AI Driven Verbatim Coding – The Real World

Tim Brandwood & Pat Molloy

Who are we?

We're a start-up that makes smart MR software

codeit
Coding. Only Better
from digital taxonomy

Uses a set of AI techniques, seamlessly blended, to code open-ended text answers

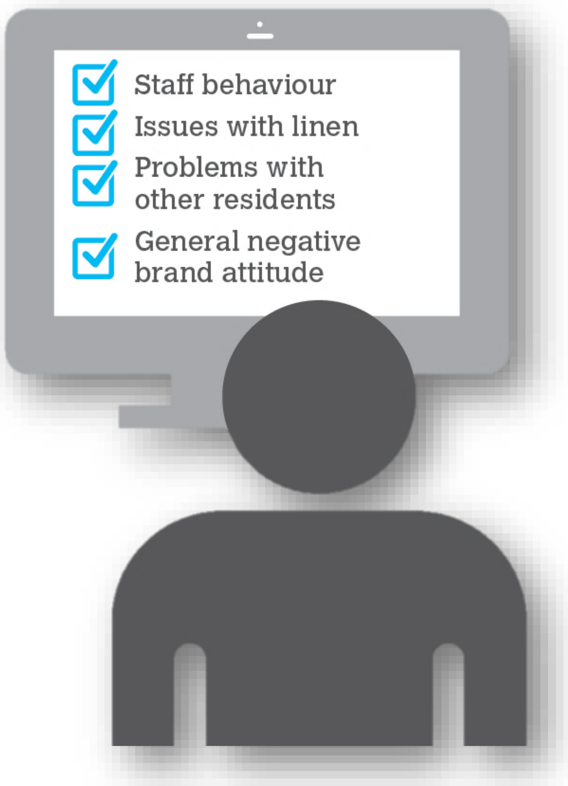
loadit
Data Processing Made Simple
from digital taxonomy

ETL for mere mortals! Automate repetitive data processing tasks with a simple drag and drop workflow

Open Ended Questions

Q: Why did you not enjoy your stay at Hotel X?

A: Staff were rude, sheets were dirty, noisy stag party. I hate Hotel X



- **30%** Staff behaviour
- **28%** Issues with linen
- **15%** Problems with other residents
- **8%** General negative brand attitude

The case for open ended questions

Open Ended Questions

Purpose: To encourage participation, gain new information, clarify meaning, and increase understanding.

Pros and Cons

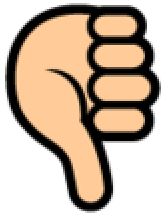
- Unlimited response possibilities
- Enables the collection of finer detail
- Uncover genuinely new/unexpected material
- Allow respondents to self-express and encourage more creative thinking
- Gather more detailed answers to potentially complicated questions
- Perhaps get some insight into thought processes
- No forced response pre-conceptions for respondent
- Make use of the expressive language to uncover sentiment/feeling

- Expensive to process
- Slow to process
- Requires human experts to make sense of answers
- Variability / Subjectivity of human experts
- Difficult to analyse/aggregate /interpret
- Creating effective/unambiguous codeframes is hard

The story so far…

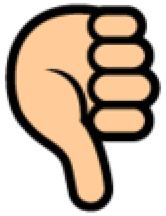
- In May 2017 we talked about:

Off the shelf APIs



Too generic and not tailored to the exact requirements of Market Research coding

Simple text matching rules

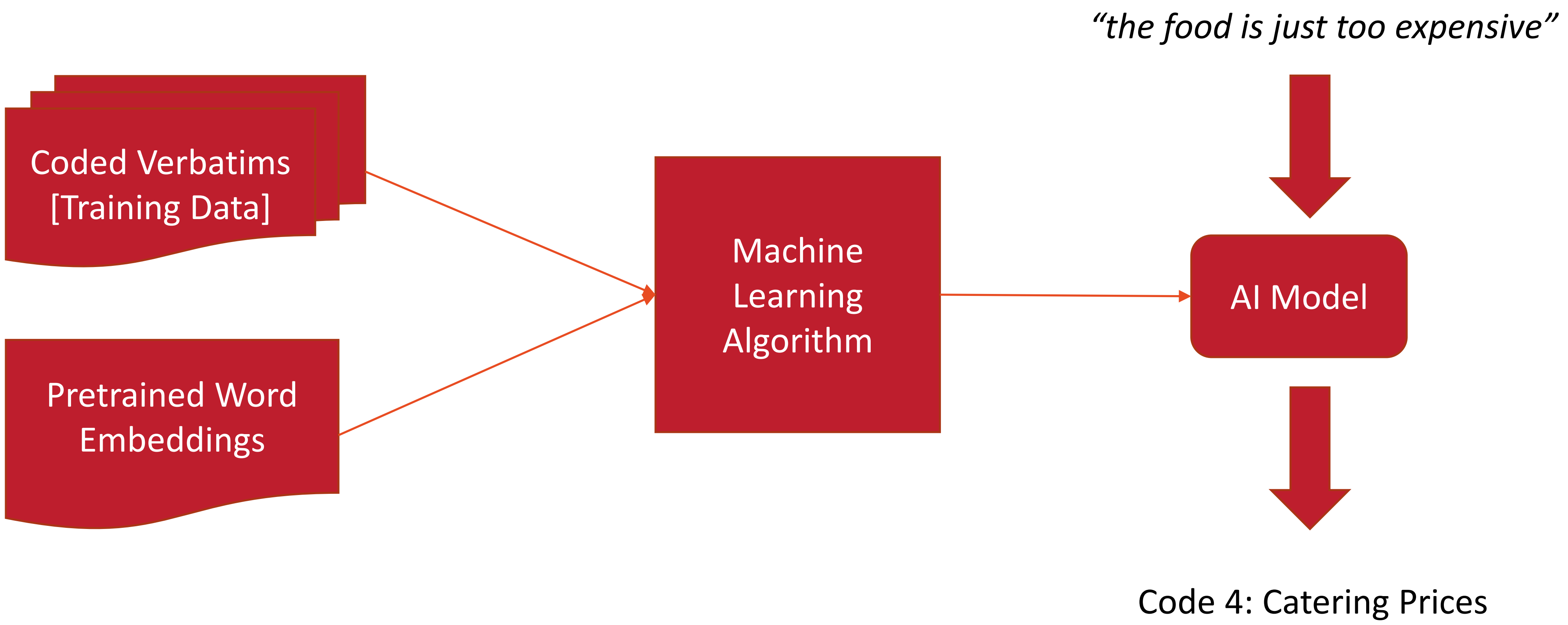


Risk of false positives/negatives
Quickly get out of hand and become unmaintainable

Machine Learning

Proved to give good results

Machine Learning Process



Machine Learning Process

"Reasonably priced food, not rip-off pricing."

"Cheaper soft drinks..."

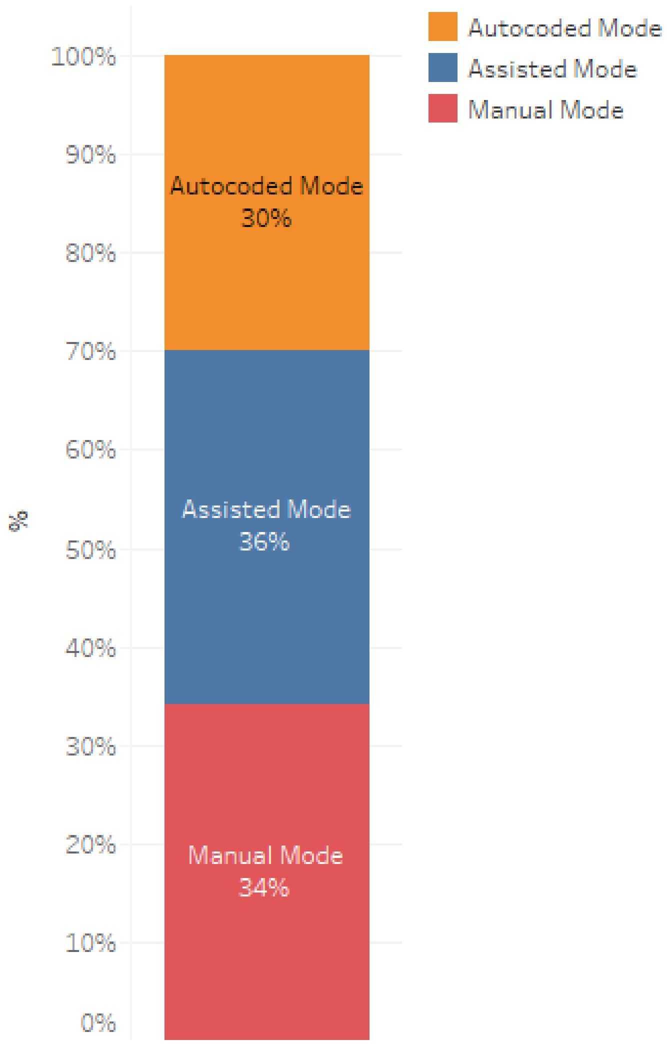
"More affordable food options."

"cheaper tea and coffee outlets"

"reduce price of food"

"Food less costly"

Machine Learning



- Overall effect is a useful timesaver
- The Machine Learning is able to help with around two thirds of items
- So where is the Machine Learning less able to help…?

Very short responses

Machine Learning is less able to cope with very short responses (e.g. brand coding)

- Miscoding.

  A 3% – 5% error rate is common in human coded training data.

- Typos/Unknown Words

  Even small typos can make certain words unrecognizable to the Machine Learning

  Certain words may be valid, but lie outside of the list of “known words” for the Machine Learning

  In longer verbatims, other words within a phrase can help compensate for misspellings or unknown words

- Overkill

  Machine Learning is a like using a hammer to crack a nut for this situation

Very short responses

Miscoding – “McDonalds”

Code	Count
McDonalds	3000
Burger King	50
KFC	20
Nandos	15
Five Guys	10
Honest Burger	5

Very easy to find the most commonly occurring code for this text and apply automatically to future instances

But….need to be careful if the question is multicoded…

Very short responses

Typos/Unknown Words

mcdonald's
Mac Donald
mcdonald
mc donald
mc
mc'donald
mec donald
mec donalw
Mec dondalds
mecdonalds

Very easy to define a text matching rule that catches all of these examples:

`mc|donal`

Dealing with more complicated verbatims

- Long responses, especially without clear structural delimiters will cause the Machine Learning on its own to struggle.

  - *Watching films at home is comfortable, allows for drink and toilet breaks and I can laugh as loud as I want.*

  - *It's more comfortable and you don't have to get dressed up or go out and everything is on hand in your own home plus modern tv and sound systems are like being in the cinema*

  - *Cinema is too expensive, and you don't know whether you'll be stuck with annoying people who'll ruin the experience, at home however you can guarantee a good environmment, whatever food you want, and toilet breaks without missing important bits!*

Dealing with more complicated verbatims

- NLP/Text Analytics can be used to extract key segments from longer verbatims

  - *Watching films at home is comfortable, allows for drink and toilet breaks and I can laugh as loud as I want.*

  - *It's more comfortable and you don't have to get dressed up or go out and everything is on hand in your own home plus modern tv and sound systems are like being in the cinema*

  - *Cinema is too expensive, and you don't know whether you'll be stuck with annoying people who'll ruin the experience, at home however you can guarantee a good environmment, whatever food you want, and toilet breaks without missing important bits!*

Dealing with more complicated verbatims

- Segments can be coded and be used to train the system and build further coding rules

home is comfortable => Home Comforts

comfortable => Comfortable

toilet breaks => Can pause / take breaks / flexible

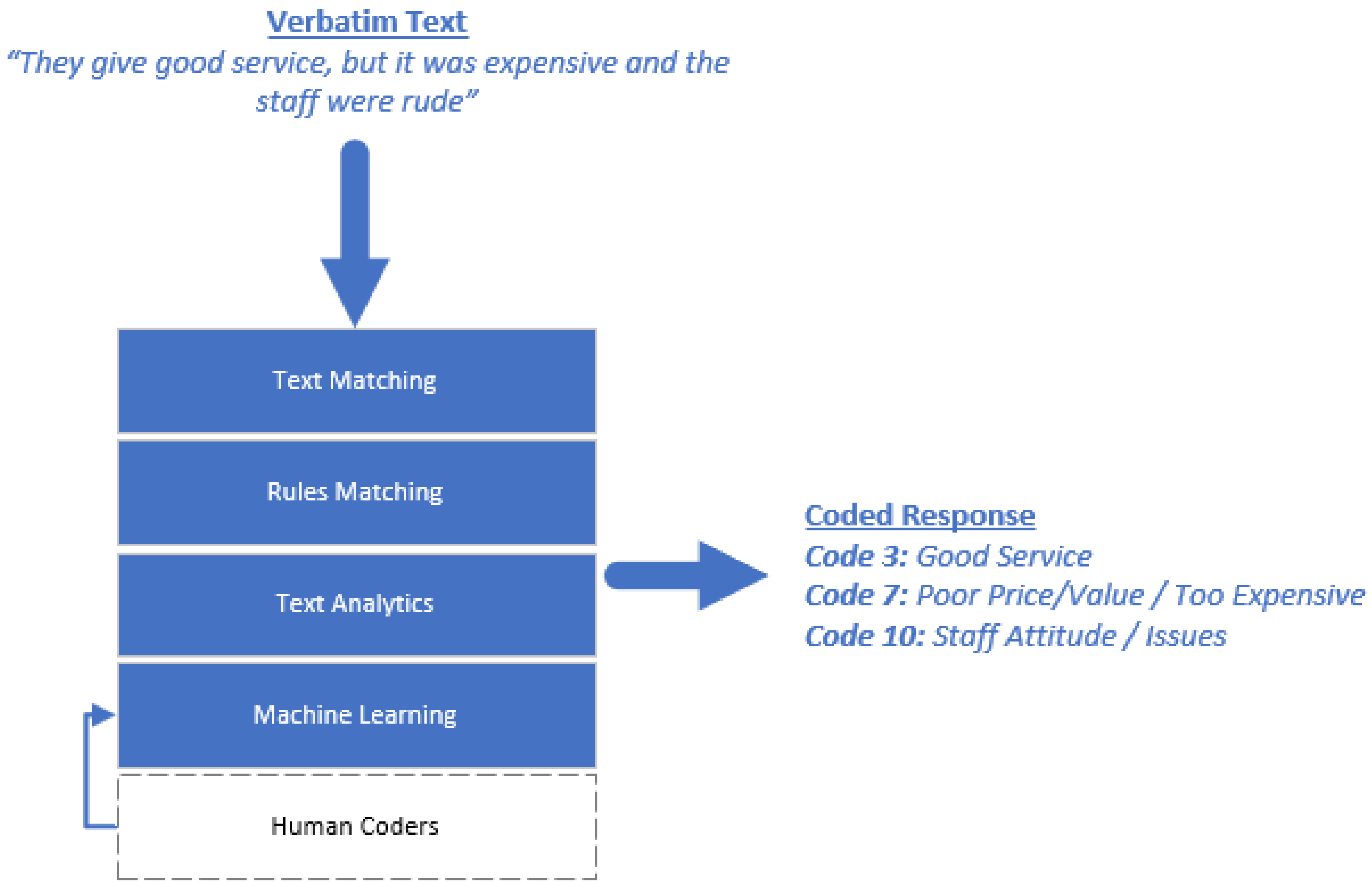
Cinema is too expensive => Cheaper

etc..

People!

- People are still essential to the coding process
- Coders are needed to categorize segments and approve rules suggested by the system
- Coders are still needed to “fill in the blanks” and code items not caught by the other methods.

The Codeit "Blended AI" Approach



Real-World Example

- Network Research – Insurance Renewal Exit Interview
- How to code ~46,000 verbatims gathered via IVR

Real-World Example

- Step 1 – Apply Text Analytics to extract key topics and auto-generate codeframe

Code	Label
4	Competitive
5	Competitive Price
6	Competitive Quote
9	Easy
10	Easy To Deal With
11	Easy To Understand
16	Excellent
17	Excellent Customer Service
18	Excellent Service

28,537 (57%) items coded

Real-World Example

- Step 2 – Train AI and apply autocoding

Code	Label
4	Competitive
5	Competitive Price
6	Competitive Quote
9	Easy
10	Easy To Deal With
11	Easy To Understand
16	Excellent
17	Excellent Customer Service
18	Excellent Service

30,468 (66%) items coded

Real-World Example

- Step 3 – Coder Refines the Codeframe

Code	Label
8	Competitive/Competitive Price/Competitive Quote
9	Easy/Easy To Deal/Easy To Understand/Simple
10	Excellent Customer Service/Excellent Service

30,256 (65%) items coded

Real-World Example

- Step 4 – Coder manually codes a sample of uncoded items
- Items are selected based on AI uncertainty
- An additional 3,000 items are coded => 72%

Real-World Example

- Step 5 – Train AI and apply autocoding
- An additional 5,096 items are coded => 83%
- This was deemed sufficient for analysis purposes
- Overall human time was 6% of the amount of time normally required to code this volume of verbatims
- Overall results have a traditional "curated" coding feel despite being largely coded by machine

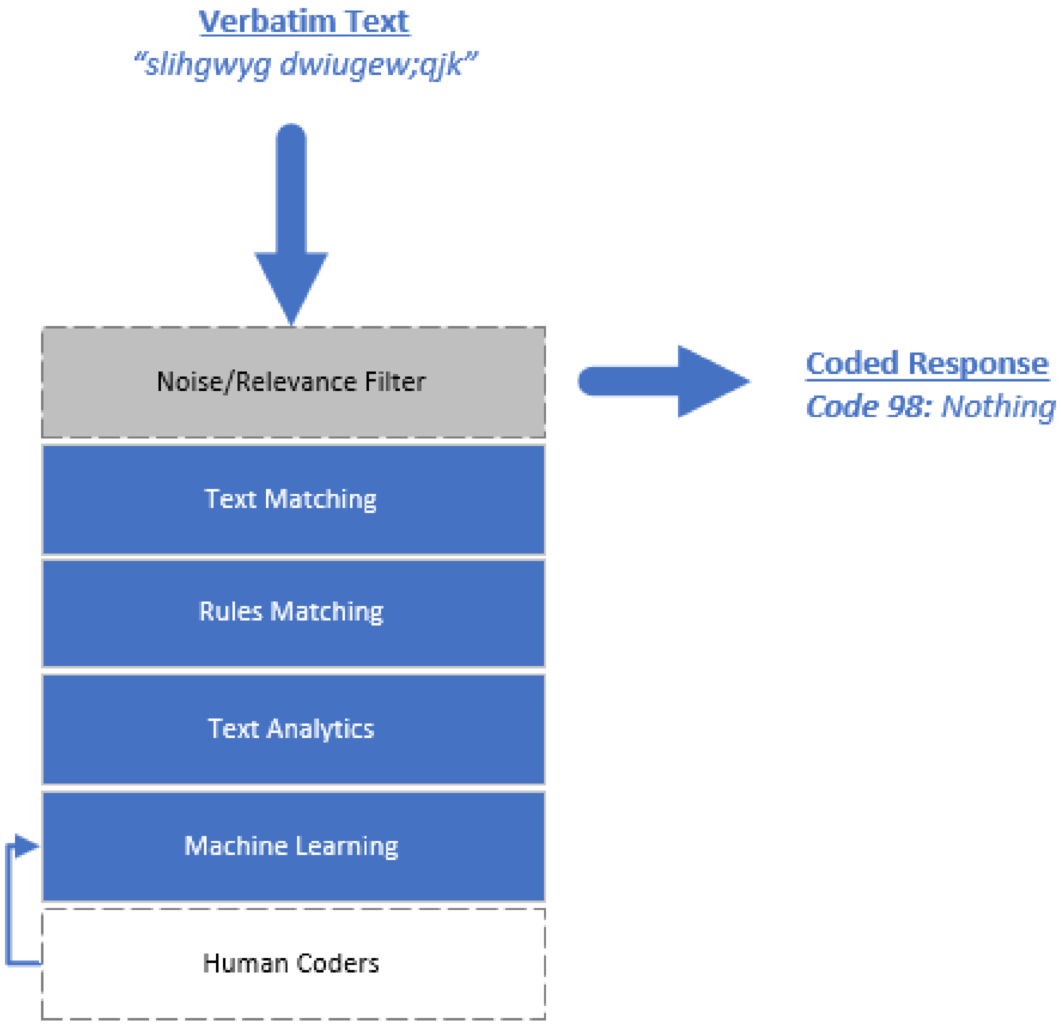
What next for AI in the Survey World?

- *Filtering "Nonsense"*
  e.g. Keyboard mashing
  Swearing
  Gobbledegook

- *Filtering for Relevance*
  e.g. Off Topic
  Particularly important when coding social media data

```
-------------------------------------------
----  Welcome to Junk-o-Matic 2000  --------
-------------------------------------------
Enter some text to check (enter 'exit' to quit): This is a verbatim that I would be very happy to have as a response to my open ended question!
Results:
============
VVS: 100

Enter some text to check (enter 'exit' to quit): On the other hand this is simply crap crap crap and more crap
Results:
============
RepeatedWordsPct: 23
VVS: 0
BadWordsFlag: 1

Enter some text to check (enter 'exit' to quit): sd;flkjsdf;jasfjneerer
Results:
============
VVS: 17
GarbageWordsFlag: 1

Enter some text to check (enter 'exit' to quit): pipppy doobly dobblen all gymbly woth the frome
Results:
============
VVS: 63

Enter some text to check (enter 'exit' to quit): df fg
Results:
============
VVS: 0
GarbageWordsFlag: 1

Enter some text to check (enter 'exit' to quit): noggin plips
Results:
============
VVS: 31

Enter some text to check (enter 'exit' to quit): 23-4582-9348=19481=
Results:
============
VVS: 3
GarbageWordsFlag: 1

Enter some text to check (enter 'exit' to quit): On the other hand this is quite good
Results:
============
VVS: 63
```

What next for AI in the Survey World?

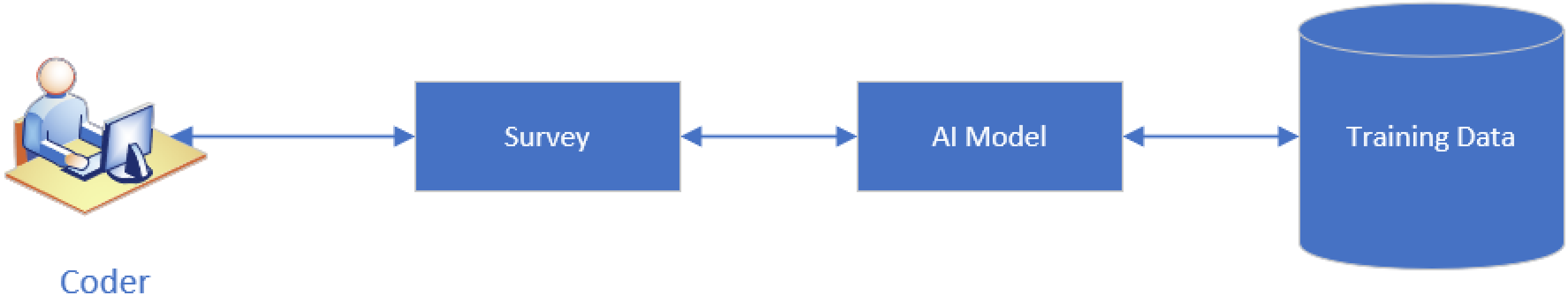


What next for AI in the Survey World?

- *"Life begins at a billion examples"*

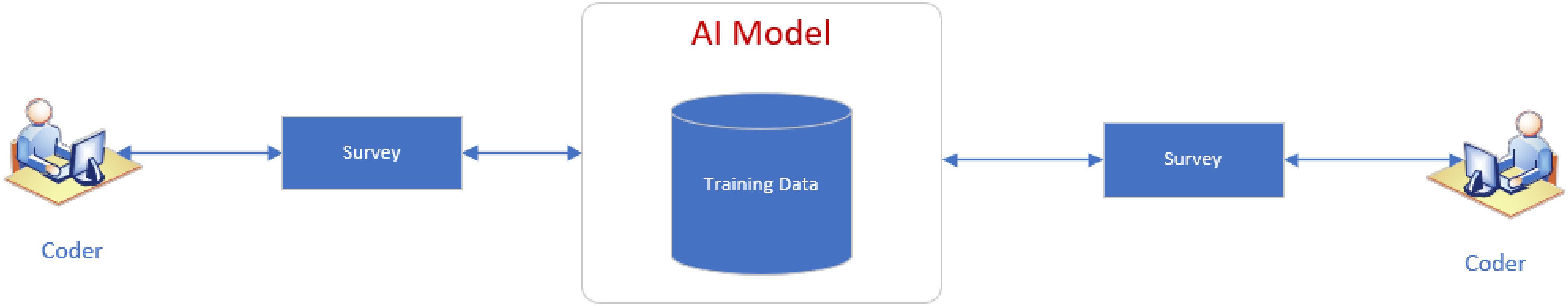
- *Current training data and AI model is siloed*

- *Quantity of data in each silo is relatively small*



What next for AI in the Survey World?

- *Codeit architecture allows for the "pooling" of data to train the AI*
- *Pooling data within a company, across companies or industry-wide(!) would massively increase the power of AI for Market Research*



Conclusions

- It is a mistake to look for a single “silver bullet”
- For best results it will often be necessary to blend techniques to suit the task at hand
- It is a mistake to think of AI as simply replacing humans.
  For now, AI is optimized if it is used to assist humans rather than trying to do everything for them.
- This “new wave” of AI is only 5 years old .. Plenty of room for further innovation in models and techniques

"Any questions?"

tim@digitaltaxonomy.co.uk
pat@digitaltaxonomy.co.uk

www.digitaltaxonomy.co.uk
@digitaltaxonomy