

Data Science and Technology Entrepreneurship

Data and Startups, Data Scoring Methods
Evaluating Startup Ideas

Sameer Maskey
Week 1

Course Information

- ▶ Data Science and Technology Entrepreneurship
- ▶ One of the first joint courses that brings together MBA students and CS/Engineering students in the same class
 - ▶ Computer Science Code - 6998-004
 - ▶ Business School Code - B8848-001
- ▶ Time : 4:10 to 6pm, Wednesday
- ▶ Room: Registrar assigned - 327 Mudd (small classroom)
- new room being assigned
- ▶ Next 3 weeks will be in Warren Hall

Course Information

- ▶ Short Introduction :
 - ▶ Sameer Maskey, PhD
 - ▶ PhD, 2008, Computer Science, Columbia University
 - ▶ Usually I teach
 - ▶ “Statistical Methods/Machine Learning for Natural Language Processing”
 - ▶ Machine Learning algorithms for language problems that use a lot of unstructured data
 - ▶ Speech to Speech Translation, Question Answering, Summarization
 - ▶ Founder, Machine Learning/Artificial Intelligence based startup
- ▶ Teaching Assistants :
 - ▶ Morgan Ulinski - mulinski@cs.columbia.edu (Computer Science)
 - ▶ Jigar Patel - jpatel13@gsb.columbia.edu (Business School)

Mentors/Advisors

1. Amol Sarva, Co-Founder, Peek, Virgin Mobile USA
2. Alok Ranjan, Co-Founder, CEO, ifood.tv
3. Ben Siscovick - General Partner, IA Ventures
4. Charlie O'Donnell - Partner, Brooklyn Bridge Ventures
5. David Lerner, Director, Columbia Venture Lab
6. Ella Gorgla, CEO, I-Ella.com
7. Jeb Miller, General Partner, Jafco Ventures
8. Jose Cabo - Founder, Olapic
9. James Wahba, Founder, Projective Space
10. Hrishu Dixit, CTO, LearnVest
11. Jerry Neumann, Partner, Neu Venture Capital
12. Kathryn Finney, Founder, Digital Undivided
13. Luis Sanz, Founder Olapic
14. Maryam Kamvar, Research Scientist, Google
15. Paul Tumpowsky, Chairman, InSITE
16. Sharib Khan, Co-Founder, Trial-X
17. Scott Ungerer, Founder and MD, EnerTech Capital
18. Shari Coulter Ford, Executive Director, NYC TechConnect
19. Stephen Messer, Vice-Chairman, Collective-i
20. Ted Shergalis, CSO, X+I
21. Wim Sweldens, Technology, Innovation & Business Leader

Goal of the Class

- ▶ Help you build a startup! (if it's possible in one semester?)
- ▶ Along the way show you how to use data science algorithms for your startup
- ▶ MBA + CS student teams
- ▶ 21 Mentors/Advisors will guide you through an entrepreneurial experience
- ▶ Participate in pitch days to get feedback
- ▶ Identify opportunity, Experiment, Build product, Validate customers, Iterate, Raise capital
- ▶ Visits to Incubation spaces

Teams

- ▶ Minimum (1 MBA + 1 CS)
- ▶ Maximum (2 MBA + 2 CS)
- ▶ If you want to build a team with more than 4 students please let me know

Course Stages

Stage 1 (3 weeks – Jan 30 – March Feb 20) Problem definition, Data collection, Customer development, Business Model Canvas, Data science methods for testing your hypothesis

Stage 2 (5 weeks – Feb 4 – March 10) Minimum Viable Product development, Quantifying customer feedback with classification and clustering techniques

Stage 3 (2 weeks – March 11 – March 31) Agile development, Data analysis of feature surveys, Sequential prediction algorithms (costs, revenue, traction)

Stage 4 (2 weeks – April 1 – April 29) Launching the product, Data driven marketing techniques, A/B testing

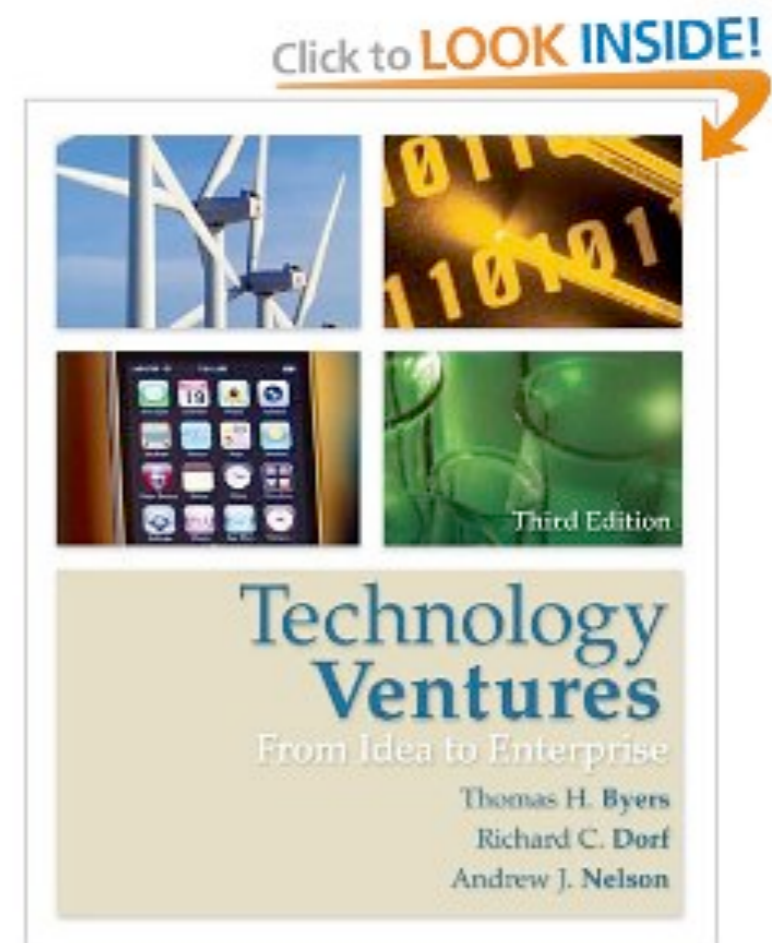
Stage 5 (2 weeks – April 1 – May 5) Try to raise funds with VC network provided in the class

Grading and Academic Integrity

- ▶ No midterms
- ▶ No Final Exams
- ▶ Short Assignments (15%)
- ▶ First Pitch Day Presentation (15%)
- ▶ Mid-semester Demo/Update Presentation (20%)
- ▶ Final Presentation/Demo/Pitch Presentation (45%)
- ▶ Class Participation (5%)

Textbook

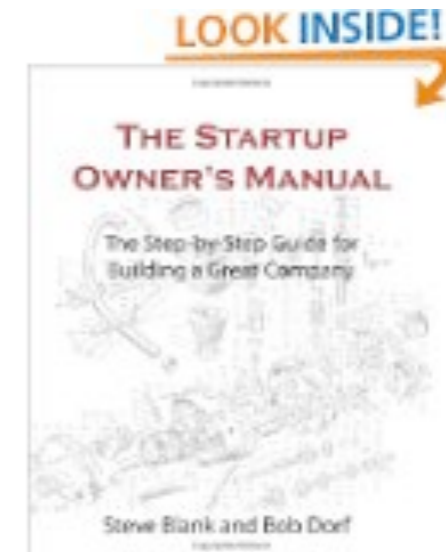
- ▶ Technology Ventures: From Idea to Enterprise, 3rd edition
- ▶ Thomas Byers (Author), Richard Dorf (Author), Andrew Nelson (Author)



Additional Readings

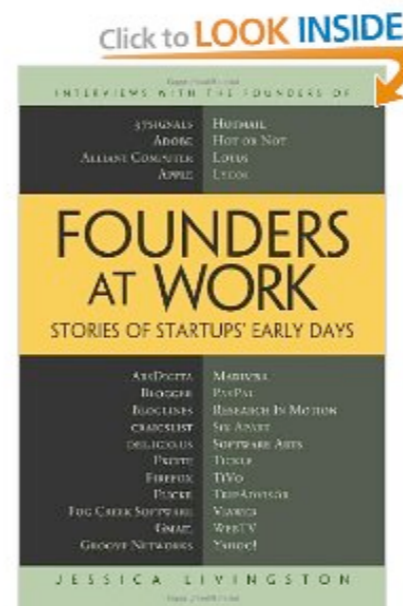
- ▶ The Startup Owner's Manual: The Step-By-Step Guide for Building a Great Company

- ▶ Steve Blank and Bob Dorf



- ▶ Founders at Work

- ▶ Jessica Livingston



DSTE Platform

- ▶ Platform to help CS and MBA students interact better
- ▶ You can start new topics for discussion
 - ▶ (please do!)
- ▶ All events will be posted here
- ▶ Calendar updated regularly

Canvas Platform

- ▶ MBA students are familiar with it
- ▶ Importing of CS students a manual process
- ▶ We won't be using it for now
- ▶ May use it for submissions after class roster is final

Important Dates/Pitch Days

- ▶ Tonight : 6:00 pm onwards - CS + MBA Mixer Event - Uris 1st-Hepburn Lounge Terrace
- ▶ DSTE First Pitch Day - Feb 20 at 4pm, Warren Lobby Feldberg Space
- ▶ DSTE Student-Mentor/Advisor Mixer Day - Feb 20 at 7pm, Warren Lobby Feldberg Space (depends on how many mentors agree for mixer)
- ▶ DSTE Incubation Space Visit - sometime in March
- ▶ DSTE Midway Pitch/Update Day - Mar 27 at 4pm, Uris 1st-Hepburn Lounge Terrace
- ▶ DSTE Final Demo/Pitch/Conference Day - May 7, 10:00 am, Uris 1st-142

Name Tags for Mixer Event

- ▶ Name
- ▶ Graduation Year
- ▶ Department
- ▶ If you already have an idea add *
- ▶ If you already have a running business add **

Extra Classes on Web Programming

- ▶ 3 Extra Lectures on Fridays (will start in end of Feb)
- ▶ For MBA students who want to learn basic computer science/programming
- ▶ For EE students who do not have a lot of web programming experience
- ▶ For CS students who want to know more on web programming
- ▶ NOT for experienced programmers
- ▶ Morgan Ulinski will be giving these lecture

Guest Lecture

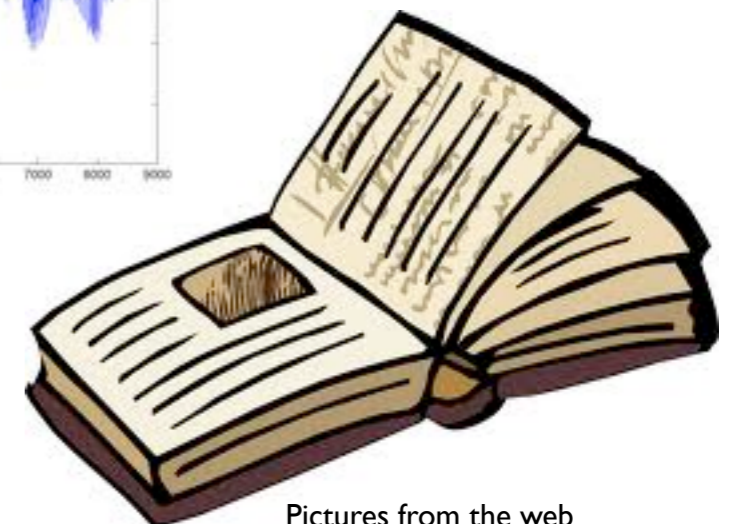
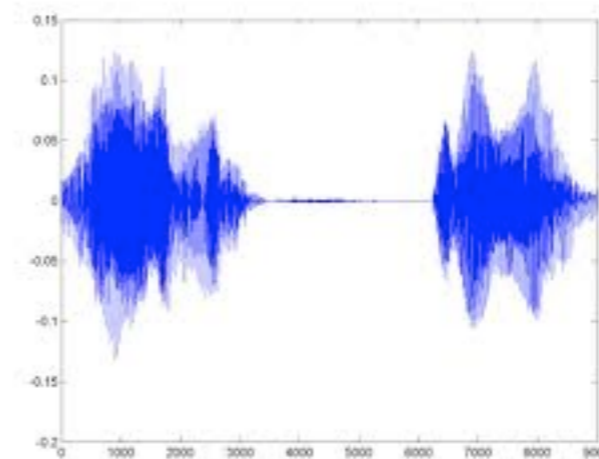
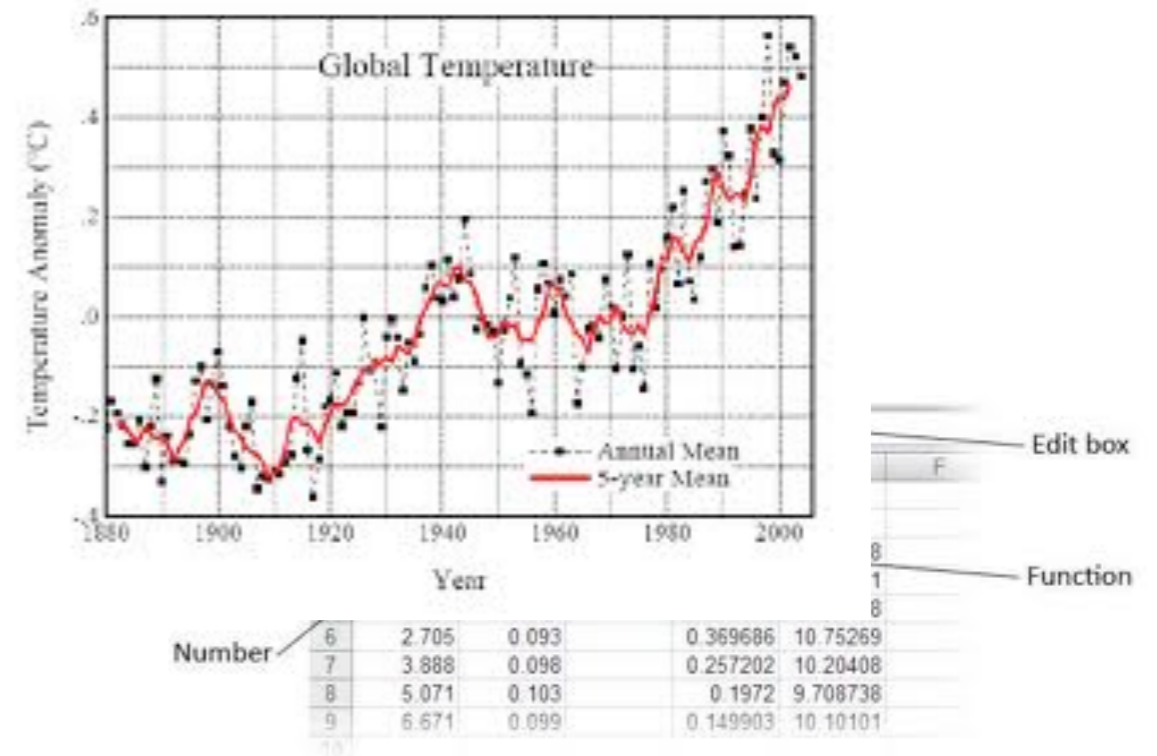


▶ **Wim Sweldens**

- ▶ Independent technology, innovation, and business leader
- ▶ Until end of 2012 President of Alcatel-Lucent's Wireless Division
- ▶ Earlier In his role as VP of Network Technology, Wim took the company's Applications Enablement strategy from paper to practice
- ▶ Founder and leader of Alcatel-Lucent Ventures, a strategic new business incubator and investment division inside of Bell Labs, Wim managed the lifecycle, from concept to commercialization, for eight ventures, ranging from enterprise & consumer applications to groundbreaking mobile innovations for operators

Types of Data

- ▶ Qualitative Data
 - ▶ “Weather is nice outside”
- ▶ Quantitative Data
 - ▶ Discrete - Red, Blue, Orange
 - ▶ Continuous, Temperature in Central Park



Data is available in many forms

Pictures from the web

Data Analysis

- ▶ You can do various kinds of analysis with data depending on your goal
- ▶ Descriptive Analysis
 - ▶ Describe data, e.g. census
- ▶ Exploratory Analysis
 - ▶ Discover connections
- ▶ Inferential Analysis
 - ▶ Use small set of data to explain larger population
- ▶ Predictive Analysis
 - ▶ Use some data X to predict Y
- ▶ Causal Analysis
 - ▶ randomized trials
- ▶ more methods ...



Analyzing Very Large Datasets

- ▶ Previously mentioned analysis methods have been done for decades now
- ▶ What has changed?
 - ▶ Amount of data
 - ▶ Speed data is generated everyday
 - ▶ Types of data
 - ▶ Noise
- ▶ Trying to run analysis methods in traditional approaches may fail on these large datasets - Big data problem

Big Data

- ▶ **Volume**

- ▶ 12 Terabytes of Twitter data created everyday!

- ▶ **Velocity**

- ▶ 5 million trades a day

- ▶ **Variety**

- ▶ sensors, audio, video, call logs, text

- ▶ **Value/Veracity**

- ▶ accuracy of data on producing valuable insights

Analysis Methods and Big Data

- ▶ How should we go about performing one of the analysis methods we mentioned before with such vast amount of data?

- ▶ Aviation Big Data and Analysis Methods

- ▶ Flight Quest Challenge by GE

- ▶ https://www.youtube.com/embed/sFZ4hWzsunc?feature=player_embedded

- ▶ Improve aviation efficiency

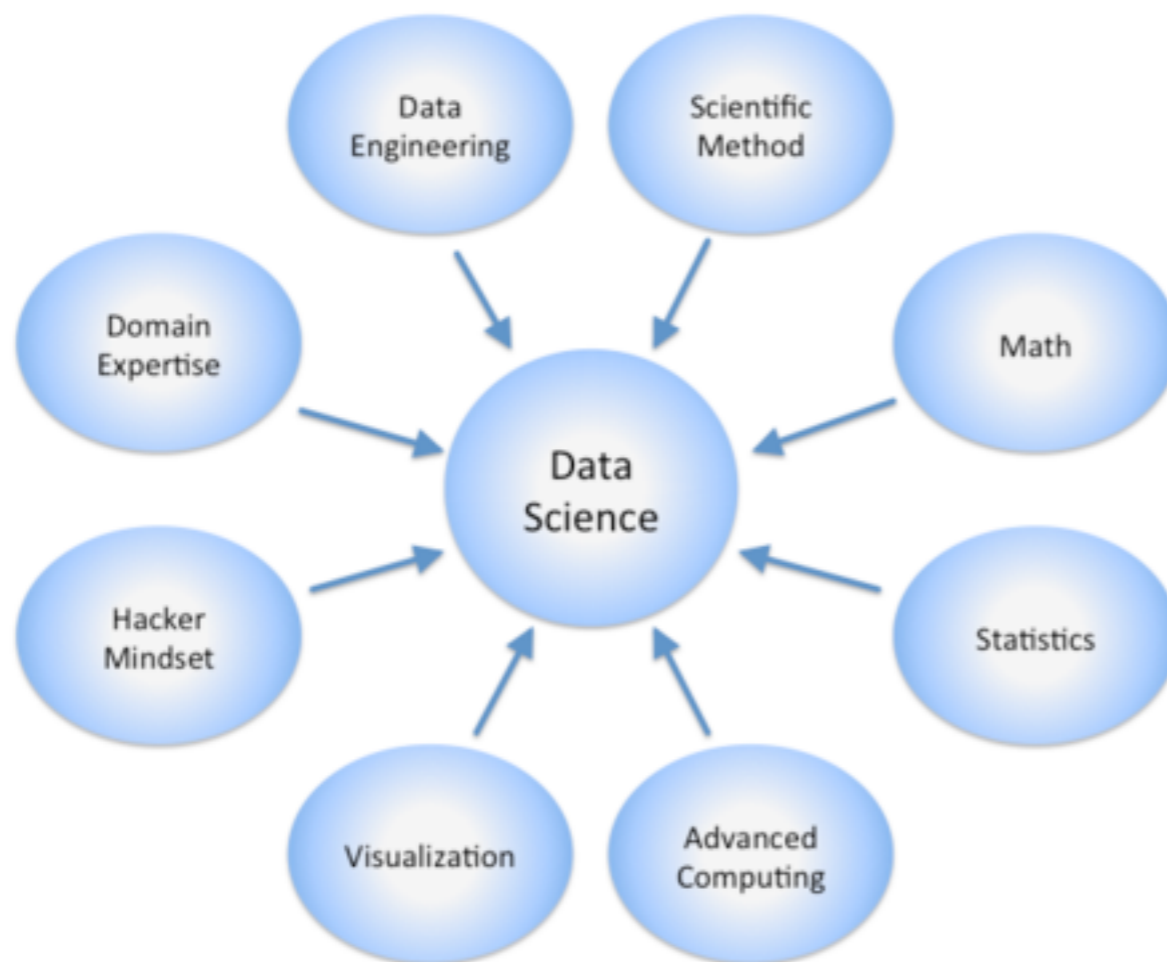
- ▶ Data description : flight number, origin, destination, take-off time, arrival time, latitude and longitude at frequent interim waypoints along the journey, and weather and wind data.
 - ▶ Thousands of data points generated every second
 - ▶ Need to come up with algorithm that reduces delays
 - ▶ How should one go about coming up with such algorithm?



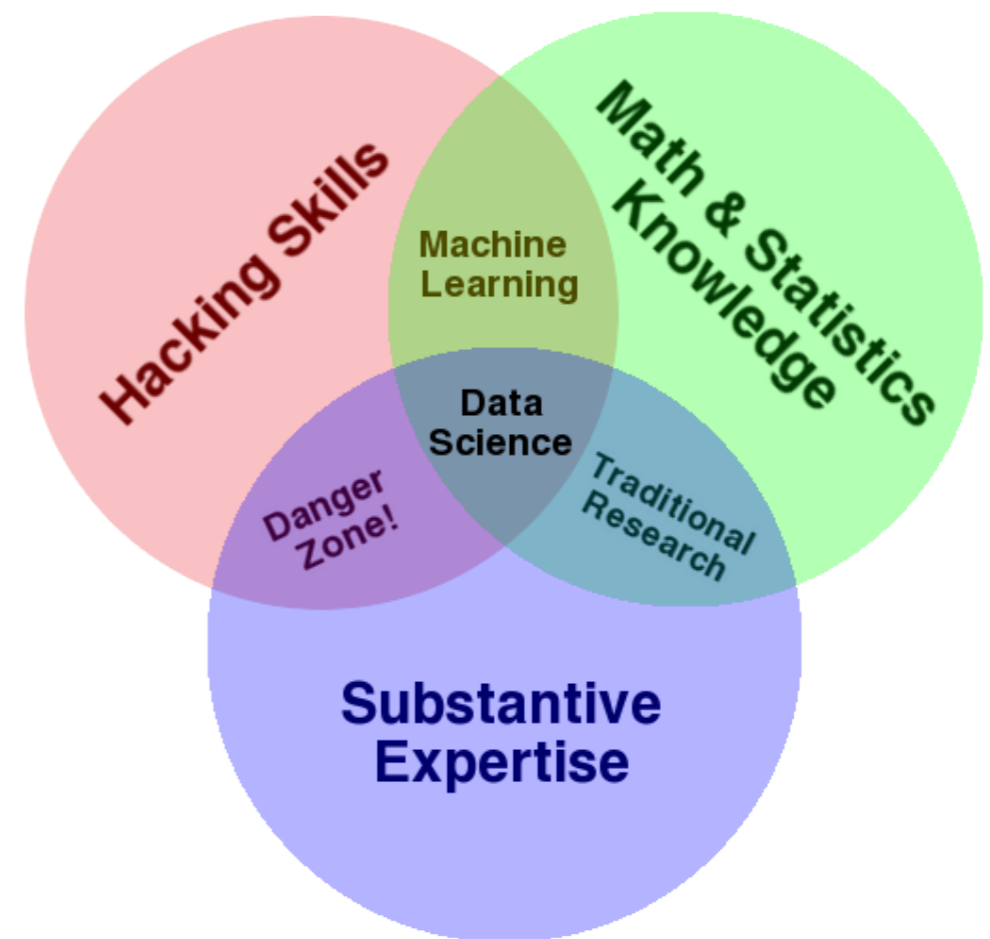
Data
Science

Data Science

- ▶ There isn't one standard definition of data science
- ▶ New growing field that brings together Machine Learning, Statistics, Business, Visualization



From Wikipedia



Copyright - Drew Conway

Data Science

- ▶ Data Science helps
 - ▶ in finding knowledge from large amount of data
 - ▶ knowledge should be of some value
- ▶ Ideal data scientist will know
 - ▶ Programming
 - ▶ Machine Learning & Statistics
 - ▶ Experience in processing large data sets
 - ▶ Domain Expertise
 - ▶ Knows what question to ask
 - ▶ Visualization expert

Very hard to find one
person with all these qualities

Data Science for Business

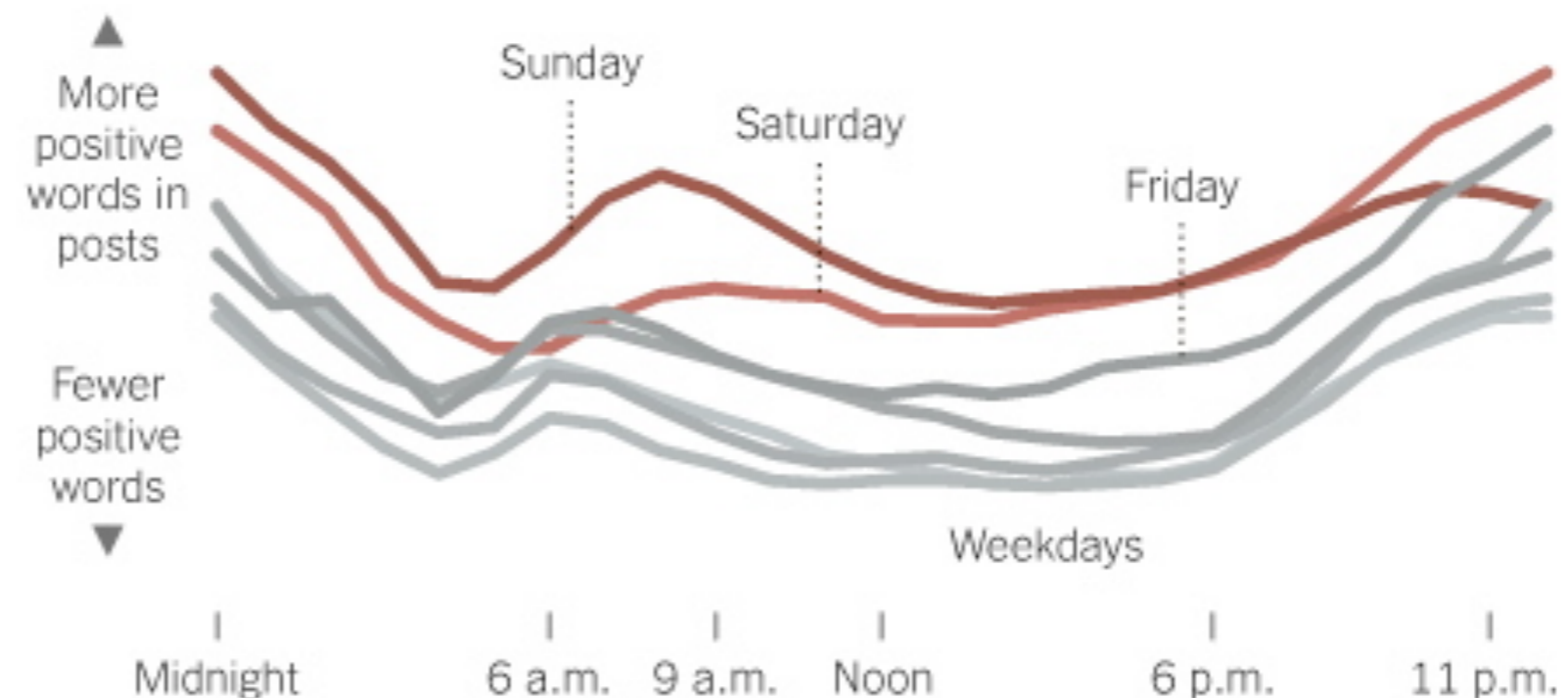
- ▶ How can we use data science for business?
- ▶ Let's look at a hypothetical example
 - ▶ Company Zoolaster announces deals on their product every week
 - ▶ Assume : They realize that the consumer is likely to buy Product Zoola if they see the deal when they are happy
 - ▶ Problem : Zoolaster wants to know which day and what time of the day to announce the deal
 - ▶ (Assume that all of their consumers see the deal right away)
 - ▶ Question : Which day of a week should Zoolaster send their product announcement
 - ▶ In other words how can Zoolaster find out when their consumers are in the best mood?

Using Twitter Data for Mood Analysis

- ▶ Our hypothetical company Zoolaster could have mined large amount of twitter data and come up with an answer to their question
- ▶ In fact processing 500 million tweets a study showed that people are most happy on Sunday around 9am

Studying Moods Through Twitter

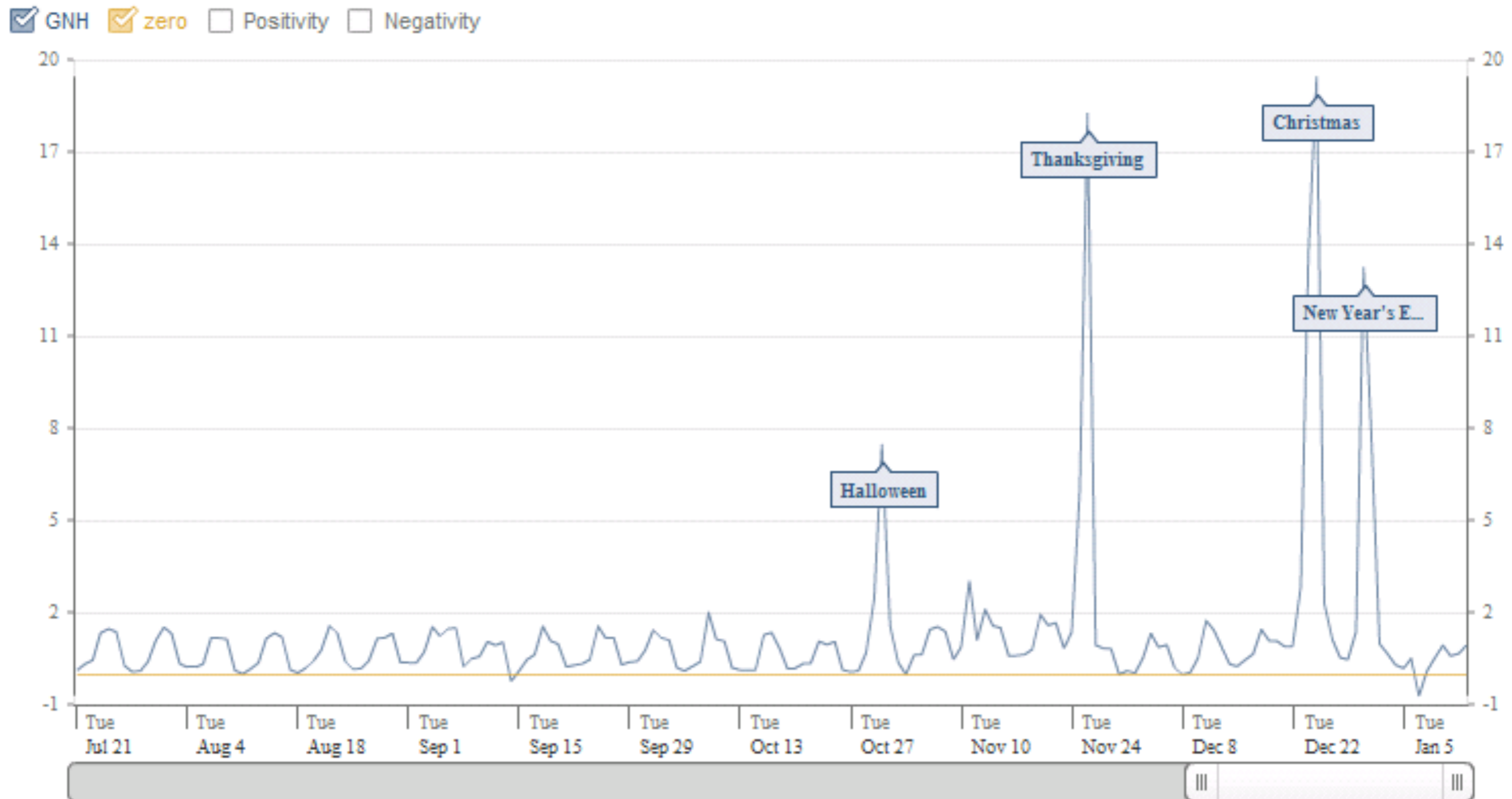
A textual analysis of more than 500 million Twitter messages found people around the world tend to express more positive emotions in the morning and evening, and are most positive on weekends. The recurring daily pattern suggests moods are influenced by sleep and circadian rhythms.



Source: Science

THE NEW YORK TIMES

Facebook's Gross Happiness Index



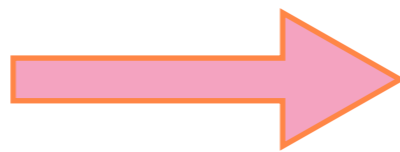
Source : Facebook Blog

Facebook Blog Explains

“The result was an index that measures how happy people on Facebook are from day-to-day by looking at the number of positive and negative words they're using when updating their status. When people in their status updates use more positive words - or fewer negative words - then that day as a whole is counted as happier than usual.”

Data Science and Business

- ▶ Examples we saw just now were experiments that performed sentiment analysis to provide insights into the state of customers
- ▶ How did data scientist in Facebook come up with “Gross National Happiness Index”



~18.5 for Nov 24

Big Data
Millions of status updates
Unstructured Text

Happiness Score
that can be used by
Marketing Manager

Use of Data Science Methods in Business

▶ LinkedIn

- ▶ People You May Knows
- ▶ Identified number of connections it takes for a long-term engagement

▶ Netflix

- ▶ Signup process
- ▶ Encourage to add movies to your queue
- ▶ Once you add certain number of movies likelihood of you being a long term customer goes up

Use of Data Science Methods in Business

▶ Zynga

- ▶ Monitors their users constantly
- ▶ Analyze how users interact with games to find out what makes a game successful

▶ Financial Services

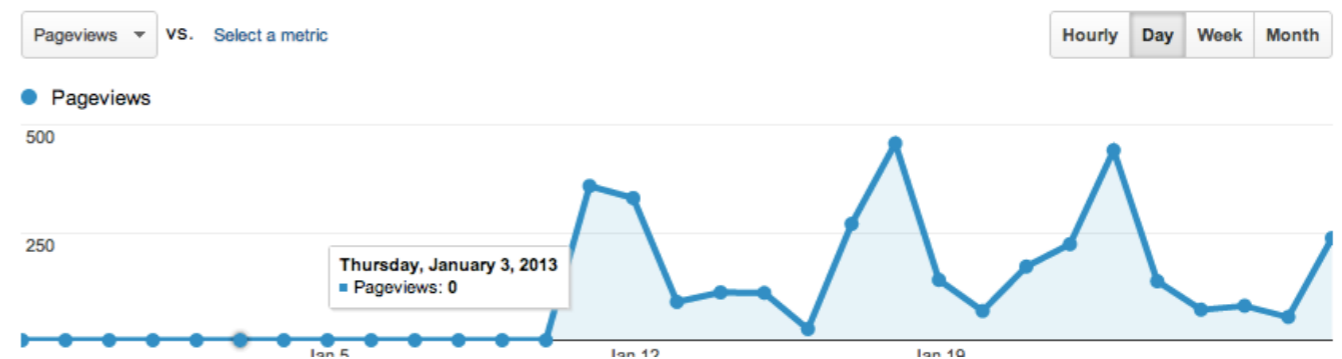
- ▶ Fraud detection

▶ OkCupid

- ▶ Marketing Analytics with viral blogs
- ▶ Facial attitude and new contacts blog

Data Science and Technology Startups

- ▶ Technology Startups can generate a lot of data
- ▶ For example a web startup with 500K users can generate a lot of data every user action is stored
 - ▶ Visits
 - ▶ Click through rates
 - ▶ Search logs
 - ▶ User generated content
 - ▶ Time spent on individual pages
 - ▶ Mouse movement behavior
 - ▶ Many more individual data points
- ▶ Mining this large set of data generated every day for identify various types of pattern about users could lead to increased engagement



Data Science and Technology Startup

- ▶ Let's do another hypothetical example
- ▶ Using data science for a web startup that sells products online
- ▶ Want to increase the click through rate on related items?
- ▶ In other words, want to build a very simple minded recommendation engine

Example : User Data and Engagement

- ▶ Assume you are running a shopping site and you want to produce top 5 items to recommend like Amazon

Inspired by Your Shopping Trends



Bluetooth USB 2.0 Micro Adapter Dongle
Generic
★★★★☆ (817)
~~\$19.99~~ \$4.49



RF Wireless Laser Pointer with Page...
Generic
★★★★☆ (179)
\$6.87



Satechi SP400 Smart-Pointer 2.4Ghz RF...
★★★★☆ (359)
~~\$49.99~~ \$34.99



August LP103R Red Laser Presentation...
★★★★☆ (8)
~~\$16.99~~ \$9.95



Logitech Wireless Presenter R400
★★★★☆ (216)
~~\$49.99~~ \$36.63

[▶ View your shopping cart](#)

- ▶ Data you have
 - ▶ Click through rates and data item




User Data History

► Data from User's history

Product Clicks

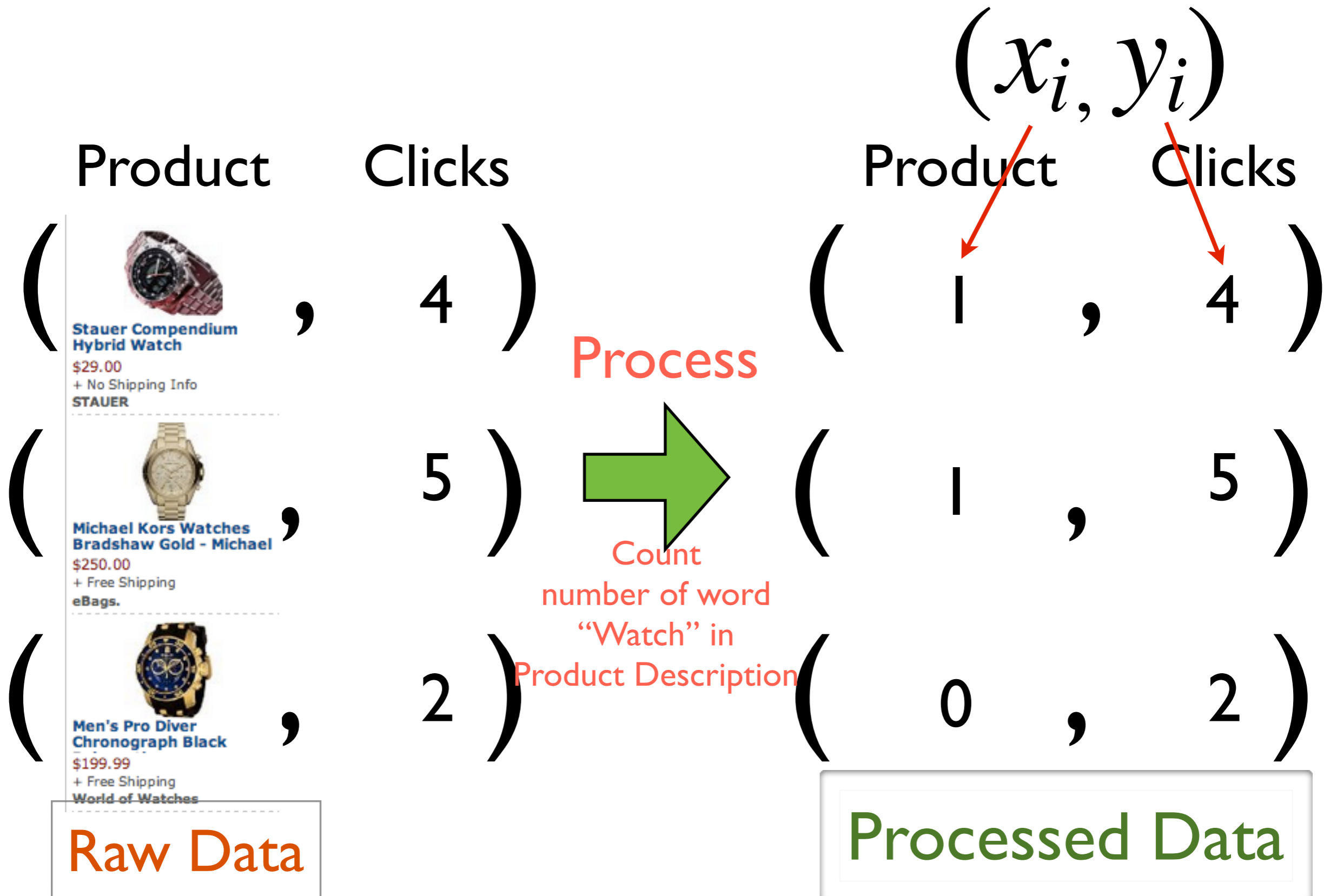
 Stauer Compendium Hybrid Watch \$29.00 + No Shipping Info STAUER	4
 Michael Kors Watches Bradshaw Gold - Michael \$250.00 + Free Shipping eBags.	5
 Men's Pro Diver Chronograph Black \$199.99 + Free Shipping World of Watches	2

User Data History

Product	Clicks
 Stauer Compendium Hybrid Watch \$29.00 + No Shipping Info STAUER	4
 Michael Kors Watches Bradshaw Gold - Michael \$250.00 + Free Shipping eBags.	5
 Men's Pro Diver Chronograph Black \$199.99 + Free Shipping World of Watches	2

- ▶ Can you use this data to build a simple model that can predict the number of clicks for a new product?
- ▶ Imagine you have such data for millions of users

Process Raw Data



Scoring Problem

- ▶ Given a large amount of data we want to predict a score that represents the number of times the item will be clicked

(1,4) (x_1, y_1)

(1,5) (x_2, y_2)

(0,2) (x_i, y_i)

(1,3)

(0, 1)

$(x_n, ?)$

- What kind of modeling technique can we use?

(1,3) $(x_{1000000}, y_{1000000})$

Data to Scores

- ▶ We want to find a function that given our x it would map it to y
- ▶ One such function is

$$f(x) = \theta_0 + \theta_1 x$$

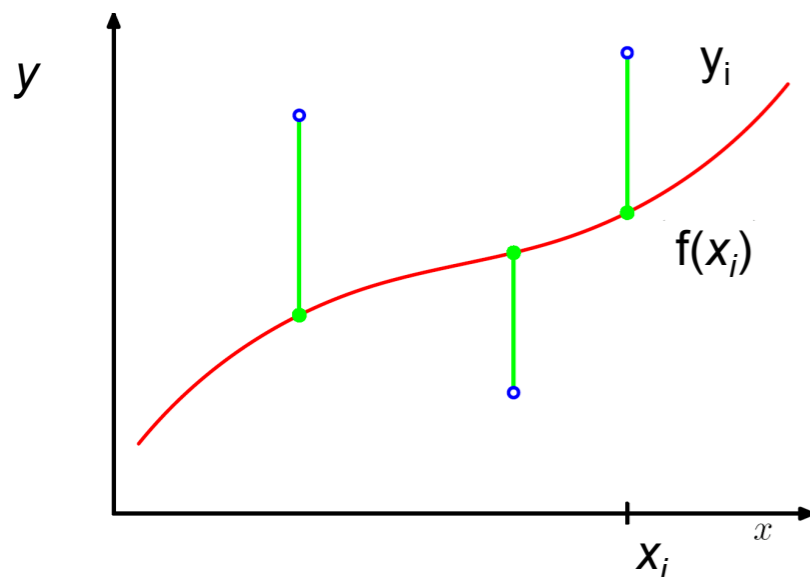
- ▶ Different values of theta give different functions
- ▶ Best theta so that we make least error on predictions when compared with given y

What kind of
Modeling Technique
can we use?

Minimize Loss : Predicted vs True

► Regression Model

- Our function $f(x)$ approximates y
- Given a true value of y we can compute the error $f(x)$ made against true y
- For any point x_i we can compute such error by $y_i - f(x_i)$; or by squared error $\{y_i - f(x_i)\}^2$
- But we have N points so the total error/Loss L on squared error would be



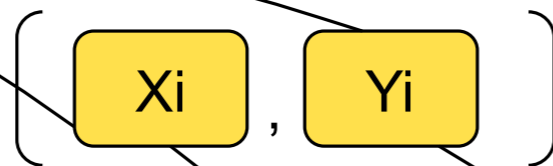
$$L = \sum_{i=1}^N (y_i - f(x_i))^2$$

Data to Scores

► Raw Data => Processed Data => Model => Prediction Score

- Given our training data

- (1, 4)
- (0, 2)
- .
- .
- .
- .
- (1, 9)



Training Our Regression Model:

Just need to implement for loop that computes numerators and denominators in equations here. And we get optimal thetas

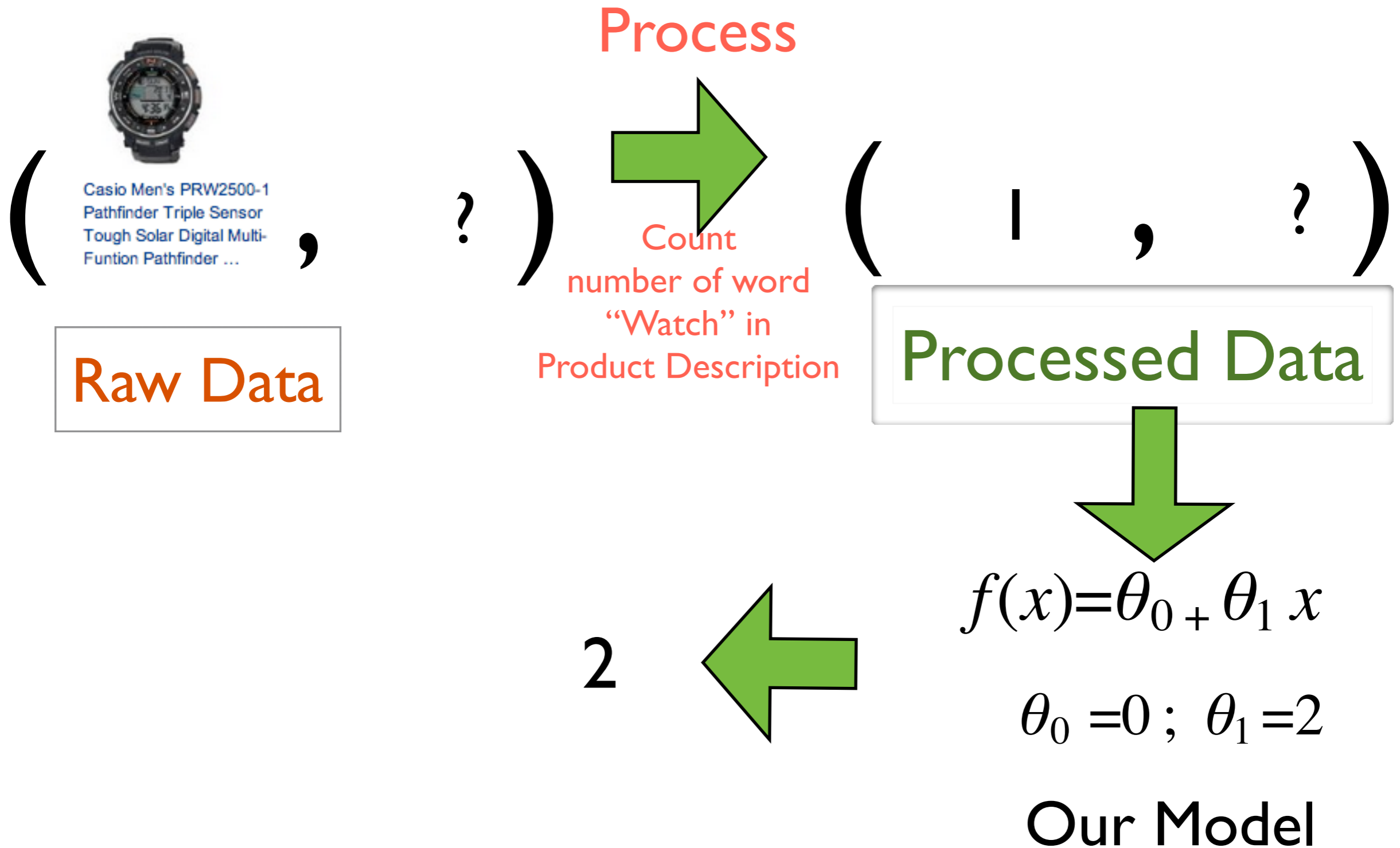
$$\theta_1 = \frac{\sum_{i=1}^N x_i y_i - \frac{1}{N} \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{\sum_{i=1}^N x_i^2 - \frac{1}{N} \sum_{i=1}^N x_i \sum_{i=1}^N x_i}$$

For Prediction/Testing:

Given optimal thetas, plug in the x value in our equation to get y

$$\theta_0 = \frac{1}{N} \sum_{i=1}^N y_i - \frac{1}{N} \theta_1 \sum_{i=1}^N x_i$$

Data to Predicted Scores



Data to Scores

- ▶ We looked at how we can use a linear regression model to predict if a product is likely to be clicked
- ▶ Converted raw data (product information) into insights (likelihood of clicks)
- ▶ The method of scoring raw data can be useful in many different steps of building startups
 - ▶ Testing value proposition
 - ▶ Customer development
 - ▶ Testing distribution channels
 - ▶ and more ...

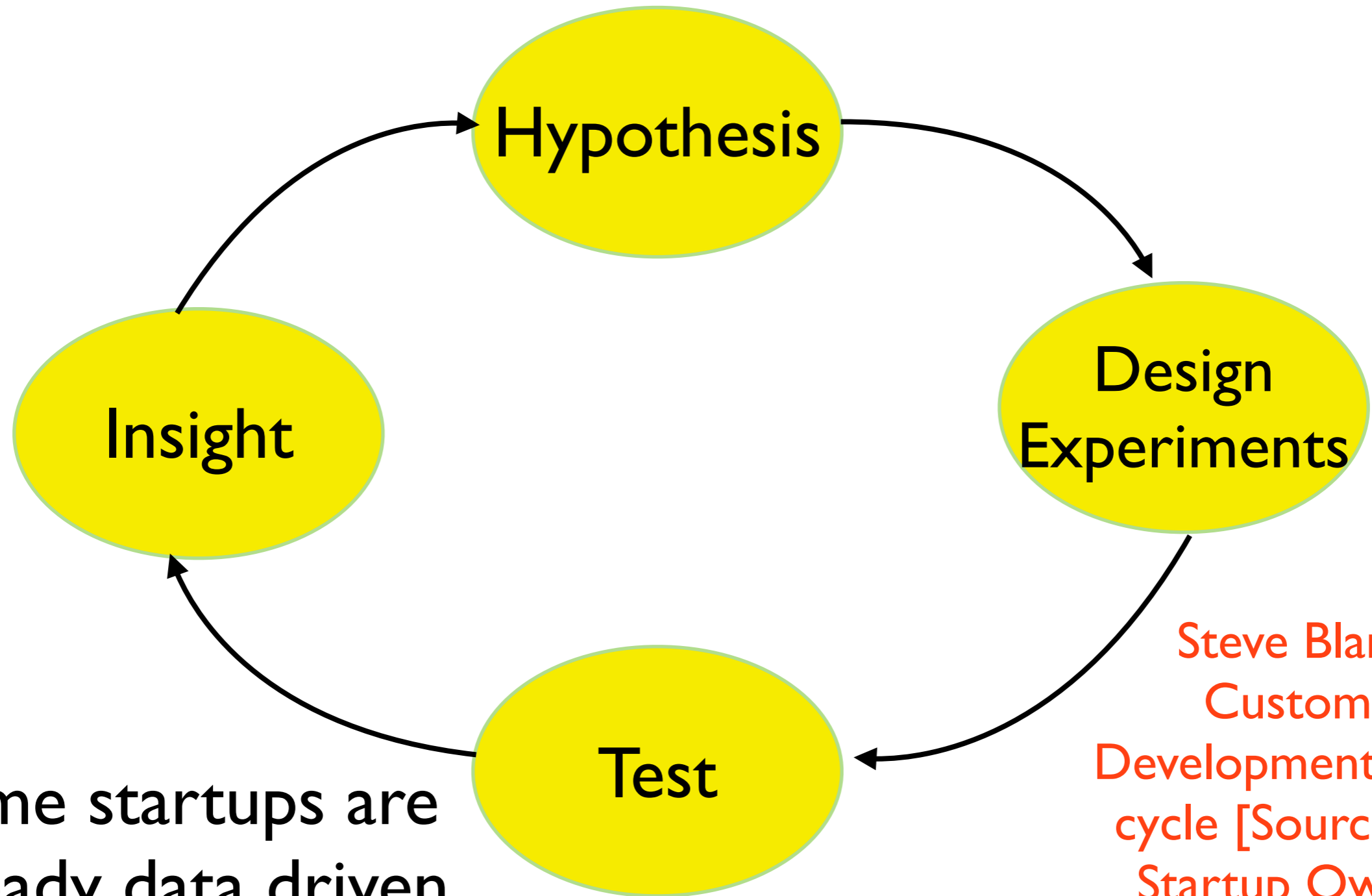
Data Driven Decision Making

- ▶ You have a bunch of ideas
- ▶ You decide to pursue one of them
- ▶ Which one should you pursue?
- ▶ Can data driven methods be used to help you decide what idea to pursue?
- ▶ Does data driven methods even payoff?

Data Driven Decision Making

- ▶ Study at MIT Sloan Management and Wharton showed an increase of 5 to 6 percent in output and productivity for those who adopted “data driven decision making” [Source NYTimes]
- ▶ Quote for NYTimes :The companies that are guided by data analysis, Mr. Brynjolfsson says, are “harbingers of a trend in how managers make decisions.”
- ▶ Can we use data driven decision making for Technology Startups?

Data Driven Decision and Startups



Some startups are already data driven

Steve Blank's
Customer
Development Insight
cycle [Source :The
Startup Owner's
Manual]

Data, Decisions and Startups

- ▶ In this course, we will follow lean startup concepts and customer development process (Steve Blank)
- ▶ We will also learn Data Science topics that can be applied to
 - ▶ improve data collection process
 - ▶ better design experiments
 - ▶ test hypothesis
 - ▶ get better insights to data

Data Science and Business

- ▶ We will look at four main ways to analyze data that are all useful in decision making for businesses in general
- ▶ Data to Scores
- ▶ Data to Classes
 - ▶ Discriminative Methods
 - ▶ Generative Methods
- ▶ Data to Clusters

Assignment - I

- ▶ Form a team
- ▶ Name the team
- ▶ Write a short summary about your business concept (~10 sentences)
 - ▶ Problem addressed, Your proposed idea/solution, Value Proposition, Prospective customers, Team skills
- ▶ Write 5 bullets on possible data points you can collect to test your value proposition
- ▶ Due Next Friday @ 6pm

Submitting Assignments

- ▶ Stay tuned for how to submit the assignment
- ▶ Canvas or DSTE

Reading Assignments

- ▶ Technology Ventures book
 - ▶ Chapter 2: Opportunity and Concept Summary
- ▶ Chapter 1 and 2, Startup Owner's Manual
- ▶ What is Data Science? By Mike Loukides
 - ▶ http://www.cloudera.com/content/dam/cloudera/Resources/PDF/What_is_Data_Science_OReilly.pdf
- ▶ Data Scientist :The Sexiest Job of 21st Century By Thomas H. Davenport and D.J. Patil
 - ▶ <http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/2>

Guest Lecture

- ▶ Next is Guest Lecture on Evaluating Startup Ideas and Entrepreneurship Experience by Wim Sweldens