**SLIDE-1**

**Data Mining Supervised Learning**

**SLIDE-2**

**Data Scientist: The Sexiest Job of the 21st Century**

*-- Harvard Business Review (2012)*

Why Machine Learning?

**SLIDE-3**

**Welcome to the Information Age …**

**Drowning in data and starving for Knowledge**

**SLIDE-4**

**Data in every domain…**

* **Web** – web content, link structure, search clicks…
* **Retail** – customer details, point of sales, inventory, …
* **Census** – demographics, population indicators
* **Medical** –literature, biological data, diagnosis, drug trials,…
* **Science** – literature, scientific measurements,…
* **Remote Sensing** – optical, infrared, hyper-spectral
* **Space** – Mars rovers, Hubble telescope, Radio astronomy
* **Financial** – stocks, currencies, financial news, commodities,…
* **Insurance** – customer records, claims history, process data,…
* **Telecom** – customer records, call history, payment history,…
* **Intelligence** – phone taps, intelligence reports, surveillance,…

SLIDE-5

How BIG is Big Data?

**TWITTER** 🡪 175 million tweets every day, 465 million accounts. (Early 2012)

**YOUTUBE** 🡪 YouTube users upload 100 hours of video every minute

**WWW** 🡪 800 new websites are created every minute

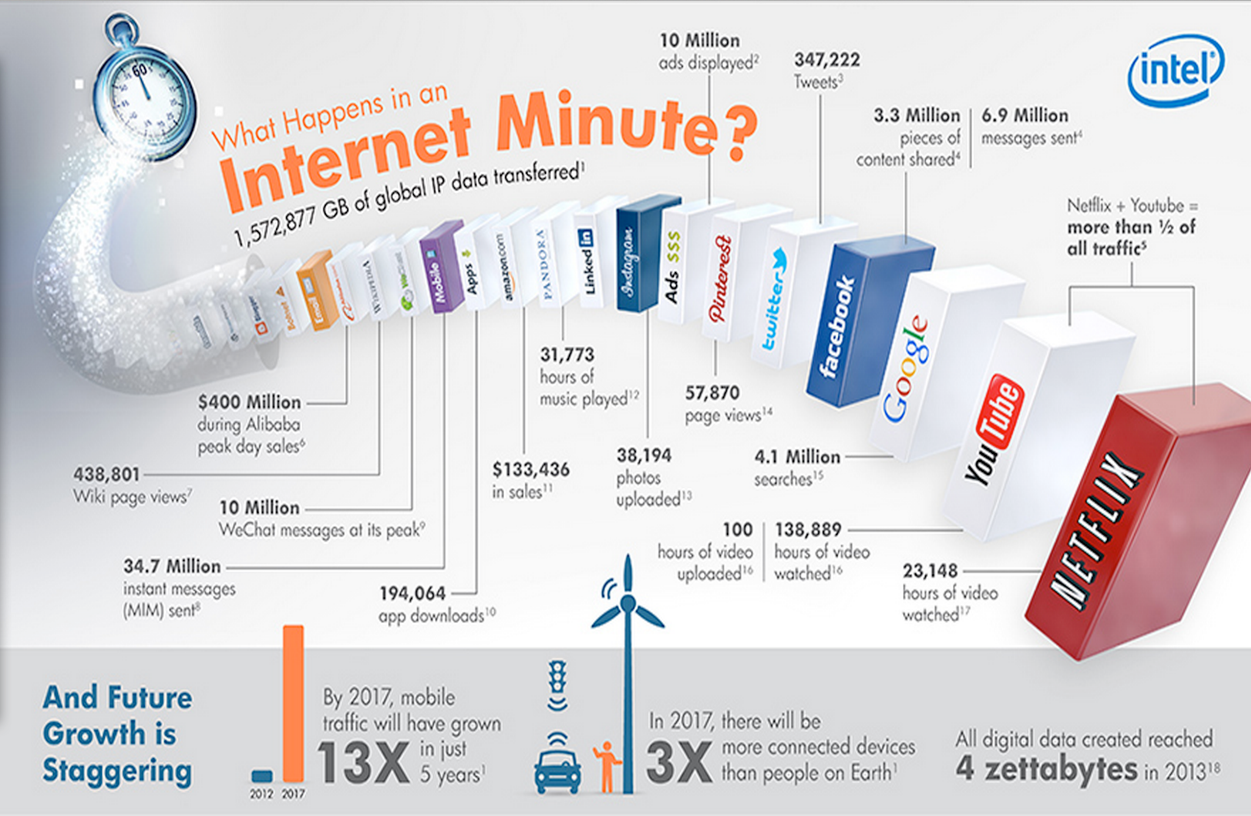
**FACEBOOK** 🡪 100 terabytes of data uploaded daily

30 Billion Pieces of content shared every month 30+ Petabytes of user data

**GOOGLE** 🡪Processing 20 petabytes a day (2008)

**WALMART 🡪** more than 1 million customer transactions every hour

SLIDE-6



Source: http://www.intel.in/content/www/in/en/communications/internet-minute-infographic.html

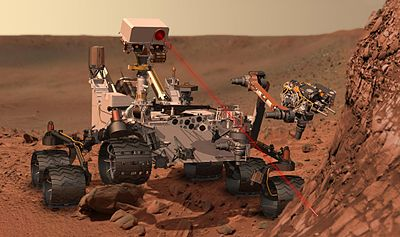
**SLIDE-7**

* This data explosion is enabled by…
* Better “Sensors” – Higher Resolution, Bands that are More Spectral, Quick Experimental Turnaround, and Crowd sourcing…
* Higher Bandwidth Communication – Faster Networks and Routers, Better Compression technologies…
* Larger Warehouses – Cheaper Storage, Multi-Level Caching, Scalable Database/Data warehousing technologies…
* Massive Crunching Power – Faster Multi-core processors, Parallel Distributed Computing, Map Reduce paradigms…
* Advances in Machine Learning and Data Mining –Sophisticated Learning frameworks, Distributed Data mining…

SLIDE-8

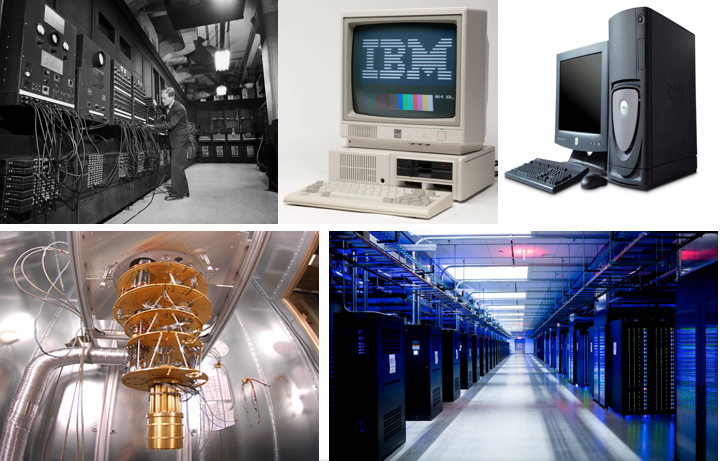
The evolution of “**Sensors**”



SLIDE-9

THE EVOLUTION OF COMPUTING



SLIDE-10

* The evolution of **Information Technology**
* The **INDEXING** era Collection, Storage, Querying, and Retrieval
* The **INTERPRETATION** era Visualizations, Analysis, Insights, and Semantics
* The **INTELLIGENCE** era Learning, Prediction, Decisions, and Thinking!

SLIDE-11

Two Goals of Machine Learning

* PREDICTION
* INTERPRETATION

SLIDE-12

1. What is the nature of Mind?
2. What is learning? Intelligence? Thinking? Creativity? Consciousness?
3. How does the brain do it?
4. Reading 🡪 Understanding (book, paper, review, poetry, joke)
5. Listening 🡪 Hearing (speech, music, nature, )
6. Looking 🡪 Seeing (images, videos, maps, environment)

SLIDE-13



Generalization: The ability to predict or assign a label to a “new” observation based on the “model” built from experience.

SLIDE-14

What is “**Thinking**”?

**Semantics**: Ability to assign **meaning** to all the **parts** simultaneously based on (the still ambiguous) **context** so the **whole** makes sense!

SLIDE-15

2. Driving “Decisions” from data!

* Which position should a Page/Ad be shown for a query?
* Should the home loan/credit card transaction be approved?
* Which Video to show next on YouTube?
* Who might user X connect/follow on LinkedIn/Twitter?
* What articles/movies/songs customer X might be interested in?
* Which coupons/offers will a customer respond to here/now?
* Where to open the next store? How to price the next gadget?
* Which medicine/treatment will cure this disease?
* Who will be the right person for this Job?
* Which crop is best for this soil/weather?

SLIDE-16

From **Data** to **Decisions**…

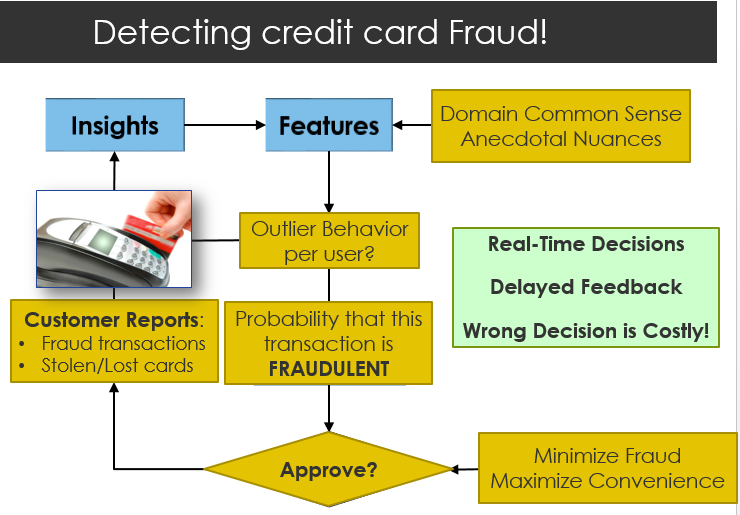
**Insights 🡪Features🡨 Domain Knowledge**

**Data Models**

**Feedback Predictions**

**Decision Business Objectives**

SLIDE-17



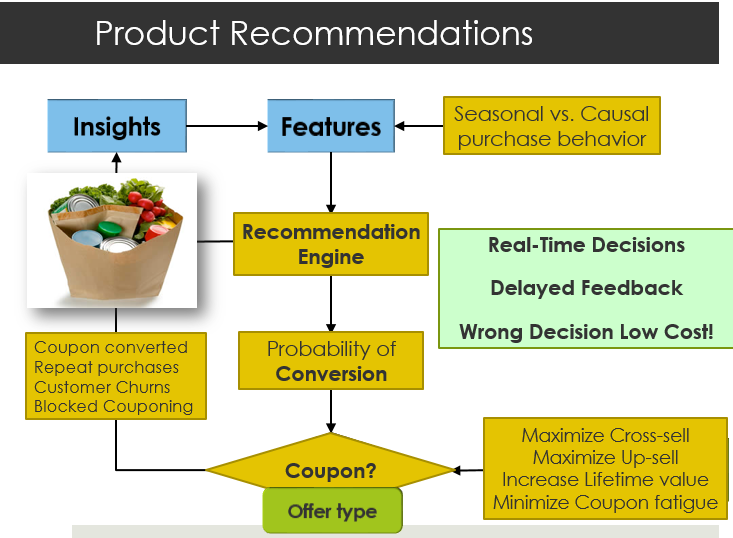
**SLIDE-18**



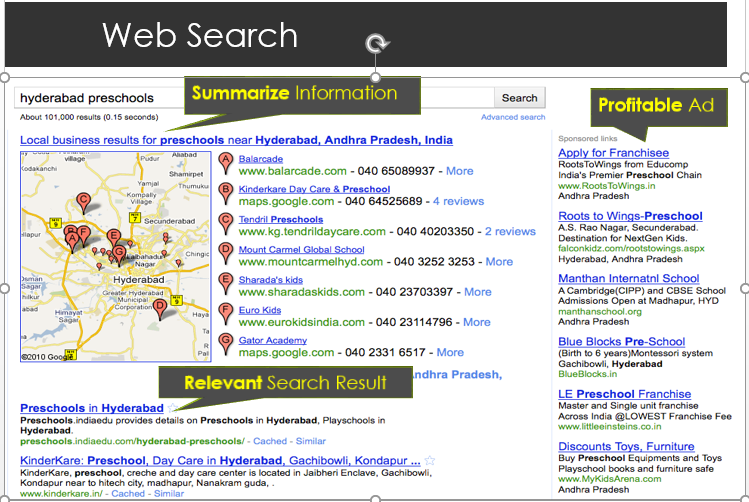
SLIDE-19



SLIDE-20



SLIDE-21



SLIDE-22



SLIDE-23



SLIDE-24

Key Data Scientist Questions!

What DECISIONS does my business make?

e.g. Which offer to send to which customer?

On what BASIS do I make those decisions?

e.g. Past purchase behaviors of those/similar customers?

How do I quantify SUCCESS of my decisions?

e.g. What fraction of offers get converted?

What data should I collect to EVALUATE my decisions?

e.g. Did customers redeem the coupons – after how long, how often.

What data should I collect to IMPROVE my decisions?

e.g. Point of sales data, Social data, Reviews, etc.

How do I improve my MODELS from the data I collect?

More data, More features, Better Modeling, More Customization,…

SLIDE-26

**Machine Learning Paradigms**

**Unsupervised Learning**

Find structure in data. (Clusters, Density, Patterns)

**Supervised Learning**

Find mapping between features to labels

**Semi-Supervised Learning**

Using unlabeled data to improve Supervised Learning Models

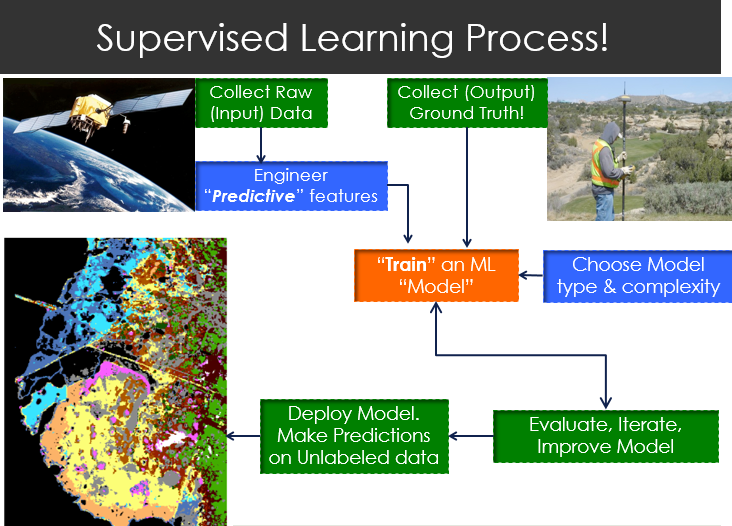
**Active Learning**

Find which model to get labeled next to maximize value

**Reinforcement Learning**

Learn local/early strategies from global/delayed “rewards”

SLIDE-27



SLIDE-28

**Supervised Learning Paradigms**

**Regression**

Predict a numerical VALUE.

**Classification**

Predict a categorical CLASS.

**Recommendation**

Predict user PREFERENCE from a large pool of options.

**Retrieval**

Predict RELEVANCE of an entity to a “query”

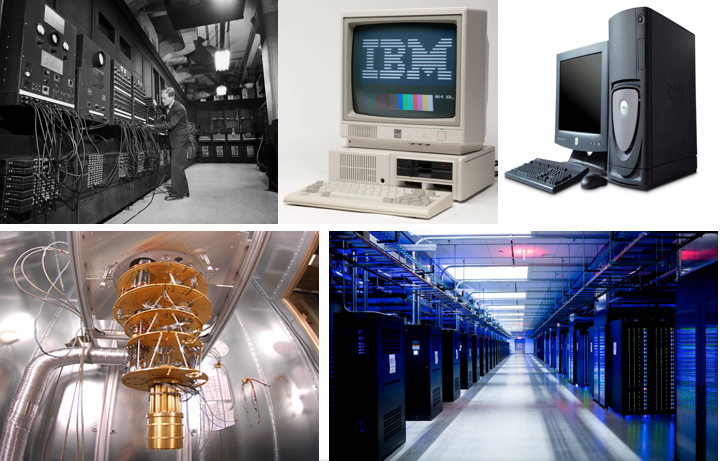
SLIDE-29

**Which Supervised Learning Problem is this?**

* Does this MRI scan show Cancer?
* What will this stock value be tomorrow?
* How much credit limit should I give a customer X?
* Is customer X about to Churn?
* What activity (*stealing* or *shopping*) is user doing in a Video?
* What objects (*car, tree, horse, …*) are present in an image?
* How many iPads will I sell in Mumbai between April – July?
* Is this a positive or negative TWEET about a person/product?
* How high the value of my house will go in 3 years?
* Which song are you humming the tune of?
* Who will be the best match for you (say shaadi.com)

**SLIDE-30**

Which Supervised Learning Problem is this?



* Which crop is best for this soil/weather?
* Which position should a Page/Ad be shown for a query?
* Should the home loan/credit card transaction be approved?
* Which Video to show next on YouTube?
* Who might user X connect/follow on LinkedIn/Twitter?
* Which product the customer will buy next?
* How much should Android One / iPhone 6 be sold for in India?
* Which medicine/treatment will cure this disease?
* Who will be the right person (resume) for this Job (JD)?
* What movies customer X might be interested in?