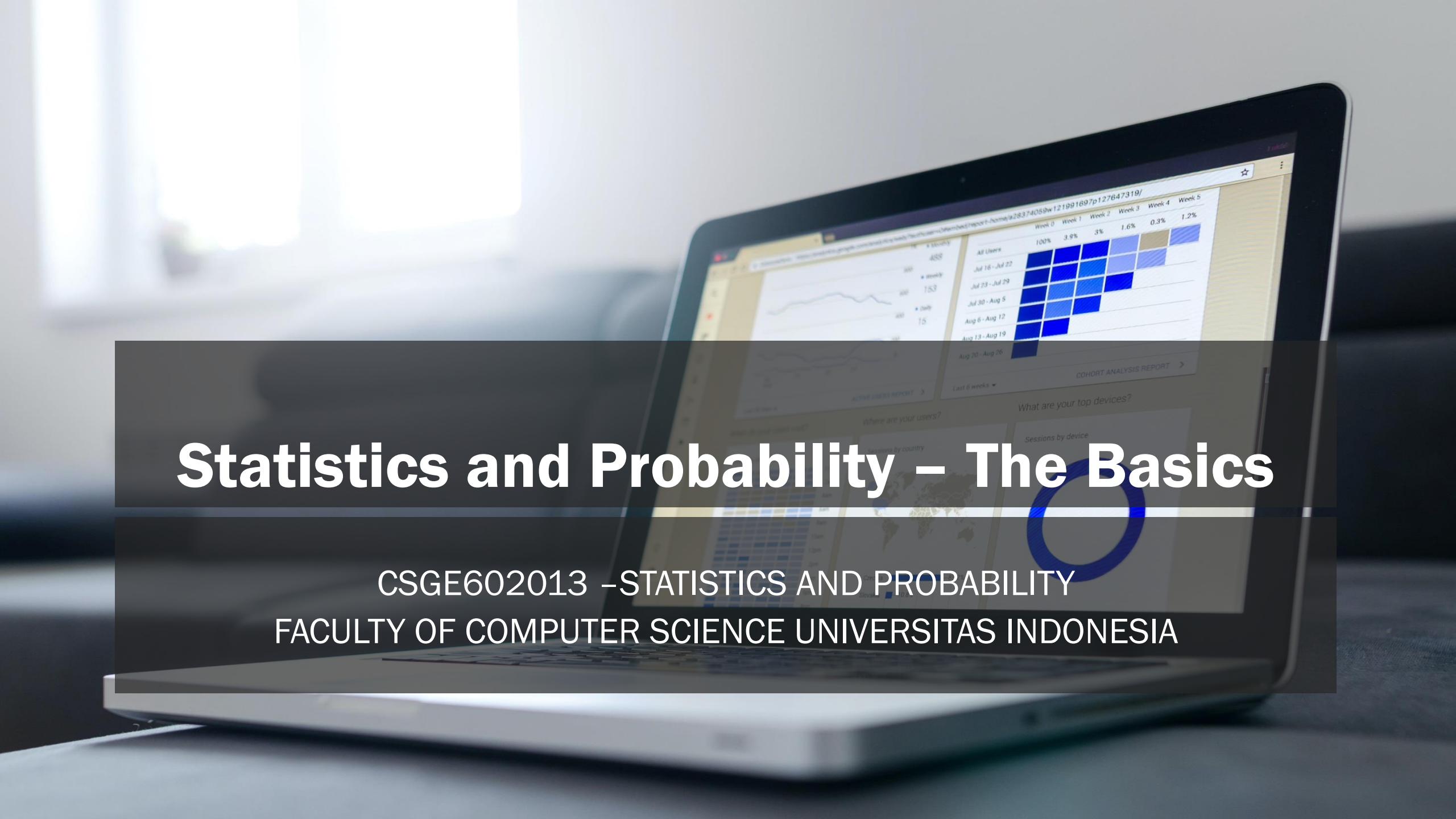


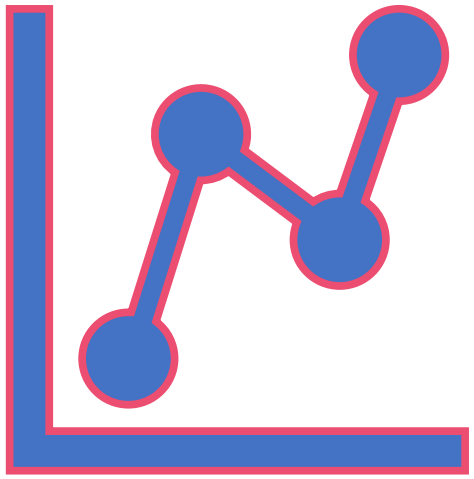
Statistics and Probability – The Basics

CSGE602013 –STATISTICS AND PROBABILITY

FACULTY OF COMPUTER SCIENCE UNIVERSITAS INDONESIA

References

- Introduction to Probability and Statistics for Engineers & Scientists, 4th ed., Sheldon M. Ross, Elsevier, 2009.
- Probability and Statistics for Engineers & Scientists, 3rd Edition. Anthony J. Hayter, Thomson Higher Education

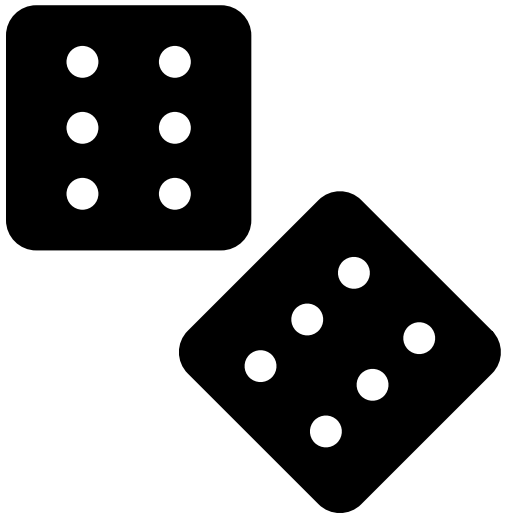


WHAT IS STATISTICS?

Statistics

- The art of learning from **data**
- Includes:
  - The **collection**, **description,** and **analysis** of data
- Why?
  - To draw conclusions

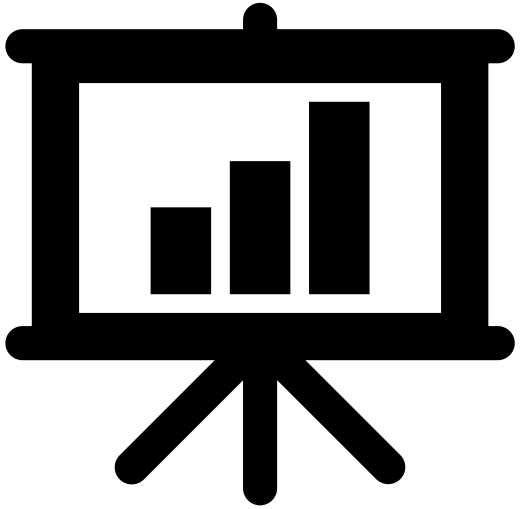
Data → Information



WHAT IS PROBABILITY?

Probability

- Branch of mathematics that has been developed to deal with uncertainty (random events).
- The concept of probability of a particular event is subject to various meanings, interpretation, and context
  - Depends on the other supporting data!
  - Depends on **context!**



WHY STATISTICS AND PROBABILITY?

ALAN TURING

While we develop a system for determining how much intelligence to act on. Which attacks to stop, which to let through. Statistical analysis. The minimum number of actions it'll take to win the war, but the maximum number we're able to take before the Germans get suspicious.

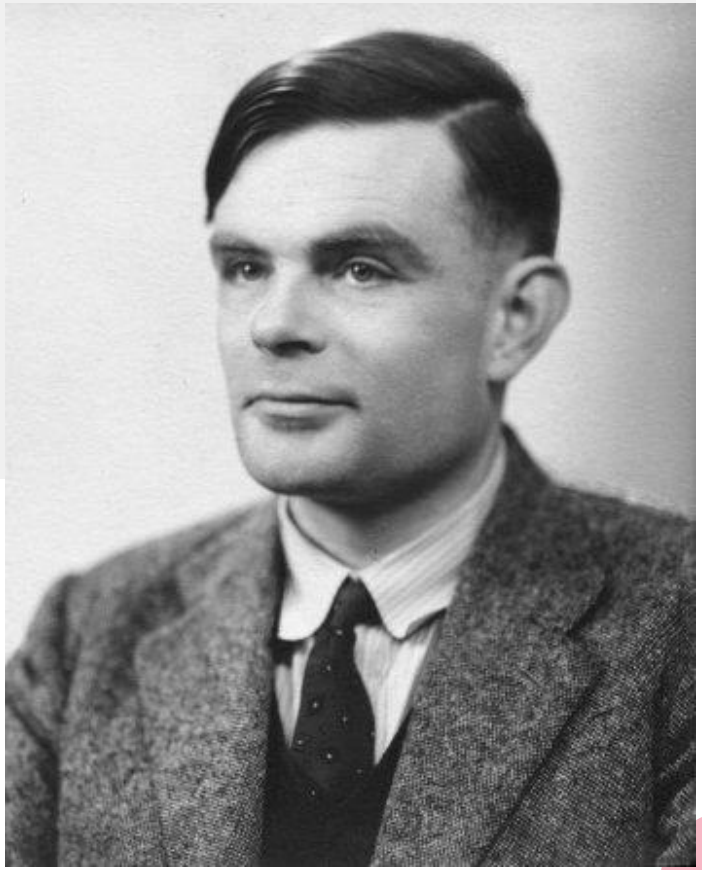
STEWART MENZIES

You're going to trust this all to statistics?

To maths?

ALAN TURING

Correct.



Dialog Script of the film "The Imitation Game"
In http://stats.stackexchange.com/

Photo: http://en.wikipedia.org/wiki/Alan_Turing

Probability & Statistics for Computer Science

Machine Learning

Data Mining

Text Mining

Natural Language Processing

Simulation

Cryptography

Robotics & AI

Algorithms

Image Processing

Computer Graphics

Computer Vision

Software Testing

Probability & Statistics for Information Systems

- Modern Information Systems are associated with huge amounts of data
- Probability and statistics provide strong theories and tools to all aspects of data analysis in the wide discipline of information systems.

Risk Management

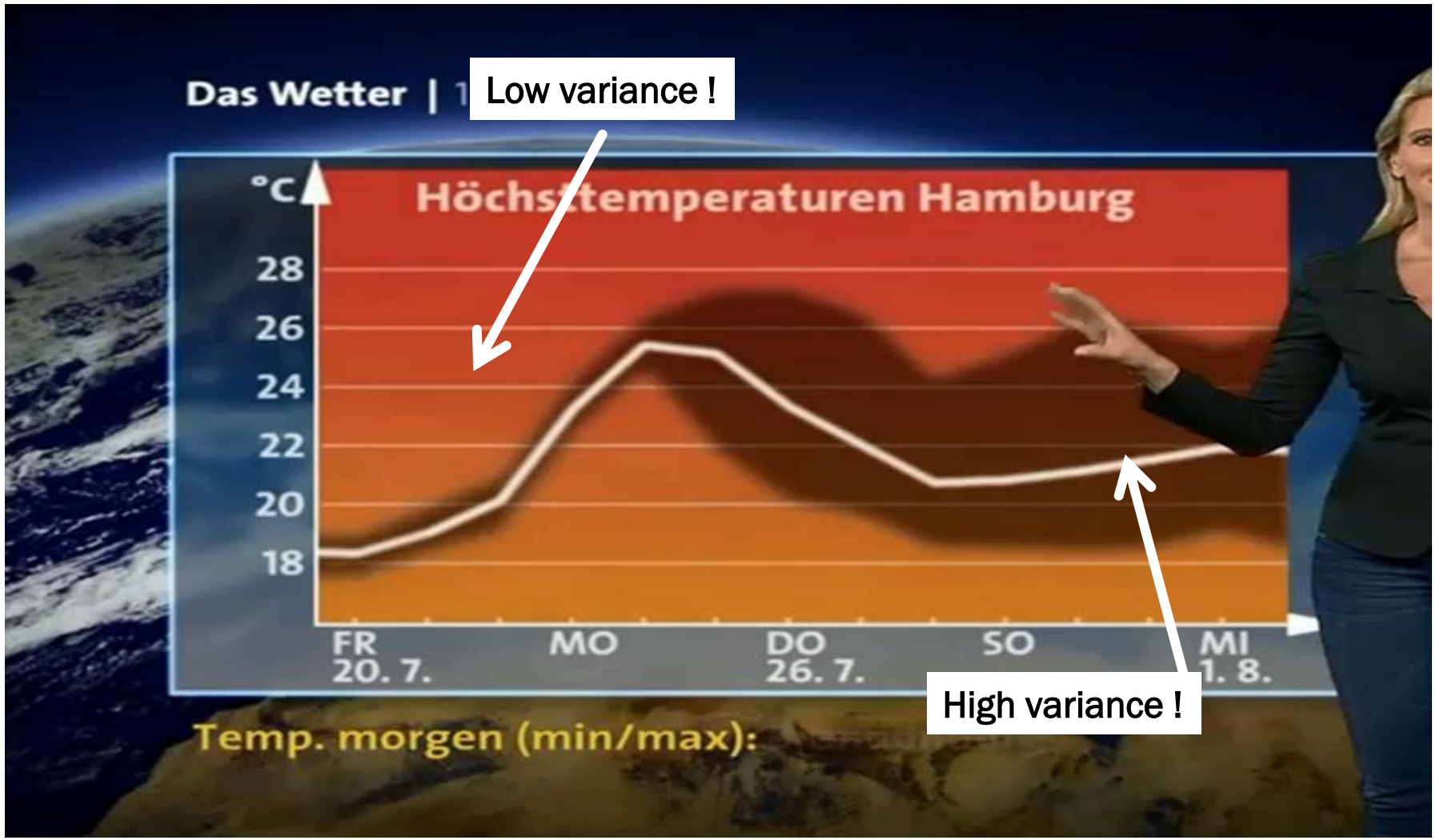
Requirements Engineering

Information Systems Security

Information Systems Project Simulation

Business Intelligence

Weather Forecast for the **next 15 days** !



http://dwjunkie.wordpress.com/2013/07/24/data-sciencewhats-the-big-deal-about-it/

Math equation could help find missing Malaysian plane

Bayes' Theorem helped researchers locate Air France Flight 447's black box in 2011

March 12, 2014 1:37PM ET

by **Ehab Zahriyeh** - @EhabZ

http://america.aljazeera.com/articles/2014/3/12/mathematical-equationcouldhelpfindmissingmalaysianplane.html

Machine Learning

Machine learning provides mechanisms to learn from data.

- There exists underlying **statistical model** on our data
- We estimate the parameter of our model based on **observable data**
- We use that to make decisions

Example of application:

- Classification (SPAM filtering, Handwriting Recognition)
- Prediction (Elections, Market analysis)
- Natural Language Processing
- …

Machine Learning (an Example)

For example, you have the following data obtained from previous experience.

Gender	Weather	GPA	Outfit Color	Lunch
Male	Rain	4	Red	Meetball
Male	Sunny	4	Blue	Chicken noddle
Female	Rain	3	Black	Lamb Sate
Female	Rain	4	Blue	Meetball

Create an algorithm that receives the input of **the table** and produces a prediction function **F**.

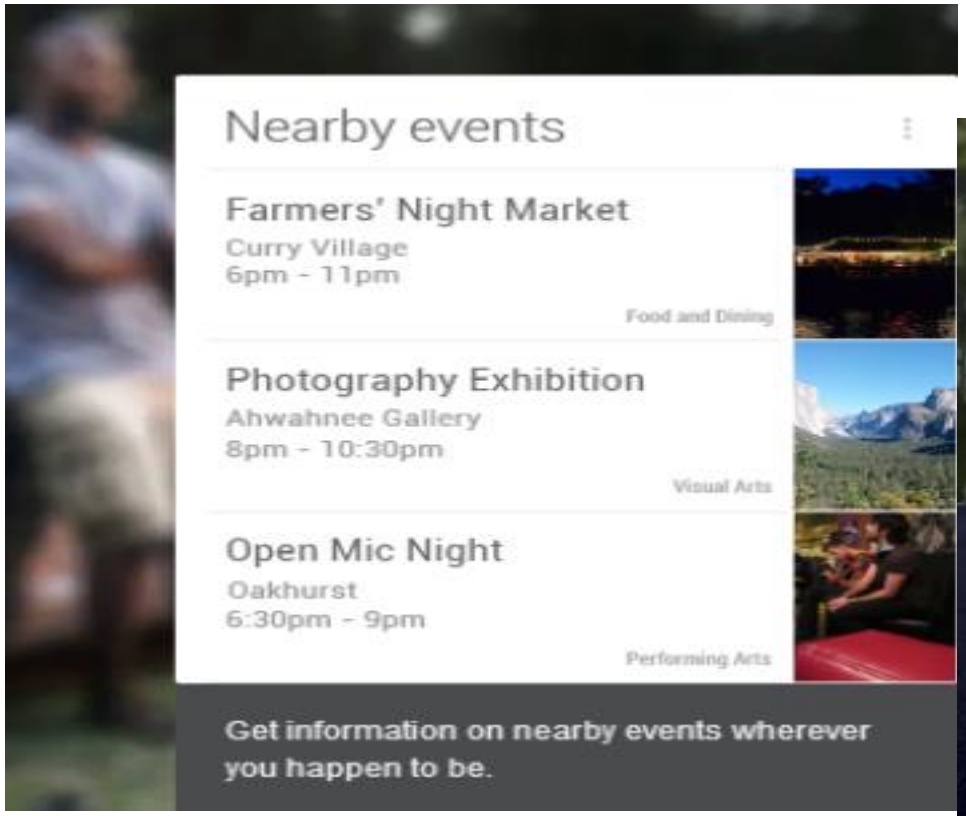
The prediction function is used to answer the following questions: If it rains today and there is a man in black clothes and has a GPA = 4, what kind of lunch is right for that person?

Application: Face Detection

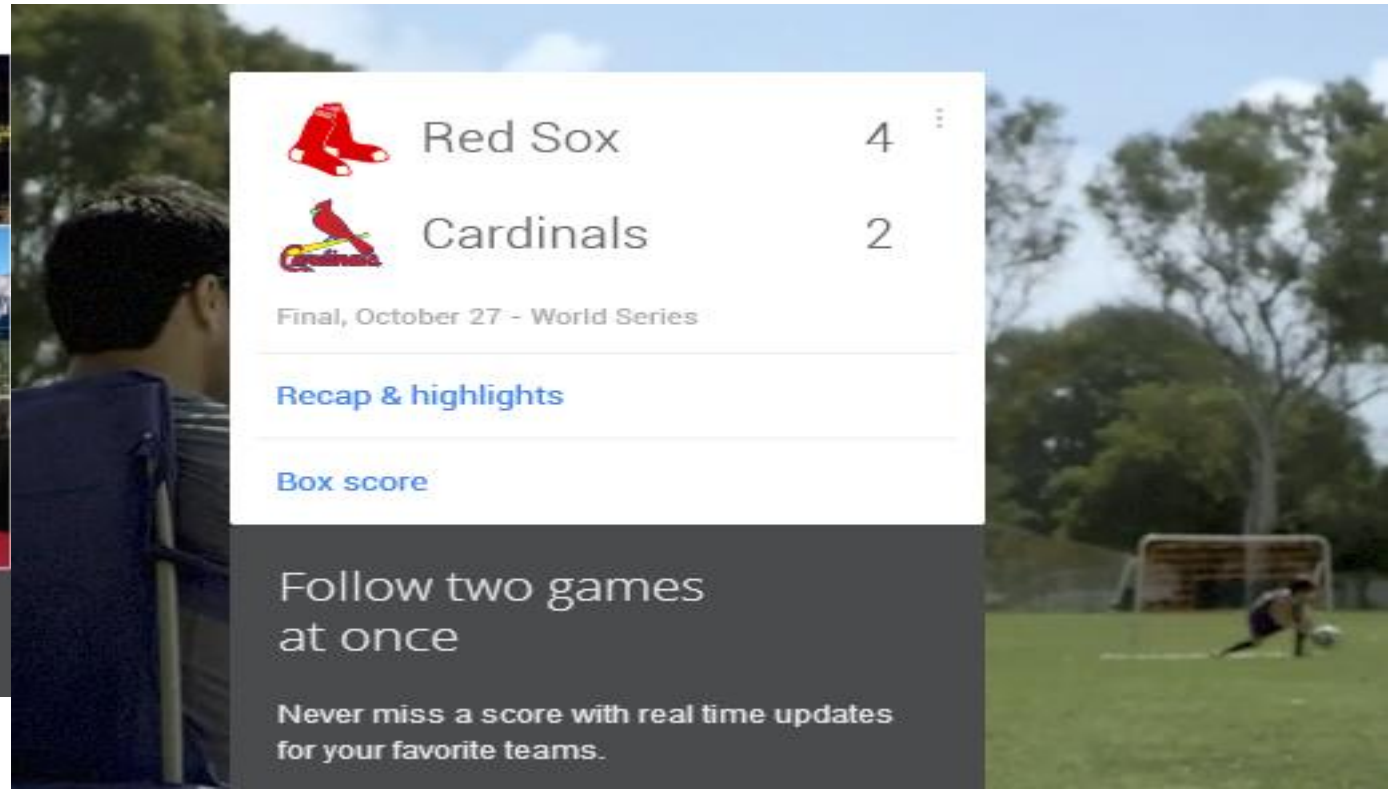
http://www.briancbecker.com/blog/projects/facebook-face-recognition/

B. C. Becker, E. G. Ortiz. "*Evaluation of Face Recognition Techniques for Application to Facebook*". IEEE International Conference on Automatic Face and Gesture Recognition 2008.

Google Now

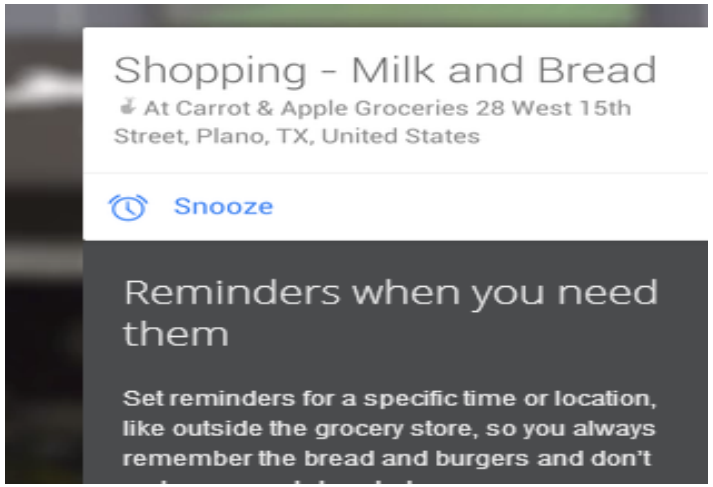


http://www.google.com/landing/now/



personalized assistant that can **predict your needs, wants, and deep desires !**

Google Now



How to do that ?

Google uses your **private data**

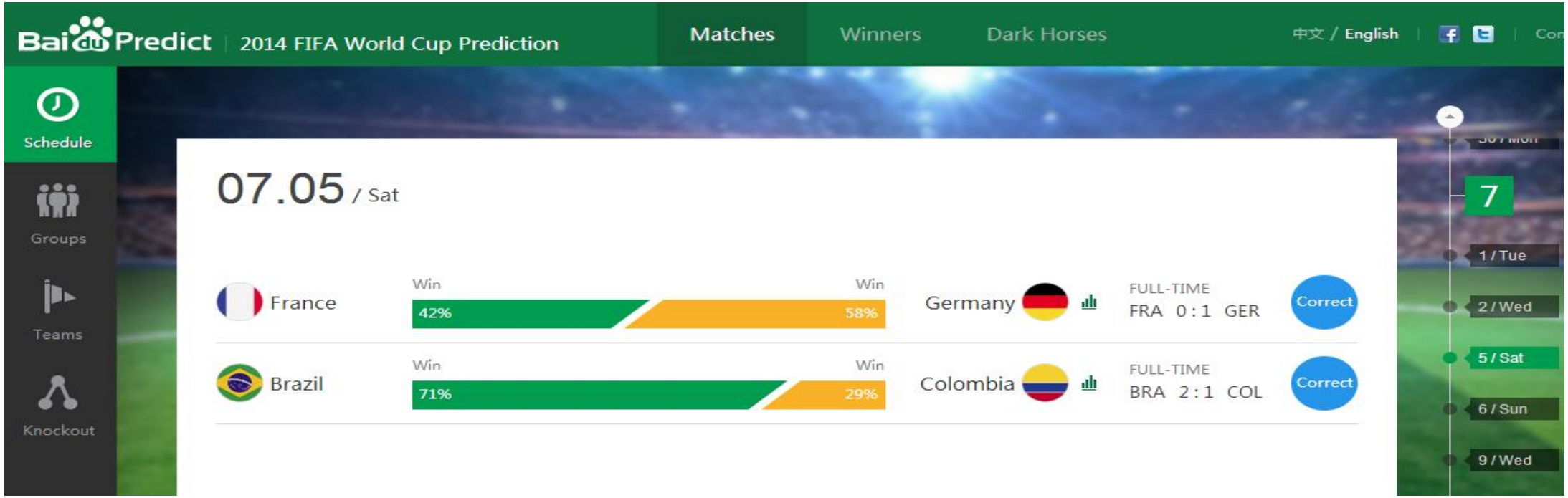
- people you know, documents, images, hangouts
- accessing your location, e-mail, daily calendar, and other info

in order to keep tabs on things like search preferences, appointments, flight reservations, payments and hotel bookings.

We need **statistics & math** to do that !

Deep Learning

Baidu big winner in World Cup !

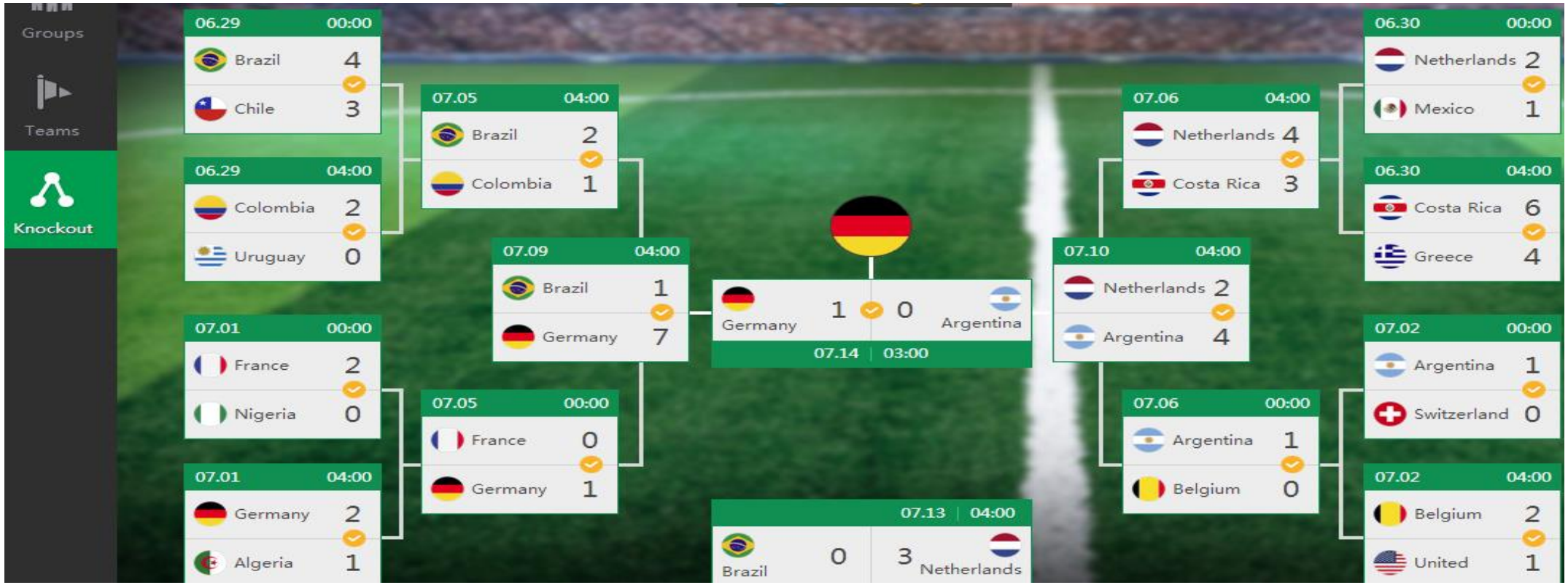


http://europe.chinadaily.com.cn/business/2014-07/14/content_17763137.htm

Deep Learning

Baidu said that its World Cup prediction model is based on data from as many as 37,000 matches played by 987 teams over the past five years.

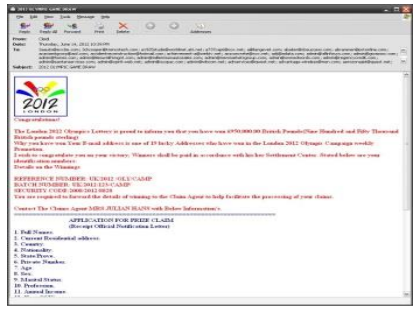
five factors: the teams' strength, home advantage, recent game performance, overall World Cup performance.



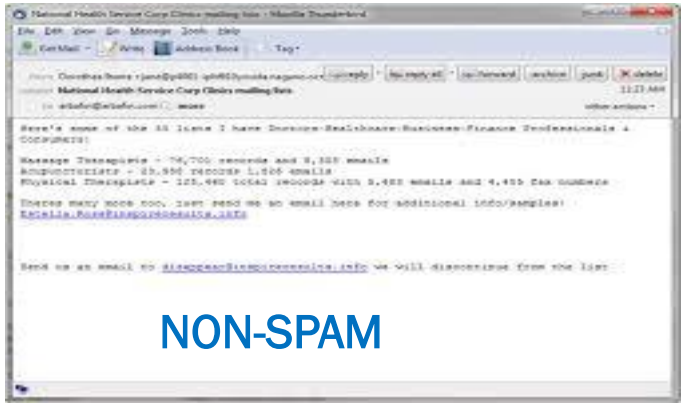
Application: SPAM Filtering

Training data (sample)

SPAM EMAIL



We want to check whether or not this email is SPAM ?



NON-SPAM

Statistical Model

P(SPAM | that email) = 0.8
P(NON-SPAM | that email) = 0.2

We can say, that email is SPAM ☺

Simple case is based on *Naive Bayes Classifier*

Application: Statistical Machine Translation

Parallel Corpus

Saya suka makan sup I like to eat soup
Dia pergi ke depok She goes to depok
Saya cinta dia I love him
Aku suka berbelanja I love shopping
Mereka suka makan They love eating
Saya pergi berbelanja di hari libur I go shopping on holiday

I love him

INPUT

i	saya	3
	aku	1
like	suka	1
love	suka	2
	cinta	1
she	dia	1
...		

Statistical Translation Model

Saya suka dia

OUTPUT

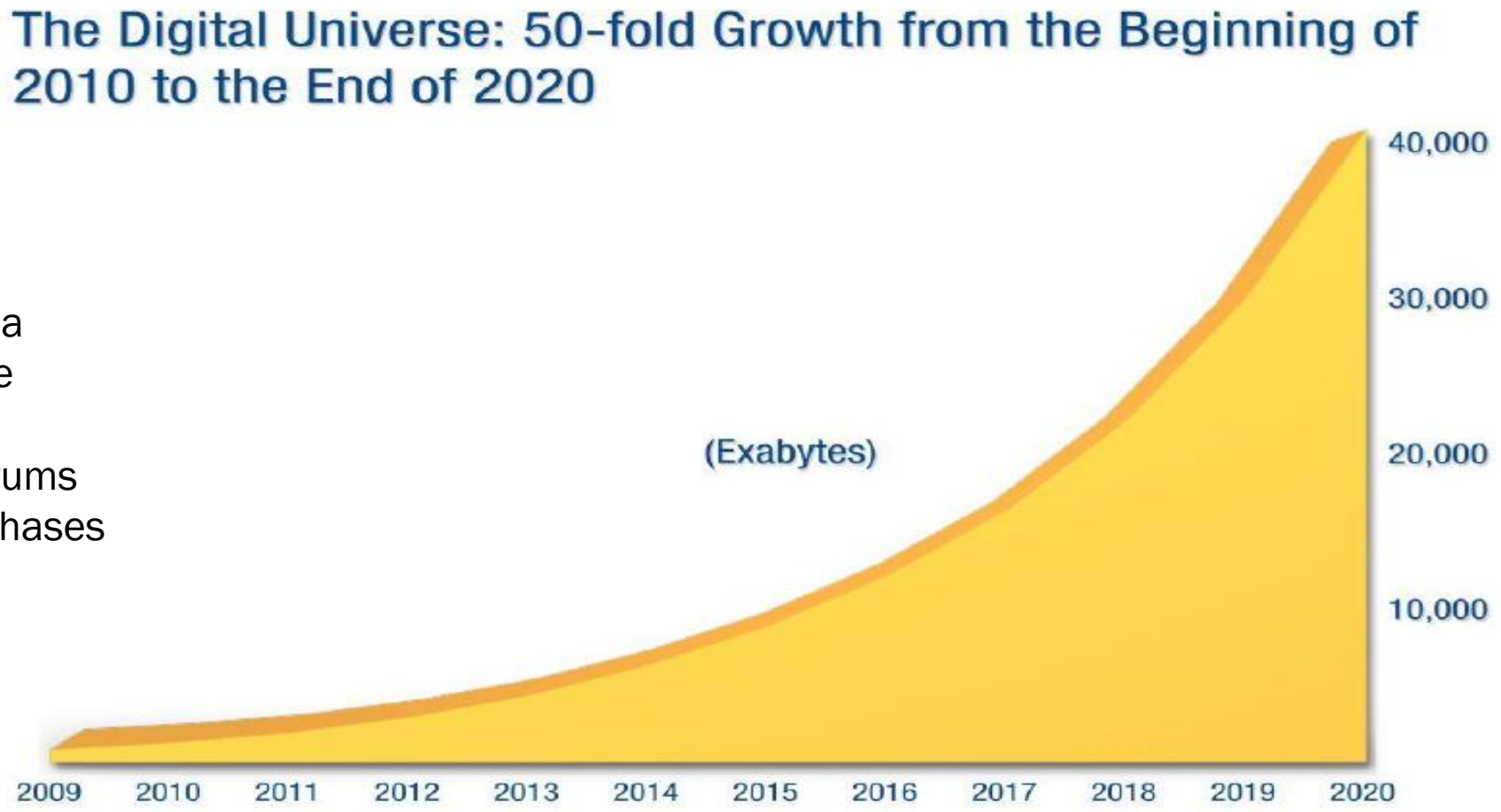
This is the simple case of SMT ☺

Data Scientist

"The **SEXIEST** Job of The 21st Century",
Thomas H. Davenport

Image: http://dwjunkie.wordpress.com/2013/07/24/data-sciencewhats-the-big-deal-about-it/

Digital Universe

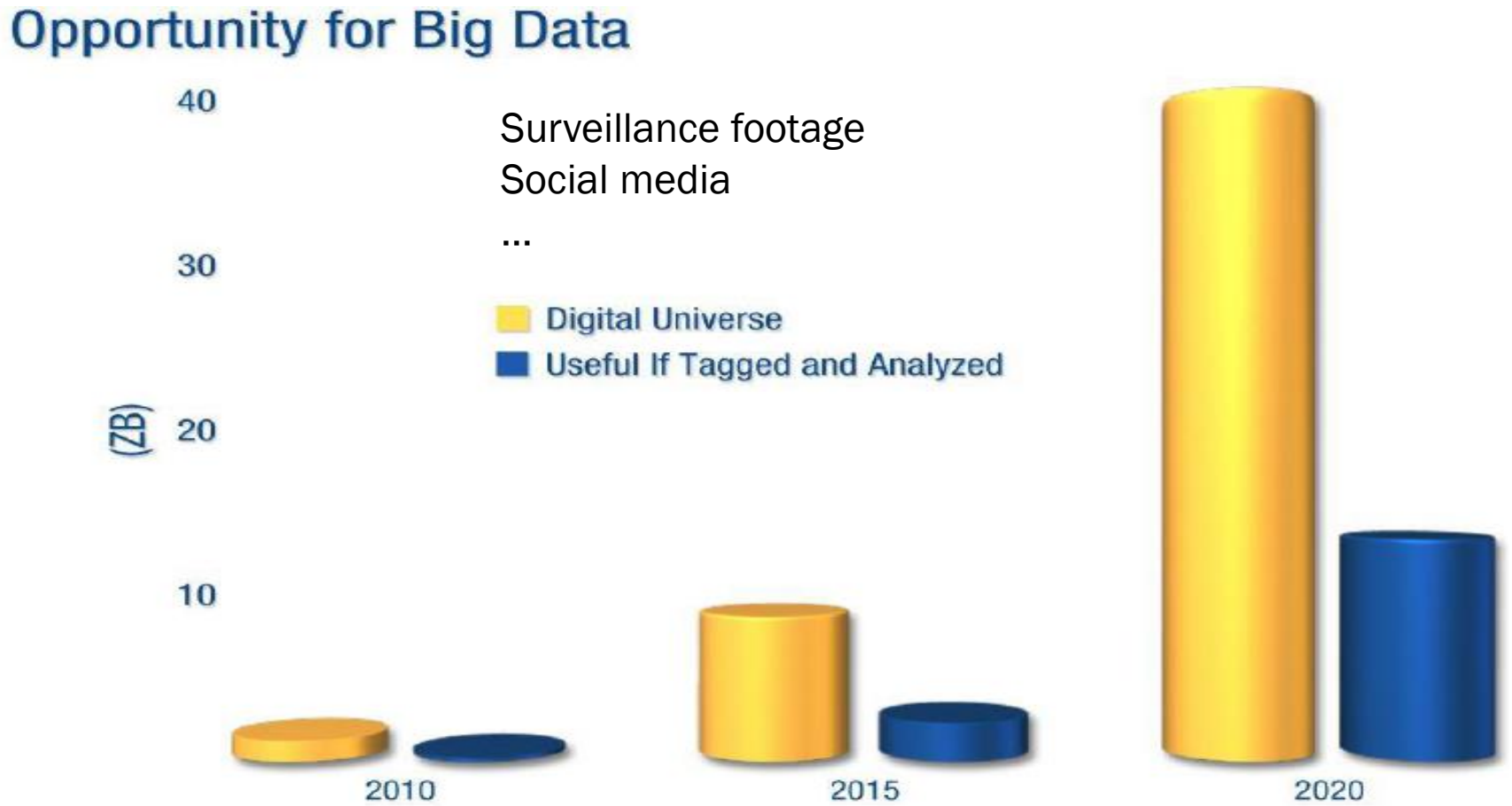


Example:
Social Media
News Article
Weblogs
Internet Forums
Online Purchases
...

Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Big Data

Big Data is part of digital universe. If it is tagged and analyzed, it will **provide useful knowledge** !



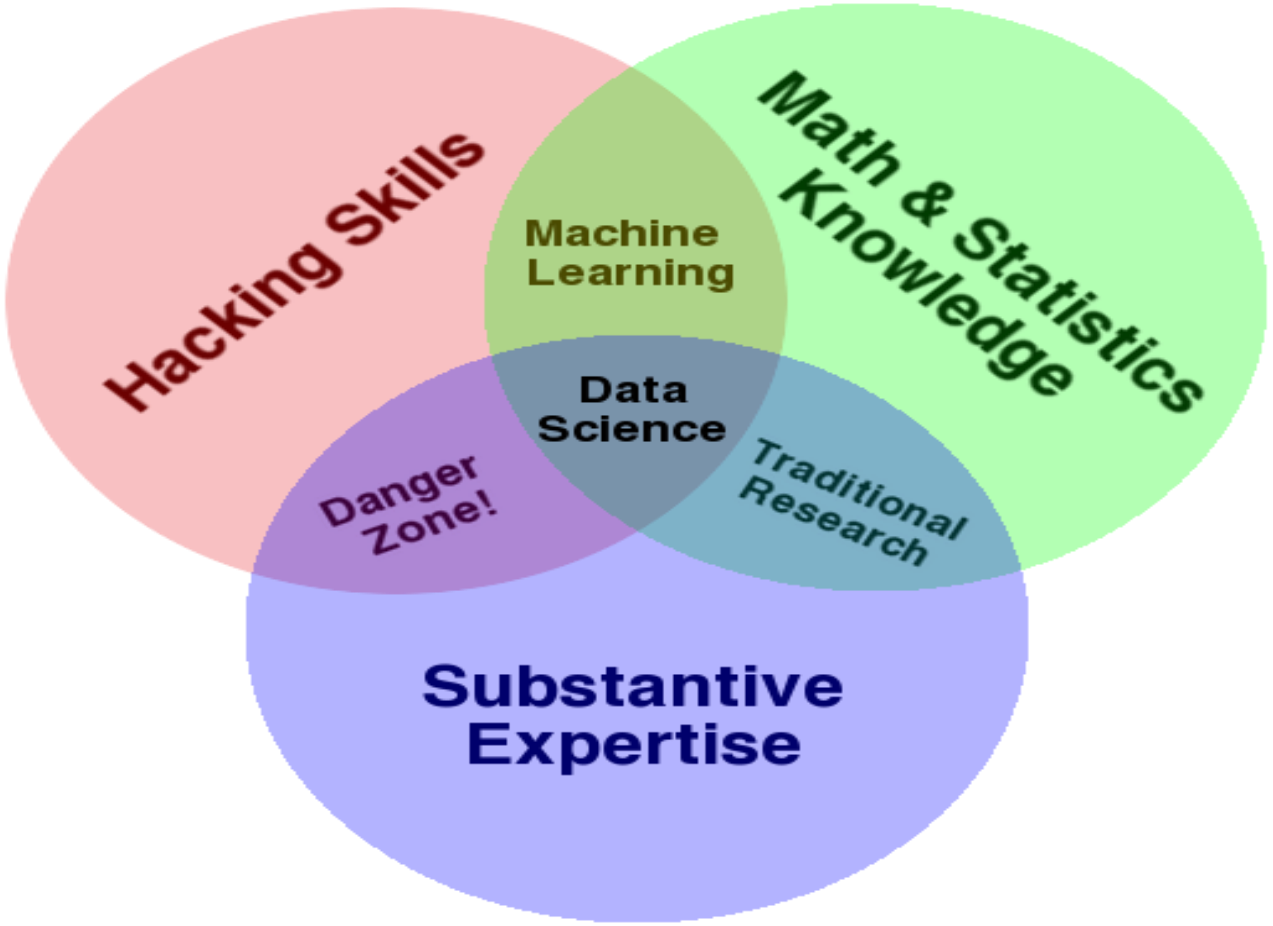
Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Big Data Gap

in practice, **only 3%** of the potentially useful data is **tagged**, and even **less is analyzed**.

Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

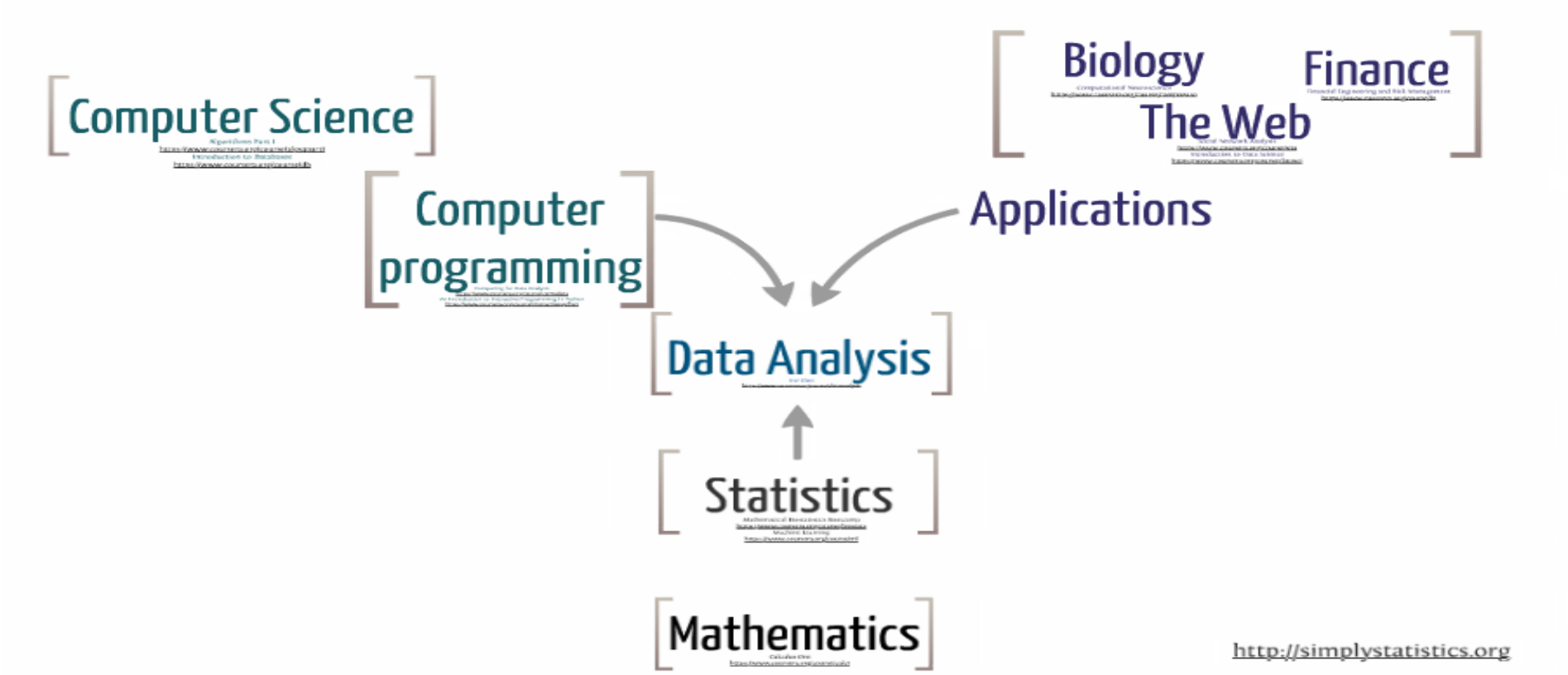
Data Science Venn Diagram



By **Drew Conway** Data Consulting, LLC. 2013

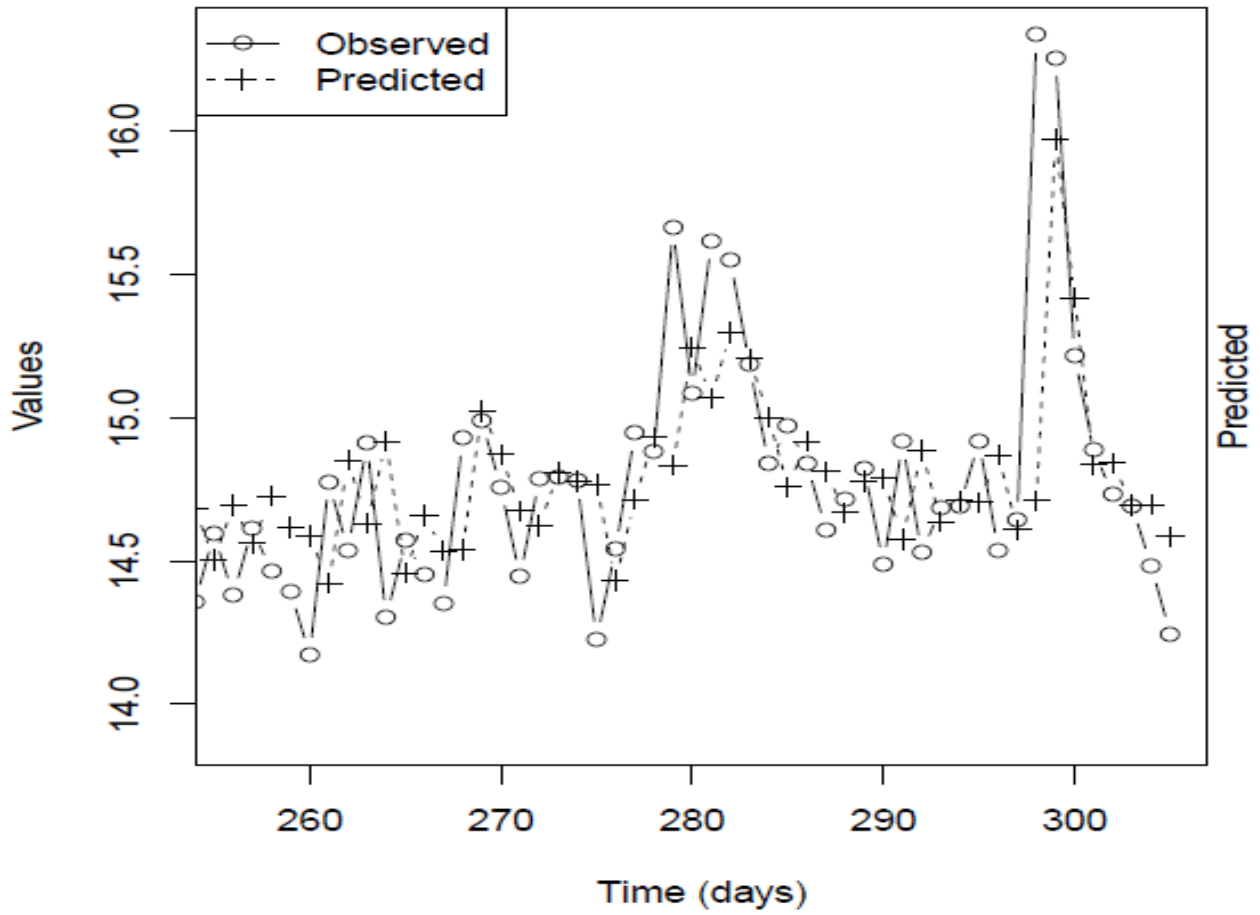
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Simple Data Analysis



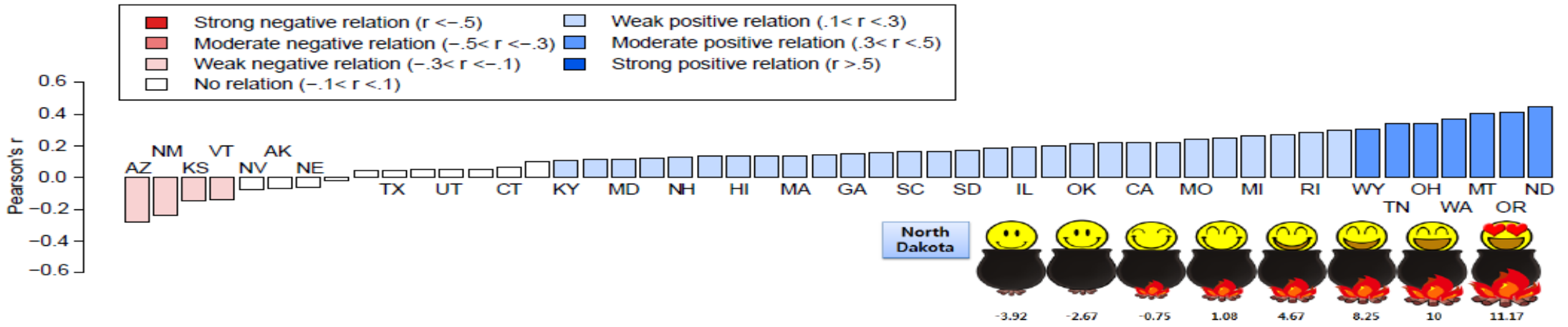
http://digitheadslabnotebook.blogspot.com/2013/02/data-analysis-class.html

Tweets for predicting stock market

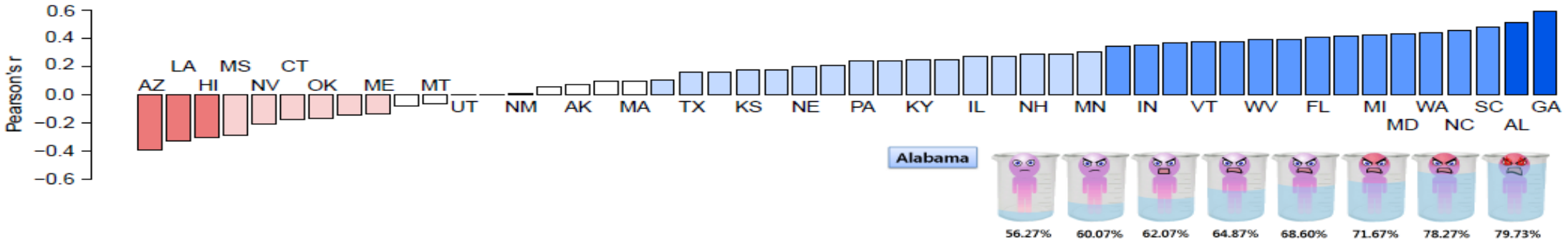


Nuno Oliveira, Paulo Cortez, Nelson Areal: **On the Predictability of Stock Market Behavior Using StockTwits Sentiment and Posting Volume.** EPIA 2013: 355-365

Mood & Weather



(a) Correlation between temperature and positive affect



(b) Correlation between humidity and negative affect

Park et al., Mood and Weather: Feeling the Heat ?, ICWSM 2013 (poster paper).

Twitter Can Predict Election ?

Table 4: Share of tweets and election results

Party	All mentions		Election	
	Number of tweets	Share of Twitter traffic	Election result*	Prediction error
CDU	30,886	30.1%	29.0%	1.0%
CSU	5,748	5.6%	6.9%	1.3%
SPD	27,356	26.6%	24.5%	2.2%
FDP	17,737	17.3%	15.5%	1.7%
LINKE	12,689	12.4%	12.7%	0.3%
Grüne	8,250	8.0%	11.4%	3.3%
			MAE:	1.65%

** Adjusted to reflect only the 6 main parties in our sample*

Mean Average Error

6 Parties in German election 2009

Tumasjan et al., Predicting Elections with Twitter: What 140 Characters Reveal about Political Statements, ICWSM 2010.

Jakarta: the most active *twitter* city

Table 1: Top 20 cities by percent of Twitter Decahose georeferenced tweets 23 October 2012 to 30 November 2012.	
City	**Percentage georeferenced tweets**
Jakarta	2.86
New York City	2.65
São Paulo	2.62
Kuala Lumpur	2.10
Paris	2.03
Istanbul	1.60
London	1.57
Rio de Janeiro	1.39
Chicago	1.28
Madrid	1.17
Los Angeles	1.14
Singapore	1.05
Houston	1.04
Mexico City	1.03
Philadelphia	0.99
Dallas	0.91
Manila	0.90
Brussels	0.88
Tokyo	0.85
Moscow	0.77

http://firstmonday.org/article/view/4366/3654

Social Media as early indicator of an unemployment spike

Challenge

Can social media add depth to unemployment statistics ?

Solution

1. Collect digital data (social media, blogs, forums, news articles) related to unemployment.
2. Perform sentiment analysis to categorize the mood of these online conversations.
3. Correlate volume of mood-related conversation to official unemployment statistics.

Source: IQ (Intelligence Quarterly), Journal of Advanced Analytics, 4Q 2013

SHARE QUESTION

Like 928

Tweet 110

QUESTION TOPICS

Gender Relations

Girls and Young Women

Interpersonal Interaction

Women

What does it mean when a girl smiles at you every time she sees you?

I get lots of smiles and a few hugs, the advantage of being 99 and still driving nice wheels, nite/day. A Happy Bachelor!

Follow Question 630 Comments 18+

156 ANSWERS ASK TO ANSWER

Mark Eichenlaub, graduate student in physics
17.5k upvotes by Abdul Rahman, Carlos Whitt, Oshea Waite, (more)

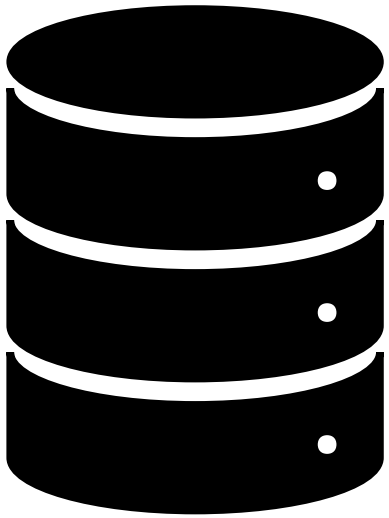
It's simple. Just use Bayes' theorem.

The probability she likes you is

$$P(like|smile) = \frac{P(smile|like)P(like)}{P(smile)}$$

$P(like|smile)$ is what you want to know - the probability she likes you given the fact that she smiles at you.

Just a Joke ! ☺



BASIC CONCEPTS

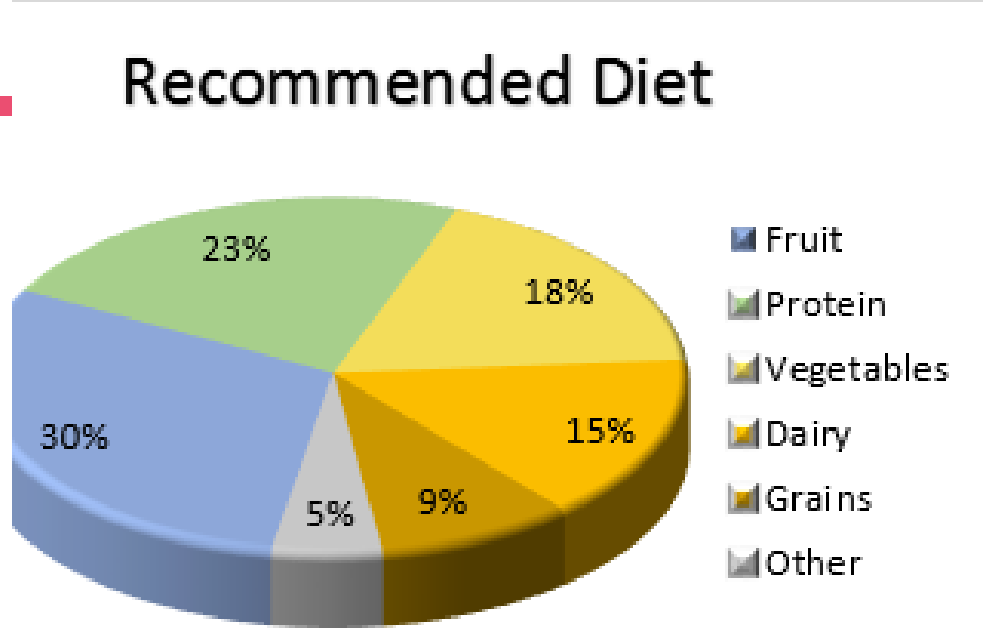
Two Parts of Statistics

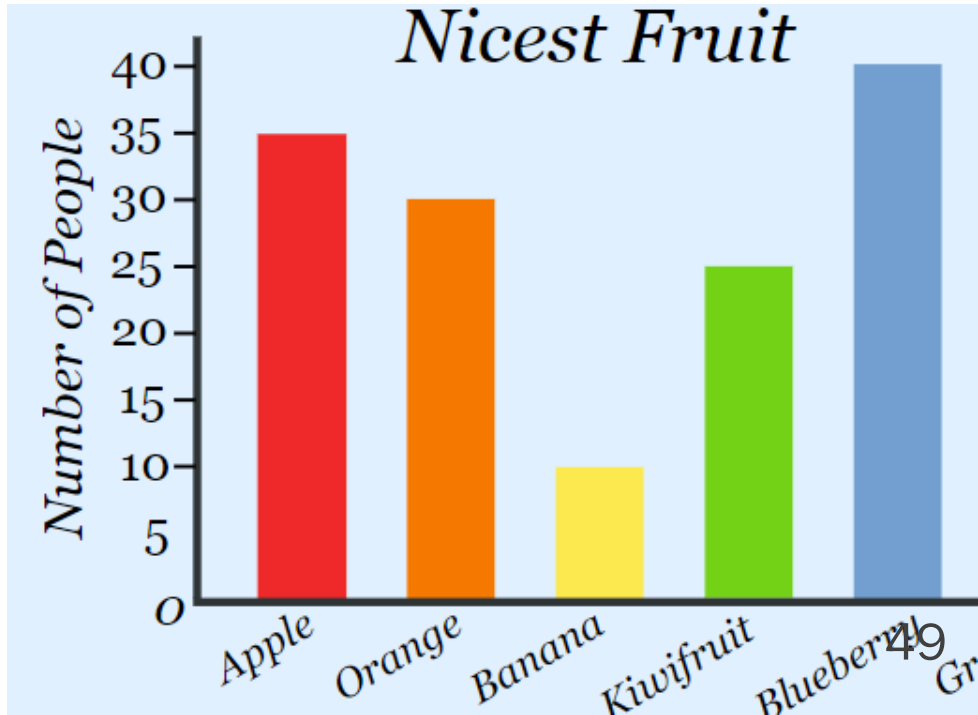
DESCRIPTIVE STATISTICS

INFERENTIAL STATISTICS

DESCRIPTIVE STATISTICS

- Gives description (presentation) of data
  - Output: tables or graphs.
- Gives summarization of data
  - Output: numerical quantity from data (mean, median, variance, mode, etc.)





INFERENTIAL STATISTICS

- Involves techniques for drawing conclusions
- Making inferences about a **population** from the **samples**.

SOME DEFINITIONS

Data & Data Set

Population & Sample

Parameters & Statistics

Variables

Scale of measurement

Data & Data Set

Data & Data Set

Data (plural)

Measurements or observations

Data Set

A collection of measurements or observations

Datum (singular)

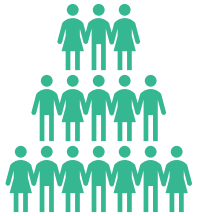
A Single measurement or observation and is commonly called as score or raw score.

Population & Sample

Population & Sample

- Let's study the habits of **all** UI students

Population & Sample



Population

- A total collection of elements being studied
- Complete set of individuals, objects, or scores of interest

Sample

- Population is often too large to examine
- Sample is a group of subjects selected from a population
- The sample must be informative about the total population (representative of that population).
  - Completely RANDOM!

Parameters & Statistics

Parameters

- Descriptive measures of a population
- Quantities that describe a population characteristics.
- Usually unknown, why ? ☺
  - Ex: The mean of all UI students' GPA.

Statistics

- Descriptive measures of a sample
  - Ex: The mean of 100 UI students' GPA.
- Mean statistic is then used to make statistical inferences about the parameter, i.e., population's mean.

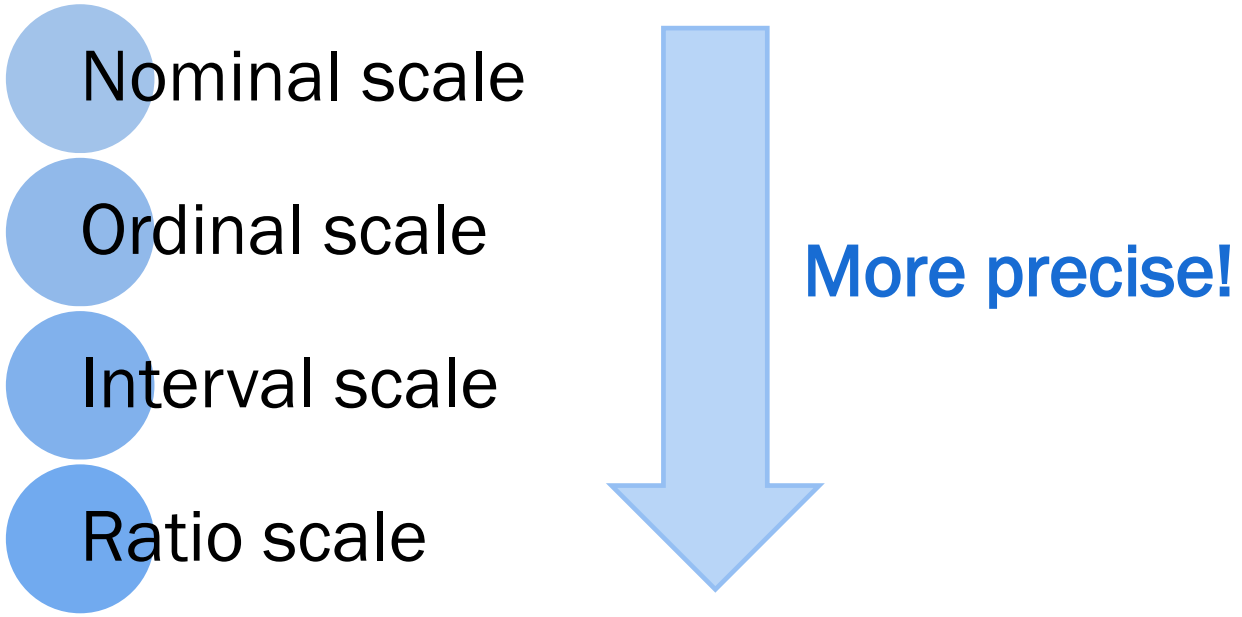
A parameter is to a population as a statistic is to a sample

Measuring the Data

- The data collected on variables are the result of **measurement.**
- **Measurement is a process of assigning numbers to characteristics according to a defined rule.**
- Not all measurements are measured the same:
  - Precise: the person is six feet, five inches.
  - Less-precise: the person is tall.
- Precision of measurement of a variable is important in determining what statistical method should be used to analyze the data in a study.

Scale of Measurement

- Measurement scales of variables are classified in a hierarchy based on their degree of precision.



Nominal Scale

- Least precise measurement scale.
- Data categories are **mutually exclusive**; that is, an object can belong to only one category.
- Data categories have no logical order.
- Example:
  - Gender
  - Color of eyes
  - Blood types

Ordinal Scale

- Data categories are **mutually exclusive**
- Data categories have **some logical order.**
- Data categories are scaled according to the amount of the particular characteristics they possess.
- Differences in the amount of the measured characteristic are indiscernible.
- Example: Your Grade : A, B, C, D, E.
  - We cannot infer: difference between A and B = difference between D and E ?

Interval Scale

- Data categories are **mutually exclusive** and have **logical order**.
- Data categories are scaled according to the amount of the characteristics they possess.
- **Equal differences** are represented by equal differences in the numbers assigned to the categories.
- Point 0 is just another point on the scale.
- Example: Temperature
  - Difference between 23’C and 20’C is the same with difference between 100’C and 97’C, i.e., 3’C.

Ratio Scale

- Most precise measurement scale.
- Data categories are mutually **exclusive** and have **logical order**.
- Data categories are scaled and the **equal differences** are represented by equal differences in the numbers.
- **Point 0 reflects an absence of the characteristics.**
- Example: Weight, Height
  - We cannot say 50'C is twice as warm as 25'C.
  - But, 50 KG really weights twice as much as 25 KG

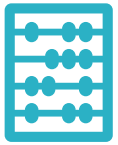
Variables

- Feature characteristic or attribute that can take on different values for different members of a group being studied.
- Types of variables 1:

Quantitative variable

Qualitative variable

- Type of variables 2:



Discrete variable

Continuous variable

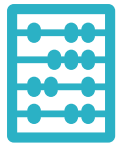
Qualitative & Quantitative Var.

Quantitative variable

Qualitative variable

- Qualitative (Nominal) Variable
  - A variable measured on the nominal or ordinal scale
  - Measurement consists of unordered or ordered discrete categories.
  - Example: blood group, color
- Quantitative Variable
  - A variable measured on the interval or ratio scale
  - Described by a number
  - Example: weight & height of people, time till cure

Discrete & Continuous Var.



Discrete variable

Continuous variable

- Discrete Variable
  - Variable can only take one of a finite or countable number of values
  - Example: a number of admissions at a hospital
- Continuous Variable
  - A measurement which can take any value in an interval of the real line
  - Example: Weight, Height, etc.