CIS8695
Big Data Analytics

Ling Xue

Computer Information Systems
J Mack Robinson College of Business
Georgia State University

Data Analytics

Georgia State, Leading U.S. in Black Graduates, Is Engine of Social Mobility

Georgia State, once seen as a night school for white businessmen, has
and a raft of data-driven

‹ Back to New at McKinsey Blog

How McKinsey volunteers used data to help the children of Atlanta

Atlanta Braves Win Horizon Summit Data Visualization Contest

GSU: Data Analytics For Graduating Students

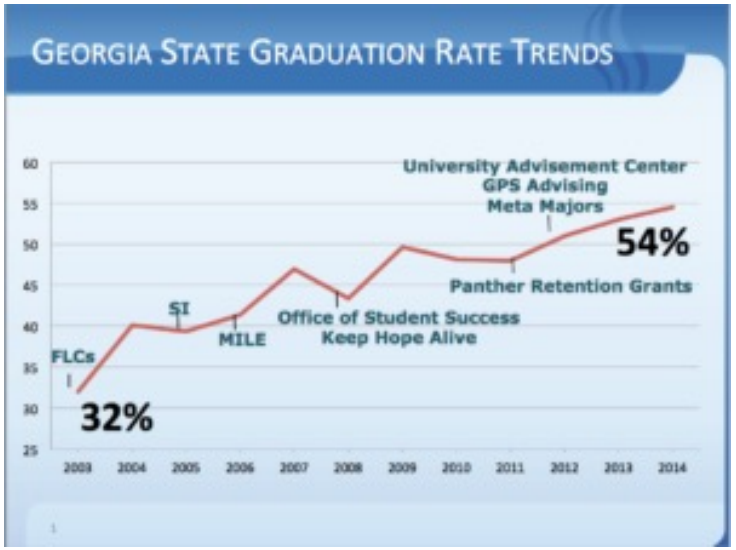
- State of Georgia: 10 worst for graduating black males from high school (Schott Foundation for Public Education 2015)
- GSU:
  - founded "as a night school for white businessmen," (a college spokeswoman). The school remained segregated until the 1960s.
  - 50,000+ students, 60% are nonwhite, and many are from working-class and first-generation families
- As of 2016, there were only 92 African-American faculty members totaled out of 1,624 part and full-time positions at Georgia State. And non-instructional African-American personnel amounted to 1,842 out of 3,437 positions. So with, Black people comprising 39 percent of GSU's consumer base, having only 11 percent of faculty doesn't match up
- Limited resources: 700 students per advisor

GSU: Data Analytics For Graduating Students

- Working with the help of an outside consulting firm, EAB, GSU analyzed 2.5 million grades earned by students in courses over 10 years to create a list of factors that hurt chances for graduation.
- EAB then built an early-warning system, which GSU calls GPS, for Graduation and Progression Success.
- The system is updated daily and includes more than 700 red flags aimed at helping advisers keep students on track to graduation
- E.g., an adviser gets an alert when:
  - A student does not receive a satisfactory grade in a course needed in his or her major.
  - A student does not take a required course within the recommended time.
  - A student signs up for a class not relevant to his or her major.
- Hired additional academic advisers, bringing its caseload down to 300:1. 90% of the cost of the project was in staffing. The system prompted 51,000 in-person meetings between students and advisers annually, 3-4 times more than was happening before.

GSU Data Analytics For Graduating Students: Results

- Graduation rates are up 6 percentage points since 2013.
- Graduates are getting that degree an average half a semester sooner than before, saving an estimated $12 million in tuition.
- Low-income, first-generation and minority students have closed the graduation rate gap.
- And those same students are succeeding at higher rates in tough STEM majors.



- Challenges: the possibility of discrimination, invasions of privacy and groups of students being stigmatized. Lack of transparency when decision-making is turned over to an opaque computer program.

Trends: Data, Analytics, ML, DL, AI

- Improvements in computing power and capacity
- Explosion of data
- Progress in algorithms
- Proliferation of software tools

Analytics, ML, DL, AI

- Making sense of data → Analytics
- Scaling up analytics tasks, improving efficiency → Machine Learning
- Improving accuracy → Deep Learning
- Automation of learning and actions → Artificial Intelligence

- All about making machine "smart" to handle data

Definition of Big Data: 4Vs

- **Volume: amount of data**
  - Transaction-based data stored through the years.
  - Unstructured data streaming in from social media or smartphone, etc.
  - Sensor and machine-to-machine data being collected (e.g, IoT).
- **Variety: different types of data being generated (currency, dates, numbers, text, etc.)**
  - Data today comes in all types of formats.
  - Structured, numeric data in traditional databases.
  - Information created from line-of-business applications.
  - Unstructured text documents, email, video, audio, stock ticker data and financial transactions.
  - Managing, merging and governing different varieties of data is very complex.
- **Velocity: flow rate-the speed at which it is being generated and changed**
  - Data is streaming in at unprecedented speed and must be dealt with in a timely manner.
  - To be useful, we need to deal with this large amount of data in near-real time.
  - Reacting quickly enough to deal with data velocity is a big challenge.
- **Veracity: Variability in data content**
  - data flows can be highly inconsistent with periodic peaks.
  - Is something trending in social media?
  - Daily, seasonal and event-triggered peak data loads can be challenging to manage.

Business Challenges

- Most companies are capturing only a fraction of the hugh potential value from data
- Organizational barriers for extracting data value, e.g.,
  - Struggle to incorporate data-driven insights into day-to-day business processes
  - Challenge to attract and retain the right talent
- Data, analytics, ML and AI are changing the basis of competition
- Data, analytics, ML and AI underpin disruptive business models
- Data, analytics, ML and AI are pushing the boundaries
  - New systems enabled by ML fosters greater automation
  - Breakthroughs in the processing of natural languages, images, videos, etc.
  - Deep learning may replace human work activities

Data & Competitive Advantages

1. How much value is added by data relative to the stand-alone value of product?
2. How fast can the insights from data be incorporated into products?

3. How quickly does the marginal value of data-enabled learning drop off?
4. How fast does the relevance of the data depreciate?

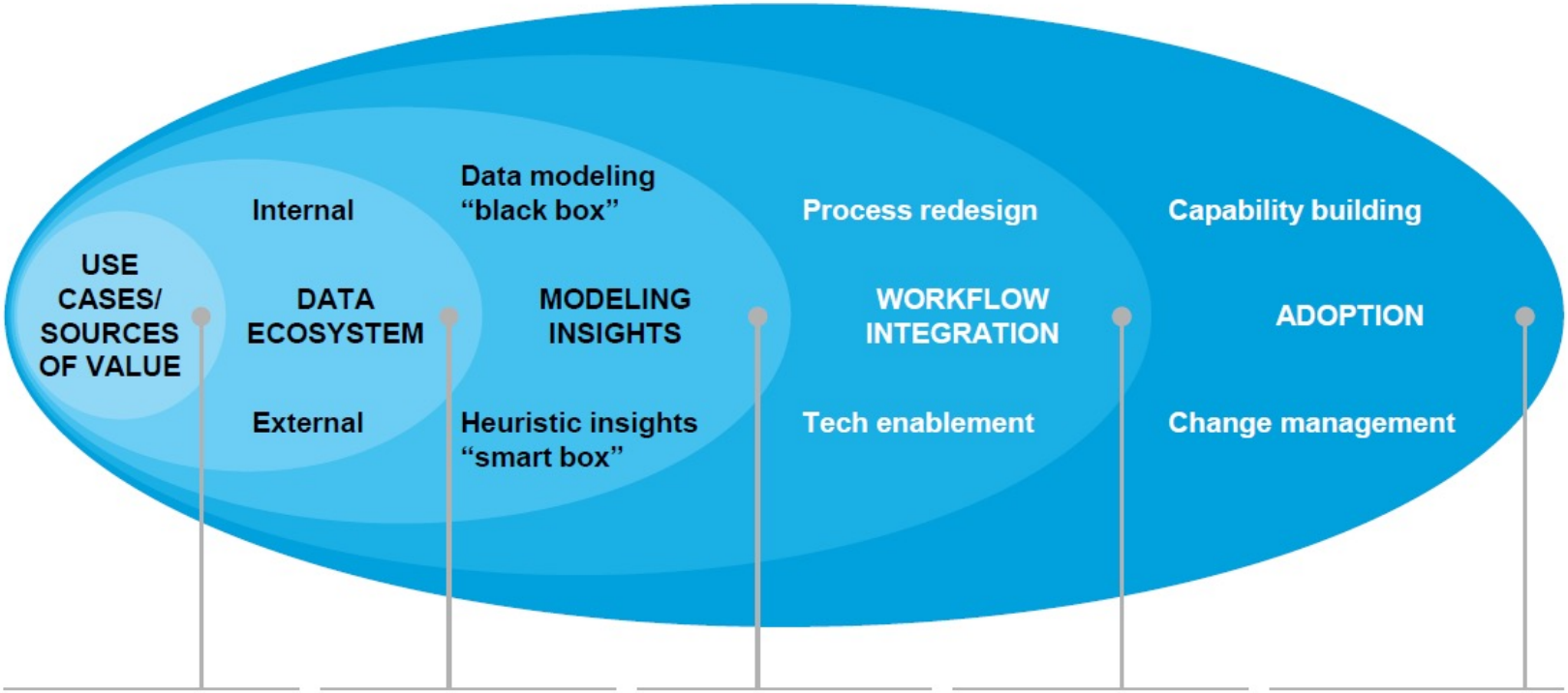
5. Is the data proprietary, meaning it can't be purchased from other sources, easily copied, or reversed-engineered?
6. How hard is it to imitate product improvements that are based on data?

7. Does the data from one user help improve the product for the same user or for others?

Cutting-Edge Technical Challenges: AI

- Tremendous human efforts needed: e.g., understanding and labeling of data
- Appropriate data for training: domain-specific
- "Blackbox" and interpretability
- Building generalized learning techniques

Successful Data/Analytics Transformation



- Clearly articulating the business need and projected impact
- Outlining a clear vision of how the business would use the solution

- Gathering data from internal systems and external sources
- Appending key external data
- Creating an analytic "sandbox"
- Enhancing data (deriving new predictor variables)

- Applying linear and nonlinear modeling to derive new insights
- Codifying and testing heuristics across the organization (informing predictor variables)

- Redesigning processes
- Developing an intuitive user interface that is integrated into day-to-day workflow
- Automating workflows

- Building frontline and management capabilities
- Proactively managing change and tracking adoption with performance indicators

SOURCE: McKinsey Analytics; McKinsey Global Institute analysis

Analytics Talent Market

- Average wage for analytics-related jobs increases 16% annually, versus 2% increase for other jobs
- A shortfall of
  - **Data scientists**
  - **Business translators**
- Data scientists
  - While data science graduates increase by 7% annually, demand increases by 12%
  - A shortfall of approximate 250,000 in the near future
  - A countervailing force: data preparation work will largely be automated
- Business translators
  - Data savvy, knowledgeable about functions, organizations, and industries; **be able to ask the right questions and derive right insights**; combine a strong understanding of data with user interface, user experience, and visualization skills for decision-making; not outsourceable
  - A need of 2-4 millions in the next decade
  - Current graduation rate can only meet a need of 1 million

Data Ecosystems

- Not all data is equal
  - Behavioral data, transactional data, environmental data, geospatial data, references materials & knowledge, public records, etc.
  - Data comes from web, social media, mobile/wearable devices, sensors, payment systems, cameras, human entry, etc.
- Three categories of players in data ecosystems
  - Data generation and collection: hold value when access is limited (e.g., physical barriers)
  - Data aggregation: adds value when combining data is technically and organizationally challenging
  - Data analysis: value generated from the **insights; combining analytics tools with business insights**.

Predictive Analytics & Machine Learning

- **Techniques:**
  - Regression (e.g., linear, logistic, ridge, lasso, etc.)
  - Clustering (e.g., k-means, hierarchical)
  - Association rules
  - Dimensionality reduction
  - Classification trees
  - Support vector machines
  - Conventional neural networks
  - Text classification and topic modeling
  - Deep learning networks
- **Problem types**
  - Classification
  - Prediction
  - Generation

Predictive Analytics & Machine Learning

Classification	Classify/label visual objects	Identify objects, faces in images and video
	Classify/label writing and text	Identify letters, symbols, words in writing sample
	Classify/label audio	Classify and label songs from audio samples
	Cluster, group other data	Segment objects (e.g., customers, product features) into categories, clusters
	Discover associations	Identify that people who watch certain TV shows also read certain books
Prediction	Predict probability of outcomes	Predict the probability that a customer will choose another provider
	Forecast	Trained on historical data, forecast demand for a product
	Value function estimation	Trained on thousands of games played, predict/estimate rewards from actions from future states for dynamic games
Generation	Generate visual objects	Trained on a set of artist's paintings, generate a new painting in the same style
	Generate writing and text	Trained on a historical text, fill in missing parts of a single page
	Generate audio	Generate a new potential recording in the same style/genre
	Generate other data	Trained on certain countries' weather data, fill in missing data points for countries with low data quality

SOURCE: McKinsey Global Institute analysis

Types of Analytics

- Descriptive: "What has happened?"
  - Query/drill down: What exactly is happening?
  - Ad hoc reports: How many, how often, where?
  - Standard reports
- Predictive: "What could happen?"
- Prescriptive: "What's the best outcome given our descriptive and predictive analytics?"

Data Preparation: Issues

- Types of variables:
  - Categorical variables (numerical or text): nominal variable; ordinal variable
  - Continuous variables
- Variable conversion to ***dummies***:
- Variable selection (sometimes requires dimension reduction);
- Determining sample size to use;
- Handling outliers;
- Handling missing values;
- Normalizing (standardizing) and rescaling variables.

Supervised vs Unsupervised Learning

- **Supervised Learning:**
  - Outcome of interest is known, "labeled data": target variable, outcome variable, dependent variable
  - **Training data**: from which the model/algorithm "learns"
  - **Validation data**: where the trained model is applied to for assessment of performance, and model comparison
  - **Test data**: if model selection is involved, test data is used to assess the finally selected model
    - The Outcome in validation data is also known
- **Unsupervised Learning:**
  - No outcome to be classified or predicted
  - Let the data tells the story (patterns) by itself, and no predetermined patterns to be learned from training data

Bias & Fairness

- AI can help identify and reduce the impact of human biases
- But it can also make the problem worse by baking in and deploying biases

occupation	bias
maid	59.2
waitress	52.5
midwife	50.9
receptionist	50.2
nanny	47.7
nurse	45.4
midwives	43.8
housekeeper	36.6
hostess	32
gynecologist	31.6

occupation	bias
undertaker	-73.4
janitor	-62.3
referee	-60.7
plumber	-58
actor	-56.9
philosopher	-56.2
barber	-55.4
umpire	-54.3
president	-54
coach	-53.8

- Bias can creep into AI algorithms
  - Training data may capture human biases
  - Flawed data sampling can cause biases
  - Biases may be embedded in algorithms
  - Indirect biases may be caused by the uses of attributes correlated with sensitive attributes

To Correct Biases

- Use AI to correct human biases
  - Use ML to improve predictive outcomes and select appropriate variables
  - Use AI to probe algorithms for biases
  - Use AI to focus more on the traditionally disadvantaged groups

- Address biases in AI
  - Understand and measure fairness
  - Use techniques to ensure that AI meets criteria of fairness
    - Processing data beforehand
    - Altering systems' decisions afterwards
    - Incorporating fairness into the training process
  - Incorporate multi-disciplinary perspectives, including from scholars of ethics, social scientists, and other humanities thinkers.
  - Humans and machines work together to mitigate biases. Human double-check. Results comparison. Explainability techniques.

Reasons That ML Can Go Wrong

1. Prediction is based on ***probability***, there is always a chance that it will be wrong
2. Environment can evolve and differ from what the algorithms were developed to face
   - Concept drift: the relationship between inputs and outputs is not stable over time
   - Covariate shift: the data fed into an algorithm during its use differs from the data that trained it
3. The system that ML is embedded in can be complex
4. Difficulty to deal with agency risks: stemming from things that aren't under the control of a specific business or user
5. Difficulty to deal with ***responsible algorithm design***: How should businesses balance trade-offs among, say, privacy, fairness, accuracy, and security?

Cope With ML Risks

1. Treat machine learning as if it's human
2. Think like a regulator and certify first
3. Monitor continuously
4. Deal with Open questions:
    1. Allow learning to continuously evolve or to "**lock**" their algorithms and periodically update them?
    2. What are **biases**? E.g., What data was used to train the algorithm? How representative is it of the population on which the algorithm will ultimately operate? Can we predict whether an unlocked algorithm will produce less-biased results than a locked one if we allow it to learn over time?
    3. How will the **environment** in which the offering is used change over time?
    4. On which third-party **agents**, including data sources, does the behavior of our machine-learning algorithms depend?
5. Develop principles that address your business risks