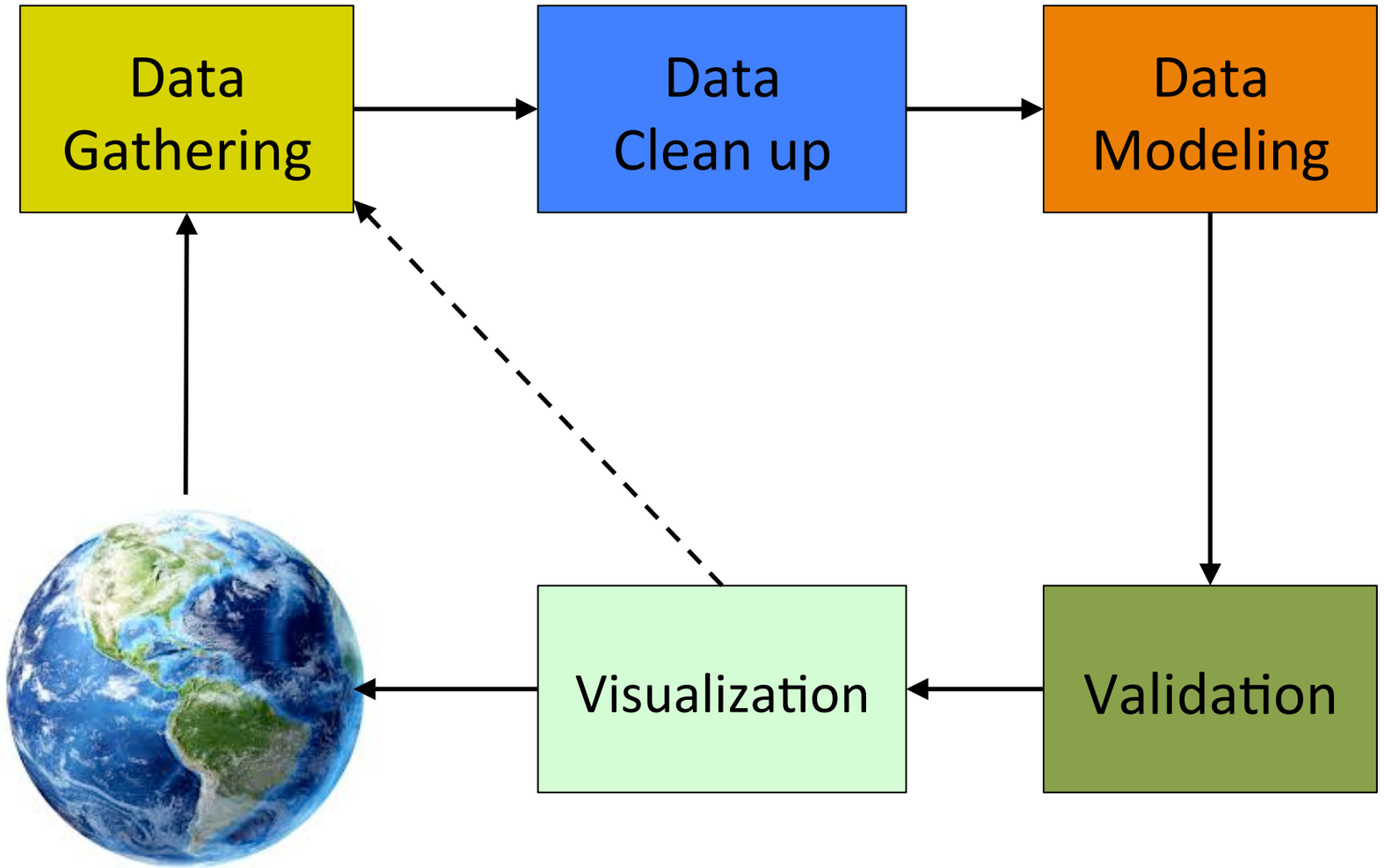
GENERAL ASSEMBLY

DATA SCIENCE

DATA PREPROCESSING, FEATURE SELECTION

GENERAL ASSEMBLY

DATA MODELING LIFE CYCLE



1. Gather Data
2. Select Data
3. Preprocess Data
4. Transform Data
5. Model Data
6. Evaluate Data
7. Visualize Data

GENERAL ASSEMBLY

The following methodology is from Chapter 2 (by Kevin Fink) of “Bad Data Handbook,” edited by Q. Ethan McCallum.

The principles here are commonly used. For practice, see:

- “Using a spreadsheet to clean up a dataset”, http://schoolofdata.org/handbook/recipes/cleaning-data-with-spreadsheets/

GENERAL ASSEMBLY

METHODOLOGY FOR ASSESSING DATA

1. UNDERSTAND THE DATA STRUCTURE

2. VALIDATE THE FIELDS

3. VALIDATE THE VALUES

4. INTERPRET COLUMNS VIA SIMPLE STATISTICS

5. VISUALIZE COLUMNS VIA HISTOGRAMS

source: Q. Ethan McCallum, “BAD DATA HANDBOOK”

GENERAL ASSEMBLY

1. UNDERSTAND THE DATA STRUCTURE

Ensure the data is in a common format that you can read in.

The most common formats are:

- Columnar
- XML
- JSON
- Excel

```
waco tourism, $0.99
calibre cpa, $1.99,,,,,
c# courses,$2.99 ,,,,,
cad computer aided dispatch, $1.49 ,,,,,
cadre et album photo, $1.39 ,,,,,
cabana beach apartments san marcos, $1.09,,,
"chemistry books, a level", $0.99
cake decorating classes in san antonio, $1.59 ,,,,,
k & company, $0.50
p&o mini cruises, $0.99
c# data grid,$1.79 ,,,,,
advanced medical imaging denver, $9.99 ,,,,,
canadian commercial lending, $4.99 ,,,,,
cabin vacation packages, $1.89 ,,,,,
c5 envelope printing, $1.69 ,,,,,
mesothelioma applied research, $32.79 ,,,,,
ca antivirus support, $1.29 ,,,,,
"trap, toilet", $0.99
c fold towels, $1.19 ,,,,,
cabin rentals wa, $0.99
```

GENERAL ASSEMBLY

```
waco tourism, $0.99
calibre cpa, $1.99,,,,,
c# courses,$2.99 ,,,,,
cad computer aided dispatch, $1.49 ,,,,,
cadre et album photo, $1.39 ,,,,,
cabana beach apartments san marcos, $1.09,,,
"chemistry books, a level", $0.99
cake decorating classes in san antonio, $1.59 ,,,,,
k & company, $0.50
p&o mini cruises, $0.99
c# data grid,$1.79 ,,,,,
advanced medical imaging denver, $9.99 ,,,,,
canadian commercial lending, $4.99 ,,,,,
cabin vacation packages, $1.89 ,,,,,
c5 envelope printing, $1.69 ,,,,,
mesothelioma applied research, $32.79 ,,,,,
ca antivirus support, $1.29 ,,,,,
"trap, toilet", $0.99
c fold towels, $1.19 ,,,,,
cabin rentals wa, $0.99
```

GENERAL ASSEMBLY

2. VALIDATE THE FIELDS

- Understand the units or meaning of a field
  - Revenue: gross or net? distance - miles or km?

- Look at the values. Ensure they make sense in the context of the field.
  - Currency: should be decimals with 2-4 digits after the decimal.
  - User Agent: should be string.
  - IP address: should be integers/dotted quads.

- Look for missing/empty values.
  - NULL/undefined/NA/NaN/0.

GENERAL ASSEMBLY

3. VALIDATE THE VALUES

Write a script or use Excel to validate:

- For categorical/enumerable fields, do all values fall into the proper set?
  - Month field: should contain 0-11, 1-12
- For numeric fields, are all values numbers?
- + For fixed-format fields (e.g. emails, IP addresses), do all values match a regular expression?
  - IP address: ^(?:[0-9]{1,3}\.){3}[0-9]{1,3}$

GENERAL ASSEMBLY

```
waco tourism, $0.99
calibre cpa, $1.99,,,,,
c# courses,$2.99 ,,,,,
cad computer aided dispatch, $1.49 ,,,,,
cadre et album photo, $1.39 ,,,,,
cabana beach apartments san marcos, $1.09,,,
"chemistry books, a level", $0.99
cake decorating classes in san antonio, $1.59 ,,,,,
k & company, $0.50
p&o mini cruises, $0.99
c# data grid,$1.79 ,,,,,
advanced medical imaging denver, $9.99 ,,,,,
canadian commercial lending, $4.99 ,,,,,
cabin vacation packages, $1.89 ,,,,,
c5 envelope printing, $1.69 ,,,,,
mesothelioma applied research, $32.79 ,,,,,
ca antivirus support, $1.29 ,,,,,
"trap, toilet", $0.99
c fold towels, $1.19 ,,,,,
cabin rentals wa, $0.99
```

GENERAL ASSEMBLY

3. INTERPRET COLUMNS VIA SIMPLE STATS

Most stats packages automate most of these:

- For numeric fields, does min/max make sense in the context of the field?
  - min value of a counter should be at least 0,
  - max value of a counter should rarely if ever be in billions
  - ratios such as CTR should generally not exceed 1

- Do financial values have reasonable upper bounds?
  - For PPC/CPC, hundreds of dollars would be the upper limit

- Does the average value of a field (or median/mode) make sense?
  - If the sale price is $10, but the average sale is $999, there may be something wrong. In this case, is the right unit dollars or cents?

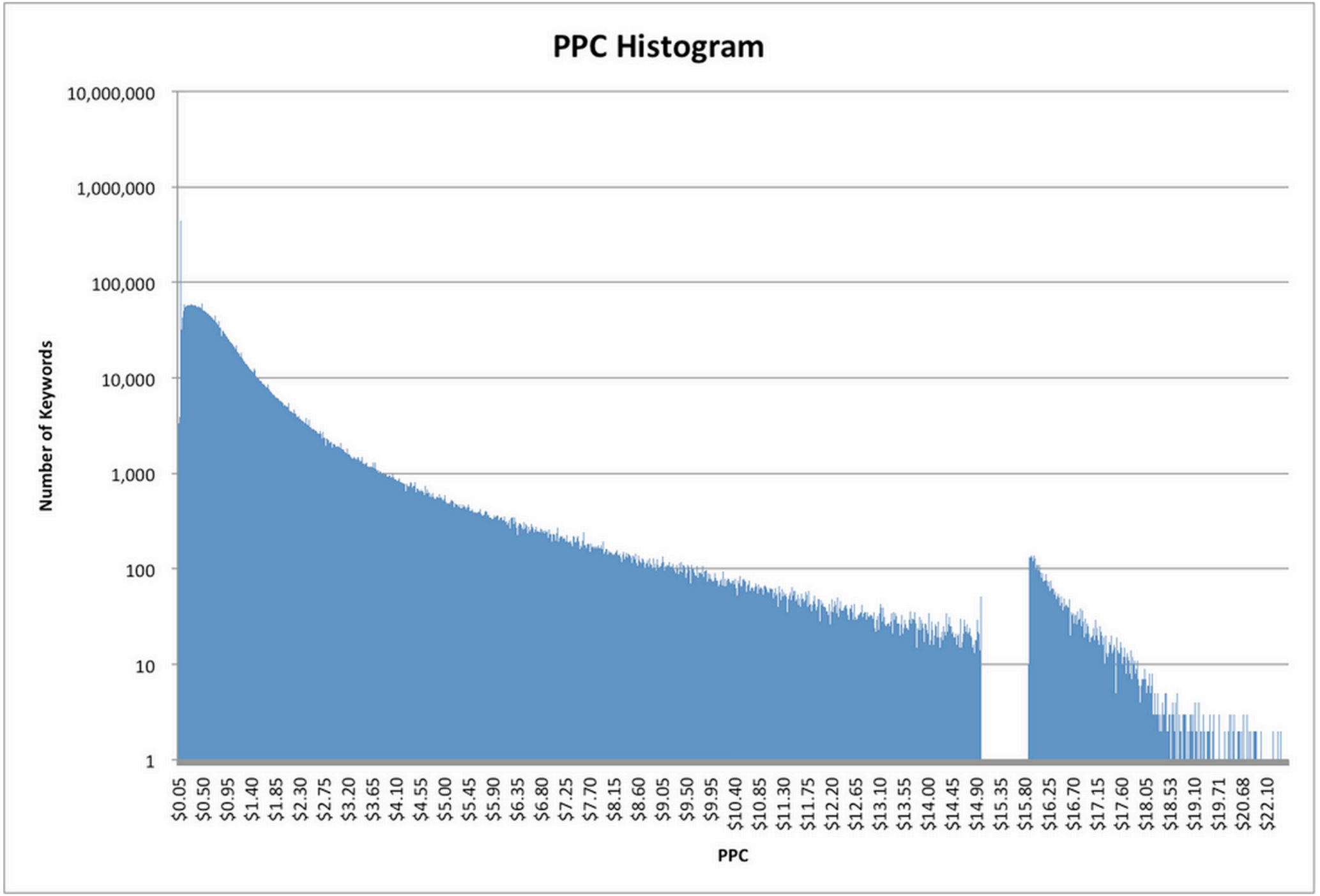
GENERAL ASSEMBLY

4. VISUALIZE COLUMNS VIA HISTOGRAMS

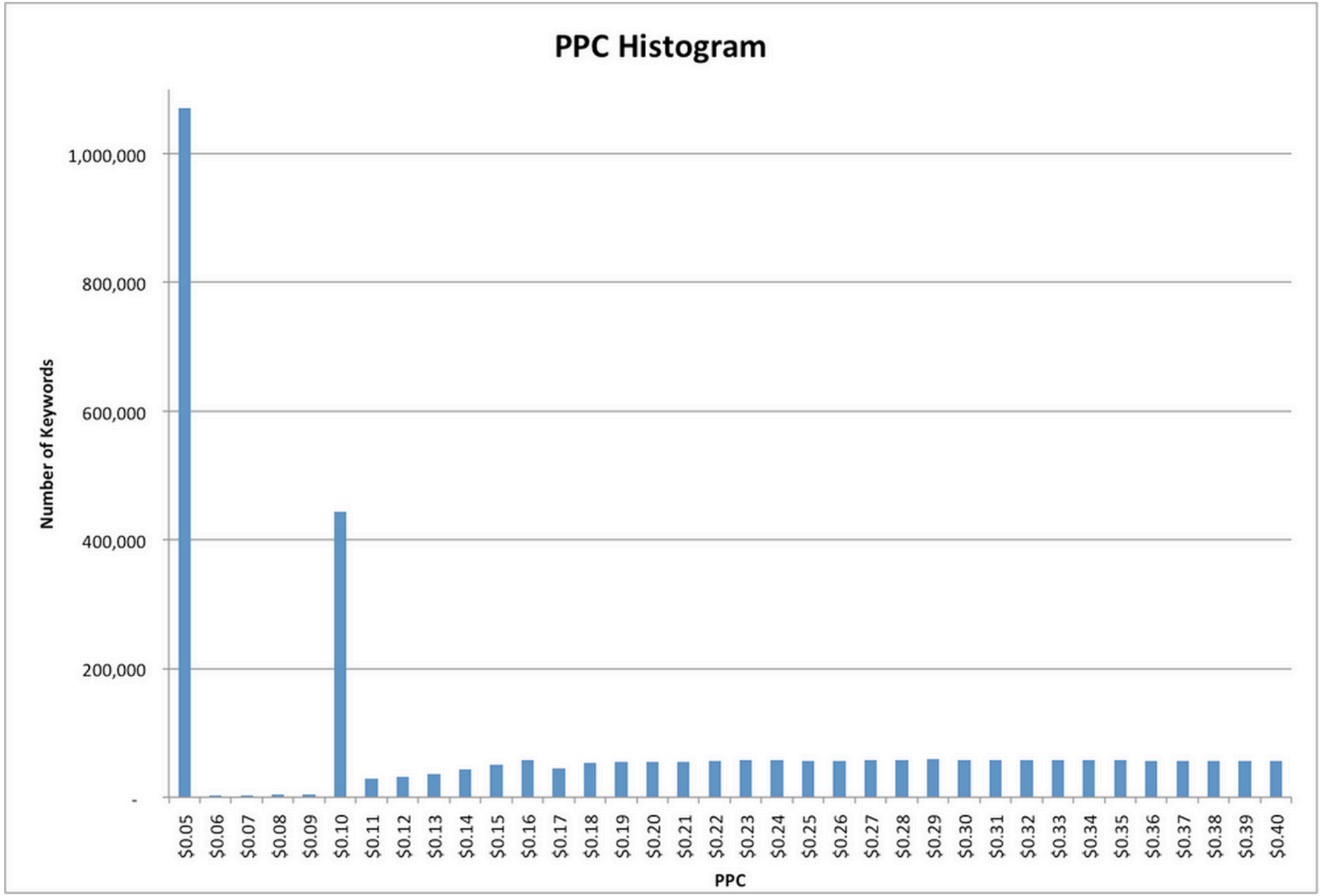
- Create a histogram of the values in a field.
  - Suppose data is referral keywords with billions of keywords. This cannot be analyzed using min/max/average, but a histogram can summarize it. (# referrals per keyword, or by putting them into bins 1-10 referrals, 11-20 referrals, etc.)

- Histograms also provide approximations of the distribution function -- may be flat, Gaussian, or have a long tail. But, a discontinuity may indicate a problem with the data.

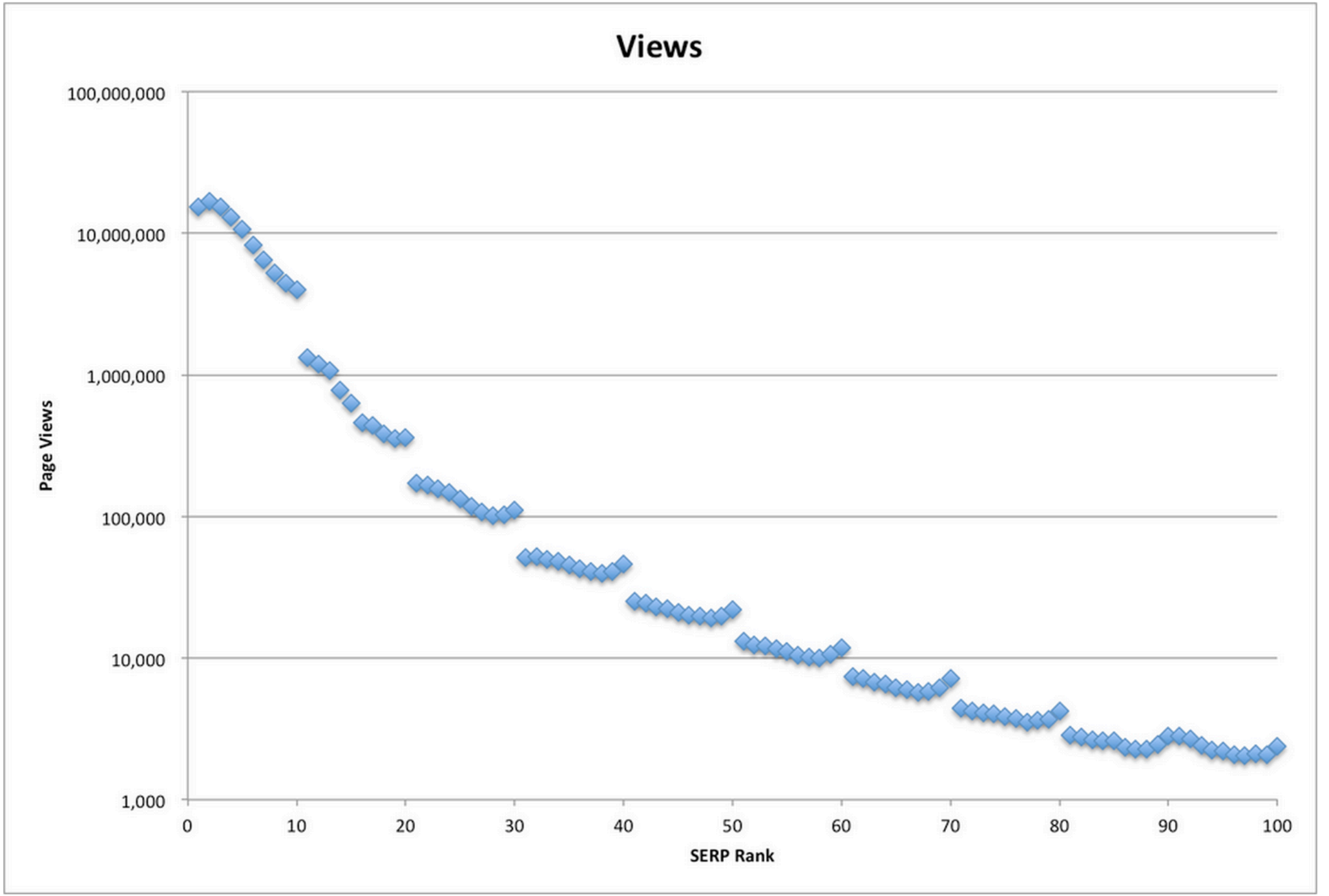
GENERAL


PPC Histogram
Number of Keywords
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
$0.05
$0.50
$0.95
$1.40
$1.85
$2.30
$2.75
$3.20
$3.65
$4.10
$4.55
$5.00
$5.45
$5.90
$6.35
$6.80
$7.25
$7.70
$8.15
$8.60
$9.05
$9.50
$9.95
$10.40
$10.85
$11.30
$11.75
$12.20
$12.65
$13.10
$13.55
$14.00
$14.45
$14.90
$15.35
$15.80
$16.25
$16.70
$17.15
$17.60
$18.05
$18.53
$19.10
$19.71
$20.68
$22.10
PPC

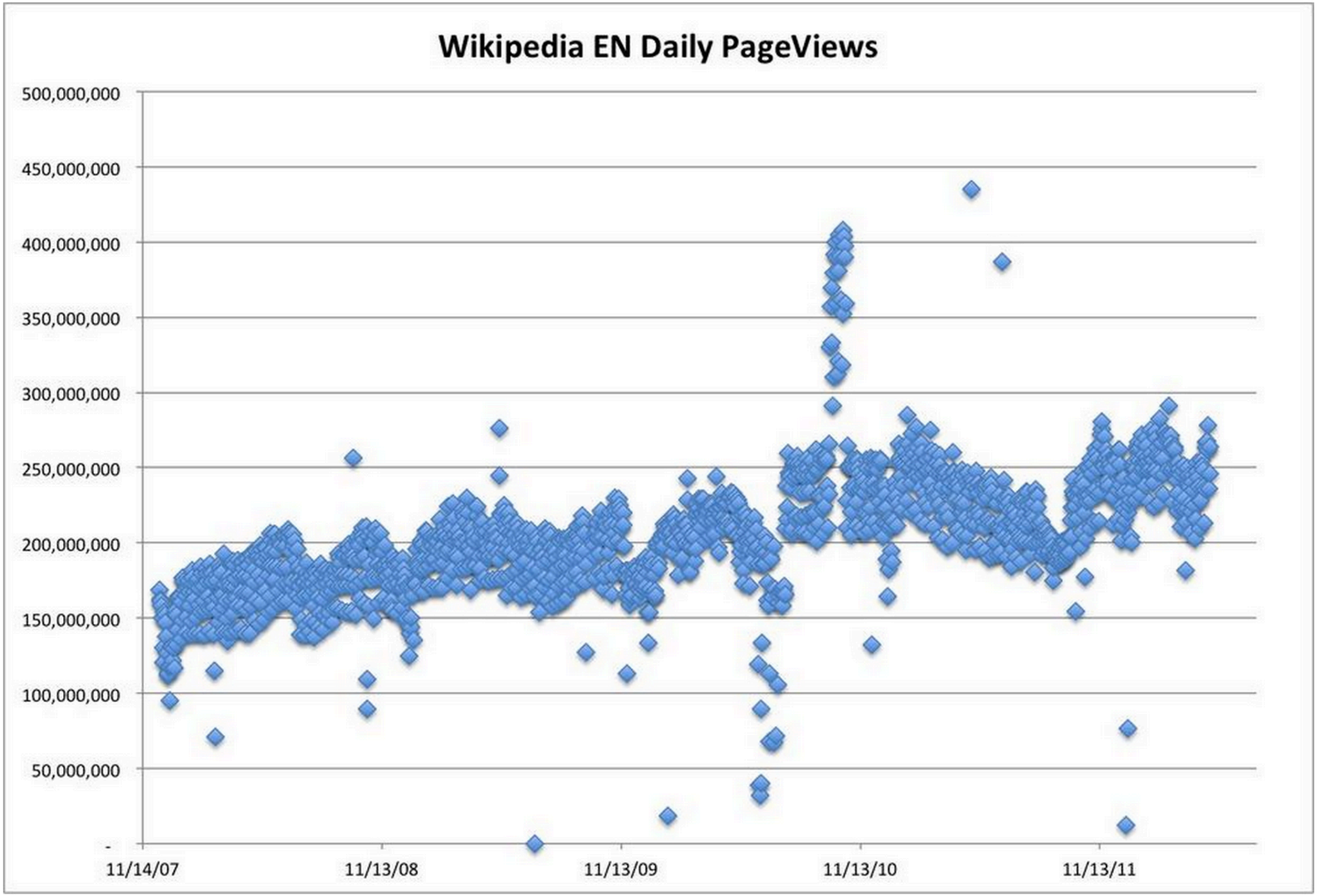
GA GENERAL A


PPC Histogram
Number of Keywords
1,000,000
800,000
600,000
400,000
200,000
-
$0.05
$0.06
$0.07
$0.08
$0.09
$0.10
$0.11
$0.12
$0.13
$0.14
$0.15
$0.16
$0.17
$0.18
$0.19
$0.20
$0.21
$0.22
$0.23
$0.24
$0.25
$0.26
$0.27
$0.28
$0.29
$0.30
$0.31
$0.32
$0.33
$0.34
$0.35
$0.36
$0.37
$0.38
$0.39
$0.40
PPC

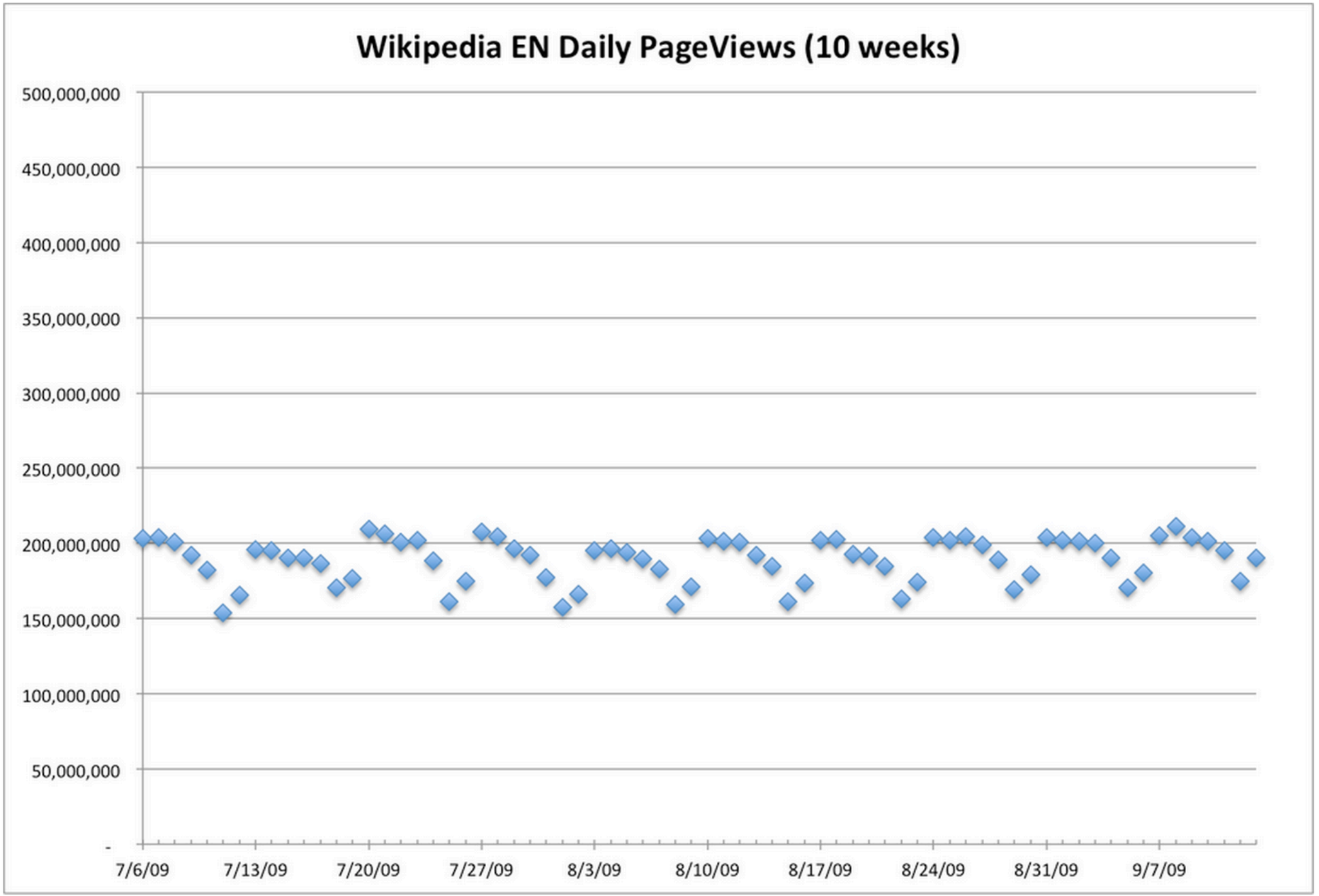
GA GENERAL


Views
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
Page Views
0
10
20
30
40
50
60
70
80
90
100
SERP Rank

GENER


Wikipedia EN Daily PageViews
500,000,000
450,000,000
400,000,000
350,000,000
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
-
11/14/07
11/13/08
11/13/09
11/13/10
11/13/11

GA GENE


Wikipedia EN Daily PageViews (10 weeks)
500,000,000
450,000,000
400,000,000
350,000,000
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
-
7/6/09
7/13/09
7/20/09
7/27/09
8/3/09
8/10/09
8/17/09
8/24/09
8/31/09
9/7/09

GENERAL ASSEMBLY

DATA PREPARATION

1. SAMPLE LARGE DATASETS

2. INFER MISSING DATA

3. STANDARDIZE OR NORMALIZE NUMERIC VALUES

4. REDUCE DIMENSIONALITY

"Predictive Analytics: Data Preparation",
http://horicky.blogspot.com/2012/05/predictive-analytics-data-preparation.html

GENERAL ASSEMBLY

STANDARDIZE OR NORMALIZE NUMERIC VALUES

Many machine learning algorithms are affected by scale. Standardizing/normalizing your features ensures they are all on the same scale and weighted evenly in the resulting model.

Normalization. Rescale to the range [0, 1].
This is great for pixel values and neural network inputs.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization (Z-Score Normalization). Ensures the feature data has a mean of 0 and standard deviation of 1.
This is a requirement for many learning algorithms, and it minimizes scaling effects. Preferred to do before PCA if the variables are not the same scale, since PCA relies on maximizing variance.

$$z = \frac{x - \mu}{\sigma}$$

GENERAL ASSEMBLY

SKLEARN EXAMPLE

```
from sklearn import preprocessing

std_scale = preprocessing.StandardScaler().fit(X)
X_std = std_scale.transform(X)

minmax_scale = preprocessing.MinMaxScaler().fit(X)
X_minmax = minmax_scale.transform(X)
```

See http://sebastianraschka.com/Articles/2014_about_feature_scaling.html for a great walkthrough!

GENERAL ASSEMBLY

DATA PREPARATION

5. ADD DERIVED ATTRIBUTES

6. DISCRETIZE NUMERIC VALUES INTO CATEGORIES

7. BINARIZE CATEGORICAL ATTRIBUTES

8. SELECT, COMBINE, AGGREGATE DATA

9. TRANSFORM DATA TO BE NORMALLY DISTRIBUTED

“Predictive Analytics: Data Preparation”,
http://horicky.blogspot.com/2012/05/predictive-analytics-data-preparation.html

GENERAL ASSEMBLY

FEATURE ENGINEERING

Feature

An attribute useful for your modeling task.

GENERAL ASSEMBLY

GROUP EXERCISE

Brainstorm multiple ways of "featurizing" each of the below. You may make assumptions about the question asked or intended model:

1. **A categorical attribute "Item_Color" that can be Red, Blue, or Unknown.**
2. **A Date-Time item in ISO 8601 form: "2014-09-20T20:45:40Z"**
3. **A numerical quantity "Item_Weight" with value such as 6289.**