets with the family to conduct a qulitative assessment of the family's housing needs and vouchers are offered based on their findings. This is one of multiple ways in which ACS can identify families potentially eligible for housing assistance.

Tool Purpose

The City has allocated 100 housing vouchers to families receiving ACS Prevention Services. The shelter application model identifies the likelihood of a family in prevention services applying for homeless shelter within 12 months beyond the current prevention case. The model uses a machine learning methodology and is calculated for all children in a prevention case. ACS Prevention Services uses the results as one of several ways of identifying possible families that service providers can assist in applying for shelter.

Populations Impacted

Individuals

Data Analyzed

Training Data	ACS trained the model on ACS historic administrative data regarding preventive services started between 2014 and 2020. An 80/20 split of data to train on 80% and test on 20% ensuring that no family appears in both sets. The training set contains 140,242 observations between Jan 2014 and December 2020. The test set consisted of 34,508 observations between Jan 2014 and December 2020.
Input Data	Predictions are based on administrative data about prior and current child welfare involvement at the start of a case. This includes SCR investigations and time spent in foster care. Only ACS administrative data are used in the model.
Output Data	Rank ordered list of open prevention cases involving children whose families have the highest likelihood of applying for a homeless shelter within 12 months of starting a prevention service.

Prevention Score Card

First Used: Sep 2021

Computation Type: Ranking	**Purpose Type**: Performance evaluation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

Predictions of Repeat Maltreatment (identifying the likelihood of being involved in a future indicated investigation within the next 24 months at the start of service) are based on a machine learning methodology and are calculated for all children receiving prevention services from ACS prevention service providers.

Tool Purpose

The Repeat Maltreatment model is used to make predictions on day 10 from the start of the prevention case to assess the risk of the family at the beginning of the service. A prevention case is assigned a numeric likelihood of an indicated investigation based on a New York State Central Register (SCR) within 24 months from the start of a prevention service.

The prevention providers are assessed for their performance based on the service needs/risk levels of the families they've served during the previous fiscal year.

The programs were sorted and ranked based on their average risk, and then divided into four quartiles by rank order: the top 25 percent of programs are classified as the Very High-Risk Cohort, the next 25 percent of programs as the High-Risk Cohort, the next 25 percent as the Medium-Risk Cohort, and the lowest 25 percent as the Low-Risk Cohort. Assignment to a cohort is not a way of performance assessment of the program but to group prevention service providers for fair comparisons based on the risk level of families they served.

Populations Impacted

Individuals

Data Analyzed

Training Data	ACS trained the model on ACS historic administrative data about closed investigations from July 2009 to June 2016. Training set included about 158,787 observations. The model was tested on closed investigations from July 2016 to June 2018 with 46,969 observations.
Input Data	Predictions are based on administrative data about prior and current child welfare involvement at the start of a case. This includes SCR investigations and time spent in foster care. Only ACS administrative data are used in the model.
Output Data	The model is used for generating a scorecard of prevention service providers by categorizing prevention programs based on the average risk profile of the cases they served during the assessment year. These groupings of program cohorts provide context for understanding the Scorecard, as it allows for performance comparison of programs that accepted and served families with similar risk profiles.

Department of Education (DOE)

Eureka! Chatbot

First Used: Aug 2023

Computation Type:	**Purpose Type**: NA
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

The Azure Cognitive Services technology and chatbot (internally branded as "Eureka!" has been configured and deployed in August 2023 to be the first response to calls to the NYCPS IT Service Desk. It accesses scripts to handle four common reasons for a user to call or contact the service desk - Password Reset, Create a Ticket, Ticket Status, Request for Information. The chatbot accesses pre-defined scripts to respond to user voice or text input. The user's request is either serviced, completed and closed by the chatbot, or the user is given the option (at any time) to connect to a live agent.

Tool Purpose

The tool is used to respond to common IT service desk requests - Password Reset, Create a Ticket, Ticket Status, Request for Information. Users can access the tool by phone, by computer through the DOE Support Hub application, and from links from MS Teams and other DOE systems, such as TeachHub.

Populations Impacted

Individuals

Data Analyzed

Training Data	Pre-defined scripts designed to respond to four common requests to the IT Service Desk.
Input Data	A voice call or text-based chat session initiated by a user and responded to by the Eureka! chatbot before being handled by a human Service Desk agent.
Output Data	The chatbot generates responses to user-entered prompts based on the training data, or forwards the inquiry to a human Service Desk agent. Since its launch in August 2022, the chatbot handles an average of 1,500 calls and 300 web-based inquiries each day. Approximately 30 percent of the voice calls and 10 percent of the web-based queries have been handled completely by Eureka! without being forwarded to a human Service Desk agent.

Vendor Involvement

Vendor Name: Nagarro and Microsoft

Developed by an IT services vendor (Nagarro) using Microsoft Cognitive services.

MySchools

First Used: Aug 2018

Computation Type:	**Purpose Type**: NA
Identifying Information: Yes	**Updated in 2024**: Yes

Tool Description

The tool utilizes the Gale-Shapley deferred acceptance algorithm to match applicants to schools. This algorithm has been in existence for many years, used internationally for various purposes. Perhaps most common is its use in the National Resident Matching Program for medical school students.

Deferred acceptance works as an iterative series of steps: students and programs are tentatively matched in each step, but nothing is finalized until the algorithm terminates (hence the deferred). 1. Each student "proposes" to their first choice • Programs assign seats to students one at a time • When all seats are filled, programs may reject previously accepted students in favor of new applications from students they prefer (e.g., students with a better lottery number) • Remaining students are rejected 2. Students rejected in the last step "propose" to the next choice on their list 3. The algorithm terminates when all students are matched or have proposed to all the programs they listed

Tool Purpose

MySchools is an application used to house online school directories, collect application choices, and run the admissions matching algorithm that is used for all centralized admissions processes (3K, pre-K, Gifted & Talented, middle school, and high school). The tool encompasses a family-facing portal, a school-facing portal, and an administrative portal.

Populations Impacted

Individuals

Individuals impacted include: Students

Data Analyzed

Training Data	The algorithm was already widely recognized for its advantages prior to adoption in New York City. The DOE consulted with a team of researchers at MIT who had been closely involved in its initial creation when we adopted it.
Input Data	Student biographical information (e.g., home address, poverty status, home language), student academic information (e.g., course grades, state test scores), and student school records (e.g., sending school).
Output Data	The algorithm outputs a school match for each student.

Vendor Involvement

Vendor Name: Blenderbox

We have a five year contract with the agency Blenderbox who designed the application and implemented the algorithmic matching functionality. The work is meant to transition to be run in-house, by the Division of Instructional and Information Technology (DIIT) within the Department of Education, by the end of the contract. The team at DIIT has already begun to takeover maintenance and development of the tool.

Update Description

A feature was added to determine "Probability of Acceptance at a specific school" for a future high school student.

NYCDOE APPR Measures of Student Learning (MOSL) Growth Model

First Used: Sep 2013

Computation Type:	**Purpose Type**: NA
Identifying Information: No	**Updated in 2024**: No

Tool Description

The growth model uses a variety of student-level (assessment scores, English Language Learner, Disability, and Economic Disadvantage indicators), classroom-level (e.g. percent Students With Disabilities), and school-level data (e.g. percent English Language Learners, percent Students With Disability, average prior achievement, school type) to estimate/predict a student's score on one of many possible course-culminating assessments. These predicted scores are used to either 1) identify "peer groups" of students, from which student growth percentiles (SGPs) are determined, or 2) compared to actual scores to determine student credit values. These units (SGPs or credit values) are then weight-averaged to generate an educator-level result - the MOSL Rating. The MOSL Rating is combined with the MOTP Rating to produce an Overall Rating. Per state law 3012-d, annual ratings "shall be a significant factor in HR decisions." This is often implemented by making ratings a qualifying/disqualifying element in decision-making concerning employment, tenure, salary, and other professional opportunities.

Tool Purpose

In accordance with New York state law and New York State Education Department (NYSED) regulations, the Department developed and maintains a "growth model" to produce Measures of Student Learning (MOSL) ratings for use in annual professional performance reviews (APPR) for teachers and principals. The MOSL ratings are combined with Measures of Teaching/Leadership Practice (MOTP/MOLP) ratings to produce an annual Overall Rating for each eligible educator.

Populations Impacted

Individuals

Data Analyzed

Training Data	The growth model process is employed in both retrospective and prospective ways. In the retrospective version, the results are determined entirely within-sample. In the prospective version, the coefficients of the model are estimated on multiple prior years of data.
Input Data	The growth model makes use of three types of data: (1) students' end-of-year assessment scores, (2) enrollment and attendance records that link students to teachers and schools, and (3) historical academic and demographic information used to identify groups of similar students.
Output Data	The model outputs an estimate of a student's score on a course-culminating assessment.

Vendor Involvement

Vendor Name: Education Analytics

Education Analytics provides technical assistance and quality assurance for the growth model.

NYCDOE APPR Measures of Teaching/Leadership Practice (MOTP/MOLP) Calculation

First Used: Oct 2013

Computation Type:	**Purpose Type**: NA
Identifying Information: No	**Updated in 2024**: No

Tool Description

Throughout a school year, evaluators observe teachers/principals multiple times and use a rubric to provide a numerical rating on one or more rubric components. These rubric component scores are then weight-averaged according to collectively bargained rules to produce an MOTP/MOLP Rating. The MOTP/MOLP Rating is combined with the MOSL Rating to produce an Overall Rating for each eligible educator. Per state law 3012-d, annual ratings "shall be a significant factor in HR decisions." This is often implemented by making ratings a qualifying/disqualifying element in decision-making concerning employment, tenure, salary, and other professional opportunities.

Tool Purpose

In accordance with New York state law and New York State Education Department (NYSED) regulations, the Department developed and maintains databases and calculation rules to produce Measures of Teaching/Leadership Practice (MOTP/MOLP) ratings for use in annual professional performance reviews (APPR) for teachers and principals. The MOTP/MOLP ratings are combined with Measures ofStudent Learning (MOSL) ratings to produce an annual Overall Rating for each eligible educator.

Populations Impacted

Individuals

Data Analyzed

Training Data	Pilot data prior to program launch was used to inform the weights assigned to various rubric components. However, the weights are ultimately determined via collective bargaining.
Input Data	Rubric component numerical ratings.
Output Data	The model outputs a score for teachers and principals.

Open Gen AI and Teaching Assistant Tool

First Used: May 2023

Computation Type:	**Purpose Type**: NA
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

The generative AI system using large language models was a system custom-built by DIIT using advanced Microsoft technologies to create a set of generative AI tools. To date, two tools have been built. One is named "Open Gen AI" - it accesses a large language model (currently GPT 3.5) to provide responses to a broad range of prompts. The other is named "Teaching Assistant for Algebra" - it accesses specific Algebra-focused content to provide responses to prompts related to Algebra.

Tool Purpose

The tool is used to generate responses to prompts entered by a student or teacher, requesting the generative AI tool to compose a text response to a text input.

Populations Impacted

Individuals

Data Analyzed

Training Data	For the Open Gen AI tool, the ChatGPT large language model is used as the "trained data". For the Teaching Assistant for Algebra tool, the LLM has been trained exclusively on curriculum from Illustrative Math.
Input Data	Prompts provided by the users of the system.
Output Data	The output data for the Open Gen AI tool is the response generated by the ChatGPT LLM. The output data for the Teaching Assistant is the response generated by specifically developed LLM using the Illustrative Math curriculum.

Vendor Involvement

Vendor Name: Microsoft

Microsoft provided technical guidance for their emerging generative AI technology and built some small module of code for the specific NYCPS Gen AI and Teaching Assistant use cases.

Department of Environmental Protection (DEP)

Idling Complaints Program

First Used: Aug 2022

Computation Type:	**Purpose Type**: NA
Identifying Information: No	**Updated in 2024**:

Tool Description

A contractor helped create an AI tool that analyzes the audio and visual aspects of pictures and videos submitted by citizens of alleged car idling complaint occurrences that are in violation of New York City air pollution laws.

Tool Purpose

The analysis from the tool makes a recommendation to staff reviewers whether the submitted evidence support an occurrence of car idling in violation of New York City laws. The tool also provides a level of confidence in its recommendation. The tool does not make the review decision in the Idling Complaints system. It is still entirely up to the staff to decide whether to take the tool's recommendation or not.

Populations Impacted

Individuals

Data Analyzed

Training Data	Videos and pictures of cars idling submitted by citizens, along with staff decisions on whether the picture/video constituted as an idling violation.
Input Data	Videos and pictures submitted by citizens through our web portal.
Output Data	Recommendation, confidence level, description of its decision from the tool.

Vendor Involvement

Vendor Name: Acuvate

Acuvate developed the AI tool that performs the automated analysis of the submitted evidence.

Department of Health and Mental Hygiene (DOHMH)

Bowtie2

First Used: Jun 2022

Computation Type: Matching	**Purpose Type**: Risk management
Identifying Information: No	**Updated in 2024**: No

Tool Description

Aligns sequencing data to a reference sequence. Bowtie2 aligns sequencing data to a reference using Burrows-Wheeler transformations. It is geared to use with Illumina sequencing data.

Tool Purpose

Bowtie2 is an intermediate step in the workflow to analyze COVID variants in wastewater.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genomes are sequenced.

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	Sequence reads (fastq) for single or paired-end runs (sequence reads can be considered strings).
Output Data	Aligned reads in SAM format

BWA

First Used: Jul 2017

Computation Type: Matching	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: No

Tool Description

Aligns sequencing data to a reference sequence.

Tool Purpose

Burrows-Wheeler Aligner (BWA) is aligning sequence data to reference using Burrows-Wheeler transformations. This tool is optimal for low-divergent genomic data and short read data; such as Illumina sequence data. This tool is used to predict the order in which the fragments generated by sequencers are pieced together to form a complete genomic sequence data. This tool is used for Legionella and PulseNet sequencing analyses.

Populations Impacted

Individuals; Other

Individuals impacted include: Individuals whose sequence data is being analyzed

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	Sequence reads (fastq) for single or paired-end runs (sequence reads can be considered strings).
Output Data	Aligned reads in SAM format

ChoiceMaker (CM)

First Used: Jun 2003

Computation Type: Matching	**Purpose Type**: Data management
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

ChoiceMaker (CM) is a record-matching tool that identifies duplicate records belonging to the same individual.

Tool Purpose

CM is used by BOI and Healthy Homes to identify duplicate immunization and lead records. The outputs produced by CM are used in ongoing manual and automated deduplication processes (record merging).

Populations Impacted

Individuals

Individuals impacted include: Individuals of all ages with a record in the CIR.

Data Analyzed

Training Data	The CM model was trained on human decisions.
Input Data	CM uses demographic data (e.g.; names; date of birth; address; identifiers) and health event data (e.g.; date and type of event) from BOI's Citywide Immunization Registry (CIR) and Healthy Homes' LeadQuest registry in its evaluation.
Output Data	The program outputs a series of record pairs and a match probability for each pair.

Vendor Involvement

Vendor Name: HLN Consulting

A vendor was involved in the development of the program initially. CM is now available as an open-source program. The DOHMH implementation is maintained by HLN Consulting.

GATK

First Used: Oct 2017

Computation Type: Classification	**Purpose Type**: Risk management
Identifying Information: No	**Updated in 2024**: No

Tool Description

A suite of tools for variant calling and filtering after sequence alignment. It uses naive Bayesian to qualify aligned bases as sequence or erroneous data; which would be excluded from the final genomic sequence.

Tool Purpose

Used to identify mutations and call upon differences from the reference; which is used to generate the predicted complete sequence.

Populations Impacted

Individuals; Other

Individuals impacted include: Pathogens whose genomes are sequenced

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	Sets of known variant sites.
Input Data	Fasta; uBam; SAM/BAM/CRAM; VCF.
Output Data	Bam; txt; vcf.

Guppy

First Used: Jun 2020

Computation Type: Classification	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: No

Tool Description

Converts electric signals to predict a nucleotide and enables filtering of low-quality calls.

Tool Purpose

This is a tool designed specifically for Oxford Nanopore Technology (ONT) data. This is a neural network based basecaller; a tool that determines nucleotide bases of a genetic material; that converts electric signals into strings to represent genomic data. In addition to basecalling; the tool also performs filtering of low-quality reads; a stretch of sequenced genetic material. This is the initial step that converts electric signals to fragments of sequence data; which can then be used for COVID-19 sequencing analysis.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genomes are sequenced.

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	The default models within Guppy are trained on a mixture of native and amplified DNA/RNA; from multiple organisms including plant; animal; bacterial and viral genomes.
Input Data	DNA/RNA strand passing through the nanopore. Raw data is stored as .fast5 files
Output Data	.fast5 files; fastq; or BAM files.

Vendor Involvement

Vendor Name: Oxford Nanopore Technologies

Developed and maintains the tool.

ICE - Immunization Calculation Engine

First Used: 1997

Computation Type: Forecasting	**Purpose Type**: Information presentation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

The Immunization Calculation Engine (ICE) is an immunization evaluation and forecasting system; whose default immunization schedule supports all routine childhood; adolescent; and adult immunizations based on the recommendations of the Advisory Committee on Immunization Practices (ACIP). ICE is free and open-source available through https://cdsframework.atlassian.net/wiki/spaces/ICE/overview.

Tool Purpose

ICE is used by the Bureau of Immunization to evaluate a patient's immunization history and generate appropriate immunization recommendations.

Populations Impacted

Individuals

Individuals impacted include: Anyone who needs a vaccine

Data Analyzed

Training Data	N/A
Input Data	ICE uses demographic data (e.g. date of birth) and vaccination data (e.g.; immunization date; vaccine group and type) in the evaluation process. Data used are stored in the Citywide Immunization Registry (CIR).
Output Data	The program returns recommendations on whether a patient has completed a vaccine series or is due for vaccines.

Vendor Involvement

Vendor Name: HLN Consulting

A vendor was involved in the development of the program and continues to be involved in program enhancements. ICE is also available as an open-source program. The DOHMH implementation is maintained by HLN Consulting.

Improving Foodborne Disease Outbreak Detection by Incorporating Complaints Identified in Social Media Data

First Used: Nov 2016

Computation Type: Scoring	**Purpose Type**: Triage
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

Restaurant associated foodborne disease outbreaks are often identified through complaints received via New York City's 311 non-emergency information system; however not all individuals report to 311. The New York City Department of Health and Mental Hygiene (NYC DOHMH) in collaboration with Columbia University developed a text classifier program which monitors Yelp and Twitter data to identify complaints of foodborne illness which was supported by grants from the Alfred P Sloan Foundation and the National Science Foundation. As of April 2023; the tool no longer uses data from Twitter/X due to API changes.

Tool Purpose

The model uses data from Yelp restaurant reviews and previously used Twitter data that was available on Twitter's publicly available API. Twitter (X) removed free access to their publicly available API in April 2023; so these data are no longer included in our analyses. The classifiers assign a "sick score" to each Yelp review indicating the likelihood that the review pertains to foodborne illness. The sick score is based on whether the review contains key words indicative of foodborne illness ("e.g. vomit"); the Yelp classifier also incorporates if the review indicates that multiple people became sick and if the review indicates a time between eating at a restaurant and illness onset (incubation period) that is consistent with foodborne illness. Each review with a sick score greater than or equal to a threshold value are reviewed and annotated by the NYC Health Department's foodborne disease epidemiology and environmental health staff to determine if the review was actually reporting foodborne illness possibly associated with a New York City restaurant; if yes; Yelp messages are sent to Yelp reviewers; requesting that they contact the NYC Health Department. Data from annotations are used to improve classifier performance. Foodborne disease complaints identified through Yelp are combined with foodborne disease complaints reported to 311 to improve efficiency of outbreak detection.

Populations Impacted

Individuals

Individuals impacted include: 1) Public who dine at NYC restaurants and are Yelp users. 2) NYC Restaurants.

Data Analyzed

Training Data	Training data was used in the development of the Yelp classifiers. The training data consisted of restaurant reviews obtained from Yelp by Columbia University; a subset of these data were joined with annotations provided by NYC Health Department staff. The annotations of restaurant reviews focused on the following: 1) if the review indicated foodborne illness; 2) if the incident occurred in the past 30 days; 3) if multiple people were sick and 4) if the incubation period was consistent with foodborne illness. The training data is periodically updated (with annotations from the NYC Health Department) to improve the classifiers.
Input Data	Yelp reviews of New York City restaurants are pulled from a privately available application programming interface (API) provided by Yelp.
Output Data	The output data includes a "sick score" that the classifiers assign to each Yelp review indicating the likelihood that the review pertains to foodborne illness.

Vendor Involvement

Vendor Name: Columbia University

New York City Health Department staff; including Bureau of Communicable Disease; Office of Environmental Investigations; and Division of Informatics and Information Technology & Telecommunications and Columbia University are involved in making decisions about the tool. Columbia University Department of Computer Science professors and doctoral students maintain the classifier. The project was previously funded by the Alfred P Sloan Grant; for which The Fund for Public Health in New York provided administrative support and grant management to the New York City Health Department. This support and management ended at the completion of the grant in 2021.

IQTREE

First Used: May 2020

Computation Type: Clustering	**Purpose Type**: Risk management
Identifying Information: No	**Updated in 2024**: No

Tool Description

IQTREE uses maximum-likelihood regression to create phylogenetic trees from genomes.

Tool Purpose

Produced phylogenetic trees are used to help rule in or out outbreaks of COVID or other organisms.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genomes are sequenced.

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	FASTA; NEXUS; CLUSTALW; PHYLIP.
Output Data	Readable report; ML tree in NEWICH format; log file for entire run.

kSNP4

First Used: Mar 2022

Computation Type: Clustering	**Purpose Type**: Risk management
Identifying Information: No	**Updated in 2024**: Yes

Tool Description

kSNP3 can use multiple algorithms (maximum-likelihood; parsimony; neighbor-joining) to infer phylogenetic trees from genomes.

Tool Purpose

Produced phylogenetic trees are used to help rule in or out outbreaks of bacteria.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genomes are sequenced.

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	Fasta
Output Data	ML tree in NEWICH format; log & configuration files; Fasta file

Update Description

it used to be kSNP3, and now it got updated to kSNP4

MAFFT

First Used: Jan 2021

Computation Type: Matching	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: No

Tool Description

Aligns multiple sequencing data.

Tool Purpose

MAFFT (for Multiple Alignment using Fast Fourier Transform) includes several algorithmic methods; including guided tree; scoring matrices; and sequence alignment algorithms to realign multiple genomic sequencing data. The realignment tool is used to locally re-arrange sequence data to make all sequences comparable but the same genomic coordinates. This is used in all sequencing analysis prior to building a phylogenetic tree or distance tree.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genomes are sequenced.

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	Sequences can be in GCG; FASTA; EMBL (Nucleotide only); GenBank; PIR; NBRF; PHYLIP or UniProtKB/Swiss-Prot (Protein only) format
Output Data	Fasta or Clustalw

Minimap2

First Used: May 2020

Computation Type: Matching	**Purpose Type**: Data management

Identifying Information: No	**Updated in 2024**: No

Tool Description

Aligns sequencing data to a reference sequence.

Tool Purpose

Minimap2 uses optimal chaining scores to align sequencing data to reference genomes. This tool is faster and more optimal for long read sequences; such as Oxford Nanopore Technologies (ONT) data. This tool is used to predict the order in which the fragments generated by sequencers are pieced together to form a complete genomic sequence data. This tool is used for COVID-19 and monkeypox virus (MPXV) sequencing analyses.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genomes are sequenced.

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	Sequence reads (fastq) for single or paired-end runs (sequence reads can be considered strings).
Output Data	Aligned reads in SAM format

Pangolin

First Used: Jul 2021

Computation Type: Clustering	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: Yes

Tool Description

Assigns lineage names to SARS-CoV-2.

Tool Purpose

Pangolin uses a combination of several methods; including random forest tree; classification methods; and maximum parsimony to assign lineage names to SARS-CoV-2 genomic sequences to bin sequences that are more likely to be similar. This is a tool that designates a name based on a nomenclature for COVID-19 sequence data.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genome is sequenced

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	Trained on a data set of genomes that have been designated to Pango lineages using whole genome information.
Input Data	Fasta files.
Output Data	.csv fiile with taxon name and lineage assigned.

Update Description

Pangolin is updated as new sequences are published, and as the virus evolves. Currently, it is version 4.3

PHYLOViZ

First Used: Oct 2017

Computation Type: Clustering	**Purpose Type**: Risk management
Identifying Information: No	**Updated in 2024**: No

Tool Description

For representing the possible evolutionary relationships between strains; PHYLOViZ uses the goeBURST algorithm; a refinement of eBURST algorithm by Feil et al.; and its expansion to generate a complete minimum spanning tree (MST).

Tool Purpose

Used to generate the minimum spanning tree relationships.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genomes are sequenced.

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	txt; NEWICK; FASTA.
Output Data	Minimmum spanning tree.

PulseNet 2.0

First Used: Sep 2017

Computation Type: Clustering	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: Yes

Tool Description

A suite of tools used to align and analyze bacterial genomes.

Tool Purpose

PulseNet 2.0 is used to 1) assemble the bacterial genome (since the sequencing process involves fragmenting the bacterial DNA and then amplifying it into millions of pieces) 2) identify the genus; species; and serotype of the bacterial isolate 3) perform quality control checks to ensure the sequence meets certain quality standards 4) perform core and whole genome multi-locus sequence typing (cg/wgMLST; a technique used to type bacteria based on their genetic code) 5) perform cluster analysis for cases related to one another based upon case definitions recommended by the Centers for Disease Control and Prevention (CDC). This information is then communicated to partners including foodborne epidemiologists at the Bureau of Communicable Disease; who investigate all reported cases of foodborne disease; with those investigations potentially resulting in restaurant inspections; closures; and food recalls.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens in lab test samples that are sent for sequencing.

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	Fastq.
Output Data	txt; Excel.

Vendor Involvement

Vendor Name: CDC

Developed and maintains the tool.

Update Description

was put on the cloud and accessed through CDC SAMS site

Spades

First Used: Oct 2017

Computation Type: Matching	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: No

Tool Description

Spades uses several algorithms to simplify genomic read data into de Brujin graphs and finds overlaps to assemble genomes.

Tool Purpose

Spades is an intermediate step in the workflows of bacterial analyses.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genomes are sequenced.

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	Fastq.
Output Data	Fastas and other files for corrected reads; scaffolds; contigs; paths in GFA format; fastg assembly graph.

Vsearch

First Used: Jun 2022

Computation Type: Matching	**Purpose Type**: Risk management
Identifying Information: No	**Updated in 2024**: No

Tool Description

Vsearch uses the Needleman-Wunsch algorithm to merge read pairs and align and dereplicate sequences to detect chimeric genomic sequences.

Tool Purpose

Vsearch is an intermediate step in the workflow to analyze COVID variants in wastewater.

Populations Impacted

Individuals; Other; Biological sample

Individuals impacted include: Pathogens whose genomes are sequenced

Others impacted include: Sequence data can belong to any species

Data Analyzed

Training Data	N/A
Input Data	Sequence reads (fastq; Fasta) for single or paired-end runs (sequence reads can be considered strings).
Output Data	FASTA; FASTQ; tables; alignments; SAM.

Department of Investigation (DOI)

Facial Recognition Technology

First Used: Mar 2019

Computation Type: Matching	**Purpose Type**: NA
Identifying Information: Yes	**Updated in 2024**: Yes

Tool Description

The tool analyzes an uploaded image or video and searches and compares it with lawfully possessed images to generate a pool of possible matches. If possible matches are identified, a trained DOI examiner visually analyzes and evaluates potential matches to assess reliability of a match consistent with agency policy and applicable laws. A match serves as an investigative lead for additional investigative steps and does not constitute a positive identification.

Tool Purpose

Facial recognition is a digital technology that DOI uses to analyze uploaded images or videos of people and objects obtained during an investigation by comparison with lawfully possessed images. Facial recognition generates possible matches of an object or individual from this analysis and comparison. The purpose of the tool is to assist DOI investigations of matters within its jurisdiction including fraud and other criminal activity.

Populations Impacted

Individuals

Data Analyzed

Training Data	Vendor uses publicly available open source media data.
Input Data	Images.
Output Data	Images.

Vendor Involvement

Vendor Name: Not disclosable

Out-of-the-box products. The vendors provide ongoing technical assistance. Confidentiality agreements are in place with the vendors.

Update Description

General system updates by vendor to improve quality of search results and to fix bugs.

Department of Records and Information Services (DORIS)

Records365 from Record Point

First Used: Jan 2022

Computation Type: Classification	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: No

Tool Description

Records365 is a SAAS tool for managing agency's electronic records' lifecycle from creation to disposition.

Tool Purpose

The machine learning function trains the documents managed in the system to automatically categorize the record series and triggers the proper record retention.

Populations Impacted

Other

Others impacted include: agency documents

Data Analyzed

Training Data	Various types of documents that are managed in the system.
Input Data	Various types of documents that are managed in the system.
Output Data	record categorization of each trained document.

Vendor Involvement

Vendor Name: Record Point

They helped us with the implementation of the module including training and support.

Department of Social Services (DSS)

Homebase Risk Assessment Questionnaire (RAQ)

First Used: Jun 2012

Computation Type: Scoring	**Purpose Type**: Resource allocation
Identifying Information: Yes	**Updated in 2024**: Yes

Tool Description

Homebase applicants answer screening questions about their current housing situation, history of disruptive experiences, shelter history, and other domains. Each of the answers is assigned a number of points, and applicants that reach a certain point threshold are eligible for deeper Homebase services, such as financial assistance and case management. Workers are able to override a limited number of model decisions with permission of a supervisor.

The Homebase program was created to prevent households from entering the DHS shelter system. Since NYC has a range of antipoverty programs and the number of households entering shelter is small compared to the pool of New Yorkers who have an eviction filing each year, the Agency had to ensure that the households who most needed additional homelessness prevention services were being enrolled in Homebase programs. Research showed that staff were not accurately able to predict who would or would not enter the DHS shelter system and that using a risk assessment would provide a better way to match resources to the families who would benefit the most.

Tool Purpose

The Homebase program was created to prevent households from entering the DHS shelter system. Since NYC has a range of antipoverty programs and the number of households entering shelter is small compared to the pool of New Yorkers who have an eviction filing each year, the Agency had to ensure that the households who most needed additional homelessness prevention services were being enrolled in Homebase programs. Research showed that staff were not accurately able to predict who would or would not enter the DHS shelter system and that using a risk assessment would provide a better way to match resources to the families who would benefit the most.

Populations Impacted

Individuals

Individuals impacted include: Households seeking Homebase assistance

Data Analyzed

Training Data	The RAQ was developed based on analysis of data on Homebase enrollees from 2004 to 2008, conducted in conjunction with a team of academic researchers, to determine predictive factors for those entering shelter. It was updated in 2023 based on analysis led by NYC DSS researchers of 2013-2016 Homebase data.
Input Data	Factors include, among others: personal characteristics such as age and pregnancy; educational attainment and employment status; housing issues such as eviction, discord, a move in the past year; past and recent experience of homelessness.
Output Data	The tool produces a score that is used to assess eligibility for full versus brief Homebase services.

Vendor Involvement

Vendor Name: Multiple researchers

DHS contracted with researchers to evaluate years of Homebase administrative data to develop a risk assessment. The DSS research team then led an updated analysis that led to tool revisions. The published research papers are listed below: https://ajph.aphapublications.org/doi/10.2105/AJPH.2013.301468 https://www.journals.uchicago.edu/doi/abs/10.1086/686466?mobileUi=0&journalCode=ssr https://www.tandfonline.com/doi/abs/10.1080/10511482.2022.2077801

Update Description

The RAQ was developed based on analysis of data on Homebase enrollees from 2004 to 2008, conducted in conjunction with a team of academic researchers, to determine predictive factors for those entering shelter. It was updated in 2023 based on analysis led by NYC DSS researchers of 2013-2016 Homebase data.

NYC Enterprise Data Solutions Service (EDS)

First Used: Mar 2024

Computation Type: Matching	**Purpose Type**: Data management
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

The NYC Enterprise Data Solutions Service (EDS) automatically matches and links person records from multiple source systems, regardless if they share a common identifier such as a social security number. EDS supports cross-agency data integration and interoperability, particularly across health and human service agencies. EDS works by processing business rules that standardize name and geocode address data before applying a series of deterministic and probabilistic matching rules that generate detailed or summarized reports of record linkages.

Tool Purpose

In 2024, PEU moved the operations of its Rent Freeze team (focused on enrollment in SCRIE, DRIE, SCHE, DHE) to a new case/client management system within our custom Salesforce instance. Prior to migrating data from the prior system (based in our EveryAction tool), PEU needed to identify duplicate clients within EveryAction and between EveryAction and Salesforce. The primary reason for duplication is that we use Salesforce to support a number of other teams that might have interacted with Rent Freeze clients as well. We used the EDS tool developed by our partners at NYC Opportunity and maintained by OTI to identify client records that were highly likely to be the same client. PEU developed its own process to combine outreach and case history information for the records identified as related to the same client to create a single record in PEU's Salesforce system across programs. Using the EDS tool helped us provide better service to our clients by providing our specialists with richer information about the client's interaction with the Rent Freeze, Tenant Helpline, Tenant Support Unit, and GetCovered teams.

Populations Impacted

Individuals

Data Analyzed

Training Data	n/a
Input Data	Client contact information, including name, phone numbers, email addresses, home addresses
Output Data	The same client information but with common identifier numbers

SmartVAN / TargetSmart

First Used: Nov 2019

Computation Type: Matching	**Purpose Type**: Data management
Identifying Information: Yes	**Updated in 2024**: Yes

Tool Description

The Mayor's Public Engagement Unit (PEU) uses SmartVAN to manage outreach across a range of projects. SmartVAN provides functionality to create lists of potential clients to contact, collect personal information and survey responses from clients, and conduct outreach via phone banks and canvassing. SmartVAN also contains a frequently updated commercial dataset, provided by TargetSmart, of New York City residents and their demographic, contact, and other information. PEU uses this preloaded data to create outreach lists when data on existing clients or from partner agencies is unavailable or insufficient to meet the scope of the outreach project.

Tool Purpose

In 2023, PEU has the TargetSmart data within SmartVAN on a number of projects. PEU frequently uses the data to create lists of residents who live within certain zip codes that PEU wants to target for outreach. For example, PEU created lists using TargetSmart data to conduct door-knocking and phone banking outreach to New Yorkers identified as potentially eligible for the DOF Rent Freeze program based on TargetSmart data. In cases like these, TargetSmart's determination of who lives in which zip codes as well as estimated income affects whether New Yorkers receive PEU outreach. Additionally, the algorithm that TargetSmart uses to match phone numbers to individuals impacts the type of outreach that New Yorkers receive.

Populations Impacted

Individuals

Data Analyzed

Training Data	Training data is part of vendor's proprietary processes.
Input Data	Input data is part of vendor's proprietary processes.
Output Data	The algorithmically-derived data that PEU accesses is the output of proprietary algorithmic processes developed and operated by TargetSmart. These algorithmic processes include matching multiple input datasets to determine residency, contact information, and demographics on New York City residents. SmartVAN also includes a number of algorithmically-determined likelihood scores, including scores for the likelihood that a household contains children under 18, etc.

Vendor Involvement

Vendor Name: EveryAction and TargetSmart

EveryAction and TargetSmart jointly provide the SmartVAN product. EveryAction is the software provider. TargetSmart is the data provider. TargetSmart is the entity who applies algorithmic techniques. EveryAction provides access to this data through their platform.

Update Description

We get regular updated lists of New Yorkers from the vendor directly into the EveryAction platform we use for outreach.

Department of Transportation (DOT)

Midtown in Motion / ACDSS

First Used: Jul 2011

Computation Type: Forecasting	**Purpose Type**: Performance evaluation
Identifying Information: No	**Updated in 2024**: Yes

Tool Description

The "Midtown in Motion" (MIM) program is used to enhance multimodal mobility and mitigate the developing levels of congestion in the Midtown Core of Manhattan from 1st to 9th Avenues, from 57th to 34th Streets, inclusive. The method is to rebalance the traffic approaching the Midtown Core by actively changing the signal timing along the avenues outside the zone, based on the calculated Median Travel Times data collected from the NYCDOT RFID readers along the Manhattan arteries inside the core area.

Tool Purpose

Performance evaluation, risk management, triage of corridors and arteries in midtown Manhattan

Populations Impacted

Geographic space

Data Analyzed

Training Data	N/A
Input Data	pre-stored library of timing plans
Output Data	Median Travel Times

Vendor Involvement

Vendor Name: KLD Engineering , P.C

KLD Engineering, P.C is the developer of the Tool, and currently handle the maintenance contract

Update Description

Ongoing maintenance

Fire Department (FDNY)

EMD Schedule Optimization Tool

First Used: Jun 2021

Computation Type: Forecasting	**Purpose Type**: Resource allocation
Identifying Information: No	**Updated in 2024**: No

Tool Description

The purpose of the tool is to provide Emergency Medical Dispatchers (EMD) staff a tool to optimally allocate call takers during a 24-hour period. The tool uses an expected number of incoming calls and the number of personnel scheduled to work in order to allocate the call takers to different shifts such that the supply of call takers exceeds the demand for call takers.

Tool Purpose

The algorithm requires two datasets. First, the tool requires the average number of medical calls per hour for a 24-hour period. Second, the tool requires a user to specify the number of call takers assigned to each tour. Based on these two inputs, the tool provides a projection of supply (call takers) versus demand (medical calls). Additionally, the tool can take the total number of available staff and optimally allocate them across tours to maximize the minimum difference between supply and demand. Based on these outputs, EMD officers can identify times during the day when call taker utilization is high and reallocate staff to accommodate.

Populations Impacted

Other

Others impacted include: FDNY radio and assignment dispatcher employees

Data Analyzed

Training Data	This is an optimization model and was not "trained" using training data. The algorithm relies on actual historical data to determine average hourly medical calls.
Input Data	The tool requires an hourly count of medical calls arriving during a 24-hour period. Additional "data" requirements are input from the user depending on user-driven scenarios. For example, a user could specify five eight-hour tours per day (at different start times) rather than existing four tours (two eight-hour tours and two 12-hour tours).
Output Data	The algorithm outputs a projection of supply (call takers) versus demand (medical calls). Additionally, the tool can take the total number of available staff and optimally allocate them across tours to maximize the minimum difference between supply and demand.

EMS Hospital Suggestion Algorithm

First Used: Mar 2007

Computation Type: Ranking	**Purpose Type**: NA
Identifying Information: No	**Updated in 2024**: No

Tool Description

The EMS Hospital Suggestion Algorithm is used to determine the closest, appropriate hospital to the incident location based on the needs of a patient requiring transport.

Tool Purpose

The algorithm computes a list of hospitals in order of closest to furthest in time for each medical condition category as currently established. (For example, there is a list of hospitals computed in order of closest in time for all hospitals that accept General Emergency Department patients and for all hospitals that accept special conditions, such as burns). Depending on the medical needs category of the patient, the algorithm produces a pre-determined list of hospitals which is based on the location of the patient and then made available to the crew as a list of "closest, most appropriate hospitals."

Populations Impacted

Geographic space

Data Analyzed

Training Data	The EMS Hospital Suggestion algorithm relies on automatic vehicle location data from ambulances transporting patients to hospitals between 2018 and 2019 to calibrate a network analysis model that derives incident to hospital transport times. The order of suggested hospitals are then compared with five years of historical EMS hospital transport data from before the COVID-19 pandemic (2015-2019) to validate and correct the network model.
Input Data	The inputs for the algorithm include the location and medical condition of the patient.
Output Data	The algorithm outputs the closest, most appropriate hospitals.

EMS Unit Suggestion Algorithm

First Used: Mar 2007

Computation Type: Ranking	**Purpose Type**: Resource allocation
Identifying Information: No	**Updated in 2024**: No

Tool Description

The Emergency Medical Services (EMS) Unit Suggestion Algorithm is used to determine which order of geographic regions (known as atoms) to search in order for the EMSCAD system to select an appropriate EMS unit for dispatch to an incident.

Tool Purpose

The algorithm computes a list of geographic regions (known as atoms) in order of closest to furthest in travel time for each atom in the city. This list of ordered atoms is the output of an algorithm that relies on a calibrated network model to derive travel time estimates. The output is an excel file which is converted into an EMSCAD-compatible file and loaded into the system for real-time unit selection capabilities. The file is generated and implemented as a 24/7 source file, meaning, the recommended search order is not currently varying by time of day. The Department is intending to implement time-of day search orders in the near future.

Populations Impacted

Geographic space

Data Analyzed

Training Data	The EMS Unit Suggestion algorithm relies on historical FDNY CAD trip time data which is used to calibrate a network analysis model which derives atom-to-atom transport times.
Input Data	The input for the algorithm is a geographic location.
Output Data	The algorithm outputs a recommended EMS unit for dispatch.

Vendor Involvement

Vendor Name: Deccan International

This algorithm and the resulting output file that is used in our EMS CAD system to suggest atom order for unit search is currently provided by a vendor, Deccan International.

RBIS (Risk Based Inspection Program); ALARM (A learning Approach to Risk Modeling)

First Used: Nov 2019

Computation Type: Scoring	**Purpose Type**: Risk management
Identifying Information: No	**Updated in 2024**: Yes

Tool Description

A Learning Approach to Risk Modeling (ALARM) creates risk scores for each building in the city. These scores are used to schedule our Fire Operations building inspections within the inspectable population of buildings in the city (~330,000 Building Identification Numbers (BINs)), as a part of the Risk-Based Inspection Program (RBIS).

Tool Purpose

ALARM is a combined approach using machine learning and risk ratios to assess the risk of a building for structural fire ignition (probability) and civilian fire injury/death (impact). The machine learning algorithm takes incident data, housing characteristics, and 311 data and creates a probability of structural fire ignition. This is combined with a civilian injury or death risk ratio for the building which is based on building characteristics, incident data and nearby felony crimes to create a risk score (range is one to nine), with one being highest risk and nine being lowest risk. Buildings are prioritized within each of the nine risk scores according to the residential population in each building.

Populations Impacted

Property; Other

Others impacted include: Civilian Fire Injuries/Fatalities

Data Analyzed

Training Data	Each month the team uses a five-year incident dataset and reserves 99 percent of the data to train the ignition model and 80 percent of the data to train the impact model.
Input Data	The ALARM risk score utilizes data from our fire and Emergency Medical Services (EMS) dispatch system, building characteristic data, 311 calls, felony crimes, census data and civilian injury data.
Output Data	The tool outputs a risk score from one (highest risk) to nine (lowest risk).

Update Description

The fire ignition and injury/death models are recalculated monthly with fresh training data to create updated variable weights.

Mayor’s Office (MO)

Adobe Photoshop

First Used: Jan 2024

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

NYC Media uses Adobe Photoshop to edit images. Adobe Photoshop uses generative AI to allow users to edit images without manual work.

Tool Purpose

NYC Media uses Adobe Photoshop to make slight edits to some images that appear in some content that is produced in-house and broadcast on the City’s television edits. For example, to comply with FCC regulations for non-commercial educational stations, we may use an AI tool to blur a company’s logo on a t-shirt. As another example, we may use the AI tool to add visual interest, for example, to add legs in a picture that is cropped at the waist.

Populations Impacted

Individuals

Individuals impacted include: Viewers who watch the City’s television channels; some individuals who appear in MOME’s television content

Data Analyzed

Training Data	According to Adobe’s website, “generative AI Image models were trained on licensed content, such as Adobe Stock, and public domain content where copyright has expired.”
Input Data	Images copyrighted by the City of New York or licensed to the City of New York pursuant to an agreement that authorizes edits and, if involving images of people, content that is covered by a written consent form.
Output Data	Visual content that is broadcast on the City’s television stations.

Vendor Involvement

Vendor Name: Adobe

Adobe regularly updates the Photoshop software

Adobe Premiere Pro

First Used: 2021

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

NYC Media uses Adobe Premiere Pro to edit video content broadcast on the City's television stations. Within Adobe Premiere Pro, we use AI-powered tools to help generate closed captions of some video content that is edited in-house prior to broadcast.

Tool Purpose

We use AI-powered tools to help a human editor generate closed captions of some video content that is edited in-house prior to broadcast on the City's television channels.

Populations Impacted

Individuals

Individuals impacted include: Individuals who watch content on NYC Media's television channels; individuals who speak in content on NYC Media's television channels

Data Analyzed

Training Data	According to Adobe's website, "Speech to Text is powered by a combination of Adobe proprietary technology — including Adobe Sensei machine learning— and 3rd-party technologies."
Input Data	Spoken words in video programs
Output Data	Closed captions

Vendor Involvement

Vendor Name: Adobe

Adobe provides regular software updates

AI Transcription on Teams

First Used: Apr 2024

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

Microsoft Teams has a build in feature that uses AI to create a transcript of the meeting.

Tool Purpose

The tool provided a transcript of a meeting which was then reviewed by ENDGBV staff for accuracy. The transcript was then e-mailed to meeting participants.

Populations Impacted

Individuals

Individuals impacted include: Participants in the meeting might provide their name and business affiliations.

Data Analyzed

Training Data	N/A
Input Data	The words spoken during a meeting are captured by the tool.
Output Data	The tool provides text of the words spoken during the meeting.

Vendor Involvement

Vendor Name: Microsoft

Microsoft provides this tool as part of Teams.

AppTek OmniCaption 300 closed captioning appliance

First Used: Nov 2022

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

The tool uses AI-enabled automatic speech recognition to create closed captions of live television content.

Tool Purpose

MOME is using the AppTek Omni 300 closed captioning appliance to provide closed captioning of live-broadcasted events (e.g., City Council hearings) and content that is cablecast on NYC World.

Populations Impacted

Individuals

Individuals impacted include: People who watch live City Council and mayoral content televised on NYC Gov and other content televised on NYC World; people who appear in the content that is televised.

Data Analyzed

Training Data	According to AppTek's website, the "OmniCaption 300 closed captioning appliance was developed for and trained on broadcast news, sports, weather and other programming."
Input Data	The input data are words spoken by people during live broadcasts of public hearings, meetings, and events and content on NYC World.
Output Data	Closed captions that reflect the written text of the input data (spoken words)

Vendor Involvement

Vendor Name: AppTek

AppTek provides support for the Omni 300 closed captioning system

ChatGPT

First Used: Mar 2024

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: No	**Updated in 2024**: No

Tool Description

ChatGPT is an advanced AI language model that can understand and generate human-like text based on the input it receives. It can assist with a wide range of tasks, including answering questions, providing recommendations, and engaging in meaningful conversations.

Tool Purpose

ENDGBV staff have used ChatGPT to complete the following tasks: Assist in gathering information for literature reviews for ENDGBV reports; assist in creating SQL code to run data reports from the FJC database and other data sources; and assist in creating potential job interview questions based on content in the job description.

The literature review and data from the SQL code was used in public facing reports.

Populations Impacted

Data Analyzed

Training Data	N/A
Input Data	Text prompts were provided to ChatGPT.
Output Data	ChatGPT provided text responses to the text prompts.

ElevenLabs Speech Synthesis - MO - COS

First Used: Jun 2023

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

ElevenLabs creates the most realistic, versatile and contextually-aware AI audio, providing the ability to generate speech in hundreds of new and existing voices in over 20 languages.

Tool Purpose

The tool was used to generate audio recordings of mayor’s voice delivering hiring hall/Jobs NYC messages in various languages to be used for robo-calls.

Populations Impacted

Individuals

Individuals impacted include: phone call recipients

Data Analyzed

Training Data	Submitted audio recordings of mayor's speech, as well as a live language sample, into the program.
Input Data	Script translated in desired languages (Spanish, Yiddish, Haitian Creole).
Output Data	Audio recording (MP3) of mayor's voice speaking the script in the desired language (Spanish, Yiddish, Haitian Creole).

Vendor Involvement

Vendor Name: ElevenLabs

Utilized two existing features offered by ElevenLabs (VoiceLab and Speech Synthesis).

Methodology for Poll Site Language Assistance - MO - CEC

First Used: Nov 2020

Computation Type: Ranking	**Purpose Type**: Resource allocation
Identifying Information: No	**Updated in 2024**: Yes

Tool Description

Since no dataset is currently available that reliably captures the number of limited English proficient (LEP) registered voters for all program languages, the Civic Engagement Commission (CEC) uses the percentage of LEP citizens of voting age (CVALEP) as a substitute or proxy measure of need. CEC ranks the program-eligible languages in order of magnitude of CVALEP and distributes poll sites to each language based on its ranking (excluding CVALEP persons that speak languages served by the NYC Board of Elections (NYCBOE) in certain New York City counties). The number of poll sites that will receive services in any given language will depend on each language's share of the total CVALEP in the population eligible to be served. For example, according to U.S. Census data, approximately 207,926 New Yorkers are CVALEP and speak a language that is served by this program. This proportionality approach allows CEC to balance goals of including diverse language communities as well as fair access to the total number of eligible voters within each language community. The program provides interpreters in program-eligible languages at poll sites based on U.S. Census data showing concentrations of CVALEP individuals who speak these languages and reside around each poll site. For each language, poll sites are chosen in descending order of concentration of CVALEP, until the language's share is met. This process is repeated for each language, thereby including the poll sites with the highest concentration of CVALEP for each program-eligible language until that language's share is met, and the total number of poll sites for which resources are allocated is reached. It may be possible, based on analysis of data, to reassign poll sites to languages with greater need; however, each language will receive a minimum of at least one poll site. Models used included the Thiessen polygon method to create a Voronoi diagram to determine CVALEP estimates.

Tool Purpose

This is a methodology for determining how the New York City Civic Engagement Commission (NYCCEC) will provide interpretation services at poll sites for limited English proficient voters. The methodology explains how the NYCCEC will identify the languages and locations in which interpretation services will be offered during the November 2024 election and beyond. These services supplement the interpretation assistance provided by NYC Board of Elections in several languages. Under the Charter, the NYCCEC can only provide interpretation services in a language if: (1) it is a designated citywide language; or (2) it is spoken by a greater number of LEP New Yorkers than the lowest ranked designated citywide language and at least one poll site has a significant concentration of speakers of such language with LEP. This methodology ensures service for all languages that are eligible under the Charter.

Populations Impacted

Individuals

Data Analyzed

Training Data	N/A
Input Data	For citywide estimates, this methodology uses current data from the American Community Survey (ACS) 2017-2021 5-year estimates. This methodology also uses the American Community Survey Census Tract 2017-2021 5-year Public Use Microdata Samples for poll site level analysis, which tracks resident New Yorkers at the neighborhood level. In addition, the methodology uses data from the Board of Elections on the location of election districts and poll sites.
Output Data	The algorithm estimates the number of citizens of voting age with Limited English Proficiency for each program-eligible language who could report to each polling site.

Update Description

We are now using 2017-2021 data from the American Community Survey 5 year data. The differences between new and previous data are not statistically significant, therefore the distribution of services will not be affected based on these differences.

Scorecard Blockface Sampling Algorithm - MO - Operations

First Used: Mar 2022

Computation Type: Sampling	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: Yes

Tool Description

The Scorecard program sends inspectors across New York City to rate street and sidewalk cleanliness. The sampling algorithm creates a monthly list of blocks for inspectors to visit and rate.

The primary goal of the algorithm is to produce a sample of blockfaces that is statistically sound and geographically representative. This list is used to rate street and sidewalk cleanliness citywide, as well as by borough and DSNY district.

Tool Purpose

The algorithmic tool is used to produce a random sample set blocks/streets for crews to conduct litter inspections. Results are then used by DSNY to develop, and evaluate policies related to cleaning and enforcement programs.

Populations Impacted

Geographic space

Data Analyzed

Training Data	N/A
Input Data	The blockface sample is selected from the Pavement Edge File, which is part of the New York City Planimetric Database managed by the Office of Technology and Innovation. Blockface sampling is weighted equally across all geographic areas and includes extra sampling of blockfaces in Business Improvement Districts (BIDs). It also takes into account the linear miles of street within a DSNY District.
Output Data	A count of blockfaces that are statistically representative of our target areas throughout the city.

Vendor Involvement

Vendor Name: Legacy Mayor's Office of the Chief Technology Officer

The sampling algorithm was developed by the former Mayor's Office of the Chief Technology Officer in partnership with the Mayor's Office of Operations.

Update Description

We introduced equal weighting across all selected sample sizes. And, included new inspections geographies - NTAs (Neighborhood Tabulation Areas)

StratifySelect

First Used: Nov 2022

Computation Type: Matching	**Purpose Type**: Resource allocation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

StratifySelect is used to select a group of people from a pool of applicants such that the selected group A) matches target demographics and B) is as randomized as possible given the need to match targeted demographics. Traditional stratified sampling can create cases in which some individuals have a near-zero chance of being selected. This method uses a new technique of (1) explicitly computing a maximally fair output distribution and then (2) sampling from that distribution to select the final panel, achieving a fairer distribution of probabilities per applicant while maintaining fidelity to the demographics of the borough. The tool compares applicant demographic data to the data from the American Community Survey to achieve a representative sample. See this paper in Nature for details on the methodology: https://www.nature.com/articles/s41586-021-03788-6

Tool Purpose

The group selected by StratifySelect will be invited to participate in the CEC's Borough Assemblies. Each group contains a prioritized list including "backup options" in case some of the invitees decline or are unable to participate.

Populations Impacted

Individuals

Individuals impacted include: New York City residents (all boroughs)

Data Analyzed

Training Data	N/A
Input Data	Fully anonymized data from interested people (age, gender, race, and Hispanic identity, borough, zip code) is used as input. We emphasize that no identifying data is stored, shared and/or transmitted during the entire process.
Output Data	The output is a subset of the same data such that the output group is both randomly selected and representative of each borough in these categories, as defined by public census data.

Zoom

First Used: 2020

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

Zoom is a virtual meeting platform; Zoom has an auto closed-caption function that uses AI.

Tool Purpose

MOME uses Zoom for public hearings on rulemaking and for public webinars. We use Zoom's auto transcript function and captioning function. (Note: We provide ASL and human-typed CART services as a reasonable accommodation upon request.) If we publish a transcript after the Zoom meeting, a human reviews and corrects it.

Populations Impacted

Individuals

Individuals impacted include: People who participate in rulemaking hearings and webinars; people who read transcripts of those hearings and webinars

Data Analyzed

Training Data	According to Zoom's website, "Zoom does not use any customer audio, video, chat, screen sharing, attachments, or other communications-like customer content (such as poll results, whiteboard, and reactions) to train Zoom's or its third-party artificial intelligence models."
Input Data	Speech at MOME's rulemaking hearings and agency webinars.

Output Data	Text in a transcript and captions.

Vendor Involvement

Vendor Name: Zoom

Regular updates to the application

New York Police Department (NYPD)

Evolv Express Weapons Detection System

First Used: Jul 2024

Computation Type: Forecasting	**Purpose Type**: Data management
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

Electromagnetic weapons detection system

Tool Purpose

Electromagnetic weapons detection devices emit ultra-low frequency, electromagnetic pulses (similar to those used in retail loss prevention) that pass through objects moving through the system. Sensors process the relayed information and the system uses this data to determine if it detects a potential firearm. The system is equipped with video cameras that are part of a real-time image-aided alert system that will indicate the presence of a firearm to monitoring personnel.

Populations Impacted

Individuals; Property; Geographic space

Individuals impacted include: General Population

Data Analyzed

Training Data	Training data is proprietary to the vendor.
Input Data	Individuals walk through two towers which emit ultra-low frequency, electromagnetic pulses that pass through objects on the individual's person. Sensors process the information generated and the system uses this data to determine if it detects a potential firearm.

Output Data	If a potential firearm is detected, the system will capture a still image and an approximately three-second video of the individual moving through the system. The system will alert monitoring personnel that a potential firearm has been detected and wirelessly transmit the still image and video to a tablet being monitored by personnel. A cube will appear on both the still image and video clip, indicating the location of the potential firearm being worn or carried by the individual. The location of a cube is discerned by the system based on the electromagnetic data processed by the system sensors.

Vendor Involvement

Vendor Name: EVOLV

Software developed and maintained by vendor

Facial Recognition Technology

First Used: Oct 2011

Computation Type: Matching	**Purpose Type**: NA
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

Tool which may help investigators identify unknown subjects in law enforcement investigations.

Tool Purpose

Facial recognition is a digital technology that NYPD uses to compare images obtained during investigations with lawfully possessed arrest and parole photos. The tool analyzes an uploaded image, known as a probe image, and searches and compares against the image repository. The purpose of the tool is to enhance law enforcement's ability to investigate criminal activity as well as identify deceased persons and missing persons. When used in combination with human analysis and additional investigation, facial recognition technology is a valuable tool in solving crimes and increasing public safety.

Populations Impacted

Individuals

Individuals impacted include: General Population

Data Analyzed

Training Data	Training data is proprietary to the vendor.

Input Data	If NYPD investigators obtain a still image depicting a face of an unknown individual during an investigation, the image can be submitted for facial recognition analysis in accordance with NYPD facial recognition policy. Known as a probe image, NYPD facial recognition software compares the image to a controlled and limited group of lawfully obtained photos called the photo repository.
Output Data	The facial recognition software will generate a pool of possible match candidates for review by trained Facial Identification Section investigators.

Vendor Involvement

Vendor Name: Dataworks

Software developed and maintained by vendor

Patternizr

First Used: Dec 2016

Computation Type: Matching	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: Tool was created in CY2024

Tool Description

Aids crime analysis in detection of potential crime patterns.

Tool Purpose

Patternizr compares features of crimes and finds ones that are similar and may be part of a crime pattern. Analysts will look at the candidate crimes and suggest the formation of crime patterns to a pattern identification module. If a pattern is formed, detectives often consolidate the investigative efforts (e.g., one detective investigates all the crimes in the pattern.) The report filters non-normal trends into a spreadsheet and displays year-over-year counts of crimes that have non-normal trends. The tool requires a human user to evaluate the output data to see if complaints identified as similar are, in fact, connected to a pattern.

Populations Impacted

Individuals; Property; Geographic space; Other

Others impacted include: Crime classification

Data Analyzed

Training Data	Separate models were trained for each of three different crime types (burglaries, robberies, and grand larcenies). These crime types have a sufficient corpus of prior manually identified patterns for use as training examples. This corpus consists of approximately 10,000 patterns between 2006 and 2015 from each crime type. A portion of this corpus includes complaint records where the same individual was arrested for multiple crimes of the same type within a span of two days.
Input Data	The input data is a candidate crime and its features. A complaint describes details of the crime, including the date and time (which can be a range if the precise time of occurrence is unknown), location, crime subcategory, modus operandi, and suspect information. This information is used to calculate the five types of crime-to-crime similarities used as features by Patternizr: location, date-time, categorical, suspect and unstructured text.
Output Data	Probability that a complaint is connected to a pattern.

ShotSpotter

First Used: Mar 2015

Computation Type: Matching	**Purpose Type**: Data management
Identifying Information: No	**Updated in 2024**: Yes

Tool Description

Provides acoustic gunshot detection to assist with emergency call response

Tool Purpose

Provides acoustic gunshot detection to assist with emergency call response. The tool supports patrol operations in alerting units to potential gunfire and enhances investigations involving firearms.

Populations Impacted

Geographic space

Data Analyzed

Training Data	Training data is proprietary to the vendor.
Input Data	Specialized software analyzes audio signals for potential gunshots.

Output Data	The tool determines the location of the sound source, and once classified as potential gunfire sends the incident to acoustic experts for additional analysis. Notifications are sent for confirmed gunfire. ShotSpotter activations may result in evidence collection that can enhance case investigations. Problematic locations identified through alerts may require additional resource deployment and/or investigations.

Vendor Involvement

Vendor Name: Shotspotter

Software developed and maintained by vendor

Update Description

routine maintenance

Office of Chief Medical Examiner (OCME)

STRMix

First Used: Jan 2017

Computation Type:	**Purpose Type**: Data management
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

STRmix™ combines sophisticated biological modeling and standard mathematical processes to interpret a wide range of complex DNA profiles. Using well-established statistical methods, the software builds millions of conceptual DNA profiles. It grades them against the evidential sample, finding the combinations that best explain the profile. A range of likelihood ratio options are provided for subsequent comparisons to reference profiles. Using a Markov Chain Monte Carlo engine, STRmix™ models any types of allelic and stutter peak heights as well as drop-in and drop-out behavior. It does this rapidly, accessing evidential information previously out of reach with traditional methods. STRmix™ is supported by comprehensive empirical studies with its mathematics readily accessible to DNA analysts, so results are easily explained in court.

Tool Purpose

STRMix is a probabilistic genotyping tool that is used to analyze mixtures of DNA profiles to help associate the crime scene evidence to potential victims or suspects of crimes.

Populations Impacted

Individuals; Biological sample

Data Analyzed

Training Data	Training data was not used in the sense of AI software. The OCME performed thousands of tests using the software to validate it for optimum use with our current laboratory standard operating procedures and genetic analyzers.
Input Data	Forensic DNA profiles from crime scenes as well as the DNA profiles from victims and suspects of crimes.
Output Data	The output is a deconvolution of genotype probability distribution that lists all of the accepted genotype sets and their associated weights. These weights can take any value from 0 to 1.

Vendor Involvement

Vendor Name: NicheVision Forensics, LLC

The software has been developed by New Zealand Crown Institute of Environmental Science and Research (ESR) with Forensic Science South Australia. The developer assisted OCME in analyzing and interpreting our data during the validation of the software.

Office of Technology and Innovation (OTI)

311 AI Voice Pilot

First Used: Mar 2024

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: No	**Updated in 2024**: Tool was created in CY2024

Tool Description

311 AI Voice pilot is an LLM-powered voice call solution that provided information and access to services for residents, businesses, and visitors.

Tool Purpose

The 311 AI Voice was a pilot program that enabled NYC311 to test a generative AI voice application with 311 customers, for 21 days. On a daily basis, customers can reach the 311 call center, by dialing 311, 212-NEW-YORK, or 211. The pilot was kept on a small scale, available only to customers who dialed 211. The 311 AI Voice provided information to residents, businesses, and visitors on a wide range of inquiries, as well as providing updates on city programs, events, and notifications using NYC311's content.

Populations Impacted

Individuals

Data Analyzed

Training Data	Training data is proprietary to the vendor.
Input Data	The customer made voice inquiries when contacting NYC311.

Output Data	311 AI Voice provided the customer with information/responses from the content within NYC311 Content API.

Vendor Involvement

Vendor Name: Microsoft, Nuance

Microsoft provided an LLM-powered solution to aid in handling customers who call 311 for information and services. The focus was handling voice interactions.

MyCity Chatbot

First Used: Sep 2023

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: No	**Updated in 2024**: Yes

Tool Description

The NYC MyCity chatbot is a beta AI-powered chatbot that provides information and access to services for residents and businesses in New York City.

Tool Purpose

The NYC MyCity chatbot is a beta AI-powered chatbot that provides information and access to services for residents and businesses in New York City. It's currently focused on two main areas: Business Services and MyCity Basics. The chatbot provides information on starting or operating a business in New York City, answers questions about permits, licenses, regulations, and other business requirements, and connects users with relevant resources and support services. It also offers information on various city services and benefits, and helps users find resources related to childcare, career, and other areas. The chatbot is using Microsoft's Azure AI technology and OpenAI's ChatGPT 4-o LLM.

Populations Impacted

Individuals; Group, organization, or business

Data Analyzed

Training Data	Training data is proprietary to the vendor.
Input Data	Text queries are input by the user on the MyCity portal.
Output Data	The tool produces text responses with references based on information from Business Services and MyCity Basics.

Vendor Involvement

Vendor Name: Microsoft, EY

Microsoft provides Cloud-based ChatGPT services, and EY is the professional services vendor.

Update Description

Security updates and LLM upgraded to use ChatGPT 4-o.

Omnichannel Language Translation

First Used: Jan 2024

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: No	**Updated in 2024**: Tool was created in CY2024

Tool Description

The Omnichannel language translation tool delivers multi-language capability for the 311 text channel. The tool supports the 10 designated citywide languages to enable 311 agents to interact with customers in their language.

Tool Purpose

The algorithmic tool converts the customer's text inquiry in their chosen language into English, allowing the text agent to understand, research and reply to the inquiry. The tool converts the agent's English language response to the customer's chosen language among the 10 designated Citywide languages.

Populations Impacted

Individuals

Individuals impacted include: Customers who contact 311 via text/SMS

Data Analyzed

Training Data	The training data is proprietary to Microsoft.
Input Data	Text/SMS inquiries from customers via 311-NYC.
Output Data	Responses to customer inquiries in the language the customer used.

Vendor Involvement

Vendor Name: Microsoft

Omnichannel is part of the Microsoft suite available to OTI as part of the Dynamics CRM platform. Microsoft supported the design, development, and testing of the tool preparation and deployment.

Zoom Automated Captions

First Used: Mar 2021

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: Yes	**Updated in 2024**: No

Tool Description

Creates virtual closed captioning/live transcription during Zoom meetings

Tool Purpose

Captions are provided to attendees of Zoom meetings held by the NYC Open Data team at the Office of Data Analytics in conjunction with the civic tech non-profit BetaNYC under the Open Data Week and Open Data Ambassador initiatives. The full transcription of the event is then added to the meeting recordings, which are uploaded on YouTubce. The purpose of the captions, both for the live event and the recording, is to improve meeting accessibility.

Populations Impacted

Individuals

Individuals impacted include: Attendees of Open Data Week and Open Data Ambassadors meetings who have spoken during the meeting or whose names have been mentioned during the meeting.

Data Analyzed

Training Data	Not known.
Input Data	Live audio from Zoom meeting
Output Data	VTT format file including captions/transcript of meeting

Vendor Involvement

Vendor Name: BetaNYC

BetaNYC is our collaborator on the Open Data Week and Open Data Ambassador initiatives. They own and operate the Zoom account that is used for meetings under these initiatives, have access to the transcription files, and use these when editing and uploading video recordings onto YouTube.

School Construction Authority (SCA)

GitHub Copilot

First Used: May 2023

Computation Type: Data generation	**Purpose Type**: NA
Identifying Information: No	**Updated in 2024**: No

Tool Description

GitHub Copilot is an AI-powered code assistant that provides suggestions for whole lines or blocks of code in a wide range of programming languages. It leverages a vast codebase and machine learning to improve coding efficiency, helping programmers by autocompleting code snippets and offering context-appropriate code suggestions.

Tool Purpose

GitHub Copilot is primarily used by some of our software developers as an advanced coding assistant within our agency. Its role is to augment and streamline the coding process for our software development projects. By providing real-time code suggestions and completions, it reduces the time developers spend on routine coding tasks, allowing them to focus on more complex aspects of software development.

The tool functions by analyzing the context of the code being written and suggesting relevant, syntactically correct code snippets. This includes generating code for standard programming patterns, filling in boilerplate code, and offering solutions to simple programming queries. It's important to note that while GitHub Copilot assists in the coding process, final decisions on the code's implementation and its use in any software or application rest solely with our human developers. The tool's suggestions are always reviewed and potentially modified by our team to ensure they meet our specific requirements and standards. Therefore, GitHub Copilot acts as a support tool in the decision-making process of software development rather than a decisive entity.

Populations Impacted

Individuals; Group, organization, or business

Data Analyzed

Training Data	GitHub Copilot was developed by OpenAI and trained using a large corpus of public source code available on GitHub. This training data includes a wide variety of code in multiple programming languages, along with associated comments and documentation. The data encompasses a broad range of coding styles, patterns, and solutions across different software development projects.
Input Data	When in use, GitHub Copilot analyzes the code that a developer is currently writing. This input data consists of the programming language syntax, structure, and any comments or context within the code file. The tool also takes into account the specific coding task, patterns, and functions that the developer is working on. This real-time data is essential for the tool to provide relevant and context-appropriate coding suggestions.
Output Data	The output data from GitHub Copilot includes suggested lines or blocks of code that align with the input data provided by the developer. These suggestions are generated based on the patterns, structures, and coding practices learned from the training data. The output is designed to seamlessly integrate with the existing code, offering syntactically correct and contextually relevant code completions.

Vendor Involvement

Vendor Name: GitHub, a subsidiary of Microsoft

n/a

Secure Conversational AI

First Used: Oct 2024

Computation Type: Data generation	**Purpose Type**: Information presentation
Identifying Information: No	**Updated in 2024**: Tool was created in CY2024

Tool Description

A secure, ChatGPT-like assistant powered by Azure OpenAI Services, offering advanced capabilities such as summarization, brainstorming, and contextual information retrieval. Designed to enhance employee productivity while ensuring data confidentiality within SCA's infrastructure.

Tool Purpose

The algorithmic tool is used to support employees by providing capabilities such as summarization, brainstorming, and contextual information retrieval. Its purpose is to streamline tasks, improve access to information, and enhance productivity while maintaining data confidentiality

Populations Impacted

Group, organization, or business

Data Analyzed

Training Data	N/A
Input Data	User-provided queries or prompts, which may include text-based questions, descriptions, or tasks requiring assistance with summarization, brainstorming, or information retrieval.
Output Data	Text-based responses generated by the tool, including summaries, ideas, contextual information, or task solutions, tailored to the user's input