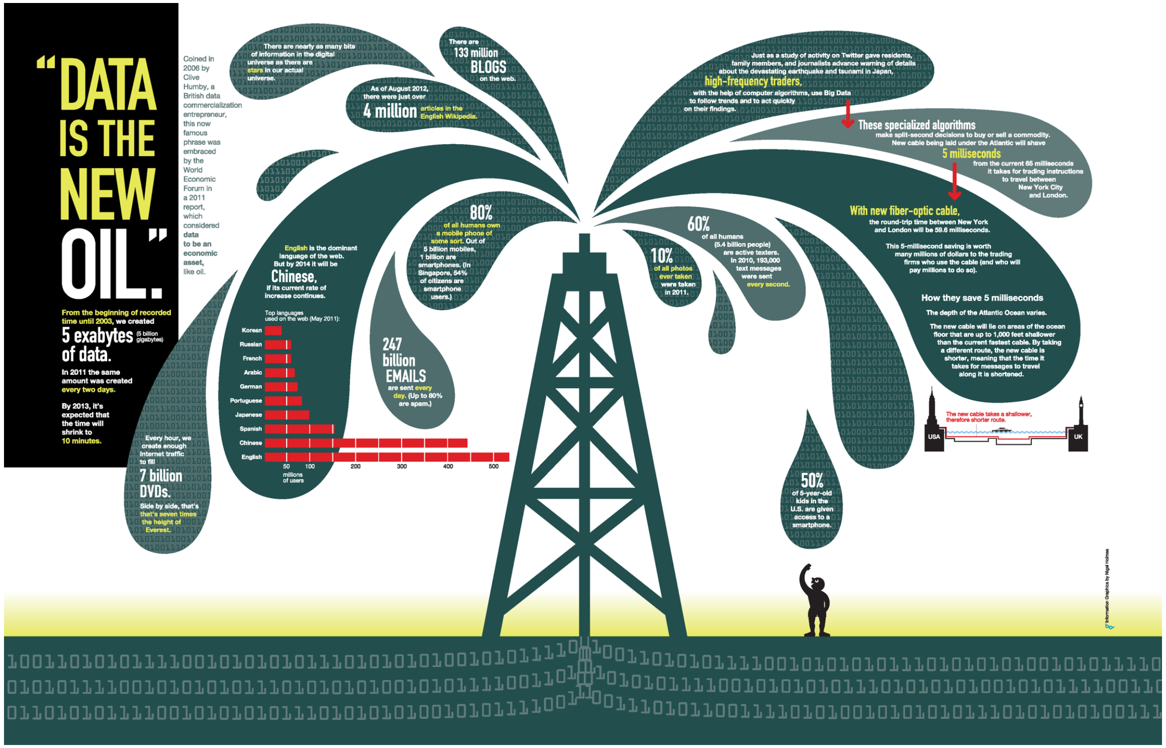
Introduction to Data Mining

Anis Yazidi, Professor

OsloMet

"Data is the New Oil"
– World Economic Forum 2011



"DATA IS THE NEW OIL."

From the beginning of recorded time until 2003, we created 5 exabytes (5 billion gigabytes) of data.

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

There are 133 million BLOGS on the web.

As of August 2012, there were just over 4 million articles in the English Wikipedia.

English is the dominant language of the web. But by 2014 it will be Chinese, if its current rate of increase continues.

80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)

247 billion EMAILS are sent every day. (Up to 80% are spam.)

Every hour, we create enough Internet traffic to fill 7 billion DVDs. Side by side, that's that's seven times the height of Everest.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, high-frequency traders, with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave 5 milliseconds from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 milliseconds

The depth of the Atlantic Ocean varies.

The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.

The new cable takes a shallower, therefore shorter route.

60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

10% of all photos ever taken were taken in 2011.

50% of 5-year-old kids in the U.S. are given access to a smartphone.

Top languages used on the web (May 2011):

"Retailers who can use the full power of Big Data can increase their income by 60% "- Forbes

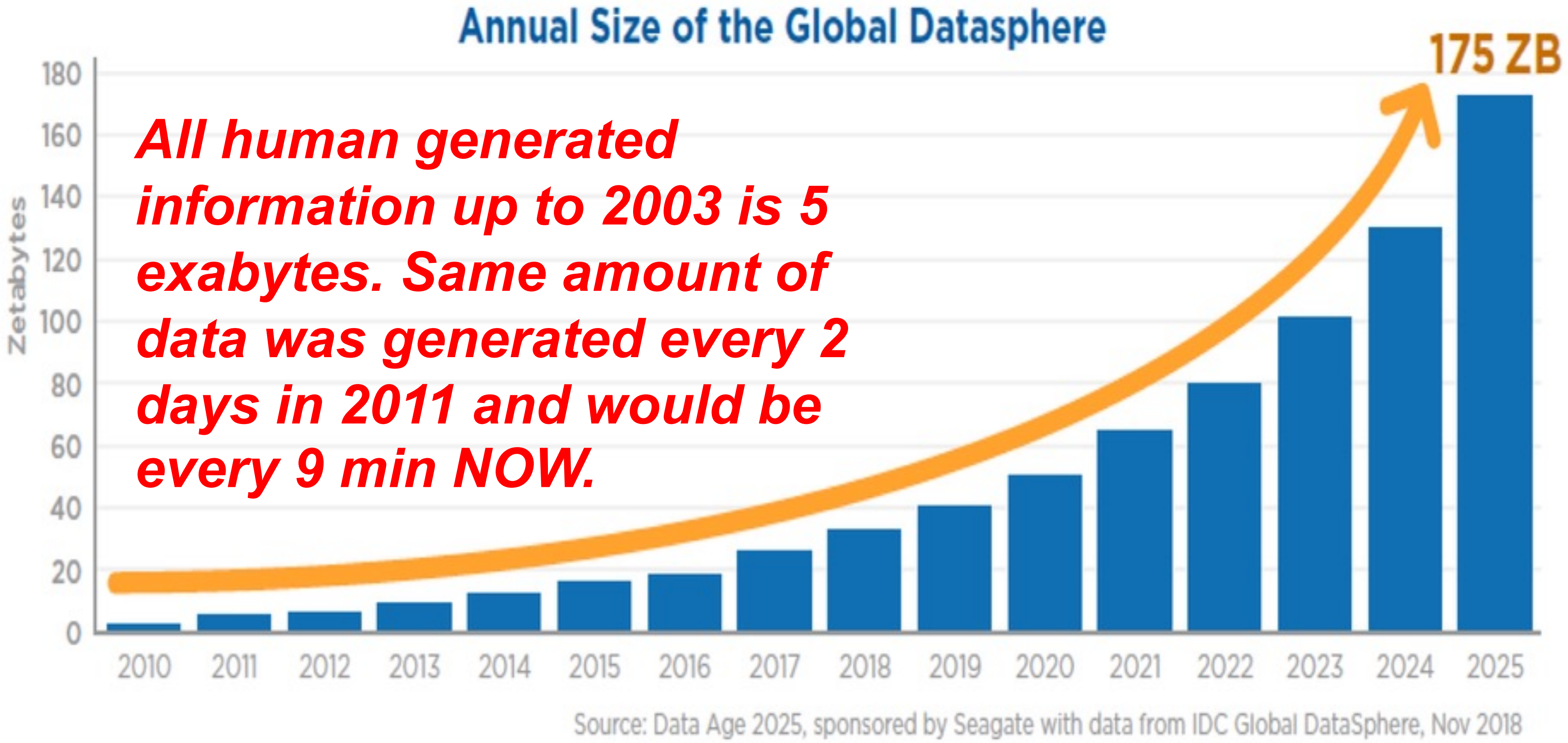
“Data is the New Oil”

Oil	Data
Crude oil: It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to.	Raw Data: must be broken down, analyzed for it to be changed to a value.

The Economist
Crunch time in France
Ten years on: banking after the crisis
South Korea’s unfinished revolution
Biology, but without the cells
MAY 6TH–12TH 2017
The world’s most valuable resource
Data and the new rules of competition
Google
amazon
Microsoft
UBER
TESLA
osoft

We live in the era of “Big Data”

Global data volume in Zettabytes



All human generated information up to 2003 is 5 exabytes. Same amount of data was generated every 2 days in 2011 and would be every 9 min NOW.

Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

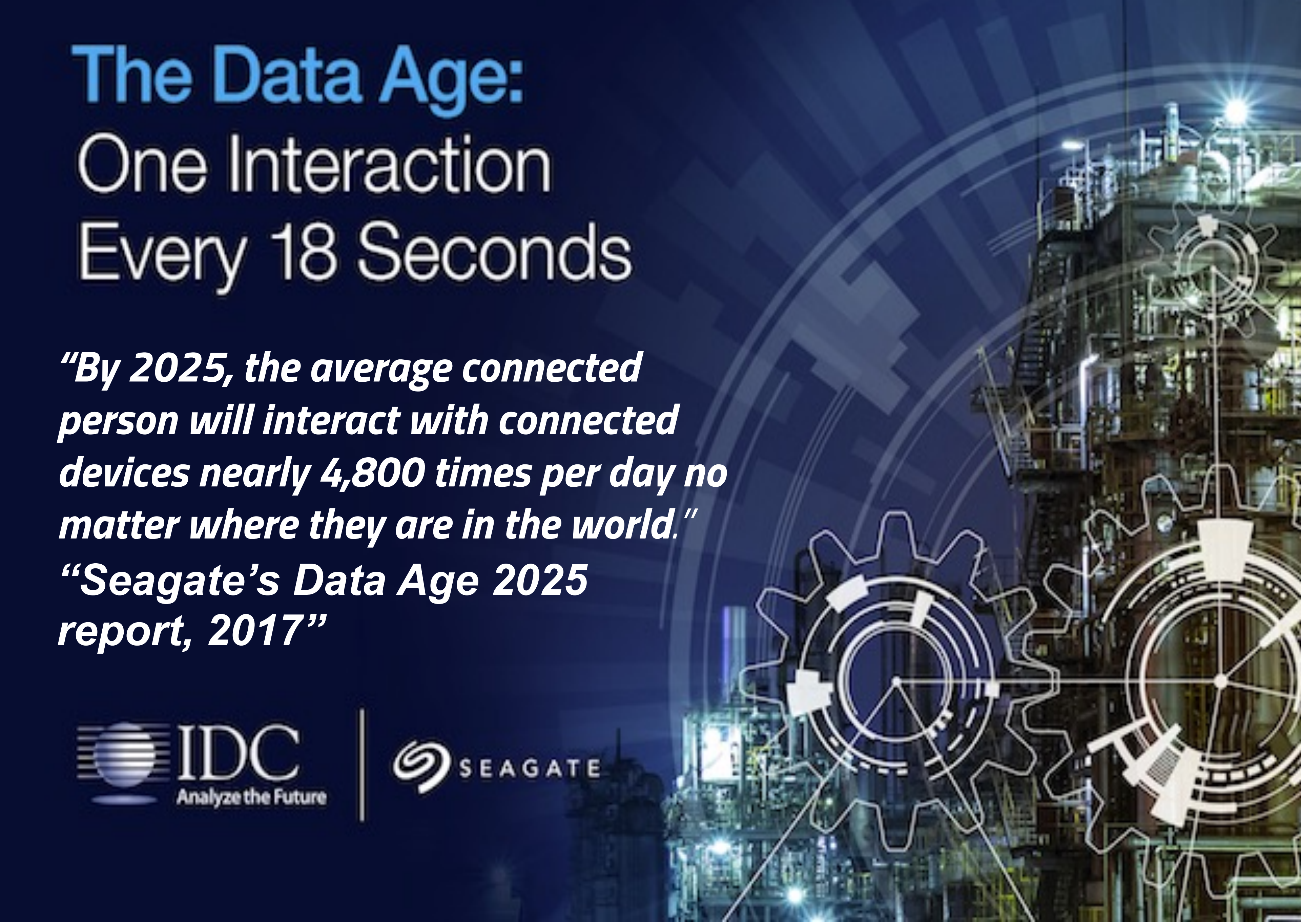
What is a Zettabyte?

- It is difficult for our minds to imagine such such a large number.
- One Zettabyte = trillion = million million gigabytes

Illustration of how large 175ZB is

stack of DVDs that could get us to the moon 23 times or circle Earth 222 times.

https://www.zmescience.com/space/the-moon-is-moving-away-from-earth-and-this-is-making-days-longer/

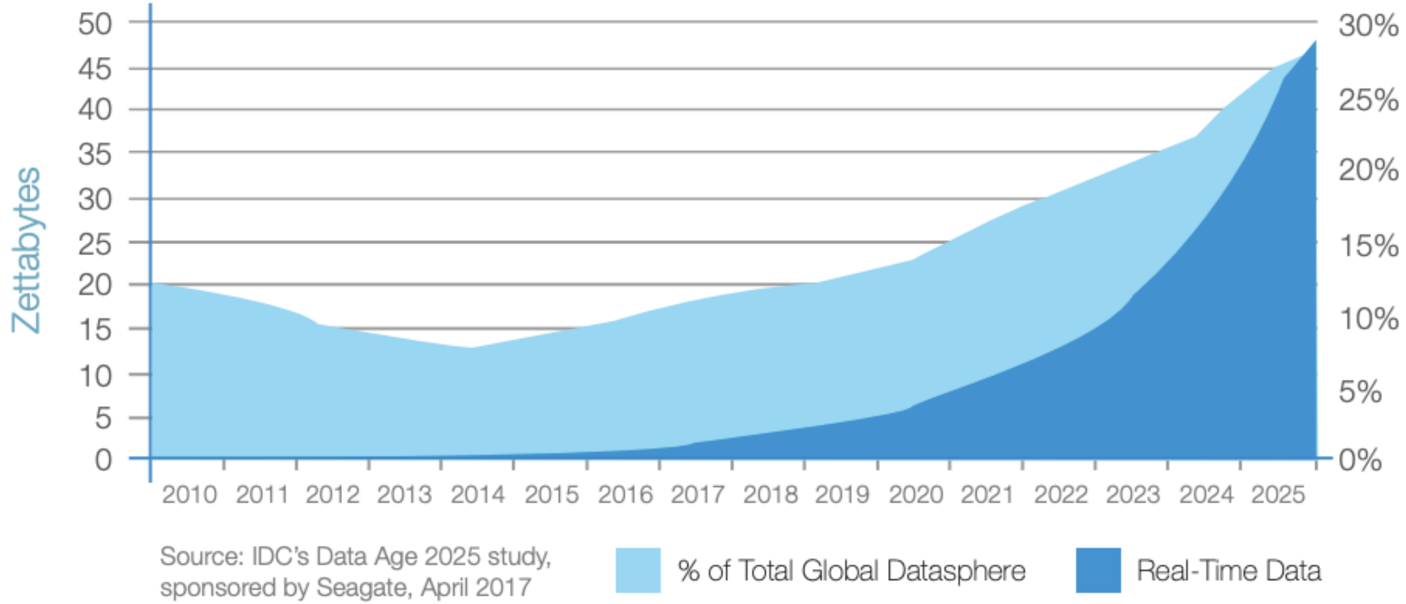


The Data Age:
One Interaction Every 18 Seconds

"By 2025, the average connected person will interact with connected devices nearly 4,800 times per day no matter where they are in the world."

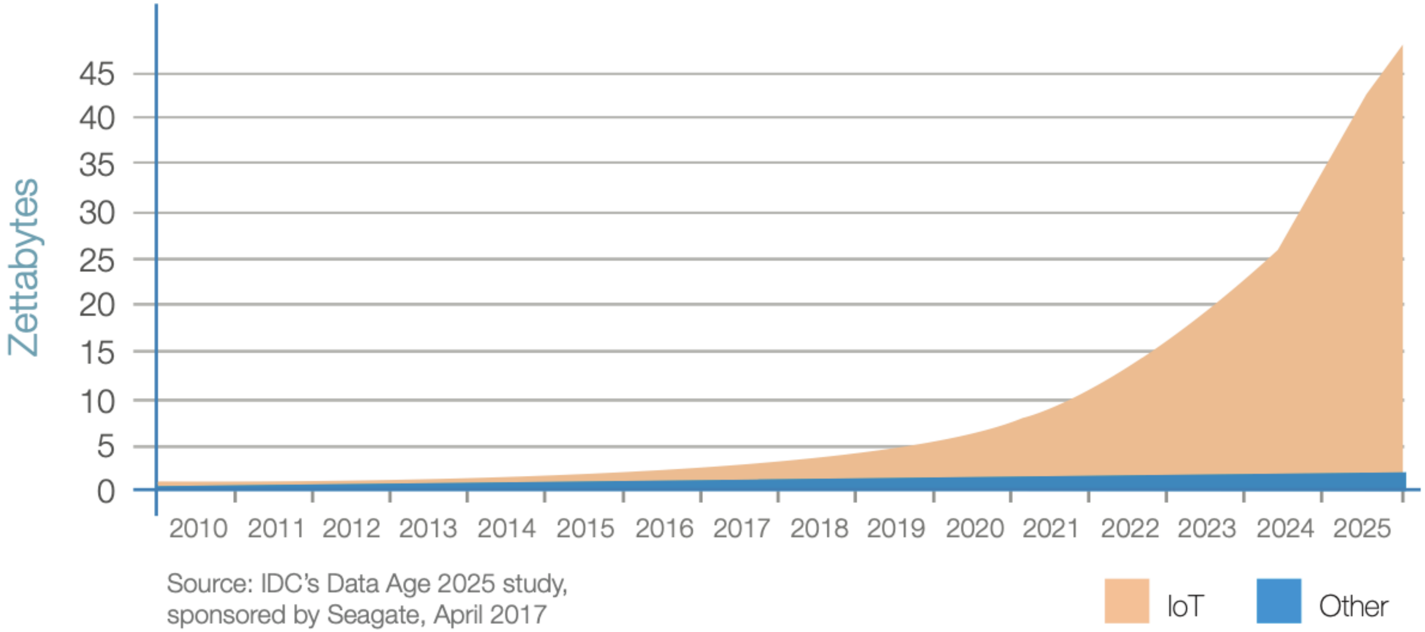
"Seagate's Data Age 2025 report, 2017"

Over 25 % will be real-time data



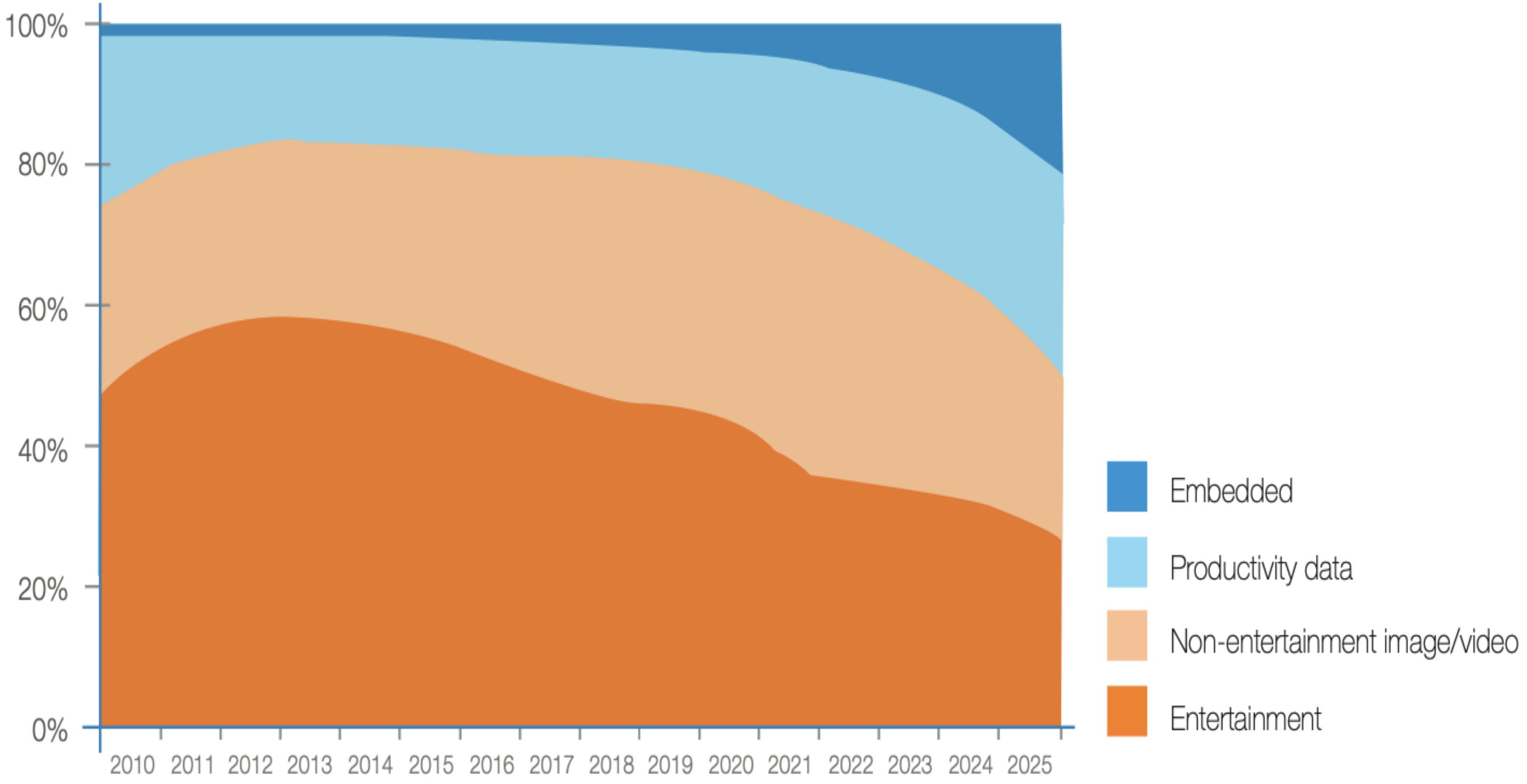
Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

95 % of real-time data will be IoT



Source: IDC's Data Age 2025 study,
sponsored by Seagate, April 2017

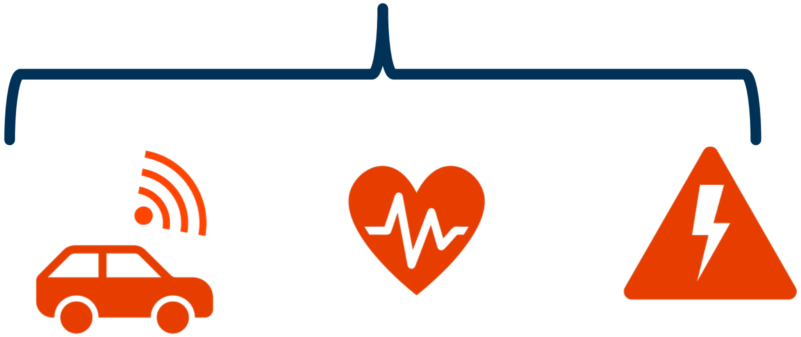
The mix of data creation by type is changing



Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

20% of the data will be critical and 10% of that will be hypercritical

Life Critical



Autonomous Cars

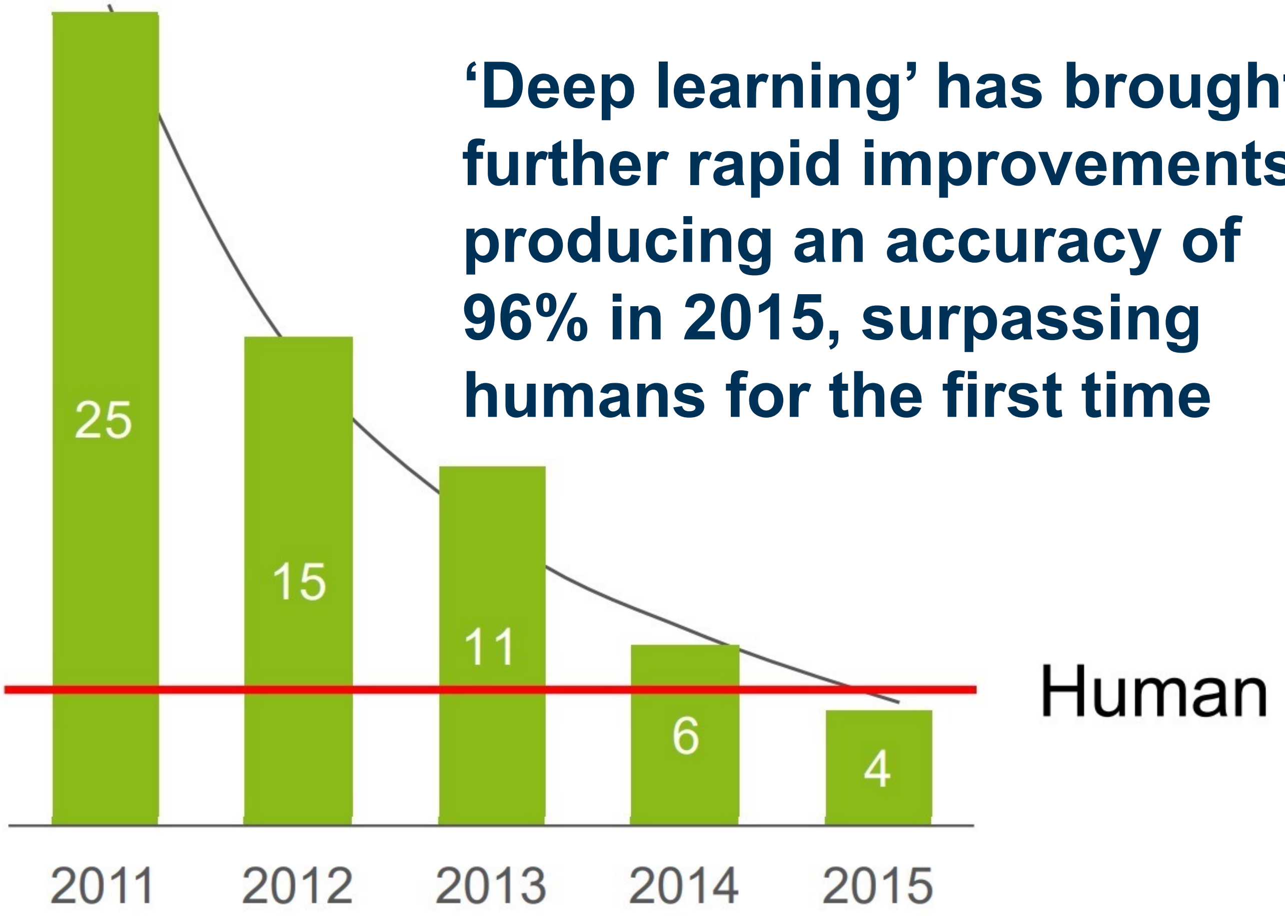
Remote monitoring patients devices

Power Grid operations

Critical: Data known to be necessary for the expected continuity of users' daily lives

Hypercritical: Data with direct and immediate impact on the health and wellbeing of users. (Examples include commercial air travel, medical applications, control systems. Heavy in metadata and data from embedded systems.)

First Inflection Point

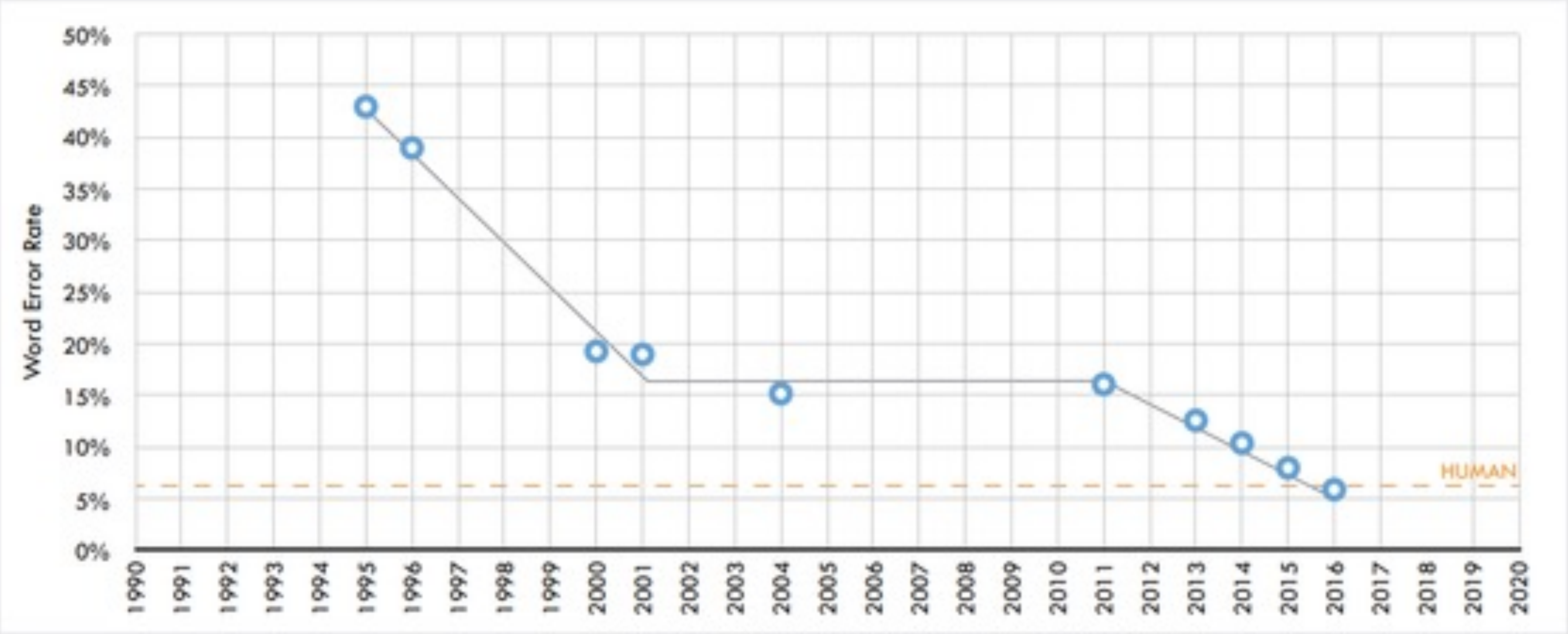


‘Deep learning’ has brought further rapid improvements, producing an accuracy of 96% in 2015, surpassing humans for the first time

Error rates on ImageNet visual challenge, %

Second Inflection Point

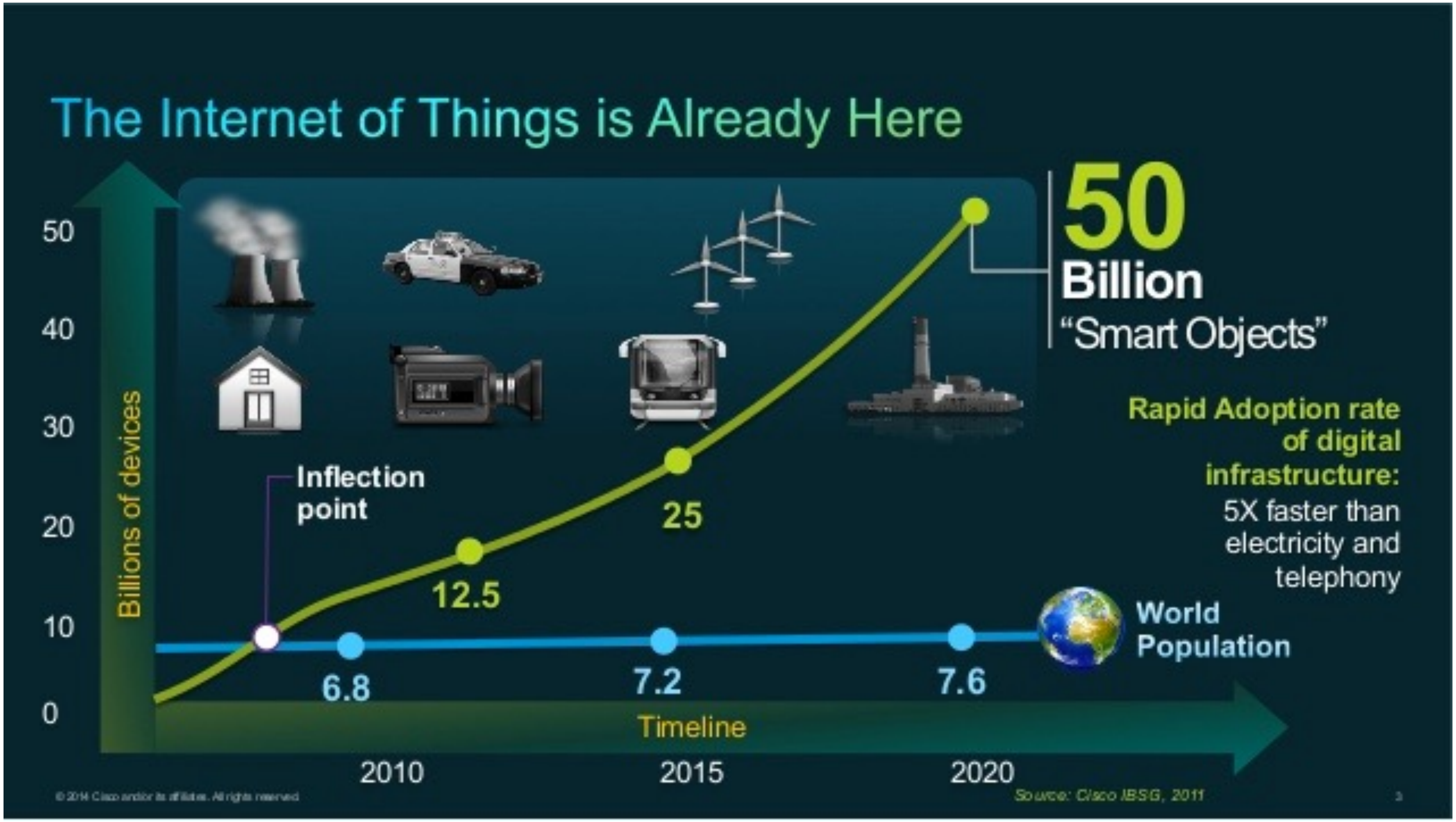
Speech to Text Transcription Error Rate (Switchboard)



Source: ARK Investment Management LLC

More IoT data: Third Inflection point

- Cisco predictions:



The Internet of Things is Already Here

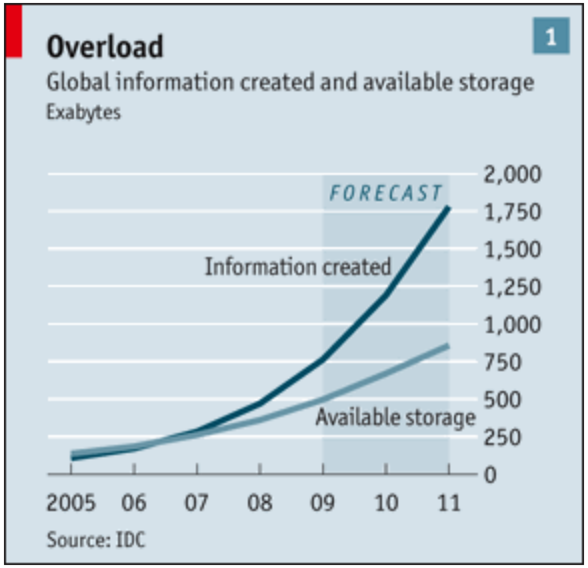
50 Billion "Smart Objects"

Rapid Adoption rate of digital infrastructure: 5X faster than electricity and telephony

Source: Cisco IBSG, 2011

Fourth Inflection point

- We live in an era of “Big Data”:
  - Technology are producing increasingly large data streams every second.
- We can no longer afford to store all the data produced by devices

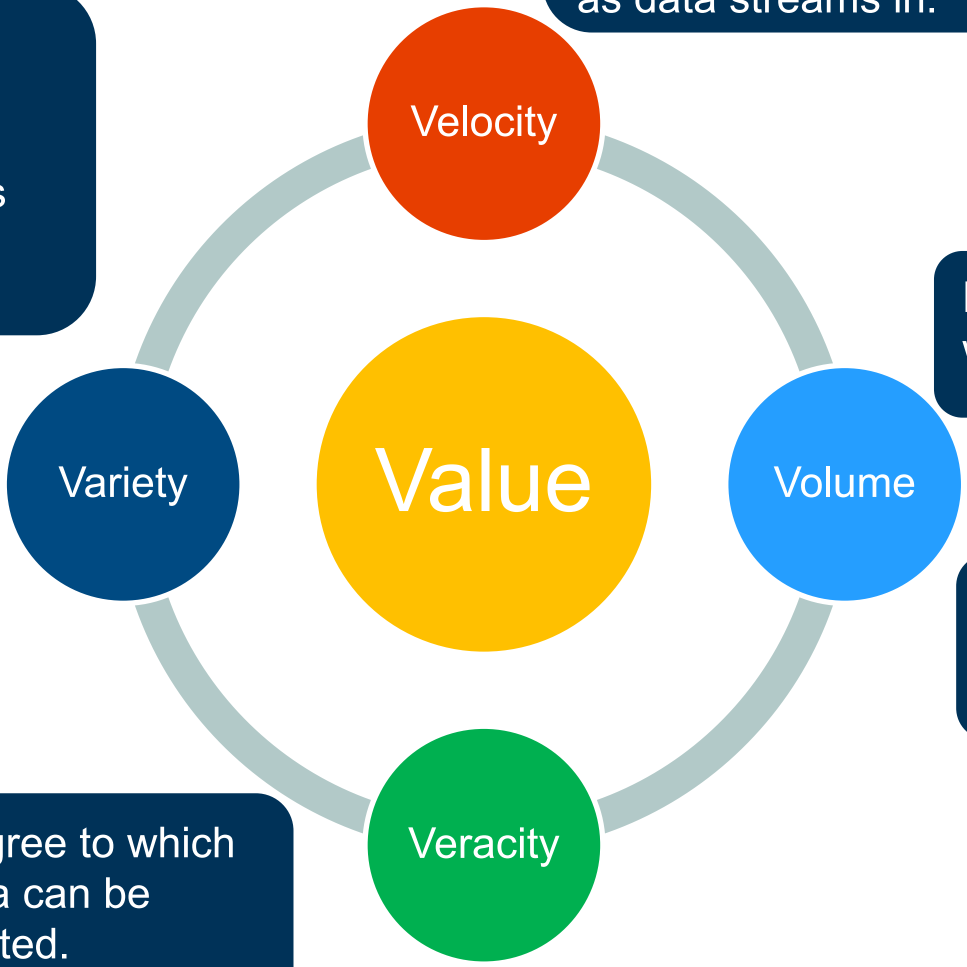


The economist

The 5 Vs

Speed at which the data is emanating. For time sensitive industries, meaningful insights must be extrapolated as data streams in.

Structured, unstrcuted. New and actionable insights can be found when various data types are analyzed together.



Dimensions over which data spans.

Amount of data from myriad sources.

Degree to which data can be trusted.

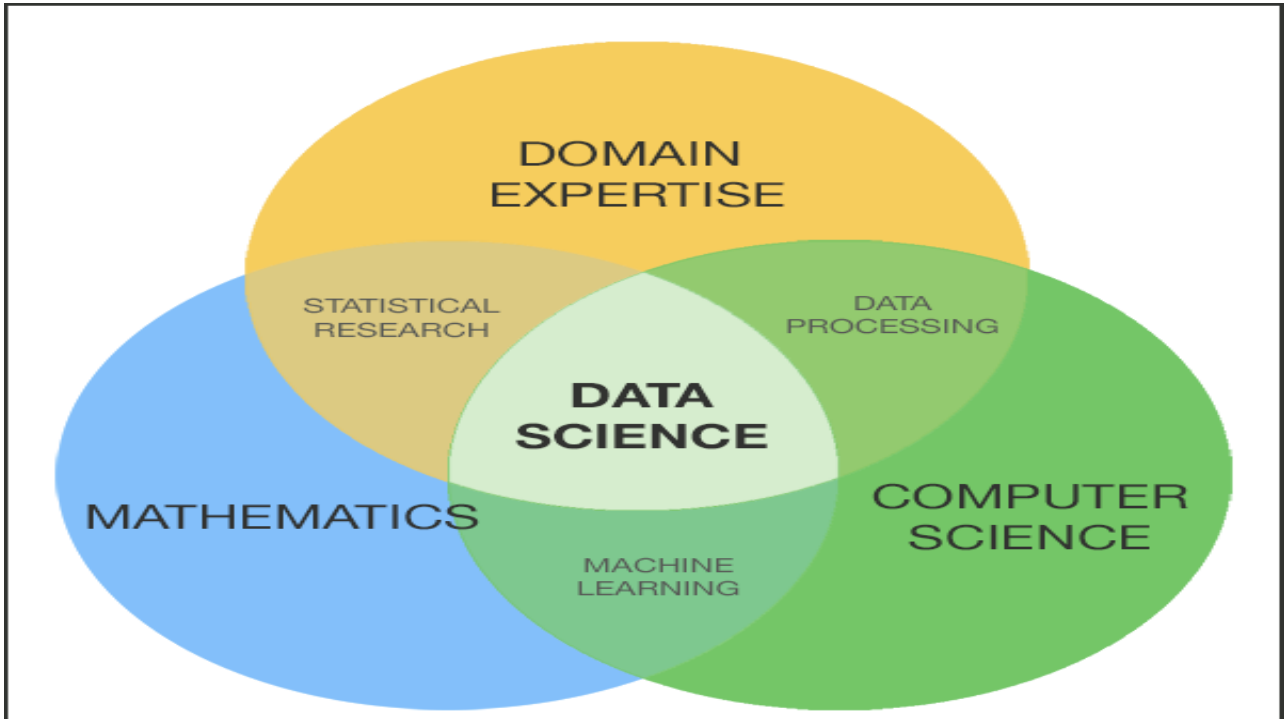
How to get the Big V: “Value”?

Data Science is the answer

- ***“Data science is the discipline that deals with collecting, preparing, managing, analysing, interpreting and visualising large and complex datasets.***”

  Data Science Institute at Imperial College London.
  Guo Yike, and David Johnson.

The data science Venn diagram



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Domain Expertise

- Data scientist need to have knowledge be related to the domain that the project/analysis is in.
  - Example if the data scientist is working in a risk management department, he will need to understand
    - the specific business definitions
    - regulations
    - accounting policies & international standards process etc.
- Domain experts can ask:
  - more specific questions
  - that will yield measurable, actionable improvements

Data Scientist

- A data scientist should be a good programmer!
- A data scientist should have solid quantitative skills!
- A data scientist should excel in communication and visualization skills!
- A data scientist should have a solid business understanding!
- A data scientist should be creative!

Techniques

- Term often used interchangeably with data mining, knowledge discovery, predictive/descriptive modeling, …
- Essentially refers to extracting useful business patterns and/or mathematical decision models from a preprocessed data set
- **Predictive analytics**
  - Predict the future based on patterns learnt from past data
  - Classification versus regression
- **Descriptive analytics**
  - Describe patterns in data
  - Clustering, Association rules, Sequence rules

Enablers

- Rapid advances in Machine learning algorithms, and especially within the subfield of deep learning.
- Intense development of **machine learning software and libraries:**
  - This is improving the quality of the algorithms and making the tools easier to use, lowering the barriers to entry for aspiring data scientists.

Online Website: key drivers

- Online Websites are largest drivers of data science tools
  - open sourcing their own technologies that they use in their workflow to the public.

- The most well-known open source projects in deep learning are:
  - PyTorch (Facebook)
  - Tensorflow (Google)
  - both coincidentally being the de-facto standard for Deep Learning.

Application areas

E-commerce

- Product recommendation
- Sentiment analysis
- Analyzing reviews

Transport

- Self-driving cars
- Enhanced driving experience
- Enhanced safety for passengers

Manufacturing

- Monitoring systems
- Predictive maintenance
- Anomaly detection

Finance

- Customer Segmentation
- Strategic decision making
- Algorithmic trading

Healthcare

- Medical Image analysis
- Drug Discovery
- Genomics
- Virtual assistants and chatbots

Banking

- Loan risk modelling
- Fraud detection
- Risk analysis

Apriori algorithm for association rule mining

- Proposed by Agrawal et al in 1993.
- It is an important data mining model studied extensively by the database and data mining community.
- Assume all data are categorical.
- Considered as the most influential data mining paper

The model: data

- $I = \{i_1, i_2, \ldots, i_m\}$: a set of *items*.
- Transaction *t* :
  - t a set of items, and $t \subseteq I$.
- Transaction Database *T*: a set of transactions $T = \{t_1, t_2, \ldots, t_n\}$.

Example transactions database

Transaction	Items
1	stella, hoegaarden, diapers, baby food
2	coke, stella, diapers
3	cigarettes, diapers, baby food
4	chocolates, diapers, hoegaarden, apples
5	tomatoes, water, leffe, stella
6	spaghetti, diapers, baby food, stella
7	water, stella, baby food
8	diapers, baby food, spaghetti
9	baby food, stella, diapers, hoegaarden
10	apples, chimay, baby food

Association rules

- Purpose
  - Detect frequently occurring patterns between items

- Example Applications
  - Which products\services are frequently bought together?
  - Which web pages are frequently visited together?
  - Which terms often co-occur in a text document?

Market Basket Analysis

baby food, diapers ⇨ stella

1. Put them closer together in the store.
2. Put them far apart in the store.
3. Package baby food, diapers and stella.
4. Package baby food, diapers and stella + poorly selling item.
5. Raise the price on one, and lower it on the other.
6. Do not advertise baby food, diapers and stella together

The model: rules

- A transaction *t* contains *X*, a set of items (itemset) in *I*, if $X \subseteq t$.
- An association rule is an implication of the form:

  $X \rightarrow Y$, where $X, Y \subset I$, *and* $X \cap Y = \varnothing$

  baby food, diapers ⇨ stella

- An itemset is a set of items.
  - E.g., X = {milk, bread, cereal} is an itemset.
- A *k*-itemset is an itemset with *k* items.
  - E.g., {milk, bread, cereal} is a 3-itemset

Rule strength measures

- Support: The rule holds with support *sup* in *T* (the transaction data set) if sup% of transactions contain $X \cup Y$.
  - $sup = \Pr(X \cup Y)$.
- Confidence: The rule holds in *T* with confidence *conf* if *conf*% of tranactions that contain *X* also contain *Y*.
  - $conf = \Pr(Y \mid X)$
- An association rule is a pattern that states when *X* occurs, *Y* occurs with certain probability.

Support and Confidence

- Support count: The support count of an itemset *X*, denoted by *X.count*, in a data set *T* is the number of transactions in *T* that contain *X*. Assume *T* has *n* transactions.
- Then,

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

More on association rule mining

- Clearly the space of all association rules is exponential, O(2^m), where m is the number of items in *I*.
- The mining exploits sparseness of data, and high minimum support and high minimum confidence values.
- Still, it always produces a huge number of rules, thousands, tens of thousands, millions, ...

Goal and key features

- **Goal:** Find all rules that satisfy the user-specified *minimum support* (minsup) and *minimum confidence* (minconf).
- **Key Features**
  - Completeness: find all rules.
  - No target item(s) on the right-hand-side

Associations: Support and Confidence

Transaction	Items
1	stella, hoegaarden, diapers, baby food
2	coke, stella, diapers
3	cigarettes, **diapers, baby food**
4	chocolates, diapers, hoegaarden, apples
5	tomatoes, water, leffe, stella
6	spaghetti, diapers, baby food, stella
7	water, stella, baby food
8	**diapers, baby food,** spaghetti
9	baby food, stella, diapers, hoegaarden
10	apples, chimay, baby food

E.g. itemset {baby food, diapers, stella } has support = 3/10 or 30%

Association Rule: baby food, diapers ⇨ stella has confidence of 3/5 or 60%

Frequent Itemset Property

- ***Frequent Itemset Property:***

Any subset of a frequent itemset is frequent.

- Contrapositive:

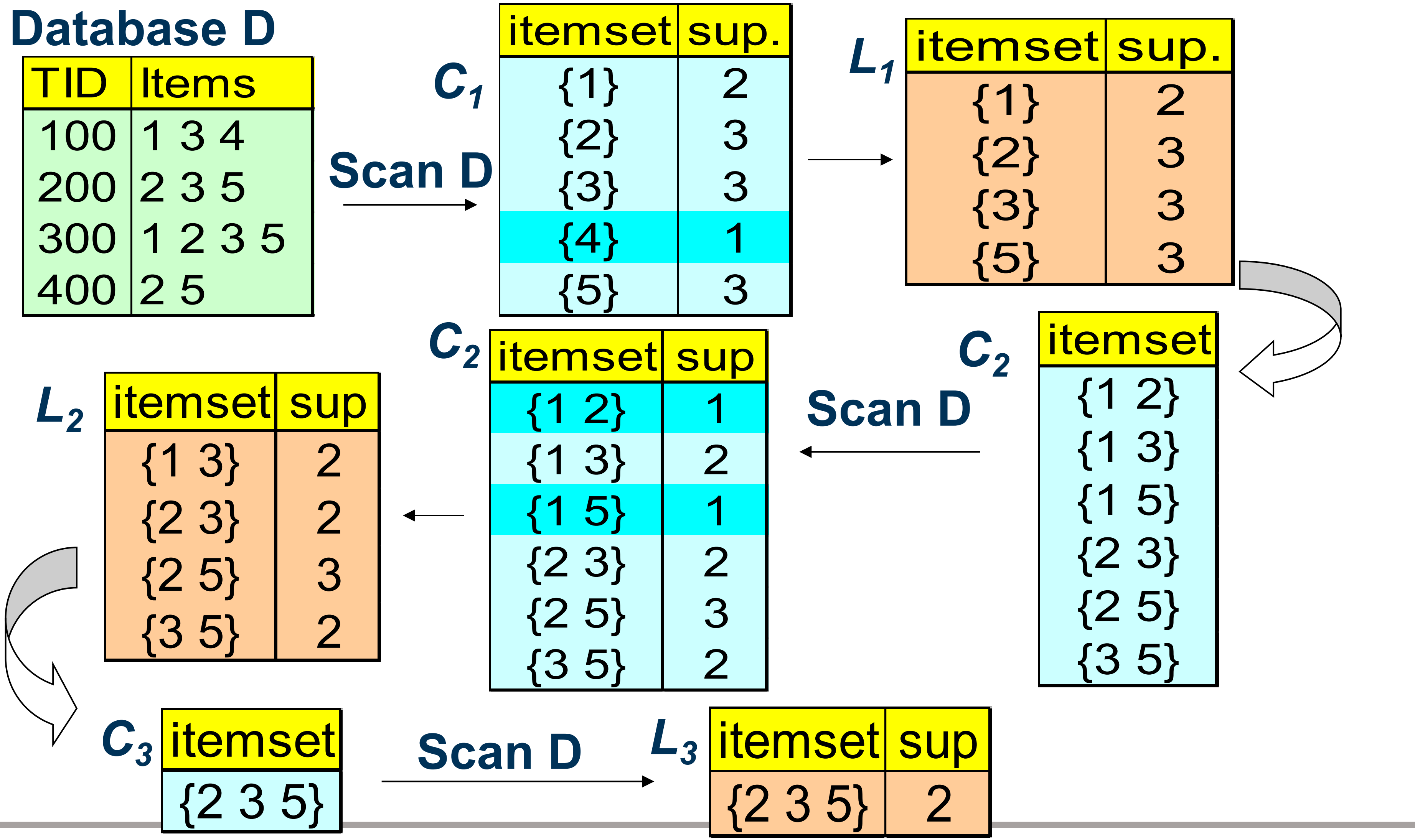
If an itemset is not frequent,

none of its supersets are frequent.

Mining Frequent Itemsets

- Find the *frequent itemsets*: the sets of items that have minimum support
  - A subset of a frequent itemset must also be a frequent itemset
    - i.e., if {*AB*} is a frequent itemset, both {*A*} and {*B*} should be a frequent itemset
  - Iteratively find frequent itemsets with cardinality from 1 to *k (k*-itemset*)*
- Use the frequent itemsets to generate association rules.

The Apriori Algorithm — Example



Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

Scan D

L_3

itemset	sup
{2 3 5}	2

Terminology

frequent itemset: doesn’t mean an itemset with many items. It means one whose support is at least minimum support.

L_k : the set of all large *k*-itemsets in the DB.

C_k : a set of *candidate* large *k*-itemsets. In the algorithm we will look at, it generates this set, which contains all the *k*-itemsets that might be large, and then eventually generates the set above.

The Apriori Algorithm: Basic idea

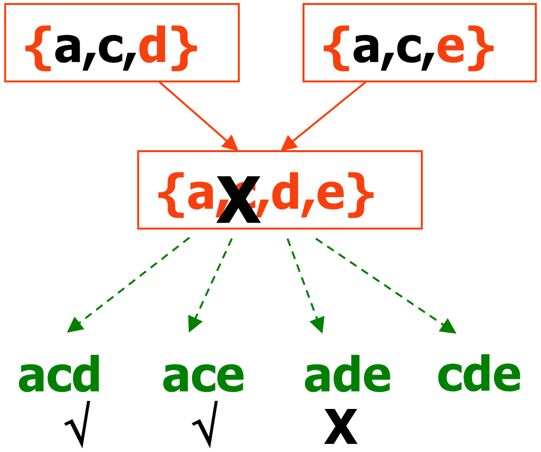
- Join Step: C_k is generated by joining $L_{k-1}$with itself
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
- Pseudo-code:

```
$C_k$: Candidate itemset of size k
$L_k$ : frequent itemset of size k

$L_1$ = {frequent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
    $C_{k+1}$ = candidates generated from $L_k$;
    for each transaction $t$ in database do
            increment the count of all candidates in $C_{k+1}$
            that are contained in $t$
    $L_{k+1}$  = candidates in $C_{k+1}$ with min_support
    end
return $\cup_k L_k$;
```

Example of Candidates Generation

- L_3={*abc, abd, acd, ace, bcd*}
- ***to generate*** C_4 , ***Self-joining*** : L_3*L_3
  - ***abcd*** from ***abc*** and ***abd***
  - ***acde*** from ***acd*** and ***ace***
- ***Pruning:***
  - ***acde*** is removed because ***ade*** is not in $\boldsymbol{L_3}$
- C_4={*abcd*}



Generating candidate itemsets …

Suppose these are the only 3-itemsets all have >10% support:

{a, b, c}

{a, e, g}

{e, f, h}

{e, f, k}

{p, q, r}

.. How do we generate candidate 4-itemsets that *might* have 10% support?

Generating candidate itemsets …

Suppose these are the only 3-itemsets all have >10% support:

{a, b, c}

{a, e, g}

{e, f, h}

{e, f, k}

{p, q, r}

One possibility:

1. **note all the items involved:**

   {a, b, c, e, f, g, h, k, p, q, r}

2. **generate all subsets of 4 of these:**
   {a,b,c,e}, {a,,b,c,f}, {a,b,c,g}, {a,b,c,h}, {a,b,c,k}, {a,b,c, p},… etc … there are 330 possible subsets in this case !

Generating candidate itemsets …

Suppose these are the only 3-itemsets all have >10% support:

{a, b, c}

{a, e, g}

{e, f, h}

{e, f, k}

{p, q, r}

One possibility:

1. **note all the items involved:**

   {a, b, c, e, f, g, h, k, p, q, r}

2. **generate all subsets of 4 of these:**

{a,b,c,e}, {a,b,c,f}, {a,b,c,g}, {a,b,c,h}, {a,b,c,k}, {a,b,c, p},… etc … there are 330 possible subsets in this case !

But, hold on: we can easily see that {a,b,c,e} couldn't have 10% support – because {a,b,e} is *not* one of our 3-itsemsets

Generating candidate itemsets …

Suppose these are the only 3-itemsets all have >10% support:

{a, b, c}

{a, e, g}

{e, f, h}

{e, f, k}

{p, q, r}

One possibility:

1. **note all the items involved:**

   {a, b, c, e, f, g, h, k, p, q, r}

2. **generate all subsets of 4 of these:**

{a,b,c,e}, {a,b,c,f}, {a,b,c,g}, {a,b,c,h}, {a,b,c,k}, {a,b,c, p},… etc … there are 330 possible subsets in this case !

But, hold on: the same goes for several other of these subsets …

A neat Apriori trick

{a, b, c}

{a, e, g}

{e, f, h}

{e, f, k}

{p, q, r}

i. enforce that subsets are always arranged ‘lexicographically’ (or similar), as they are already on the left

ii. Only generate *k*+1-itemset candidates from *k*-itemsets that differ in the last item.

So, in this case, the only candidate 4-itemset would be:

{e, f, h, k}

A neat Apriori trick

{a, b, c, e}

{a, e, g, r}

{a, e, g, w}

{e, f, k, p}

{n, q, r, t }

{n, q, r, v }

{n, q, s, v }

i. enforce that subsets are always arranged 'lexicographically' (or similar), as they are already on the left

ii. Only generate *k*+1-itemset candidates from *k*-itemsets that differ in the last item.

And in this case, the only candidate 5-itemsets would be: {a, e, g, r, w}, {n, q, r, t, v}

A neat Apriori trick

This trick

- **guarantees to capture the itemsets that have enough support,**
- **will still generate some candidates that don’t have enough support, so we still have to check them in the ‘pruning’ step,**
- **is particularly convenient for implementation in a standard relational style transaction database; it is a certain type of ‘self-Join’ operation.**

Last step: Generating rules from frequent itemsets

- Frequent itemsets ≠ association rules
- One more step is needed to generate association rules
- For each frequent itemset *X*,

  For each proper nonempty subset *A* of *X*,
  - Let $B = X - A$
  - $A \rightarrow B$ is an association rule if
    - Confidence($A \rightarrow B$) ≥ minconf,

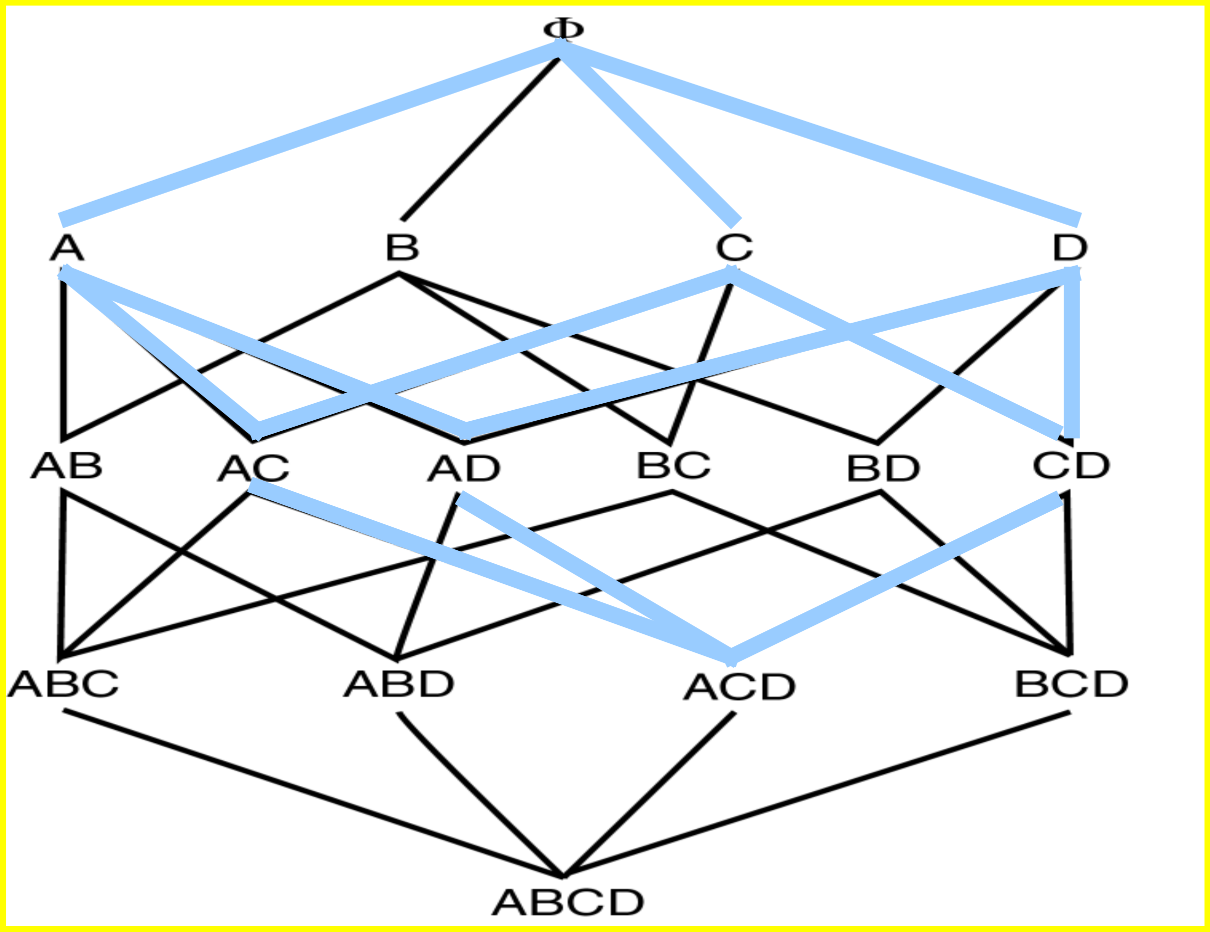
      support($A \rightarrow B$) = support($A \cup B$) = support(X)

      confidence($A \rightarrow B$) = support($A \cup B$) / support(A)

Generating rules: an example

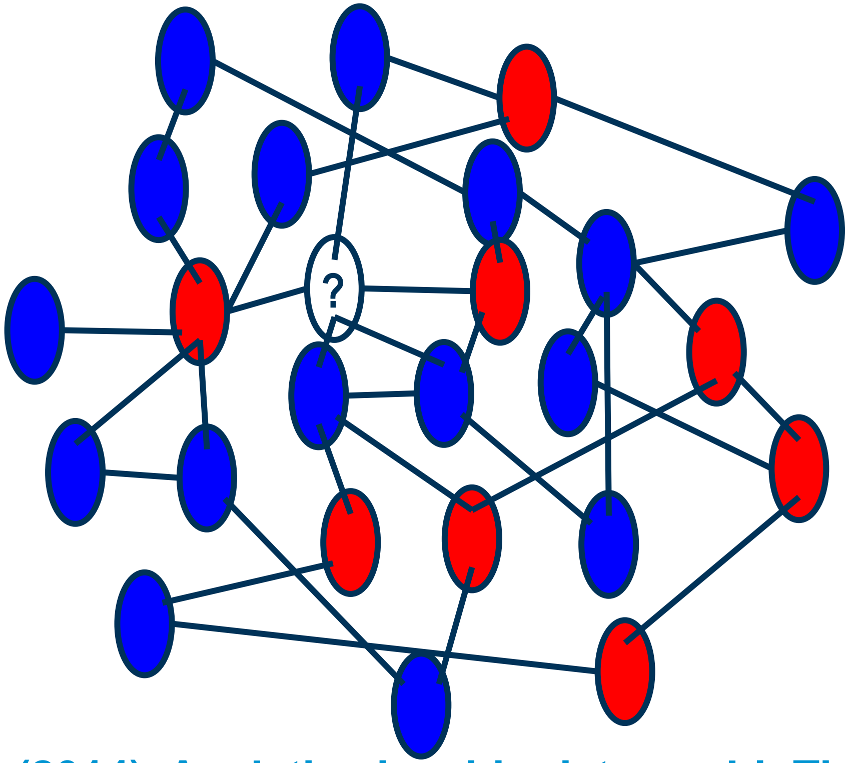
- Suppose {2,3,4} is frequent, with sup=50%
  - Proper nonempty subsets: {2,3}, {2,4}, {3,4}, {2}, {3}, {4}, with sup=50%, 50%, 75%, 75%, 75%, 75% respectively
  - These generate these association rules:
    - 2,3 → 4, confidence=100%
    - 2,4 → 3, confidence=100%
    - 3,4 → 2, confidence=67%
    - 2 → 3,4, confidence=67%
    - 3 → 2,4, confidence=67%
    - 4 → 2,3, confidence=67%
    - All rules have support = 50%

Frequent Itemset Property



Social Network Analytics

- Networked data
  - Telephone calls
  - Facebook, Twitter, LinkedIn, …
  - Web pages connected by hyperlinks
  - Research papers connected by citations
  - Terrorism networks
- Applications
  - Product recommendations
  - Web page classification
  - Fraud detection
  - Terrorism detection

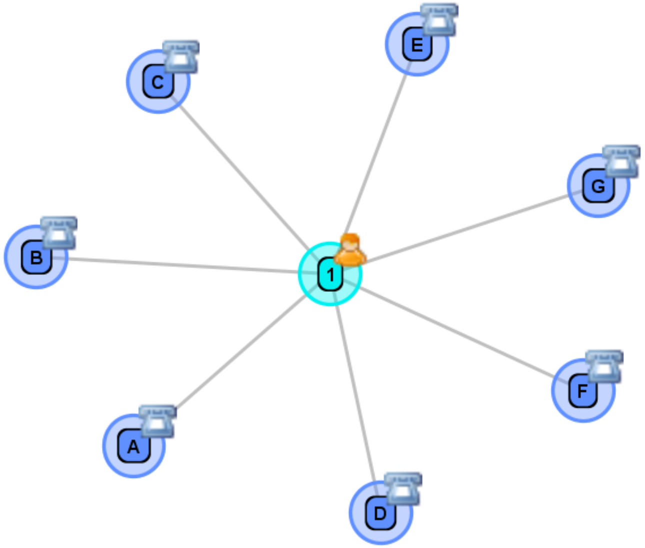


Baesens (2014), Analytics in a big data world: The essential guide to data science and its applications

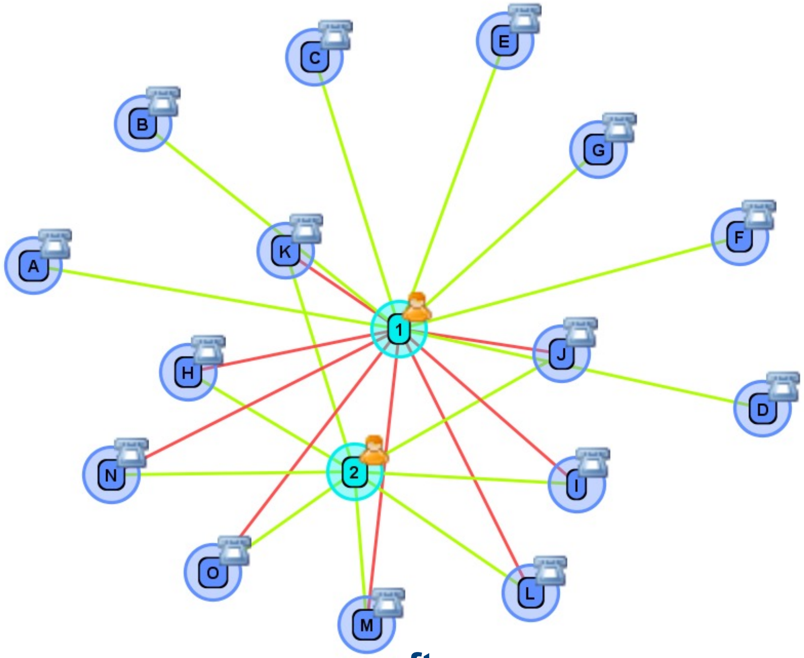
Fraud Analytics

Identify theft:

- Before: person calls his/her frequent contacts
- After: person also calls new contacts which *coincidentally* overlap with another persons contacts.



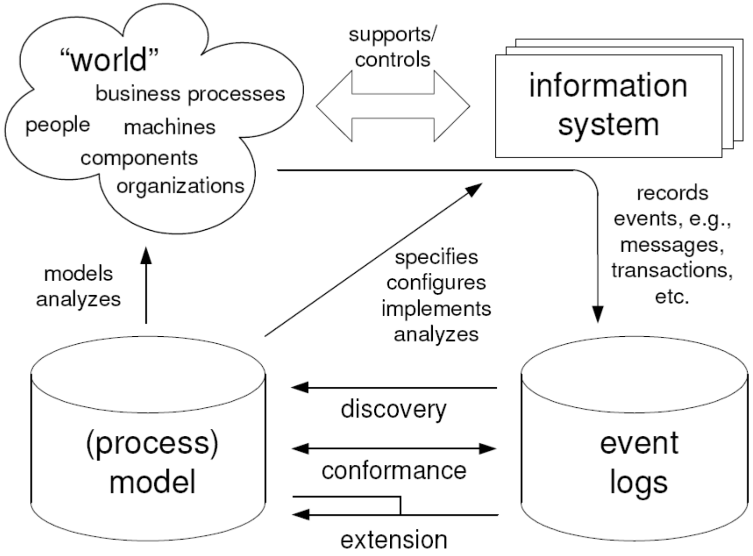
before



after

Process Analytics

- Extracting knowledge from event logs of information systems
  - Control flow perspective
  - Organizational perspective
  - Information perspective



De Weerdt, De Backer, Vanthienen, Baesens (2012), A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs