BIG DATA

POSTACADEMISCHE OPLEIDING

7 maart 2018 – 13 juni 2018

UNIVERSITEIT GENT

INLEIDING

'Big Data' kunnen worden omschreven als gegevenscollecties die niet efficiënt met traditionele gegevensbeheer en -verwerkingstechnieken kunnen worden behandeld. Bepalende factoren daarbij zijn de grotere datavolumes, de grotere snelheden waarmee de data worden aangeboden en de grotere variëteit aan dataformaten en de kwaliteit van de data. De tendens naar `Big Data' wordt gevoed door de almaar groeiende beschikbaarheid van digitale informatie uit nieuwsbronnen, multimedia, sensors, ... en gaat gepaard met nieuwe uitdagingen om deze data efficiënt te kunnen verzamelen, opslaan, beheren, analyseren en presenteren.

Het inzetten van geavanceerde technologieën die specifiek zijn afgestemd op het verwerken van zeer grote hoeveelheden data, kan bedrijven helpen om beter tegemoet te komen aan de steeds groter wordende informatienoden die vaak vereist zijn om gegevensanalyse nog beter te kunnen onderbouwen. Een beter inzicht in de beschikbare data en een optimale exploitatie ervan levert de beste garantie om met meer kennis van zaken belangrijke beslissingen te onderbouwen en daar dan ook een concurrentieel voordeel mee te behalen.

GETUIGSCHRIFT

U ontvangt een getuigschrift, indien u deelneemt aan de volledige opleiding en slaagt voor het bijbehorende examen.

DOELPUBLIEK

U krijgt inzicht in de problematiek die gepaard gaat met 'Big Data' en in de beschikbare ICT-oplossingen die momenteel voorhanden zijn. Er wordt aangetoond hoe de aangereikte oplossingen werken, wat hun beperkingen en voordelen zijn en waar en wanneer ze het beste kunnen worden ingezet.
Voor de lessen wordt bewust gekozen voor een sterke academische aanpak waarbij de accenten liggen op het verwerven van kennis in de breedte zonder daarbij productgebonden te zijn. Deze lessen worden aangevuld met een aantal gastpresentaties waarbij aandacht wordt besteed aan praktische voorbeelden.

De opleiding is dusdanig opgevat dat deze toegankelijk is voor iedereen die ietwat vertrouwd is met informatica. Er wordt gewerkt rond hoorcolleges die handelen rond vier thema's: gegevensbeheer, gegevensanalyse, visualisatie en ethische en juridische aspecten. Deze vier thema's worden aangevuld met twee praktische getuigenissen met demo's.

WETENSCHAPPELIJKE COÖRDINATIE

Prof. dr. Guy De Tré, Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent

LESGEVERS

- Antoon Bronselaer, Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent
- Hans Constandt, Ontoforce
- Peter Craddock, DLA Piper
- Thomas Demeester, Vakgroep Informatietechnologie, Universiteit Gent
- Wesley De Neve, Vakgroep Elektronica en Informatiesystemen, Universiteit Gent
- Guy De Tré, Vakgroep Telecommunicatie en Informatieverwerking, Universiteit Gent
- Jan Fostier, Vakgroep Informatietechnologie, Universiteit Gent
- Fréderic Godin, Vakgroep Elektronica en Informatiesystemen, Universiteit Gent
- Alain Houf, Intersystems
- Peter Lambert, Vakgroep Elektronica en Informatiesystemen, Universiteit Gent
- Raf Schoefs, DLA Piper
- Pieter Vandekerckhove, Add Perspective
- Dirk Van den Poel, Vakgroep Marketing, Universiteit Gent
- Baptist Vandersmissen, Vakgroep Elektronica en Informatiesystemen, Universiteit Gent
- Katrien Verbert, Departement Computerwetenschappen, KU Leuven
- Ruben Verborgh, Vakgroep Elektronica en Informatiesystemen, Universiteit Gent
- Bruno Volckaert, Vakgroep Informatietechnologie, Universiteit Gent

PROGRAMMA

1. GEGEVENSBEHEER

Inleiding en NoSQL
In de introductie wordt aandacht besteed aan de oorsprong van de term 'Big Data'. Aspecten zoals de interpretatie, het belang, de problematiek en de kritiek op 'Big Data' worden besproken. Daarna komen de verschillende vormen en karakteristieken (Volume, Variety, Velocity en Veracity) van 'Big Data' aan bod.
Er wordt gekeken naar de tekortkomingen en beperkingen van traditionele databanksystemen en er wordt dieper ingegaan op mogelijke oplossingen. Vervolgens worden de belangrijkste NoSQL databankoplossingen ('Not only' SQL) gesitueerd en bestudeerd. Zowel key/value stores, documentdatabanken, column stores als graafdatabanken worden daarbij behandeld.

Lesgever: Guy De Tré
Datum: 7 maart 2018

Datakwaliteit
Er wordt vertrokken vanuit de vaststelling dat modaliteiten voor de garanties van datakwaliteit bij NoSQL opslagsystemen van ondergeschikt belang zijn ten voordele van een betere throughput. De kwaliteitscontrole van data wordt dan ook verschoven naar het niveau van applicaties die gebruik maken van data.
Er wordt toegelicht hoe men in een applicatie een dergelijk controlemechanisme kan inbouwen op een efficiënte manier. Er wordt overlopen hoe meting van data kwaliteit wordt aangepakt in de context van Big Data. Hierbij wordt vertrokken van het concept "meten". De formele aspecten van dit concept worden toegelicht in de context van data kwaliteit. Bijzondere aandacht gaat uit naar situaties waar meten wordt bemoeilijkt door onzekerheid over de data. Tot slot wordt een overzicht gegeven van de problematiek van data ontdubbeling.

Lesgever: Antoon Bronselaer
Datum: 14 maart 2018

Linked Open Data
De evolutie van het World Wide Web tot een globaal en wereldwijd platform voor de meest uiteenlopende digitale diensten, heeft er toe geleid dat een goede representatie van data op het web steeds belangrijker wordt.
Tijdens deze lesavond worden verschillende technologieën behandeld met het oog op het machine-leesbaar maken van data en informatie op het web (cf. semantisch web). Het koppelen van verschillende databronnen (Linked Data) zal daar een belangrijke rol in spelen. Daarnaast worden ook de principes van Open Data besproken, met als doel het uniform beschikbaar stellen van data via het web. Linked Open Data leidt aldus tot een aantal specifieke uitdagingen op het vlak van web-gebaseerde data-analyse en interpretatie.

Lesgevers: Peter Lambert en Ruben Verborgh
Datum: 21 maart 2018

Information retrieval en document classificatie
Eerst wordt er een inleiding gegeven op 'Information Retrieval' (IR), de technologie die toelaat om grote digitale collecties van ongestructureerde teksten efficiënt te doorzoeken. De bedoeling is eerder om een overzicht te geven van de basisconcepten zoals indexering en retrieval modellen, dan om hier heel technisch op in te gaan.
Nadien worden de mogelijkheden bekeken om in dergelijke collecties structuur aan te brengen door de documenten op specifieke manieren te gaan groeperen. Er worden enkele basisconcepten aangebracht rond machinaal leren in het algemeen, en rond classificatie en clustering. Vervolgens worden een paar belangrijke algoritmen besproken, en toegelicht via enkele applicaties.

Lesgever: Thomas Demeester
Datum: 28 maart 2018

2. GEGEVENSANALYSE

Het MapReduce programmeermodel
MapReduce is een generiek raamwerk dat bijzonder geschikt is om op eenvoudige wijze analyses uit te voeren op enorme hoeveelheden data, gebruik makend van een parallel computersysteem. De basisconcepten rond 'Mappers' en 'Reducers' komen aan bod, alsook enkele veelgebruikte ontwerppatronen. Deze laatste worden gestaafd aan de hand van eenvoudig te begrijpen voorbeelden. De Hadoop-implementatie van MapReduce wordt besproken, alsook het verwante Hadoop Distributed File System (HDFS). De doelstelling is dat deelnemers na deze sessie in staat zijn Hadoop MapReduce toe te passen op eigen problemen. Voorkennis van de programmeertaal Java is aan te raden, maar niet noodzakelijk.

Lesgever: Jan Fostier
Datum: 18 april 2018

Gedistribueerde gegevensverwerking
Voor de efficiënte verwerking van Big Data is men grotendeels naar volledig gedistribueerde vormen van gegevensverwerking overgestapt. Tijdens deze lesavond wordt u wegwijs gemaakt in enkele van de belangrijkste architecturen voor gedistribueerde gegevensverwerking (streamverwerking, Lambda architectuur, Delta architectuur, Kappa architectuur, etc.). De opbouw en werking van deze architecturen worden besproken, hoe deze geheel of gedeeltelijk mappen op bestaande technologieën / implementaties (Apache Storm, Apache Hadoop, Apache Samza, Druid, Apache Spark, Apache Kafka, etc.) en wat hun belangrijkste voor- en nadelen zijn. Dit geheel wordt aangevuld met voorbeelden van gedistribueerde architecturen die technologiereuzen zoals LinkedIn, Twitter, etc. geadopteerd hebben om hen om te laten gaan met de enorme hoeveelheid data die ze dagelijks moeten verwerken.

Lesgever: Bruno Volckaert
Datum: 25 april 2018

Deep learning
Kunstmatige neurale netwerken zijn in staat om het menselijk leerproces na te bootsen door het veranderen van de sterkte van gesimuleerde neurale verbindingen, een eigenschap die ervoor zorgt dat deze netwerken uiterst effectief zijn in het automatisch terugvinden van patronen in grote hoeveelheden data (deep learning). Dit heeft onlangs geleid tot een aantal doorbraken op het vlak van taalverwerking en audiovisuele analyse.
Voortbouwend op een aantal basisconcepten uit het domein van machinaal leren, wordt er in deze les bijzondere aandacht besteed aan het gebruik van meerlagige neurale netwerken, alsook aan de technieken die de inzet van deze netwerkarchitecturen praktisch haalbaar hebben gemaakt. Vervolgens wordt er stilgestaan bij verschillende toepassingen op het vlak van taalverwerking en audiovisuele analyse, illustrerend hoe meerlagige neurale netwerken kunnen aangewend worden om kennis te extraheren uit grote hoeveelheden ruizige data. Tot slot wordt er eveneens een overzicht gegeven van toekomstige uitdagingen op het vlak van onderzoek en ontwikkeling in het domein van deep learning.

Lesgevers: Wesley De Neve, Fréderic Godin en Baptist Vandersmissen
Datum: 2 mei 2018

Big Data Analytics
Volgende onderdelen worden behandeld:
- Inleiding tot 'Analytics' ('Descriptive', 'predictive' en 'prescriptive analytics')
- De Spark (Streaming) software stack (Berkeley Data Analytics Stack (BDAS))
- Gebruik van MLlib (machine learning library) binnen Spark voor 'Analytics'
- Bespreking van enkele gebruikscases van de Spark software stack voor 'Analytics'
- Hands-on demo van de UGent-implementatie van de Berkeley open source software stack.

Lesgever: Dirk Van den Poel
Datum: 9 mei 2018

3. VISUALISATIE

Getuigenissen uit de praktijk
De eerste spreker is Hans Constandt, CEO en medeoprichter van ONTOFORCE. ONTOFORCE is een Belgisch bedrijf dat oplossingen aanbiedt voor 'information flow' en management. In deze lezing wordt een gebruiksvriendelijk data zoekplatform voorgesteld dat gedreven is door semantische technologie en in staat is om intelligente links te bouwen tussen de steeds groeiende hoeveelheid van interne en externe gegevens. Tevens wordt een toepassing voor de farmaceutische industrie toegelicht.
In de tweede lezing zal Alain Houf, sales engineer bij Intersystems, de NoSQL oplossing van Intersystems toelichten en onder andere illustreren aan de hand van een case-study over de Gaia-missie van de Europese Ruimtevaartorganisatie ESA die de melkweg nader in kaart moet brengen.

Datum: 16 mei 2018

Visualisatie
Een uitgelezen manier om mensen te helpen om Big Data te exploreren en te begrijpen, is het visualiseren van de data: we zijn immers vaak erg goed in staat om patronen, tendensen, uitschieters, … te begrijpen met behulp van visualisaties.
Mede aan de hand van een groot aantal concrete voorbeelden wordt uiteengezet hoe een goede interactieve visualisatie kan worden opgebouwd, wat de typische misvattingen zijn, hoe visualisaties kunnen misbruikt worden, … Er wordt ook een overzicht gepresenteerd van een aantal typische technieken en hulpmiddelen voor interactieve informatie-visualisatie.

Lesgever: Katrien Verbert
Datum: 23 mei 2018

4. ETHISCHE EN JURIDISCHE ASPECTEN

Ethische aspecten
Big Data is goud waard in een kenniseconomie maar wat als de economie zelf draait op toepassingen van Big Data?
Waarom wordt Big Data nog steeds als product gezien, terwijl de toepassingen oneindig zijn als het als een nieuw periodiek element beschouwd wordt?
Big Data is deel van onze meest menselijke en intieme sfeer. Het 2013 Onlife Manifesto van de Europese Commissie getuigt hierover. Een nieuwe maatschappelijke visie is essentieel voor pioniers.
Tijdens dit college staan de volgende leerdoelen voorop:
– Begrijpen dat Big Data niet ethisch neutraal is en hoe ontwikkelaars bepaalde waarden mee "programmeren"
– Verschillende ethische perspectieven kunnen toepassen op Big Data technologie om de onderliggende waarden kritisch te analyseren
– Een filosofische tool kunnen toepassen om ervoor te zorgen dat Big Data technologie ontwikkeld wordt zodat gebruikers een ethisch bewustzijn ontwikkelen

Lesgever: Pieter Vandekerckhove
Datum: 30 mei 2018

Juridische aspecten
U wordt wegwijs gemaakt in de juridische uitdagingen rond Big Data projecten zoals privacy, concurrentierecht, eigendomsrechten, discriminatie en andere topics.
U krijgt tevens een aantal vuistregels omtrent het op een juridisch correcte manier aanvatten en uitvoeren van Big Data projecten.

Lesgevers: Peter Craddock en Raf Schoefs
Data: 6 en 13 juni 2018

MEER INFO EN INSCHRIJVEN
www.ugain.ugent.be/bigdata

PRAKTISCH

PRIJS

Deelnameprijs omvat lesgeld, hand-outs, frisdranken, koffie en broodjes. Betaling geschiedt na ontvangst van de factuur. Alle facturen zijn betaalbaar dertig dagen na dagtekening. Alle vermelde bedragen zijn vrij van BTW.
Voor iedere module kan er afzonderlijk ingeschreven worden.

Module 1: Gegevensbeheer	€ 600
Module 2: Gegevensanalyse	€ 600
Module 3: Visualisatie	€ 300
Module 4: Ethische en juridische aspecten	€ 450
Volledige opleiding	**€ 1.755**

KORTING

- Indien minstens één deelnemer van een bedrijf inschrijft voor de volledige opleiding, wordt voor alle bijkomende gelijktijdige inschrijvingen van hetzelfde bedrijf een korting van 20% verleend. Facturatie geschiedt dan d.m.v. een gezamenlijke factuur.
- 10% korting op de in de tabel vermelde prijzen voor leden AIG, VBIG en Agoria.
- Aangepaste prijzen voor personeel van UGent en geassocieerde hogescholen.
- Kortingen zijn niet cumuleerbaar.

ANNULERING

Raadpleeg onze annulatievoorwaarden op www.ivpv.ugent.be/annulatievoorwaarden

KMO-PORTEFEUILLE

Universiteit Gent aanvaardt betalingen via de KMO-portefeuille (www.kmo-portefeuille.be; gebruik authorisatiecode DV.0103194).

TIJDSTIP EN LOCATIE

- De lessen worden gegeven van **18u tot 21u30**, in 2 delen, gescheiden door een broodjesmaaltijd en vinden plaats aan de **Universiteit Gent, UGent Academie voor Ingenieurs, Technologiepark 904, 9052 Zwijnaarde.**
- Data onder voorbehoud van wijzigingen om onvoorziene omstandigheden.

MEER INFO EN INSCHRIJVEN

www.ugain.ugent.be/bigdata

ORGANISATIE
Universiteit Gent
UGain (UGent Academie voor Ingenieurs)
Technologiepark 904, 9052 Zwijnaarde
Tel: +32 9 264 55 82
E-mail: ugain@ugent.be

UNIVERSITEIT GENT

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN

FACULTEIT INGENIEURSWETENSCHAPPEN EN ARCHITECTUUR