01MBID Fundamentos de la tecnología Big Data

Universidad Internacional de Valencia

Sesión 7

De:
Planeta Formación y Universidades

Agenda

- **Encuesta**
- **Fechas**
- **Tema 4: Beneficios y Riesgos**
- **Tema 5: Criterios de calidad**
- **Clase de dudas extra**
- **Dudas**

Agenda

- **Encuesta**
- **Fechas**
- **Tema 4: Beneficios y Riesgos**
- **Tema 5: Criterios de calidad**
- **Clase de dudas extra**
- **Dudas**

Encuesta

Agenda

- Encuesta
- **Fechas**
- **Tema 4: Beneficios y Riesgos**
- **Tema 5: Criterios de calidad**
- **Clase de dudas extra**
- **Dudas**

Fechas

1ª Convocatoria

Examen: 11/11/2024 de 12:00 a 14:00 o de 20:00 a 22:00
Actividad: ~~10/11/2024~~ 17/11/2024 a las 23:59h
Foro: 10/11/2024 a las 23:59h
AEC: 30/10/2024 de 00:00 a 23:59 (20 minutos, 10 preguntas)

Casos justificados hasta mañana 07/11/2024.

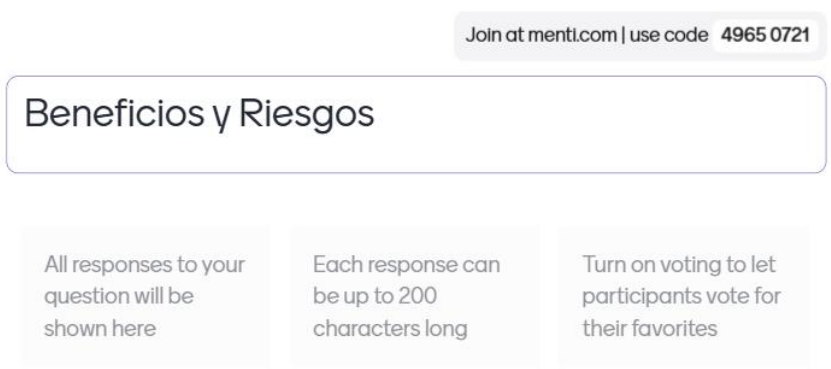
2ª Convocatoria

Examen: 24/03/2025 de 12:00 a 14:00 o de 20:00 a 22:00
Actividad: 23/03/2025 a las 23:59h
No hay foro ni AEC

Agenda

- Encuesta
- Fechas
- **Tema 4: Beneficios y Riesgos**
- **Tema 5: Criterios de calidad**
- **Clase de dudas extra**
- **Dudas**

Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos



Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

Response summary

Beneficios y Riesgos

- Beneficios: Automatización, facilidad de toma decisiones Riesgos: Falta de privacidad y seguridad
- Puede ayudar mucho en areas como Agricultura, Medicina, IoT, como riesgo que puede hacer que tengamos menos privacidad
- hace falta registros en tiempo y forma
- riesgo: seguridad
- Beneficio: mayor procesamiento de transferencia de datos y disponibilidad de los datos.. Riesgos: Datos inexactos o incompletos. Costos elevados.
- Beneficios: Manejar grandes cantidades de datos para toma de decisiones, automatización de tareas, acceso a la información Riesgos: Seguridad de los datos, privacidad, credibilidad de los datos
- riesgo: filtración de datos personales
- Riesgos: No tener el gobierno del dato bien estructurado
- beneficios: poder procesar muchos datos distintos
- Beneficios: mejora de la oferta
- Riesgo: privacidad y preocupación de la ciudadanía
- Riesgo de calidad de los datos, y beneficio: saber tener información de una muy buena parte de la población
- riesgos la calidad de los datos seguridad de los datos benficios mejora y experecia al cliente analizar grandes volumenes de datos reduccion de costos

Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

Beneficios
Decisiones basadas en datos (que nos aportan valor)
Mejorar la toma de decisiones / Optimización de la toma de decisiones
Diferenciarse de los competidores
Beneficios económicos
Conocer a nuestros clientes
“Conocer el futuro” Oráculo
“Tiempo real”
Procesamiento de datos variados (heterogéneos [estructurados/no estructurados/semi])

Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

Beneficios

Aportar Valor, Extraer información relevante, Valor agregado, Obtener información

Rapidez para tomar decisiones o entender necesidades, Descubrir cosas que ni nos preguntábamos, Ayuda a la toma de decisiones, Solucionar problemas

-Tiempo real -Predicción del futuro

Capacidad de procesar gran cantidad de datos

Flexibilidad, Agrupar información heterogénea

Costo/Beneficio

Capaz de manejar grandes volúmenes de datos

Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

Riesgos
Veracidad (necesarios criterios para garantizarla)
Falta de consistencia debido a la velocidad de procesamiento
Privacidad (LOPD) / Almacenamiento inseguro
Inversión inicial / Retorno de la inversión
Calidad de los datos (datos no fiables) (diferentes fuentes / coexistencia)
Volumen + Variedad -> ¿Velocidad?
Resultados **no** superiores a los obtenidos con técnicas tradicionales
Depender de terceros (marketing)

Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

Riesgos
Miedo al cambio
Robo de la información por 3eros, Seguridad, Ciberseguridad, Vulnerabilidad Informática, Monopolio de la informació
Información Falsa, No confiable , Veracidad, Sesgos en los datos
Privacidad, Información sensible, Gobernanza sobre los resultados
Conectividad / Acceso a los datos
Demasiados datos
Escalabilidad

Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

Privacidad

Pedir permiso, Consentimiento informado.

Riesgo para las empresas al no poder acceder a datos relevantes.

Aún tomando medidas para resguardar la privacidad, por ejemplo anonimización de los datos, hay maneras de cruzar los datos anonimizados con otras base de datos para identificar nuevamente los individuos.

Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

Riesgo legales

- Sigue las normas
- Asegúrate que es legal usar
- Se capaz de reproducir los resultados

Más detalles en la asignatura:
Riesgo, seguridad y legislación en sistemas de información

Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

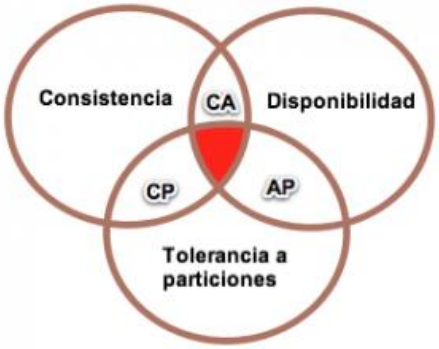
Calidad

Difícil de garantizar por la propia naturaleza de Big Data:

- Datos masivo
- Datos creciendo exponencialmente
- Velocidad de procesamiento

Usar BD públicas, especializadas, que garanticen cierta calidad

Teorema de CAP



Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

Bias / Sesgo

Los datos no son malos pero tenemos que ser concientes de sus características

Tomar medidas en lo posible para corregirlo

Beneficios y riesgos inherentes a la aplicación de técnicas de procesamiento masivo de datos

Bias

Agenda

- Encuesta
- Fechas
- Tema4: Beneficios y Riesgos
- **Tema5: Criterios de calidad**
- **Clase de dudas extra**
- **Dudas**

Criterios de calidad de datos en Big Data

Handbook on Data Quality Assessment Methods and Tools

eurostat

EUROPEAN COMMISSION
EUROSTAT

Table 1: List of Standard Quality Indicators (Eurostat 2005d)

Quality component	Indicator	1=Key 2=Supportive 3=Advanced
Relevance	R1. User satisfaction index	3
	R2. Rate of available statistics	1
Accuracy	A1. Coefficient of variation	1
	A2. Unit response rate (un-weighted/weighted)	2
	A3. Item response rate (un-weighted/weighted)	2
	A4. Imputation rate and ratio	2
	A5. Over-coverage and misclassification rates	2
	A6. Geographical under-coverage ratio	1
	A7. Average size of revisions	1
Timeliness and Punctuality	T1. Punctuality of time schedule of effective publication	1
	T2. Time lag between the end of reference period and the date of first results	1
	T3. Time lag between the end of reference period and the date of the final results	1
Accessibility and Clarity	AC1. Number of publications disseminated and/ or sold	1
	AC2. Number of accesses to databases	1
	AC3. Rate of completeness of metadata information for released statistics.	3
Comparability	C1. Length of comparable time-series	1
	C2. Number of comparable time-series	1
	C3. Rate of differences in concepts and measurement from European norms	3
	C4. Asymmetries for statistics mirror flows	1
Coherence	CH1. Rate of statistics that satisfies the requirements for the main secondary use	3

Criterios de calidad de datos en Big Data

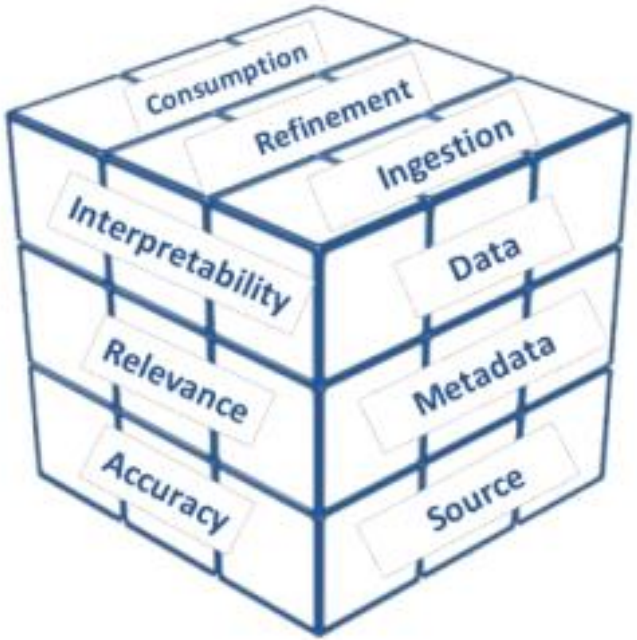
Data Quality Review

Table 1.2 Data quality dimension, metrics and standard benchmarks

DIMENSION 1: COMPLETENESS OF REPORTING An assessment of each dimension should be conducted for each of the recommended core indicators: antenatal care, immunization, HIV, TB and malaria. Additional indicators can be selected according to the priority and focus of the data quality assessment.			
Data quality metric	**Definition**		
	National level		Subnational level
Completeness of district reporting	% of expected district monthly reports (previous 1 year) that are actually received		Number and % of districts that submitted: 1) at least 9 out of 12 expected monthly reports; 2) 100% of expected monthly reports
Timeliness of district reporting	% of submitted district monthly reports (previous 1 year) that are received on time (i.e. by the deadline for reporting)		Number and % of districts that submitted on time at least 75% of the monthly reports received at national level from the district
Completeness of facility reporting	% of expected facility monthly reports (previous 1 year) that are actually received		Number and % of districts with at least 9 out of 12 monthly facility reports received
			Number and % of facilities that submitted 100% of expected monthly reports
Timeliness of facility reporting	% of submitted facility monthly reports (previous 1 year) that are received on time (i.e. by the deadline for reporting)		Number and % of districts that received on time at least 75% of monthly facility reports that were submitted
Completeness of indicator data (% of data elements that are non-zero values, % of data elements that are non-missing values) Carry out each analysis separately	ANC first visit		Number and % of districts with < 90% 1) non-zero values; 2) non-missing values
	3rd dose DTP-containing vaccine		Number and % of districts with < 67% 1) non-zero values; 2) non-missing values
	Currently on ART		Number and % of districts with < 90% 1) non-zero values; 2) non-missing values
	Notified cases of all forms of TB		Number and % of districts with < 75% 1) non-zero values; 2) non-missing values
	Confirmed malaria cases		Number and % of districts with < 90% 1) non-zero values; 2) non-missing values
Consistency of reporting completeness	Each information system	Evaluate the trend in completeness of reporting from district to national level over the past 3 years	Evaluate the trend in completeness from facility to district level over the past 3 years

Criterios de calidad de datos en Big Data

Data Characteristics and Data Usefulness



Interpretability	Data:	How easy to find meaning in the data?
	Metadata:	How easy to understand and use the metadata?
	Source:	How useful is source/supplier documentation & support?
Relevance	Data:	How does the data relate to business information needs?
	Metadata:	Does the metadata help to find and work with the data?
	Source:	Does the source/supplier understand my industry?
Accuracy	Data:	How well does the data represent the real world?
	Metadata:	How well does the metadata describe the data?
	Source:	How trustworthy is the source/supplier of the data?

* https://www.eckerson.com/articles/a-data-quality-framework-for-big-data

Criterios de calidad de datos en Big Data

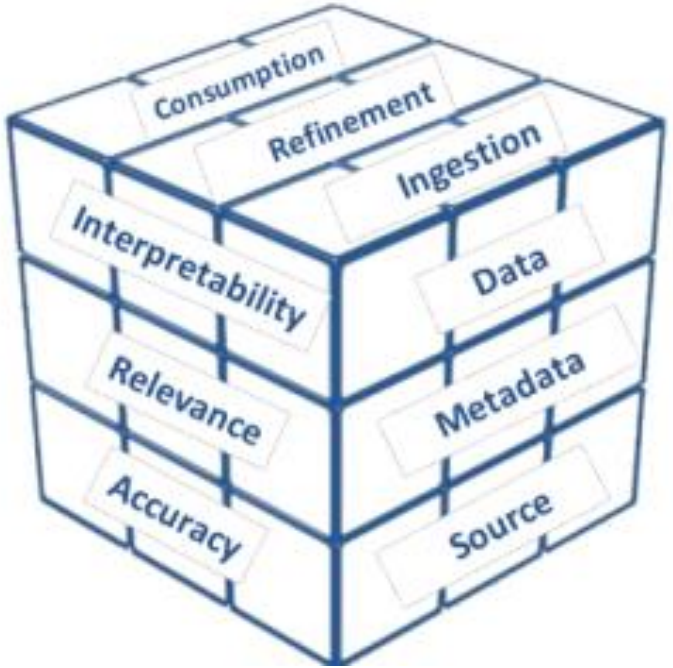
Data Processing and Data Usefulness

Interpretability	Ingestion:	How clear is meaning when data is first received?
	Refinement:	Does processing improve or amplify meaning?
	Consumption:	How clear is meaning at time of use?
Relevance	Ingestion:	Are representative use cases known?
	Refinement:	Does processing add context for the data?
	Consumption:	How well suited is the data to the use case goals?
Accuracy	Ingestion:	How trustworthy is the data when first received?
	Refinement :	Does processing increase or quantify data accuracy?
	Consumption:	How trusted is the data at time of analysis & reporting?

* https://www.eckerson.com/articles/a-data-quality-framework-for-big-data

Criterios de calidad de datos en Big Data

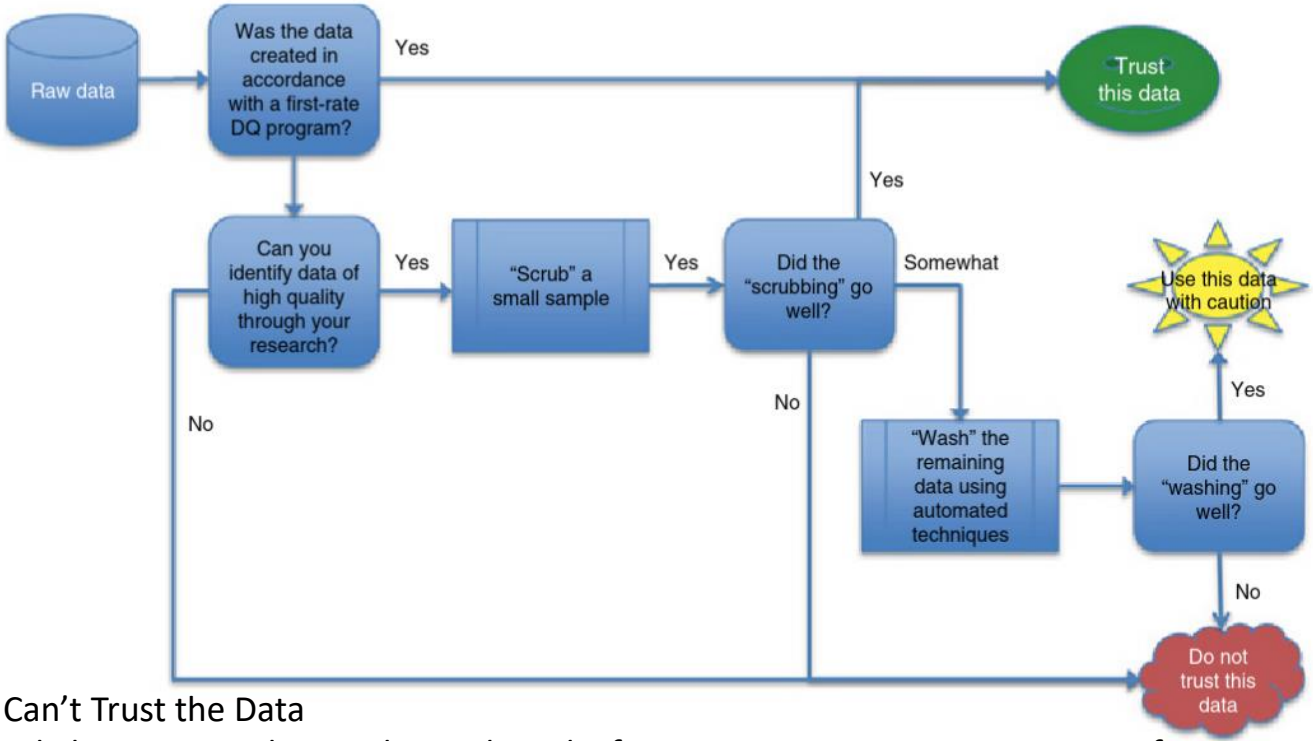
Data Processing and Data Usefulness



Ingestion	Data:	How visible are data patterns and anomalies?
	Metadata:	Does supplied metadata assert or imply quality criteria?
	Source:	Does source/supplier provide quality metrics or estimates?
Refinement	Data:	How well does processing show patterns and anomalies?
	Metadata:	Does the metadata help to estimate or judge data quality?
	Source:	Does the source/supplier provide data quality guidelines?
Consumption	Data:	Is data quality sufficient for the needs of the use case?
	Metadata:	Does metadata help to tune for noise in the data?
	Source:	Does the source/supplier provide quality metrics?

* https://www.eckerson.com/articles/a-data-quality-framework-for-big-data

Criterios de calidad de datos en Big Data



* Sorry, but You Can't Trust the Data

Kenett, Ron. S, and Thomas C. Redman. *The Real Work of Data Science : Turning Data into Information, Better Decisions, and Stronger Organizations*, John Wiley & Sons, Incorporated, 2019. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5741750.

Criterios de calidad de datos en Big Data

Without delving too deeply into details, to be judged of high quality, data must meet three distinct criteria (Redman 2016):

- It must be “right:” correct, properly labeled, deduplicated, and so forth.
- It must be “the right data:” unbiased, comprehensive, relevant to the task at hand.
- It must be “(re)presented in the right way.” For example, people can’t read bar codes, locally used acronyms may confuse others, and so forth.

* Kenett, Ron. S, and Thomas C. Redman. *The Real Work of Data Science : Turning Data into Information, Better Decisions, and Stronger Organizations*, John Wiley & Sons, Incorporated, 2019. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5741750.

Criterios de calidad de datos en Big Data

"Thus, no sector, government agency, or department is immune to the ravages of extremely poor data quality."

* Kenett, Ron. S, and Thomas C. Redman. *The Real Work of Data Science : Turning Data into Information, Better Decisions, and Stronger Organizations*, John Wiley & Sons, Incorporated, 2019. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=5741750.

Criterios de calidad de datos en Big Data

- **Completeness:** The proportion of stored data against the potential of "100% complete";
- **Uniqueness:** Nothing will be recorded more than once based upon how that thing is identified;
- **Timeliness:** The degree to which data represent reality from the required point in time;
- **Validity:** Data are valid if it conforms to the syntax (format, type, range) of its definition;
- **Accuracy:** The degree to which data correctly describes the "real world" object or event being described;
- **Consistency:** The absence of difference, when comparing two or more representations of a thing against a definition.

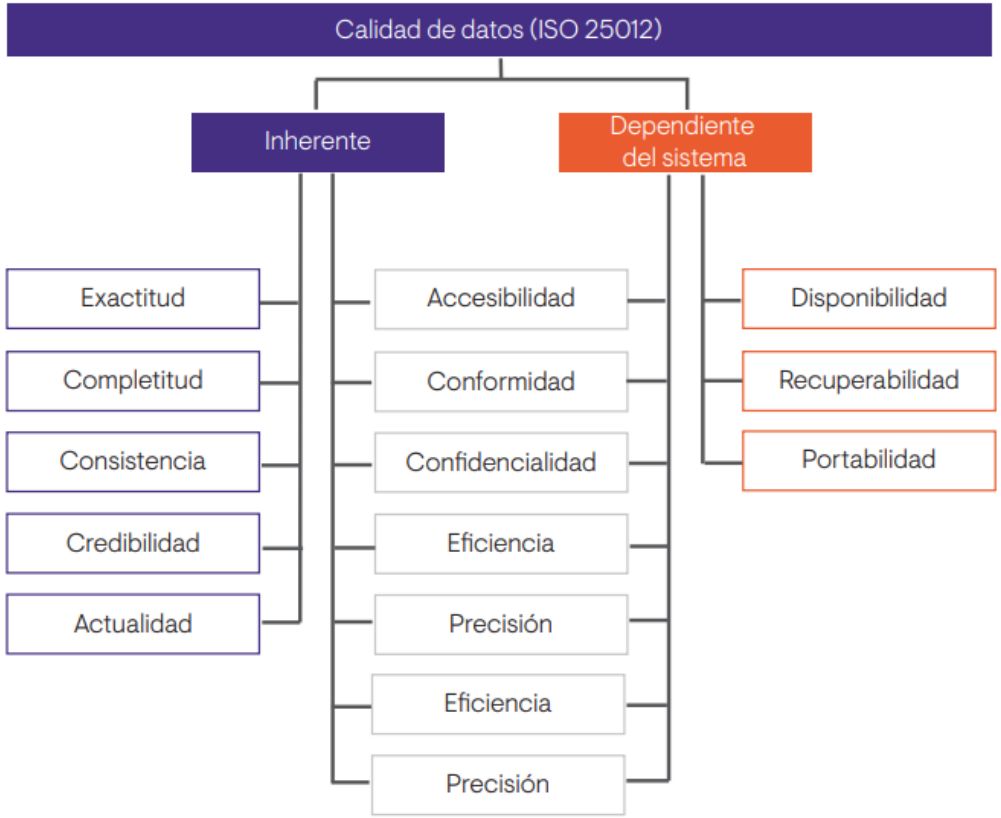
* Ramasamy, Anandhi, and Soumitra Chowdhury. "Big Data Quality Dimensions: A Systematic Literature Review." *JISTEM-Journal of Information Systems and Technology Management* 17 (2020).

Figure 1. Data Quality Dimensions adapted from (DAMA, 2013)

Criterios de calidad de datos en Big Data: ISO 25000, ISO25012

1. **Calidad de datos inherente:** se compone de características propias de los datos que miden el grado de cumplimiento de requisitos establecidos en función de los procesos internos de la organización.
2. **Calidad de datos dependiente del sistema:** en este caso, son atributos de calidad que se evalúan a partir del uso de los datos en un sistema *software* y representan el grado de cumplimiento de los requisitos de calidad con los datos siendo utilizados en tal solución.

Criterios de calidad de datos en Big Data: ISO 25000, ISO25012



*Manual Capítulo 4

Universidad Internacional de Valencia

Agenda

- Encuesta
- Fechas
- Tema 4: Beneficios y Riesgos
- Tema 5: Criterios de calidad
- **Clase de dudas extra**
- **Dudas**

¿ Clase de dudas extra 07/11/2024 ?

¿ Jueves 16:00h – 17:00h ? **España-Península**	**10h en Ecuador, Colombia, Perú, Panamá** **9:00h México - Ciudad de México** **12:00h Chile, Argentina, Brasil-Brasilia** **...**
¿ Jueves a las 02:00h – 03:00h ? **España-Península**	**20h en Ecuador, Colombia, Perú, Panamá** **19h México - Ciudad de México** **22h Argentina, Brasil-Brasilia, Chile** **...**

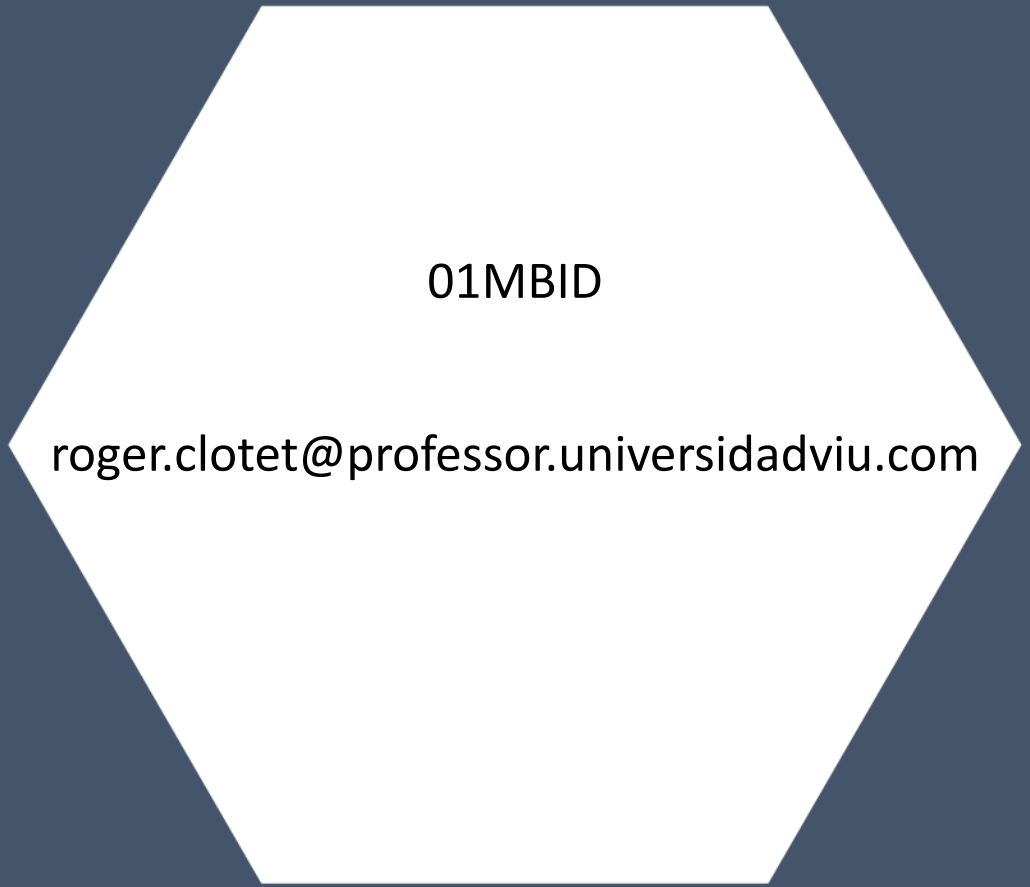
¿ Clase de dudas extra 14/11/2024 ?

¿ Jueves 16:00h – 17:00h ? **España-Península**	**10h en Ecuador, Colombia, Perú, Panamá** **9:00h México - Ciudad de México** **12:00h Chile, Argentina, Brasil-Brasilia** **...**
¿ Jueves a las 02:00h – 03:00h ? **España-Península**	**20h en Ecuador, Colombia, Perú, Panamá** **19h México - Ciudad de México** **22h Argentina, Brasil-Brasilia, Chile** **...**

Agenda

- Encuesta
- Fechas
- Beneficios y Riesgos
- Criterios de calidad
- Clase de dudas extra
- **Dudas**

Dudas


01MBID
roger.clotet@professor.universidadviu.com

Gracias

Universidad Internacional de Valencia