# National Judicial Academy, Bhopal

Welcome to a session on

***Data and Information Management***

Prof. Madhukar Dayal
IIM Indore

# Request...

Please turn your mobile / smart phones to silent mode for the duration of this class.

THANKS !!

# Briefly about me...

- Selected in UPSC's SCRA 1987 exam (after class XII).
- Mechanical Engineering: 4 years (at Indian Railways Service of Mechanical Engineers, Jamalpur, Feb-88 to Feb-92).
- Joined IR as Gazetted officer (IRSME).
- Served from 1992 to 2012, 20+ years in.
- VR in 2012 (after 20+ years of service).
- Fellow (Computers & Information Systems), IIM Ahmedabad.
- Faculty at IIM Indore since ~4 years.

# Briefly about me...

Teach:

- Spreadsheet Modeling.
- Information Technology and Systems for Managers (an application challenges oriented course covering BPR, ERP, CRM, SCM, Social media).
- Modern Computing Applications for Businesses.
- DBMS & OLTP (technical, PhD level).
- Computer Networking (technical, PhD level).

Research:

- High Performance Compute Cluster (HPCC) algorithms and applications.
- Advanced IT systems (selection, implementation and adoption challenges).
- Big Data (applications and policy).

# Data…

What is data ?

Where do we get data from ?

How do we get this data ?

Why do we collect data ? What do we do with it ?

Is data useful ?

# Data…

Example data of a class of students: Registration No, Name, Age, Gender, State, Education, University.

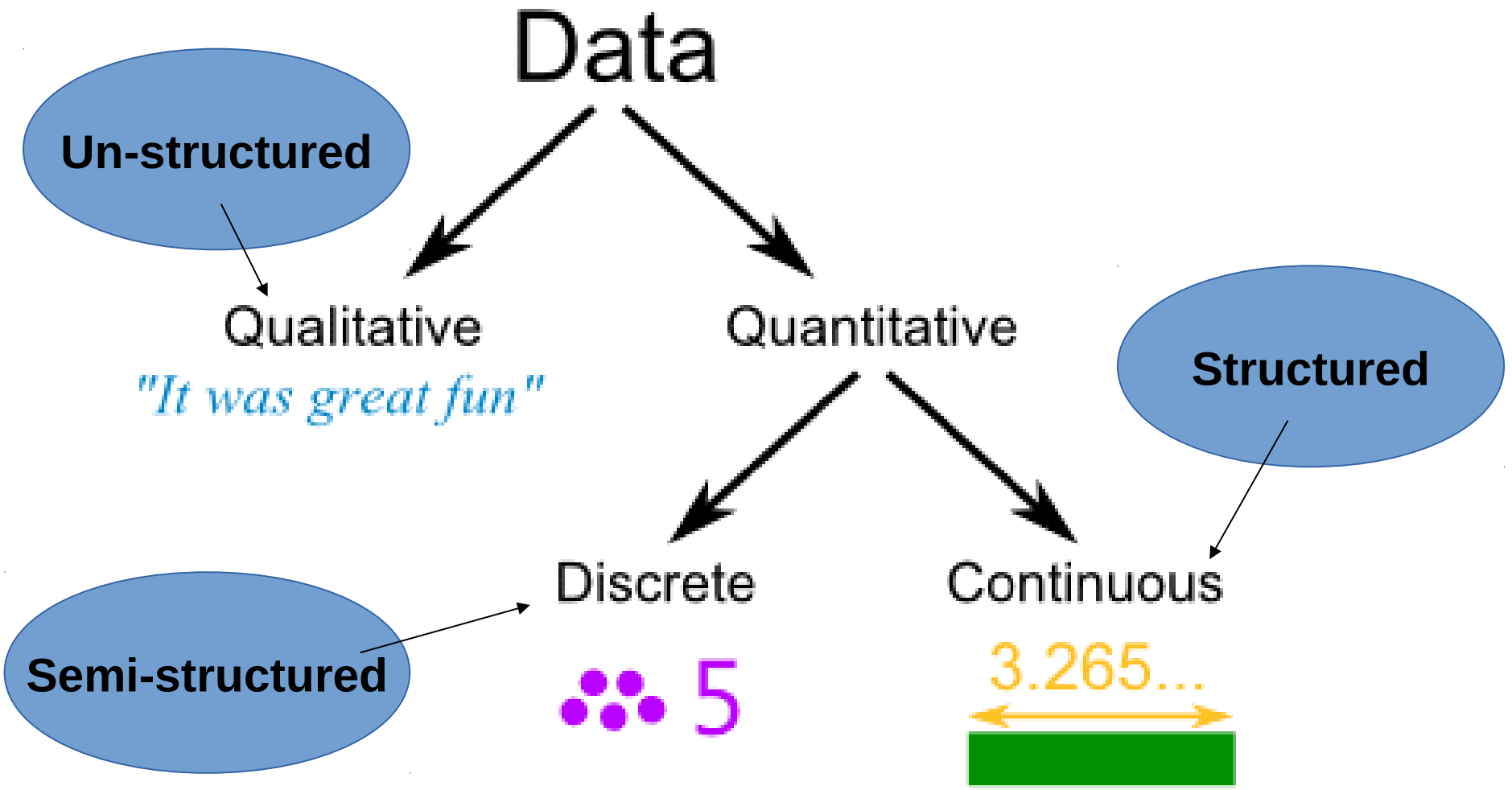
Data is voluminous. Not of much use.

Its "aggregation" (summarization) is useful for us.

When properly processed, data gives us "Information". Information is useful.

What are the types of data we see ?

# Types of data...




Types of data….

# Data…

Data leads to information.

Data → Information.

What do we do with information ?

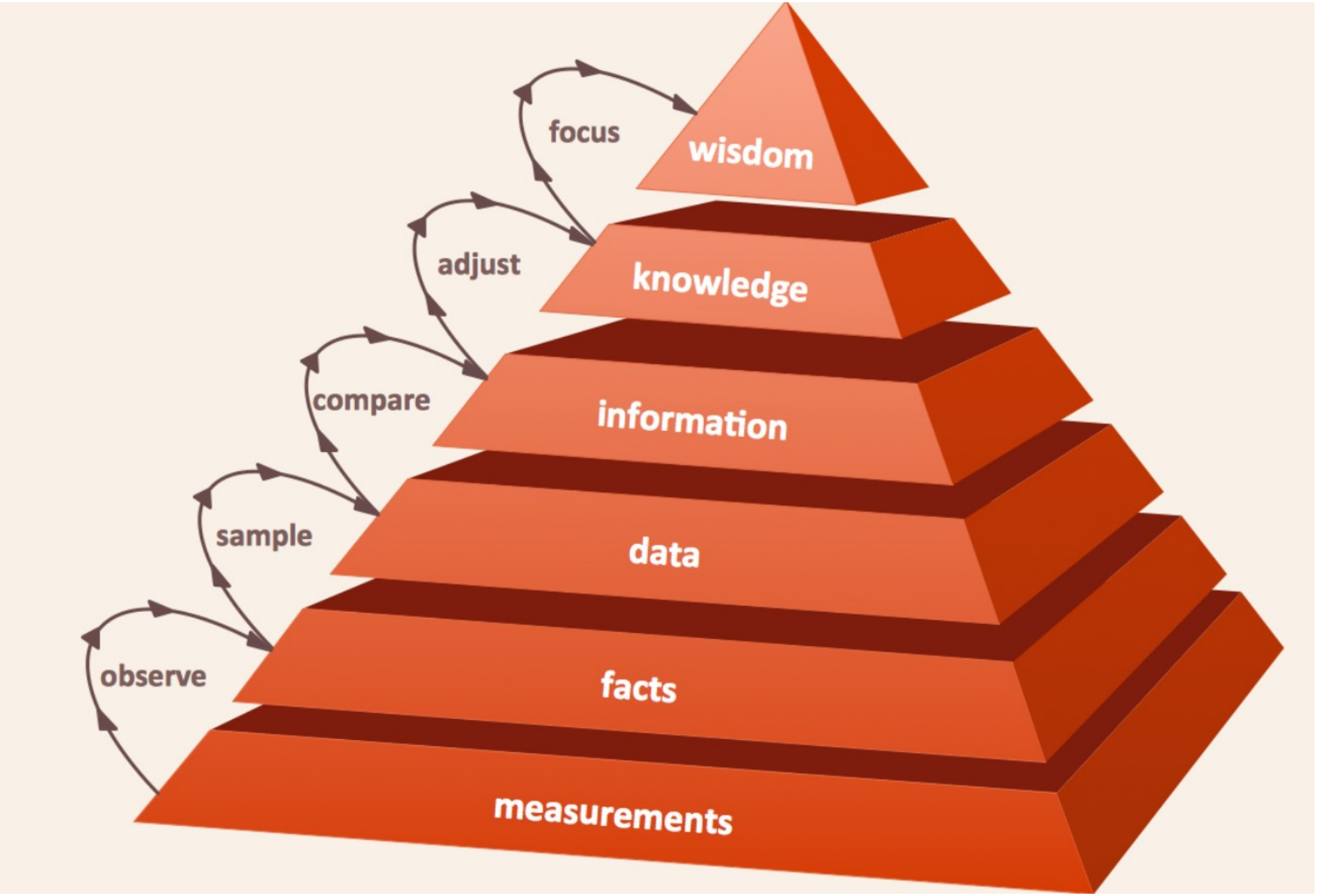
# Data…

Related and relevant information when properly compiled, analysed, interpreted, integrated, and presented becomes "Knowledge".

Accumulation of "Knowledge" by humans leads to "Wisdom".

So, summarising…

# Purpose of data




Data (analysis and interpretation) leads to information.
Information (collection and aggregation) leads to knowledge.
Knowledge (integration and assimilation) leads to wisdom.

# Volume of data…

Structured data: easy to collect, store, analyse.

Semi-structured data: difficult to analyse.

Un-structured data: very difficult !

Today, data comes with a large…
…volume.
…variety,
…velocity.
Known as: *Big Data*.

# Volume of data…

- Data collected from:
- Your mobile phones – where you go, how long you stay, where you pay, what you buy, etc.

- Your Internet usage: which website, which page, where clicked, how long stayed, what purchased, email sent to whom, etc.

- Sensors – weather (temp, wind velocity) everywhere on Earth.
- Sensors – fitted on birds, animals.
- Nano-sensors – sprayed on ants, where they go, what they do.

# Big Data Storymap

EMC²

## CURRENT STATE

### BI/DW CHALLENGES

- Batch oriented, inflexible, high latency
- Brittle & labor intensive
- Performance & scalability challenged
- Aggregated, silo'd, structured data

**BUSINESS CHALLENGES**

- Rigid architecture impedes immediately available business opportunities
- Retrospective reporting doesn't guide business decisions
- Social, mobile, or machine-generated data insights not available
- Aggregation destroys valuable insights

TRADITIONAL TOWN

DATA WAREHOUSE

ETL

BUSINESS INTELLIGENCE

BUSINESS MONITORING

BUSINESS INSIGHTS

BUSINESS OPTIMIZATION

DATA MONETIZATION

BUSINESS METAMORPHOSIS

BUSINESS-MODEL-MATURATION

## FUTURE STATE

- Big Data analytics continuously drive business innovations
- Real-time insights optimize operations and boost profits
- Unprecedented customer insights improve products and user experience

Value Creation City

BIG DATA VALUE CHAIN "MILE"

PRODUCT DEVELOPMENT

PROCUREMENT

MANUFACTURING

FINANCE

SALES & MARKETING

DISTRIBUTION

CUSTOMER EXPERIENCE

Continuously Refine Analytic Models

Publish Insights in Apps

Operationalize Big Data

Acquire & Prepare Data

Capture new Business Requirements

Measure Decision Effectiveness

Implement Big Data infrastructure

Incorporate insights into apps

Define data strategy

Build analytic models

Identify targeted business initiatives

Determine required insights

VALUE JOURNEY

CONDUCT DATA SCIENTIST TRAINING & CERTIFICATION

VOLUME

VARIETY

VELOCITY

Records Terabytes

Transactions Petabytes

Tables, Files

Structured

Unstructured

Semi-Structured

Real-Time

Near Real-Time

Batch

BIG DATA RIVER

DEEP DATA

DIVERSE DATA

FAST DATA

Risk of Not Preparing for BIG DATA...

- Competitors Innovate Faster
- Profits & Margins Degenerate
- Market Share Declines
- Missed Business Opportunities

Big Data concepts by William Schmarzo
Visualization concept by Mark A. Lawson and Glenn Steinhandler

# Volume of data...

- Short video: changes in mankind due to technology today !!

# Volume of data...

- Rise of new engineering discipline: “Data Science”.
- New jobs like “Data Scientist”.
- Performing “Data warehousing” and “Data Mining”.
- Analysing: “Knowledge Discovery in Databases” (KDD).
- Using: High Performance Compute Clusters (super computers).

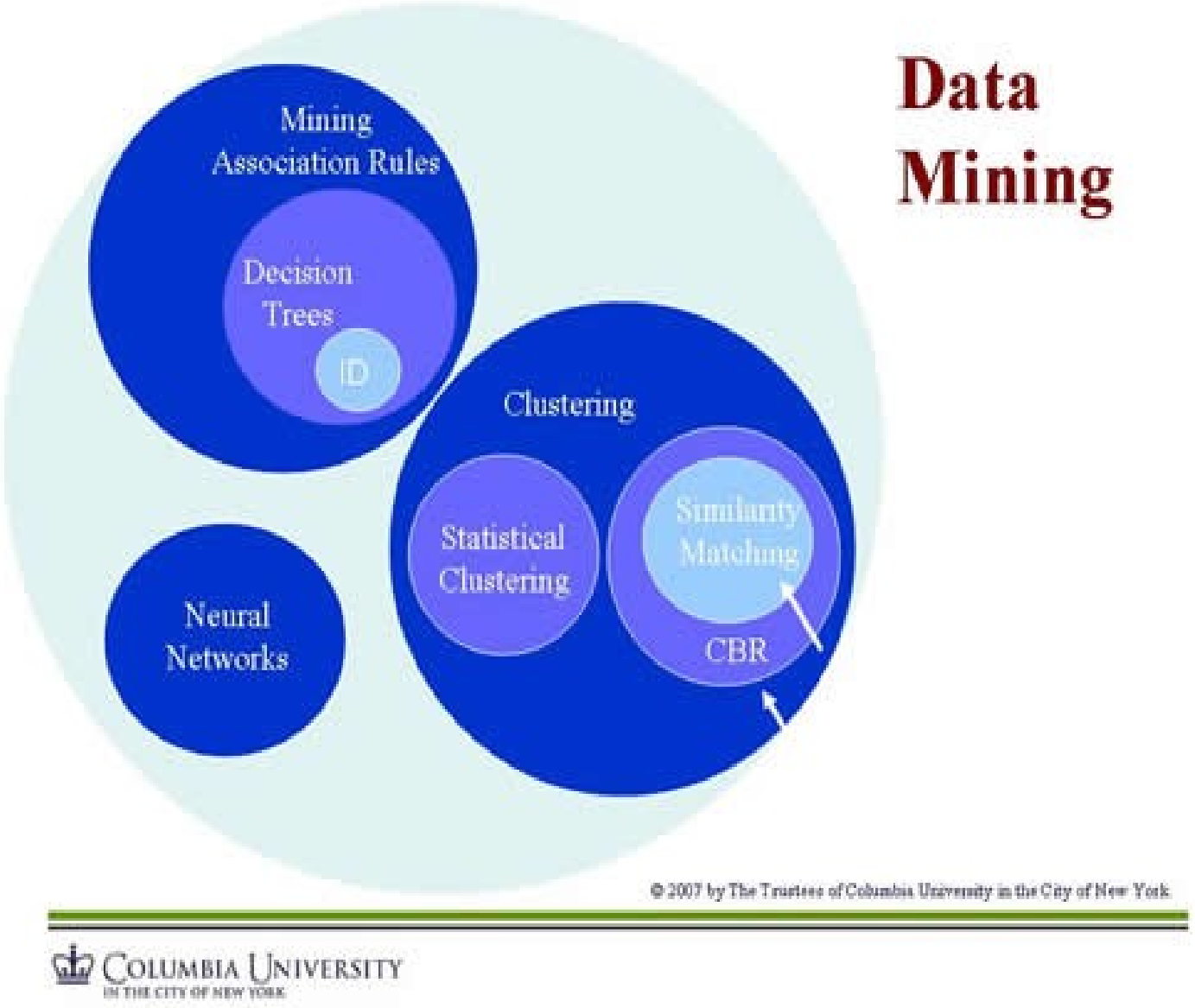
# World’s most powerful HPCC


World’s top supercomputer: Tihane-2 (China), 33.86 petaflops, 16000 nodes, 3,120,000 cores, 88 GB RAM at each node

# World’s most powerful HPCC

World’s top supercomputer: Tihane-2 (China), 33.86 petaflops, 16000 nodes, 3,120,000 cores, 88 GB RAM at each node

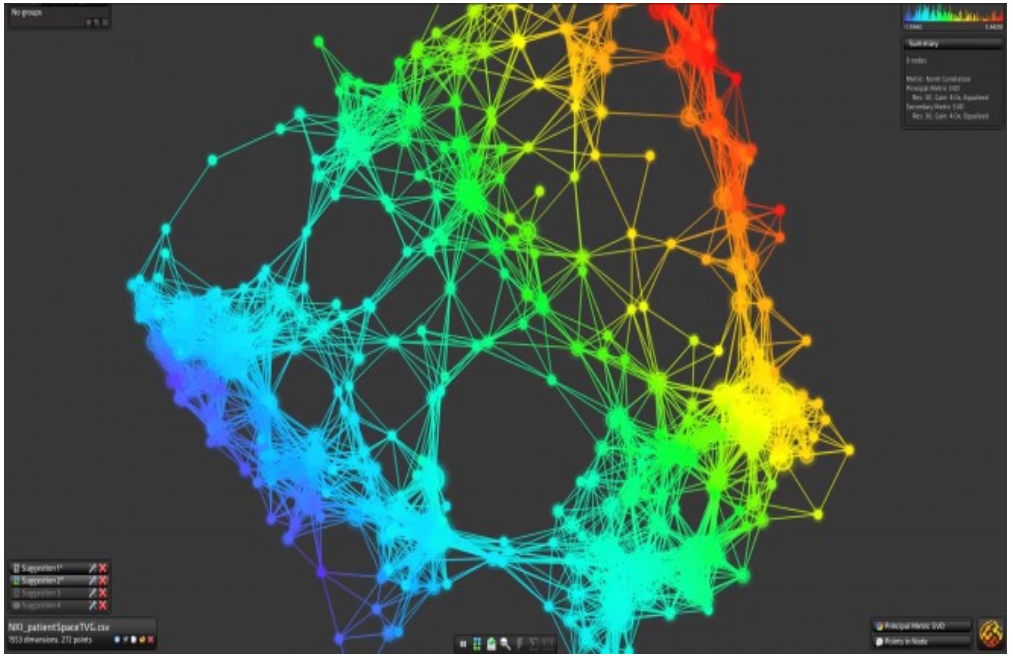
# Data Mining

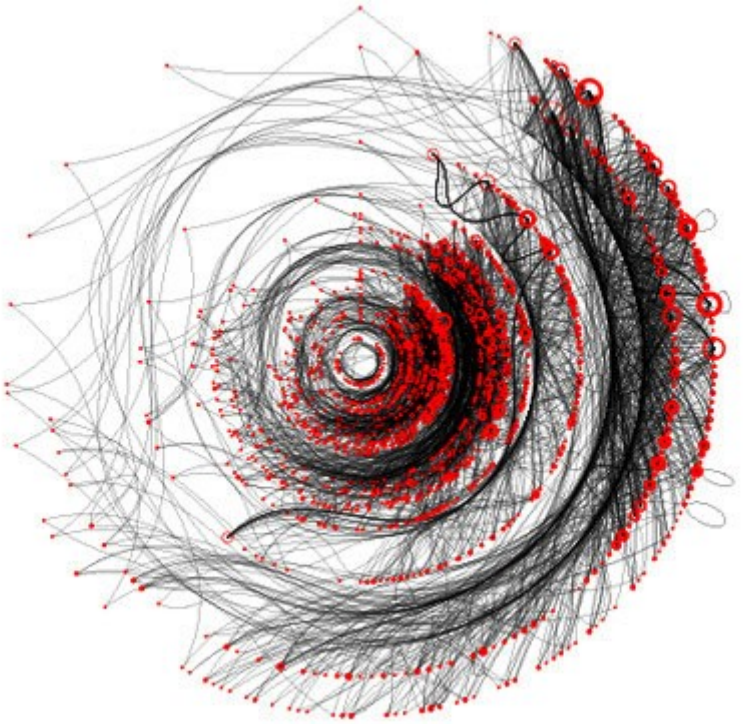



Data Mining (image from Columbia University)

# Data Mining

Data Visualisation (to catch patterns)



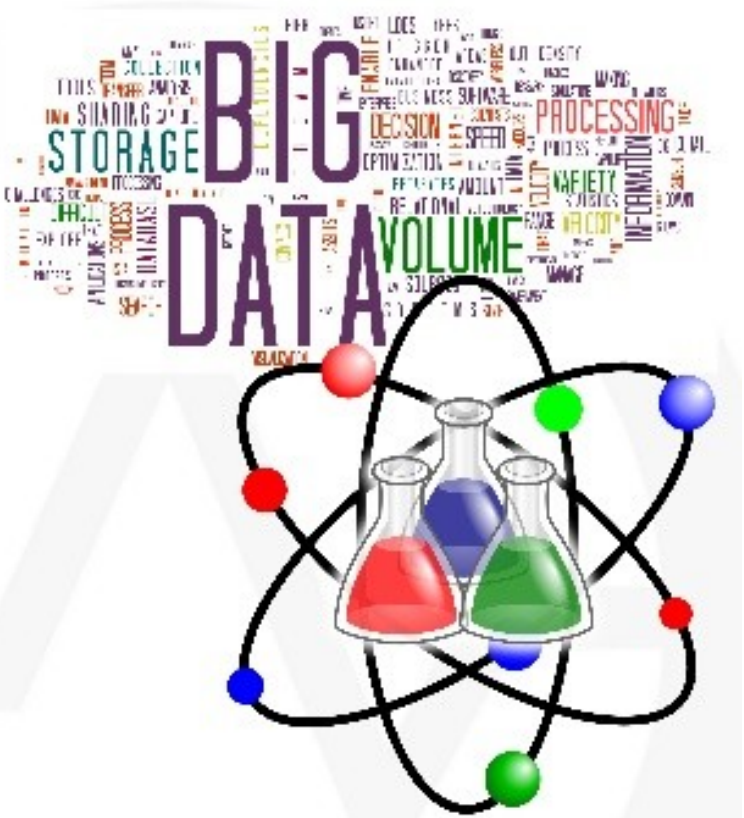


# Data Science...

## What is Data Science?

**Extraction of knowledge from large volumes of data** that are structured or unstructured.

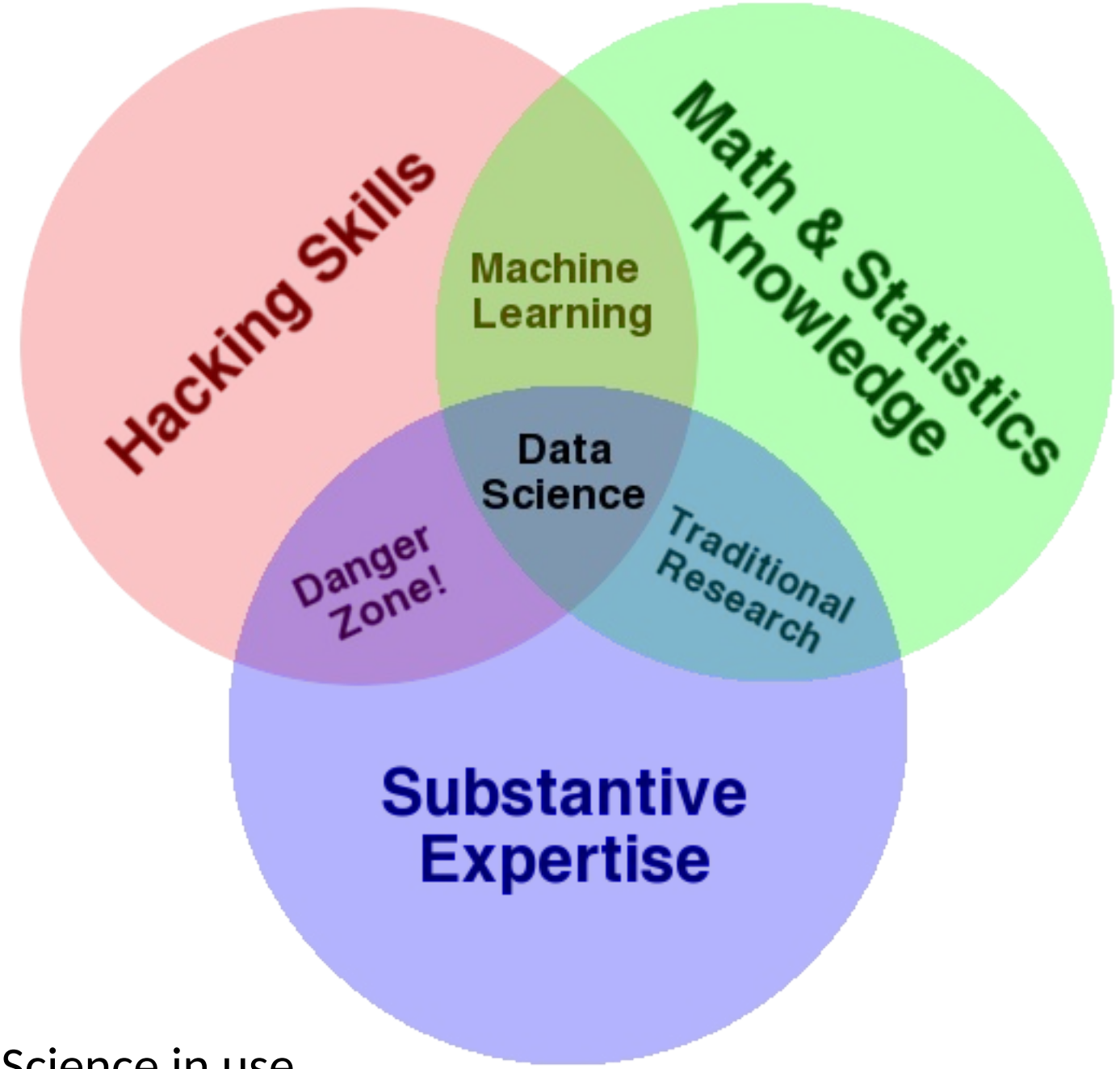
It is a continuation of the fields **data mining** and **predictive analytics**





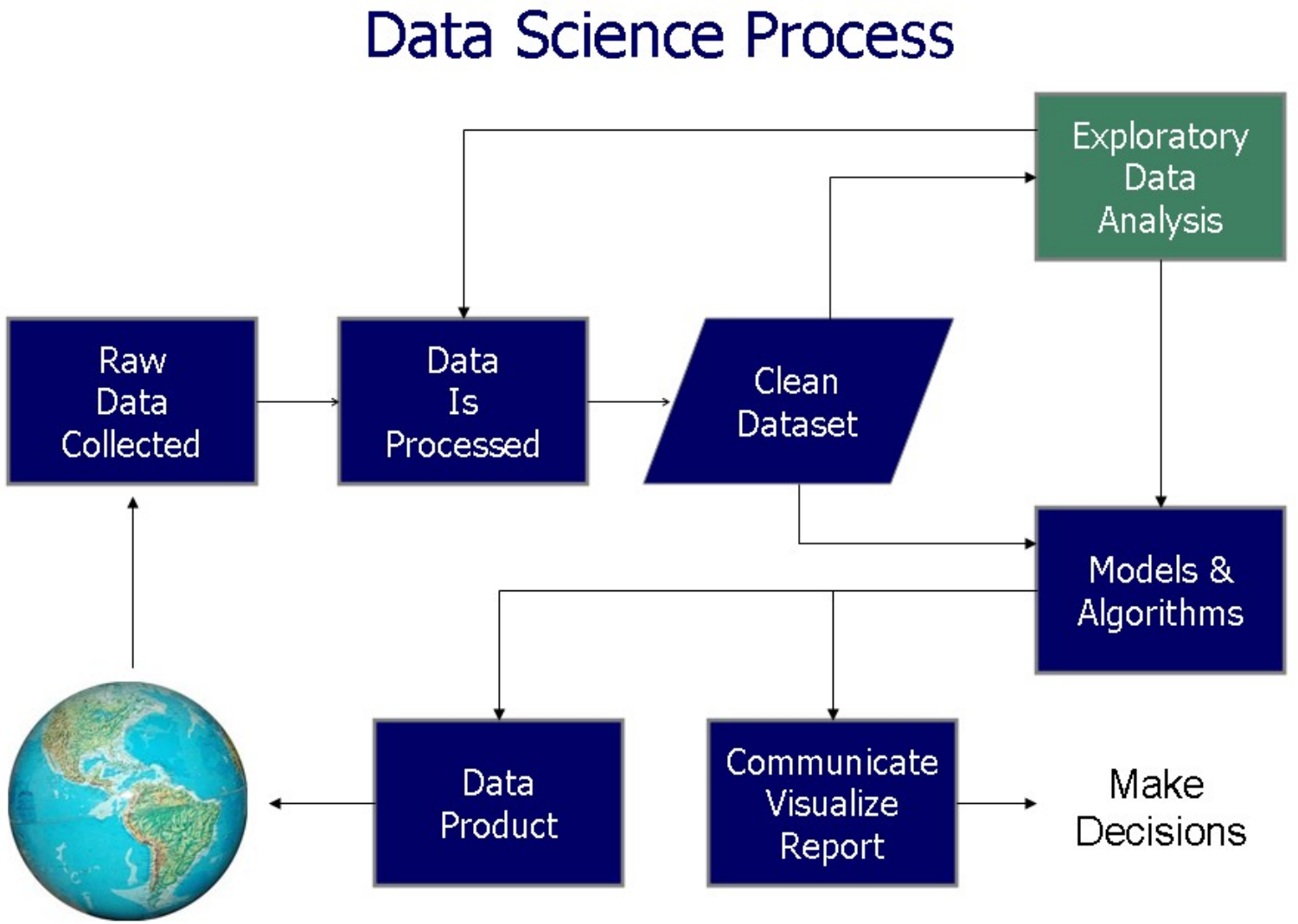

Structured data (<5%), semi-structured (<10%), un-structured or *big data* (85+%).

# Data science in use



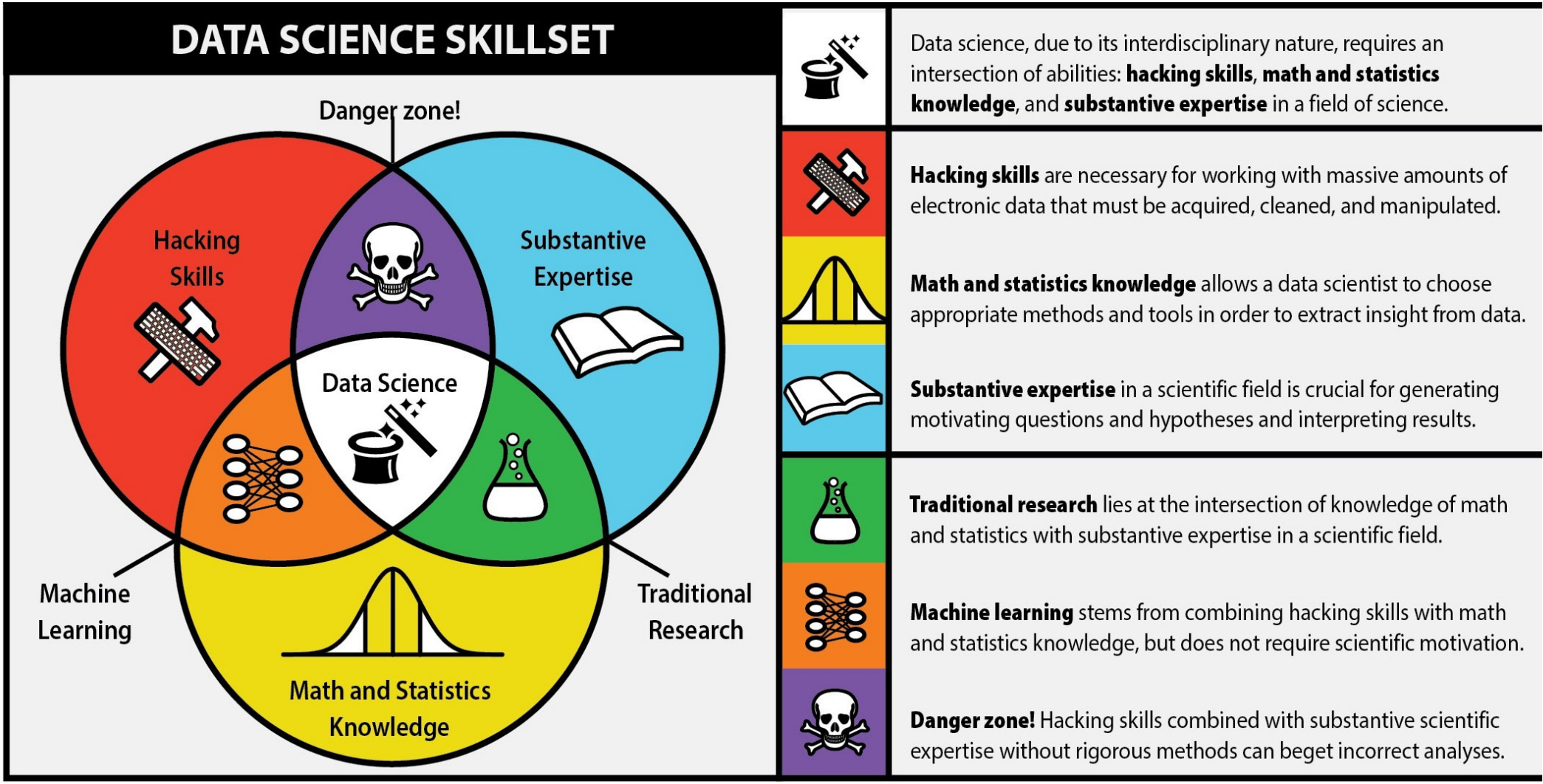

Data Science in use….

# Data science process




Data science process….

# Data science explained




Data science explained….

# Information…

- How do we manage information ?

- We use various Information Systems.

Transaction Processing System (TPS)
Management Information System (MIS)
Enterprise Resource Planning system (ERP)
Library Information System (LIS)
and, many others.

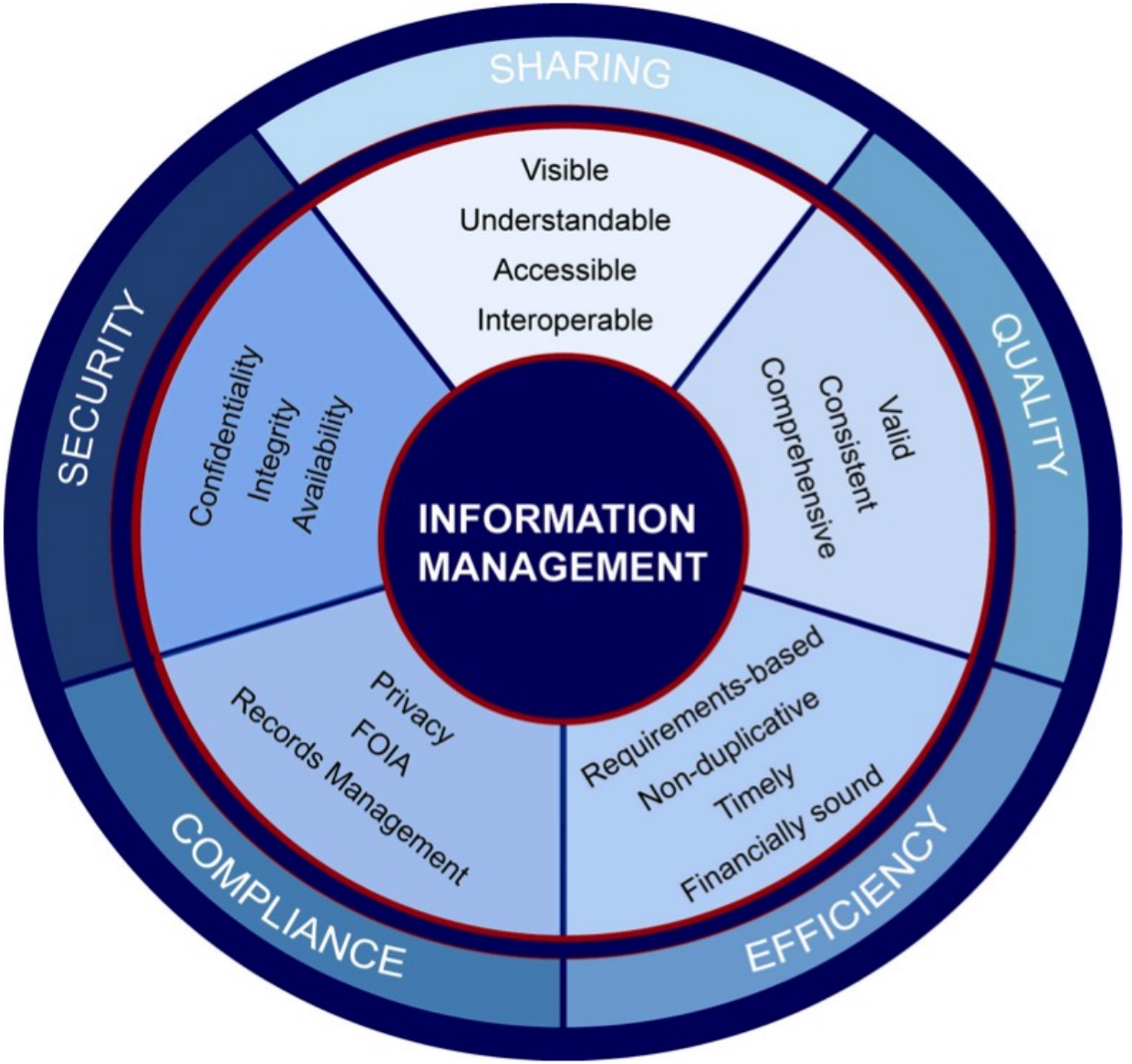
- For structured data: relational DBMS.
- For un-structured data: IBM InfoSphere, IBM InfoStream, Hadoop (several others too).

# Information…

- As research shows, there are a few important aspects of good "Information Management".

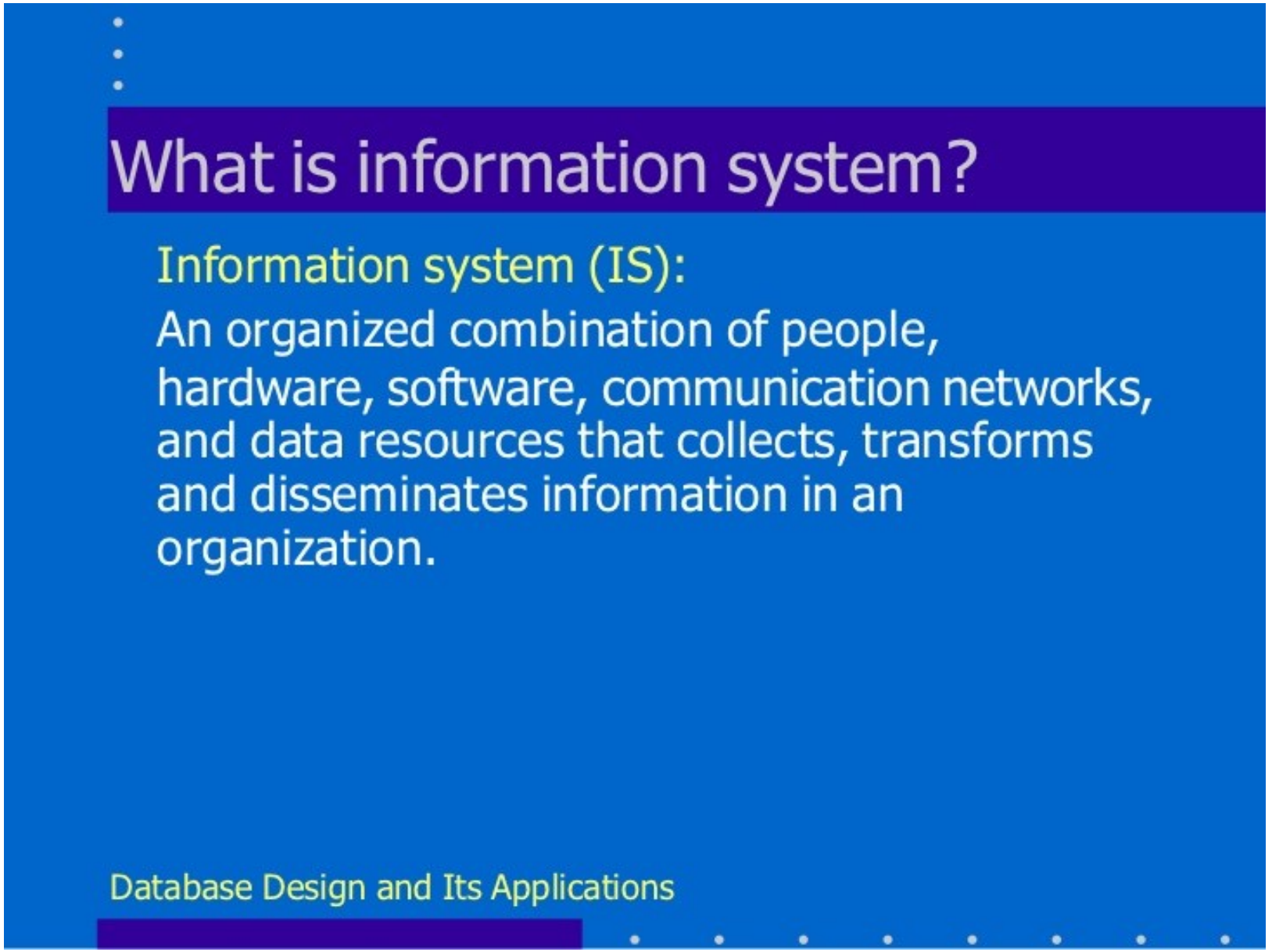
Efficiency (of collection, storage, retrieval…)
Quality (completeness, correctness, reliability…)
Compliance (with need, law, …)
Security (authentic access, prevention of theft and corruption)
Sharing (timely, as much as needed, …)

# Information management




Information management….

# Information system




Information system….

# Future (and current) uses…

- How are these technologies being used ?
- What are the new (and current) developments ?
- In the context of Judicial systems…?

# Future (and current) uses...

- New applications include...

- Text mining:

- Computerised Language processing:

An example of translation by computers:
Pope (on being sked to go and work in Africa for children):
"The spirit is strong but the flesh is weak".
Translated by computer and back: "The vodka is strong but the meat is rotten".

- Google Translate

# Future (and current) uses...

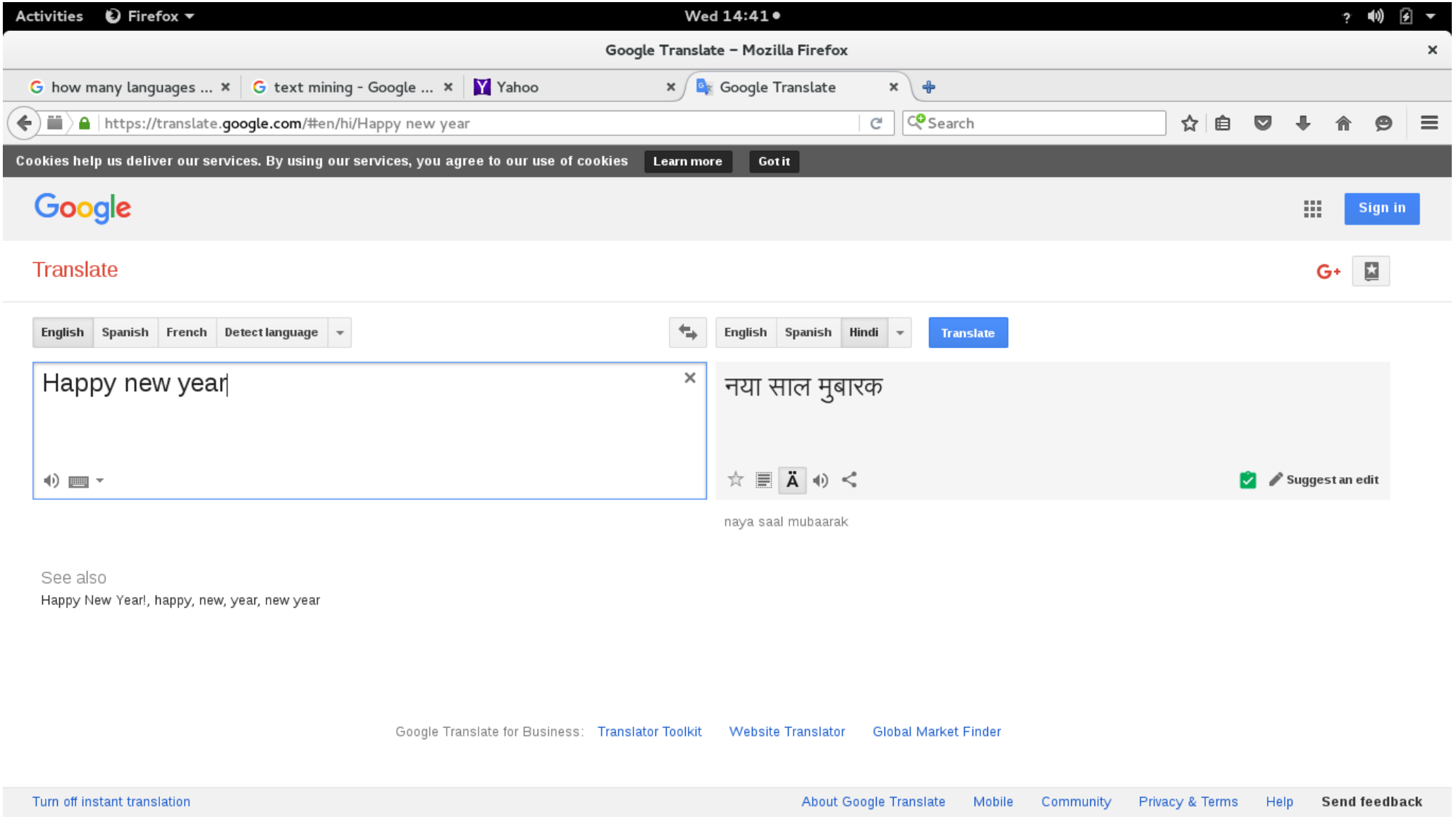
- Google Translate:

- Voice input for 15 languages.
- Translation of a typed word or phrase in over 50 languages.
- Translation can be spoken out loud in: >23 languages.

- Very few Indian Languages: Bengali, Gujarati, Hindi, Punjabi, Sindhi, Tamil, Telugu, Urdu.

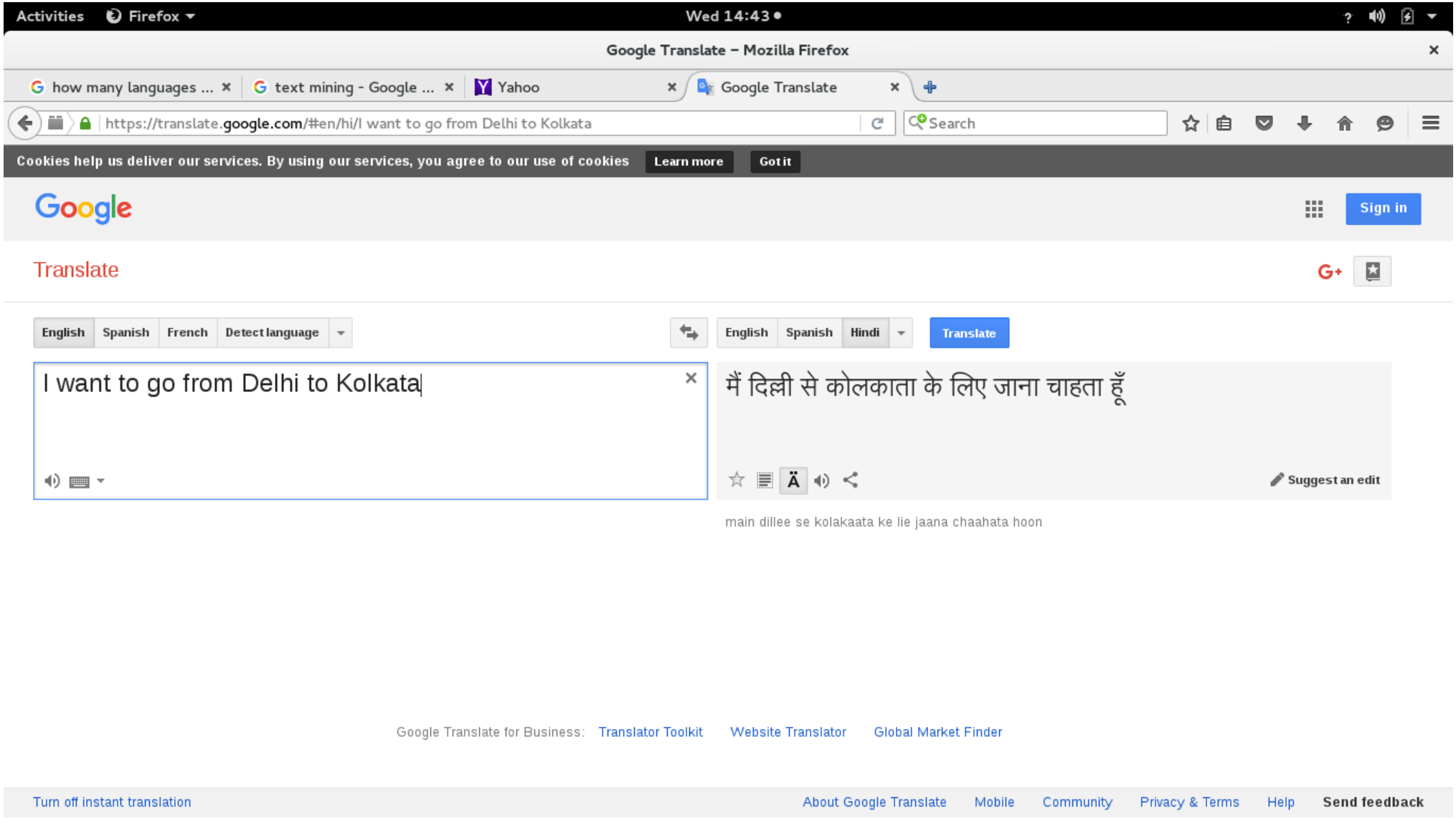
- See: https://translate.google.com/

# Future (and current) uses...




A Google Translate screenshot...

# Future (and current) uses...




Another Google Translate screenshot...

# Problems…

- One word → multiple uses, many meanings.
- Syntax and semantic problems.
- Definitions of technical / legal terms.

- Legal documents – the most difficult.
- Requires: language being read, language to translated to, understanding of all legal terms.
- Also requires: computer proficiency, keyboard familiarity.

- Expensive labour, expensive technologies.
- Modern aids available: can speak to type automatically.

# Available software...

- Free / Open source:

(For more: see wiki/List_of_speech_recognition_software or just search in Google)

- Basic engines: CMU Sphinx, HTK, Julius, Kaldi.
- Usable Applications: Simon, Jasper project.

- Speechnotes (though commercial, available free).
- For mobile: many are available, but none is open source.

- How to do it for dozens of Indian languages ?
- Manually ?
- Huge investment for R&D in technology is needed.

# Available software...

- Problems with simple scanning:

- Scan is an image. Can not be read or searched.
- R&D is needed for "character recognition".

- Typed character easy (optical character recognition).

- Handwritten difficult – ICR (Intelligent Character Recognition) can be used.
- Complex "artificial intelligence" and "neural networks" technology is needed.
- First a learning material is given, corrections for computer are made. Thereafter, achieves 97%+ results.

# Available software...

- CDAC is working on language translation and character recognition in India (with several partners).

That's all folks !!
Thank you !!!

Any Questions?