irevolution.net

TOPIC

TRANSPORTATION - SENSORS - AND THE SMART CITY
GEOG 491/591 ADVANCED GIS

DATE **4-6-15**

LECTURE **#3 - DATA FOR THE SMART CITY**

- **DATA FROM SENSORS**
  What part does data play in the ecosystem of the smart city?

- **BIG DATA**
  How does Big Data fit into a “Sensored Smart City”

- **BALANCE**
  balance the practical, nuts and bolts aspects of data with the conceptual and philosophical

TOPIC

GOALS - OBJECTIVES - OUTCOMES

DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

Data

From Wikipedia, the free encyclopedia

For data in computer science, see Data (computing). For other uses, see Data (disambiguation).

Data (/ˈdeɪtə/ DAY-tə, /ˈdætə/ DA-tə, or /ˈdɑːtə/ DAH-tə)[1] is a set of values of qualitative or quantitative variables; restated, pieces of data are individual pieces of information. Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs or images. Data as a general concept refers to the fact that some existing information or knowledge is *represented* or *coded* in some form suitable for better usage or processing.

Raw data, i.e., unprocessed data, refers to a collection of numbers, characters and is a relative term; data processing commonly occurs by stages, and the "processed data" from one stage may be considered the "raw data" of the next. Field data refers to raw data that is collected in an uncontrolled in situ environment. Experimental data refers to data that is generated within the context of a scientific investigation by observation and recording.

The word "data" has been considered the plural of "datum" until recently, but is now generally used in the singular as a mass noun.[2][*dubious – discuss*]

Contents [hide]

Meaning of data, information and knowledge [edit]

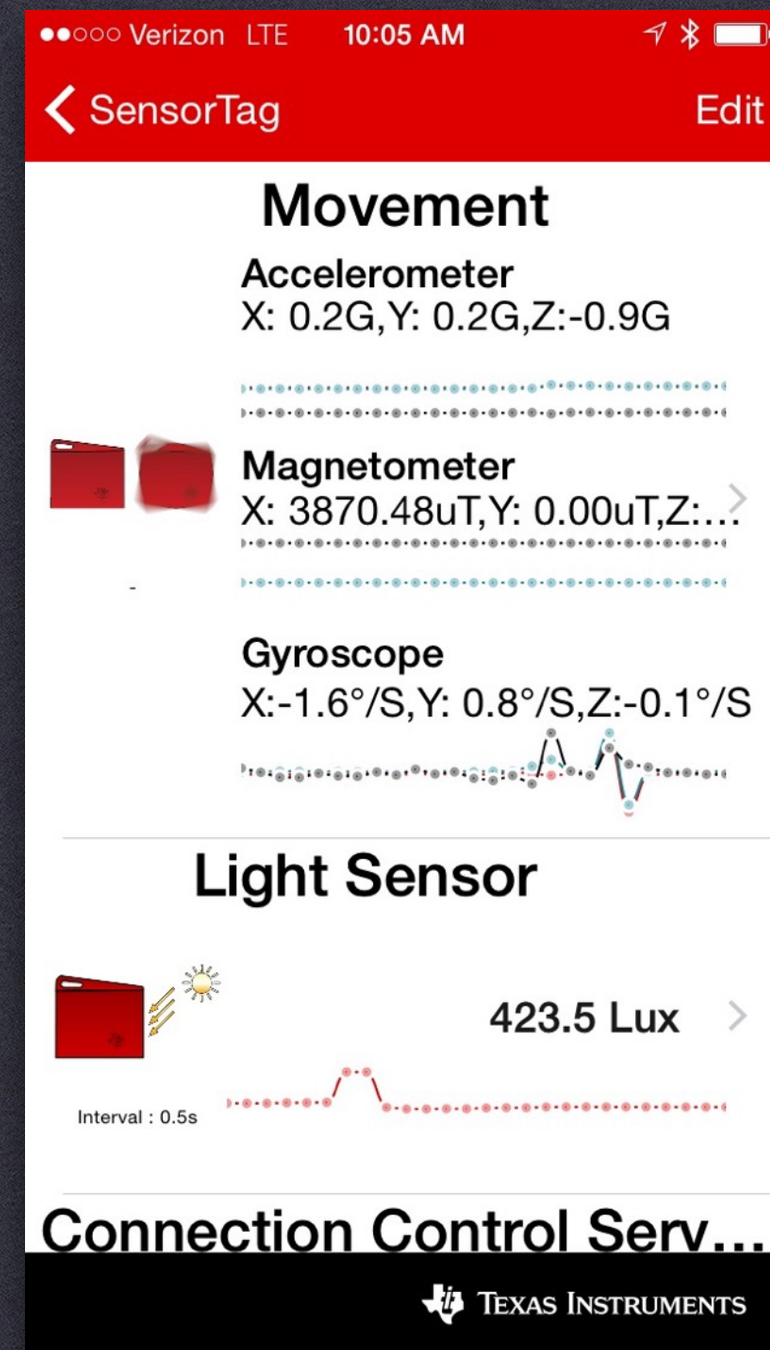
Data, information and knowledge are closely related terms, but each has its own role in relation to the other. Data is collected and analyzed to create information suitable for making decisions,[3] while knowledge is derived from extensive amounts of experience dealing with information on a subject. For example, the height of Mt. Everest is generally considered data. This data may be included in a book along with other data on Mt. Everest to describe the mountain in a manner useful for those who wish to make a decision about the best method to climb it. Using an understanding based on experience climbing mountains to advise persons on the way to reach Mt. Everest's peak may be seen as "knowledge".

TOPIC

DATA

DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

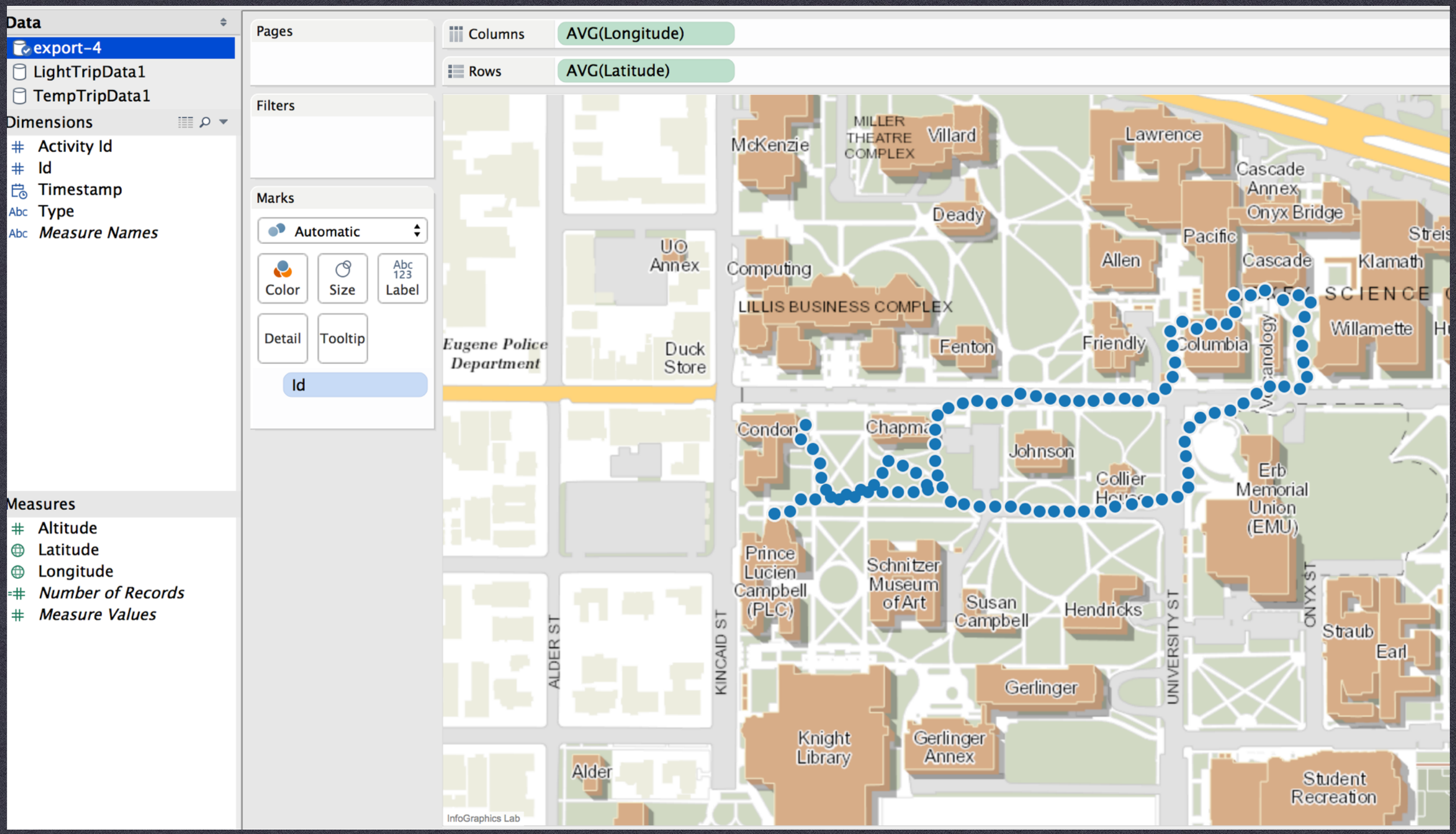


TOPIC

A SENSOR EXPERIMENT

DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

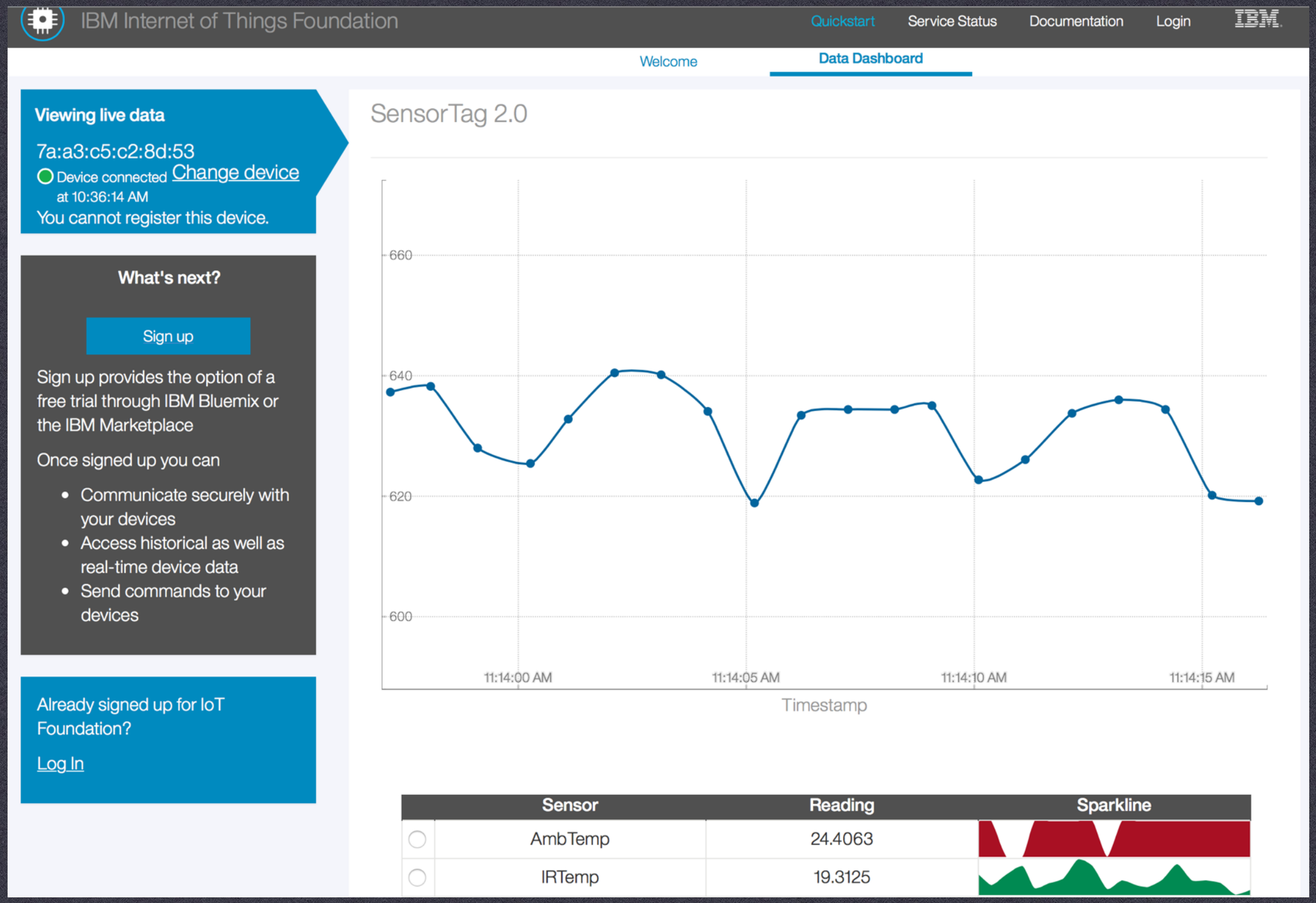


TOPIC

A SENSOR EXPERIMENT

DATE **4-6-15**

LECTURE **#3 - DATA FOR THE SMART CITY**



TOPIC

A SENSOR EXPERIMENT

DATE **4-6-15**

LECTURE **#3 - DATA FOR THE SMART CITY**

- THE REAL-TIME CITY? BIG DATA AND SMART URBANISM
  GEOJOURNAL
- BIG DATA BASIC CONCEPTS AND BENEFITS EXPLAINED
  TechRepublic

TOPIC

ASSIGNED READINGS FOR TODAY

DATE	4-6-15	LECTURE	#3 - DATA FOR THE SMART CITY

- **OPEN DATA VERSUS CLOSED DATA**

  Informing internally versus opening to the public

- **RAW VERSUS VISUALIZED DATA**

  What exactly are you sharing? How useful is it?

- **HISTORICAL CHALLENGES — QUANTITATIVE DATA AND QUALITATIVE DATA**

  relied on samples, are generated on a non-continuous basis, the number of variables are quite small, are aggregated to a relatively coarse spatial scale, and are often limited in access.

  questionnaire surveys, case studies, city audits, inter- views and focus groups, and ethnographies—that capture a relatively limited sample of data that are tightly focused, time and space specific, restricted in scope and scale, and relatively expensive to generate and analyze, t

- **DEFINITION**

  big data consists of massive, dynamic, varied, detailed, inter-related, low cost datasets that can be connected and utilized in diverse ways, thus offering the possibility of studies shifting from: data-scarce to data-rich; static snapshots to dynamic unfoldings; coarse aggregation to high resolution; relatively simple hypotheses and models to more complex, sophisticated simulations and theories

TOPIC

SMART CITY DATA - CONCEPTS (KITCHIN ARTICLE)

DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

- **POLITICS OF BIG URBAN DATA**

  "raw data is an oxymoron" - Data inflected by social privilege and social values

- **TECHNOCRATIC GOVERNANCE AND CITY DEVELOPMENT**

  "It's not me , it's the data" (Haque).

- **CORPORATIZATION OF CITY GOVERNANCE AND CORPORATE LOCK-IN**

  "smart cities are currently more vendor push than city government pull based" (Shaffers)
  The comparison to US auto industry and urban development.

- **BUGGY, BRITTLE AND HACKABLE CITIES**

  "What if the seeds of smart cities' own destruction are already built into their DNA? What if the smart cities of the future are buggy and brittle? What are we getting ourselves into?"

- **THE PANOPTIC CITY**

  Big Brother?

TOPIC

SMART CITY DATA - CHALLENGES (KITCHIN ARTICLE)

DATE	4-6-15	LECTURE	#3 - DATA FOR THE SMART CITY

- **VOLUME**

  90% of all data ever created, was created in the past 2 years.

- **VELOCITY**

  Velocity is the speed at which the data is created, stored, analyzed and visualized.

- **VARIETY**

  The wide variety of data requires a different approach as well as different techniques to store all raw data.

- **VERACITY**

  Is the data reliable - accurate - actionable - correct

- **VARIABILITY**

  Meaning is changing (rapidly)

- **VISUALIZATION**

  Making all that vast amount of data comprehensible in a manner that is easy to understand and read

- **VALUE**

  What's it worth? The value is in the analyses done on that data and how the data is turned into information and eventually

TOPIC

BIG DATA - SEVEN V'S

Why The 3V's Are Not Sufficient To Describe Big Data, Mark van Rijmenam, 2013

DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

- **VOLUME**

  …the amount of data in the world will double every two years. By 2020, we will have 50 times the amount of data as that we had in 2011. The sheer volume of the data is enormous and a very large contributor to the ever expanding digital universe is the Internet of Things with sensors all over the world in all devices creating data every second. The era of a trillion sensors is upon us.

  If we look at airplanes they generate approximately 2.5 billion Terabyte of data each year from the sensors installed in the engines. Self-driving cars will generate 2 Petabyte of data every year. Also the agricultural industry generates massive amounts of data with sensors installed in tractors. Shell uses super-sensitive sensors to find additional oil in wells and if they install these sensors at all 10.000 wells they will collect approximately 10 Exabyte of data annually. That again is absolutely nothing if we compare it to the Square Kilometer Array Telescope that will generate 1 Exabyte of data per day.

  In the past, the creation of so much data would have caused serious problems. Nowadays, with decreasing storage costs, better storage solutions like Hadoop and the algorithms to create meaning from all that data this is not a problem at all.

People	Sensor Count	Sensor Type	Second	Minute	Hour	Day	Week	Month	Year	
1	1	Light	0.040666667	2.44	146.4	3513.6	24595.2	98380.8	1180569.6	
2	1	Light	0.081333333	4.88	292.8	7027.2	49190.4	196761.6	2361139.2	
4	1	Light	0.325333333	19.52	1171.2	28108.8	196761.6	787046.4	9444556.8	
16	1	Light	5.205333333	312.32	18739.2	449740.8	3148185.6	12592742.4	151112908.8	
256	1	Light	1332.565333	79953.92	4797235.2	115133644.8	805935514	3223742054	38684904653	
65536	1	Light	87331001.69	5239.860101	314391.6061	7545398.546	52817789.8	211271159.3	2535253911	GB
4294967296	1	Light	3.75084E+17	22505027.77	1350301666	32407239989	226850680	907402719.7	10888832636	TB
1.84467E+19	1	Light	6.91907E+36	4.15144E+17	2.49087E+28	5.97808E+29	4.1847E+27	1.67386E+28	2.00864E+29	
3.40282E+38	1	Light	2.35444E+75	1.41266E+56	8.47598E+66	2.03424E+68	1.424E+66	5.69586E+66	6.83503E+67	
1.15792E+77	1	Light	2.7263E+152	1.6358E+133	9.8145E+143	2.3555E+145	1.649E+143	6.5954E+143	7.9144E+144	

TOPIC

BIG DATA - VOLUME

DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

- # VELOCITY

  The Velocity is the speed at which the data is created, stored, analyzed and visualized. In the past, when batch processing was common practice, it was normal to receive an update from the database every night or even every week. Computers and servers required substantial time to process the data and update the databases. In the big data era, data is created in real-time or near real-time. With the availability of Internet connected devices, wireless or wired, machines and devices can pass-on their data the moment it is created.

  The speed at which data is created currently is almost unimaginable: Every minute we upload 100 hours of video on Youtube. In addition, every minute over 200 million emails are sent, around 20 million photos are viewed and 30.000 uploaded on Flickr, almost 300.000 tweets are sent and almost 2,5 million queries on Google are performed.

- **VARIETY**

  In the past, all data that was created was structured data, it neatly fitted in columns and rows but those days are over. Nowadays, 90% of the data that is generated by organization is unstructured data. Data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data. The wide variety of data requires a different approach as well as different techniques to store all raw data.

  There are many different types of data and each of those types of data require different types of analyses or different tools to use. Social media like Facebook posts or Tweets can give different insights, such as sentiment analysis on your brand, while sensory data will give you information about how a product is used and what the mistakes are.

LightTripData1.csv

A89 | 4/3/2015 10:37:00 AM

	A	B
178	4/3/15 10:37	154.4
179	4/3/15 10:37	155.1
180	4/3/15 10:37	160.6
181	4/3/15 10:37	163
182	4/3/15 10:37	166.2
183	4/3/15 10:37	167
184	4/3/15 10:37	163.5
185	4/3/15 10:37	160
186	4/3/15 10:37	160
187	4/3/15 10:37	160.5
188	4/3/15 10:37	161
189	4/3/15 10:37	161.1
190	4/3/15 10:37	161.1
191	4/3/15 10:37	161
192	4/3/15 10:37	161
193	4/3/15 10:37	161
194	4/3/15 10:37	160.6
195	4/3/15 10:37	160.6
196	4/3/15 10:37	161
197	4/3/15 10:37	162.9
198	4/3/15 10:37	168.6
199	4/3/15 10:37	172.2
200	4/3/15 10:37	176.8
201	4/3/15 10:37	181.3
202	4/3/15 10:37	181.6
203	4/3/15 10:37	200.8
204	4/3/15 10:37	219.8
205	4/3/15 10:37	248.7
206	4/3/15 10:37	278.7
207	4/3/15 10:37	279.8
208	4/3/15 10:37	288.5
209	4/3/15 10:38	297.1
210	4/3/15 10:38	148.6
211	4/3/15 10:38	0.1
212	4/3/15 10:38	0.1
213	4/3/15 10:38	80.5
214	4/3/15 10:38	160.5
215	4/3/15 10:38	160
216	4/3/15 10:38	160.6

LightTripData1.csv

H4

	A	B	C	D	E	F	G
1	Id	ActivityId	Timestamp	Latitude	Longitude	Altitude	Type
2	1	540732196	0	44.045318	-123.07793	135.730966	start
3	2	540732196	12.0269206	44.045232	-123.07797	136.027982	gps
4	3	540732196	20.0879846	44.045169	-123.07786	136.395443	gps
5	4	540732196	25.0294575	44.045083	-123.0778	136.856539	gps
6	5	540732196	33.0836148	44.04499	-123.07779	137.192921	gps
7	6	540732196	42.0208181	44.044902	-123.07774	137.373266	gps
8	7	540732196	49.9173852	44.044847	-123.07763	137.56561	gps
9	8	540732196	57.0878653	44.044864	-123.07749	137.838498	gps
10	9	540732196	64.0414666	44.044888	-123.07737	138.030109	gps
11	10	540732196	73.0534728	44.044895	-123.07724	138.29139	gps
12	11	540732196	82.0156465	44.044892	-123.0771	138.319483	gps
13	12	540732196	91.0227354	44.04489	-123.07696	138.221066	gps
14	13	540732196	99.0729927	44.044913	-123.07683	138.114615	gps
15	14	540732196	107.022555	44.044998	-123.07675	137.88823	gps
16	15	540732196	115.033602	44.045094	-123.07677	137.567979	gps
17	16	540732196	125.088812	44.045195	-123.07676	137.234527	gps
18	17	540732196	132.019504	44.045287	-123.07677	137.073594	gps
19	18	540732196	138.026347	44.045379	-123.07675	137.051832	gps
20	19	540732196	144.027632	44.045438	-123.07665	137.026414	gps
21	20	540732196	149.014103	44.045463	-123.07652	137.189517	gps
22	21	540732196	156.042277	44.045475	-123.0764	137.241807	gps
23	22	540732196	164.013975	44.045468	-123.07627	137.583232	gps
24	23	540732196	174.014337	44.04548	-123.07612	138.028228	gps
25	24	540732196	180.036365	44.045525	-123.07601	138.373477	gps
26	25	540732196	188.017956	44.045511	-123.07587	138.591117	gps
27	26	540732196	193.058234	44.045498	-123.07574	138.726504	gps
28	27	540732196	200.094243	44.045478	-123.07561	138.825392	gps
29	28	540732196	206.053038	44.045473	-123.07549	138.814955	gps
30	29	540732196	214.205534	44.045474	-123.07535	138.788854	gps
31	30	540732196	221.002581	44.045487	-123.07521	138.768746	gps
32	31	540732196	227.029392	44.045488	-123.07508	138.731936	gps
33	32	540732196	232.033267	44.045477	-123.07495	138.651053	gps
34	33	540732196	238.034987	44.0455	-123.07482	138.528678	gps
35	34	540732196	244.039194	44.04556	-123.07471	138.263774	gps
36	35	540732196	254.193152	44.04564	-123.07464	138.074521	gps
37	36	540732196	261.162273	44.045729	-123.07463	138.078699	gps
38	37	540732196	267.087858	44.045829	-123.07465	138.107785	gps

export-4.csv

Normal View | Ready

- ## VISUALIZATION

  This is the hard part of big data. Making all that vast amount of data comprehensible in a manner that is easy to understand and read. With the right analyses and visualizations, raw data can be put to use otherwise raw data remains essentially useless. Visualizations of course do not mean ordinary graphs or pie charts. They mean complex graphs that can include many variables of data while still remaining understandable and readable.

  Visualizing might not be the most technologically difficult part; it sure is the most challenging part. Telling a complex story in a graph is very difficult but also extremely crucial. Luckily there are more and more big data startups appearing that focus on this aspect and in the end, visualizations will make the difference. Visualizations help organizations answer questions they did not know to ask.

Home | Layout | Tables

Edit

Paste | Fill | Clear | Calibri (Body) | B I U

A89

	A	B	C
3	4/3/15 10:36	21	
4	4/3/15 10:36	20.9	
5	4/3/15 10:36	21.2	
6	4/3/15 10:36	21.4	
7	4/3/15 10:36	21.2	
8	4/3/15 10:36	21	
9	4/3/15 10:36	21	
10	4/3/15 10:36	21	
11	4/3/15 10:36	20.1	
12	4/3/15 10:36	19.2	
13	4/3/15 10:36	19.2	
14	4/3/15 10:36	19.3	
15	4/3/15 10:36	19.4	
16	4/3/15 10:36	19.4	
17	4/3/15 10:36	19.4	
18	4/3/15 10:36	21.2	
19	4/3/15 10:36	23	
20	4/3/15 10:36	23	
21	4/3/15 10:36	23	
22	4/3/15 10:36	92.6	
23	4/3/15 10:36	162.2	
24	4/3/15 10:36	113.4	
25	4/3/15 10:36	64.6	
26	4/3/15 10:36	64.6	
27	4/3/15 10:36	64.6	
28	4/3/15 10:36	64.6	
29	4/3/15 10:36	64.6	
30	4/3/15 10:36	64.6	
31	4/3/15 10:36	35.8	
32	4/3/15 10:36	6.9	
33	4/3/15 10:36	6.9	
34	4/3/15 10:36	6.9	
35	4/3/15 10:36	6.9	
36	4/3/15 10:36	6.9	
37	4/3/15 10:36	6.9	
38	4/3/15 10:36	6.9	
39	4/3/15 10:36	6.9	
40	4/3/15 10:36	6.9	
41	4/3/15 10:36	6.9	

Home | Layout | Tables

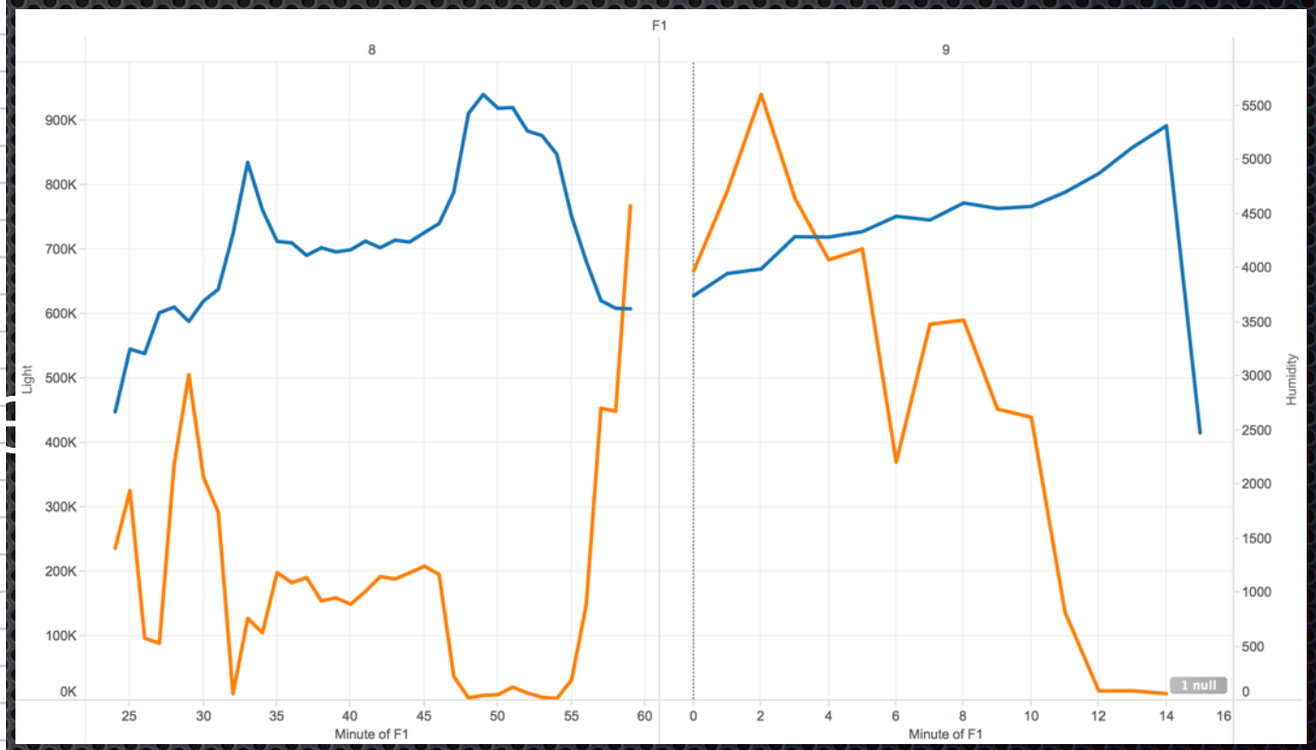
Edit

Paste | Fill | Clear | Calibri (Body) | B I U

A1

	A	B	C
1	4/3/15 8:24	0	0
2	4/3/15 8:24	1010.5	
3	4/3/15 8:24	1010.5	
4	4/3/15 8:24	1010.5	
5	4/3/15 8:24	1010.5	
6	4/3/15 8:24	1010.5	
7	4/3/15 8:24	1010.5	
8	4/3/15 8:24	1010.1	
9	4/3/15 8:24	1010.5	
10	4/3/15 8:24	1010.5	
11	4/3/15 8:24	1010.5	
12	4/3/15 8:24	1010.5	
13	4/3/15 8:24	1010.5	
14	4/3/15 8:24	1010.5	
15	4/3/15 8:24	1010.1	
16	4/3/15 8:24	1010.5	
17	4/3/15 8:24	1010.5	
18	4/3/15 8:24	1010.1	
19	4/3/15 8:24	1010.1	
20	4/3/15 8:24	1010.5	
21	4/3/15 8:24	1010.1	
22	4/3/15 8:24	1010.5	
23	4/3/15 8:24	1010.1	
24	4/3/15 8:24	1010.1	
25	4/3/15 8:24	1010.1	
26	4/3/15 8:24	1010.1	
27	4/3/15 8:24	1010.1	
28	4/3/15 8:24	1010.5	
29	4/3/15 8:24	1010.1	
30	4/3/15 8:24	1010.5	
31	4/3/15 8:24	1010.1	
32	4/3/15 8:24	1010.5	
33	4/3/15 8:24	1010.5	
34	4/3/15 8:24	1010.1	
35	4/3/15 8:24	1010.1	
36	4/3/15 8:24	1010.5	
37	4/3/15 8:24	1010.5	
38	4/3/15 8:24	1010.1	
39	4/3/15 8:24	1010.5	

Barometer.csv



- **VERACITY**

  Having a lot of data in different volumes coming in at high speed is worthless if that data is incorrect. Incorrect data can cause a lot of problems for organizations as well as for consumers. Therefore, organizations need to ensure that the data is correct as well as the analyses performed on the data are correct. Especially in automated decision-making, where no human is involved anymore, you need to be sure that both the data and the analyses are correct.

  If you want your organization to become information-centric, you should be able to trust that data as well as the analyses. accountability.

- # VARIABILITY

  Big data is extremely variable. Brian Hopkins, a Forrester principal analyst, defines variability as the “variance in meaning, in lexicon”. He refers to the supercomputer Watson who won Jeopardy. The supercomputer had to “dissect an answer into its meaning and […] to figure out what the right question was”. That is extremely difficult because words have different meanings and all depends on the context. For the right answer, Watson had to understand the context.

  Variability is thus very relevant in performing sentiment analyses. Variability means that the meaning is changing (rapidly). In (almost) the same tweets a word can have a totally different meaning. In order to perform a proper sentiment analyses, algorithms need to be able to understand the context and be able to decipher the exact meaning of a word in that context. This is still very difficult.

mappiness.org.uk

Available on the App Store

Home News Meters Maps More info

We don't want to wake you in the night! If you're outside the UK please click here for important time zone info.

mappiness maps happiness across space in the UK

mappiness is a free app for your iPhone

It's part of a research project at the London School of Economics

We'd love to have you on board!

how does it work?

- **You get mappiness from the App Store, open it, and sign up**
- We beep you once (or more) a day to ask how you're feeling, and a few basic things to control for: who you're with, where you are, what you're doing (if you're outdoors, you can also take a photo)
- The data gets sent back — anonymously and securely — to our data store, along with your approximate location from the iPhone's GPS

what's in it for you?

- **Interesting information about your own happiness**, which you can download or see charted inside the app — including when, where and with whom you're happiest
- The warm glow of helping increase the sum of human knowledge

what's in it for us?

- **We're particularly interested in how people's happiness is affected by their local environment** — air pollution, noise, green spaces, and so on — which the data from mappiness will be absolutely great for investigating
- We'll be publishing the results in academic journals and elsewhere — starting with this paper in *Global Environmental Change*.

mappiness.org.uk LSE

Available on the App Store

Home News Meters Maps More info

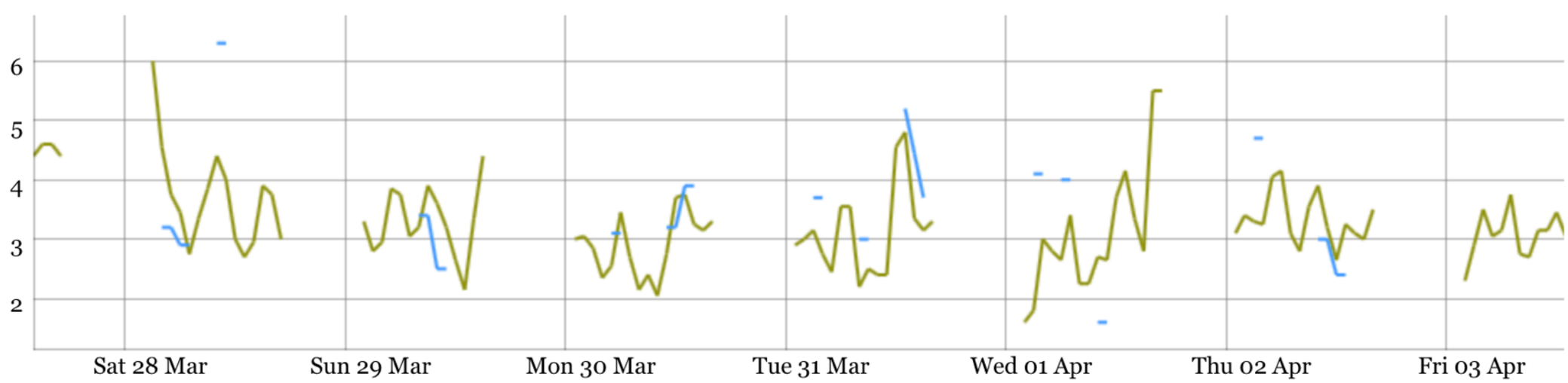
when are we happy?

The **hedonimeters** on the right display mappiness users' happiness in real-time, compared against the all-time average.

Below, happiness levels are charted hour by hour over the past week.

UK

London



Click and drag to zoom. Double-click to zoom back out.

- **VALUE**

  All that available data will create a lot of value for organisations, societies and consumers. Big data means big business and every industry will reap the benefits from big data. McKinsey states that potential annual value of big data to the US Health Care is $ 300 billion, more than double the total annual health care spending of Spain. They also mention that big data has a potential annual value of € 250 billion to the Europe's public sector administration. Even more, in their well-regarded report from 2011, they state that the potential annual consumer surplus from using personal location data globally can be up to $ 600 billion in 2020. That is a lot of value.

  Of course, data in itself is not valuable at all. The value is in the analyses done on that data and how the data is turned into information and eventually turning it into knowledge. The value is in how organizations will use that data and turn their organization into an information-centric company that relies on insights derived from data analyses for their decision-making.

- # LUX

  The primary sensor data type for ambient light sensors is illuminance in lux (lumens per square meter). The principles outlined in this topic are based on taking lux values as input and reacting to that data in a program.
  Lux readings are directly proportional to the energy per square meter that is absorbed per second. Human perception of light levels is not so straightforward. Human perception of light is complicated because our eyes are constantly adjusting and other biological processes are affecting our perception. However, we can think of this perception from a simplified perspective by creating several ranges of interest with known upper and lower thresholds.

  The following example data set represents rough thresholds for common lighting conditions, and the corresponding lighting step. Here, each lighting step represents a change in lighting environment.

Lighting condition	From (lux)	To (lux)	Mean value (lux)	Lighting step
Pitch Black	0	10	5	1
Very Dark	11	50	30	2
Dark Indoors	51	200	125	3
Dim Indoors	201	400	300	4
Normal Indoors	401	1000	700	5
Bright Indoors	1001	5000	3000	6
Dim Outdoors	5001	10,000	7500	7
Cloudy Outdoors	10,001	30,000	20,000	8
Direct Sunlight	30,001	100,000	65,000	9

TOPIC

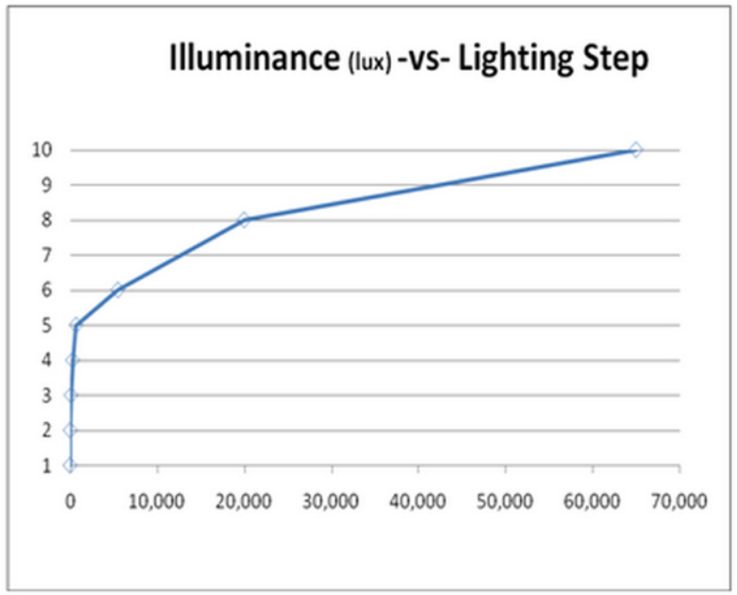
SENSOR DATA - LIGHT

https://msdn.microsoft.com/en-us/library/windows/desktop/dd319008%28v=vs.85%29.aspx

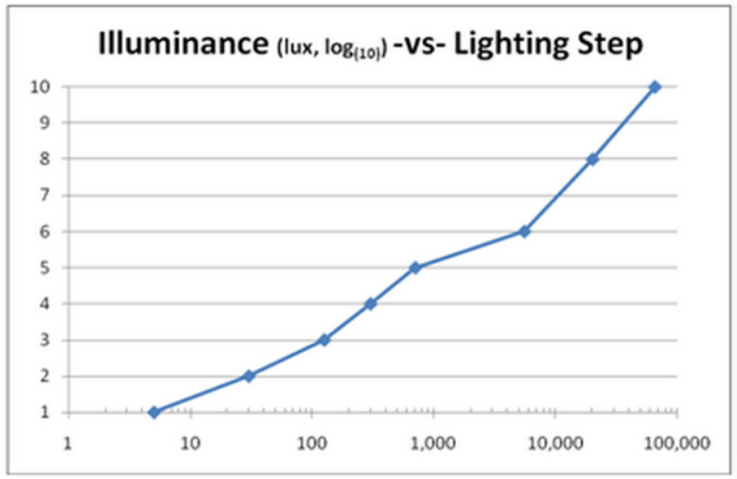
DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

If we visualize this data by using the mean values from this table, we see that the lux-to-lighting-step relationship is not linear, as show in the following graph.



However, if we view this data by using a logarithmic scale on the x-axis, we can see that a roughly linear relationship emerges.



An Example Transform

Based on the sample data set for ambient light sensors previously provided, you could arrive at the following equation to map lux values to human perception. In this example, the expected values range from 0 lux to 1,000,000 lux.

$$Light_{Normalized} = \frac{Log_{10}(x)}{5.0}$$

This equation results in values that vary in a roughly linear fashion between 0.0 and 1.0. This result indicates how human-perceived lighting changed based on the example data set that was shown previously.

TOPIC

SENSOR DATA - LIGHT

https://msdn.microsoft.com/en-us/library/windows/desktop/dd319008%28v=vs.85%29.aspx

DATE 4-6-15

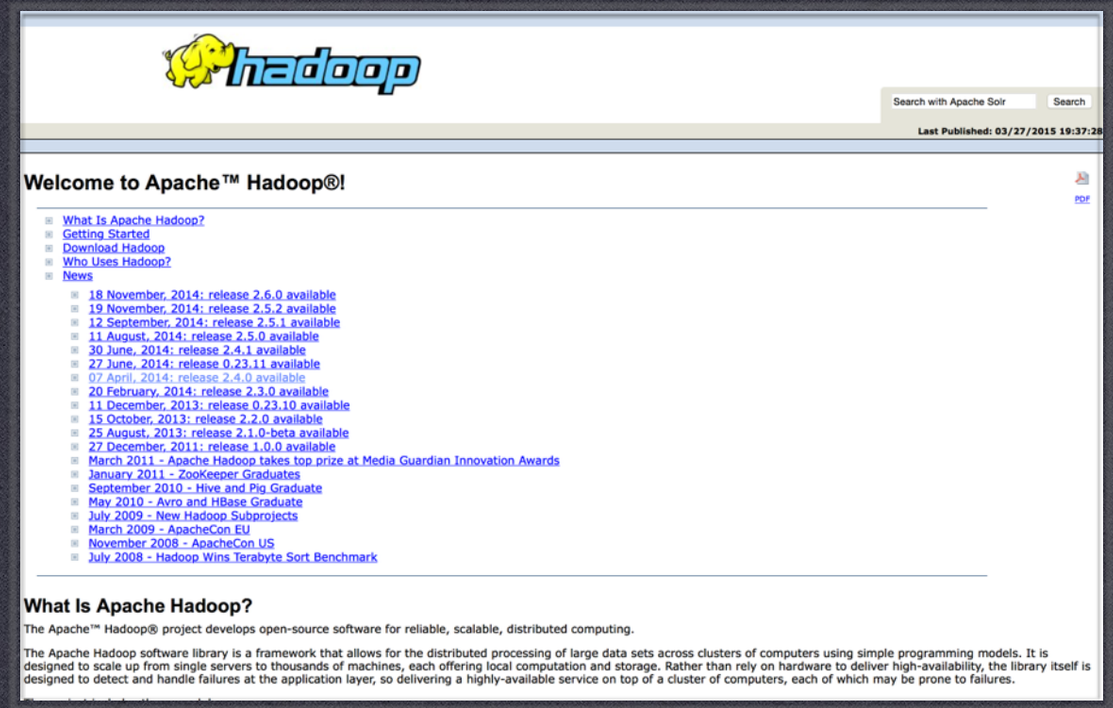
LECTURE #3 - DATA FOR THE SMART CITY

- **APACHE HADOOP**
  Wikipedia definition

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

- **STORAGE**
  Hadoop Distributed File System (HDFS)

- **PROCESSING**
  Map Reduce



TOPIC

HOW DO WE STORE AND PROCESS BIG DATA?

DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

Chris Stucchio Publications Work Code Blog

MapReduce explained in 41 words

Fri 05 August 2011 hadoop / mapreduce

Follow @stucchio Tweet 0

Like Share Be the first of your friends to like this.

g+1 +12 Recommend this on Google

Goal: count the number of books in the library.

Map: You count up shelf #1, I count up shelf #2.

(The more people we get, the faster this part goes.)

Reduce: We all get together and add up our individual counts.

TOPIC

MAP REDUCE

DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

▲ yummyfajitas 1339 days ago

The explanation I gave to my mom:

We want to count all the books in the library. You count up shelf #1, I count up shelf #2. That's map. The more people we get, the faster it goes.

Now we get together and add our individual counts. That's reduce.

[edit: created a blog post containing this explanation. Maybe it will be useful after this HN conversation vanishes. http://crazybear.posterous.com/mapreduce-explained-in-41-wor...]

▲ wisty 1339 days ago

That's the first iteration of the explanation.

Next iteration: But what if you want to count up non-fiction and fiction? There's just a slightly different "reduce" step, because people have to add up both of them. You could also break it down by subject, or author, or pretty much anything else.

Next iteration: Of course, you don't have to just add. You can do anything that has mathematical properties similar to addition (my mom teaches math at high-school). You could add up the set of all the offense words found, since addition and set addition have similar mathematical properties. You couldn't take the average number of pages, because some people will find more books than others, and you wouldn't get a fair average. But you might be able to keep track of a couple of attributes that can be summed up, and use that to calculate the average after the job is done (for example, total number of pages, and total number of books).

▲ dcosson 1339 days ago

Now if only everyone's mom understood the concept of a linear set operation... :)

A good example that would also help someone appreciate the power of map/reduce might be something like "how many times does each letter of the alphabet appear in the library?" For the map step I take shelf 1 and you take shelf 2, but now for each book we'll have to write down something like "a:300,b:50,c:90,..." and so on. But now there are more things to "reduce" than there are books (since we now have a value for the number of occurrences of each letter in the alphabet for every book). Well, there's nothing stopping us from splitting up the reduce step as well - we could split it up by shelf again (where you add the totals for each letter on your shelf and I'll do mine, then we'll do one final reduce of at the end), or you can add up letters a-m and I'll do n-z.

▲ th0ma5 1339 days ago

I remember a Stanford lecture video that said "well, if we called it Map, Sort, Group-reduce then no-one would use it :P" but that's a minor naming issue to describe just the subtly you are exactly describing, which is to just not be entirely random in the specifics of how the MapReduce is executed.

▲ lgarron 1339 days ago

Why wouldn't you be able to get the average? Return page count + total pages for each shelf and then take the ratio of "reduced" sums.

▲ michael_dorfman 1339 days ago

Re-read the comment. It discusses exactly what you are proposing.

TOPIC

MAP REDUCE

DATE 4-6-15

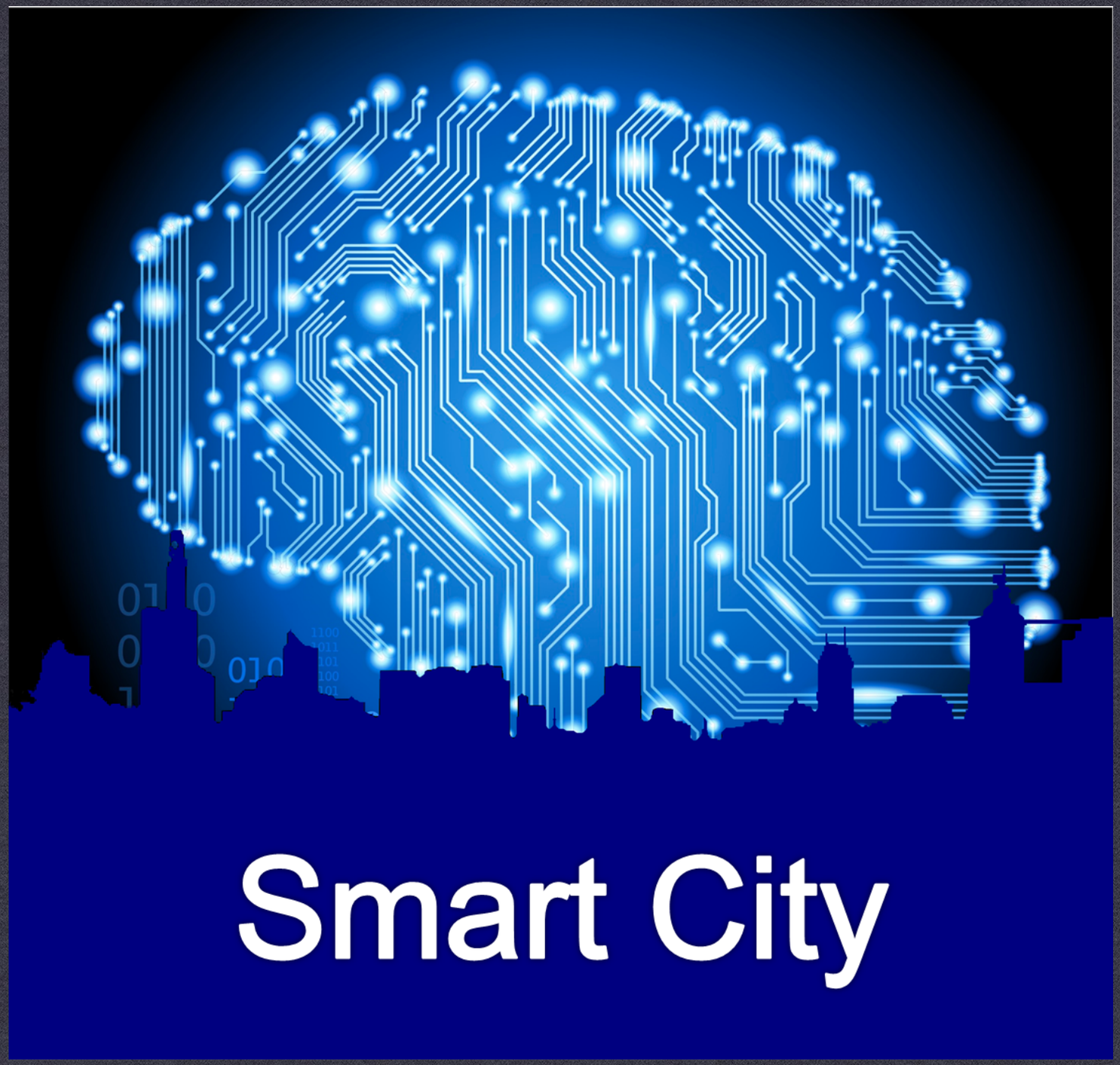
LECTURE #3 - DATA FOR THE SMART CITY

TOPIC

MAP REDUCE

DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY

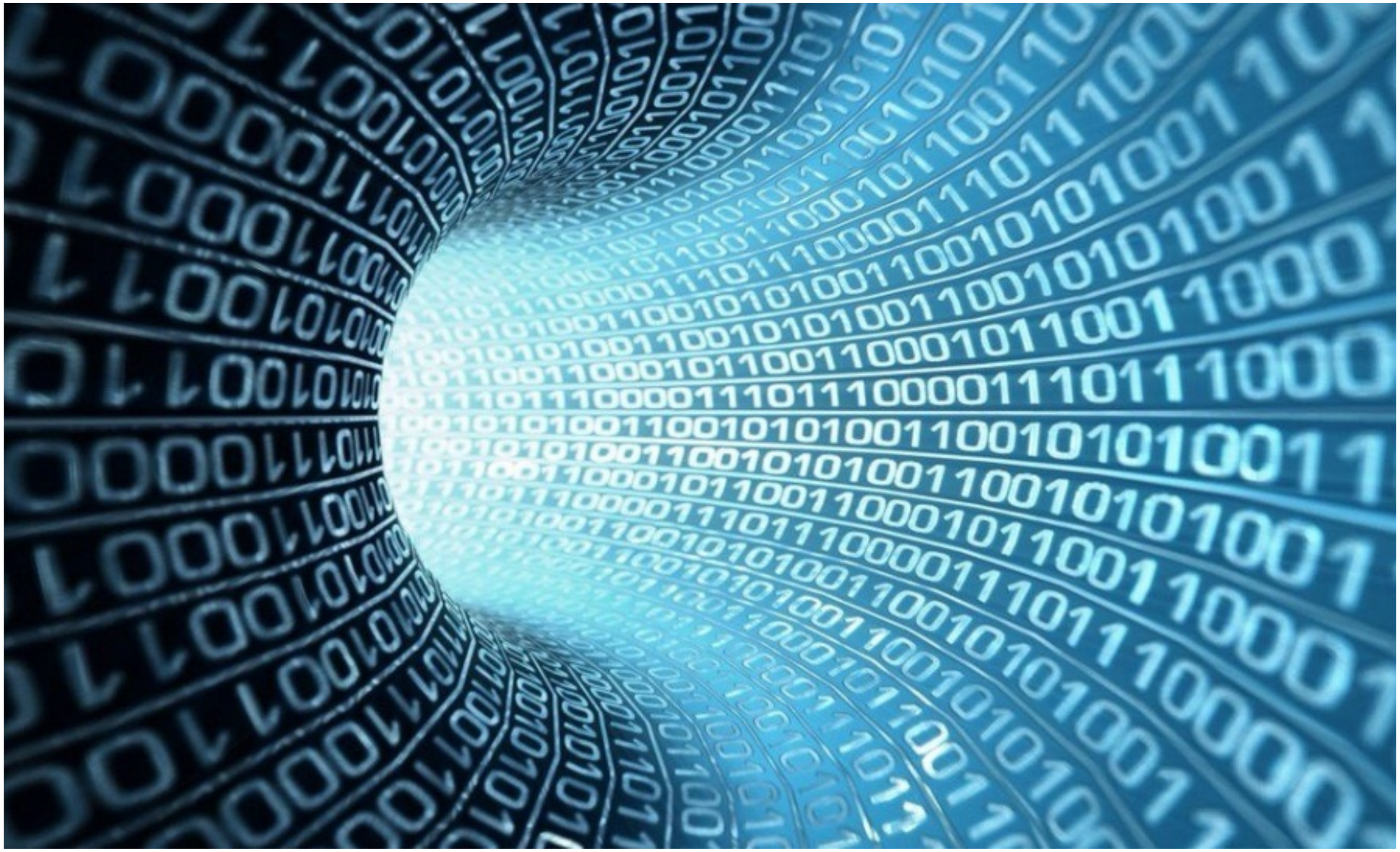


TOPIC

BIG DATA - IMAGES AND INFOGRAPHICS

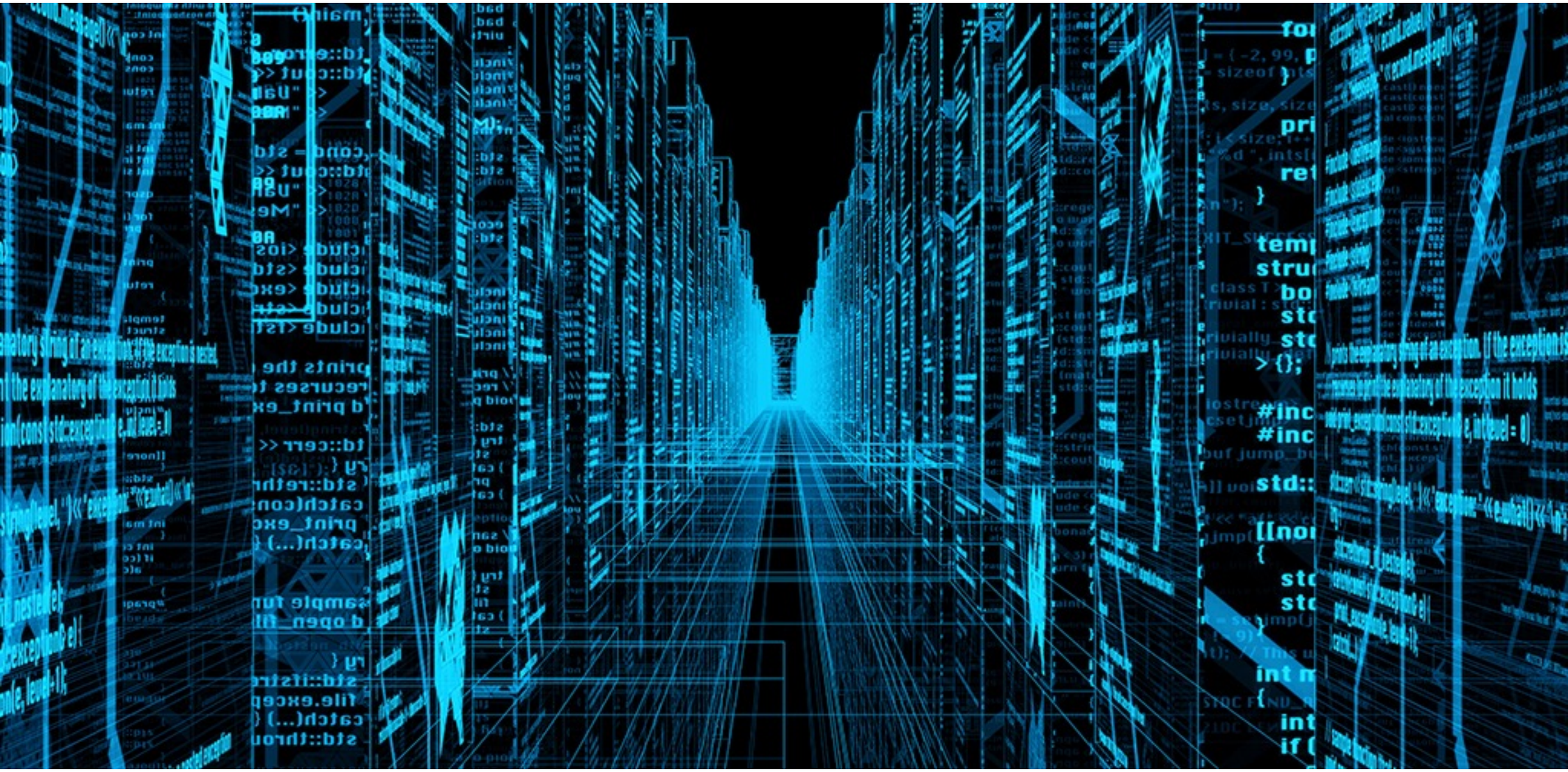
DATE 4-6-15

LECTURE #3 - DATA FOR THE SMART CITY



www.savortex.com


SMS
SMS
www.slideshare.net



www.wipro.com

BIG DATA

Big Data is data that is too large, complex and dynamic for any conventional data tools to capture, store, manage and analyze.

The right use of Big Data allows analysts to spot trends and gives niche insights that help create value and innovation much faster than conventional methods.

The "three V's", i.e the Volume, Variety and Velocity of the data coming in is what creates the challenge.

VOLUME

Amount of Big Data stored across the world (in petabytes)

VARIETY

PEOPLE TO PEOPLE
NETIZENS, VIRTUAL COMMUNITIES, SOCIAL NETWORKS, WEB LOGS...

PEOPLE TO MACHINE
ARCHIVES, MEDICAL DEVICES, DIGITAL TV, E-COMMERCE, SMART CARDS, BANK CARDS, COMPUTERS, MOBILES...

MACHINE TO MACHINE
SENSORS, GPS DEVICES, BAR CODE SCANNERS, SURVEILLANCE CAMERAS, SCIENTIFIC RESEARCH...

VELOCITY

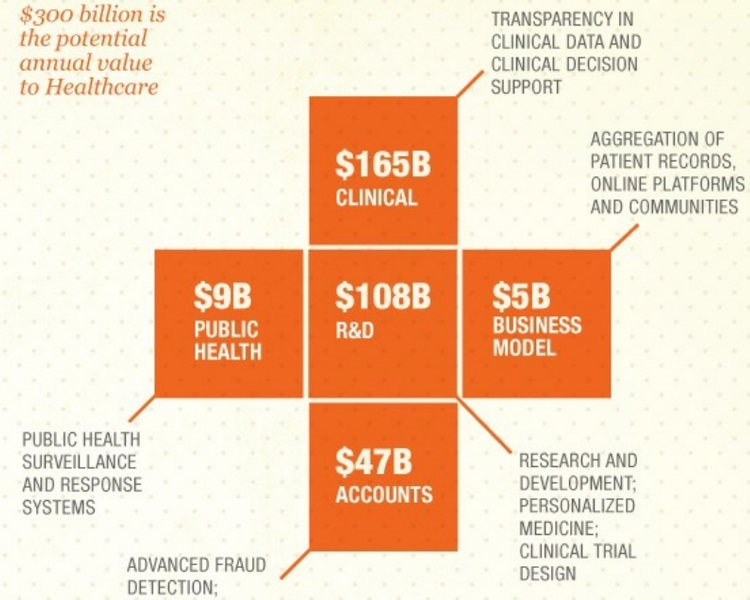
2.9 MILLION
EMAILS SENT EVERY SECOND

20 HOURS
OF VIDEO UPLOADED EVERY MIN

50 MILLION
TWEETS PER DAY

CASE STUDY - Healthcare

$300 billion is the potential annual value to Healthcare

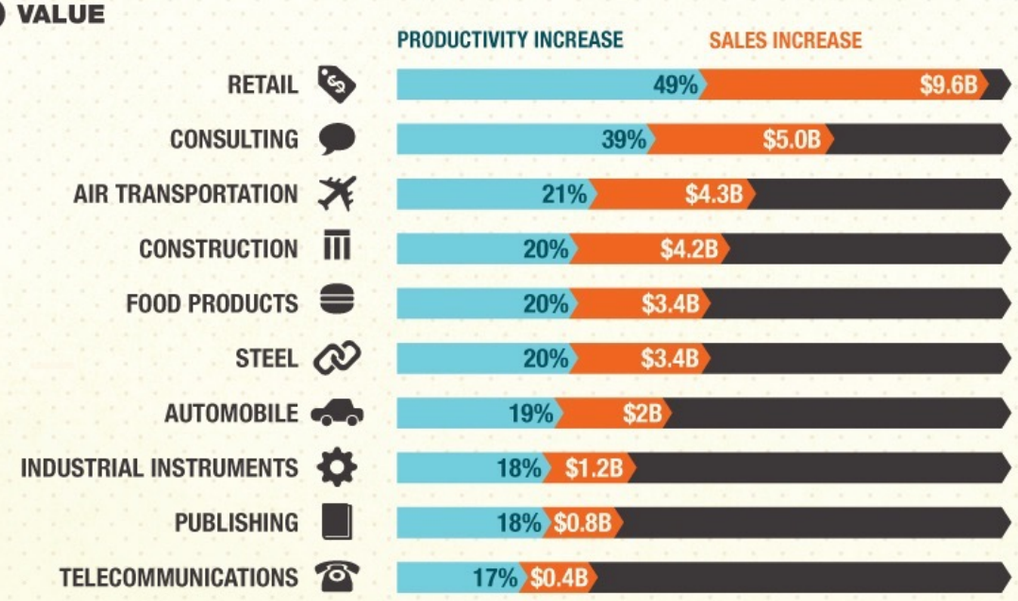


57.6% OF ORGANIZATIONS SURVEYED SAY THAT BIG DATA IS A CHALLENGE

72.7% CONSIDER DRIVING OPERATIONAL EFFICIENCIES TO BE THE BIGGEST BENEFIT OF A BIG DATA STRATEGY

50% SAY THAT BIG DATA HELPS IN BETTER MEETING CONSUMER DEMAND AND FACILITATING GROWTH

VALUE



40% PROJECTED GROWTH IN GLOBAL DATA CREATED PER YEAR

5% PROJECTED GROWTH IN GLOBAL IT SPENDING PER YEAR

The estimated size of the digital universe in 2011 was 1.8 zettabytes. It is predicted that between 2009 and 2020, this will grow 44 fold to 35 zettabytes per year. A well defined data management strategy is essential to successfully utilize Big Data.

Sources - 1 Reaping the Rewards of Big Data - Wipro Report 2 Big Data: The Next Frontier for Innovation, Competition and Productivity - McKinsey Global Institute Report 3 comScore, Radicati Group 4 Measuring the Business Impacts of Effective Data - study by University of Texas, Austin 5 US Department of Labour.

DO BUSINESS BETTER

NYSE:WIT | OVER 130,000 EMPLOYEES | 54 COUNTRIES | CONSULTING | SYSTEM INTEGRATION | OUTSOURCING

WIPRO Applying Thought

Visualization: Seventhinc

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

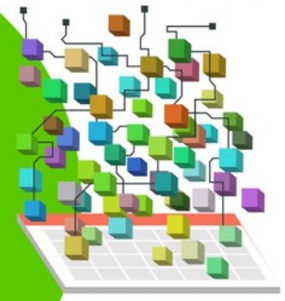
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day



6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

www.ibm.com

BIG DATA & ANALYTICS

Core to Nex-Gen Manufacturing

With more Asia Pacific manufacturers looking at making better and faster business decisions based on evidence, Big Data technologies can help them analyze larger volumes of data from a variety of sources and deliver the analysis at greater velocity. IDC observes that regional manufacturers are taking a cautious approach and the use of Analytics tools today is basic. Here is the 2013 landscape in Asia Pacific.

Asia Pacific manufacturers

3 key highlights

4 of the top 10 **automotive producing** countries are from Asia Pacific

+370%

China's share of exports has nearly **quadrupled** over the past 15 years, rising from 3% in 1995 to 11% in 2012

x2

The average **labor cost in China** has nearly doubled in the past five years

Top 3 regional business challenges

- Increasing competition
- Increasing internal costs
- Barriers to expand into new markets

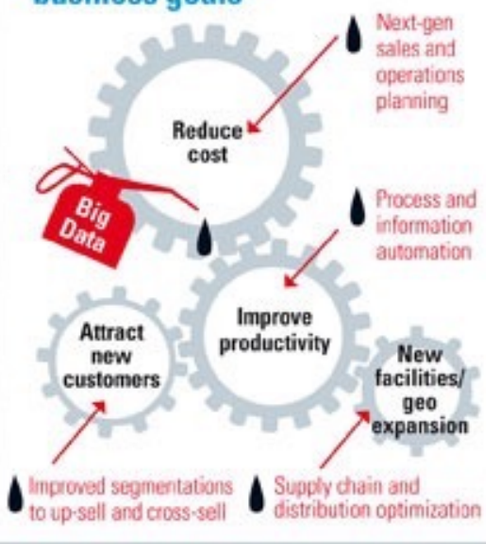
What Big Data means

Awareness

2/5 are still not aware of Big Data

Data that manufacturers are analyzing: Transactions (47%), text (45%) and machine (32%)

Big Data helps meet business goals



Common use cases

Today

37% The % of Asia Pacific manufacturers using Big Data and Analytics technologies to improve **production quality management**

29% The % of Asia Pacific manufacturers using Big Data and Analytics technologies for **inventory management**

Enabling technologies

- Production reporting tools (46%)
- Supply Chain analytics (35%)
- Hadoop (8%)

Industry leader use cases

Condition-based monitoring
Reduce downtime from unscheduled maintenance

Supply chain resilience
Decrease production delays due to global disruptions

Shop floor automation
Increase production velocity, productivity and consistency

Top 3 IT challenges

37% Not having the right people and skills

20% Cost of Technology infrastructure

17% Deciding what data is relevant

Moving forward

In the next 2 years

Asset Oriented Value Chain (AOVC) and **Engineering Oriented Value Chain (EOVC)** manufacturers will use Big Data and Analytics technologies to gain **Real time Asset Management**

Brand Oriented Value Chain (BOVC) manufacturers will use Big Data and Analytics technologies to improve **real time sales and operations planning**

Technology Oriented Value Chain (TOVC) manufacturers will use Big Data and Analytics technologies to improve **Supply Chain Management**

Enabling technologies

- M2M, Telematics
- Predictive analytics
- Data warehouse, Analytics appliances

Source
- IDC Big Data Pulse Survey
- IDC Manufacturing Insights Industry Survey
- IDC Big Data and Analytics CIS
- IDC Manufacturing Insights
- IMF
- China Ministry of Commerce

Research team
- Chris Holmes
- Daniel-Zoe Jimenez
- William Lee
- Craig Stires

Art direction
- ST Leng

IDC Analyze the Future

Smarter Cities: Turning Big Data Into Insight

City Planning and Operations

$1 Trillion
global annual savings could be attained by optimizing public infrastructure.
Source: McKinsey

$57 Trillion
in infrastructure investments will be needed between 2013-2030.
Source: McKinsey

Transportation Analytics

50 Hours
of traffic delays per year are incurred, on average, by travelers.

30 Billion
people all over the world travel approximately 30 billion miles per year. By 2050, that figure will grow to over 150 billion miles.

Water Management

60%
of water allocated for domestic human use goes to urban cities.

$14 Billion
in potable water is lost every year because of leaks, theft and unbilled usage.
Source: World Bank

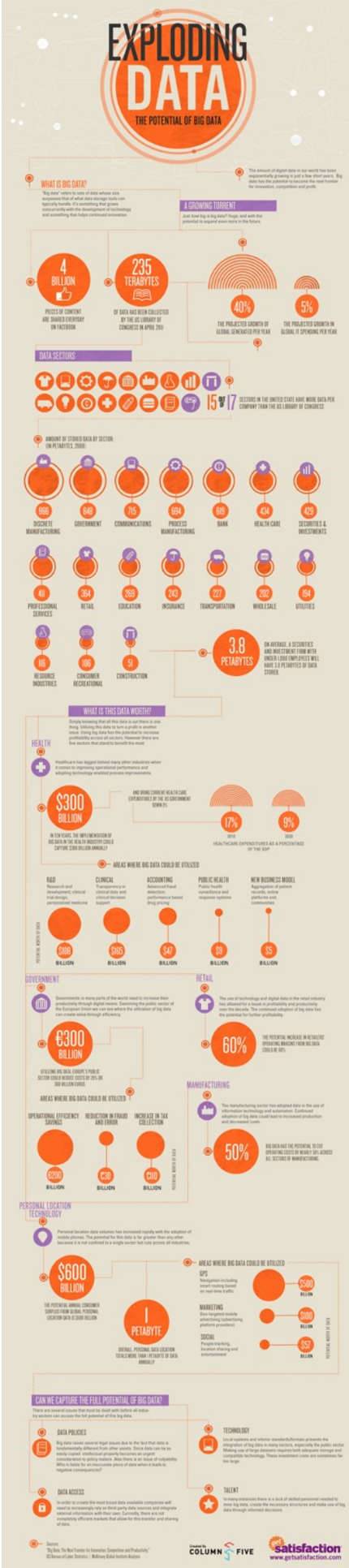
Open Cloud

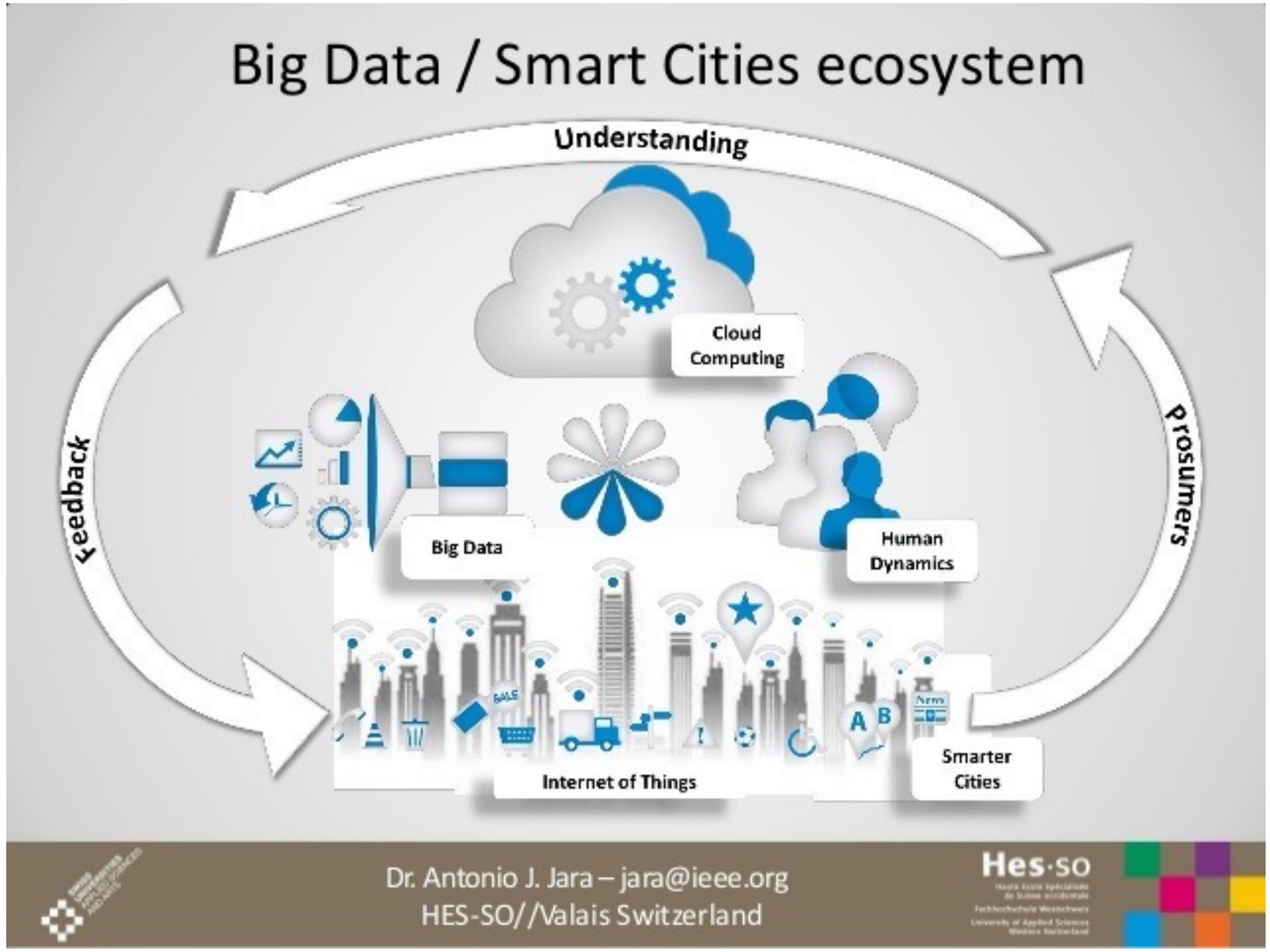
37,000
cloud experts support IBM's industry team alone.

$6 Billion
has been invested by IBM in more than a dozen acquisitions to accelerate its cloud initiatives.

IBM Intelligent Operations software is designed with cities, for cities, to provide the tools to monitor, visualize and analyze vital city services such as water and wastewater systems, transportation, infrastructure planning, permit management and emergency response.

IBM


EXPLODING
DATA
THE POTENTIAL OF BIG DATA
WHAT IS BIG DATA?
A GROWING TORRENT
4 BILLION
235 TERABYTES
40%
5%
DATA SECTORS
3.8 PETABYTES
WHAT IS THIS DATA WORTH?
$300 BILLION
17%
8%
AREAS WHERE BIG DATA COULD BE UTILIZED
R&D
CLINICAL
ACCOUNTING
PUBLIC HEALTH
NEW BUSINESS MODEL
$108 BILLION
$165 BILLION
$40 BILLION
$33 BILLION
$35 BILLION
GOVERNMENT
RETAIL
€300 BILLION
60%
MANUFACTURING
50%
PERSONAL LOCATION TECHNOLOGY
$600 BILLION
1 PETABYTE
CAN WE CAPTURE THE FULL POTENTIAL OF BIG DATA?
DATA POLICIES
TECHNOLOGY
DATA ACCESS
TALENT
COLUMN FIVE
satisfaction
www.getsatisfaction.com


Big Data / Smart Cities ecosystem
Understanding
Feedback
Prosumers
Cloud Computing
Big Data
Human Dynamics
Internet of Things
Smarter Cities
Dr. Antonio J. Jara – jara@ieee.org
HES-SO//Valais Switzerland
Hes·so

- **WIRELESS SENSOR NETWORKS – AN INTRODUCTION.GEOJOURNAL**
  INTECH
- **SMARTSANTANDER – THE CITY OF SANTANDER'S SENSOR NETWORK FACILITY**
  Smart Santander

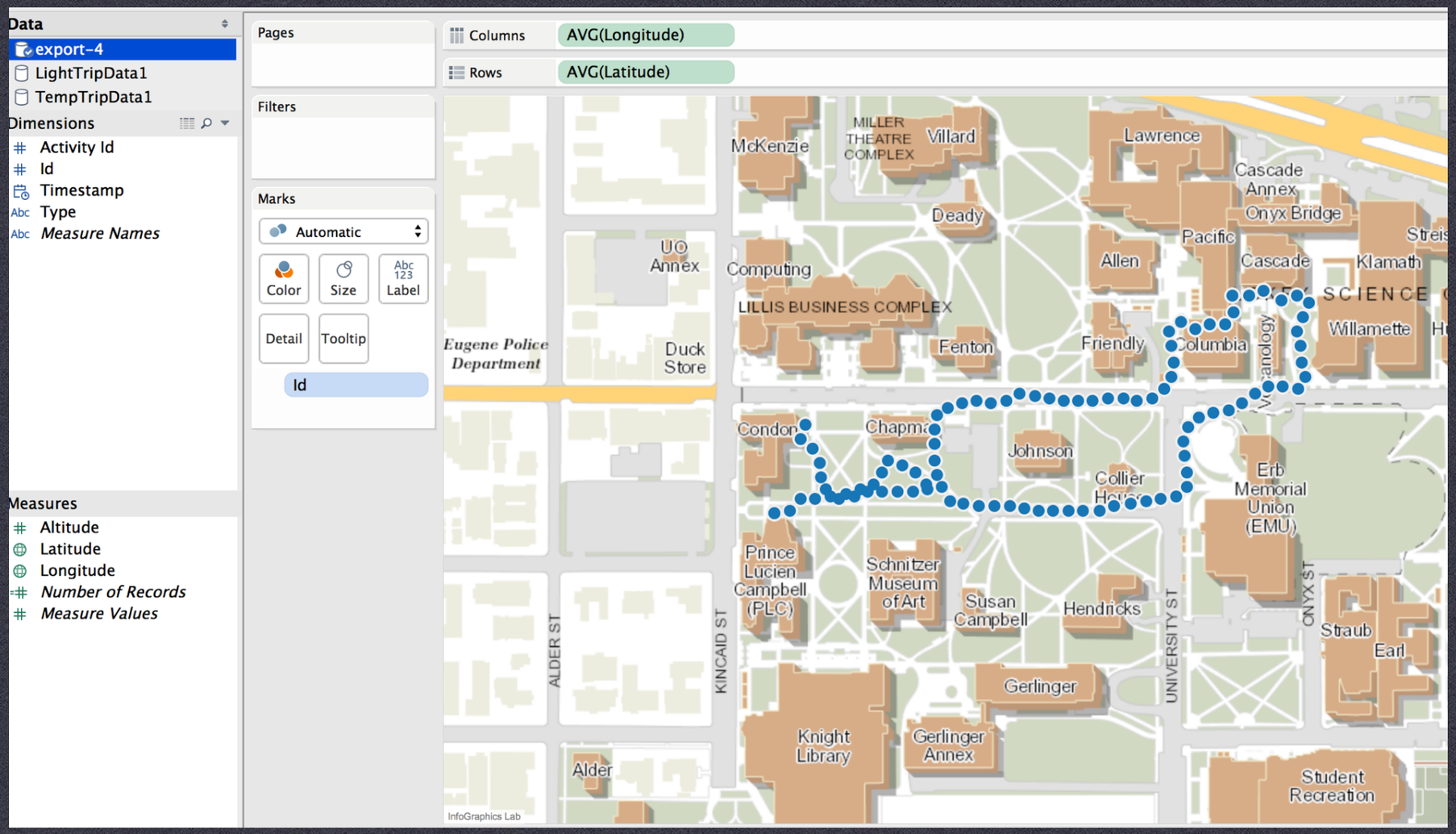
TOPIC

ASSIGNED READINGS FOR NEXT LECTURE

DATE	LECTURE
4-6-15	#3 - DATA FOR THE SMART CITY

MAPPING YOUR LOCATION SENSOR DATA

Run Keeper



TOPIC

THIS WEEK'S LAB

DATE **4-6-15**

LECTURE **#3 - DATA FOR THE SMART CITY**