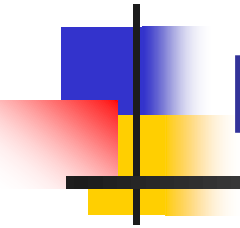
Advanced Web Analytics: Harnessing the Predictive Power



By Dr. Ash Pahwa

Lesson 1: Introduction to Analytics

Lesson 1.1: What is Predictive Analytics + CRISP/DM

Lesson 1.2: Data Problems in Predictive Analytics

Outline: Lesson 1.2: Data Problems in PA

- Data characteristics for PA
- Data Problems
- Web Analytics
- Google Analytics
- Key metrics

Data Characteristics for PA

- 90% of a successful outcome is contingent on having good data. What does "good" mean in this context?
  - **Reliable** - the effects can be reproduced.
    - Contrary example:
      - Using sales data from March-April to predict November-December data.
  - **Valid** - the data measures what you want it to measures.
    - Contrary example:
      - Number of page visits shows how popular the page is.
      - How well a job candidate does on a brainteaser indicates whether he will be a good hire.
  - Other data issues are small.

Data Issues

- **Missing data** - A small amount of missing data is not a problem, since the process of statistical modeling will smooth over these bumps.
  - However a large amount of missing data will invalidate the model
  - Missing not at random: this can be very serious.
  - Not to be confused with sparse data. Recommendation systems, such as Netflix, have sparse data.
- **Sparse data** - An example is a matrix with users on one dimension and movies on the other. The entries are ratings. Most people haven't seen most of the movies, so the matrix will be mostly empty.

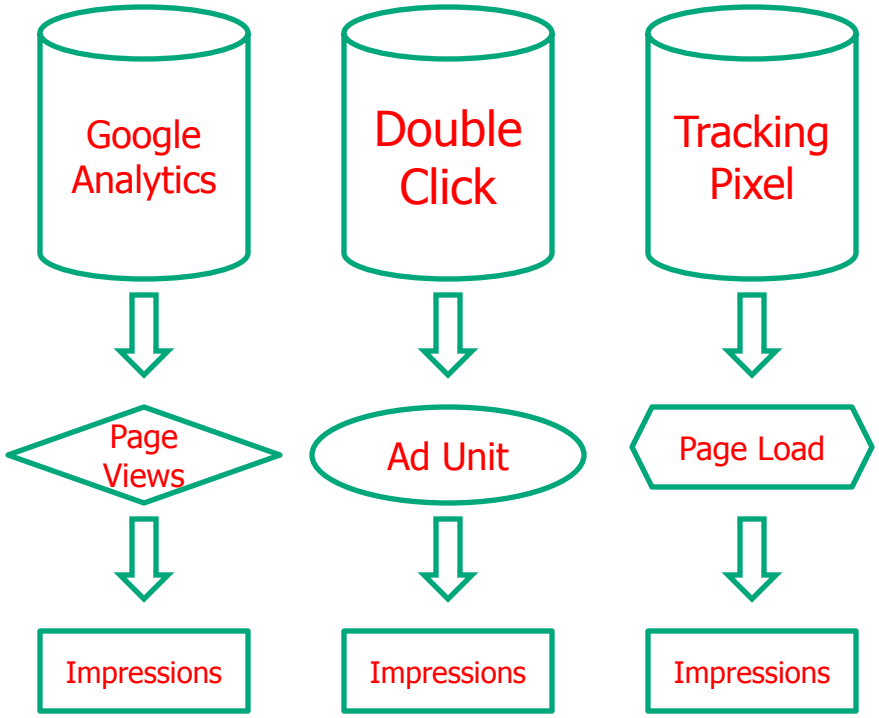
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Country	January	February	March	April	May	June	July	August	Septembe	October	Novembe	Decembe
2	Angola	11027	4608	9715	8381	5881	10286	9311	2775	3354	9935	7317	3973
3	Burundi	4988	6822			13264	4150	8469	2212	6620	7883	11637	10439
4	Chad	8711	8189	12772	13158	4491	2458	13498			7975	7945	8441
5	Congo	7139	13737	5152	12218		4390	6764		11791	12331	8443	5769
6	Egypt	9474	1278	8594	13938	1354	10004	10240	14054	2510	1360	3830	14432
7	Ethiopia	10590	10572	4933	6058	13121	10772	10602	11251	8648	9382	14768	14022
8	Gabon	7003	2462	5243	8851				13258	7038	14862	4153	9738
9	Ivory Coast				6354	12059	3591	5827	14469	13454	4326	11593	14240
10	Kenya	7234	9542	12112	14368	12704	10580	7558	5355	7864	11395	3114	8491
11	Libya	10099	11447	12909	7378	12713	13599	13203	8052	6800	12004	2028	6341
12	Madagasc	2013	6180		6785	13269	11403	5693	8438	8088	7647	5806	
13	Morocco	5463	10133	4515	6198	13884	14120	2120			5933	8445	12781
14	Namibia	7810	7507	9088	14838	12787	6350	7306	13090	1911	8260	12865	7507
15	Somalia	7505	13614	2231	5130	10979	12869	1449	4973	3609	14918	4638	3051
16	Swaziland	8035	12596	5948	6166	6198	13517	9700	9470	5804	9105	12671	3518
17	Zambia	7042	7553	6785	7141	6368	2492	2557	5078	14694	5739	4812	14291

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Country	January	February	March	April	May	June	July	August	Septembe	October	Novembe	December
2	Angola												
3	Burundi					11491							
4	Chad												
5	Congo												
6	Egypt												
7	Ethiopia												
8	Gabon												
9	Ivory Coast												
10	Kenya						4417						
11	Libya												
12	Madagascar												
13	Morocco										5284	9331	
14	Namibia												
15	Somalia												
16	Swaziland												
17	Zambia		2767										

Data Issues

Inaccurate data

- If a metric is measured by 3 different systems, you are likely to have 3 different values.
- It matters when the values are way off, or you have reason to believe the differences are systematic.
- This is a problem often faced by companies that run online advertising campaigns, and rely on both the publisher, the advertiser and ad exchanges to supply data



Data Issues

- **Nonstationary** (over time) data: When data doesn't have constant statistical properties over time, it is non-stationary.
- **Overfitting** to historical data - there is a balance to be had between fitting existing data and generalizing future non-observed events. Even regular validation steps in model building cannot completely remove this effect
- Relying on automated processes amplifies "**winner's curse**". Usually an effect of overfitting. The prediction from a model is more optimistic than what's actually observed. For example, the winner has almost certainly overpaid in an auction.
- **Complexity.** If something goes wrong in a complex model, harder to find the problem.

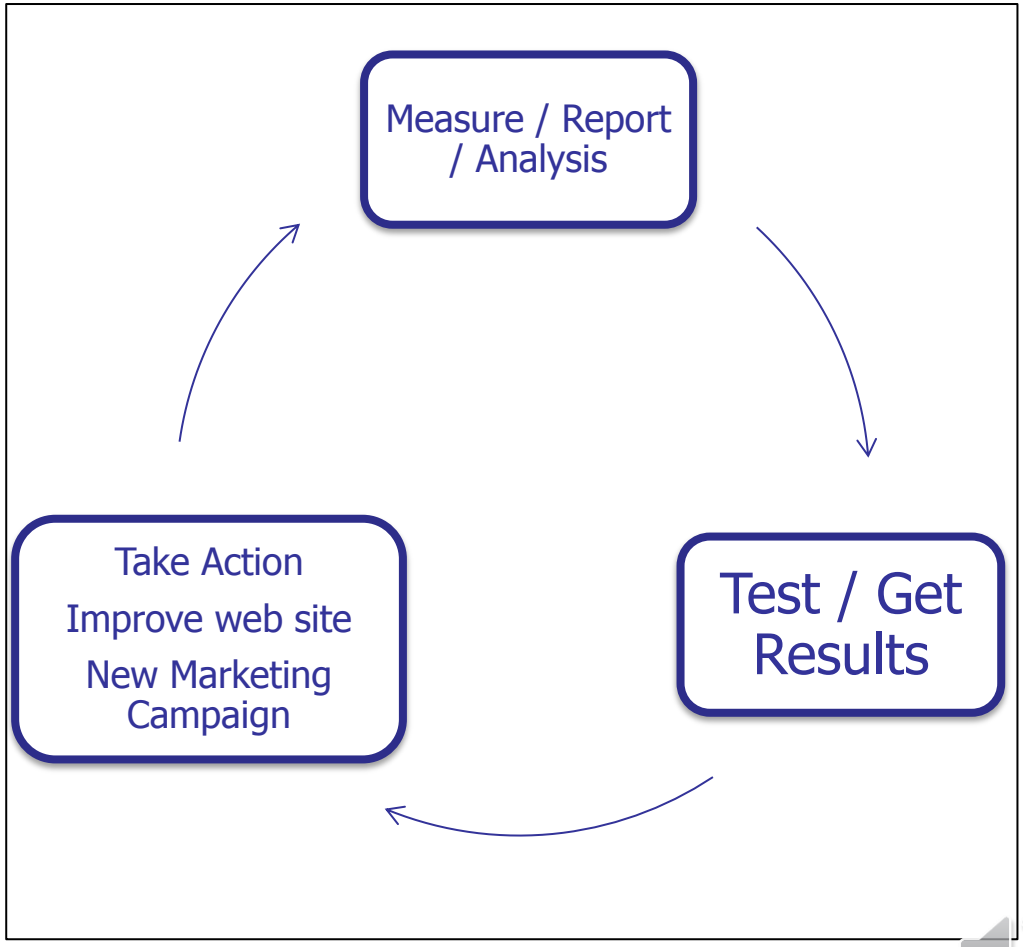
What is Web Analytics?

- Wikipedia definition
  - Web Analytics is the
    - measurement,
    - collection,
    - analysis and
    - reporting

    of internet data for purposes of understanding and optimizing web usage.
  - Web analytics is not just a tool for measuring website traffic
    - but can be used as a tool for
      - Business research
        - Which products are selling best
      - Market research
        - Which marketing campaigns are most effective

Web Analytics is a Continuous Process

- Web Analytics
- Continuously repeat the following actions
  - Measure
  - Report
  - Analysis
  - Test
  - Improve



Google Analytics

- Client side Data Collection
  - Page Tagging
- It's free
- It works
- 60% to 70% of all websites use Google Analytics

Google Analytics Competitors

Service		
Omniture	www.omniture.com	Owned by Adobe
CoreMetrics	www.coremetrics.com	Owned by IBM
Webside Story	Changed name to Visual Sciences Bought by Omniture 2007 Bought by Adobe 2009	
Web Trends	www.WebTrends.com	
Click Tracks	www.ClickTracks.com	
IndexTools	Yahoo Web Analytics Service	
Urchin	Previous version of Google Analytics product	
Gatineau	Microsoft unannounced Web Analytics service	

- Prices range from $200 to $2000 per month for these services

- Google Analytics is free
- Yahoo Web Analytics is free

Google Analytics Architecture

- When a visitor arrives at a website GATC embedded in a page is executed
- JavaScript code (18K) is downloaded
- First party cookie is created to identify a visitor

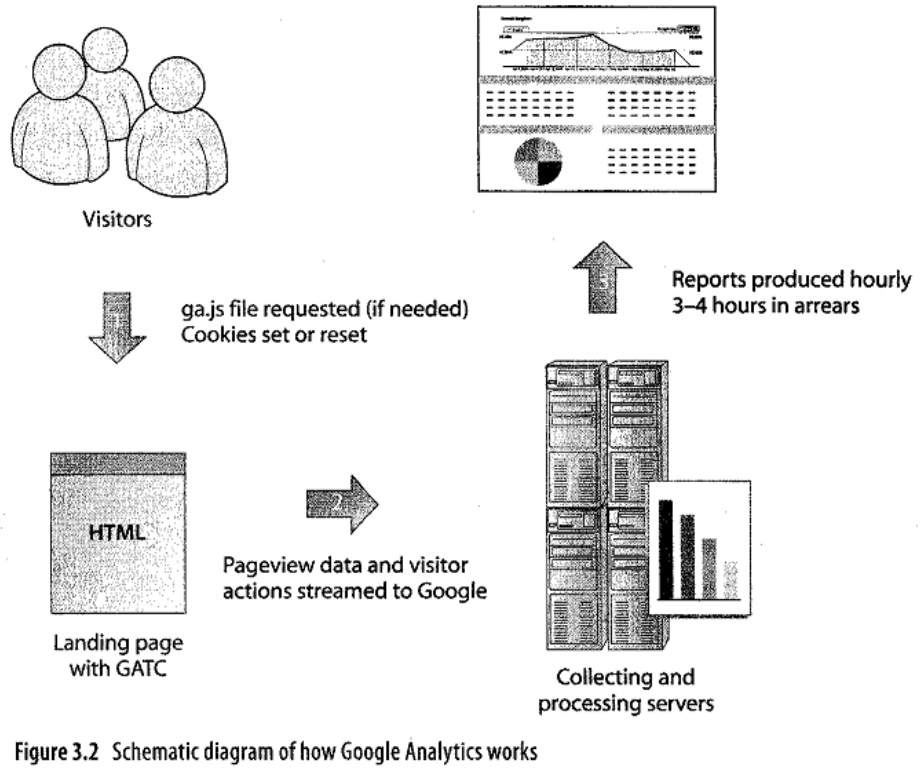


Figure 3.2 Schematic diagram of how Google Analytics works

- Each page view, GATC sends information to Google data collection service via a 1x1 pixel GIF image
- Each hour GA processes and collects data and updates your GA reports
- Reports are typically delayed by 3-4 hours

Key Metrics

- What should we measure and Why

  What advantages will I derive if I know those metrics

- Every Web Analytics Package provides many metrics

- We need to understand the semantics of those metrics
  1. Why are we measuring it
  2. What this number represents
  3. How can we use this number to improve our website or our business
  4. How is this number computed
  5. How accurate are those numbers

Criteria to Judge a Metric

- There are many metrics, which ones to use
- Evaluation Criteria
  - Simple
    - Easy to understand
  - Relevant
    - Related to your business
  - Timely
    - Can be easily computed within a short amount of time
  - Useful
    - Can be used on a daily basis

Key Metrics

- PageView
- Visit
- Visitors
- Unique pageviews
- Total visitors
- New versus returning visitors
- Time on page
- Time on site
- Length of a visit
- Average time on a page
- Average time on a site
- Flash based sites
- Visit duration
- Unique visitors
- Bounce rate
- Exit rate
- Conversion rate
- Traffic sources
  - Direct
  - Referring sites
  - Search Engines
- Good Traffic sources
- Keyword Report
- Keyword and Landing Pages
- Landing Pages
- Navigation summary
- Entrance Path Report
- Landing page / Entrance Report

Summary: Lesson 1.2: Data Problems in PA

- Data characteristics for PA
- Data Problems
- Web Analytics
- Google Analytics
- Key metrics