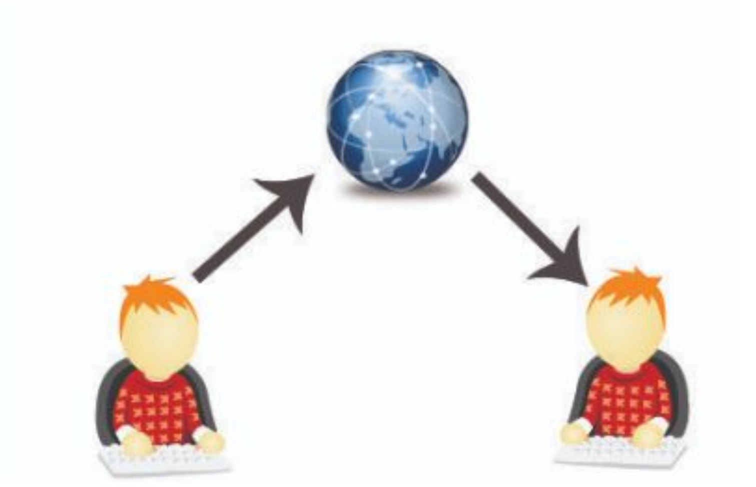
Introduction to Web X.0

Module 1

Evolution of Web X.0

Web 1.0

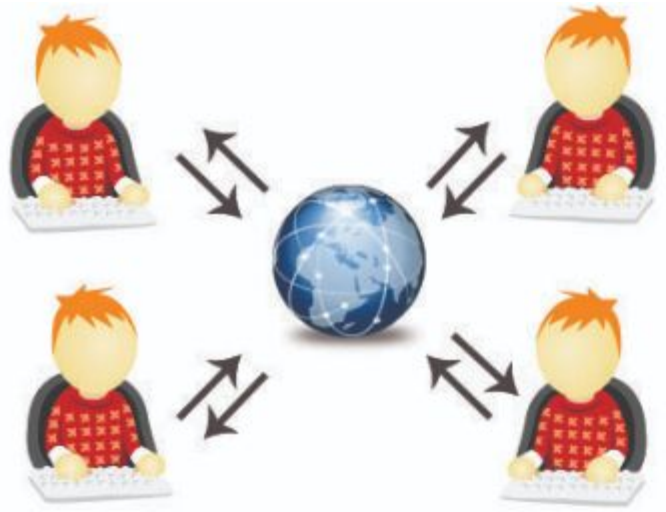
- It is 1st generation of the **WWW**.
- **Web 1.0** is one-way broadcasting, meaning only the owner of the site can publish information.
- It was the content web, static information.
- coined by Tim Berners-Lee as "**read only**" web
- focus was on content delivery rather than interaction and production.
- A system of internet servers that support specially formatted document by using **HTML**.



- WWW is not synonymous with internet.
    - The Internet is a massive network of networks, a networking infrastructure. While WWW or simply Web, is a way of accessing information over the medium of the internet.
    - Internet connects millions of computers together globally. WWW is an information sharing model that is built on top of the browsers

Web 2.0

- It is 2nd generation of the www that focused on the ability for people to **collaborate and share** information online.
- It signifies a **conversation** between the original author of the content and all those who can comment or participate.
- wisdom, people-centric, participative, and **read-write** web
- It refers to the transition form static HTML web pages to a more **dynamic web**.
- It is also regarded as the social web. Eg: Wikipedia, facebook, youtube.



- Web 2.0 is the **business revolution** in the computer industry caused by the move to the internet as platform
  - Recommendation Systems
- But Security was an issue!

Web 3.0

- Web 3.0 is an extension of Web 2.0.
- Owning data with **Decentralized** web
- Non-fungible tokens, or NFTs (.nft, .crypto, .bitcoin, .x, .blockchain)
  - Records in Blockchain
  - Need not renew
  - Use for website, social media handle
  - decentralized hosting (IPFS)/ Serverless hosting
  - Introduces Security
- Data on the web are **Semantic** based, rather than keyword

- web 3.0 is also known as the "**Semantic web**".
- It includes several technologies in order to arrange and structure data you can find on the internet to make it **available and usable by programs and software** thanks to a metadata system.
- The purpose is to make the **web readable by machines and not only by humans.**
- It is also about language, or ontology, for recording how the linked data relates to real world objects, allowing a 'machine' to `understand' the semantic meaning and the difference between Jaguar (car) and jaguar (animal)

- It starts with the Resource Description Framework (RDF) which gives the specifications of such a metadata data model. It is also the Web Ontology Language (OWL) and notations like RDF Schema (RDFS).

Web 1.0	Web 2.0	Web 3.0
Hypertext Web	Social Web	Semantic Web
Read-only	Read-and-Write	Executable
One Directional	Bi-directional	Multi-user virtual environment
Companies publish content	People publish content	People build applications where people interact and publish content
Company focus	Community focus	Individual focus
Owning content	Sharing content	Consolidating content
WebForms	Web Applications	Smart applications
Personal Websites	Blogs and Social Media Profiles	SemiBlog, Haystack
Page Views	Cost Per Click	User Engagement
Banner Advertising	Interactive Advertising	Behavioral Advertising

Web 4.0

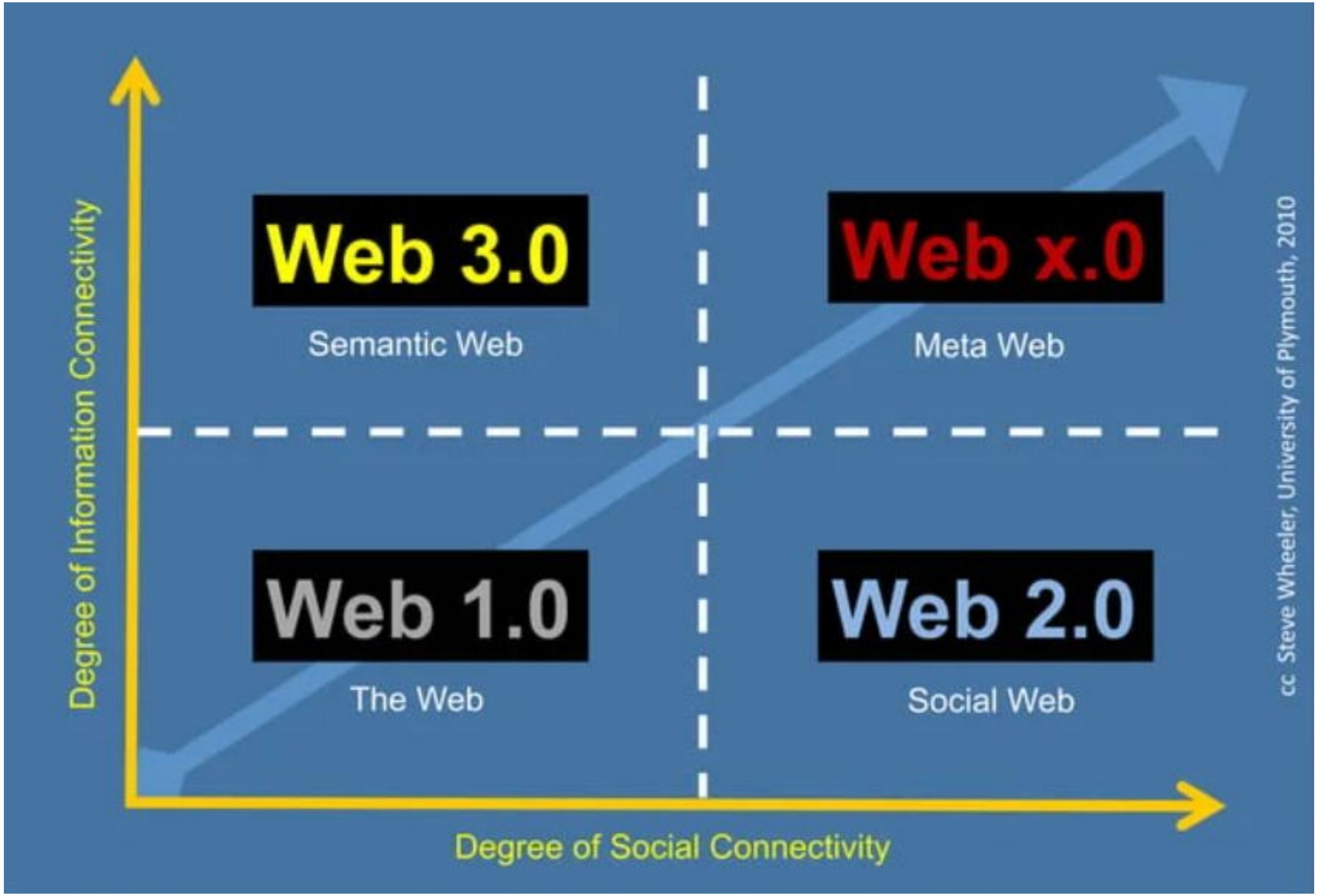
- Here is addition of more technology rather than only content
- Connect device to humans
  - Machine Learning
  - VR
  - IOT/ Internet of Everything
  - Artificial Intelligence
- Tools to make it more secure
- Intelligent Web
- Eg: facial recognition, an electronic agent will recognize users by their voice through an Internet-connected device, Personalized search, auto suggestion

- Web 4.0 also known as the **Symbiotic web**.
  - It will aim at the interaction between humans and machines in symbiosis. Humans and machines will have a **symbiotic relationship**, where machines will be capable of learning from human interactions and providing insights that can assist in decision-making.
- In Web 4.0 services are autonomous, proactive, self-learning, collaborative, and content-generating agents.
- Fully mature semantic and reasoning technologies, as well as AI are the foundation.
- They use and support Web databases through intelligent agents.
- Examples include services that interact with sensors and implants, natural language services, or virtual reality services.

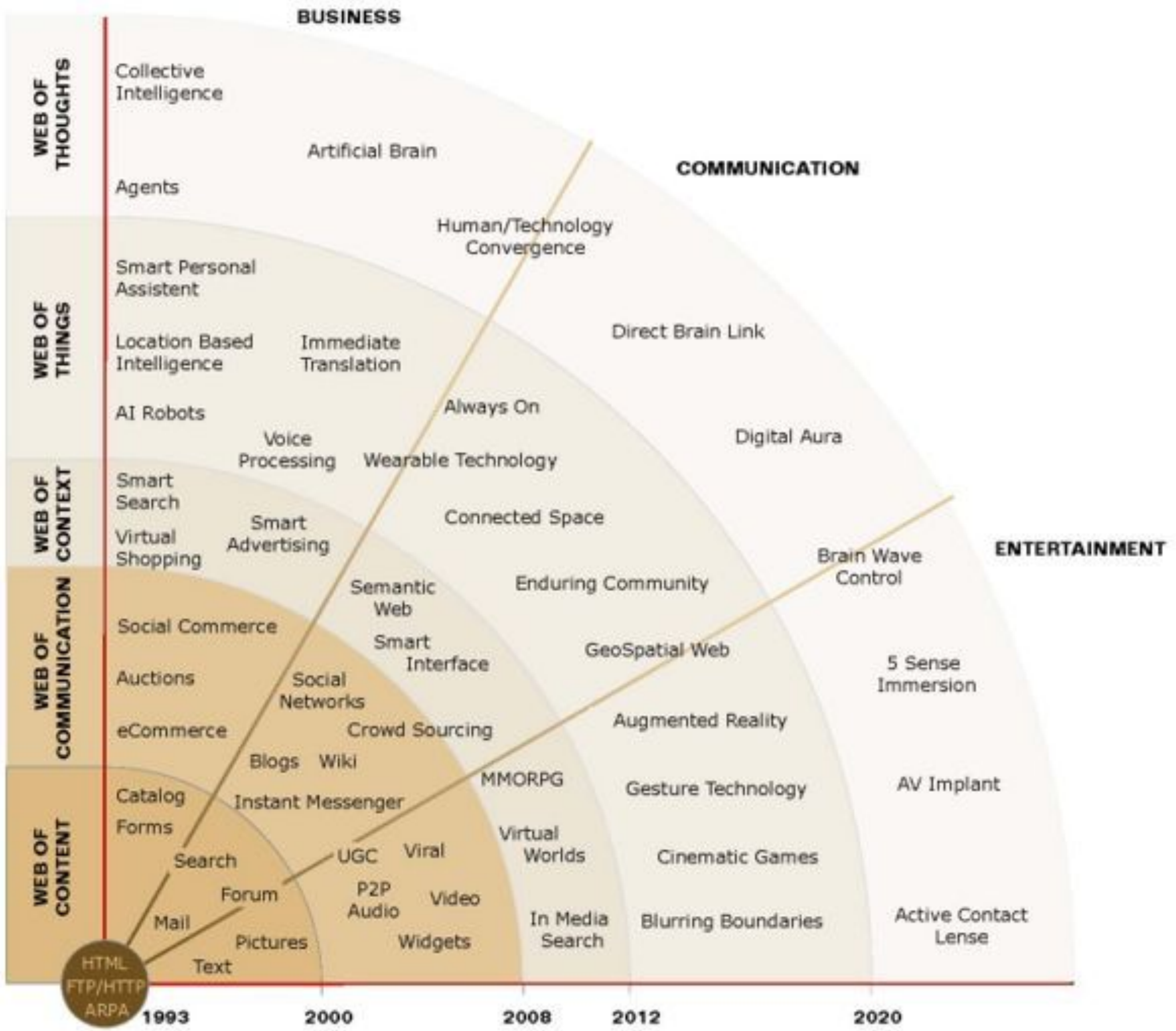
Web 5.0

- focus on empowering users to **reclaim control and ownership** of their data
- Former Twitter CEO Jack Dorsey recently announced his vision for a **new decentralized web** platform that is being called Web 5.0 and is being built with an aim to return “**ownership of data and identity**” to individuals.
  - The **Block Head (TBH)**. Web 5.0 is aimed at “building an extra decentralized web that puts one in control of ones’s data and identity”.
  -

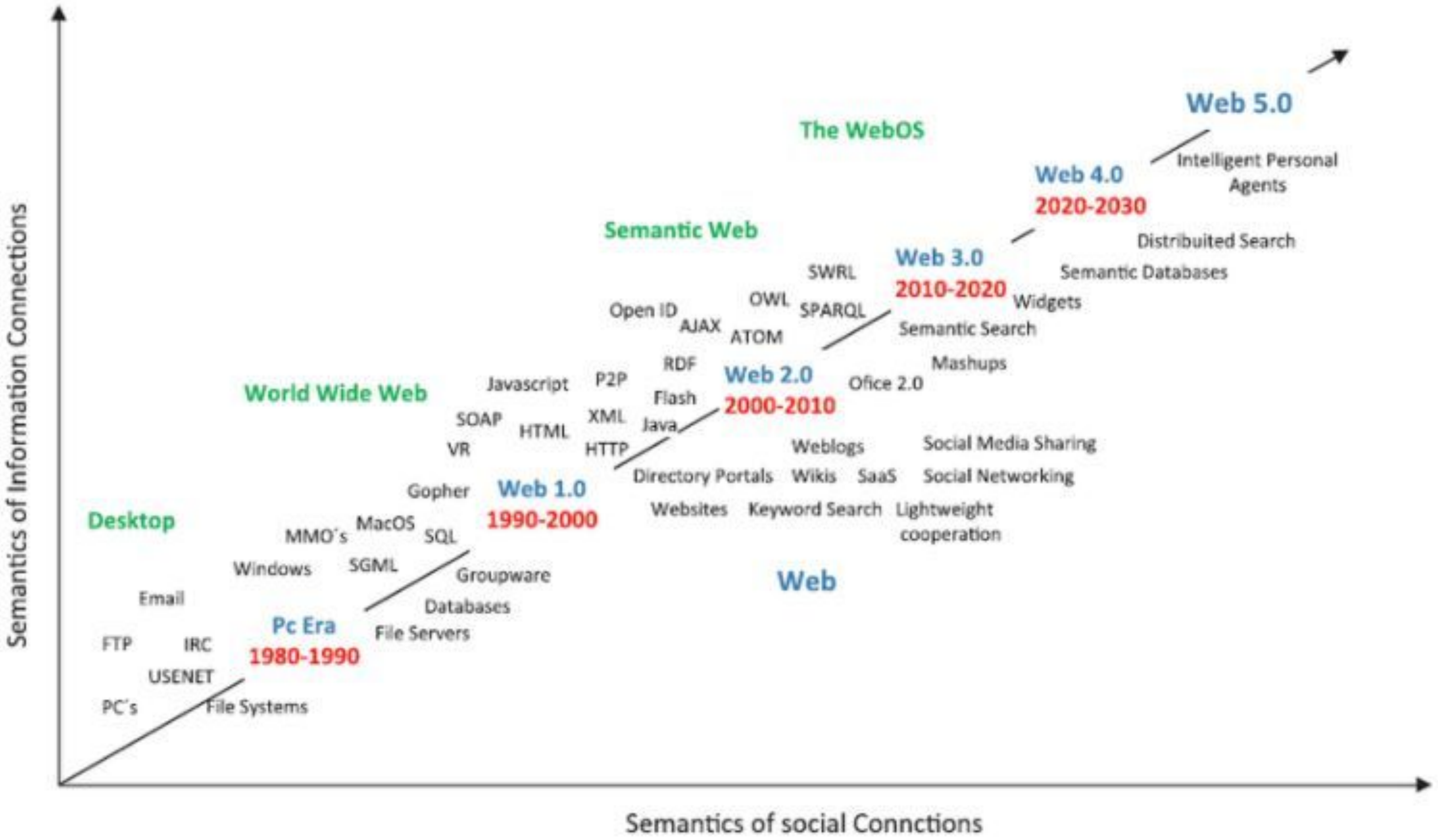
- linked, emotive and symbolic web that communicates with humans.
  - It is based on emotional association with humans, paving the way for more personalized experience, which will attract more people ever.
  - Eg: a personal assistant
- think of possibilities with web 5.0 examples
  - such as a website that can map the emotions of a person.
  - The website can use facial recognition as a tool for mapping the person’s emotions when the user wears a headphone.
  - Users can interact with the website, and the website would respond accordingly based on emotions of the person.


Web 3.0
Semantic Web
Web x.0
Meta Web
Web 1.0
The Web
Web 2.0
Social Web
Degree of Information Connectivity
Degree of Social Connectivity
cc Steve Wheeler, University of Plymouth, 2010

THE WEB EXPANSION
FROM WEB OF THINGS TO WEB OF THOUGHTS



©TrendONE 2008 by Nils Müller
www.TrendONE.de All rights reserved

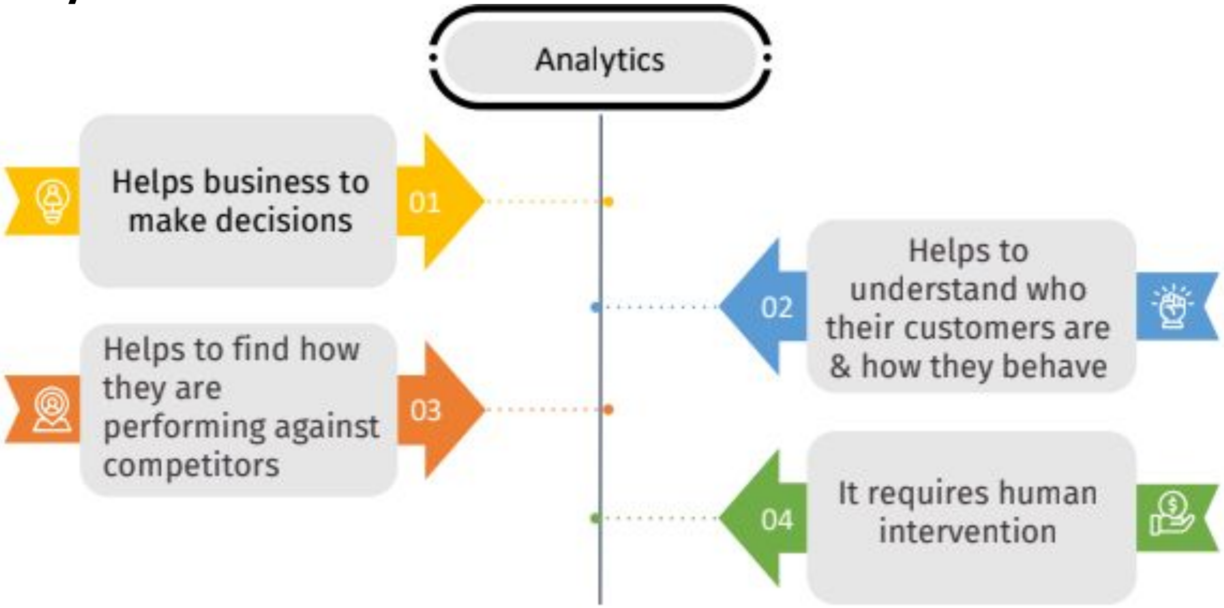


Web 5.0
The WebOS
Intelligent Personal Agents
Web 4.0
2020-2030
Semantic Web
Distribuited Search
SWRL
Web 3.0
2010-2020
Semantic Databases
OWL
Open ID
SPARQL
Widgets
AJAX
ATOM
Semantic Search
RDF
Web 2.0
Mashups
World Wide Web
Javascript
P2P
Office 2.0
Flash
2000-2010
SOAP
XML
HTML
Java
VR
HTTP
Weblogs
Social Media Sharing
Directory Portals
Wikis
SaaS
Social Networking
Gopher
Web 1.0
Websites
Keyword Search
Lightweight cooperation
Desktop
MacOS
1990-2000
MMO´s
SQL
Windows
SGML
Groupware
Web
Email
Databases
Pc Era
File Servers
FTP
IRC
1980-1990
USENET
PC´s
File Systems
Semantics of Information Connections
Semantics of social Conntions

Web Analytics 2.0:

Web Analytics

- Why analysis?
  - examination of the different parts or details of something
  - Break down and study for better understanding
- What is Analytics?
  - Systematic computational analysis or Science of Analysis
- Why Web Analysis?



- Web analytics allow an organization to **measure, collect, and analyze data from the visitors** of an organization’s website.
- Web analytics is used primarily for **marketing research**.
- You can view the effectiveness of your website and exactly **where the traffic is coming from**.
- You can also screen the results of different types of advertising.
- Web analytics is used to boost your nonprofit organization’s brand.
- It is a way to analyze your website’s usage statistics in order to better reach your target audience, and customize web content to their needs.

In all ……

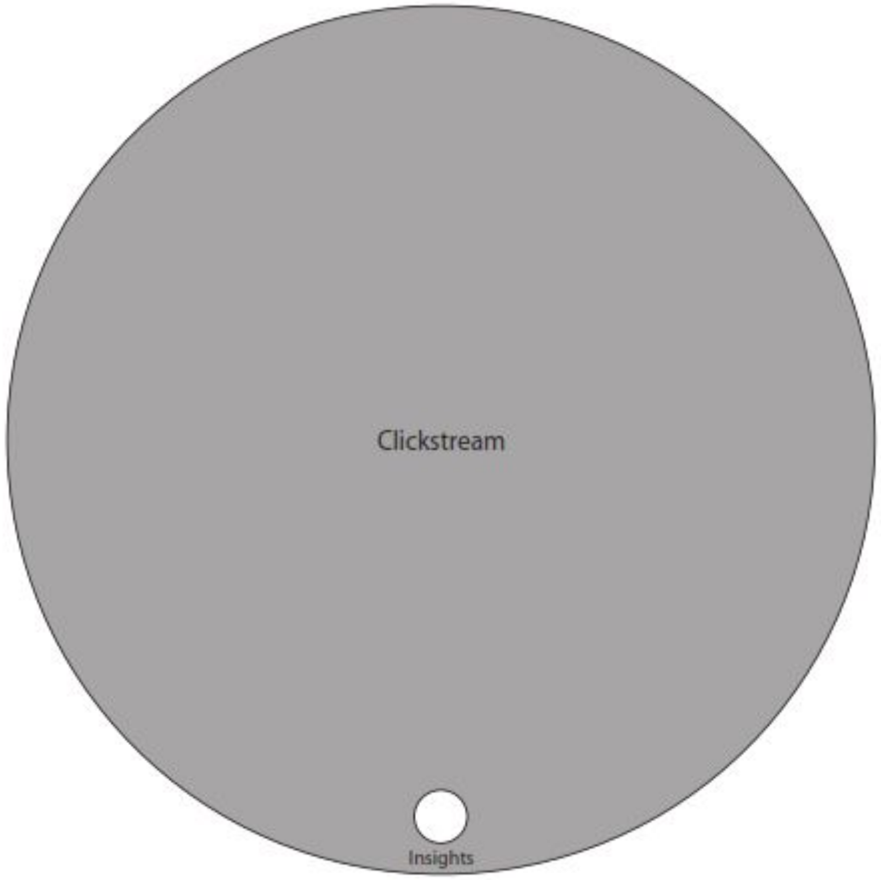
- **Understanding user behavior.** Website analytics tools can provide insights into user behavior, highlighting popular areas of a site and identifying pain points for users.
- **Identifying opportunities for improvement.** By analyzing data from website analytics tools, it is easier to identify opportunities for improving user experience. This includes optimizing content for search engines, improving the overall functionality of the site or adjusting layout or design.
- **Measuring the effectiveness of marketing efforts**. With website analytics tools it’s more simple to measure the effectiveness of marketing efforts, such as email campaigns or social media posts.

You can track the number of visitors and conversions from your social media channels to identify the most successful marketing strategy and allocate resources accordingly. To generate more follows from target customers on social media you can also make a custom QR code that will be shared with potential clients.

- **Setting and tracking goals.** Website analytics tools allow you to set and track specific goals for your website, such as increasing the number of conversions or improving the average time spent on the site. By regularly reviewing data from these tools, it is possible to measure the progress toward these goals and identify areas for improvement.

- Paradox of data: a lack of it means you cannot make complete decisions, but even with a lot of data, you still get an infinitesimally small number of insights.
- Tools from Google Analytics, omniture sitecatalyst, webtrends, clicktracks or Xiti

Rethinking Web Analytics: Meet Web Analytics 2.0



The old paradigm of Web Analytics 1.0

- limits of clickstream data
- we have the *what:*
    - *what pages did people* view on our website?
    - what products did people purchase?
    - what was the average time spent?
    - what sources did they come from?
    - what keywords or campaigns produced clicks?

what this, and what that, and what not?

But what data is missing the why……….

The Awesome World of Clickstream Analysis: Metrics

- *exploring the awesome world of Clickstream analysis with the building blocks of metrics and key performance indicators (KPIs)*.
- **The Critical Eight:**
    - **Visits:** Someone came to the website. Also sometimes called a session.
    - **Unique Visitors**: An approximation of the number of people who have visited the site. Not a perfect measure; it assumes each visitor is an individual. Also must be carefully parsed to not count each visitor multiple times over a given period; metrics vendors must account for this, and it is a computational expense.
    - **Time on Page**: Time spent on an individual page. Difficult to measure; metrics vendors must adjust code to measure this.

- **Time on Site**: Total time spent on the site during a session, related to Time on Page. Has the same issues as Time on Page.
- **Bounce Rate:** Percentage of sessions spent on a site with a single page view. Not a positive metric; it means that people are not satisfied and leave immediately. Can be used to find bad pages on your site, and determine which referring sites and search terms are not working. Note an exception for blogs
- **Exit Rate:** How many people left the site from a certain page. Everyone has to leave eventually, so Bounce Rate is better metric for that. But the related "Abandonment Rate" can measure users who complete a multipage task.

- **Conversion Rate:** Mathematically, it’s Outcomes/Unique Visitors (or Visits). Used by ecommerce sites to determine if people are buying, buying after multiple visits, or just looking.
- **Engagement:** The depth of the visit, impossible to gauge without surveys (especially the question of “likely to recommend”).

Web Analytics 2.0 is:

- the analysis of **qualitative and quantitative** data from your website and the competition,
- to drive a **continual improvement** of the online experience that your customers, and potential customers have,which **translates into your desired outcomes** (online and offline).
- Web Analytics 2.0 answers
    - What is happening on my website?
    - What are the outcomes?
    - Why is it happening?
    - Competitor Analysis
    - What actions should be taken?

The What: Clickstream

- *what is collecting, storing, processing, and analyzing your* website's **click-level data**
- it helps you measure pages and campaigns and helps you analyze all kinds of site behavior: Visits, Visitors, Time on site, PageViews, Bounce rate, sources, and more.
- click-level data is data you get from Webtrends, Google Analytics, and other clickstream tools.
- you will have a lot of data—in the order of gigabytes in a few months and more if you store history.

The How Much: Multiple Outcomes Analysis

- If the reports don’t focused on **measuring outcomes**, then it is of no use
- focusing deeply and specifically on measuring outcomes means connecting customer behavior to the bottom line of the company.
- Tie outcomes to profits and to the bonuses of your report recipients.
- A website attempts to deliver just three types of outcomes:
  - increase revenue.
  - reduce cost.
  - improve customer satisfaction/loyalty.

- Having clearly defined website goals is necessary and a great starting point, but not sufficient for true user insight.
- Analysts should recognize that users and companies may have many different reasons for visiting or hosting a website
- All those objective outcomes need to be measured to see if the site is really driving the desired outcomes

The Why: Experimentation and Testing

- you can run experiments live on your site with various ideas and let your customers tell you what works best
- Failing online is cheap and fast!
- Testing online - different promotional offers via email or search ads before you finalize your strategy and launch it using print, catalog, or TV ads
- Tools: Google website optimizer, Omniture’s Test&Target, Autonomy’s Optimost, Sitespect

- Some testing techniques
    - A/B testing - Test the design of your website, including text, graphics, buttons, banner ads, everything.
    - Google Website Optimizer - Automates the process of running multivariate tests to help you quickly find the optimal.

The Why: Voice of Customer

- your web analytics tool can report only what it can record. what your customers wanted but did not see was not recorded.
- ask website visitors through surveys:

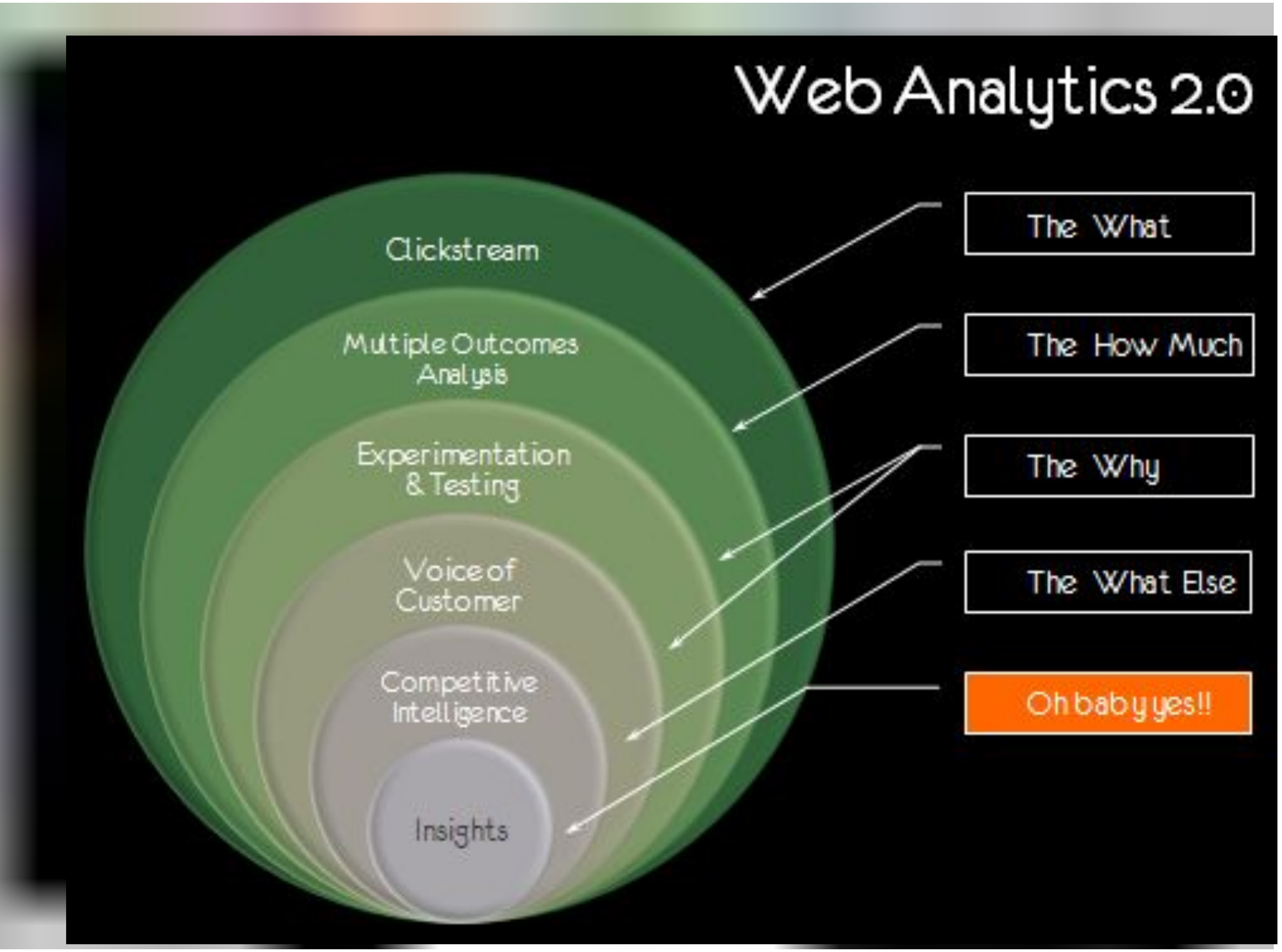
  (1) why are you here?;

  (2) were you able to achieve what you came for?;

  (3) if not, why not? The results can be tied back to analytics data and may reveal customers' true motivations.

The What Else: Competitive Intelligence

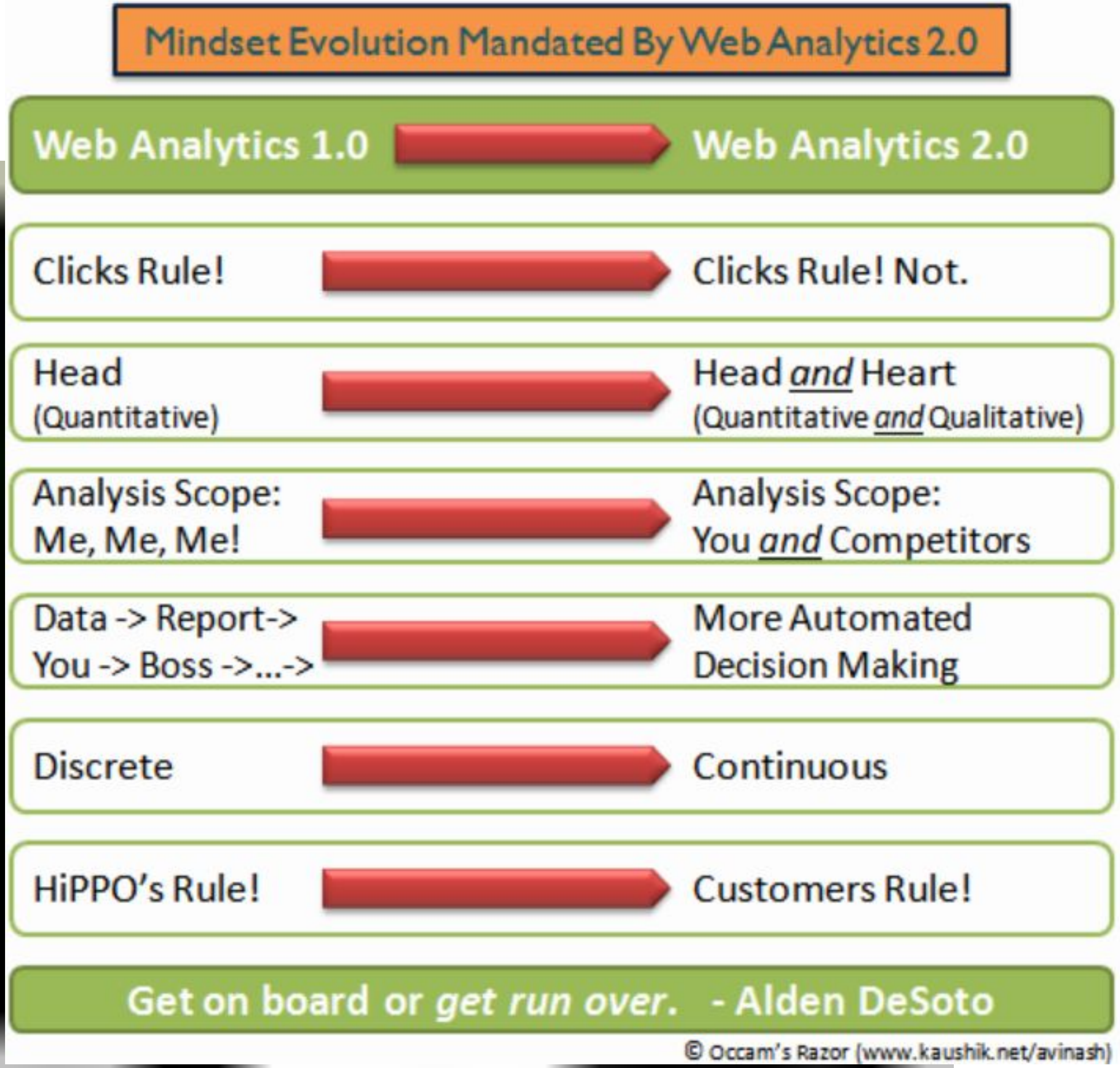
- Knowing how you are performing against your competition - helps you improve, it helps you identify new opportunities, and it helps you stay relevant
- Your competitors may be running campaigns or launching products/features that are impacting your site’s performance (could be either up or down).



Web Analytics 2.0.

- Clickstream answers the ***what.***
- *Multiple outcomes Analysis answers the* ***how***
- *Experimentation and Testing and Voice of* customer *help explain the* ***why***
- competitive intelligence answers the ***what else***

Strategic Shift with Web Analytics 2.0



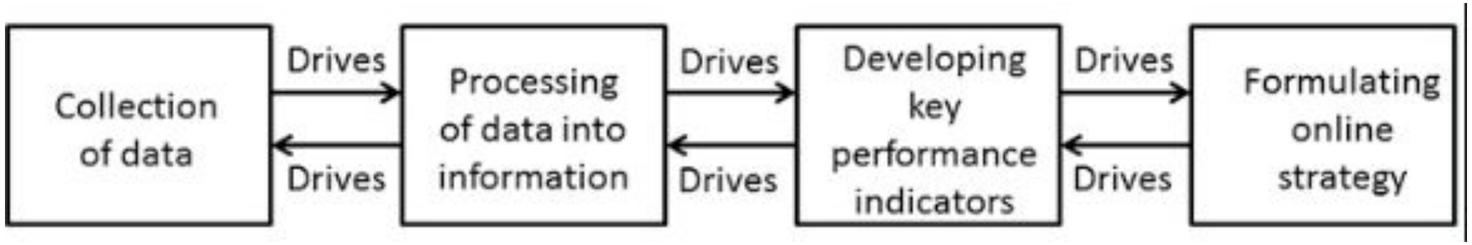
Digital Analytics

- Digital Analytics is the science of analysis with focus on Internet data.
  - Source : Website – Web Analytics, Mobile apps, Social Media, Third party sources
- It is the analysis of quantitative and qualitative data, to gain insights and drive

continual improvement.

- Big Data, Cloud and Diagnostic Analytics are the trends in Digital Analytics

Web Anaytics is the **measurement**, **collection**, **analysis** and **reporting** of **Internet data** for the purpose of **understanding** and **optimizing** web usage



How to Choose a Web Analytics Tool:

- Collect business requirement
- Collect technical requirements
- Documentation
- Prepare Request for Proposal (RFP)
- Send RFP to Vendor
- Analyse reply
- Select one vendor
- Implement

- Collect business requirement
  - support both your existing and future business requirements.
  - Find core objectives of your business and create a list of your desired business outcomes. Next, break down your business objectives into measurable analytics goals.
  - Business Objectives like - goals, strategy, KPI's, reports, reporting schedule and so on
  - Decide what exactly it is you want to track and improve in your business.

- Some of the metrics you might want to measure include website traffic, online sales, SEO ranking, sales lead quality, customer engagement, subscriber growth, inbound links and time spent on site.
- choose a web analytics software that can provide the data you need for your business. In addition, the solution should be scalable to accommodate the future growth and expansion of your business.

- Collect technical requirements
  - site architecture, servers, scripts, pages, IT needs and so on
- Documentation
  - All who need to access the web analytics data should be contacted (inside and outside the company)
  - Document their needs
- Prepare Request for Proposal (RFP)
  - A **Request for Proposal** is a document that solicits proposal, often made through a bidding process, by an agency or company interested in procurement of a commodity, service, or valuable asset, to potential suppliers to submit business proposals

- Put up all the business and technical requirements along with vendors financial stability

- Send RFP to Vendor
  - Send RFP to Vendor and set a schedule for reply
- Analyse reply
  - Check feasibility
  - How much support will you need during the tool evaluation phase, start-up and maintenance?
  - Is there a downloadable evaluation version or hosted demo available for a trial period?

- Verify if start-up or maintenance support is part of the package or incremental.
  - What amount of technical support and problem response will you need?
  - Is ongoing or more advanced training available?
  - Is training available online anytime or periodically in person?
- Select one vendor who meet your requirements
- Implement the tool

The Optimal Strategy for Choosing Your Web Analytics Soul Mate

- Tools are Data rich but information poor, we need an **analyst**
- Follow 10/90 rule
    - for every $10 you spend on your analytics tool and implementation, you should spend $90 on intelligent digital analysts who can convert your data into actionable insights
- **Step 1: Three Critical Questions to Ask Yourself Before You Seek an Analytics Soul Mate!**
    - “Do I want reporting or analysis?”
    - “Do I have IT strength, business strength, or both?”
    - “Am I solving just for Clickstream or for Web Analytics 2.0?”

- **Step 2: Ten Questions to Ask Vendors Before You Marry Them**
  - "What is the difference between your tool/solution and **free tools** from Yahoo! and Google?"" Are you planning a software version? "
  - "What data **capture mechanisms** do you use?"
  - "Can you calculate the total cost of ownership for your tool?"
  - "What kind of **support** do you offer? What do you include for free, and what costs more? Is it free 24/7?"
  - "What features in your tool allow me to segment the data?"
  - "What options do I have for **exporting** data from your system into our company's system?"
  - "Why did the last two clients you **lost cancel** their contracts? Who are they using now?

- **Step 3: Identifying Your Web Analytics Soul Mate (How to Run an Effective Tool Pilot)**
  - **Usability**:
    - Determine the **accessibility/intuitiveness** of the tool.
    - Establish whether your target audiences (for example, business, data analyst, and IT) can actually use and **customize** the tool set and reporting or whether you must get dedicated resources to create the necessary reporting and dashboards on their behalf.
    - Get a feel for the extent of **training** needed.

- **Functionality**
  - Test the functionality in **realistic business situations**: does it really do what it said on the tin?
  - Can you use out-of-the-box reports/features and page tagging, or do you need to customize and extend data collection to meet your needs? (You may need to run a handful of scenarios with vendors.)
  - Test **potential interoperability** with your other systems/data sources.

- **Step 4: Negotiating the Prenuptials: Check SLAs for Your Web Analytics Vendor Contract**

The Key to Glory:Measuring Success

- Actionable Outcome KPIs
- **Key Performance Indicators (KPI)** are financial and non-financial metrics used to help an organization define and measure progress toward organizational goals.
- Most important measures are: **outcomes** / **conversions**
- Implement your **Goals** in the web analytics tool
- Focus on the **critical few** - 2 to 4 metrics - most important factors for your business

- **Key Performance Indicators (KPIs) for business:**
  - **Task Completion Rate** - is the percentage of Visitors to your website who rate if they were able to complete the primary purpose for their visit.
  - **Share of Search** - is the percentage of traffic you get from search engines compared to your key competitors.
  - **Visitor Loyalty and Recency** - Visitor Loyalty measures the distribution of the number of Visits by each Visitor to your site. Recency measures the gap between two Visits by the same Visitor.

- **RSS/Feed Subscribers** - measures the raw number of people who have signed up for your website or blog’s RSS (really simple syndication) feed.
- **% of Valuable Exits** - measures the percentage of people who leave your website by clicking something of value to you.

- **For ecommerce site** use the following Abandonment Rate metrics
    - **Cart Abandonment** (percentage) This is 1 minus (the total Visitors who start checkout divided by the total number of add to cart clicks).
    - **Checkout Abandonment** (percentage) This is 1 minus (the total Visitors who complete checkout divided by the total number of people who start checkout).
    - **Days and Visits to Purchase** - Days to Purchase shows the distribution of the number of days it takes someone to make a purchase on your website. Visits to Purchase shows the number of Visits until purchase.

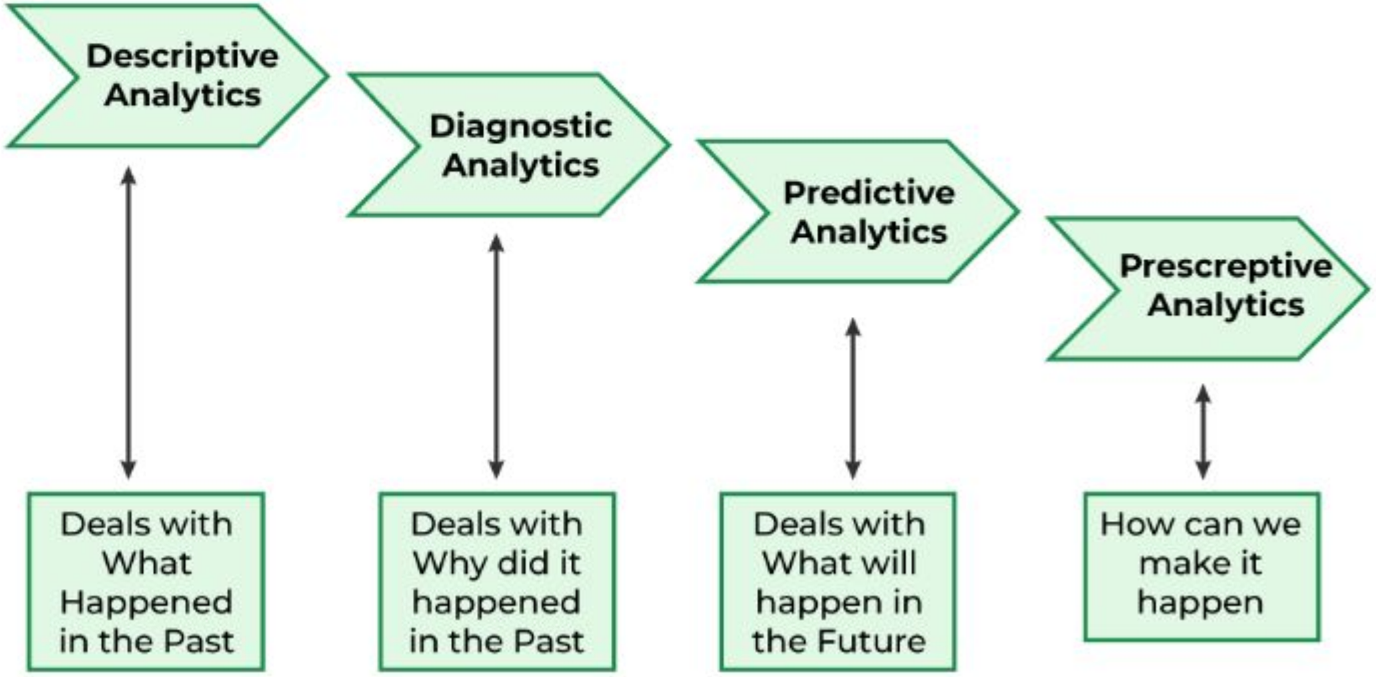
- **Average Order Value** - It’s the total amount of revenue divided by the total number of orders received.
- Primary Purpose - Identify the Convertible - Opportunities
- Measuring Macro (targeted/main) and Micro (other) Conversions Examples
    - Technical Support Website
        - Macro - Task Completion Rate
        - Micro - Call Avoidance, Content Consumption, Tickets Opened, Sales, Net Promoters
    - Multipurpose Ecommerce Website
        - Macro - Completed orders
        - Micro - Visitors create accounts, software downloads, Contact Us

- Measuring success for **Non-Ecommerce Website** (news website or a blog or an academic journal website)
    - **Visitor Loyalty** - Visitor Loyalty tells you how often Visitors visit your website during the reporting period. Matric - average Visits per visitor.
    - **Visitor Recency** - Visitor Recency tells you how long it has been since a visitor last visited your website
    - **Length of Visit** - Length of Visit measures the quality of visit as represented by the length of a visitor session in seconds.
    - **Depth of Visit** - Depth of Visit measures the distribution of the number of pages in each Visit to the website, during a given reporting period

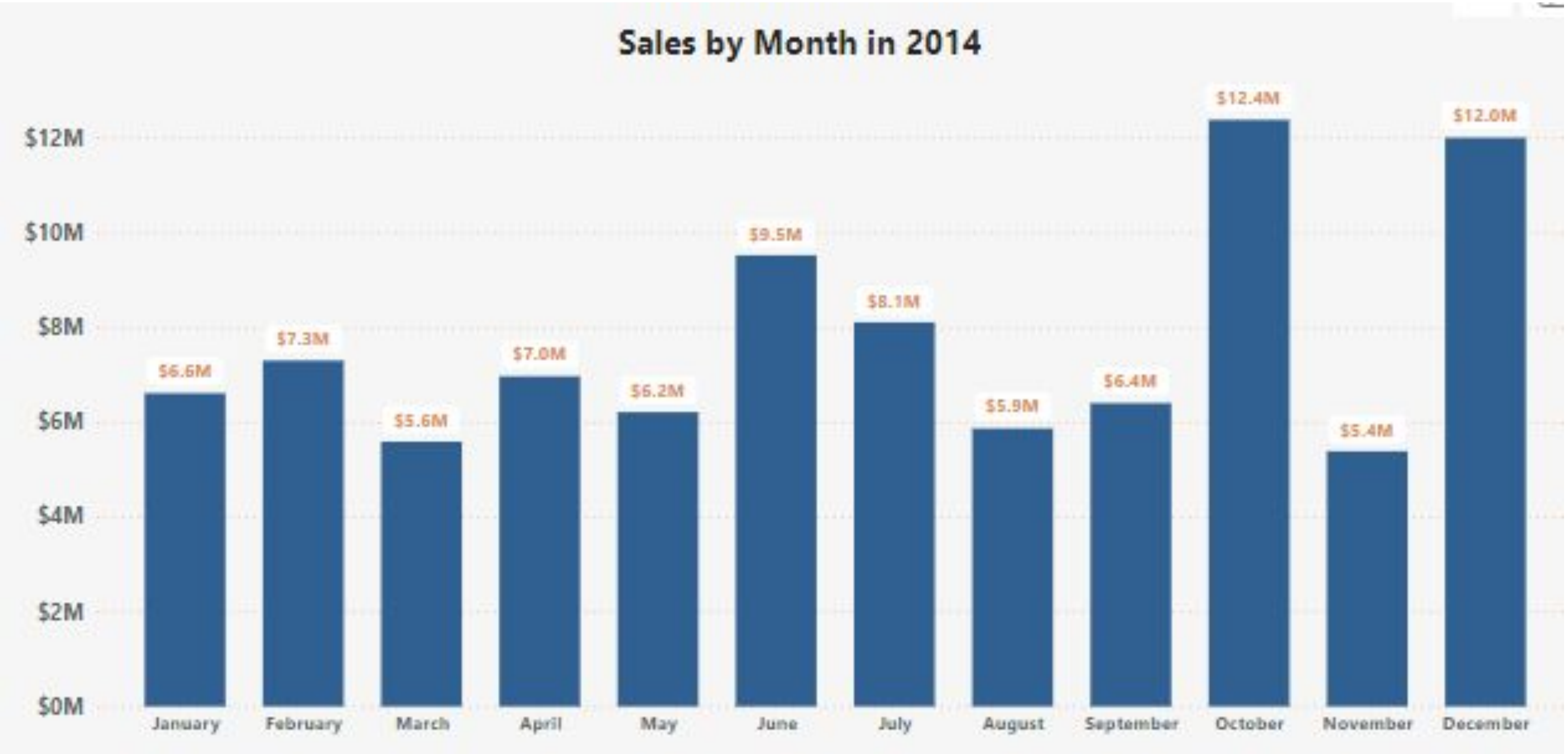
- **Measuring success for B2B website**
  - Percentage of Visits that viewed the Product Folder directories
  - Percentage of selection and solution guide downloads
  - Number of free samples ordered
  - Number of new accounts opened
  - Number completed videos watched
  - Percentage of solutions posted by the same member ID
  - Usage of decision support tools

Different Types of Analytics

- Understanding and leveraging different types of analytics provided added value to a business to improve its organization-level operational capabilities.
- Choosing the right forms of data, analytics can make most of the unstructured or structured data they own
- Types of Analytics
  - Descriptive
  - Diagnostic
  - Predictive
  - Prescriptive.



- **Descriptive Analytics (What happened?)**
  - Descriptive analytics is the branch of analytics that involves the interpretation and summarization of **historical data** to understand **what** has happened in the past.
  - This is the baseline and the place that all organizations should start
  - It looks at the events of the past and tries to identify specific patterns within the data.
  - Visualizations commonly used for Description Analytics include pie charts, bar charts, tables, or line graphs.



- The chart below shows sales from 12 months through this we can identify the trend of sales.
- Observe increase in sales in October and December, with a decrease in November.

Diagnostic Analytics (Why did it happen?)

- This is a form of advanced analytics that examines data or content to answer the question, “Why did it happen?”.
- This is the second step as you must first understand what happened to be able to identify why it happened.
- Once an organization achieves descriptive insights, they can apply diagnostics with a little more work.
- It is characterized by techniques such as drill-down, data discovery, data mining, and correlations.

Example you hover over you can see a breakdown by segment. You can now see which segments contributed the most to an increase in sales.

(with reference to example in slide 40)

- Diagnostic Analysis:
  - Explore factors that may have influenced variations in sales performance. For example, analyze marketing campaigns, promotions, or external events that coincided with peaks or drops in sales.
- Reporting:
  - Generate a monthly sales performance report that includes visualizations and key insights. Share this report with stakeholders, including executives, marketing teams, and sales teams.

Predictive Analytics (What is likely to happen?)

- Once an organization can effectively understand what occurred and why it happened, they can move up to the next tier in analytics, Predictive.
- Predictive Analytics is another type of advanced analytics that looks to use data and information to answer the question "What is likely to happen?".
- Predictive Analytics involves techniques such as regression analysis, forecasting, multivariate statistics, pattern matching, predictive modeling, and forecasting.
- These techniques are harder for organizations to accomplish as they require large amounts of high-quality data. Additionally, these techniques require a deep understanding of statistics and programming languages such as R and Python.

(with reference to example in slide 40)

- Once validated, use the trained model to make predictions for future monthly sales performance. This involves i**nputting relevant data** for upcoming months into the model to generate forecasts.
- Predictive Model will suggest the **impact** of the next marketing campaign on customer engagement using historical data.
- Conduct scenario analyses by adjusting input variables to observe how changes in factors like marketing spend or promotions might impact future sales predictions.

Prescriptive Analytics (What should be done?)

- The final level and most advanced level of analytics and most difficult level to achieve
- Prescriptive Analytics is a method of analytics that analyzes data to answer the question “What should be done?”.
- Characterized by techniques such as graph analysis, simulation, complex event processing, neural networks, recommendation engines, heuristics, and machine learning.
- involves using data, algorithms, and business rules to recommend actions that can optimize or improve future outcomes.

- The reliability of Prescriptive Analytics depends heavily on the accuracy of
  - how well an organization as has accomplished each level of analytics.
  - quality of data
  - appropriate data architecture to facilitate it and the expertise needed to implement this architecture.
- The value that it brings is that an organization will be able to make decisions based on highly analyzed facts rather than instinct.
- Meaning they are more likely to guarantee the desired result, such as increasing revenue.

- (with reference to example in slide 40)
- Scenario 1: If the predictive model suggests a potential drop in sales due to decreased marketing effectiveness, the prescriptive analytics recommends increasing marketing spend for targeted campaigns.
- Scenario 2: If the model predicts a surge in demand for a particular product category, the prescriptive analytics recommends optimizing inventory levels for those products.
- By applying prescriptive analytics to monthly sales performance analysis, businesses can move beyond merely predicting future outcomes to actively recommending the best actions to achieve desired results.

Web3.0 and Semantic Web

- The development of Semantic Web is well underway with a **goal that it would be possible for machines to understand the information on the web rather than simply display**.
- The major obstacle to this goal is the fact that most information on the **web is designed solely for human consumption**. This information should be structured in a way that machines can understand and process that information.
- The key technological threads that are currently employed in the development of Semantic Web are: eXtensible Markup Language (**XML**), Resource Description Framework (**RDF**), DAML (**DARPA Agent Markup Language**).

- The concept of machine-understandable documents does not imply "*Artificial Intelligence*". It only indicates a machine's ability to solve well-defined problems by performing well-defined operations on well-defined data.
- Most of the web's content today is designed for humans to read , and not for computer programs to process meaningfully.
- Computers can
  - parse the web pages.
  - perform routine processing (here a header, there a ink, etc.)

- In general, they have **no reliable way to understand and process the semantics**.
- The **Semantic Web** will bring structure to the **meaningful content of the web of web pages,** creating an environment where software agents roaming from page to page carry out sophisticated tasks for users.
- The Semantic Web is not a separate web.
- **Tim Berner Lee**'s idea
  - To unlock the potential of the web
  - Link data and relation between the data
  - **Web of data rather than Web of Documents**

- Tim Beerner’s 3 Rules of Semantic Web
  1. A URI should **point to the data.**
  2. Anyone accessing the URL should **get data back.**
  3. Relationships in the data should **point to additional** URLs with data.
- Berners-Lee et al. **describe** the Semantic Web as “an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.”
- Information is structured in a way that computers can draw relationships between resources.
- Those relationships are what is known as **linked data.**

The Semantic Web - the Web of Data

- The meaning of information (Semantics) is made explicit by **formal (structured) and standardized knowledge representations (Ontologies)**.

- Thus it will be possible,
  - to process the meaning of information automatically
  - to relate and integrate heterogeneous data
  - to deduce implicit (not evident) information from existing (evident) information in an automated way

- The Semantic Web is kind of a **global database** that contains a **universal network of semantic propositions**.

three main topics that provide conceptual underpinnings for the Semantic Web:

- Building models: the quest for describing the world in abstract terms

to allow for an easier understanding of a complex reality.

- Computing with knowledge: the endeavor of constructing reasoning ma-

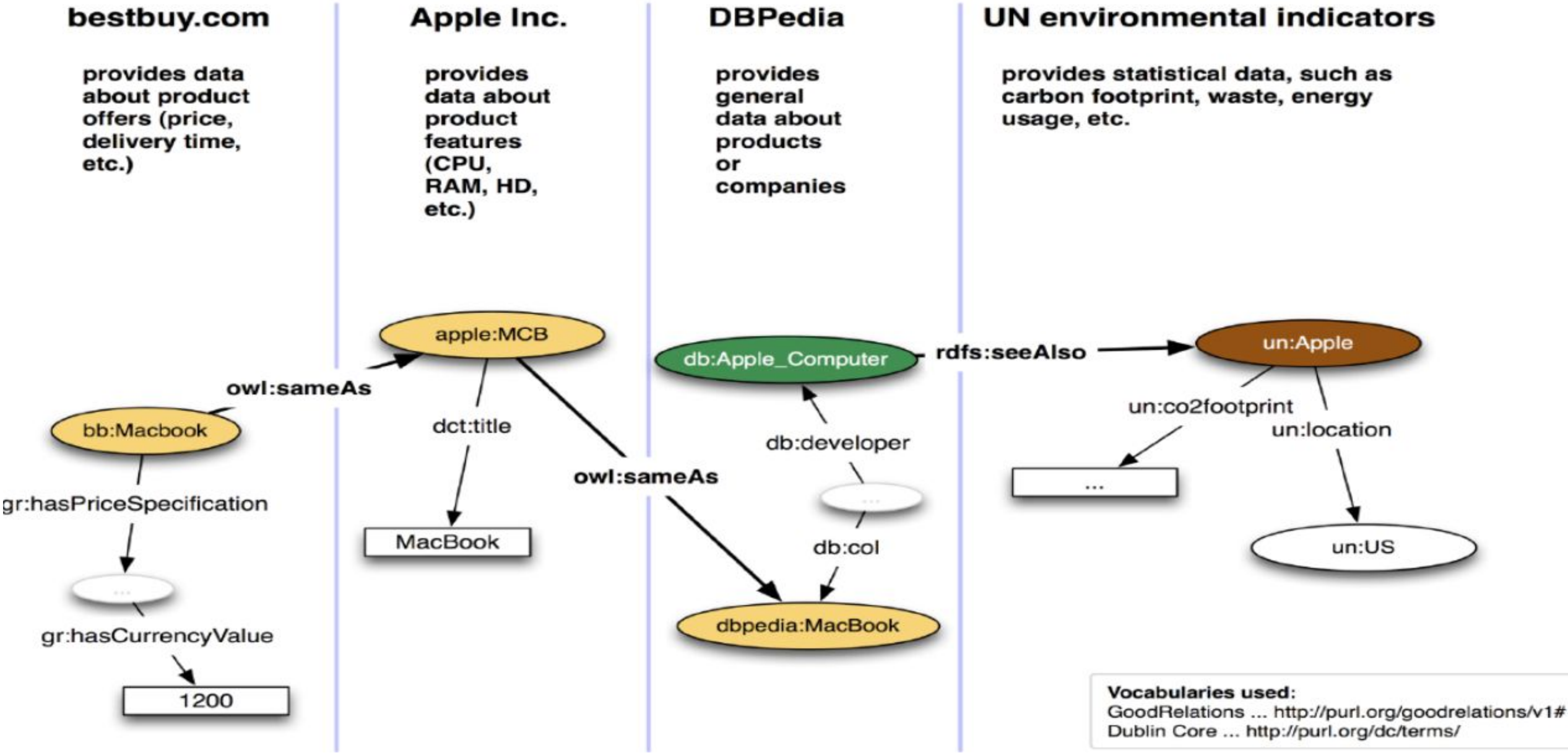
chines that can draw meaningful conclusions from encoded knowledge.

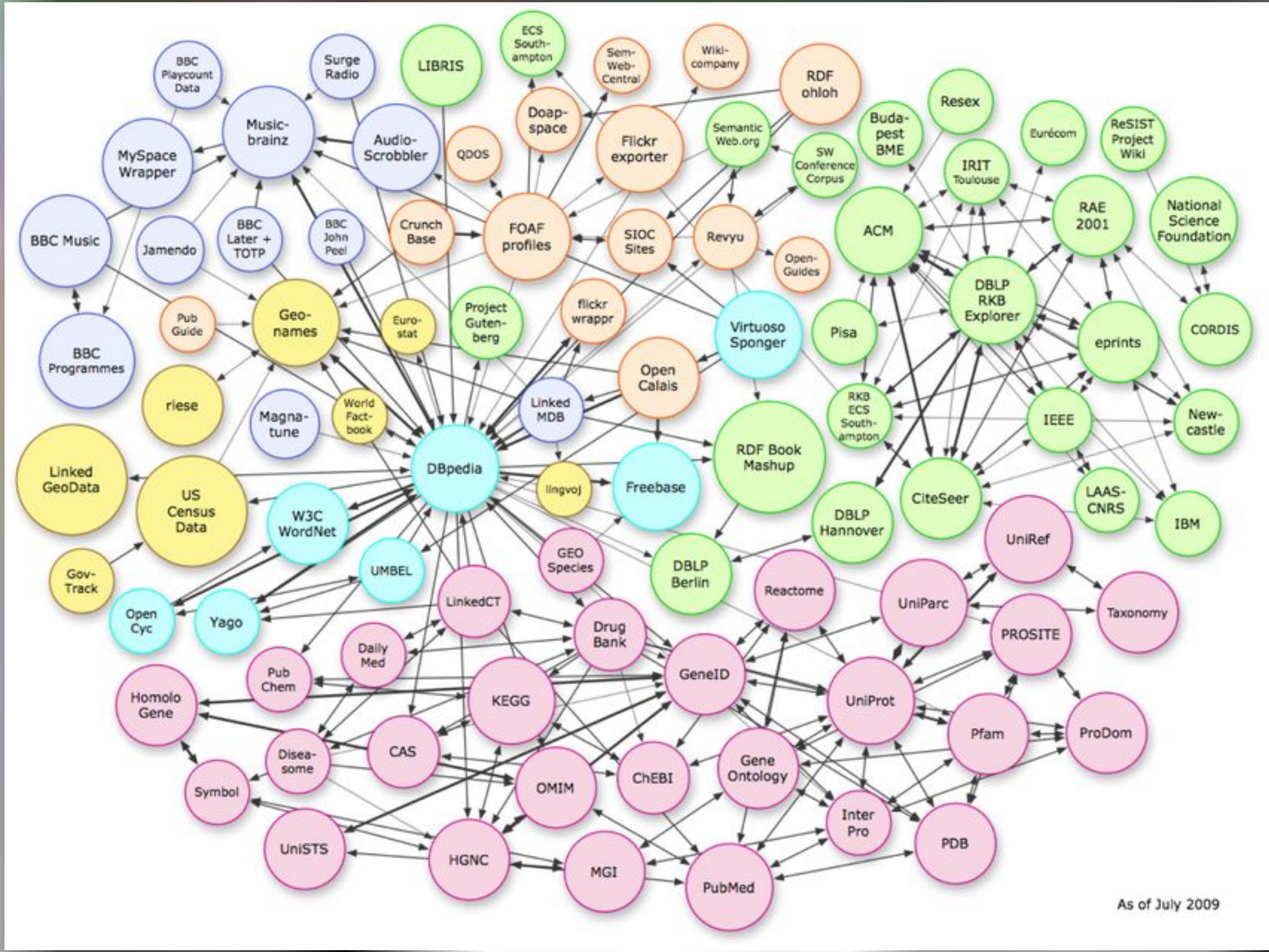
- Exchanging information: the transmission of complex information re-

sources among computers that allows us to distribute, interlink, and reconcile knowledge on a global scale.

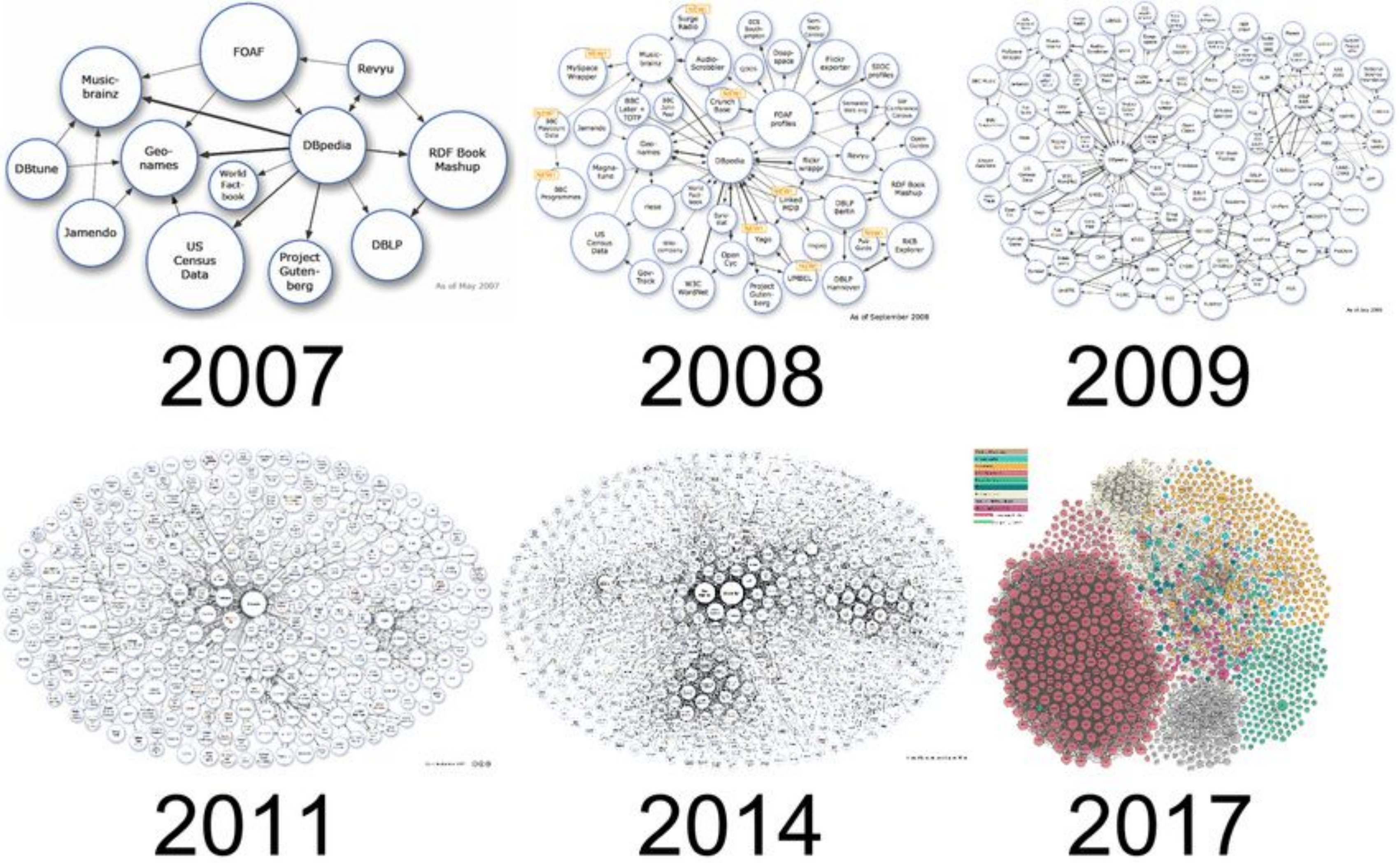
Linked Data

- Define **relationship** between data
- Enable **accessibility and integration** of different datasets on the Web
- Using a standard format for data and relationship: **Linked Data triples**


bestbuy.com
provides data about product offers (price, delivery time, etc.)
Apple Inc.
provides data about product features (CPU, RAM, HD, etc.)
DBPedia
provides general data about products or companies
UN environmental indicators
provides statistical data, such as carbon footprint, waste, energy usage, etc.
apple:MCB
owl:sameAs
bb:Macbook
gr:hasPriceSpecification
...
gr:hasCurrencyValue
1200
dct:title
MacBook
owl:sameAs
db:Apple_Computer
rdfs:seeAlso
un:Apple
db:developer
...
db:col
dbpedia:MacBook
un:co2footprint
un:location
...
un:US
Vocabularies used:
GoodRelations ... http://purl.org/goodrelations/v1#
Dublin Core ... http://purl.org/dc/terms/


BBC Playcount Data
Surge Radio
LIBRIS
ECS Southampton
Sem-Web-Central
Wiki-company
RDF ohloh
Music-brainz
Audio-Scrobbler
MySpace Wrapper
QDOS
Doap-space
Flickr exporter
Semantic Web.org
SW Conference Corpus
Buda-pest BME
Resex
Eurécom
ReSIST Project Wiki
IRIT Toulouse
BBC Music
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
Open-Guides
ACM
RAE 2001
National Science Foundation
BBC Programmes
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
DBLP RKB Explorer
eprints
CORDIS
riese
Magna-tune
World Fact-book
Linked MDB
Open Calais
RKB ECS South-ampton
IEEE
New-castle
Linked GeoData
US Census Data
DBpedia
lingvoj
Freebase
RDF Book Mashup
LAAS-CNRS
W3C WordNet
CiteSeer
DBLP Hannover
IBM
UniRef
Gov-Track
UMBEL
GEO Species
DBLP Berlin
Open Cyc
Yago
LinkedCT
Drug Bank
Reactome
UniParc
Taxonomy
PROSITE
Daily Med
Pub Chem
GeneID
Homolo Gene
KEGG
UniProt
Pfam
ProDom
CAS
Disea-some
OMIM
ChEBI
Gene Ontology
Symbol
Inter Pro
PDB
UniSTS
HGNC
MGI
PubMed
As of July 2009

It gets bigger



- Semantic Web allows machines to make **inferences**
- Linked data enables for
  - New relationships between data can be created automatically
  - Analysing the content of new data
  - Discovering potential inconsistencies

Vocabularies

- **Basic building blocks** for making Inference
- Helps to **resolve ambiguity** between different datasets, define new relationships, organize Knowledge.
- Ontology and Vocabulary are used interchangeably, generally Complex is referred as Ontology and simpler ones Vocabulary
- Eg: A bookseller may want to integrate data coming from different publishers. The data can be imported into a common RDF model, eg, by using converters to the publishers' databases. However, one database may use the term "author", whereas the other may use the term "creator". To make the integration complete, and extra definition should be added to the RDF data, describing the fact that the relationship described as "author" is the same as "creator".

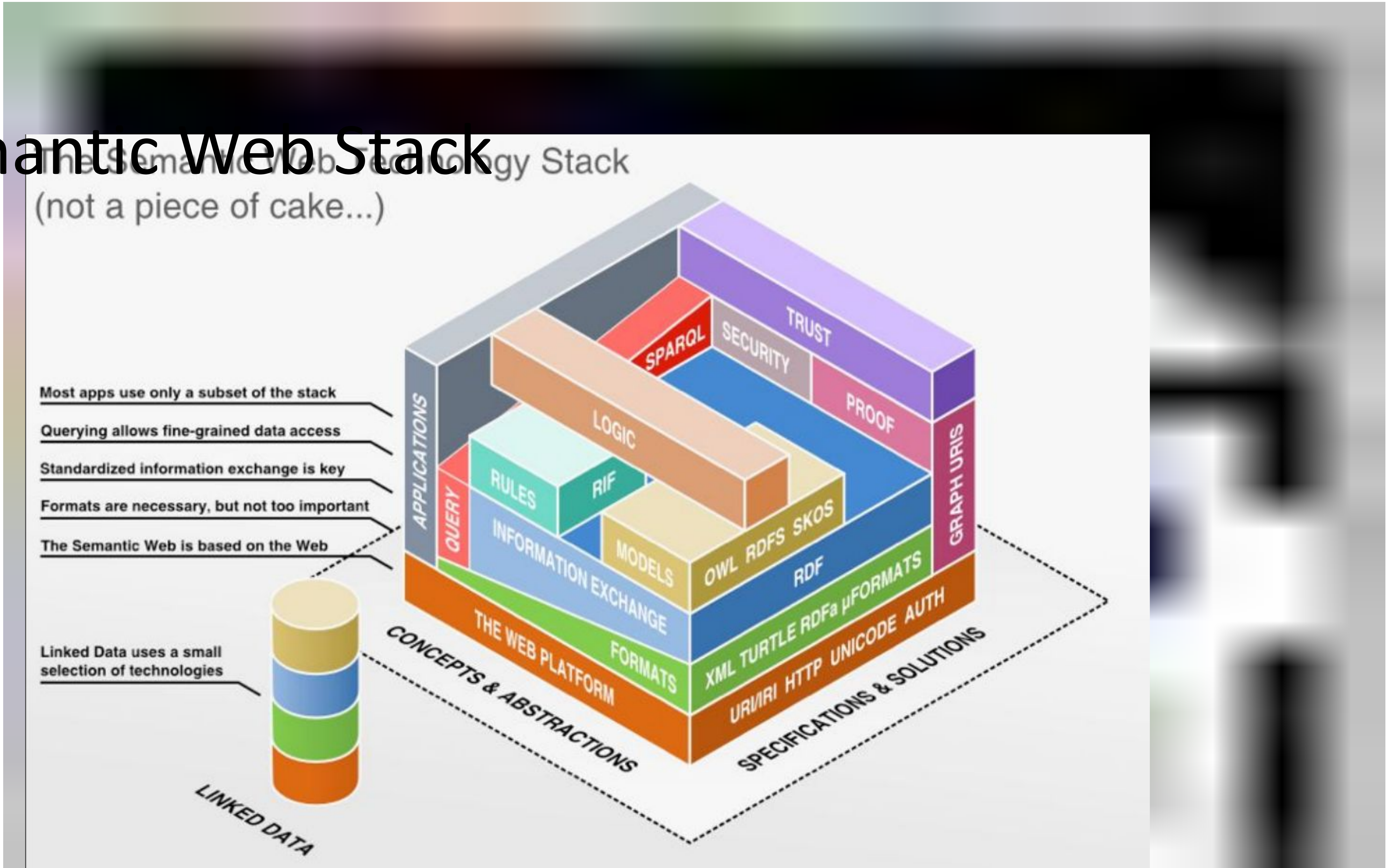
Semantic Web is…..

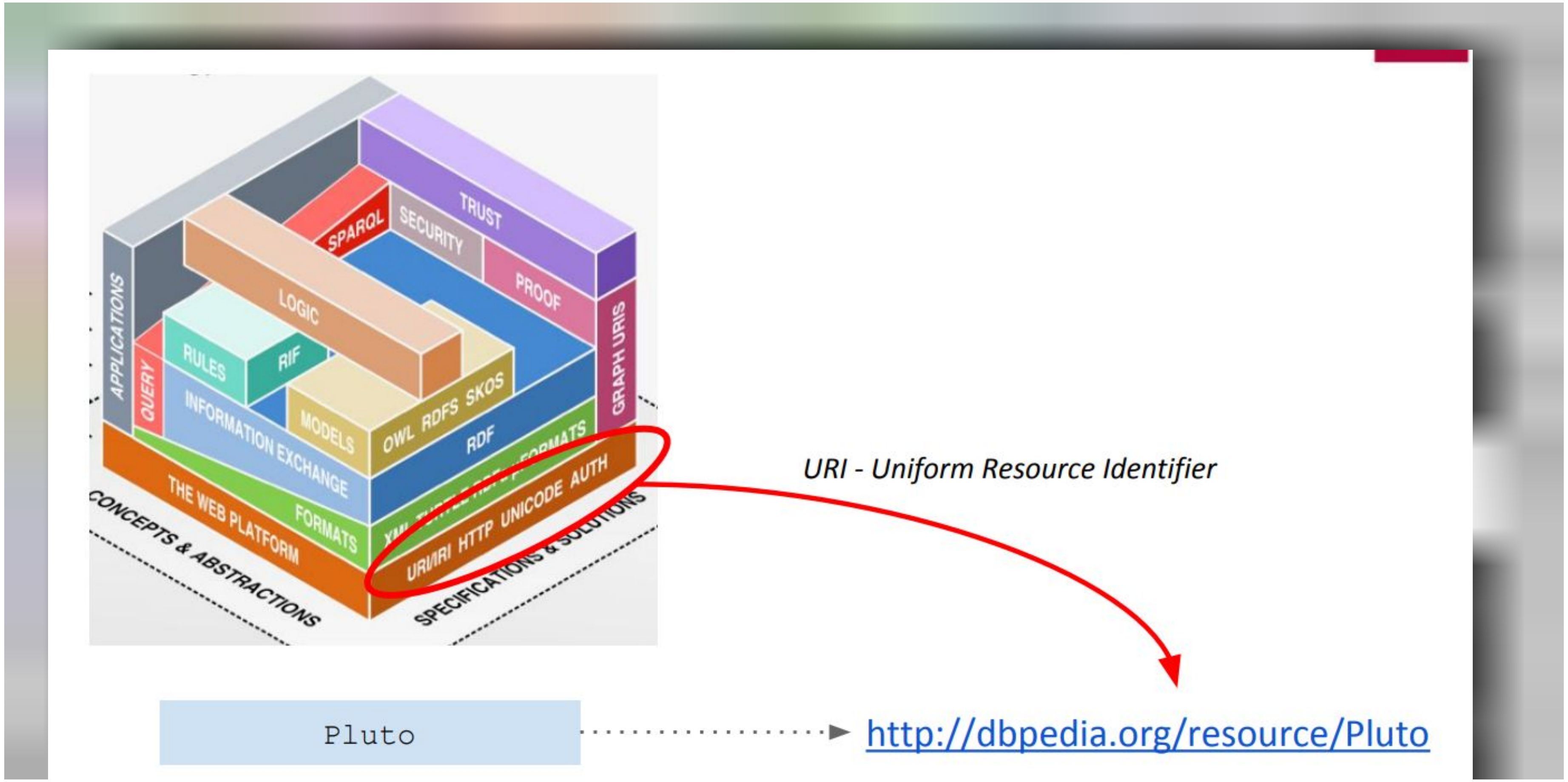
- Assembles data from many authoritative sources in ONE place
- Standardized data collections for improved access and collaboration
- Allows computers to understand the meaning of data, enabling humans to work more effectively

Characterised as

- Machine-interpretable, structured, interlinked, open access data repositories
- Globally edited adaptive information resources
- Unique web resource identifiers for every bit of information (e.g., each table data cell of a table has a unique identifier)

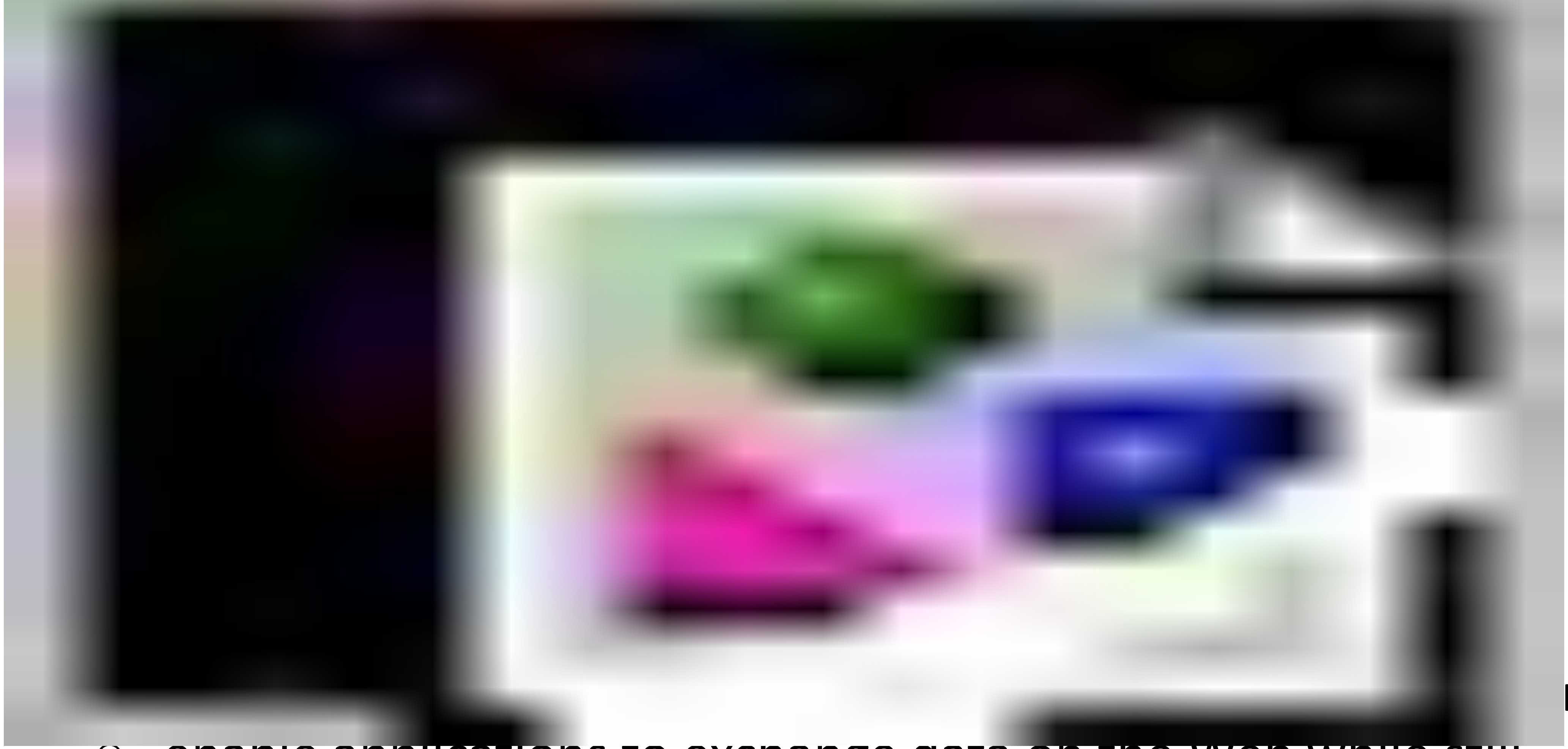
Semantic Web Stack





- URIs and Unicode.

URIs (uniform resource identifiers) are unique identifiers for resources of all types—from schemas to people. A major component of the base layer, URIs are metadata (eg:ISBNs (international standard book numberrs) or Social Security numbers in the context of the Web.



n.

- enable applications to exchange data on the Web while still preserving their original meaning.
- As opposed to HTML and XML, the main intention now is not to display documents correctly, but rather to allow for further processing and re-combination of the information contained in them.

- RDF is one of the main building bloksAll knowledge is stored in the form of RDF triples
- The RDF family further supports interoperability at the semantic level so that agents, can make logical inferences, based on metadata, to perform tasks.

- Knowledge about the entities is expressed usually in relation with other entities, hence a Model
  - Schema Definition

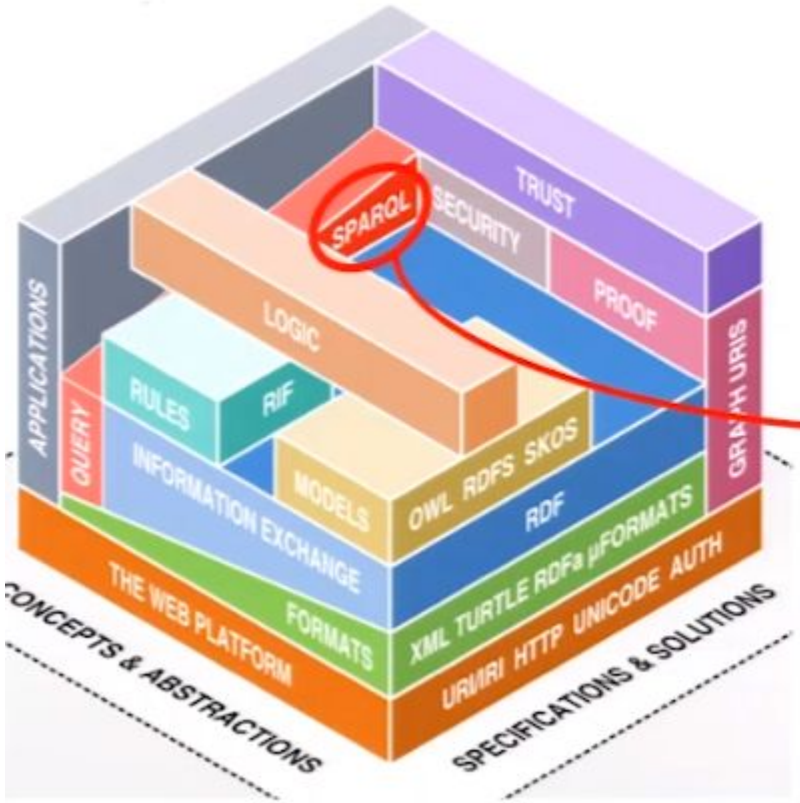
- Ontology vocabulary
  - formal description of knowledge as a set of concepts within a domain and the relationships that hold between them
  - contains simple descriptive to complex classificatory schemas are to be created and registered so that agents can intelligently interpret data, make inferences, and perform tasks.

- components such as individuals (instances of objects), classes, attributes and relations as well as restrictions, rules and axioms
- ontologies do not only introduce a sharable and reusable knowledge representation but can also add new knowledge about the domain
- ontology data model can be applied to a set of individual facts to create a knowledge graph
  - a collection of entities, where the types and the relationships between them are expressed by nodes and edges between these nodes

- provide users with the necessary structure to link one piece of information to other pieces of information on the Web of Linked Data.
- they specify common modeling representations of data from distributed and heterogeneous systems and databases, ontologies enable database interoperability, cross-database search and smooth knowledge management.

- OWL
    - ontologies using ontology languages such as the Web Ontology Language (OWL)
    - OWL is a semantic web computational logic-based language, designed to represent rich and complex knowledge about things and the relations between them

SPARQL



Look for all **space missions in the Solar System** which have become a **satellite** of their target

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX dbc: <http://dbpedia.org/resource/Category:>

SELECT distinct ?s ?o
FROM <http://dbpedia.org/>
WHERE{
?s dcterms:subject/skos:broader*
    dbc:Discovery_and_exploration_of_the_Solar_System ;
    dbp:satelliteOf ?o .
}
```

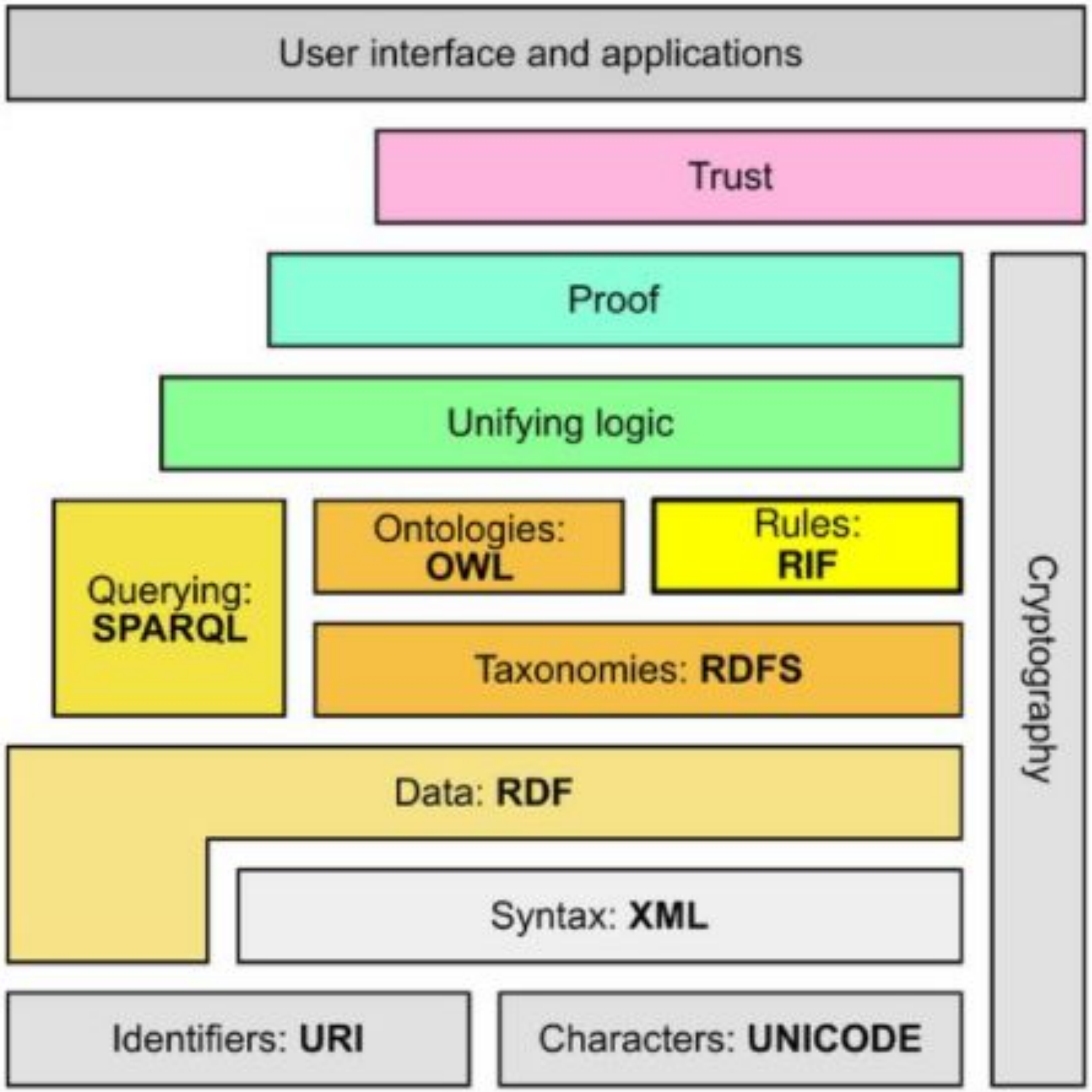
- XML + NS + XMLschema.

Extensible Markup Language (XML) and more recently XML schemas facilitate the creation, use, and syntactic interoperability of metadata vocabularies.

NS (namespaces), which are identified via URIs, secure semantic interoperability among metadata vocabularies.

Semantic Web Stack

- Provides architecture for the Semantic web and it deals in relationships related to the components.
- provides the functions to be used in the components and provides the content structure.
- Syntax of the XML is used to provide with the technologies and it provides the process to be made standardized.
-



- RDF and RDFschema.

The RDF family further supports interoperability at the semantic level. RDF developments comprise the base Web language, so that agents, like Pete’s and Lucy’s discussed above, can make logical inferences, based on metadata, to perform tasks.

- Logic.

We make logical inferences in our performance of daily tasks. For example: If N denotes new unread email in an email inbox, then if an N appears by a particular message, the message is new unread email. This inference is based on evidence provided by the letter N. The logic layer of the Semantic Web works on this basic principle through First Order Predicate Logic. An agent can derive a logical conclusion (or reason) in the process of completing a task based on what are essentially “facts” rendered from semantically encoded metadata. Other types of logic may also be applicable in the Semantic Web.

Proof and Trust. The last two horizontal layers build off of the logic layer, proof being a validation of the “evidence” stemming from the inferential logic activity and trust relating to the integrity of the proof, which may be traced back down through the other layers in Berners-Lee’s diagram. The functionality of these two layers is highly dependent on creation of accurate and trustworthy metadata.

■ Digital signature. Digital signatures run horizontal to the RDF family up through the proof layer and support the notion of trust. Developments in the area of digital signatures are progressing, and could eventually help validate the integrity of metadata that an agent will use for reasoning and task completion

N-Triples and Turtle

What are the challenges faced by the technology?

- Vastness: this includes the large group of pages that is being accessed by the users using the existing technology.
- This consists of any automated system that is good in reasoning and deals with the very high inputs.
- Vagueness: it occurs due to the queries that are being provided by the content providers.
- If the query terms are matched then the knowledge can be combined together to find the knowledge.
- Uncertainty: includes uncertain value that can provide the correspondence using the different probability.
-

- Inconsitency: is the very big challenge that provides logical contradictions between the ontologies.
- It combines the resources to answer the questions that are being raised by the theories and sources

Resource Description Framework (RDF)

- RDF is a language for the representation of resources:
  - A resource can be anything
  - Within the context of the Web relevance is given to Web resources, i.e. anything that can be located via a URL (Uniform Resource Locator)
  - The basic building block is the statement (or triple)

- One of the main applications: data integration

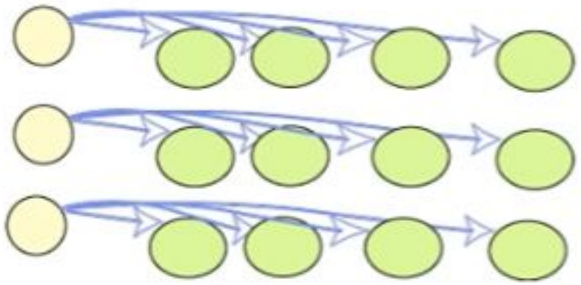
RDF Overview

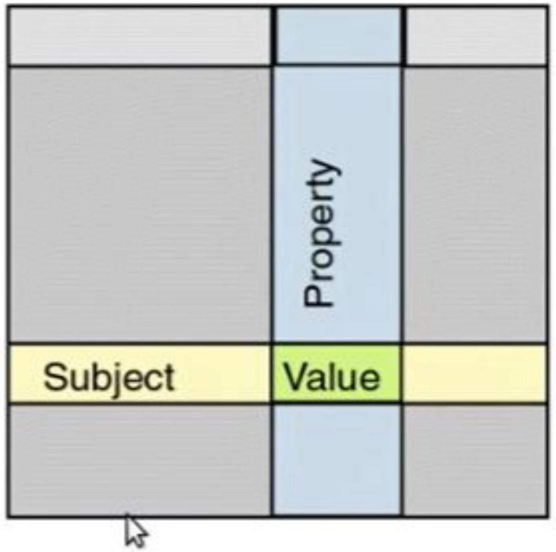
- W3C Recommendation since 1998
- RDF is a data model
- Originally used for metadata for web resources, then generalized for other applications
- Universal, machine readable exchange format
- Data structured in graphs (vertices, edges)

RD2RDF

- The adaptation of the relational model to the Web give rise to RDF
- From *tuples* to *triples*

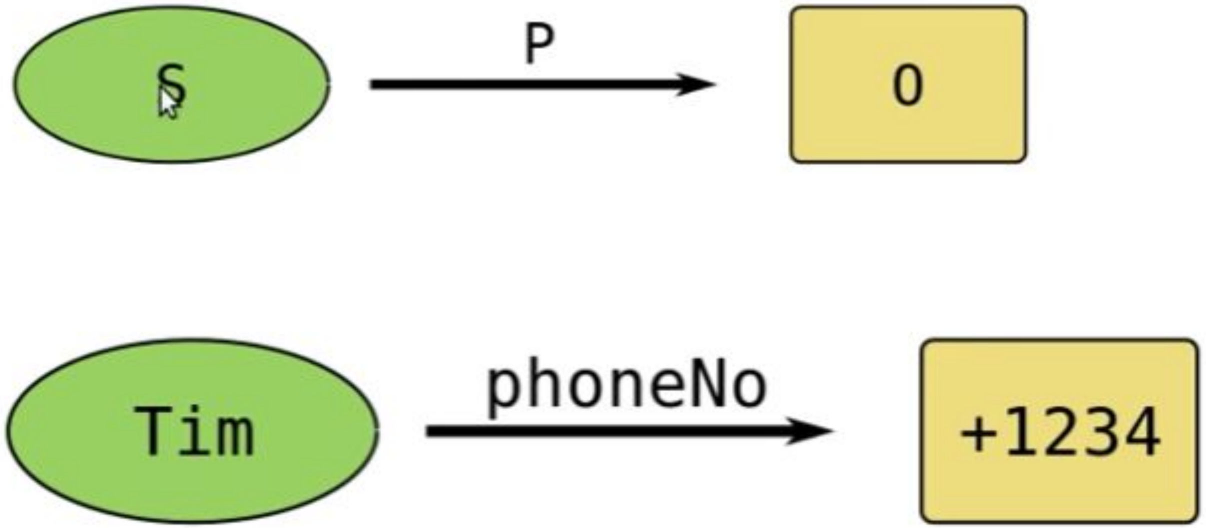
- Any relational data can be represented as triples
  - Row Key --> Subject
  - Column --> Property
  - Value --> Value



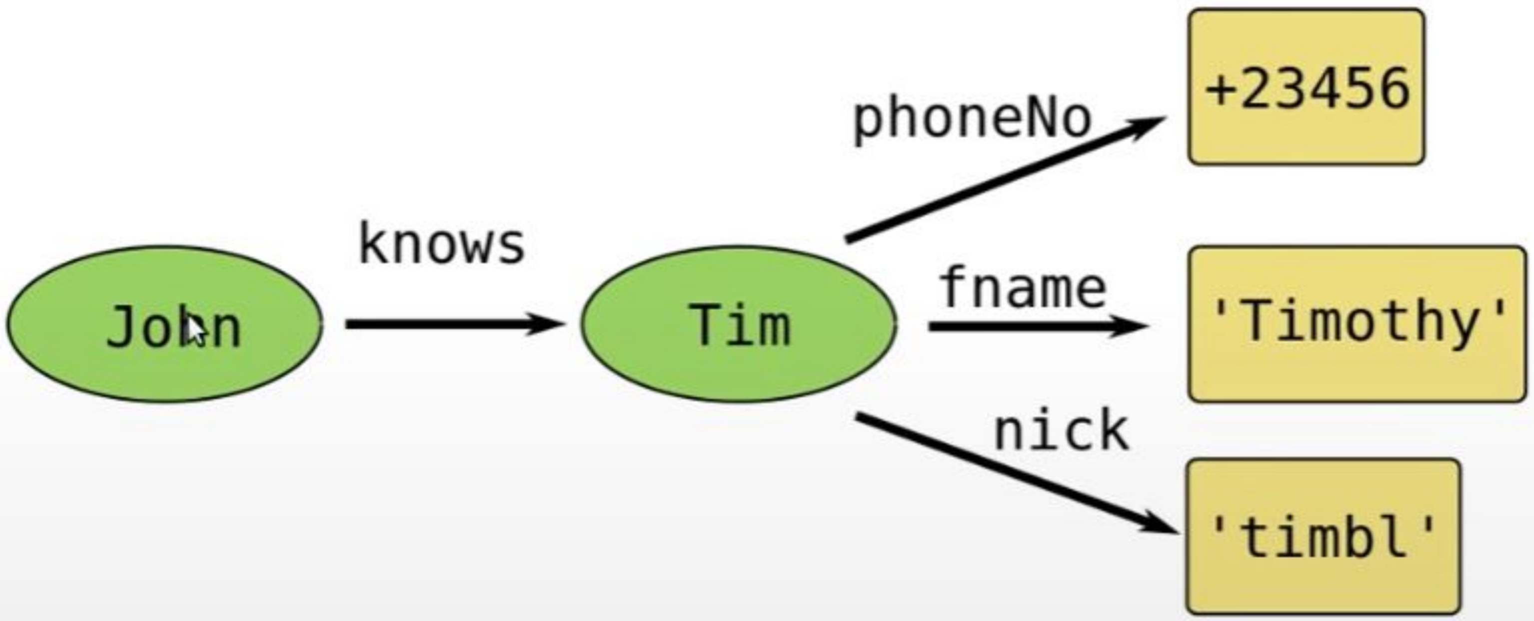


Triples

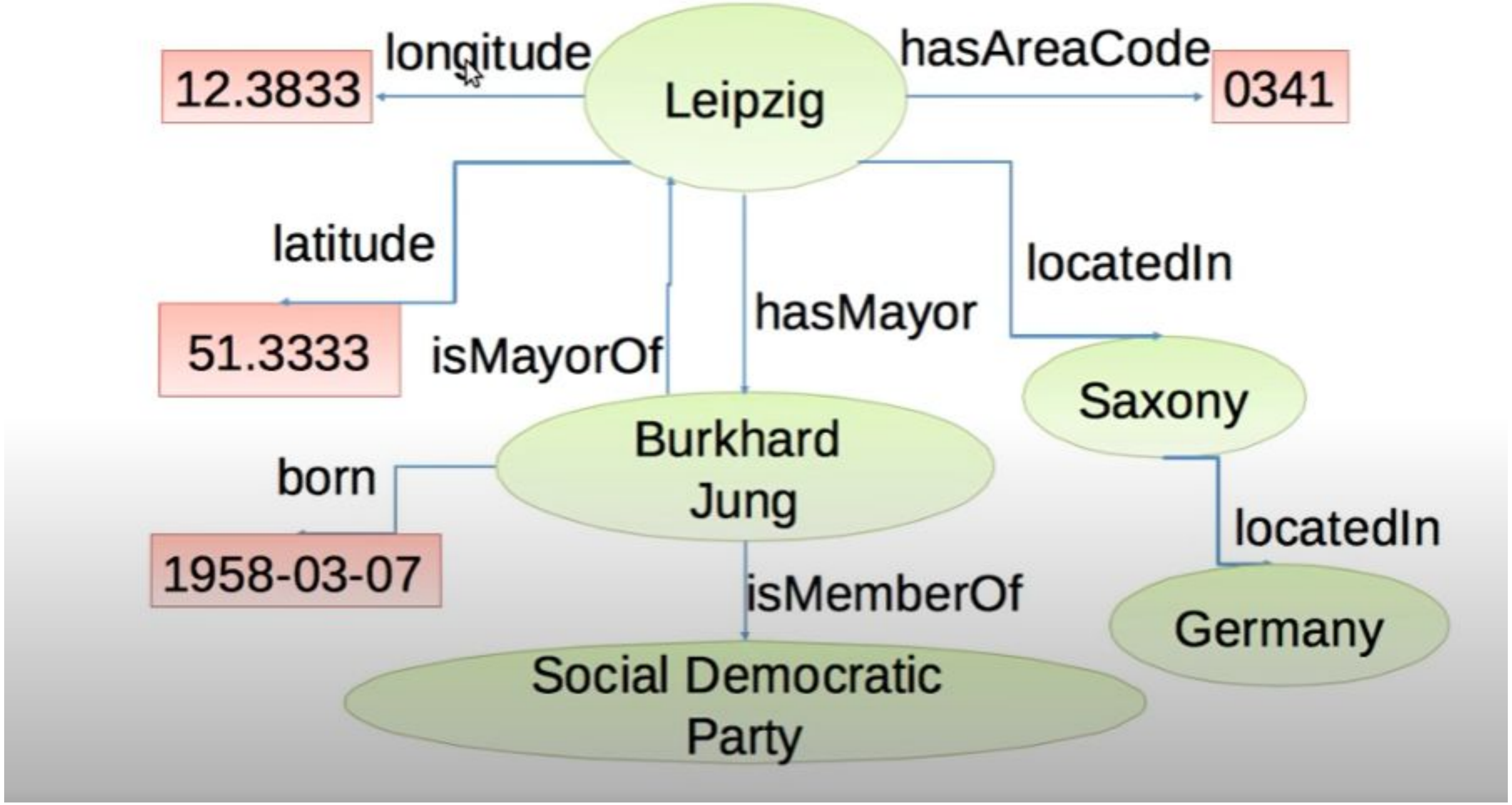
Triples are the statements about things (resources), using URIs and Literal values



Graph



Example of an RDF graph



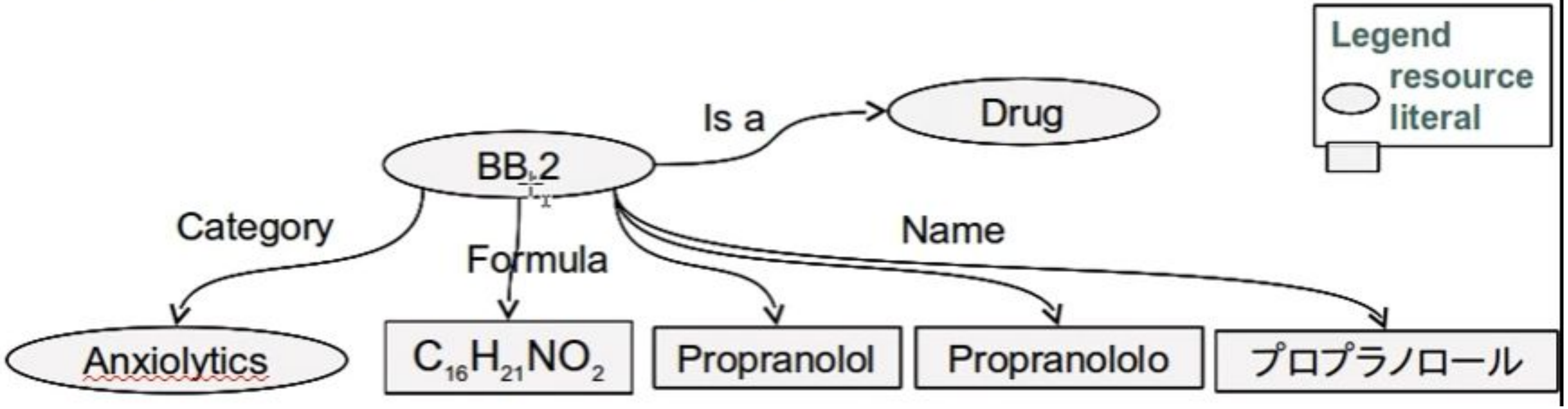
Representing RD in RDF (almost)

- E.g., drug data

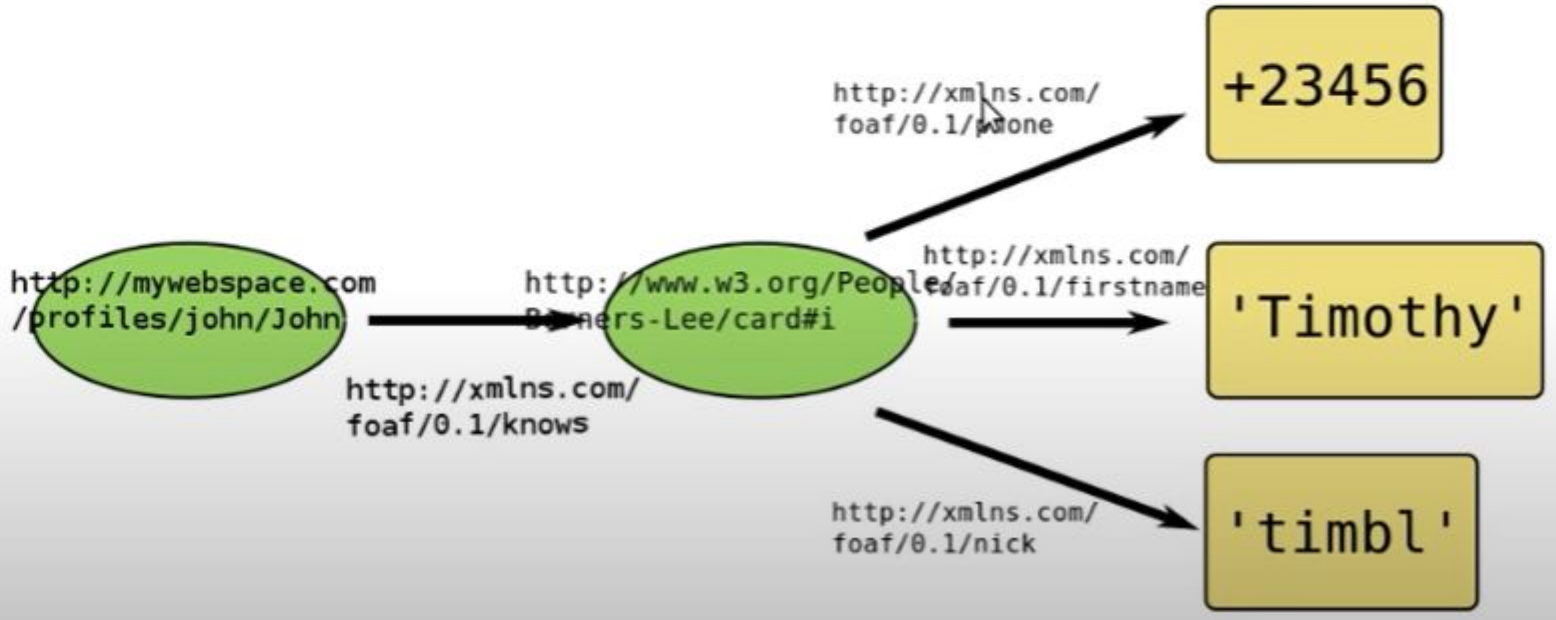
Drug	Category	Formula
BB.2	Anxiolytics	$C_{16}H_{21}NO_2$

Drug	Name
BB.2	Propranolol
BB.2	Propranololo
BB.2	プロプラノロール

- Represented in RDF (almost)

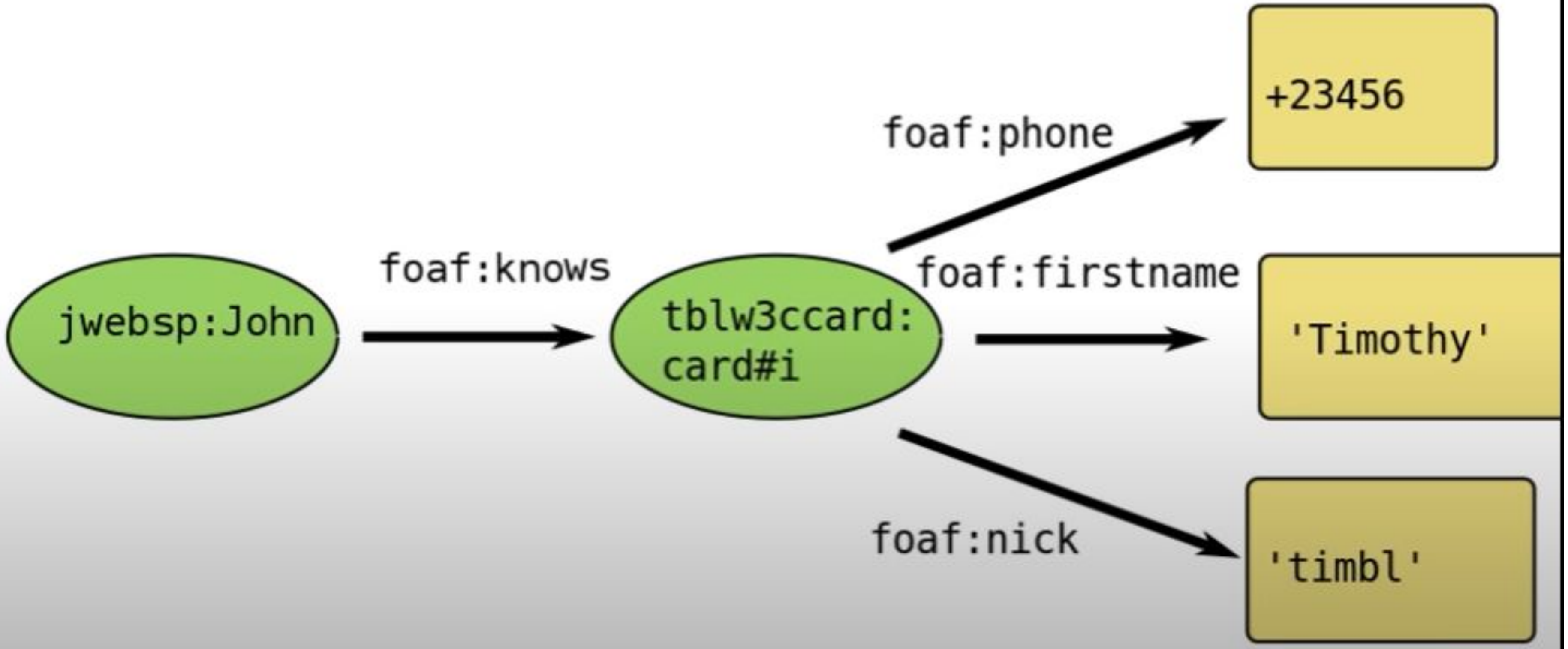


Graphs with URIs



Prefixes

```
@prefix wbsp: <http://mywebspace.com/profiles/john/>.
@prefix tblw3ccard: <http://www.w3.org/People/Berners-Lee/>
@prefix foaf: <http://xmlns.com/foaf/0.1/>
```



Parts of the RDF graph

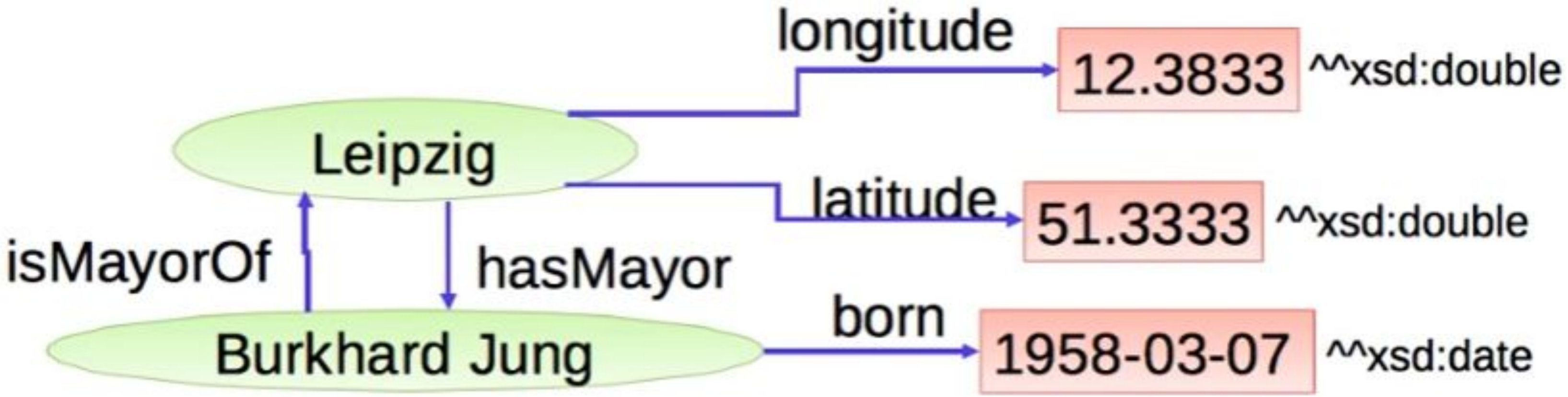
- **URI**s - used to reference resources unambiguously
- **Literals** – used to describe data values with no clear identity like "100 km/h”
- **Blank nodes** facilitate existential quantification for an individual with certain properties without naming it

RDF Triple

- Allowed assignments for components of an RDF triple
  - Subject: URI or blank node
  - Predicate: URI (a.k.a. property)
  - Object: URI, blank node or literal
- Node and edge labels should be unambiguous, so that the original graph is reconstructable from triple list

Literals

- Used to model data values
- Representation as strings
- Interpretation through datatype
- Literals without datatype are treated as strings
- Literals may never be the origin of a node of an RDF graph



Most popular RDF formats

- **N-Triples** - a text format focusing on simple parsing
- **Turtle** - a text format focusing on human readability
- **Notation 3** - a text format with advanced features beyond RDF
- **RDF/XML** - the official XML serialization of RDF
- **RDF/JSON** - a proposal for serializing RDF in JSON
- **JSON-LD** - another proposal for expressing RDF in JSON
- **RDFa** - a mechanism for embedding RDFa in (X)HTML

Turtl

Turtle Syntax

- Language to serialize RDF Triples to strings
- Turtle - Terse RDF Triple Language
- URIs in angle brackets <http://dbpedia.org/resource/Leipzig>
- Literals in quotes "Leipzig"@de, "51.333332"^^xsd:float
- A triple is separated by a dot
- White spaces and line breaks are ignored outside of identifiers
- Status: W3C Candidate Recommendation 19 February 2013, http://www.w3.org/TR/turtle/

EXAMPLE:

```
<http://dbpedia.org/resource/Leipzig>
<http://www.w3.org/2000/01/rdf-schema#label>
  "Leipzig"@de .
```

- In Turtle one can use abbreviations.

Syntax: @prefix abbr ':' <URI> .

E.g. @prefix dbr: <http://dbpedia.org/resource/> .

- One can transform the previous example to:

```
@prefix dbr: <http://dbpedia.org/resource/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema> .
dbr:Leipzig rdfs:label "Leipzig"@de .
```

- Triples with the same subject can be grouped together:

```
@prefix rdf:
...
@prefix geo:
dbr:Leipzig dbp:hasMayor dbr:Burkhard_Jung ;
  rdfs:label "Leipzig"@de ;
  geo:lat "51.333332"^^xsd:float ;
  geo:long "12.383333"^^xsd:float .
```

- Even triples with the same subject and predicate can be grouped together

```
@prefix dbr: .
@prefix dbp: .
dbr:Leipzig dbp:locatedIn dbr:Saxony, dbr:Germany;
      dbp:hasMayor dbr:Burkhard_Jung .
```

Example

https://www.oreilly.com/library/view/learning-sparql-2nd/9781449371449/ch04.html

RDFS and OWL

- What is a vocabulary?
  - In semantic web development, a vocabulary is a set of terms stored using a standard format that people can reuse
  - A vocabulary of property names typically has its own namespace to make it easier to use it with other sets of data
- We can use existing vocabularies of properties (Where do we find them?)
- Vocabularies are usually stored using the RDF Schema and OWL standards

- Its full title is "RDF Vocabulary Description Language: RDF Schema"
- It gives people a way to describe vocabularies
- You describe these vocabularies with RDF
- This metadata is just as accessible to your SPARQL queries as the data itself
- Dublin Core is a standard set of basic metadata terms
- Here are a few of the triples from the RDF Schema vocabulary description of the Dublin Core vocabulary

```
@prefix dc:   <http://purl.org/dc/elements/1.1/> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

dc:creator
    rdf:type rdf:Property ;
    # a rdf:Property ;
    rdfs:comment "An entity primarily responsible for making the resource."@en-US ;
    rdfs:label "Creator"@en-US .
```

- RDF Schema is itself a vocabulary with a schema whose triples declare facts
  - For example, that *rdfs* : *label* and *rdfs* : *comment* are properties
- Just like the Dublin Core schema excerpt we saw before declares that *dc* : *creator* is a property
- RDF Schema lets you define new classes of resources
- For example, the following shows how one might declare ab:Musician and ab:MusicalInstrument classes for address book data

```
@prefix ab:   <http://learningsparql.com/ns/addressbook#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

ab:Musician
     rdf:type rdfs:Class ;
     rdfs:label "Musician" ;
     rdfs:comment "Someone who plays a musical instrument" .

ab:MusicalInstrument
     a rdfs:Class ;
     rdfs:label "musical instrument" .
```

- There's a lot more metadata that we can assign when we declare a class
  - for example that it's subclass of another one
- In ex044.ttl the author declares an $ab : playsInstrument$ property
- The $rdfs : domain$ property means that if I use this $ab : playsInstrument$ property in a triple, then the subject of the triple is an $ab : Musician$
- The $rdfs : range$ property means that the object of such a triple is an $ab : MusicalInstrument$

```
@prefix ab: <http://learningsparql.com/ns/addressbook#> .

ab:i0432 ab:firstName     "Richard" ;
       ab:lastName       "Mutt" ;
       ab:postalCode     "49345" ;
       ab:city           "Springfield" ;
       ab:homeTel        "(229) 276-5135" ;
       ab:streetAddress  "32 Main St." ;
       ab:region         "Connecticut" ;
       ab:email          "richard49@hotmail.com" ;
       ab:playsInstrument ab:vacuumCleaner .
```

- Once we've added the new triple shown at the end of *ex*045.*ttl*, an RDFS-aware SPARQL processor knows that Richard Mutt (or, more precisely, resource ab:i0432) is now a member of the class *ab* : *Musician*
- This is because *ab* : *playsInstrument* has a domain of *ab* : *Musician*.
- Because *ab* : *playsInstrument* has a range of *ab* : *MusicalInstrument*, *ab* : *vacuumCleaner* is now a member of the *ab* : *MusicalInstrument* class, even if it never was before
- **Using semantic web standards, by adding one property to the metadata about the existing resource ab:i0432, that resource becomes a member of a class that it wasn't a member of before**

- The ability of RDF resources to become members of classes based on their data values has made semantic web technology popular in areas such as medical research and intelligence agencies
- Researchers can accumulate data with little apparent structure and then see what structure turns out to be there
- That is, which resources turn out to be members of which classes, and what their relationships are

- RDFS lets you define classes as subclasses of other ones, and (unlike object-oriented systems) properties as subproperties of other ones
- This broadens the possibilities for how you can use SPARQL to retrieve information
- For example, if we say that $ab:Musician$ was a subclass of $foaf:Person$ and then queried an RDFS-aware processor for the phone numbers of all the $foaf:Person$ instances in a dataset that included the $ex044.ttl$ and $ex045.ttl$ data, the processor would give me Richard's phone number, because by being in the class of musicians Richard is also in the class of persons

OWL

- The W3C's Web Ontology Language, abbreviated as OWL
- OWL builds on RDFS to let you define ontologies.
- What is an Ontology?
  - Ontologies are formal definitions of vocabularies that allow you to define complex structures as well as new relationships between your vocabulary terms and between members of the classes that you define
  - Ontologies often describe very specific domains such as scientific research areas so that scientists from different institutions can more easily share data
- An ontology defined with OWL is a collection of triples

- Without defining a large, complex ontology, many semantic web developers use just a few classes and properties from OWL to add metadata to their triples.
- For example, how much information do you think the following dataset has about resource ab:i9771, Cindy Marshall? (ex046.ttl)
- Notice:
  - the ab:spouse property is defined in this dataset as being *symmetric*
  - the ab:patient property is the *inverse* of the ab:doctor property
  - Now we know who Cindy's spouse and doctor are, even though these facts are not explicitly included in the dataset

- Without defining a large, complex ontology, many semantic web developers use just a few classes and properties from OWL to add metadata to their triples.
- For example, how much information do you think the following dataset has about resource ab:i9771, Cindy Marshall? (ex046.ttl)
- Notice:
  - the ab:spouse property is defined in this dataset as being *symmetric*
  - the ab:patient property is the *inverse* of the ab:doctor property
  - Now we know who Cindy's spouse and doctor are, even though these facts are not explicitly included in the dataset

```
@prefix ab:   <http://learningsparql.com/ns/addressbook#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:  <http://www.w3.org/2002/07/owl#> .

ab:i0432
  ab:firstName "Richard" ;
  ab:lastName  "Mutt" ;
  ab:spouse    ab:i9771 .

ab:i8301
  ab:firstName "Craig" ;
  ab:lastName  "Ellis" ;
  ab:patient   ab:i9771 .

ab:i9771
  ab:firstName "Cindy" ;
  ab:lastName  "Marshall" .

ab:spouse
  rdf:type owl:SymmetricProperty ;
  rdfs:comment "Identifies someone's spouse" .

ab:patient
  rdf:type rdf:Property ;
  rdfs:comment "Identifies a doctor's patient" .

ab:doctor
  rdf:type rdf:Property ;
  rdfs:comment "Identifies a doctor treating the named resource" ;
  owl:inverseOf ab:patient .
```

Download Protege

https://protege.stanford.edu/products.php#desktop-protege

- Of all the W3C semantic web standards, OWL is the key one for putting the "semantic" in "semantic web"
- The term "semantics" is sometimes defined as the meaning behind words, and those who doubt the value of semantic web technology like to question the viability of storing all the meaning of a word in a machine-readable way
- As we saw, though, we don't need to store all the meaning of a word to add value to a given set of data
- For example, simply knowing that "spouse" is a symmetric term made it possible to find out the identity of Cindy's spouse, even though this fact was not part of the dataset

References