FIT1043 Introduction to Data Science

Module 2: Data Models in Organisations

Lecture 4: application areas and case studies

Monash University

Unit Schedule: Modules

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	**3**	**data business models**
	4	**application areas and case studies**
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	7	resources and standards
	8	resources case studies
5.	9	data analysis theory
	10	data analysis process
6.	11	issues in data management
	12	data management frameworks

Discussion: Python Language

- easy to learn
- flexible and multi-purpose
- great libraries
- well designed computer language
- good visualization for basic analysis

Discussion: Python versus R

- both are free
- R developed by statisticians for statisticians, huge support for analysis
- Python by computer scientists for general use
- R is better for stand-alone analysis and exploration
- Python allows for easier integration with other systems
- Python easier to learn and extend than R (better language)
- R has vectors and arrays as first class objects; similar to Matlab!
- R currently less scalable.

See *In data science, the R language is swallowing Python* by Matt Asay, recent blog in *Infoworld*.

Application Areas
(ePub section 2.5)

Consider different application areas:

- case studies from NIST:
  - provides a broad framework for analysing applications
- McKinsey Global Institute (MGI) report on big data:
  - study of different application areas

Application Areas:
Case studies (from NIST)

provides a broad framework for analysing applications

NIST Case Studies

- we will now review the NIST analysis

Caveat: many of the questions asked we wont look at properly until later modules

- so you may not be able to complete this kind of analysis now, but this should show you where we are headed

NIST Analysis

data sources: where does the data comes from?

data volume: how much there is?

data velocity: how does the data change over time?

data variety: what different kinds of data is there?

data veracity: is the data correct? what problems might it have?

software: what software needed to do the work?

analytics: what statistical analysis & visualisation is needed?

processing: what are the computational requirements?

capabilities: what are key requirements of the operational system?

security/privacy: what security/privacy requirements are there?

lifecycle: what ongoing requirements are there?

other: are there other notable factors?

NIST Analysis, cont.

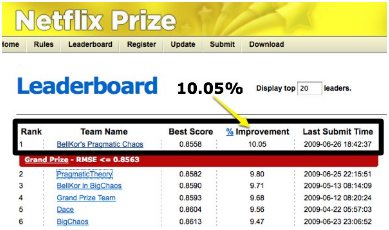
Generally, we can relate the NIST categories to materials in our modules:

- **Module 2:** capabilities; lifecycle; other
- **Module 3:** data sources; data volume; data velocity; data variety; data veracity; processing;
- **Module 4:** software; other
- **Module 5:** analytics;
- **Module 6:** security/privacy; lifecycle

Case Study: Netflix Movies

Netflix

- on demand internet streaming, and flat-rate DVD rental
- over 50 million subscribers in the US by 2014
- international market
- video recommendation!
- established the *Netflix Prize* in 2006-2009 as a crowdsourced way of testing out algorithms



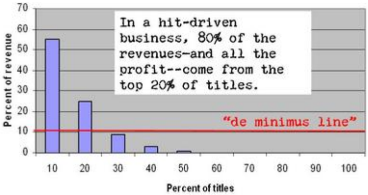
Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:06:52

By Ivongala (Own work) [Public domain], via Wikimedia Commons

Netflix: Background

Analysis follow the NIST Big Data WG Netflix analysis in *Volume 3, Use Cases and General Requirements*, case 7 on page 8, A-24 and elsewhere

- Pareto principle, or 80/20 rule:
  - top 20% of films watched 80% of time
  - standard video store stocked less than 20% of available titles in order to make the most money



from The real meaning of 80/20

- By adopting an Amazon style business model, Netflix could afford to rent the remaining 80%, the so-called *long tail*

Netflix: Analysis

data sources: user movie ratings, user clicks, user profiles

data volume: in 2012: 25 million users, 4 million ratings/day, 3 million searches/day, video cloud storage of 2 petabytes

data velocity: video titles change daily, rankings/ratings updated

data variety: user rankings, user profiles, media properties

software: *Hadoop*, *Pig*, *Cassandra*, *Teradata*

analytics: personalised recommender system

processing: analytic processing, streaming video

capabilities: ratings and search per day, content delivery

security/privacy: protect user data; digital rights

lifecycle: continued ranking and updating

other: mobile interface

Case Study: Electronic Medical Records (EMR)

Clinical Data

CLINICAL HISTORY: Cough, congestion.

COMMENTS:

PA and lateral views of chest reveals no evidence of active pleural or pulmonary paren
are diffusely increased interstitial lung markings consistent with chronic bronchitis. Und
is not excluded. The cardiac silhouette is enlarged. The mediastinum and pulmonary ve
is tortuous. Degenerative changes are noted in the thoracic spine.

IMPRESSION:

1. No evidence of acute pulmonary pathology.
2. Enlarged cardiac silhouette.
3. Tortuous aorta.
4. Diffusely increased interstitial lung markings consistent with chronic bronchitis. Unde
not excluded.
5. Consider follow up with Chest CT if clinically warranted.

Claims and Cost Data

Stanford HEALTH CARE
STANFORD MEDICINE

P.O. BOX 742188
Los Angeles, CA 90074-2188

Customer Service
Toll Free: 1-800-549-3720, or
650-498-5850
9am-4pm
Monday-Friday

See Reverse Side for Patient Billing Details

Pay Online: stanfordhealthcare.org/billing
To activate your MyHealth account, visit myhealth.stanfordhealthcare.org/activation and enter the access code: XX0XX-0XXXX-XXX00

Guarantor ID# 000000000

YOUR PHYSICIAN STATEMENT

Page 1 of 1

Summary of Services and Amounts Due

Service Date	Provider	Description	Charges	Credits	Insurance Balance	Patient Balance
Patient: SAMPLE, SAMPLE S		Visit #000000000 Service Line: General Surgery				
07/01/2014	Badger, James T, MD	99201 EVAL/MGMT OF NEW PATIENT	165.00		0.00	82.50
08/07/2014		UNINSURED DISCOUNT ADJ		-82.50		
		Totals:	165.00	-82.50	0.00	82.50

Electronic Medical Records

follows NIST Big Data WG Electronic Medical Records analysis in *Volume 3, Use Cases and General Requirements*, case 16 in page 14, A-45 and elsewhere

- clinical data and claims/cost data is available per patient, per hospital
- large variety of sources of data
- systematic errors and difference in standards across institution

Task: segment patients into different types (“phenotypes”) to use in subsequent cohort studies

- case study is for Indiana Network for Patient Care

EMR: Analysis

data sources: clinical and claims data

data volume: 1000 centres, 12 million patients, 4 billion clinical events

data velocity: approx. 1 million clinical events/day

data variety: free text, lab results, pathology, outpatient, *etc.*

data veracity: different standards in different places

software: Hadoop, *Hive*, Teradata, PostgreSQL, MongoDB

analytics: visualisation for data checking; standardisation of incoming data; general data analysis

processing: analytic processing, handling the volume

capabilities: models to support subsequent cohort studies

security/privacy: privacy and confidentiality required

lifecycle: full data management required

Case Study: Medical Imaging

Medical Imaging Data

Task: Produce Analysis

CLINICAL HISTORY: Cough, congestion.
COMMENTS:
PA and lateral views of chest reveals no evidence of active pleural or pulmonary paren
are diffusely increased interstitial lung markings consistent with chronic bronchitis. Und
is not excluded. The cardiac silhouette is enlarged. The mediastinum and pulmonary ve
is tortuous. Degenerative changes are noted in the thoracic spine.
IMPRESSION:
1. No evidence of acute pulmonary pathology.
2. Enlarged cardiac silhouette.
3. Tortuous aorta.
4. Diffusely increased interstitial lung markings consistent with chronic bronchitis. Unde
not excluded.
5. Consider follow up with Chest CT if clinically warranted.

Medical Imaging

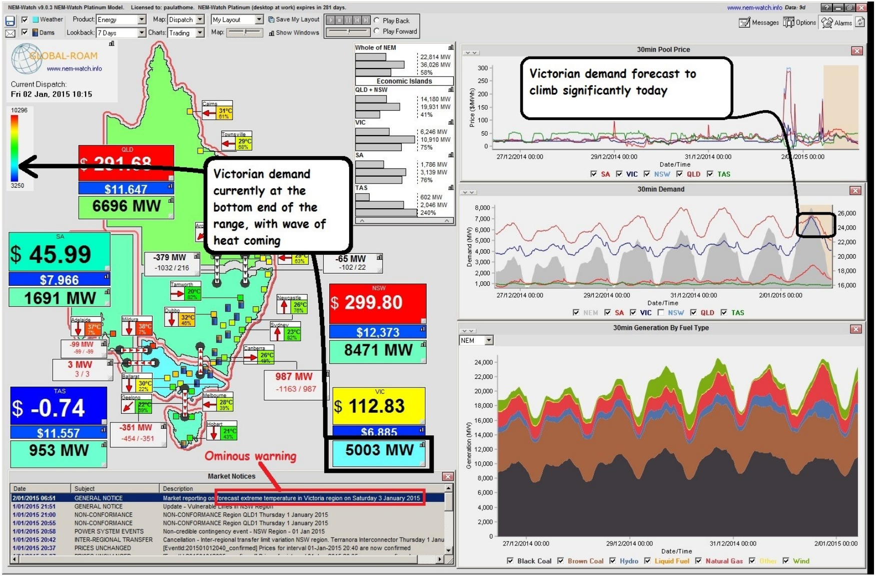
follow NIST Big Data WG Pathology Imaging in *Volume 3, Use Cases and General Requirements*, case 17 in page 14, A-48 and elsewhere

- biomedical data for imaging is high resolution and some is 3D
- interpretation of images done by trained experts
- requires significant training in interpretation
- many different kinds of instruments each requiring different interpretations
- millions produced daily in the USA

EMR: Analysis

- data sources: biomedical image data
- data volume: approx. 1 million events/day nationally
- data variety: X-rays, CT scans, microsopes, ...
- data veracity: current interpretation is often text based, so prone to text errors
- software: advanced image processing and machine learning systems
- analytics: computational image processing, supervised learning from images
- processing: handling the large volume, distributed and high throughput
- capabilities: produce initial analysis for experts
- security/privacy: privacy and confidentiality required
- lifecycle: full data management required

Case Study: Electricity Demand Forecasting



from [Analysing trends in VIC electricity demand]

Electricity Demand Forecasting

from NIST Big Data WG Electricity Demand Forecasting in
Volume 3, Use Cases and General Requirements, case 51 in page 43 and A-134

- near realtime usage available thanks to smart meters
- with solar cells, consumers do energy generation too, but it is unpredictable
- main electricity generation must be planned
- brownouts and blackouts need to be prevented
- see [1]Australian Energy Market Operator (AEMO) and [2]their electricity site

[1] http://www.aemo.com.au/
[2] http://www.aemo.com.au/Electricity

Electricity Demand Forecasting: Analysis

data sources: utilities, smart meters, weather data, grids

data volume: city scale: 10GB/day

data velocity: updates every 15 minutes

data variety: time series, networks, spatial data

data veracity: occasional dropouts

software: advanced timeseries processing, spatial analysis

analytics: forecasting models

processing: handling the forecasting volume

capabilities: produce forecasts at different scales (hourly, daily)

security/privacy: privacy and confidentiality required

lifecycle: full data management required

Application Areas: McKinsey Global Institute report on big data

study of different application areas

Application Areas

We present details from McKinsey Global Institute report on Big Data from 2011, *"Big data: The next frontier for innovation, competition, and productivity"*

According to the MGI report, the main application areas of Big Data are:

1. Health
2. Government
3. Retail
4. Manufacturing
5. Location Technology

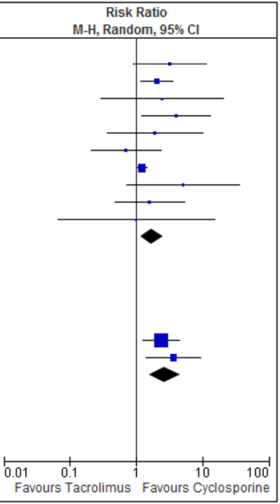
NB. What happened to Science? MGI is an industry organisation.

Application Areas: Health

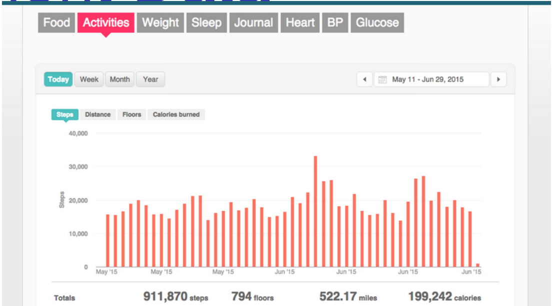
Lets view
"A Data Driven Approach to Diagnosing and Treating Disease", on VideoLectures.NET (video, see time 00:00-11.27)

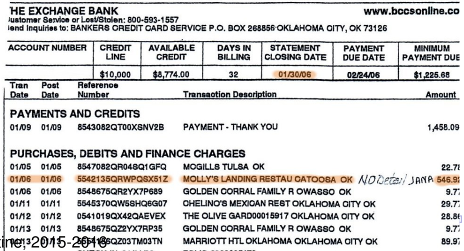
Pharmaceutical R&D Data

Study or Subgroup	Tacrolimus Events	Tacrolimus Total	Cyclosporine Events	Cyclosporine Total	Weight	Risk Ratio M-H, Random, 95% CI
1.14.1 Occurrence of diabetes at 12 months (RCT)						
Campos (2002)	10	85	3	81	7.5%	3.18 [0.91, 11.13]
Ekberg (2007)	34	403	17	408	20.5%	2.02 [1.15, 3.57]
Hardinger (2005)	5	134	1	66	3.0%	2.46 [0.29, 20.66]
Mayer (1997)	25	303	3	145	8.2%	3.99 [1.22, 12.99]
Murphy (2003)	4	52	2	50	4.7%	1.92 [0.37, 10.04]
Scantlebury (1991)	4	20	4	14	8.0%	0.70 [0.21, 2.34]
Silva (2007)	136	212	111	212	34.7%	1.23 [1.04, 1.44]
Vicenti (1996)	17	67	1	20	3.5%	5.07 [0.72, 35.81]
Wang (2000)	5	25	4	32	8.0%	1.60 [0.48, 5.35]
Yang (1999)	1	30	1	30	1.9%	1.00 [0.07, 15.26]
Subtotal (95% CI)		**1331**		**1058**	**100.0%**	**1.72 [1.17, 2.52]**
Total events	241		147			
Heterogeneity: Tau² = 0.10; Chi² = 13.94, df = 9 (P = 0.12); I² = 35%						
Test for overall effect: Z = 2.75 (P = 0.006)						
1.14.2 Occurrence of diabetes at 36 months (Cohort)						
Ekberg (2009)	30	251	13	255	70.0%	2.34 [1.25, 4.39]
Jurewicz (2003)	18	115	5	117	30.0%	3.66 [1.41, 9.54]
Subtotal (95% CI)		**366**		**372**	**100.0%**	**2.68 [1.59, 4.53]**
Total events	48		18			
Heterogeneity: Tau² = 0.00; Chi² = 0.59, df = 1 (P = 0.44); I² = 0%						
Test for overall effect: Z = 3.69 (P = 0.0002)						



Patient Behaviour and Sentiment Data





HE EXCHANGE BANK www.bccsonline.co
ustomer Service or Lost/Stolen: 800-593-1557
end Inquiries to: BANKERS CREDIT CARD SERVICE P.O. BOX 268856 OKLAHOMA CITY, OK 73126

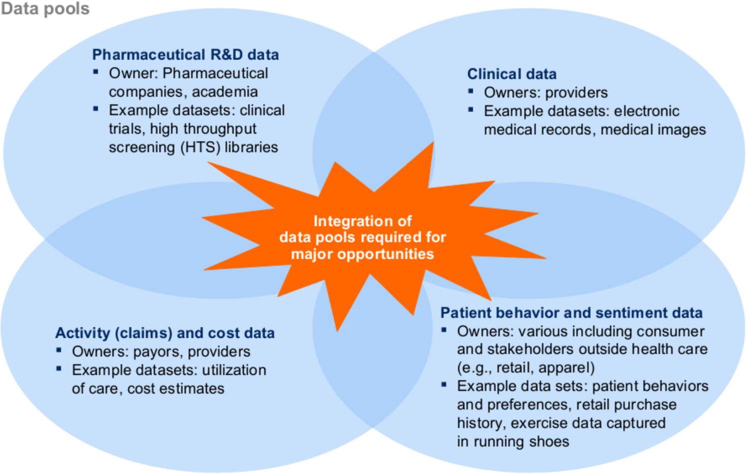
ACCOUNT NUMBER	CREDIT LINE	AVAILABLE CREDIT	DAYS IN BILLING	STATEMENT CLOSING DATE	PAYMENT DUE DATE	MINIMUM PAYMENT DUE
	$10,000	$8,774.00	32	01/30/06	02/24/06	$1,225.68

Tran Date	Post Date	Reference Number	Transaction Description	Amount
PAYMENTS AND CREDITS				
01/09	01/09	8543082QT00XSNV2B	PAYMENT - THANK YOU	1,458.09
PURCHASES, DEBITS AND FINANCE CHARGES				
01/05	01/05	8547082QR04SQ1GFQ	MOGILLS TULSA OK	22.7
01/06	01/06	5542135QRWPQSX51Z	MOLLY'S LANDING RESTAU CATOOSA OK NO Detail JANA	546.9
01/06	01/06	8548675QR2YX7P689	GOLDEN CORRAL FAMILY R OWASSO OK	9.7
01/11	01/11	5545370QW5SHQ6G07	CHELINO'S MEXICAN REST OKLAHOMA CITY OK	29.7
01/12	01/12	0541019QX42QAEVEX	THE OLIVE GARD00015917 OKLAHOMA CITY OK	28.8
01/13	01/13	8548675QZ2YX7RP35	GOLDEN CORRAL FAMILY R OWASSO OK	9.7
01/13		2654186QZ03TM03TN	MARRIOTT HTL OKLAHOMA OKLAHOMA CITY OK	89.9

Health Applications

Four distinct big data pools exist in the US health care domain today with little overlap in ownership and low integration

Data pools



SOURCE: McKinsey Global Institute analysis

Clinical operations

Comparative effectiveness research: to study patient characteristics and the cost and outcomes of treatments

Clinical decision support systems: compare treatment against guidelines, to alert to drug interactions, *etc.*

Transparency about medical data: to help patients make more informed health care decisions, *e.g.*, cesarean birth

Remote patient monitoring: support cronically ill, and at home care, to reduce subsequent hospital use

Advanced analytics applied to patient profiles: to identify who would benefit from proactive care or lifestyle changes

Payment/Pricing

Automated systems: for insurance fraud detection and checking the accuracy and consistency of payors' claims

Health Economics: performance-based pricing plans based on real-world patient outcomes data to arrive at fair economic compensation

Research and Development

Predictive modeling: rationale drug design

Statistical tools and algorithms to improve clinical trial design: in the clinical phases of the R&D process

Analyzing clinical trials data: to identify additional indications and discover adverse effects

Personalized medicine: understading genetic variation and individual treatment response

Analyzing disease patterns: analyzing disease patterns and trends to model future demand/costs, and make strategic R&D investment decisions

Application Areas: Government

lets review *"The Mayor's Geek Squad"* from New York Times

The Mayor's Geek Squad: Feedback

- Lesson: analysis can be the basis for financial allocation decisions
- New York City was really the right city at the right time. Importantly that the mayor is a major advocate
- other cities:
  - infrastructures readiness, academic supports, citizen habits, and supporting mayor, cross-departmnent cooperation
- data:
  - tollway, myki, traffic cameras, bike rental... social media, people "checking-in", or hashtag in tweets, facebook or instagram
- what is the nature of data availability?
  - city should provide enough public data to start from
  - help the local governing bodies plan urban development
  - could support international students better!

Government Applications

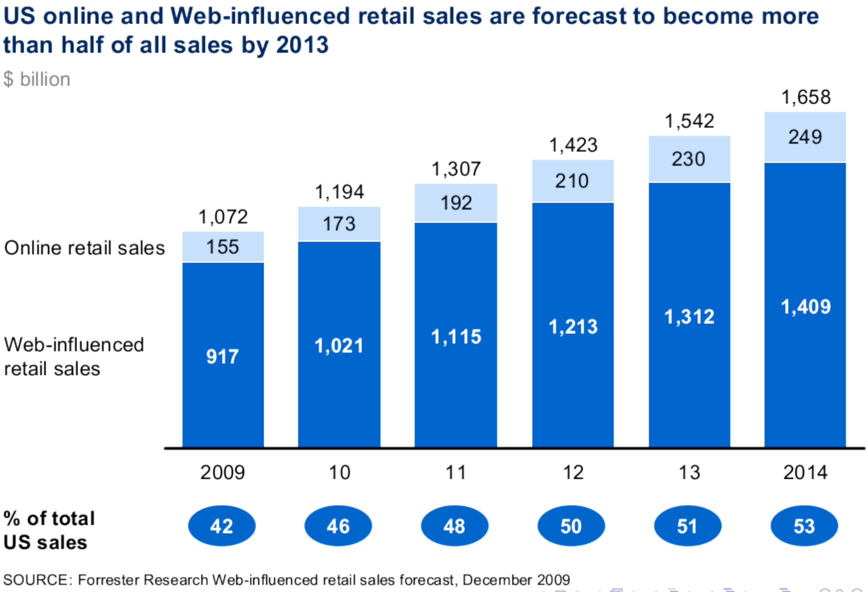
- Creating transparency
- Enabling experimentation to discover needs, expose variability, and improve performance
- Segmenting populations to customize actions
- Replacing/supporting human decision making with automated algorithms
- Innovating new business models, products, and services with big data

Application Areas: Retail

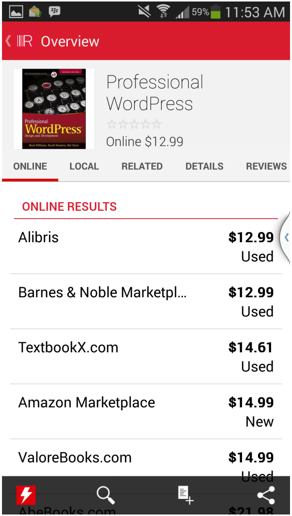
Understanding the Internet

- a major driver for new experiences in retail
- see *"How People Spend Their Time Online"* by GO-Gulf (infographic on a blog)

Retail Pressure



Retail Pressure, cont.



price comparison apps like this (from Ebay) let you be choosier

Retail Applications

Marketing: Cross-selling; Location based marketing; In-store behavior analysis; Customer micro-segmentation; Sentiment analysis; Enhancing the multichannel; consumer experience

Merchandising: Assortment optimization; Pricing optimization; Placement and design; optimization

Operations: Performance transparency; Labor inputs optimization;

Supply chain: Inventory management; Distribution and logistics; optimization; Informing supplier negotiations

Application Areas: Manufacturing

Operations Analysis

The abundance and growth of machine data, which can include anything from IT machines to sensors and meters and GPS devices, is another major driver of big data solutions. In its raw format, many organizations are unable to leverage machine data. Yet disregarding this data means that organizations are making business decisions based only on a subset of available information. Leveraging machine data and combining it with existing enterprise data enables a new generation of applications that are able to analyze and gain insight from large volumes of multi-structured machine data—which in turn improves business results.

Raw Logs & Machine Data

Enterprise Data

Capture a Complete View
Access large volumes of machine, operational and transactional data and combine with other enterprise data.

Get the Context
Overcome complexities to perform advanced analysis and provide context across different data sets.

Get Insights From Analytics
Release intelligence trapped in your data, allowing agile interpretation and action.

WHAT DO YOU NEED TO SUCCEED?

THE RESULTS

Empower the C-Suite
Reassure decision makers that they are acting with full knowledge & understanding of *all* available data.

Improve Reliability
Perform root cause analysis on data to more easily identify and preempt system failures, keeping customers happy.

Speed Operations
Help departments proactively minimize the problems and bottlenecks that stymie the flow of operations.

Monitor & React
Visualize streaming data to monitor the end-to-end infrastructure and deliver real-time alerts.

Learn more at **IBM.com/BigData**

Manufacturing Applications

We have identified the following big data levers across the manufacturing value chain

	R&D and design	Supply-chain mgmt	Produc-tion	Market-ing and sales	After-sales service
1 Build consistent **interoperable, cross-functional R&D and product design databases** along supply chain to enable concurrent engineering, rapid experimentation and simulation, and co-creation	✓				
2 **Aggregate customer data and make them widely available** to improve service level, capture cross- and up-selling opportunities, and enable **design-to-value**	✓			✓	
3 Source and share data through **virtual collaboration sites** (**idea marketplaces** to enable crowd sourcing)	✓			✓	
4 Implement advanced **demand forecasting and supply planning** across suppliers and using external variables		✓	✓	✓	
5 Implement **lean manufacturing and model production virtually (digital factory)** to create process transparency, develop dashboards, and visualize bottlenecks			✓		
6 Implement **sensor data-driven operations analytics** to improve throughput and enable mass customization			✓		
7 Collect **after-sales data from sensors** and feed back in real time to trigger after-sales services and detect manufacturing or design flaws			✓	✓	✓

SOURCE: McKinsey Global Institute analysis

Application Areas: Location

Mobile location-based services (LBS) and applications have proliferated

Mobile LBS applications continue to proliferate[1]

People locating (e.g., safety family/ child tracking, friend finder)	at&t FamilyMap; Pink Map
Location check-in/ sharing on social community applications	foursquare; loopt; Places — Who. What. When. And now **Where.**
City/regional guide, neighborhood service search	 LON lonely planet
Location-enabled entertainment, e.g., mobile gaming, geo-tagged photo/travel	GEOCACHING.COM; *Geo-tagged photo*

Revenue through the "Freemium" model

Revenue model for these mobile LBS applications will be a mix of

- Free services/applications supported by advertising revenue
  - Sponsor links for mobile location-enabled (e.g., nearby point of interest) search
  - Advertising embedded in mobile applications
- Mobile apps requiring premiums for download or subscription
  - Onetime charge to download apps from mobile marketplaces
  - Recurring subscription fees for services/content
  - Add-on charges, e.g., purchase of virtual items in mobile games

1 Navigation and other applications for non-individual usage have been assessed separately and are not included here.

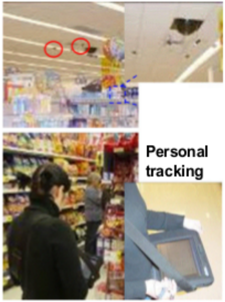
SOURCE: Press search; McKinsey Global Institute analysis

How does location tracking work?

Input | **Output**

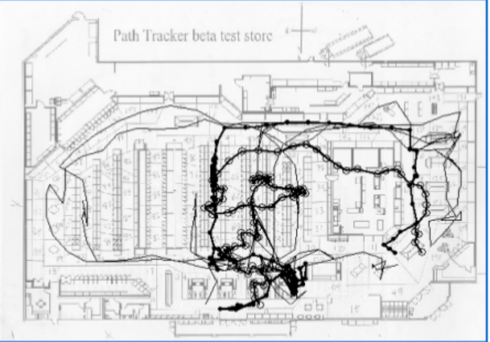
RFID Tag

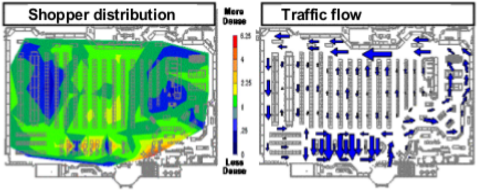
+ Mobile phones



Video

Personal tracking





SOURCE: Press and literature search

Location Applications

Location-based applications and services for individuals:

- smart routing
- automotive telematics
- mobile phone location-based services

Organizational use of individual personal location data:

- geo-targeted advertising
- electronic toll collection
- insurance pricing
- emergency response

Macro-level use of aggregate location data:

- urban planning.
- retail business intelligence
- some new business models

Next: Module 3

Data Types and Storage