Real-time Meatspace Data Science

Data Intelligence Conference

Jason Walsh
Data Engineer
Jason.Walsh@uphs.upenn.edu
@rightlag

Michael Becker
Sr. Data Scientist
Michael.Becker2@uphs.upenn.edu
@beckerfuffle

6/24/2017

Penn Medicine

Data Scientist or Data Janitor?

Amplitude

A blog about building better products

Stop Treating Your Data Scientist Like a Janitor

By Archana Madhavan | July 19 | 0 Comments

63 12

So you just hired someone to fill an opening for the sexiest job of 2016, a data scientist at your up-and-coming Silicon Valley startup. What if, a few months down the line, you realize your data scientists are actually spending the majority of their time working on something decidedly... unsexy?

Stop Treating Your #DataScientist Like a Janitor

CLICK TO TWEET

Data cleanups aren't fun

Data Scientist or Data Janitor?

Amplitude

A blog about building better products

Stop Treating Your Da… Scientist Like a Janit…

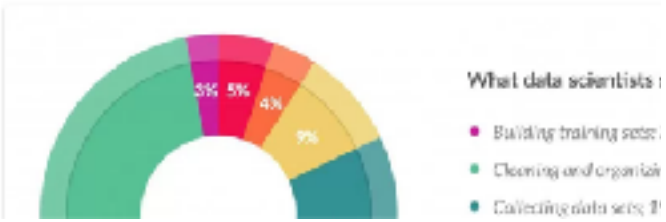
By Archana Madhavan | July 19 | 0 Comments

63 12

So you just hired someone to fill an opening for the sexiest job of 201… coming Silicon Valley startup. What if, a few months down the line, yo… actually spending the majority of their time working on something d…

Stop Treating Your #DataScientist Like…

Data cleanups aren't fun



MAR 23, 2016 @ 09:33 AM 19,156

The Little Black Book of Billionaire Secrets

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Gil Press, CONTRIBUTOR

I write about technology, entrepreneurs and innovation. FULL BIO

Opinions expressed by Forbes Contributors are their own.

TWEET THIS

- data scientists found that they spend most of their time massaging rather than mining or modeling data.
- 76% of data scientists view data preparation as the least enjoyable part of their work

A new survey of data scientists found that they spend most of their time massaging rather than mining or modeling data. Still, most are happy with having the sexiest job of the 21st century. The survey of about 80 data

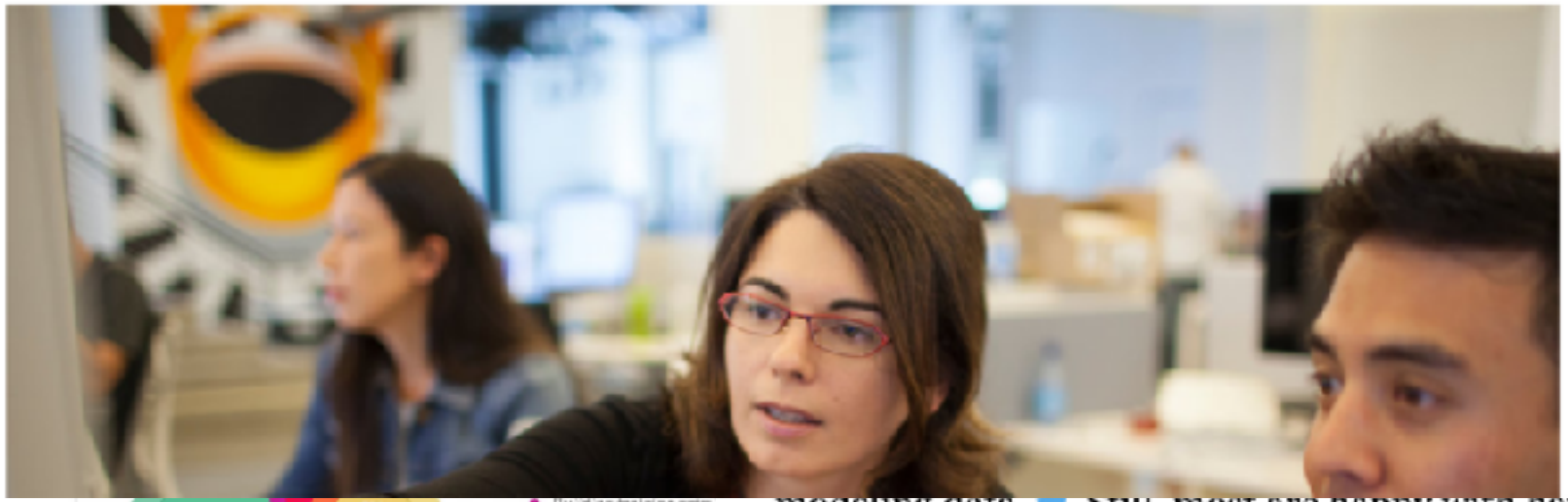
Data Scientist or Data Janitor?

ONS HOME SEARCH The New York Times SUBSCRIBE NOW LOG

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014



modeling data. Still, most are happy with having the sexiest job of the 21st century. The survey of about 80 data

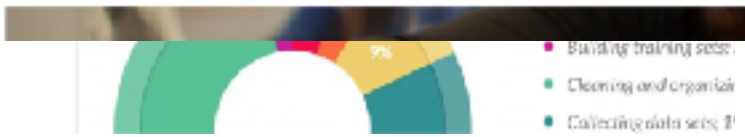
Data Scientist or Data Janitor?

Big Data Borat

@BigDataBorat

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

Retweets	Likes
509	281



modeling data. Still, most are happy with having the sexiest job of the 21st century. The survey of about 80 data

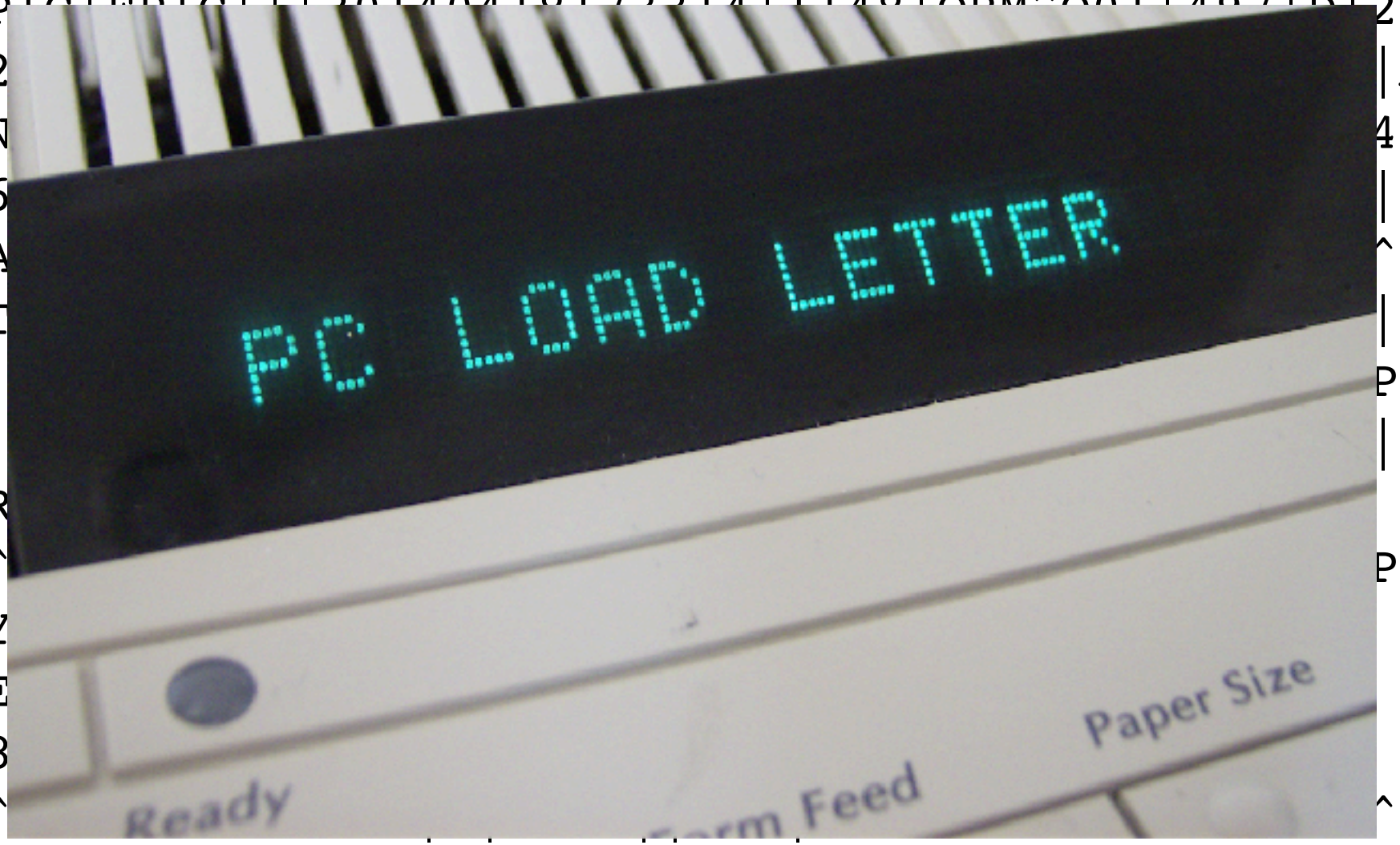
Health Care Data “Standards”

HL7

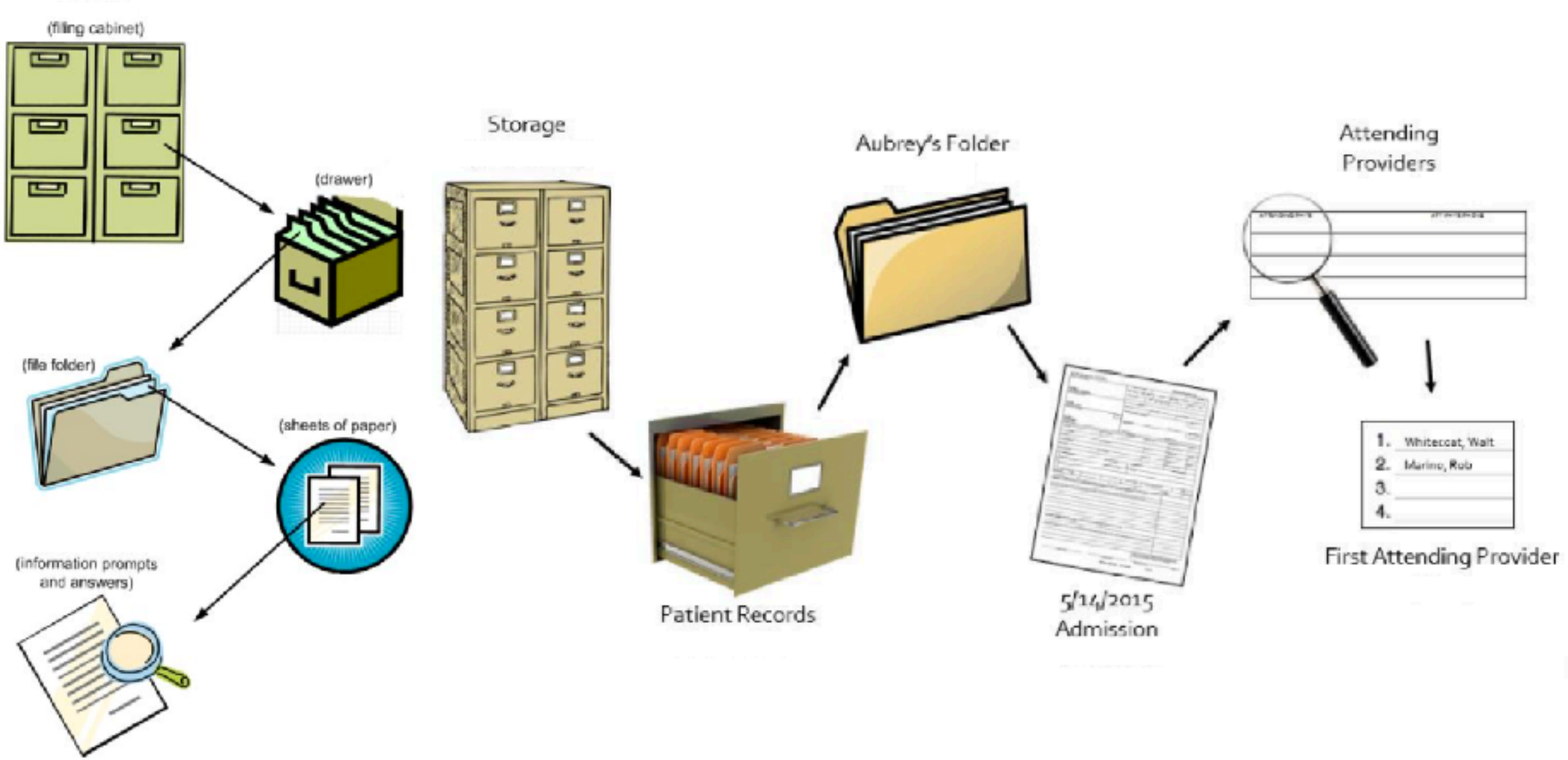
```
MSH|^~\&|EPIC|EPIC|||20140418173314|1148|ORM^O01|497|D|2.3|| PID|
1||20891312^^^^EPI||APPLESEED^JOHN^A^^MR.^||19661201|M||AfrAm|505
S. HAMILTON AVE^^MADISON^WI^53505^US^^DN |DN|(608)123-4567|
(608)123-5678||S|| 11480003|123-45-7890||||^^^WI^^ PD1|||
FACILITY(EAST)^^12345|1173^MATTHEWS^JAMES^A^^^ PV1|||^^^CARE
HEALTH SYSTEMS^^^^^||| |1173^MATTHEWS^JAMES^A^^^||||||||||||
610613|||||||||||||||||||||||||||||V ORC|NW|987654^EPIC|
76543^EPC||Final||^^^20140418170014^^^^||20140418173314|
1148^PATTERSON^JAMES^^^^||1173^MATTHEWS^JAMES^A^^^|
1133^^^222^^^^^|(618)222-1122|| OBR|1|363463^EPC|1858^EPC|
73610^X-RAY ANKLE 3+ VW^^^X-RAY ANKLE ||||||||||||
1173^MATTHEWS^JAMES^A^^^|(608)258- 8866|||||||Final||
^^^20140418170014^^^^||||6064^MANSFIELD^JEREMY^^^^||
1148010^1A^EAST^X-RAY^^^|^| DG1||I10|S82^ANKLE FRACTURE^I10|ANKLE
FRACTURE||
```

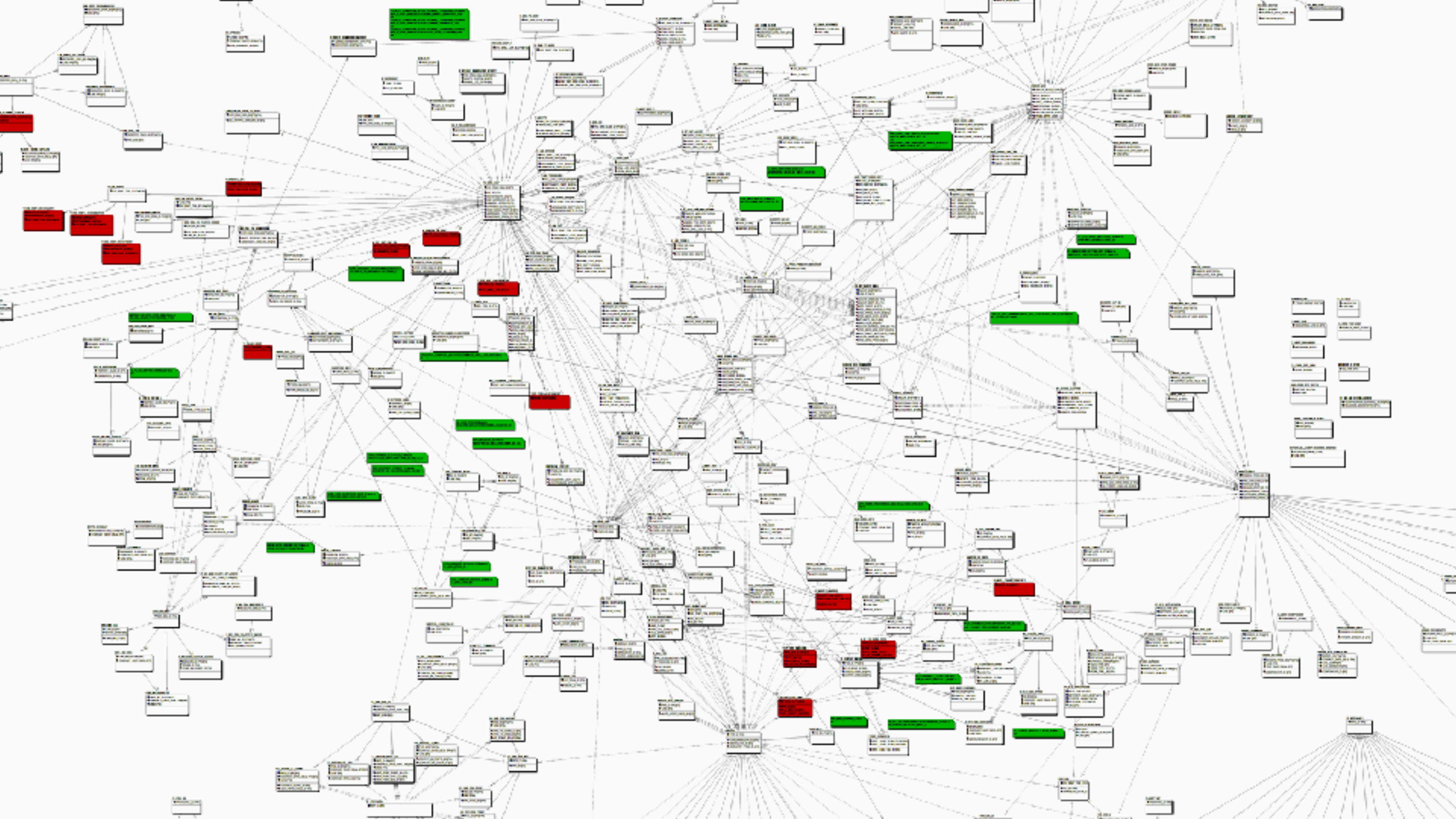
HL7

```
MSH|^~\&|EP                                                    2.3|| PID|
1||20891312                                                   |AfrAm|505
S. HAMILTON                                                   4567|
(608)123-56                                                   ||
FACILITY(EA                                                   ^CARE
HEALTH SYST                                                   |||||
610613||||                                                    PIC|
76543^EPC||                                                   |
1148^PATTER
1133^^^222^                                                   PC|
73610^X-RAY
1173^MATTHE
^^^20140418
1148010^1A^                                                   ^I10|ANKLE
FRACTURE||
```



“Enterprise” Clinical Databases






My eyes! The goggles do nothing!

Health Care Data Issues

Inconsistent Data Types

Inconsistent Data Types

- unable to take temperature, patient just ate ice

Inconsistent Data Types

- unable to take temperature, patient just ate ice
- Pt c/o feeling hot, "like I have a fever". Deferred temp reassessment for 10min-pt recently sipped ice water

Inconsistent Data Types

- **unable to take temperature, patient just ate ice**
- **Pt c/o feeling hot, "like I have a fever". Deferred temp reassessment for 10min-pt recently sipped ice water**

An urge to eat non-nutritive substances, a condition called pica, **can** occur in **pregnant** women with low iron levels. On the other hand, **ice cravings**, known medically as pagophagia, don't necessarily **mean you** have a nutrient deficiency, which a blood test **can** diagnose. Apr 15, 2015

Inconsistent Data Types

- unable to take temperature, patient just ate ice
- Pt c/o feeling hot, "like I have a fever". Deferred temp reassessment for 10min-pt recently sipped ice water

> An urge to eat non-nutritive substances, a condition called pica, **can** occur in **pregnant** women with low iron levels. On the other hand, **ice cravings**, known medically as pagophagia, don't necessarily **mean you** have a nutrient deficiency, which a blood test **can** diagnose. Apr 15, 2015

- unable to check temp with 3 different thermometers -- pt skin warm and dry -- bear hugger applied and will recheck temp soon
- BEAR HUGGER APPLIED

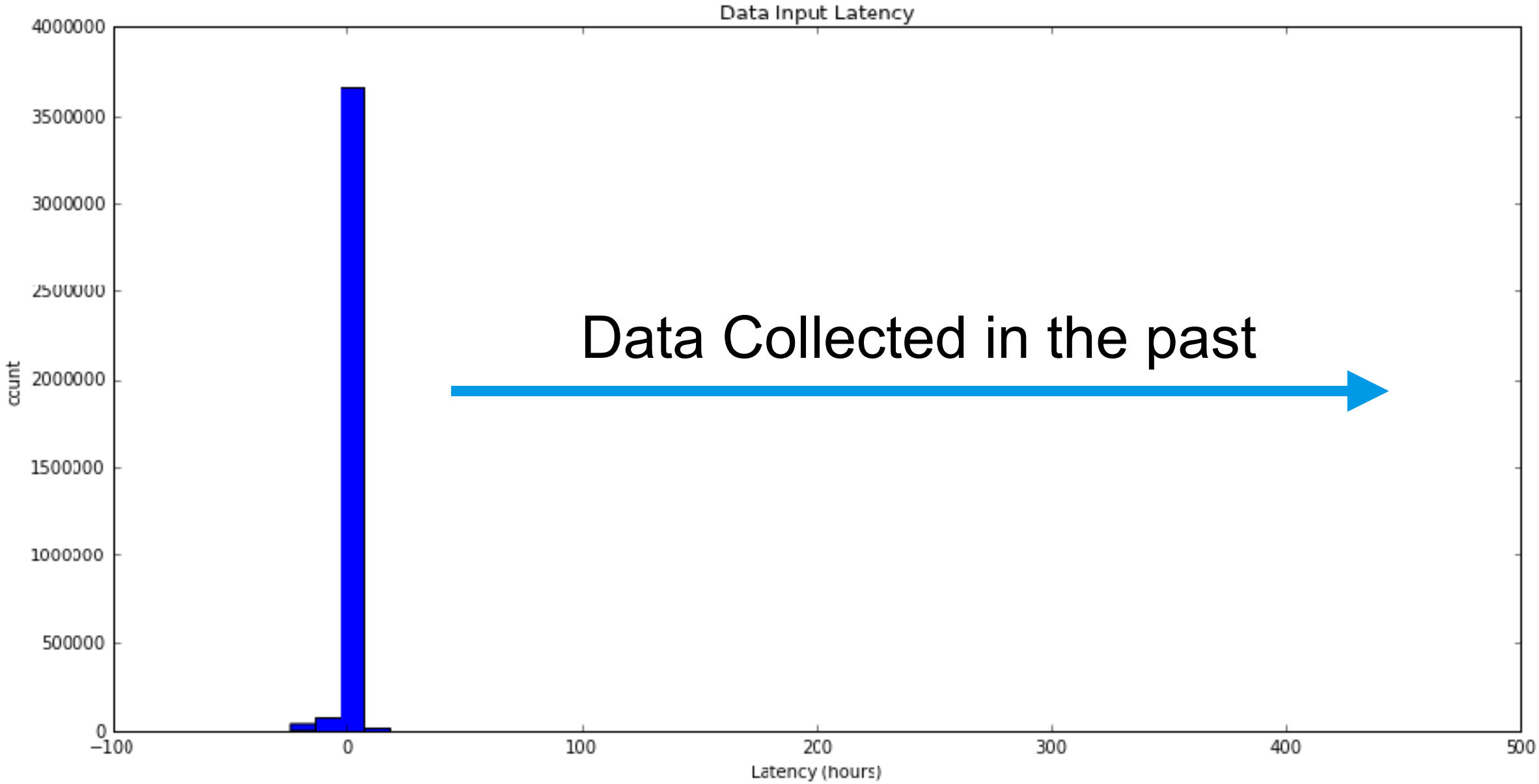
Inconsistent Data Types

- unable to take temperature, patient just ate ice
- Pt c/o feeling hot, "like I have a fever". Deferred temp reassessment for 10min-pt recently sipped ice water

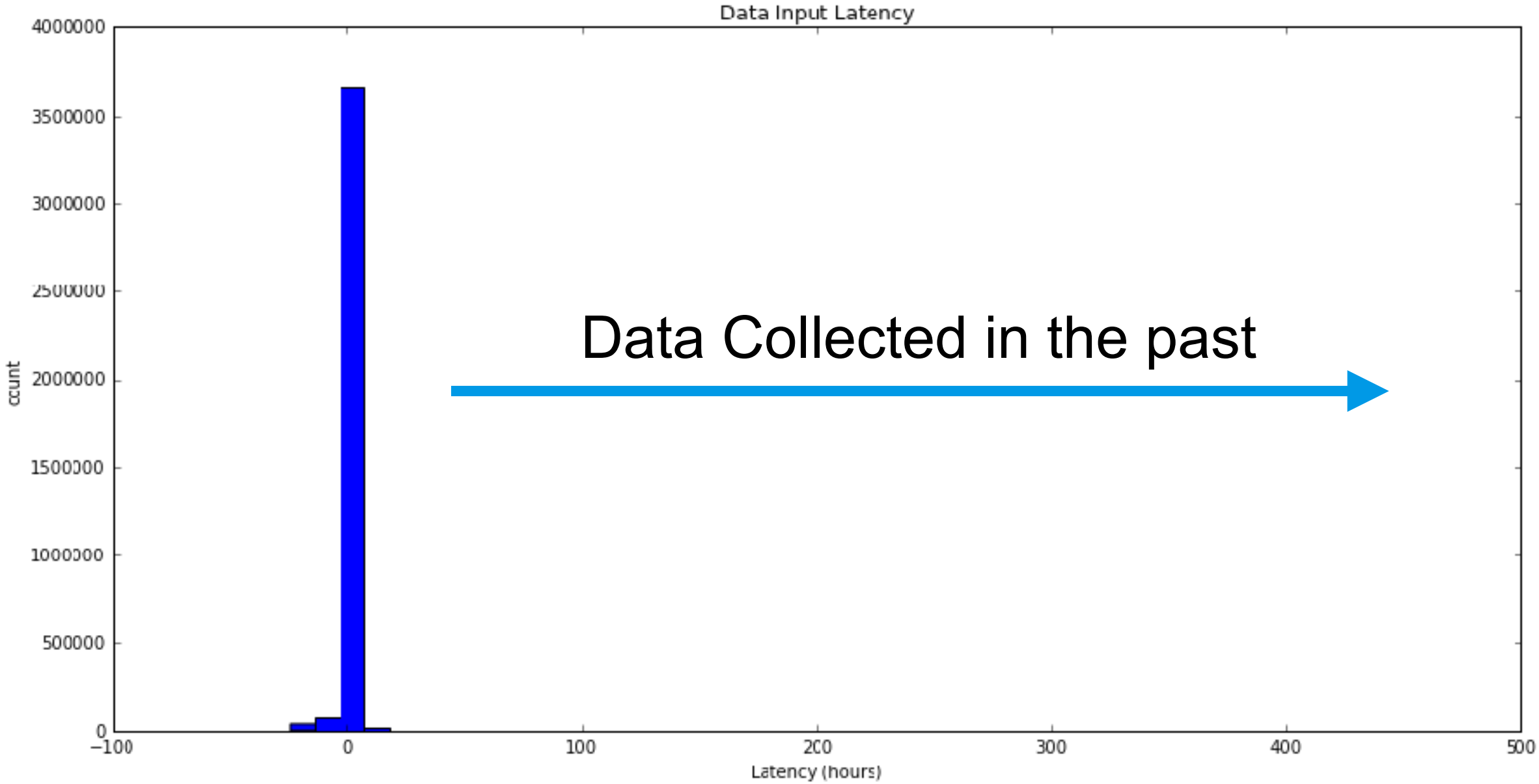
> An urge to eat non-nutritive substances, a condition called pica, **can** occur in **pregnant** women with low iron levels. On the other hand, **ice cravings**, known medically as pagophagia, don't necessarily **mean you** have a nutrient deficiency, which a blood test **can** diagnose. Apr 15, 2015

- unable to check temp with 3 different thermometers -- pt skin warm and dry -- bear hugger applied and will recheck temp soon
- BEAR HUGGER APPLIED

Future Data


Data Input Latency
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
count
−100
0
100
200
300
400
500
Latency (hours)
Data Collected in the past

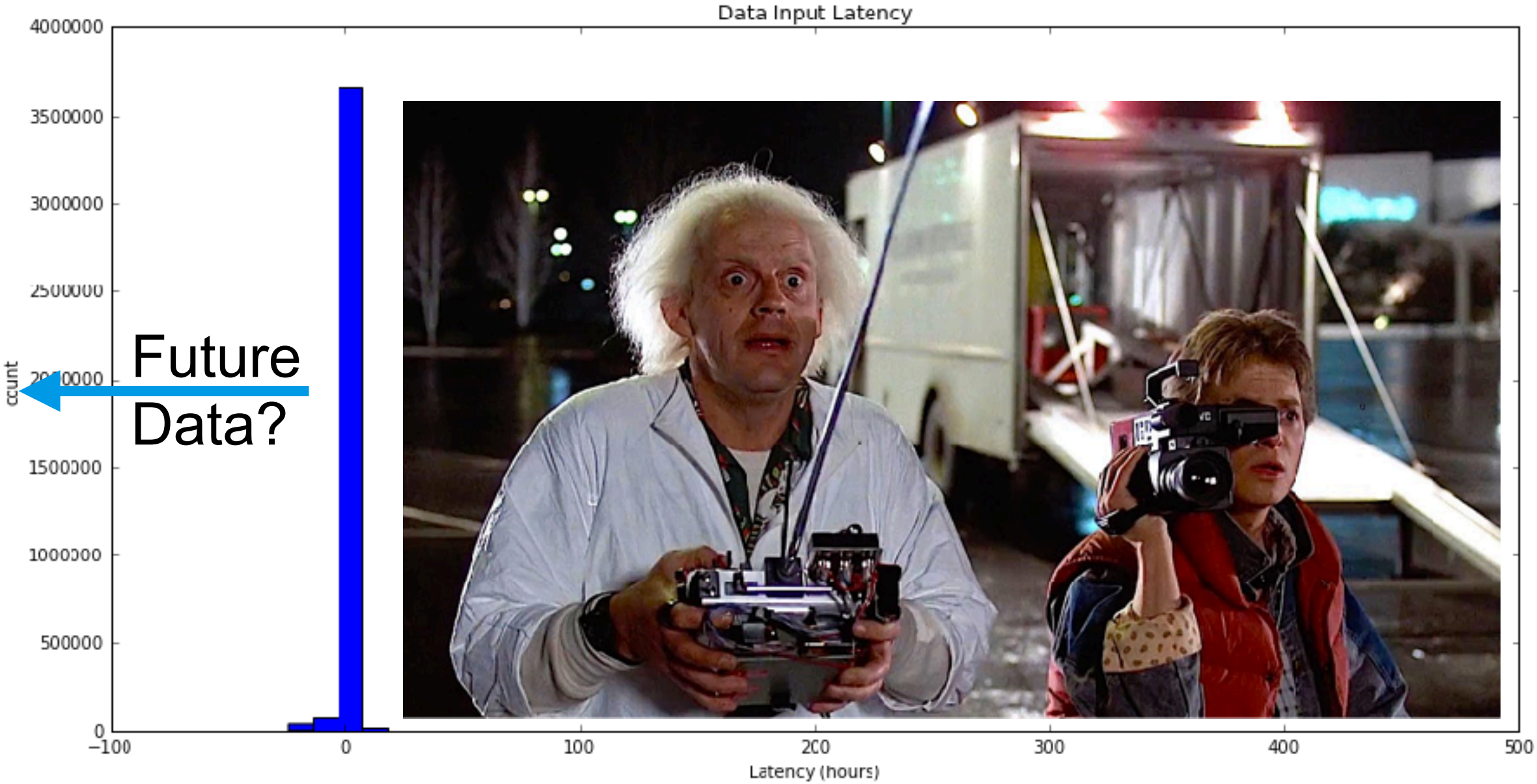
Future Data


Data Input Latency
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
count
−100
0
100
200
300
400
500
Latency (hours)
Data Collected in the past

Future Data


Data Input Latency
4000000
3500000
3000000
2500000
1500000
1000000
500000
0
count
−100
0
100
200
300
400
500
Latency (hours)
Future Data?
Data Collected in the past

Future Data


Data Input Latency
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
count
−100
0
100
200
300
400
500
Latency (hours)
Future Data?

More Data Issues

- **Multiple identifiers for the same patient**

More Data Issues

- Multiple identifiers for the same patient
- Throwing out historic data

More Data Issues

- Multiple identifiers for the same patient
- Throwing out historic data
- Data collection biases

More Data Issues

- Multiple identifiers for the same patient
- Throwing out historic data
- Data collection biases
- Inconsistent database models

More Data Issues

- Multiple identifiers for the same patient
- Throwing out historic data
- Data collection biases
- Inconsistent database models
- Mostly human entered data… in a time crunch data is lost

More Data Issues

- Multiple identifiers for the same patient
- Throwing out historic data
- Data collection biases
- Inconsistent database models
- Mostly human entered data… in a time crunch data is lost
- Little validation of input

More Data Issues

- Multiple identifiers for the same patient
- Throwing out historic data
- Data collection biases
- Inconsistent database models
- Mostly human entered data… in a time crunch data is lost
- Little validation of input

Little understanding of business impact

Penn Signals

Penn Signals

- A single data schema for data in stream and at rest

Penn Signals

- A single data schema for data in stream and at rest
  - Enables fast transition from modeling to applications

Penn Signals

- A single data schema for data in stream and at rest
  - Enables fast transition from modeling to applications
- Storing event stream allows you to recreate the state of a patient at any time in their stay

Penn Signals

- A single data schema for data in stream and at rest
  - Enables fast transition from modeling to applications
- Storing event stream allows you to recreate the state of a patient at any time in their stay
  - Allows for deterministic modeling

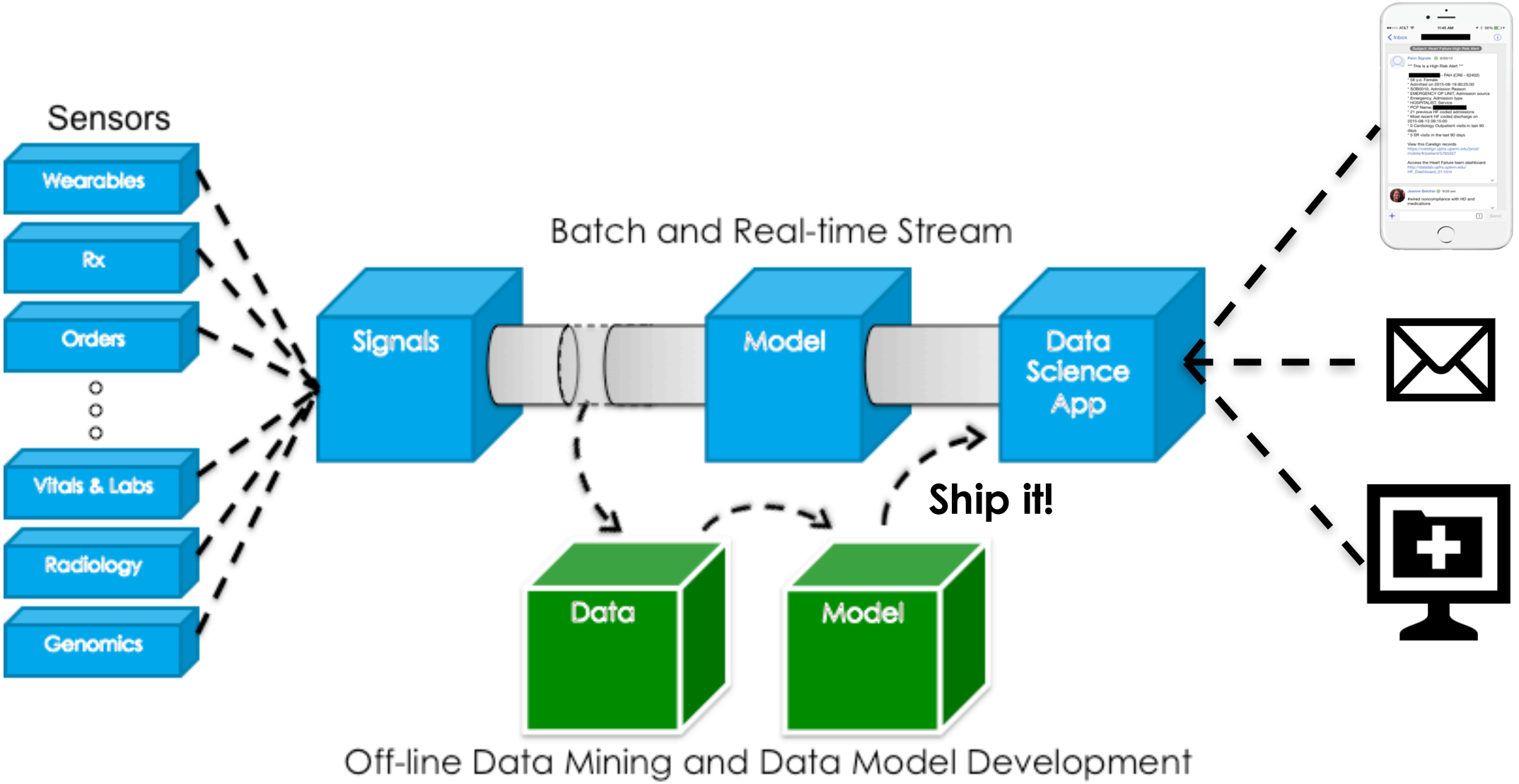
Penn Signals

- A single data schema for data in stream and at rest
  - Enables fast transition from modeling to applications
- Storing event stream allows you to recreate the state of a patient at any time in their stay
  - Allows for deterministic modeling
  - Crucial for real-time applications

Penn Signals

- A single data schema for data in stream and at rest
  - Enables fast transition from modeling to applications
- Storing event stream allows you to recreate the state of a patient at any time in their stay
  - Allows for deterministic modeling
  - Crucial for real-time applications
- Type enforcement

Penn Signals Architecture



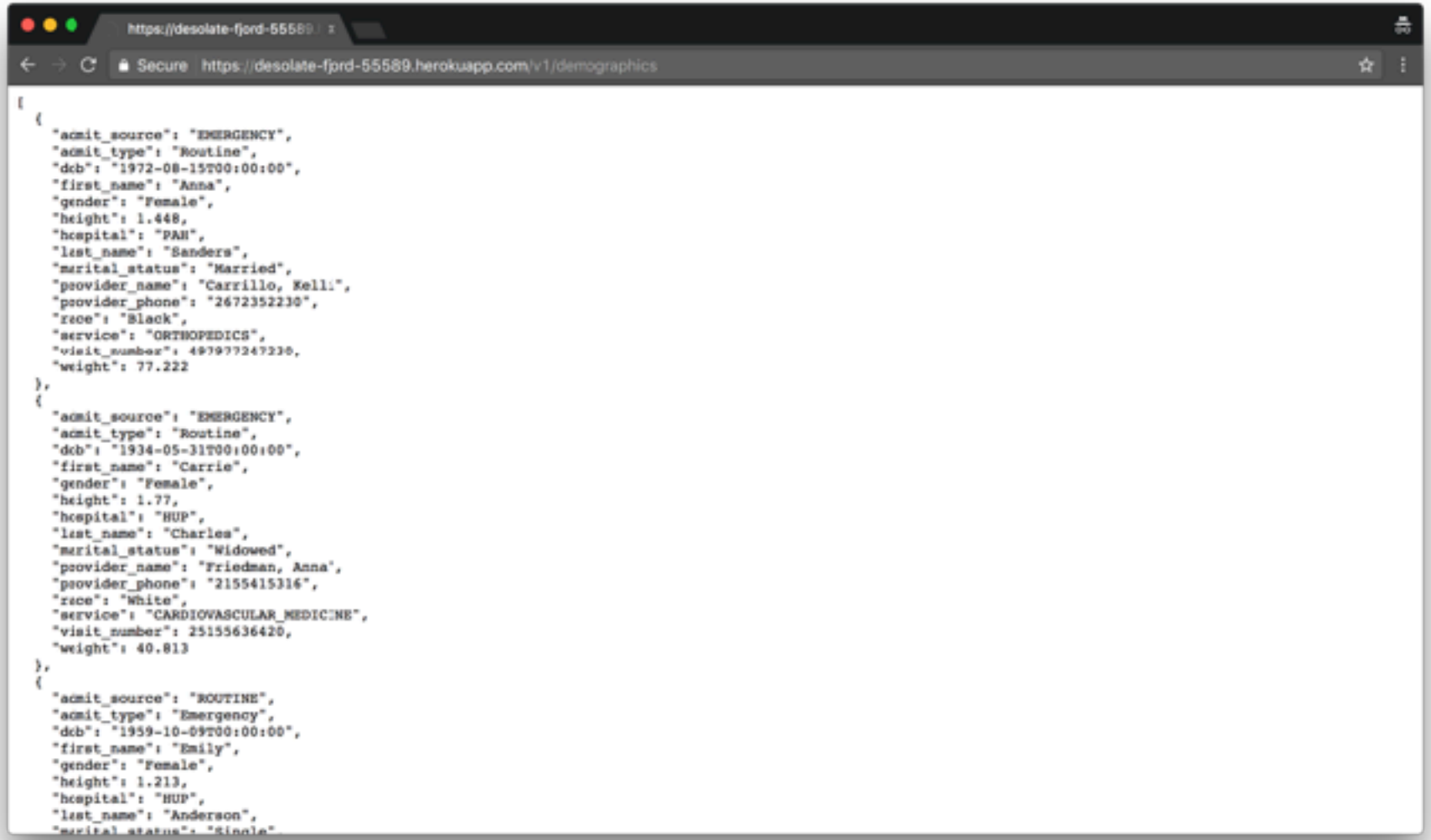
Penn Signals Stack

Upstream Resources

- Enterprise data warehouses
- Web services
- Streaming telemetry data

fuzzy-engine

- https://github.com/pennsignals/fuzzy-engine
- returns JSON formatted HTTP responses containing randomized patient data



Can we validate upstream data?

```
{
    "id": 1,
    "name": "A green door",
    "price": 12.50,
    "tags": ["home", "green"]
}
```

🤔🤔🤔🤔

JSON Schema Vocabulary

```
{
    "$schema": "http://json-schema.org/draft-04/schema#",
    "title": "Product",
    "description": "A product from Acme's catalog",
    "type": "object",
    "properties": {
        "id": {
            "description": "The unique identifier for a product",
            "type": "integer"
        },
        "name": {
            "description": "Name of the product",
            "type": "string"
        },
        "price": {
            "type": "number",
            "minimum": 0,
            "exclusiveMinimum": true
        },
        "tags": {
            "type": "array",
            "items": {
                "type": "string"
            },
            "minItems": 1,
            "uniqueItems": true
        }
    },
    "required": ["id", "name", "price"]
}
```

JSON Schema Validation Keywords

strings

- `maxLength`
- `minLength`
- `pattern`

arrays

- `items`
- `additionalItems`
- `maxItems`
- `minItems`
- `uniqueItems`

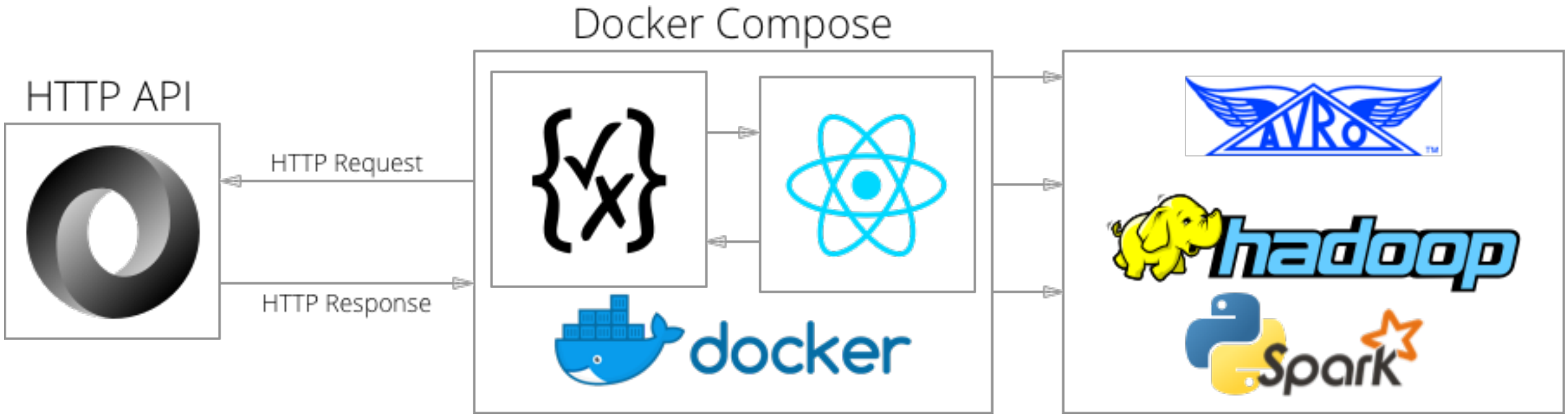
That’s cool, but are there documents that describe web services?

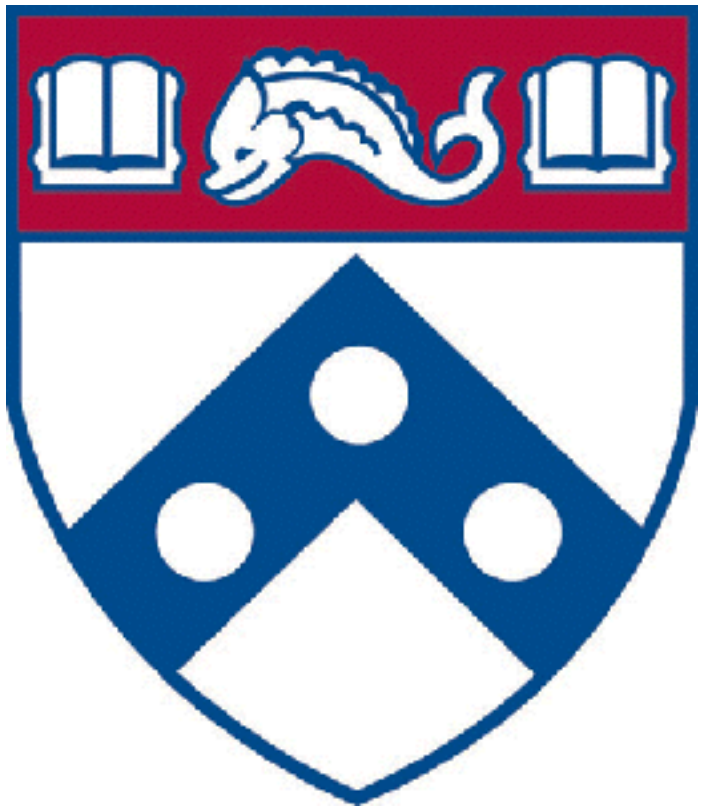
YES!

http://swagger.io/

Putting it all together

Process


Docker Compose
HTTP API
HTTP Request
HTTP Response
docker
AVRO
hadoop
Spark



Data Scientist or Data Janitor?

Big Data Borat

@BigDataBorat

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

Retweets	Likes
509	281

About us

- Twitter @beckerfuffle & @rightlag
- Follow our team @PennDataScience
- Updates on Penn Signals can be found at predictivehealthcare.pennmedicine.org